CLC **Microbial Genomics** Module

USER MANUAL

User manual for

CLC Microbial Genomics Module 1.0

Windows, Mac OS X and Linux

June 12, 2015

This software is for research purposes only.

CLC bio, a QIAGEN Company
Silkeborgvej 2
Prismet
DK-8000 Aarhus C
Denmark

Contents

Chapter 1

Introduction to the CLC Microbial Genomics Module

1.1 The concept of the CLC Microbial Genomics Module

The majority of microbial species present in the human body or indeed anywhere in the environment have never been isolated, cultured or sequenced, due to our inability to reproduce necessary growth conditions in the lab. Therefore, there are huge amounts of organismal and functional novelty still to be discovered. Two central questions in microbial community analysis ask: Which microbial species are present in a sample from a given habitat, and at what frequencies? Microbiome analysis takes advantage of DNA molecular techniques and sequencing technology in order to comprehensively retrieve specific regions of microbial genomic DNA useful for taxonomical identification. For bacteria, the most widely used regions are parts of the 16S rRNA gene. In a microbiome analysis workflow, total genomic DNA is extracted from the sample(s) of interest, a region of the 16S gene is PCR amplified, and the resulting amplicon is sequenced using an NGS machine. The bioinformatics task is then to assign taxonomy to the reads and tally their occurrences. Due to the incomplete nature of bacterial taxonomy and presence of sequencing errors in the NGS reads, a common approach is to cluster reads at some level of similarity into pseudospecies called Operational Taxonomical Units (OTUs), where all reads within e.g. 97% similarity are clustered together and represented by a single sequence. PCR amplification can introduce artefacts in the form of chimeric sequences, where template swapping results in a sequence having two or more parental templates. These chimeras can be identified during the clustering step.

The features of the CLC Microbial Genomics Module 1.0 include:

- Trimming and merging of reads.
- Clustering of reads into OTUs.
- Generation of OTU tables.
- Working with metadata.
- Visualization options (stacked barplots, zoomable sunbursts).
- Multiple sequence alignment using MUSCLE.

- Phylogenetic tree construction using maximum likelihood.
- Estimation of alpha diversity.
- Rarefaction analysis.
- Estimation of beta diversity.
- Principal coordinates analysis
- PERMANOVA test
- Statistical tests for differential abundance

The primary output of the clustering and tallying process is an OTU table, listing the abundances of OTUs in the samples under investigation, as well as new features allowing clear visualization of the results. Secondary analyses include estimations of alpha and beta diversities, in addition to statistical tests for differential abundance.

1.2 Contact information

The CLC Workbench is developed by:

CLC bio, a QIAGEN Company
Silkeborgvej 2
Prismet
8000 Aarhus C
Denmark

http://www.clcbio.com

VAT no.: DK 28 30 50 87

Telephone: 45 70 22 32 44
Fax: +45 86 20 12 22

E-mail: info@clcbio.com

If you have questions or comments regarding the program, you can contact us through the support team as described here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Getting_help.html.

Chapter 2

System requirements and installation

2.1 System Requirements

To work with the CLC Microbial Genomics Module 1.0 you will need to have a CLC Genomics Workbench version 8.0 or higher or the Biomedical Genomics Worbench 2.1 or higher installed on your computer. With exception of the two editors below, the system requirements of CLC Microbial Genomics Module 1.0 are the same as the ones required for the CLC Genomics Workbench:

- Windows Vista, Windows 7, Windows 8 or Windows Server 2008.
- Mac OS X 10.7 or later.
- Linux: RHEL 5.0 or later. SUSE 10.2 or later. Fedora 6 or later.
- 2 GB RAM required.
- 4 GB RAM recommended.
- 1024 x 768 display required.
- 1600 x 1200 display recommended.
- Intel or AMD CPU required.

The **PCoA 3D viewer** requirements are the same as the 3D Molecule Viewer:

- **System requirements**
  - A graphics card capable of supporting OpenGL 2.0.
  - Updated graphics drivers. Please make sure the latest driver for the graphics card is installed.
- **System Recommendations**
  - A discrete graphics card from either Nvidia or AMD/ATI. Modern integrated graphics cards (such as the Intel HD Graphics series) may also be used, but these are usually slower than the discrete cards.

  - A 64-bit workbench version.

The **Sunburst viewer** makes use of JavaFX and may not work on older Linux kernels. Un updated list of requirements for JavaFX can be found at http://www.oracle.com/technetwork/java/javafx/downloads/supportedconfigurations-1506746.html.

2.2 Installation of modules

Modules are installed using the Plugins and Resources Manager[1], which can be accessed via the menu in the Workbench

Help | Plugins and Resources ( )

or via the **Plugins ( )** button on the Toolbar.

From within the Plugins and Resources Manager, choose the Download Plugins tab and click on the CLC Workbench Client Plugin. Then click on the button labeled **Download and Install**.

If you are working on a system not connected to the internet, then you can also install the plugin by downloading the cpa file from the plugins page of our website

http://www.clcbio.com/clc-plugin/

Then start up the Plugin manager within the Workbench, and click on the button at the bottom of the Plugin manager labeled **Install from File**.

You need to restart the Workbench before the module is ready for use.

2.3 Workbench Licenses

When you have installed the CLC Microbial Genomics Module, and start it for the first time or after installing a new major release, you will meet the license assistant, shown in figure 2.1.

To manually start up the License Manager for a Workbench plugin, first open the Plugin Manager (see Section 2.2), select the relevant plugin or module, and press the button labeled *Import a new license*.

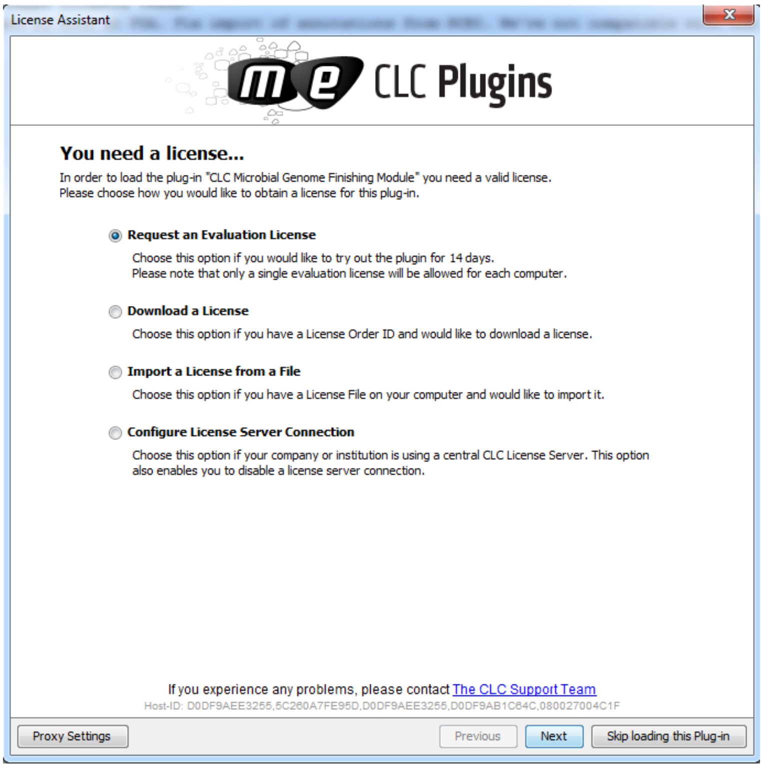
To install a license, you must be running the program in administrative mode [2].

The following options are available. They are described in detail in the sections that follow.

- **Request an evaluation license.** Request a fully functional, time-limited license (see below).
- **Download a license.** Use the license order ID received when you purchase 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.

[1] In order to install plugins on many systems, the Workbench must be run in administrator mode. On Windows Vista and Windows 7, you can do this by right-clicking the program shortcut and choosing "Run as Administrator".

[2] How to do this differs for different operating systems. To run the program in administrator mode on Windows Vista, or 7, right-click the program shortcut and choose "Run as Administrator.



You need a license...

In order to load the plug-in "CLC Microbial Genome Finishing Module" you need a valid license.
Please choose how you would like to obtain a license for this plug-in.

● **Request an Evaluation License**
Choose this option if you would like to try out the plugin for 14 days.
Please note that only a single evaluation license will be allowed for each computer.

○ **Download a License**
Choose this option if you have a License Order ID and would like to download a license.

○ **Import a License from a File**
Choose this option if you have a License File on your computer and would like to import it.

○ **Configure License Server Connection**
Choose this option if your company or institution is using a central CLC License Server. This option also enables you to disable a license server connection.

If you experience any problems, please contact The CLC Support Team
Host-ID: D0DF9AEE3255,5C260A7FE95D,D0DF9AEE3255,D0DF9AB1C64C,080027004C1F

Figure 2.1: *The license assistant showing you the options for getting started.*

- **Configure license server connection**. If your organization has a CLC License Server, select this option to configure the connection to it.

Select the appropriate option and click on button labeled **Next**.

To use the Download option in the License Manager, your machine must be able to access the external network. If this is not the case, please see section 2.5.5.

2.3.1 Request an evaluation license

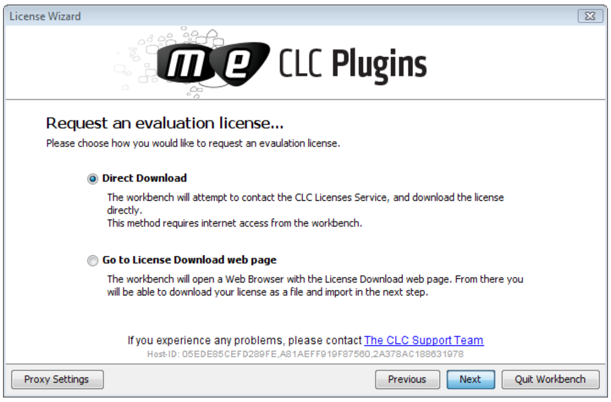
We offer a fully functional version of the CLC Microbial Genomics Module for evaluation purposes, free of charge. Each user is entitled to 14 days demo of the CLC Microbial Genomics Module. If you are unable to complete your assessment in the available time, please send an email to sales@clcbio.com to request an additional evaluation period.

When you choose the option **Request an evaluation license**, you will see the dialog shown in figure 2.2.

In this dialog, there are two options:

- **Direct download**. Download the license directly from CLC bio. This method requires that the Workbench has access to the external network.

- **Go to license download web page**. In a browser window, show the license download web page, which can be used to download a license file. This option is suitable in situations where, for example, you are working behind a proxy, so that the Workbench does not have direct access to the CLC Licenses Service.



Request an evaluation license...

Please choose how you would like to request an evaulation license.

● Direct Download

The workbench will attempt to contact the CLC Licenses Service, and download the license directly.
This method requires internet access from the workbench.

○ Go to License Download web page

The workbench will open a Web Browser with the License Download web page. From there you will be able to download your license as a file and import in the next step.

If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

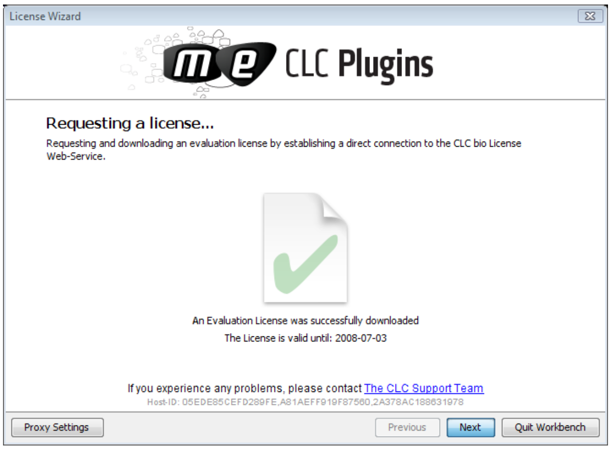
Figure 2.2: *Choosing between direct download or going to the license download web page.*

If you select the option to download a license directly and it turns out that the Workbench does not have direct access to the external network, (because of a firewall, proxy server etc.), you can click **Previous** button to try the other method.

After selection on your method of choice, click on the button labeled **Next**.

Direct download

After choosing the **Direct Download** option and clicking on the button labeled **Next**, the dialog shown in figure 2.3 appears.



Requesting a license...

Requesting and downloading an evaluation license by establishing a direct connection to the CLC bio License Web-Service.

An Evaluation License was successfully downloaded
The License is valid until: 2008-07-03

If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 2.3: *A license has been downloaded.*

A progress for getting the license is shown, and when the license is downloaded, you will be able to click **Next**.

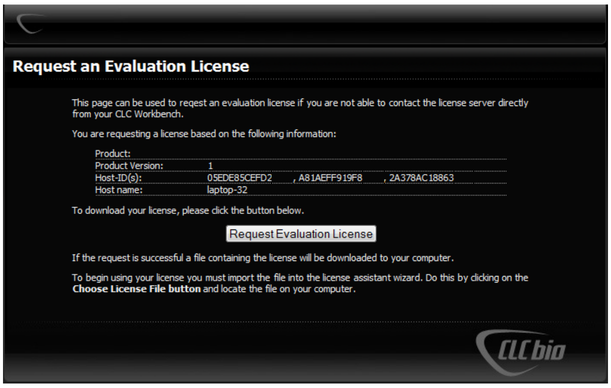
Go to license download web page

After choosing the *Go to license download web page* option and clicking on the button labeled **Next**, the license download web page appears in a browser window, as shown in 2.4.

Click the **Request Evaluation License** button. You can then save the license on your system.

Back in the Workbench window, you will now see the dialog shown in 2.5.

Click the **Choose License File** button and browse to find the license file you saved. When you



Request an Evaluation License

This page can be used to reqest an evaluation license if you are not able to contact the license server directly from your CLC Workbench.

You are requesting a license based on the following information:

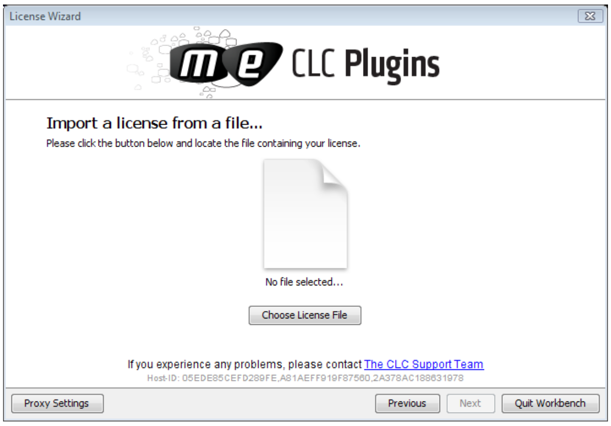
Product:	
Product Version:	1
Host-ID(s):	05EDE85CEFD2 , A81AEFF919F8 , 2A378AC18863
Host name:	laptop-32

To download your license, please click the button below.

If the request is successful a file containing the license will be downloaded to your computer.

To begin using your license you must import the file into the license assistant wizard. Do this by clicking on the **Choose License File button** and locate the file on your computer.

Figure 2.4: *The license download web page.*



Import a license from a file...

Please click the button below and locate the file containing your license.

No file selected...

If you experience any problems, please contact The CLC Support Team

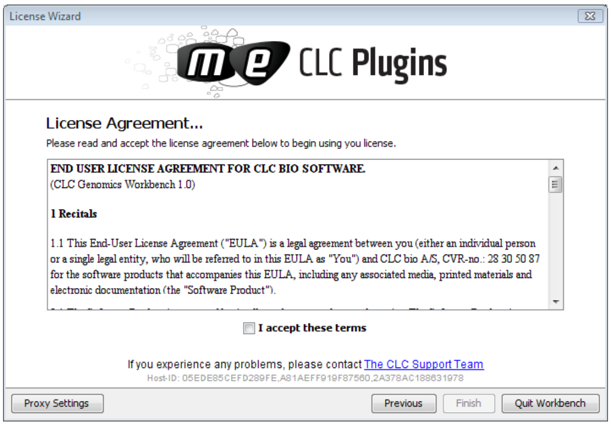
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 2.5: *Importing the license file downloaded from the web page.*

have selected the file, click on the button labeled **Next**.

Accepting the license agreement

Part of the installation of the license involves checking and accepting the end user license agreement (EULA). You should now see the a window like that in figure 2.6.



License Agreement...

Please read and accept the license agreement below to begin using you license.

END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE.
(CLC Genomics Workbench 1.0)

1 Recitals

1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").

☐ I accept these terms

If you experience any problems, please contact The CLC Support Team

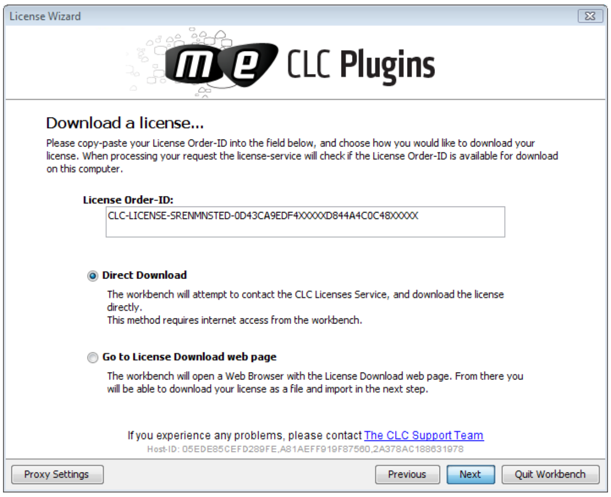
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 2.6: *Read the license agreement carefully.*

Please read the EULA text carefully before clicking in the box next to the text **I accept these terms** to accept, and then clicking on the button labeled **Finish**.

2.3.2 Download a license using a license order ID

Using a license order ID, you can download a license file via the Workbench or using an online form. When you have chosen this option and clicked **Next** button, you will see the dialog shown in 2.7. Enter your license order ID into the text field under the title License Order-ID. (The ID can be pasted into the box after copying it and then using menus or key combinations like Ctrl+V on some system or ⌘ + V on Mac).


License Wizard

CLC Plugins

Download a license...

Please copy-paste your License Order-ID into the field below, and choose how you would like to download your license. When processing your request the license-service will check if the License Order-ID is available for download on this computer.

License Order-ID:
CLC-LICENSE-SRENMNSTED-0D43CA9EDF4XXXXXD844A4C0C48XXXXX

● **Direct Download**
The workbench will attempt to contact the CLC Licenses Service, and download the license directly.
This method requires internet access from the workbench.

○ **Go to License Download web page**
The workbench will open a Web Browser with the License Download web page. From there you will be able to download your license as a file and import in the next step.

If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 2.7: *Enter a license order ID for the software.*

In this dialog, there are two options:

- **Direct download**. Download the license directly from CLC bio. This method requires that the Workbench has access to the external network.
- **Go to license download web page**. In a browser window, show the license download web page, which can be used to download a license file. This option is suitable in situations where, for example, you are working behind a proxy, so that the Workbench does not have direct access to the CLC Licenses Service.

If you select the option to download a license directly and it turns out that the Workbench does not have direct access to the external network, (because of a firewall, proxy server etc.), you can click **Previous** button to try the other method.

After selection on your method of choice, click on the button labeled **Next**.

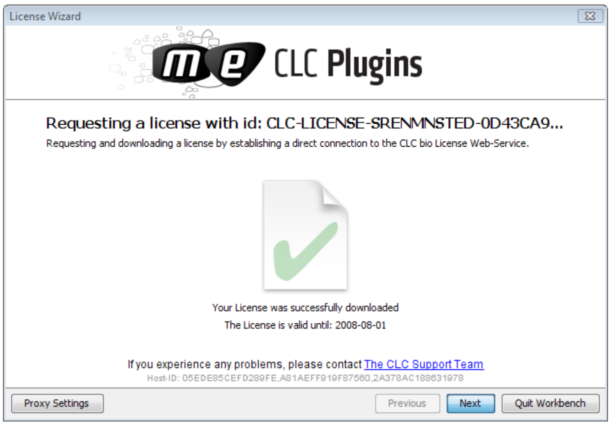
Direct download

After choosing the **Direct Download** option and clicking on the button labeled **Next**, the dialog shown in figure 2.8 appears.

A progress for getting the license is shown, and when the license is downloaded, you will be able to click **Next**.

Go to license download web page

After choosing the *Go to license download web page* option and clicking on the button labeled **Next**, the license download web page appears in a browser window, as shown in 2.9.



Requesting a license with id: CLC-LICENSE-SRENMNSTED-0D43CA9...

Requesting and downloading a license by establishing a direct connection to the CLC bio License Web-Service.

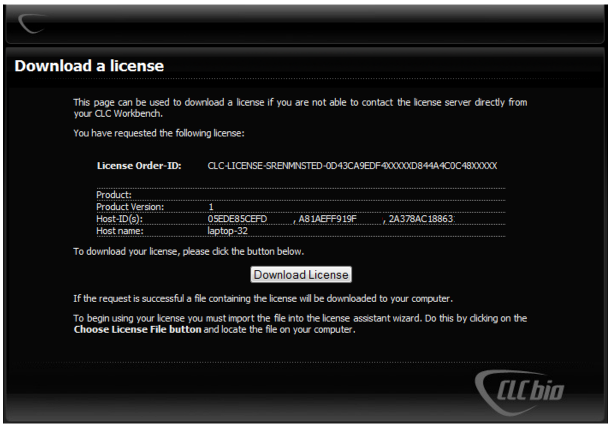
Your License was successfully downloaded

The License is valid until: 2008-08-01

If you experience any problems, please contact The CLC Support Team

Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 2.8: *A license has been downloaded.*



Download a license

This page can be used to download a license if you are not able to contact the license server directly from your CLC Workbench.

You have requested the following license:

License Order-ID:	CLC-LICENSE-SRENMNSTED-0D43CA9EDF4XXXXXD844A4C0C48XXXXX
Product:	
Product Version:	1
Host-ID(s):	05EDE85CEFD , A81AEFF919F , 2A378AC18863
Host name:	laptop-32

To download your license, please click the button below.

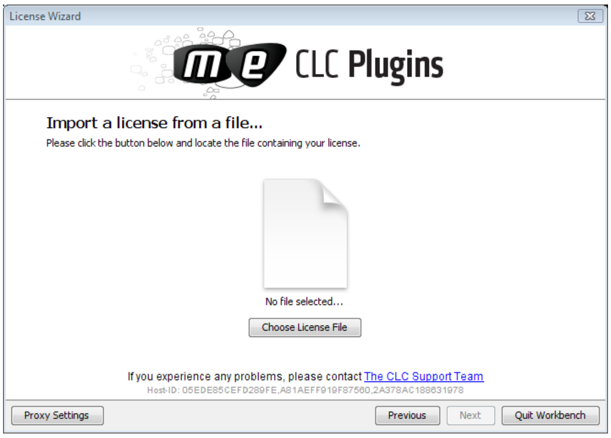
If the request is successful a file containing the license will be downloaded to your computer.

To begin using your license you must import the file into the license assistant wizard. Do this by clicking on the **Choose License File button** and locate the file on your computer.

Figure 2.9: *The license download web page.*

Click the **Request Evaluation License** button. You can then save the license on your system.

Back in the Workbench window, you will now see the dialog shown in 2.10.



Import a license from a file...

Please click the button below and locate the file containing your license.

No file selected...

If you experience any problems, please contact The CLC Support Team

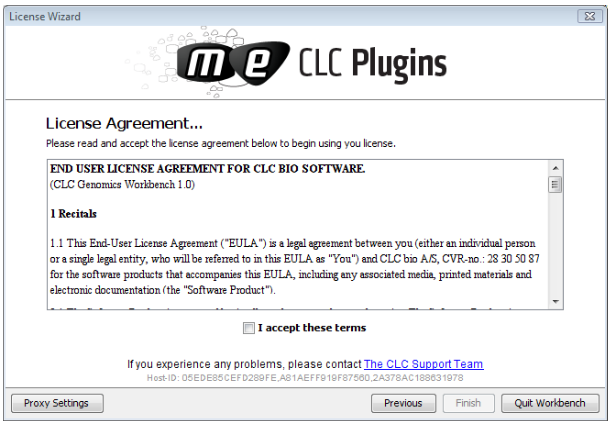
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 2.10: *Importing the license file downloaded from the web page.*

Click the **Choose License File** button and browse to find the license file you saved. When you have selected the file, click on the button labeled **Next**.

Accepting the license agreement

Part of the installation of the license involves checking and accepting the end user license agreement (EULA). You should now see the a window like that in figure 2.11.



License Agreement...

Please read and accept the license agreement below to begin using you license.

END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE.
(CLC Genomics Workbench 1.0)

1 Recitals

1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").

□ **I accept these terms**

If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

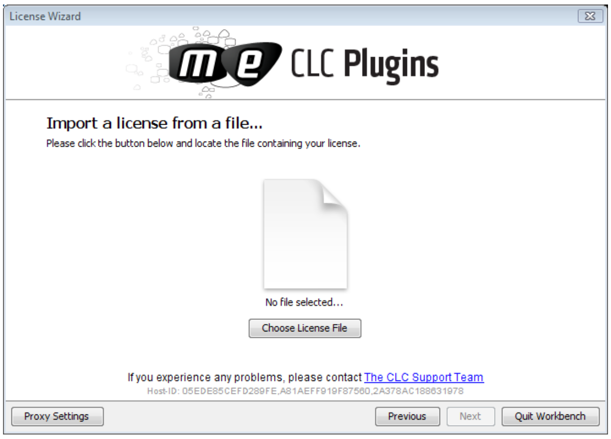
Figure 2.11: *Read the license agreement carefully.*

Please read the EULA text carefully before clicking in the box next to the text **I accept these terms** to accept, and then clicking on the button labeled **Finish**.

2.3.3 Import a license from a file

If you already have a license file associated with the host ID of your machine,it can be imported using this option.

When you have clicked on the **Next** button, you will see the dialog shown in 2.12.



Import a license from a file...

Please click the button below and locate the file containing your license.

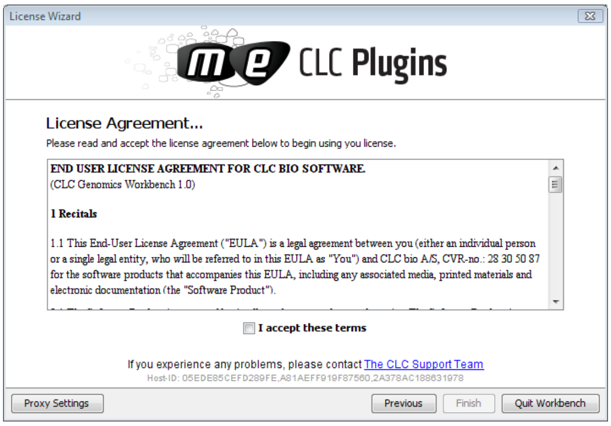
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 2.12: *Selecting a license file .*

Click the **Choose License File** button and browse to find the license file. When you have selected the file, click on the **Next** button.

Accepting the license agreement

Part of the installation of the license involves checking and accepting the end user license agreement (EULA). You should now see the a window like that in figure 2.13.



License Agreement...

Please read and accept the license agreement below to begin using you license.

END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE.
(CLC Genomics Workbench 1.0)

1 Recitals

1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").

□ **I accept these terms**

If you experience any problems, please contact The CLC Support Team

Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

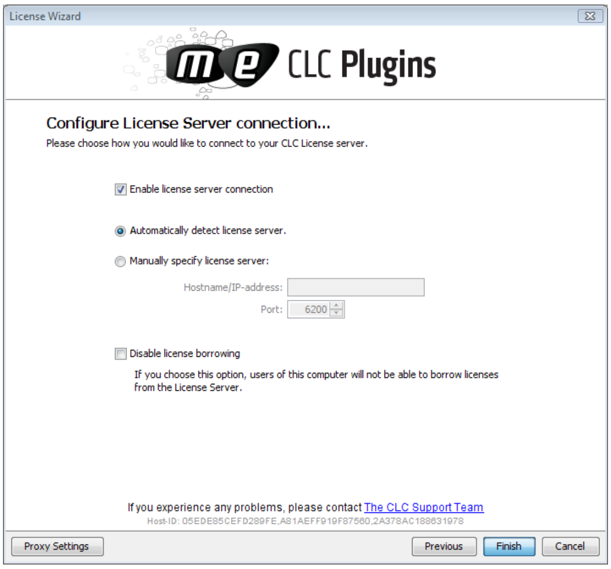
Figure 2.13: *Read the license agreement carefully.*

Please read the EULA text carefully before clicking in the box next to the text **I accept these terms** to accept, and then clicking on the button labeled **Finish**.

2.3.4 Configure license server connection

If your organization is running a CLC License Server, you can configure your Workbench to connect to it to get a license.

To do this, select this option and click on the **Next** button. A dialog like that shown in figure 2.14 then appears. Here, you configure how to connect to the CLC License Server.



Configure License Server connection...

Please choose how you would like to connect to your CLC License server.

☑ Enable license server connection

● Automatically detect license server.

○ Manually specify license server:

Hostname/IP-address:

Port: 6200

□ Disable license borrowing

If you choose this option, users of this computer will not be able to borrow licenses from the License Server.

If you experience any problems, please contact The CLC Support Team

Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 2.14: *Connecting to a CLC License Server.*

- **Enable license server connection.** This box must be checked for the Workbench is to contact the CLC License Server to get a license for CLC Workbench.
- **Automatically detect license server.** By checking this option the Workbench will look for a CLC License Server accessible from the Workbench[3].

[3] Automatic server discovery sends UDP broadcasts from the Workbench on a fixed port, 6200. Available license

- **Manually specify license server**. If there are technical limitations such that the CLC License Server cannot be detected automatically, use this option to provides details of machine the CLC License Server software is on, and the port used by the software to receive requests. After selecting this option, please enter:
  - **Host name**. The address for the machine the CLC Licenser Server software is running on.
  - **Port**. The port used by the CLC License Server to receive requests.
- **Disable license borrowing on this computer**. If you do not want users of the computer to borrow a license from the set of licenses available, then (see section 2.3.4), select this option.

Borrowing a license

A network license can only be used when you are connected to the a license server. If you wish to use the CLC Workbench when you are not connected to the CLC License Server, you can *borrow* an available license for a period of time. During this time, there will be one less network license available on the for other users. The Workbench must have a connection to the CLC License Server at the point in time when you wish to borrow a license.

The procedure for borrowing a license is:

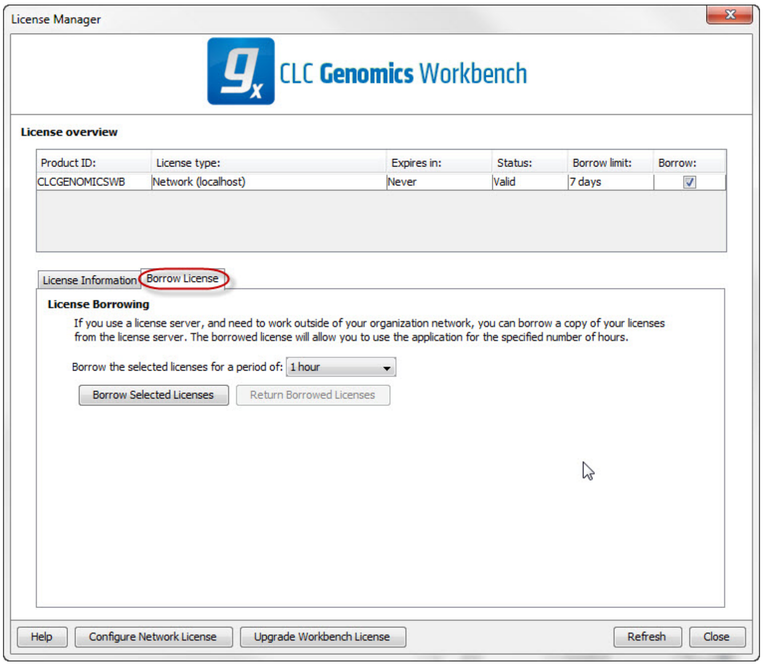
1. Go to the Workbench menu option:

   Help | License Manager
2. Click on the "Borrow License" tab to display the dialog shown in figure 2.15.
3. Use the checkboxes at the right hand sideof the table in the License overview section of the window to select the license(s) that you wish to borrow.
4. Select the length of time you wish to borrow the license(s).
5. Click on the button labeled **Borrow Licenses**.
6. Close the License Manager when you are done.

You can now go offline and work with the CLC Workbench. When the time period you borrowed the license for has elapsed, the network license you borrowed is made available again for other users to access. To continue using the CLC Workbench with a license, you will need to connect the Workbench to the network again so it can contact the CLC Licene Server to obtain one.

Note! Your CLC License Server administrator can choose to disable to the option allowing the borrowing of licenses. If this has been done, you will not be able to borrow a network license using your Workbench.

servers respond to the broadcast. The Workbench then uses TCP communication for to get a license, assuming one is available. Automatic server discovery works only on local networks and will not work on WAN or VPN connections. Automatic server discovery is not guaranteed to work on all networks. If you are working on an enterprise network on where local firewalls or routers cut off UDP broadcast traffic, then you may need to configure the details of the CLC License server manually instead.



License Manager

CLC Genomics Workbench

License overview

Product ID:	License type:	Expires in:	Status:	Borrow limit:	Borrow:
CLCGENOMICSWB	Network (localhost)	Never	Valid	7 days	☑

License Information | Borrow License

License Borrowing

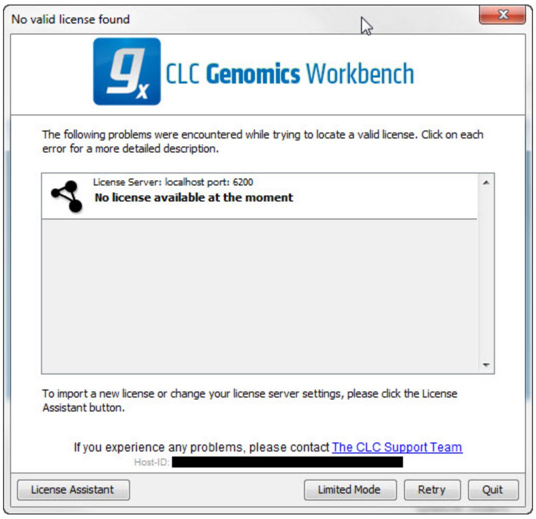
If you use a license server, and need to work outside of your organization network, you can borrow a copy of your licenses from the license server. The borrowed license will allow you to use the application for the specified number of hours.

Borrow the selected licenses for a period of: 1 hour

Figure 2.15: *Borrow a license.*

Common issues when using a network license

No license available at the moment If all the network licenses or CLC Workbenchare in use, you will see a dialog like that shown in figure 2.16 when you start up the Workbench.



No valid license found

CLC Genomics Workbench

The following problems were encountered while trying to locate a valid license. Click on each error for a more detailed description.

License Server: localhost port: 6200
No license available at the moment

To import a new license or change your license server settings, please click the License Assistant button.

If you experience any problems, please contact The CLC Support Team

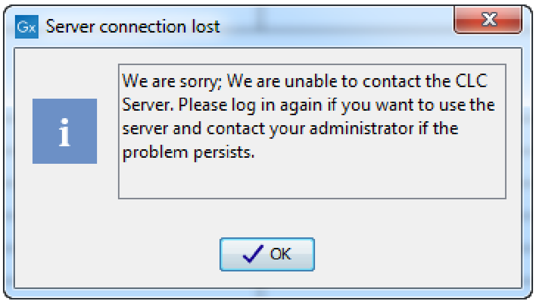
Host-ID: [redacted]

Figure 2.16: *This window appears when there are no available network licenses for the software you are running.*

This means others are using the network licenses. You will need to wait for them to return their licenses before you can continue to work with a fully functional copy of software. If this is a frequent issue, you may wish to discuss this with your CLC License Server administrator.

Clicking on the **Limited Mode** button in the dialog allows you to start the Workbench with functionality equivalent to the CLC Sequence Viewer. This includes the ability to access your CLC data.

Lost connection to the CLC License Server If the Workbench connection to the CLC License Server is lost, you will see a dialog as shown in figure 2.17.



We are sorry; We are unable to contact the CLC Server. Please log in again if you want to use the server and contact your administrator if the problem persists.

Figure 2.17: *This message appears if the Workbench is unable to establish a connection to a CLC License server.*

If you have chosen the option to **Automatically detect license server** and you have not succeeded in connecting to the License Server before, please check with your local IT support that automatic detection will be possible to do at your site. If it is not possible at your site, you will need to manually configure the CLC License Server settings using the License Manager, as described earlier in this section.

If you have successfully contacted the CLC License Server from your Workbench previously, please consider discussing this issue with your CLC License Server administrator or your local IT support, to make sure that the CLC License Server is running and that your Workbench can connect to it. There may be situations where you wish to use a different license or view information about the license(s) the Workbench is currently using. To do this, open the License Manager using the menu option:

Help | **License Manager** ( )

The license manager is shown in figure 2.18.

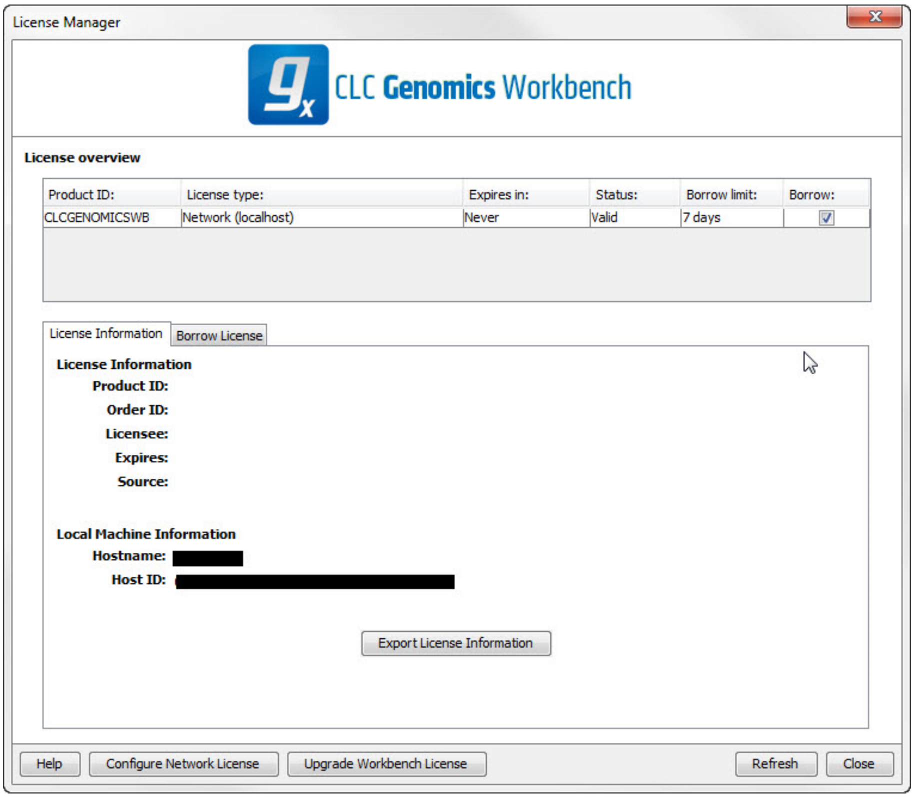
This dialog can be used to:

- See information about the license (e.g. what kind of license, when it expires)
- Configure how to connect to a license server (**Configure License Server** the button at the lower left corner). Clicking this button will display a dialog similar to figure 2.14.
- Upgrade from an evaluation license by clicking the **Upgrade license** button. This will display the dialog shown in figure 2.1.
- Export license information to a text file.
- Borrow a license

If you wish to switch away from using a network license, click on the button to **Configure License Server** and uncheck the box beside the text **Enable license server connection** in the dialog. When you restart the Workbench, you can set up the new license as described in section 2.3.

2.3.5 Download a static license on a non-networked machine

To download a static license for a machine that does not have direct access to the external network, you can follow the steps below:



License overview

Product ID:	License type:	Expires in:	Status:	Borrow limit:	Borrow:
CLCGENOMICSWB	Network (localhost)	Never	Valid	7 days	☑

License Information
Product ID:
Order ID:
Licensee:
Expires:
Source:

Local Machine Information
Hostname: [redacted]
Host ID: [redacted]

Figure 2.18: *The license manager.*

- Install the CLC Microbial Genomics Module on the machine you wish to run the software on.
- Start up the software as an administrative user and find the host ID of the machine that you will run the CLC Workbench on. You can see the host ID the machine reported at the bottom of the License Manager window in grey text.
- Make a copy of this host ID such that you can use it on a machine that has internet access.
- Go to a computer with internet access, open a browser window and go to the relevant network license download web page:
- For Workbenches released from January 2013 and later, (e.g. the Genomics Workbench version 6.0 or higher, and the Main Workbench, version 6.8 or higher), please go to:

  https://secure.clcbio.com/LmxWSv3/GetLicenseFile
- Paste in your license order ID and the host ID that you noted down in the relevant boxes on the webpage.
- Click 'download license' and save the resulting .lic file.
- Open the Workbench on your non-networked machine. In the Workbench license manager choose 'Import a license from a file'. In the resulting dialog click 'choose license file' to browse the location of the .lic file you have just downloaded.

  If the License Manager does not start up by default, you can start it up by going to the Help menu and choosing License Manager.
- Click on the **Next** button and go through the remaining steps of the license manager wizard.

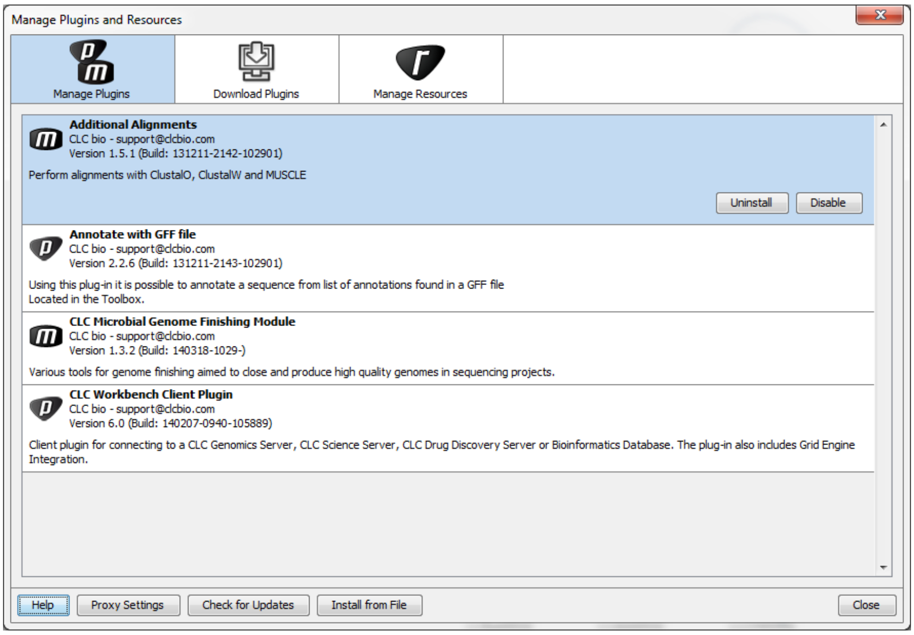
2.4 Uninstall of modules

Modules are uninstalled using the plugin manager:

Help in the Menu Bar | Plugins and Resources... ( )

or **Plugins ( ) in the Toolbar**

This will open the dialog shown in figure 2.19.



Additional Alignments
CLC bio - support@clcbio.com
Version 1.5.1 (Build: 131211-2142-102901)

Perform alignments with ClustalO, ClustalW and MUSCLE

Annotate with GFF file
CLC bio - support@clcbio.com
Version 2.2.6 (Build: 131211-2143-102901)

Using this plug-in it is possible to annotate a sequence from list of annotations found in a GFF file
Located in the Toolbox.

CLC Microbial Genome Finishing Module
CLC bio - support@clcbio.com
Version 1.3.2 (Build: 140318-1029-)

Various tools for genome finishing aimed to close and produce high quality genomes in sequencing projects.

CLC Workbench Client Plugin
CLC bio - support@clcbio.com
Version 6.0 (Build: 140207-0940-105889)

Client plugin for connecting to a CLC Genomics Server, CLC Science Server, CLC Drug Discovery Server or Bioinformatics Database. The plug-in also includes Grid Engine Integration.

Figure 2.19: *The plugin manager with plugins installed.*

The installed plugins and modules are shown in this dialog. To uninstall:

Click the CLC Microbial Genomics Module | Uninstall

If you do not wish to completely uninstall the module but you don't want it to be used next time you start the Workbench, click the **Disable** button.

When you close the dialog, you will be asked whether you wish to restart the workbench. The module will not be uninstalled until the workbench is restarted.

2.5 How to install a Server plugin

If you wish to use the tools and functionalities of the CLC Microbial Genomics Module with a CLC Genomics Server, you must purchase a Microbial Genomics Extension license and install it on your CLC Server as explained in the following steps:

1. Install *plugin licenses* to each machine with the CLC Server software installed, as described below.
2. Install *the Server plugin* on only the master CLC Server in the server setup, as described in section 2.5.7.

3. Restart *all* CLC Servers in the setup. How to stop and start CLC Servers is covered in the CLC Server manual at http://www.clcsupport.com/clcgenomicsserver/current/admin/index.php?manual=Starting_stopping_server.html.

There are three different server setups. A short description of each setup and a summary of the plugin licensing requirements are below.

- **Single server setup** - A single machine is running the CLC Server software. Jobs are submitted to this server, which receives and executes them. In this setup, a single machine acts both as a master and an executor of jobs. Here, a single static license for the plugin is installed in the CLC Server software.
- **Job node setup** - More than one machine is running the CLC Server software. The system acting as the master server receives job requests and then submits these jobs to other machines, the job nodes, for execution. Here, a single static license is installed *on each machine* running the CLC Server software. That is, a static license is installed on the master node and on each job node.
- **Grid setup** - One machine runs the CLC Server software and receives job requests. It then submits these to a third party scheduler. The scheduler then chooses an appropriate grid machine, or node, to submit a given job to for execution. Here, a a single static license for the plugin is installed on the master server, and the same number of network plugin licenses as there are network gridworker licenses needs to be made available by installing these in the *CLC License Server* software.

For a more detailed description of the different server setups, please refer to the CLC Server manual at: http://clcsupport.com/clcgenomicsserver/current/admin/index.php?manual=Job_Distribution.html

2.5.1 Static license installation

In each of the server models described above, a static license is installed in the CLC Server on a master machine. In the case of a job node setup, static licenses are also installed on each machine acting as a job node.

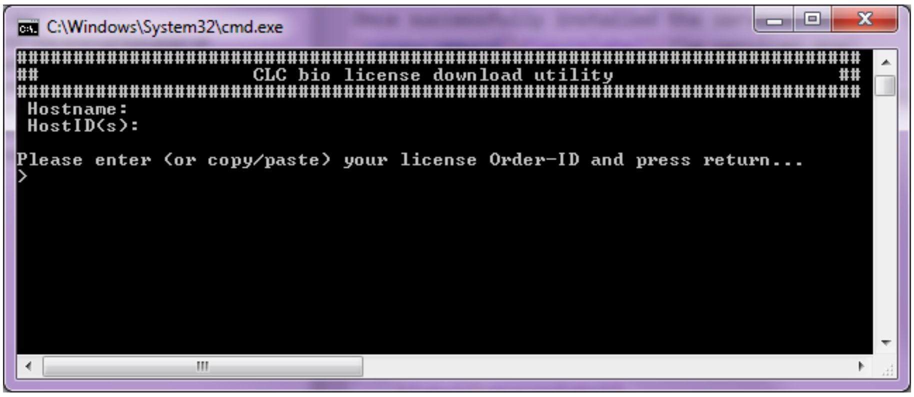
Static licenses for the Server CLC Workbench are downloaded and installed into the `licenses` folder in the *CLC Server* installation area. Downloading a license is similar for all supported platforms, but varies in certain details. Please see the platform-specific instructions below for details on how to download a license file on the system you are running the *CLC Server* on. See section 2.5.5 for a description on how to download a license for a machine that does not have access to the internet.

For the master machine and for each machine in a job node setup:

1. Log on to the machine that is running the *CLC Server*.
2. Move into the *CLC Server* installation directory, where the license download script can be found.
3. Download and install the CLC Workbench license as described in the relevant section below.

2.5.2 Windows license download

License files are downloaded using the `licensedownload.bat` script. To run the script, right-click on the file and choose **Run as administrator**. This will present a window as shown in figure 2.20.



```
############################################################################
##                   CLC bio license download utility                     ##
############################################################################
 Hostname:
 HostID(s):

Please enter (or copy/paste) your license Order-ID and press return...
>
```

Figure 2.20: *Download a license based on the Order ID.*

Paste the Order ID supplied by CLC bio (right-click to **Paste**) and press Enter. Please contact support@clcbio.com if you have not received an Order ID.

Note that if you are *upgrading* an existing license file, this needs to be deleted from the `licenses` folder. When you run the `downloadlicense.command` script, it will create a new license file.

2.5.3 Mac OS license download

License files are downloaded using the `downloadlicense.command` script. To run the script, double-click on the file. This will present a window as shown in figure 2.21.

Paste the Order ID supplied by CLC bio and press Enter. Please contact support@clcbio.com if you have not received an Order ID.

Note that if you are *upgrading* an existing license file, this needs to be deleted from the `licenses` folder. When you run the `downloadlicense.command` script, it will create a new license file.

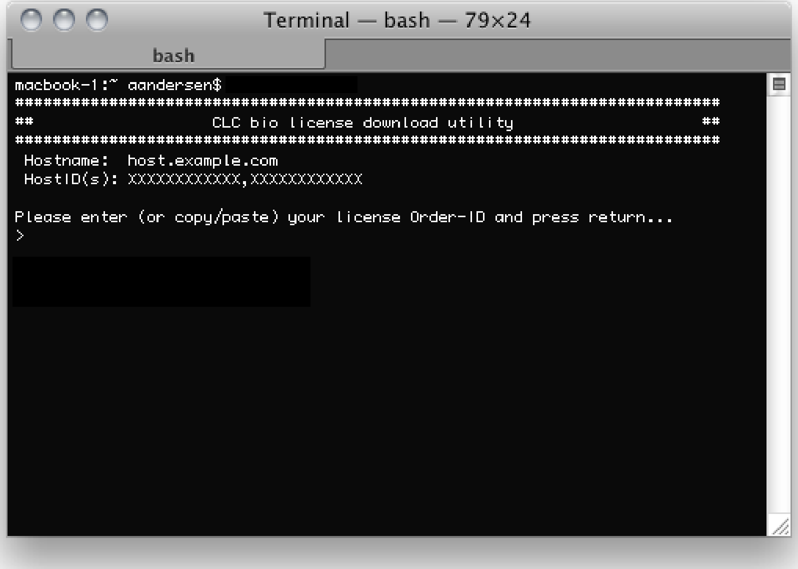
2.5.4 Linux license download

License files are downloaded using the `downloadlicense` script. Run the script and paste the Order ID supplied by CLC bio. Please contact support@clcbio.com if you have not received an Order ID.

Note that if you are *upgrading* an existing license file, this needs to be deleted from the `licenses` folder. When you run the `downloadlicense` script, it will create a new license file.

2.5.5 Download a static license on a non-networked machine

To download a static license for a machine that does not have direct access to the external network, you can follow the steps below after the Server software has been installed.



```
macbook-1:~ aandersen$ [redacted]
###########################################################################
##                   CLC bio license download utility                    ##
###########################################################################
 Hostname:  host.example.com
 HostID(s): XXXXXXXXXXXX,XXXXXXXXXXXX

Please enter (or copy/paste) your license Order-ID and press return...
>

[redacted]
```

Figure 2.21: *Download a license based on the Order ID.*

- Determine the host ID of the machine the server will be running on by running the same tool that would allow you to download a static license on a networked machine. The name of this tool depends on the system you are working on:

  - Linux: downloadlicense
  - Mac: downloadlicense.command
  - Windows: licensedownload.bat

  When you run the license download tool, the host ID for the machine you are working on will be printed to the terminal.

- Make a copy of this host ID such that you can use it on a machine that has internet access.

- Go to a computer with internet access, open a browser window and go to the relevant network license download web page:

  https://secure.clcbio.com/LmxWSv3/GetLicenseFile

- Paste in your license order ID and the host ID that you noted down earlier into the relevant boxes on the webpage.

- Click on 'download license' and save the resulting .lic file.

- Take this file to the machine with the host ID that you used when downloading the license file. Place it in the folder called 'licenses' that can be found within the CLC Server installation directory.

- Restart the CLC Server software.

2.5.6 Network license installation

Network licenses are necessary to run CLC Microbial Genomics Module analysis tasks on grid nodes. Network licenses are made available using a separate piece of software called the *CLC License Server*. This software is normally run as a service. CLC client software, such as Workbenches and gridworkers, contact the CLC License Server to obtain a network license when needed. For a description of how to download and install a license on a *CLC License Server*, please refer to the following section in the *CLC License Server* manual: http://clcsupport.com/clclicenseserver/current/index.php?manual=License_download.html

The same number of network plugin licenses as there are CLC gridworker licenses for the CLC Server setup are required. A license order ID is used when downloading a single license file. This license file includes information about how many network licenses are associated with the license order ID.

2.5.7 Server plugin download, installation and removal

1. Download the CLC Workbench plugin for the CLC Server as a .cpa file from http://www.clcbio.com/clc-plugin/#Server.

2. Install the plugin .cpa file on the *master* CLC Server using the Server web administrative interface. The plugin **should only be installed on the master server** in all server setup models. It does not need to be manually installed on any machine acting as an execution node.

   To install the plugin:

   (a) Go to the Plugins section under the **Admin** (⚙) tab (see figure 2.22).

   (b) Click on the **Browse** button and locate the .cpa file for the plugin to install.

   Logging into the web administrative interface is described in the CLC Server manual at:

   http://www.clcsupport.com/clcgenomicsserver/current/admin/index.php?manual=Logging_into_administrative_interface.html.

3. Restart the master CLC Server.

   Starting, stopping and restarting the CLC Server software is described in the CLC Server manual started at:

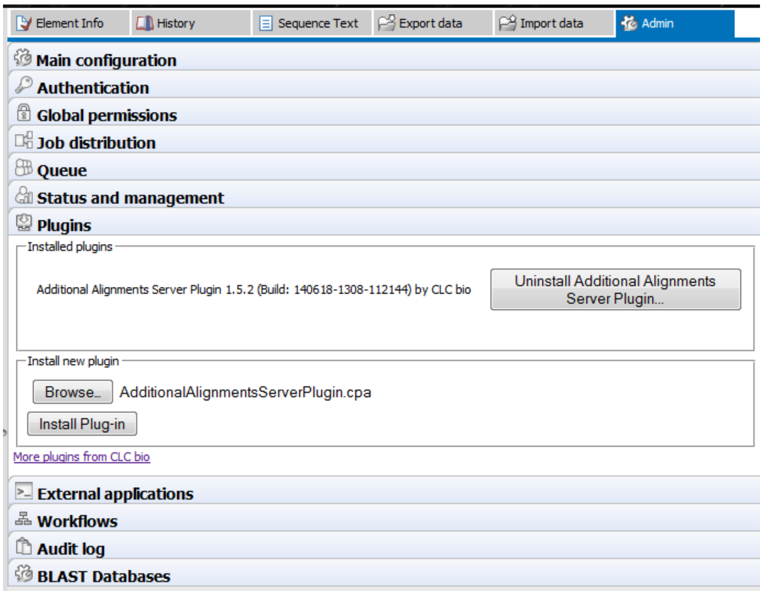
   http://www.clcsupport.com/clcgenomicsserver/current/admin/index.php?manual=Starting_stopping_server.html

4. **For job node setups only:**

   (a) Wait until the *master* CLC Server is up and running normally. Then restart each *job node* CLC Server so that the plugin is ready to run on each node.

   (b) In the web administrative interface on the *master* CLC Server, check that the plugin is enabled for each job node. This is described in more detail in the CLC Server manual at:

   http://www.clcsupport.com/clcgenomicsserver/current/admin/index.php?manual=Configuring_your_setup.html



Additional Alignments Server Plugin 1.5.2 (Build: 140618-1308-112144) by CLC bio

AdditionalAlignmentsServerPlugin.cpa

Figure 2.22: *Installing and uninstalling CLC Server plugins is done via the Plugins section of the web administrative interface.*

To uninstall a CLC Server plugin, simply click on the button that has Uninstall on its label next to the relevant plugin.

Chapter 3

OTU clustering

3.1 Download OTU Reference Database

OTU reference databases contain representative OTU sequences and their taxonomy. They are needed to perform reference-based OTU clustering. Three popular reference OTU databases, clustered at various similarity percentages, can be downloaded using the **Download OTU Reference Database** tool:

- Greengenes: 16S rRNA gene from ribosomal Small Subunit for Prokaryotic taxonomic assignment clustered OTUs at different percentages. http://greengenes.secondgenome.com/downloads
- Silva/Arb SSU: 16S/18S rRNA from ribosomal Small Subunit for Prokaryotic and Eukaryotic taxonomic assignment clustered OTUs at different percentages. http://www.arb-silva.de/download/archive/qiime/
- UNITE: ITS spacer clustered OTUs at different percentages for fungal taxonomic assignment. https://unite.ut.ee/repository.php

To run the tool, go to **Toolbox** | **OTUclustering ()** | **Download OTU Reference Database ()**, select the database needed and specify where to save it. When using this tool, the databases downloaded are automatically formatted. If you wish to format your own database, use the tool called **Format Reference Database**.

3.2 Format Reference Database

In addition to the download of reference databases using the **Download OTU Reference Database** tool, you can format your own databases by running the Format Reference Database tool:

Toolbox | **OTUclustering ()** | **Format Reference Database ()**.

This tool takes as input QIIME-format fasta files (typically clustered into OTUs at 99, 97, 94, and 90 percent similarity) and corresponding taxonomy mapping files. Each line of the taxonomy file should contain an OTU name and its taxonomy, where taxonomy levels are separated with semicolons. For example, the following line

```
o123  k__Bacteria;p__Bacteroidetes;c__Sphingobacteria;o__;f__;g__;s__.
```

indicates that the OTU o123 belongs to the class Sphingobacteria and that its taxonomy is specified only up to the class level.

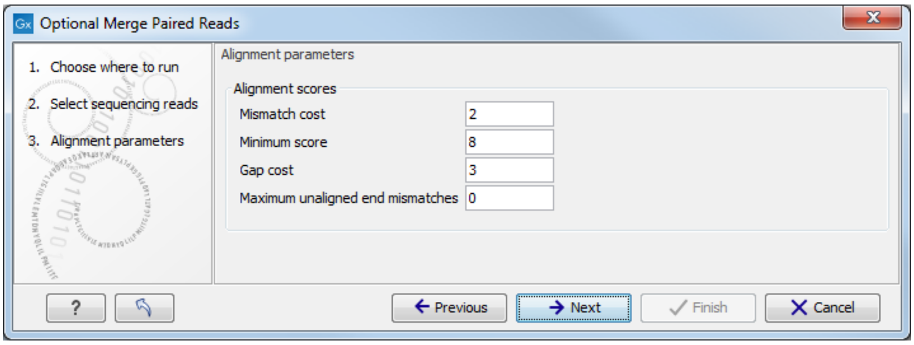
3.3 Optional Merge Paired Reads

In order to use the highest quality sequences for clustering, it is recommended to merge paired read data. If the read length is smaller than the amplicon size, forward and reverse reads are expected to overlap in most of their 3' regions. Therefore, one can merge the forward and reverse reads to yield one high quality representative using the Optional Merge Paired Reads tool. The Optional Merge Paired Reads tool will merge paired-end reads according to some pre-selected merge parameters: the overlap region and the quality of the sequences. For example, for a designed 150 bp overlap, a maximum score of 150 is achievable, but as the real length of the overlap is unknown, a lower minimum score should be chosen. Also, some mismatches and InDels should be allowed, especially if the sequence quality is not perfect. You can also set penalties for mismatch, gap and unaligned ends.

To run the Optional Merge Paired Reads tool, go to **Toolbox** | **OTUclustering** ( ) | **Optional Merge Paired Reads** ( ).

Select any number of sequences as input. The tool accepts both paired and unpaired reads but will only merge the paired reads while returning the unpaired ones as "not merged" reads in the output. Note that paired reads have to be in forward-reverse orientation. After merging, the merged reads will always be in the forward orientation.

Click Next to open the dialog shown in figure 3.1.



Optional Merge Paired Reads

1. Choose where to run
2. Select sequencing reads
3. Alignment parameters

Alignment parameters

Alignment scores

Mismatch cost: 2

Minimum score: 8

Gap cost: 3

Maximum unaligned end mismatches: 0

Figure 3.1: *Alignment parameters.*

In order to understand how these parameters should be set, an explanation of the merging algorithm is needed: Because the fragment size is not an exact number of base pairs and is different from fragment to fragment, an alignment of the two reads has to be performed. If the alignment is *good and long enough*, the reads will be merged. *Good enough* in this context means that the alignment has to satisfy some user-specified score criteria (details below). Because of sequencing errors that typically are more abundant towards the end of the read, the alignment is not expected always to be perfect, and the user can decide how many errors are acceptable. *Long enough* in this context means that the overlap between the reads has to be non-coincidental. Merging two reads that do not really overlap, leads to errors in the downstream analysis, thus it is very important to make sure that the overlap is big enough. If only a few bases overlap was required, some read pairs will match by chance, so this has to be avoided.

The following parameters are used to define what is *good enough* and *long enough*.

- **Mismatch cost**: The alignment awards one point for a match, and the mismatch cost is set by this parameter. The default value is 2.

- **Minimum score**: This is the minimum score of an alignment to be accepted for merging. The default value is 8. As an example: with default settings, this means that an overlap of 11 bases with one mismatch will be accepted (10 matches minus 2 for a mismatch).

- **Gap cost**: This is the cost for introducing an insertion or deletion in the alignment. The default value is 3.

- **Maximum unaligned end mismatches**: The alignment is local, which means that a number of bases can be left unaligned. If the quality of the reads is dropping to be very poor towards the end of the read, and the expected overlap is long enough, it makes sense to allow some unaligned bases at the end. However, this should be used with great care which is why the default value is 0. As explained above, a wrong decision to merge the reads leads to errors in the downstream analysis, so it is better to be conservative and accept fewer merged reads in the result. Please note that even with the alignment scores above the minimum score specified in the tool setup, the paired reads also need to have the number of end mismatches below the "Maximum unaligned end mismatches" value specified in the tool setup to be qualified for merging.

The main result will be two sequence lists for each sample selected as input to the tool: one containing the merged reads (labeled as "merged"), and one containing the reads that could not be merged (labeled as "not merged"). Note that low quality can be one of the reasons why a pair cannot be merged. Hence, the list of reads that could not be paired is more likely to contain more reads with errors than the one with the merged reads.

3.4 Fixed Length Trimming

In order to compare sequences and cluster them, they all need to be of exact same length. All reads which are shorter than the cut-off are discarded, and reads longer than that are trimmed back to the chosen length.

To run the tool, go to **Toolbox** | **OTUclustering ()** | **Fixed Length Trimming ()** and select the sequences you would like to trim.

In the next wizard window you can enter manually the desired length for the trimmed reads. Alternatively, the **Fixed length Trimming** algorithm can calculate the trimming cut-off value as the mean length of the merged reads minus one standard deviation. If this option is chosen, it is important that all samples are trimmed at the same time as the mean and standard deviation for the combined reads in all samples needs to be estimated at once.

You can also offset one adapter or barcode by typing the nucleotide sequence in the Primer offset window. Exact matching is performed but ambiguous symbols are allowed. For more than one adapter, we recommend to perform prior to the Fixed Length trimming a **Trim Sequences** step from the NGS Core Tools in which you can import an adapter list (see the CLC Microbial Genomics Module tutorial, or the Adapter Trimming section of the CLC Genomics Workbench manual (http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Adapter_trimming.html).

3.5 Filter Samples Based on the Number of Reads

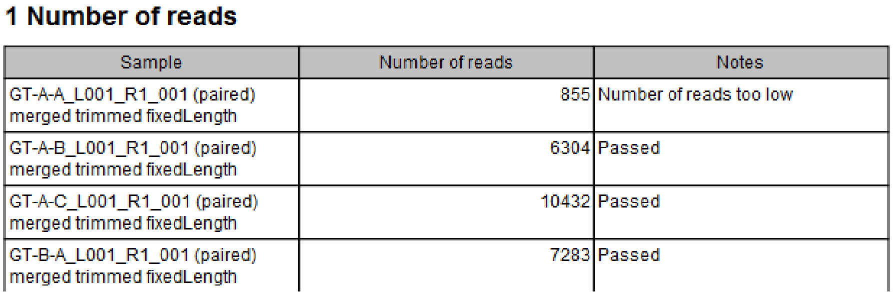
In order to cluster accurately samples, they should have comparable coverage. Sometimes, however, DNA extraction, PCR amplification, library construction or sequencing has not been entirely successful, and a fraction of the resulting sequencing data will be represented by too few reads. These samples should be excluded from further analysis using the **Filter Samples Based on the Number of Reads** tool.

To run the tool, go to **Toolbox** | **OTUclustering ( )** | **Filter Samples Based on the Number of Reads ( )**.

The tool requires that the input reads from each sample must be either all paired or all single. This check ensures that the samples are comparable, as the number of reads before merging paired reads is twice as great as the number of merged reads. The preferred way to run this tool with OTU sequencing data is to use merged reads obtained from the **Optional Merge Paired Reads** tool.

The threshold for determining whether a sample has sufficient coverage is specified by the parameters **minimum number of reads** and **minimum percent from the median**. The algorithm filters out all samples whose number of reads is less than the **minimum number of reads** or less than the **minimum percent from the median** times the median number of reads across all samples.

The primary output is a table describing how many reads are in a particular sample and if they passed or failed the quality control (see figure 3.2).



1 Number of reads

Sample	Number of reads	Notes
GT-A-A_L001_R1_001 (paired) merged trimmed fixedLength	855	Number of reads too low
GT-A-B_L001_R1_001 (paired) merged trimmed fixedLength	6304	Passed
GT-A-C_L001_R1_001 (paired) merged trimmed fixedLength	10432	Passed
GT-B-A_L001_R1_001 (paired) merged trimmed fixedLength	7283	Passed

Figure 3.2: *Output table from the Filter Samples Based on the Number of Reads tool.*

In the next wizard window you can decide to **Copy samples with sufficient coverage** as well as to **Copy the discarded samples**. Copying the samples with sufficient coverage will give you a new list of sequences that you can use in your following analyses because it does not contain the reads of poor quality that failed the Remove the samples with Low Coverage analysis.

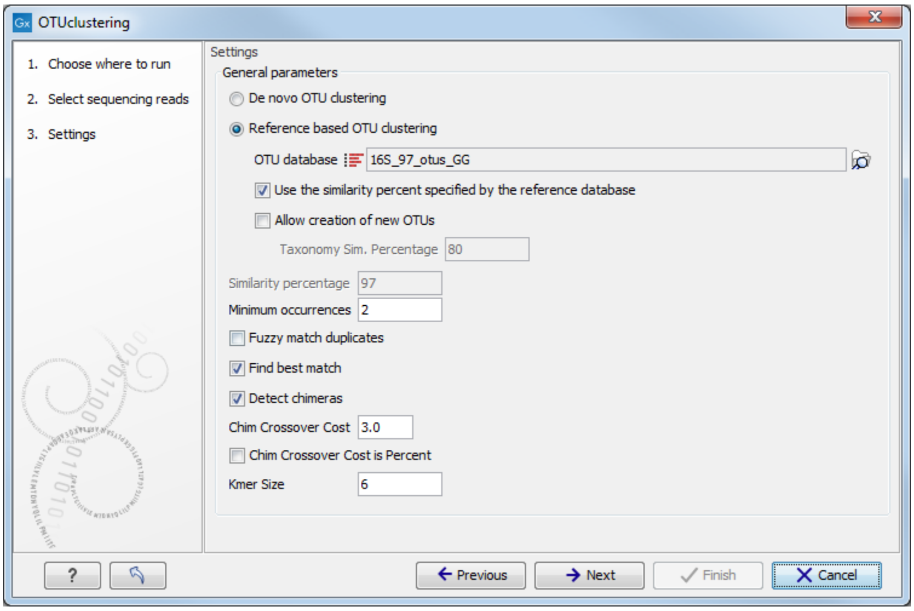
3.6 OTUclustering

The OTUclustering tool clusters a collection of fixed length trimmed reads to operational taxonomy units.

To run the tool, go to **Toolbox** | **OTUclustering ( )** | **OTUclustering ( )**.

3.6.1 OTUclustering parameters

After having selected the sequences you would like to cluster, the wizard offers to set some general parameters (see figure 3.3).



OTUclustering

1. Choose where to run
2. Select sequencing reads
3. Settings

Settings

General parameters

○ De novo OTU clustering

● Reference based OTU clustering

OTU database: 16S_97_otus_GG

☑ Use the similarity percent specified by the reference database

□ Allow creation of new OTUs

Taxonomy Sim. Percentage: 80

Similarity percentage: 97

Minimum occurrences: 2

□ Fuzzy match duplicates

☑ Find best match

☑ Detect chimeras

Chim Crossover Cost: 3.0

□ Chim Crossover Cost is Percent

Kmer Size: 6

Figure 3.3: *Outputtable from the Remove Low Coverage Samples tool.*

You can choose to perform a **De novo OTU clustering**, or you can perform a **Reference based OTU clustering**. The following parameters can then be set:

- **OTU database** Specify here the reference database to be used for Reference based OTU clustering. Reference databases can be created by the **Download OTU Reference Database** or the **Format Reference Database** tools.

- **Use the similarity percent specified by the reference database** Allows to use the same similarity percent value (see below) that was used when creating the reference database. This parameter is available only when performing a reference based OTU clustering. Selecting this parameter will disable the similarity percent parameter.

- **Allow creation of new OTUs** Allows sequences which are not already represented at the given similarity distance in the database to form a new cluster, and a new centroid is chosen. This parameter can be set only when performing a Reference based OTU clustering. Disallowing the creation of new OTUs is also known as closed reference OTU picking.

- **Taxonomy similarity percentage** Specifies the similarity percentage to be used when annotating new OTUs. This parameter is available only when **Allow creation of new OTUs** is selected.

- **Similarity percentage**: Specifies the required percentage of identity between a read and the centroid of an OTU for the read to join the OTU cluster.

- **Minimum occurrences**: Specifies the minimum number of duplicates for specific read-data before it will be included in further analyses. For instance, if set to 2, at least two reads with the same exact nucleotides needs to exist in the input for the data to propagate to further analysis. Other data will be thrown away. This can for instance be used to filter out singletons. Note that matches does not need to be exact when the **Fuzzy match duplicates** option is used.

- **Fuzzy match duplicates**: Specifies how duplicates are defined. If the option is not selected two reads are only duplicates if they are exactly equal. If the option is selected, two reads are duplicates if they are almost equal, i.e. all differences are SNVs and there are not too many of them ($\leq 2\%$). This pseudo-merging is done by lexicographically sorting the input and looking in the neighborhood of the read being processed. The reads are processed from most abundant (in a completely equivalent sense) to the least. In this way two singletons can for instance be pseudo-merged together and be included for further study despite the **Minimum occurrences** option having specified 2. Upon further analysis a group can be split into several OTUs if not all members are within the specified threshold from the ''OTU-leader''.

- **Find best match**: If the option is not selected, the read becomes a member of the first OTU-database entry found within the specified threshold. If the option is selected all database entries are tested and the read becomes a member of the best matching result. Note that ''first'' and ''all'' are relative terms in this case as kmer-searches are used to speed up the process. ''All'' only includes the database entries that the kmer search deems close enough, i.e., database entries that cannot be within the specified threshold will be filtered out at this step. ''First'' is the first matching entry as returned by the kmer-search which will sort by the number of kmer-matches.

- **Detect chimeras**: Chimeric sequences are frequent artifacts of PCR reactions. They are detected by assessing whether a sequence is likely to have two different and more frequent "parent" sequences in the current collection of OTUs, meaning that two fragments of the sequence map to different sequences.

- **Chimera crossover cost**: The cost of doing a chimeric crossover, i.e. the higher the cost the less likely it is that a read is marked as chimeric.

- **Chimera crossover cost is percent**: Whether the above cost is an absolute value or a percentage value (i.e. an absolute value will be automatically calculated based on the read length).

- **Kmer size**: The size of the kmer to use in regards to the kmer usage in finding the best match.

Chimera detection is performed by kmer searches as follows:

- All database entries and the read being processed are split into 4 equally sized portions with an additional 3 half-way-shifted to cover the merge-points. This results in 7 different kmer-search-options. Each of these is queried for matches within some threshold, and only the results are processed further. A database entry has to fit well in at least one of these 7 portions for the database-entry to be relevant.

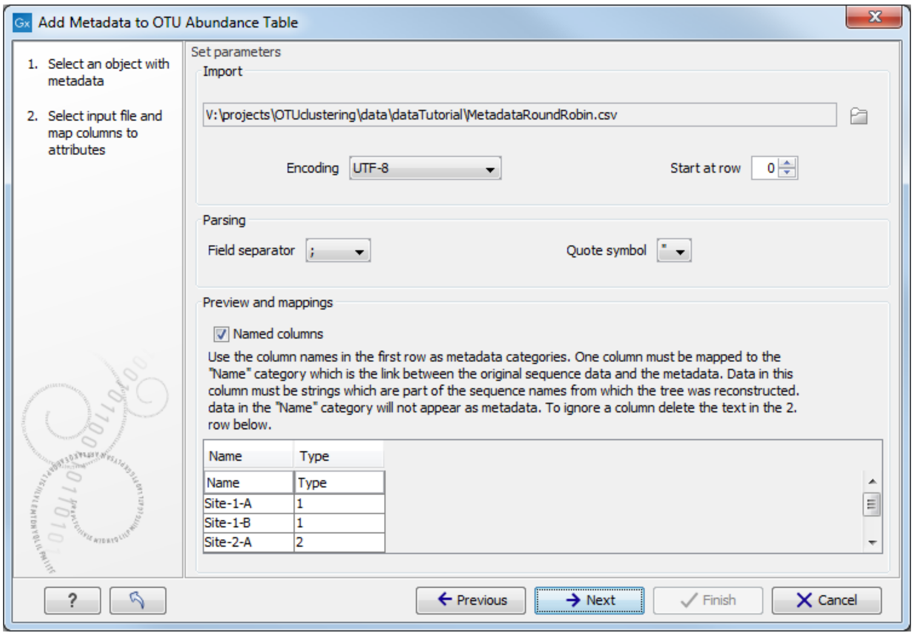
- Given the 7 sets of database entries some entries may just be present in one of the sets. For these entries, some may be duplicates in the region they represent. The duplicates are filtered out and only an arbitrary representative of the duplicates is kept.

The OTU clustering tool produces several outputs: a sequence list of the OTU centroids and/or of the Chimeras, and abundance tables with the newly created OTUs and/or the chimeras. Each table give abundance of the OTU or chimeras at each site, as well as the total abundance for all samples.

3.6.2 Add Metadata to OTU Abundance Table

In order to enhance the visualization functionality of the OTU abundance table, it is useful to decorate it with metadata on the samples. To run the tool go to: **Toolbox** | **OTUclustering** (**) | **Add Metadata to OTU Abundance Table** (**)**

Choose an OTU table as input. In the next wizard window you can select a file describing the metadata on your local computer (figure 3.4).


Add Metadata to OTU Abundance Table
1. Select an object with metadata
2. Select input file and map columns to attributes
Set parameters
Import
Parsing
Preview and mappings
?
Previous
Next
Finish
Cancel

V:\projects\OTUclustering\data\dataTutorial\MetadataRoundRobin.csv

Encoding UTF-8

Start at row 0

Field separator ;

Quote symbol "

☑ Named columns

Use the column names in the first row as metadata categories. One column must be mapped to the "Name" category which is the link between the original sequence data and the metadata. Data in this column must be strings which are part of the sequence names from which the tree was reconstructed. data in the "Name" category will not appear as metadata. To ignore a column delete the text in the 2. row below.

Name	Type
Name	Type
Site-1-A	1
Site-1-B	1
Site-2-A	2

Figure 3.4: *Setting up metadata parameters.*

The metadata should be formatted in a tabular file format (.xls, .xlsx, .csv). The first row of the table should contain column headers. There should be one column called ''Name'' and the entries in this column should match the names of the reads selected for OTUclustering. This column is used to match row in the table with samples present in the OTU table, so if the names do not match you will not be able to aggregate your data at all. There can be as many other columns as needed, and these information can be used as grouping variables to improve visualization of the results or to perform additional statistical analyses. If you wish to ignore a column without deleting it from your file, simply delete the text in the header row.

3.6.3 Remove OTUs with Low Abundance

Low abundance OTUs can eliminated from the OTU table if they have fewer than a given count across all the samples in the experiment.

To run the tool, go to **Toolbox** | **OTUclustering** (**) | **Remove OTUs with Low Abundance** (**).

Choose an OTU table as input, select the filtering parameters and save the table. The threshold for determining whether an OTU has sufficient abundance is specified by the parameters **minimum combined abundance** and **minimum combined abundance (% of all the reads)**. The algorithm filters out all OTUs whose combined abundance across all samples is less than the minimum combined abundance or whose combined abundance is less than the minimum combined abundance (% of all the reads) across all samples. The default value for the Minimum combined abundance is set at 10.

3.6.4 Visualization of the OTU abundance table

The OTU clustering tool produces several outputs: a sequence list of the OTU centroids or of the Chimeras, and abundance tables with the newly created OTUs or the chimeras. Each table give abundance of the OTU or chimeras at each site as well as the total abundance for all samples. There are a number of ways of visualizing the contents of an OTU abundance table:

- **Table view** (▦) The table display the following columns:
  - **Name** The name of the OTU, specified by either the reference database or by the OTU representative.
  - **Taxonomy** The taxonomy of the OTU, as specified by the reference database when a database entry was used as Reference.
  - **Combined Abundance** The total number of reads belonging to the OTU across all samples.
  - **Min** Minimum abundance across all samples
  - **Max** Maximum abundance across all samples
  - **Mean** Mean abundance of all samples
  - **Median** Median abundance of all samples
  - **Std** Standard deviation of all samples
  - **Abundance for each sample** The number of reads belonging to the OTU in a specific sample.
  - **Sequence** The sequence of the centroid of the OTU.

  Under the tab Data in the right side panel, you can switch between absolute counts and relative abundances. You can also combine absolute counts and relative abundances by taxonomic levels by selecting the appropriate phylum in the **Aggregate taxonomy** drop-down menu. Finally, if you have previously annotated your table with Metadata (see section 3.6.2), you can **Aggregate sample** by the groups previously defined in your metadata table. This is useful when analyzing replicates from the same sample origin.

- **Stacked Bar Chart and Stacked Area Chart** (▥) In the Stacked Bar (figure 3.5) and Stacked Area Charts (figure 3.6), the metadata can be used to aggregate groups of columns (samples) by selecting the relevant metadata category in the right hand side panel. Also, the data can be aggregated at any taxonomy level selected. The relevant data points will automatically be summed accordingly.

  Holding the pointer over a colored area in any of the plots will result in the display of the corresponding taxonomy label and counts. The slider **Filter level** allows the setting of a viewing filter on the minimum counts for individual viewing instead of being aggregated in the "Other" category displayed in the top field of the chart. One can select which taxonomy level to color, and change the default colors manually. Colors can be be specified at the same taxonomy level as the one use to aggregate the data or at a lower level. When lower taxonomy levels are chosen in the data aggregation field, the color will be inherited in alternating shadings. Using the bottom right-most button (**Save/restore settings** (☰)), the settings can be saved and applied in other plots, allowing visual comparisons across analyses.

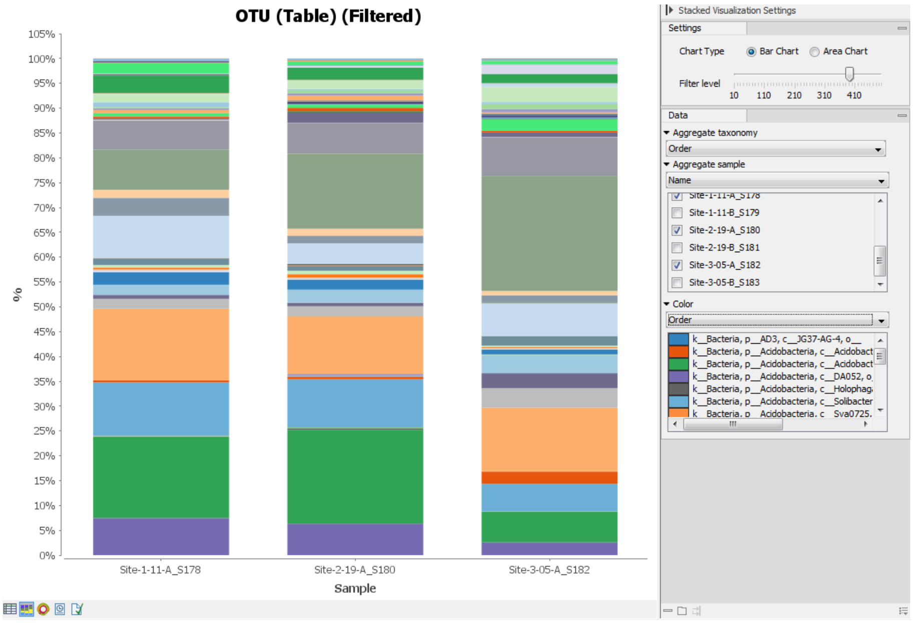


Chart Type ● Bar Chart ○ Area Chart

Filter level

Aggregate taxonomy: Order

Aggregate sample: Name

☑ Site-1-11-A_S178
☐ Site-1-11-B_S179
☑ Site-2-19-A_S180
☐ Site-2-19-B_S181
☑ Site-3-05-A_S182
☐ Site-3-05-B_S183

Color: Order

- k__Bacteria, p__AD3, c__JG37-AG-4, o__
- k__Bacteria, p__Acidobacteria, c__Acidobact
- k__Bacteria, p__Acidobacteria, c__Acidobact
- k__Bacteria, p__Acidobacteria, c__DA052, o
- k__Bacteria, p__Acidobacteria, c__Holophag
- k__Bacteria, p__Acidobacteria, c__Solibacter
- k__Bacteria, p__Acidobacteria, c__Sva0725.

Figure 3.5: *Stacked bar of the microbial community at the order level for 3 different sites.*

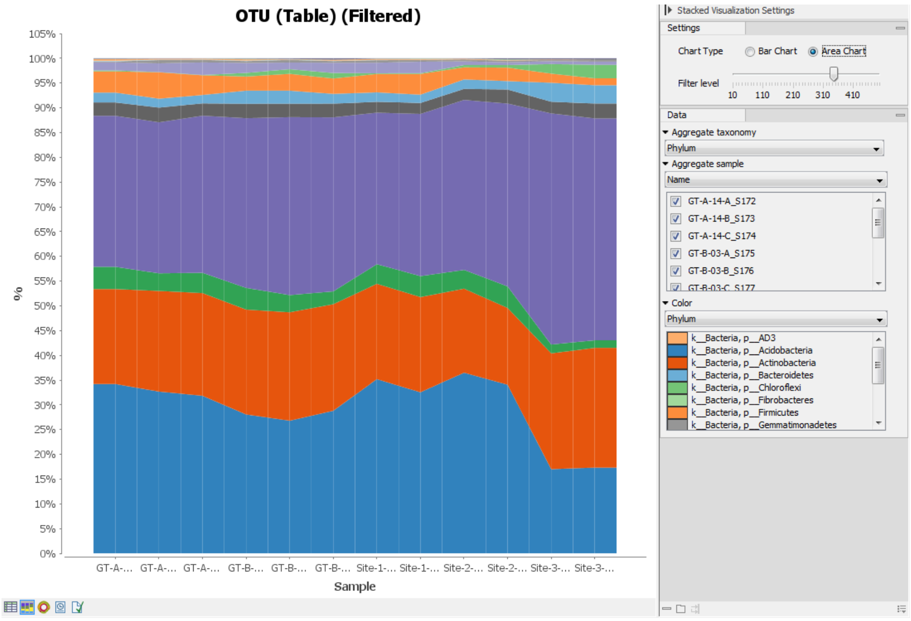


Chart Type ○ Bar Chart ● Area Chart

Filter level

Aggregate taxonomy: Phylum

Aggregate sample: Name

☑ GT-A-14-A_S172
☑ GT-A-14-B_S173
☑ GT-A-14-C_S174
☑ GT-B-03-A_S175
☑ GT-B-03-B_S176
☑ GT-B-03-C_S177

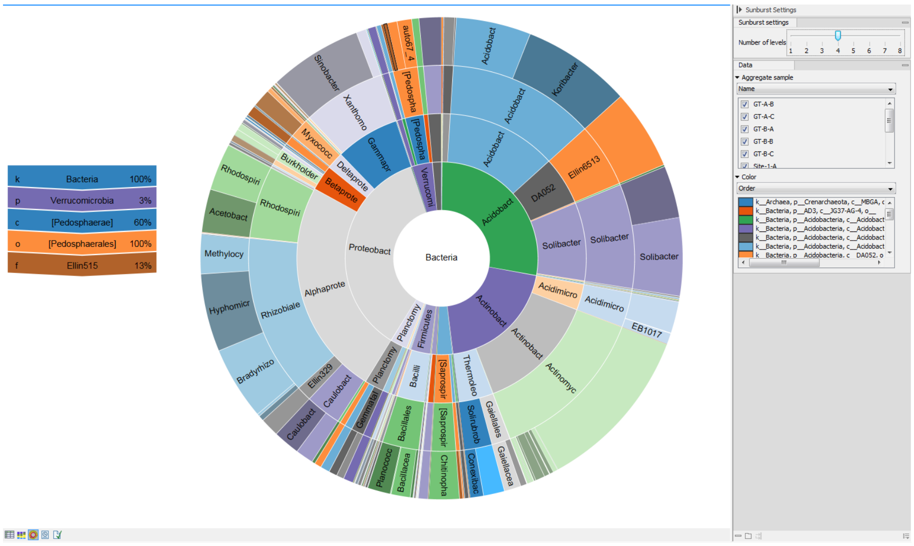
Color: Phylum

- k__Bacteria, p__AD3
- k__Bacteria, p__Acidobacteria
- k__Bacteria, p__Actinobacteria
- k__Bacteria, p__Bacteroidetes
- k__Bacteria, p__Chloroflexi
- k__Bacteria, p__Fibrobacteres
- k__Bacteria, p__Firmicutes
- k__Bacteria, p__Gemmatimonadetes

Figure 3.6: *Stacked area of the microbial community at the phylum level for 12 different sites.*

- **Zoomable Sunbursts** (◎) The Zoomable Sunburst viewer lets the user select how many taxonomy level counts to display, and which level to color. Lower levels will inherit the color in alternating shadings. Taxonomy and relative abundances are displayed in a legend to the left of the plot when hovering over the sunburst viewer with the mouse. The metadata can be used to select which sample or group of samples to show in the sunburst (figure 3.7).

  Clicking on a lower level field will render that field the center of the plot and display lower level counts in a radial view. Clicking on the center field will render the level above the current view the center of the view (figure 3.8).



k	Bacteria	100%
p	Verrucomicrobia	3%
c	[Pedosphaerae]	60%
o	[Pedosphaerales]	100%
f	Ellin515	13%

Sunburst Settings

Sunburst settings

Number of levels: 1 2 3 4 5 6 7 8 (slider at 4)

Data

Aggregate sample

Name

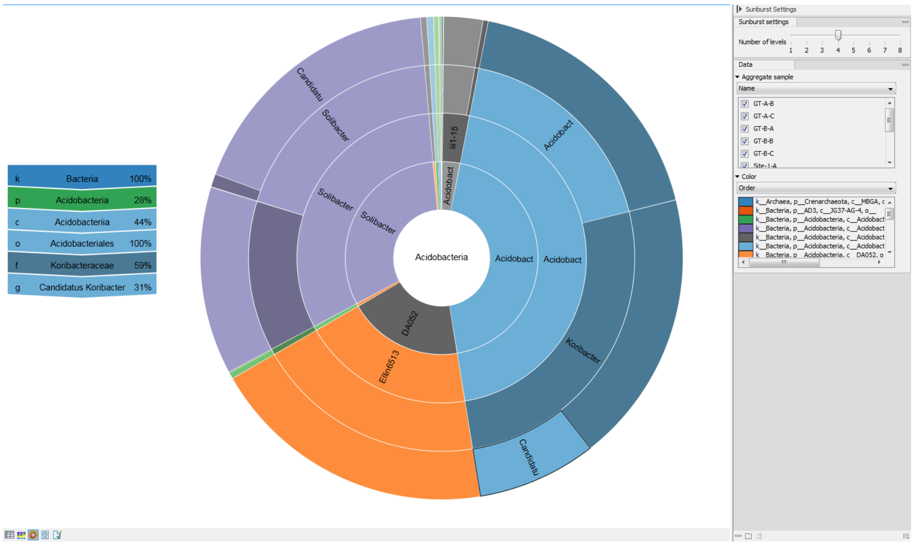
- ☑ GT-A-B
- ☑ GT-A-C
- ☑ GT-B-A
- ☑ GT-B-B
- ☑ GT-B-C
- ☑ Site-1-A

Color

Order

- k__Archaea, p__Crenarchaeota, c__MBGA, c
- k__Bacteria, p__AD3, c__JG37-AG-4, o__
- k__Bacteria, p__Acidobacteria, c__Acidobact
- k__Bacteria, p__Acidobacteria, c__Acidobact
- k__Bacteria, p__Acidobacteria, c__Acidobact
- k__Bacteria, p__Acidobacteria, c__Acidobact
- k__Bacteria, p__Acidobacteria, c__DA052, o

Figure 3.7: *Sunburst view of the microbial community showing all taxa belonging to the kingdom bacteria.*



k	Bacteria	100%
p	Acidobacteria	28%
c	Acidobacteriia	44%
o	Acidobacteriales	100%
f	Koribacteraceae	59%
g	Candidatus Koribacter	31%

Sunburst Settings

Sunburst settings

Number of levels: 1 2 3 4 5 6 7 8 (slider at 4)

Data

Aggregate sample

Name

- ☑ GT-A-B
- ☑ GT-A-C
- ☑ GT-B-A
- ☑ GT-B-B
- ☑ GT-B-C
- ☑ Site-1-A

Color

Order

- k__Archaea, p__Crenarchaeota, c__MBGA, c
- k__Bacteria, p__AD3, c__JG37-AG-4, o__
- k__Bacteria, p__Acidobacteria, c__Acidobact
- k__Bacteria, p__Acidobacteria, c__Acidobact
- k__Bacteria, p__Acidobacteria, c__Acidobact
- k__Bacteria, p__Acidobacteria, c__Acidobact
- k__Bacteria, p__Acidobacteria, c__DA052, o

Figure 3.8: *Sunburst view of the microbial community zoomed to show all taxa belonging to the phylum bacteroidetes.*

Chapter 4

Estimation of Alpha and Beta diversity

Two levels of diversity are typically considered in microbial ecology: alpha- and beta-diversity. Alpha-diversity estimates describe the number of species (or similar metrics) in a single sample, whereas beta-diversity compares the number of species (or similar metrics) across samples [Whittaker, 1972].

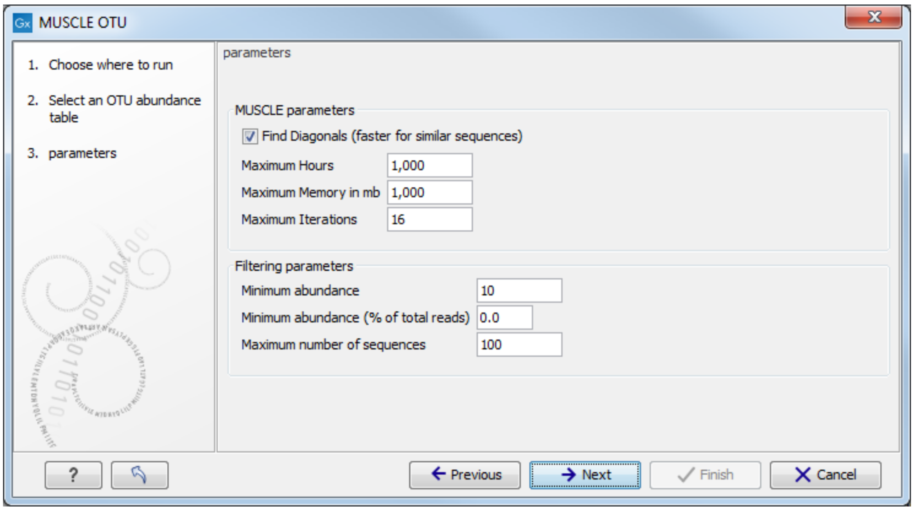
Some measures of estimate alpha and beta diversity require a phylogenetic tree of all OTUs. The phylogenetic tree is reconstructed based on a multiple sequence alignment (MSA) of the OTU sequences. Therefore, as a pre-requisite, a MSA needs to be created and a phylogeny reconstructed.

4.1 Align OTUs with MUSCLE

In order to estimate Alpha and Beta diversity, you must use first use the Align OTUs with MUSCLE tool of the CLC Microbial Genomics Module:

Toolbox | OTUclustering ( ) | Align OTUs using MUSCLE ( ).

Choose an OTU abundance table as input. the next wizard window allows you to set up the alignment parameters with MUSCLE (figure 4.1).



MUSCLE parameters

☑ Find Diagonals (faster for similar sequences)

Maximum Hours 1,000

Maximum Memory in mb 1,000

Maximum Iterations 16

Filtering parameters

Minimum abundance 10

Minimum abundance (% of total reads) 0.0

Maximum number of sequences 100

Figure 4.1: *Set up parameters for aligning sequences with MUSCLE.*

- **Find Diagonals**: you can decide on some restrictive parameters for your analysis: the **Maximum Hours** the analysis should last, the **Maximum Memory in mb** that should be used for the analysis, or the **Maximum Iterations** the analysis should make. The latter is set to 16 by default.
- **Filtering Parameters**: The algorithm filters out all OTUs whose combined abundance across all samples is less than the **minimum combined abundance** or whose combined abundance is less than the **minimum combined abundance (% of all the reads)** across all samples. The default value for the Minimum combined abundance is set at 10. Moreover, you can specify the **Maximum number of sequences to be aligned**, so that only the sequences with the highest combined abundances will be used. Note that reducing the number of sequences will speed up the alignment and the construction of phylogeny trees.

Note that by default only the top 100 most abundant OTUs are aligned using MUSCLE and used to reconstruct the phylogeny tree in the next step. This phylogenetic tree is used for calculating the phylogenetic diversity and the UniFrac distances, so these measures disregard the low abundance OTUs by default. If more OTUs are to be included, the default settings for the MUSCLE alignment need to be changed accordingly.

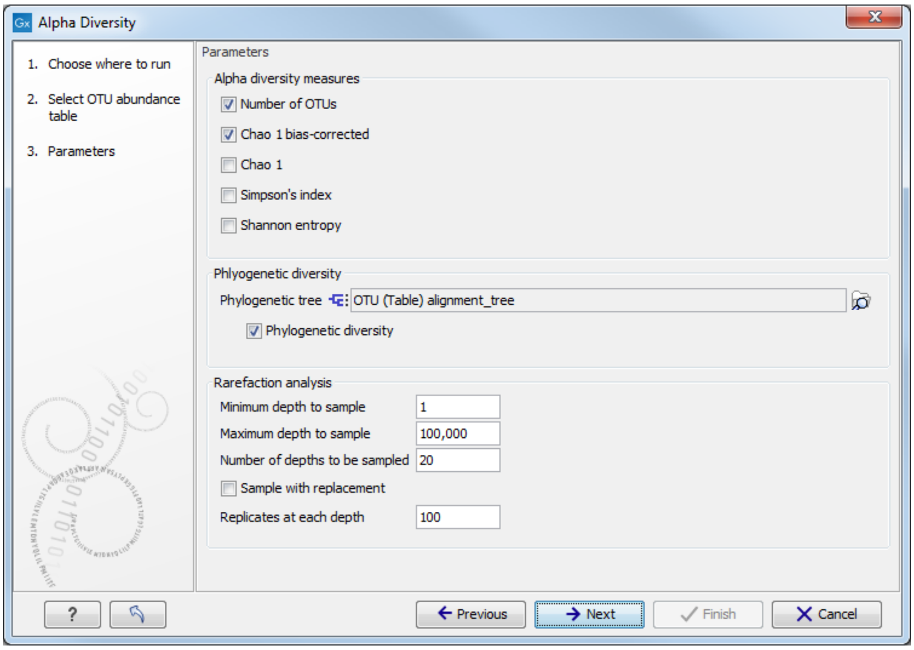
For further analysis with the Alpha and Beta diversity tools, save the alignment and construct a phylogeny using the Maximum Likelihood Phylogeny tool from CLC Workbench core tools in **Toolbox** | **Classical Sequence Analysis ()** | **Alignments and Trees ()** | **Maximum Likelihood Phylogeny ()**. Users of the Biomedical Genomics Worbench will find the tool under **Toolbox** | **OTUclustering ()**. For more information, see http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Maximum_Likelihood_Phylogeny.html.

4.2 Alpha Diversity

Alpha diversity is the diversity within a particular area or ecosystem; usually expressed by the number of species (i.e., species richness) in that ecosystem. Alpha diversity estimates are dependent on sampling depth, and hence rarefaction analysis is integral to this analysis step.

To run the tool go to **Toolbox** | **OTUclustering ()** | **Alpha Diversity ()**. Choose an OTU table to use as input. The next wizard window offers you to set up different analysis parameters (figure 4.2). For example, you can select which diversity measures to calculate (see section 4.2.1), specify the appropriate phylogenetic tree for computing phylogenetic diversity, and parameterize the rarefaction analysis.

The rarefaction analysis is done by sub-sampling the OTU abundances in the different samples at different depths. The range of depths to be sampled is defined by the parameters **Minimum depth to sample** and **Maximum depth to sample**. If the maximum depth is set to 0, the number of reads of the most abundant sample is used. The number of different depths to be sampled is specified by the **Numbers of depths to be sampled parameters**. For example, if you choose to sample 5 depths between 1000 and 5000, the algorithm will sub-sample each sample at 1000, 2000, 3000, 4000, and 5000 reads. At each depth, the algorithm subsamples the data several times, according to the **Replicates at each depth**. You can choose whether the sampling should be performed with or without replacement by setting the **Sample with replacement** parameter.



Parameters

Alpha diversity measures

☑ Number of OTUs

☑ Chao 1 bias-corrected

□ Chao 1

□ Simpson's index

□ Shannon entropy

Phylogenetic diversity

Phylogenetic tree: OTU (Table) alignment_tree

☑ Phylogenetic diversity

Rarefaction analysis

Minimum depth to sample: 1

Maximum depth to sample: 100,000

Number of depths to be sampled: 20

□ Sample with replacement

Replicates at each depth: 100

Figure 4.2: *Set up parameters for the Alpha Diversity tool.*

4.2.1 Alpha diversity measures

The available diversity measures are:

- Number of OTUs: The number of OTUs observed in the sample.
- Chao 1 bias-corrected: Chao1-bc $= D + \frac{f_1(f_1-1)}{2(f_2+1)}$.
- Chao 1: Chao1 $= D + \frac{f_1^2}{2f_2}$.
- Simpson's index: SI $= 1 - \sum_{i=1}^{n} p_i^2$.
- Shannon entropy: H $= \sum_{i=1}^{n} p_i \log_2 p_i$.

where n is the number of OTUs; D is the number of distinct OTUs observed in the sample; f_1 is the number of OTUs for which only one read has been found in the sample; f_2 is the number of OTUs for which two reads have been found in the sample; and p_i is the fraction of reads that belong to OTU i.

If a phylogenetic tree is provided as input, the following distance is also available:

- Phylogenetic diversity: $PD = \sum_{i=1}^{n} b_i I(p_i > 0)$

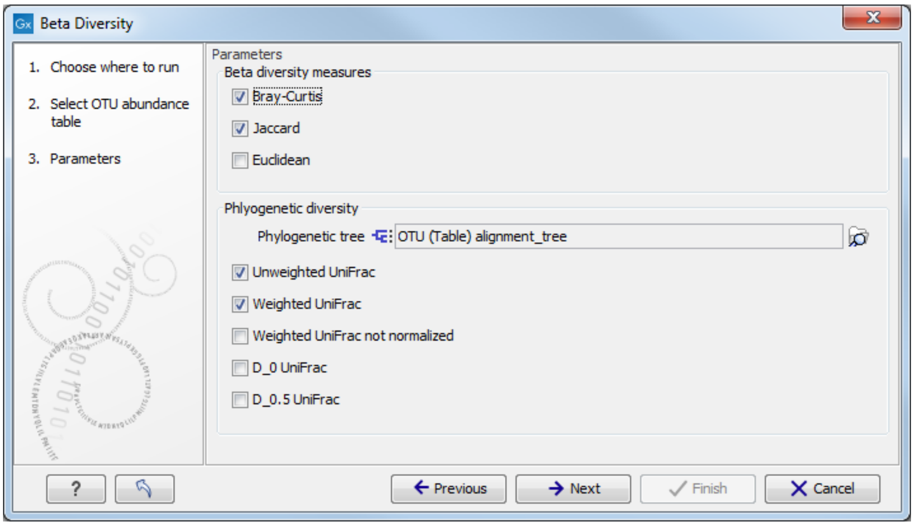
where n is the number of branches in the phylogenetic tree, b_i is the length of branch i; p_i is the proportion of taxa descending from branch i; and the indicator function $I(p_i > 0)$ and $I(p_i^B > 0)$ assumes the value of 1 if any taxa descending from branch i is present in the sample or 0 otherwise.

4.3 Beta Diversity

Beta diversity examines the change in species diversity between ecosystems. The analysis is done in two steps. First, the tool estimates a distance between each pair of samples (see the list of available distances below). Once the distance matrix is calculated, the beta diversity analysis tool performs Principal Coordinate Analysis (PCoA) on the distance matrices. These can be visualized by selecting the PCoA icon ( ) in the bottom of the Beta Diversity results ( ).

To run the tool, open **Toolbox** | **OTUclustering ( )** | **Beta Diversity ( )** and select an OTU abundance table before clicking on the button labeled Next.

The following wizard window is shown in figure 4.3.



Beta Diversity

1. Choose where to run
2. Select OTU abundance table
3. Parameters

Parameters

Beta diversity measures

☑ Bray-Curtis

☑ Jaccard

☐ Euclidean

Phlyogenetic diversity

Phylogenetic tree: OTU (Table) alignment_tree

☑ Unweighted UniFrac

☑ Weighted UniFrac

☐ Weighted UniFrac not normalized

☐ D_0 UniFrac

☐ D_0.5 UniFrac

Figure 4.3: *Set up parameters for the Alpha diversity tool.*

As in section 4.2, you must have previously aligned the OTUs and constructed a phylogeny that can be used as input in the Beta Diversity tool.

4.3.1 Beta diversity measures

The following beta diversity measures are available:

- Bray-Curtis: $B = \frac{\sum_{i=1}^{n} |x_i^A - x_i^B|}{\sum_{i=1}^{n} (x_i^A - x_i^B)}$
- Jaccard: $J = 1 - \frac{\sum_{i=1}^{n} \min(x_i^A, x_i^B)}{\sum_{i=1}^{n} \max(x_i^A, x_i^B)}$
- Euclidean: $E = \sum_{i=1}^{n} \sqrt{(x_i^A - x_i^B)^2}$

where n is the number of OTUs and x_i^A and x_i^B are the abundances of OTU i in samples A and B, respectively.

If a phylogenetic tree is provided as input, the following distances are also available:

- Unweighted UniFrac: $d^{(U)} = \frac{\sum_{i=1}^{n} b_i \left| I(p_i^A > 0) - I(p_i^B > 0) \right|}{\sum_{i=1}^{n} b_i}$
- Weighted UniFrac: $d^{(W)} = \frac{\sum_{i=1}^{n} b_i \left| p_i^A - p_i^B \right|}{\sum_{i=1}^{n} b_i (p_i^A + p_i^B)}$
- Weighted UniFrac not normalized: $d^{(w)} = \sum_{i=1}^{n} b_i \left| p_i^A - p_i^B \right|$
- D_0 UniFrac: The generalized UniFrac distance $d^{(0)} = \frac{\sum_{i=1}^{n} b_i \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{n} b_i}$
- D_0.5 UniFrac: The generalized UniFrac distance $d^{(0.5)} = \frac{\sum_{i=1}^{n} b_i \sqrt{p_i^A + p_i^B} \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{n} b_i \sqrt{p_i^A + p_i^B}}$

where n is the number of branches in the phylogenetic tree, b_i is the length of branch i; p_i^A and p_i^B are the proportion of taxa descending from branch i for samples A and B, respectively; and the indicator functions $I(p_i^A > 0)$ and $I(p_i^B > 0)$ assume the value of 1 if any taxa descending from branch i is present is samples A and B, respectively, or 0 otherwise. The unweighted UniFrac distance gives comparatively more importance to rare lineages, while the weighted UniFrac distance gives more important to abundant lineages. The generalized UniFrac distance $d^{(0.5)}$ offers a robust tradeoff [Chen et al., 2012].

4.4 PERMANOVA Analysis

PERMANOVA Analysis (PERmutational Multivariate ANalysis Of VAriance, also known as non-parametric MANOVA [Anderson, 2001]), can be used to measure effect size and significance on beta diversity for a grouping variable. For example, it can be used to show whether OTU abundance profiles of replicate samples taken from different locations vary significantly according to the location or not. The significance is obtained by a permutation test.

To perform a PERMANOVA analysis, go to:

Toolbox | OTUclustering () | PERMANOVA Analysis ().

Choose an OTU abundance table as input. In the next wizard window you can specify the phylogenetic tree reconstructed from the alignment of the most abundant OTUs in the previous step and select the beta-diversity and the phylogenetic diversity measures you wish to use for this analysis (see section 4.3.1 for definitions).

The output of the analysis is a report which contains 2 tables for each beta diversity measure used:

- A table showing the metadata variable used, its groups and the results of the test (pseudo-f-statistic and p-value)
- A PERMANOVA analysis for each pair of groups and the results of the test (pseudo-f-statistic and p-value). Bonferroni-corrected p-values (which correct for multiple testing) are also shown.

4.5 Convert to Experiment

This tool takes an OTU abundance table as input and assigns categories to the samples, allowing users to use the thus-generated table to perform statistical tests. To use the tool, go to:

Toolbox | OTUclustering ( ) | Convert to Experiment ( )

Choose an OTU table as input, and define which metadata group is to be consider as factors.

The tool output is a table labeled (experiment). The first column is the name of the group used as factor in the analysis (its taxonomy and an ID number). For each OTUs there will be the following data:

- Range: The difference between the highest and the lowest expression value for the feature over all the samples.
- IQR: The inter-quantile range of the values for a feature across the samples, that is, the difference between the 75%-ile value and the 25%-ile value.
- Difference: The difference for a two-group experiment between the mean of the expression values across the samples assigned to group 2 and the mean of the expression values across the samples assigned to group 1. Thus, if the mean expression level in group 2 is higher than that of group 1 the 'Difference' is positive, and if it is lower the 'Difference' is negative. For experiments with more than two groups the 'Difference' contains the difference between the maximum and minimum of the mean expression values of the groups, multiplied by -1 if the group with the maximum mean expression value occurs before the group with the minimum mean expression value (with the ordering: group 1, group 2, ...).
- Fold change: For a two-group experiment the 'Fold Change' tells you how many times bigger the mean expression value in group 2 is relative to that of group 1. If the mean expression value in group 2 is bigger than that in group 1 this value is the mean expression value in group 2 divided by that in group 1. If the mean expression value in group 2 is smaller than that in group 1 the fold change is the mean expression value in group 1 divided by that in group 2 with a negative sign. Thus, if the mean expression levels in group 1 and group 2 are 10 and 50 respectively, the fold change is 5, and if the and if the mean expression levels in group 1 and group 2 are 50 and 10 respectively, the fold change is -5. For experiments with more than two groups, the 'Fold Change' column contains the ratio of the maximum of the mean expression values of the groups to the minimum of the mean expression values of the groups, multiplied by -1 if the group with the maximum mean expression value occurs before the group with the minimum mean expression value (with the ordering: group 1, group 2, ...).
- Taxonomy: The taxonomy of the OTU, as specified by the reference database when a database entry was used as Reference.
- Expression values and Means for each site: The expression values represents the number of reads belonging to the OTU in a specific sample, and Means is the mean of these expression values.

You can create subexperiment by selecting only a some of the OTUs from your experiment table (the most abundant ones for example).

Once you have created your experiment table, you can perform several statistical analyses with the following tools of the CLC Genomics Workbench. Note that your data needs to be normalized before performing a hierarchical clustering of samples or features. To find additional statistical analyses, go to **Toolbox** | **Transcriptomics Analysis** () in CLC Genomics Workbench or **Expression Analysis** () in Biomedical Genomics Worbench and choose the following tools:

- **Statistical Analysis** () | **Empirical Analysis of DGE** ()

  For more information about Empirical analysis of DGE, see http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Empirical_analysis_DGE.html.

- **Transformation and Normalization** | **Normalize** ()

  For more information about Normalization, see http://www.clcsupport.com/clcgenomicsworkbe current/index.php?manual=Normalization.html).

- **Quality control** () | **Hierarchical Clustering of Samples** ()

  For more information about Hierarchical Clustering of Samples, see http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Result_hierarchical_clustering_samples.html.

4.6 Importing OTU abundance tables

It is possible to import a csv or excel file as an OTU abundance table, by going to **File** | **Import** () | **Standard Import...** () and force the input as type "OTU abundance table(.xls, .xlsx, .csv)".

This importer allows users to perform statistical analyses on abundance tables that were not generated by OTUclustering tool.

For example, Terminal Restriction Fragment Length Polymorphism (TRFLP) data can be imported and treated similarly as OTU abundance tables. However, all sequence-based actions cannot be applied to this data (i.e., multiple sequence alignment, tree reconstruction and phylogenetic tree measure estimation).

Chapter 5

Workflows

In CLC Genomics Workbench and Biomedical Genomics Worbench, you can link tools to one another to be processed in sequential order enabling repeated execution of a workflow. Working with workflows is described in detail in http://www.clcbio.com/files/tutorials/Workflow-intro.pdf.

The CLC Microbial Genomics Module contains two workflows that you can start here:

Toolbox | **Workflows** | **Data QC and OTU Clustering** or **Estimate Alpha and Beta Diversities**.

To explore a workflow and see the tools it is made of, select the workflow and right click on its name to select the **Open Copy of Workflow** option. They require that you provide the necessary input files and edit the parameter settings, and the workflow will output all relevant results. As many secondary analyses require metadata, the metadata has to be assigned to an OTU table between execution of the two workflows as described in 3.6.2.

5.1 Data QC and OTU clustering workflow

The **Data QC and OTU clustering** workflow consists of 5 tools being executed sequentially (figure 5.1). The only necessary input to run the workflow are the reads you want to cluster. You also have the option to provide a list of the primers that were used to sequence these reads if you wish to perform the adapters trimming step with the **Trim Sequences** tool.

The first tool is the **Optional Merge Paired Reads** that will output 2 sets of sequences, the merged reads and the not paired reads. Both will be used as input in the **Trim Sequences** tool together with the sequencing primer list. This tool provides a list of trimmed sequences that will be the input of the **Fixed Length Trimming** tool. Again, the output is a list of trimmed sequences, used as input file for the **Filter Samples Based on the Number of Reads** tool. The results of the filter are compiled in a report and the tool generates a sequence list that does not contain the reads of poor quality. This "filtered" list will be used for the final tool of the workflow, the **OTUclustering** tool. This tool will give 2 outputs: a sequence list of the OTU centroids and an abundance table with the newly created OTUs, their abundance at each site as well as the total abundance for all samples.

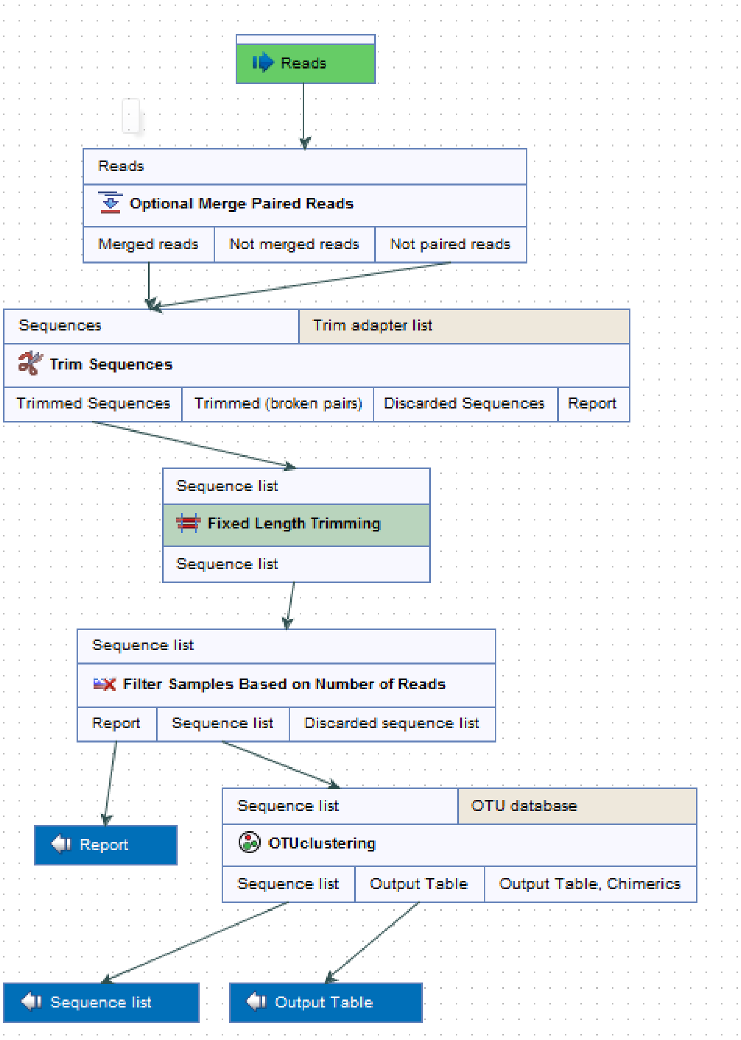


Figure 5.1: *Layout of the Data QC and OTU clustering workflow.*

5.2 Estimate Alpha and Beta Diversities workflow

The **Estimate Alpha and Beta Diversities** workflow consists of 5 tools and requires only the OTU table as input file (figure 5.2). Remember to add metadata to your output table before starting the workflow.

The first tool of the workflow is the **Filter OTUs Based on the Number of Reads**. The output is a reduced abundance table that will be used as input for 3 other tools:

- **Align OTUs with MUSCLE**, a tool that will produce an alignment used to reconstruct a Maximum Likelihood Phylogeny (see http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Maximum_Likelihood_Phylogeny.html), which will in turn output a phylogenetic tree also used as input in the following 2 tools.
- **Alpha diversity** tool
- **Beta diversity** tool

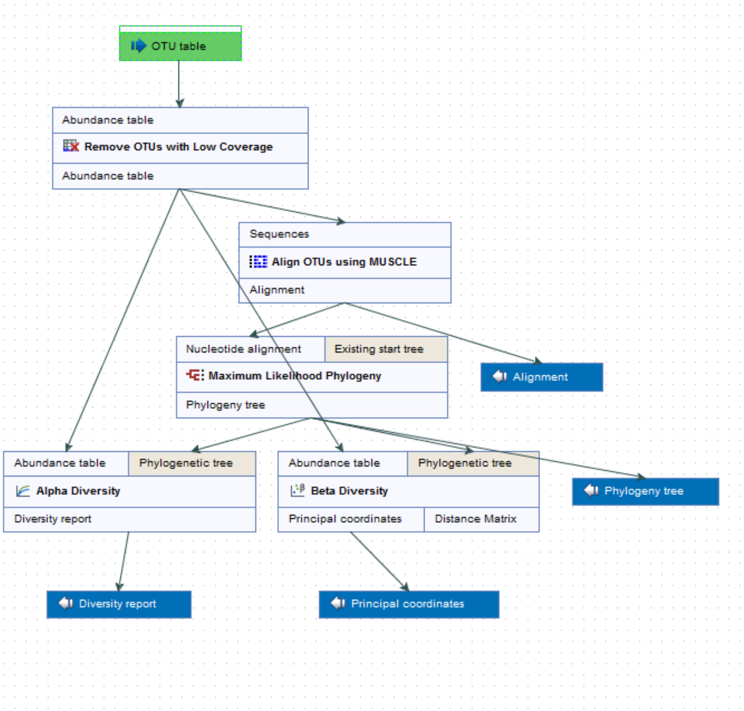


Figure 5.2: *Layout of the Alpha and Beta Diversities workflow.*

Running this workflow will therefore give 4 outputs: an alignment of the OTUs, a phylogenetic tree of the OTUs, a diversity report for the alpha diversity and a PCoA for the beta diversity.

Bibliography

[Anderson, 2001] Anderson, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46.

[Chen et al., 2012] Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., and Li, H. (2012). Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, 28(16):2106–13.

[Whittaker, 1972] Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, pages 213–251.

Index