Customize the Genome Browsers produced by G-OnRamp

Yating Liu

07/2017

# 1. Introduction

In addition to running the G-OnRamp workflow with default settings, one of the key features of Galaxy is the ability to modify an existing workflow (e.g., change tool parameters, add or remove tools) using the Workflow Canvas. From this tutorial, you will learn how to:

* Modify tool parameters
* Modify the workflow by adding and removing tools
* Add or remove evidence tracks from the Hub Archive Creator

Note that this tutorial assumes that the reader is already familiar with the basic concepts of Galaxy and of G-OnRamp. It will modify the “**G-OnRamp: D. biarmipes F element**” workflow that we have previously created in the “Introduction to G-OnRamp Walkthrough”.

# 2. Modify G-OnRamp workflow

**Log into your account on the G-OnRamp Galaxy instance at** <http://cloud5.galaxyproject.org/>. (Note that this instance will only be available during the G-OnRamp workshop.) **Click on the “Workflow” menu item in the menu bar to access the list of available workflows. Click on the down** arrow for the “G-OnRamp: D. biarmipes F element” workflow and click on “Copy” to create a copy the G-OnRamp workflow for editing (Figure 1).

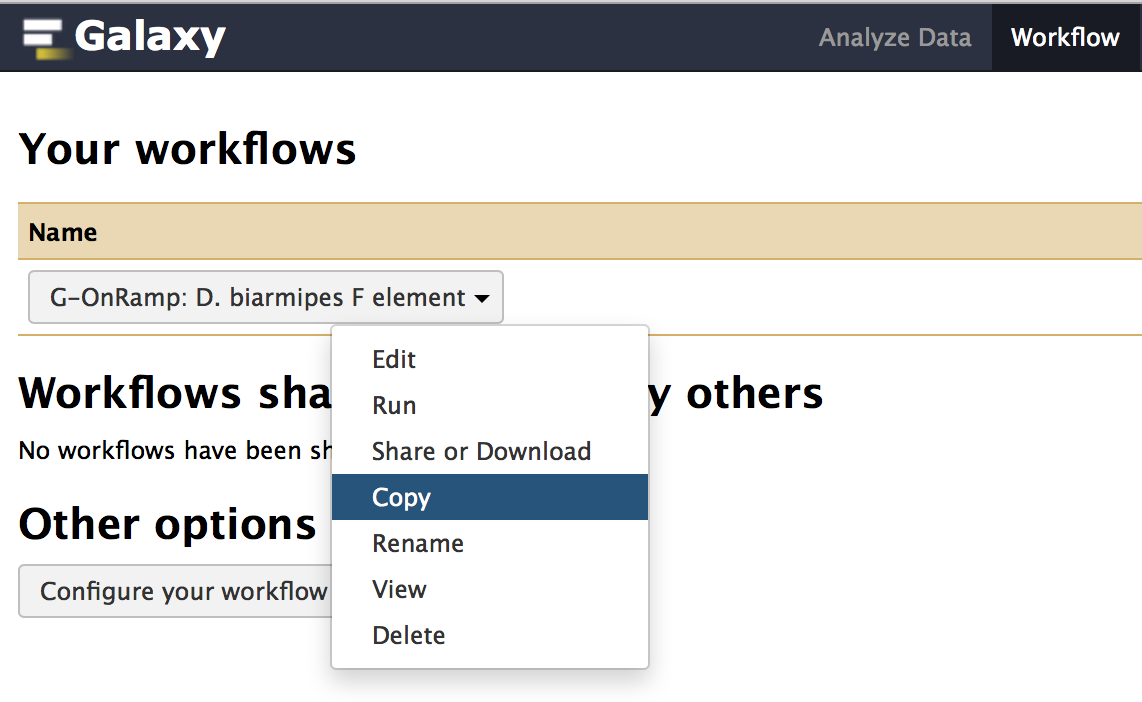


Figure 1: Click on the "Copy" option to create a new copy of the G-OnRamp workflow.

Rename the new copy of the workflow (i.e. Copy of ‘G-OnRamp: D. biarmipes F element’) to “Customized G-OnRamp” using the “Rename” option in the drop-down menu. **Click on the down** arrow for the “Customized G-OnRamp” workflow and click on “Edit” to open the Workflow Canvas (Figure 2).

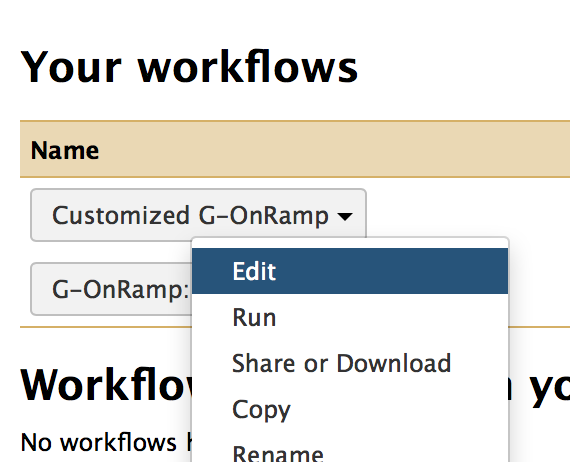


Figure 2: Click on the "Edit" option to open the “Customized G-OnRamp” workflow in the Workflow Canvas.

## 2.1 Big picture

The entire workflow is shown in Figure 3. Each box represents a tool. The “>” symbol on the left side of the box denotes an input dataset for the tool. The “>” symbol on the right side of the box denotes the output dataset produced by the tool. A tool could have multiple input datasets (e.g., HISAT) and output datasets (e.g., Augustus). The “noodles” between the tools correspond to how data are processed by the different tools within the workflow. Each noodle shows how the output dataset from one tool is used as the input dataset for another tool. For example, Figure 4 shows the connection between the “Input dataset” and the Augustus gene predictor. The output from the “Input dataset” tool serves as the “Genome Sequence” input for Augustus.

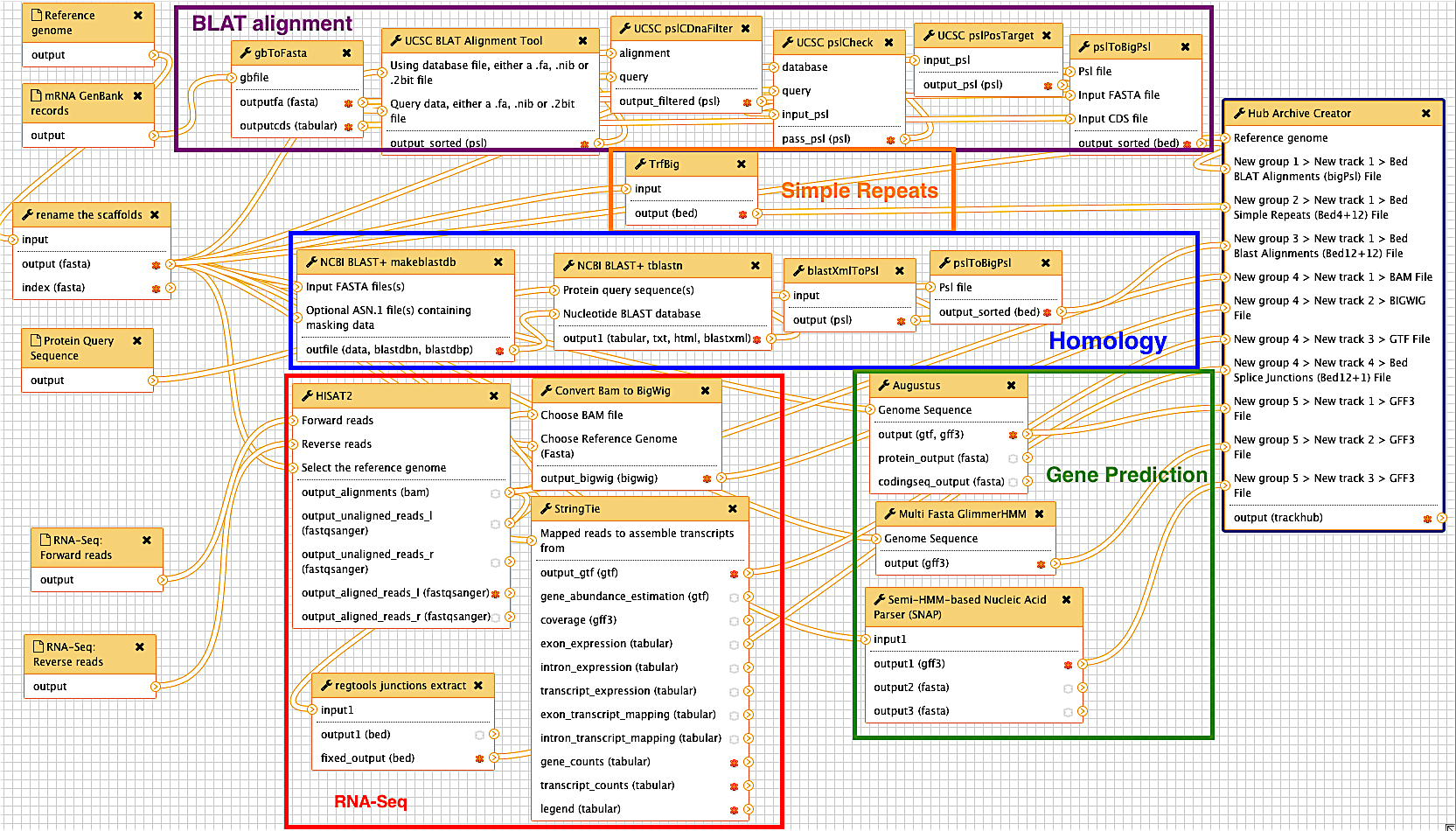


Figure 3: The entire G-OnRamp workflow shown in the Workflow Canvas

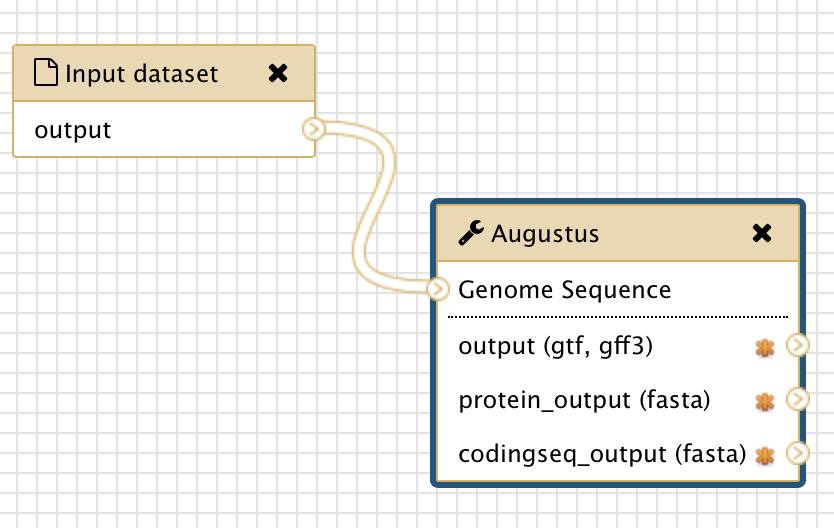


Figure 4: The connection between the “Input dataset” tool and the Augustus tool (red arrow)

## 2.2 Modify tool parameters

You can click on each tool in the Workflow Canvas to learn more about the tool (i.e. what it does and how to use it). You can use the “Details” panel on the right to examine and change the tool parameters (Figure 5). If the analysis workflow was originally derived from a History, then the settings of each tool within the workflow will reflect the parameters used in that analysis. For example, because we used paired-end RNA-Seq data in the “Introduction to G-OnRamp Walkthrough”, the “Individual paired reads” option is selected. If you have unpaired RNA-Seq data, you need to click on the down arrow and select the “Individual unpaired reads” option under the “Single end or paired reads” field. (Note that you do not need to change the parameter here because we will use paired-end RNA-Seq data in this walkthrough.)

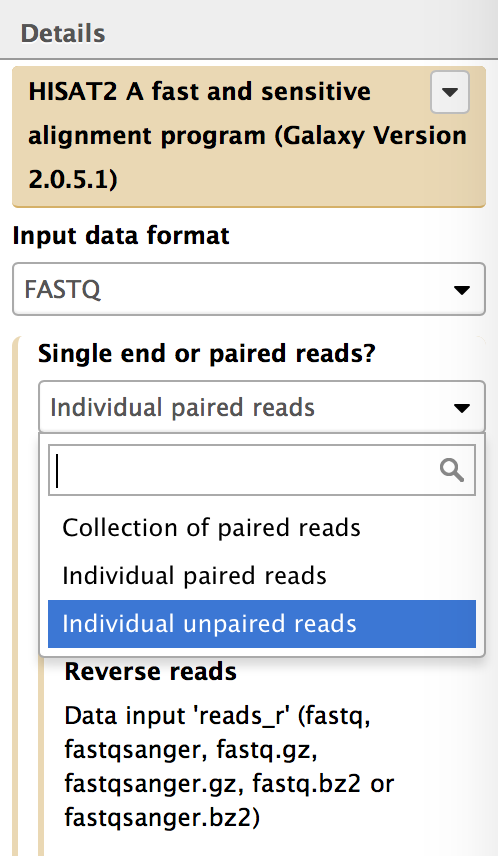


Figure 5: Click on HISAT and use the Details panel to edit the parameter settings for the HISAT.

If there is an arrow on the left side of a parameter, then the parameter could be set when you run the workflow (i.e., set at Runtime). For instance, if you open the Details panel of the Augustus tool, you will see that the “Model Organism” parameter is hidden. This is because we designed the workflow so that you must specify the model organism to use at runtime. If you click on the down arrow on the left side of the “Model Organism” field, a drop-down menu will appear where you can set the “Model Organism” parameter in advance. If you would like to set this parameter at runtime, click on the up arrow to hide the field (Figure 6).

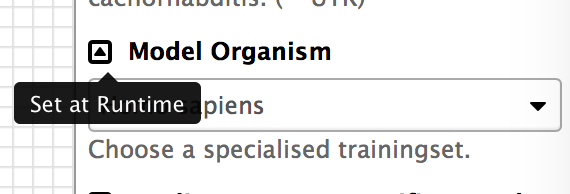


Figure 6: Click on the up arrow icon on the left side of the "Model Organism" field to set this parameter at runtime (red arrow).

## 2.3 Add an evidence track to the Hub Archive Creator

You can also use the Workflow Canvas to add or delete a tool. For example, if you want to add the results from WindowMasker (which identifies simple repeats and low complexity sequences) to the G-OnRamp workflow, you can use the search field in the “Tools” panel to search for “WindowMasker”. The tools WindowMasker\_ustat and WindowMasker\_mkcounts will appear in the Tools panel (Figure 7). These tools correspond to the two stages (ustat and mk\_counts, respectively) used by WindowMasker to identify repeats in DNA sequences. Click on both links to add the WindowMasker\_ustat and WindowMasker\_mkcounts tools to the Workflow Canvas (Figure 8).



Figure 7: Use the search field in the Tools panel to search for the WindowMasker tools.

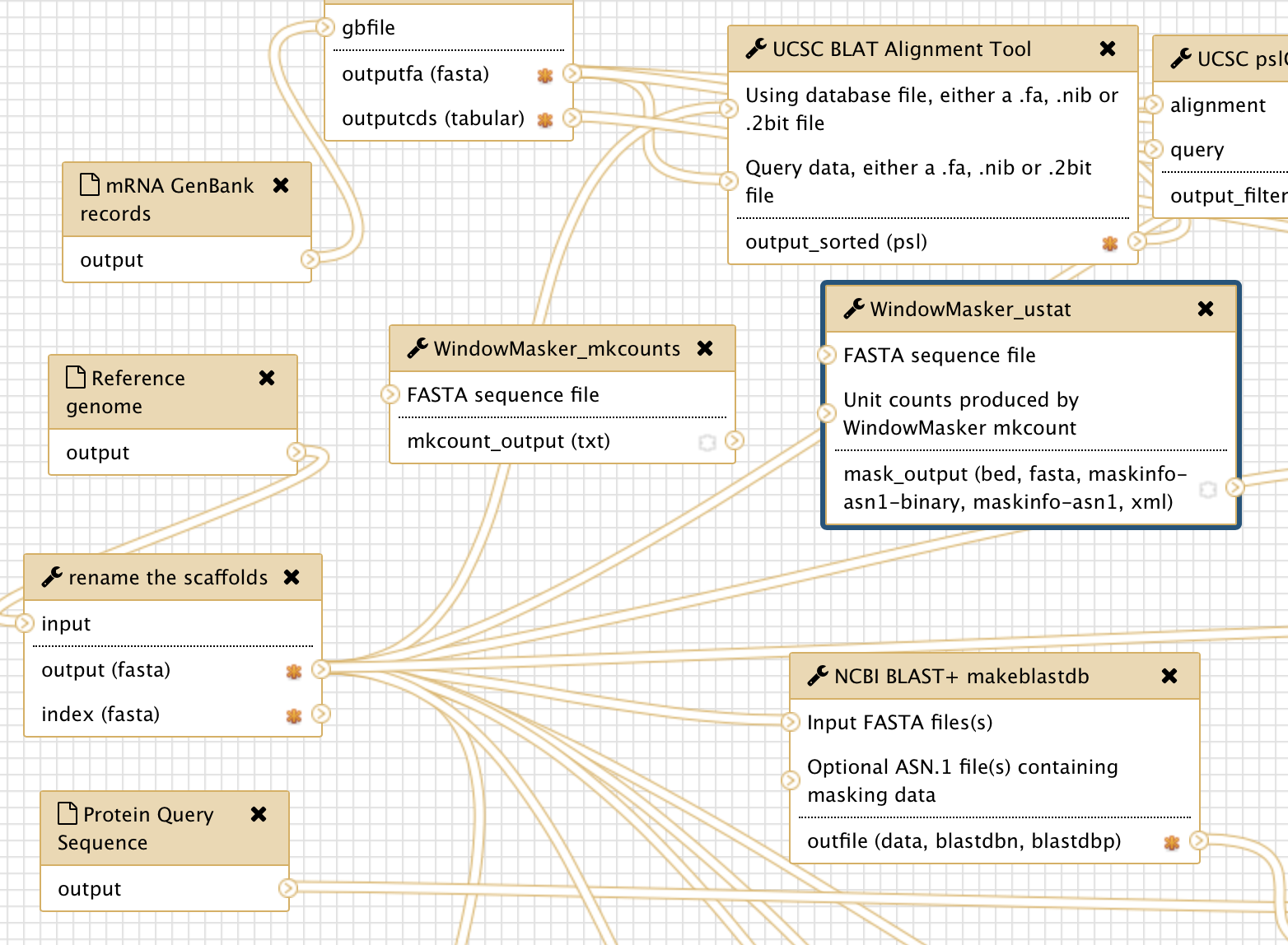


Figure 8: WindowMasker\_ustat and WindowMasker\_mkcounts are added to the Workflow Canvas (red arrows).

The next step would be to incorporate the WindowMasker\_ustat and WindowMasker\_mkcounts tools with the rest of the G-OnRamp workflow. This is accomplished by specifying the input and output datasets for these WindowMasker tools.

The first stage of the WindowMasker analysis is mk\_counts (WindowMasker\_mkcounts), which constructs a unit counts table for a genome assembly. The unit counts correspond to the frequency of short sequences with length k (k-mers) in the genome assembly. There is a “>” symbol on the left side of the “FASTA sequence file” field in WindowMasker\_mkcounts, which indicates that it requires the genome sequences in FASTA format. Here we will use the genome sequences that have been renamed by “rename the scaffolds” tool, which shortens the scaffold names to less than 32 characters. (Older versions of the UCSC Genome Browser impose a 31-character limit on the scaffold names.)

To establish a new connection between the “rename the scaffolds” and the WindowMasker\_mkcounts tools, click on the “>” symbol on the right side of “output (fasta)” in the “rename the scaffolds” tool and drag it to the “>” symbol on the left side of “FASTA sequence file” in the WindowMasker\_mkcounts tool. As you drag the connection, the connection will appear as a green “noodle” (Figure 9).

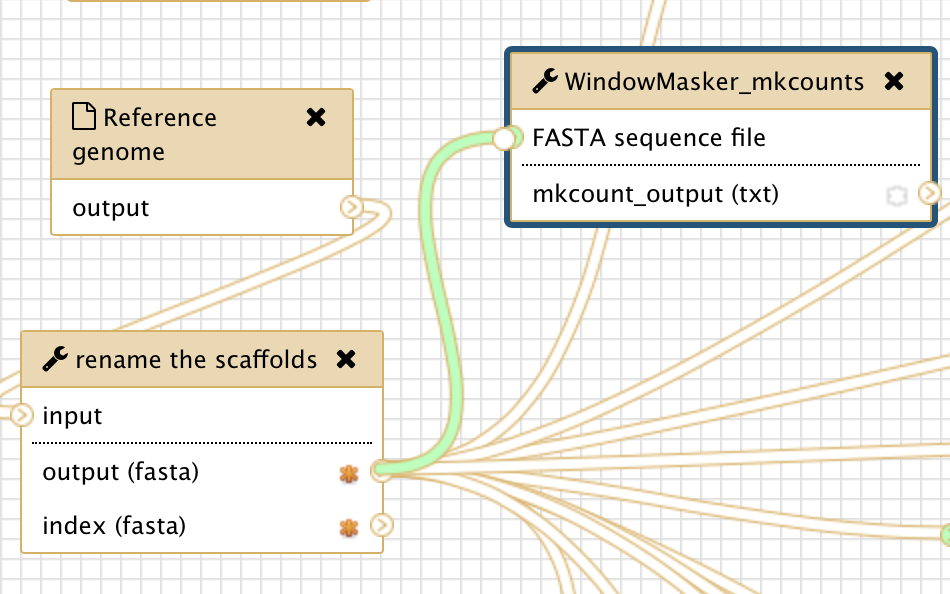


Figure 9: Connect the output from the “rename the scaffolds” tool to the input for the WindowMasker\_mkcounts tool.

When you release the mouse next to the “Genome Sequence” field on the left side of the WindowMasker\_mkcounts box, it will establish the connection between these two tools.

The second stage of the WindowMasker analysis is ustat (WindowMasker\_ustat), which uses the unit counts produced by WindowMasker\_mkcounts to mask repetitive regions within a DNA sequence. The WindowMasker\_ustat tool requires a FASTA sequence file as an input dataset. Repeat the steps above to create a connection between “output (fasta)” of the “rename the scaffolds” tool with the “FASTA sequence file” input for the WindowMasker\_ustat tool (red arrow in Figure 10).

WindowMasker\_ustat also needs the unit counts table produced by WindowMasker\_mkcounts as an input. Connect the “mkcount\_output” in WindowMasker\_mkcounts with the “Unit counts produced by WindowMasker mkcount” in WindowMasker\_ustat (blue arrow in Figure 10).

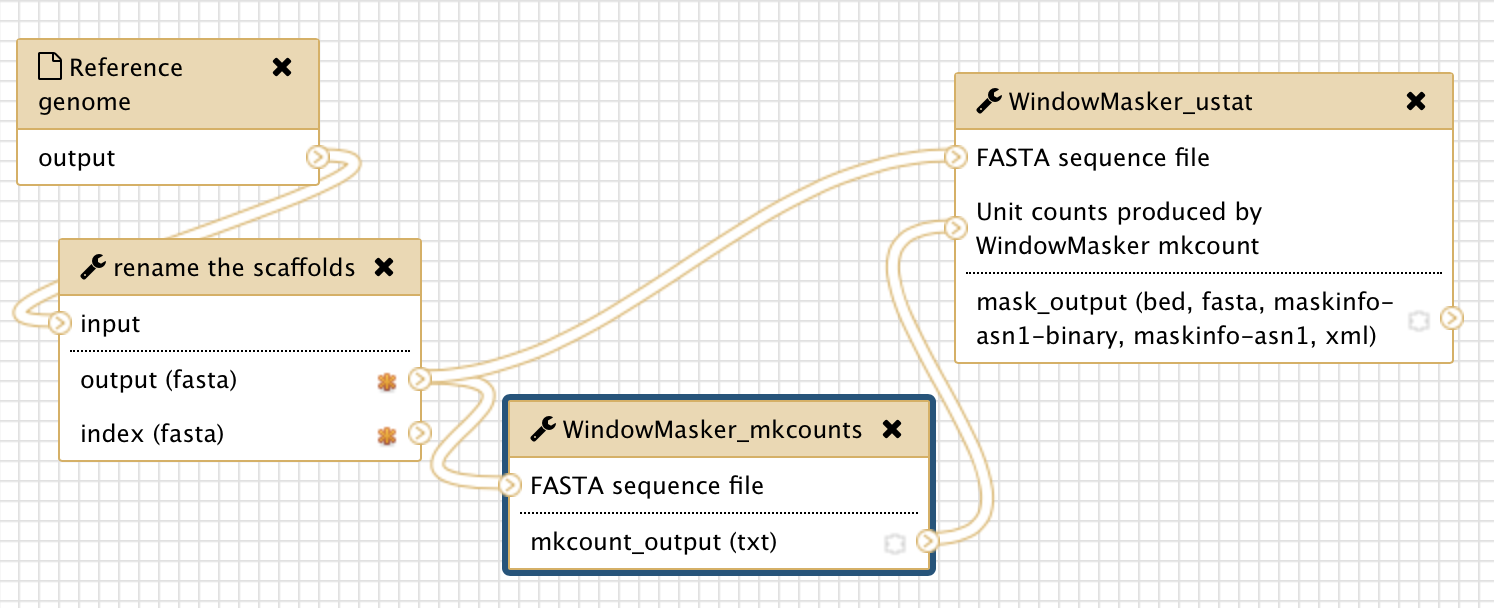


Figure 10: Connect the “output (fasta)” from the “rename the scaffolds” tool to the “FASTA sequence file” inputs for the WindowMasker\_mkcounts and WindowMasker\_ustat tools (red arrow). In addition, connect the “mkcount\_output (txt)” from the WindowMasker\_mkcounts tool to the “Unit counts produced by WindowMasker mkcount” input for the WindowMasker\_ustat tool (blue arrow).

Using the same approach, you can connect the output of WindowMasker to the Hub Archive Creator. However, because all the input connections to the Hub Archive Creator are already connected to the output connections from the other tools, we need to add another input connection to the Hub Archive Creator before we can create the connection with the output from the WindowMasker\_ustat tool.

Click on the Hub Archive Creator box in the Workflow Canvas and then examine the Details panel on the right. Scroll down to near the bottom of the Details panel and click on the “Insert New group” button to create a new group (Figure 11). Enter “Variation and Repeats” in the “Group name” field and then click on “Insert New track” to create the WindowMasker track (Figure 12).

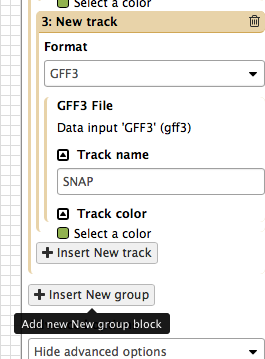


Figure 11: Click on the "Insert New group" button in the Details panel to add a new group in the Hub Archive Creator.

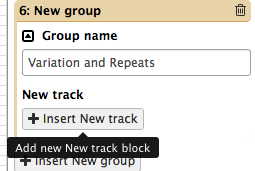


Figure 12: Change the name of the new group to "Variation and repeats", and then click on the "Insert New track" button to create a new input connection for the Hub Archive Creator.

In order to establish a connection between two tools, the datatype of the output dataset from the first tool must be the same as the datatype of the input dataset for the second tool. You can see the output format of WindowMasker by clicking on WindowMasker\_ustat and open the Details panel (Figure 13). In this case, the output format for WindowMasker\_ustat is BED.

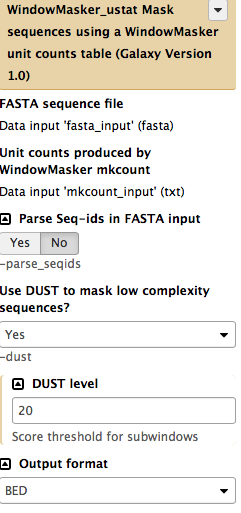


Figure 13: Open the Details panel of the WindowMasker\_ustat tool and verify that the output format is set to BED.

Consequently, we need to select the “BED” format for the new input connection in the Hub Archive Creator. (Figure 13, red arrow) You can specify the name of the track as “WindowMasker” and select a color (*e.g.*, light blue) for that track (Figure 14).

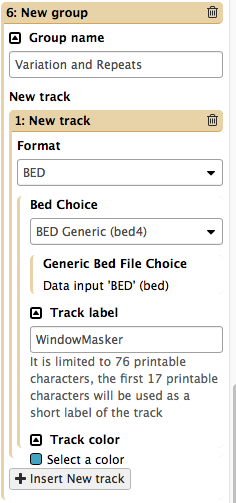


Figure 14: Select the BED format in the drop-down menu, then change the “Track label” to “WindowMasker” and the “Track color” to “light blue”.

A “New group 6 > New track 1 > Generic Bed File Choice” entry will appear in the Hub Archive Creator tool (Figure 15).

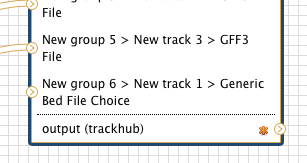


Figure 15: The Hub Archive Creator with a new BED input connection (red arrow)

After you have created the new BED input connection, you can connect the output of the WindowMasker\_ustat tool to the input of the Hub Archive Creator. Click on the “>” symbol next to the “mask\_output” field in the WindowMasker\_ustat tool, drag it to the “>” symbol next to the “New group 6 > New track 1 > Generic Bed File Choice” field in the Hub Archive Creator tool and then release the mouse (Figure 16).

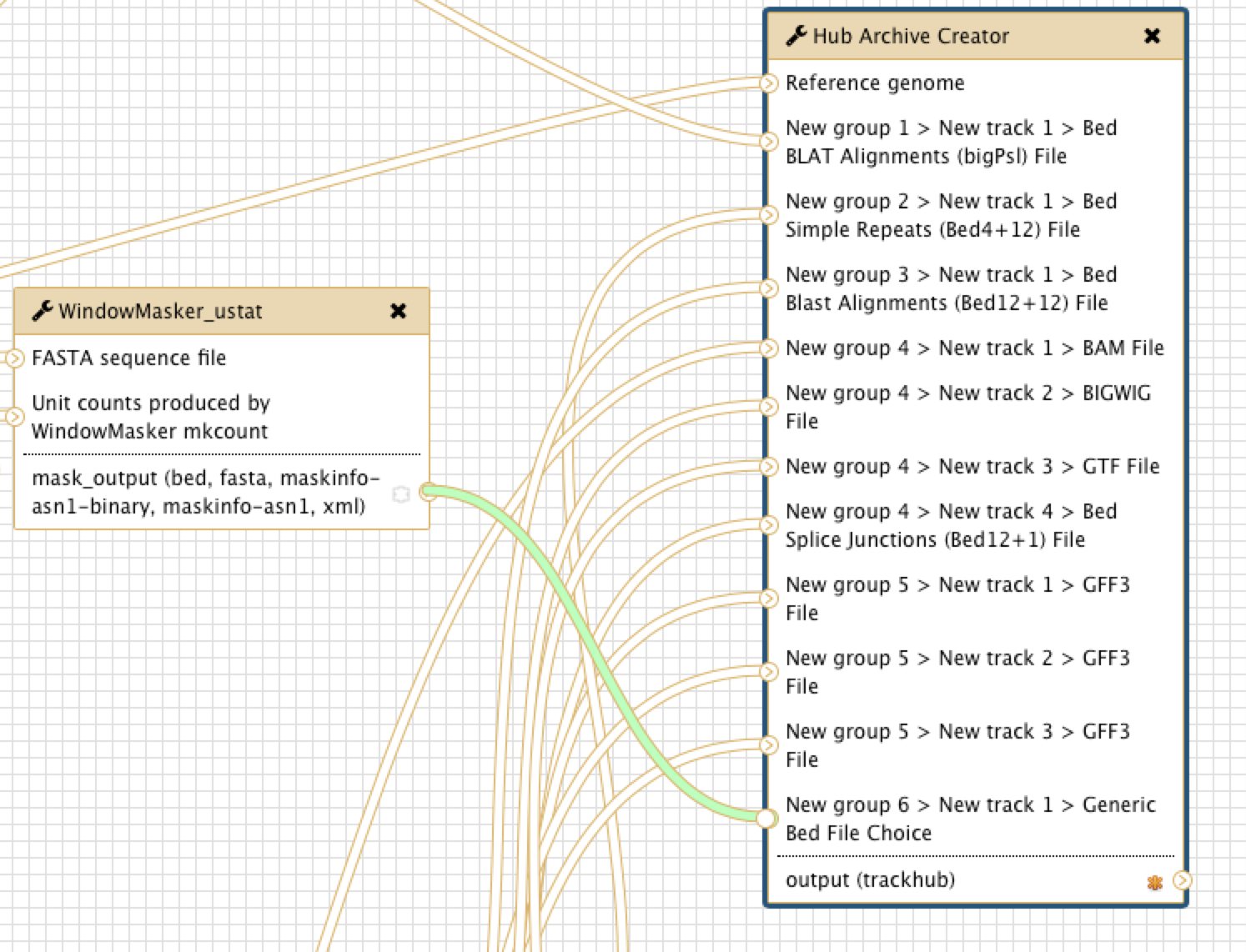


Figure 16: Connect the output of WindowMasker\_ustat to the Hub Archive Creator

## 2.4 Show or hide an output dataset

To simplify the display in the History panel, the output of each tool is hidden from your History by default. To show a dataset in the History, you can mark the dataset as a workflow output by clicking on the “\*” symbol. All unmarked datasets will be hidden from your History. For example, StringTie will produce eight output files. However, because only the “output\_gtf (gtf)” output is marked (“\*” symbol in dark orange) in the workflow, only the GTF file will appear in your History after you run the workflow (Figure 17). The other seven output datasets will be hidden from the History. This feature is particularly useful when you are working with large workflows that produce many temporary datasets.

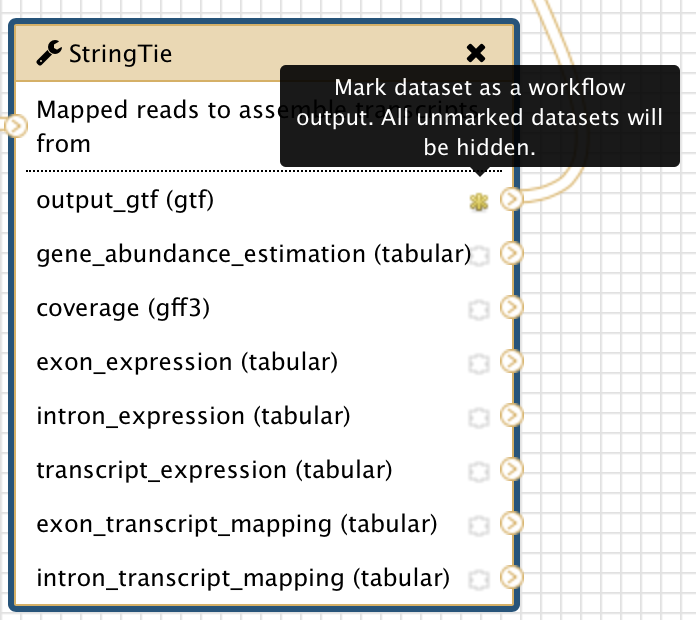


Figure 17: The yellow star next to “output\_gtf (gtf)” output for the StringTie tool indicates that this dataset has been marked as a workflow output

You can show the outputs from WindowMasker\_mkcounts and WindowMasker\_ustat tools in the History by clicking on the “\*” symbol next to “mkcount\_output(txt)” and “mask\_output” (Figure 18).

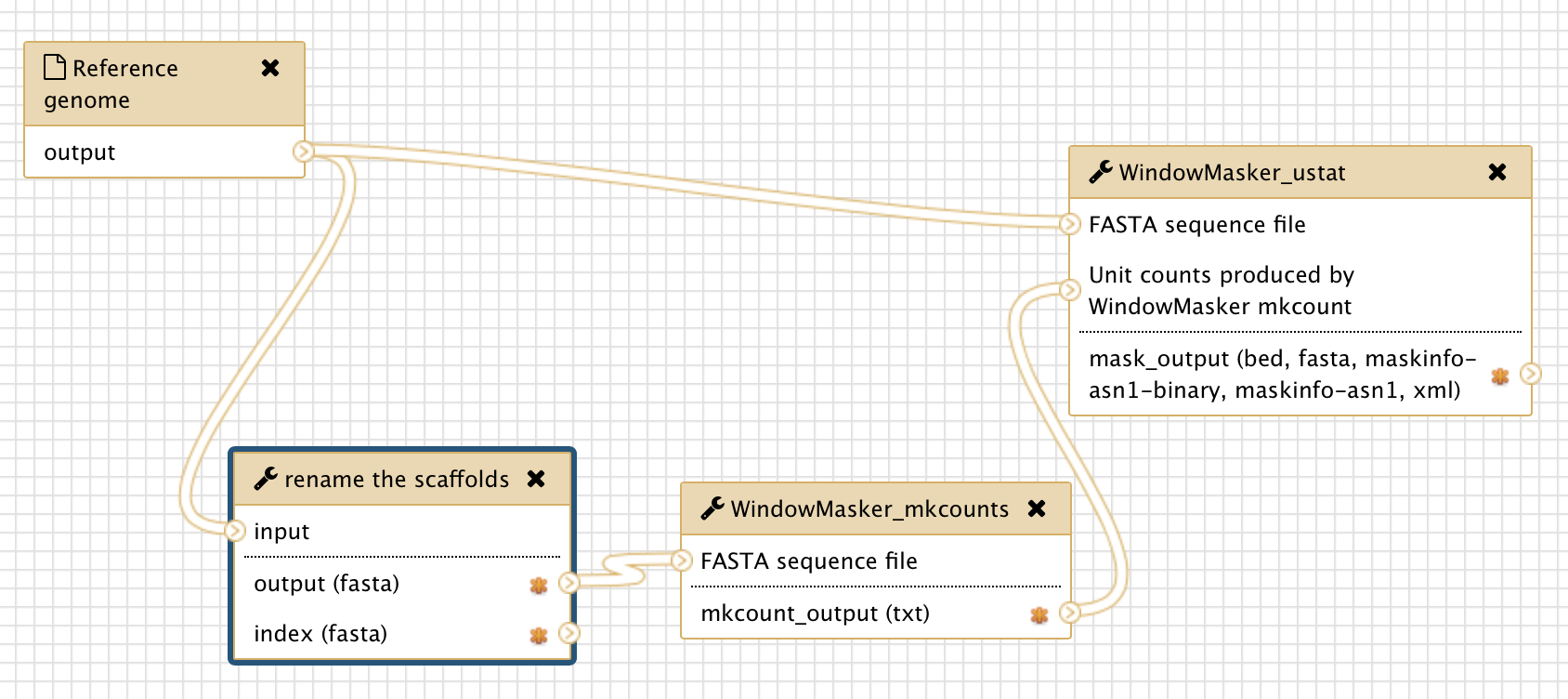


Figure 18: Show the outputs from WindowMasker tools in the History by clicking on the "\*" symbols (red arrows) next to the output files.

## 2.6 Save the changes to the workflow

Remember to save your changes before you leave the Workflow Canvas page. Click on the settings icon at the top right corner of Workflow Canvas and then click on “Save” (Figure 19).

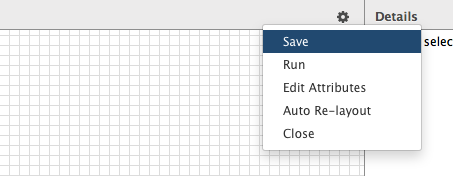


Figure 19: Save the changes that you have made to the workflow

# 3. Exercise: edit the workflow to use the masking data from WindowMasker for the TBLASTN search

## 3.1 Make a copy of the current workflow

Make a copy of “Customized G-OnRamp” workflow and rename the new workflow as “G-OnRamp: WindowMasker” (Figure 20). Click on the down arrow next on the new workflow and then select the Edit option to go to the Workflow Canvas.

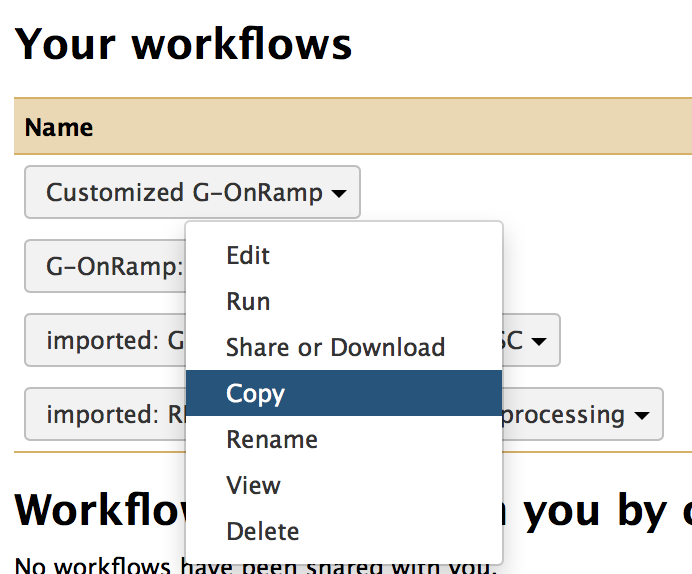


Figure 20: Copy the current workflow and rename it as “G-OnRamp: WindowMasker”.

## 3.2 Edit the “G-OnRamp: WindowMasker” workflow

Below are some hints on how to construct this workflow:

1. You need to add another WindowMasker\_ustat tool and then incorporate it into the workflow.
2. You need to modify the parameter settings for the WindowMasker\_ustat tool. Select the “maskinfo ASN.1 text” option as the Output format for this tool (Figure 21).

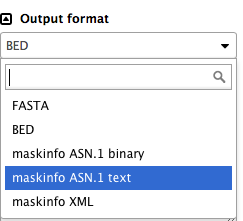


Figure 21: Choose the “maskinfo ASN.1 text” option under the “Output format” field

1. You need to use the maskinfo output from WindowMasker\_ustat tool as the input to the NCBI BLAST + makeblastdb tool.
2. Don’t forget to connect the “output (fasta)” from the “rename the scaffolds” tool to the “FASTA sequence file” inputs for the WindowMasker\_ustat tool (Figure 22).
3. (Optional) You can also add a WindowMasker track to the Hub Archive Creator (see section 2.3 of this walkthrough for details).

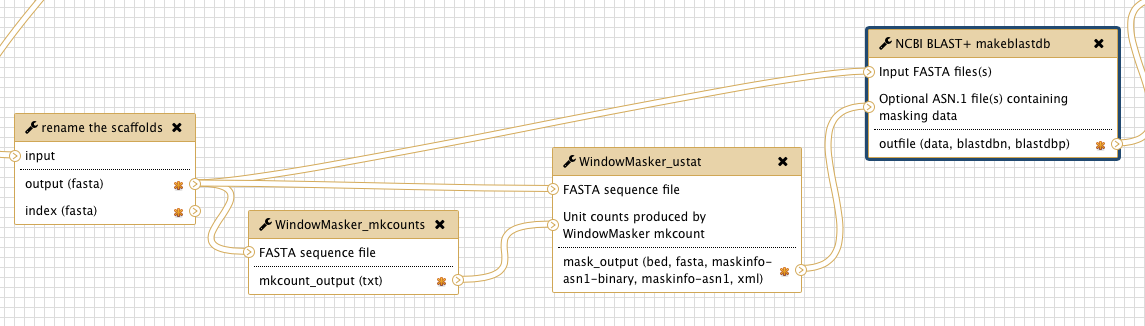


Figure : Use the “mkcount\_output” from “WindowMasker\_mkcounts” and the “output(fasta)” from “rename the scaffolds” as inputs for WindowMasker\_ustat. Then connect “mask\_output” (in “maskinfo ASN.1 text” format) to the “Optional ASN.1 file(s)” input for the “NCBI BLAST+ makeblastdb” tool.

Don’t forget to save the changes before you leave the Workflow Canvas.

## 3.3 Upload your datasets and run the workflow

Create a new History and import the datasets in the “Intro\_G-OnRamp” folder in the Data Libraries. Run the “G-OnRamp: WindowMasker” workflow. Make sure to select the correct input datasets (Figure 23). Don’t forget to change the model organisms for the gene prediction tools (Figure 24) and change the name of your new Assembly Hub.

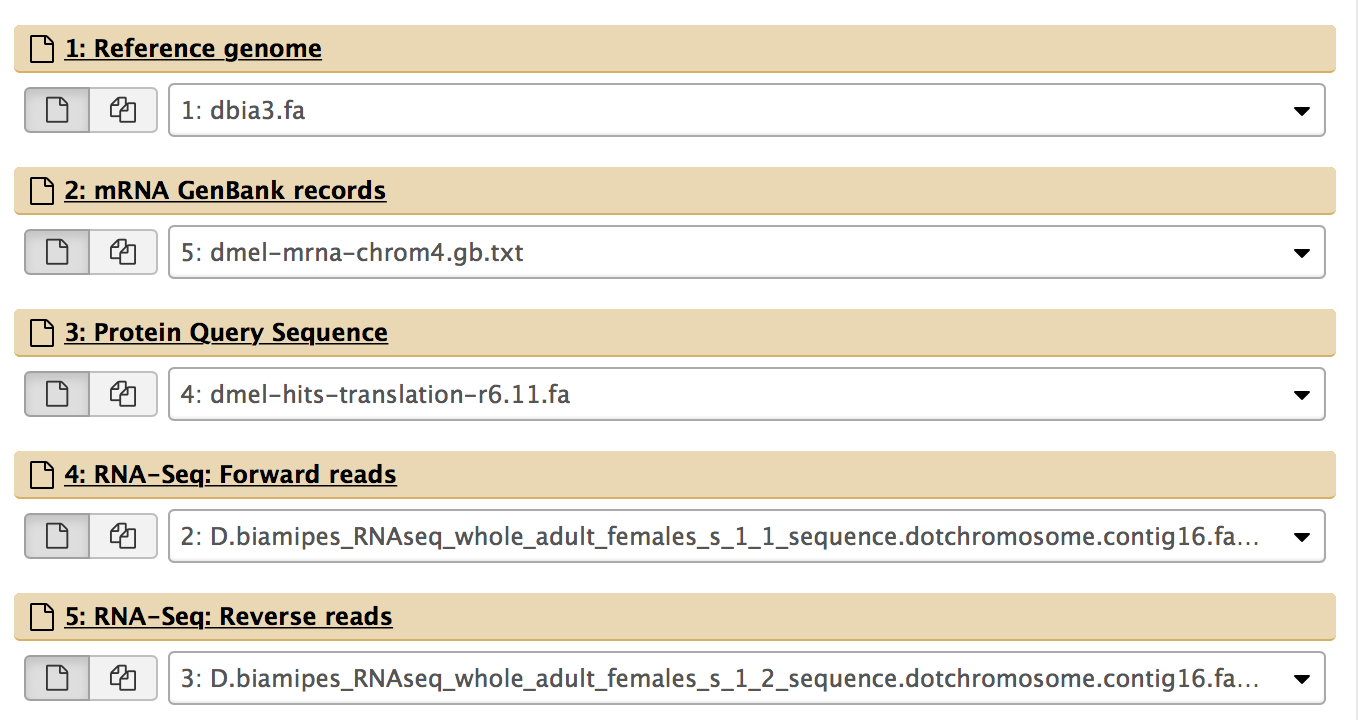


Figure : Make sure the input datasets that you have specified are the same as those shown in this Figure.

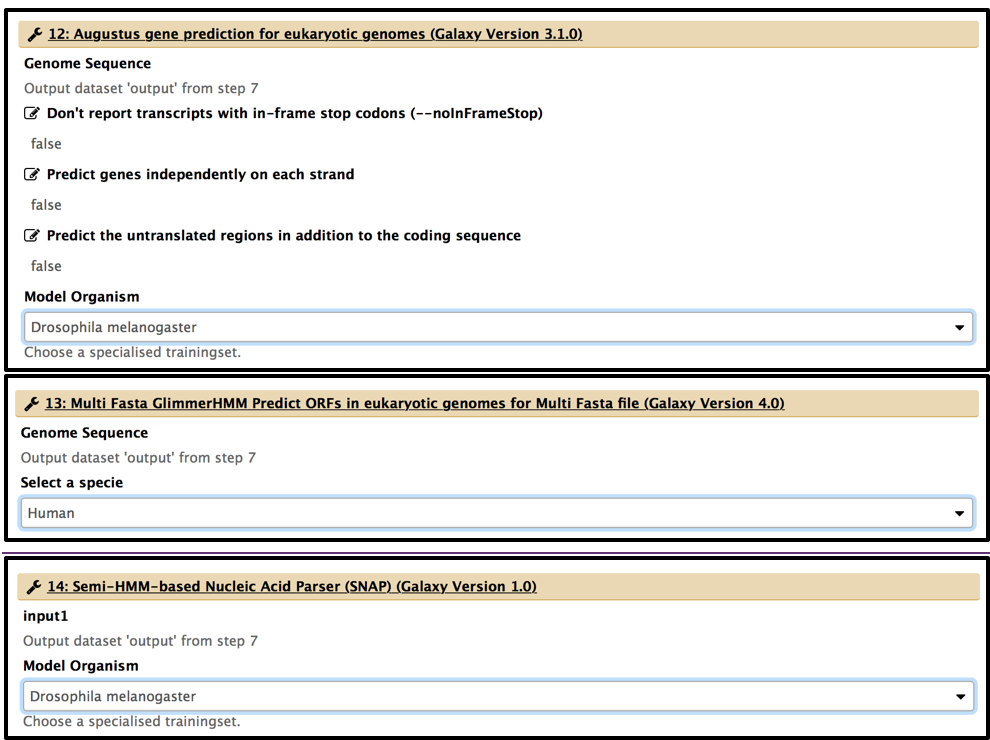


Figure 24: Change the model organisms for the gene prediction tools (i.e. Augustus, GlimmerHMM, and SNAP).

Run the workflow. Finally, you can view the UCSC Genome Browser produced by the modified G-OnRamp workflow as different evidence tracks on the UCSC Genome Browser (Figure 25).



Figure 25: After the “G-OnRamp: WindowMasker” workflow creates the Assembly Hub, you can view the Dbia3 genome assembly with the WindowMasker evidence track for scaffold\_16 on the UCSC Genome Browser.

## 3.4 Remove an evidence track from the Hub Archive Creator

Make a copy of “G-OnRamp: WindowMasker” workflow and rename the new workflow as “G-OnRamp: Remove WindowMasker track”. Click on the down arrow next on the new workflow and then select the Edit option to go to the Workflow Canvas.

To delete a tool from the workflow, click on the “x” at the top right corner of that tool. The tool and its connections will be removed from the workflow. For example, click on the “x” at the top right corner of the two WindowMasker tools to remove them from the workflow (Figure 26).

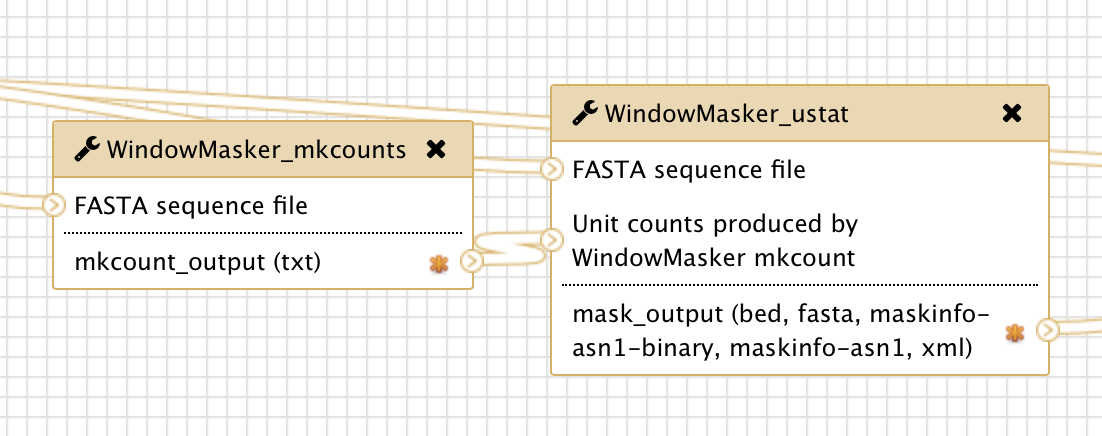


Figure 26: Click on the “x” at the top right corner to delete the WindowMasker\_mkcounts and the WindowMasker\_ustat tools from the workflow.

When you delete the WindowMasker tools, the connections between WindowMasker\_ustat and the Hub Archive Creator will also be removed. However, the “New group 6 > New track 1 > Generic Bed File Choice” entry in Hub Archive Creator will need to be removed manually.

Click on the Hub Archive Creator tool and scroll down to the last “Variation and Repeats” group in the Details panel. Because there is only one entry in that group, we can delete the WindowMasker entry by removing the whole group. Click on the “Trash” icon at the top right corner of the “Variation and Repeats” group to delete the group (Figure 27). Remember to save the changes to the workflow before you leave the Workflow Canvas.

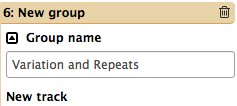


Figure 27: Click on the trash icon to remove the “Variation and Repeats” group from the Hub Archive Creator.

# 4. Exercise: use G-OnRamp to create the JBrowse

Import the “G-OnRamp workflow for JBrowse” from Shared Data. Create a new History and import the datasets from the “Intro\_G-OnRamp” folder in the Data Libraries. Run the JBrowse workflow for the test datasets.

After all the steps in the G-OnRamp workflow have completed (which will take a few minutes), you can view the JBrowse for the *D. biarmipes* F element by clicking on the eye icon to open a html file (Figure 28), and then click on the “View JBrowse Hub” link (Figure 29).



Figure 28: Click on the eye icon for the dataset generated by the JBrowse Archive Creator to access the web page with a link to the JBrowse Hub.

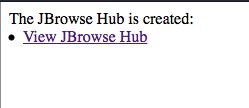


Figure : Click on the “View JBrowse Hub” link to view the JBrowse for the *D. biarmipes* F element.

Because the RNA-Seq dataset used in this walkthrough only contains the RNA-Seq reads that mapped to scaffold\_16, we will examine this scaffold on JBrowse. Select “scaffold\_16” from the drop-down menu on the top menu bar (Figure 30).

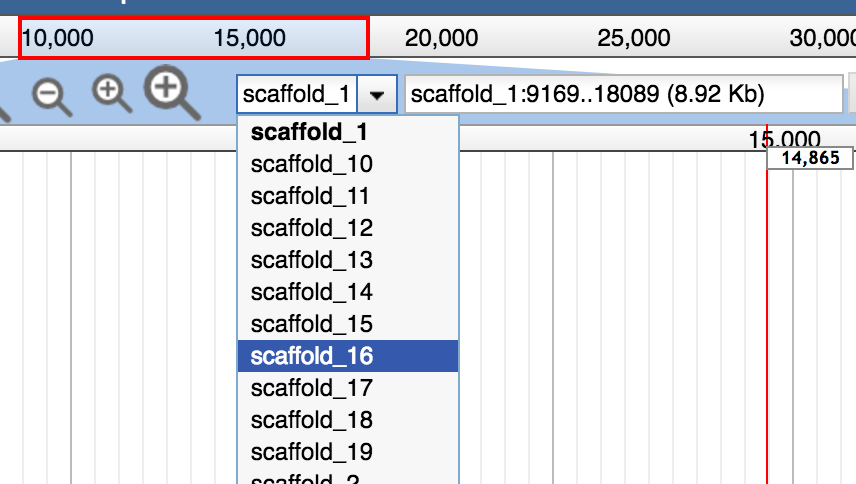


Figure : Select scaffold\_16 from the drop-down menu.

The “Available Tracks” panel on the left side of the window shows the list of available evidence tracks. Select the checkbox next to the name of the evidence track to show the evidence track in JBrowse (Figure 31).

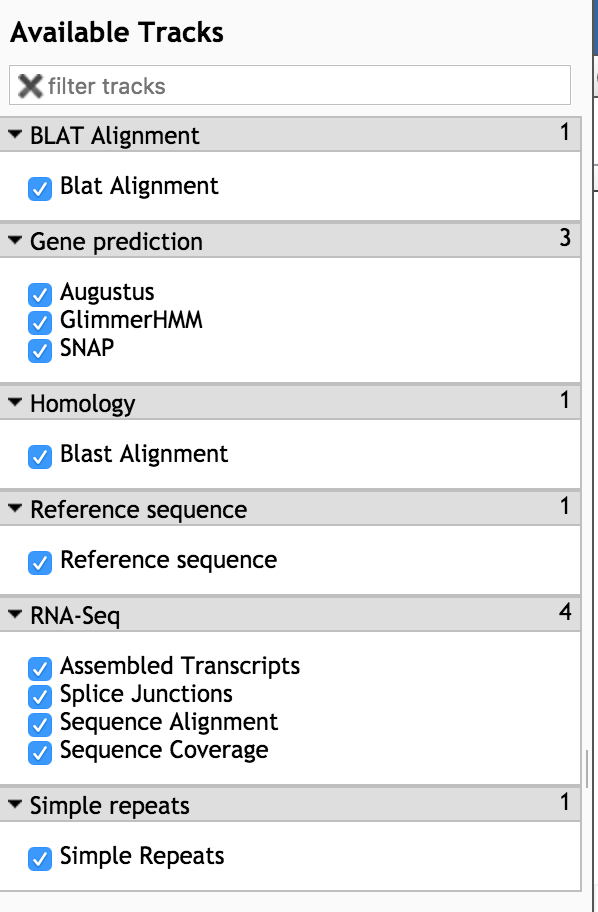


Figure : Use the checkboxes in the “Available Tracks” panel to control which evidence tracks to show. (Based on the settings shown in this screenshot, all the evidence tracks produced by the JBrowse Archive Creator will be shown.)

Finally, you can then see the results of the G-OnRamp workflow as different evidence tracks on JBrowse. You can view the pairwise alignments in the “Blat Alignment” and “Blast Alignment” tracks if you right click on the feature, and then select the “View alignment” option from the drop-down menu (Figure 32).

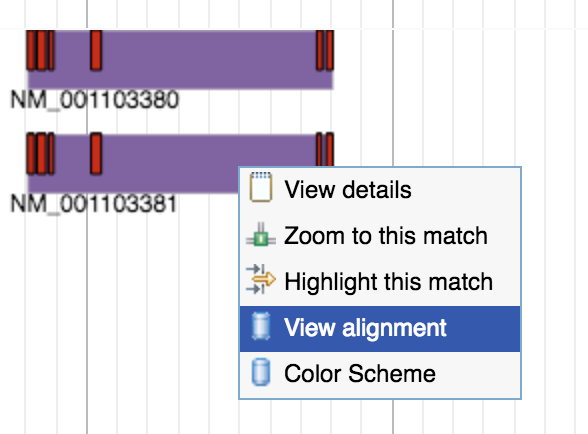


Figure : Right click on a feature in the “Blat Alignment” or “Blast Alignment” tracks. Then select the "View alignment" option to see the pairwise alignments.

You can view the translated protein sequence of a gene prediction (i.e., from Augustus, GlimmerHMM, and SNAP) if you right click on the feature, and then select the “View translated protein” option from the drop-down menu (Figure 33).

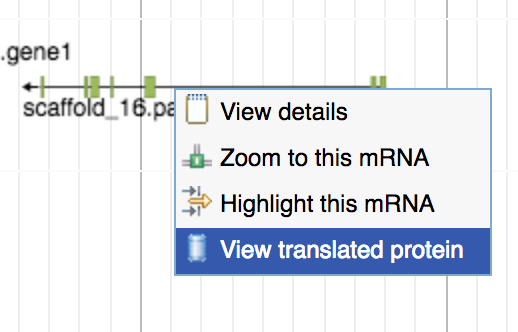


Figure : Use the “View translated protein” option to view the translated protein sequence of a feature in the gene prediction tracks.