User manual for
QIAGEN CLC Cloud Module 23.0.1
Windows, macOS and Linux

October 5, 2023

This software is for research purposes only.
Contents

1 Introduction to the CLC Genomics Cloud .......................... 5
   1.1 Overview of CLC Genomics Cloud ................................... 5
   1.2 CLC Genomics Cloud requirements .................................. 6
   1.3 Contact information ......................................................... 7
   1.4 Installing the CLC Cloud Module ...................................... 8
       1.4.1 Licensing modules ....................................................... 9
       1.4.2 Uninstalling modules .................................................... 10
   1.5 Installing the Cloud Server Plugin ................................. 11

2 Setting up CLC Genomics Cloud resources on AWS ............... 13
   2.1 Setting up AWS resources ............................................... 13
   2.2 Adding more AWS Batch queues for CLC jobs ....................... 15
   2.3 Adding more AWS IAM users .............................................. 16
   2.4 Deleting or redeploying CLC Genomics Cloud infrastructure on AWS .... 16

3 Using a CLC Workbench to work on the cloud ...................... 18
   3.1 Prerequisites for working on the cloud using a Workbench ........ 18
   3.2 Configuring an AWS Connection in the Workbench ................. 19
   3.3 Submitting analyses to AWS from a Workbench ..................... 20
   3.4 Working with AWS S3 using the Remote Files tab .................. 21
   3.5 Accessing results from the Processes tab ........................... 25

4 Using a CLC Server to work on the cloud ........................... 26
   4.1 Benefits of submitting jobs via a CLC Server ....................... 26
   4.2 AWS Connections in the CLC Server .................................. 27
   4.3 Configuring cloud presets ............................................... 28
Chapter 1

Introduction to the CLC Genomics Cloud

This manual describes the CLC Genomics Cloud solution, including establishing the necessary infrastructure on AWS, installation and use of QIAGEN CLC Cloud Module 23.0.1 and Cloud Server Plugin 23.0.1, and information about submitting CLC jobs to the cloud for execution.

CLC Genomics Cloud setups require CLC software version 23 or above.

1.1 Overview of CLC Genomics Cloud

With access to a CLC Genomics Cloud setup, analyses can be submitted from a CLC Workbench or a CLC Server to run on AWS EC2 instances, with results saved to AWS S3. Access to AWS S3 is built into CLC software, making it simple to upload data to and download data from the cloud. Data to be analyzed can be local or on AWS S3.

![Diagram of CLC Genomics Cloud setup]

Figure 1.1: The CLC Genomics Cloud is a combination of CLC software and AWS resources.

With access to a CLC Server there is substantial potential for extending the functionality of CLC software using External Applications, whereby non-CLC tools can be launched using CLC software. In combination with a CLC Genomics Cloud setup, external applications can be launched directly
from a CLC Workbench, via the Toolbox or in a workflow context, to run on the cloud. See chapter 5 for details.

All traffic to and from AWS is encrypted using a minimum of TLS version 1.2. AWS credentials provided in a CLC Workbench are stored, obfuscated, within user configuration files. Those provided in a CLC Server are stored, obfuscated, in the server configuration files. All EC2 instances run in a security group that rejects incoming traffic that was not initiated from the EC2 host.

A CLC Genomics Cloud setup minimally consists of:

1. An AWS account with the relevant resources configured.
2. A CLC Workbench with an AWS Connection configured for access to that AWS account.
3. The QIAGEN CLC Cloud Module installed on the CLC Workbench.

Benefits of using a CLC Server in addition to the above are described in chapter 4.

**CLC Genomics Cloud costs**

The cost of a CLC Genomics Cloud setup consists of the costs of CLC software licenses and the costs of the AWS resources used (EC2 instances used for running analyses, AWS S3 used for data storage). There are no costs associated with setting up the CLC Genomics Cloud environment on AWS and there are no idle costs.

To submit analyses to run on AWS EC2 instances, you need the following CLC licenses:

- A license for a CLC Workbench. Jobs can also be submitted via a licensed CLC Server.
- A license for the CLC Cloud Module.

Pricing relevant to the AWS resources used by a CLC Genomics Cloud can be found at:

- [AWS EC2 On-demand instance pricing](https://aws.amazon.com/ec2/pricing/on-demand/)
- [AWS S3 pricing](https://aws.amazon.com/s3/pricing/)

You are billed directly by AWS for use of their resources.

**1.2 CLC Genomics Cloud requirements**

The requirements for a fully functional CLC Genomics Cloud setup are:

- Access to an AWS account where the relevant infrastructure is available. Establishing the necessary infrastructure on AWS requires administrative privileges (see chapter 2).
• A licensed **CLC Workbench**, with the CLC Cloud Module installed and licensed.
  Installation of the CLC Cloud Module is described in section 8.
  If the CLC Cloud Module is installed but not licensed, or you are working with a Workbench in Viewing Mode, you can interact with files on AWS S3, but analyses cannot be submitted to run on AWS.

• If analyses will be submitted via a **CLC Server**, the Cloud Server Plugin must be installed and cloud presets configured (see chapter 4).

The versions of your CLC software must be 23.0 or above.

General system requirements for CLC software are provided on https://digitalinsights.qiagen.com/technical-support/system-requirements/


### 1.3 Contact information

QIAGEN CLC Cloud Module is developed by:

QIAGEN Aarhus
Silkeborgvej 2
Prismet
8000 Aarhus C
Denmark

https://digitalinsights.qiagen.com/

Email: ts-bioinformatics@qiagen.com

The QIAGEN Aarhus team continuously improves products with your interests in mind. We welcome feedback and suggestions for new features or improvements. How to contact us is described at: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Contact_information_citation.html.

You can also make use of our online documentation resources, including:

• **Core product manuals** https://digitalinsights.qiagen.com/technical-support/manuals/

• **Plugin manuals** https://digitalinsights.qiagen.com/products-overview/plugins/

• **Tutorials** https://digitalinsights.qiagen.com/support/tutorials/

• **Frequently Asked Questions** https://qiagen.my.salesforce-sites.com/KnowledgeBase/KnowledgeNavigatorPage
1.4 Installing the CLC Cloud Module

**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 1.2). 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.

![Figure 1.2: Plugins and modules available for installation are listed in the Plugin Manager under the Download Plugins tab.](image)

**Accepting the license agreement**

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 QIAGEN CLC Cloud Module, 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.

### 1.4.1 Licensing modules

When you have installed the QIAGEN CLC Cloud Module and start a tool from that module for the first time, the License Assistant will open (figure 1.3).

The License Assistant can also be launched by opening the Workbench Plugin Manager, selecting the installed module from under the Manage Plugins tab, and clicking on the button labeled Import License.

To install a license, 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.

![Figure 1.3](image)

Figure 1.3: The License Assistant provides options for licensing modules installed on the Workbench.

The following options are available:

- **Request an evaluation license.** Request a fully functional, time-limited license.
- **Download a license.** Use the license order ID received when you purchased the software to download and install a license file.
- **Import a license from a file.** Import an existing license file, for example a file downloaded from the web-based licensing system.
• **Configure License Server connection.** If your organization has a CLC Network License Manager (or CLC License Server), select this option to configure the connection to it.


To download licenses, including evaluation licenses, your machine must have access to the external network. To install licenses on non-networked machines, please see [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Download_static_license_on_non_networked_machine.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Download_static_license_on_non_networked_machine.html).

### 1.4.2 Uninstalling modules

Plugins and modules are 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...**

This will open the Plugin Manager (figure 1.4). Installed plugins and modules are shown under the Manage Plugins tab of the Plugins Manager.

![Plugin Manager](image)

**Figure 1.4:** 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.
1.5 Installing the Cloud Server Plugin

To submit jobs to run on the cloud via a CLC Server, the Cloud Server Plugin needs to be installed.

**Download and install server plugins and server extensions**

Plugins, including server extensions (commercial plugins), are installed by going to the Extensions tab in the web administrative interface of the single server, or the master node of a job node or grid node setup, and opening the Download Plugins area (figure 1.5).

![Figure 1.5: Installing plugins and server extensions is done in the Download Plugins area under the Extensions tab.](image)

If the machine has access to the external network, plugins can be both downloaded and installed via the CLC Server administrative interface. To do this, locate the plugin in the list under the Download Plugins area and click on the Download and Install... button.

To download and install multiple plugins at once on a networked machine, check the "Select for download and install" box beside each relevant plugin, and then click on the Download and Install All... button.

If you are working on a machine without access to the external network, server plugin (.cpa) files can be downloaded from: https://digitalinsights.qiagen.com/products-overview/plugins/ and installed by browsing for the downloaded file and clicking on the Install from File... button.

The CLC Server must be restarted to complete the installation or removal of plugins and server extensions. All jobs still in the queue at the time the server is shut down will be dropped and would need to be resubmitted. To minimize the impact on users, the server can be put into Maintenance Mode. In brief: running in Maintenance Mode allows current jobs to run, but no new jobs to be submitted, and users cannot log in. The CLC Server can then be restarted when desired. Each time you install or remove a plugin, you will be offered the opportunity to enter Maintenance Mode. You will also be offered the option to restart the CLC Server. If you choose not to restart when prompted, you can restart later using the option under the Server maintenance tab.
Managing installed server plugins

Installed plugins are managed, and can be uninstalled, from under the Manage Plugins area (figure 1.6), under the Extensions tab.

The list of tools delivered with a server plugin can be seen by clicking on the Plugin contents link to expand that section. Workflows delivered with a server plugin are not shown in this listing.

Figure 1.6: Managing installed plugins and server extensions is done in the Manage Plugins area under the Extensions tab. Clicking on Plugin contents opens a list of the tools delivered by the plugin.

Plugin compatibility with the server software

The version of plugins and server extensions installed must be compatible with the version of the CLC Server being run. A message is written under an installed plugin’s name if it is not compatible with the version of the CLC Server software running.

When upgrading to a new major version of the CLC Server, all plugins will need to be updated. This means removing the old version and installing a new version.

Incompatibilities can also arise when updating to a new bug fix or minor feature release of the CLC Server. We recommend opening the Manage Plugins area after any server software upgrade to check for messages about the installed plugins.
Chapter 2

Setting up CLC Genomics Cloud resources on AWS

2.1 Setting up AWS resources

We recommend using the CloudFormation template described below to set up the standard resources needed for a CLC Genomics Cloud setup. Some configuration can be done when using this template. Additional AWS Batch queues can be added using another CloudFormation template afterwards. That template provides more configuration options (see section 2.2).

Overview of standard CLC Genomics Cloud infrastructure on AWS

The cgc-standard.json CloudFormation template defines the resources needed for a CLC Genomics Cloud, which include:

- **Three AWS Batch queues** The queues are named cgc-small, cgc-medium, and cgc-large. When a user launches a job to run on the cloud from their CLC Workbench, they select a queue to send the job to from a drop down list of the available queues. Details of each of these queues are provided in section 8.1.

- **An S3 bucket for system files** The name of this bucket begins with "cgc-system-". This bucket is used by the CLC software for system files, including read mapper indexes. It is not intended for storing sample data or results. It is not listed when browsing using CLC software. Details about the retention policies on this bucket are provided in section 8.3.

  **Note:** S3 buckets for holding input data and results need to be created directly in AWS. They are not created by CloudFormation templates provided by QIAGEN.

- **Two AWS IAM users** One with properties supporting submission of analyses, and the other allowing only access to AWS S3 buckets. This is described in more detail below.

  When working with a CLC Workbench, access to the AWS Batch queues is determined by the access rights of the AWS IAM user configured in the AWS connection. With a CLC Server, access to AWS Batch queues can be fine tuned by setting group permissions on cloud presets, using the web administrative interface of the CLC Server, as described in section 4.3.
Creating CLC Genomics Cloud infrastructure on AWS

To set up the standard infrastructure on AWS for handling CLC jobs:

- Log into the AWS console as a user with privileges that allow the infrastructure described above to be created.
- Set the region to the one the AWS resources should be established in.
- Copy the URL below:
- Go to CloudFormation and click on Create stack.
- Review the max vCPUs setting for each queue.
  By default, the values allow for up to 10 EC2 instances to be launched. I.e. up to 10 jobs can be run in parallel. This value can be increased or decreased. We recommend it is not decreased below the number of cores designated for each job. These values can be found in the setting details provided in section 8.1.
- Review the Disk size setting for each queue.
  This value refers to the space available for each job when it is running.
- Specify the URL for custom plugin repository if you have custom CLC plugins providing tools that should run on the cloud.
  See section 8.4 for the requirements for a custom plugin location.
  This field should be left blank if you require access only to plugins and modules distributed by QIAGEN. These are the plugins and modules listed in CLC Workbench and CLC Server Plugin Manager, or on the QIAGEN website:
  https://digitalinsights.qiagen.com/products-overview/plugins/

AWS IAM users

When stack creation is complete, go to the Outputs tab of the main stack to find the credentials for the AWS IAM users created (figure 2.1).

AWS Connections using the "SubmitterUser" (CgcSubmitterUser-<EnvironmentId>) credentials allow CLC analyses to be submitted an AWS Batch queue for analysis. This user also has full access to AWS S3, read access to CloudWatch logs, and can list CloudFormation resources.

AWS Connections using the "BrowserUser" (CgcBrowserUser-<EnvironmentId>) credentials support listing S3 buckets and accessing bucket contents. Jobs cannot be submitted to run on AWS using these credentials.

AWS IAM user credentials are entered in AWS Connections in CLC Workbenches or the CLC Server, described in section 3.2 and section 4.2, respectively.

The full policy for each user can be viewed in the Identity and Access Management (IAM) area of the AWS Console.
AWS S3 buckets for storing input data and results

One or more AWS S3 buckets must be created for holding input data and results. These buckets must be created in the same AWS account and region that the AWS Batch queues were established in. Please refer to AWS documentation for details:


2.2 Adding more AWS Batch queues for CLC jobs

To add another AWS Batch queue for running CLC analyses using on-demand EC2 instances:

- Log into the AWS console as a user with privileges that will allow the infrastructure described above to be created.
- Set the region to the one the AWS resources should be established in.
- Copy the URL below:
- Go to CloudFormation and click on Create stack.
- Configure the settings for the new queue. This includes:
  - **vCPUs per job** We recommend setting this to the number of vCPUs available on the instance types you specify. If there are different numbers of vCPUs on the instance types specified, provide the lowest value.
  - **Total available memory per job** We recommend that this is set to the available amount of memory on the instance types you specify, minus 1000 MB. So for instance types with 32 GB of memory, the recommended value would be 31000.
- **Amount of memory allocated for Java per job** We recommend that this is set to approximately half the total memory, up to a maximum of 50GB. So, with 31000 MB of total memory allocated, a reasonable value here would be 15500 MB. With 128000 MB total memory, a reasonable value would be 50000MB.

- **max vCPUs** This value controls the number of EC2 instances that can be started up in parallel in this queue. For example, if the vCPUs per job were set to 16, then the value of 160 here would mean a maximum of 10 EC2 instances could be running at the same time in this queue.

- **Allowed instance types** One or more EC2 instance types that can be started up when jobs are sent to this queue. Linux 64-bit (x86_64 / AMD64) instance types are supported. See the AWS documentation for details of the types available, [https://aws.amazon.com/ec2/instance-types/](https://aws.amazon.com/ec2/instance-types/) and their costs, [https://aws.amazon.com/ec2/pricing/on-demand/](https://aws.amazon.com/ec2/pricing/on-demand/).

- **Disk size** Specify the size of the disk to allocate to each EC2 instance.

AWS Batch queues made using this template can be used by the AWS IAM "SubmitterUser" created using the standard template (see section 2.1).

When an AWS Batch queue is created using the CloudFormation templates provided by QIAGEN, it is mapped to an individual compute environment configured with the latest Amazon ECS-optimized AMI available at that time. Thus, the AMI version used when jobs are sent to queues established at different times may be different. See appendix section 8.2 for further details.

### 2.3 Adding more AWS IAM users

To create additional AWS IAM users with relevant permissions, choose the option to "Copy permissions from existing user" when creating the IAM user, and select one of the users created using the standard CloudFormation template as the user to copy.

Please refer to AWS documentation for further details:


### 2.4 Deleting or redeploying CLC Genomics Cloud infrastructure on AWS

**Deleting an AWS Batch queue**

To delete a particular AWS Batch queue, go to

AWS Batch | Job queues

Select the queue you wish to delete and click on the Disable button.

This action may take a while. Once complete, select the queue again, and click on the Delete button.

**Deleting CLC Genomics Cloud infrastructure**
if you have run only the cgc-standard.json CloudFormation template to create your CGC Environment (and have not run other CloudFormation templates), then deleting the existing CGC infrastructure is usually a three step process:

1. **Empty the contents of the cgc-system bucket** Go to the S3 area of the AWS Console, select the cgc-system bucket and click on the **Empty** button. Non-empty system buckets cannot be deleted as part of the stack deletion.

2. **Delete the stack** Go to the CloudFormation area of the AWS Console, select the main stack, and click on the **Delete** button.

3. Delete old job definitions If any job was run using the older CGC resources, delete the old job definitions. Go to **AWS Batch | Job definitions**, select the job definition files with names that start with cwe and click on **Deregister**.

If stacks for additional job queues have been created, delete these stacks before carrying out the steps above. These additional stacks are linked as dependencies to the Infrastructure stack, which is a nested part of the CGC Environment.

**Redeploying CLC Genomics Cloud Infrastructure**

To redeploy CGC infrastructure on AWS:

1. **Delete the existing CGC resources** Please refer to the paragraph above.

2. **Deploy new CGC resources** Follow the instructions provided earlier in this chapter.

New user credentials will need to entered for AWS Connections in CLC Workbenches and CLC Servers. Cloud presets on CLC Servers will also need to be configured.
Chapter 3

Using a CLC Workbench to work on the cloud

This chapter describes configuring and using a CLC Workbench to submit jobs for execution on AWS, as well as working with data on AWS S3, including uploading data to AWS S3.

3.1 Prerequisites for working on the cloud using a Workbench

Submitting analyses to run on AWS

To submit jobs to run on AWS from a CLC Workbench you need:

- A licensed CLC Workbench.
- The CLC Cloud Module installed on the Workbench and licensed (see chapter 1.4).
- An AWS Connection configured with AWS IAM user credentials for a user with permission to submit analyses to run on AWS (see section 3.2).

An option called CLC Genomics Cloud is enabled in launch wizards when the setup is ready (see section 3.3).

Working with AWS S3 only

If you only wish to browse and access data stored on AWS S3 from your CLC Workbench you need:

- A CLC Workbench with the CLC Cloud Module installed. Licenses for the Workbench and module are not required (see chapter 1.4).
- An AWS Connection configured with AWS IAM user credentials with appropriate permissions to work with AWS S3 buckets in your account. These could be the same user credentials used for submitting jobs to run on AWS, or an IAM user with permissions limited to working with files on AWS S3 (see section 3.2).
3.2 Configuring an AWS Connection in the Workbench

To configure an AWS connection, go to:

Connections | AWS Connections

After a connection has been configured, it will look like that in figure 3.1, where AWS connections are listed, along with information about their status. These can be edited or removed, if desired. The status is indicated using colors. Green indicates the connection is valid and ready for use.

Connections to a CLC Genomics Cloud are indicated in the CGC column. To submit analyses to the CLC Genomics Cloud, the CLC Cloud Module must be installed and a license for that module must be available.

![Configuration dialog for AWS connections.](image)

Click on the Add AWS Connection button to configure an AWS connection. Enter the following information (figure 3.2):

- **Connection 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**: A description of the AWS account (optional).
- **AWS access key ID**: The access key ID for programmatic access for your AWS IAM user.
- **AWS secret access key**: The secret access key for programmatic access for your AWS IAM user.
- **AWS region**: An AWS region. Select from the drop-down list.
- **AWS partition**: The AWS partition for your account.

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

AWS credentials entered are stored, obfuscated, in Workbench user configuration files.
3.3 Submitting analyses to AWS from a Workbench

Once the prerequisites described in section 3 are in place, the CLC Genomics Cloud option will be enabled in the launch wizards of CLC workflows and tools\(^1\). The available AWS Batch queues are listed in a drop-down menu (figure 3.3). After selecting a queue, mouse over its name to see information about the settings for that queue (figure 3.4).

\(^1\)Only tools that can be used within workflows can be submitted to run on a CLC Genomics Cloud. Tools that are not workflow-enabled cannot be run on the cloud.
transfer the data to AWS S3, so this phase can take some time. See section 6.1 for further details.

### 3.4 Working with AWS S3 using the Remote Files tab

#### Browsing data in AWS S3 using the Remote Files tab

The S3 buckets available via the configured **AWS Connections** are listed under the **Remote Files** tab, to the right of the Navigation Area tab.

![Remote Files tab](image)

Figure 3.5: AWS S3 buckets you have access to are available under the Remote Files tab.

If more than one AWS Connection is available, for example, if the **CLC Workbench** is connected to a **CLC Server** with a valid AWS Connection, a drop-down menu appears above the bucket listing, allowing you to select the AWS Connection to use.

#### Uploading data to AWS S3 using the Remote Files tab

To upload data from your Navigation Area to AWS S3, right-click on a folder in the Remote Files tab and choose the option **Upload to this folder** (figure 3.7).
Figure 3.6: This Workbench has a valid AWS Connection and is connected to a CLC Server with a valid AWS Connection. A drop-down menu allows selection of the connection to use. The S3 buckets available using the credentials in that AWS Connection are then listed in the Remote Files tab.

Figure 3.7: To upload data from your Navigation Area to AWS S3, open the Remote Files tab, right-click on the folder you wish to upload data to, and select the option "Upload to This Folder". Upload is sequential. Information about the data upload is shown in the Processes tab, at the bottom left of the Workbench (figure 3.8).

Download results via the Remote Files tab

Right-click on an element or elements under the Remote Files tab to download from AWS S3 (figure 3.9).

To see all the outputs of a particular job, double-click on a workflow-result.json file. All the results can then be downloaded and opened from that list, or individual elements can be selected and downloaded. The Execution Log is also available from this list (see figure 3.10).

AWS charges for downloaded data from S3. See section 6.2 for further information about downloading data from AWS S3.
CHAPTER 3. USING A CLC WORKBENCH TO WORK ON THE CLOUD

Figure 3.8: After choosing to upload data to S3, the progress of the upload is reported in the Processes tab.

Figure 3.9: Right click on a file or files in the Remote Files tab and choose to Download and Open or Download and Save those elements.
Figure 3.10: Double-click on a workflow-result.json file in the Remote Files tab in the Workbench to reveal a list of all results from that job, as well as the Execution Log. All items can be downloaded and opened from this menu, or individual items can be selected and downloaded.
3.5 Accessing results from the Processes tab

Analysis results can be accessed via the menu available via the small triangle to the right of the process in the Processes tab (figure 3.11). Under the Download and Open Results option, you can select individual results to download and open, or you can download and open all results. When closing an element opened this way, you will be prompted for whether to save it. If it is not saved, it will need to be downloaded from AWS S3 again to be viewed.

Figure 3.11: Select "Download and Open Results" from the menu for a process to see a list of all the results as well as the Execution Log.

This same list is available via the Remote Files tab by double-clicking on a workflow-result.json file.

When you select Find Results from under the Processes tab, the Remote Files tab will be opened, with the relevant folder opened and the outputs highlighted.

**Note:** When launching a job in Batch mode, a process labeled "Preparing N batch units" will be present, with N being the number of batch units. This process records the preparation of the individual jobs in the batch.

The Processes tab is described in more detail in section 7.1.

AWS charges for downloaded data from S3. See section 6.2 for further information about downloading data from AWS S3.
Chapter 4

Using a CLC Server to work on the cloud

Jobs can be submitted to the CLC Genomics Cloud via a CLC Genomics Server that has the Cloud Server Plugin installed and valid cloud presets configured.

In this chapter, we cover:

- Benefits of submitting to the cloud via a CLC Server (section 4.1)
- Cloud preset configuration (section 4.3)

Installing server plugins is described in section 1.5.

4.1 Benefits of submitting jobs via a CLC Server

Working with a CLC Server brings several benefits when interacting with a CLC Genomics Cloud setup, including:

- **Audit trail** Jobs submitted to the cloud are logged, just as other jobs sent to the server. The audit log contains details of the job such as who submitted it, to which AWS Batch queue, whether the job succeeded and failed, etc.

- **User management** A single set of AWS IAM credentials can be entered in an AWS Connection in the CLC Server, instead of configuring credentials in each individual CLC Workbench. All jobs sent to the cloud will use those AWS credentials. The CLC Server user that submitted a job is recorded in the server audit log. CLC Server user management supports connections to LDAP or Active Directory, as described at https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Authentication_options.html.

- **Access control** Access to an AWS Connection can be limited to certain groups, allowing fine control over access to particular accounts or resources. In addition, access to AWS Batch queues is specified in CLC Genomics Cloud presets, configured in the CLC Server web administrative interface. Access to cloud presets can also be limited to specified groups.

- **Extend functionality using external applications** Non-CLC tools can be launched using CLC software by configuring them as external applications. Following the steps described in
chapter 5, external applications can be launched directly from a CLC Workbench, via the Toolbox or in a workflow context, to run on the cloud.

- **Close the CLC Workbench after job or data transfer has started** When launching analyses on the cloud using data from a CLC Server import/export area, you close down your CLC Workbench, or disconnect from the CLC Server, and transfers to S3 associated with the job (e.g. the workflow, parameter settings and any data held on the CLC Server or in import/export directories) will continue to be transferred. If submitting directly from a CLC Workbench, the process must proceed beyond data transfer steps before it can be closed.

To submit jobs to a CLC Genomics Cloud setup via a CLC Server, choose the launch wizard option "CLC Genomics Cloud via CLC Server" (figure 4.1).

![Figure 4.1: When a CLC Workbench with a licensed CLC Cloud Module installed is connected to a CLC Server configured for access to a CLC Genomics Cloud, the option to submit jobs to the cloud via the server is enabled.](image)

### 4.2 AWS Connections in the CLC Server

Access to AWS accounts is configured under:

**Configuration | External data | AWS Connections**

Click on the Add AWS Connection button. The following information should be entered in the configuration dialog:

- **Connection 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.

- **AWS access key ID**: The access key ID for programmatic access for your AWS IAM user.

- **AWS access secret key**: The secret access key for programmatic access for your AWS IAM user.

- **AWS Region**: An AWS region. Select from the drop-down list.

- **AWS Partition**: The AWS partition for your account.
AWS connections are used when:

- Accessing AWS S3 locations, to import data from or export data to.
- Submitting analyses to a CLC Genomics Cloud setup, if available on that AWS account.

All traffic to and from AWS is encrypted using a minimum of TLS version 1.2. AWS credentials entered are stored, obfuscated, in the server configuration files.

AWS connections are listed, along with information about their status. These can be edited or deleted. The status is indicated using colors. Green indicates the connection is valid and is ready for use.

Connections to a CLC Genomics Cloud, indicated in the CGC column, require the Cloud Server Plugin to be installed, and access to a CLC Genomics Cloud.

![AWS Connections](image)

**Figure 4.2:** Configure AWS Connections. A green dot in the S3 column indicates a valid connection for accessing S3 locations. A green dot in the CGC column indicates a valid connection to a CLC Genomics Cloud, meaning analyses can be submitted to run on AWS via this CLC Server.

Access to AWS connections can be limited to specified groups using options available under the Global permissions tab in the CLC Server web administrative interface.

Clicking on the Browse S3 locations link opens the relevant tab under Element info.

### 4.3 Configuring cloud presets

When submitting a job to a CLC Genomics Cloud setup via a CLC Server, the submitter specifies a cloud preset to use. Each preset is configured with the name of a single AWS Batch queue. When a given preset is selected, jobs are sent to the corresponding AWS Batch queue.

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

**Creating and editing CLC Genomics Cloud presets**

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

**Extensions (تكوين) | CLC Genomics Cloud ( rộng)**

Click on the Add Preset... button.
The preset name is what the user of client software specifies when launching their analysis. The AWS Batch queue is where jobs will be submitted to when this preset is selected. AWS Batch queue settings are configurable within AWS.

An AWS Connection must have been configured before cloud presets can be configured.

### 4.4 Browse AWS S3 locations

In the **Browse S3 locations** area, you can browse data held in the S3 buckets you have access to via the AWS Connections configured in the **External data** area, under the Configuration tab. When an item is selected, options are enabled to download or delete it. When a folder is selected, options are enabled to upload to that folder, download from that folder, create a new folder or to delete the folder and its contents (figure 4.3). The URLs provided at the bottom are particularly useful if you are using the **CLC Server Command Line Tools**.

If you choose to download, a window pops up listing the selected files and providing the total download size (figure 4.4). You can download to a CLC location or to a CLC Server import/export area. When CLC data has been downloaded to a CLC location, it can be opened and inspected using a CLC Workbench connected to the CLC Server, or used for further analysis. For other file types, a CLC Server import/export directory would usually be selected.

**Figure 4.3:** Browse S3 locations accessible via AWS Connections configured in the CLC Server. Data can be uploaded, downloaded, deleted and new folders created using functionality in this tab.

**Note:** AWS charges for downloading data from AWS S3.

If the message "No active S3 locations found." is visible, it means that either no AWS Connections have been configured, or that there are no AWS Connections valid for the user logged into the CLC Server.

**Note:** If you are using a CLC Genomics Cloud on AWS, non-admin users may see one or more entries called **CGC system bucket**. These cannot be browsed. They are used for system files, and are not intended for storing sample data or results. When logged in as an admin user, these buckets are listed with their full names, which start with **cgc-system**. They can be browsed but there is generally no need to do so.
Figure 4.4: When downloading data from S3, a list of the selected files and the total size of the download is displayed. After clicking on the relevant download button, you will be prompted for a specific location to download to.

## 4.5 Submitting workflows using the CLC Server Command Line Tools

### Submitting workflows to the cloud


The cloud preset to use is specified using the `-L` option. To see the list of cloud presets available, run the `clcserver` command with no arguments or with an incomplete set of arguments, as described on [http://resources.qiagenbioinformatics.com/manuals/clcservercommandlinetools/current/index.php?manual=Basic_usage.html](http://resources.qiagenbioinformatics.com/manuals/clcservercommandlinetools/current/index.php?manual=Basic_usage.html).

### Specifying inputs and the destination for results

Inputs for analyses to be run on the cloud can be in CLC Server file locations or in remote locations. Input data files in an AWS S3 bucket are specified using an s3 URL.

Output destinations should be specified using an s3 URL.

Data in CLC Server file locations and CLC format files in remote locations can be specified directly as input. That is, on-the-fly import is not needed in these cases.

Data in supported formats other than CLC format require the use of on-the-fly import. The precise command options depend on the importer to be used. To reveal the importers available for use with a given workflow, run the `clcserver` command with the name of the workflow specified using the `-A` option. For example,

---

1. Support of URLs to remote locations without specifying the location as a clccloudfile was introduced in CLC Genomics Server 23.0.3.
2. Support for supplying CLC format files in remote locations directly as input to workflows to be executed on the cloud was introduced in CLC Genomics Server 23.0.3 and Cloud Server Plugin 23.0.1. With earlier versions, CLC format data in remote locations, like other data formats, had to be specified using on-the-fly import.
One of the options returned will usually relate to on-the-fly import. E.g. for some template workflows, this would be the `--reads-import-command`, with a list of the available importers. An example of one of these is `ngs_import_illumina`. Running an incomplete command of this form would then reveal the on-the-fly import options relevant for that importer:

```
clcserver -S <server> -U <username> -W <password or token> -A <workflow name> --reads-import-command ngs_import_illumina
```

**Accessing AWS CloudWatch logs via the command line**

The **CLC Server Command Line Tools** command `-A cgc_read_aws_logs` supports the retrieval of AWS CloudWatch logs for jobs run on a **CLC Genomics Cloud**.

The messages returned from jobs run on the cloud include the information needed to access the AWS CloudWatch log for that job. The AWS CloudWatch information retrieved is the same as that returned when the "Execution Log" is opened in the CLC Workbench, either via the Processes tab or via options under the Remote Files tab, as described in section or chapter 3.5.
Chapter 5

Integrating third party tools into CLC software

Third party applications can be integrated with CLC software by configuring them as *external applications* using a *CLC Server*. Workflows containing external applications can be submitted for execution on the cloud, like any other workflow.


By default, only administrative users can configure external applications, but permissions can be granted more broadly, as described under "Web admin access" on [https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Controlling_access_server_server_tasks_external_data.html](https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Controlling_access_server_server_tasks_external_data.html)

Resources referred to in an external application must be available when it is run on the cloud. For containerized external applications that can be successfully run on a *CLC Server*, this requirement will generally be met. For standard external applications, the tools referred to in the configuration, including any referred to in an included script in that configuration, need to be available in the Linux image used when the job is run on AWS. See appendix section 8.2 for information about the compute resources in a *CLC Genomics Cloud* setup.

**Note:** When running a workflow containing external applications available on a *CLC Server*, there must be a connection to that *CLC Server* when launching the workflow. This can mean using a *CLC Workbench* connected to the *CLC Server*, or submitting workflows installed on the *CLC Server* using the *CLC Server Command Line Tools*. 

---

32
Chapter 6

General information about inputs and results

This chapter covers general considerations about the location of input data, including reference data, and working with results of CLC analyses run on the cloud.

6.1 General information about input data for cloud analyses

Cloud-based analyses use input data in AWS S3. Data not already in S3 is transferred from the local system to the cloud at the start of an analysis job. Thus, it is often most efficient to use data already in AWS S3 for these analyses.

CLC files can be uploaded to AWS S3 directly via the CLC Workbench, as described in section 3.4.

Raw data, for example, FASTQ files, can be uploaded to AWS S3 using third party tools, or via the web interface of a CLC Server, as described at https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Browse_AWS_S3_locations.html.

QIAGEN reference data is already available on AWS S3, so no transfer of this data is needed. This is discussed further in section 6.1.1.

The few situations where using locally held data is the only option are described in section 6.1.2.

6.1.1 Reference data for analyses on the cloud

When running workflows configured to use QIAGEN reference data, no transfer of that data will take place, and there is no need to explicitly copy it to your own AWS S3 account. Under certain conditions, no local copy of this reference data is needed either. This is described in more detail later in this section.

In cases where other reference data will be used, the following steps are recommended:

- Connect Input elements to all input channels that require reference data (figure 6.1).
  When a workflow Input element is connected, you are presented with the option of using on-the-fly import. This allows you to select data in an AWS S3 bucket (figure 6.2).
- Upload reference data to AWS S3 before launching analyses.
Figure 6.1: Using the workflow on the left, data for the References field can only be selected from a CLC Location. The workflow on the right has an Input element connected to the References input channel. Using that workflow, files can be selected from an AWS S3 bucket, or from other accessible places, including CLC Locations.
Figure 6.2: An Input element is connected to the References input channel. On-the-fly import of a CLC format file has been specified by selecting the "Select files for import" option and "CLC Format" from the drop-down list of formats. The relevant AWS Connection has been selected from the drop-down list of locations. A CLC file was then selected for use as the reference genome.

QIAGEN reference data in workflows

QIAGEN reference data elements\(^1\) are already present in AWS S3 (figure 6.3) and thus do not need to be uploaded to your own S3 bucket when running workflows configured to refer to them. This includes many of the template workflows delivered with the software, and thus also workflows derived from those.

When the conditions listed below are met, there is also no need for a local copy of QIAGEN reference data when launching workflows to run on the cloud.

- All reference data parameters must be configured with a single QIAGEN reference data element and/or configured with a workflow role that is specified in one or more QIAGEN reference sets.


\(^1\)“QIAGEN reference data” refers to data sets or data elements provided by QIAGEN, available from under the QIAGEN references tab of the Reference Data Manager. Template workflows, delivered with the software, are commonly configured to use QIAGEN reference data.
All workflow parameters referring to QIAGEN reference data are locked. Where reference data elements need to be specified when launching, those elements, by definition, must be unlocked, and so a locally accessible copy of the relevant reference data will be needed. This is the case, for example, for many QIseq analysis workflows provided by the Biomedical Genomics Analysis plugin, where there are drop-down menus for selecting target region and target primer sets. If QIAGEN reference elements are selected, they are not copied to S3 from your system. Instead, the copy provided by QIAGEN in AWS S3 will be used for analyses run on the cloud.

Workflows are being submitted via a CLC Workbench. When submitting workflows that make use of QIAGEN reference data via a CLC Server, the reference data elements must be present in the CLC Server CLC_References location. Any QIAGEN reference elements selected will, however, not be copied to S3 from your system. The copy provided in AWS S3 by QIAGEN will be used for analyses run on the cloud.

Note: To view track lists that refer to reference data elements, those elements must be available locally.

Figure 6.3: QIAGEN Reference Sets do not need to be available locally (left hand image) for them to be available when launching a workflow to run on the cloud (right hand image).

6.1.2 Using locally held data for analyses on AWS S3

Two considerations when using data stored locally for analysis on AWS S3 are:

1. Transfer of data from a non-cloud location, such as from CLC Locations, from local directories, or from CLC Server import/export directories, takes time, and when submitting jobs from the CLC Workbench, any necessary data transfer must be complete before the software can be closed.

2. Data uploaded to AWS S3 as part of job submission is not saved to AWS S3 for subsequent use. Later analyses using the same input data would involve another transfer of that data to the cloud.

Cases where locally held data must be used:
• Launching a tool from the CLC Workbench Toolbox to run on a CLC Genomics Cloud\(^2\). This tool will be wrapped in a workflow for you, before the job is sent to the cloud. When launching analyses this way, you can only select data held in CLC Locations. Thus, the first step when a tool has been launched will always be to upload the input data to AWS S3. Launching tools can be useful when running small tests, but for analysis of large data sets, or where data needs to be imported, we recommend creating a workflow containing the tool of interest and launching that instead.

• When specifying reference data elements in certain workflow designs, as described in section 6.1.1

• Running analyses that require a CLC Metadata Table with data associated to it using a workflow that does not include an Iterate element. An example would be a workflow with the Differential Expression tool at the top of the workflow.

Workflows containing Differential Expression can be run using data already on AWS S3 when that workflow contains an Iterate element upstream. An example illustrating this is the RNA-Seq and Differential Gene Expression Analysis template workflow, delivered with the CLC Genomics Workbench, described at http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_Differential_Gene_Expression_Analysis_workflow.html. If the batch units for the RNA-Seq analysis section of that workflow are defined using metadata, that same metadata will be used as input to the Differential Expression step.

### 6.2 General information about outputs from cloud analyses

All results generated on the cloud are stored on the cloud. To view results in a CLC Workbench, the data must be downloaded to a local CLC location. This is easily done via a CLC Workbench via the Remote Files tab, described in section 3.4, and via the Processes tab, described in section 3.5.

**AWS charges for downloading data.** In CLC Workbenches, when the download size exceeds 1 GB, you are prompted for confirmation that you wish to proceed. The size required to trigger this warning can be changed in the General section of the Workbench Preferences (figure 6.4). When downloading via the CLC Server web interface, a window pops up listing the selected files and providing the total download size. After selecting whether to download to a CLC location or to a CLC Server import/export area, you are prompted to select a specific location to download to. For details, see https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Browse_AWS_S3_locations.html.

For downloading large amounts of data, we recommend using dedicated software for interacting with AWS S3. CLC data downloaded using external software can be imported into the CLC Workbench using Standard Import, described at http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Standard_import.html.

**Download all contained tracks when downloading track lists to view**

Track lists are small elements that refer to tracks. When downloading track lists to view, the track list and all the tracks contained must be downloaded at the same time.

\(^2\)Only tools that can be used within workflows can be submitted to run on a CLC Genomics Cloud. Tools that are not workflow-enabled cannot be run on the cloud.
Figure 6.4: Download cost warning dialogs are shown whenever the size of a download from AWS S3 exceeds the value configured in the Workbench Preferences. By default, the value is set to 1000 MB.

If this is not done, one or more of the tracks in the track list will not be resolved. The relevant links between track lists and the tracks referred to cannot be made if these elements are downloaded separately.
Chapter 7

Monitoring and reviewing CLC cloud jobs

CLC jobs run on the cloud are monitored using CloudWatch. Information about jobs can be found via CLC software as well as the AWS Console.

Key files for investigating cloud jobs are:

- **workflow-result.json** This file contains a list of the outputs from an analysis and accessing it via a CLC Workbench or CLC Server.

- **The analysis log** This is the usual source of information about analyses, where the focus is on the analysis steps run. This log provides information useful when investigating problems with the analysis itself. The filename include the name of the analysis followed by "log", for example, "Prepare Raw Data log" (figure 7.2).

- **Execution Log** This file contains the log stream information from Cloud Watch for the job. This file can be opened from the list provided in the Processes tab (figure 7.1), or by double-clicking on the workflow-result.json file in the Remote Files tab (figure 7.2). For very long logs, the first 10,000 entries from the log stream are included, as well as a few entries from the end of the log stream. If the full log is needed, it can be accessed directly in AWS CloudWatch.

  The Execution Log can also be retrieved, as long as the information is still available on AWS.

  The Execution Log is primarily for troubleshooting problems with a job on the cloud.

In this chapter, we describe monitoring jobs using:

- A CLC Workbench: section 7.1
- A CLC Server: section 7.2
- The AWS Console: section 7.3

7.1 Using a CLC Workbench to monitor and review cloud jobs

Information about the status of queued, running and completed jobs submitted from a CLC Workbench to the cloud, either directly, or via a CLC Server, appears in the Processes tab in the
lower left side of the Workbench. When the CLC Workbench is closed and then re-opened, AWS Batch is queried, and processes for jobs are listed and updated with their progress. Any cloud job that has not previously been actively cleared from the process list will be shown.

When jobs have completed, results can be found, or downloaded and opened, using options in the menu found under the small icon to the right of the process. The standard log for an analysis (e.g. "Prepare Raw Data log" in the figures below) and the Execution Log are available from this menu (figure 7.1).

![Figure 7.1: Results and log files for a given analysis run on the cloud can be accessed from the Processes tab in the Workbench.](image)

To clear processes from the list in the Processes tab, right-click on an empty area and select the option **Remove Finished Processes**.

After processes have been cleared, the same list of results, and access to the Execution Log is available by double clicking on a workflow-result.json file under the Remote Files tab (figure 7.2).
Figure 7.2: Double-clicking on a workflow-result.json file in the Remote Files tab in the Workbench provides access to results and log files for a given analysis run on the cloud.

7.2 Using a CLC Server to monitor and review cloud jobs

Information about queued and running jobs submitted via the CLC Server can be seen in the Queue tab, under the Management tab of the CLC Server web administrative interface. Detailed job information for completed processes is available from the Audit log, also under the Management tab of the web administrative interface. Cloud jobs have a C in the Status column (figure 7.3).

Figure 7.3: Jobs sent to the cloud via a CLC Server are recorded in the audit log. A C in the status column indicates a job sent to the cloud. Clicking on the links (red text) opens up information about the job.

In the audit log, click on a link in the row for a job that contains the word "Done" to open a window containing detailed information about the job. Drill down through the Server Command Output to get to buttons for downloading the AWS log, i.e. the Cloud Watch log stream for that job, as well as buttons for downloading results or downloading a copy of the workflow that was run.
Figure 7.4: In the Command done window, buttons are provided under Server Command Output that can be used to view the AWS log, i.e. the CloudWatch log stream, as well as for downloading results or the workflow run.

The same information can be obtained by navigating to a workflow-result.json file in the Browse S3 locations area under the Element info tab of the CLC Server web interface and clicking on the View workflow result... button (figure 7.5).

By default, the Audit Log and Queue tabs in the CLC Server web interface are only available to administrative users. Access can be granted to users of other groups, as described at https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Controlling_access_server_server_tasks_external_data.html.
CHAPTER 7. MONITORING AND REVIEWING CLC CLOUD JOBS

7.3 Using the AWS Console to monitor and review cloud jobs

CLC jobs run on AWS are monitored using CloudWatch. The log group `cgc` contains log streams for each job. The AWS_BATCH_JOB_ID field, near the top of the log information, links the log stream to the job ID.

You can navigate to the job log in CloudWatch from the AWS Batch page in the AWS Console. The Job queue overview in the Batch dashboard contains links to information about jobs in each queue, according to their status (e.g. Running, Succeeded, Failed). Click on one of these links to see a list of the jobs in that category for that queue (figure 7.6). From there, click on the link for an individual job and then click on the logstream link, in the Job information section. This takes you to the information about that job in CloudWatch.

Figure 7.5: Buttons to download workflow results and the AWS log, i.e. the CloudWatch log stream, can be revealed by navigating to a workflow-result.json file in the Browse S3 locations area and clicking on the "View workflow result..." button.
Figure 7.6: Information about jobs run on a CLC Genomics Cloud setup can be found using AWS functionality, such as the AWS Batch overview, shown here. Clicking on the links on this page allow you to drill down to details about individual jobs.
8.1 Standard AWS Batch queues for CLC Genomics Cloud

The three AWS Batch queues defined in the CloudFormation template at:


have the following settings:

cgc-small

<table>
<thead>
<tr>
<th>Setting</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC2 instance type</td>
<td>c5a.4xlarge</td>
</tr>
<tr>
<td>vCPUs per job</td>
<td>16</td>
</tr>
<tr>
<td>Max vCPUs</td>
<td>160</td>
</tr>
<tr>
<td>Total available memory per job</td>
<td>31000 MB</td>
</tr>
<tr>
<td>Java available memory per job</td>
<td>15500 MB</td>
</tr>
<tr>
<td>Disk space</td>
<td>100 GB</td>
</tr>
</tbody>
</table>

cgc-medium

<table>
<thead>
<tr>
<th>Setting</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC2 instance type</td>
<td>c5a.8xlarge</td>
</tr>
<tr>
<td>vCPUs per job</td>
<td>32</td>
</tr>
<tr>
<td>Max vCPUs</td>
<td>320</td>
</tr>
<tr>
<td>Total available memory per job</td>
<td>63000 MB</td>
</tr>
<tr>
<td>Java available memory per job</td>
<td>31000 MB</td>
</tr>
<tr>
<td>Disk space</td>
<td>300 GB</td>
</tr>
</tbody>
</table>

cgc-large
### 8.2 AWS Batch queues, compute environments and AMIs

AWS Batch queues created using the CloudFormation templates provided by QIAGEN are each mapped to an individual compute environment. Each compute environment is configured to use the latest Amazon ECS-optimized AMI available at the time it is established. Thus, in a CLC Genomics Cloud setup, all jobs sent to a given AWS Batch queue will use the same AMI version.

AWS Batch does not upgrade the AMIs in existing compute environments. Thus, compute environments established at different times may contain different AMI versions. An example where this could happen in a CLC Genomics Cloud context would be where a new AWS Batch queue is created some time after the initial infrastructure was set up, and there had been an AMI update released in the meantime.

We generally do not expect performance issues using supported, older AMI versions. From a security perspective, AWS Batch compute environments established using the CloudFormation templates provided by QIAGEN are placed in a security group that rejects incoming traffic that was not initiated from the EC2 host.

**Which AMI is being used?**

The image ID for the AMI used for a job is recorded near the top of the CloudWatch log (aka Execution Log). Information about that public image, such as the Amazon Linux AMI version and the deprecation time for that AMI can be found via the AWS Console:

Console | EC2 | Images | AMIs


**Accessing a newer AMI**

To update the AMI being used, a new compute environment is needed. An easy way to achieve this is to create a new AWS Batch queue, as described in section 2.2. Queues that are no longer needed can be deleted, as described in section 2.4.

### 8.3 System file retention policies

By default, system files in the CGC system S3 bucket are deleted after 30 days, and read mapper indexes are deleted 14 days after creation. These defaults can be changed by editing the expire-jobs and expire-read-mapper-index-files lifecycle rules associated with the CGC system bucket.
8.4 Making custom plugins available for cloud analyses

Tools delivered by custom CLC plugins are supported for cloud analyses when:

- Server plugins are placed in a plugin repository that has the expected file organization (described below).
- The server plugins have the name `<workbench_plugin_id>.cpa`.
- The location of the custom plugin repository is configured when setting up the standard CLC Genomics Cloud infrastructure on AWS, as described at section 2.1.

When the above requirements have been met, analyses using tools delivered by a custom plugin can be run on the cloud. The corresponding plugin in the custom repository is used for the cloud-based execution.

Custom plugin repository structure

The custom plugin repository must have the structure and contents defined below. The term `<workbench_plugin_id>` refers to the value of that field in the plugin.properties file.

- A top level folder per plugin. The folder name must be `<workbench_plugin_id>`.
- A subfolder named `server`.
- Within the `server` folder, a subfolder with the name of the `version of the plugin`. e.g. "23.0" or "22.0.2".
  
  This name must be the version of the plugin. This may be different to the version or your CLC Workbench or CLC Server.
- The plugin .cpa file and any other relevant files (e.g. plugin.properties) must be placed in this folder.
  
  The plugin .cpa file must be named as `<workbench_plugin_id>.cpa`.

For example, a server plugin with the `<workbench_plugin_id>` value `example_plugin` and version 23.0, the following structure in the plugin repository would be expected:

```
example_plugin/
  server/
    23.0/
```

---

1 Only tools that can be used within workflows can be submitted to run on a CLC Genomics Cloud. Tools that are not workflow-enabled cannot be run on the cloud.
When a job is launched to run on the cloud, the exact version of the plugin in the CLC Workbench or CLC Server must be available in the repository. If it is not, the job will fail.