Customize the Genome Browsers produced by G-OnRamp

Yating Liu

06/2017

# 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).

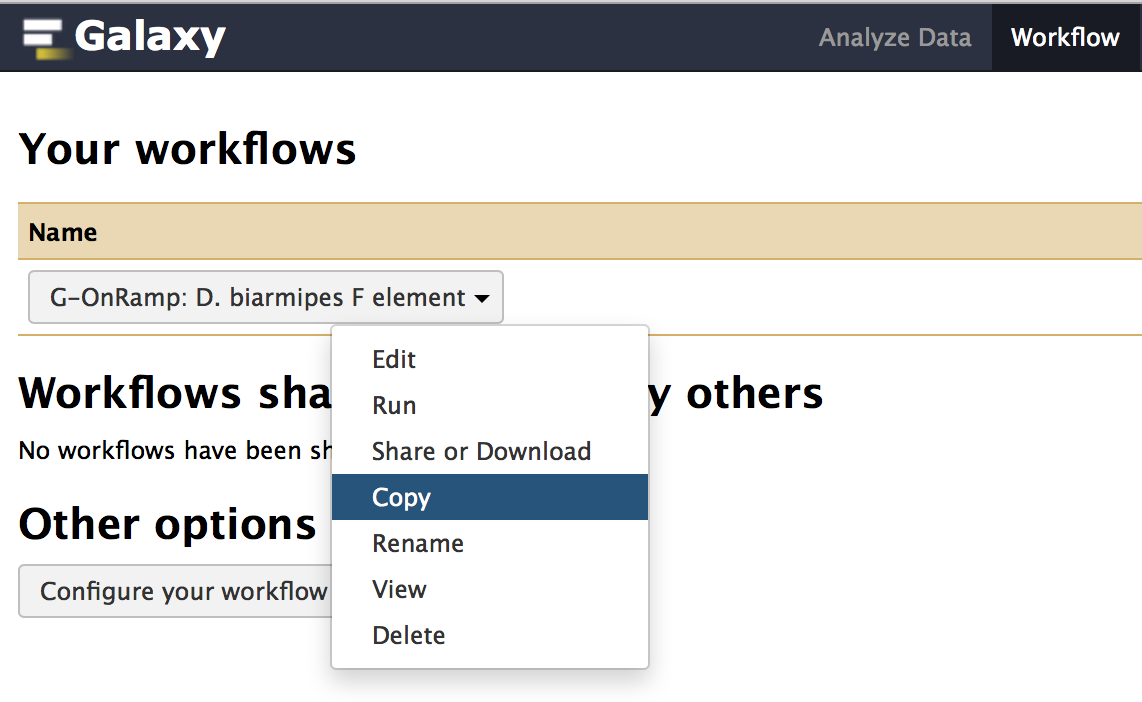


Figure : Click on the "Copy" option to create a new copy of the G-OnRamp workflow.

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).

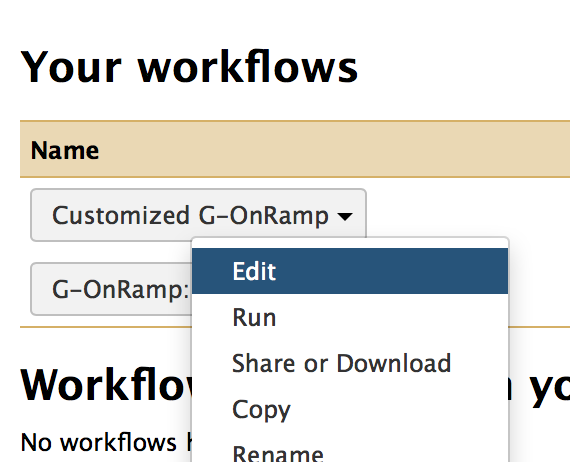


Figure : Click on the "Edit" option to open the “Customized G-OnRamp” workflow in the Workflow Canvas.

## 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.

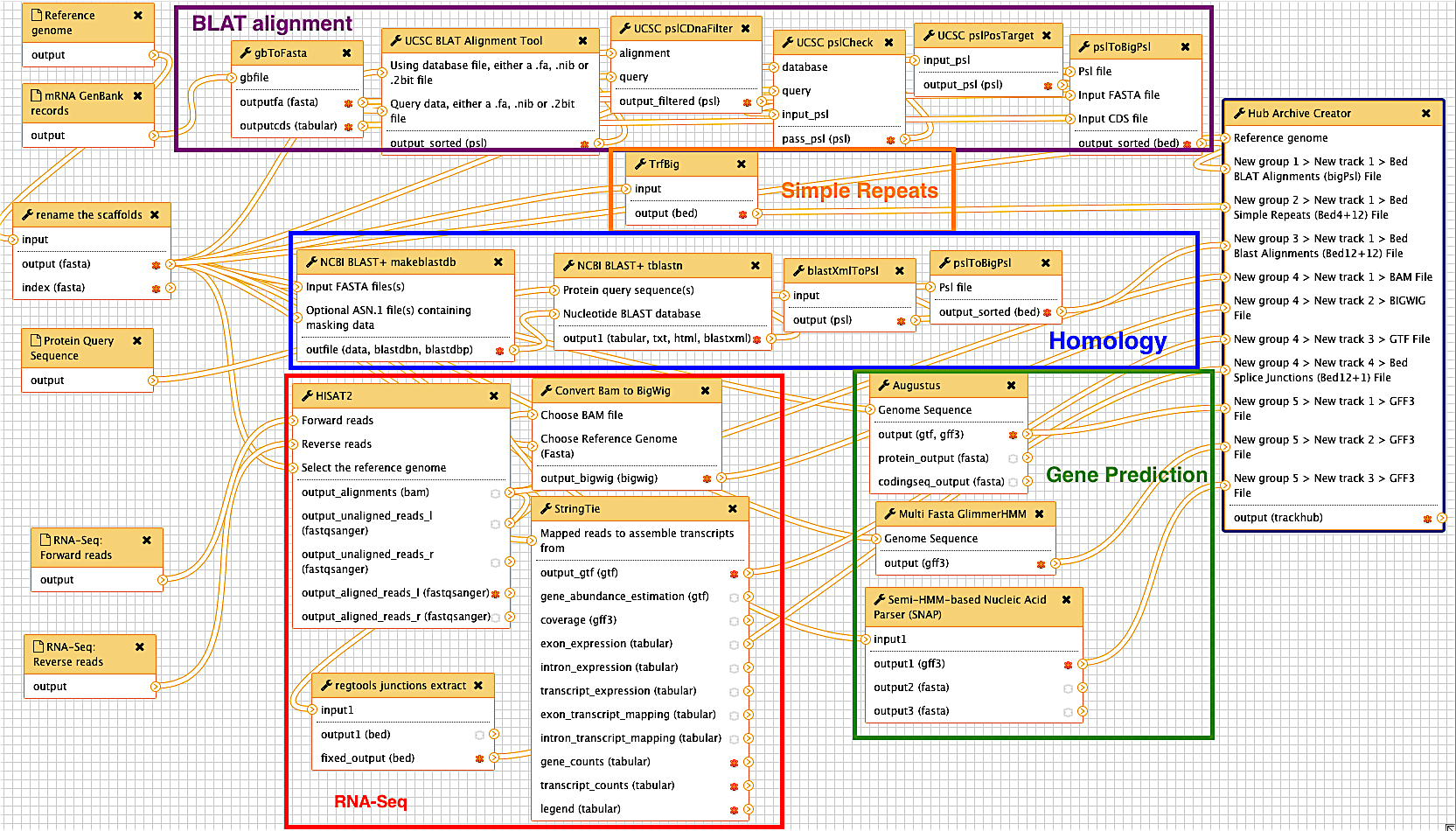


Figure : The entire G-OnRamp workflow shown in the Workflow Canvas

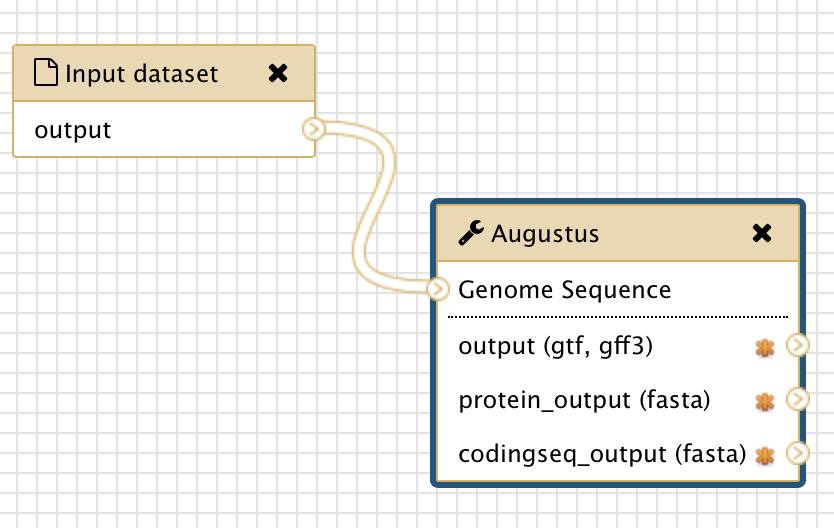


Figure : The connection between the “Input dataset” tool and the Augustus tool (red arrow)

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

If there is an arrow on the left of a parameter, the parameter is allowed to be set at the runtime. For instance, if you open the Details panel of Augustus tool, you will see the setting for “Model Organism” is hidden. It is because we choose to set the parameter at the runtime. If you click on the down arrow on the left of “Model Organism” field, you can see an input box shows up and you can set the parameter in advance by selecting a model organism from the drop-down menu. If you want to set it at the runtime, click on the up arrow to hide the field (Figure 6). When you set it at the runtime, you need to specify the model organism when you run the workflow. You can use a different model organism each time you run the workflow with no need for changing the parameter in the Workflow Canvas.

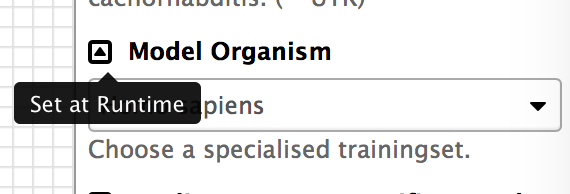


Figure 6: Click on the up arrow on the left of "Model Organism" to set this parameter at the runtime.

## 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 Tools panel (Figure 7). Click on both links to add the WindowMasker\_mkcounts and WindowMasker\_ustat tools to the Workflow Canvas (Figure 8).



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

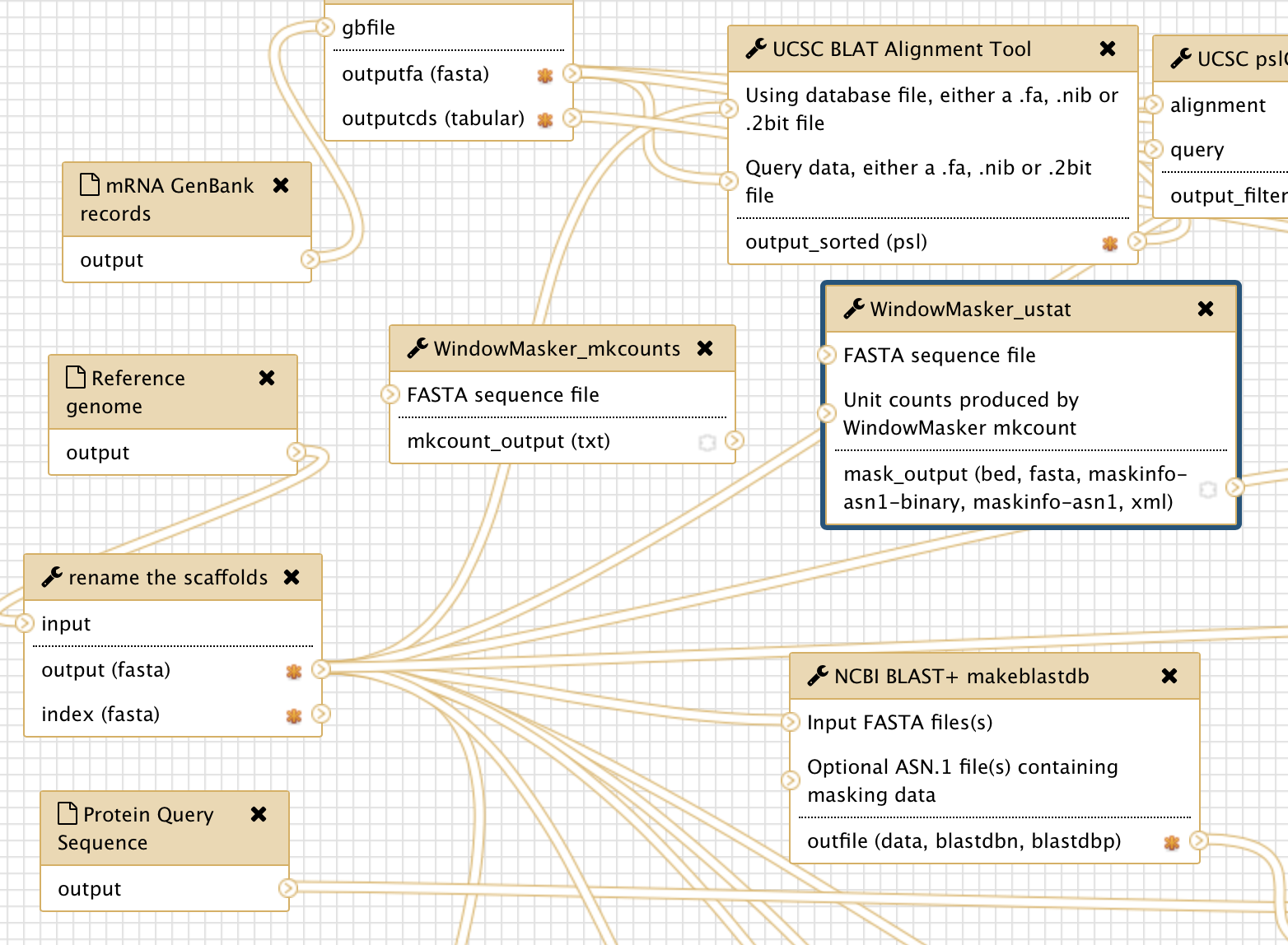


Figure : WindowMasker\_ustat and WindowMasker\_mkcounts are added to the Workflow Canvas (red arrows).

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.

WindowMasker\_mkcounts is used to construct WindowMasker unit counts table for WindowMasker\_ustat. There is a “>” symbol on the left side of “FASTA sequence file” 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).

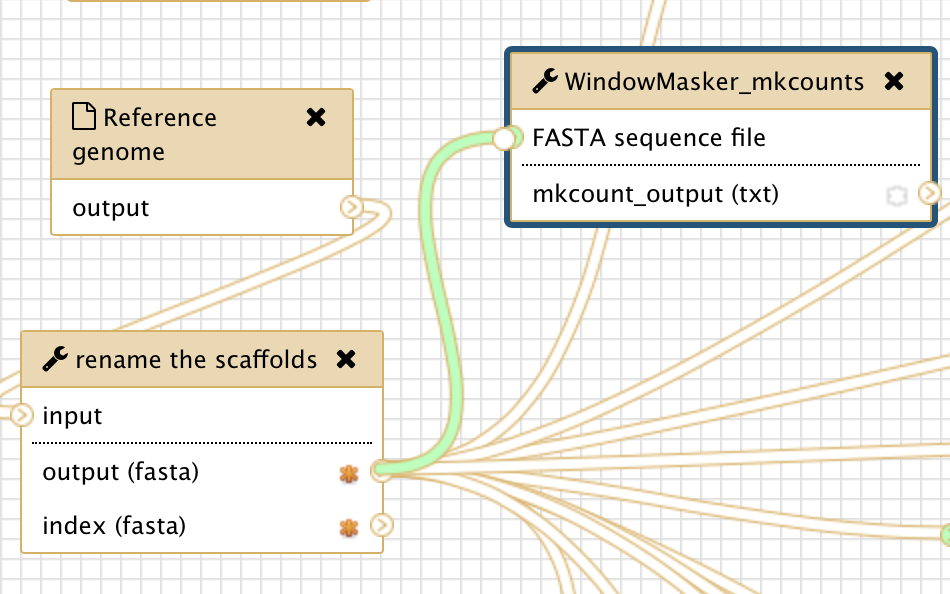


Figure : Connect the output from the “rename the scaffolds” tool to the input for the WindowMasker\_mkcounts tool.

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. WindowMasker\_ustat also needs an input of FASTA sequence file, 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).

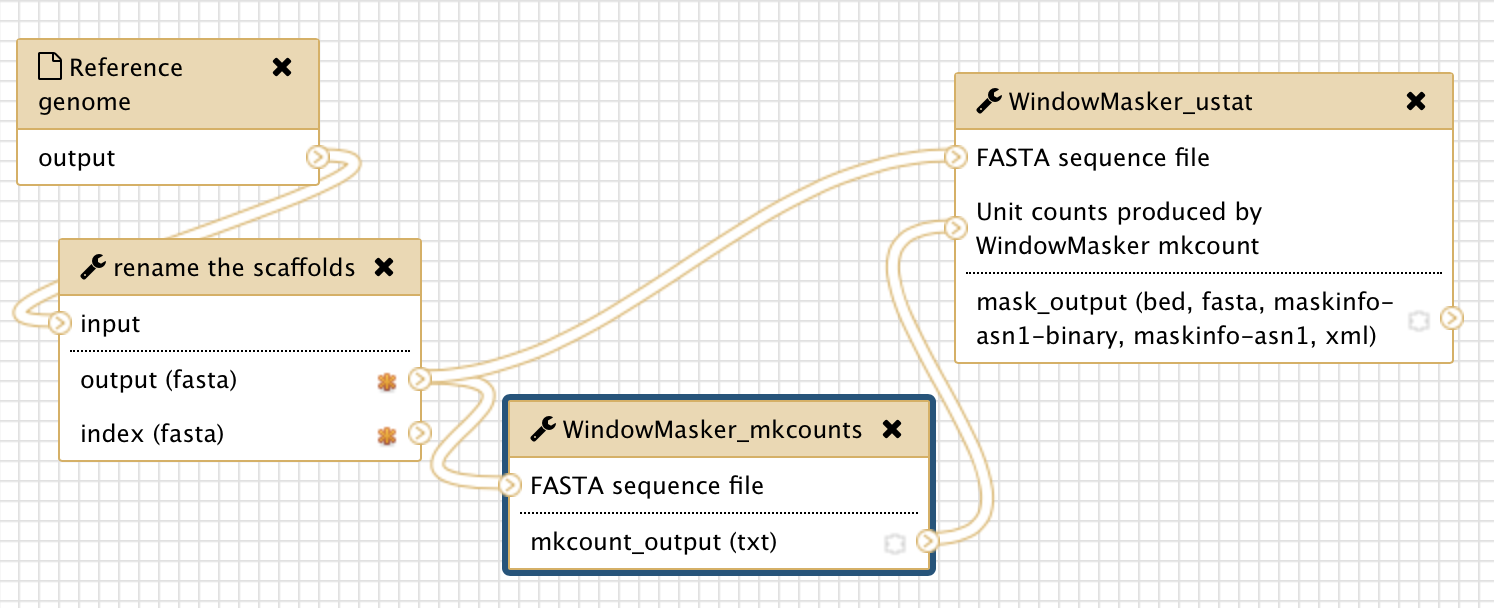


Figure : Connect the “output (fasta)” from the “rename the scaffolds” tool to the “FASTA sequence file” inputs for the WindowMasker\_mkcounts and WindowMasker\_ustat tools (red arrow). In addition, connect “mkcount\_output (txt)” from the WindowMasker\_mkcounts tool with the “Unit counts produced by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

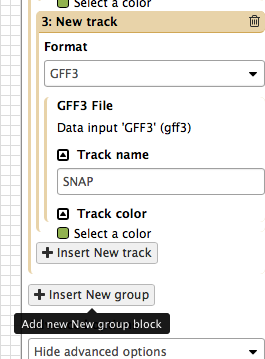


Figure : Click on the "Insert New group" button in the Details panel to add a new group in the Hub Archive Creator.

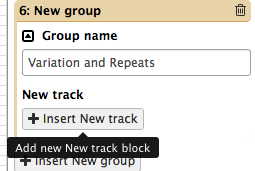


Figure : Change the name of the new group to "Variation and repeats", and then click on the "Insert New track" button to create a new input connection for the Hub Archive Creator.

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

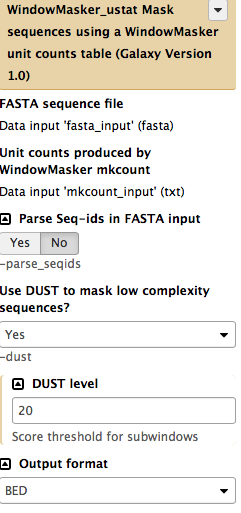


Figure : Open the Details panel of the WindownMasker\_ustat tool and verify that the output format is set to BED.

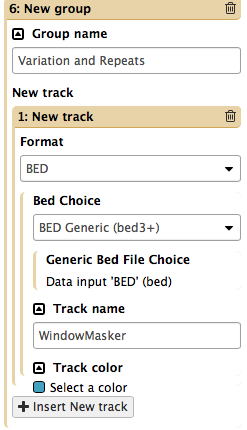


Figure : Select the BED format in the drop-down menu, then change the “Track name” to “WindowMasker” and the track color to “light blue”.

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

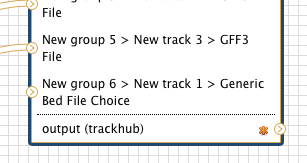


Figure : The Hub Archive Creator with a new GFF3 input connection

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).

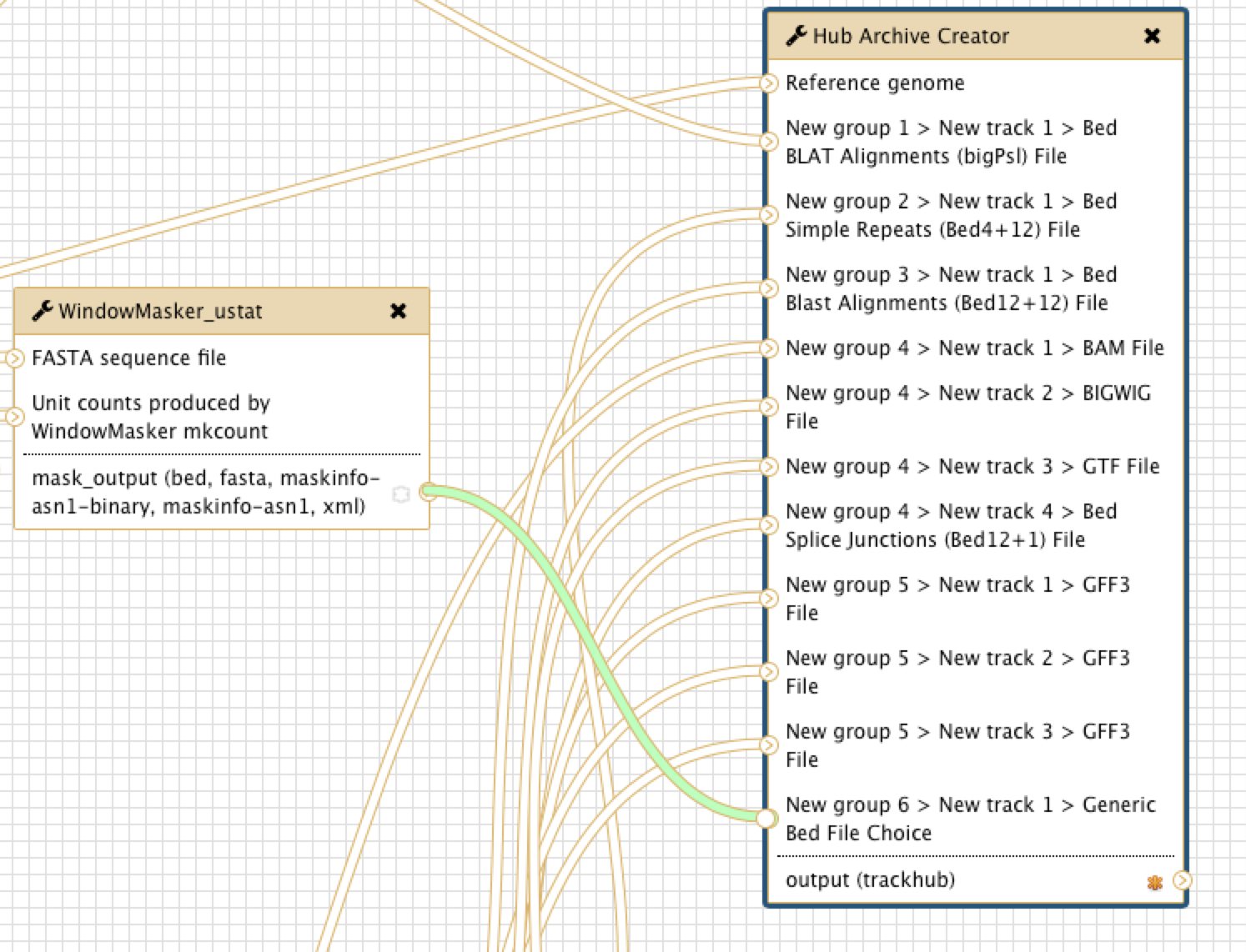


Figure : Connect the output of WindowMasker\_ustat to the Hub Archive Creator

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

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 17).

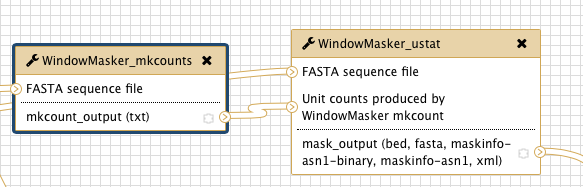


Figure : Click on the “x” at the top right corner to delete the WindowMasker\_mkcounts and the WindowMasker\_ustat tools from the workflow.

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 entry (Figure 18).

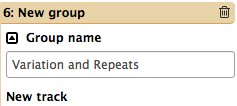


Figure : Click on the trash icon to remove the “Variation and Repeats” group from the Hub Archive Creator.

## 2.5 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 in the workflow, only the gtf file will appear in your History after you run the workflow (Figure 19). 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.

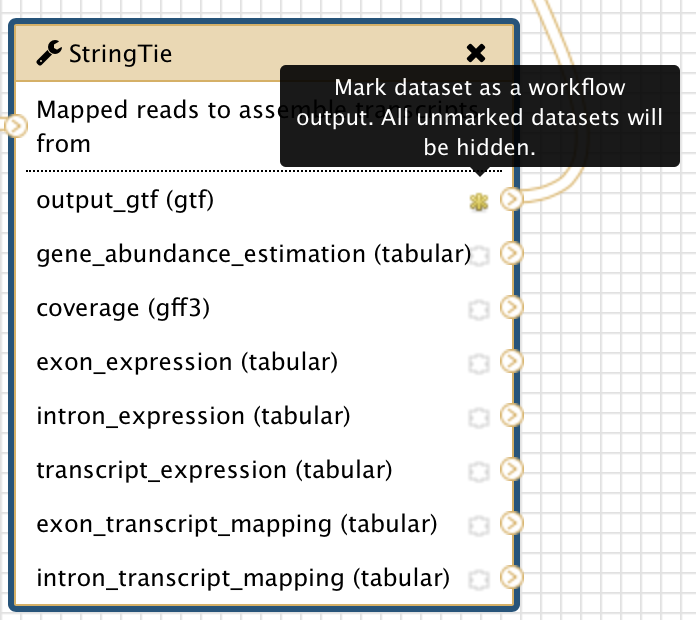


Figure : The yellow star next to “output\_gtf (gtf)” output for the StringTie tool indicates that this dataset has been marked as a workflow output

## 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 20).

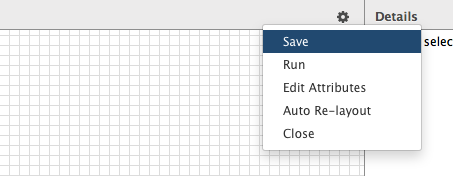


Figure 20: Save the changes that you have made to the workflow

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

## 3.1 Make a copy of the current workflow

Make a copy of “G-OnRamp:D. biarmipes F element” workflow and rename the new workflow as “G-OnRamp: WindowMasker” (Figure 21). Click on the down arrow next on the new workflow and then select the Edit option to go to the Workflow Canvas.

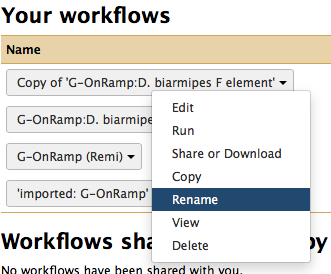


Figure : Copy the current workflow and rename it as “G-OnRamp: WindowMasker”

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

Below are some hints on how to construct this workflow:

1. You need to add two WindowMasker tools and incorporate them to 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 22).

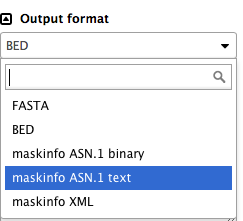


Figure : Choose the “maskinfo ASN.1 text” option under the “Output format” field

1. You need to use the maskinfo output from WindowMasker\_ustat tool as the input to the NCBI BLAST + makeblastdb tool.
2. (Optional) You can also add a WindowMasker track to the Hub Archive Creator. Hint: you will need to add another WindowMasker\_ustat tool and specify the Output format as BED.

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

## 3.3 Upload your datasets and run the workflow

Create a new History and import the datasets in Intro\_G-OnRamp in the Data Libraries. Run the “G-OnRamp: WindowMasker” workflow. Don’t forget to change the model organisms for the gene prediction tools and change the name of your new assembly hub. 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 23).



Figure : 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.

# 4. Exercise: run G-OnRamp to create JBrowse genome browser

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.