Geneious 9.0

Biomatters Ltd

December 10, 2015

Contents

Chapter 1

Getting Started

The best way to get started with Geneious is to try out some of our tutorials. The **Tutorial** option under the Help menu in Geneious provides an inbuilt tutorial with a basic introduction to the major features of Geneious. Additional tutorials on specialized functions can be downloaded from our website http://geneious.com/tutorials.

For additional information and help with troubleshooting, please visit the Geneious support website at http://support.geneious.com.

1.1 Downloading & Installing Geneious

Geneious is free to download from http://geneious.com/download. If you are using Geneious for the first time you will be offered a free trial. If you have already purchased a license you can enter it when Geneious starts up.

To download the latest version of Geneious, click on http://geneious.com/download (or type it in to your internet browser), choose the version of Geneious you want to download and click **Download**. If you have a 64-bit Linux or Windows machine, ensure you check the "64-bit" box.

Geneious is compatible with the three most common operating systems: Windows, Mac, and Linux. Check that you have one of the following OS versions before you launch Geneious:

Operating System	System requirements
Windows	XP/Vista/7/8/10
Mac OS	10.6/10.7/10.8/10.9/10.10
Linux	CentOS 6/RHEL6/Ubuntu Desktop LTS

We recommend at least the following specifications for running Geneious (note that these are minimum requirements - for working with large datasets such as NGS sequences you will need a higher-spec machine):

- Processor: Intel x86/x86_64
- Memory: 2048MB or more
- Hard-disk: 2GB or more free space
- Video: 1024x768 resolution or higher

Geneious also needs Java 1.6 or higher to run. Geneious comes bundled with the correct version of Java for your OS. If you require a version of Geneious that does not include Java, please contact Geneious Support.

Once Geneious has downloaded, double-click on the Geneious icon to start installing the program. While this is happening, you will be prompted for a location to install Geneious. Please check that you are satisfied with the location before continuing.

If you are using Mac OS X you will only have to double-click on the disk image that is downloaded, then drag the Geneious application to your Applications folder. Don't run Geneious from the mounted disk image as there are no write permissions on this. You must drag the icon into your Applications folder and run it from there.

1.2 Geneious setup

1.2.1 User preferences

User preferences can be changed by going to **Tools** → **Preferences**. This window can also be opened using the shortcut keys Ctrl+Shift+P (Windows/Linux) or command+Shift+P (Mac OS X). In the user preferences you can change data storage, memory and connection settings, install plugins, customize the appearance and behaviour of Geneious, define shortcut keys, and set up sequencing profiles. Many of these options are explained in more detail in the next sections.

The tabs in the Preferences window are as follows:

General

This tab contains general setup options:

- **Data storage location** - shows the location of your Geneious database (see section 1.2.2)

- **Search history** - allows you to clear your search history
- **Check for updates** - Geneious can check for the release of new versions every time it is started. If a new version has been released Geneious will tell you and give you a link to download it. Geneious will also check for beta versions if you have enabled this option. A beta version is a version that is released before the official release for the purposes of testing. It may therefore be less stable than official releases.
- **Max memory available to Geneious** - allows you to change the RAM allocated to Geneious (see section 1.2.4)
- **Advanced** - allows editing of advanced preferences. You should not alter these unless you know what you are doing.
- **Connection settings** - set how Geneious connects to the internet (see section 1.2.5)

Plugins and Features

Install plugins and customize features in this tab, see section 1.2.6. This tab can also be accessed via **Tools** → **Plugins**.

Appearance and Behavior

In the **Appearance** panel you can change the way the main toolbar and the document table look, and also show or hide tips and the memory usage bar.

In the **Behavior** panel you can change the way newly created documents are handled, such as where they should be saved to and whether they are selected straight away. You can choose whether to store document history, and create active parent/descendant links (see section 6)

Reset questions allows you to reset the questions where you have previously told Geneious to remember your preference. If you have checked ”remember my preference” in a dialog window in Geneious that window will no longer appear. You can click the Reset questions button to get these windows to appear again.

In the **Viewer** panel you can set whether the same view settings are saved across documents of the same type.

Keyboard

This section contains a list of Geneious functions and allows you to define keyboard shortcuts for them. Shortcuts that are already defined are highlighted in blue. Setting shortcuts can

help you quickly navigate through Geneious without using the mouse and also allows you to redefine shortcuts to ones you may be familiar with from other programs.

Double click on a function to bring up a window to enter your new keyboard shortcut. If you use one that is already assigned, Geneious will tell you what function currently has that shortcut.

NCBI

Here you can set the URL for the NCBI BLAST database, and specify which field of the Genbank document should be copied to the "Name" field in Geneious.

Sequencing

This tab has options for the management of trace files and assemblies, allowing you to set thresholds that assign sequences as low, medium or high quality. To change the default parameters or set up a new binning profile, click the Default profile then click Edit/View. The following options are available:

- **Confidence:** Set the threshold values of base call confidence used to determine if a base call is low, medium or high quality. This affects the binning parameters described below as well as the Base Call Quality color scheme in the Sequence View.
- **Sequence binning options:** Specifies the requirements for individual traces to be binned as medium or high quality overall. To see the Bin for a trace, turn on the Bin column under Table Columns in the View menu.
- **Assembly binning options:** Specifies the requirements for assembly documents to be binned as medium or high quality overall. To see the Bin for an assembly, turn on the Bin column under Table Columns in the View menu.

To create a new profile, change the parameters how you wish then enter a new Profile name and click OK.

Other options for managing quality bins are:

- **Track binning history in meta-data:** When turned on, meta-data will be added to traces when they are trimmed (see the Properties view tab). This meta-data will then be updated every time the trace is re-trimmed, maintaining a history of the trimming.
- **Enable per folder/document binning:** When turned on, the **Set Binning Parameters** option is added to the Sequence menu. This allows you to select an individual folder or set of documents and set the binning parameters to use on those documents instead of the global ones set in the Preferences.

1.2.2 Choosing where to store your data

Geneious stores your data in a folder called Geneious X.Y Data (where X and Y are the version of Geneious you are using), which is stored separately from the application itself. When Geneious first starts up you will be asked to choose a location for this folder. The default location in the user's home directory is normally the best option. Although it's possible to store your data on a network or USB drive so you can access it from other computers, this is not recommended because it can have adverse effects on performance. Please do not use a DropBox folder to store your data. This may corrupt your data.

To store your data somewhere different to the default, simply click the 'Select' button in the welcome window and choose an empty folder on your drive where you would like to store your data.

To change the location of your Geneious database at a later date, go to **Tools** → **Preferences** → **General**. The **Data Storage Location** shows the current location of your database (see Figure 1.1). Click the **Browse** button to select a new location. Geneious will offer to either copy your existing data across to a new location, or open an empty database at that location. Geneious will remember this new location when you exit.

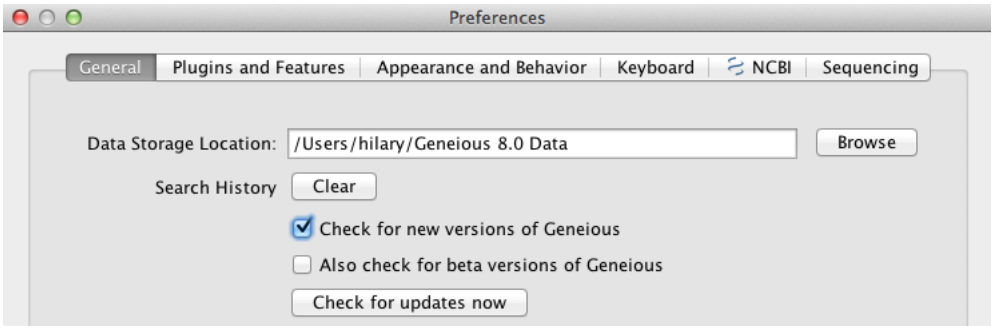


Figure 1.1: Setting the location of your local documents

Note that if you uninstall Geneious your data will not be wiped as the application itself is stored in a different location from the data folder.

1.2.3 Sharing files or the local database

The best solution for sharing files with other users is to set up a shared database (see Chapter 20) in a central location which all users can connect to. Alternatively, you can share files using the collaboration plugin (see Chapter 19), but users all need to be running Geneious at the same time for the files to be accessed. You can also export documents in .geneious format and transfer these files via a USB or network drive for others to use.

We do not recommend storing databases on Dropbox or network drives as a way of sharing files.

1.2.4 Changing the memory available to Geneious

Geneious runs in a Java Virtual Machine. When this JVM starts, it will be allocated a certain amount of RAM and the program can use less than that but never more. In Geneious R8 and higher, on Windows and Linux machines, an appropriate amount of RAM based on your total RAM will automatically be allocated. In Geneious R7 and earlier, and on all mac versions, the default memory allocation is 1GB of RAM on 64-bit machines, and 700MB of RAM on 32-bit machines. To change the amount of memory allocated to Geneious, go to **Tools→Preferences**, and in the General tab increase the **Max memory available to Geneious** setting. On Windows you need to run Geneious as an administrator to change this setting (close Geneious, then right-click on the Geneious icon and choose "Run as administrator").

On a 32-bit machine you can only allocate a maximum of 1GB, so if you need more RAM than this you'll need a 64-bit machine. You should never allocate the total memory of your computer to Geneious, as you need to leave some RAM available for your operating system. As a general rule of thumb, on a 64-bit machine with up to 8GB of RAM in total you can allocate half the RAM to Geneious. On machines with more than 8GB RAM you should leave 3-4GB spare for your operating system and allocate the rest to Geneious.

To see how much memory Geneious is using, check the **Memory Usage Bar** under the Sources panel. You may need to turn this on under **Tools → Preferences → Appearance and Behavior**. Clicking this bar will force a garbage collection and free up memory within the JVM.

1.2.5 Connecting to the internet from within Geneious

In order to activate a license, download plugins and search external databases like NCBI, Geneious needs to be able to connect to the internet. If you have a firewall preventing direct access, or are behind a proxy server you may need to manually configure your connection settings.

You can do this as follows:

1. Check the browser you are using. These instructions are for Explorer, Safari, Chrome and Firefox.
2. Open your default browser.
3. Use the steps in Figure 1.3 for each browser to find the connection settings.
4. Now, in Geneious, go to **Tools → Preferences** and click on the **General** tab. There are five options in the drop-down options under **Connection settings** (Figure 1.2):

- **Use direct connection**. Use this setting when no proxy settings are required.
- **Use browser connection settings**. This allows Geneious to automatically import the proxy settings. This may not work with all web browsers.
- **Use HTTP proxy server**. This enables two text fields : Proxy host and Proxy port. This information is in your browser's connection settings. Use this if your proxy server is an HTTP proxy server. Please see step 3.
- **Use SOCKS proxy server - Autodetect Type**. This enables two text fields : Proxy host and Proxy port. This information is in your browser's connection settings. Use this if your proxy server is a SOCKS proxy server. Please see step 3.
- **Use auto config file**. This enables one text field called "Config file location". These details can also be found in your browser's settings.

5. Set the proxy host and port settings under the General tab to match those in your browser.
6. If your proxy server requires a username and password you can specify these by clicking the 'Proxy Password...' button directly below.

Note. If you are using any other browser, and cannot find the proxy settings, please use the Support Button in the Geneious toolbar to contact Geneious support.

Figure 1.2: Proxy settings in Preferences

1.2.6 Installing plugins and customizing features

You can extend the functionality of Geneious with a variety of plugins. These can be downloaded from our website, or managed via the plugins and features preferences (Figure 1.4) in Geneious. To access the plugins preferences, go to **Tools** → **Plugins**.

This window contains a list of available plugins in the top window, which lists plugins available for download which aren't already installed. To install a plugin, click the Install button, and for more information about the plugin, click Info. Plugins with a star are note-worthy plugins, as chosen by the Geneious team.

Figure 1.3: Checking browser settings

If you have downloaded a plugin file from our website (or obtained one from another source in .gplugin format) you can install it by clicking **Install plugin from a gplugin file** and browsing to the location of the file. You can also install plugins by dragging the gplugin file into Geneious.

If you are running Geneious as an administrator then any plugin you install will be installed for all users on the same computer. If you are not running as an administrator then plugins will only be installed for the current user account. When upgrading plugins, Geneious may display a message indicating that Geneious needs to be restarted in order to complete the plugin upgrade. If Geneious was being run as an administrator, it needs to be restarted as an administrator to complete the plugin upgrade.

Note that on Windows you cannot drag and drop plugins in to Geneious while running as an administrator.

Installed Plugins lists all the plugins you current have installed. Click the uninstall button next to a plugin to remove it.

Other options for managing plugins are:

- **Check for plugin updates now:** Checks if there are any new versions available for the plugins you have installed.
- **Automatically check for updates to installed plugins:** If checked, Geneious will check for new versions of your installed plugins each time the program is started.
- **Tell me when new plugins are released:** Changes the way the program notifies you about new plugin releases.

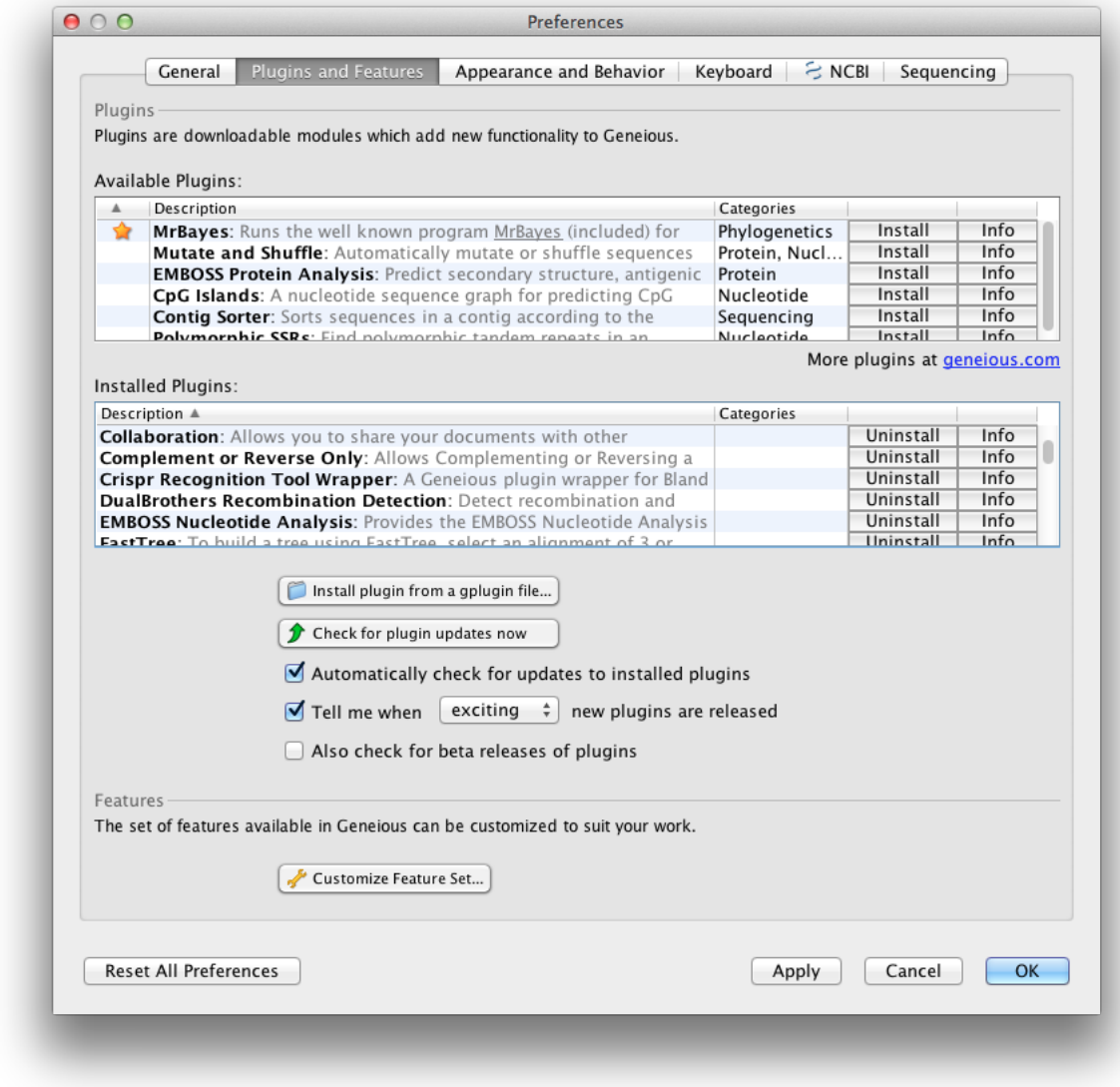


Figure 1.4: The plugins preferences in Geneious

- **Also check for beta releases of plugins:** Plugins are sometimes initially released as a beta for the purposes of testing before the officially release. Check this to be notified about the release of beta plugins.

Customizing features

To see a list of all the features in Geneious, click **Customize feature set**. Features can be turned on or off by checking or un-checking the Enabled box next to each feature. You might like to turn off the Tree Builder and Tree Viewer plugins if you don't do phylogenetics for example.

1.3 Upgrading to new versions

To upgrade existing Geneious installations, simply download and install the new version to the same location. Your existing data and license will be automatically loaded in the new version. If you are upgrading to a new major version (e.g. 8.0 to 8.1, or 8.1 to 9.0) Geneious will update your data folder to the new format (creating a new folder with the upgrade's version number in the name), and will offer to keep a copy of your old data folder in case you wish to downgrade.

1.3.1 Downgrading to an earlier version

If you chose to keep a copy of your old data folder when you upgraded, you can easily downgrade if you prefer to use the earlier version, or if your license isn't valid for the latest version. Downgrading requires that the new version of Geneious is uninstalled first to avoid there being vestiges of the old copy in place. Once this is done, the old version can be reinstalled and Geneious will start up and see the old data folder. For instructions on how to import data from a newer version into an old version see the section below on File compatibility.

1.3.2 File compatibility

Geneious data files are backwards compatible to version 6.0. Thus, files that were created in version 6.1 or higher can be exported in .geneious format and opened in any version back to 6.0. If you are using an earlier version than version 6.0 you won't be able to open files in .geneious format that were created in a newer version. Entire Geneious databases are not backwards compatible, so when you upgrade you should accept the offer to keep a backup of your existing database. If you then need to downgrade to an earlier version you can swap back to your old database, and if you are on Geneious 6.0 or above, export any files you changed in the new version in .geneious format and import them into your old database. If you wish to export data back into a version earlier than 6.0, you will need to export the files in a common format such as fasta or genbank.

1.4 Licensing

Licenses can be managed under the **Help** menu in Geneious, using the following options:

1.4.1 Activate License

In this window you can activate a personal license or choose to connect to a license server. The options are as follows:

- **Use license key**. If you have purchased a personal license you can enter the details here to activate it. Make sure you enter the licensee name exactly as it appears in the email in which you received your activation ID/registration key. An internet connection is required to activate personal licenses, and you may need to configure your firewall/proxy settings to enable access to http://licensing.biomatters.com on port 80.
- **Use license server**. If your organization has purchased a floating license administered through a FLEXnet license server, this is where you enter the details required to connect to the license server. Ask your system administrator for the host name and port of the license server.
- **Use Sassafras KeyServer**. If your organization has purchased a floating license administered through Sassafras KeyServer, select this option. Your system administrator needs to configure KeyAccess to point to the KeyServer license server.

1.4.2 Install FLEXnet

This installs the FLEXnet license manager which is necessary for activating a personal license. This is normally installed automatically when Geneious is installed, but Geneious will tell you if need to run this when you activate your license. Only an administrator on your computer can do this but it only needs to be done once from one user account. Once this has been done, any non-admin user can activate their license on the machine. The admin should not activate licenses for users as this will prevent the user from activating the license themselves.

1.4.3 Borrow Floating License

This item is only available to users for a floating license administered through a FLEXnet license server. Borrowing a license allows you to borrow one of the seats of a floating license so you can use it even when disconnected from the network. Since this decreases the number of seats available for other users, borrowing can only occur with the authorization of the system administrator. If your borrowing is approved, the system administrator will provide you with a ”borrow file” authorizing the borrow. To borrow a license, check ”Borrow” in the menu,

and navigate to this file when prompted by Geneious. Borrowed licenses have an expiry date, when they will automatically be returned to the server, but if you are finished with the license before the expiry date, please uncheck "Borrow" in the menu while connected to the network in which the license server resides, so that the license is returned to the server and is available to other users again.

1.4.4 Release License

Personal licenses can only be activated on a maximum of three computers at one time. If you no longer need to have Geneious available on a computer where you have activated it, you can release the license so it is available for use on another computer. Licenses can only be released twice in a 6 month period so do not do it unnecessarily.

If you're using a floating license, you can release it allowing another user to access it without you having to shut Geneious down. Once you've released the license, Geneious will enter restricted use mode.

1.4.5 Buy Online

This item will open the Geneious store in your browser, where you can upgrade licenses, purchase new licenses or contact our sales team.

1.5 Troubleshooting

For help with troubleshooting or to request a feature, please contact the Geneious support team. You can do this either by clicking the Support button located in the toolbar in Geneious, or going to our support website at http://support.geneious.com and submitting a request. Clicking the Support button will automatically send through some system information to help us assist you. If you are submitting a request through our website, please include details of your operating system and the version of Geneious you are using and as much information as possible about the nature of your problem (including screenshots to illustrate the issue if appropriate).

The Geneious support website also contains a comprehensive Knowledge Base with solutions to common problems and tips for getting the most out of Geneious, as well as a User Forum where you can post questions to the Geneious community.

You can access the support website and download user manuals, license agreements and release notes from within Geneious by going to **Help** → **Online Resources**.

Chapter 2

The Geneious main window

Figure 2.1 shows the main Geneious window. This has five important areas or 'panels'.

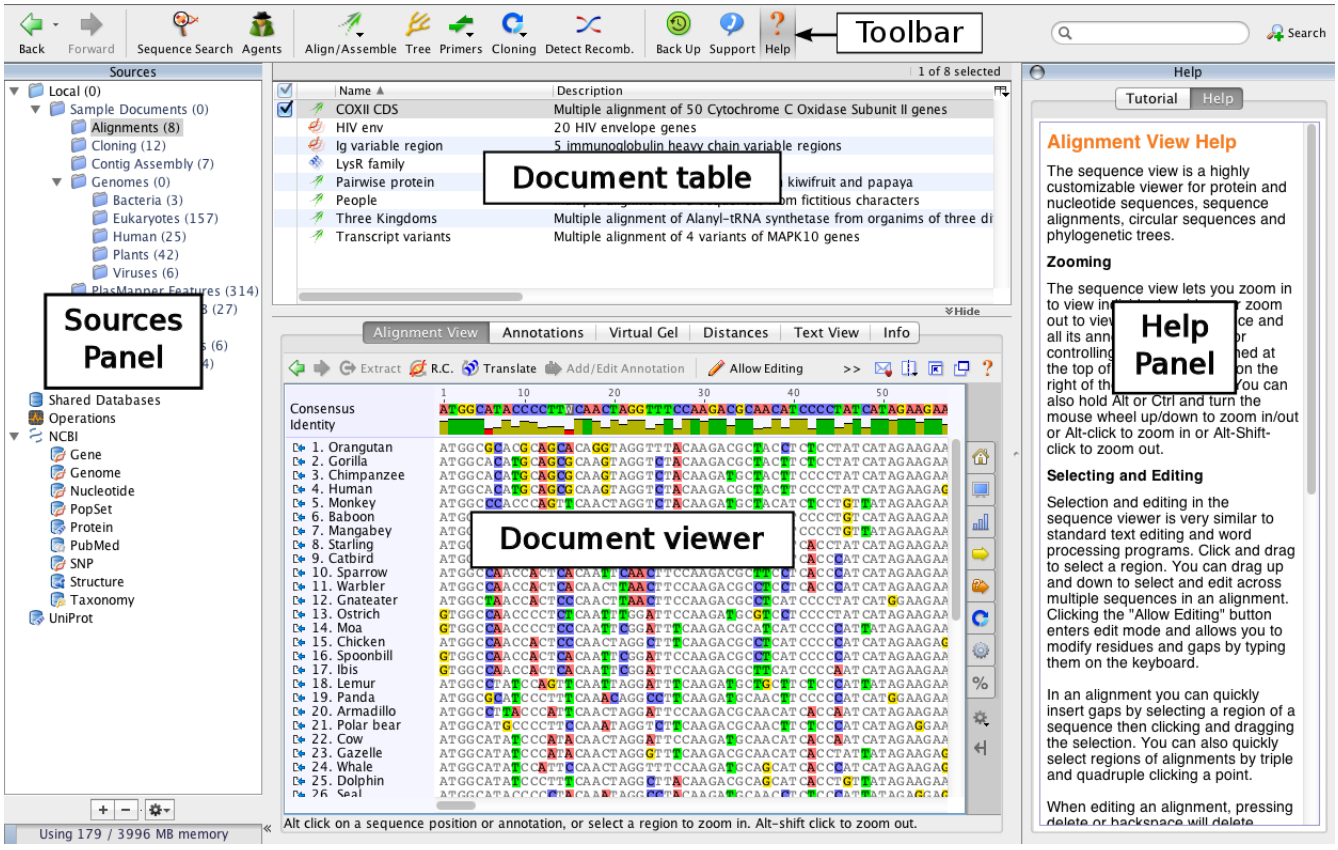


Figure 2.1: The main window in Geneious

2.1 The Sources Panel

The Sources Panel displays your stored documents and contains the services Geneious offers for storing and retrieving data. The plus (+) symbol indicates that a folder contains sub-folders. A minus (-) indicates that the folder has been expanded, showing its sub-folders. Click these symbols to expand or contract folders.

Geneious Sources Panel allows you to access:

- Your Local Documents.
- NCBI databases - Gene, Genome, Nucleotide, PopSet, Protein, Pubmed, SNP, Structure and Taxonomy.
- An EMBL database - Uniprot.
- Shared databases, if set up.
- Your contacts' Geneious databases, if collaboration is installed.

All these services will be described in detail later in the manual. You can view options for any selected service with the right mouse button, or by clicking the Options button at the bottom of the Sources Panel in Mac OS X.

For more information on managing folders in the Sources panel, see section 4.1.

2.2 The Document Table

The Document Table displays summaries of each document in a selected folder or folders, presented in table form. A local folder may contain any mixture of documents, such as DNA sequences, protein sequences, journal articles, sequence alignments, and trees (Figure 2.2). If you cannot see all of the columns in the document table you may want to close the help panel to make more room.

For information on how to search and filter documents in the Document Table, see section 4.2.

Selecting a document in the Document Table will display its details in the Document View Panel. Selecting multiple documents will show a view of all the selected documents if they are of similar types, e.g. selecting two sequences will show both of them in the sequence view.

The easiest way to select multiple documents is by clicking on the checkboxes down the left-hand side of the table. Standard keyboard controls can also be used (Shift and Ctrl/command click).

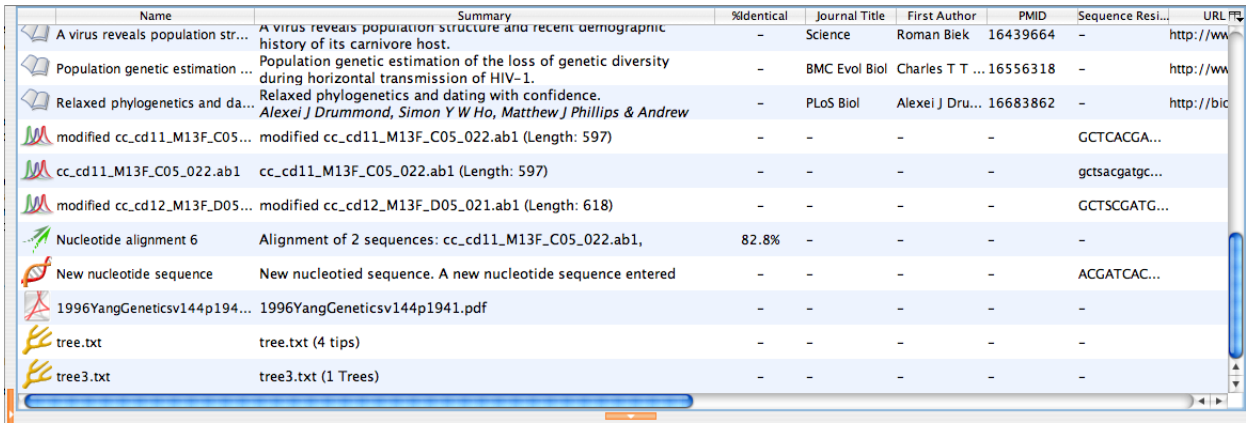


Figure 2.2: The document table, when browsing the local folders

Double-clicking a document in the Document Table displays the same view in a separate window.

To view the functions available for any particular document or group of documents, right-click (Ctrl+click on Mac OS X) on a selection of them. These options vary depending on the type of document.

Document Table features

Editing. Values can be typed into the columns of the table. This is a useful way of editing the information in a document. To edit a particular value, first click on the document and then click on the column which you want to edit. Enter the appropriate new information and press enter. Certain columns cannot be edited however, eg. the NCBI accession number.

Copying. Column values can be copied. This is a quick method of extracting searchable information such as an accession number. To copy a value, right-click (Ctrl+click on Mac OS X) on it, and choose the "Copy `name`" option, where `name` is the column name.

Sorting. All columns can be alphabetically, numerically or chronologically sorted, depending on the data type. To sort by a given column click on its header. If you have different types of documents in the same folder, click on the "Icon" column to sort then according to their type.

Managing Columns. You can reorder the columns to suit. Click on the column header and drag it to the desired horizontal position.

You can also choose which columns you want to be visible by right-clicking (Ctrl-click on Mac OS X) on any column header or by clicking the small header button in the top right corner of the table. This gives a popup menu with a list of all the available columns. Clicking on a column will show/hide it. Your preference is remembered so if you hide a column it will

remain hidden in all areas of the program until you show it again.

As well as items to show/hide any of the available columns, there are a few more options in this popup menu to help you manage columns in Geneious.

- **Lock Columns** locks the state of the columns in the current table so that Geneious will never modify the way the columns are set up. You can still change the columns yourself however.
- **Save Current State...** allows you to save the the current state if the columns so you can easily apply it to other tables. You can give the state a name and it will then appear in the Load Column State menu.
- **Load Column State** contains all of the columns states you have saved. Selecting a column state from here will immediately apply that state to the current table and lock the columns to maintain the new state. Use **Delete Column State...** to remove unwanted columns states from this menu.

Note. New columns can be added to the document table by adding Meta-Data to documents (see section 5.7 - Meta-Data).

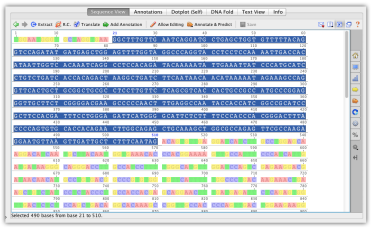
2.3 The Document Viewer Panel

The Document Viewer Panel shows the contents of any document clicked on in the Document Table, allowing you to view sequences, alignments, trees, 3D structures, journal article abstracts and other types of documents in a graphical or plain text view (Figure 2.3). Options for controlling the look and layout of a given document are displayed in the right-hand panel. These options vary depending on what type of document you are viewing. For detailed information on specific types of viewers, please refer to the sections below:

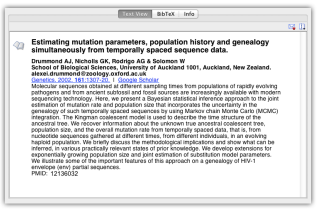
Sequence/Alignment Viewer - section 5.2
3D Structure Viewer - section 7.2
Dotplot Viewer - section 9.1
Tree Viewer - section 12.5
Journal Article Viewer - section 3.3.3

To view large documents, you can open them in a new window by double clicking.

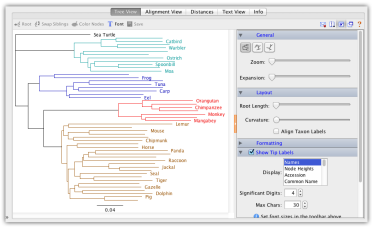
In the document viewer panel there are two tabs that are common to most types of documents: **Text view** and **Info**. **Text view** shows the document's information in text format. The exception to this rule occurs with PDF documents where the user needs to either click the **View Document** button or double-click to view it. Under the **Info** tab, you can view document **Properties** (meta-data, section 5.7), **History** (section 4.6) and **Lineage** (section 6.2).



(a) Nucleotide sequence



(b) Journal Article



(c) Phylogenetic tree

Figure 2.3: Three document viewers

2.3.1 General viewer controls

There are several general options which are available on all viewers, which are shown in the toolbar at the top right of the viewer. Some of these can also be accessed through the **View** menu.

Share: Allows you to share the current visualization on Twitter, Facebook or email.

Split View: Provides several options for splitting the view so that multiple views are shown simultaneously for one document. When the view is split, selection of annotations and regions of the sequence are synchronized across the viewers. To close split views click the button which is also on the right of the toolbar.

Expand View: Expands the document view panel to fill the main window by hiding the sources panel on the right and the document table above. Clicking this again will return the layout to it's original state.

New Window: Opens another view of the current document in a separate window. This allows you to have several documents open at once and gives more space for viewing. This can also be achieved by double clicking in the document table.

Help: Opens the Help Panel (section 2.4) and displays some short help for the current viewer.

2.4 The Help Panel

The Help Panel has a **Help** tab and a **Tutorial** tab.

The Help tab provides you with information about the service you are currently using or the viewer you are currently viewing. The help displayed in the help tab changes as you click on different services and choose different viewers.

The Tutorial is aimed at first-time users of Geneious and has been included to provide a feel for how Geneious works. It is highly recommended that you work through the tutorial if you haven't used Geneious before.

The Help panel can be closed at any time by clicking the button in its top corner, or by toggling the 'Help' button in the Toolbar.

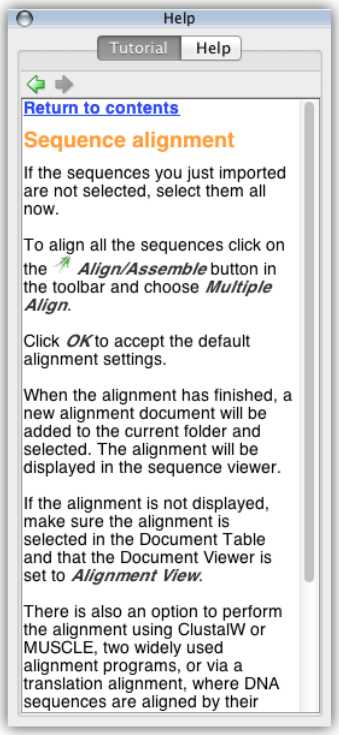


Figure 2.4: The Help Panel

2.5 The Toolbar

The toolbar contains several icons that provide shortcuts to common functions in Geneious, including **BLAST** (Sequence Search), **Agents** that search databases for new content even while you sleep, **Align/Assemble**, **Tree** building, **Primer Design** and **Help**.

The **Back** and **Forward** options help you move between previous views in Geneious and are analogous to the back and forward buttons in a web browser. The ▽ option shows a list of previous views. The other features that can be accessed from the toolbar are described in later sections.

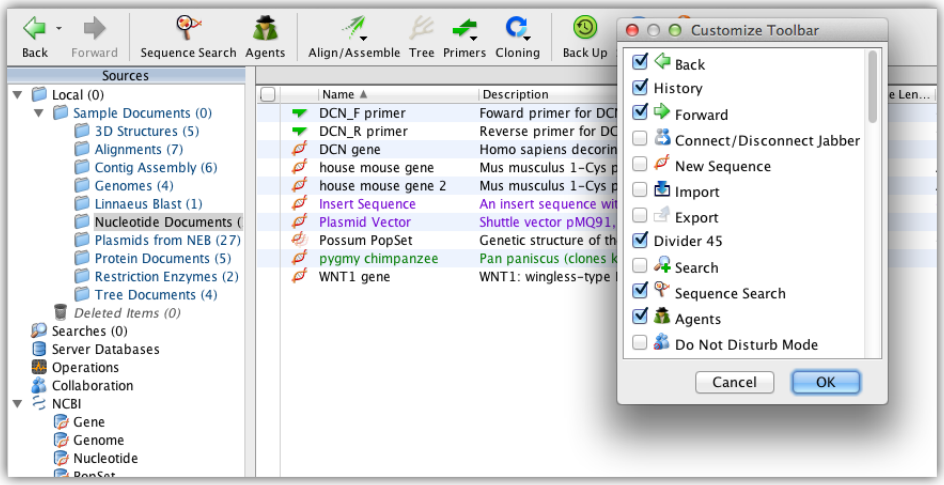


Figure 2.5: The Toolbar

The toolbar can be customized by right-clicking (Ctrl-click on Mac OS X) on it. This gives a popup menu with the following options:

- **Show Labels:** Turns the text labels on or off.
- **Large Icons:** Switches between large and small icons.
- **Customize:** Lists all available toolbar buttons. Selecting/deselecting buttons will show/hide the buttons in the toolbar.

2.5.1 Status bar

Below the Toolbar, there is a grey status bar. This bar displays the status of the currently selected service. For example, when you are running a search, it displays the number of matches, and the time remaining for the search to finish.

2.6 Geneious menu bar options

All of the functions in Geneious can be accessed from the menu bar above the Toolbar. This is split into seven main menus **File**, **Edit**, **View**, **Tools**, **Sequence**, **Annotate & Predict** and **Help**.

2.6.1 File Menu

This contains standard options for managing files, including import/export (see chapter 3), deleting, saving and backing up files (see 4.5), creating, renaming and moving folders. It also contains options for printing and saving image files (see 3.6.2).

2.6.2 Edit Menu

This menu contains the standard editing functions for transferring information from within documents to other locations, both within and outside of Geneious: **Cut**, **Copy**, **Paste**, **Delete** and **Select All**.

This menu also contains options for finding and renaming documents and their contents:

- **Find in Document** can be used to find text or numbers in a selected document, see section 4.2.3.
- **Find Next and Find Previous** finds the next or previous match for the text you specified in the Find in Document dialog.
- **Find Duplicates**, see section 4.3.
- **Batch Rename**, see section 4.4.
- **Go to base/residue**, see section 5.2.2.

2.6.3 View Menu

This contains several options and commands for changing the way you view data in Geneious:

- **Back**, **Forwards** and **History** allow you to return to documents you had selected previously.
- **Search** is discussed in section 4.2.
- **Agents** are discussed in section 3.4.
- **Next unread document** selects the next document in the current folder which is unread.
- **Table Columns** contains the same functionality as the popup menu for the document table header. See section 2.2 for more details.
- **Open document in new window** Opens a new window with a view of the currently selected document(s).

- **Expand document view** expands the document viewer panel in the main window out to fill the entire main window. Selecting this again to return to normal.
- **Split Viewer Left/Right** creates a second copy of the document viewer with the two views laid out side by side.
- **Split Viewer Top/Bottom** creates a second copy of the document viewer with one on top of the other.
- **Document Windows** Lists the currently open document windows. Selecting one from this menu will bring that document window to the front.

2.6.4 Tools Menu

- **Align/Assemble** - see section 9.2 and section 10 respectively
- **Tree** - see section 12
- **Primers** - see section 13
- **Cloning** - see section 14
- **BLAST** - Perform a BLAST search (such as NCBI Blast) to find sequences that are similar to the currently selected sequence(s). See section 15
- **Add/Remove Databases** - options for setting up and configuring NCBI and custom BLAST, see section 15.4.3.
- **Extract Annotations** - see section 8.1.4.
- **Strip Alignment Columns** - see section 9.3.4.
- **Concatenate Sequences or Alignments** - see section 5.3.1.
- **Generate Consensus Sequence** - see section 9.4.
- **Workflows** - access built-in Workflows and create new ones, see chapter 16.
- **Plugins** - Takes you to the Plugins menu where you can install or uninstall plugins.
- **Preferences** - see section 1.2.1

Many plugin options will also appear in this menu when installed, such as **Classify Sequences** (Sequence Classifier plugin), **Submit to Genbank** (Genbank submission plugin) and **Collaboration** (see chapter 19)

2.6.5 Sequence Menu

This contains several operations for manipulating nucleotide and protein sequences, including processing NGS reads prior to assembly.

- **New Sequence**: Create a new nucleotide or protein sequence (including oligos) from residues that you can paste or type in. See section 5.1.
- **Extract Region**: Extract the selected part of a sequence or alignment into a new document.
- **Reverse Complement**: Reverse sequence direction and replace each base by its complement. See section 5.4.
- **Translate**: Creates a new protein document from the translated DNA, see section 5.5.
- **Back Translate**: Creates nucleotide version of the selected protein document, see section 5.5.2.
- **Circular Sequences**: Sets whether the currently selected sequences are circular. This effects the way the sequence view displays them as well as how certain operations deal with the sequences (eg. digestion). See section 5.2.3.
- **Free End Gaps Alignment**: Sets whether the currently selected alignment has free end gaps. This effects calculation of the consensus sequences and statistics.
- **Change Residue Numbering...**: Changes the "original residue numbering" of the selected sequence. On a linear sequence, this is used to indicate that a sequence is a subsequence of some larger sequence. On a circular sequence, this is used to shift the origin of the sequence.
- **Convert between DNA and RNA**: Changes all T's in a sequence to U's or vice versa, depending on the type of the selected sequence. Once this is performed, click "Save" in the Sequence View to make the change permanent.
- **Set Paired Reads**: Sets up paired reads for assembly. See section 10.1.4.
- **Merge Paired Reads**: Merges paired reads using BBMerge, see section 10.1.5.
- **Remove Duplicate Reads**: Uses Dedupe to remove duplicate sequences from NGS datasets, see section 10.1.2.
- **Error Correct and Normalize**: Uses BBNorm to error correct and normalize NGS reads, see section 10.1.3.
- **Set Read Direction**: Marks sequences as forward or reverse reads so the correct reads are reverse complemented by assembly.
- **Separate Reads by Barcode** separates multiplex or barcode data (e.g. 454 MID data). See section 10.1.6

- **Group Sequences into a List** creates a sequence list containing copies of all of the selected sequences. See section 5.1.1.
- **Extract Sequence from List** copies each sequence out of a sequence list into a separate sequence document.

2.6.6 Annotate & Predict Menu

This menu contains many tools for finding, predicting and annotating regions of interest in sequences and alignments. Plugins that involve sequence prediction and annotation will also appear in this menu when installed.

- **Trim Ends**: Trims vectors, primers and/or poor quality sequence. see section 10.1.1.
- **Annotate from Database**: Annotates sequences with similar annotations from your database. See section 8.2.3
- **Find ORFs**: Finds all open reading frames in a sequence and annotates them
- **Search for Motifs**: Searches for motifs in PROSITE format. Uses "fuzznuc" and "fuzzpro" from EMBOSS.
- **Find Variations/SNPs**: Finds variable positions in assemblies and alignments. See section 11.1.1.
- **Find Low/High Coverage**: Finds regions with low or high read coverage in assemblies. See section 10.4.
- **Calculate Expression Levels**: Calculates expression levels (RPKM, FPKM, TPM) for a single sample. See section 11.2.
- **Compare Expression Levels**: Compares expression levels between two samples. See section 11.2.
- **Transfer Annotations**: Copies annotations to the reference and/or consensus sequence of an alignment or assembly. See section 8.2.3.
- **Compare Annotations**: Compares annotations across up to 3 annotation tracks or documents. See section 8.3.

2.6.7 Help Menu

This consists of the standard Help options offered by Geneious.

- **Help** shows and hides the Help panel

- **Tutorial** shows and hides the Tutorial panel
- **Online Resources** gives access to a variety of resources on our website
- **Check for Updates** checks for new versions of Geneious
- **Contact Support** allows you to contact our Support team through Geneious
- **Activate License** lets you activate a license or connect to a license server
- **Install FLEXnet** installs the FLEXnet licensing service which is necessary to use FLEXnet licenses
- **Borrow Floating License** lets you borrow a license from a FLEXnet server, if the maintainer of the server has provided you with a Borrow File
- **Release Licenses** releases any floating license you are currently holding and returns any local FLEXnet licenses to our server so they can be activated on a difference machine
- **Buy Online** sends you to our online store
- **About Geneious** gives details about the version of Geneious you are running, and licensing information

Chapter 3

Importing and Exporting Data

Geneious is able to import raw data from different applications and export the results in a range of formats. If you are new to bioinformatics, please take the time to familiarize yourself with this chapter as there are a number of formats to be aware of.

3.1 Importing data from the hard drive to your Local folders

To import files from local disks or network drives, click **File** → **Import** → **From file**. This will open up a file dialog. Select one or more files and click **Import**. If Geneious' automatic file format detection fails, select the file type you wish to import (Figure 3.1). Files can also be dragged and dropped from your hard drive directly into Geneious and Geneious will automatically determine the file type. The different file formats that Geneious can import are described in detail in the next section.

Files imported from disk are imported directly into the currently selected local folder within Geneious. If no folder is selected, Geneious will open a dialog which lets you specify a folder.

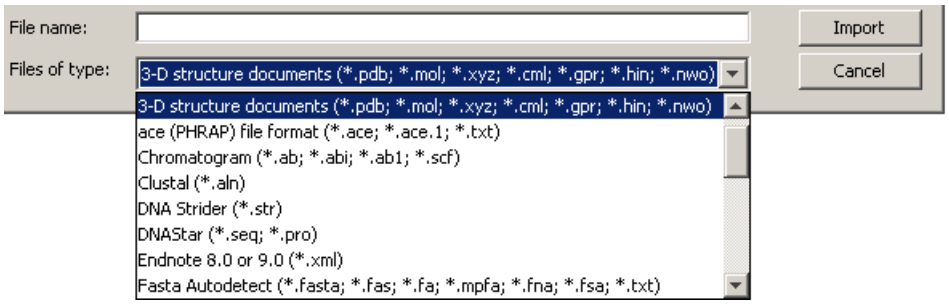


Figure 3.1: File import options

In Geneious R9 and above, it is possible to import an entire folder and all its subfolders and files into Geneious in one step. To do this, choose **File** → **Import** → **From Directory...**. If the folder has subfolders, the folder structure will be retained when it is imported into Geneious.

3.2 Data input formats

Geneious version 9.0 can import the following file formats:

Format	Extensions	Data types	Common sources
BED	*.bed	Annotations	UCSC
Common Assembly Format	*.caf	Contigs	Sequencher
Clustal	*.aln	Alignments	ClustalX
CSFASTA	*.csfasta	Color space FASTA	ABI SOLiD
DNAStar	*.seq, *.pro	Nucleotide & protein sequences	DNAStar
DNA Strider	*.str	Sequences	DNA Strider (Mac program), ApE
Embl/UniProt	*.embl, *.swp	Sequences	Embl, UniProt
Endnote (8.0 or 9.0) XML	*.xml	Journal article references	Endnote, Journal article websites
FASTA	*.fasta, *.fas, *fasta.gz etc.	Sequences, alignments	PAUP*, ClustalX, BLAST, FASTA
FASTQ	*.fastq, *.fq, *fastq.gz etc.	Sequences with quality	Illumina and other NGS sequencers
GCG	*.seq	Sequences	GCG
GenBank	*.gb, *.xml	Nucleotide & protein sequences	GenBank
Geneious	*.xml, *.geneious	Preferences, databases	Geneious
Geneious Education	*.tutorial.zip	Tutorial, assignment etc.	Geneious
GFF	*.gff	Annotations	Sanger Artemis
MEGA	*.meg	Alignments	MEGA
Molecular structure	*.pdb, *.mol, *.xyz, *.cml, *.gpr, *.hin, *.nwo	3D molecular structures	3D structure databases and programs
Newick	*.tre, *.tree, etc.	Phylogenetic trees	PHYLIP, Tree-Puzzle, PAUP*, ClustalX
Nexus	*.nxs, *.nex	Trees, Alignments	PAUP*, Mesquite, MrBayes & MacClade
PDB	*.pdb	3D Protein structures	SP3, SP2, SPARKS, Protein Data Bank
PDF	*.pdf	Documents, presentations	Adobe Writer, LaTeX, Miktex
Phrap ACE	*.ace	Contig assemblies	Phrap/Consed
PileUp	*.msf	Alignments	pileup (gcg)
PIR/NBRF	*.pir	Sequences, alignments	NBRF PIR
Qual	*.qual	Quality file	Associated with a FASTA file
Raw sequence text	*.seq	Sequences	Any file that contains only a sequence
Rich Sequence Format	*.rsf	Sequences, alignments	GCGs NetFetch
Comma/Tab Separated Values	*.csv, *.tsv	Spreadsheet files	Microsoft Excel
SAM/BAM	*.sam, *.bam	Contigs	SAMtools
Sequence Chromatograms	*.ab1, *.scf	Raw sequencing trace & sequence	Sequencing machines
VCF	*.VCF	Annotations	1000 Genomes Project
Vector NTI sequence	*.gb, *.gp	Nucleotide & protein sequences	Vector NTI
Vector NTI/AlignX alignment	*.apr	Alignments	Vector NTI, AlignX
Vector NTI Archive	*.ma4, *.pa4, *.oa4, *.ea4, *.ca6	Nucleotide & protein sequences, enzyme sets and publications	Vector NTI
Vector NTI/ContigExpress	*.cep	Nucleotide sequence assemblies	Vector NTI
Vector NTI database	VNTI Database	Nucleotide & protein sequences, enzyme sets and publications	Vector NTI

BED annotations

The BED format contains sequence annotation information. You can use a BED file to annotate existing sequences in your local database, import entirely new sequences, or import the annotations onto blank sequences.

CLUSTAL alignment

The Clustal format is used by ClustalW and ClustalX, two well known multiple sequence alignment programs.

Clustal format files are used to store multiple sequence alignments and contain the word clustal at the beginning. An example Clustal file:

```
CLUSTAL W (1.74) multiple sequence alignment

seq1 ----------------------KSKERYKDENGGNYFQLREDWWDANRETVWKAITCNA
seq2 ---------------YEGLTTANGXKEYYQDKNGGNFFKLREDWWTANRETVWKAITCGA
seq3 ----KRIYKKIFKEIHSGLSTKNGVKDRYQN-DGDNYFQLREDWWTANRSTVWKALTCSD
seq4 ------------------------SQRHYKD-DGGNYFQLREDWWTANRHTVWEAITCSA
seq5 --------------------NVAALKTRYEK-DGQNFYQLREDWWTANRATIWEAITCSA
seq6 ------FSKNIX--QIEELQDEWLLEARYKD--TDNYYELREHWWTENRHTVWEALTCEA
seq7 -------------------------------------------------KELWEALTCSR

seq1 --GGGKYFRNTCDG--GQNPTETQNNCRCIG----------ATVPTYFDYVPQYLRWSDE
seq2 P-GDASYFHATCDSGDGRGGAQAPHKCRCDG---------ANVVPTYFDYVPQFLRWPEE
seq3 KLSNASYFRATC--SDGQSGAQANNYCRCNGDKPDDDKP-NTDPPTYFDYVPQYLRWSEE
seq4 DKGNA-YFRRTCNSADGKSQSQARNQCRC---KDENGKN-ADQVPTYFDYVPQYLRWSEE
seq5 DKGNA-YFRATCNSADGKSQSQARNQCRC---KDENGXN-ADQVPTYFDYVPQYLRWSEE
seq6 P-GNAQYFRNACS----EGKTATKGKCRCISGDP----------PTYFDYVPQYLRWSEE
seq7 P-KGANYFVYKLD-----RPKFSSDRCGHNYNGDP---------LTNLDYVPQYLRWSDE
```

CSFASTA format

ABI .csfasta files represent the color calls generated by the SOLiD sequencing system.

DNAStar sequences

DNAStar .seq and .pro files are used in Lasergene, a sequence analysis tool produced by DNAStar.

DNA Strider sequences

Sequence files generated by the Mac program DNA Strider, containing one Nucleotide or Protein sequence.

EMBL/Swiss-Prot sequences

Nucleotide sequences from the EMBL Nucleotide Sequence Database, and protein sequences from UniProt (the Universal Protein Resource)

EndNote 8.0/9.0 XML

EndNote is a popular reference and bibliography manager. EndNote lets you search for journal articles online, import citations, perform searches on your own notes, and insert references into documents. It also generates a bibliography in different styles. Geneious can interoperate with EndNote using Endnote's XML (Extensible Markup Language) file format to export and import its files.

FASTA sequences

The FASTA file format is commonly used by many programs and tools, including BLAST, T-Coffee and ClustalX. Each sequence in a FASTA file has a header line beginning with a ">" followed by a number of lines containing the raw protein or DNA sequence data. The sequence data may span multiple lines and these sequence may contain gap characters. An empty line may or may not separate consecutive sequences. Here is an example of three sequences in FASTA format (DNA, Protein, Aligned DNA):

```
>Orangutan
ATGGCTTGTGGTCTGGTCGCCAGCAACCTGAATCTCAAACCTGGAGAGTGCCTTCGAGTG

>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQK

>Chicken
CTACCCCCCTAAAACACTTTGAAGCCTGATCCTCACTA------------------CTGT
CATCTTAA
```

FASTQ sequences

FASTQ format stores sequences and Phred qualities in a single file.

GenBank sequences

Records retrieved from the NCBI website (http://www.ncbi.nlm.nih.gov) can be saved in a number of formats. Records saved in GenBank or INSDSeq XML formats can be imported into Geneious.

Geneious format

The Geneious format can be used to store all your local documents, meta-data types and program preferences. A file in Geneious format will usually have a `.geneious` extension or a `.xml` extension. This format is useful for sharing documents with other Geneious users and backing up your Geneious data.

Geneious tutorial

This is an archive containing a whole bundle of files which together comprise a Geneious education document. This format can be used to create assignments for your students, bioinformatics tutorials, and much more. See chapter 17 for information on how to create such files.

GFF annotations

The GFF format contains sequence annotation information (and optional sequences). You can use a GFF file to annotate existing sequences in your local database, import entirely new sequences, or import the annotations onto blank sequences.

MEGA alignment

The MEGA format is used by MEGA (Molecular Evolutionary Genetics Analysis).

Molecular structure

Geneious imports a range of molecular structure formats. These formats support showing the locations of the atoms in a molecule in 3D:

- **PDB format** files from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Database
- ***.mol format** files produced by MDL Information Systems Inc
- ***.xyz format** files produced by XMol
- ***.cml format** files in Chemical Markup Language
- ***.gpr format** ghemical files
- ***.hin format** files produced by HyperChem
- ***.nwo format** files produced by NWChem

Newick tree

The Newick format is commonly used to represent phylogenetic trees (such as those inferred from multiple sequence alignments). Newick trees use pairs of parentheses to group related taxa, separated by a comma (,). Some trees include numbers (branch lengths) that indicate the distance on the evolutionary tree from that taxa to its most recent ancestor. If these branch lengths are present they are prefixed with a colon (:). The Newick format is produced by phylogeny programs such as PHYLIP, PAUP*, Tree-Puzzle and PHYML. Geneious can import and export trees (including bootstrap values and branch lengths) in Newick format.

Nexus tree

The Nexus format was designed to standardize the exchange of phylogenetic data, including sequences, trees, distance matrices and so on. The format is composed of a number of blocks such as TAXA, TREES and CHARACTERS. Each block contains pre-defined fields. Geneious imports and exports files in Nexus format, and can process the information stored in them for analysis.

If you want to export a tree in a format that preserves **bootstrap** values make sure you export with metacomments enabled, otherwise the bootstraps will be lost.

PDB structure

Protein Databank files contain a list of XYZ co-ordinates that describe the position of atoms in a protein. These are then used to generate a 3D model which is usually viewed with Rasmol or SPDB viewer. Geneious can read PDB format files and display an interactive 3D view of the protein structure, including support for displaying the protein's secondary structure when the appropriate information is available.

PDF

PDF stands for Portable Document Format and is developed and distributed by Adobe Systems. It contains the entire description of a document including text, fonts, graphics, colors, links and images. The advantage of PDF files is that they look the same regardless of the software used to create them. Some word processors are able to export a document into PDF format. Alternatively, Adobe Writer can be used. You can use Geneious to read, store and open PDF files.

ACE/PHRAP assembly

Ace is the format used by the Phrap/Consed package, created by the University of Washington Genome Center. This package is used mainly to assemble sequences.

GCG PileUp alignment

The PileUp format is used by the pileup program, a part of the Genetics Computer Group (GCG) Wisconsin Package.

PIR/NBRF sequences

Format used by the Protein Information Resource, a database established by the National Biomedical Research Foundation

Qual quality/Phred scores

Quality file which must be in the same folder as the sequence file (FASTA format) for the quality scores to be used.

Unformatted sequence

A file containing only a sequence.

RSF rich sequences

RSF (Rich Sequence Format) files contain one or more sequences that may or may not be related. In addition to the sequence data, each sequence can be annotated with descriptive sequence information.

CSV/TSV (Comma/Tab Separated Values) sequences

Sequences such as primer lists are often stored in spreadsheets. Geneious has an importer that can be given the field values for a spreadsheet file exported in CSV or TSV format, and it will import them and convert them to documents as well as preserving the additional field contents. It can handle nucleotide and amino acid sequences, as well as primers and probes. For more information on importing primers from a spreadsheet, see the PCR Primers section.

SAM/BAM alignment

SAM and BAM format are produced and used by SAMtools. SAM/BAM files contain the results of an assembly in the form of reads and their mappings to reference sequences.

Chromatograms

Sequence chromatogram documents contain the results of a sequencing run (the trace) and a guess at the sequence data (base calling).

Informally, the trace is a graph showing the concentration of each nucleotide against sequence positions. Base calling software detects peaks in the four traces and assigns the most probable base at more or less even intervals.

VCF variant calls

The VCF format contains sequence annotation information. You can use a VCF file to annotate existing sequences in your local database, import entirely new sequences, or import the annotations onto blank sequences.

Vector NTI®

Geneious supports the import of several Vector NTI formats:

- ***.gb and *.gp formats** These formats are used in Vector NTI for saving single nucleotide and protein sequence documents. They are very similar to the GenBank formats with the same extensions, although they contain some extra information.
- ***.apr format** This format is used for storing alignments and trees made with AlignX, Vector NTI's alignment module.
- ***.ma4, *.pa4, *.oa4, *.ea4 and *.ca6 formats** These are the archive formats which Vector NTI uses to export whole databases.
- ***.cep format** This format is produced by the ContigExpress module and Geneious will import sequences (including the positions of the base calls), traces, qualities, trimmed regions, annotations and editing history for individual reads and contigs.

3.3 Importing files from public databases

Geneious is able to communicate with a number of public databases hosted by the National Centre for Biotechnology Information (NCBI), as well as the UniProt database. You can access these databases through the web at http://www.ncbi.nlm.nih.gov and http://www.uniprot.org/ respectively. These are all well known and widely used storehouses of molecular biology data - more information on each is given in the sections below.

You can search these databases through Geneious by selecting the NCBI or Uniprot database folders at the bottom of the Sources tree, and entering your search term. Press Enter or the Search button to initiate the search. If you get a connection error, you may need to configure your network connections manually, as described in Section 1.2.5.

For advanced search options, click the **More Options** button. This allows you to search for specific terms in specific fields of the Genbank or Uniprot documents, such as specific organisms or author names. By clicking the '+' icon you can search in multiple fields at once, and choose to match either "Any" of the fields (if only one of the fields needs to match), or "All" of the fields (if all of the fields must match). For more on advanced search options, see section 4.2.

If you have a list of known accession numbers that you wish to download, you can enter these in the Search box separated by a comma. For consecutive accessions, enter the first and last numbers separated by a colon, and append [accn] to this. E.g. Entering "AB000001:AB000009[accn]" will download all accessions between AB000001 and AB000009.

The results will appear in the Document table as they are found. The Search button changes to a Stop button while the search is running, and this can be clicked at any time to terminate the search. As the results are downloaded, you will see a small padlock icon in the status bar above the Document Table, which indicates that these items cannot be modified in any way. **You must drag the file into a folder in your local database if you wish to retain the file and/or modify it.** If you don't drag the documents from a database search into your local folders the results will be lost when Geneious is closed. For more information on how to move files between folders in your database, see 4.1.1.

Note: When searching the Genome, Gene or PopSet databases, the documents returned are only summaries. To download the whole genome, select the summary(s) of the genome(s) you would like to download and the click the **Download** button inside the document view or just above it. Alternatively you can choose **Download Documents** in the **File** menu and in the popup menu when document summary is right-clicked (Ctrl+click on Mac OS X). The size of these files is not displayed in the Documents Table. Be aware that whole genomes can be very large and can take a long time to download. You can cancel the download of document summaries by selecting **Cancel Downloads** from any of the locations mentioned above.

3.3.1 UniProt

This database is a comprehensive catalogue of protein data. It includes protein sequences and functions from Swiss-Prot, TrEMBL, and PIR.

3.3.2 NCBI (Entrez) databases

NCBI was established in 1988 as a public resource for information on molecular biology. Geneious allows you to directly download information from nine important NCBI databases and perform NCBI BLAST searches (Table 3.1).

Table 3.1: NCBI databases accessible via Geneious

Database	Coverage
Gene	Genes
Genome	Whole genome sequences
Nucleotide	DNA sequences
PopSet	sets of DNA sequences from population studies
Protein	Protein sequences
PubMed	Biomedical literature citations and abstracts
SNP	Single Nucleotide Polymorphisms
Structure	3D structural data
Taxonomy	Names and taxonomy of organisms

Entrez Gene. Entrez Gene is NCBI's database for gene-specific information. It does not include all known or predicted genes; instead Entrez Gene focuses on the genomes that have been completely sequenced, that have an active research community to contribute gene-specific information, or that are scheduled for intense sequence analysis.

The Entrez Genome database. The Entrez genome database has been retired. For backwards compatibility Geneious simulates searching of the old genome database by searching the Entrez Nucleotide database and filtering the results to include only genome results.

The Entrez Nucleotide database. This database in GenBank contains 3 separate components that are also searchable databases: "EST", "GSS" and "CoreNucleotide". The core nucleotide database brings together information from three other databases: GenBank, EMBI, and DDBJ. These are part of the International collaboration of Sequence Databases. This database also contains RefSeq records, which are NCBI-curated, non-redundant sets of sequences.

The Entrez Popset database. This database contains sets of aligned sequences that are the result of population, phylogenetic, or mutation studies. These alignments usually describe evolution and population variation. The PopSet database contains both nucleotide and protein sequence

data, and can be used to analyze the evolutionary relatedness of a population.

The Entrez Protein database. This database contains sequence data from the translated coding regions from DNA sequences in GenBank, EMBL, and DDBJ as well as protein sequences submitted to the Protein Information Resource (PIR), SWISS-PROT, Protein Research Foundation (PRF), and Protein Data Bank (PDB) (sequences from solved structures).

The PubMed database. This is a service of the U.S. National Library of Medicine that includes over 16 million citations from MEDLINE and other life science journals. This archive of biomedical articles dates back to the 1950s. PubMed includes links to full text articles and other related resources, with the exception of those journals that need licenses to access their most recent issues.

Entrez SNP. In collaboration with the National Human Genome Research Institute, The National Center for Biotechnology Information has established the dbSNP database to serve as a central repository for both single base nucleotide subsitutions and short deletion and insertion polymorphisms.

The Entrez Structure database. This is NCBI's structure database and is also called MMDB (Molecular Modeling Database). It contains three-dimensional, biomolecular, experimentally or programmatically determined structures obtained from the Protein Data Bank.

Entrez Taxonomy. This database contains the names of all organisms that are represented in the NCBI genetic database. Each organism must be represented by at least one nucleotide or protein sequence.

3.3.3 Literature searching

Geneious allows you to search for relevant literature in NCBI's PubMed database. The results of this search are summarized in columns in the Document Table and include the PubMed ID (PMID), first and last authors, URL (if available) and the name of the Journal.

When a document is selected, the abstract of the article is displayed in the Document Viewer (Text View tab) along with a link to the full text of the document if available, and a link to Google Scholar, both below the author(s) name(s) (Figure 3.2).

As well as the abstract and links, Geneious also shows the summary of the journal article in BibTex format in a separate tab of the Document Viewer (BibTex tab). BibTex is the standard LaTeX bibliography reference and publication management data format and the information in the BibTex screen can be imported directly into a LaTeX document when creating a bibliography. Alternatively, a set of articles in Geneious can be directly exported to an EndNote 8.0 compatible format. This is usually done when creating a bibliography for Microsoft Word documents.

Note: If the full text of the article is available for download in PDF format, it can also be stored in Geneious by saving it to your hard drive and then importing it. This will allow full-text searches to be performed on the article. To view a `.pdf` document either double click on the

document in the Documents Table or click on the **View Document** button. This opens the document in an external PDF viewer such as Adobe Acrobat Reader or Preview (Mac OS X). On Linux, you can set an environmental variable named "PDFViewer" to the name of your external PDF viewer. The default viewers on Linux are `kpdf` and `evince`.

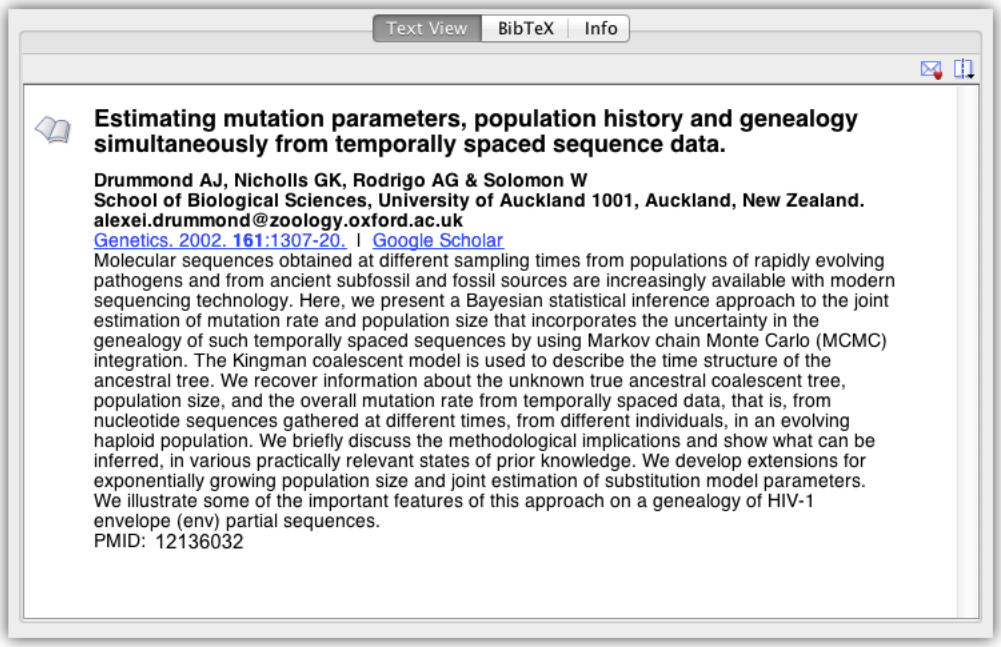


Figure 3.2: Viewing bibliographic information in Geneious

3.4 Agents

Databases searches can be automated using an **Agent**, allowing you to continuously receive the latest information on genomes, sequences, and protein structures. Each agent is a user-defined, automated search. You can instruct an agent to search any Geneious accessible database at regular intervals (e.g. weekly) including your contacts on Collaboration. This simple but powerful feature ensures that you never miss that critical article or DNA sequence. To manage agents click on the agent icon in the toolbar. An agent has to be set up before it can be used.

3.4.1 Creating agents

To set up an Agent click the **Agents** icon and the **Create** button. You now need to specify a set of search criteria including the database to search, key words to search on, search frequency and the folder you wish the agent to deliver its results to.

The search frequency may be specified in minutes, hours, days or weeks. You can only use whole numbers.

Selecting **Only get documents created after today** will cause the agent to check what documents are currently available when the agent is created. Then when the agent searches it will only get documents that are new since it was created. This is useful if, for example, you have already read all publications by a particular author and you want the agent to only get new publications.

Alternatively you can click the **Create Agent...** button which is available in some advanced search panels. This will use the advanced search options you have entered to create the agent.

The easiest way to organize your search results is to create a new folder and name it appropriately. You can do that by navigating to the parent folder in the **Deliver to** box and clicking **New Folder**, or by creating a new folder beforehand as follows:

1. Right-click (Ctrl+click on Mac OS X) on the **Sample Documents** or **Local** folders. This brings up a popup menu with a **New Folder...** option.
2. Create a new folder and name it according to the contents of the search. (For example, type "CytB" if searching for cytochrome b complex.)
3. Once created, select the new folder. You can now select the **Create** or **Create and Run**. The agent will then be added to the list in the agent dialog and it will perform its first search if you clicked Create and Run. Otherwise it will wait until its next scheduled search.

All downloaded files are stored in the destination folder and are marked "unread" until viewed for the first time.

3.4.2 Checking agents

Once you have created one or more agents, Geneious allows you to quickly view their status in the agents window which is accessible from the toolbar. Your agents' details are presented in several columns: **Enable**, **Action**, **Status** and **Deliver To**.

Enable: This column contains a check box showing whether the agent is enabled. **Action**: This summarizes the user-defined search criteria. It contains:

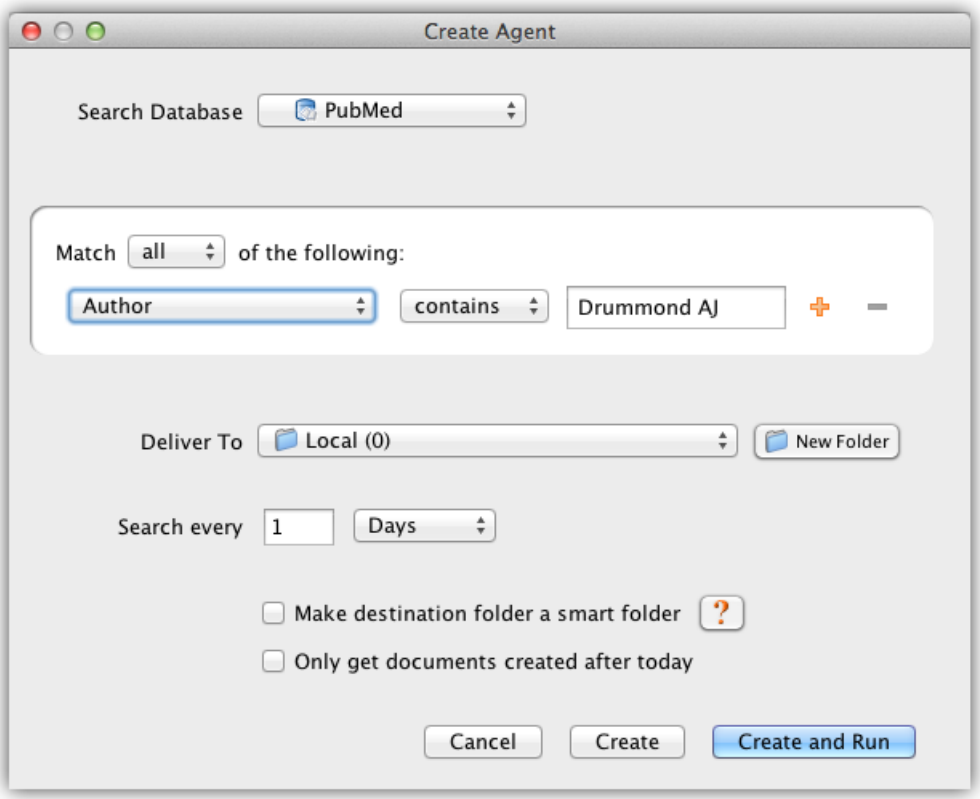


Figure 3.3: The Create Agent Dialog

1. Details of the database accessed. For example, Nucleotide and Genome under NCBI.
2. The search type the Agent performed, e.g. "keyword".
3. The words the user entered in the search field for the Agent to match against.

Status: This indicates what the Agent is currently doing. The status will be one of the following:

- "Next search in x time" e.g. 18 hours. The agent is waiting until its next scheduled search and it will search when this time is reached.
- "Searching." These are shown in bold. The agent is currently searching.
- "Disabled." The agent will not perform any searches.
- "Service unavailable." The agent cannot find the database it is scheduled to search. This will happen if the database plugin has been uninstalled or if for example the Collaboration contact is offline currently.
- "No search scheduled" The agent is enabled but doesn't have a search scheduled. To correct this click the "Run now" button in the agent dialog to have it search immediately and schedule a new search.

Deliver To: This names the destination folder for the downloaded documents. This is usually your Local Documents or one of your local folders.

Note. If you close Geneious while an agent is running, it will stop in mid-search. It will resume searching when Geneious is restarted.

3.4.3 Manipulating an agent

Once an agent has been set up, it can be disabled, enabled, edited, deleted and run. All these options are available from within the Agents dialog.

- **Enable** or **disable** an agent by clicking the check box in the Enable column.
- **Run Now** Cause the agent to search immediately
- **Cancel** If the agent is currently searching, click this to stop the search.
- **Edit** Click this to change an agent's database, search criteria, destination or search interval.
- **Delete** Delete the agent permanently. Any documents retrieved by the agent will remain in your local documents.

3.5 Exporting files

To export files from Geneious, select the file or files you want to export and click **File** → **Export** → **Selected Documents**. Each data type has several export options, as detailed in the table below. Any set of documents may be exported in Geneious native format, and these files are back-compatible to Geneious version 6.0 (e.g. Files exported in .geneious format from Geneious R7 or R8 can be imported to Geneious R6 or later).

If you wish to export the parents and/or descendants of a particular file at the same time, choose **File** → **Export** → **with Parents/Descendants**.

Data type	Export format options
DNA sequence	FASTA, Genbank XML, Genbank flat, Geneious
Amino acid sequence	FASTA, Genbank XML, Genbank flat, Geneious
Chromatogram sequence	ABI, Geneious
Sequence with quality	FastQ, Qual, Geneious
Annotation	GFF, BED, Geneious
Multiple sequence alignment	Phylip, FASTA, NEXUS, MEGA, Geneious
Assembly	Phrap ACE, Geneious, SAM/BAM
Variant calls	VCF (single sample only)
Phylogenetic tree	Phylip, FASTA, NEXUS, Newick, MEGA, Geneious
PDF document	PDF, Geneious
Publication	EndNote 8.0, Geneious
Graphs	CSV, WIG
Document Properties	CSV, TSV, Geneious

Documents imported in any chromatogram or molecular structure format can be re-exported in that format as long as no changes have been made to the document.

Both fasta and fastq files can be exported in compressed (fasta.gz and fastq.gz) format for smaller file size. If exporting paired reads in fasta format, an option to export the forward and reverse reads to separate files is available by choosing 'Fasta Paired Files' as the file type option.

3.5.1 Export to comma-separated (CSV) or tab-separated (TSV) files

The values displayed in the document table can be exported to a csv file which can be loaded by most spread sheet programs. When choosing to export in csv format Geneious will also present a list of the available columns in the table (including hidden ones) so you can choose which to export.

3.5.2 Exporting multiple files

There are several options for export of multiple files from Geneious:

- **Export to a single file**: Multiple files can be exported to a single file by selecting all the files you wish to export and going to **File** → **Export** → **Selected Documents**. This will combine all the files you selected into a single file for export.
- **Batch Export**: This option exports each selected file as a different document. E.g. you can select several sequence documents and use Batch Export to export each sequence as an individual fasta file. The options for batch export let you specify the format and folder to export to as well as the extension to use. Each file will be named according to the Name column in Geneious.
- **Export Folder**: To export an entire folder to a single file, click on the folder in the Sources panel and go to **File** → **Export** → **Export Folder**. Note that folders can only be exported in .geneious format.

3.6 Printing and Saving Images

Geneious allows you to print (or save as an image) the current display for any document viewer. This includes the sequence viewer, tree view, dotplot, and text view.

3.6.1 Printing

Choose **Print** from the file menu. The view is printed without the options panel. It is recommended to turn on **Wrap sequence** and deselect **Colors** before printing. Wrapping prints the sequence as seen in the sequence viewer and the font size is chosen to fill the horizontal width of the page. The following options are available:

Portrait or landscape. Controls the orientation of the page.

Scale. Can be used to decrease or increase the size of everything in the view, while still printing within the same region of the page. For many types of document views, this will cause it to wrap to the following line earlier, usually requiring more pages.

Size. Controls the size the printed region on the paper. Effectively, increasing the size, reduces the margins on the page.

3.6.2 Saving Images

Choose **Save as image file** from the file menu. The following options are available:

Size. Controls the size of the image to be saved. Depending on the document view being saved, these may be fixed or configurable. For example, with the sequence viewer, if wrapping is on, you are able to choose the width at which the sequence is wrapped, but if wrapping is off, both the width and height will be fixed.

Format. Controls image format. Vector formats (PDF, SVG and EMF) are ideal for publication because they won't become pixelated. Raster formats (PNG and JPG) are easier to share, great for emailing and posting on the web. If you wish to edit the file outside of Geneious, SVG or EMF format should be used. SVG files can be edited in tools such as Adobe Illustrator or Inkscape, and EMF files can be edited on Windows using PowerPoint, or LibreOffice Draw on Mac or Linux (the Mac version of PowerPoint can't modify EMF files). With SVG or EMF it is possible to ungroup components of the graphic for editing, and because they are vector graphics they will scale without becoming pixelated.

Resolution. Only applies to raster formats (PNG and JPG) and is used to increase the number of pixels in the saved image. We recommend increasing the resolution to at least 300% for printing PNG or JPG files.

Chapter 4

Managing your Local Documents

4.1 Organizing your local documents

Geneious documents are stored in a hierarchal arrangement of folders under the **Local** folder in the Sources Panel. Clicking on a folder will display its contents in the Document Table. Next to each folder name in the hierarchy is the number of documents it contains in brackets. When the Local folder or a sub-folder is collapsed (minimized), the brackets next to the folder shows how many files are contained in that folder as well as all of its sub-folders. In addition, if some of the documents in a folder are unread, the number of unread documents will also appear in the brackets.

The document types available are listed in Table 4.1.

To create a new folder in Geneious, select the **Local** folder or a sub-folder icon in the Sources panel and either right-click (Ctrl+click on Mac OS X) and select **New folder** from the popup menu, or go to **File** → **New Folder**. This will open a dialog where you can name your new folder. This folder will then be created within the folder you originally selected.

You can also **delete**, **rename**, **move**, **export** or **change the color** of a folder by right-clicking on the folder (or control-click on MacOSX) and selecting the option you require from the menu. These options are also available under the **File** menu. In Mac OS X, you can also use the plus (+) and minus (-) buttons located at the bottom of the service panel to create and delete folders.

4.1.1 Moving files around

Files can be moved between folders in a number of ways:

Drag and drop. This is quickest and easiest. Select the documents that you want to move. Then, while holding the mouse button down, drag them over to the desired folder and release. If you

Table 4.1: Geneious document types

Document type	Geneious Icon
Nucleotide sequence	
Oligo sequences	
Enzyme Sets	
Chromatogram	
Contig	
Protein sequence	
Phylogenetic tree	
3D structure	
Sequence alignment	
Journal articles	
PDF	
Other documents	

dragged documents from one local folder to another, this action will move the documents – so that a copy of the document is not left in the original location. In external databases such as NCBI the documents will be copied, leaving one in its original location.

Drag and copy. While dragging a document over to your folder, hold the Ctrl key (Alt/Option key on Mac OS X) down. This places a copy of the document in the target folder while leaving a copy in the original location. This is useful if you want copies in different folders. Folders themselves can also be dragged and dropped to move them but they cannot be copied.

The Edit menu. Select the document and then open the Edit menu on the menu bar. Click on **Cut** (Ctrl+X/command+X), or **Copy** (Ctrl+C/command+C). Select the destination folder and **Paste** (Ctrl+V/command+V) the document into it.

4.1.2 Deleting Data and the Deleted Items folder

The **Deleted Items** folder is located underneath the local document folders in the Sources panel. When a folder or document is deleted, Geneious moves the data to the Deleted Items folder instead of erasing it immediately. This means the data can be recovered if it was deleted by mistake. Pressing the **Delete** key is the easiest way to move the selected folder or documents to the Deleted Items folder.

To recover documents or folders from **Deleted Items** you can either drag and drop them to another folder or use **Restore from Deleted Items** (**Put Back from Deleted Items** on Mac OS) in the File menu to automatically move them to folder they were deleted from.

The Deleted Items folder should be cleared periodically to keep hard drive space free. This can be done by selecting **Erase All Deleted Items** from the File menu. Geneious will warn you if the Deleted Items folder contains a large amount of data.

To erase a document immediately without moving it to **Deleted Items**, use **Erase Document Permanently** in the File menu (or press Shift+Delete).

Many of these actions can also be accessed by right clicking on a folder or document.

4.2 Searching and filtering local documents

The local database can be searched by clicking on the folder you wish to search, then clicking the **Search** icon (Search) in the top right of the toolbar. Enter the desired term(s) in the text field and press enter or click the **Search** button. Once a search starts the document table will initially be emptied, then results will appear in the table as they are found (see Figure 4.1). The **Search** button changes to a **Cancel** button while a search is in progress and this may be clicked at any time to terminate the search. Feedback on a search progress is presented in the status bar directly below the toolbar.

To exit the search window and return to browsing click the orange X in the search options bar, or press the Escape key while the cursor is in the search text field.

Important: You must use quotation marks (“”) if “!”, “@”, “$”, and blank spaces (“ ”) are part of your search criteria. No quotation marks lead to unreliable results.

Autocompletion of search words

Geneious remembers previously searched keywords and offers an auto-complete option. This works in a similar way to Google or predictive text on your mobile phone. If you click within the search field, a drop-down box will appear showing previously used options.

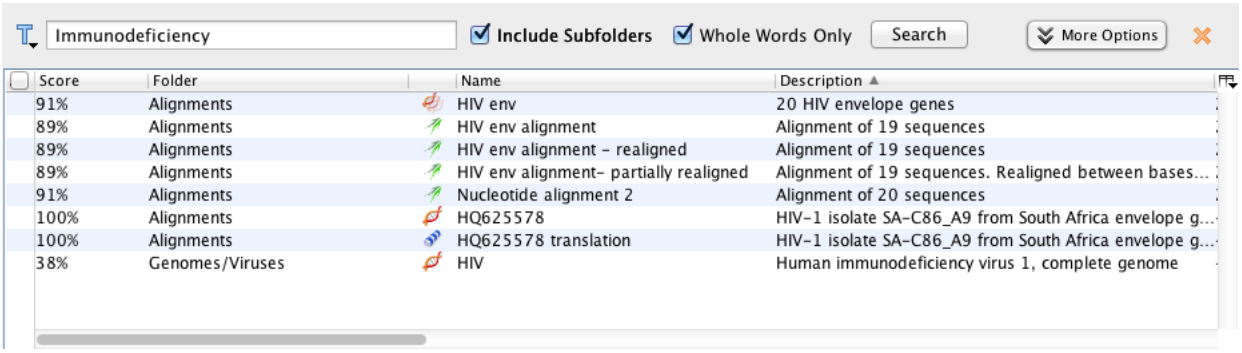


Figure 4.1: Searching the Document Table

Wild card searches

When you are looking for all matches to a partial word, use the asterisk (*). Asterisks can be placed at the beginning, end or in the middle of a word. For example, typing "oxi*" would return matches such as oxidase, oxidation, oxido-reductase, and oxide. Searching for CO*I would return matches for COI and COXI. Similarly, you can use a question mark '?' to represent any single character. This feature is available only for local documents and not NCBI or UniProt searches.

4.2.1 Advanced Search options

To access advanced search click the **More Options** button inside the basic search panel. To return to basic search click the **Fewer Options** button. Switching between advanced and basic will not clear the search results table.

The advanced search allows you to search for specific terms in specific fields of your documents. The fields available for a search can be found in the left-most drop-down box; all fields potentially available on your local documents are listed here. If you have defined a new type of meta-data in Geneious, and that meta-data field has been added to a document, then this field will also be available to search.

Advanced Search also provides you with a number of options for restricting the search on a field depending on the field you are searching against. For example, if you are using numbers to search for "Sequence length" or "No. of nodes" you can further restrict your search with the second drop-down box:

- "is greater than" (>)
- "is less than" (<)

- "is greater than or equal to" ($\geq$)
- "is less than or equal to" ($\leq$)

Likewise if you are searching on the "Creation Date" search field you have the following options

- "is before or on"
- "is after or on"
- "is between"

When searching your local folders you have the option of searching by "Document type". The second drop-down list provides the options "is" and "is not". The third drop-down lists the various types of documents that can be stored in Geneious such as "3D-Structure", "Nucleotide sequence", and "PDF" (see Figure 4.2).

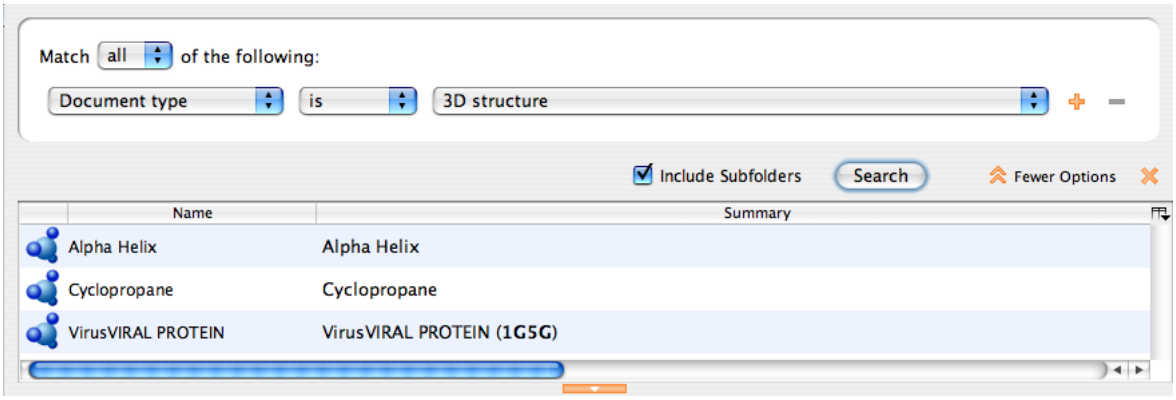


Figure 4.2: Document type search options

And/Or searches

The advanced options lets you search using multiple criteria. By clicking the "+" button on right of the search term you can add another search criteria. You can remove search criteria by clicking on the appropriate "-" button. The "Match all/any of the following" option at the top of the search terms determines how these criteria are combined:

Match "Any" requires a match of one or more of your search criteria. This is a broad search and results in more matches.

Match "All" requires a match all of your search criteria. This is a narrow search and results in fewer matches.

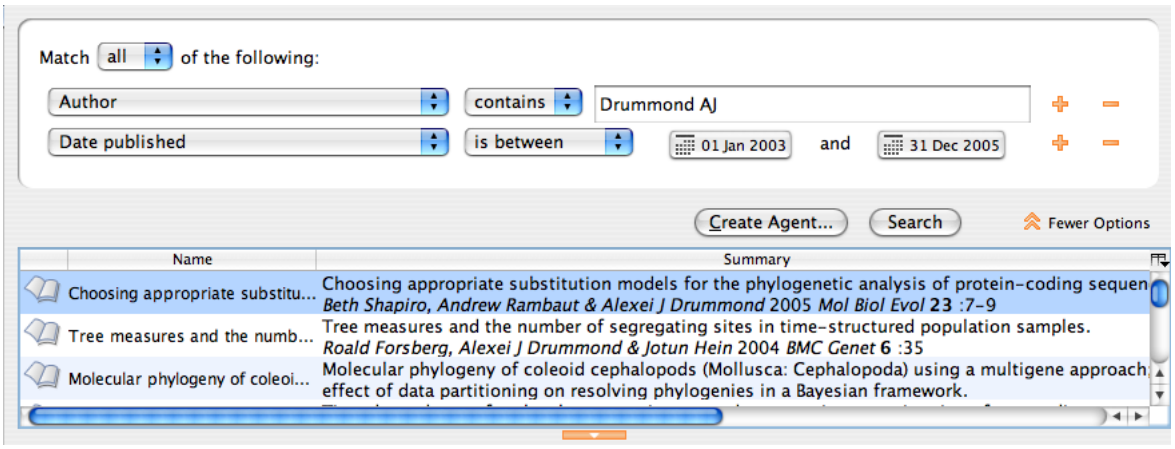


Figure 4.3: Advanced Search

4.2.2 Similarity ("BLAST-like") searching

It is possible to search your local documents not only for text occurrences but by similarity to sequence fragments. Click the small arrow at the bottom of the large T to the left of the search dialog, select **Nucleotide similarity search** or **Protein similarity search** and enter the sequence text. Geneious will try to guess the type of search based on the text, so that simply entering or pasting a sequence fragment may change the search type automatically.

The search locates documents containing a similar string of residues, and orders them in decreasing order of similarity to the string. The ordering is based on calculating an E-value for each match. You can read more about the E-value in chapter 15.

For the search to be successful, you need to specify a minimum of 11 nucleotides and 3 amino acids. Note that search times depend on the number and size of your sequence documents, and so may take a long time to complete.

4.2.3 Find in Document

The **Find in Document** option under the **Edit** menu allows you to search for a particular motif in your sequences, annotations or document names. For example, you can search for a particular string of nucleotides in sequences or alignments, or search for annotations and sequences by name or number. The search can optionally be made case sensitive. Matching regions are selected in the document at the end of the search.

The shortcut for this function is Ctrl+F. To find the next match for the text specified in the dialog you can use F3 or Ctrl+G, and to find the previous match use Ctrl+Shift+G or Shift+F3.

4.2.4 Filtering and Similarity sorting

The Filter () allows you to instantly identify documents in the document table matching chosen keywords. It is located in the top right hand corner of the Main Toolbar.

Type in the text you are searching for and Geneious will display all the documents that match this text and hide all other documents in the Document Table. To view all the documents in a folder, clear the Filter box of text or click the  button.

It is also possible to sort sequences in a given folder by their similarity. To use this function, select a single sequence in the document table and right-click, then choose **Sort**. **Sort by similarity** will rank all other sequences by their similarity to the selected sequence. The most similar sequence is placed at the top and the least similar sequence at the bottom. This also produces an E-value column describing how similar the sequences are to the selected one. The **Remove Sort by Similarity** option will remove the E-value column and return the table to its previous sorting.

Filtering on-the-fly

Filtering can be used while searching for documents via public databases, filtering data as it is being downloaded. Type in the appropriate text in the Filter Box and only those documents that match both the original criteria (as specified by the search terms) and the "Filter" text will be displayed. This is an effective way of filtering within your search results.

4.3 Find Duplicates

Find Duplicates, under the **Edit** menu, is used to identify duplicate copies of sequences and other documents. Duplicates can be identified by sequence name, database ID (e.g. accession) or by the residues/bases, and the **Search Scope** can be set so that it checks within either a selected set of documents, all documents in a folder or in the sequences of a single alignment or sequence list.

When searching for duplicates within sequences of a single alignment or sequence list, two options are available for displaying results once the search has run:

- **Select earlier duplicates in list**: This will select all but one copy of a duplicated document, allowing the duplicates to easily be deleted or moved to another folder leaving one copy behind.
- **Extract unique sequences**: Unique sequences will be extracted to a new sequence list, and the sequence names modified to show the duplicate count for that sequence. For

large data sets, or removing duplicates in paired reads, or removing non-exact duplicates, see Remove Duplicate Reads using BBTools.

If you are searching for duplicates within a folder or multiple select documents, you can choose to select either the most recently or least recently modified copy.

Remove Duplicate Reads

For identifying non-exact duplicates, removing exact duplicates from large data sets, or removing duplicates on paired read data sets, use **Remove Duplicate Reads...** from the **Sequence** menu. This tool runs Dedupe from the BBTools suite.

For a detailed explanation of any Dedupe setting, hover the mouse over the setting, or click the help (question mark) button next to the custom options under **More Options**.

4.4 Batch Rename

Batch rename is located under the **Edit** menu and can be used to edit any field in multiple documents at once. It can also be used to batch edit any property of sequences within an alignment or sequence list.

Existing fields can be replaced with a combination of values from other fields (e.g. Name replaced with Organism and/or Accession), and fixed text can be added to the beginning or end of existing fields.

The advanced options (under the **More Options** button) enables the use of regular expressions to replace a specific part of one field with another property or text string. Click the **Help** button in this section for more information on formatting expressions.

4.5 Backing up your local documents

It is important to keep frequent back ups of your data because computers can fail suddenly and unexpectedly. A computer can be replaced, but your data is much harder to replace. The best way to back up all of your data and settings in Geneious is to use the *Back Up* button in the toolbar or select *Back Up Data* in the File menu (Figure 4.4) .

Note: Due to the way the local database works, it is important that Geneious is not accessing the database when a backup is taken. For example, Mac users with Time Machine will have backups taken during the day but if Geneious is running when those backups are taken, they will not be suitable for restoring from. However, backups taken overnight when Geneious isn't running should be fine.

Backing up your data directory manually is not recommended because the Geneious database structure is complex and many programs will fail to back it up properly.

The back up command has two options:

- **Export selected folder:** This will export the selected folder (including all subfolders) to a Geneious format file. This allows you to back up an individual project within your database. The backup can also be imported in to an existing database by drag and drop. If you have finished working on a project it is a good idea to back it up in this way then delete it from inside Geneious to keep the size of your database down and improve the performance of Geneious. You should keep archive backups in addition to these because this backup will miss your settings and data outside the selected folder.
- **Archive all data and settings:** This is equivalent to creating a zip archive of your entire Geneious data directory which includes all your data, preferences, searches and agents. This option will cause Geneious to cease working on the local database while it creates the archive. This type of backup cannot be directly imported in to an existing database, when it is loaded everything in Geneious will revert to how it was when you took the backup.

Backups should be stored on another drive, or can be left to general system backups safely since they are made when Geneious is in a non-running state. These backups can also be safely moved around including to other machines.

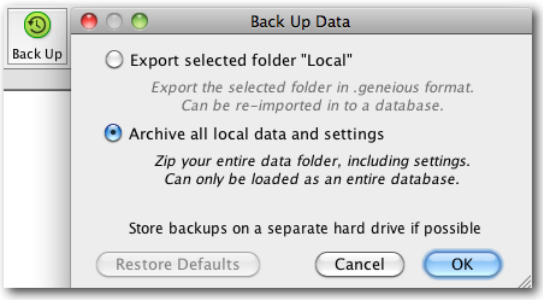


Figure 4.4: Using the backup tool

4.5.1 Restoring a backup

- **Geneious format file** The easiest way to restore this is to drag and drop the Geneious file in to the folder in Geneious where you want it to go. Alternatively you can use *Restore Backup* in the File menu and the backup will be added under the *Local* folder in your current database.

- **Archive all data and settings:** It is strongly recommended that you use *Restore Backup* in the file menu to load the zip file rather than unzipping it manually. Some operating systems may not be able to unzip the data correctly. The *Restore Backup* command will unzip your backed up data directory to a folder of your choosing which you can then load immediately. If you choose not to load it immediately you can switch to the restored data directory by going to *Preferences* in the Tools menu and changing the *Data Storage Location* on the General tab.

4.6 Document History

The history of an document can be viewed by going to the **Info** tab above the sequence viewer, and choosing **History**. This displays information on how the document was created, plus a record of each time it has been modified. The exact information displayed is flexible, but is the entries will always include the time and user responsible for the edit. An entry may also reference other documents via hyperlinks, and has the ability to display a re-creation of the options used.

Saving of history can be disabled for performance or privacy reasons by going to the Appearance and Behaviour tab in Preferences, see section 1.2.1.

Chapter 5

Creating, viewing and editing sequences

5.1 Creating new sequences

New sequences can be imported from existing files as described in chapter 3, or they can be created manually by going to **Sequence** → **New Sequence**. Here you can paste or type in the residues for your new sequence, then enter the Name, Description and Organism for your sequence if required (see figure 5.1). Geneious will automatically determine whether your sequence is nucleotide or protein based on the composition of the bases you enter. You can change this by clicking the **Type** option. If your sequences are oligonucleotides, choose **Primer** or **Probe** as the type.

To create a new sequence from an existing sequence, select the region of sequence that you want then click the **Extract** button above the sequence viewer, or go to **Sequence** → **Extract Regions**. This will create a new sequence document containing the selected sequence.

5.1.1 Sequence lists

Sequence lists make it easier to manage large numbers of sequences by grouping related sequences into a single document. When you import files containing multiple sequences you will be asked if you want to store those sequences in a list. To existing sequences in your database into a list, or combine two lists into one, select the sequences or sequence lists you want to group and go to **Sequence** → **Group Sequences into a List**. Note that this copies your sequences into a list and retains the original sequence documents.

To extract sequences from a list, select the sequence(s) you want to extract and go to **Sequence** → **Extract Sequences from List**. This will copy each sequence out of the list into a separate

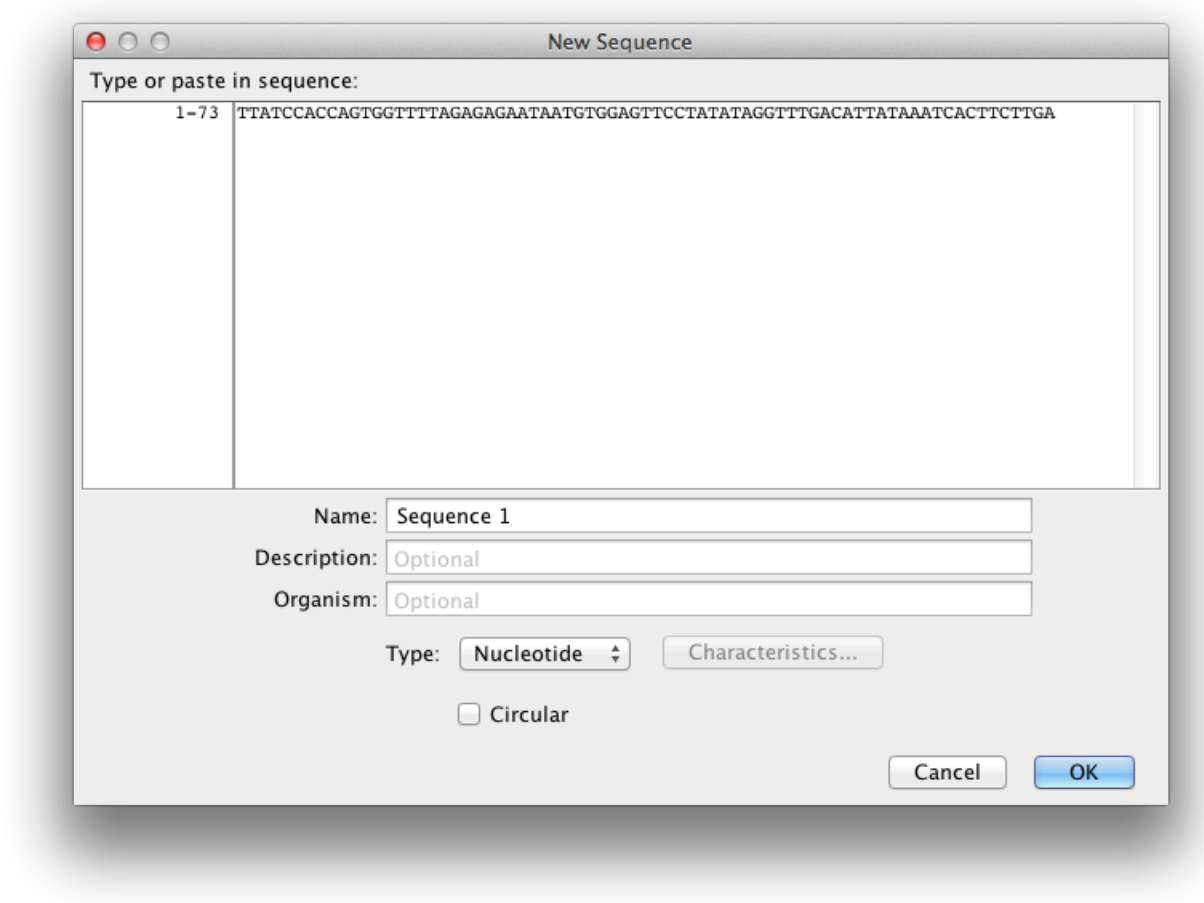


Figure 5.1: Entering a new sequence in Geneious

document, while retaining the original sequence within the list. To remove a sequence from a list entirely, select it and click the Delete button on your keyboard, or go to **Delete** under the **Edit** menu.

5.2 The Sequence Viewer

Sequences are displayed in the viewer below the document table. Annotations (Chapter 8), translations (section 5.5) and analysis graphs (section 5.2.7) are also displayed in this viewer.

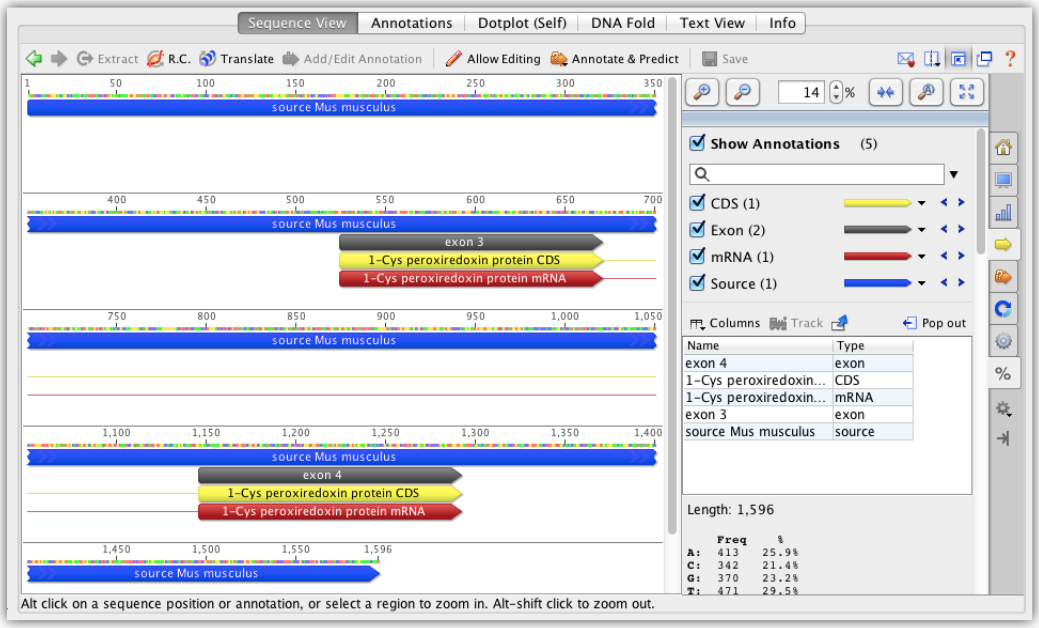


Figure 5.2: A view of an annotated nucleotide sequence in Geneious

5.2.1 Zoom level

Controls for zooming in and out on sequences are located at the top of the side panel, to the right of the sequence viewer. The plus and minus buttons increase and decrease the magnification of the sequence by 50%, or by 30% if the magnification is already above 50%. To zoom in or out by a smaller amount, hold down the alt and/or shift key while clicking the plus or minus button.

zooms in to fit the selected region in the available viewing area.

zooms to 100%. The 100% zoom level allows for comfortable reading of the sequence.

zooms out so as to fit the entire sequence in the available viewing area.

Zooming can also be quickly achieved by holding down the zoom modifier key, which is the Ctrl key on Windows/Linux or the Alt/Option key on Mac OS X, and clicking as described below. When the zoom key is pressed a magnifying glass mouse cursor will be displayed.

- Hold the zoom key and left click on the sequence to zoom in.
- Hold the zoom key and Shift key to zoom out.
- Hold the zoom key and turn the scroll wheel on your mouse (if you have one) to zoom in

and out.

- Hold the zoom key and click on an annotation to zoom to that annotation

You can also pan in the Sequence View by holding Ctrl+Alt (command+Alt on Mac OS X) and clicking on the sequence and dragging.

5.2.2 Selecting part of a sequence

Within the sequence viewer, there are several ways to select part of a sequence, or select a subset of sequences from a list, alignment or assembly:

- **Mouse dragging**: Click and hold down the left mouse button at the start position, and drag to the end position. By using the Ctrl (Windows/Linux) or command (Mac) keys it is possible to select multiple regions of a sequence or alignment, as described in the keyboard shortcuts below.
- **Select from annotations**: When annotations are available, click on any annotation to select the annotated residues. As with mouse dragging, multiple selections are supported.
- **Click on sequence name**: This will select the whole sequence.
- **Select all**: Use the keyboard shortcut Ctrl+A (+A on Mac) to select everything in the panel.

Keyboard shortcuts for selection of sequences:

- To quickly select a single residue, double-click on it.
- To select a block of residues within a single sequence, triple click.
- To select a block of residues across multiple sequences, quadruple click.
- To select a block of 10 residues, hold down Shift and Cntrl (alt/option on a Mac) and press the keyboard arrow.
- To select a specific region of sequence, click at beginning of the region you want to select, hold down Shift and then click the end.
- To modify the right-hand end of a selection, hold down Shift/alt (command on a Mac) and use the right/left arrows to select more or fewer bases. Holding down Shift and Cntrl (alt/option on a Mac) modifies the selection by 10 bases at a time.

- To select the same region across multiple sequences in an alignment, select the region you want in the first sequence, then hold down Shift / alt (command on a mac) and click the down arrow to apply the selection to the sequences underneath. Holding down Shift /Cntrl (alt/option on a mac) while pressing the down arrow will select the sequences in batches of 10.

Go to position

To jump to a particular base in a sequence you can use **Go to base** under the **Edit** menu (for amino acid sequences, this appears as Go to Residue). This allows for the instant navigation to a particular nucleotide or amino acid coordinate for any sequence in the current document selection. It also allows the selection of a particular region of sequence, either for individual sequences or across sequence lists and alignments, or the selection of particular sequences out of a sequence list. Formatting examples are given in the setup dialog. **Go to Position** also appears next to the sequence viewer when in genome view (see section 5.2.4)

5.2.3 Circular sequences

When a circular sequence is selected, the default view is to display the sequence as circular. The view can be rotated by using the scrollbar at the bottom or by turning the mouse wheel. Even though a sequence is circular, you can display it as a linear sequence using the **Linear view on circular sequence** checkbox under the **Layout** section of Advanced.

To change a linear sequence into a circular sequence, select the sequence then go to **Sequence → Circular Sequence**, then click **Save**. This will join the ends of your sequence up to create a circular sequence, but does not check for overlapping ends. Circularization will affect how some operations (such as restriction digest and map to reference) deal with the sequences.

5.2.4 Genome View

The genome view (Figure 5.3) is displayed when sequences larger than 100,000 bp are selected (either as individual sequences or within a sequence list).

The genome viewer contains additional controls which allow for the efficient navigation of large sequences:

1. The **Go to Position** button allows for the instant navigation to a particular nucleotide coordinate for any sequence in the current document selection. It also allows the selection of a particular region of sequence, or the selection of particular sequences out of a sequence list.

2. A **minimap** is shown above the sequence viewer which shows a representation of the entire sequence plus its underlying annotations. The portion of the sequence currently visible in the

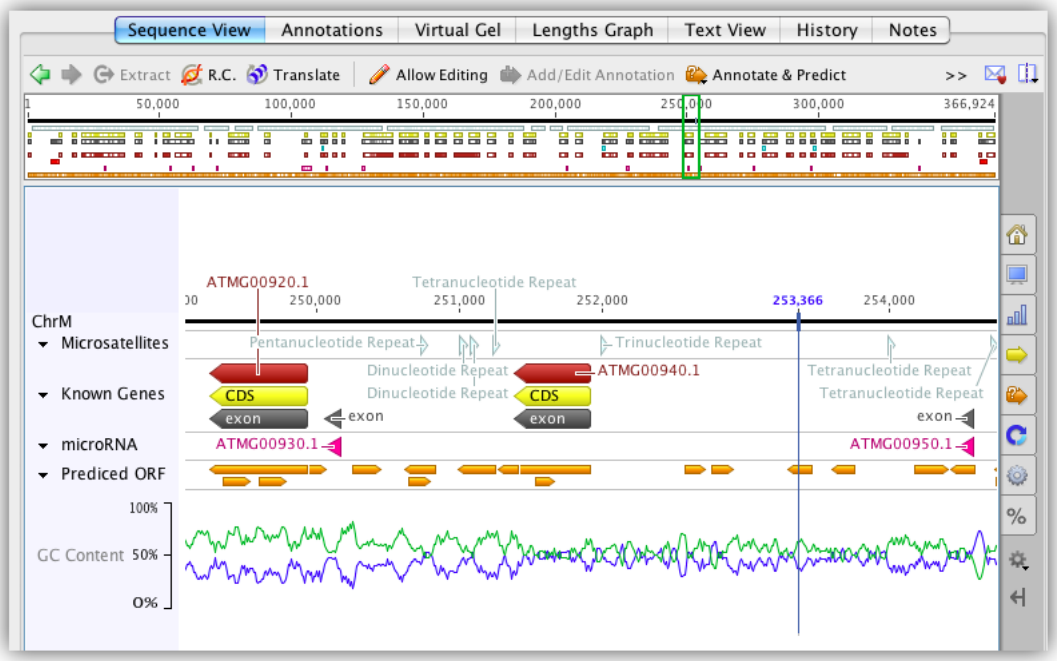


Figure 5.3: The minimap and sequence view, of a chromosome with gene and variation annotations, under the genome viewer configuration

viewing window is highlighted on the minimap, showing the relative position of the visible section to the overall sequence.

The minimap can also be used to quickly navigate around the visible sequence. Clicking on a section of the minimap will jump the sequence viewer to center on that position. Double-clicking the minimap will zoom further in on the clicked section. Finally, highlighting a section of the minimap using a click-drag-release action will display the highlighted region in the sequence viewer.

5.2.5 The Side Panel Controls

The panel to the right of the sequence viewer allows you to control what is displayed in the sequence viewer (e.g. translations, consensus sequences, graphs and annotations), displays sequence statistics, and provides functions for finding annotations, ORFs, and restriction sites on your sequences. A brief description of each tab is given below:

General Options

Contains the color options (see section 5.2.6), check-boxes to turn on and off main aspects of the sequence view and options for what to display as the name of each sequence.

Display

Contains options for displaying the translation and/or complement of a sequence, and turning off the original nucleotide sequence. See sections 5.4 and 5.5 for more information. This tab is not displayed for protein sequences.

Graphs

This option is visible when viewing nucleotide or protein sequences, chromatogram traces, sequence alignments or assemblies, and includes graphs for GC content, Identity, Coverage, and Quality. The graphs available for display depend on the type of sequence you are viewing. More detail on graphs for nucleotide and protein sequences is given in section 5.2.7 and for alignment and contig graphs is given in section 9.3.2.

Annotations

On sequences containing annotations this tab will show a yellow arrow. It contains controls for turning on and off annotations of each type, customising the way each type is displayed, and filtering based on annotation name or type. See chapter 8 for more information on working with annotations.

Live Annotate and Predict

Contains real-time annotation generators such as Annotate from Database, Find ORFs and Transfer Annotations. To use one of these, turn on the check-box at the top of the

generator you want to use and annotations will immediately be added to the sequence. You can then change settings for the generator and the annotations will change on the sequence in real-time as you do. If you want to save the annotations permanently on the sequence click Apply.

Restriction Analysis

This behaves similarly to the Live Annotate & Predict section above. Please refer to chapter 14 for full details.

Advanced

Contains advanced options for controlling the look of sequences and alignments, including wrapping, numbering, annotation placement and font sizes. See section 5.2.8 for more information.

Statistics

Displays statistics about the sequence or alignment currently being viewed, such as length, molecular weight and nucleotide, codon and amino acid frequencies. See section 5.2.9 for more information.

5.2.6 Sequence Colors

The colors of nucleotide and amino acid sequences can be set under the **General Options** tab by clicking the drop down menu next to **Colors**.

Coloring schemes differ depending on the type of sequence. For example, the **Polarity** and **Hydrophobicity** color schemes are available only for protein sequences. Alignments and assemblies can be colored by similarity, read direction and paired distance (if paired reads are used), in addition to standard options.

To change the colors that a particular color scheme uses, click the **Edit** button then click each base and select the new color you wish to use.

Similarity Color Scheme

The similarity scheme is used for quickly identifying regions of high similarity in an alignment.

In order for a column to be rendered black (100% similar) all pairs of sites in the column must have a score (according to the specified score matrix) equal to or exceeding the specified threshold.

So for example, if you have a column consisting of only K (Lysine) and R (Arginine) and are using the Blosum62 score matrix with a threshold of 1, then this column will be colored entirely black because the Blosum62 score matrix has a value of 2 for K vs R.

If you raised the threshold to 3, then this column would no longer be considered 100% similar. If the column consisted of 9 K's and 1 R, then continuing with the threshold value of 3, the 9 K's which make up 90% of the column would now be colored the dark-grey (80%→100%) range while the single R would remain uncolored.

If instead the column consisted of 7 K's and 3 R's (still with threshold 3) then 70% of the column is now similar so those 7 K's would be colored the lighter grey (60%→80%) range.

Alternatively, going back to the default threshold value of 1, and with a column consisting of 7 K's, 2 R's and 1 Y, now since the 7 K's and 2 R's have similarity exceeding the threshold whereas the Y is not that similar to K and R, the K's and R's will be colored dark grey since they make up 90% of the column.

Hydrophobicity color scheme

This colors amino acids from red through to blue according to their hydrophobicity value, where red is the most hydrophobic and blue is the most hydrophilic. The values the color scale is based on are given in Figure 5.4. These values are taken from http://biochem.ncsu.edu/faculty/mattos/CrystallographyTutorial/AminoAcids.htm

Polarity color scheme

This colors amino acids according to their polarity as follows:

Yellow: Non-polar (G, A, V, L, I, F, W, M, P)
Green: Polar, uncharged (S, T, C, Y, N, Q)
Red: Polar, acidic (D, E)
Blue: Polar, basic (K, R, H)

5.2.7 Graphs

The **Graphs** tab enables you to display a range of additional metrics on your sequences. The type of graphs which are available depend on the type of sequence you are viewing, and are listed in the sections below. For information on alignment graphs, see section 9.3.2. The number control to the right of each graph controls the height of that graph (in pixels).

Sliding window size. Many types of graph use a sliding window when calculating values. This calculates the value of the graph at each position by averaging across a number of surrounding positions. When the value is 1, no averaging is performed. When the value is 3, the value of the graph is the average of the residue value at that position and the values on either side.

Amino acid		Hydrophobicity	
Phe	F	1.000	**Red**
Leu	L	0.943	
Ile	I	0.943	
Tyr	Y	0.880	
Trp	W	0.878	
Val	V	0.825	
Met	M	0.738	
Pro	P	0.711	
Cys	C	0.680	
Ala	A	0.616	
Gly	G	0.501	**Purple**
Thr	T	0.450	
Ser	S	0.359	
Lys	K	0.283	
Gln	Q	0.251	
Asn	N	0.236	
His	H	0.165	
Glu	E	0.043	
Asp	D	0.028	
Arg	R	0.000	**Blue**

Figure 5.4: Hydrophobicity values for amino acids and corresponding color scale

Nucleotide sequence graphs

GC content. This plots a graph of the GC content of the sequence within a window of specified length as the window is moved along the sequence. If Frame Plot is checked, the graph shows the GC content in the 3rd codon position only for each frame, where frame 1 = red, frame 2 = green and frame 3 = blue.

Chromatogram. This is available with chromatogram traces, and displays the chromatogram trace above the sequence. If **Qual** is checked the quality scores are displayed as a blue graph overlaid on the chromatogram. For more information on viewing chromatograms, see section 5.6.

Stylized DNA Helix. Shows the bases in your sequence as a rotating DNA helix. To turn off the rotation, uncheck the "Animated" option.

Protein sequence graphs

Amino Acid Charge. This runs the EMBOSS `charge` tool to plot a graph of the charges of the amino acids within a window of specified length as the window is moved along the sequence.

Hydrophobicity. This displays the Hydrophobicity of the residue at every position, or the average Hydrophobicity when there are multiple sequences.

pI. pI stands for Isoelectric point and refers to the pH at which a molecule carries no net electrical charge. The pI plot displays the pI of the protein at every position along the sequence, or the average pI when multiple sequences are being viewed.

5.2.8 Advanced sequence view options

Advanced options for controlling the look of sequences and alignments are under the  **Advanced** tab. These options are as follows:

Layout:

- *Wrap sequence*. This wraps the sequences in the viewing area.
- *Linear view on circular sequences*. This forces circular sequences to be shown linearly.
- *Spaces every 10 bases*. If you are zoomed in far enough to be able to see individual residues, then an extra white space can be seen every 10 (or whatever number you choose) residues when this option is selected.

Properties

- *Numbering*. Enables the display of base position number above the sequence residues. For alignments and assemblies, options are available for displaying the numbering of consensus, reference, alignment and/or all original sequences.
- *Mini-map*. Enables the display of a mini-map at the top of the sequence viewer which highlights the currently displayed location in the entire sequence.
- *Outline residues when zoomed out*. This adds a fine line around the sequence which can help with clarity and printing.

Annotations

- *Labels*. This option changes where the labels are displayed on the annotation: "Inside", "Outside", "Inside or Outside" and "None".
- *Overlay when zoomed out*. When only a single annotation covers a region, it will be placed on top of the sequence.
- *Compress annotations*. This option reduces the vertical height of the annotations on display. This reduces the space occupied by annotations by allowing them to overlap and increases the amount of the sequence displayed on the screen.
- *Show arrow tips*. Displays the directional indicator for an annotation as a large arrow tip.
- *Hide excessive labels*. This will reduce screen clutter by removing annotation labels which are too frequent.

Sizes

Here you can set the font size for bases, labels, names and numbering.

5.2.9 Statistics

The % **Statistics** tab displays statistics about the sequence(s) being viewed. If only part of the sequence/alignment or assembly is selected then the statistics displayed will correspond to the highlighted part. The length of the sequence or part of the sequence is displayed next to the Statistics option.

- *Residue frequencies*: This section lists the residues and their frequencies for both DNA and amino acid sequences, for both single sequences and alignments/assemblies. It gives the frequency of each nucleotide or amino acid over the entire length of the sequence, including gaps. If there are gaps, then a second percentage frequency is calculated ignoring gap characters. The G+C content for nucleotide sequences is shown as well for easy reference (see GC content, below).

- *Amino acid and codon frequencies*: These are listed for nucleotide sequences based on the current translation options. Click **Options** to change the translation options. For codon usage statistics, the frequency of all 64 codons (with their associated amino acid) will be displayed. If any CDS contains non-standard start codons then some of the 64 codons may be split into 2 entries based on whether they translate to methionine or their standard translation.

- *Rough Tm*: A rough calculation of the melting point for a nucleotide sequence using the following calculations:

  If the sequence is less than 14bp in length, $RoughTm = 4 \times GCcount + 2 \times ATcount$

  If the sequence is greater than 13bp in length, $RoughTm = 64.9 + 41 \times (GCcount - 16.4) \div length$

- *Molecular Weight*: For **protein** sequences, the following values are used for the amino acids:

  A=71.0788 R=156.1875 N=114.1038 D=115.0886 C=103.1388 E=129.1155 Q=128.1307 G=57.0519 H=137.1411 I=113.1594 L=113.1594 K=128.1741 M=131.1926 F=147.1766 P=97.1167 S=87.0782 T=101.1051 W=186.2132 Y=163.1760 V=99.1326 U=150.0388 O=237.3018

  For **DNA** sequences, the following values are used:

  A=313.21 T=304.2 G=329.21 C=289.18

  The DNA molecular weight assumes no modification of the terminal groups of the sequence.

  If the sequence is a single-stranded, synthesised oligonucleotide (e.g. by primer extension), the value is adjusted for the removed phosphate group by using:

  Molecular Weight = calculated molecular weight - 61.96

  If the sequence is a single-stranded sequence cut by a restriction enzyme, the value is adjusted for the extra 5'-monophosphate left by most restriction enzymes by using:

  Molecular Weight = calculated molecular weight - 61.96 + 79.0

  For dsDNA, these values are adjusted for both strands.

  For **RNA** sequences, the following values are used:

  A=329.21 U=306.2 G=345.21 C=305.18

  The RNA molecular weight assumes no modification of the terminal groups of the sequence. For a 5'-triphosphate group, weights are adjusted using

  Molecular Weight = calculated molecular weight + 159.0

- *Isoelectric Point*: Calculates the isoelectric point of a protein as per this method, but using the following values for the amino acids: D=-3.9 E=-4.1 C=-8.5 Y=-10.1 H=6.5 K=10.8 R=12.5

- *Extinction Coefficient*: Calculates the extinction coefficient of a protein as per this paper, using the following values for the amino acids and assuming all cysteines are paired in a disulfide bridge (making cystine): C=62.5 (only counting up to an even number) W=5500 Y=1490

Statistics for multiple sequences (lists, alignments, assemblies)

- *Sequences*: The number of sequences in the document, or in the currently selected region.
- *Identical sites*. When viewing alignments or assemblies this considers only those columns in the alignment that have at least 2 nucleotides/amino acids/gaps that are not free end gaps and are not columns consisting entirely of gaps. A column not meeting this requirement is not even counted as non-identical for the percentage calculation. A column meeting this requirement is considered identical if it contains no internal gaps and all the nucleotides/amino acids are identical. Ambiguity characters are not interpreted, so a nucleotide column of A and R is not considered identical.
- *Pairwise % Identity*. When viewing alignments or assemblies this gives the average percent identity over the alignment. This is computed by looking at all pairs of bases at the same column and scoring a hit (one) when they are identical, divided by the total number of pairs. Ambiguity characters are interpreted, meaning a nucleotide A vs a nucleotide R is considered to have 50% identity.

  For both *Identical sites* and *Pairwise % Identity*, the statistics are calculated from the subset of sequences and nucleotides/amino acids selected. If just a single sequence is selected, the statistics are calculated as if all sequences are selected over the selected columns. The consensus sequence is always excluded from calculation of both of these values.
- *Coverage of Bases*. When viewing a contig assembly this gives the mean, standard deviation, minimum and maximum of the coverage of each base in the consensus sequence. For small contigs the coverage is further broken down into coverage by reads mapped to the forward and reverse strands. For large contigs, separate forward/reverse coverage can't be efficiently calculated, so is displayed as ?. If your contig has a reference sequence, then the percentage of the ungapped reference sequence that is covered by at least 1 read is also displayed.

  Selecting a sub-region of your contig will display statistics for just that region, including calculation of separate forward/reverse coverage on large contigs.
- *[Ungapped] Lengths of Sequences*. Displays the mean, standard deviation, minimum and maximum of the lengths of the sequences.
- *Confidence (mean)*. When viewing sequences containing quality scores (e.g. chromatograms or NGS reads) this gives the mean of the confidence scores for the currently selected base calls. Confidence scores are provided by the base calling program (not Geneious) and give a measure of quality (higher means a base call is more likely to be correct). An untrimmed value is also displayed if the selected region contains trims.

- *Expected Errors*. When viewing sequences containing quality scores, this gives the approximate number of errors that are statistically expected in the currently selected region. This is calculated by converting the confidence score for each base call to the error probability using the formula $10^{(-Q/10)}$. For example, a base with a quality score of 30 will have an error probability of 0.001. The expected errors value is then calculated by summing up the error rates for each base. This also has a value for the untrimmed selection if the region contains trims.

GC content

For documents that are created or modified in Geneious 8.1 or later, the GC content can also be viewed in the **%GC** column in the document table.

The %GC column shows the percentage of A, C, G, T, U, S, W nucleotides that are either G, C, or S. Ambiguous bases that contain a mixture of GC and non-GC bases (e.g. R, Y, M, K) are excluded from the calculation. This field is available on all nucleotide sequences, contigs, alignments, and sequence lists that were created or had their sequences last modified in Geneious 8.1 or later. For contigs and alignments, the consensus sequence and reference sequence (if any) are excluded from the calculation.

For sequences within an alignment, contig or list, the %GC column only shows the overall value for the alignment. To see a table of GC percentages for all individual sequences within an alignment or contig, the sequences need to be extracted to stand-alone sequences. Alternatively, individual values can be viewed in the statistics panel by clicking on the name of the sequence to select it.

Sequences in a list or alignment can be sorted by GC content by right clicking in the sequence viewer and choosing **Sort** → **%GC**.

5.3 Editing sequences

To edit sequence(s) or an alignment click the **Allow Editing** toolbar button.

You can manually enter or delete sequence, or use any of the standard editing operations, such as **Copy** (Ctrl/command+C), **Cut** (Ctrl/command+X), **Paste** (Ctrl/command+V), **Paste Without Annotations** (Shift+Ctrl/command+V), **Paste Reverse Complement** and **Undo** (Ctrl/command+Z). All operations are under the main **Edit** menu, or can be accessed by right-clicking in the sequence view and selecting the option from the popup window.

To **insert** sequence, click at the position where you want to insert the sequence and type or paste it in. Normally existing residues will be shifted to the right when you insert sequence; to get them to shift to the left hold down the Shift key as you insert the sequence. To **overwrite** sequence, select the sequence you wish to overwrite and type or paste the new sequence in.

To **select sequences or regions of sequence**, there are several options:

- To quickly select a single residue, double-click on it.
- To select a block of residues within a single sequence, triple click.
- To select a block of residues across multiple sequences, quadruple click.
- To select a block of 10 residues, hold down Shift and Cntrl (alt/option on a Mac) and press the keyboard arrow.
- To select a specific region of sequence, click at the beginning of the region you want to select, hold down Shift and then click the end.
- To select the same region across multiple sequences in an alignment, select the region you want in the first sequence, then hold down Shift / alt (command on a mac) and click the down arrow to apply the selection to the sequences underneath. Holding down Shift /Cntrl (alt/option on a mac) while pressing the down arrow will select the sequences in batches of 10.

Sequences can be **reordered** within an alignment by clicking the sequence name and dragging.

Sequences can be **removed** from an alignment by right-clicking (Ctrl+click on Mac OS X) on the sequence name and choosing the **remove sequence** option. Alternatively, select the entire sequence (by clicking on the sequence name) and press the delete key.

To **delete a region** of a sequence or alignment, select the region and press the delete or backspace key. Normally this will move residues on the right into the deleted area. To move the residues on the left into the deleted area, hold down the Shift key while deleting.

To **drag a sequence** to the left or right, select the region you want to move, then click it again and drag it to the position you want. Dragging will either move residues over existing gaps or open new gaps when necessary. Dragging a selection consisting entirely of gaps moves the gaps to the new location.

After editing is complete, click **Save** to permanently save the new contents.

5.3.1 Concatenating sequences

To join several sequences end-on-end, select all the sequences and go to **Tools** → **Concatenate Sequences or Alignments**. This creates a single sequence document from the input sequences. The order in which sequences are concatenated can be chosen in the setup dialog box, and the resulting sequence can be circularized if required by checking **Circularize sequences**. If one or more of the component sequences was an extraction from over the origin of a circular sequence, you can choose to use the numbering from that sequence, thus producing a circular sequence with its origin in the same place as the original circular sequence. Overhangs will be taken into account when concatenating.

5.4 Complement and Reverse Complement

To display the complement of a sequence alongside the original sequence, check the **Complement** box in the **Display** tab. To display only the complement, and not the original nucleotide sequence, uncheck **Nucleotides**. Note that these options are only for display purposes - if you wish to create a separate document to work with the complement (not reverse complement) of a sequence, you will need to install the **Complement or Reverse** plugin by going to **Tools** → **Plugins**. Once the plugin is installed, go to **Sequence** → **Complement only** to create a new document containing the complement.

To **reverse complement** a nucleotide sequence (i.e. reverse the sequence direction and replace each base by its complement), click the **R.C** button above the sequence viewer, or go to **Reverse Complement** under the **Sequence** menu. You can also access this option right-clicking in the sequence viewer and selecting it from the popup menu. When you click **Save** after reverse complementing, the tag *(reversed)* will be added to the sequence name.

When only part of a sequence is selected, you can choose to either reverse-complement only the selected region and extract it to a new sequence document, or reverse complement the entire sequence. On alignment or contig documents you can reverse complement individual sequences within the alignment or assembly by selecting that sequence, and choosing **reverse complement selected sequence**.

5.5 Translating sequences

The protein translation can be viewed alongside the nucleotide sequence by checking the **Translation** option in the **Display** tab. Select the genetic code and reading frame(s) you require. You can also choose to translate relative to selection or annotations such as CDS (Figure 5.5). In an alignment, the sequence frame can be calculated relative to the individual sequences, the alignment, the consensus or a specific reference sequence. On a contig or alignment, the translation can be displayed on the consensus and reference sequence only, or it can be displayed on all sequences.

If you wish to view only the translation and turn off the nucleotide sequence, uncheck **Nucleotides**. However, this is only for display purposes: if you wish to work with the translation in downstream analysis you must extract it to a separate document using the **Translate** button above the sequence viewer. The **Translate** button will create new protein document from the translated DNA, using your choice of reading frame and genetic code. This option can also be accessed from the **Sequence** menu.

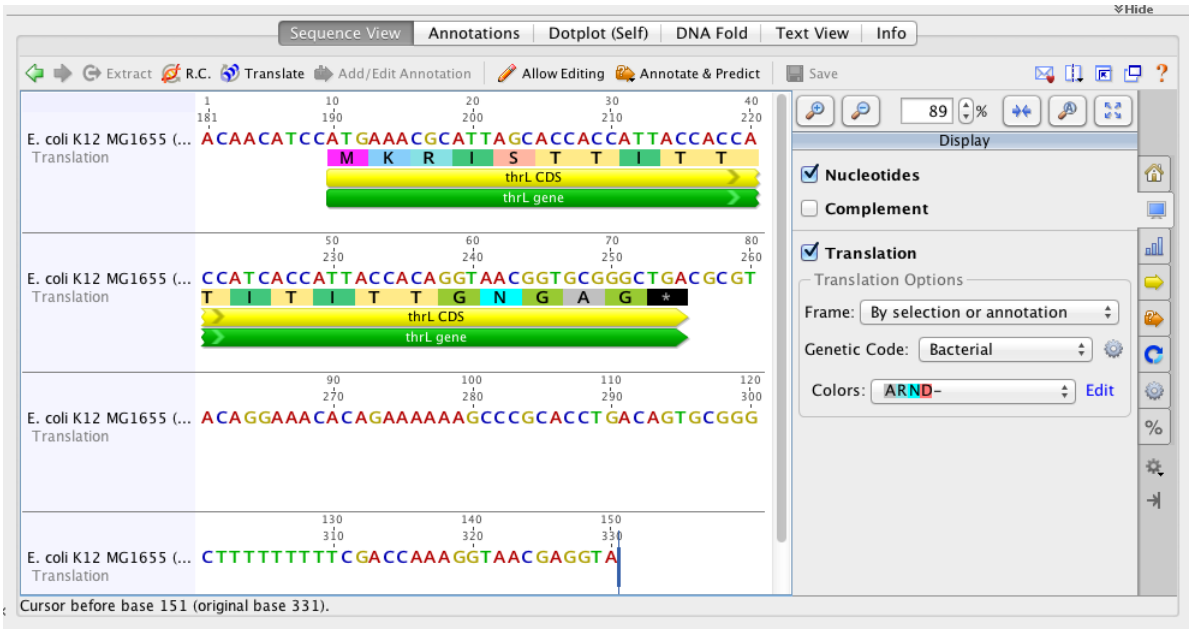


Figure 5.5: Translating a CDS

5.5.1 Genetic Codes

Geneious supports a range of genetic codes which can be chosen from the drop-down menu in the Translation options. To set a default genetic code which will apply to all documents in your database, click the Settings cog next to the drop-down menu. You can also enter a custom genetic code here by clicking Add and editing an existing genetic code template.

5.5.2 Back Translating

To create a nucleotide sequence from a protein document, go to **Sequence** → **Back Translate**. Ambiguous back-translation uses a specific genetic code to produce a nucleotide sequence with ambiguous bases, so that every possible codon is represented for each amino acid. Unambiguous back-translation uses codon usage tables for selected organisms to produce a nucleotide sequence where the most frequently used codon for that organism is used for each amino acid.

5.6 Viewing chromatograms

Geneious can view chromatogram information from files imported in .ab1 or .scf format. If the chromatograms are not visible, check **Chromatograms** under the **Graphs** tab (see Figure 5.6).

Chromatogram files are produced from sequencing machines such as the Applied Biosystems 3730 DNA analyzer. The raw output of a sequencing machines is known as a **trace**, a graph showing the concentration of each nucleotide against sequence positions. The raw trace is processed by a "Base Calling" software which detects peaks in the four traces and assigns the most probable base at more or less even intervals. Base calling may also assign a quality measure for each such call, typically in terms of the expected probability of making an erroneous call. Geneious does not perform base-calling itself: this information is already contained in the .ab1 or .scf file.

Chromatogram peaks for individual bases can be turned off by checking the A/G/C/T boxes in the Graphs tab. Note that since the distance between bases as inferred from the trace varies the trace may be either contracted or expanded compared with the raw data. The vertical scale of the chromatogram can be adjusted by clicking and dragging on the graph itself. The total height of the graph can be adjusted by increasing the number displayed next to the graph on the right of the Sequence View.

Quality. The quality scores associated with a chromatogram can be viewed by checking the Qual box under the Chromatogram graph options. This displays a quality measure (typically Phred quality scores) for each base as assessed by the base calling program. The quality is shown as a shaded blue bar graph overlaid on top of the chromatogram. Note that those scores represent an estimate of error probability and are on a logarithmic scale - the highest bar represents a one in a million (10^{-6}) probability of calling error while the middle represents a probability of only a one in a thousand (10^{-3}).

Sequence Logo. When checked, bases letters are drawn in size proportional to call quality, where larger implies better quality or smaller chance of error. Note that the scale is logarithmic: the largest base represents a one in a million (10^{-6}) or smaller probability of calling error while half of that represents a probability of only a one in a thousand (10^{-3}).

On large contigs (over 100,000 bp long), the sequence logo can't be efficiently calculated in regions of over 1000 fold coverage, in which case the sequence logo will display ?.

To view the raw chromatogram traces, click the **Chromatograms** tab above the sequence viewer. In this view, the exact location of the base call can be viewed by checking **Mark calls**. To view sequence logos indicating base quality in this view, check **Scale by confidence**. The **Trace** options for X and Y scales allow you to zoom in on the X or Y axes, respectively.

Figure 5.6: A sequence alignment containing chromatograms, with quality scores enabled

5.6.1 Binning by quality

Chromatograms can be binned on the basis of their quality scores into Low, Medium or High quality bins. The parameters for each of these bins are set under **Tools** → **Preferences** → **Sequencing** (see section 1.2.1 for more information). To see the bin for a trace, add the **Bin** column to the document table by going to **View** → **Table Columns**. You can also view the percentage of bases that are low, medium or high quality by adding **LQ%**, **MQ%**, and **HQ%** columns to the document table.

5.7 Meta-data

Meta-data is additional information you can add to any of your local documents, for example sample collection history, organism identification, primers used etc. Meta-data is added in the **Properties** view under the **Info** tab. This tab displays standard properties of documents such as the name and description, plus any meta-data you have added. Any meta-data that you add can appear in a column in the document table, and can be treated as a user-defined field for use in sorting, searching and filtering your documents.

When multiple documents are selected, the Properties view displays all of the fields and meta-data belonging to the selected documents. When all documents have the same value for a field, it is displayed in the viewer. If the documents have different values, or some of the selected documents do not have a value, then the field will show that it represents multiple values. Changes made to the fields will apply to all selected documents.

You can add meta-data to any of your local documents, including molecular sequences, phylogenetic trees and journal articles. You cannot add meta-data to search results from NCBI or EMBL etc until the documents are copied into one of your local folders.

5.7.1 Adding meta-data

To add meta-data to your document, select the **Add Meta-Data** button on the toolbar and then choose from the available types. Selecting a meta-data type will create an empty instance of that type. To fill meta-data values just start typing into the fields. See section 5.7.3 for information on how to create a new meta-data type.

5.7.2 Editing Meta-Data

To edit existing meta-data fields, simply click on the field and enter your data. Some fields may have constraints (which you can edit in the Edit Meta-Data Types dialog, *(see 5.7.4)*. If the data you have entered does not conform to the constraints of the field, it will be displayed in red

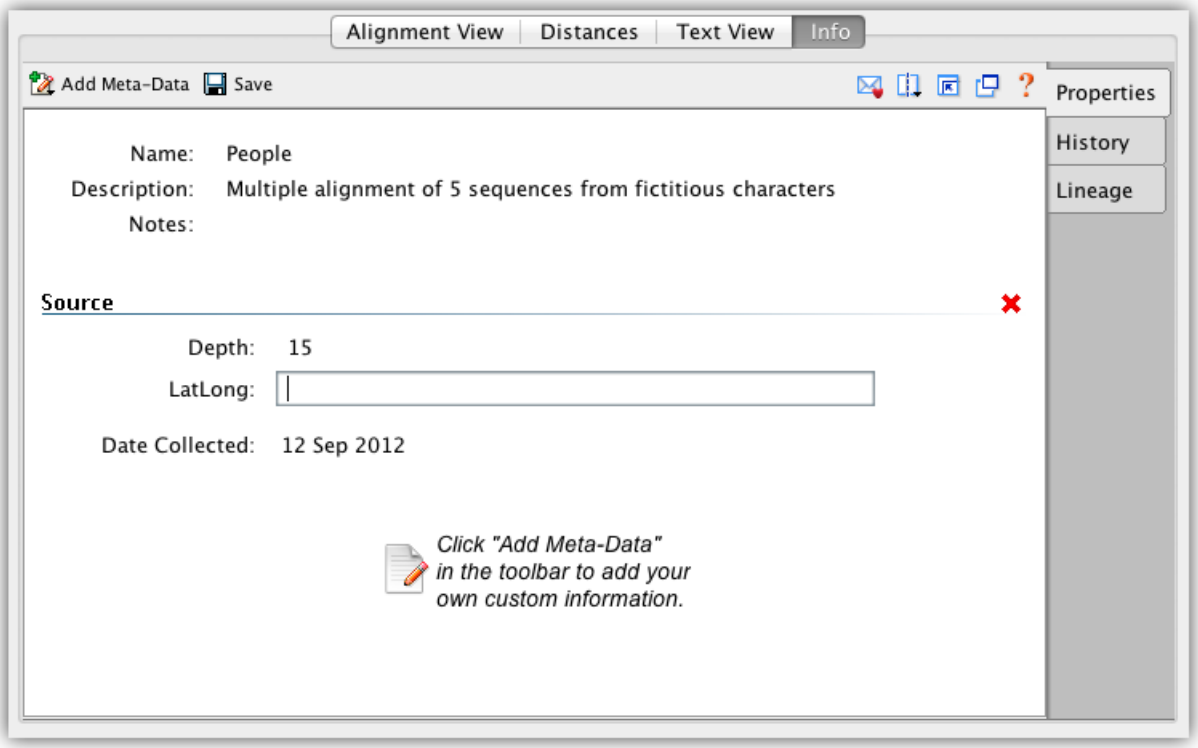


Figure 5.7: The Properties View

and you will be shown the field's constraints in a tooltip.

***Tip: To enter a new line in a text field, press* Shift+enter *or* Ctrl+enter**

5.7.3 Creating a new Meta-Data type

Geneious does not restrict you to the meta-data types that it comes with. You can create your own types to store any information you want.

To create a new type, click **Edit meta-data types**, then click the **Create** button at the bottom left of the panel. This creates a new type, with one empty field, and displays it in the panel to the right.

Note. The **Name** and **Description** fields distinguish your meta-data type from other user-defined types. They do not have any constraints.

Next, you need to decide what values your Meta-Data Type will store by specifying its fields:

Field name. This defines what the field will be called. It will be displayed alongside columns such as Description and Creation Date in the Documents Table. You can have more than one Field in a single Meta-Data Type - to add or remove a field from the type, click the + or - buttons to the right of the field.

Field type. This describes the kind of information that the column contains such as Text, Integer, and True/False. The full list of choices in Geneious is shown in figure 5.9.

Constraints. These are limiting factors on the data and are specific to each field type. For example, numbers have numerical constraints – is greater than, is less than, is greater or equal to, and is less or equal to. These can be changed to suit. The constraints for each field can be viewed by clicking the "View Constraints" button next to the field. This will show a pop-up menu with the constraints you have chosen. *(see figure 5.8)*

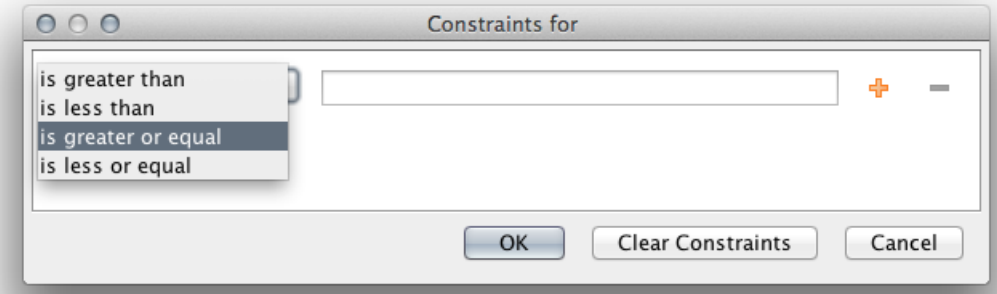


Figure 5.8: The Edit Constraints window

5.7.4 Editing Meta-Data Types

To edit meta-data types, e.g. by adding and removing fields, click **Edit Meta-Data Types**. Select the meta-data type you want to edit, and then add, remove or edit the fields as described in section 5.7.3.

5.7.5 Using Meta-Data

The main purpose of meta-data is to add user defined information to Geneious documents. However, meta-data can be searched for and filtered as well. Also, documents can be sorted according to meta-data values.

Searching - Once meta-data is added to a document, it is automatically added to the standard

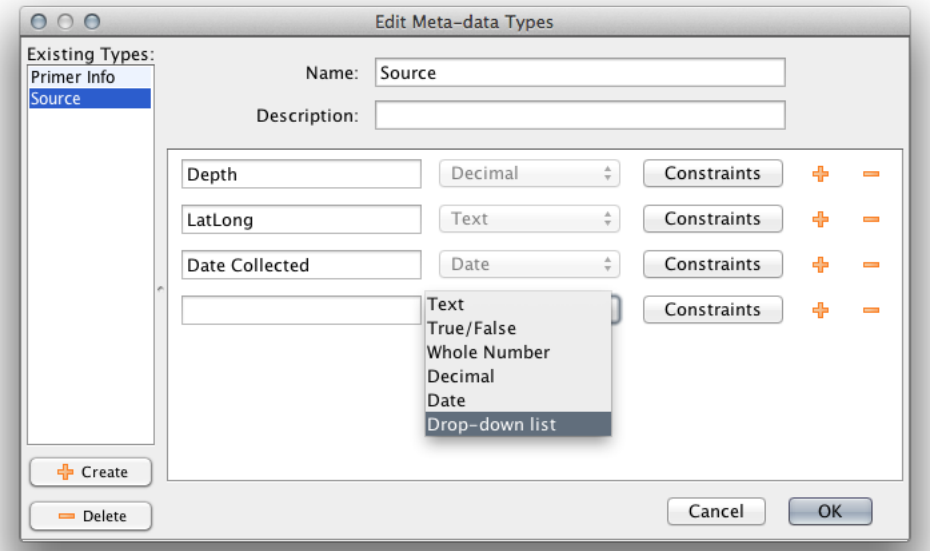


Figure 5.9: The Edit Meta-Data Types window

search fields. These are listed under the **Advanced Search** options in the Document Table (see section 4.2.1). From then on, you can use them to search your Local Documents. If you have more than one Field in a meta-data type, they will all appear as searchable fields in the search criteria.

Filtering - Meta-data values can be used to filter the documents being viewed. To do so, type a value into the **Filter Box** in the right hand side of the Toolbar (see section 4.2.4). Only matching documents will be shown.

Sorting - Any meta-data fields added to documents will also appear as columns in the Document Table. These new columns can be used to order the table.

Chapter 6

Parent / Descendant tracking

Many documents in Geneious are the output of an operation run on a set of input documents. The input documents of the operation are known as the **parents** of the output, and the output documents the **descendants** (or **children**) of the input. Those parent documents may themselves be the descendants of other documents, each with their own parents, and so on. In many situations it is useful to preserve this hierarchy, so that future alterations, for example the re-calling of a base, or the addition of a new annotation, can be transferred downstream to the molecules affected by this change in a parent.

An **active link** between a child and its parents means that when you modify any of the parent documents, you will be given the choice of propagating these changes to the child. When this modification affects a part of the parent involved in creating the child, the change will be immediately visible in the child. Modifications include things like editing the residues of a sequence, adding new annotations, or changing the meta-data associated with the document.

Propagating a change to a parent document causes Geneious to rerun every operation that links that parent actively to one or more child documents, with the altered parent document (and any other parents) as input. Geneious stores the options that the operation was originally run with so that it can reproduce the original operation's conditions exactly, and afterwards matches up the newly regenerated child documents with any former children, and replace their contents where possible.

Occasionally, one or more of the parent documents has been altered to a point where an operation can no longer be rerun, or a necessary parent document is inaccessible. In this case, Geneious will inform you of the failure, and attempt to be as specific as possible about the cause of the failure (Figure 6.1)

Inactive links do not propagate changes from parent to child. Inactive links are created in two different ways; firstly, when you choose not to propagate changes, that active link becomes *temporarily* inactive. Secondly, if an operation does not support creation of active links, or was told not to create them, all links between its parents and children will be *permanently* inactive.

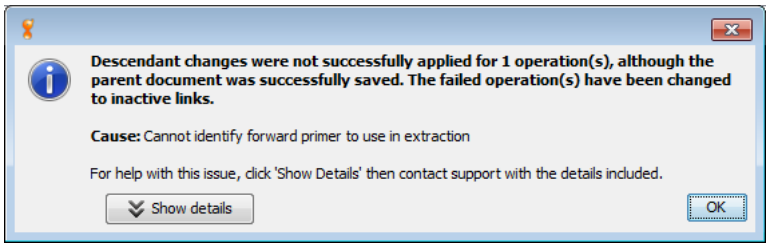


Figure 6.1: Failure to propagate an Extract PCR product operation due to a missing forward primer

All operations in Geneious at least create inactive links.

The following operations in Geneious can produce actively linked documents:

- Cloning: Digest into Fragments...
- Cloning: Insert into Vector...
- Cloning: Ligate sequences...
- Cloning: Gateway
- Primers: Extract PCR product
- Sequence Viewer: Extract
- Sequence Viewer: Translate

Note: Extract and Translate will not create active links by default. To do so, you must select the **Actively link source and extracted documents** checkbox in the relevant dialog (see Figure 6.2), otherwise they will be created with permanently inactive links.

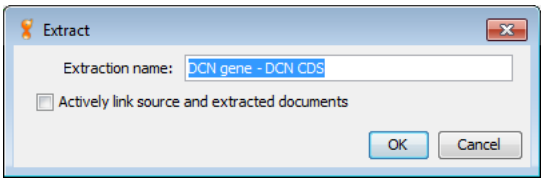


Figure 6.2: Extract dialog with active link checkbox

6.1 Editing Linked Documents

When you make changes to a document that is the parent of another document, you will be given the opportunity to either propagate the changes, deactivate the link (which can later be reactivated, see Lineage View, Section 6.2), or save the changed document as a new copy (Figure 6.3). You may also simply back out of this process by choosing to cancel, which will return you to your unsaved changes. Note that if you choose to deactivate the link, this dialog will not be displayed upon subsequent saves of the parent document, unless the link is reactivated again at some future time.

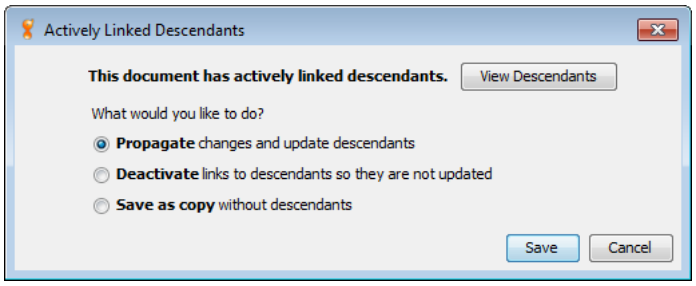


Figure 6.3: Actively Linked Descendants dialog

In order to aid with your decision making, the dialog allows you to view the document's descendants in a smaller, cut down version of the Lineage View. Pressing the **View Descendants** button will bring up this view (Figure 6.4).

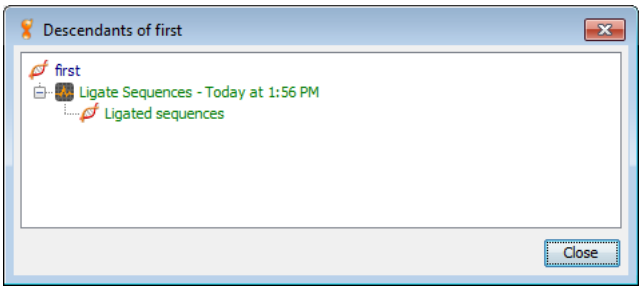


Figure 6.4: Descendants view

When you choose to begin editing a document with actively linked **parents** in the Sequence View, you will immediately be warned that in order to save your changes you will need to deactivate this link. As with the Actively Linked Descendants view, you will be given the opportunity to view the document's lineage. Editing a document that is a descendant of other documents is usually unintentional; however, in some circumstances you may simply be interested in the output documents of an operation (not the parent-descendant relationship), and as such you may hide this dialog (Figure 6.5).

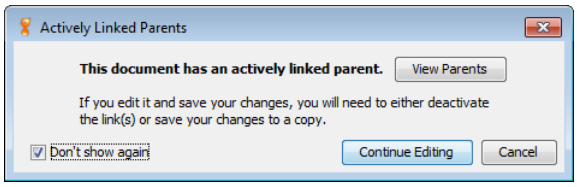


Figure 6.5: Actively Linked Parents dialog

Upon conclusion of your editing, you will again be prompted to either deactivate links or save a copy.

6.2 The Lineage View

Every document that is linked (actively or otherwise) to another document has a tab called **Lineage** in the Info View tab. The lineage view allows you view parent-descendant relationships, manage links, and navigate between documents (Figure 6.6).

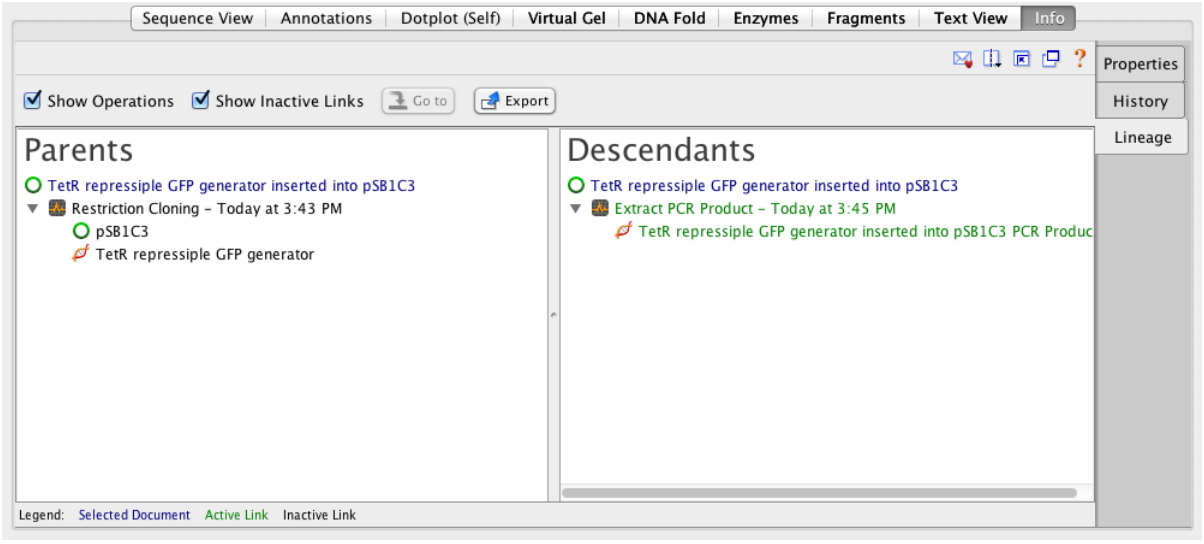


Figure 6.6: The Lineage View

All active links appear as green text, whilst inactive links appear as black text and the document currently being viewed (and which is the root of the parents tree and the descendents tree) appears in blue. Each's document's name is displayed along with an icon (similar to the document table) denoting what type of sequence it is.

Also displayed in the viewer are the operations that generated each set of children, along with

the time at which the operation was run and the type of operation. If preferred, these operations can be hidden by unchecking the **Show Operations** checkbox, providing a layout which is akin to Vector NTI®. You can also choose to view only inactive links by unchecking the **Show Inactive Links** checkbox. **This will hide all inactively linked documents, as well as *those* documents' parents or descendants**. This means that you will only be viewing documents that are directly affected by one currently being viewed.

You can reactivate temporarily deactivated links from the view by right-clicking (Windows, Linux) or control-clicking (MacOS) on a document and choosing **Activate link to parent** from the context menu. Alternatively you can reactivate links to all children at once by choosing **Show Operations** and right- or control-clicking, then selecting **Reactivate all links for this operation**. You may also manually deactivate links in this fashion (Figure 6.7).

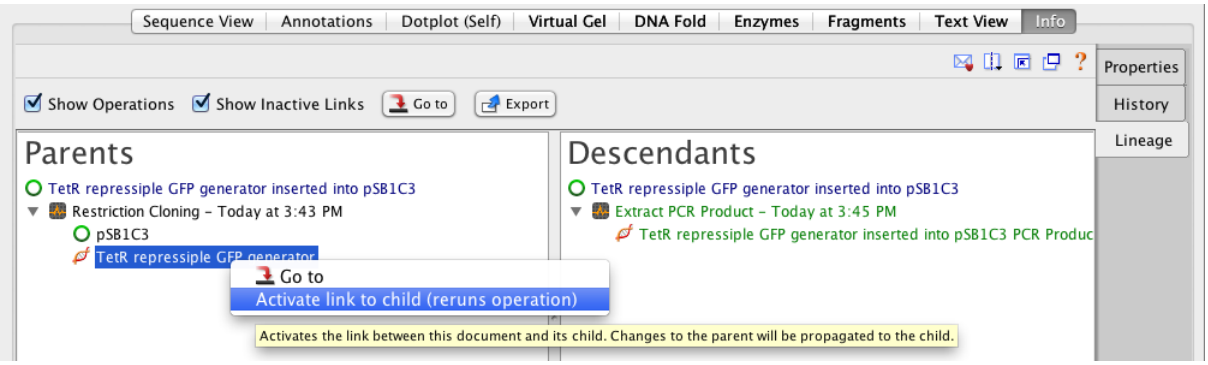


Figure 6.7: Context Menu

Note: reactivating links immediately reruns the operation; depending on the size and type of the operation, this can be time consuming. Also note that reactivating will cause any unsaved changes to any direct or indirect descendants to be overwritten, since this involves a complete recompute from the parent documents. You will be warned about this before Geneious allows you to reactivate.

Finally, you may export the currently selected document (highlighted in blue in the view) directly, via the **Export** button. Doing so will bring up a dialog (Figure 6.8). From here you can choose to export parents or descendants only, or both, as well as choose to export only those documents that are actively linked in the hierarchy. Similarly to how unchecking the "Show Inactive Links" checkbox works, unchecking "Inactively linked documents" here will mean that the export will stop as soon as it finds in inactively linked parent or descendant (depending on the relevant direction), and stop exporting down that branch of the lineage.

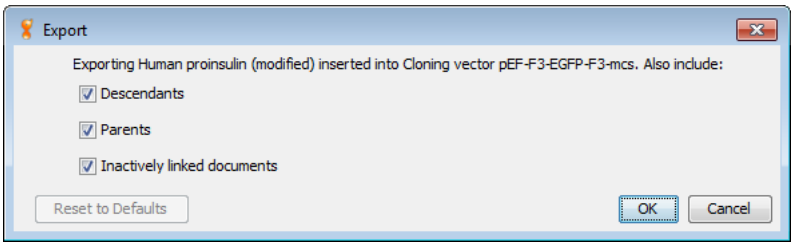


Figure 6.8: Export Dialog

Chapter 7

RNA, DNA and Protein structure viewer

7.1 RNA/DNA secondary structure fold viewer

This viewer will appear when the selected nucleotide sequence is less than 3000bp long. If the sequence is DNA, the tab will be labelled **DNA Fold** and if it is RNA it will be labelled **RNA Fold** (Figure 7.1)

The fold prediction is performed by the Vienna package `RNAfold` tool. Information on the options for this tool can be found at the following web page: http://www.tbi.univie.ac.at/~ivo/RNA/RNAfold.html.

The **View Options** allow you to turn off/on and color the bases, flip the coordinates, highlight the start (blue) and end (red) of the sequence and rotate the model. As with other viewers, you can zoom in on the model and drag the view around, or use the scrollwheel using the same keyboard modifiers as the sequence viewer. Selection is synchronized between the sequence view and the fold view. In addition, when in split view mode, the fold viewer will scroll to the selected area when zoomed in.

By default, **Color by probability** is used where red bases are the ones with the strongest probability of the bases being paired with each other in paired regions, or being unpaired in unpaired regions. Green is the middle ground and blue is the lowest probability. Color by probability is only available when using the Partition Function.

Compute Options will rerun `RNAfold` when you change their settings, so depending on the size of the sequence there may be a noticeable recompute time.

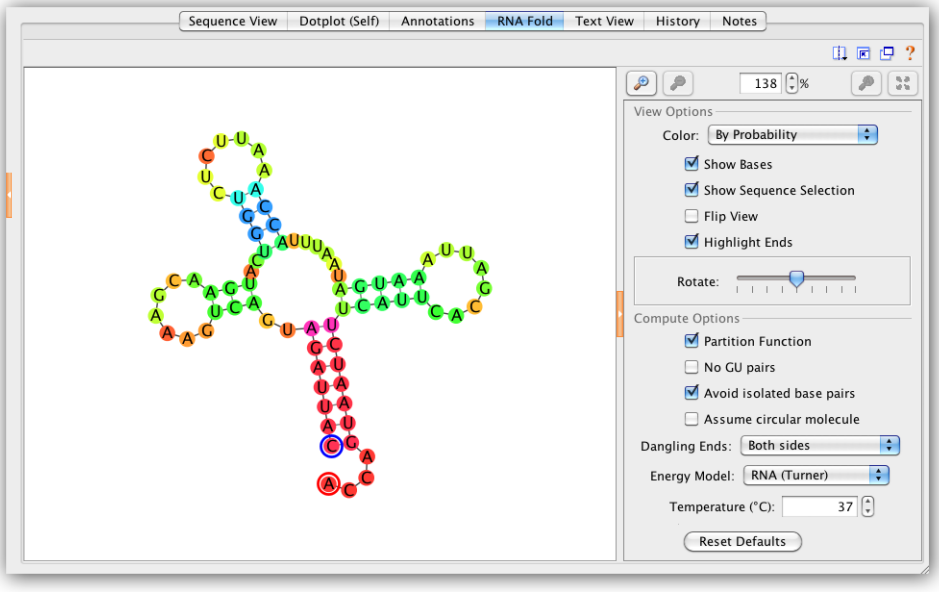


Figure 7.1: A view of an mRNA secondary structure prediction in Geneious

7.2 3D protein structure viewer

For molecular structure documents, such as PDB documents, this displays an interactive three dimensional view of the structure.

7.2.1 Structure View Manipulation

- Click and drag the mouse to rotate the structure.
- Hold the Alt or Shift key then click and drag to zoom in/out
- Hold the Ctrl key then right-click and drag to pan, or, if you are using a Mac, click and hold, press Ctrl and Alt/Option then drag to pan.

7.2.2 Selection Controls

To the right of the structure are controls that let you control the selected part of the structure.

- If the structure you are viewing contains more than one model, the **model** combo box will you choose between them.

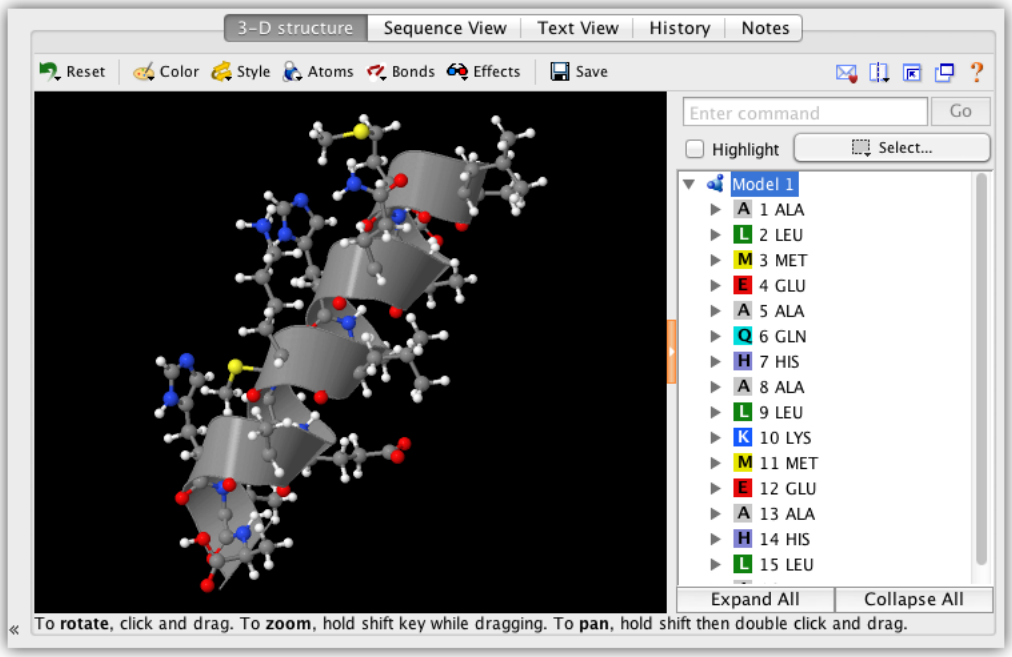


Figure 7.2: A view of a 3D protein structure in Geneious

- The **select** button lets you select all, none or the nonselected region of the structure, as well as by element, group type or secondary structure.
- The **highlight selected** checkbox lets you select whether to highlight the selected atoms in the structure view.
- The **structure tree** shows the atoms in the structure in a tree format. Click on regions in the tree to select those regions. You can also Shift-click and Ctrl-click to select multiple regions at once.
- The **command box** lets you type in arbitrary jmol scripting commands. To see some examples, select one of the pre-populated options in the box's drop-down. For a complete description of the commands you can use, see http://www.stolaf.edu/academics/chemapps/jmol/docs.

7.2.3 Display Menu

At the top of the viewer is the display menu. Here you can modify the appearance of the structure.

- **Reset** lets you reset the position of the structure, reset the appearance of the structure to the default, or reset the appearance of the structure to its appearance when it was last saved.
- **Color** lets you change the color scheme of the selected region of the atom.
- **Style** lets you change the style of the selected region of the molecule eg to spacefill or cartoon view.
- **Atoms** lets you hide atoms or change their size in the selected region of the molecule. You can also choose whether to show hydrogen atoms and atom symbols.
- **Bonds** lets you hide bonds or change their size in the selected region of the molecule. Covalent/ionic bonds, hydrogen bonds and disulfide bonds can be affected separately.
- **Effects** lets you toggle spin, antialiasing, stereo and slabbing effects for the whole molecule.
- **Save** saves the current appearance of the molecule.

Chapter 8

Working with Annotations

8.1 Viewing, editing and extracting annotations

Annotations are used to describe and visualize features, such as coding regions, restriction sites and repetitive elements, on sequences and alignments. Annotations can either be annotated directly on a sequence in the sequence viewer, or they can be grouped logically into **tracks**. A track is a collection of one or more annotation types. Tracks are stacked vertically underneath the sequence in question, with a separate line for each track and its annotations.

An annotation may have one or more properties or qualifiers associated with them. These can be added at the time an annotation is created, or at a later date by editing the annotation. To view the properties of a given annotation, mouse over it in the sequence viewer. This will display a tooltip, listing the Name, Type, Length, Interval and Sequence for that annotation plus any additional qualifiers (see Figure 8.1)

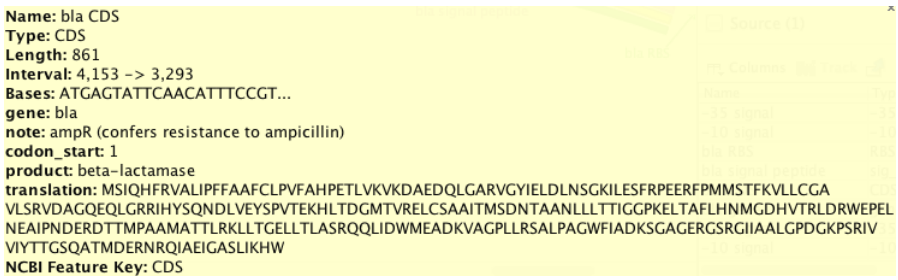


Figure 8.1: Annotation properties and qualifiers, displayed by mousing over an annotation

8.1.1 Viewing and Customizing Annotations: The Annotations and Tracks tab

If a sequence contains annotations, the **Annotations and Tracks** tab to the right of the sequence viewer will show a yellow arrow, and all the annotation types present on the sequence will be listed in this panel (see figure 8.2). Annotations that are directly on the sequence are listed first, followed by annotations on tracks. Tracks with only one annotation type will show a single listing, whilst tracks with multiple annotations will show a list of the annotation types.

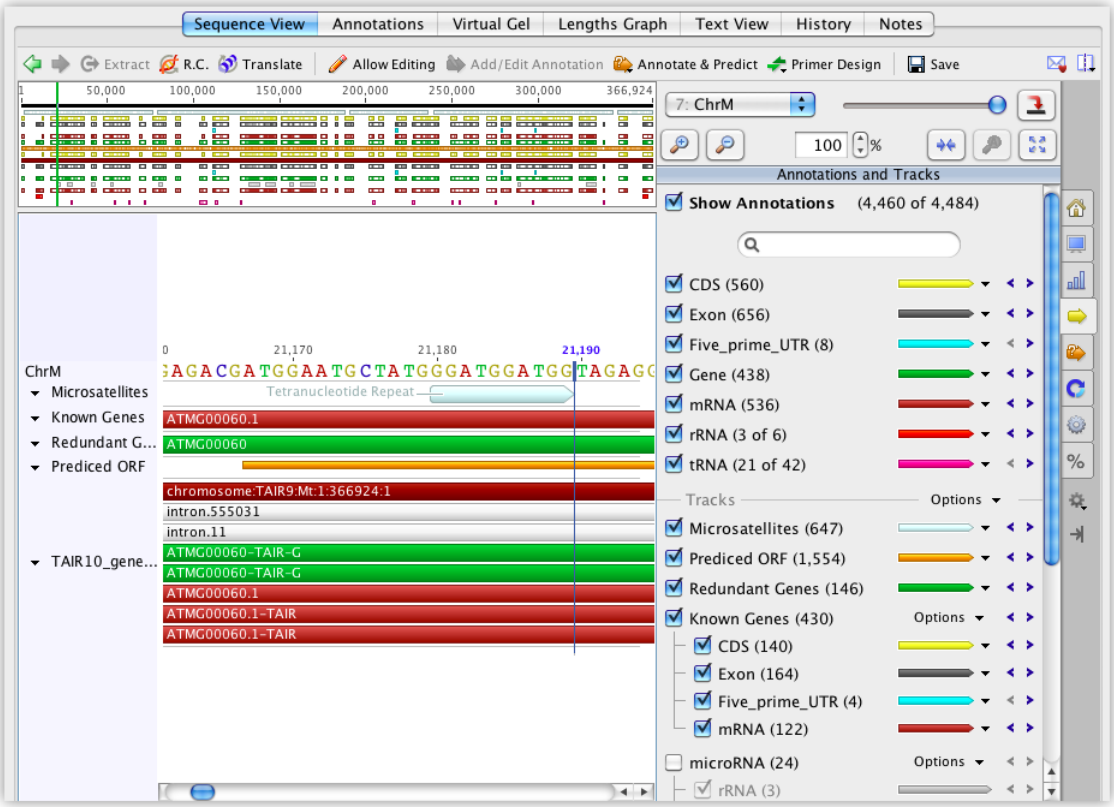


Figure 8.2: The annotations options in the sequence viewer

Individual annotation types can be turned on or off using the checkboxes to the left of the annotation type, or all annotations can be turned off by unchecking **Show Annotations** at the top of the panel. Note that turning an annotation type off does not remove it from the sequence, it only hides it from view.

Directly beneath the Show Annotations box is a filter text field. Typing a term in this field will highlight any annotations that contain the entered text in their name or qualifiers. To filter for a term in a specific field, click the down arrow to the right of the box and choose the field and term you wish to search for.

To quickly find and move between instances of a particular annotation type on a sequence, click the small left/right buttons  to the right of each annotation type. This will move the selection in the sequence view to the next or previous instance of that annotation type. This is useful for navigating large genomes or assemblies.

To **customize** the way an annotation type is displayed, click on the preview of the annotation arrow in the Annotation and Tracks window. This will bring up a popup menu containing the following options:

- *Show only [type] annotations:* Turns off all other annotation types and shows only this type.
- *Show [type] labels:* When this is unchecked, only the annotation arrow is shown on the sequence and not the annotation name.
- *[type] labels >:* Allows you to customize which property of the annotation is displayed as the name.
- *Show above sequence:* Moves the annotation so that it is sitting above the sequence, rather than on or below it.
- *Edit Color:* Allows you to change the color of the annotation arrow.
- *Edit all [type] annotations:* Allows you to change the name, type or properties of a given annotation type, or move the annotation to or from a track. This applies the change to all annotations of this type.
- *Delete all [type] annotations:* Deletes all annotations of a particular type.

The popup menu for individual tracks has an additional option **Color by / Heatmap**. This will color annotations on that track according to the contents of a qualifier field, enabling the creation of annotation heatmaps by using a score value (or some other metric) stored in the qualifier of an annotation.

The way annotations are drawn on the sequence can be further customized in the **Advanced** tab of the sequence viewer (see section 5.2.8).

8.1.2 The Annotations Table

The **Annotations** tab appears above the sequence viewer whenever sequences containing annotations are selected. It displays each annotation as a row in a table, with columns corresponding to the qualifiers for the annotations. Selection of annotations is synchronised with other viewers, such as the sequence viewer and dotplot.

To change what is displayed in the annotations table, use the buttons above the table:

- *Types* allows you to specify what annotation types are displayed in the table.

  The **Select One** button in the menu is a quick way to view just one type while also selecting the relevant columns for that type. Relevant columns are deemed to be ones where at least one annotation of that type has a value for the column.

- *Tracks* allows you to specify what tracks are displayed in the table.

- *Columns* allows control over which columns are visible in the table.

To further **filter** what is visible in the table, use the filter box at the top right of the table. Filtering is only done against the currently visible columns for each annotation.

To export the visible rows and columns of your annotations table, click **Export table**. This exports the table to a CSV (comma-separated values) file.

The **Extract** and **Translate** buttons will create a new document from the selected annotation(s). **Extract** extracts the region of the selected annotation to a new document. **Translate** translates the nucleotides in the region of the selected annotation into amino acids, using your choice of reading frame and genetic code, and saves it to a new protein document.

Annotations table functions can also be accessed via a popup menu when right-clicking on one or more selected annotations in the table. This menu contains options for copying the selected value, extracting, translating, showing on sequence, editing and deleting the selected annotations. The show on sequence function in this list will show the selected annotations in the sequence viewer.

8.1.3 Editing annotations

Annotations can be edited by selecting them either on the sequence or from the annotations table, and clicking **Edit Annotations**. This brings up a window where the annotation name, type, location, properties and intervals can be edited (see Figure 8.3).

To move the annotation onto a track, click the **Track** option and either choose an existing track, or type in the name of a new track you want to create. To move the annotation from a track to the sequence itself, choose **No track** in this setting.

In the **Properties** section, properties can be added, edited, removed or moved up and down in the list by clicking the buttons to the right. The annotation color can also be set in this section by clicking the color boxes. The **Override color** sets the color for that particular annotation only but does not change the color of other annotations of that type. To change the color of all annotations of that type, click the **Color** box.

The **Intervals** section shows where the annotation is located on the sequence. To change the location, the direction, or mark an annotation as partial, select the interval and click Edit. The direction of the annotation can be changed by clicking the colored arrow. Check **truncated left**

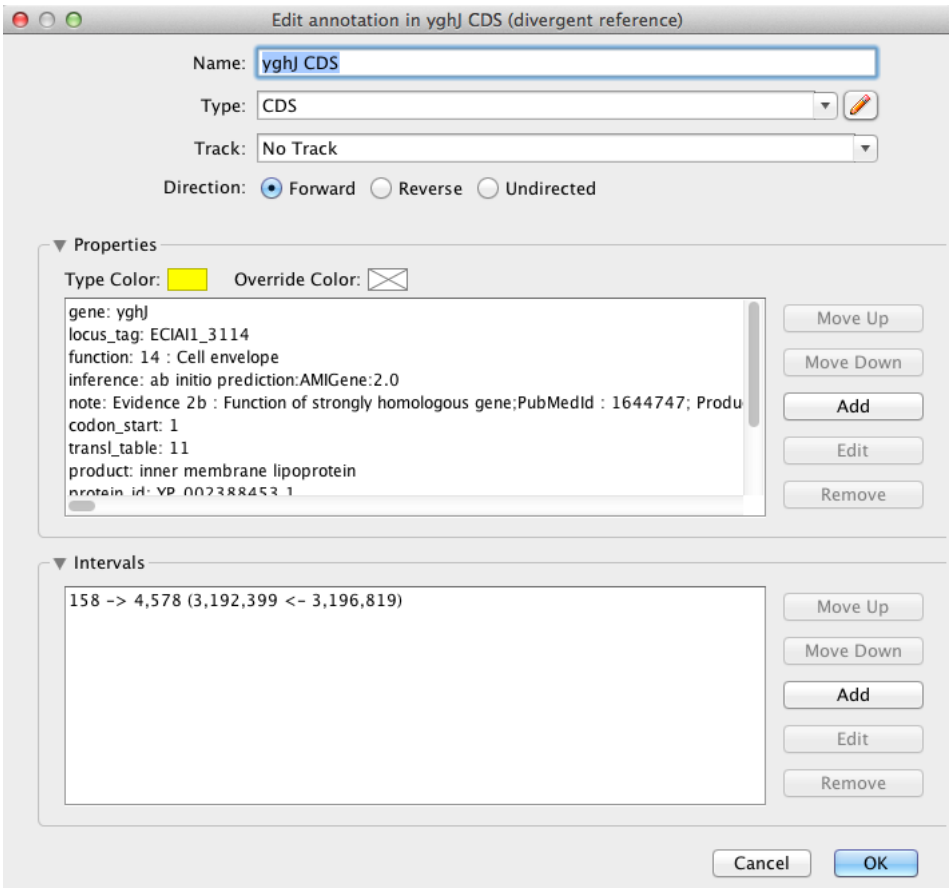


Figure 8.3: The Edit Annotations Window

end to mark an annotation as partial at the left end (e.g. if a gene is incomplete at the left end when viewed on the sequence), or **truncated right end** to denote that the feature is incomplete at the right end.

Bulk editing

Multiple annotations can be edited at once by selecting the annotations while holding down the cntrl/command key, and clicking Edit Annotations. Where properties contain different values in different annotations, *Multiple values* will be shown next to that property. Editing a property will change the value to be the same for all annotations. Intervals cannot be edited when multiple annotations are selected.

8.1.4 Extracting Annotations

To extract an annotation to a separate document, select it either on the sequence or in the Annotations Table and click **Extract**. If you want to actively link the extracted annotation to the source document (so that changes on the source document are propagated to the extracted document), check **Actively link source and extracted documents**.

If the annotation you are extracting contains multiple intervals, the intervals can be concatenated into a single sequence. If this option is not selected, then each interval in the annotation will be extracted to a separate sequence and grouped into a list.

Bulk extraction

Bulk extraction of annotations can be done in two ways:

1. Select all the annotations you want to extract, either on the sequence or in the Annotations table and click **Extract**. As with multi-interval annotations, you are given the option to concatenate all the annotations into a single sequence.
2. Go to **Extract Annotations** under the **Tools** menu. Using this interface, all annotations on the selected sequences which match certain criteria (e.g. a particular annotation type or gene name) can be extracted in bulk, without needing to select the annotations on the sequence first. To define what annotations to extract, select the value of the annotation type or property (qualifier) that you want to extract by in the chooser (see Figure 8.4). To set more than one criteria, click the + button to add an additional row of options, and choose to either **Match all...** or **Match any...** of the criteria.

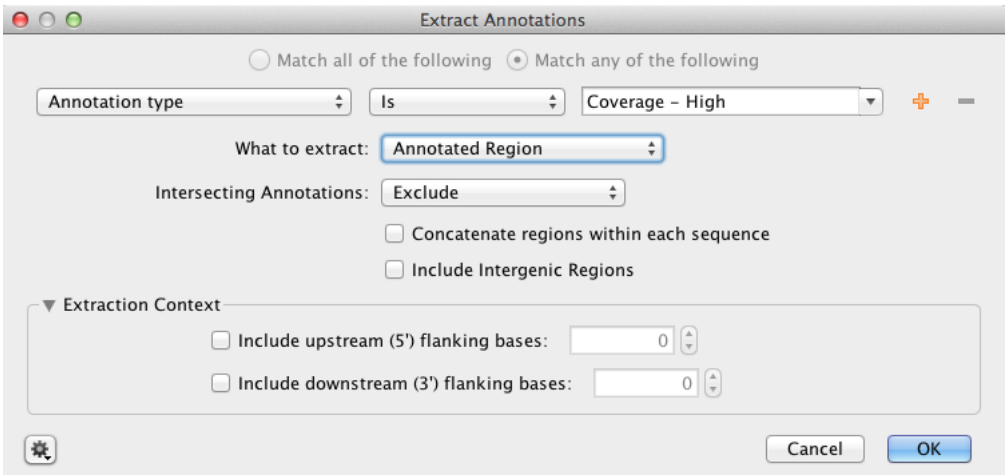


Figure 8.4: The Extract Annotations Interface

What to Extract allows you to set which part of the sequence to extract (e.g. just the annotated region, or the entire sequence) based on the criteria you have set. To extract regions of sequence upstream or downstream of the annotated sequence, enter the length of the additional sequence you want to extract under **Extraction Context**. **Intersecting Annotations** allows you to set what to do with other annotations that don't match the criteria, but which overlap with the matched region.

If there are multiple annotations on one sequence which match the criteria, these can be concatenated into one sequence by checking **Concatenate regions within each sequence**.

8.2 Adding annotations

Geneious has many functions for adding annotations to sequences. They can be added manually, imported from external sources, transferred from other sequences, or added as part of structure or gene prediction steps. Each of these options is described in the sections below. For more information on viewing and editing annotations, see sections 8 and 8.1.2.

8.2.1 Manual creation of annotations

To create an annotation on your sequence or alignment, select the region of the sequence where you wish to place the annotation and click the **Add Annotation** button. In the Add Annotation dialog enter an annotation name and select a existing type or type a new one. If you wish to put the annotation on a track rather than directly on the sequence, either choose an existing track from the drop-down menu, or type in a new track. Expand the **Properties** section to enter additional properties for that annotation. In the **Interval** section you can adjust the position of the annotation, add an additional interval, or mark the annotation as truncated at the 5′ or 3′ end.

8.2.2 Importing annotations from external sources

BED files, GFF files and VCF files contain annotation information which can be imported into Geneious. These files often do not contain the sequence itself, so when you import the file you will be prompted for the reference sequence as shown in Figure 8.5.

Note that if you choose **"use a sequence in the selected folder"**, or you have a sequence list or more than one sequence selected and choose **"use a sequence in the selected documents"**, then your sequence's name in Geneious must match the sequence name in the first column of the BED, GFF or VCF file. If you select a single sequence and choose **"use a sequence in the selected documents"** then your annotations will be imported onto that sequence regardless of whether it matches the sequence name in the file.

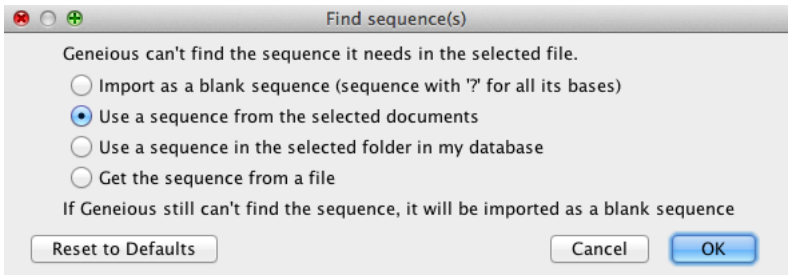


Figure 8.5: Selecting a reference sequence when importing BED, GFF or VCF files

For more information on these file formats, see:

- BED format: http://genome.ucsc.edu/FAQ/FAQformat.html#format1
- GFF format: http://www.sanger.ac.uk/resources/software/gff/spec.html
- VCF format: http://samtools.github.io/hts-specs/VCFv4.2.pdf

Note that in BED format the first base is numbered 0 rather than 1, and Geneious accounts for this when it imports the file so your annotations will be shifted 1 bp to the right compared to their positions listed in the BED file.

8.2.3 Transferring annotations from other sequences

Copy to...

You can copy annotations from one sequence to other sequences in the same alignment or assembly document by right-clicking on the annotation and choosing **Annotation** → **Copy to**. This will give you the option of either transferring the annotation to the consensus sequence, reference sequence (if there is one), or any of the other sequences in the alignment or assembly. You can also use the associated options: **Copy all x to...** to copy all annotations of the currently selected annotation type on the selected sequence; and **Copy all in selected region to...** to copy all annotations in the selected sequence. In each case, the annotation(s) will be copied across regardless of the similarity between the sequences.

Transfer Annotations

Using Transfer Annotations, you can copy annotations to the reference and/or consensus sequence of an alignment or assembly. This function can be accessed from the **Annotate and**

Predict menu, or in the **Live Annotate and Predict** tab in the sequence viewer. To set a sequence as a reference sequence Ctrl-click on it and choose **Set as reference sequence**.

Annotations will only be transferred where the annotated sequence and the reference/consensus sequence have at least the specified minimum similarity. All transferred annotations will be annotated with a **Transferred From** qualifier indicating the names of the sequences the annotation came from (sorted in order of decreasing similarity), and a **Transferred Similarity** qualifier which indicates the percentage similarity of the most similar sequence the annotation was transferred from.

Annotations of the same type and covering the same interval that would be transferred from multiple sequences are merged together such that the name of the transferred annotation will consist of the names of all contributing annotations sorted in order of decreasing similarity. Similarly, if contributing annotations have different qualifier values, the resulting qualifier value will consist of all contributing qualifier values sorted in order of decreasing similarity.

The percentage similarity is the sum of the similarity values for each position as a fraction of the sum of the maximum similarity values of the bases/residues in each position. For example if one sequence is LLK and the other is LIK using the Blosum62 matrix, L/L scores 4, I/I scores 4, L/I scores 2, K/K scores 5. Therefore the total score is 4+2+5=11 out of a maximum of 4+maximum(4,4)+5=13 for a percentage similarity of 85%. Gaps (if allowed) are scored as the lowest value from the score matrix (e.g. -4 for Blosum62).

Annotate from Database

Annotate from Database allows you to annotate your sequence with particular genes or motifs from a custom annotation database. The annotation database is a folder (called the **Source** folder) within your Geneious database which contains sequences with existing annotations (e.g. specific genes you are interested in searching for on other sequences). This function uses a BLAST-like algorithm to search for annotations in the Source folder that match your sequence. Annotations which match with the given similarity are copied to your sequence. The default annotation folder is the **PlasMapper Features** folder from the Sample Documents which contains features from the database used by PlasMapper. By default this also annotates Gateway recombination sites. This can be turned off in the Advanced options.

To use this function:

1. Create your annotation database by placing your sequences containing annotations of interest in a separate folder in Geneious.
2. Select the sequence(s) you want to annotate and check **Annotate From...** in the **Live Annotate and Predict** tab in the sequence viewer (or choose **Annotate and Predict** → **Annotate From Database** from the menu).
3. Click on the folder name next to the label **Source:**, and in the window that appears, select

the folder containing your annotation database, then click OK.

4. Adjust the similarity setting if necessary.
5. If you have any short annotations, click on the **Advanced** button and adjust the **Index Sequence Length**.
6. Click **Apply**.

This will apply any matching annotations that have been found in your annotation database to your sequence(s). If you wish to only apply only some of the annotations to your sequence, select the annotations you want (either directly on the sequence or in the annotations table) before you click Apply. This will apply only the selected annotations onto the sequence rather than all of them.

8.2.4 Annotation of sequence features using EMBOSS tools

The EMBOSS tools which were available under the Annotate and Predict menu in previous Geneious versions are now available as separate plugins. **EMBOSS protein analysis** includes **antigenic** to predict antigenic regions, **sigcleave** to predict signal cleavage sites, and **garnier** to predict secondary structures. The **EMBOSS Nucleotide Analysis** plugin includes **tfscan** to search for transcription factors, and **tcode** which provides a protein coding prediction graph.

When these plugins are installed, the protein coding prediction graph will be available under the Graphs tab, and the other options will show up under the Annotate and Predict menu. Further information on these applications is available at http://emboss.open-bio.org/

8.3 Compare Annotations

Geneious can compare annotations across up to 3 annotation tracks or documents, highlighting annotations which are common or unique depending on which criteria you choose.

To use this function, select the sequence or sequences containing annotations you wish to compare, and go to **Annotate and Predict** → **Compare Annotations**. In the Annotation Types panel, select the annotations you wish to compare. The default setting is for pairwise comparison; if you wish to do three-way comparison select **Set C**. Choose the annotation type and location for each set, and the type of comparison.

Comparison options are:

- **Names must match:** The names of the annotations must be the same. When allowing partial matches for polymorphism annotations, just the name of the matching region of the annotation must match.

- **All properties must match:** All properties in addition to the name must be the same to be considered a match
- **Allow intervals to partially match:** When one annotation partially overlaps another, you can choose to return either **Partial Annotations** or **Full Annotations**. Partial Annotations will return annotations spanning or excluding only the matching region. Full Annotations will return annotations spanning the entire length of the applicable source annotation. Uncheck this option if you only want to return annotations which match across the full length of both sets.

Results

A new annotation track will be created showing the results of the comparison. The results panel in the compare annotations set up allows you to choose which comparison to return:

- **A-B-C** returns annotations found only in set A
- **B-A-C** returns annotations found only in set B
- **C-A-B** returns annotations found only in set C
- **A&B&C** returns annotations common to all sets. For a pairwise comparison, when the annotation is common to both sets the result will have the name and type of annotation from set A but share properties from both sets if selected. For a three-way comparison properties from set A will be used.

More than one of these options can be selected at once, either by checking the box next to the option, or clicking the appropriate section of the venn diagram to the right. Each result comparison is displayed on a separate track on the original sequence, and a preview of these tracks is given in the Example panel.

Example 1 - finding polymorphisms within a gene or feature (e.g. CDS, restriction site)

To return a track containing polymorphism annotations that are within another gene or feature, such as a coding sequence, select the Polymorphism annotation type for Set A, and choose which track it is on (or select anywhere if there are no tracks, or you want to include polymorphisms on any track). Then choose the other annotation type (e.g. CDS) for Set B. Uncheck **Names must match** as in general polymorphism annotation names do not match those of other annotation types, and check **Allow intervals to partially match and produce partial annotations**. Under Results, choose to return **A & B (annotations common to both sets)**. This will return a new track containing annotations of the type in Set A (polymorphisms) that are contained within Set B (CDS) annotations. See Figure 8.6.

Example 2 - finding polymorphisms in a child which are not present in either parent

To return polymorphisms which are unique to the offspring, set Annotation Type as Polymorphism for Sets A, B and C, and set A and B as the parent tracks, and C as the child track, as in Figure 8.7. Check **Names must match** and **Allow intervals to partially match and produce partial annotations**. Choose **C-A-B** in the results display. This will return a new track containing annotations found in Set C (child), but not in Sets A and B (parents).

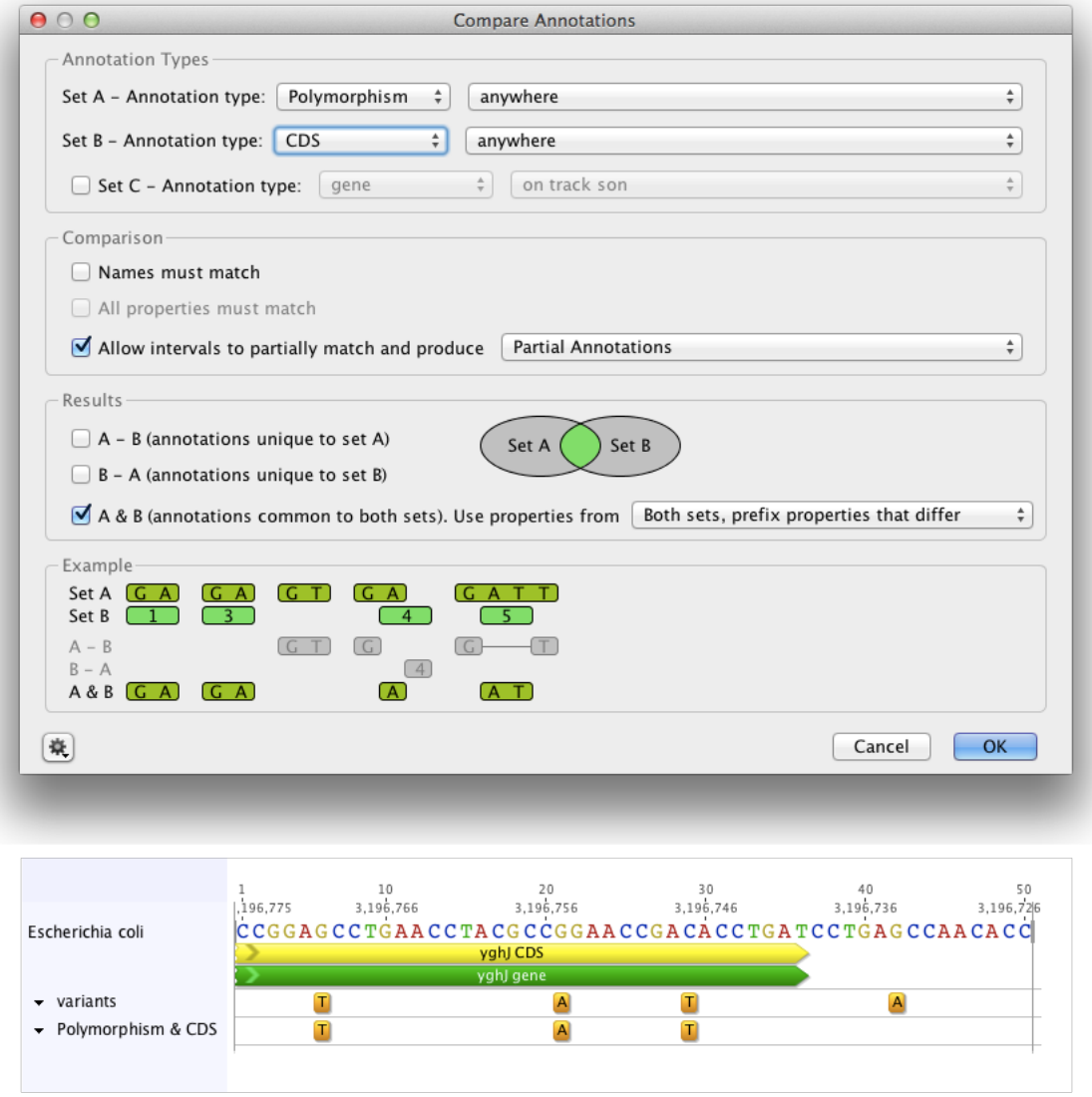


Figure 8.6: Example 1: Compare annotations setup (above) and results (below) for finding polymorphisms within a CDS

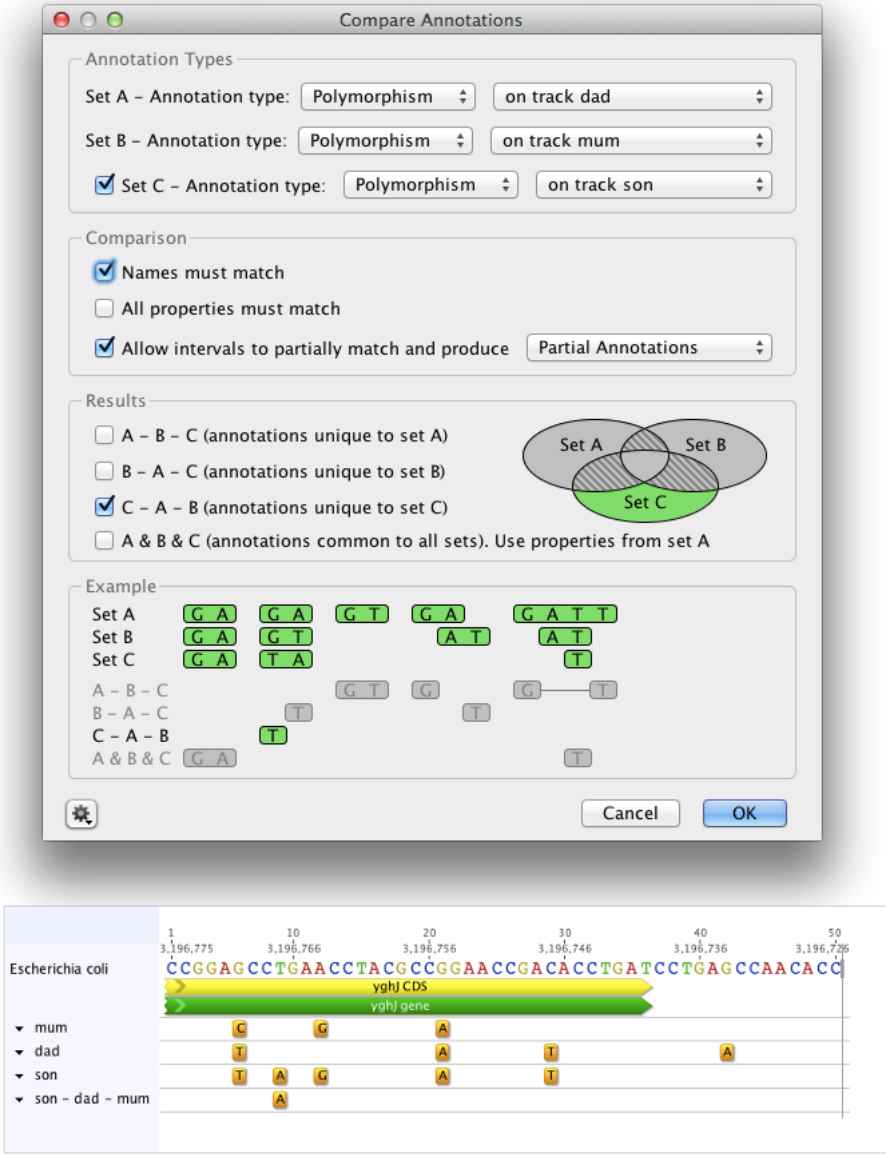


Figure 8.7: Example 2: Compare annotations setup (above) and results (below) for finding polymorphisms in a child which are not present in either parent

Chapter 9

Sequence alignments

9.1 Dotplots

A dotplot compares two sequences against each other and helps identify similar regions. Using this tool, it can be determined whether a similarity between the two sequences is global (present from start to end) or local (present in patches).

To view a dotplot in Geneious, select two nucleotide or protein sequences in the Document Table and select Dotplot in the tab above the sequence viewer (Figure 9.1). If a single nucleotide or protein sequence is selected then a dot plot showing a comparison of the sequence to itself can be viewed.

The Geneious dotplot offers two different comparison engines based on the EMBOSS `dottup` and `dotmatcher` programs. You can choose which program to use by setting the sensitivity under **Data Source**, the panel to the right of the dot plot. The **Low Sensitivity/Fast** setting uses `dottup`, and the **High Sensitivity/Slow** setting uses `dotmatcher`. More information on these programs can be found by going to `http://emboss.sourceforge.net`.

The dotplot is drawn from top-left to bottom-right. The **Minimap** in the panel to the right of the viewer aids navigation of large dotplots by showing the overall comparison and a box indicating where the dotplot window sits.

The **Colors** for the Dotplot can be selected at the top of the settings panel. The Classic scheme will color the dot plot lines according to the length of the match, from blue for short matches, to red for matches over 100 bp long.

For nucleotide comparisons, the **reverse complement** can also be viewed, where matches with one of the sequences reverse complemented are displayed. These matches are shown by lines running from the bottom left to top right. When a pairwise alignment is selected, the path that the alignment takes through the dot plot can be displayed by checking **Pairwise alignment path**. This is shown as a light blue line running through the dot plot. Both of these options can

be found under the **Display** section.

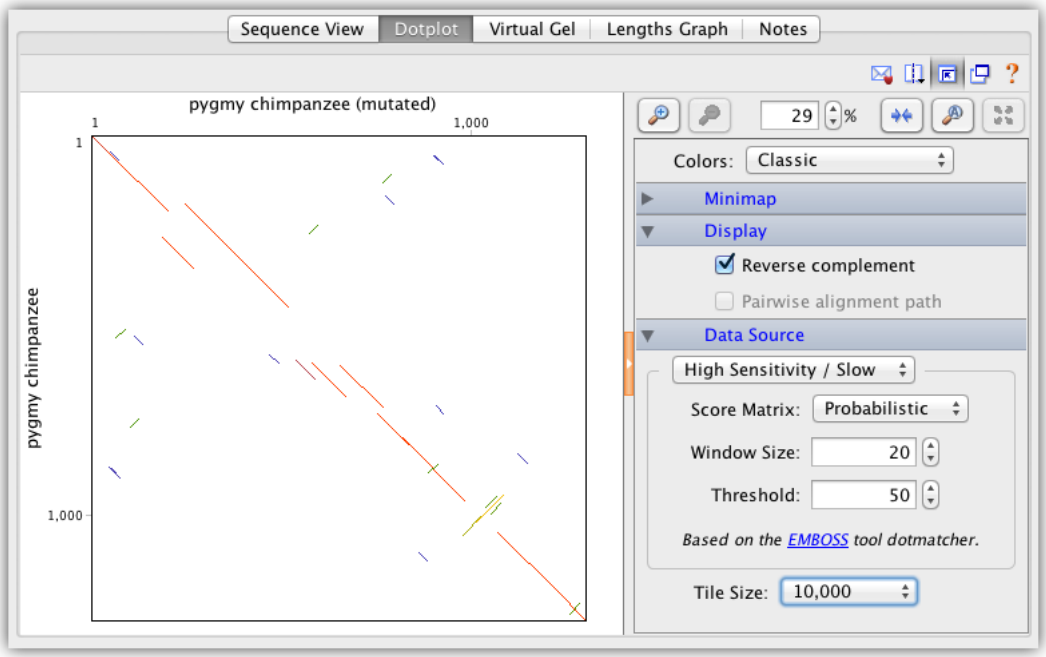


Figure 9.1: A view of dotplot of two sequences in Geneious

Interpreting a Dotplot

- Each axis of the plot represents a sequence.
- A long, largely continuous, diagonal indicates that the sequences are related along their entire length.
- Sequences with some limited regions of similarity will display short stretches of diagonal lines.
- Diagonals on either side of the main diagonal indicate repeat regions caused by duplication.
- A random scattering of dots reflects a lack of significant similarity. These dots are caused by short sub-sequences that match by chance alone.

For more information on dotplots, refer to the paper by Maizel and Lenk 1981

9.2 Sequence Alignments

Over evolutionary time, related DNA or amino acid sequences diverge through the accumulation of mutation events such as nucleotide or amino acid substitutions, insertions and deletions.

A *sequence alignment* is an attempt to determine regions of homology in a set of sequences. It consists of a table with one sequence per row, and with each column containing homologous residues from the different sequences, e.g. residues that are thought to have evolved from a common ancestral nucleotide/amino acid. If it is thought that the ancestral nucleotide/amino acid got lost on the evolutionary path to one descendant sequence, this sequence will show a special gap character "–" in that alignment column.

9.2.1 Pairwise sequence alignments

There are two types of pairwise alignments: **local** and **global** alignments.

A **local alignment** is an alignment of two sub-regions of a pair of sequences. This type of alignment is appropriate when aligning two segments of genomic DNA that may have local regions of similarity embedded in a background of a non-homologous sequence.

A **global alignment** is a sequence alignment over the entire length of two or more nucleic acid or protein sequences. In a global alignment, the sequences are assumed to be homologous along their entire length.

Scoring systems in pairwise alignments

In order to align a pair of sequences, a scoring system is required to score matches and mismatches. The scoring system can be as simple as "+1" for a match and "-1" for a mismatch between the pair of sequences at any given site of comparison. However substitutions, insertions and deletions occur at different rates over evolutionary time. This variation in rates is the result of a large number of factors, including the mutation process, genetic drift and natural selection. For protein sequences, the relative rates of different substitutions can be empirically determined by comparing a large number of related sequences. These empirical measurements can then form the basis of a scoring system for aligning subsequent sequences. Many scoring systems have been developed in this way. These matrices incorporate the evolutionary preferences for certain substitutions over other kinds of substitutions in the form of log-odd scores. Popular matrices used for protein alignments are BLOSUM and PAM[1] matrices.

Note: The BLOSUM and PAM matrices are substitution matrices. The number of a BLOSUM matrix indicates the threshold (%) similarity between the sequences originally used to create the

[1]MO. Dayhoff (ed.), Atlas of protein sequence and structure, vol. 5, National biomedical research foundation Washington DC, 1978

matrix. BLOSUM matrices with higher numbers are more suitable for aligning closely related sequences. For PAM, the lower numbered tables are for closely related sequences and higher numbered PAMs are for more distant groups.

When aligning protein sequences in Geneious, a number of BLOSUM and PAM matrices are available.

Algorithms for pairwise alignments

Once a scoring system has been chosen, we need an algorithm to find the optimal alignment of two sequences. This is done by inserting gaps in order to maximize the alignment score. If the sequences are related along their entire sequence, a global alignment is appropriate. However, if the relatedness of the sequences is unknown or they are expected to share only small regions of similarity, (such as a common domain) then a local alignment is more appropriate.

An efficient algorithm for global alignment was described by Needleman and Wunsch 1970, and their algorithms was later extended by Gotoh 1982 to model gaps more accurately. For local alignments, the Smith-Waterman algorithm is the most commonly used. See the references at the links provided for further information on these algorithms.

Pairwise alignment in Geneious

Like a dot plot, a pairwise alignment is comparison between two sequences with the aim of identifying which regions of two sequences are related by common ancestry and which regions of the sequences have been subjected to insertions, deletions, and substitutions.

To run a pairwise alignment in Geneious, select the two sequences you wish to align and choose **Align/Assemble** → **Pairwise align...**. The options available for the alignment cost matrix will depend on the kind of sequence.

- Protein sequences have a choice of PAM and BLOSUM matrices.
- Nucleotide sequences have choices for a pair of match/mismatch costs. Some scores distinguish between two types of mismatches: transition and transversion. Transitions ($A \leftrightarrow G$, $C \leftrightarrow T$) generally occur more frequently than transversions. Differences in the ratio of transversions and transversions result in various models of substitution. When applicable, Geneious indicates the target sequence similarity for the alignment scores, i.e. the amount of similarity between the sequences for which those scores are optimal.
- Both protein and nucleotide pairwise alignments have choices for gap open / gap extension penalties/costs. Unlike many alignment programs these values are not restricted to integers in Geneious.

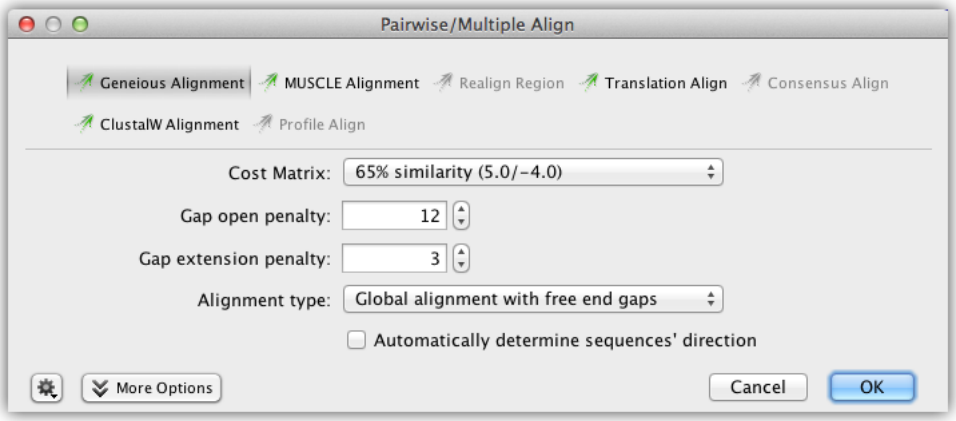


Figure 9.2: Options for nucleotide pairwise alignment

The score of a pairwise alignment is:

$matchCount \times matchCost + mismatchCount \times mismatchCost$

For each gap of length n, a score of $gapOpenPenalty + (n - 1) \times gapExtensionPenalty$ is subtracted from this.

Where

- $gapOpenPenalty =$ The "gap open penalty" setting in Geneious.
- $gapExtensionPenalty =$ The "gap extension penalty" setting in Geneious.
- $matchCost =$ The first number in the Geneious cost matrix.
- $mismatchCost =$ The second number in the Geneious cost matrix.
- $matchCount =$ The number of matching residues in the alignment.
- $mismatchCount =$ The number of mismatched residues in the alignment.

When doing a **Global alignment with free end gaps**, gaps at either end of the alignment are not penalized when determining the optimal alignment. This is especially useful if you are aligning sequence fragments that overlap slightly in their starting and ending positions, e.g. when using two slightly different primer pairs to extract related sequence fragments from different samples. You can also do a **Local Alignment** if you want to allow free end overlaps, rather than just free end gaps in one alignment.

9.2.2 Multiple sequence alignments

A multiple sequence alignment is a comparison of multiple related DNA or amino acid sequences. A multiple sequence alignment can be used for many purposes including inferring the presence of ancestral relationships between the sequences. It should be noted that protein sequences that are structurally very similar can be evolutionarily distant. This is referred to as distant homology. While handling protein sequences, it is important to be able to tell what a multiple sequence alignment means – both structurally and evolutionarily. It is not always possible to clearly identify structurally or evolutionarily homologous positions and create a single "correct" multiple sequence alignment (Durbin et al 1998).

Multiple sequence alignments can be done by hand but this requires expert knowledge of molecular sequence evolution and experience in the field. Hence the need for automatic multiple sequence alignments based on objective criteria. One way to score such an alignment would be to use a probabilistic model of sequence evolution and select the alignment that is most probable given the model of evolution. While this is an attractive option there are no efficient algorithms for doing this currently available. However a number of useful heuristic algorithms for multiple sequence alignment do exist.

Progressive pairwise alignment methods

The most popular and time-efficient method of multiple sequence alignment is progressive pairwise alignment. The idea is very simple. At each step, a pairwise alignment is performed. In the first step, two sequences are selected and aligned. The pairwise alignment is added to the mix and the two sequences are removed. In subsequent steps, one of three things can happen:

- Another pair of sequences is aligned
- A sequence is aligned with one of the intermediate alignments
- A pair of intermediate alignments is aligned

This process is repeated until a single alignment containing all of the sequences remains. Feng & Doolittle were the first to describe progressive pairwise alignment. Their algorithm used a guide tree to choose which pair of sequences/alignments to align at each step. Many variations of the progressive pairwise alignment algorithm exist, including the one used in the popular alignment software ClustalX.

Multiple sequence alignment in Geneious

To run a multiple alignment in Genieous, select all the sequences you wish to align and click **Align/Assemble** →**Multiple align...**. Select **Geneious** as the alignment algorithm. The Geneious

multiple alignment algorithm uses progressive pairwise alignment. The neighbor-joining method of tree building is used to create the guide tree.

As progressive pairwise alignment proceeds via a series of pairwise alignments, this function in Geneious has all the standard pairwise alignment options. In addition, Geneious also has the option of refining the multiple sequence alignment once it is done. "Refining" an alignment involves removing sequences from the alignment one at a time, and then realigning the removed sequence to a "profile" of the remaining sequences. The number of times each sequence is realigned is determined by the **refinement iterations** option in the multiple alignment window. The resulting alignment is placed in the folder containing the original sequences.

A profile is a matrix of numbers representing the proportion of symbols (nucleotide or amino acid) at each position in an alignment. This can then be pairwise aligned to another sequence or alignment profile. When pairwise aligning profiles, mismatch costs are weighted proportional to the fraction of mismatching bases and gap introduction and gap extension costs are proportionally reduced at sites where the other profile contains some gaps.

In some cases building a guide tree can take a long time since it requires making a pairwise alignment between each pair of sequences. The **build guide tree via alignment** option may speed this part by taking a different route. First make a progressive multiple alignment using a random ordering, and use that alignment to build the guide tree. Notice that while this usually speeds up the process, it may not if the sequences are very distant genetically.

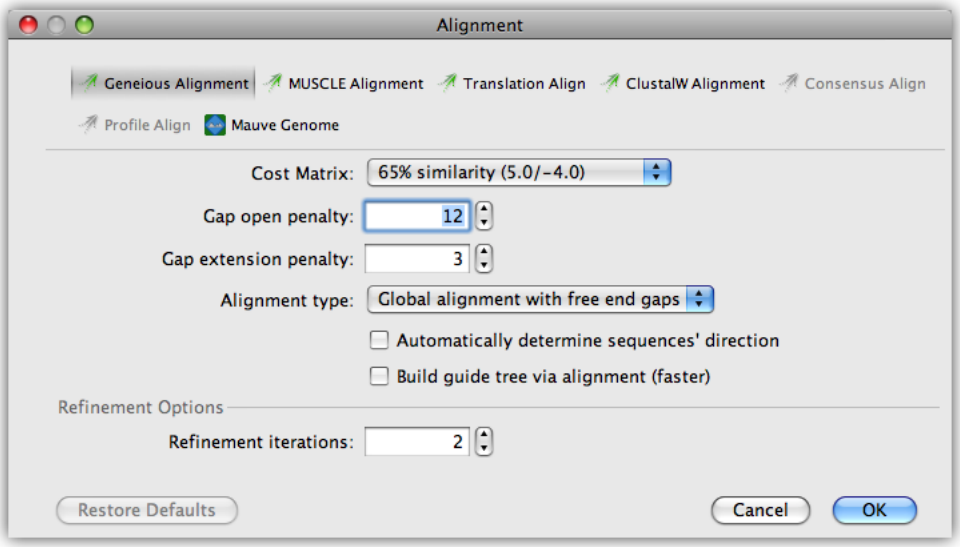


Figure 9.3: The multiple alignment window

You can also do a multiple alignment via **translation** and back, as with pairwise alignment (see section 9.2.6)

9.2.3 Sequence alignment using ClustalW

ClustalW is a widely used program for performing sequence alignment. Geneious allows you to run ClustalW directly from inside the program without having to export or import your sequences.

To perform an alignment using ClustalW, select the sequences or alignment you wish to align, then select **Align/Assemble** → **Multiple Align...**. Select **ClustalW** as the alignment type, and the options available for a ClustalW alignment will be displayed.

The options are:

- *Cost Matrix:* Use this to select the desired cost matrix for the alignment. The available options here will change according to the type of the sequences you wish to align. You can also click the 'Custom File' button to use a cost matrix that you have on your computer (the format of these is the same as for the program BLAST).
- *Gap open cost* and *Gap extend cost:* Enter the desired gap costs for the alignment.
- *Free end gaps:* Select this option to avoid penalizing gaps at either end of the alignment. See details in the Pairwise Alignment section above.
- *Preserve original sequence order:* Select this option to have the order of the sequences in the table preserved so that the alignment contains the sequences in the same order.
- *Additional options:* Any additional parameters accepted by the ClustalW command line program can be entered here. Refer to the ClustalW manual for a description of the available parameters.
- *Custom ClustalW executable:* The ClustalW2 executable is bundled with Geneious so there is no need to download this yourself. However, if you wish to use a different executable from the one within Geneious, you can choose this here.

You can also do a Clustal alignment via translation and back (see section 9.2.6).

After entering the desired options click **OK** and ClustalW will be called to align the selected sequences or alignment. Once complete, a new alignment document will be generated with the result as detailed previously.

9.2.4 Sequence alignment using MUSCLE

MUSCLE is public domain multiple alignment software for protein and nucleotide sequences. MUSCLE stands for multiple sequence comparison by log-expectation.

To perform an alignment using MUSCLE, select the sequences or alignment you wish to align and select **Align/Assemble** → **Multiple Align...**. Select **MUSCLE** as the alignment type, and the options available for a MUSCLE alignment will be displayed.

For more information on muscle and its options, please refer to the original documentation for the program: http://www.drive5.com/muscle/muscle.html.

9.2.5 Other sequence alignment plugins for Geneious

MAFFT

MAFFT (Multiple Alignment using Fast Fourier Transform) is a fast multiple alignment program suitable for large alignments. To use MAFFT, you must first download the plugin by going to **Plugins** under the **Tools** menu and selecting **MAFFT Multiple Alignmen** from the list of available plugins. Click the Install button to install it, and then click OK to close the Plugins window. To run MAFFT, select the sequences or alignment you wish to align, select the **Align/Assemble** button from the Toolbar and choose **Multiple Alignment**. MAFFT should now be showing as an option for the type of alignment.

For more information on MAFFT and its options, please refer to the original documentation for the program: http://mafft.cbrc.jp/alignment/software/.

Mauve

The Mauve aligner allows you to construct whole genome multiple alignments in the presence of large-scale evolutionary events such as rearrangement and inversion. To use Mauve, you must first download the plugin by going to **Plugins** under the **Tools** menu and selecting **Mauve** from the list of available plugins. Click the Install button to install it, and then click OK to close the Plugins window. To run Mauve, select the sequences or alignment you wish to align, select the **Align/Assemble** button from the Toolbar and choose **Align Whole Genomes**.

An alignment produced by Mauve is displayed in the Mauve genome alignment viewer, which allows you to easy see aligned blocks of sequence and genome rearrangements. Note that this is not a regular Geneious alignment document and you cannot run downstream analyses such as tree building from this document. To run downstream analyses you must first extract the local alignment blocks. To do this, switch to the Alignment View tab above the sequence viewer and if you have more than one local alignment block, choose which one you wish to extract in the General tab to the right of the sequence viewer. Then select all the sequences in that alignment and click **Extract**. Choose the option to extract the sequences as an alignment, and a separate alignment document will be created in the document table.

For more information on Mauve and its options, please refer to the original documentation for the program: http://darlinglab.org/mauve/mauve.html.

LASTZ

LASTZ is designed for pairwise alignments of whole genomes and can efficiently align chromosomal or genomic sequences millions of nucleotides in length. To use LASTZ, you must first download the plugin by going to **Plugins** under the **Tools** menu and selecting **LASTZ** from the list of available plugins. Click the Install button to install it, and then click OK to close the Plugins window. To run LASTZ, select the sequences or alignment you wish to align, select the **Align/Assemble** button from the Toolbar and choose **Align Whole Genomes**.

For more information on LASTZ and its options, please refer to the original documentation for the program: http://www.bx.psu.edu/~rsharris/lastz/

9.2.6 Translation alignment

In a translation alignment, nucleotide sequences are translated into protein, the alignment is performed on the protein sequence, and then translated back to nucleotide sequence. Translation alignments can be performed with any of the alignment algorithms in Geneious, such as the Geneious aligner, MUSCLE or ClustalW.

In the translation alignment options you can set the genetic code and translation frame for the translation as well as the alignment algorithm you wish to use. Under **More Options** you can set the parameters such as cost matrix, gap open penalty and gap extension penalty for the alignment.

9.2.7 Combining alignments and adding sequences to alignments

Consensus Alignment allows you to align two or more alignments together (and create a single alignment) and align a new sequence in to an existing alignment. Select the sequences or alignment you wish to align and select the **Align/Assemble** button from the Toolbar and choose **Multiple Alignment**. Consensus alignment allows you to choose which alignment algorithm to use for aligning the consensus sequences. All of the pairwise and multiple alignment algorithms are available. The consensus sequence used for each alignment is a 100% consensus with gaps ignored.

9.3 Alignment viewing and editing

Alignments are displayed in the viewer below the document table, in the same way as individual sequences. See section 5.2 for details on the sequence viewer, including basic controls such as zooming in and out, wrapping sequences, setting colors, and selecting individual or multiple sequences from an alignment. For a description of alignment statistics available in the

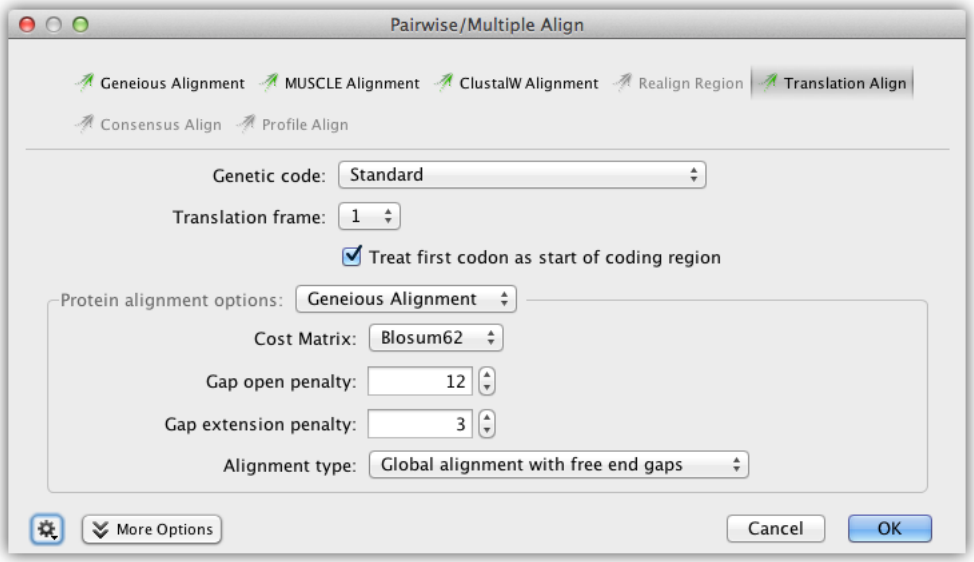


Figure 9.4: Options for nucleotide translation alignment

% **Statistics** tab, see section 5.2.9.

To **edit** an alignment, you must first click the **Allow Editing** button on the toolbar above the sequence viewer. Alignments are edited in the same way as for individual sequences - for details on editing operations and shortcut keys, see section 5.3. If the consensus sequence of an alignment or assembly is edited, the changes are applied to all sequences in the alignment, with the exception of the reference sequence.

9.3.1 Highlighting

Identical or variable sites in an alignment can be highlighted by checking the **Highlighting** option under the **Display** tab. The drop down menus allow you to choose what to highlight, and whether the consensus or reference sequence should be used for the comparison. The options for what to highlight are:

- **Disagreements**: This greys out residues that are identical to the consensus or reference in that column, allowing you to quickly locate variable sites in the alignment.
- **Agreements**: Greys out residues that are not indentical to the consensus or reference allowing you to quickly locate conserved sites in the alignments.
- **Ambiguities**: Greys out non-ambiguous residues.

- **Gaps**: Greys out non-gap positions.
- **Transitions/transversions**: Greys out residues that are not transitions/transversions compared to the consensus or reference sequence. When highlighting transitions/transversions, it is recommended you turn on the ignore gaps consensus option or some residues may be wrongly highlighted due the consensus displaying N for sites that contain gaps and non-gaps.

Go to next disagreement/agreement/transition/transversion/ambiguity goes to the next highlighted feature as described in the previous section.

If **Use dots** is checked, non-highlighted residues are displayed as dots instead of being greyed out.

9.3.2 Alignment graphs

In addition to the basic graphs available for individual sequences, the following graphs are available for alignments and assemblies:

Coverage: The height of the graph at each position represents the number of sequences which have a non-gap character at that position. The coverage graph is made up of three bar graphs overlaid on each other: a blue graph shows the minimum coverage, a black graph shows the mean coverage and a yellow graph (underneath the blue and black graphs) shows the maximum coverage. The minimum graph is drawn over the top of the mean color graph, but if necessary the minimum color graph will be reduced in height so that a single pixel of the mean color graph is always visible at each position. Thus, for sequences which are zoomed in so that the horizontal width of each site is one pixel or more, then the graph will be shown in blue with a black line across the top, denoting the coverage at that position. For large alignments which are zoomed out so that the horizontal width of each site is less than one pixel (i.e. each pixel represents more than one site in the alignment), all three bars are visible, showing the minimum, mean and maximum coverage of bases within that pixel (see Figure 9.5).

To highlight regions above or below a particular coverage level, check **Highlight above...** or **Highlight below...** and a bar will appear below the coverage graph across regions which fit these criteria. The "Highlight above" bar is blue, and the "Highlight below" bar is yellow. Regions where the alignment or assembly is made up of sequences in a single direction (e.g. forward or reverse sequences only) can be highlighted by checking **Highlight single strand**.

The scale bar to the left of the graph shows minimum and maximum coverage for the entire alignment or assembly, as well as a tick somewhere in between for the mean coverage.

Sequence Logo: This displays a sequence logo, where the height of the logo at each site is equal to the total information at that site and the height of each symbol in the logo is proportional to its contribution to the information content. When zoomed out far enough such that the horizontal width of each site is less than one pixel, then the height is the average of the information

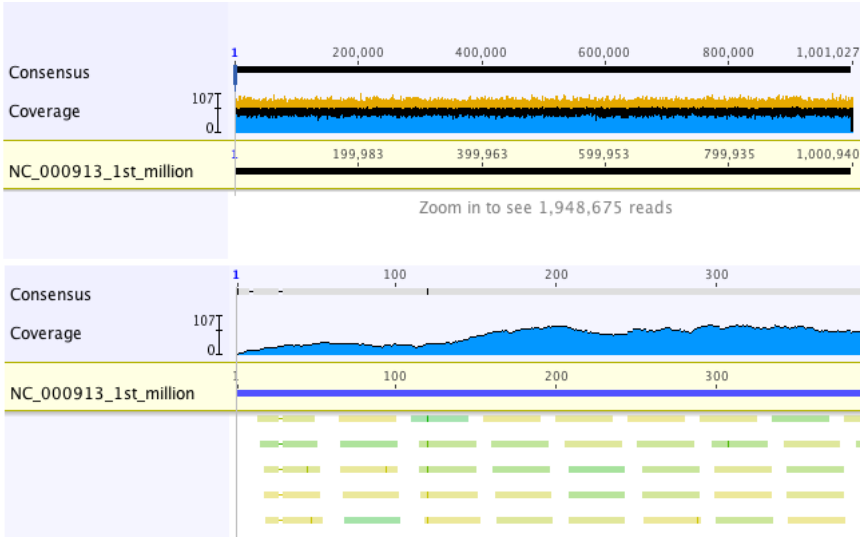


Figure 9.5: The coverage graph for an assembly, shown zoomed out in the top panel, and zoomed in below

over multiple sites. When gaps occur at at some sites, the height is scaled down further to be proportional in height to the number of non-gap residues.

Identity: This displays the identity across all sequences for every position. Green means that the residue at the position is the same across all sequences. Yellow is for less than complete identity and red refers to very low identity for the given position (Figure 9.6).

9.3.3 Concatenating alignments

Concatenating alignments works in the same way as concatenating sequences (see section 5.3.1). However, for alignments to be concatenated, the sequences names must be identical between alignments so that the correct sequences are matched up.

9.3.4 Strip alignment columns

To strip columns out of your alignment, go to **Tools** → **Strip alignment columns**. This will create a new alignment document from your original alignment, with the sites you have chosen removed. Options are available for stripping columns containing all or some gaps, identical sites, identical sites plus transitions, and ambiguous sites.

For the purposes of phylogenetic analyses, you can also use this tool to create alignments composed of 1st, 2nd and/or 3rd codon positions using the options to **Strip every third codon** or **Strip two adjacent columns per codon**.

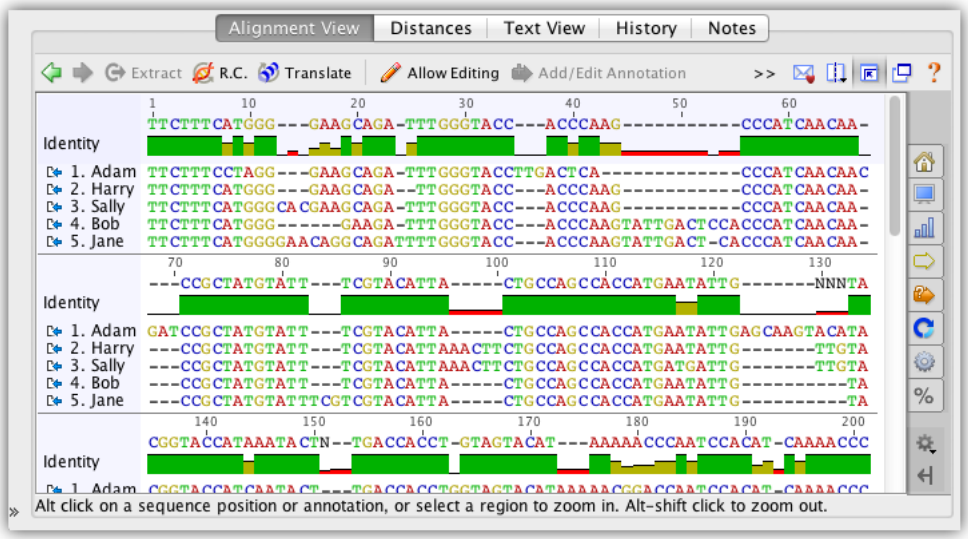


Figure 9.6: The identity graph for an alignment of nucleotide sequences

Trimmed regions in sequences are not included when deciding when to strip a column.

9.4 Consensus sequences

To display a consensus sequence on your alignment, check the **Consensus** option under the **Display** tab.

The consensus sequence is displayed above the alignment or assembly, and shows which residues are conserved (are always the same), and which residues are variable. A consensus is constructed from the most frequent residues at each site (alignment column), so that the total fraction of rows represented by the selected residues in that column reaches at least a specified threshold. IUPAC ambiguity codes (such as R for an A or G nucleotide) are counted as fractional support for each nucleotide in the ambiguity set (A and G, in this case), thus two rows with R are counted the same as one row with A and one row with G. When more than one nucleotide is necessary to reach the desired threshold, this is represented by the best-fit ambiguity symbol in the consensus; for protein sequences, this will always be an X.

For example, assume a column contains 6 A's, 3 G's and 1 T. If the consensus threshold is set to 60% or below, then the consensus will be A. If the consensus threshold is set to between 60% and 90%, then the consensus will be R. If the consensus threshold is set to over 90%, then the consensus will be D.

In the case of ties, either all or none of the involved residues will be selected. For example, if the above case instead had 6 A's, 2 G's and 2 T's, then for a consensus threshold of 60% or below, an A will be called. Above a threshold of 60%, a D will be called.

When *Ignore Gaps* is checked, the consensus is calculated as if each alignment column consisted only of the non-gap characters; otherwise, the gap character is treated like a normal residue, but mixing a gap with any other residue in the consensus always produces the total ambiguity symbol (N and X for nucleotides and amino acids, respectively).

When the aligned sequences contain quality information in the form of chromatograms or fastq data, you can select *Highest Quality* to calculate a majority consensus that takes the relative residue quality into account. This sums the total quality for each potential base call, and if the total for a base exceeds 60% of the total quality for all bases, then that base is called.

You can also choose to map the quality of the sequences onto the consensus. Choose *Highest* to map the quality of the highest quality base at each column onto the consensus. Select *Total* to map the sum of the contributing bases, minus the sum of the non-contributing bases.

For example: if there are two G's and three A's in a column, with the G's having qualities of 16 and 24, and the A's having qualities of 40, 42, and 50 respectively, then because $(40+42+50) >$ 60% of $(40 + 42 + 50 + 16 + 24)$, then an A will be called for the consensus. This consensus A will have a quality of $(40 + 42 + 50) - (16 + 24) = 92$ if using *Total* or 50 if using *Highest*.

A more complicated example for *Highest Quality* consensus calling using *Total*: Assume a column contains 2 A's with qualities of 30 and 25, 1 G with quality 30 and 1 T with quality 15. Because the total qualities of the A's is 55 out of 100 for the column, this is not higher than the 60% threshold to call an A. With the G included, the total quality is 30 + 25 + 30 = 85, which is higher than the 60% threshold, so a consensus call of R will be made. The quality assigned to this R will be the sum of the bases that agree with the consensus call minus the bases that disagree, which is $30 + 25 + 30 - 15 = 70$.

For alignments or contigs with a reference sequence, the *If no coverage call* setting can be used to control what character the consensus sequence should use when the reference sequence has no coverage. Options available are -, X/N, ? or Ref. A '?' represents an unknown character, potentially a gap. If Ref is selected, then the consensus is assigned whatever character the reference sequence has at that position. Note that if any sequence in the alignment/contig has an internal gap in it, that is still considered valid coverage at that position, and this setting will not apply.

Choose *Call N if Quality below* to change consensus bases to N's if the quality is below the threshold that you set. This is particularly useful for exporting sequences to file formats which do not preserve quality (for example FASTA).

To work with the consensus sequence in a downstream analysis, you must first **Extract** it from your alignment. To do this, click on **Consensus** to select the entire sequence, then click **Extract** to extract it to a new sequence document. Alternatively, go to **Tools** → **Generate Consensus Sequence**. This operation allows you to choose the options for how your consensus sequence

is called (as described above), and then saves it to a separate document. If your consensus sequence contains ? characters where there are regions of no/low coverage in your assembly, you can split the consensus sequence at these bases to generate multiple sequences by checking the option to **Split into separate sequences around ? calls**

Chapter 10

Assembly and Mapping

Assembly is normally used to merge overlapping fragments of a DNA sequence into a contig which can be used to determine the original sequence. The contig essentially appears as a multiple sequence alignment of the fragments. After some manual editing of the contig to resolve disagreements between fragments which result from read errors, the consensus sequence of the contig is extracted as the sequence being reconstructed.

Contig assembly is also used to align a large number of reads of the same sequence (from different individuals). This is done to find small differences between reads or SNPs (Single Nucleotide Polymorphisms). In this type of analysis the consensus sequence of the contig is not the interesting part, the differences between fragments is. This can also be done against a known reference sequence when differences between each of the fragments and the reference are of interest.

10.1 Read processing

10.1.1 Trim Ends

Trimming low quality ends of sequences is normally performed before assembling a contig. This is because the noise introduced by low quality regions and vector contamination can produce incorrect assemblies.

To trim vectors, primers and poor quality bases using the Geneious tools, select the sequences you wish to trim and choose **Annotate and Predict** → **Trim Ends**. This option can also be performed at the assembly step, by checking the trim sequences option in the assembly set-up. Geneious R9 also provides a plugin for trimming using the BBDuk algorithm from the BBTools suite. This is can be installed by going to **Tools** → **Plugins**.

Trim Ends can soft or hard trim your sequences. If you wish to soft trim, choose to **Annotate**

new trimmed regions in the Trim Ends set up. The trimmed sequence will then remain visible but will be annotated with "Trimmed" annotations. Sequence annotated with a trimmed annotation is ignored by the assembler when constructing a contig and will not be included in the consensus sequence calculation. So although the trimmed regions are visible, they do not affect the results of assembly at all. Soft trims can be adjusted as needed, or deleted completely. Dragging the ends of the trim annotation will make the newly untrimmed sequence visible and part of the consensus (Figure 10.1). If you wish to remove the trimmed sequence completely (hard trim), choose **Remove new trimmed regions from sequences**.

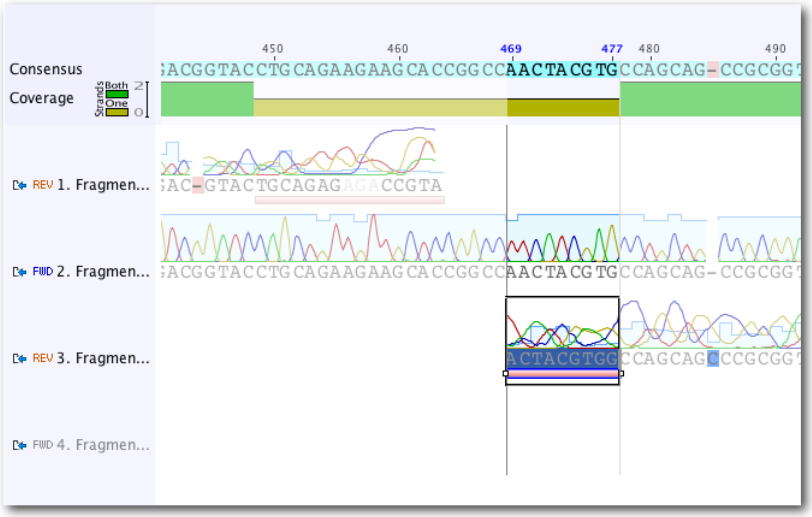


Figure 10.1: Click and drag the trims to adjust

If you choose to trim your sequences at the assembly step, the sequences are trimmed and assembled in one operation and you will not be able to view the trimming that Geneious uses before assembly is performed. However, the trimmed regions will still be available and adjustable after assembly is complete. If you choose to trim your sequences prior to assembly, select **Use Existing Trim Regions** when you set up the assembly.

Trimmed annotations can also be created manually using the annotation editing in the sequence viewer. If you create annotations of type "trimmed" and save them, then Geneious will treat them the same as ones generated automatically and they will be ignored during assembly. Trimmed annotations can also be modified in this way before or after assembly.

Trim Ends options

- **Annotate new trimmed regions:** Calculate new trimmed regions and annotate them - the trimmed regions will be ignored when performing assembly and calculating the consensus sequence.

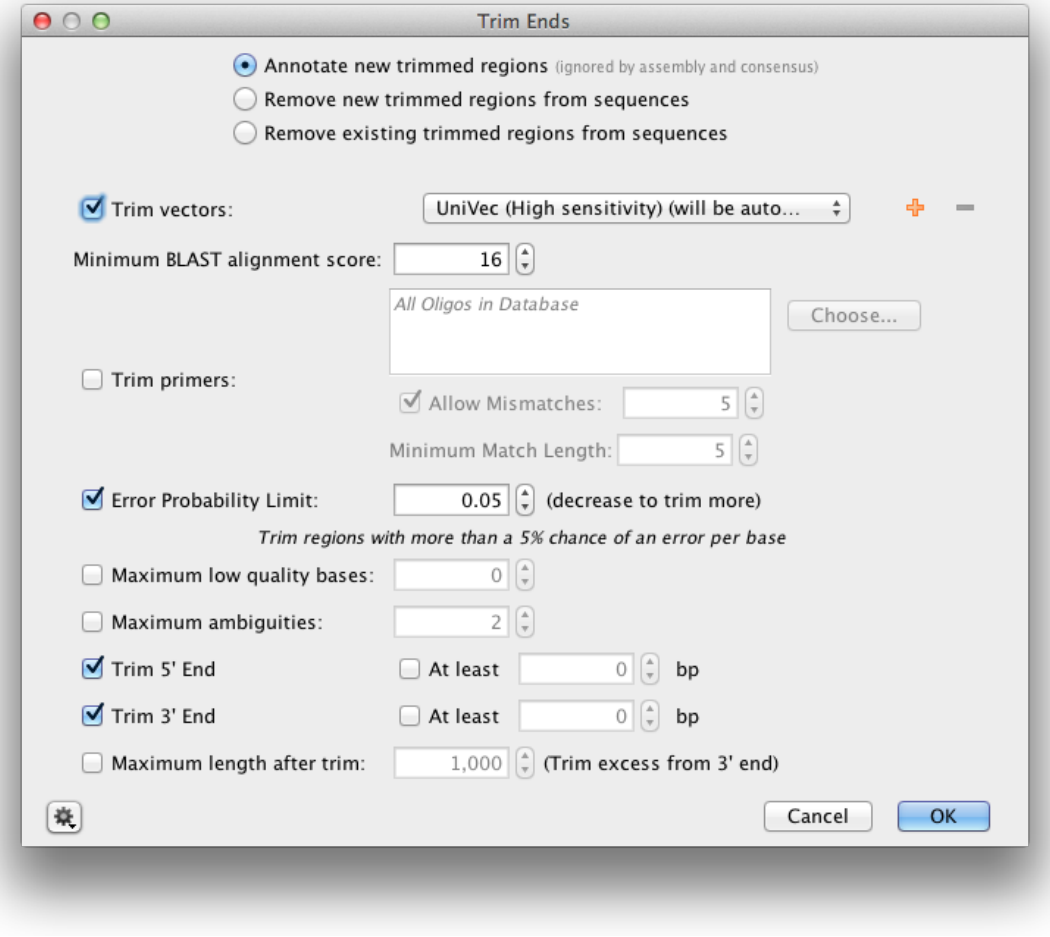


Figure 10.2: Trimming options

- **Remove new trimmed regions from sequences:** Calculate new trimmed regions and remove them from the sequence(s) completely. This can be undone in the Sequence View before the sequences are saved.
- **Remove existing trimmed regions from sequences:** This is only available when there are already trimmed regions on some of the sequences. This will remove the existing trimmed regions from the sequences permanently; no new trimmed regions are calculated.
- **Trim vectors:** Screens the sequences against UniVec or your own custom BLAST database to locate any vector contamination and trim it. This uses an implementation similar to NCBI's VecScreen to detect contamination (http://www.ncbi.nlm.nih.gov/projects/VecScreen/). Multiple databases can be selected to trim from by clicking the **+** sign.
- **Trim primers:** Screens the sequences against primers in your local database.
- **Error Probability Limit:** Available for chromatogram documents which have quality (confidence) values. The ends are trimmed using the modified-Mott algorithm (see below) based on these quality values (Richard Mott personal communication).
- **Maximum low quality bases:** Specifies the maximum number of low quality bases that can be in the untrimmed region. Low Quality is normally defined as confidence of 20 or less. This can be adjusted on the Sequencing and Assembly tab of Preferences.
- **Maximum Ambiguities:** Finds the longest region in the sequence with no more N's than the maximum ambiguous bases value and trims what is not in this region. This should be used when sequences have no quality information attached.
- **Trim 5′ End and Trim 3′ End:** These can be set to specify trimming of only the 3′ or 5′ end of the sequence. A minimum amount that must be trimmed from each end can also be specified.
- **Maximum length after trim:** If the untrimmed region is longer than the specified limit then the remainder will be trimmed from the 3′ end of the sequence until it is this length.

The Modified Mott algorithm

The modified-Mott algorithm for trimming ends based on quality operates as follows:

For each base, it subtracts the base error probability from an error probability cutoff value (default 0.05) to form the base score. The base error probability is calculated from the quality score (Q), such that P(error)=$10^{(Q/-10)}$. This means that low quality bases have high error probabilities and thus may have a negative base score.

E.g. For Q10, P(error)= 0.1, For Q30, P(error)=0.001

So with an error probability cutoff of 0.05, a base with Q10 has a base score of 0.05-0.1= -0.05, and a base with Q30 would have a base score of 0.05-0.001=0.049.

The trimming algorithm then calculates the running sum of the base score across the sequence. If the sum drops below zero it is set to zero. The part of the sequence not trimmed is the region between the first positive value of the running sum and the highest value of the running sum (i.e. the highest scoring segment of the sequence). Everything before and after this region is trimmed.

10.1.2 Removing duplicate reads

To remove duplicate reads from NGS datasets, use **Remove Duplicate Reads...** under the **Sequence** menu. This function runs Dedupe, and will remove duplicate sequences that are either exact matches, subsequences, or sequences within some percent identity. It can also find overlapping sequences and group them into clusters. For a detailed explanation of any Dedupe setting, hover the mouse over the setting, or click the help (question mark) button next to the custom options under **More Options**.

10.1.3 Error correction and normalization of reads

Prior to de novo assembly, it can sometimes be useful to error correct the data or to normalize coverage by discarding reads in regions of high coverage. This functionality is available using **Error Correct & Normalize Reads...** from the **Sequence** menu.

This function uses BBNorm from the BBtools suite, and it requires Java 7 or later in order to run. For a detailed explanation of any BBNorm setting, hover the mouse over the setting, or click the help (question mark) button next to the custom options under **More Options**. For more information, see the BBNorm page on SeqAnswers.

10.1.4 Setting paired reads

To assemble paired read (or mate pair) data, prior to assembly you first need to tell Geneious the reads are paired. The assembler will then automatically used the paired data unless you turn off the advanced option to **Use paired distances**.

To set up paired reads, you need to select the document(s) containing the paired reads and select **Set Paired Reads** from the **Sequence** menu. Depending on your data source, reads could be in parallel sets of sequences, or interlaced, so you need to tell Geneious which format. Geneious will guess and select the appropriate option based on the data you have selected, so most of the time you can just use the default value for this. However, you must make sure you select the correct **Relative Orientation** for your data. Different sequencing technologies orientate their paired reads differently. All paired read data will have a known expected distance between

each pair. It is important you set this to the correct value to achieve good results when assembling. If you don't know what the relative orientation or expected distance is between the reads you should ask your sequencing data provider.

When you click 'OK', if you chose to pair by parallel lists of sequences, Geneious will create a new document containing the paired reads. If you chose to pair an interlaced list of sequences (or modify settings for some already paired data), Geneious will just modify the existing list of sequences to mark it as paired.

If you choose to split reads based on the presence of a linker sequence (e.g. for 454 data) the original sequences will be unmodified and the split reads will be created in a new document. The default behaviour is to ignore sequences shorter than 4 bp either side of the linker, but this can be customized from the **Edit Linkers** option in the paired reads options.

Polonator sequencing machine reads can be split using the **Split each read in half** option.

10.1.5 Merging paired reads

If paired reads were sequenced with an insert size shorter than twice the read length then pairs may overlap with each other, in which case it can be useful merge to each pair into a single longer read. After setting up paired reads (see section 10.1.4), use **Merge Paired Reads...** from the **Sequence** menu to merge them.

This function uses BBMerge from the BBtools suite. For a detailed explanation of any BBMerge setting, hover the mouse over the setting, or click the help (question mark) button next to the custom options under **More Options**.

Alternatively, an experimental plugin for merging paired reads using FLASH is available from the Geneious website plugins page.

10.1.6 Splitting multiplex/barcode data

Multiplex or barcode data (e.g. 454 MID data) can be separated using **Separate Reads by Barcode** from the **Sequence** menu. This function copies all sequences matching a given barcode to a correspondingly named sequence list document.

Default settings are provided for 454 standard and Titanium MID barcodes (with or without Adaptor B trimming), and Rapid MID barcodes. These settings recognise standard MID sequences provided by 454 and use their names when appropriate.

To enter a custom barcode set, select **Custom** settings, then under **Barcode set** choose **Edit barcode sets**. Then click **Add** to add your list of barcodes. To specify fixed sequences either side of the barcode, enter these in the **Adaptor** and **Linker** sections.

If you only want to extract sequences with a single, specific barcode sequence (e.g. a primer),

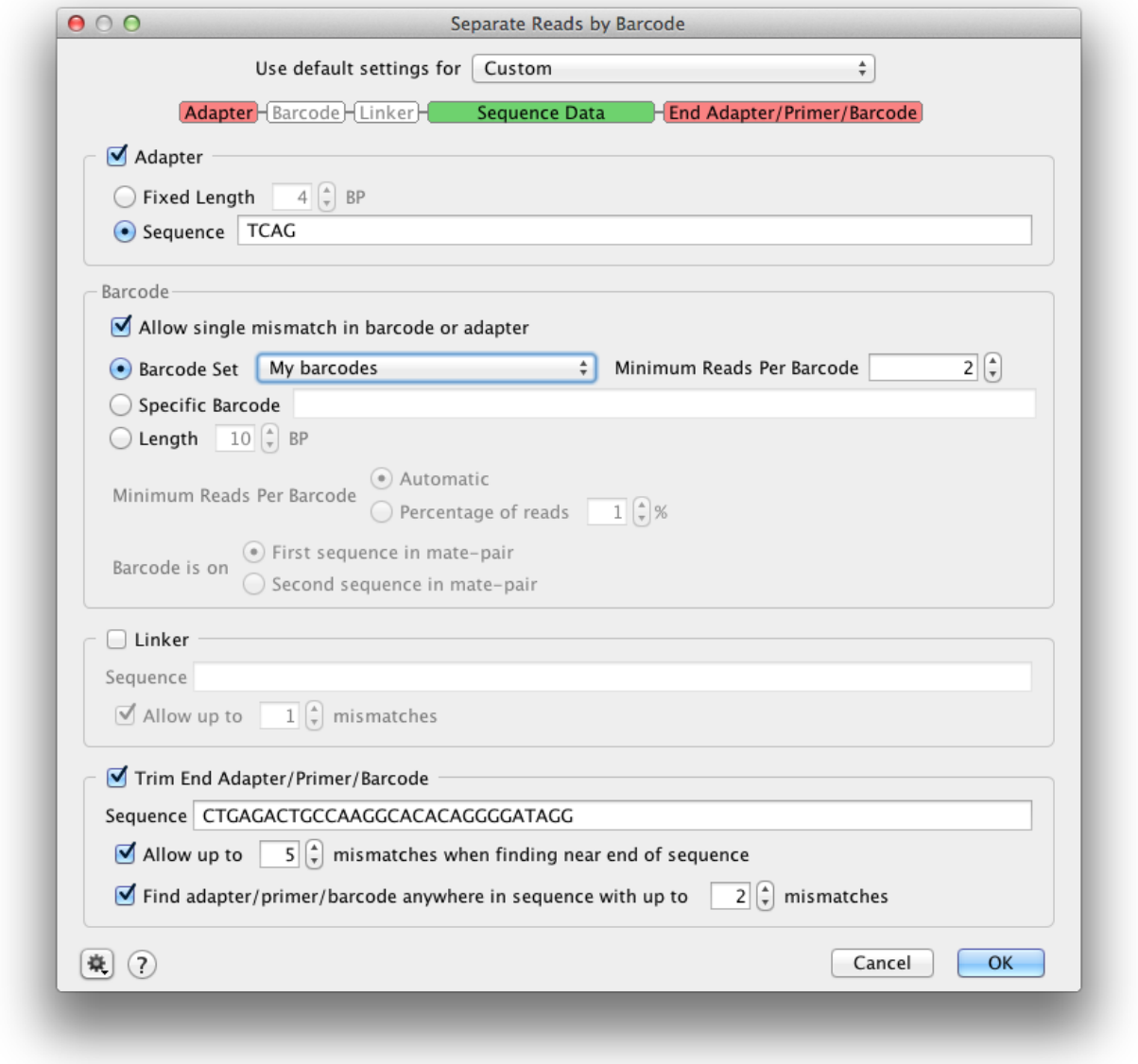


Figure 10.3: Options for separating reads by barcodes, using a custom barcode set, a fixed adaptor and a fixed end primer

check **Specific barcode** and enter the sequence. Alternatively if you do not know your barcode sequences, you can just enter the length of your barcodes and Geneious will automatically identify what the barcodes are.

Separate reads by barcodes only sorts by barcodes at the 5‘ end of the sequence, but primers, adapters or barcodes on the 3‘ end of the sequence can be trimmed off by checking **Trim End Adaptor/Primer/Barcode**. Either enter a specific sequence to trim, or add your 3‘ barcode to your custom barcode set and use the text $[END\_BARCODE]$ in the sequence box. Primer trimming can also be performed after separating by barcodes using **Trim Ends** (see section 10.1.1).

For further information on splitting barcode data, hover the mouse over any of the settings in the **Separate Reads by Barcode** options window.

10.2 De novo assembly

This can be used to assemble a small number of Sanger sequencing reads (i.e. forward and reverse reads of the same sequence), or millions of reads generated by next-generation sequencing machines. To assemble a contig firstly select all of the sequences and/or contigs you wish to assemble in the document table then click **Align/Assemble** in the toolbar and choose **De Novo Assemble.** The basic options for de novo assembly will then be displayed.

The options available here are as follows:

- **Assemble by (aka Assemble by Name):** If you have selected several groups of fragments which are to be assembled separately, you can specify a delimiter and an index at which the identifier can be found in all of the names. Sequences are grouped according to the identifier and each group is assembled separately. If a reference sequence is specified, it is used for all groups. eg. For the names A03.1.ab1, A03.2.ab1, B05.1.ab1, B05.2.ab1 etc where “A03” and “B05” are the identifiers you would choose “Assemble by 1st part of name, separated by . (full stop)”
- **Use % of data**: This option is will show with large datasets and enables you to assemble a subset of your data, rather than the full dataset. For example, if you enter 20% here, then the first 20% of reads in a sequence list will be assembled and the rest will be ignored. This is useful in situations where the full dataset is too large for the size of genome being assembled.
- **Assembly method:** In this section you can choose from the built-in Geneious assembler, or Tadpole, Velvet, MIRA and CAP3 assemblers if you have these plugins installed. Click the question mark button next to the method to see a list of the advantages and disadvantages of each assembler. The sensitivity setting specifies a trade off between the time it takes to assemble and the accuracy of the assembly. Higher sensitivity is likely to result in more reads being assembled.

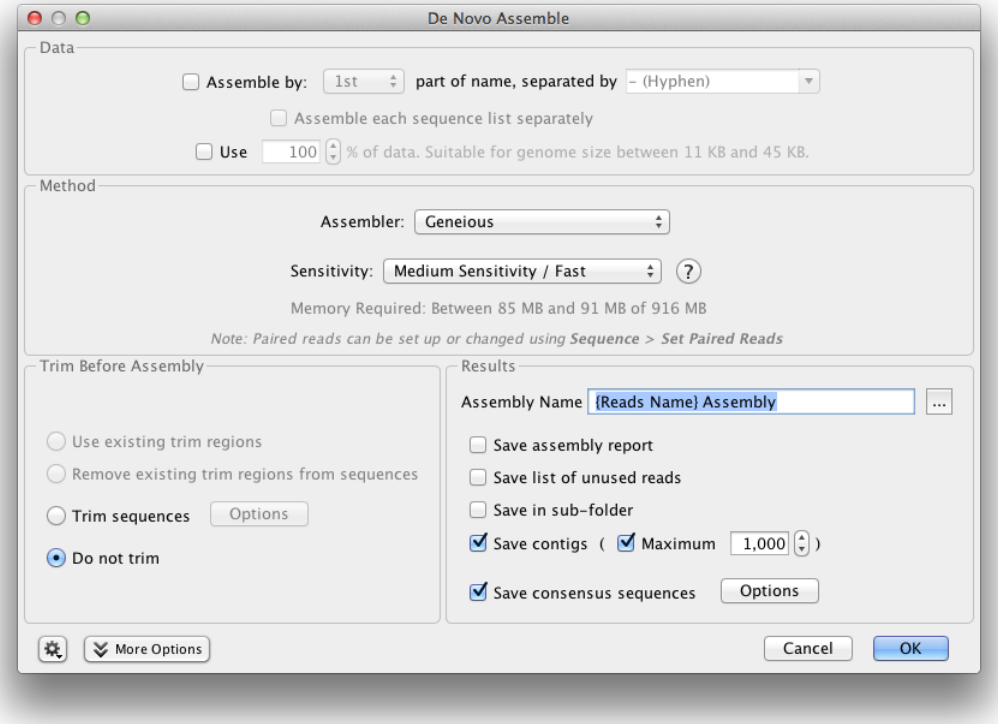


Figure 10.4: Basic de novo assembly options

- **Trim Sequences:** Select how to trim the ends of the sequences being assembled. See section 10.1.1.
- **Results:** Allow you to choose an assembly name and what to return in your results. By default, only the assembled contigs are saved, but you can also choose to return an assembly report, lists of used or unused reads and the consensus sequences. The assembly report summarises the assembly statistics and lists which fragments were successfully assembled and which contig they went in to along with a list of unassembled fragments. If Save in Subfolder is selected all the results of the assembly will be saved to a new subfolder inside the one containing the fragments. This folder will always only contain the assembly results from the one most recent assembly - it creates a new folder each time it is run.
- **More Options:** Under the advanced options you can change the parameters used by Geneious when aligning fragments together. These are fully documented if you hover the mouse over them in Geneious. To edit these settings, you must first choose**Custom Sensitivity** in the assembly method panel. For sequences which are lower quality or contain many errors, or are expected to be divergent from one another, you may need to decrease the minimum overlap identity and maximum mismatches per read, and increase the maximum gaps allowed per read.

Choose the options you require and click 'OK' to begin assembling the contig. Once complete, one or more contigs may be generated. If you got more contigs than you expect to get for the selected sequences then you should try adjusting the options for assembly. It is also possible that no contigs will be generated if no two of the selected sequences meet the overlap requirements.

Note: The orientation of fragments will be determined automatically, and they will be reverse complemented where necessary.

If you already have a contig and you want to add a sequence to it or join it to another contig then just select the contig and the contig/sequence and click de novo assembly as normal.

Scaffolding

Scaffolds are contigs which are linked together, with the missing regions between them filled by Ns. The size of the missing region is based on paired read distances. The Geneious assembler will produce scaffolds if this option is turned on under **More options**. If this setting is disabled it is because your data does not have paired reads or you haven't marked the data as paired using **Set Paired Reads** from the Sequence menu.

Unlike some assemblers where scaffolding is performed after contig formation, Geneious scaffolding is integrated into the contig assembly process. When there is strong support for scaffolding, it may take precedence over potentially conflicting standard contig formation. For this reason, Geneious can't be configured to produce both scaffolds and non-scaffolds from a single run.

De novo assembly of circular genomes

The Geneious de novo assembler can produce a circular contig if the ends of a contig match. To enable this option, click the **More Options** button and check **Circularize contigs with matching ends**. Circularization requires that pair of reads on each end of the contig match and that those reads don't intersect with each other anywhere else in the contig (i.e. pairwise contigs will never circularise).

A circular contig will contain reads at either end marked with arrows, which denotes that these reads span the origin and link back around to the other end of the assembly. The consensus sequence produced from this contig will also be circular. The Topology column in the Document table lists whether a given contig is circular or linear.

10.2.1 The de novo assembly algorithm

The sequence assembler in Geneious is flexible enough to handle read errors consisting of either incorrect bases or short indels. It can handle data from any type of sequencing machine with reads of any length, including paired-reads and mixtures of reads from different sequencing machines (hybrid assemblies).

The de novo assembly algorithm used is a greedy algorithm which is similar to that used in multiple sequence alignment.

1. For each sequence a blast-like algorithm is used to find the closest matching sequence among all other sequences.
2. The highest scoring sequence and its closest matching sequence are merged together into a contig (reverse complementing if necessary). This process is repeated, appending sequences to contigs and joining contigs where necessary.
3. For paired read de novo assembly, 2 sequences with similar expected mate distances are given a higher matching score if their mates also score well against each other. Similarly a sequence and its mate will be given a higher score if they both align at approximately their expected distance apart to an already formed contig. The effect of this heuristic is that paired read de novo assembly starts out by finding 2 sets of paired reads and forming 2 contigs. Each of these 2 contigs will contain 1 sequence from each pair and the 2 contigs are expected to be separated by the expected mate distance. Assembly proceeds from there either adding new paired reads to the contigs or forming new pairs of contigs which eventually merge together. Due to the nature of this algorithm, paired read de novo assembly in Geneious only works well if you have high coverage of paired reads - a hybrid assembly of mostly unpaired data with a few paired reads will not make good use of the paired read data, but this is expected to improve in future versions.
4. Each contig generated by a gapped de novo assembly has some minor fine tuning performed on it both during assembly and upon completion. For each gapped position in a

sequence, a base adjacent to the gap is shuffled along into the gap if it is the same base as the most common base in other sequences in the contig at that position. After doing this if any column now consists entirely of gaps that column is removed from the contig

5. Other minor heuristics are applied throughout the assembly to improve the results.

6. Both the Geneious de novo and reference assemblers use a deterministic method (even when spreading the work cross multiple CPUs) such that if you rerun the assembler using the same settings and same input data it will always produce the same results.

10.3 Map to reference

Assembling to reference is used when you have known sequence and you wish to compare a number of reads of the same sequence with it to locate differences or SNPs. To perform assembly to a reference sequence select the sequences and the reference sequence and click **Align/Assemble** and choose **Map to Reference**. Choose the name of the sequence you wish to use as the reference in the Reference Sequence chooser and click **OK**. One contig will be produced per reference and this will display the reference sequence at the top of the alignment view with all other sequences below it.

The options available in the Map to Reference setup dialog are similar to those for de novo assembly (see section 10.2), with a few differences detailed in the subsequent sections. In the Methods panel, you can choose between the standard Geneious assembler, Geneious for RNAseq (R9+), or the BBMap, Bowtie or Tophat mappers if you have these plugins installed.

See chapter 11 for details on identifying differences or SNPs in your assembly.

10.3.1 Choosing reference sequences

In version 8.1 and above, it is not necessary to select the reference sequence prior to choosing Map to Reference. Instead, the reference sequence can be chosen from within the Map to Reference setup options by clicking **Choose...**. This brings up a document chooser from which you can select single or multiple reference sequences from any folder in your database.

Multiple reference sequences may be selected either by choosing a sequence list containing your reference sequences, or by selecting multiple sequence documents using the **Choose...** button. Single reference sequences may be pre-selected prior to choosing the Map to reference... option, however when using the Choose... button, only those documents selected within the Choose... dialog will be used as reference sequences (all other sequences will be mapped).

If multiple reference sequences are selected each read will be mapped to the sequence with the best match only, and will produce one contig per reference. Batch assemblies where each read

gets mapped to each reference sequence can be done by using **Workflows → Map reads to each reference sequence**.

10.3.2 Fine tuning

When aligning to reference the sequences are not aligned to each other, each of them is instead aligned to the reference sequence independently and the pairwise alignments are combined into a contig. However, an iterative fine tuning step can be enabled, which makes reads that overlap from the initial assembly stage align better to each other. Fine tuning causes reads to align better to each other around indels which improves the accuracy of consensus and variant calling. For more information, click the help (question mark) button next to the fine tuning options in the Map to Reference setup dialog.

If you just wish to use a reference sequence to help construction of the contig where the reads extend beyond the length of the reference then you have two options. With iterative fine tuning, reads can extend a bit further past the ends of the reference sequence on each iteration so make sure you set the number of iterations high enough. Or you could select all sequences including the reference and use the De Novo assembler.

10.3.3 Deletion and structural variant discovery (DNA mapping)

Geneious can discover structural rearrangements and arbitrarily large deletions from paired or unpaired reads by analyzing how fragments of each read align to different regions of the reference sequence(s). To enable this option, check **Find structural variants and deletions of any size**. If you only want to find deletions up to a specified size, check **Find large deletions up to...**

For this operation, Geneious makes two passes during mapping. On the first pass each read mapped will generate candidate junctions (sites for structural variants) based on where fragments of the read align to different regions of the reference sequence(s). The more reads that support a candidate junction, the more likely it will be used during the second pass. The second pass involves mapping reads using the discovered junctions. By default, at least 2 reads must support the discovery of a junction in order for it to be used during the next pass. This threshold can be adjusted under **More Options** by changing the **Minimum support for structural variant discovery** setting.

Junctions used during the second mapping pass are annotated on the reference sequence under a track named after the reads. Each junction annotation has the following properties:

- **Junction Type:** This will be *Deletion* for deletions up to 1000 bp. For longer deletions or structural variants, this will be *Rearrangement*, with *(inversion)* potentially appended.
- **Intervals:** Each junction consists of two annotation intervals of a single nucleotide in

length which cover the final nucleotide used in the reference sequence before the read jumps and continues elsewhere. For junctions of type *Deletion*, the junction is represented as a single annotation with two linked intervals. For junctions of type *Rearrangement* the junction site is split into two separate annotations, each with a jagged edge on one side of the interval to indicate the side which jumps elsewhere.

- **Deletion Size:** This is present when *Junction Type* is *Deletion*.
- **Reads supporting discovery:** Indicates the number of reads that supported discovery of this junction during the first pass. This may be lower than the advanced minimum support setting in cases where other reads supported discovery of a slightly offset version of this junction, which allows this junction to be retained on the next pass.
- **Reads using:** Indicates the number of reads that used this junction as part of their mapping during the second pass.
- **Junction Source & Junction Destination:** Clickable links to the junction positions in the reference sequence. When the destination is a different reference sequence, this is prefixed with the sequence name followed by a colon.
- **Color:** Annotations are colored from blue to green based on increasing values of *Reads supporting discovery*.

Reads spanning junctions may be represented in one of two possible ways. For deletions under 1,000 bp, the deletion is represented as a gap in the read. For longer deletions or for structural variants, two copies of the read appear in the contig where the fragment of the read extending past the junction is marked as trimmed.

10.3.4 RNAseq mapping

To map RNA sequence reads to a genome with introns, choose **Geneious for RNAseq** as the **Mapper** in the Map to Reference setup dialog. This function can map reads that span existing annotated introns, or discover novel introns and fusion genes.

This function works in the same way as **deletion and structural variant discovery** (section 10.3.3) for DNA mapping, by analyzing how fragments of each read align to different regions of the reference sequence(s), and creating a junction annotation at the point where the read is split. By default, at least 2 reads must support the discovery of a junction in order for it to be annotated. This threshold can be adjusted under **More Options** by changing the **Minimum support for intron/fusion gene discovery** setting.

If **Span annotated mRNA introns** is checked, junctions will be created from existing annotations on the reference sequence. Reads are still allowed to map anywhere, but will be allowed to freely span these junctions if that produces the best mapping.

To only find introns up to a certain size, check **Find novel introns up to...**; to find introns of any size, or structural rearrangements that may indicate a fusion gene, use **Find fusion genes and novel introns**.

As for deletion and structural variant discovery, junctions are annotated on the reference sequence under a track named after the reads. Each junction has the following properties:

- **Junction Type:** For introns under 2,000,000 bp, this will be *Intron*. For longer introns or structural variants, this will be *Fusion*, with *(inversion)* potentially appended.
- **Intervals:** Each junction consists of two annotation intervals of a single nucleotide in length which cover the final nucleotide used in the reference sequence before the read jumps and continues elsewhere. For junctions of type *Intron*, the junction is represented as a single annotation with two linked intervals. For introns that have common start and finish nucleotides, the intervals will be assigned an appropriate direction. For junctions of type *Fusion* the junction site is split into two separate annotations, each with a jagged edge on one side of the interval to indicate the side which jumps elsewhere.
- **Intron Size:** This is present when *Junction Type* is *Intron*.
- **Reads supporting discovery:** Indicates the number of reads that supported discovery of this junction during the first pass. This may be lower than the advanced minimum support setting in cases where other reads supported discovery of a slightly offset version of this junction, which allows this junction to be retained on the next pass.
- **Reads using:** Indicates the number of reads that used this junction as part of their mapping during the second pass.
- **Junction Source & Junction Destination:** Clickable links to the junction positions in the reference sequence. When the destination is a different reference sequence, this is prefixed with the sequence name followed by a colon.
- **Color:** Annotations are colored from blue to green based on increasing values of *Reads supporting discovery*.

Reads spanning junctions may be represented in one of two possible ways. For introns under 50 bp, the intron is represented as a gap in the read. For longer introns or for fusion genes, two copies of the read appear in the contig where the fragment of the read extending past the junction is marked as trimmed.

10.3.5 The map to reference algorithm

The reference assembly algorithm used is a seed and expand style mapper followed by an optional fine tuning step to better align reads around indels to each other rather than the reference sequence. Various optimizations and heuristics are applied at each stage, but a general outline of the algorithm is

1. First the reference sequence(s) is indexed to create a table making a record of all locations in the reference sequence that every possible word (series of bases of a specified length) occurs.
2. Each read is processed one at a time. Each word within that read is located in the reference sequence and that is used as a seed point where the matching range is later expanded outwards to the end of the read.
3. If a read does not find a perfectly matching seed, the assembler can optionally look for all seeds that differ by a single nucleotide.
4. Before the seed expansion step, all seeds for a single read that lie on the same diagonal are filtered down to a single seed.
5. During seed expansion, when mismatches occur a look-ahead is used decide whether to accept it as a mismatch or to introduce a gap (in either the reference sequence or read)
6. The mapper handles circular reference sequences by indexing reference sequence words spanning the origin and allowing the expansion step to wrap past the ends
7. All results are given a score based on the number of mismatches and gaps introduced. Normally the best scoring (or a random one of equally best scoring) matches are saved although there is an option to map the read to all best scoring locations
8. Paired reads are given an additional score penalty based on their distance from their expected distance so that they prefer mapping close to their expected distance with as few mismatches as possible, but they can also map any distance apart if an ideal location is not found.
9. The final optional fine tuning step at the end, shuffles the gaps around so that they reads better align to each other rather than the reference sequence.
10. For details on how mapping qualities are calculated, see See section 10.4.

For further details and for a comparison of the Geneious reference assembler to other software, see the Geneious Mapper white paper.

10.4 Viewing Contigs

Contigs in Geneious are viewed (and edited) in exactly the same way as alignments. There are several features in the sequence viewer which are worth taking special note of when viewing contigs:

- The consensus sequence is normally of particular interest and this is always displayed at the top of the sequence view (labeled Consensus).

- When all sequences in a contig (or alignment) have quality information attached then you can select the **Highest Quality** consensus type. This almost removes the need for manually editing the contig because this consensus chooses the base with the highest total quality at each position. See section 9.4 for more information on how this is calculated.
- There is a **Base Call Quality** color scheme which is selected by default for alignments of all chromatograms. This assigns a shade of blue to each base based on its quality. Dark blue for confidence < 20, blue for 20 - 40 and light blue for > 40. The consensus is also colored with this scheme where the confidence of a given base in the consensus is equal to the maximum confidence from the bases at that site in the alignment.
- There is a **Mapping Quality** color scheme for reads mapped to a reference sequence. A mapping quality represents the confidence that the read has been mapped to the correct location. For a read with mapping quality Q, the probability that it has been incorrectly mapped is $10^{(-Q/10)}$. For example, a read with a mapping quality score of 20 has a 1% chance of having been incorrectly mapped. Reads that could be mapped to multiple locations will have a maximum mapping quality score of 3, which indicates it had at least a 50% probability of mapping elsewhere. Mapping qualities have a maximum value of 254 for consistency with the SAM/BAM format. If a sequence has no mapping quality (i.e the document was produced in a version of Geneious prior to 8.1 or imported from a SAM/BAM file that didn't have mapping quality) then it will be colored gray. Mapping quality for the sequence under the mouse is also displayed in the status bar. All mappers use heuristics to calculate mapping qualities. For unpaired reads, the Geneious mapper assigns a mapping quality of 20*(the number of additional mismatches in the second best location the read maps to). For paired reads the individual unpaired mapping qualities are calculated, but these are increased by up to 20 depending on how close the best pair is to the expected insert distance compared with the second best pair.
- The sequence logo graph has an option to "Weight by quality". This is very useful for identifying low quality regions and resolving conflicts.

Finding regions of low/high coverage

In addition to the coverage graph which gives you a quick overview of coverage, under the **Annotate & Predict** toolbar is the **Find Low/High Coverage** feature. This feature annotates all regions of low/high coverage which you can then navigate through using the little left and right arrows next to the coverage annotations in the controls on the right. You can set the threshold low/high coverage by either specifying an absolute number of sequences or a number of standard deviations from the mean coverage.

The find low/high coverage tool can also be used to record the minimum, mean, and maximum coverage of each annotation of a particular type on the reference sequence. To do this, in the **Only Find In** section of the options, turn on **Annotations in reference sequence of type** and choose **Create annotations of same type on reference sequence**.

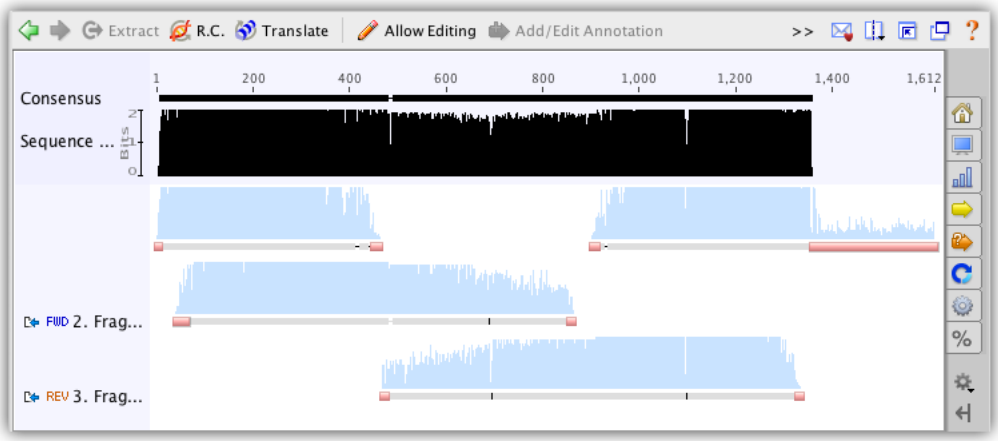


Figure 10.5: The overview of a contig

Viewing Contigs of Paired Reads

In order to view a contig of paired reads, you first need to have set up the paired data before assembling - see 10.1.4. Once you have your paired read assembly, the contig viewer adds an option to **Link paired reads** in the advanced section of the controls on the right. This means that pairs of reads will be laid out in the same row with a horizontal line connecting them. Reads separated by more than 3 times their expected distance are not linked by default unless the **Link distant reads** setting is turned on.

The horizontal line between paired reads is colored according to how close the separation between the reads is to their expected separation. Green indicates they are correct, yellow and blue indicate under or over their expected separation and red indicates the reads are incorrectly orientated.

The reads themselves can also be configured to be colored in this way if you use the **Paired Distance** color scheme from the general (top section in the controls on the right) settings. The colors used and the sensitivity for deciding if reads are close enough to their expected distance can be configured from the **Options** link when the Paired Distance color scheme is selected.

You can hover the mouse of any read in a contig and the status bar will indicate the expect separation and expected separation between the reads.

10.5 Editing Contigs

Editing a contig is exactly the same as editing an alignment in Geneious. After selecting the contig, click the **Allow Editing** button in the sequence viewer and you can modify, insert and delete characters like in a standard text editor.

Editing of contigs is done to resolve conflicts between fragments before saving the final consensus. The normal procedure for this is to look through the disagreements in the contig (as described above) and change bases which you believe are bad calls to be the base which you believe is the correct call. This is often decided by looking at the quality for each of the bases and choosing the higher quality one. Geneious can do this automatically for you if you use the **Highest Quality** consensus.

Bases in the consensus sequence can also be edited which will update every sequence at the corresponding position to match what is set in the consensus.

You can also manually move a read mapped to a reference sequence to a specific position in the contig. To do this, select the read and right-click, then choose **Move read to position..**, and enter the position where you want the left-most base in the read to sit.

Figure 10.6: Highlight disagreements and edit to resolve them

10.6 Extracting the Consensus

Once you are satisfied with a contig you can save the consensus as a new sequence by clicking on the name of the consensus sequence in your contig and clicking the **Extract** button. You can also generate consensus sequences for single or multiple contig documents by selecting the documents and going to **Tools** → **Generate Consensus Sequence**.

Chapter 11

Analysis of assemblies and alignments

11.1 Finding polymorphisms

To easily identify bases which do not match the consensus, turn on **Highlighting** in the consensus section of the sequence viewer options. Select the options **Disagreements** to **Consensus** (this is the default setting for nucleotide alignments). When this is on, any base in the sequences which matches the consensus at that position is grayed out and bases not matching are left colored.

With this on you can quickly jump to each disagreement by pressing Ctrl+D (command+D on Mac OS X) or by clicking the **Next Disagreement** button in the sequence viewer option panel to the right. Each disagreement can then be examined or resolved.

You can also use this feature if you have aligned to a reference sequence and you are interested in finding differences between each sequence and the reference (or SNPs).

11.1.1 Find Variations/SNPs

Manually investigating every little disagreement can be time consuming on larger contigs. The **Find Variations/SNPs** feature from the **Annotate & Predict** menu will annotate regions of disagreement and can be configured to only find disagreements above a minimum threshold to screen out disagreements due to read errors. This feature can also be configured to only find disagreements in coding regions (if the reference sequence has CDS annotations present) and can analyze the effects of variations on the protein translation to allow you to quickly identify silent or non-silent mutations. It can also calculate p-values for variations and filter only for variations with a specified maximum P-Value.

For full details of how the various settings work in the Variation/SNP finder, hover the mouse over them in Geneious to read the tooltips or click one of the '?' buttons.

P-values

The p-value represents the probability of a sequencing error resulting in observing bases with at least the given sum of qualities. The lower the p-value, the more likely the variation at the given position represents an real variant. Click the down arrow next to the exponent of the **Maximum Variant P-Value** setting to increase the number of variants found.

When calculating P-Values:

- The contig is assumed to have been fine tuned around indels
- Ambiguity characters are ignored (other characters in the column are still used)
- Homopolymer region qualities are reduced to be symmetrical across the homopolymer. For example if a series of 6 G's have quality values 37, 31, 23, 15, 7, 2 then these are treated as though they are 2, 7, 15, 15, 7, 2. This is done because variations may be called at either end of the homopolymer and because reads may be from different strands.
- Gaps are assumed to have a quality equal to the minimum quality on either side of them (after adjusting for homopolymers)
- When finding variations relative to a reference sequence, the p-value calculated is for the variant, not the change. In other words the p-values calculated are independent of the reference sequence data.
- The approximate p-value method calculates the p-value by first averaging the qualities of each base equal to the proposed SNP and averaging the qualities of each base not equal to the proposed SNP.
- Example: Assume you have a column where the reference sequence is an A and there are 3 reads covering that position.

  1 read contains an A in the column and the other 2 reads contain a G. All 3 reads have quality 20 (= 99% confidence) at this position. We want to calculate the p-value for calling a G SNP in this column.

  Since the quality values are all equal, the p-value is the probability of seeing at least 2 G's if there isn't really a variant here. In other words, the probability of seeing 2 G's by chance due to a sequencing error plus the probability of seeing 3 G's by chance due to a sequencing error, which is calculated using the binomial distribution: ${}^3C_2 * 0.01^2 * 0.99 + {}^3C_3 * 0.01^3$ (${}^NC_K$ is a binomial coefficient)

False SNPs due to strand-bias (when sequencing errors tend to occur only on reads in a single direction) can be eliminated by specifying a value for the **Minimum Strand-Bias P-value** setting. A **Strand-Bias P-Value** property is added to each SNP to indicate the probability of seeing a strand bias at least this extreme assuming that there is no strand bias. SNPs with a smaller strand bias p-value will be excluded from the results when using this setting.

Click the up arrow next to the exponent of the **Minimum Strand-Bias P-Value** setting to increase the number of variants found. If there are any forward/reverse or reverse/forward style paired reads, then variants with strand bias which are less than 1.5 times the insert size from either end of the contig will not be filtered out.

Results display

The results of the Variant/SNP finder are added to the reference sequence in the assembly or alignment as an annotation track. Clicking **Save** and clicking **"Yes"** when prompted to **apply the changes to the original sequences** will add this annotation track onto the original reference sequence file. If there is no reference sequence for the alignment or assembly the annotations are added to the consensus sequence.

The results are also displayed in the annotations table and the following columns can be displayed:

- **Change**: Indicates the reference sequence nucleotides followed by the variant nucleotides. For example 'C → A'
- **Coverage**: The number of reads that cover the SNP region in the contig. The coverage includes both the reads containing the SNP and other reads at that position.
- **Reference Frequency**: The percentage of reads that agree with the reference sequence at that position. This field will only be present if at least 1 read agrees with the reference sequence.
- **Variant Frequency**: The percentage of reads that have the variation at that position. For variations that span more than a single nucleotide, the variant frequency may appear as a range (e.g. 47.8% – 51.7%) to indicate the minimum/maximum variant frequency over that range.
- **Polymorphism Type**: This may be one of the following.

  SNP (Transition): a single nucleotide transition change from the reference sequence

  SNP (Transversion): a single nucleotide transversion change from the reference sequence

  SNP: At a single position, there are multiple variations from the reference sequence

  Substitution: A change of 2 or more adjacent nucleotides from the reference sequence

  Insertion: 1 or more nucleotides inserted relative to the reference sequence

  Deletion: 1 or more nucleotides deleted relative to the reference sequence

  Mixture: multiple variations from the reference sequence which are not all the same length

For variations inside coding regions (CDS annotations) the following fields can be displayed:

- **Codon Change**: indicates the change in codon. Essentially this is the same as the 'Change' field, but extended to include the full codon(s). For example 'TTC → TTA'

- **Amino Acid Change**: indicates the change (if any) in the amino acid(s) by translating the codon change. For example 'F → L'

- **Protein Effect**: summarizes the change on the protein as either a substitution, frame shift, truncation (stop codon introduced) or extension (stop codon lost)

11.2 Analyzing Expression Levels

11.2.1 Calculating Expression Levels

The **Calculate Expression Levels** feature from the **Annotate & Predict** menu calculates normalised expression measures from mapped RNA-seq data. RPKM, FPKM and TPM are calculated for each CDS annotation on the reference sequence of a contig and the results are displayed as a heat map annotation track.

If you have multiple reference sequences for each sample (e.g. reads mapped to multiple chromosomes), all contigs from a single sample should be selected and run in a single step.

To calculate differential expression between samples you need to run **Calculate Expression Levels** for each sample separately and then compare the results using **Compare Expression Levels**.

Counting

The three metrics are calculated by normalizing the count of reads that map to each CDS annotation. If a read at least partially intersects at least one interval from a CDS annotation, then it will be treated as though that read mapped to that CDS annotation.

For reads that map to multiple locations, or reads that map to a location that intersect multiple CDS annotations, these may either be counted as partial matches, excluded from the calculations, or counted as full matches to each location they map to. For example if a read maps to two locations, then it will be counted as if 0.5 reads mapped to each of the two locations.

When calculating statistics, reads that don't map or map outside of an annotation CDS annotation are ignored.

RPKM

Reads per kilobase per million normalizes the raw count by transcript length and sequencing depth.

RPKM = (CDS read count * 10^9) / (CDS length * total mapped read count)

FPKM

Same as RPKM except if the data is paired then only one of the mates is counted, ie. fragments are counted rather than reads.

TPM

Transcripts per million (as proposed by Wagner et al 2012) is a modification of RPKM designed to be consistent across samples. It is normalized by total transcript count instead of read count in addition to average read length.

TPM = (CDS read count * mean read length * 10^6) / (CDS length * total transcript count)

Results

Results are displayed as an annotation track on the reference sequence. By default, annotations are colored based on the TPM property, ranging from blue for 0, through to white for the mean TPM, up to red for the highest TPM for any gene in the sample. In the results view, by clicking on the little down arrow to the left of the track's name, you can choose to color by a different property.

The values for RPKM, FPKM and TPM, as well as the raw read counts, are entered as properties on the annotation and can be displayed by mousing over an annotation. To export these values as a table, switch to the Annotations tab above the sequence viewer then click the **Track** button and choose the **Expression** track to display. Then click the Columns button and add the columns for FPKM, RPKM, TPM and/or the raw counts. Once you have the columns you need, you can export the table in .csv format by clicking **Export table**.

11.2.2 Comparing Expression Levels

Geneious is able to find differentially expressed genes between two samples. Comparison between more than two samples is not currently supported.

To compare expression levels, you must first run **Calculate Expression Levels** from the **Annotate & Predict** menu on each contig assembly of the two datasets. These assemblies must use the same reference sequence. This will save an **Expression Level** track for each contig on the reference sequence. To compare these tracks, select the reference sequence document and go to **Annotate & Predict** → **Compare Expression Levels**.

Like **Calculate Expression Levels**, which must be run on all contigs from a single sample at once to produce correct results in all cases, **Compare Expression Levels** should also be run on all reference sequences at once. However, **Compare Expression Levels** can optionally be run on just one reference sequence at a time when using a normalization system other than **Median of Gene Expression Ratios**.

Values to Compare

Either read counts, fragment counts, or transcript counts from each annotation can be compared. Since a single transcript can produce multiple reads and fragments, the number of reads and fragments produced aren't independent events so the confidence values produced by comparing these are unlikely to be accurate. For this reason we recommend comparing samples using transcript counts.

Normalization

Since different samples produce different quantities of transcripts, in order to compare values between samples, the counts need to be normalized using one of the following methods.

- **Total Count**: The counts in each gene are scaled according to the total number of transcripts mapped to all genes. For example, if one sample has twice as many transcripts mapped as the other sample, then the counts for each gene need to be halved to make them comparable with the other sample.
- **Median Expression**: The expression level of all expressed genes from the sample are calculated and the median values of these from each sample are used to normalize. For example, if one sample has a median twice as high as the other sample, then the counts for each gene need to be halved to make them comparable with the other sample.
- **Total Count Excluding Upper Quartile**: The expression level of all expressed genes from the sample are calculated and the total number of reads, fragments, or transcripts from the lowest 75% of those are totaled. Values are normalized between samples based on this total.
- **Median of Gene Expression Ratios**: For each gene the ratio of the expression level between samples is calculated. Then the median ratio across all expressed genes is used as

the normalization scale. This normalization method is the same as that implemented by DESeq.

All of these normalization methods (and more) are described and compared by Dillies et al 2012, and they recommend using 'Median of Gene Expression Ratios' rather than the other three normalization methods implemented here. One reason for this is that a few highly expressed genes can greatly affect the total number of transcripts produced, so this can distort the fraction of the total reads that contribute to genes with lower expression. The choice of normalization method determines the **Differential Expression Ratio** for each gene.

P-Value Calculation

In addition to calculating the differential expression ratio, it is useful to know whether or not that differential expression is statistically significant. This is represented by a p-value. A number of advanced methods have been published for the calculation of p-values based on a range of assumptions. Many of these are compared by Soneson & Delorenzi 2013 and they conclude that no single method is optimal under all circumstances and that very small samples sizes impose problems for all evaluated methods.

In this basic differential expression plugin in Geneious we have implemented a simple statistical test based on the assumption that the gene which each observed transcript came from is an independent event.

For a given gene, the probability that a randomly selected transcript would come from that gene is calculated as **number of transcripts mapped to that gene/total number of transcripts from that sample**. This probability is normalized, the mean probability between the two samples calculated, and this mean un-normalized for each sample. This produces an expected probability that a randomly selected transcript from this sample comes from that gene, assuming that this gene is not differentially expressed.

The Binomial Distribution is used to calculate the probability that an observed count at least as extreme as the observed one would be seen, assuming this non-differentially expressed mean probability. The probabilities from each sample are multiplied together to form the p-value.

Results

Results are displayed as an additional **Expression level** annotation track on the reference sequence. Each annotation produced will have the following properties:

- **Differential Expression p-value**: A p-value indicating the confidence that this gene is differentially expressed.

- **Differential Expression Absolute Confidence**: The negative base 10 log of the p-value. This field is useful for filtering on. For example in the **Filter** box on the results, you could type "absolute confidence>6" (without quotes) to only show genes that we are fairly confident have differential expression. Since filtering works on partial property names, in most cases just a shorter filter like "abs>6" is sufficient unless your genes have existing numeric properties containing the text "abs".
- **Differential Expression Confidence**: The negative base 10 log of the p-value, but adjusted to be negative for genes that are under expressed in sample 2 compared to sample 1, or positive for over expressed genes. The results are colored based on this property, from blue for under expressed genes, through to white for genes that are not differentially expressed, through to red for genes that are over expressed. Confidence coloring reaches maximum intensity at +-8.
- **Differential Expression Ratio**: The ratio of the normalized values between the two samples, but ratios less than one are replaced with -1/value. This results in a value greater than 1 for over expressed genes and less than -1 for under expressed genes. When one sample has no or very low expression, the ratio is capped at +/- 1,000,000. Ratio coloring reaches maximum intensity at +-5.
- **Differential Expression Log2 Ratio**: The base 2 log of the ratio of the normalized values between the two samples. When one sample has no or very low expression, the log2 ratio is capped at +/- 1,000,000.

Although results are colored by **Differential Expression Confidence** by default, you may prefer to switch to **Differential Expression Ratio** coloring once you have filtered at an appropriate confidence level (e.g using an "abs>6" filter). To do this click the arrow to the left of the track name and choose **Color by / Heatmap...**.

To export these values as a table, switch to the Annotations tab above the sequence viewer then click the **Track** button and choose the **Expression** track to display. Then click the Columns button and add the results columns you wish to display. You can export the table in .csv format by clicking **Export table**.

Chapter 12

Building Phylogenetic trees

Geneious provides inbuilt algorithms for Neighbour-joining (Saitou & Nei 1987) and UPGMA (Mitchener & Sokal 1957) methods of tree reconstruction, which are suitable for preliminary investigation of relationships between newly acquired sequences. For more sophisticated methods of phylogenetic reconstruction such as Maximum Likelihood and Bayesian MCMC, external plugins for specialist software are available (see section 12.3.4 for a full list). These can be downloaded from the plugins page on our website or within Geneious by going to**Plugins** under the **Tools** menu.

12.1 Phylogenetic tree representation

A phylogenetic tree describes the evolutionary relationships amongst a set of sequences. They have a few commonly associated terms that are depicted in Figure 12.1 and are described below.

Branch length. A measure of the amount of divergence between two nodes in the tree. Branch lengths are usually expressed in units of substitutions per site of the sequence alignment.

Nodes or internal nodes of a tree represent the inferred common ancestors of the sequences that are grouped under them.

Tips or leaves of a tree represent the sequences used to construct the tree.

Taxonomic units. These can be species, genes or individuals associated with the tips of the tree.

A phylogenetic tree can be rooted or unrooted. A rooted tree consists of a root, or the common ancestor for all the taxonomic units of the tree. An unrooted tree is one that does not show the position of the root. An unrooted tree can be rooted by adding an outgroup (a species that is distantly related to all the taxonomic units in the tree).

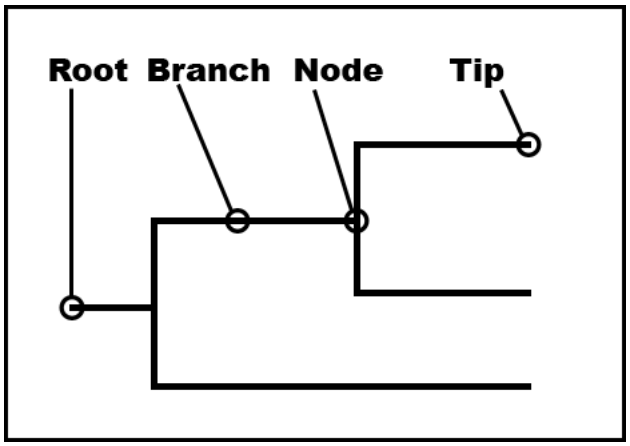


Figure 12.1: Phylogenetic tree terms

For information on viewing and formatting trees in Geneious, see Section 12.5.

12.2 Tree building in Geneious

To build a tree, select an alignment or a set of related sequences (all DNA or all protein) in the Document table and click the **Tree** icon or choose this option from the Tools menu.

If you are building a simple tree (Neigbour joining or UPGMA) using the Geneious tree builder, the tree can be built directly from a set of unaligned sequences, as the alignment will be built as part of the tree-building process. For more advanced trees, or if you wish to bootstrap your trees you must build an alignment first and use that as input for your tree. You can also select an existing tree document (which contains an alignment) and build another tree from that, as the alignment will simply be extracted from the existing tree and used build the new tree.

The following options are available in the tree-building dialog for the Geneious tree builder. For more information on these options see sections 12.3 to 12.4.

- **Genetic distance model**. This lets the user choose the kind of substitution model used to estimate branch lengths. If you are building a tree from DNA sequences you have the choices “Jukes Cantor”, “HKY” and “Tamura Nei”. If you are building a tree from amino acid sequences you only have the option of “Jukes Cantor” distance correction.
- **Tree building method**. There are two methods under this option – Neighbor joining and UPGMA.
- **Outgroup**. Choose which sample to use as an outgroup, or leave it as ”no outgroup” to build an unrooted tree.

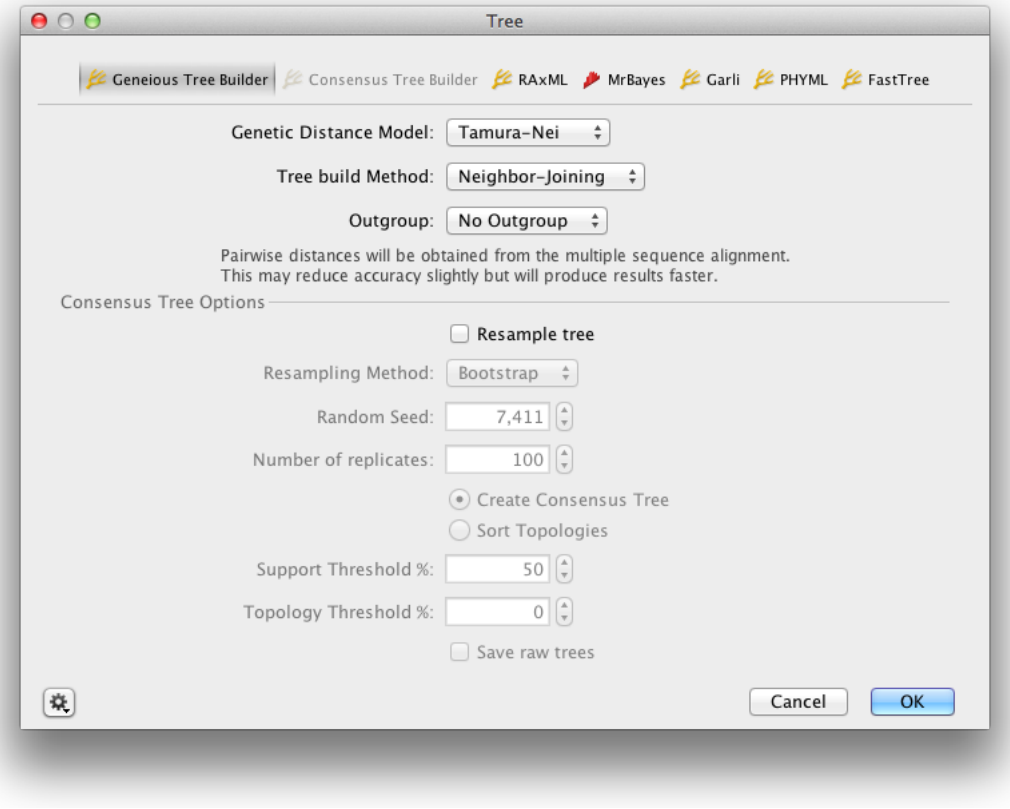


Figure 12.2: Tree building options in Geneious

- **Resample tree**. Check this to perform resampling.
- **Resampling method**. Either bootstrapping or jackknifing can be performed when resampling columns of the sequence alignment.
- **Number of samples**. The number of alignments and trees to generate while resampling. A value of at least 100 is recommended.
- **Create Consensus Tree**. Choose this to create a consensus tree from the resampled data.
- **Sort Topologies**. Produce trees which summarise the topologies resulting from resampling.
- **Support threshold**. This is used to decide which monophyletic clades to include in the consensus tree, after comparing all the trees in the original set.
- **Topology Threshold**. The percentage of topologies in the original trees which must be represented by the summarizing topologies.
- **Save raw trees**. If this is turned on then all of the trees created during resampling will be save in the resulting tree document. The number of raw trees saved will therefore be equal to the number of samples.

12.3 Tree building methods and models

12.3.1 Neighbor-joining

In this method, neighbors are defined as a pair of leaves with one node connecting them. The principle of this method is to find pairs of leaves that minimize the total branch length at each stage of clustering, starting with a star-like tree. The branch lengths and an unrooted tree topology can quickly be obtained by using this method without assuming a molecular clock (see Saitou & Nei 1987).

12.3.2 UPGMA

This clustering method is based on the assumption of a molecular clock. It is appropriate only for a quick and dirty analysis when a rooted tree is needed and the rate of evolution is does not vary much across the branches of the tree (see Mitchener & Sokal 1957).

12.3.3 Distance models or molecular evolution models for DNA sequences

The evolutionary distance between two DNA sequences can be determined under the assumption of a particular model of nucleotide substitution. The parameters of the substitution model

define a rate matrix that can be used to calculate the probability of evolving from one base to another in a given period of time. This section briefly discusses some of the substitution models available in Geneious. Most models are variations of two sets of parameters – the **equilibrium frequencies** and **relative substitution rates**.

Equilibrium frequencies refer to the background probability of each of the four bases A, C, G, T in the DNA sequences. This is represented as a vector of four probabilities $\pi_A, \pi_C, \pi_G, \pi_T$ that sum to 1.

Relative substitution rates define the rate at which each of the transitions ($A \leftrightarrow G$, $C \leftrightarrow T$) and transversions ($A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$, $G \leftrightarrow T$) occur in an evolving sequence. It is represented as a 4x4 matrix with rates for substitutions from every base to every other base.

Additionally, **gaps** are not penalized when using the Geneious Tree Builder. Comparisons involving any gaps are ignored when calculating the distance matrix.

Jukes Cantor

This is the simplest substitution model. It assumes that all bases have the same equilibrium base frequency, i.e. each nucleotide base occurs with a frequency of 25% in DNA sequences and each amino acid occurs with a frequency of 5% in protein sequences. This model also assumes that all nucleotide substitutions occur at equal rates and all amino acid replacements occur at equal rates (see Jukes and Cantor 1969).

HKY

The HKY model assumes every base has a different equilibrium base frequency, and also assumes that transitions evolve at a different rate to the transversions (see Hasegawa et al 1985).

Tamura-Nei

This model also assumes different equilibrium base frequencies. In addition to distinguishing between transitions and transversions, it also allows the two types of transitions ($A \leftrightarrow G$ and $C \leftrightarrow T$) to have different rates (see Tamura & Nei 1993).

12.3.4 Advanced Tree Building methods

The following plugins are available for running maximum likelihood or Bayesian phylogenetic analyses in Geneious:

- **MrBayes**: For Bayesian estimation of phylogenies, runs MrBayes 2.0.9 (http://mrbayes.sourceforge.net/)
- **PhyML**: Builds maximum likelihood trees using PhyML 3.0 (http://www.atgc-montpellier.fr/phyml)
- **GARLI (Genetic Algorithm for Rapid Likelihood Inference)**: Builds maximum likelihood trees from alignments of 4 or more sequences using Garli 2.0 (https://www.nescent.org/wg_garli/Main_Page).
- **RAxML (Randomized Axelerated Maximum Likelihood)**: Rapid maximum likelihood tree-building using RAxML 7.2.8 (http://www.exelixis-lab.org/). Allows partitioned datasets.
- **FastTree**: Approximately-maximum-likelihood phylogenetic trees from alignments of nucleotide or protein sequences using Fasttree 2.1.5 (http://www.microbesonline.org/fasttree/). Ideal for large alignments.
- **PAUP***: Builds maximum parsimony and maximum likelihood trees. You must own or purchase a copy of PAUP* to use this plugin (http://paup.csit.fsu.edu/)

These can be downloaded from the plugins page on our website or within Geneious by going to **Plugins** under the **Tools** menu. For more information on running these programs, please consult the user manual for the source software.

12.4 Resampling – Bootstrapping and jackknifing

Resampling is a statistical technique where a procedure (such as phylogenetic tree building) is repeated on a series of data sets generated by sampling from one original data set. The results of analyzing the sampled data sets are then combined to generate summary information about the original data set.

In the context of tree building, resampling involves generating a series of sequence alignments by sampling columns from the original sequence alignment. Each of these alignments (known as pseudoreplicates) is then used to build an individual phylogenetic tree. A consensus tree can then be constructed by combining information from the set of generated trees or the topologies that occur can be sorted by their frequency (see below).

Bootstrapping is the statistical method of resampling with replacement. To apply bootstrapping in the context of tree building, each pseudo-replicate is constructed by randomly sampling columns of the original alignment with replacement until an alignment of the same size is obtained (see Felsenstein 1985).

Jackknifing is a statistical method of numerical resampling based on deleting a portion of the original observations for each pseudo-replicate. A 50% jackknife randomly deletes half of the columns from the alignment to create each pseudo-replicate.

12.4.1 Consensus trees

A consensus tree provides an estimate for the level of support for each clade in the final tree. It is built by combining clades which occurred in at least a certain percentage of the resampled trees. This percentage is called the consensus support threshold. A 100% support threshold results in a **Strict consensus tree** which is a tree where the included clades are those that are present in all the trees of the original set. A 50% threshold results in a **Majority rule consensus tree** that includes only those clades that are present in the majority of the trees in the original set. A threshold less that 50% gives rise to a **Greedy consensus tree**. In constructing a **Greedy consensus tree** clades are first ordered according to the number of times they appear (i.e. the amount of support they have), then the consensus tree is constructed progressively to include all those clades whose support is above the threshold *and* that are compatible with the tree constructed so far.

The length of the consensus tree branches is computed from the average over all trees containing the clade. The lengths of tip branches are computed by averaging over all trees.

Note: The above definitions apply to rooted trees. The same principles can be applied to unrooted trees by replacing "clades" with "splits". Each branch (edge) in an unrooted tree corresponds to a different split of the taxa that label the leaves of this tree.

12.4.2 Creating a consensus tree of existing trees

Select a tree set document (e.g. a set of bootstrap replicate trees) and choose **Tree** then **Consensus Tree Builder** at the top of the setup dialog. Check **Create Consensus Tree** and choose the **Support Threshold %** you wish to use. This will create a consensus tree using the trees already in the document (no resampling will be performed) and it will either be added to the tree document or saved as a separate tree document.

12.4.3 Sort topologies

This will produce one or more trees sorted by topology, summarizing the results of resampling, check **Sort topologies** under the **Consensus Tree Builder** options. The frequency of each topology in the set of original trees is calculated and the topologies are sorted by their frequency. A number of these topologies, based on the topology threshold, will be output as summary trees. The summary trees have branch lengths that are the average of the lengths of the same branch from trees with the same topology.

The **topology threshold** determines what percentage of the original tree topologies must be represented by the summarizing topologies. The most common topology will always be output as the first summary tree. If the frequency (%) of this does not meet the threshold then the next most frequent topology will be added, and so on until the total frequency of the topologies reaches the threshold value.

A topology threshold of 0 will result in only the most common topology being output, a threshold of 100 will result in all topologies being output.

12.5 Viewing and formatting trees

Once the tree is built it will appear in the Document Viewer window (Figure 12.3). When viewing a tree a number of other view tabs may be available depending on the information at hand. The **Alignment View** tab will be visible if the tree was built from a sequence alignment using Geneious. The **Text View** shows the tree in text format (Newick).

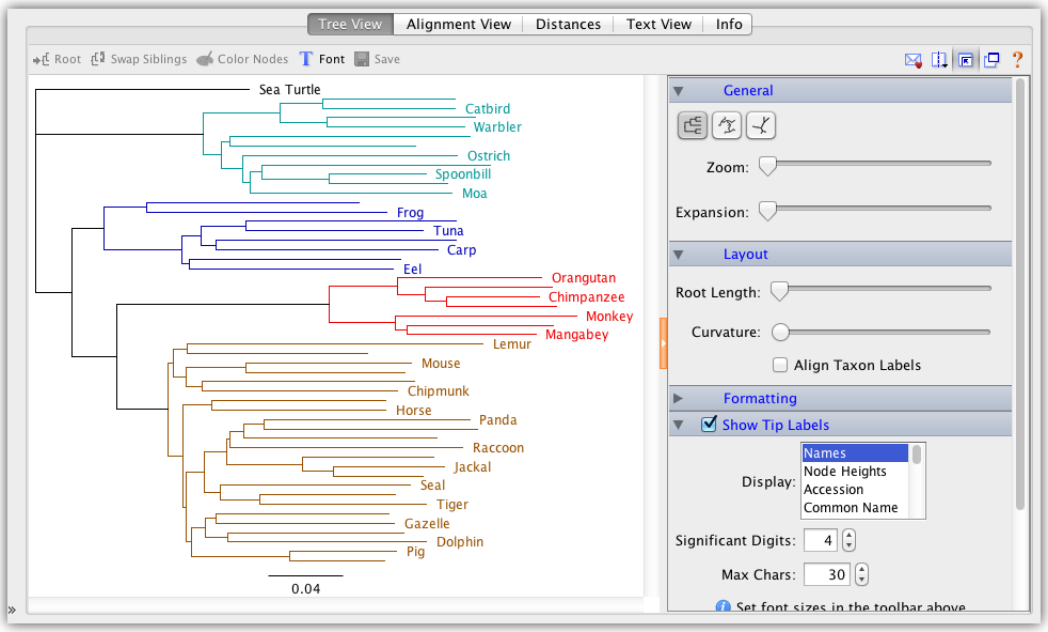


Figure 12.3: A view of a phylogenetic tree in Geneious

The tabs to the right of the tree viewer contain options for controlling the look of the tree, and the information displayed on it. The toolbar above the tree provides additional formatting options, and allows you to change the root, or rearrange the tree. The subsequent sections provide more detail on these options.

12.5.1 Current Tree

If you are viewing a tree set, this option will be displayed. Select the tree you want to view from the list.

12.5.2 General

The **General** tab has 3 buttons showing the different possible tree views: rooted, circular, and unrooted. The **Zoom** slider controls the zoom level of the tree while the **Expansion** slider expands the tree vertically (in the rooted layout).

12.5.3 Layout

This has different options depending on the layout that you select above:

- **Root Length** Sets the length of the visible root of the tree *(Rooted and Circular trees)*
- **Curvature** Adds curvature to the tree branches *(Rooted view only)*
- **Align Taxon Labels** Aligns the tip labels to make viewing a large tree easier *(Rooted view only)*
- **Root Angle** Rotates the tree in the viewer *(Circular and Unrooted views)*
- **Angle Range** Compresses the branches into an arc *(Circular view only)*

12.5.4 Formatting

The following options are available for formatting branches:

- **Flip the tree horizontally** flips the tree so that branches go from right to left, rather than left to right.
- **Transform branches** allows the branches to be equal like a cladogram, or proportional. Leaving it unselected leaves the tree in its original form.
- **Ordering** orders branches in increasing or decreasing order of length, but within each clade or cluster.
- **Show root branch** displays the position of the root of the tree (has no effect in the unrooted layout).
- **Line weight** can be increased or decreased to change the thickness of the lines representing the branches.
- **Show selected subtree only** shows only the part of the tree that is selected (or the entire tree if there is no selection).

12.5.5 Show Tips, Node and Branch Labels

If you are unfamiliar with tree structures, please refer to Figure 12.1 for a diagram of tips, nodes and labels.

Show tip labels: This refers to labels on the tips of the branches of the tree. Tip labels can be any of the fields on your document, and can be set in the **Display** option. To select multiple fields to display at the tips, hold down the command/control key while selecting.

Show node labels: This refers to labels on the internal nodes of the tree. If you are viewing a consensus tree, you can display consensus support % here, or you can display the node heights.

Show branch labels: This refers to labels the branches of the tree. You can display substitutions per site (branch lengths) here, or for a consensus or bootstrapped tree you can display consensus support % or bootstrap support %.

For node and branch labels, the font can be set using the **Font Size** options in the tab. The tree viewer will shrink the font size of some labels if they cannot all fit in the available space. The lower end of the range specifies the minimum size that the tree viewer is allowed to shrink the label font to. The font sizes for the tip labels are set using the **Font** button in the toolbar above the tree viewer. **Significant Digits** sets how many digits to display if the value the node is displaying is numeric.

12.5.6 Automatically collapse subtrees

This option is available in version 9 onwards, and enables groups of similar nodes to be collapsed into a single node that represents that subtree. The maximum distance within the subtrees is determined by the Subtree Distance slider. Use this option to help navigate trees with many nodes and tips.

Collapsed nodes are labeled with the name of one of the tips, a count of how many tips the subtree contains, and the maximum distance between the top of the subtree and any of the tips within it. Double-clicking a node in a tree will force it to expand or contract. **Automatically Collapse Subtrees** will not override this state. To reset the state of double-clicked nodes in the tree, click **Reset state of X nodes**. X is the number of nodes with a manually expanded or collapsed state.

12.5.7 Show scale bar

This displays a scale bar at the bottom of the tree view to indicate the length of the branches of the tree. It has three options: **Scale range**, **font size** and **line weight**. Setting the scale range to 0.0 allows the scale bar to choose its own length, otherwise it will be the length that you specify.

12.5.8 Statistics

Displays information on the number of nodes and number of tips in the tree.

12.5.9 The Toolbar

The buttons on the toolbar along the top of the viewer allow you to edit the tree.

Click on a node in the tree viewer to select the node and its clade. Double-click the node to collapse/un-collapse the clade in the view. Once you have selected a clade in the view, you can edit it using the following toolbar buttons:

- **Color Nodes**: allows you to choose a new color for the selected clade.
- **Font**: allows you to change the font for the tip labels.
- **Root**: allows you to re-root the tree on the selected node.
- **Swap Siblings**: allows you to swap the position of the sibling clades of the selected node.

In version 9 onwards, the toolbar also contains a **Search** box that allows you to search for particular tip labels. If a match is found, this tip is displayed on the tree and all other tips are greyed out. If you wish to search by a field that is not currently displayed on a tip label, you need to change the field under **Show Tip Labels** first.

Chapter 13

PCR Primers

Geneious provides several operations for designing and working with PCR Primers and DNA or hybridisation probes. PCR Primers and DNA or hybridisation probes can be designed for or tested on existing nucleotide sequences or alignments. A PCR product can be extracted from a sequence that has been annotated with both a forward and a reverse primer. 5′ extensions consisting of restriction enzymes or arbitrary sequence may also be added to primer documents.

In addition Geneious can determine the primer characteristics for a primer sized sequence and convert it into a primer. Characteristics can also be determined for any number of primer sized selections made in the Sequence View.

To use any one of these primer operations simply select the appropriate nucleotide sequences and either select **Primers** from the Tools menu or right-click (Ctrl+click on Mac OS X) on the document(s) and select **Primers**. A popup menu will appear showing the operations valid for your current selection.

13.1 Design New Primers

Geneious uses Primer3 to design PCR primers. The Primer Design dialog allows you to set options for where your PCR primers should sit, what size product to return and characteristics such as primer length and melting temperature.

Two options are available for primer design: **Design New** or **Design with Existing**. **Design New** designs a pair of forward and reverse primers. You can specify if you wish to design with or without a matching probe. **Design with Existing** can design a partner primer to match an existing one, for example a reverse primer for a forward or vice versa. It also allows you to design a probe to match a pair of primers.

If any documents were selected which either are primer sequences or contain primer annota-

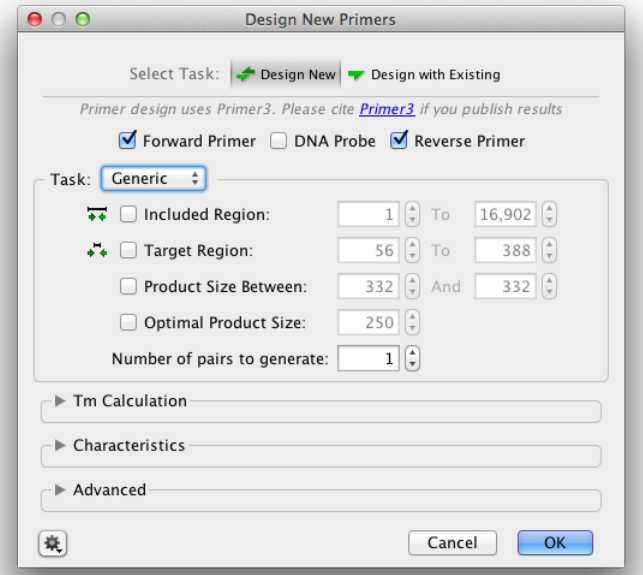


Figure 13.1: The primer design dialog

tions then these will be made available for selection as primers in a drop-down box. Selected sequences are treated as primer or probe sequences if they are 150bp in length or less.

For each of these options, Generic or Cloning primers can be designed.

13.1.1 Generic primers

This option will design standard PCR primers according to the region input options you select. These options allow you to specify what part of a sequence you wish to amplify. Most options are optional and can be enabled or disabled with the associated check boxes beside them. If you have selected a region in the sequence before opening the primer dialog then this region will automatically be used for Included Region and Target Region. All of these are expressed in base pairs from the beginning of the sequence and are as follows:

- **Included Region:** Specifies the region of the sequence within which primers are allowed to fall. This must surround the target region and allows you to choose a small region on either side of the target in which primers must lie.
- **Target Region:** Specifies which region of the sequence you wish to amplify and unless the advanced options allow otherwise, the forward and reverse primers must fall somewhere

outside this region.

- **Product Size:** Specifies the range of sizes which the product of a primer pair can have. The product size is the distance in bp between the beginning of the forward primer to the end of the reverse primer.
- **Optimal Product Size:** Specifies the preferred size of the product. Setting this will mean primer pairs that have a product size close to this will be chosen over those that do not. Warning: Setting these options can cause the primer design process to take considerably longer to complete.

The final option in this section is **Number of Pairs to Generate** which specifies how many candidate pairs of primers and DNA probes to generate and is compulsory. Setting this to 1 will give you only the primer pair which was considered best by the set parameters.

13.1.2 Cloning primers

This option allows you to design primers to amplify a specific region. Only the included region can be set, and the primers will be designed to the very ends of this region so that the entire region is included in the PCR product. This option is useful for amplifying an entire CDS for creating an insert for cloning.

13.1.3 Tm calculation

This section gives the formulas used to calculate the melting point of oligos. Under **Formula** you can choose between two different tables of thermodynamic parameters and methods for melting temperature calculation:

- **Breslauer et al. 1986**. This is used by old versions of Primer3 (until version 1.0.1), and uses the formula for melting temperature calculation suggested by Rychlik et al. 1990.
- **SantaLucia 1998**. This is the recommended value.

Three different Salt Correction Formula options are available:

- **Schildkraut and Lifson 1965** . This is used by old versions of Primer3 (until version 1.0.1)
- **SantaLucia 1998**. This is the recommended value.
- **Owczarzy et al. 2004**.

13.1.4 Characteristics

The Characteristics section allows you to set absolute limits on properties of primers and probes such as melting point and GC content. Optimum values can also be specified. For details on individual options hover your mouse over them and a popup box will describe the function of the option.

Characteristics can be set for either **Primers** or **DNA Probes**, depending on the task you have chosen. The **Primer** section is available if one of **Forward Primer** or **Reverse Primer** is being designed or tested and **DNA Probe** is available if a DNA Probe is being designed or tested. These two sections are quite similar; the DNA probe section has a subset of the options available in the primer section. This is because primers are usually chosen in pairs and so several options can be set for how pairs are chosen.

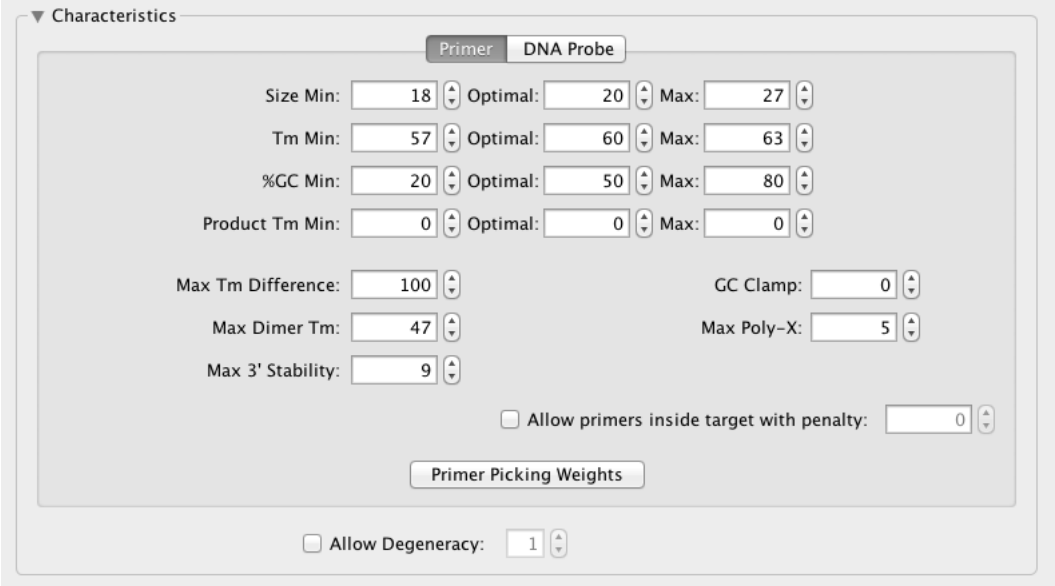


Figure 13.2: Primer characteristics options

13.1.5 Primer Picking Weights

At the bottom of the Characteristics panel there is a **Primer Picking Weights** button. Clicking this brings up a second dialog containing many more options. The purpose of all of these options is to allow you to assign penalty weights to each of the parameters you can set in the options. The weight specified here determines how much of a penalty primers and probes get when they do not match the optimal options. The higher the value the less likely a primer or probe will be chosen if it does not meet the optimal value.

Some of the weights allow you to specify a "Less Than" and "Greater Than". This is for options which allow you to specify an optimum score such as GC content. These weights are used when looking at primers whose value for this option falls below and above the optimum respectively. The other weights are applied no matter in which direction they vary.

For details on individual options in the Primer Picking Weights dialog, again hover your mouse over the option to see a short description.

13.1.6 Degenerate Primer Design

A degenerate primer contains a mix of bases at one or more sites. They are useful when you only have the protein sequence of your gene of interest so want to allow for the degeneracy in the genetic code, or when you want to isolate similar genes from a variety of species where the primer binding sites may not be identical. You can design degenerate primers in Geneious by using either a sequence containing ambiguous bases or an alignment as the template and checking the **Allow degeneracy** box. The degeneracy value that you specify is the maximum number of primers that any primer sequence is allowed to represent. For example, a primer which contains the nucleotide character N once (and no other ambiguities) has a degeneracy of 4 because N represents the four bases A,C,G and T. A primer that contains an N and an R has degeneracy $4 \times 2 = 8$ because R represents the two bases A and G.

13.1.7 Advanced Options

In the Advanced panel there are options to add 5′ extensions to primers and to specify a mispriming library.

A 5′ extension can be your own sequence, a restriction enzyme or Gateway site, or a combination of these. For more information see section 13.8.

A mispriming library is a set of sequences (usually repeats) which the primers should not bind to. Four inbuilt libraries are available for selection, or you can upload a custom library of sequences in fasta format. For more information on the inbuilt libraries, see the Primer3 help page.

13.1.8 Batch Primer Design

Multiple primer pairs can be designed at once by selecting multiple regions within a single sequence and opening **Design New Primers**. When multiple regions are selected, a checkbox **Use all selections** will appear next to Target and Included regions. Check this option for either Target or Included regions to design primers to all selected regions in one step.

Geneious Primer Characteristics	Primer3 Web Interface	Primer3 Command Line
%GC	Primer GC%	PRIMER_{LEFT,RIGHT}_GC_PERCENT
Tm	Primer Tm	PRIMER_{LEFT,RIGHT}_TM
Hairpin	Max Self Complementarity (Any)	PRIMER_{LEFT,RIGHT,INTERNAL_OLIGO}_SELF_ANY
Primer-Dimer	Max 3' Self Complementarity	PRIMER_{LEFT,RIGHT,INTERNAL_OLIGO}_SELF_END
Monovalent Salt Concentration	Concentration of monovalent cations	PRIMER_SALT_CONC
Divalent Salt Concentration	Concentration of divalent cations	PRIMER_DIVALENT_CONC
DNTP Concentration	Concentration of dNTPs	PRIMER_DNTP_CONC
Sequence	Seq	PRIMER_{LEFT,RIGHT}_SEQUENCE
Product Size	Product Size Ranges	PRIMER_PRODUCT_SIZE
Pair Hairpin	PAIR ANY COMPL	PRIMER_PAIR_COMPL_ANY
Pair Primer-Dimer	PAIR 3' COMPL	PRIMER_PAIR_COMPL_END
Pair Tm Diff	Max Tm Difference	PRIMER_PRODUCT_TM_OLIGO_TM_DIFF

Table 13.1: Geneious primer characteristics and their Primer3 counterparts

13.1.9 Output from Primer Design

Once the task and options have been set, click the **OK** button to design the primers. A progress bar may appear for a short time while the process completes. When complete, primers and probes will be added as annotations on the sequences. The annotations will be labelled with the base number the primer starts at, followed by either F (forward primer), R (reverse primer), or P (probe). Primers will be coloured green and probes red.

Detailed information such as melting point, tendency to form primer-dimers and GC content can be seen by hovering the mouse over the primer annotation. The information will be presented in a popup box. Alternatively, double clicking on an annotation will display its details in the annotation editing dialog. Table 13.1 shows how the values in the Geneious primer annotation map to the original Primer3 values.

If you are designing primers off an alignment the primer will be designed on the consensus sequence by default. To design primers for every sequence in the alignment and have the primers annotated separately on each sequence or on a few selected sequences, choose **Design primers on "Every Sequence", or "Selected Sequences"** in the alignment options at the bottom of the Design New Primers window. This option is only available in Geneious 8.1 onwards.

The best way to save a primer or DNA probe for further testing or use is to select the annotation for that primer and click the **Extract** button in the sequence viewer. This will generate a separate, short sequence document which just contains the primer sequence and the annotation (so it retains all the information on the primer). In the case of the reverse primer it will automatically be reverse complemented.

To **delete primers** that you don't want, just select the primer annotation and click the Delete button. You will then be given the option to delete the pair of that primer at the same time.

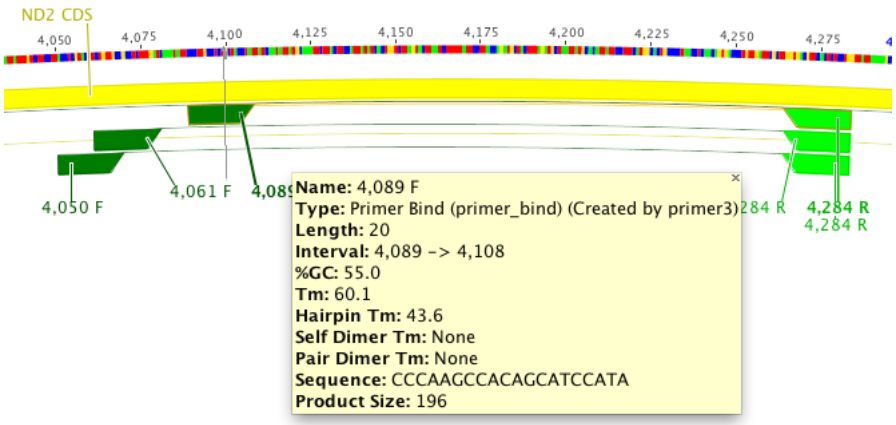


Figure 13.3: Primer design output

13.1.10 When no primers can be found

If no primers or DNA probes that match the specified criteria can be found in one or more of the sequences then a dialog is shown describing how many had no matches and for what reasons.

To see why no primers or DNA probes were found for particular sequences, click the 'Details' button at the bottom of the dialog. The dialog will then open out to display a list of all the sequences for which no primers or DNA probes were found. For each of the sequences the following information is listed:

- Which of Forward Primer, Reverse Primer, Primer Pair and/or DNA Probe could not be found in the sequence

- For each of these, specific reasons for rejection are listed (eg. "Tm too high" or "Unacceptable product size") along with a percentage which expresses how many of the candidate primers or probes were rejected for this reason.

After examining the details you can choose take no action or continue and annotate the primer and/or DNA probes on the sequences which were successfully designed for.

13.2 Manual primer design

It is possible to create PCR primers by adding a primer annotation directly onto a sequence. This is especially useful for cloning applications as generally the primers must bind to a specified set of bases at the beginning and end of the gene to be cloned. To manually add a primer, select the region of sequence where you wish the primer to bind and click **Add Annotation**. Make the annotation type **primer bind** and you will then see primer-specific options and characteristics as in Figure 13.4. Changing the primer binding site position in the Add annotation window will automatically update the primer sequence and characteristics. A 5′ extension can also be added directly onto a primer in this step by clicking the button next to "Extension". See section 13.8 for more information on adding 5′ extensions.

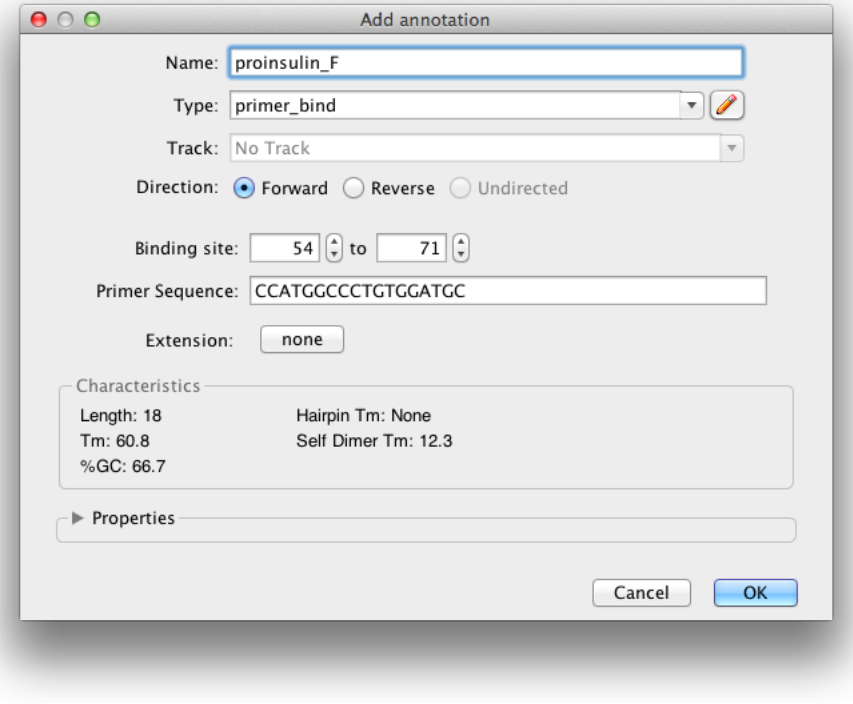


Figure 13.4: Create a primer by adding a primer annotation

13.3 Importing primers from a spreadsheet

You can import primers and probes directly into Geneious from Comma/Tab-Separated Values documents. You can either import them from the **Import** → **From File** menu, or simply paste

the contents of the document into Geneious.

When Geneious has successfully recognized the file as CSV or TSV, you will see the following dialogue (Figure 13.5).

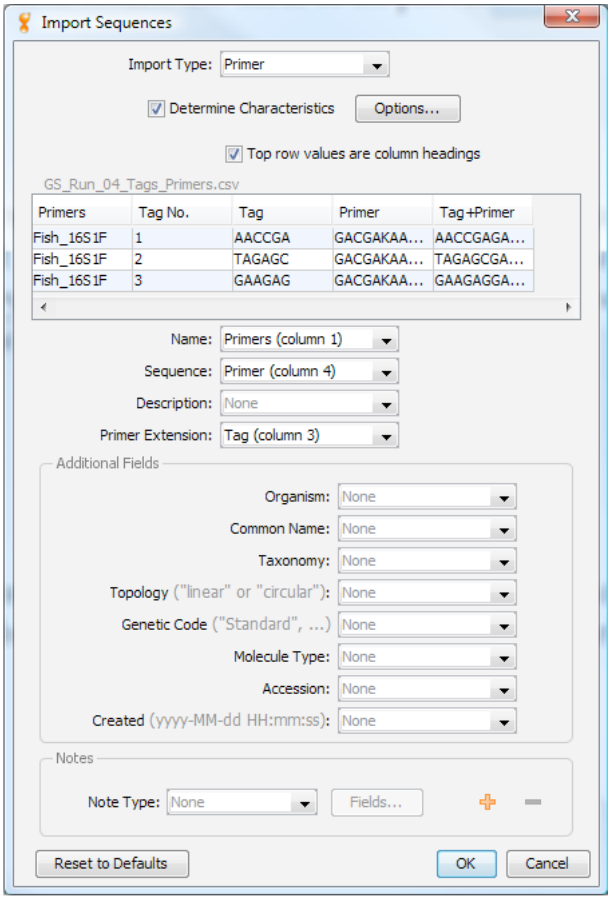


Figure 13.5: Importing primers from a spreadsheet

You will be asked which type of sequence you are importing. When you choose to import primers or probes, you will receive some options that allow you to determine characteristics for them as an extra step.

Immediately below this is a preview of the first few rows of data, and a checkbox that allows you to tell Geneious that the top row is a heading row and should be ignored.

Below the preview is a list of common and additional fields, along with dropdown boxes. These boxes allow you to specify which column contains which piece of data – often, one or more of these won't be applicable and can be left as **None**. Note that at minimum, you must specify a **Sequence** field.

Lastly, any add additional data in the form of meta-data. Clicking the dropdown box next to **Meta Data** at the bottom of the dialog will allow you to import values to meta-data, and clicking the + or − will allow you to insert or remove additional meta-data types. Next, click the **Fields...** button to bring up a dialog.

An additional set of dropdown boxes will allow you to specify again which columns of data contain the fields which comprise this meta-data type. This includes custom meta-data types that you have created and saved in the past.

When you're ready, hit **OK** to begin importing. When Geneious is done, you may be presented with the option of grouping the sequences you imported into a sequence list. This option is recommended if you're importing very large sets of sequences.

13.4 Primer Database

The Primer Database consists of all the oligonucleotide documents that exist in your Local or Shared Databases. The **oligonucleotide** document type is a short nucleotide sequence representing either a primer or a probe. The text view lists the primer characteristics (Tm, GC *etc*). These properties can be shown in the document table. Tm is shown by default, but you can turn on others by right clicking on the table header.

Oligo sequences are created via one of the following methods:

- Extract a primer/probe annotation from a sequence
- Select **Sequence** → **New Sequence** from the menu and choose Primer or Probe as the type of the new sequence
- Select one or more existing primer sequences (maybe ones imported from a file) then click **Primers** → **Convert to Oligo** to transform them into oligo type sequences
- Import primer sequences from a comma separated file (.csv) and choose Primer or Probe as the sequence type (see section 13.3).

If you select a target sequence and go to **Test with Saved Primers** or **Design Primers** → **Design With Existing**, Geneious will find all oligo sequences in your database and offer them as options in the list of oligo sequences. There is no need to select them along with the target sequence before starting the operation.

The meta-data type **Primer Info** can be used to note the fridge location *etc* of a particular primer.

13.5 Test with Saved Primers

Primers and probes can also be quickly tested against large numbers of sequences to see which ones the primers will bind to. To test primers select the target sequences you want to test for compatibility with primers and choose **Primers → Test with Saved Primers**.

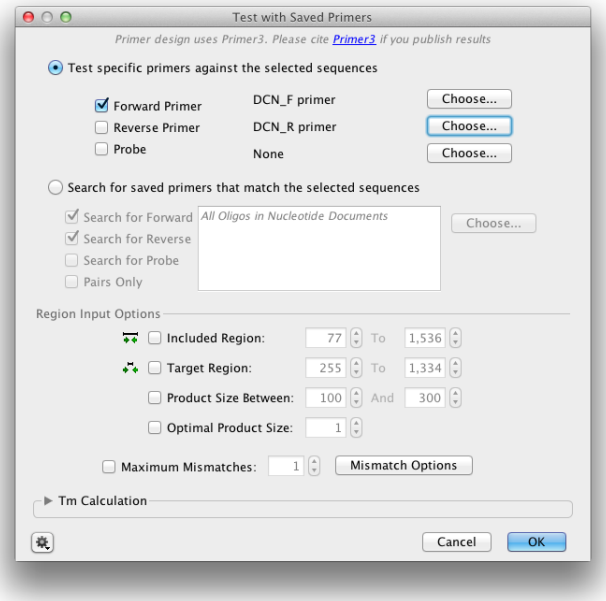


Figure 13.6: The primer test dialog

There are two ways in which Geneious can test your selection of primers and probes. The first option in the dialog tests a specific forward and reverse primer pair and/or probe. Clicking the **Choose** buttons next to forward, reverse and probe options will bring up your primer database, allowing you to select any primer in your database for testing.

The second option allows you to specify multiple primers and probes to test all selected sequences against. Click the Choose button then hold down CTRL-click (command-click on Mac) to select multiple primers and probes from many different locations in your database. Alternatively, you can select one or more folders to test with all the primers and probes inside them, or click the Use All button to use every primer in your database.

As with the first option, you can choose which types of primers you'd like to test for, by selecting the checkboxes on the left. Note that each primer you select will be considered in both the forward and reverse roles, if you have checked both Search for Forward and Search for Reverse. One final checkbox, Pairs Only, forces primers and probes to be considered as pairs (with the probe inside), otherwise they can be found anywhere in the sequence with no constraints.

All of the same options available for designing primers also apply to testing so if the primers are expected to bind to quite different regions of the test sequences the primer binding region may have to be extended and the target region option can be omitted.

By default, only primers that match the target sequence exactly will be found. If you wish to allow a limited number of mismatches between the primer and target sequence you can specify this under **Maximum Mismatches**. You can limit the position in which mismatches are allowed by clicking the **Mismatch Options** button.

Click the **OK** button and testing will commence. Once complete, a dialog will present the results. This dialog tells you how many of the sequences were compatible with the specified primers and probes and provides details and choices very similar to the one described in section 13.1.10. The compatible primers can be annotated onto the sequences in a similar manner to that when designing primers. Additionally if the primer sequences were not already annotated with a primer annotation they will be annotated during testing.

13.6 Characteristics for Selection

The **Characteristics for Selection** option will determine the primer characteristics of a selection of sequence within a larger sequence. Select a region of 36bp or less in the Sequence View and choose **Primers** → **Characteristics for Selection**. The primer characteristics will then be added as an annotation over the exact region that was selected. This will also work on multiple selected regions in the Sequence View. Hold the Ctrl key while clicking and dragging to select multiple regions simultaneously.

13.7 Convert to Oligo

Geneious can convert nucleotide sequences into primers. This is necessary for sequences to show up in the oligo database.If your sequences are less than 36 bp long, this operation will also determine the primer characteristics of the sequences, such as melting point. To do this, select your sequences and choose **Primers** → **Convert to Oligo** from the popup menu that appears. If you select just two sequences you have the additional option of determining their pair characteristics. Determining the pair characteristics of two primer sequences can be used to see if two sequences can pair and how well they do so.

Note: Sequences longer than 36 bp can still be converted into oligo sequences, but their characteristics cannot be calculated as this is the size limit that Primer3 sets for accurate Tm calculations.

13.8 Primer Extensions

You can add a primer extension to an existing **oligonucleotide** sequence by selecting **Primers → Add 5′ Extension**. You can add your own sequence, a restriction site, and/or Gateway cloning site. Multiple extensions can be added in one go, and the preview window in the 5′ extension dialog box shows how these extensions will be arranged on the primer. The order of 5′ additions can be edited by dragging and dropping them in this window. Primer extensions can also be added at the time the primer is designed from within the **Design New Primers** setup options.

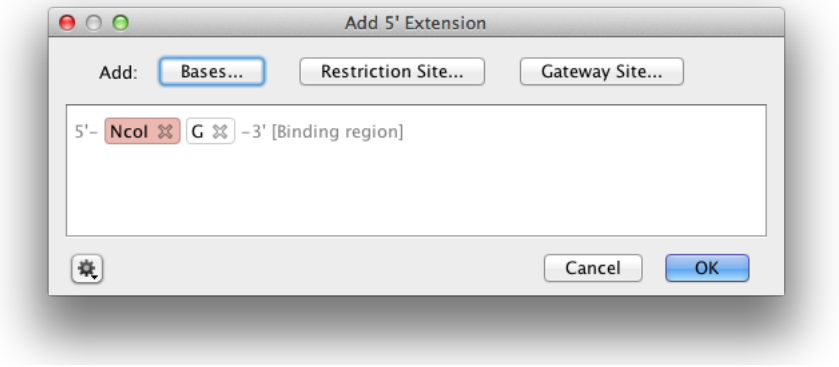


Figure 13.7: Adding 5′ extensions to a primer

Once added, the 5′ extension is shown in bold on the primer sequence and is not covered by the primer annotation, as shown in Figure 13.8. These extensions will not change the binding region of the primer and will be ignored when primer testing is conducted against potential target sequences.

If the primer is annotated onto a sequence following testing, the extension sequence is shown in the list of the annotation's qualifiers. If the primer or a PCR product is extracted from this annotation, the result will include the extension.

13.9 Extract PCR Product

Once primers are annotated on a sequence, the resulting PCR product can be extracted by selecting **Primers → Extract PCR Product**. If only a single pair of primers are annotated on a sequence then these will automatically be chosen as the Forward and Reverse primer. If multiple primers are annotated on a sequence, then the drop down menus allow you to choose which primers to use for extracting a single PCR product, or alternatively you can choose to Extract PCR products from all primers.

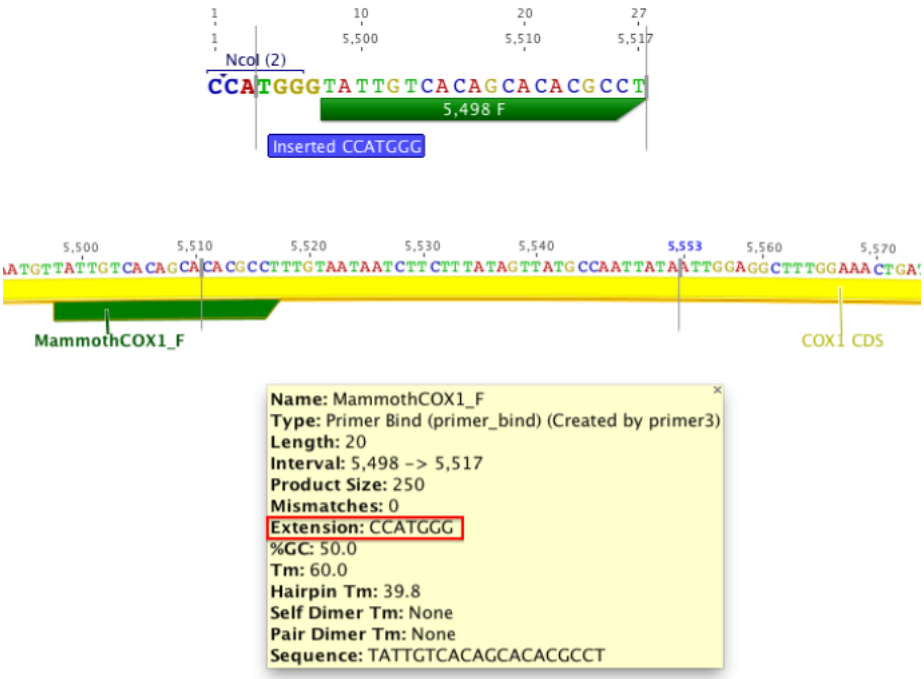


Figure 13.8: Primer annotation with 5′ extension

When you click **OK** a new sequence document is produced containing only the sequence spanning (and including) the PCR primers. Any 5′ extensions on the primers will also be included.

13.10 More Information

The Primer feature in Geneious is based on the program Primer3 (http://bioinfo.ut.ee/primer3/).

If you use the primer design feature of Geneious for publication we request that you cite Primer3 as:

Steve Rozen and Helen J. Skaletsky (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386 Source code available at http://sourceforge.net/projects/primer3/.

Further information on the functionality of the primer design feature can be found in the primer3 documentation available here: http://primer3.ut.ee/primer3web_help.htm. Please note that some controls have been changed, renamed or removed from Geneious, but most of the primer3 functionality is available.

Chapter 14

Cloning

The cloning features in Geneious allow you to simulate several different types of cloning, including basic restriction cloning, Topo cloning, Gateway cloning and Gibson Assembly. You can also find restriction sites on your sequence of interest, and simulate digestion and ligation reactions. The following options are available under the **Cloning** menu:

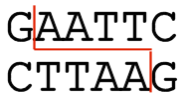
- **Find Restriction Sites...** finds and annotates restriction sites on your sequence of interest from a candidate set of restriction enzymes. See section 14.1.
- **Digest into fragments...** allows you to generate the actual fragments that would be created in a digestion experiment using restriction enzymes. See section 14.2.
- **Ligate Sequences...** lets you ligate two or more fragments, with or without overhangs.
- **Find non-cutting enzymes...** returns a list of enzymes which do not cut your sequence.
- **Restriction Cloning...** allows you to choose a digested fragment or a sequence with two restriction site annotations to use as an insert, and insert them into a vector (circular sequence). Geneious can do the work of working out what cut sites on the vector are compatible with the overhangs on the insert, with some additional information from you. See section 14.3.
- **Gibson Assembly...** provides a one-step interface to perform a Gibson Assembly or similar operation (a isothermal ligase independent, restriction free overlap extension PCR cloning). You can chose to insert one or multiple inserts into one or multiple vectors and specify the insertion order. The operation automatically creates the necessary primers and the products you will get and generates a report document. See section 14.4.
- **Gateway Cloning** automatically detects Att sites on the insert and vector(s), and performs a one-step Gateway® Cloning reaction. See section 14.5.

- **Golden Gate** simulates the digestion and ligation of multiple sequences with the same Type IIS enzyme in a single reaction. See section 14.6.
- **TOPO cloning...** automatically detects TOPO vectors amongst the selected sequences and inserts the fragments into these vectors using a BLUNT- TA- or Directional Cloning approach. See section 14.7
- **Find CRISPR sites...** annotates specific CRISPR guide sequences in your sequence of interest, and scores them based on offsite targets in your specified genome. See section 14.8.

The following sections give more detail on each option.

14.1 Find Restriction Sites

Restriction Enzymes[1] cut a nucleotide sequence at specific positions relative to the occurrences of the enzyme's *recognition sequence* in the sequence. For example, the enzyme *Eco*RI has the recognition sequence GAATTC and cuts both the strand and the antistrand sequence after the G inside the recognition sequence[2], leaving a single-stranded overhang (*sticky end (overhang)*):



The option **Find Restriction Sites...** from the **Tools** → **Cloning** menu or the context menu allows you to find and annotate restriction sites on a nucleotide sequence.

You can configure the following options:

- **Candidate Enzymes** lets you select a set of restriction enzymes from which you want to draw the ones to use in the analysis. This includes the options to use commonly used, or all known commercially available restriction enzymes. If you have created your own restriction enzyme set from your local database then this will also be listed (see below for how to create such a document).
- **Enzymes must match X to Y times:** only returns restriction enzymes which cut the sequence X to Y times. Results for enzymes that cut the sequence more or less than this will be discarded. If you set X to be 0, when this operation is complete, it will report which candidate enzymes do not cut the sequence.

[1] The restriction enzyme information included in Geneious was obtained from Rebase, available for free at http://rebase.neb.com.

[2] Like many restriction enzymes *Eco*RI is methylation dependent and cuts only if the second A in the recognition sequence is not methylated to N6-methyladenosine.

- **Specifying cut regions**: To specify a region of sequence where you want the enzyme to cut or not cut, choose one of the options below, and use the base counters to specify a sequence range that the options apply to. If you have selected a region of sequence in the sequence viewer, clicking the refresh arrow next to the base number counters will copy the selected region to this setting. The following options are available:
  - **Cut Anywhere**: Returns enzymes which cut anywhere in the entire sequence. It is not possible to select a subregion with this setting.
  - **Must only cut between**: Returns enzymes which only cut between the specified bases.
  - **Must cut between (may cut outside)**: Returns enzymes which cut between the specified bases, and may also cut outside that region.
  - **Must not cut between**: Returns enzymes which only cut outside the specified bases.
- **Advanced**: This displays a table of all enzymes in your candidate set, including their recognition site, overhang, and methylation information (Figure 14.1). Only the enzymes selected in this table will be considered in the analysis; initially, all rows are selected. You can click on the column headers to sort the table ascending or descending by that column, and you can Shift+click and Ctrl+click to select a range of rows and to toggle the selection of a row, respectively.
- To create custom enzyme set, select the enzymes you want in the Advanced window, then click **Save Selected Enzymes** and give the set a name. This set will then be available in the **Candidate Enzymes** lists.

After configuring your options, click **Apply** to record the restriction enzyme site annotations on the sequence. The annotation shows the enzyme's recognition site, and the cut site. Once the document is saved, two new tabs will appear above the sequence view: **Enzymes** displays the list of enzymes and their cut positions; **Fragments** displays a list of fragments that would be produced from the restriction digests. These tables can be exported as .csv files for subsequent processing with other software such as e.g. Microsoft Excel®.

To select the region between two cut sites on a sequence, Shift+click on the two restriction site annotations in the sequence view.

14.2 Digest into fragments

The option **Digest into fragments...** from the **Tools** → **Cloning** menu or the context menu allows you to generate the nucleotide sequences that would result from a digestion experiment. You can digest multiple nucleotide sequences at a time. If the digestion results in overhangs, these will be recorded as annotations on the fragments.

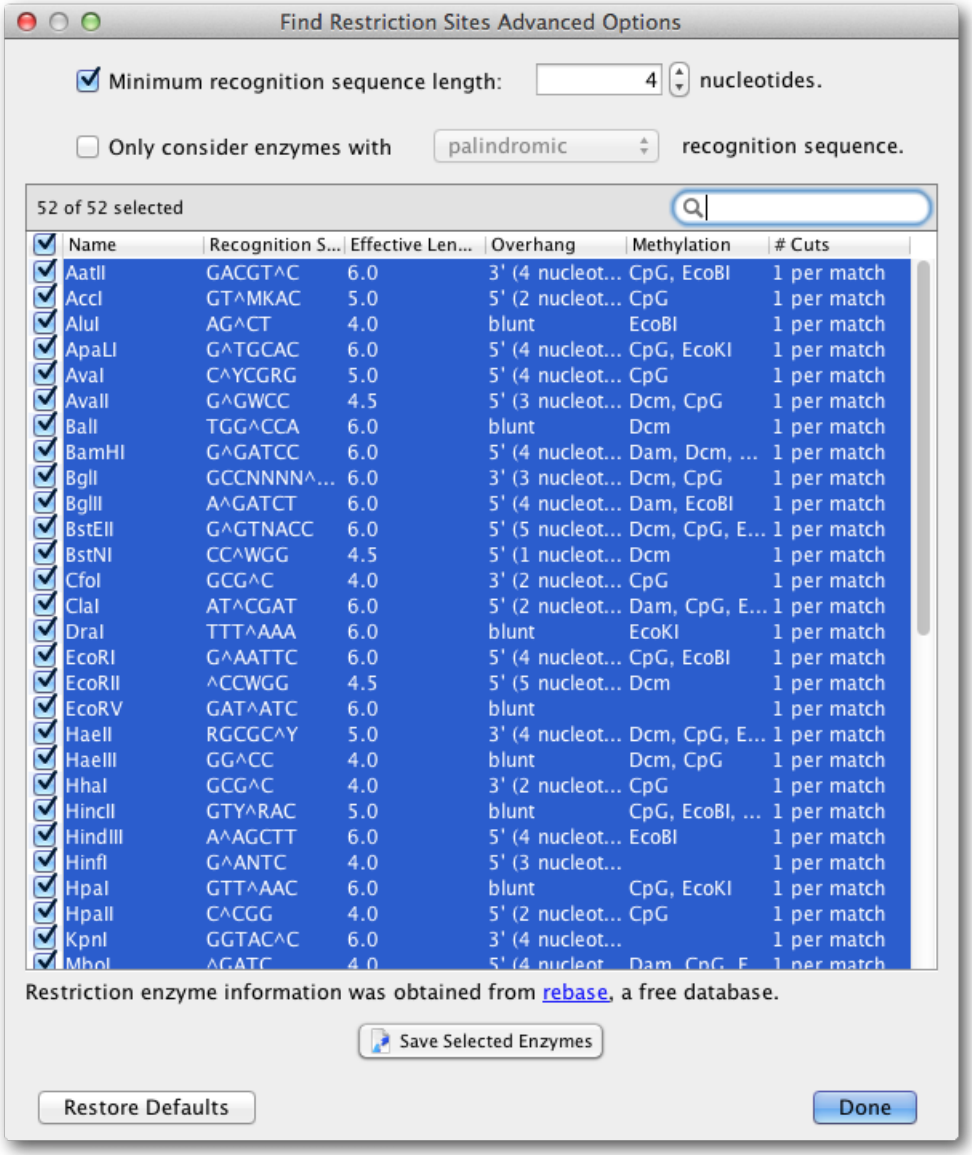


Figure 14.1: **Find Restriction Sites** restriction enzymes table accessible under the **Advanced** option.

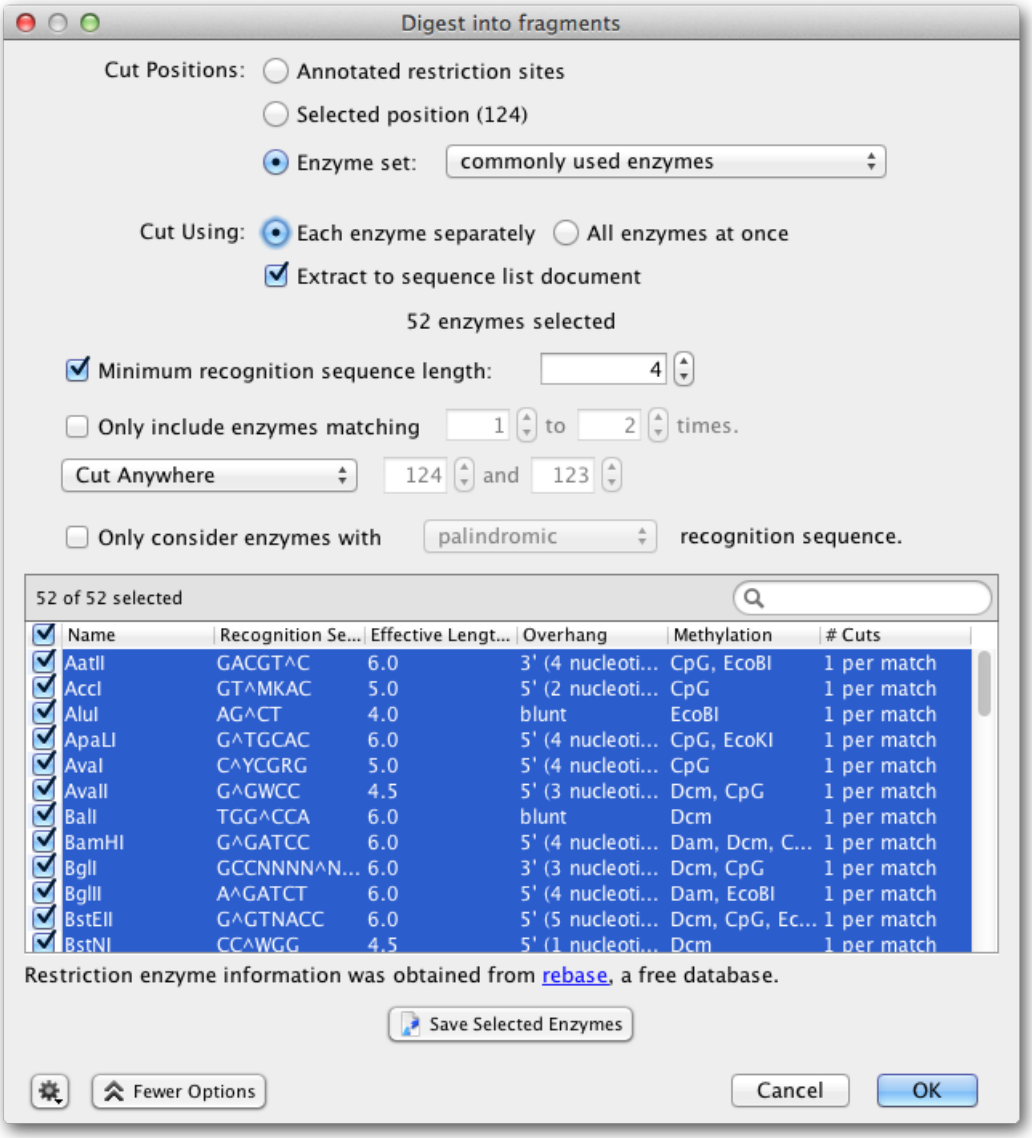


Figure 14.2: **Digest into fragments** options dialog, with extended options showing.

- If you have selected only one nucleotide sequence document and it has annotated restriction sites, you can select **Annotated cut positions** to cut the document on these sites. When this option is selected, the options to filter the enzymes by their effective recognition sequence length or number of hits are disabled. You can select a subset of the annotated enzyme sites under **More Options**.
- If you do not have annotated cut sites already on your document, you can choose **Enzyme set** and select which enzyme(s) you wish to use. This runs Find Restriction Sites first, but does not generate restriction site annotations on your original sequence. See the section on Find Restriction Sites (14.1) for more detail.
- Where multiple enzymes are selected, you can either digest the original sequence by **Each enzyme separately**, which returns a separate sequence list of the fragments produced for each enzyme, or by **All enzymes at once**, which digests by all selected enzymes in one operation.

14.3 Restriction Cloning

To do basic restriction cloning, inserting a digested fragment into a vector, select **Restriction Cloning** from the **Cloning** menu. The insert must be one of the following:

- A fragment which has already been digested. This fragment cannot have any restriction site annotations on it. The entire fragment will be inserted into the vector. Overhangs will be taken into account.
- A sequence with two restriction site annotations. The fragment resulting from digesting this sequence (and discarding the fragments from the ends) will be inserted into the vector.

The vector must be a circular sequence. You do not need to annotate the restriction sites used to cut the vector in advance; the Restriction Cloning operation will do that for you.

This operation cannot deal with some aspects of molecular cloning such as triple ligation and the blunting or filling in of overhangs. If you want to do a cloning operation outside the scope of this operation, you will need to annotate restriction sites on the sequences involved, digest the fragments, modify them in the sequence viewer if necessary and then ligate them back together as a set of discrete steps.

Insert Options

You cannot alter the insert used in the operation from the options, but you can select what direction to insert in: forward or reverse. If the insert fragment has complementary overhangs

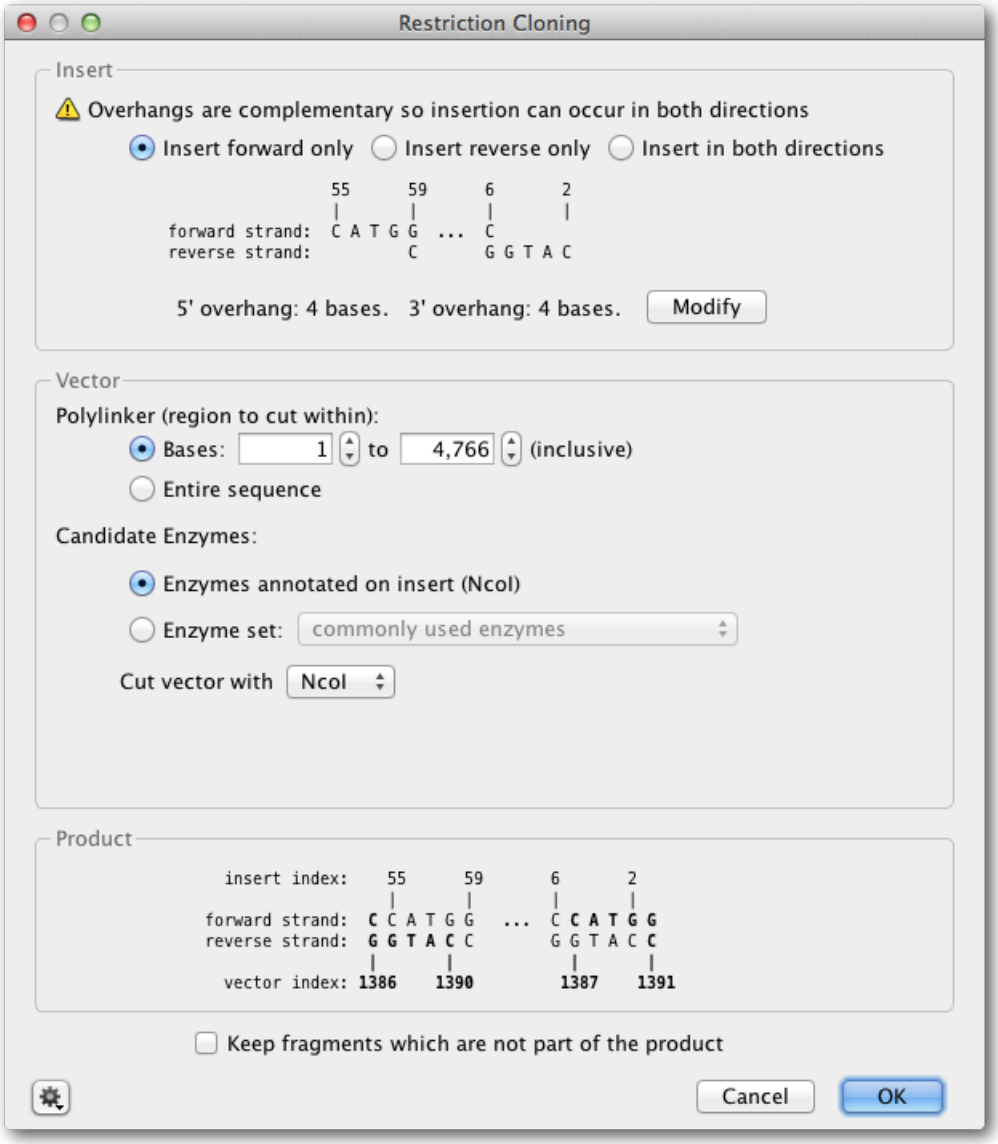


Figure 14.3: **Restriction cloning** options dialog

or is blunt at both ends, you can also choose to insert in both directions. In this case, two product documents will be created, one for the insert in each direction.

The insert options also present a diagram showing the bases at each end of the insert fragment.

Vector Options

- **Polylinker (region to cut within):** These options let you choose what region within the vector sequence to look for enzymes to cut within. Geneious will examine the vector sequence for enzymes that have cut sites within this region and none outside it. You can specify the polylinker in the following ways:
  - **Bases** Used to explicitly specify the range of bases to use.
  - **Entire sequence** Used to specify that you can cut anywhere within the sequence.
- **Candidate Enzymes:** These options let you choose which enzymes to look for on the vector sequence
  - **Enzymes annotated on insert** This option lets you use only the enzymes used to cut the insert fragment.
  - **Enzyme set** This option lets you use the enzymes from a predefined enzyme set, eg. the enzyme set you have created containing the enzymes you have in your lab.
- **Cut vector with:** Whenever you change the options for the polylinker or candidate enzymes, Geneious will recalculate the compatible enzymes on the vector. It will look for enzymes which meet one of the following conditions (in addition to cutting only within the polylinker and belonging to enzymes from the candidate enzyme set):
  1. A single enzyme which cuts the vector once, such that the insert can be inserted in the gap (Possible only when the insert has complementary cut sites).
  2. A single enzyme which cuts the vector twice, such that the insert can be inserted into the gap vacated by the fragment between the two cut sites
  3. Two enzymes which each cut the vector once, such that the insert can be inserted into the gap vacated by the fragment between the two cut sites

Other Options

The Product section of the options displays a diagram showing the ligation points in the insertion. The parts of the ligation points belonging to the vector appear in bold in this diagram.

Below this is a checkbox where you can choose whether to **Keep fragments which are not part of the product**. If this box is checked, a document will be created representing the fragment removed from the vector, if any. If the insert fragment was produced from a sequence with two restriction site annotations, the fragments on either side of the restriction site annotations will also be kept.

14.4 Gibson Assembly

The operation will generate sequences with compatible overlaps that can ligate to each other after a partial chew-back with a T5 exonuclease. The overlaps are created by extension overlap PCR, the corresponding primers will automatically be generated and displayed in a report document and as annotations on the resulting sequences. Figure 14.4 below shows the Gibson assembly options.

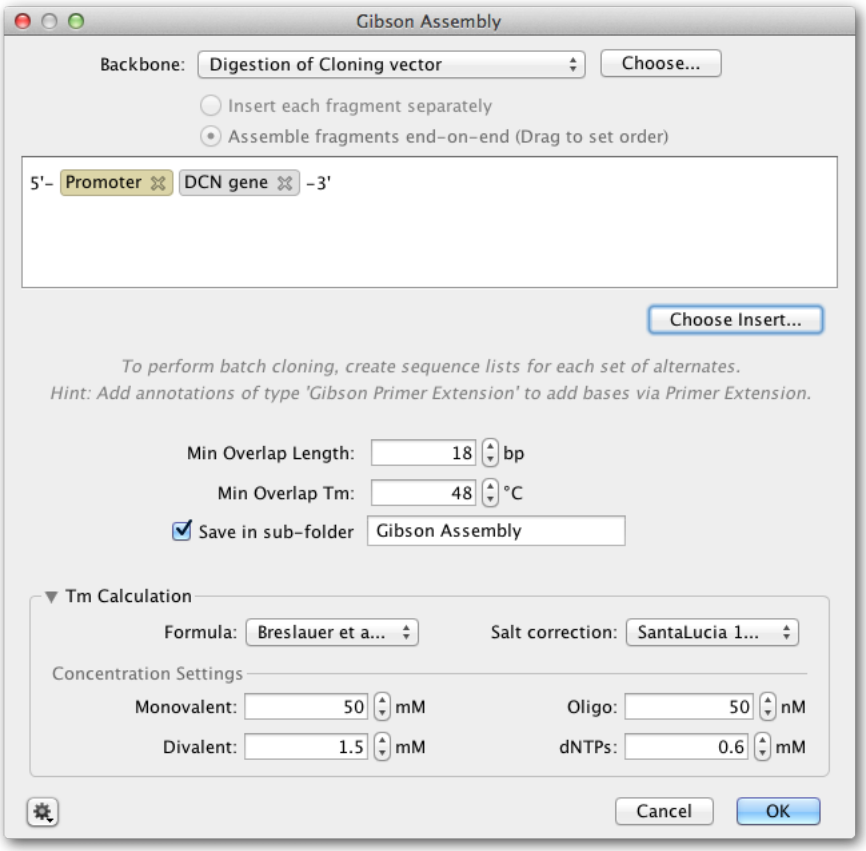


Figure 14.4: *Gibson Assembly* options dialog, with extended options showing.

- From Geneious 8.1 onwards it is not necessary to select the insert and vector prior to opening the Gibson Assembly options. If you select only a single sequence prior to selecting Gibson Assembly, then this sequence will appear as the insert. Vectors and/or other inserts can then be chosen from within the Gibson Assembly setup window by clicking the **Choose...** button. This brings up a document chooser from which you can select your vector and/or inserts from any folder in your database.
- If 'none' is selected as the Vector, the product(s) will be linear, otherwise circular.

- Insert sequences can be ordered via drag-n-drop. Sequences that have been previously grouped into lists (e.g Promoters, shown in brown-ish) will all be inserted at the specified position, generating one product per sequence at that position.
- If no grouped documents are provided in the drag-n-drop field the user can chose to insert the sequences either all at once into the vector as a sequential assembly (Assemble fragments end-on-end), or alternatively as separate inserts with one insert per generated product (Insert each fragment separately).
- The minimum overlap (length of the complementary sequence) can be specified as well as the minimum melting temperature of the complementary sequence (Min Overlap Tm).
- To calculate the Tm a collapsible field is available showing the options provided and required by `primer3`.

The operation will remove 5′ overhangs and fill up 3′ overhangs that are eventually present on the sequences, eg. when they derived from a restriction digestion. The possible different sequence combinations are created and complementary extremities that might be already present will be considered when primers get designed.

For sequences without complementary overlaps a pair of primers will be generated. If both or only one of the ends are complementary, primers for both ends will be created, since the sequence will still have to be modified by a Primer Extension PCR to make it compatible at the opposing end. If both are complementary no primers will be generated.

Extensions will be added to the primer corresponding to the neighboring sequence. Modifications that have been manually introduced at the extremities of a vector will be automatically added as part of the extension, so that they get introduced to the sequence during the PCR. Manual modifications to the ends of insert sequences should be annotated as type 'Gibson Primer Extension' in order to be added to the primer extension rather than the binding sequence.

The melting temperature is calculated for the primer binding sequence and the extension part including the modified bases. For both, primer and extension generation, the user specified Tm formula is used. In many cases the Phusion DNA polymerase is used for which it is recommended to use the Tm formula of Breslauer et al. 1986.

Primers are generated only for insert sequences, supposing that the vector should stay unmodified. For this reason the extension length of the primer extending to the vector will be twice as long (the full specified minimal overlap length) compared to extensions on primers between two inserts who share half of the specified overlap length each.

For very short or long extensions `primer3` might fail to calculate a Tm. If the sequence is too short Formula 14.1 is used, if it has a length greater than 36bp Geneious uses the Formula 14.2.

$$Tm = 2[AT] + 4[GC] \tag{14.1}$$

$$Tm = 64.9 + 41\frac{[GC] - 16.4}{[ATGC]} \tag{14.2}$$

A Report Document will be generated listing the generated products and primers in a tabular view. Errors that occurred during the primer generation process will be reflected in that report document. Furthermore any modifications (recession or maintaining overhangs, adding extensions to primers) are shown at the beginning of the document.

Geneious has a built-in parent-descendant tracking system. Whenever a change is made to a parent sequence it will ask to propagate this change to its offsprings. However, in the case that the user introduced some changes in the primer binding region or the extension region of the original sequences, these changes won't be reflected in the report document.

14.5 Gateway® Cloning

Geneious contains three operations to assist with Gateway® cloning. Gateway is a registered trademark of Invitrogen Corporation. The **Gateway** option under the **Cloning** menu will perform a BP reaction and/or an LR reaction on the selected documents. If there are a mixture of AttB/AttP and AttL/AttR sites on the input documents, it will perform a BP reaction on all documents with AttB/AttP sites, followed by an LR reaction on the results of the BP reaction and any input documents with AttL/AttR sites.

For example, to insert a PCR product with attB sites directly into a destination vector, select the PCR product, a donor vector, and a destination vector. Geneious will first produce an entry clone from the PCR product and donor vector, then react this entry clone with the destination vector to produce an expression clone.

Annotate att sites...

This operation searches for AttB, AttP, AttL and AttR sites and annotates them on your sequence.

Add AttB Sites to PCR product

This operation allows you to add AttB sites to a PCR product. It will work on the following types of document:

- A PCR product. AttB sites will be appended to the PCR product.
- A document with primer binding sites annotated. If there is more than one pair, Geneious will ask you which pair to use. The PCR product will be extracted and AttB sites appended.

14.6 Golden Gate

Golden Gate is a method to conveniently digest and ligate multiple sequences with the same Type IIS enzyme in a single reaction. The setup dialog for Golden Gate assembly is shown in Figure 14.5.

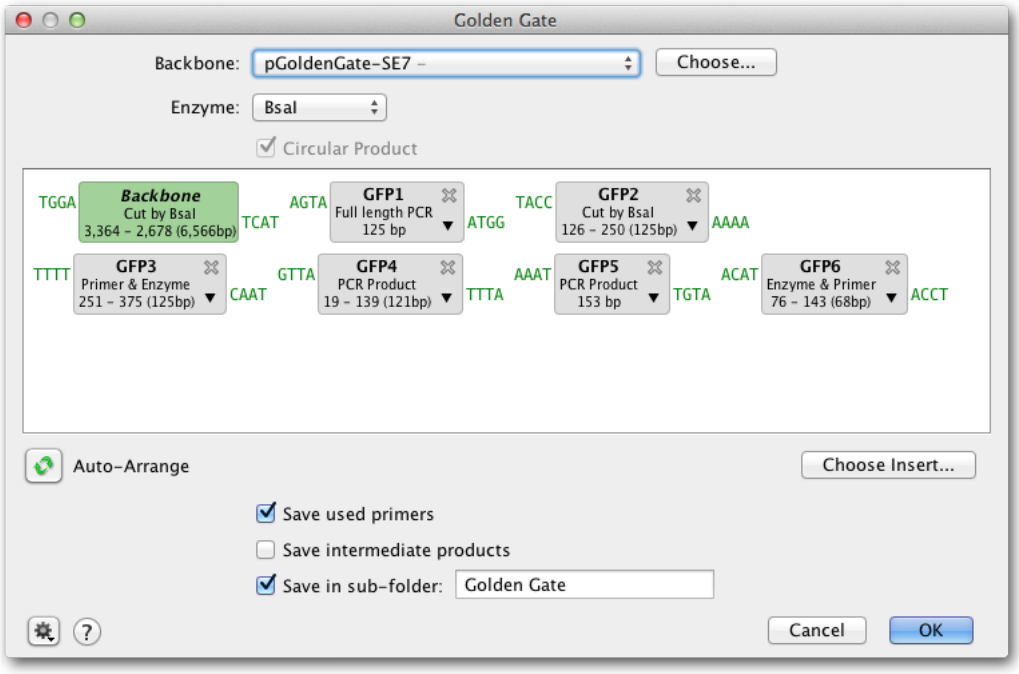


Figure 14.5: Golden Gate setup dialog, showing backbone and inserts with a range of different reaction types.

The **Backbone** is the sequence (e.g. a vector) into which the inserts are ligated. This setting will always remember the previously used backbone and can be manually changed by either using the dropdown (showing recently used backbones and any of the currently selected sequences) or selecting a backbone with the **Choose...** button. If a backbone is set the product will always be circular, and the **Circular Product** option will be disabled. If 'None' is chosen as the backbone the product can be circularized by checking Circular Product.

The **Insert(s)** can either be selected in the document table prior to opening the Golden Gate options, or can be added from within the setup options by clicking the **Choose Insert...** button. Inserts can also be removed by clicking the small X in the top right corner of the insert tag. Each insert has a small dropdown arrow, by which the sequence can be reverse complemented.

The **Enzyme** can be chosen from any of the commercially available Type IIS enzymes that have an effective recognition site of at least 4 bases and where the resulting overhang is at least 3 bases on any strand. Usually the "receiving" type IIS sites already present in your backbone

will define the site you will use (these do not have to be annotated on the backbone). Note that the Geneious Golden gate tool currently does not allow the use of multiple type IIS restriction enzymes for assembly.

Each sequence, including the backbone, is represented by a tag in the inserts field, with relevant information shown on the tag or next to it. If all overhangs are compatible, the backbone will be shown in green and the inserts in grey. Inserts or backbones with incompatible overhangs are shown in red. The information shown on the tag for each line is as follows:

- **Sequence Name** (highlighted in blue if reverse complemented)
- **Used reactions:**
  - **Pre-digested** if the sequence has existing overhangs on both ends
  - **Cut by [enzyme]** if the sequence will be cut by two enzymes
  - **PCR Product** if restriction sites for the chosen enzyme will be introduced via PCR and at least one of the existing primers on the sequence is used
  - **Full length PCR** if newly generated primers will amplify the full sequence and append restriction sites at both ends
  - **Overhang, Enzyme or Primer** for either side if a mix of either reaction is used
  - If conflicts have been detected, a warning will be shown, and the tag will be highlighted in red
- **Start, end** and **(length)** of the extraction or only the **length** if the whole sequence is used

An **Overhang** is shown on one or both sides of the tag. Only one overhang is shown if no overhang can or has to be created for the side with the missing overhang. The position (top/bottom) of the overhangs is an indication for whether it is a 3′ or 5′ overhang, the color indicates a valid match / an error state. Errors between sequences (e.g. multiples of the same overhang sequence) will be shown below the insert field.

When **Save intermediate products** is checked the PCR products and digested sequences will be saved in addition to the final sequence (where applicable).

14.6.1 Sequence ordering, rules and assumptions

Geneious will analyze your backbone (if defined), and each sequence passed to it, and will detect existing type IIS restriction sites, overhang annotations, primer annotations and blunt ends. When **Auto-arrange** is enabled, Geneious will try to identify the correct sorting of the sequences based on the available overhangs. The Auto-arrange button will be lighter grey when enabled, and darker grey when switched off. If tags are reordered or reverse complemented manually, Auto-arrange will be switched off and Geneious will not try to determine the correct order. Note that Auto-arrange will not arrange based on sequence names.

Geneious will use the existing restriction sites, overhang annotations and primer annotations to define the insert boundaries for the Golden gate recombination reaction. These are used in the order of precedence listed below (Rules 1-6), and if required, Geneious will design a primer for PCR amplification of the insert to generate the required cut site.

Rules, in order of precedence:

1. **Existing type IIS cut site(s):** If Geneious detects a pair of appropriately orientated type IIS sites with unique overhangs, then it will assume you wish to use them. Geneious will also assume that you have DNA available to use, and so will not design primers for PCR. If only one type IIS cut site is detected, then Geneious will design a primer that incorporates the site, and a second "opposite orientation" primer for PCR will be designed based on rules 3-6.

2. **Existing Overhang(s):** If Geneious detects a pair of valid overhangs compatible with the specified type IIS site, then it will assume you wish to use them. Geneious will also assume that you have this "sticky ended" DNA available and so will not design primers for PCR.

3. **Existing primer_bind annotation(s) with valid type IIS cut site(s) on the extension:** If Geneious detects a pair of inward facing primer_bind annotations with valid compatible type IIS sites then it will assume you wish to use them. Geneious will assume that you already have the corresponding primers, and new primers will not be designed for the region. If Geneious detects a single primer_bind annotation with a suitable type IIS site Geneious will assume you wish to use this and will only design a second "opposite orientation" primer, based on the appropriate rule, for PCR.

4. **Existing primer_bind annotations with other extensions:** If Geneious detects a primer_bind annotation with an extension which does not contain a valid type IIS site then the 5′ terminus of the extension will be considered the fusion point and the extension will be further extended to introduce a valid type IIS recognition site, resulting in a new primer sequence.

5. **Existing primer_bind annotation(s) without extension(s):** If a primer_bind annotation without an extension is found, then an extension will be appended to introduce a valid type IIS recognition site, resulting in a new primer sequence.

6. **Blunt ends:** If Geneious finds a blunt end, and no suitable type IIS sites or primer_bind annotations are present, then a primer with an appropriate type IIS site extension will be designed. The fusion point will be the termini(us) of the blunt end fragment.

For any of these Geneious only accepts cut sites where the recognition site will be cut out or primers that point towards each other.

Important:

- **Removal of unwanted internal Type IIS sites**: If one or more of your sequences contain the specified type IIS restriction site/s then Geneious will assume you want to use the site/s in the assembly process and design a strategy accordingly. If one or more of your sequences contain type IIS restriction sites that you do not want be involved in the assembly then you will need to engineer each site out of your fragment using PCR, taking care to avoid altering any gene product sequences.
- **Removal of unused primer annotations**: If one or more of your sequences contain primer annotations that do not define a boundary for assembly then you should remove these annotations. If you do not remove these annotations then Geneious may interpret them as boundaries of your sequence which will result in truncated inserts and incorrect assembly.

14.7 TOPO® Cloning

TOPO Cloning lets you ligate a single fragment into a Vector within only 5 minutes using the natural activity of Topoisomerase I which recognizes a specific motif **5′ - (C/T)CCTT - 3′** on the DNA. TOPO is a registered trademark of Invitrogen Corporation.

The **TOPO cloning** function under the Cloning menu allows you to insert linear fragments into either linear TOPO vectors (when a TOPO-site is present at the extremities) or into circular TOPO vectors. You can select as many sequences at once as you like, they will be ligated into each other in a batch operation.

- Three different options (TA- Blunt- or Directional cloning) are shown on the top. If Directional is selected the user can define an overlap sequence. If this field is blank it has the same effect as Blunt cloning.
- The field below shows which of the selected sequences have been detected as vectors, all other sequences are inserts.
- If any complications occur, eg. when more than one TOPO site is detected or when a linear sequence with TOPO site is selected it will print a message in this box, also showing how the corresponding sequence is processed if the user clicks OK.
- The resulting sequences will be optionally saved in a sub-folder.

14.8 CRISPR site finder

The CRISPR/Cas9 system is an RNA-guided endonuclease technology for gene editing. This system requires guide RNA (gRNA) comprised of a 20 bp target sequence next to a PAM (Protospacer Adjacent Motif) to direct the Cas9 enzyme to the cleavage site. The Find CRISPR Sites

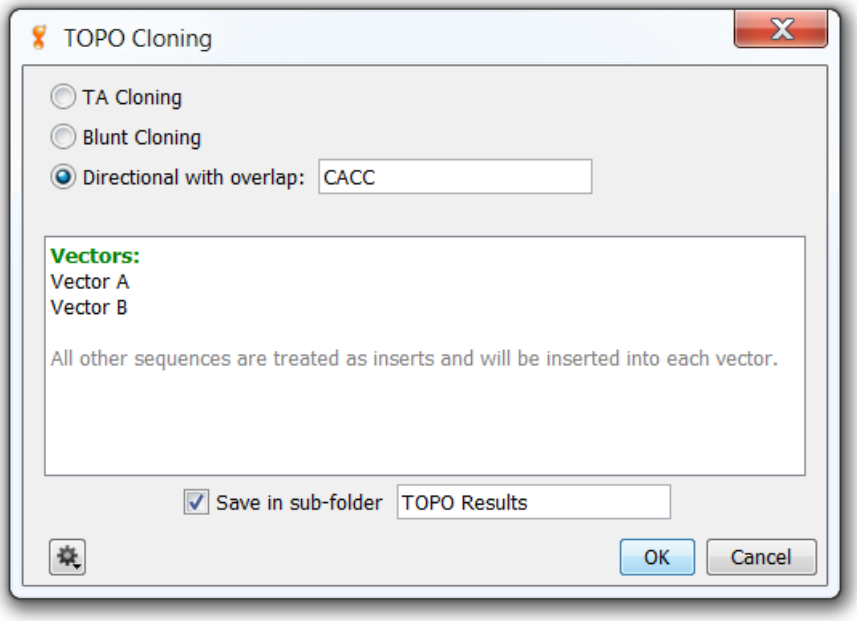


Figure 14.6: *TOPO Cloning* options dialog.

tool will search for gRNA ("CRISPR") sites in your selected sequence, and can score them based on on-target activity or potential off-target interactions.

To search for CRISPR sites, first select the sequence you want to target. This can be a full sequence document, or a selection in a sequence. For best performance the search region should be limited to 1000 bp. Open the **Find CRISPR Sites** tool in the Cloning menu. Check **Anywhere in sequence** or **Selected region** depending on if you are targeting the full sequence or a selection from it. Enter the gRNA sequence you want to search for in the Motif panel. There is a help box in the Motif section of the options panel that explains how to input your motif and gives an example. After selecting your scoring and pairing options, press **OK** to start the operation. Geneious will find potential CRISPR targets in your sequence and annotate them onto the document.

On-target Activity Score

CRISPR sites can be scored two ways. The first is by predicting the on-target activity using the method proposed by Doench et al, 2014. This scoring algorithm analyses the one- and two-base features of the gRNA, as well as the GC content, and uses an experimentally determined predictive model to score the expected activity level of the CRISPR target. Scores are between 0 and 1, with a higher score denoting higher expected activity. Note that ambiguous bases won't contribute to this score.

Off-target interaction analysis

The second scoring method compares a CRISPR site with other similar sites, and assigns a score based on how individual the CRISPR site is and how likely it is to induce off-target interactions. To enable this scoring check **Score sites through off-target analysis**. This function uses a scoring system proposed by Zhang et al, 2013, which takes into consideration the positions of mismatches between the CRISPR site and its off-target sites. Scores are between 0 and 100, with a higher score denoting less off-target activity.

Selecting **Score against an off-target database** will prompt Geneious to search sequences in the given folder for off-target interactions of your CRISPR sites. The sequence you are targeting can be inside this folder. This way you can find CRISPR sites that target a particular gene but are scored for interactions against the entire genome. The top 5 off-target sites are kept and annotated onto the CRISPR site even though all off-target interactions found are incorporated into the score. *Scoring against an off-target database will significantly increase the time taken for the operation to complete.*

Before scoring, the off-target database is searched for full exact matches of the whole target sequence. These intervals are ignored during the scoring phase because they are likely the original target itself or an expected repeat of the target. The intervals that were ignored are reported at the end of the operation in a dialog box. They can be found again later by mousing over the name of the CRISPR sites annotation track in the sequence viewer.

If **Score against an off-target database** is **not** selected, the CRISPR sites will be scored against any stretches of sequence in the original document that are not part of the selected target area.

You can control the number of mismatches allowed between a CRISPR site and its off-target sites by adjusting **Maximum mismatches allowed against off-targets**. You can also check for off-target sites with bulges and deletions by allowing some of these mismatches to be indels. **Maximum mismatches allowed to be indels** controls this. Note that increasing these numbers will increase the complexity of the search and therefore the time it takes to complete.

To help speed up off-target analysis, Geneious can filter out CRISPR sites with high-identity off-targets before performing the more intensive search for off-targets with mismatches. There are three **Strategy** options to for this:

- **Slowest - Score all sites**: No initial filtering for high identity off-target sites - scores will be calculated for all matching off-targets.
- **Normal - Discard worst sites**: Filters out CRISPR sites with off-targets that exactly match the 12 bp immediately adjacent to the PAM site..
- **Fast - Discard more sites**: Filters out CRISPR sites with off-targets that exactly match the 8 bp immediately adjacent to the PAM site..

Pair CRISPR sites

To help minimize off-target interactions, the mutant Cas9 D10A Nickase can be used with a pair of gRNAs on complementary strands of a target sequence. This induces a single-stranded break at each site, simulating a double-stranded break overall. Off-target interactions of any one gRNA target will then be only a single nick, which is repaired with high fidelity. This process is described by Zhang et al, 2013.

By selecting **Pair CRISPR sites**, only CRISPR sites that are within range of a complementary pair will be returned. You can specify the maximum overlap of the paired sites and the maximum space allowed between the paired sites. The maximum overlap and maximum space between sites are measured from the 5′, PAM-distal end of the CRISPR sites.

The optimal CRISPR pairs will be linked when they are annotated onto your target sequence, though any CRISPR site with a potential pair will be returned. Optimal pairs are decided by averaging their off-target interaction scores.

Results output

CRISPR sites are returned as annotations on your original sequence. Hovering the mouse over an annotation will bring up a tooltip showing the information about any scores calculated, plus the 5 highest-scoring off-target sites (if they are analysed). For each off-target site the score, location and off-target sequence are given. Mismatches to the CRISPR site are in red, and insertions in the off-target site are underlined and red.

The annotations are colored based on their On-target, Off-target, or Paired score. If more than one type of score is calculated you can choose which score to color the annotations by using the **Color CRISPR Sites by** option. The annotations will be colored on a gradient starting at green (for good scores), moving through yellow and ending at red (for poor scores).

14.9 Optimize Codons

The option **Optimize Codons...** from the **Annotate & Predict** menu allows you to adapt a nucleotide sequence to the genetic code and preferred codon usage of a particular expression host. The resulting sequence is optimized to avoid use of codons that rarely occur in the highly expressed genes of the expression host, thus increasing its expression level when cloned into that species. In addition, the resulting sequence can be forced to avoid cleavage sites for a set of restriction enzymes that you specify.

This tool calculates the Codon Adaptation Index (CAI) of a gene sequence as the geometric mean of the relative adaptiveness (w) of the codons in the sequence, as defined in Sharp and Li, 1987. Relative adaptiveness is calculated from Codon Usage Tables supplied by EMBOSS,

which calculate the relative codon usage of a selected set of genes in the target organism. The optimized sequence returned is the one for which CAI is maximised.

You can configure the following options (Figure 14.7):

- **Annotate Codons** creates an annotation track on the existing sequence with an annotation on each codon that is changed in the target sequence. Each annotation has qualifiers that specify the original codon, the replacement codon, and the relative adaptiveness value (W) of the replacement codon calculated for the selected target organism codon utilization table. The annotation track contains a qualifier with the CAI value of the result sequence.
- **Optimize and Save As** creates a new sequence document that contains the result codons and the same annotation track and annotations as described in the previous option. If multiple input sequences are selected the result sequences are placed in a single sequence list document. Enter the name of the new document in the text input box of this option.
- **Source Genetic Code** lets you select the genetic code to be used when translating the source sequences. If you have selected multiple source sequence documents with different genetic codes, the choice "Multiple Values" will be available to indicate that the genetic code associated with each document should be used. You can select a genetic code other that the one that is shown as the default for the selected input documents if you want to override the default.
- **Target Organism** lets you select a Codon Usage Table (from EMBOSS) for the target expression host. Each table contains the codon usage statistics for a selected set of genes of the organism.
- **Target Genetic Code** lets you specify the genetic code of the target organism. The EMBOSS CUT file format that is used for the Target Organism codon usage data does not contain a field for the genetic code. You must select the correct genetic code for the target organism that you select.
- **Threshold to be rare** lets you specify the relative adaptiveness value (W) for a codon below which the codon is considered to be "rare" and will be replaced with the highest value synonymous codon. Set the threshold to 1 to change all codons that are not the highest value in their synonymous set. Use threshold of 0 to only replace codons that code for the wrong amino acid in the target genetic code or that have to be changed to avoid restriction sites.
- **Avoid restriction sites** lets you select a set of enzymes that you intend to use to cut the target sequence. If you select an enzyme set with this option, the result sequence will not include any sites that match a recognition sequence in the set. The enzyme set choices are the same as are used in 14.1.

After configuring your options, click **OK** to start the analysis and annotate the optimized codons on the sequence or create the new sequence document.

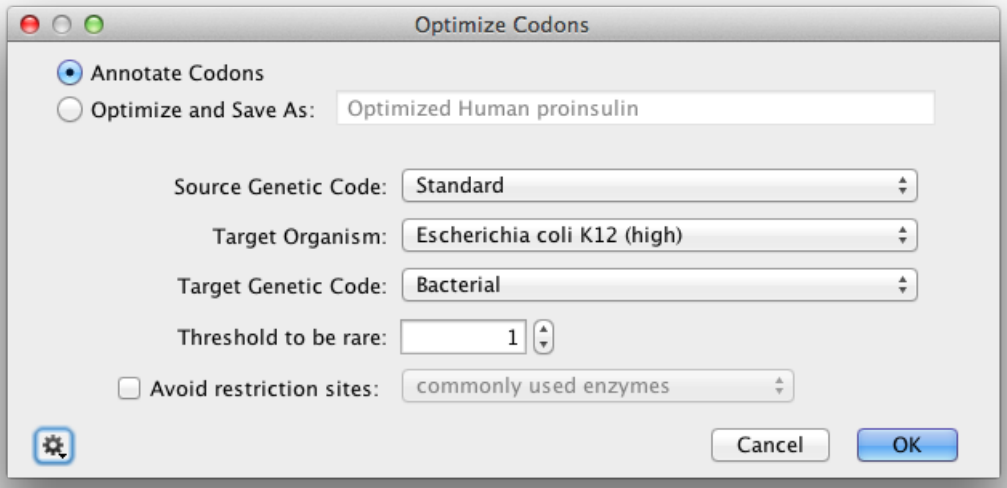


Figure 14.7: *Codon Optimization* options dialog.

Chapter 15

BLAST

BLAST stands for Basic Local Alignment Search Tool (Altschul et al 1990). It allows you to query a sequence database with a sequence in order to find entries in the database that contain similar sequences. When "BLAST-ing", you are able to specify either nucleotide or protein sequences and nucleotide sequences can be either DNA or RNA sequences. Sequences can be BLAST-ed against databases held at NCBI (see NCBI BLAST), or contained within your local Geneious database (Custom BLAST).

15.1 Setting up a BLAST search

To run a BLAST search in Geneious, select your query sequence or sequences and click the **BLAST** button in the toolbar. This operation can also be accessed by going to the Tools menu or by right-clicking (Ctrl+click on Mac OS X) on a sequence document and choosing **BLAST**.

Geneious gives you the option of searching against a database using either your currently selected sequence documents or a sequence you enter manually. If you choose to enter your sequence manually, then Geneious will display a large text box in which you can enter your query sequence as either unformatted text or FASTA format.

Select your database using the first drop-down box. Databases are grouped together under their respective services. Then choose which kind of BLAST search you wish to run under **Program**. The available programs will depend on the database you have chosen.

Geneious can perform seven different kinds of BLAST search:

- **blastn**: Compares a nucleotide query sequence against a nucleotide sequence database.
- **Megablast**: A variation on blastn that is faster but only finds matches with high similarity.

Figure 15.1: BLAST Options

- **Discontiguous Megablast**: A variation on blastn that is slower but more sensitive. It will find more dissimilar matches so it is ideal for cross-species comparison.
- **blastp**: Compares an amino acid query sequence against a protein sequence database.
- **blastx**: Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
- **tblastn**: Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
- **tblastx**: Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

Three options are available for displaying your results:

- **Hit table**: Returns one alignment for every hit against the database and displays them in a hit table. Each query displays a separate table and is also viewable as a query-centric alignment. This is suitable for less than 100 queries.

- **Query-centric alignment**: Returns one alignment for each query, showing the hits aligned against the query sequence. This is well suited for large batch searches but it doesn't display a hit table.
- **Bin into 'has hit' vs. 'no hit'** Returns two sequence lists: one containing queries which get a hit in the database, the other containing queries which don't. Details about the hits and alignments are discarded. This can be used to filter contamination (eg. human) from sequencing reads.

You can also specify how much of each matching sequence to retrieve from your database:

- **Matching region**: Just the region of the database sequence which matches the query.
- **Matching region with annotations**: The region of the database sequence which matches the query, plus any annotations on that sequence.
- **Extended region with annotations**: The matching region plus additional flanking regions upstream and downstream.
- **Full sequence with annotations**: The entire database sequence (this could be large and slow).

Geneious also allows you to specify most of the advanced options that are available in BLAST. To access the advanced options click the **More Options** button which is in the bottom left of the BLAST options. The available options vary depending on the kind of BLAST search you have selected. For details on each of the options you can hover your mouse over the option to see a short description or refer to the BLAST documentation from NCBI.

15.2 BLAST results

Once a search has started, a results subfolder will be created in the same folder as your query sequence. Search progress is shown in the document table. The search can be cancelled by clicking on the red square labelled **Stop**.

15.2.1 BLAST hit table

If you chose to return your results in a hit table, each search hit is displayed separately in the document table sorted by **bit score**. The bit score gives an indication of how good the alignment is; the higher the score, the better the alignment. In general terms, this score is calculated from a formula that takes into account the alignment of similar or identical residues, as well as any gaps introduced to align the sequences.

Search hits can also be sorted by other columns by clicking on the column header. Columns that may be useful to sort by include **E-value**, **Percent Identity**, **Query Coverage** or **Grade**. **E value** or "Expect value" represents the number of hits with at least this score that you would expect purely by chance, given the size of the database and query sequence. The lower the E-value, the more likely that the hit is real. The **Grade** column is a percentage calculated by Geneious by combining the query coverage, e-value and identity values for each hit with weights 0.5, 0.25 and 0.25 respectively. This allows you to sort hits such that the longest, highest identity hits are at the top.

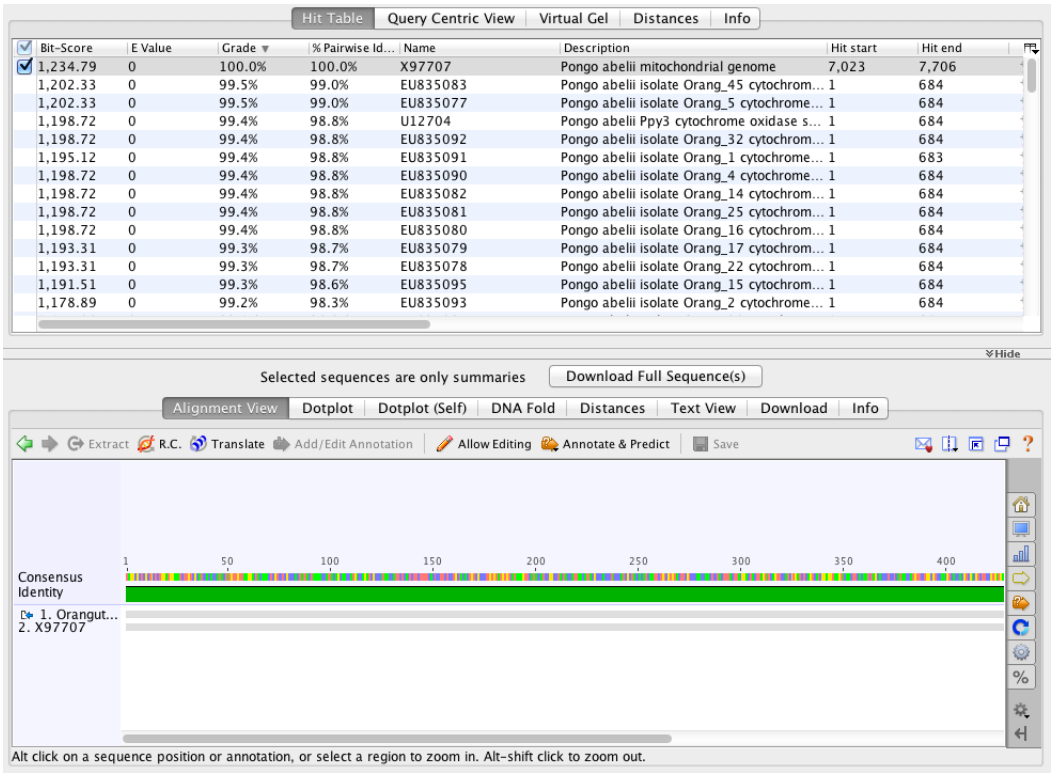


Figure 15.2: BLAST Complete

You can also download the full database sequence that corresponds to a BLAST hit. To retrieve the full sequence select a BLAST alignment and go to **File** → **Download Documents** or click the **Download Full Sequence(s)** button located above the viewer tabs. The full sequence will be available in the Sequence View tab once the download has completed and the region that matches the query sequence will be annotated as **BLAST Hit** (see Figure 15.3). In addition the annotations from the full sequence will be transferred over to the BLAST alignment and can be viewed in Alignment view.

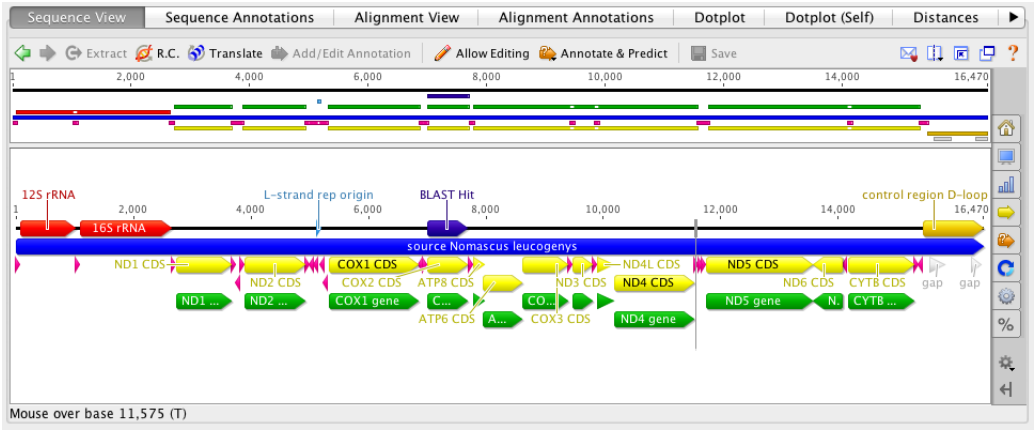


Figure 15.3: Document After Full Sequence Download

15.2.2 Query-centric view

This view displays all of the hits to your query in a single alignment. Results of single BLAST searches can be viewed in query-centric view instead of a hit table by clicking the **Query Centric View** tab at the top of the document table. Or, you can choose to only return a query-centric alignment when you set up the BLAST search. This option is particularly useful for batch BLAST, as only one alignment per query is returned and all the results are displayed in a single folder. In this view each hit sequence in each alignment is annotated with a **Search hit** annotation. If you mouse over the annotation you can bring up the values for E-value, pairwise identity, Grade etc. To display these values in a table, switch to the Annotations tab in the sequence viewer and add these columns to the table by clicking the **Columns** button.

15.3 NCBI BLAST

Geneious is able to BLAST to many different databases held at NCBI. These databases are listed in the Tables 15.1 and 15.2, and can be selected in the **Databases** drop down menu in the BLAST set up dialog. You must be able to connect to the internet from within Geneious to BLAST to NCBI, and if you are behind a proxy server you may need to enter your proxy server settings under **Tools** → **Preferences** → **Connection Settings**, as described in Section 1.2.5.

If you have a mirror of the NCBI BLAST databases you can set Geneious to use this by going to **Tools** → **Add/Remove Databases** → **Set Up BLAST Services**. This will bring up a dialog that allows you to change the setup for various search services in Geneious. Choose NCBI using the service drop-down box at the top of the dialog. Enter the URL for the mirror and click 'OK' to apply the new settings.

Table 15.1: Nucleotide BLAST databases

Database	Nucleotide searches
nr	All non-redundant GenBank+EMBL+DDBJ+PDB sequences(no EST, STS, GSS or HTGS sequences)
genome	Genomic entries from NCBI's Reference Sequence project
est	Database of GenBank + EMBL + DDBJ sequences from EST Divisions
est_human	Human subset of est
est_mouse	Mouse subset of est
est_others	Non-Human, non-mouse subset of est
gss	Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
htgs	Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2 (finished, phase 3 HTG sequences are in nr)
pat	Nucleotide sequences derived from the Patent division of GenBank
PDB	Sequences derived from the 3D-structures of proteins from PDB
month	All new/updated GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
RefSeq	NCBI-curated, non-redundant sets of sequences.
dbsts	Database of GenBank+EMBL+DDBJ sequences from STS Divisions
chromosome	A database with complete genomes and chromosomes from the NCBI Reference Sequence project.
wgs	A database for whole genome shotgun sequence entries.
env_nt	This contains DNA sequences from the environment, i.e all organisms put together

Table 15.2: Protein BLAST databases

Database	Protein searches
env_nr	Translations of sequences in env_nt
month	All new/updated GenBank coding region (CDS) translations +PDB+SwissProt+PIR released in last 30 days
nr	All non-redundant GenBank coding region (CDS) translations+PDB+SwissProt+PIR+PRF
pat	Protein sequences derived from the Patent division of GenBank
PDB	Sequences derived from 3D structure Brookhaven PDB
RefSeq	RefSeq protein sequences from NCBI's Reference Sequence Project
SwissProt	Curated protein sequences information from EMBL

You can also edit the databases that show up in Geneious by clicking on **Edit Databases** in the **Set up BLAST Services** window. This will only change the databases that Geneious displays and will not have any effect on the actual databases on the BLAST server.

15.4 Custom BLAST

Custom BLAST allows you to create your own custom database from either FASTA files or sequences in your local folders, and BLAST against it. The Custom BLAST plugin requires access to NCBI BLAST+ binary files.

15.4.1 Setting up the Custom BLAST files through Geneious

Geneious provides a download manager to help you download and extract the Custom BLAST files. To use it, go to **Tools** → **Add/Remove Databases** → **Set Up BLAST Services** and select **Custom BLAST** from the Service drop-down box (see Figure 15.4). Make sure **Let Geneious do the setup** is checked. Then click 'OK'. After a few seconds the compressed file containing all the files needed to run Custom BLAST will start downloading. You can click 'Pause' to pause the download. You can add and search Custom BLAST databases as soon as it has finished downloading and extracting. If you shut down Geneious with the file partially downloaded, you will need to start downloading it again from the beginning.

15.4.2 Setting up the Custom BLAST files yourself

If you want, you can download or otherwise acquire the NCBI BLAST+ binary files outside of Geneious. You can download them from here:

`ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/`

Choose the appropriate file for your operating system, download and extract it. You will need to let Geneious know where to look for the files once you have done this. To do this, go to **Tools** → **Add/Remove Databases** → **Set Up BLAST Services** and select **Custom BLAST** from the Service drop-down box. Enter your data location or click **Browse** to browse to the location of the files.

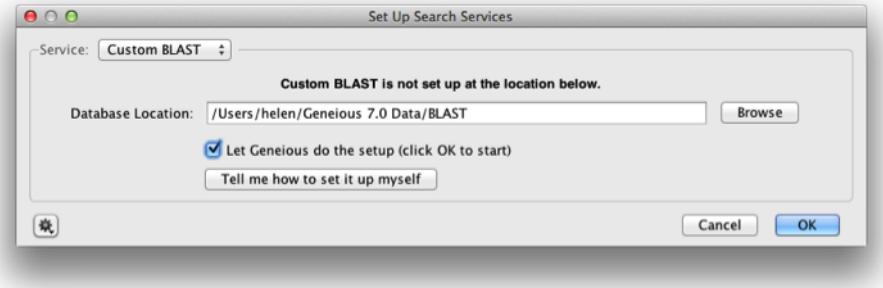
15.4.3 Adding Databases

Now that you have set up the executables, it is time to add databases to your BLAST.

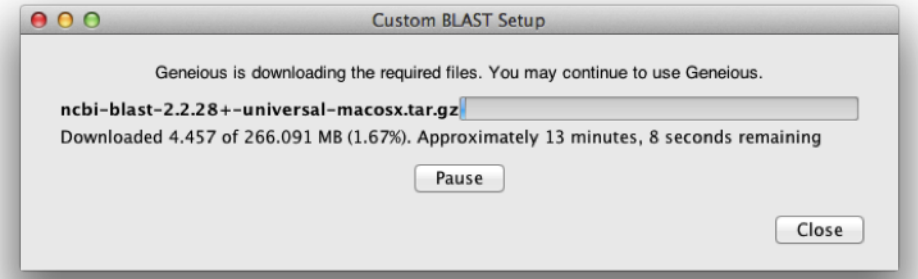
Creating a database from a fasta file

To create a database from the sequences in a FASTA file, go to **Tools** → **Add/Remove Databases** → **Add BLAST Database** and select **Custom BLAST** from the Service drop-down box (Figure 15.5). Choose to **Create from file on disk** and then click **Browse** to navigate to the FASTA file that contains the sequences you want to BLAST. Enter a name for the database and click 'OK'. There are two requirements for a FASTA file to be suitable for creating a database from:

- The FASTA file must contain only the same types of sequence (i.e. Nucleotide or Amino Acid)
- The sequences in the FASTA file must all have unique names



(a) Setup Options



(b) Downloading

Figure 15.4: Setting Up Custom BLAST

If the file meets these requirements it will be added as a database, otherwise you will be informed of the problem.

Creating a database from local documents

To create a BLAST database from sequences in your local documents folders, first select the documents that you want. Then go to **Tools** → **Add/Remove Databases** → **Add BLAST Database** and select **Custom BLAST** from the Service drop-down box. Enter a name for the database, and click 'OK'.

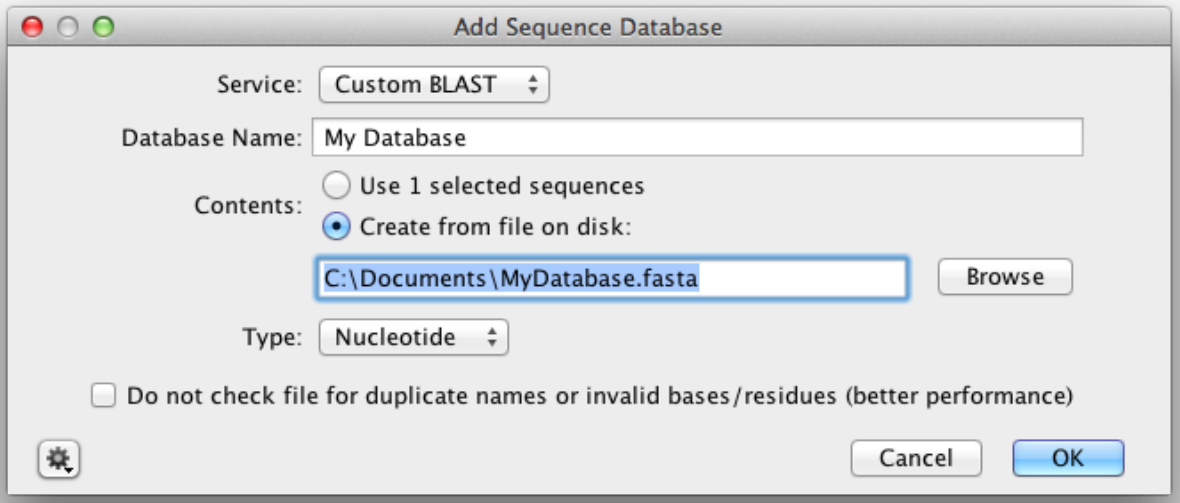


Figure 15.5: Adding a FASTA database

15.4.4 Using Custom BLAST

Once you have added one or more databases, they will appear under Custom BLAST in the BLAST database drop down (Figure 15.6). These can be used in exactly the same way as the NCBI BLAST ones.

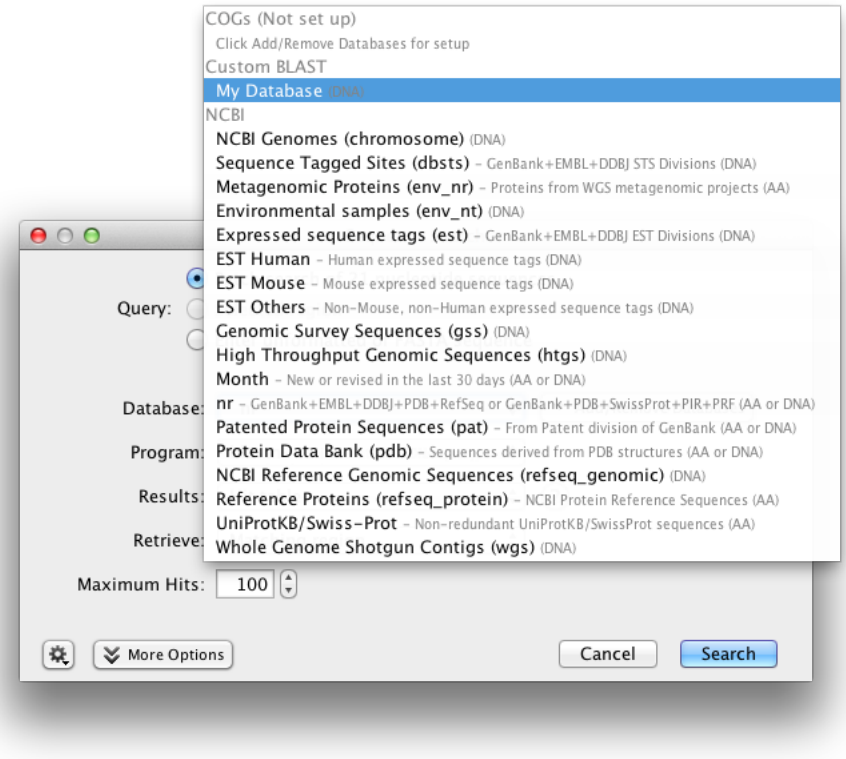


Figure 15.6: Searching a Custom BLAST database

Chapter 16

Workflows

Workflows allow you to group Operations together to reduce the number of steps required to perform an often-used combination of Operations. Options for each Operation may be preconfigured, or some or all options 'exposed' for configuration when the Workflow is run. Geneious provides a number of example Workflows for a variety of tasks that you can try. Workflows can be run or managed via the Tools menu, or if configured, via the Geneious Toolbar (See section 16.1). Workflows can be shared with other people either by exporting and importing them, or if you are connected to a shared database, by ticking the option to share them (See Figure 16.1). If you have any programming knowledge you can even add customised code in Java (See section 16.3).

16.1 Managing Workflows

Workflows can be accessed from the **Workflows** menu under the **Tools**. The Workflows menu allows you to select and run saved Workflows, manage your Workflows and create new ones. If you are going to be using Workflows routinely, then for quick access you can add Workflows to the main Geneious Toolbar. To do this, right click on the main Geneious Toolbar, choose **Customize...** and check **Workflows**.

The **Manage Workflows...** window lists all available workflows and contains options for viewing, editing, copying, deleting, exporting and importing workflows. Each Workflow listed in the Manage Workflows window with an icon next to it will be shown in the drop down under the main Workflows menu (See Figure 16.1). The order in which Workflows appear in the drop down menu can be controlled using the **Move Up** and **Move Down** buttons (see Figure 16.1). Multiple workflows can be selected using Shift- or CNTRL-click, and they can then be exported or added/removed from the drop-down list in one go.

Workflows shared by other users in a server database are indicated by an icon. These Workflows will not appear in the drop down menu unless you choose to show them. Workflows

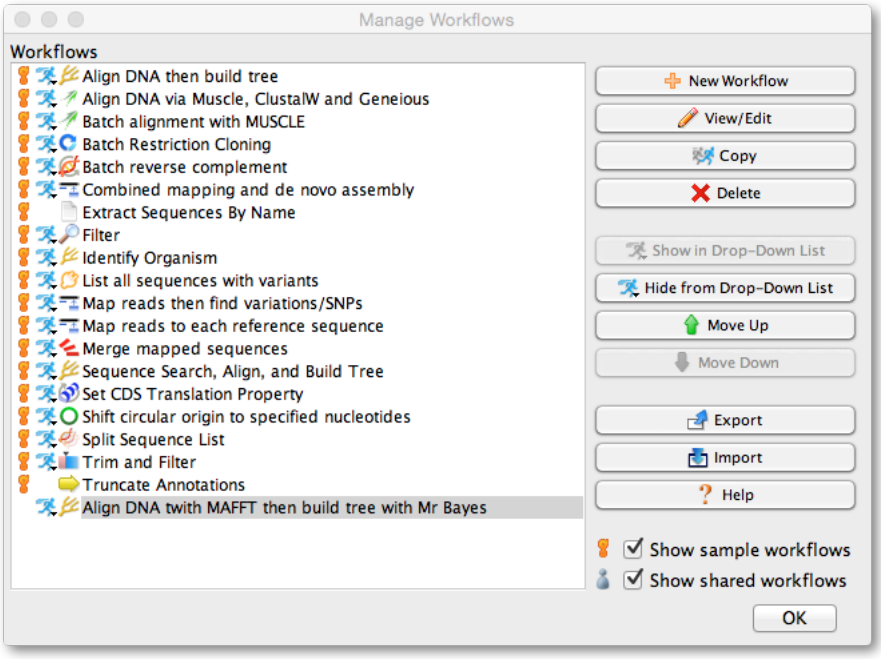


Figure 16.1: The Manage Workflows window

shared by other users in a shared database are only editable by the creator. However, you can use the **Copy** button to create your own personal editable copy of any Workflow.

16.2 Creating and editing Workflows

To create or edit a workflow, go to **Manage Workflows...** under the **Workflows** menu. Here you have the option to create a **New** Workflow, **View/Edit** an existing Workflow, or **Copy** and edit an existing Workflow. Each of these options opens an **Edit Workflow** window (See example in Figure 16.2) where you can name your Workflow, describe it's function, specify an icon for display in the Workflows menu, share via a Shared database, and build/edit your Workflow using the **Add Step**/**Delete Step** buttons.

Each Workflow is made up of one or more Steps. A Step may be an Operation (for example, perform Muscle alignment), or a special Step (for example, Group Documents). Each Step accepts one or more documents as input and produces one or more documents as output which are then used as input to the next step in the Workflow. All documents selected when the Workflow is run are provided as input to the first Step (unless specified otherwise, see section 16.2.2 for further information). The output from the final Step of a Workflow is saved in your Geneious database. Outputs from intermediate steps are not saved, unless you include the **Save Documents / Branch** option after the Step. A document is a single entry that can be

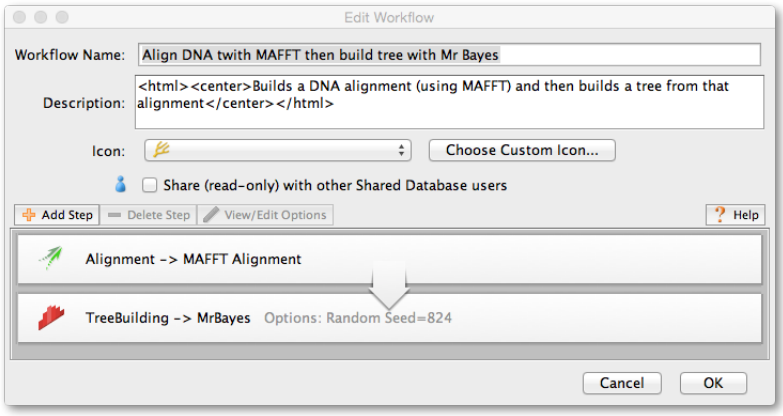


Figure 16.2: The Edit Workflow window

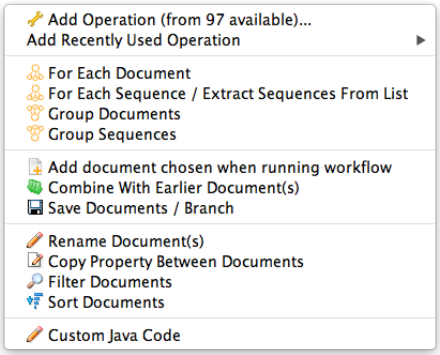


Figure 16.3: Add step Menu options for building and editing Workflows

selected from your Geneious database. For example a single document can be an alignment, a sequence list, a tree, or a stand-alone sequence.

The **Add Step** button (Figure 16.3) provides a dropdown menu with a range of Steps that can be added to your WorkFlow. The purpose/function of each type of Step is summarised in Table 16.1.

16.2.1 Configuring options for Operation and Steps

For each Operation added to your Workflow, you can edit and specify values for the configurable options available for each operation. To do this, select the Step you wish to configure, and click **View/Edit Options**. To set the Workflow up so that the options are preconfigured and cannot be changed when the workflow is run, select **Expose no options**, then select the

Table 16.1: Add Step options for Workflows

Step	Purpose/Function
Add Operation	Runs one of the standard Operations available in Geneious. This can include other Workflows
For Each Document	Runs the next Step in the Workflow independently on each document
For Each Sequence/Extract Sequences From List	Runs the next Step in the Workflow independently on each sequence, extracting sequences from a list if necessary
Group Documents	Groups results from multiple operations together so that they are all used in a single invocation of the next operation
Group Sequences	Groups sequences into a sequence list document
Add document chosen when running workflow	Prompts the user to choose a document when they first start the Workflow. This document may either be chosen from anywhere in their database, or from one of the documents selected when the workflow is run. In the latter case, the selected document will be excluded from the list of documents provided to the first Step of the Workflow. If only a single document from the selected documents matches the specified document type, then it will be automatically selected instead of asking the user. In both cases, the selected document is combined with each of the results from the previous Step in the Workflow. See section 16.2.2 for more details
Combine with Earlier Document(s)	For each result from the previous Operation, combines with the corresponding input document(s) of an earlier operation in the Workflow. The documents from the earlier Operation are added to the end of the list of documents from the previous operation
Save Documents / Branch	Optionally saves the current results. Also, optionally starts a branch beginning from an earlier Step in the Workflow
Rename Documents	Renames the document(s)
Copy Property Between Documents	Works on a pair of documents. Copies a property from one document to another document and outputs just the destination document. If only a single document is provided, outputs that document unmodified
Filter Documents	Filters some documents based on the content of their fields
Sort Documents	Sorts some documents based on the contents of their fields
Custom Java Code	Write some custom Java code to do whatever you want. See section 16.3 for more details

Operation options you want the workflow to use. To allow some or all or the options to be configured each time the workflow is run, select **Expose all options** or **Expose some options**. Exposed options can be presented in a number of ways as described below.

- **Optionally label exposed options as**: Use this to group and label all the exposed options for this Workflow Step under a labelled section rather than mixing the options for this Step in with options from other Steps in the workflow.
- **Access exposed options via button**: Rather than displaying all the options for this workflow Step in the top level options dialog, instead provide a button for the user to click on before showing the options for this Workflow Step.
- **Exposing options with dependencies**: Some options have dependencies on other options. For example when a checkbox is off, another option may become disabled. If you choose to expose a subset of the options rather than all options, the dependencies between options will be discarded. For checkbox options that have an associated value, you will probably want to expose the associated value too. Often this associated value doesn't have a label in the user interface so it will appear with its programmatic name immediately following the checkbox option in the drop down list of available options.

Note that for some Operations, not all options may be available when run in a Workflow. Some special Steps also have configurable options, which in some cases can also be exposed when the Workflow is running. For example, the **Filter Documents** Step can be exposed so that the user can set the Filter criteria when the Workflow is running.

16.2.2 Advanced document management

Grouping and separating documents

In simple Workflows all documents provided to the Workflow are grouped as a single set of documents which are used as input to a single invocation of the first Step in the Workflow. Each Workflow Step will produce one or more output documents, all of which are grouped together and used as input into the next Step in the Workflow.

However, it is possible to create Workflows where each Step in the Workflow may be invoked in parallel on different sets of documents. For example, if the first Step in the Workflow uses the **For Each Sequence / Extract Sequences From List** Step, then each input sequence is placed into a separate document and the following Step will be invoked independently on each sequence. Each call to the following Step may produce one or more documents. Each of these sets of output documents are used independently as inputs to multiple invocations of the next Workflow Step. Alternatively you could group these results together again, using **Group sequences** or **Group Documents** to use as input to a single call to the next Workflow Step. No matter what document grouping or separation (**For Each...**) Operations are used, each Step in the Workflow is always run to completion on all data sets before starting on the next Step in the Workflow.

Inputting documents into later stages of a Workflow

It is also possible to insert documents into later stages of the workflow, either from additional documents the user selects in the options when they first start the workflow (using the **Add document chosen when running workflow** Step) or to use documents generated from earlier stages in the workflow (using the **Combine With Earlier Document(s)** Step). You can create branches in your workflows by using the **Save Documents / Branch Step** often in conjunction with a **Filter Documents** Step as the first Step in each branch. For an example on branching and filtering, see the sample 'Identify Organism' Workflow.

16.3 Custom code in Workflows

Custom code allows you to create Geneious operations that do almost anything. The Workflow custom code automatically inserts the surrounding import statements for the complete Geneious API and a class framework around the methods you implement here. Additional import statements can be provided prior to the first method. Documentation for the API is available at Geneious API. For more advanced programmatic access to Geneious (for example creating importers, exporters or viewers), please download and refer to the Geneious Plugin Development Kit.

Chapter 17

Geneious Education

This feature allows a teacher to create interactive tutorials and exercises for their students. A tutorial consists of a number of HTML pages and Geneious documents. The student edits the pages and documents to answer the tutorial questions, and then exports the tutorial to submit for marking.

17.1 Creating a tutorial

Geneious Tutorials are comprised of HTML documents with linked images and geneious files. Simply create your html documents, and place them together in a folder. The first page of the tutorial should be called "index.html", and this will be loaded as the main page. Geneious will follow all hyperlinks between the pages, and external hyperlinks (beginning with `http://`) will be opened in the user's browser. If you want to include figures and diagrams in the pages, just put the image files in the same folder and reference them with `<img>` tags like a normal HTML document *(supported image formats are GIF, JPG, and PNG)*.

If you want to include Geneious documents in your tutorial, simply place them in the same folder as the html documents and they will automatically be imported into Geneious with the tutorial. If you want to link to them from the tutorial pages, create a hyperlink pointing to the file in the HTML document. For example, to create a link to the file sequence.fasta in your tutorial folder, use the HTML `<a href="sequence.fasta">click here</a>`. To open more than one document from a link, separate the filenames with the pipe *(|)* character, for example `<a href="sequence.fasta|sequence2.fasta">click here</a>`. Note that geneious files must contain only one document to be imported automatically with the tutorial.

You can add a short one-line summary by writing your summary in a file called *"summary.txt"* (case sensitive) and putting it in the tutorial folder. Make sure that the entire summary is on the first line of the file, as all other lines will be ignored.

Once you have all your files together, put the contents of the folder in a zip file with the extension *.tutorial.zip*. Subfolders within the zip file are supported in Geneious R8 and higher.

17.2 Answering a tutorial

Import the tutorial document into Geneious (use **File** → **Import** → **From file**, or drag it in). The tutorial document and any associated geneious documents will be imported into the currently selected folder. The tutorial itself will be displayed in the help pane on the right hand side of the Geneious window. If you accidentally close the help pane, you can display it by choosing **Help** from the **Help** menu.

If the tutorial requires you to enter answers, click the edit button at the top of the tutorial window and type your answer in to the space provided. Click the **Save** button when you are done.

If the tutorial has a link to a Geneious document, when you click the link the document will be opened in the document viewer. Any changes you make to this document will be preserved when you export the tutorial.

When you have finished the tutorial, export it by selecting the tutorial document and choosing **File** → **Export** → **Selected Documents** from the main menu. Make sure that **Geneious Tutorial File** is selected as the filetype, and then give it a name and click **Export**.

Chapter 18

Saving operation settings (option profiles)

Profiles allow you to save the settings for almost any analysis operation in Geneious so they can be loaded later or shared with others. Eg. the recommended trimming parameters for your organization can be saved as a profile and then shared on the Shared Database for everyone to use.

The **operation settings** button  appears in the bottom-left corner of any options window. Click on this button to reset to defaults, load a profile, save a new profile or manage your existing profiles, as described below.

Saving a profile

To create a profile, set the options up the way you want, click the operation settings button then choose **Save Current Settings**. You can then enter a name for your profile and choose whether it is shared. For a description of shared profiles see the section on sharing profiles.

When you save a profile it is attached to the particular analysis window that you have open. Eg. if you save a profile for Alignment it can only be loaded for Alignment, not for Assembly.

Loading a profile

To load a profile, click the operation settings button and choose **Load Profile** and click on the name of the profile you want to load. The settings for the operation will immediately update to reflect the profile.

Note: Sometimes when you load a profile the settings may not exactly match what was saved.

This is because the available settings can change depending on what type of documents you have selected.

Managing profiles

Click on **Manage Profiles** under **Load Profile** to see a list of profiles with options for deleting, editing, importing and exporting profiles. See sharing profiles section below for more on import and export.

Sharing profiles

There are two ways to share option profiles:

- Import and export from the **Manage Profiles** window allows you to save a file containing a particular profile. These can be emailed to other Geneious users and imported for use with their data. The easiest way to import a profile is by dragging the file directly in to Geneious.
- If a profile is marked as **Shared** (when it was created or by editing it) then the profile will be copied across to any Shared Database that you connect to. This means anyone else who connects to the same Shared Database will automatically have the profile under their Load Profile menu. *Note*: Once a profile is shared it cannot be un-shared, but it can be deleted. Also, other users can edit or delete a shared profile at any time.

Chapter 19

Collaboration

Collaboration is an external plugin.

Collaboration allows Geneious users to share the products of their research and work with each other. Based on an open Internet protocol called *XMPP* or *Jabber*, it allows you to maintain a list of contacts, so that you see who is online when you sign on yourself. You can then share documents with your online contacts, and browse and work with their documents in return. The list of contacts is stored on the server, so you can easily access an account including its contacts both at work and on your private computer.

Collaboration can work with any existing Jabber service, such as Google Talk, but we recommend using the Geneious default, `talk.geneious.com`.

You can even access several Jabber accounts at the same time, which is particularly convenient if you wish to set up and run your own Jabber server (section 19.5.3).

This chapter shows you how to:

- Create a new collaboration account
- Search for, and add contacts to your account
- Share local folders with your contacts
- Search your contacts as you would an online database
- Set up and run your own Jabber server

19.1 Managing Your Accounts

When the Collaboration plugin is installed, you will see the empty Collaboration service in the Sources Panel and the **Collaboration** submenu under **Tools**. You can open the **Add New Account** dialog by either right-clicking (Ctrl+click on Mac OS X) on Collaboration in the Sources Panel and clicking, **Add New Account** in the popup menu, or by selecting the same option from the **Collaboration** submenu.

19.1.1 Add New Account

In this dialog you are given the options of creating a new account on the server or entering the details for an existing account, e.g. if you want to access an account from an additional computer. If you choose to create a new account Geneious will attempt to automatically register your account on the server at the end of this process.

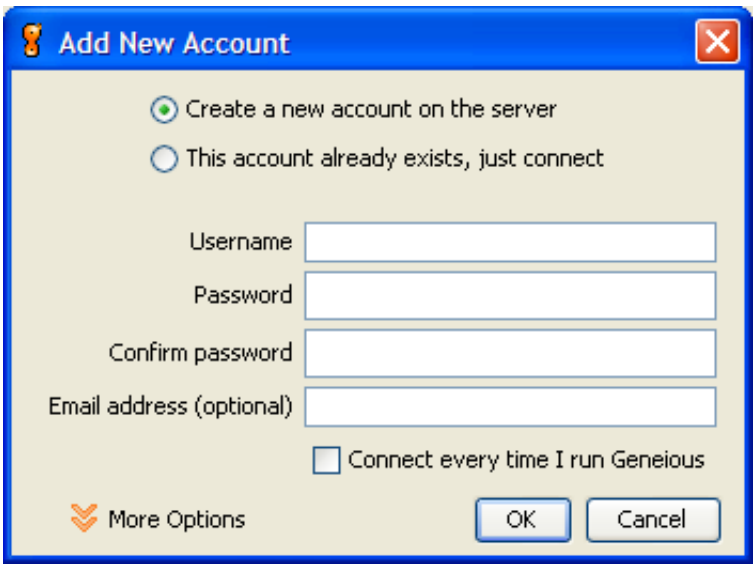


Figure 19.1: Add New Account dialog box

Choose a username and password now. Enter your password twice for a new account.

You can also optionally add an email address. Biomatters will need this if you require support regarding, e.g. reset of password or deletion of accounts.

More Options

You can change some of the defaults for new and exiting accounts:

- **Account Name** is the name displayed in the Services Panel for this account. It defaults to your username if nothing is entered
- **Server** is the server your account connects to (default: `talk.geneious.com`).
- **Jabber Service Name** is required by some other Jabber service providers, such as Google Talk. Don't enter anything here unless you know what you are doing.
- **Port Number** for Jabber servers running on a non-standard port (default: 5222).

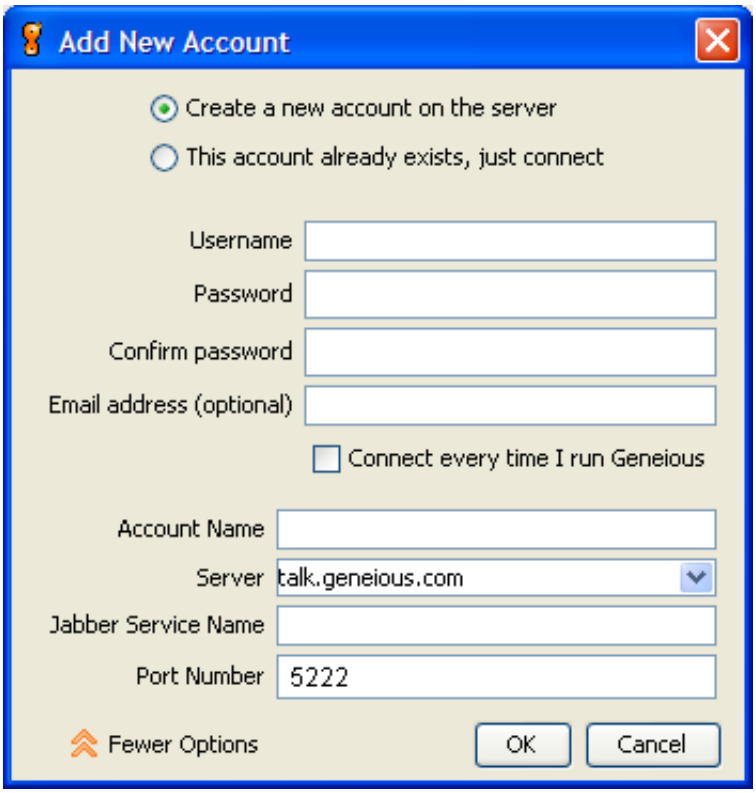


Figure 19.2: Add New Account dialog box with More Options

19.1.2 Connecting to, Editing or Removing Accounts

Account-specific options can be accessed either from the **Collaboration** submenu under **Tools**, or by right-clicking on the Collaboration account in the Sources panel to bring up the account's context menu.

To change your account configuration, including your password or email address, click **Edit Account Details**. If you change your password, Geneious will attempt to change it on the

server the next time you connect. For this purpose, Geneious internally remembers your previous password as well, so that it can still connect if you have entered your new password while disconnected.

To connect or disconnect from your collaboration account, select **Connect** or **Disconnect**.

To delete your account configuration from Geneious, select **Delete Account**. Currently, there is no option for deleting an account on the server.

19.2 Managing Your Contacts

Once you have an account and are connected you can start adding contacts. You will not be able to add contacts while an account is disconnected. Also, you will not be able to see a contact's online status until that contact has approved your request to do so.

19.2.1 Add Contact

Select your account in the Sources Panel and choose **Add Contact** from the **Collaboration** submenu or right-click (Ctrl+click on Mac OS X) on your account in the Sources Panel and choose the same option.

You will see a simple dialog with one field, Jabber ID. A Jabber ID looks like an email address and has a similar function: It uniquely identifies some other Geneious users account. You can enter a contact's Jabber ID directly into this field if you know it. To see your own Jabber ID hover your mouse over your account in the Sources Panel and it will appear in a tool-tip.

Figure 19.3: Add Contact dialog box

If the server supports it, you should also see a **Search For Contact** link. This opens a search box, with some checkboxes indicating what you are searching on. Enter all or part of the name or email of the contact you want and click the **Search** button. If any rows are returned in the results table you will be able to select one or entries and add them as contacts.

Your new contact will appear immediately in your contact list, however you will not be able

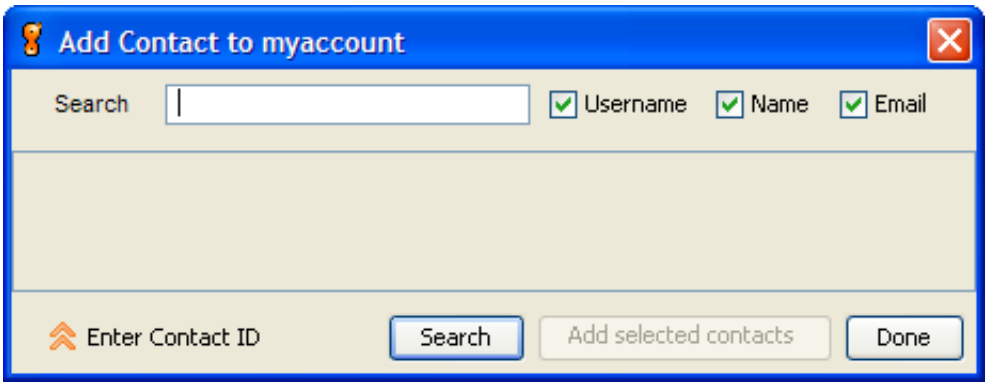


Figure 19.4: Add New Contact dialog box in searching mode

to tell whether your new contact is online until they accept you as a contact. Similarly you may occasionally see a dialog box pop up asking you, "Allow user.name@talk.geneious.com as contact?" This is another Geneious user attempting to add you as a contact in this manner.

Your contact will appear grey in your contact list when they are offline. If your contact is online, they will appear blue. A contact online in Geneious will have the orange Geneious 'G' behind them. A contact online in some other program, like a chat client, will have a speech bubble behind them.

19.2.2 Rename Contact

This option allows you to change the name that you know another contact by. This is the name the contact will appear under in the contact list and in chats; it is only visible to you.

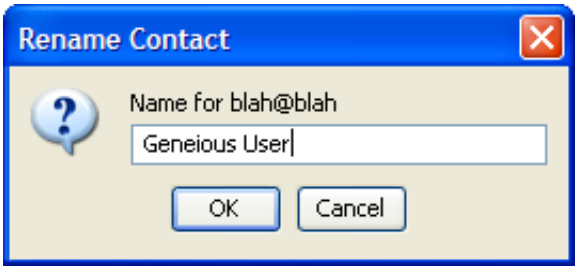


Figure 19.5: Rename Contact dialog box

19.2.3 Remove Contact

If you no longer wish to share documents with a contact, you can remove that contact by right-clicking (Ctrl+click on Mac OS X) the contact in the Sources panel and selecting **Remove Contact...**. This deletes you from their contact list as well. If you find that a contact has disappeared from your list, this may be the reason.

19.3 Sharing Documents

Select one of your local folders. Select **Share Folder** from the **File** menu. Alternatively right-click (Ctrl+click on Mac OS X) on a local folder and select the same option.

- If you share a folder all documents in that folder are shared.
- If you share a folder all sub-folders of that folder are shared.
- If you share a folder it is available to all your contacts.

19.4 Browsing, Searching and Viewing Shared Documents

Folders that your contacts have shared will appear beneath that contact just as they do in your contact's own Sources panel. You can browse these folders as you do your local folders. You can also search a shared folder just as you can a local one.

Additionally, you can search all of a contact's shared documents by clicking on the contact itself and then conducting the search. You can also search all the shared documents of all of an account's contacts by clicking on the account and conducting the search. Agents can be set up on shared folders, contacts and accounts.

You cannot search, browse or run or set up agents on a contact that is currently offline.

When you first view your contact's documents in the Document Table, the documents you see are only summaries. To view the whole document, select the summary(s) of the documents(s) you would like to view and the click the **Download** button inside the document view or just above it. **Download** options are also available in the File menu and in the popup menu when document summary is right-clicked (Ctrl+click on Mac OS X). The size of these files is not displayed in the Documents Table. You can cancel the download of document summaries by selecting **Cancel Downloads** from any of the locations mentioned above.

19.5 Chat

You can either chat with a single contact, or invite several contacts to join you in a new chat.

19.5.1 Chatting with One Contact

To start chatting with a particular contact (who may be online using Geneious or another chat client which uses the Jabber protocol), click on that contact and select **New Chat Session...** either from the **Collaboration** submenu or from the popup menu (right-click on the contact, or Ctrl+click on Mac OS X). Type your messages into the text field at the bottom of the window that pops up, and click **Send** or press the Enter key to send.

19.5.2 Chatting with Multiple Contacts

Starting a Chat Session with Multiple Contacts

To invite several contacts to join you in a new chat session, click on your account (not the contacts) and then select **New Chat Session...** from either the **Collaboration** submenu or the context menu (right-click on the account, or Ctrl+click on Mac OS X). Select the online contacts which you want to invite (you can select a range by Shift+clicking, or add contacts to the selection by Ctrl+clicking). Click 'invite' to create this new chat session.

Accepting or Declining an Invitation to Chat

When one of your contacts invites you to chat, a dialog will appear, asking you to accept or decline the chat invitation. Clicking 'Accept' will open a chat window that will allow you to chat with the contact who invited you, and with all other contacts that were invited. If you decline that invitation and enter a reason (optional), this reason will be displayed to everyone in the chat.

Sending and Viewing Messages in the Chat

The chat window displays your own and your contacts' previous messages. You can enter new messages in the field at the bottom. These messages will only be sent and become visible to your contacts once you click 'Send' or press the 'Enter' key.

To leave the chat, simply close the Chat Window.

19.5.3 Setting up and running your own Jabber server

Setting up your own Jabber server is simple and means that your documents will never leave your local network. This means that you will not have any problems with firewalls, achieve much greater download speeds, and it provides an extra security layer for the confidentiality of your documents.

Note: the communication with the Geneious Jabber server is encrypted, and that we do not log or share your data.

If you wish to set up and run your own Jabber server, we recommend using Openfire from Ignite Realtime [http://www.igniterealtime.org/projects/openfire/index.jsp] which is available for free under the Apache 2.0 Open Source License [http://www.apache.org/licenses/LICENSE-2.0.html]. Install and start the server on one computer, and then enter that computer's name or address in the **Server** field under **More Options**, when creating a new account.

Please note that Biomatters cannot provide any further support for setting up and managing your own Jabber server.

Chapter 20

Shared Databases

A Shared Database allows you to store your documents in your favorite relational (SQL) database rather than on the file system. This means that multiple users can concurrently use the same synchronized storage location without any problems. Folders in a Shared Database will show up under the **Shared Databases** icon in the Sources panel, once the user has logged into the database.

A Shared Database can be used for everything a local database is used for. This includes collaboration. Take note that unread status, agents and shared folders belong to individual users rather than the database. For example Bob may see a document as unread, but Joe will see that same document as read if he has read it.

20.1 Supported Database Systems

To use a database as a Shared Database, Geneious requires that it support transactions with an isolation level set to SERIALIZABLE. Supported databases systems include Microsoft SQL Server, PostgreSQL, Oracle and MySQL. It is possible to use other database systems if you provide the database driver, see section 20.2.1

Shared Databases have been tested using:

- Microsoft SQL Server 2008 R2
- PostgreSQL 8.4
- Oracle 11g Express Edition
- MySQL 5.1

20.2 Setting up

After a database is set up correctly, multiple users can connect to it and use it as their storage location just as if they were using their own local database.

Follow these steps to set up your database for use with Geneious.

- Install a supported database management system if you do not already have one.
- Create a new database with your desired name. Make sure that you have a user that has rights to create tables.
- Use the **Connect to a database button** to connect to your database. If the database has not been set up (usually the case if you are following these instructions) Geneious will detect this and set up the database. This will only succeed if you have permission to create tables on the database.
- Make sure any other users of the database have SELECT, INSERT, UPDATE and DELETE rights, otherwise they will not be able to use the Shared Database as intended.

There are two ways you can use your database with multiple users. The simple way is just to use the Shared Database as a shared local database. If this is all you want then you are now done with setup.

Alternatively you may want to restrict access to particular folders with groups and roles. To do this please refer to section 20.4.1.

Your database should now be ready to use with Geneious. Now all users can connect to the database by clicking on Shared Databases in the Sources panel and then clicking **Connect to a setup database**. This will bring up a dialog for the user to enter in the database details.

20.2.1 Supplying your own Database Driver

Shared Databases were designed with the supported databases in mind and packaged with database drivers for them. However Geneious allows you to supply your own jdbc database driver if you want to.

You may want to do this because you have an updated driver or because you have a driver for an unsupported database. It is not guaranteed that Shared Databases will work with another database system if you provide its driver, but it is likely that it will.

To supply your own driver open up the dialog you would normally use to connect to a database. Then click the **More Options** button.

20.3 Removing a Shared Database

To remove a Shared Database, simply right click on its top folder and choose **Remove database**.

20.4 Administration

The typical user will not have to do any administration, this section is for those in charge of the database.

20.4.1 Groups and Roles

Shared Databases support user groups and roles for managing access to documents. This means that you can restrict access of folders to privileged people. How it works is that each folder in Geneious belongs to a group. Users can belong to any number of groups and have a specified role within that group. The three roles are:

- **View** allows the user to view the contents of folders.
- **Edit** allows the user to view and edit the contents of folders.
- **Admin** allows the user special administrative functions on folders.

As of this time Geneious only uses the **Admin** role for the **Everybody** group.

By default there is only one group, the **Everybody** group. When a user logs in for the first time Geneious will put them into the Everybody group with a role of "Edit". So this means every user of the Shared Database belongs to this group with a role of "Edit" unless you enter them into the **g_user** table beforehand. You will want to give yourself the role of Admin for the Everybody group if you want to perform administrative functions within Geneious.

Unfortunately at this time there is no interface for assigning groups and roles to users. So you will need some knowledge of SQL in order to take advantage of this feature. You can create groups by adding entries into the **g_group** table in the database. Assign users groups and roles in the table **g_user_group_role**.

It is likely that if you are running in a multi user environment and taking advantage of groups and roles you will want to give only read-access of the table **g_user_group_role** to your users. This is so your users can not edit this table with SQL directly as you would do. You will also want to add all of your users into **g_user** manually so Geneious does not think that they are first time users and fail trying to insert them into the "Everybody" group due to read-only access.

20.4.2 Document Compatibility Between Versions

By default the Shared Database stores documents in the latest document format the same way the Local Database does. This can cause compatibility problems if different sets of users are using different versions of Geneious. For example if a user from Geneious 8.1 saves a new alignment document, users from Geneious 6 will be unable to read those documents.

To get around this issue the Shared Database can be configured to save documents in a format specified by an older version of Geneious if that version is still supported. In most cases this will allow for users using Geneious 8.1 or later to seamlessly share documents with users of any version back to Geneious 6.0. However be aware that if the document format has changed since the compatibility version, then users of the newer version may find their documents missing properties or bug fixes that have been added since the format changed.

This setting will apply to all users using the database with Geneious 8.1 or higher and can only be set by a user with an **Admin** role for the **Everybody** group. You will find this setting by right clicking on the root folder and choosing **Administration** and then **Set Document Compatibility Version**.

Chapter 21

Geneious Server

21.1 Introduction to Geneious Server

If your site has a Geneious Server installed you can use it to offload many of the tasks that Geneious would normally run locally on to the server, taking the processing load off your own computer. Once a job is sent to Geneious Server, it will either be processed on the server itself (a so-called standalone installation) or be handed off to a cluster running Oracle Grid Engine, LSF or PBS schedulers.

To use Geneious Server, a server-side user account is required. The server-side user account will have a server access license associated with it. Alternatively your server may have a queue licensing system, which allows a certain number of users to run jobs on Geneious Server simultaneously.

If your user account has its own access license (GSAL) then you can connect to the server and execute jobs immediately without having to wait for a queue license to become available. If your account doesn't have an access license then you can log in and submit the job to the server where it will join the queue and execute when a queue license becomes available.

21.2 Accessing Geneious Server

Assuming you have your account configured on the server, you'll need to install the necessary Geneious Server plugins. Many of your normal Geneious plugins are already server aware but there are other plugins which are different from the standard plugins, or are exclusive to Geneious Server.

Your administrator can provide you with a download location for Geneious Server plugins. You can get them either from the Geneious Server itself or they may be hosted on a network

location with the `.gplugin` files. If your institute has a web interface, get the URL from your administrator and you should see a page like figure 21.1.

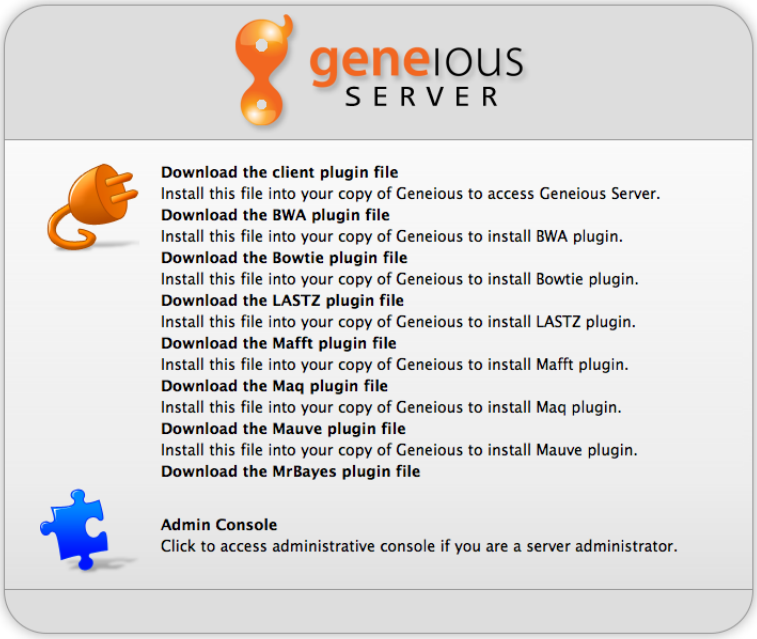


Figure 21.1: Download Geneious Server Plugins

Click on each plugin to download it and once you've downloaded all plugins, drag them from your downloads folder into Geneious. You'll probably have to restart Geneious after all plugins have been installed. Note that it may take some time for the plugins to install so give it some time. Once it is clear the plugins have all installed, restart and when Geneious comes back up you should now see the Geneious Server link in the Sources Panel. Click this and you'll see a button to log in. Use the log in button to display a dialogue requiring the hostname, username and password details which your administrator should have provided you with (Figure 21.2)

Once you've logged into the server, you will now have access to the shared database space which will appear under 'Shared Databases' in the Sources panel. We recommend you create a folder for your own documents. The benefits of this folder is that the server can see anything in there without having to get it from your Geneious client. This means large documents such as NGS sequencing data can be placed in here and the server will be able to quickly access it. Also, if you log into the server from another machine, documents you put in the Shared Database will be available unlike those of your local database. You can also see other users data so this is a good way to share your documents. This is exactly like the normal shared Databases available with Geneious, but this database is preconfigured and available as soon as you log into the server. Don't try and access it any other way using the normal shared Database plugin.

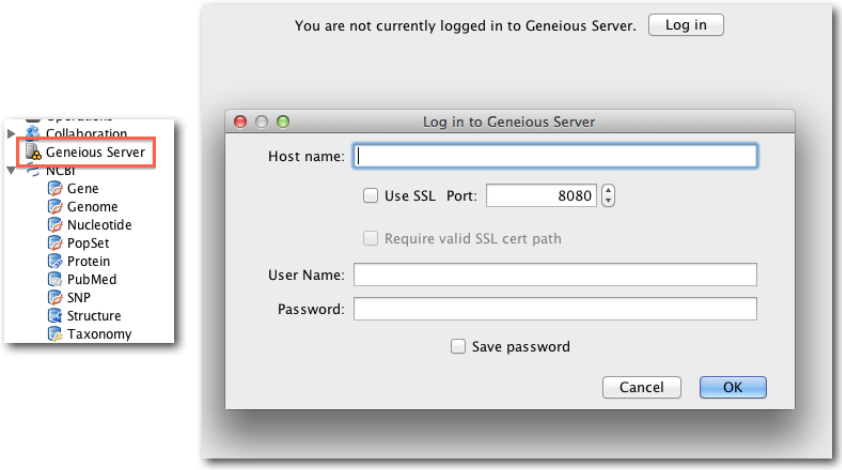


Figure 21.2: Log in to Geneious Server

21.3 Running jobs and retrieving results

Once you've logged into Geneious Server, many normal operations will now include an additional pair of buttons indicating whether the job should run on your computer or on Geneious Server (Figure 21.3). Whenever you see this choice you can choose to run the job on Geneious Server. If you're not logged in when you choose this, Geneious will prompt you to log in. The rest of the options are the same as for any local job, and the job will progress in the same way as if run locally, only using the remote resources provided by the server. If the job is likely to complete quickly, you should just run it locally but if it requires a lot of memory (more than your local machine has for instance) or if it will take a long time to process you should choose to run it on the server.

You can check the status of your job in the operations table in Geneious. You can also shut Geneious down once your job has been submitted to the server and if the job has completed when you log back in you'll be able to retrieve your results. If your jobs were running when you shut down, Geneious will request progress from the server when you restart and either show you your completed jobs, or show you the progress dialogue so you can see how far along the job has gone (Figure 21.4).

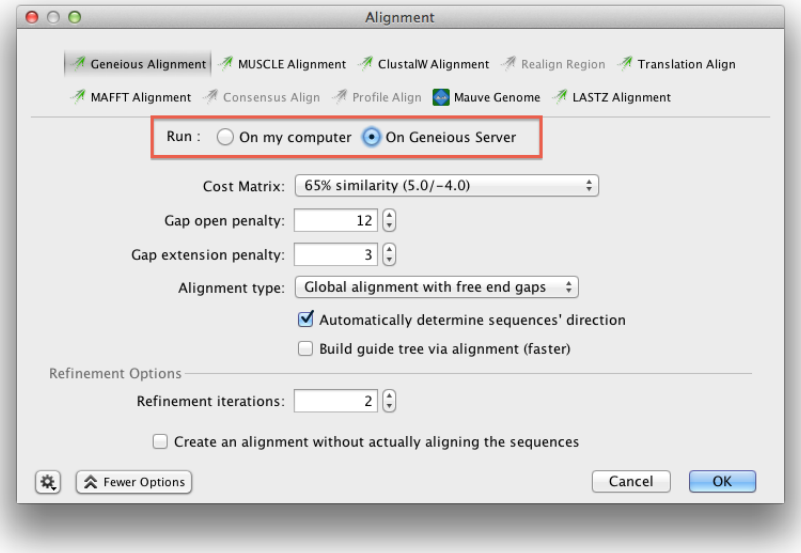


Figure 21.3: Log in to Geneious Server

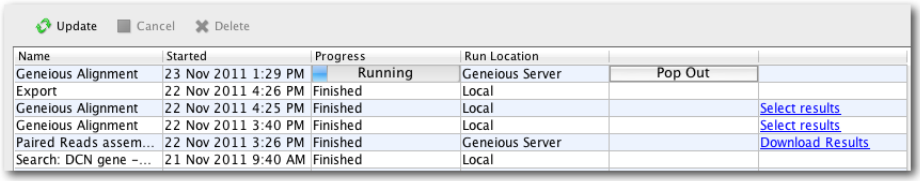


Figure 21.4: Operations table showing Geneious Server and local jobs

21.4 Geneious Server enabled plugins

This table details plugins which work with Geneious Server. Note that some of these plugins only run on Geneious Server and will not run on the local desktop implementation of Geneious.

Plugin	Local	Server
Geneious Alignment	Yes	Yes
MUSCLE Alignment	Yes	Yes
ClustalW Alignment	Yes	Yes
Realign Region	Yes	Yes
Translation Align	Yes	Yes
MAFFT Alignment	Yes	Yes
Consensus Align	Yes	Yes
Profile Align	Yes	Yes
Mauve Genome	Yes	Yes
LASTZ Alignment	Yes	Yes
Geneious Tree Builder	Yes	Yes
Consensus Tree Builder	Yes	Yes
MrBayes	Yes	Yes
PHYML	Yes	Yes
PAUP*	Yes	Yes
Geneious Assembler/Mapper	Yes	Yes
Bowtie short read mapper	Yes	Yes
BBMap short read mapper	Yes	Yes
BWA short read mapper	No	Yes
Maq short read mapper	No	Yes
SOAP2 short read mapper	No	Yes
Tophat RNAseq aligner	Yes	Yes
Velvet short read assembler	Yes	Yes
Tadpole de novo assembler	Yes	Yes
Merge Paired Reads using BBMerge	Yes	Yes
Remove Duplicate Reads using BBTools	Yes	Yes
Trim using BBDuk	Yes	Yes
Find Variations/SNPs with SAMtools	No	Yes
CustomBLAST	Yes	Yes

Chapter 22

Advanced Administration

22.1 Default data location

By default, the data location will be in the user's home directory. You can change this by setting an environment variable which will be used by the Geneious launcher such as setting a `$HOME$` variable to be where you want a user to store their data.

On Windows and Linux, edit the `Geneious.in.use.vmoptions` file in the installation directory, and add `-DdataDirectoryRoot=$HOME$/Geneious` on a new line after the other settings.

On Mac OS X, edit the `/Applications/Geneious.app/Contents/Info.plist` and find the `<key>Arguments</key>` section to match the following:

```
<key>Arguments</key>
<string>-distributionVersion
-DdataDirectoryRoot=$HOME$/Geneious</string>
```

A special `$JAVA_USER_HOME$` variable is normally used which resolves to `user.home` and is what Geneious uses by default. The program will create a `Geneious 9.0 Data` folder inside the directory you specify.

22.2 Change default preferences

22.2.1 Change preferences within Geneious

Start a fresh copy of Geneious, set it up the way you want. Shut down and then copy `Geneious 9.0 Data/user_preferences.xml` to the Geneious install directory (e.g. `C:\Program`

`Files\Geneious` on Windows XP) and rename it to `default_user_preferences.xml`

Now, when users start Geneious for the first time, they will get the configuration you set rather than the normal default.

Examples of features you can change:

- Turn off automatic updates
- Set default custom BLAST location
- Set up a shared Database
- Set up a proxy server default
- Turn off particular plugins

Any users who have already run Geneious should click the "Reset All Preferences" button in the Geneious Preferences to load these defaults.

22.2.2 geneious.properties file

Any preferences which can be set within Geneious can also be set from the geneious.properties file which can be found in the Geneious installation directory. On MacOS this file is located in Geneious.app/Contents/Java (R9 and above), or Geneious.app/Contents/Resources/app (R8 and earlier). Some examples are present in the file already- remove the hashes from the start of the lines and modify the values to use them. If you need to find out how to set other preferences using this file, please use the Support button in the Geneious toolbar to request help.

It is also possible to turn off the NCBI web services by editing this file (this is not possible via the preferences inside Geneious).

22.3 Specify license server location

This can be specified by editing the geneious.properties file in the installation directory (see 22.2.2). Scroll down to the license server settings, and change `override-property-flexnet_server.hos` and `override-property-flexnet_server.port` to the settings you require. Remove the # at the start of these lines for the setting to be used.

22.4 Deleting built-in plugins

Features of Geneious can be turned off in preferences as described in the section above on changing default preferences. If you really want to delete a feature completely so your users can't reinstate it you should shut down Geneious, go to the installation directory, into the `bundledPlugins` directory and delete the desired plugin jar files/folders.

22.5 Max memory

On Windows and Linux, edit the files `Geneious.default.64bit.vmoptions`(or `.32bit` as appropriate) and `Geneious.in.use.vmoptions` in the installation directory and change the `-Xmx` value to your preferred setting.

On Mac OS X, edit the `/Applications/Geneious.app/Contents/Info.plist` file and find the `VMOptions` section and modify the `-Xmx` setting.

It is important on Mac OS X to ensure that this value is set appropriately after an upgrade because users can often find that they have many large files in their local database preventing Geneious from starting if this value is reset to the normal default (700M on 32 bit, 1000M on 64 bit). This is an issue because the `Info.plist` file is stored in the Geneious app bundle so it gets replaced when upgrading.

22.6 Web Linking to Data in Geneious

It is possible to create web links which will open data in Geneious when they are clicked in another program. This only works on Windows and Mac OS and Geneious has to be installed on the machine where the link is clicked. Some other programs may not support these type of links.

There are two types of links supported: one which will download and import a file into Geneious and one which will select documents which are already stored in Geneious.

To create a web link that opens Geneious and imports a file from a given location, use the following form of URL:

geneious://file=<PathToFile>

eg. geneious://file=http://www.cambridgeigem.org/gbdownload/pSB1C3.gb will download and import the pSB1C3 vector from the iGEM parts registry.

The file can be of any format which Geneious is able to import (see 3.2, including plugin format. Only one file can be linked to in this way.

To create a web link that opens Geneious and selects a document which is already in your local folders or on a Shared Database, use the following form of URL:

geneious://urn=<URNofDocument>

or to select several documents:

geneious://urn=<URNofDocument1>&urn=<URNofDocument2>

To find the URN of a document, click on the small column selector button at the top-right corner of the document table and enable the URN column. You can then right-click on the URN of a document in the table and choose **Copy URN**.