

SEQUENCHER[®]

Tutorial for Windows and Macintosh

Next-Generation Sequence Alignment

© 2016 Gene Codes Corporation

Gene Codes Corporation

T C A G E N E
A G T C O D E S

Gene Codes Corporation
775 Technology Drive, Ann Arbor, MI 48108 USA
1.800.497.4939 (USA) +1.734.769.7249 (elsewhere)
+1.734.769.7074 (fax)

www.genecodes.com gcinfo@genecodes.com

Next-Generation Sequence Alignment

| | |
|---|----|
| About File Formats | 3 |
| Getting Started..... | 3 |
| Aligning Your Data with GSNAP | 4 |
| Saving Unaligned Reads From a GSNAP Run | 6 |
| Aligning your data with BWA-MEM..... | 7 |
| Reviewing the Contig in Sequencher | 10 |
| Viewing the Results in Tablet | 12 |
| Checking and Changing the Location of the Home Directory..... | 13 |
| Aligning Your Data with Maq..... | 14 |
| Conclusion | 15 |

Next-Generation Sequence Alignment

Next-Generation Sequencing requires new algorithms to process the large quantity of data produced. Whilst reads are generally shorter than those produced using capillary electrophoresis, many more reads are produced per sequencing run.

With **Sequencher**, you can choose to use **GSNAP**, **BWA-MEM**, or **Maq** to align your next-generation sequences. You can then view the contig created by **Sequencher** in the **Tablet** viewer.

Please see the [Installing DNA-Seq Tools for Sequencher](#) guide for detailed help in setting up your machine to use **Maq**, **GSNAP**, and **BWA-MEM** as well as the associated viewer, **Tablet**. **GSNAP is only supported on 64-bit operating systems.**

ABOUT FILE FORMATS

In this tutorial, you will be provided with next-generation reads in **FastQ** format. If you want to use your own data, you will need to provide the reads in **FastQ** format (which contain quality values) or **FastA** files (which do not). There are two main types of **FastQ** formats, Sanger and Illumina. You will need to know which type your FastQ files are. Note that if you are working with Illumina data that has been produced by the Casava 1.8 pipeline or later, then your data is already in Sanger **FastQ** format.

GETTING STARTED

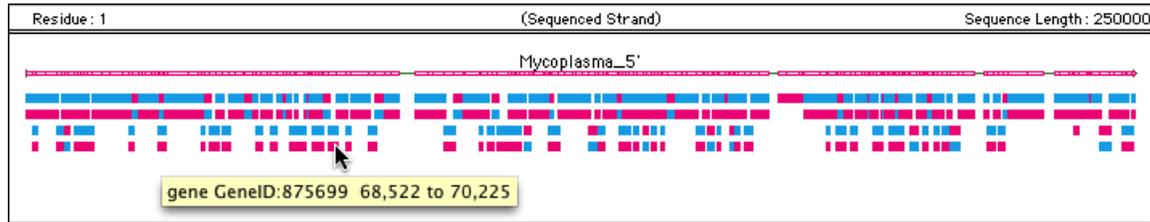
In this tutorial, you will use one of the Next-Gen algorithms in **Sequencher** to align your Next-Gen reads. You will first need to open a project. We provide a sample project that contains a reference sequence for use with the Next-Gen tutorials.

- Launch **Sequencher**.
- Go to the **File** menu and select **Import > Sequencher Project ...**
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder.
- Select the **Next Generation Sequencing** project and select **Open**.

The project contains three sequences called **Mycoplasma_5'**, **methylation_reference**, and **Excised-region**. You will be using the **Mycoplasma_5'** sequence which represents the 5 prime end of the *Mycoplasma genitalium* genome. It is already marked up with features taken from the feature table of the full-length genome.

- Double-click on the **Mycoplasma_5'** sequence to open it and then click on the **Overview** button.

You can see the features in the **Overview**. Place the cursor over a feature to see its name and location.



- Close the **Sequence Editor** window.

The first step in aligning Next-Gen sequences is to highlight your reference sequence.

- Click on the **Mycoplasma_5'** sequence in the **Project Window** to select it.
- Choose **Reference Sequence** from the **Sequence** menu.

ALIGNING YOUR DATA WITH GSNAP

Follow these steps to align your data with **GSNAP**.

- From the **Assemble** menu, select the command **Align Data Files to Ref Using > GSNAP...**

The **External Data Browser** and the **Align Using GSNAP** dialogs appear. The name of the selected sequence appears in the **Current reference seq or db** menu on the **Align Using GSNAP** dialog. You use this dialog to choose both the data files you are going to work with as well as the types of analysis you want to perform (straightforward alignment, SNP-Tolerant alignment, Methylation analysis, or RNA-Tolerant). Finally, you can also use the dialog to choose whether you want to see the results in a viewer now or review the alignment later.

When you are working with single-end data, your data will be contained in one file. In this tutorial, you are working with paired-end data so you will need to choose two files.

- Click on the **Select Reads File 1** button.
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder, then to the **NGS Data** folder.
- Select **read1.fq** and then select **Open**.
- Click on the **Select Reads File 2** button.
- Navigate to the same folder as before and select **read2.fq** and then select **Open**.
- Choose the **Sanger Standard FastQ** format from the **FASTQ Encoding** drop-down menu.

This section of the tutorial is dealing with a straightforward alignment.

- In the **Additional Analysis** groupbox, ensure the **Standard alignment** setting is selected.

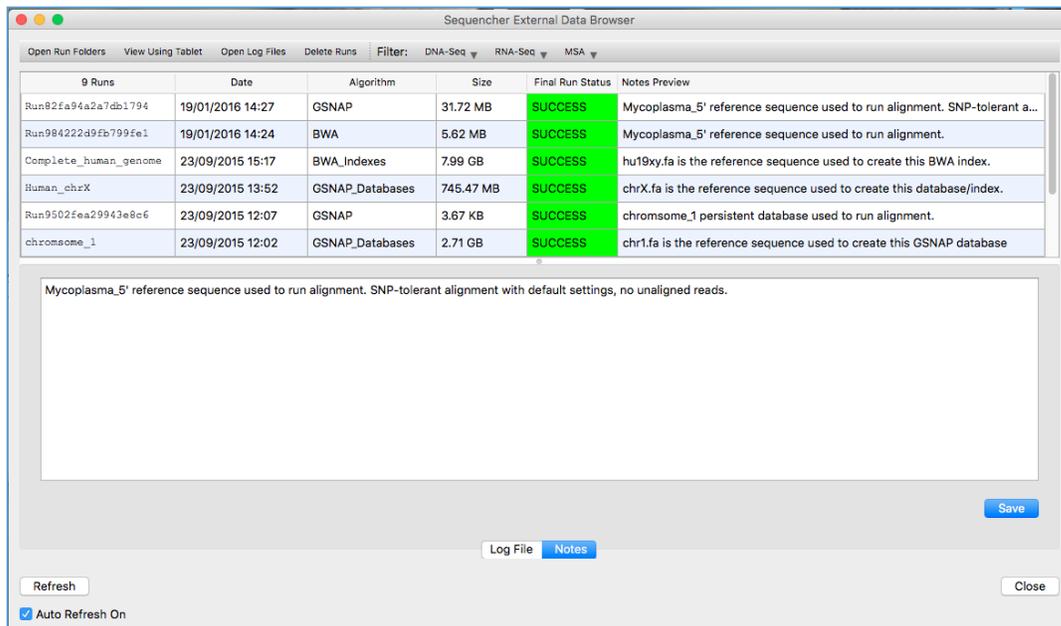
- In the **Options** groupbox, ensure the **No unaligned reads** option is selected.
- Now choose whether you want to use a viewer or not.

(Note that there are some differences when using the Viewer demo version of **Sequencher**. If you choose to view your results using Tablet before clicking the **Align** button, the alignment will proceed and your chosen viewer will open showing the contig. At the same time, a new contig will appear on the **Sequencher Project Window**. Note that if you do not choose Tablet before submitting the reads for alignment in **Sequencher's** Viewer version, you will not be able to invoke it later and will get a message to that effect.)

- Click on the **None** radio button in the View Results Using groupbox.
- Click on the **Align** button.

The alignment process begins. You will see a message saying that **GSNAP** is running. You will know the process is completed when the message disappears and a new contig appears in the **Sequencher Project Window**. If you are working with the full version of **Sequencher**, you may wish to save the project at this point. It is always good practice to save your work as often as possible.

You can also keep track of the progress of the alignment by looking at the **Log File** pane in the **External Data Browser** dialog for the run in progress. The Final Run Status column will reflect a status of **SUCCESS** when the alignment completes successfully. If the **Auto Refresh On** widget is not checked on, you can get updated progress in the Log File pane by clicking on the **Refresh** button.



SAVING UNALIGNED READS FROM A GSNAP RUN

Follow these steps to capture any unaligned reads from your **GSNAP** reference-guided alignment.

- Select the **Mycoplasma_5'** reference sequence.
- From the **Assemble** menu, select the command **Align Data Files to Ref Using > GSNAP...**

The **External Data Browser** dialog will open automatically if it isn't already. The **Align Using GSNAP** dialog will also appear. The name of the selected sequence appears in the **Current reference seq or db** menu. You use this dialog to choose the data files you are going to work with. You can also save unaligned reads on their own or in addition to the aligned reads.

- Click on the **Select Reads File 1** button.
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder, then to the **NGS Data** folder.
- Select **read1.fq** and then select **Open**.
- Choose the **Sanger Standard FastQ** format from the FASTQ Encoding drop-down menu.
- Choose the **Standard alignment** option in the **Additional Analysis** groupbox.
- Choose **Aligned reads as SAM, unaligned reads as FastA/FastQ** in the **Options** groupbox.

Input Data Files

Mycoplasma_5' Current reference seq or db

Select Reads File 1 /Applications/Sequencher 5.4...ample Data/NGS Data/read1.fq

Select Reads File 2 Optional

Sanger Standard - FASTQ Encoding

Additional Analysis

Standard alignment

SNP-Tolerant alignment

Known SNPs File

Methylation stranded. 5'->3' both strands, no reverse complement

Methylation non-stranded. 5'->3' both strands, with reverse complement

RNA-Tolerant stranded. 5'->3' both strands, no reverse complement

RNA-Tolerant non-stranded. 5'->3' both strands, with reverse complement

Options

Unaligned reads only as FastA/FastQ

Aligned reads as SAM, unaligned reads as FastA/FastQ

No unaligned reads

- Click on the **None** radio button in the View Results Using groupbox.
- Click on the **Align** button.

A new contig appears in the **Project Window**.

- Right-click the contig and select the **Open External Data Folder** menu item.
- The Run folder is opened for you.

You will see file called UNALIGNED.1 in the appropriate Run folder that contains the unaligned reads. You can import these reads into **Sequencher** by dragging the file onto the **Project Window**.

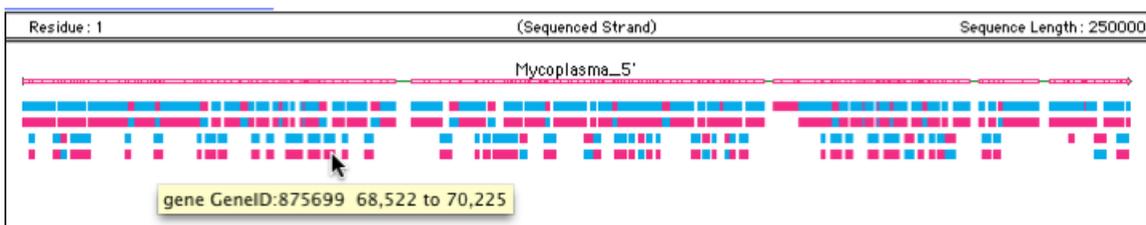
ALIGNING YOUR DATA WITH BWA-MEM

In this part of the tutorial, you will use **BWA-MEM**, another of the Next-Gen algorithms in **Sequencher**, to align your Next-Gen reads to a reference sequence. **BWA-MEM** does not have as many options as **GSNAP** but is a very fast alignment algorithm with powerful gap control parameters. You will first need to open a project. We provide a sample project that contains a reference sequence for use with the Next-Gen tutorials.

- Launch **Sequencher** if it is not already running.
- Go to the **File** menu and select **Import > Sequencher Project...**
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder.
- Select the **Next Generation Sequencing** project and select **Open**.

The project contains three sequences called **Mycoplasma_5'**, **methylation_reference**, and **Excised-region**. You will be using the **Mycoplasma_5'** sequence which represents the 5 prime end of the Mycoplasma genitalium genome. It is already marked up with features taken from the feature table of the full-length genome.

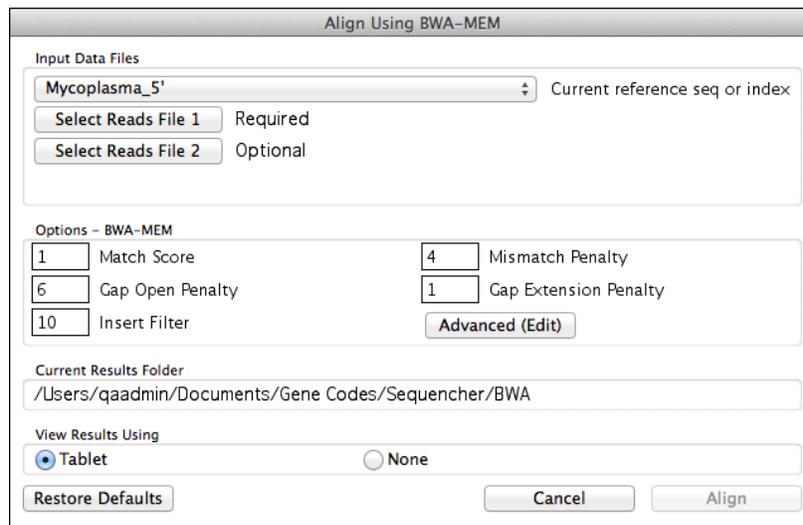
- Double-click on the **Mycoplasma_5'** sequence to open it and then click on the **Overview** button. You can see the features in the **Overview**.
- Close the **Sequence Editor** window.



The first step in aligning Next-Gen sequences is to highlight your reference sequence.

- Click on the **Mycoplasma_5'** sequence in the **Project Window** to select it.
- Choose **Reference Sequence** from the **Sequence** menu to make the sequence a reference sequence.
- From the **Assemble** menu, select the command **Align Data Files to Ref Using > BWA-MEM...**

The **External Data Browser** dialog will open automatically if it isn't already. The **Align Using BWA-MEM** dialog will also appear. The name of the selected sequence appears in the **Current reference seq or db** menu. You can set the parameters which control how gaps are treated by the alignment algorithm here. There is a penalty for opening the gap and another for extending it. This second parameter is usually set at a much lower value than the first. Having two parameters ensures that short gaps are not penalised as much as long gaps. If you choose a viewer before clicking the **Align** button, the alignment will proceed and your chosen viewer will open showing the contig. At the same time, a new contig will appear in the **Sequencher Project Window**.



You use this dialog to choose both the data files you are going to work with as well as change any parameters with which you are going to work. Finally, you can also use the dialog to choose whether you want to see the results in a viewer now or review the alignment later.

When you are working with single-end data, your data will be contained in one file. In this tutorial, you are working with paired-end data so you will need to choose two files.

- Click on the **Select Reads File 1** button.
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder, then to the **NGS Data** folder.
- Select **read1.fq** and then select **Open**.
- Click on the **Select Reads File 2** button.
- Navigate to the same folder as before and select **read2.fq** and then select **Open**.

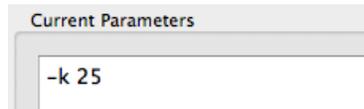
If you wish to explore the advanced parameters or wish to change some of them, follow these steps.

- Click on the **Advanced (Edit)** button.
- Turn on the **-k (Minimum seed length)** parameter by clicking its checkbox.

- Change its value to **25**.

| Argument | Value | Description |
|--|-------|-------------------------------------|
| <input checked="" type="checkbox"/> -k | 25 | Minimum seed length. The length ... |

- Note how the new values have been added to the Current Parameters preview window.



- Click on the **OK** button to dismiss this dialog.

You are now almost ready to initiate the alignment. You need to decide whether you want to view your results now or later. In this example, we are assuming that we want to look at the results now.

- Click on the **Tablet** radio button in the **View Results Using** groupbox.
- Click on the **Align** button.

The alignment process begins. You will see a message saying that **BWA** is running. You will know the process is completed when the message disappears and a new contig appears in the **Sequencher Project Window**.

The **Tablet** viewer opens automatically and your aligned reads are loaded. Once you have clicked on the contig listed in the left-hand side of the **Tablet** window, you will be able to view your reads. By setting the Minimum seed length, you have increased the stringency of the alignment. This can be readily demonstrated with the following steps.

- Click on the **Mycoplasma_5'** sequence in the **Project Window** to select it.
- From the **Assemble** menu, select the command **Align Data Files to Ref Using > BWA-MEM...**
- Click on the **Select Reads File 1** button.
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder, then to the **NGS Data** folder.
- Select **read1.fq** and then select **Open**.
- Click on the **Select Reads File 2** button.
- Navigate to the same folder as before and select **read2.fq** and then select **Open**.
- Click on the **Advanced (Edit)** button.
- Turn off the **-k** (Minimum seed length) parameter by clicking its checkbox, this will remove the check mark.
- Click on the **OK** button to dismiss this dialog.
- Click on the **Tablet** radio button in the View Results Using groupbox.
- Click on the **Align** button.

As before, the alignment process begins. You will see a message saying that **BWA** is running and you will know the process is completed when the message disappears and a new contig appears in the **Sequencher Project Window**. The **Tablet** viewer opens automatically and your aligned reads are loaded. Once you have clicked on the contig listed in the left-hand side of the **Tablet** window, you will be able to view your reads. This time there are more reads in your alignment, 149728 compared to 146781 in the previous alignment run.

This is not the only advanced way you can affect your alignments with **BWA-MEM**. Explore the other parameters and see what works with your data.

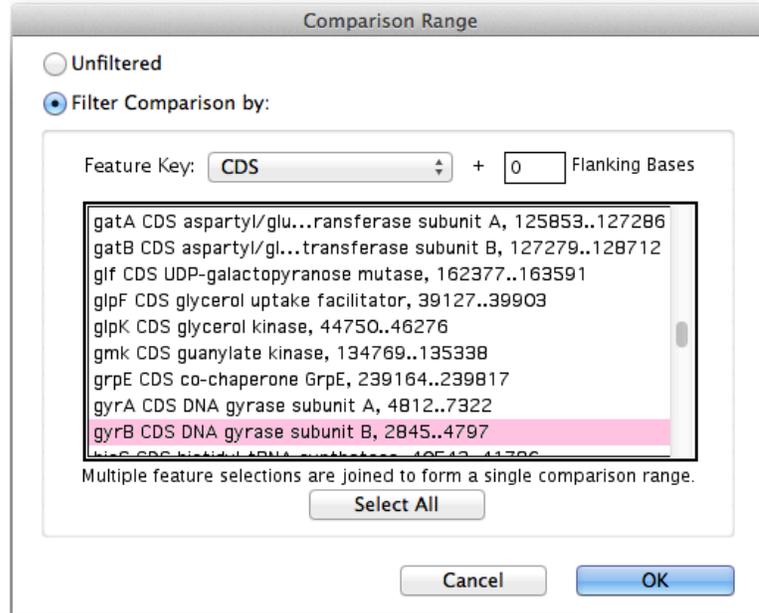
REVIEWING THE CONTIG IN SEQUENCHER

As well as being able to review the contig in an external viewer, you can also view the consensus sequence that has been aligned against the reference sequence in **Sequencher**. You have already seen how to open the contig by double-clicking on it. Now open the **Contig Editor**.

- Double-click on the contig to open it in the **Contig Editor**.
- Click on the **Bases** button to switch the **Contig Editor** from **Overview** view to **Bases** view.

You will see that the first three bases are not present in the consensus. This is due to the lack of coverage in this region.

- Close the **Contig Editor** window.
- Ensure that the contig is still selected in the **Project Window**.
- Go to the **Contig** menu and choose **Compare Consensus To Reference**.
- A new window opens containing a **Variance Table**.
- Click on the **Comparison Range** button on the button bar of this **Variance Table**.
- In the dialog that appears, click on the **Filter Comparison by** radio button.



- Scroll down the list of features until you reach **gyrB CDS DNA gyrase subunit B**.
- Select that feature.
- Click on the **OK** button.

The table shrinks in size from 179 differences between the two sequences down to two differences.

- Double-click on the **T** in the cell opposite reference sequence position 2860.
- Now click on the **Translation** button.

A new **Variance Table**, the **Translated Variance Table**, appears. Looking at the reference sequence position 2860, you can see that the codon at this position is AAA which gives rise to a Lysine in translation. In the consensus sequence, the codon is TAA which gives rise to a stop in the protein polypeptide. Notice that, if you click on the TAA in the cell opposite reference sequence position 2860 in the **Translated Variance Table**, the bases representing the stop codon are highlighted in yellow in the **Contig Editor** window in both the reference sequence and the consensus sequence of the contig. The reference sequence has features marked on it and the stop codon occurs within a feature, in this case, in the gyrB CDS feature. You can see this by looking in the **Feature Listing** window for the reference sequence by doing the following:

- Click on the **Mycoplasma_5'** reference sequence in the **Project Window**.
- Go to the **Sequence** menu and choose **Feature Listing**.
- Scroll down in the feature listing window until you locate the feature with a range of 2845 to 4797. Since position 2860 falls within this range, the codon falls within the gyrB CDS feature.
- Now close all of the windows except the **Project Window**, saving your project if you wish.

VIEWING THE RESULTS IN TABLET

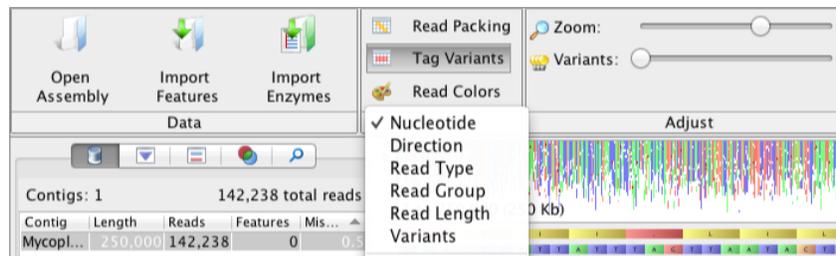
Sequencher will automatically place the results of the alignment in a **SAM/BAM** format file in the **External Data Home** directory. There will also be some log files. You can view the results of your **BWA**, **Maq**, or **GSNAP** alignment using the **Tablet** viewer. In this section, you will view the aligned reads from a **GSNAP** alignment in **Tablet**. **Tablet** has a very rich set of features, some of which are explored in this section.

- Highlight the contig in the **Project Window** whose alignment you wish to explore by clicking on it once.
- From the **Contig** menu, choose **Show NGS Data Using > Tablet**.

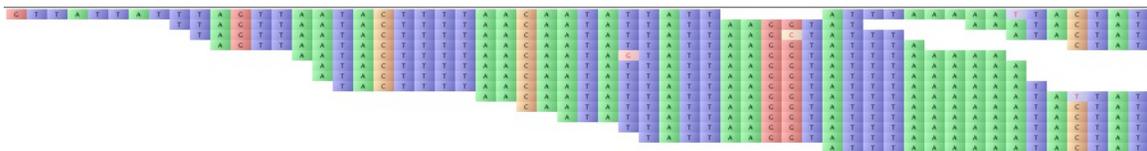
The **Tablet** viewer opens and loads the aligned reads automatically.

- Click on the contig in the list on the left-hand side of the viewer to display the aligned reads.

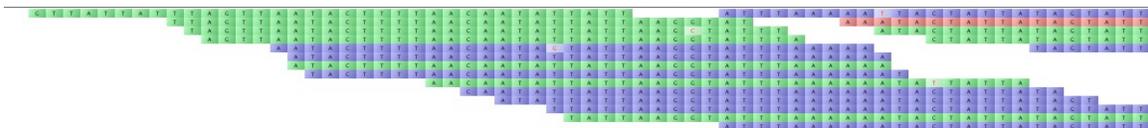
You can change the appearance of the view by clicking on one of the buttons (the look may vary with the version of **Tablet** you are using).



The reads will be displayed on the right-hand side of the window.

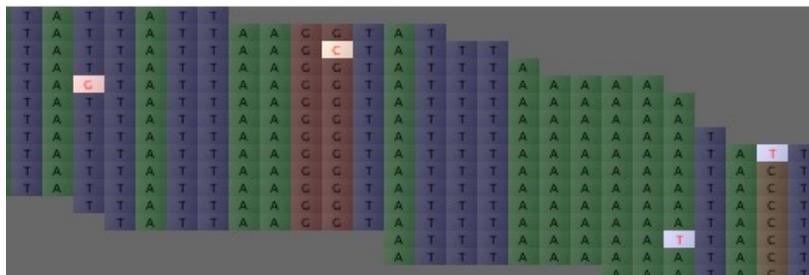


- Click on the **Read Type** button on the **Color Schemes** tab. The display changes from the default bases view to the view seen below.



In the image above, the green reads are forward reads, the blue are reverse reads, and red reads are unpaired. Depending on your version of **Tablet** and your color settings, the colors may vary.

Tablet has a nice feature that can help to highlight potential SNPs even more. Locate the Zoom and Variants sliders on the Home tab. Notice the effects as you move these sliders right and left. As you move the Variants slider to the right, the reads view becomes darker and bases are obscure except for those bases that are SNPs.

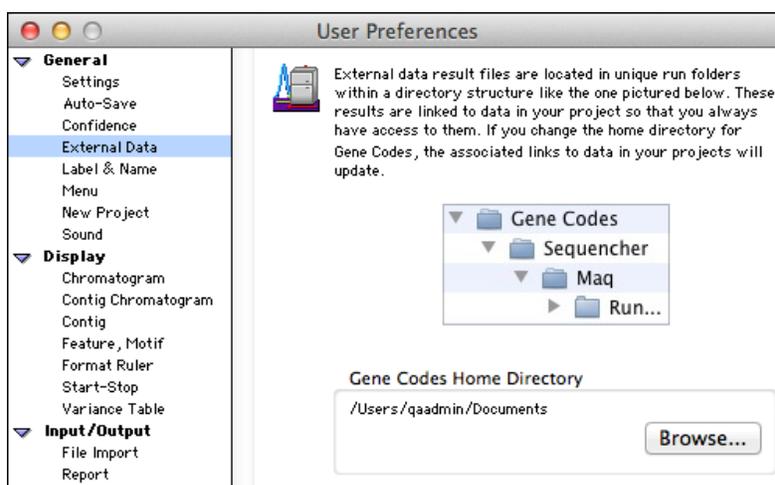


When you have finished working with **Tablet**, quit the program. You can review your alignments again by using the **Contig** menu and choosing **Show NGS Data Using > Tablet**.

CHECKING AND CHANGING THE LOCATION OF THE HOME DIRECTORY

Sequencher saves the results of any alignment or analysis in the Gene Codes Home directory. This is a default folder that is usually located within your **Documents** folder. To check the location of the directory, do the following:

- Go to **Sequencher's** user preference dialog by selecting **Window->User Preferences...**
- Select the **External Data** item.



To change the location of the Home directory, do the following:

- Click on the **Browse...** button.
- Browse to a new location.

- Select a folder.
- To confirm the new location, click on the **OK** button on Windows or the **Choose** button on Mac.

The Gene Codes Home Directory location on the **External Data** preference pane will be updated to reflect the new Home directory location.

- Close out of User Preferences.
- Quit **Sequencher**.

ALIGNING YOUR DATA WITH MAQ

In this section of the tutorial, you will use the **Maq** algorithm to align Next-Generation reads that you can view in an external viewer. You will also be able to view the consensus sequence aligned to the reference sequence.

- Launch **Sequencher**.
- Go to the **File** menu and select **Import > Sequencher Project ...**
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder.
- Select the **Next Generation Sequencing** project and select **Open**.
- Click on the **Mycoplasma_5'** sequence in the **Project Window** to select it.
- From the **Assemble** menu, select the command **Align Data Files to Ref Using > Maq...**

The **External Data Browser** dialog will open automatically if it isn't already. The **Align Using MAQ** dialog will also appear. You use this dialog to choose both the data files you are going to work with as well as the types of analysis you want to perform (straightforward alignment or alignment with SNP analysis). Finally, you can also use the dialog to choose whether you want to see the results in a viewer now or review the alignment later.

When you are working with single-end data, your data will be contained in one file. In this tutorial, you are working with paired-end data so you will need to choose two files.

- Click on the **Select File 1** button.
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder, then to the **NGS Data** folder.
- Select **read1.fq** and then select **Open**.
- Click on the **Select File 2** button.
- Navigate to the same folder as before and select **read2.fq** and then select **Open**.

This section of the tutorial is dealing with a straightforward alignment so you will not be using the Additional Analysis option. You now need to choose whether you want to use a viewer or not.

(Note that there are some differences when using the Viewer demo version of **Sequencher**. If you choose to view your results using Tablet before clicking the **Align** button, the alignment will proceed and your chosen viewer

will open showing the contig. At the same time, a new contig will appear on the **Sequencher Project Window**. Note that if you do not choose Tablet before submitting the reads for alignment in **Sequencher's** Viewer version, you will not be able to invoke it later and will get a message to that effect.)

- Click on the **Tablet** radio button in the View Results Using groupbox.
- Click on the **Align** button.

The alignment process begins. You will see a message saying that **Maq** is running. The length of time it takes to do the alignment depends on the amount of RAM, the speed of your processors, and of course, the size of the reference sequence and number of reads in your data files. You will know the process is completed when the message disappears and a new contig appears in the **Sequencher Project Window**. In addition, the **Tablet** viewer opens and loads the aligned reads automatically.

- To display the aligned reads, click on the contig in the list on the left-hand side of the viewer.

If you wish, you may save the project at this point. It is always good practice to save your work as often as possible.

CONCLUSION

In this tutorial, you have worked with two separate programs for aligning Next-Generation sequences to a reference sequence. You have learned how to use these programs to align your reads, capture unaligned reads and view the results in both **Tablet** and **Sequencher**.

For more information on using **Sequencher**, this tutorial and others are a good place to start. You can also read the manual or consult our website by visiting www.genecodes.com.