



Tutorial 2: Getting started with GIO workflows

1 Using an existing workflow

Currently available ready-made GIO workflows can be found by selecting *Published Workflows* from GIO's *Shared Data* menu. Among the pipelines listed you will see one called *Example from introductory video*. In the first part of the tutorial we'll take a look at this workflow and use it to identify proteins from some spectral data.

1.1 Importing a shared workflow

You can't use a shared workflow directly – you first need to import it to your own list of workflows. To do this:

1. Click the down arrow to the right of the workflow's name and select *Import* from the dropdown menu that appears.
2. Click *Workflow* in the GIO menu bar to see a list of your workflows – you should see the one you just imported, named "imported: Example from introductory video". Click the down arrow next to this workflow and select *Edit* to see a schematic visualisation of the workflow (see Figure 1a). This workflow essentially captures the process we went through in Tutorial 1, although there is an extra file conversion step at the beginning to cater for a wider range of data formats.

Within the *Edit* view you can explore details of the workflow, and edit parameters, by clicking on the individual tools.

1.2 Running a workflow

Let's get some data and run it through the workflow that we just imported:

1. Click *Shared Data* in GIO's menu bar and click *Data Libraries* in the dropdown list that appears.
2. From the list of available data libraries, click *NM Human PIT proteomics data*.
3. From the list of files in this library, select the *DM_6.raw* spectra file and the *human.fasta* protein database by ticking the boxes to the left of their filenames and click *Go* to import to your *History*.
4. Click *Workflow* in GIO's menu bar to go back to your list of workflows, then click the down arrow next to the *Example from introductory video* workflow and click *Run* in the dropbox list that appears.
5. In the workflow form that appears, select the files that you just imported as the two inputs (see Figure 1b).
6. Click the *Execute* button. GIO will immediately add several items to the *History* pane, representing the outputs of various workflow steps. These will gradually turn green as the workflow progresses. When the last of these history items turns green, the workflow is complete.

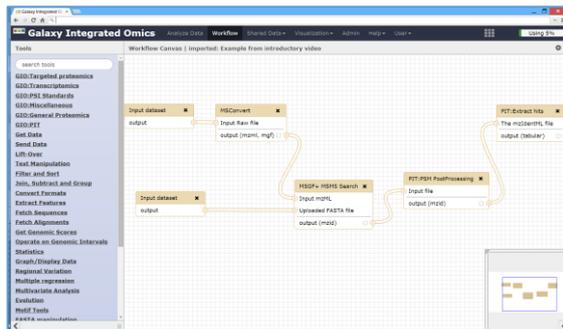


Figure 1a: Workflow shown in workflow editor.

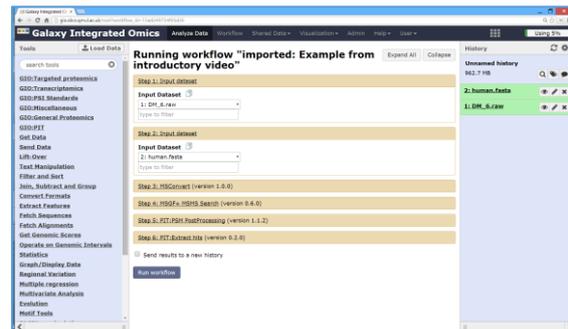


Figure 1b: Preparing to run a workflow.

2 Create your own workflow

There are three main ways in which you can create a workflow: (i) edit an existing workflow, (ii) start from scratch; (iii) generate a workflow from your *History*.

2.1 Edit an existing workflow

If you go back to the edit view of the example workflow (by going back to the workflow list, clicking the down arrow next to this workflow and selecting *Edit*), you can modify the workflow.

You can change individual parameters within the workflow by clicking on a specific tool in the workflow and modifying the parameters that appear in the rightmost pane. For example, if you click the *PIT:PSM PostProcessing* tool you can increase the false discovery rate (FDR) threshold applied by increasing the threshold value from 0.01 to 0.05. **Note that for any changes to take effect you need to save the workflow by clicking the cog icon above the workflow and selecting *Save* from the dropdown menu that appears.** If you run the workflow now, you will get slightly different results to before.

You can also remove steps from the workflow and reroute the flow of data. Let's say, for example, that your data is already in mzML format so you don't need the conversion step. In the workflow editor, just click the delete (x) icon on the top right of the MSConvert tool to remove it. You'll then need to connect the input spectral dataset directly to the *MSGF+ MSMS Search* tool – you can do this in the workflow editor by clicking on the output (>) of the *input dataset* box and dragging a connector to the input (>) of the *MSGF+ MSMS Search* box.

If you want to add a step to the workflow, you can do this simply by clicking the relevant tool in the *Tools* pane. The tool will then appear in the workflow editor, where you can drag it to where it needs to be and connect it to other tools. If you need to delete a connection to make way for the newly added tool, do this by hovering over the end of the connector (>) until a delete (x) icon appears on the connector, then click the x to make the connection disappear.

2.2 Create a workflow from scratch

To create a brand new workflow, go to the workflow list (by clicking *Workflow* in Galaxy's menu bar) and click the *Create new workflow* button that appears just above the list of workflows. You'll be asked to give the workflow a name and will then be returned to the workflow list, where your new workflow will now be listed. To start creating this workflow, click the down arrow next to the workflow's name and select *Edit* from the dropdown menu that appears. This will take you to the familiar workflow editor, which will be blank. You can now add tools, set parameters, and connect the tools together into a workflow using the methods explained in section 2.1.

2.3 Generate a workflow from your *History*

The final way to create a workflow is to generate it automatically from your *History*. The easy way to understand how to do this is to work through a simple example:

1. Start a new history by clicking the cog icon at the top of *History* pane, then selecting *Create new* from the dropdown menu that appears.
2. Import two sequence files, *adeno.fasta* and *human.fasta*, into your history from the *NM Human PIT proteomics data* shared data library. (Refer back to the first three steps of section 1.2 if you're not sure how to do this.)
3. Combine the two datasets using the built-in *Concatenate datasets* tool – you can find this in the *Text manipulation* section of the *Tools* pane. When the concatenation is complete, you should have a green item called something like *Concatenate datasets on data 2 and data 1* in the *History* pane.
4. Next, use the *PIT:Protein summary* tools (from the *GIO:PIT* tool section) to produce a text file containing a simple summary of the content of the combined FASTA file. You can click the eye (👁) icon to see what the summary looks like.
5. So we've created a simple analysis history. To create a workflow from this history, click the cog icon at the top of the *History* pane and select *Extract workflow* from the dropdown menu that appears. The central pane will change to show various options, but for the purpose of this exercise just click *Create workflow* there. The workflow will be created and added your list of workflows, from where you can run and edit it. In the editor it should look something like Figure 3.

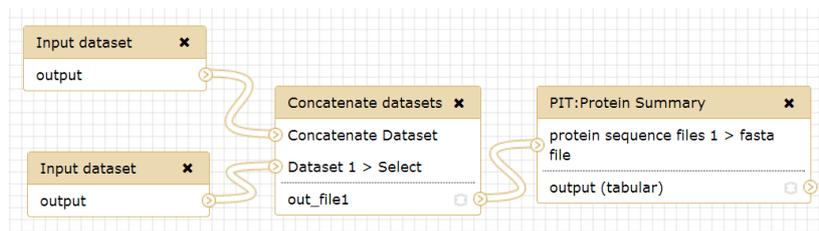


Figure 3: A simple workflow extracted from a simple history.

2.4 Tips for creating workflows

So which of the above is the best way to create a workflow? It really depends what you're trying to do. It could be that an existing workflow will almost do what you want, but you need to change some parameter or add some pre- or post-processing steps. Alternatively, if you've just completed a complicated analysis and want to record what you've done so that you (or others) can repeat it, then extracting a workflow from your *History* is the way to go.

When creating a workflow, Galaxy stops you making obvious mistakes such as using the wrong type of data as the input to a tool. However, some errors will be missed, e.g. using a FASTA file containing DNA sequences as input to a tool that expects a FASTA file of protein sequences. GIO workflows therefore need to be checked and tested for errors just like any other software pipeline.

3 Share a workflow

Having created a workflow, you may want to share it with colleagues or reviewers of work that you have done in GIO. You can do this by clicking the down arrow next to a workflow in the workflow list, selecting *Share or Publish* in the dropdown menu and then choosing how you want to share the workflow, and who you want to share it with.

4 Summary and next steps

After working through these first two tutorials you will have grasped the key concepts of GIO and the Galaxy platform on which it is built. There is a wealth of further generic Galaxy documentation available via the main Galaxy web site (wiki.galaxyproject.org). Information about the tools available within GIO can be found in GIO itself – if you click on a tool in the *Analyse Data* user interface you will find a description of the tool under its parameters.

In the next tutorial, we look at the workflows that we provide within GIO – what they do and how to use them.