QIAGEN

Tutorial

Running workflows with two inputs in batches

September 2, 2020

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Running workflows with two inputs in batches

In this tutorial, we launch a workflow with two changeable inputs in batch mode. The concepts covered can be applied to launching any workflow with multiple inputs in batch mode.

The workflow used for illustration contains the **Assemble Sequences to Reference**. Each group of data to be analyzed (batch unit) will consist of specified sequences and a specified reference genome, where both the input sequences and the reference genome can be different in each batch unit. The grouping of data is described using metadata.

At the end of the tutorial, we include references to further information about workflows and about working with metadata.

Prerequisites

For this tutorial, you must be working with *CLC Main Workbench* 20.0.2 or higher, or *CLC Genomics Workbench* 20.0.2 or higher.

The tutorial data

The tutorial data is from 3 strains of *Dehalococcoides*. Reference sequences were downloaded from GenBank and the traces were downloaded from the NCBI Trace Archive. The data was imported into a *CLC Workbench* and Trim annotations added to the sequences by running the **Trim Sequences** tool.

For illustrative purposes, we have split the sequences for 2 of the strains into multiple groups, making a total of 6 groups to be analyzed.

The tutorial data set also includes a simple workflow and metadata tables that describe the references and sequence data.

General tips

- Within the wizard windows presented when launching tools or workflows, you can use the **Reset** button to change settings to their default values.
- You can access the in-built manual by clicking on **Help** buttons or by selecting the "Help" option under the "Help" menu.

Download and import the tutorial data

1. Download the `tutorial data from our website`.
2. Open the *CLC Workbench*.
3. Import the tutorial data using the standard importer:
   (a) Go to: **File** | **Import** | **Standard Import. . .**
   (b) Select the file `workflow_batching_tutorial_data.zip`.
   (c) Leave "Automatic import" selected at the bottom and click on **Next**.
   (d) Select a folder to save the imported data to and click on **Finish**.

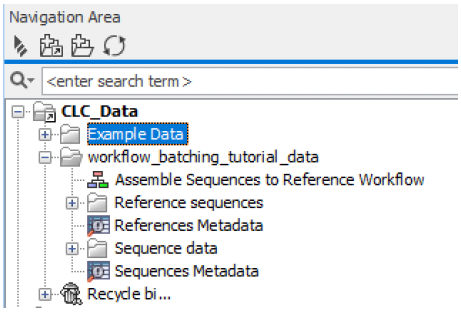

Figure 1: *After import, the tutorial data will be visible in the Navigation Area in a folder called workflow_batching_tutorial_data.*

The workflow

1. Double click on the workflow `Assemble Sequences to Reference Workflow` in the Navigation Area to open it in the viewing area (figure 2).

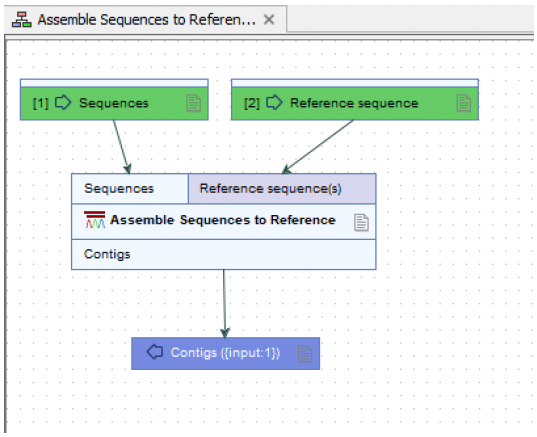

Figure 2: *This simple workflow contains two input elements, one analysis element and one output element.*

Note the 2 input elements (green boxes). When launching the workflow, we use metadata to specify which sequences and which reference sequence should be used together as inputs for each workflow run.

Expressing relationships in metadata tables

1. Double click on the two metadata tables in the Navigation Area, `References Metadata` and `Sequences Metadata` to open them in the viewing area.

`References Metadata` contains 3 columns of information about the reference sequences.

`Sequences Metadata` contains 3 columns of information about sample sequence data.

Of note in these tables:

- There is a row for each reference sequence in `References Metadata`.

- There is a row for each sample sequence in `Sequences Metadata`, and the group that the sequence belongs to is recorded in the column "Sample number".
- A column called `Species` appears in both tables, and the same terms are used in both those columns. This provides a link between the 2 tables.

In addition, each row of metadata has the relevant data element associated with it. This means that when the software reads a particular row of a metadata table, it will know which data element to use. Please refer to the manual for information about `importing metadata and associating data to metadata`.

Launching the workflow

When launching the workflow, we first select the data elements to be used in the analysis, that is, the elements with associations to the metadata.

1. Select all the rows in the `Sequences Metadata` table and click on the **Find Associated Data** button at the bottom.
2. Select all the rows in the "Metadata Elements" table that just opened, and click on the **Find in Navigation Area** button at the bottom.

   The pre-selected data (figure 3) will also be pre-selected in the workflow wizard, making this an effective way ensure the intended data is used in the analysis.

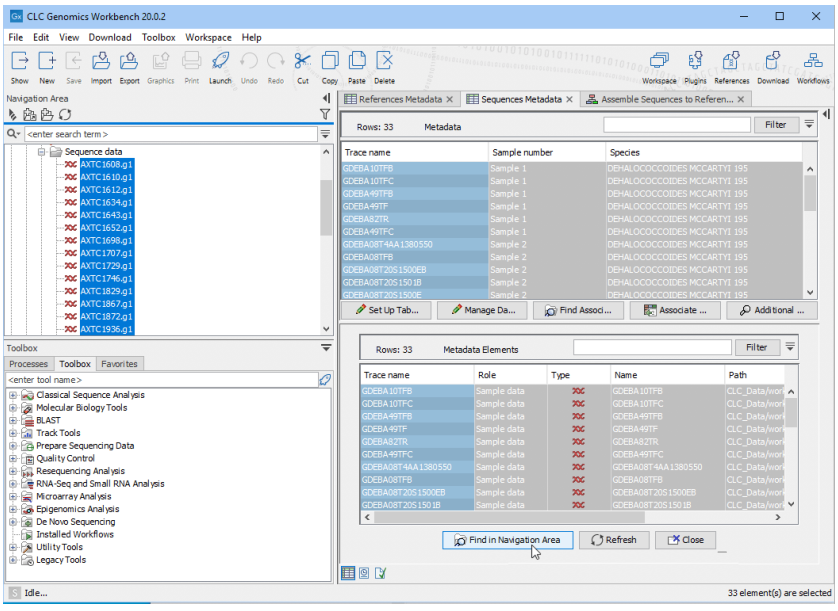

Figure 3: *Data can be pre-selected in the Navigation Area from the Metadata Elements table.*

3. Click on the `Assemble Sequences to Reference Workflow` tab to view the workflow.
4. Click on the **Run. . .** button near the bottom to launch the workflow wizard.
5. In the "Select Sequences" step of the wizard (figure 4):

(a) Check the **Batch** checkbox under the data selection area.

(b) Click on the right hand arrow to move the pre-selected data elements from the Navigation Area into the "Selected elements" area.

(c) Click on **Next**.

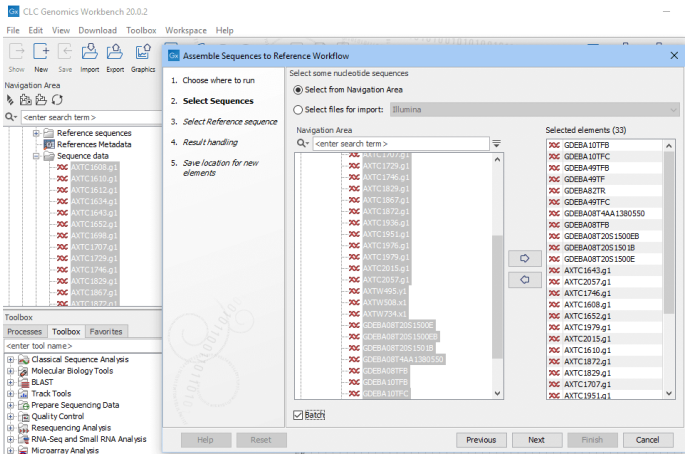

Figure 4: *All the sequences were pre-selected in the Navigation Area and then added to the "Selected elements" area, and the Batch option is checked.*

6. In the "Select Reference sequences" step:

(a) Check the **Batch** checkbox under the data selection area.

(b) Select all 3 sequences in the `Reference sequences` folder and click on **Next**.

In the "Configure batching" wizard step, (figure 5), the grouping of the data for analysis is specified.

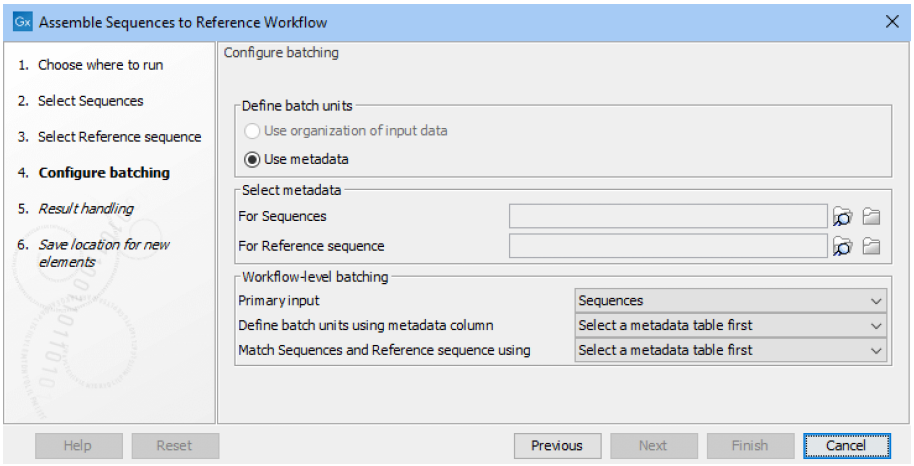

Figure 5: *The grouping of the data for analysis is specfied in the "Configure batching" step.*

Troubleshooting: If you do not see the "Configure batching" wizard step at all, or if only one field for metadata is shown, then click on the **Previous** button and ensure the "Batch" option is checked in both the previous steps.

7. Make selections so the "Configure batching" step looks like that shown in figure 6.

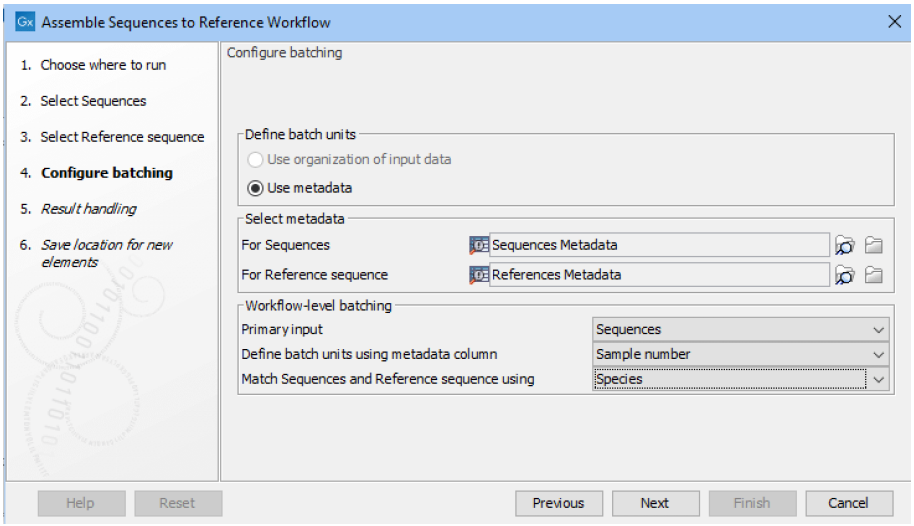

Figure 6: *By selecting relevant metadata tables and specifying relevant columns in those tables, the grouping of the data for analysis is specified. The names of the fields in this wizard step reflect the names of the input elements in the workflow.*

Specifically:

(a) In the "Select metadata" area, select `Sequences Metadata` in the "For Sequences" field and `References Metadata` in the "For Reference sequence" field.

(b) In the "Workflow-level batching" area, select `Sequences` as the "Primary input", `Sample number` for the "Define batch units using metadata column" field, and `Species` for the "Match Sequence and Reference sequence using" field.

8. Click on **Next** to get to the "Batch overview" step.

The "Batch overview" wizard step (figure 7) provides an opportunity to check that the data has been grouped as intended. Each row in the table shows a group of data that will be analyzed together. The number of rows is the number of batches, that is, the number of times the workflow will be run.

9. Click on **Next** and in the "Result handling" step, check the box beside "Create subfolders by batch unit" and check the box beside "Open log". Leave the option "Create workflow result metadata" enabled (figure 8).

10. Click on **Next**.

11. Select a folder to save the results to, and then click on **Finish**.

Reviewing the outputs

When the workflow has finished running, the results will be saved where you selected (figure 9).

Some notes about the outputs generated by this workflow run:

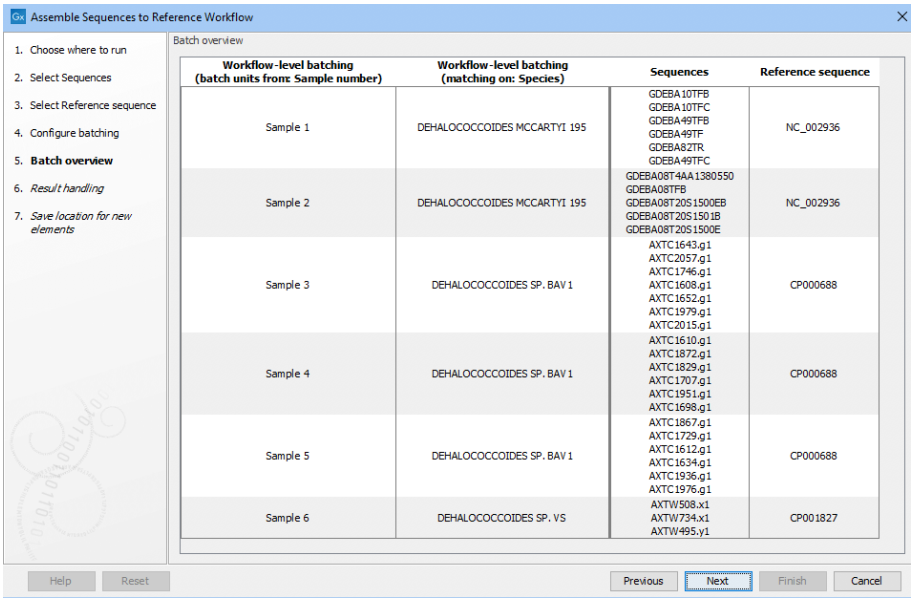

Figure 7: *The grouping of the data is presented in the "Batch overview" wizard step. The number of rows is the number of batches, and the data to be analyzed in each batch run is listed.*

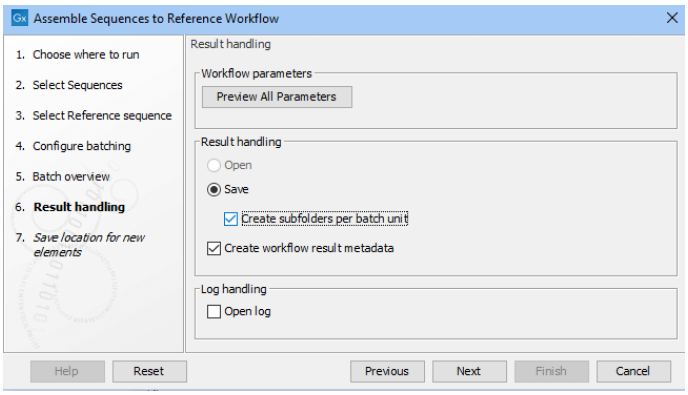

Figure 8: *With these settings in the "Result handling" wizard step, the results of each batch will be saved to individual subfolders, a workflow metadata table will be saved, and log of the progress of the workflow run will be opened when it starts running.*

- The names of the read mappings include the name of the first sequence in the list of those input in that batch. The naming `is configurable`.
- Placing results in subfolders allowed us to easily match the mapping with which sample it represents.
- The `Workflow Result Metadata` element provides another view of the results, which can be of particular use when running large numbers of batches, and where outputs of many types are generated.

Related resources

Some links to relevant information in the *CLC Workbench* manual:

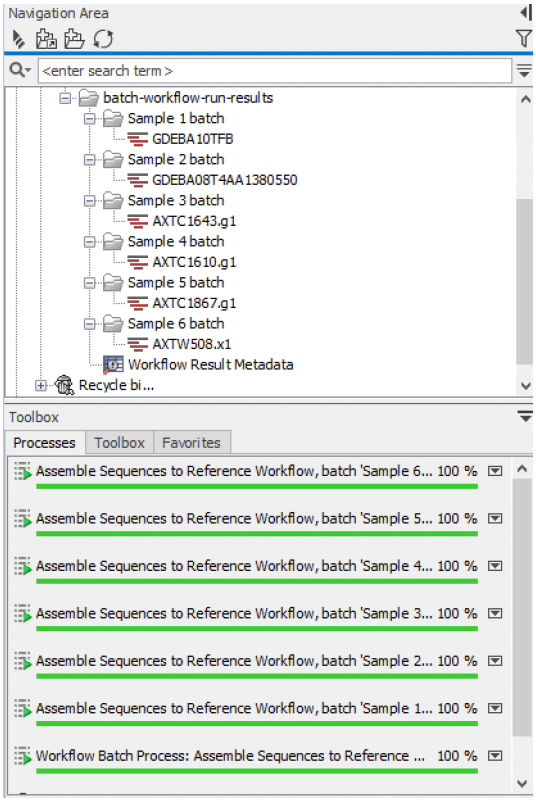

Figure 9: *A read mapping was generated for each of the 6 samples. Each mapping was put into a separate subfolder and a workflow result metadata table was produced. At the bottom, the Processes tab is shown, where the progress of the analyses can be monitored.*

- Importing metadata and associating data to metadata
- Workflows
- Assemble Sequences to Reference

We also offer a separate tutorial that focuses on generating and reviewing read mappings generated using the Assemble Sequences to Reference tool.