



# Cloud Plugin

USER MANUAL

# User manual for Cloud Plugin 22.0.1

Windows, macOS and Linux

April 12, 2022

**This software is for research purposes only.**

QIAGEN Aarhus  
Silkeborgvej 2  
Prismet  
DK-8000 Aarhus C  
Denmark



# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Prerequisites . . . . .	4
<b>2</b>	<b>Configuring AWS credentials</b>	<b>6</b>
<b>3</b>	<b>Configuring the cloud connection</b>	<b>8</b>
<b>4</b>	<b>Input data location considerations</b>	<b>10</b>
<b>5</b>	<b>Running workflows on a CLC Genomics Cloud Engine</b>	<b>12</b>
<b>6</b>	<b>Cloud Job Search</b>	<b>16</b>
6.1	Downloading all results . . . . .	18
6.2	Selective download of results . . . . .	18
<b>7</b>	<b>Cloud connections via a CLC Server</b>	<b>21</b>
7.1	Configuring the Cloud Server Plugin . . . . .	22
7.2	Configuring GCE presets . . . . .	25
<b>8</b>	<b>External applications in the cloud</b>	<b>26</b>
<b>9</b>	<b>Troubleshooting</b>	<b>28</b>
<b>10</b>	<b>Installing and uninstalling Workbench plugins</b>	<b>30</b>
10.1	Installation of Workbench plugins . . . . .	30
10.2	Uninstalling Workbench plugins . . . . .	31

# Chapter 1

## Introduction

This is the manual for Cloud Plugin 22.0.1. With this plugin installed on a *CLC Workbench*, you can configure access to a *CLC Genomics Cloud Engine (GCE)* and submit workflows to run there. The Cloud Job Search tool is also installed, allowing you to find jobs and download results from these analyses (chapter 6).

A list of prerequisites is provided in section 1.1.

Configuring connections to AWS S3 locations is described in section 2.

Configuring the connection to GCE is described in section 3.

Submitting workflows to run on GCE is described in chapter 5.

If you have access to a *CLC Genomics Server* with the Cloud Server Plugin installed and configured, then workflows can be submitted to run on GCE via the *CLC Server*, as described in chapter 7.

General information on setting up and launching workflows is provided in the *CLC Workbench* manuals: <https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflows.html>

If your organization does not yet have access to *CLC Genomics Cloud Engine*, you can read more about it and request a quote here: <https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/enterprise-ngs-solutions/qiagen-clc-genomics-cloud-engine/>

### 1.1 Prerequisites

To launch workflows to run on a *CLC Genomics Cloud Engine (GCE)* from a *CLC Workbench*, you will need the following information from your GCE administrator:

- AWS IAM user credentials in the form of an access key ID and a secret access key. These are needed for connecting to AWS to access your S3 buckets (section 2).
- The URL for your GCE setup (section 3)
- Oauth credentials for logging into GCE (section 3)
- If you will be importing data from Illumina BaseSpace, you will need an Illumina BaseSpace account.

The GCE administration manual is available from: <http://genomics-cloud-engine.s3-website-us-west-2.amazonaws.com/releases/current/>

## Chapter 2

# Configuring AWS credentials

To configure AWS accounts for data import and export, go to:

**Connections | Manage AWS S3 Locations** (🔗)

The same configuration dialog can be opened by clicking on the (🔗) icon at the bottom left of the Workbench frame.

This dialog (figure 2.1) allows you to register the credentials for one or more AWS accounts. To add an AWS account, click on **Add Amazon S3 location**. After adding one or more AWS data locations, it is possible to **Edit** or **Remove** them.

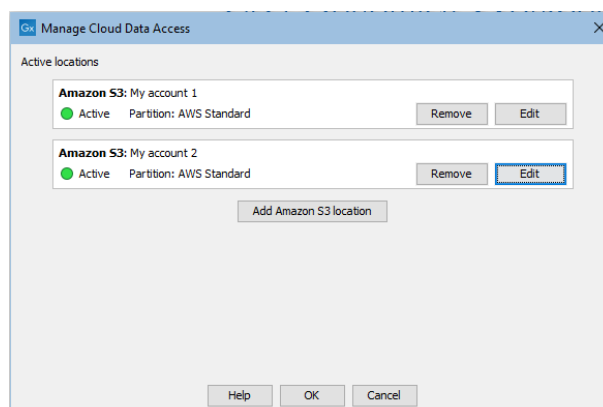


Figure 2.1: The AWS S3 Locations configuration dialog

For adding or editing AWS account credentials, the information below is required (figure 2.2). The administrator of the AWS account should be able to provide this information if you do not already have it.

**Name:** A short name of your choice, identifying the AWS account. This name will be shown as the name of the data location when importing data to or exporting data from Amazon S3.

**Description:** An optional description of the AWS account.

**AWS access key ID:** The access key ID for programmatic access, set up for the AWS IAM user.

**AWS secret access key:** The secret access key for programmatic access, set up for the AWS IAM user.

**AWS partition:** The partition under which the AWS user is registered.

The dialog continually validates the settings that have been entered. When the settings are valid, the Status box will contain the text "Valid" and a green icon will be shown. Click on **OK** to save the settings.

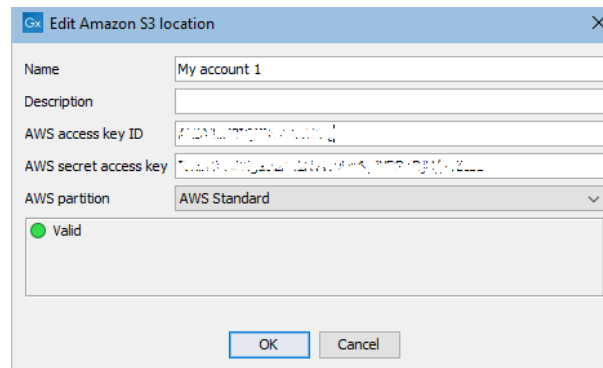



Figure 2.2: Adding an AWS account configuration dialog

When one or more AWS data locations have been added, they will be listed as data locations when importing and exporting data.

When the connection status icon at the bottom of the *CLC Workbench* looks like , a connection has been established to Amazon S3.

## Chapter 3

# Configuring the cloud connection

After one or more AWS S3 locations has been configured in your *CLC Workbench*, you can log into your *CLC Genomics Cloud Engine* by going to:

**Connections | CLC Genomics Cloud Engine Connection** (🏠)

This opens the dialog shown in figure 3.1. The following two settings must be configured:

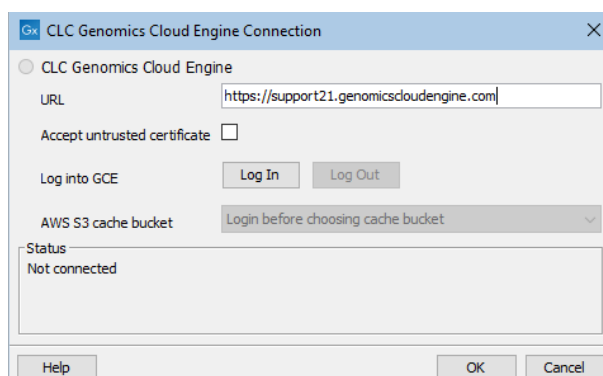


Figure 3.1: The *CLC Genomics Cloud Engine Connection* tool just after opening it from under the *Connections* menu.

**URL:** The URL pointing to the *GCE* service, as described in section 1.1.

**AWS S3 cache bucket:** The cache bucket to be used when uploading data to the cloud, as described in section 1.1.

The **Accept untrusted certificate** checkbox should be selected when *GCE* has been set up with a self-signed certificate.

### Connecting to and disconnecting from GCE

Click on the **Log In** button in the *CLC Genomic Cloud Engine Connection* dialog to open a web browser page where you can log in to *GCE* using your company credentials. Once you have logged in successfully in the browser, go back to the *CLC Workbench* and click on the **OK** button in the *CLC Genomic Cloud Engine Connection* dialog to complete the log in process.

When the authentication succeeds, the Status box will contain the user name and region for the connection, and a green icon is visible near the top (figure 3.2).



Click on the **Log Out** button in the Cloud Connection dialog and then click on the **OK** button to disconnect the *CLC Workbench* from GCE. You will be asked to confirm this is what you wish to do.

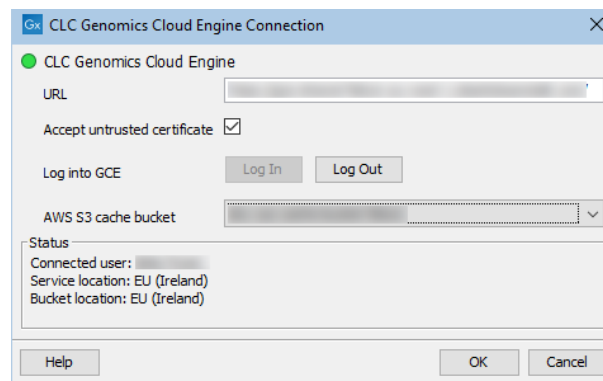



Figure 3.2: A valid **GCE** configuration is indicated by the green icon and the user details in the Status box.

The cloud connection status icon at the bottom, left hand side of the *CLC Workbench* looks like  when the connection to *CLC Genomics Cloud Engine* is ready for workflows to be submitted. Hover the mouse cursor over this icon to see details about the connection.

## Chapter 4

# Input data location considerations

Input data for your workflow may be on your local system or in the cloud, either in an S3 bucket or in Illumina BaseSpace.

- **Input data from your local system** When a workflow is launched, data is uploaded to the GCE cache bucket<sup>1</sup> unless it is already present there.

The location and time of latest modification are used to determine if the the most recent version of the data is already in the cache bucket.

See also the section below about additional considerations relating to reference data.

- **Input data from an S3 bucket**

Your AWS administrator can grant role-based access to AWS S3 buckets to GCE. This is described in the GCE Administration manual.

If you select input data from an S3 bucket that GCE does not have permissions to access directly, the files are transferred using presigned URLs. This grants *time limited* access to the selected files to GCE. By default, these presigned URLs are valid for 7 days, which is the maximum allowable by AWS at time of writing.

- There is no charge for data transfer when using data in an S3 bucket that is in the same region as GCE.
- There is a small charge for cross region traffic when using data in an S3 bucket that GCE has access to, but that is in a different region to GCE.

See the Amazon documentation for more on S3 pricing (<https://aws.amazon.com/s3/pricing/>). At time of writing, AWS does not charge for uploading data to S3, while storage in S3 and download from S3 are chargeable.

---

<sup>1</sup>The cache bucket is configured by your GCE administrator. It is a cloud-based location for the temporary storage of input data that you selected from a local system when launching a workflow. By default, files in the cache bucket are retained for 30 days after their last use. Your GCE admin can adjust this period, so please check with them if in doubt.

**Additional considerations relating to reference data**

Often, data is needed for the analysis that is not itself being acted upon by the analysis. for example, reference sequences to be mapped against, or target regions to limit the focus of the analysis. Such reference data flows into parameter input channels in workflows.

Reference data transfer costs differ depending on the data source:

- QIAGEN Reference Data Elements are already present in AWS S3 in all regions GCE is supported on.

These elements are thus not be uploaded to S3 when a workflow is launched. Rather, a copy of the data already in AWS is used.

QIAGEN reference data is available from under the QIAGEN Sets tab of the *CLC Genomics Workbench*.

- Other reference data is handled like any other data input: it is uploaded to the cache bucket *unless* the most recent version is already present there.

## Chapter 5

# Running workflows on a CLC Genomics Cloud Engine

To launch workflows on a *CLC Genomics Cloud Engine*, select a "CLC Genomics Cloud Engine" option in the first wizard step (figure 5.1).

Selecting "CLC Genomics Cloud Engine" will submit the workflow via your *CLC Workbench*. Selecting "CLC Genomics Cloud Engine (via CLC Server)" will submit the workflow via a *CLC Server*. This option is only available to select if you are connected to a *CLC Genomics Server* with the Cloud Server Plugin installed and configured. See chapter 7 for details.

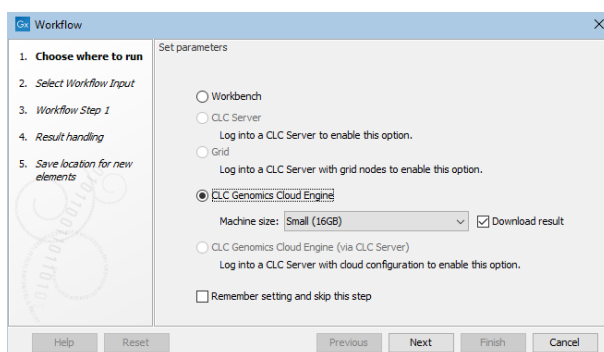


Figure 5.1: Select "CLC Genomics Cloud Engine" to submit a workflow the CLC Genomics Cloud Engine via your Workbench.

### Machine size and downloading results

Select the desired machine size in the drop-down menu (figure 5.2). The available options can be configured by your GCE administrator. Typically, the larger the machine, the greater the cost, although a job's duration will also affect costs.

When running small workflows where you wish to download all the workflow results, keep the **Download result** checkbox checked. For other situations, we recommend this option is not selected. The *CLC Workbench* must be left running until the workflow completes for the results to be downloaded automatically. Workflow results can be downloaded later using the Cloud Job Search functionality, described in chapter 6.

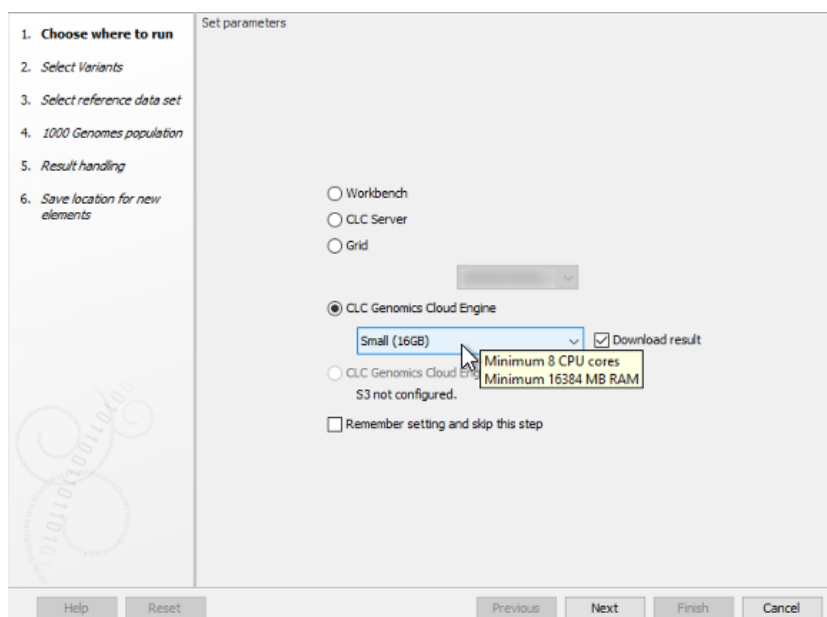


Figure 5.2: Select a machine size that reflects the requirements of the workflow being run. Hover the mouse cursor over a selected size to see more details.

### Handling input data

When launching a workflow, the data to be analyzed can be selected from the Navigation Area or from an external location. When files from an external location are chosen, they will be imported on the fly, that is, the first step in the workflow execution on GCE will be to import the data. On-the-fly import is described in more detail in the *CLC Workbench* manuals: [https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Importing\\_data\\_on\\_fly.html](https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Importing_data_on_fly.html)

How data is made accessible to GCE depends on where it is located when launching the workflow. This is described in chapter 4.

### QIAGEN reference data in workflows

Reference data provided by QIAGEN is already available on AWS in every region that GCE is supported in. This means that when these data elements are selected when launching a workflow, no data transfer occurs.

When all parameters that take reference data elements are locked, and all data referred to are QIAGEN Reference Data Elements, then the workflow can be launched to run on GCE from a *CLC Genomics Workbench* without downloading the data locally. The exception is where multiple data elements can be selected for a single parameter. In this case, the data must be present locally to be able to submit the workflow.

If any parameter for reference data in a workflow is unlocked, then the reference data elements referred to must be present locally to send the job to GCE.

**Note:** When submitting such workflows to GCE via a *CLC Server*, the data must be present in the *CLC Server CLC\_References* area, even though a copy of that data already in the cloud is actually used.

See also chapter 4 for information about data location considerations relating to reference data.

## Result handling

In the "Output location in Amazon S3" wizard step, you specify where to save the workflow results to. If you have configured more than one S3 location, you will be offered a choice of locations. Information about data to be uploaded is also displayed here, as shown in figure 5.3. Any data already present in the cloud cache will not be uploaded.

Results are always saved to Amazon S3, whether or not the "Download result" option is checked in the first wizard step when launching the workflow.

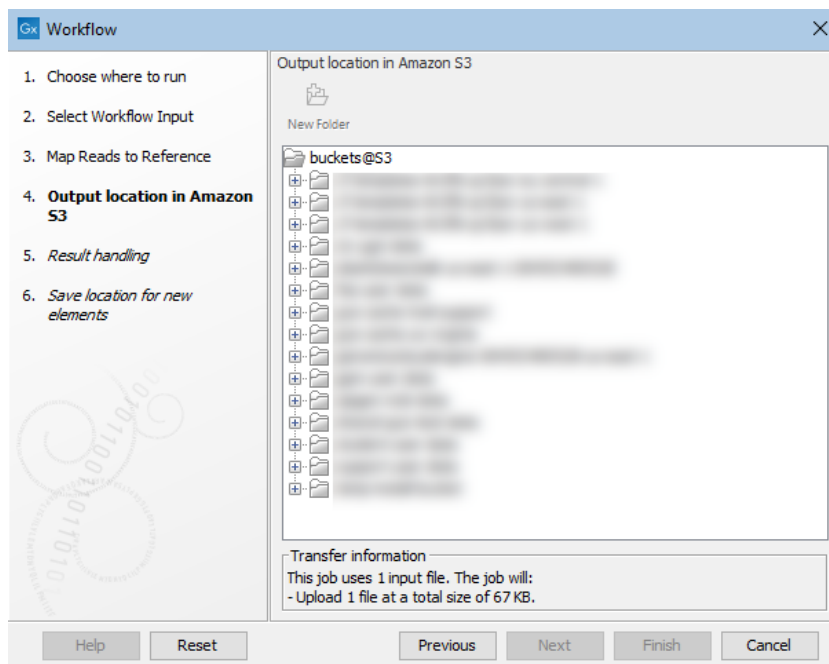


Figure 5.3: Specifying a location for saving workflow results to in Amazon S3. Information about data to be uploaded when the workflow is launched is provided near the bottom of this wizard step.

In the last wizard step, a local location must be selected for workflow outputs. If the "Download result" checkbox was not checked in the first configuration step, this location is still needed as log files are saved here in some circumstances, for example, if the workflow fails for particular reasons.



## Following the progress of workflow jobs run on the cloud

Each workflow submitted to the cloud is submitted as a *batch* consisting of *jobs*. A batch may consist of just a single job. Multiple jobs are included in a batch when:

- The "Batch" checkbox is selected in the workflow wizard, and/or
- The workflow design includes control flow elements, as described in the *CLC Genomics Workbench* manual: [https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Advanced\\_workflow\\_batching.html](https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Advanced_workflow_batching.html).

Each job within a batch is executed as a separate job in the cloud, potentially in parallel on separate server instances.

You can follow the progress of the workflow in the Processes area of the *CLC Workbench* (figure 5.4). The icon next to the process indicates the status of the job submission:

-  This icon indicates that data is being transferred to the cloud. When this icon is displayed, do not interrupt the connection to the cloud, e.g. do not disconnect from the cloud or shut down your computer.
-  This icon indicates that the job submission is complete, including any data transfer. When this icon is displayed, you can safely disconnect from the cloud, and shut down your computer if you wish.

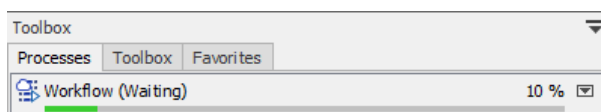


Figure 5.4: The icon next to the cloud process in the Processes area indicates the submission of this job, including data transfer, is complete.

When the job submission is complete, right-clicking the arrow next to a process and selecting "Show in Cloud Job Search" will open the batch in the Cloud Job Search (figure 5.5). See chapter 6 for further details.

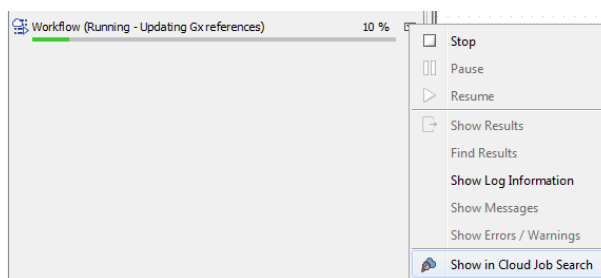


Figure 5.5: You can open an individual job in the Cloud Job Search tool by right-clicking on the arrow next to a process in the Processes area. This option is only available when the job submission to the cloud is complete, including any data transfer.

## Chapter 6

# Cloud Job Search

To open the Cloud Job Search tool, go to

**Utilities | Cloud Job Search** (🌥)

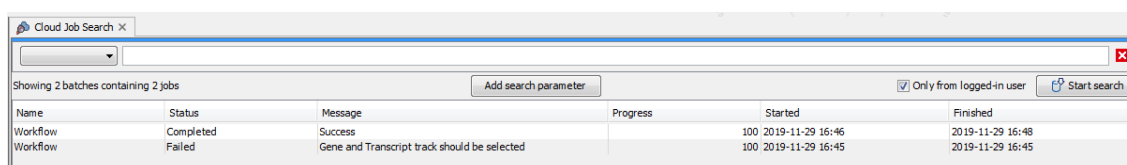
or click on the Cloud button in the right hand side of the top Toolbar.

If a search is run without specifying search criteria, the most recent batches submitted to the *CLC Genomics Cloud Engine* will be listed (figure 6.1). The number of batches returned is limited. If the jobs you are looking for do not appear in the initial set returned, restrict the search criteria using the fields at the top. To do this, click on the **Add search parameter** button and start the search using the **Start search** button on the top, right hand side.

This should result in older submissions of interest being returned.

With the "Only from logged-in user" checkbox selected, then the list retrieved will include only jobs submitted by the user currently logged into the *CLC Genomics Cloud Engine* from the Workbench. When this checkbox is not selected, jobs submitted by *any user* of the *CLC Genomics Cloud Engine* instance can be retrieved.

**Note:** The table is not refreshed automatically. The "Start search" button must be pressed to update the information, including job status information.



The screenshot shows the 'Cloud Job Search' window. At the top, there's a search bar and a dropdown menu. Below it, a status bar indicates 'Showing 2 batches containing 2 jobs'. To the right of this are buttons for 'Add search parameter', a checkbox for 'Only from logged-in user' (which is checked), and a 'Start search' button. The main area contains a table with the following data:

Name	Status	Message	Progress	Started	Finished
Workflow	Completed	Success	100	2019-11-29 16:46	2019-11-29 16:48
Workflow	Failed	Gene and Transcript track should be selected	100	2019-11-29 16:45	2019-11-29 16:45

Figure 6.1: *Cloud Job Search* allows you to search for and inspect the properties of jobs that have been submitted to the cloud.

In the side panel settings of a Cloud Job Search, you can select the columns to display and control whether batches or individual jobs are listed in the table (figure 6.2). When the **Collapse batches** option is checked, each batch is shown as a single row in the table. When unchecked, each individual job is shown in its own row.

All jobs in a batch can be canceled by right-clicking on the row in the table corresponding to the batch (or a job in a batch, when displaying individual jobs), and selecting **Cancel... | Cancel all jobs in batch** from the menu that appears.



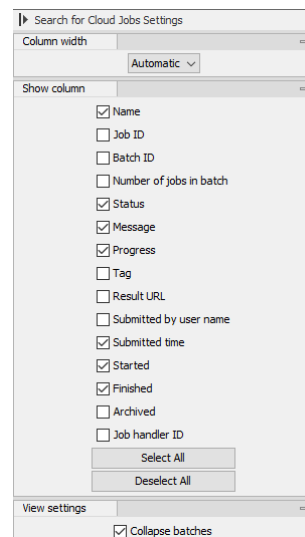


Figure 6.2: The Cloud Job Search side panel settings are used to configure what is shown in the table.

After a cloud job has completed, results can be retrieved by selecting the jobs of interest and then using the buttons at the bottom of the view (figure 6.3):

- **Download All Results** Download all results from the selected jobs, optionally including any exported files. See section 6.1 for further details.
- **Download Metadata** Download only the Workflow Result Metadata table for the selected jobs. Results can then be downloaded selectively using the Workflow Result Metadata table, as described in section 6.2.

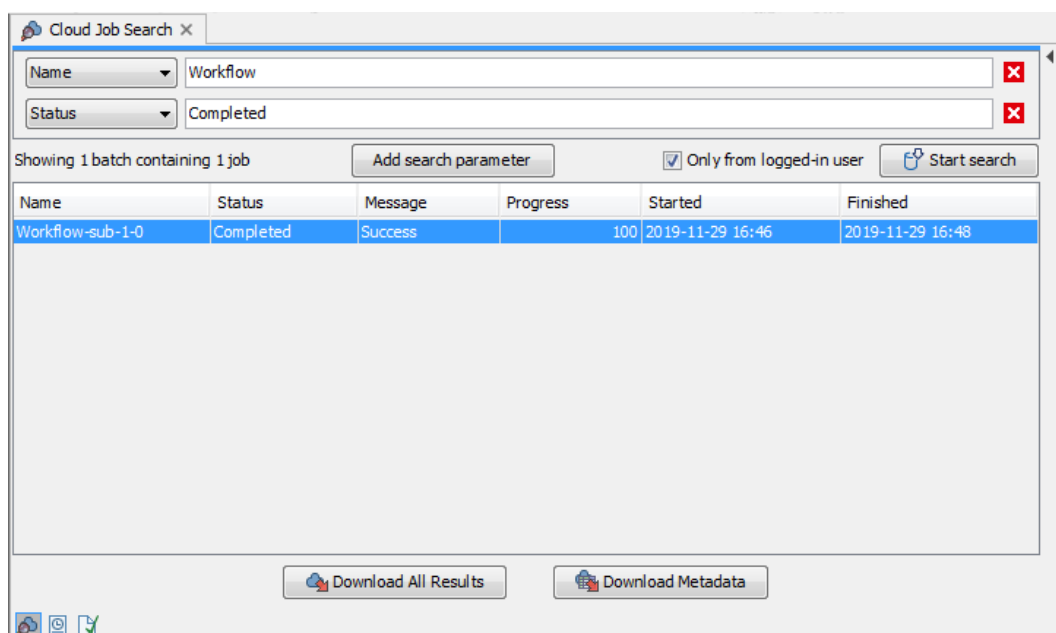


Figure 6.3: The Download All Results and Download Metadata buttons can be used to retrieve results from the cloud job submissions.

## 6.1 Downloading all results

To download all results, optionally including any exported data, select one or more rows in Cloud Job Search table, and click on the **Download All Results** button. Data elements will be downloaded into the Navigation Area in the folder structure specified in the workflow design for the output elements.

Check the **Also download exported files** checkbox in the Result handling step of the wizard to download exported files in addition to downloading the other results. You can specify the location to save the exported data to in the **Export directory** field. Note: Specifying a cloud location when downloading exported files is not supported.

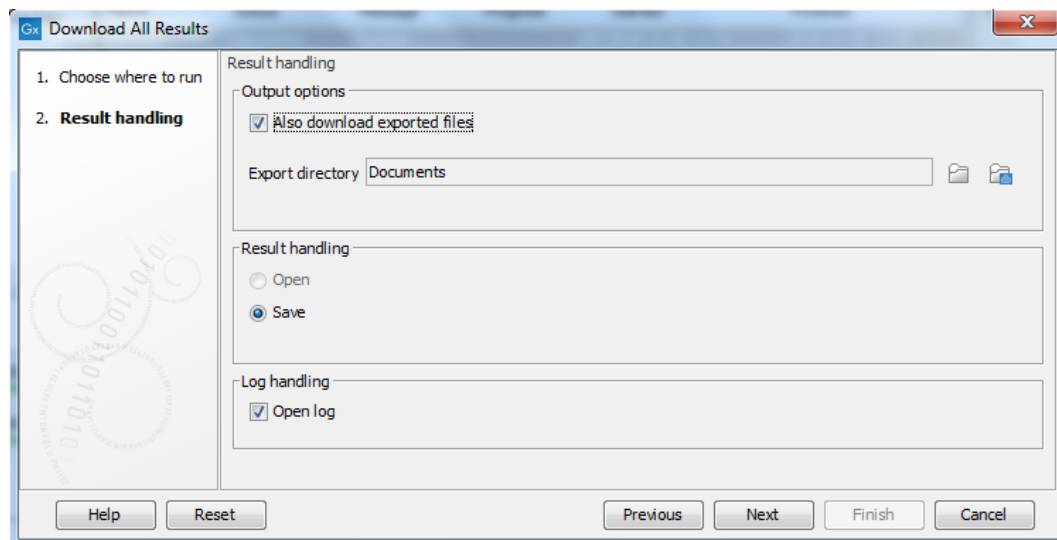


Figure 6.4: The "Also download exported files" checkbox allows you to download exported files as well as results to be saved to the **CLC Workbench** Navigation Area.

In cases where a job failed, the **Download All Results** button can be used to download logs and other technical information. More information about troubleshooting is provided in chapter 9.

## 6.2 Selective download of results

Individual results can be downloaded using the Workflow Result Metadata table, which is generated for each batch. This is particularly useful when the amount of data generated by a workflow execution is so large that it is not practical to download all the results at once.

In the Cloud Job Search table, download the Workflow Result Metadata table for selected jobs by clicking on the **Download Metadata** button.

In a Workflow Result Metadata table, each workflow output is shown in a separate row. The path to the output in the AWS S3 location is indicated in the "External path" column (figure 6.5). When one or more rows in the table are selected, the **Find Associated Data** button is enabled. Clicking on this button opens a Metadata Elements table, where the results are listed. The path to each data element is provided here. Paths starting with `s3://` are those stored in AWS S3. When such rows are highlighted, the "Download" button at the bottom of that table is enabled.

To download particular data elements from the cloud, select the desired rows in the Metadata Elements table, and click on the **Download** button. Data elements downloaded are placed in the

Batch identifier	Time Point	Run Accession	Infected With	Produced by	From output	External path
24 hours post infection-SRR3872522	24 hours post infection	SRR3872522	mock	RNA-Seq Analysis	Transcript Expression Track (RNA-Seq/{1})	s3://sky-vc-cache-bucket/CLC-processes/RNA-Seqcd7aca3-f25f-40fe-88af-3f553c49ab68-20191130-113023059/-ja5eqa
24 hours post infection-SRR3872522	24 hours post infection	SRR3872522	mock	RNA-Seq Analysis	Reads Track (RNA-Seq/{1})	s3://sky-vc-cache-bucket/CLC-processes/RNA-Seqcd7aca3-f25f-40fe-88af-3f553c49ab68-20191130-113023059/-ja5eqa

Role	Type	Name	Path
Result data		RNA-Seq/SRR3872522 (TE)	s3://sky-vc-cache-bucket/CLC-processes/RNA-Seqcd7aca3-f25f-40fe-88af-3f553c49ab68-20191130-113023059/-ja5eqa

Figure 6.5: The "External path" column in the Workflow Result Metadata table shows the path to the data in AWS S3. After the "Find Associated Data" button is clicked, a Metadata Elements table is opened, listing the associated data elements and their locations.

subfolder (if any) that was specified by the workflow design for the given output. New folders are created as required. Therefore, we recommend selecting the same folder to store different results for the same workflow. This ensures that the outputs are organized in the same folder structure that they would have been if the workflow had been executed by the *CLC Workbench*.

When a data element is downloaded, it is automatically associated with the relevant row of the Workflow Result Metadata table, and can be found by clicking on the **Find Associated Data** button again (figure 6.6). External references will not be removed from the Workflow Result Metadata table by downloading a result. Results can be downloaded this way any number of times.

Batch identifier	Time Point	Run Accession	Infected With	Produced by	From output	External path
24 hours post infection-SRR3872522	24 hours post infection	SRR3872522	mock	RNA-Seq Analysis	Transcript Expression Track (RNA-Seq/{1})	s3://sky-vc-cache-bucket/CLC-processes/RNA-Seqcd7aca3-f25f-40fe-88af-3f553c49ab68-20191130-113023059/-ja5eqa
24 hours post infection-SRR3872522	24 hours post infection	SRR3872522	mock	RNA-Seq Analysis	Reads Track (RNA-Seq/{1})	s3://sky-vc-cache-bucket/CLC-processes/RNA-Seqcd7aca3-f25f-40fe-88af-3f553c49ab68-20191130-113023059/-ja5eqa

Role	Type	Name	Path
Result data		SRR3872522 (TE)	CLC_Data/test/RNA-Seq
Result data		RNA-Seq/SRR3872522 (TE)	s3://sky-vc-cache-bucket/CLC-processes/RNA-Seqcd7aca3-f25f-40fe-88af-3f553c49ab68-20191130-113023059/-ja5eqa

Figure 6.6: When the results have been downloaded, they are automatically associated with the relevant row of the Workflow Result Metadata table, and can be located by pressing the "Find Associated Data" button again.

A connection to the *CLC Genomics Cloud Engine* is required to find the Workflow Result Metadata table and download it using the Cloud Job Search functionality. However, only the AWS S3 connection is required for downloading results via the Workflow Result Metadata table.

Thus, by saving the Workflow Result Metadata table, you can download the results from AWS S3

at a later point, as long as the data is still available in AWS S3. You do not need a connection to the *CLC Genomics Cloud Engine* or to use the Cloud Job Search for this activity.

**Note:** Exported data cannot be downloaded selectively. To download data exported by the workflow, you must use the **Download All Results** button in the Cloud Job Search tool. See section [6.1](#) for further details.

## Chapter 7

# Cloud connections via a CLC Server

Workflows can be submitted to run on the *CLC Genomics Cloud Engine*(GCE) via the *CLC Genomics Server*. This can be particularly useful when using data stored on the *CLC Server*, as data will be submitted from the server location to S3 directly, avoiding downloading it to your local system.

### Note:

- There are no user permissions on jobs in GCE. This means that GCE users will be able to find each other's jobs, for example by using the Cloud Job Search functionality in their Workbench.
- Submitting jobs to GCE can be restricted to specific groups by configuring permissions for each GCE preset (section 7.2).

To support submission of workflows to GCE, an administrator must install the Cloud Server Plugin and configure various settings via the *CLC Genomics Server* web administrative interface, as described in section 7.1.

### Launching jobs to run on a CLC Genomics Cloud Engine via a CLC Server

After the *CLC Server* has been set up to connect to GCE (section 7.1), the option "CLC Genomics Cloud Engine (via CLC Server)" becomes available to use in the workflow launch wizards in any *CLC Workbench* with the Cloud Plugin installed that are connected to the *CLC Server*.

No settings are required for the Workbench plugin. Submission will use the settings configured in the *CLC Server*.

When the "CLC Genomics Cloud Engine (via CLC Server)" option is selected, you then select a preset to use from the drop-down menu below it. Hover the mouse cursor over the preset name to see information about the machine specifications and result handling configured for that preset.

When submitting jobs to GCE this way, the *user logged into the CLC Server* is recorded as submitting the job in the *CLC Server* audit log. The credentials used when running the job on GCE are those configured in the *CLC Server* for accessing GCE. The user information in the history of data elements generated on GCE reflects this latter point.

### Importing data from Illumina Basespace via a CLC Server

When launching analyses to run on a *CLC Server* or to run on *GCE* via the *CLC Server*, you can select data from Illumina BaseSpace as input. For this, you need an Illumina BaseSpace account, but no configuration of the *CLC Server* or the Cloud Server Plugin is necessary.

### Importing data from and exporting data to Amazon S3 via a CLC Server

When launching analyses to run on a *CLC Server* or to run on *GCE* via the *CLC Server*, you can select data from an active Amazon S3 location that has been configured in the *CLC Server*. Documentation about configuring Amazon S3 locations is provided in the *CLC Server* manual, available from [https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/User\\_Manual.pdf](https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/User_Manual.pdf)

### Finding jobs and results submitted via the CLC Server using Cloud Job Search

The Cloud Job Search tool in a *CLC Workbench* with the Cloud Plugin installed can be used to find jobs and download results from analyses submitted to *CLC Genomics Cloud Engine* via a *CLC Server*. To find these jobs, the AWS locations, and the *GCE* settings specified in the *CLC Workbench* cloud configuration dialog, should either be empty, in which case the details of the configuration settings for the Cloud Server Plugin will be used, or must match those specified in the Cloud Server Plugin configuration in the *CLC Server*.

The "Only from logged-in user" option in the Cloud Job Search tool will find jobs submitted by both the user authenticated through the cloud configuration dialog and the user logged into the *CLC Server*. This is most noticeable if these user names differ.

The Cloud Job Search tool is described further in chapter 6.

## 7.1 Configuring the Cloud Server Plugin

To support submission of workflows to *GCE* via a *CLC Server*, take the steps detailed below in the *CLC Server* web administrative interface. You will need details about your *GCE* setup from your *GCE* administrator to complete these steps.

1. Configure **S3 locations** under the **External data** tab, in the **Configuration** area. Provide AWS credentials that allow access to a *GCE* cache bucket. Multiple AWS accounts can be configured, if desired.
2. Under **Direct data transfer from clients**, also under the **External data** tab, ensure direct data transfer from client systems is allowed. If you have not already done so, you will need to configure an Import/export directory for this.
3. Install the Cloud Server Plugin and restart the *CLC Server*.
4. Configure the Cloud Server Plugin settings (described below).
5. Configure *GCE* presets. These reflect the presets defined by the *GCE* administrator, which include information like machine size, and are specified when submitting jobs from client software. (section 7.2).

Related manual pages in the *CLC Genomics Server* administration manual:

- Configuring AWS S3 locations: [https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Other\\_external\\_data\\_access.htm](https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Other_external_data_access.htm)
- Installing server plugins: [https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Server\\_plugins.html](https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Server_plugins.html).
- Direct data transfer from client systems: [https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Direct\\_data\\_transfer\\_from\\_client\\_systems.html](https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Direct_data_transfer_from_client_systems.html)
- Setting up import/export directories: [https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Accessing\\_files\\_on\\_writing\\_to\\_areas\\_server\\_filesystem.html](https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Accessing_files_on_writing_to_areas_server_filesystem.html)

### Configuring Cloud Plugin settings in the CLC Server

1. Navigate to the **Extensions** tab in the *CLC Server* web administrative interface and click on the **Edit GCE connection settings** button.
2. Click on the **Edit** button next to Cloud Plugin in the window that appears (figure 7.1).

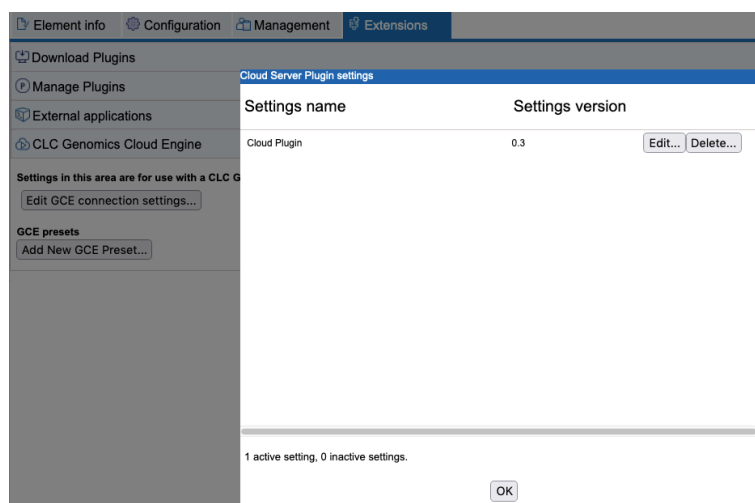


Figure 7.1: The GCE connection settings can be configured after clicking on the Edit button for the Cloud Plugin.

The settings to configure (figure 7.2) are listed below. Your GCE administrator should provide you with the necessary information. Brief notes are included to help finding the relevant information on AWS when logged into the account used to set up GCE. AWS frequently change aspects of their console, so details may vary from those described.

- **GCE S3 cloud cache bucket name:** The name of GCE cache bucket.  
GCE cache buckets are usually set up at the time when GCE was installed. An S3 bucket not configured as a GCE cache bucket should not be specified here.  
An error message is shown if the specified cannot be found. This could happen if, for example, credentials giving access to the relevant S3 location has not already been configured in the *CLC Server*, or the named bucket could not be found in an AWS account that has been configured.

- **GCE job manager rest host URI:** The URL for the GCE service.  
In the AWS Console, navigate to the Elastic Beanstalk Service. The URL is listed in the row for the Job Manager environment. Prepend this with the "https://" protocol when filling in this field.
- **GCE http oauth2 client id:** The client ID used by the *CLC Genomics Server* to authenticate using OAuth2 Client Credentials Grant.  
To get the client id, go to the Cognito service in the AWS Console, select the relevant Cognito user pool, and click on the "App Client" option. Find the "QIAGEN CLC Genomics Server" in the "App clients and applications section". Click on that app client. The value needed is under "Client ID".
- **GCE http oauth2 client secret:** The client secret used by the *CLC Genomics Server* to authenticate using OAuth2 Client Credentials Grant.  
See the notes in the point above for where to navigate to in the AWS Console. The information needed is under "Client secret". Enable "Show client secret" to reveal it.
- **GCE http oauth2 authorization server:** The access token endpoint of the authentication server used by the *CLC Genomics Cloud Engine*.  
In your Cognito user pool area in the AWS Console, click on the "Domain name" tab under "App integration". The base of the URL needed is provided under "Cognito domain." Append `/oauth2/token` to this. E.g. if the base URL was `https://gce.auth.eu-north-1-amazoncognito.com`, enter the following into this field:  
`https://gce.auth.eu-north-1.amazoncognito.com/oauth2/token`
- **Accept untrusted certificate:** Usually, production systems will have a trusted certificate, and this box is left unchecked. Check this box if the GCE service has been set up with a self-signed certificate.
- **Validate settings:** This option, enabled by default, validates the AWS and GCE settings when you press OK. Uncheck this box only if you need to temporarily store invalid settings.
- **GLOBAL\_OVERRIDABLE:** This setting has no effect for the Cloud Server Plugin and we recommend leaving it at the default settings.

3. One or more GCE presets should now be configured, as described in section 7.2.



Figure 7.2: Configuration of the Cloud Server Plugin settings

## 7.2 Configuring GCE presets

When submitting a job to *CLC Genomics Cloud Engine* via a *CLC Server*, the submitter will select a GCE preset. These presets are configured by the *CLC Server* administrator and are based on configurations in GCE<sup>1</sup>.

By default, presets are available for all users of the *CLC Server*. Access to a given preset can be restricted to specific groups using options available under the **Global permissions** tab in the *CLC Server* web administrative interface.

### Creating and editing GCE presets

To create or edit existing GCE presets, log into the *CLC Server* web administrative interface and go to:

**Extensions (  ) | CLC Genomics Cloud Engine (  ) | GCE presets**

Click on the **Add New GCE Preset...** button. The preset name is what the user of client software will use. The "GCE executor name" drop-down menu provides a list of the instance types defined by the *CLC Genomics Cloud Engine* administrator. Select one of these to associate with this preset. Optionally, provide a job tag.

The "Result handling" setting defines whether all results of jobs run using that preset should be automatically downloaded from AWS S3 to the *CLC Server*. Whether or not you specify that results should be automatically downloaded, results can be downloaded later using the Cloud Job Search tool in a *CLC Workbench*. Using that tool, results can also be selectively downloaded.

<sup>1</sup>The settings for GCE are configured in a file called `InstanceTypes.json`, described in the GCE administration manual.

## Chapter 8

# External applications in the cloud

Third party applications can be integrated into the CLC environment by configuring them as *external applications*. Containerized external applications can be used in workflows to be executed on the *CLC Genomics Cloud Engine* if the container is in the Amazon Elastic Container Repository (ECR) on the same Amazon account where the CLC Genomics Cloud Engine is deployed.

Information about configuring external applications, and exporting the configurations for use by the *CLC Genomics Cloud Engine* can be found in the **External applications** chapter of the *CLC Genomics Server* manual at [https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=External\\_applications.html](https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=External_applications.html)

### Creating and configuring workflows with external applications

Creating and editing workflows is done in the *CLC Workbench*, as described at <https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflows.html>.

Workflow elements will be available for:

- External applications available on a *CLC Genomics Server*, if the Workbench is connected to one.
- External applications described in the configuration file located on S3, where that location has been entered into the Workbench Preferences, under the Advanced tab, and the relevant AWS credentials have been configured, so you have access to the S3 bucket (see section 3).

External applications configurations are exported to S3 from a *CLC Genomics Server* by an administrative user, as described at [https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Import\\_export\\_external\\_application\\_configurations.html](https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Import_export_external_application_configurations.html)

Please ask your *CLC Genomics Server* administrator for the s3 URL to configure in the *CLC Workbench* Preferences. This is usually of particular relevance if you wish to submit workflows containing external applications to the *CLC Genomics Cloud Engine* without connecting to a *CLC Genomics Server*.

### Submitting workflows with external applications to run on GCE

Workflows can be submitted for analysis on the *CLC Genomics Cloud Engine* by a *CLC Workbench*, with or without a connection to a *CLC Genomics Server*, or using the *CLC Server Command Line Tools*.

**Submitting via a CLC Workbench without a connection to a CLC Server** The location of the exported external application configuration file in S3 must be configured in the Workbench Preferences. All the external applications included in the workflow must be described in the configuration file in S3.

**Submitting via a CLC Workbench with a connection to a CLC Server** The location of the exported external application configuration file in S3 is only needed if there are external applications in the workflow that are not present on the server, (but which are described in the configuration file on S3).

**Submitting via a CLC Server using the CLC Server Command Line Tools or a CLC Workbench** The external applications included in the workflow must be configured on the CLC Server.

## Chapter 9

# Troubleshooting

If a workflow execution fails while the *CLC Workbench* is still connected to the *CLC Genomics Cloud Engine*, the workflow process will be shown in red in the Processes area, with the message: "Errors occurred: see log". An error message will also be displayed, which may contain information about the cause of failure.

If further information is required about a failure *after* the workflow submission is completed, we recommend finding the batch using the Cloud Job Search functionality (see chapter 6), and downloading all results for the entire batch. This will download any available logs related to the failure. The exact files that will be downloaded may vary, depending on the cause of the failure. An example is shown in figure 9.1.

The type of files that may be downloaded include:

- **Workflow log** CLC data elements that can be opened within a *CLC Workbench*. These correspond to the logs produced when running a workflow within a *CLC Workbench* or on a *CLC Genomics Server*.
- **Result.json** Plain text files containing information about any outputs produced in AWS S3 for a particular job, including the exact paths to the outputs.
- **gce.log** Plain text files containing information registered by the *CLC Genomics Cloud Engine* about the workflow execution process, and technical details about any errors that have occurred.

We recommend that the Workflow log is inspected first, as it may contain useful information about simple causes of failures, such as corrupt input data or errors in the workflow configuration. For more complex cases, please email our Support team, attaching the files listed above ([ts-bioinformatics@qiagen.com](mailto:ts-bioinformatics@qiagen.com)).

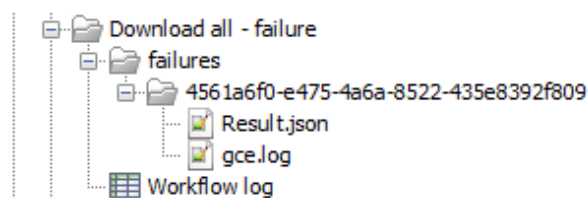


Figure 9.1: The "Download All Results" functionality in the Cloud Job Search tool allows you to download logs and other technical information if a workflow execution failed. The logs are downloaded to the Navigation Area. Some of them will be plain text files, which can be opened using any external text editing tool.


## Chapter 10

# Installing and uninstalling Workbench plugins

The following sections describe the installation and removal of plugins, including Cloud Plugin, on a *CLC Workbench*. For information about installing plugins on a *CLC Genomics Server*, please refer to [https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Server\\_plugins.html](https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Server_plugins.html) and the information in this manual about configuring the Cloud Server Plugin, provided in chapter ??.

### 10.1 Installation of Workbench plugins

**Note:** In order to install plugins and modules, the *CLC Workbench* must be run in administrator mode. On Windows, you can do this by right-clicking the program shortcut and choosing "Run as Administrator". On Linux and Mac, it means you must launch the program such that it is run by an administrative user.

Plugins and modules are installed and uninstalled using the Workbench Plugin Manager. To open the Plugin Manager, click on the **Plugins** (  ) button in the top Toolbar, or go to the menu option:

**Utilities | Manage Plugins...** (  )

The Plugin Manager has two tabs at the top:

- **Manage Plugins** An overview of your installed plugins and modules is provided under this tab.
- **Download Plugins** Plugins and modules available to download and install are listed in this tab.

To install a plugin, click on the **Download Plugins** tab (figure 10.1). Select a plugin. Information about it will be shown in the right hand panel. Click on the **Download and Install** button to install the plugin.

#### Accepting the license agreement

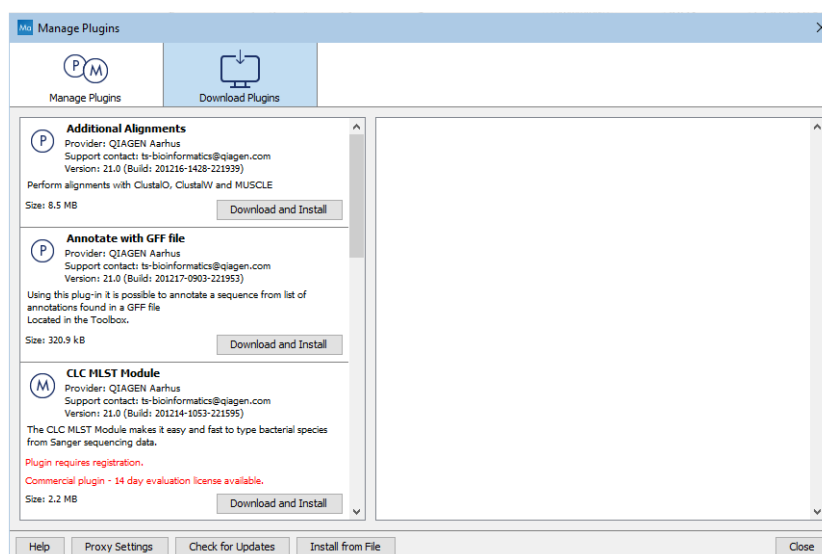


Figure 10.1: Plugins and modules available for installation are listed in the Plugin Manager under the Download Plugins tab.

The End User License Agreement (EULA) must be read and accepted as part of the installation process. Please read the EULA text carefully, and if you agree to it, check the box next to the text **I accept these terms**. If further information is requested from you, please fill this in before clicking on the **Finish** button.

### Installing a cpa file

If you have a .cpa installer file for Cloud Plugin, you can install it by clicking on the **Install from File** button at the bottom of the Plugin Manager.

If you are working on a system not connected to the internet, plugin and module .cpa files can be downloaded from <https://digitalinsights.qiagen.com/products-overview/plugins/> using a networked machine, and then transferred to the non-networked machine for installation.

### Restart to complete the installation

Newly installed plugins and modules will be available for use after restarting the software. When you close the Plugin Manager, a dialog appears offering the opportunity to restart the CLC Workbench.

## 10.2 Uninstalling Workbench plugins

Plugins and modules are uninstalled using the Workbench Plugin Manager. To open the Plugin Manager, click on the **Plugins (P)** button in the top Toolbar, or go to the menu option:

**Utilities | Manage Plugins... (P)**

This will open the Plugin Manager (figure 10.2). Installed plugins and modules are shown under the Manage Plugins tab of the Plugins Manager.

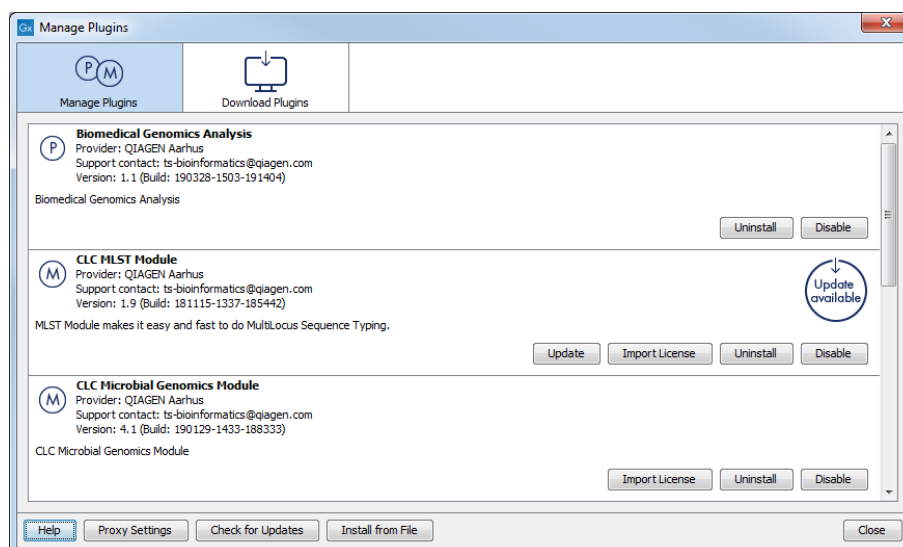


Figure 10.2: Installed plugins and modules are listed in the Plugins Manager under the Manage Plugins tab.

To uninstall a plugin or module, click on its entry in the list, and click on the **Uninstall** button.

Plugins and modules are not uninstalled until the Workbench is restarted. When you close the Plugin Manager, a dialog appears offering the opportunity to restart the *CLC Workbench*.

### Disabling a plugin without uninstalling it

If you do not want a plugin to be loaded the next time you start the Workbench, select it in the list under the Manage Plugins tab and click on the **Disable** button.