SEQUENCHER®

## Tutorial for Windows and Macintosh

# Multiplex IDs with BWA, GSNAP, and Velvet



Gene Codes Corporation
TCA GENE
AGT CODES

Gene Codes Corporation
525 Avis 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

# Multiplex IDs with BWA, GSNAP, and Velvet

# Multiplex IDs with BWA, GSNAP, and Velvet

Multiplex ID is a method for maximizing the use of sequencing reagents and reducing costs by sequencing a number of different samples at the same time. Each sample is tagged with a unique DNA identifier (barcode) so that after the sequencing run, the reads for that sample can be separated from other reads. This process is called de-multiplexing or binning. **Sequencher** includes the necessary tools not only to perform this separation so that the reads are placed in separate files according to their barcode but also to run each file using the algorithm you have chosen. **GSNAP** is only supported on 64-bit operating systems.

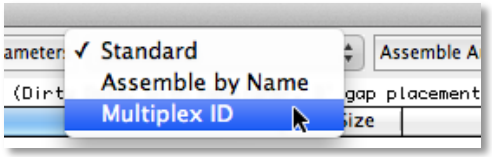
## About File Formats

As with regular alignment or assembly, your reads may be in **FastQ** or **FastA** format. As you are working with Next-Generation data, you don't need any reads to be imported into **Sequencher**; instead, you tell **Sequencher** where to find the files it needs.

You will also need to create a barcode file using a simple text editor. Each line should contain a single barcode name and barcode sequence separated by a tab.

## Aligning Multiplex ID Data with GSNAP

- Launch **Sequencher**.
- Set the **Assembly Mode** to **Multiplex ID**.

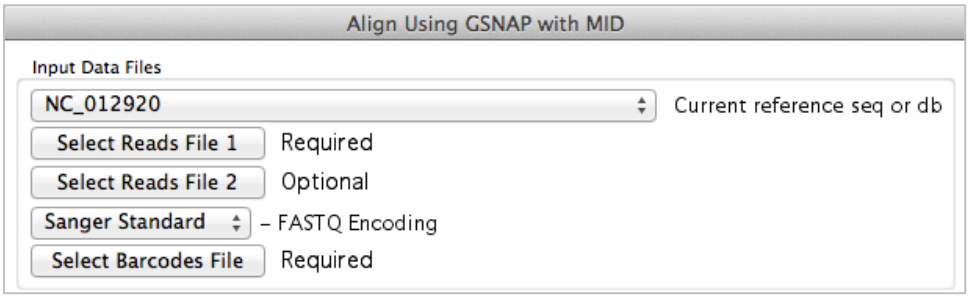



- Go to the **File** menu and select **Open Project…**
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder and choose the project called **Multiplex ID.SPF**.
- Click on the **Open** button.

The **Sequencher** project opens and contains a single reference sequence called NC_012920. This sequence is the complete human mitochondrion genome.

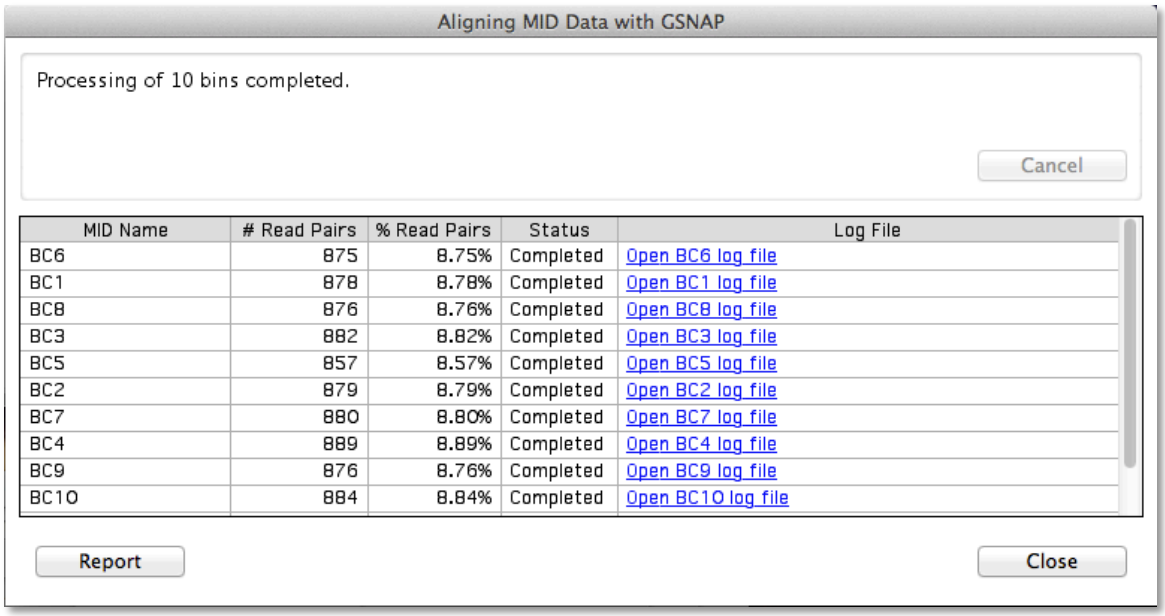
- Select the **NC_012920** sequence in the **Project Window**.
- Go to the **Assemble** menu and select **Align Data Files To Ref Using>GSNAP…**

The **Align Using GSNAP with MID** and **External Data Browser** dialogs appear. The **External Data Browser** dialog is where you will go to monitor the progress/status of the assembly. In the **Align Using GSNAP with MID** dialog, the name of the selected sequence appears in the **Current reference seq or db** menu.




- Click on the **Select Reads File 1** button.
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder, then to the **Multiplex ID Data** folder.
- Choose **MID-Reads1.fasta** and click on the **Open** button.
- Click on the **Select Reads File 2** button.
- The file picker will automatically take you to the last location you chose.
- Choose **MID-Reads2.fasta** and click on the **Open** button.
- Now click on the **Select Barcodes File** button.
- The file picker will automatically take you to the last location you chose.
- Choose **MID-list.txt** and click on the **Open** button.
- Now click on the **Align** button.

You can follow the progress of your analysis on the **Aligning MID Data with GSNAP** dialog. The top part of the dialog shows you the progress of separating your reads into their different files. The lower part of the dialog contains a status table that shows the progress of alignment for each sample. It also shows how many reads are in each file and what this represents in terms of the overall number of reads. A status column will show whether a file is in the process of being aligned or whether alignment has been completed. Once the run is complete, the **Report** and **Close** buttons will become enabled.




## Reviewing Log File

The final column in the status table lists the log files for each file.

- Click on one of the log file links to open the log file.

The log file contains details on the alignment run. Note that the top of the file contains details of the run including the sequence of the barcode for this particular run.

```
Start time: 2015-Nov-18 12:18:48

Using GSNAP version "2015-09-29".
Reference Sequence: NC_012920
Original reads file(s):
Reads file 1: /Applications/Sequencher 5.4.1/Sample Data/Multiplex ID Data/MID-Reads1.fasta
Reads file 2: /Applications/Sequencher 5.4.1/Sample Data/Multiplex ID Data/MID-Reads2.fasta
Intermediate reads file(s):
Reads file 1: /var/folders/xg/h3t9pwjs43179wmxgqr6wjmr0000gn/T/Bins21f0a2946727fa86/1/BC6.fa
Reads file 2: /var/folders/xg/h3t9pwjs43179wmxgqr6wjmr0000gn/T/Bins21f0a2946727fa86/2/BC6.fa
Quality encoding: Sanger Standard
Additional analysis: Standard alignment
Unaligned reads option: No unaligned reads
Insert threshold: 10
MID: BC6
Barcode: TCAGATATCGCGAG
```

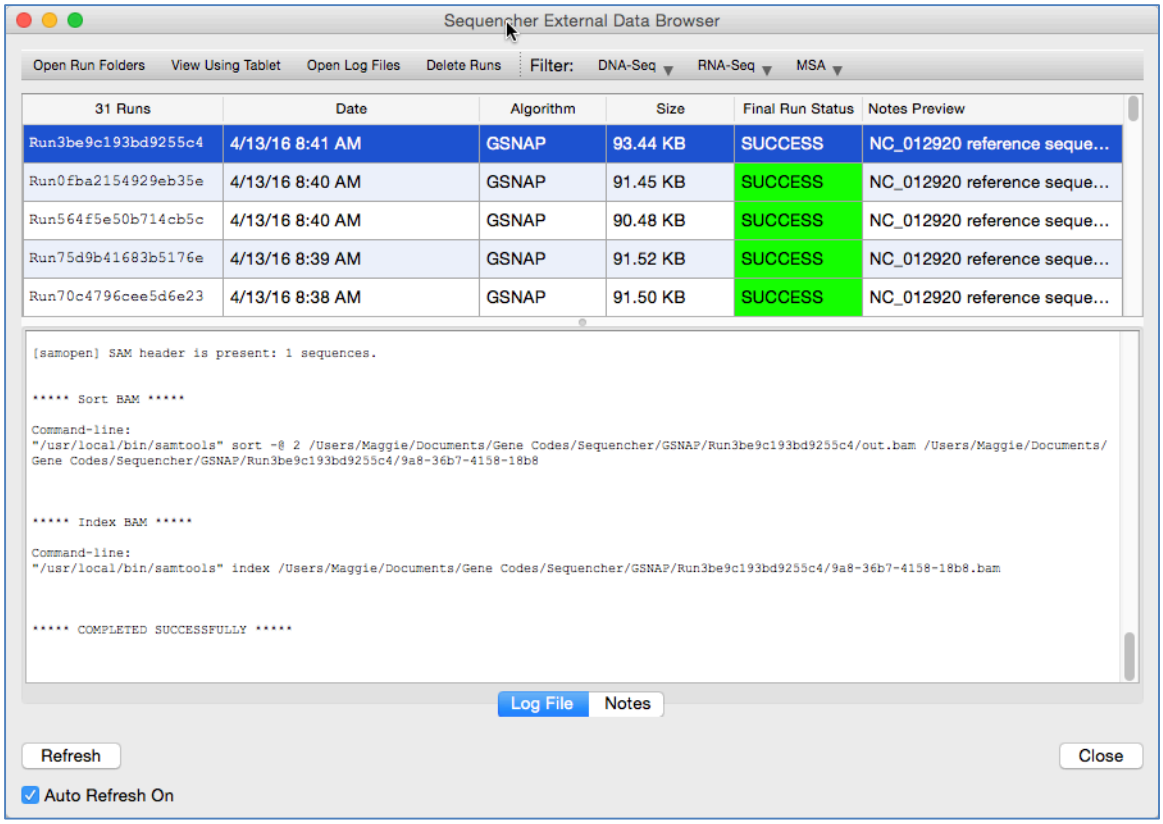
```
Start time: 2012-Oct-02 11:32:00

Using GSNAP version "2011-03-28".
Reference Sequence: NC_012920
Reads File1: /Applications/Sequencher 5.1/Sample Data/Multiplex ID Data/MID-Reads1.fasta
Reads File2: /Applications/Sequencher 5.1/Sample Data/Multiplex ID Data/MID-Reads2.fasta
MID: BC6 Barcode: TCAGATATCGCGAG
```

- Close the log file when done looking at it.

Note how many results are returned to **Sequencher's Project Window**. Each result is a contig containing the reference sequence and the consensus of a contig created by the alignment.

You can also monitor the progress of your assembly using the **External Data Browser** dialog. You can watch as the Log File pane updates. Notice that, once your alignment is complete, the status field updates to SUCCESS.

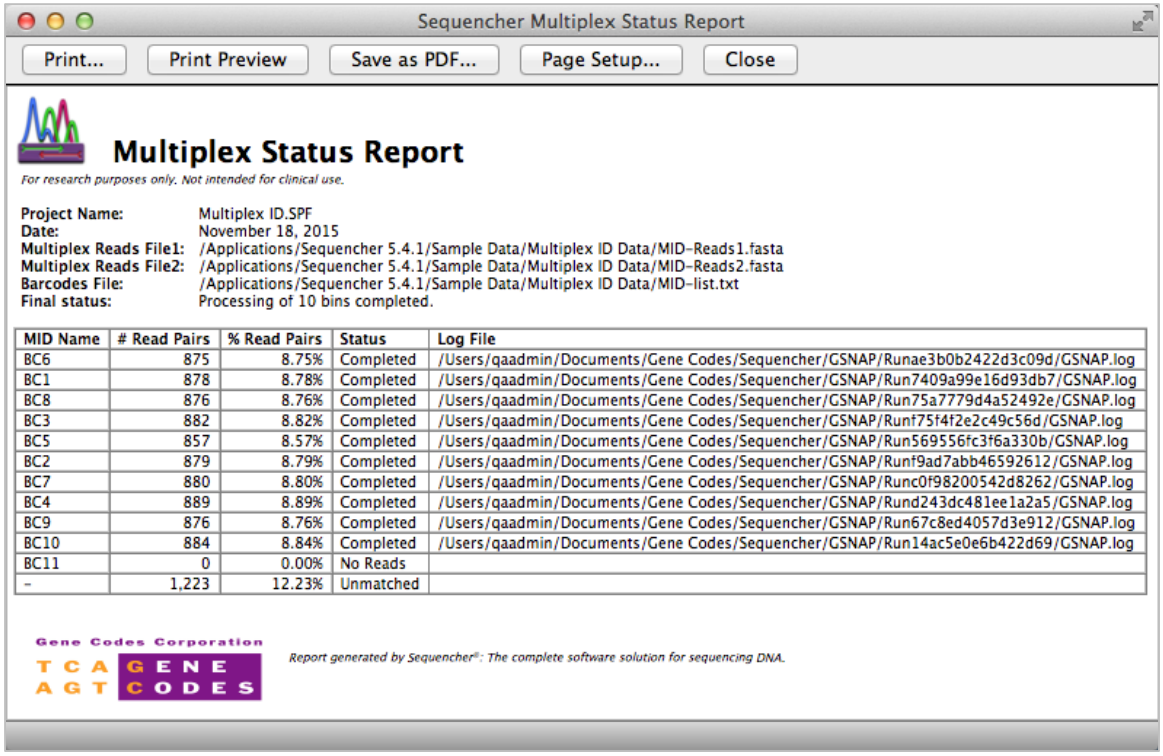



## Printing Multiplex Status Report

To write out the status information contained in the **Aligning MID Data with GSNAP** dialog, do the following:

- Click on the **Report** button at the bottom of the dialog.

This creates a preview window called **Sequencher Multiplex Status Report** that will contain the same information as the status table but also includes added information such as the date of the run and the location of the data files. This window has a button bar across the top. From this button bar, you can save the report as PDF, change the **Page Setup,** and **Print** the status report.




- Close the status report window.
- Close the **Aligning MID Data with GSNAP** dialog.

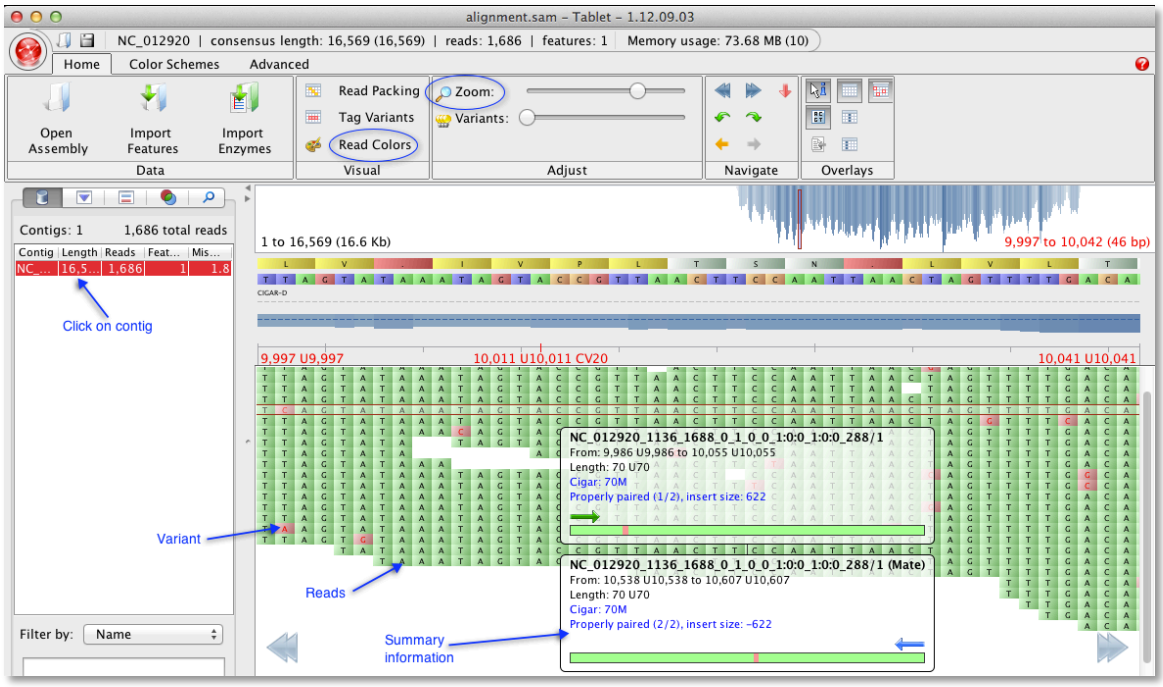
## Reviewing Your Results in Tablet

If you want to review your data in **Tablet**, click on an item of interest in the **Project Window**.

- Go to the **Contig** menu.
- Choose **Show NGS Data Using>Tablet**.

**Tablet** will launch and load your data and reference sequence automatically. On the left-hand side of the **Tablet** window, you will see a list of contig(s).

- Click once on the contig of choice (there may only be one).
- The reads will load and appear on the right-hand side of the window.
- Zoom in and out using the slider at the top of the window (top blue oval).




- Click on **Read Colors** (lower blue oval).
- Choose a class of **Read Colors** from the drop-down menu.

You will see a list of read classes in the drop-down menu such as **Nucleotide** which colors the individual bases such that all A's will be in one color and all T's will be in a different color and so on. Choosing **Direction** shows the direction of each read with coloring of the forward reads in one color and coloring of the reverse reads in a different color. **Read Group** refers to paired-end reads and **Variants** highlights any variants. In more recent versions of **Tablet**, you can customize the colors for these different classes. The view in the image above includes customized colors (pink backgrounds indicate variant bases).

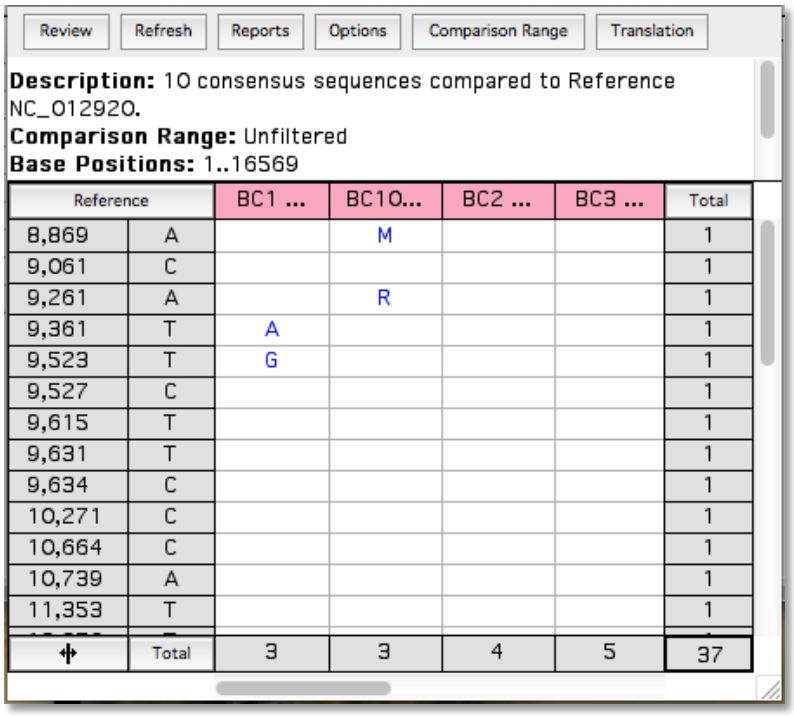
- Close the **Tablet** viewer.

## Reviewing in Sequencher

Since you have aligned your data against the same reference sequence, you can use the **Variance Table** to review the number of differences between samples.

- Select the contigs you wish to review.
- Choose the **Contig** menu and select **Compare Consensus to Reference**.

A **Variance Table** is generated and with it you can compare the position and number of variants between each contig and the reference sequence.

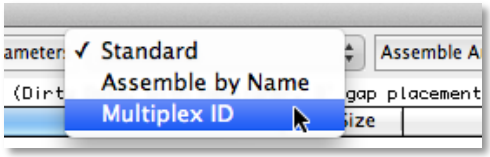



| Reference | | BC1 ... | BC10... | BC2 ... | BC3 ... | Total |
|---|---|---|---|---|---|---|
| 8,869 | A | | M | | | 1 |
| 9,061 | C | | | | | 1 |
| 9,261 | A | | R | | | 1 |
| 9,361 | T | A | | | | 1 |
| 9,523 | T | G | | | | 1 |
| 9,527 | C | | | | | 1 |
| 9,615 | T | | | | | 1 |
| 9,631 | T | | | | | 1 |
| 9,634 | C | | | | | 1 |
| 10,271 | C | | | | | 1 |
| 10,664 | C | | | | | 1 |
| 10,739 | A | | | | | 1 |
| 11,353 | T | | | | | 1 |
| | Total | 3 | 3 | 4 | 5 | 37 |

- Close the **Variance Table** window.

## Assembling Multiplex ID Data with Velvet

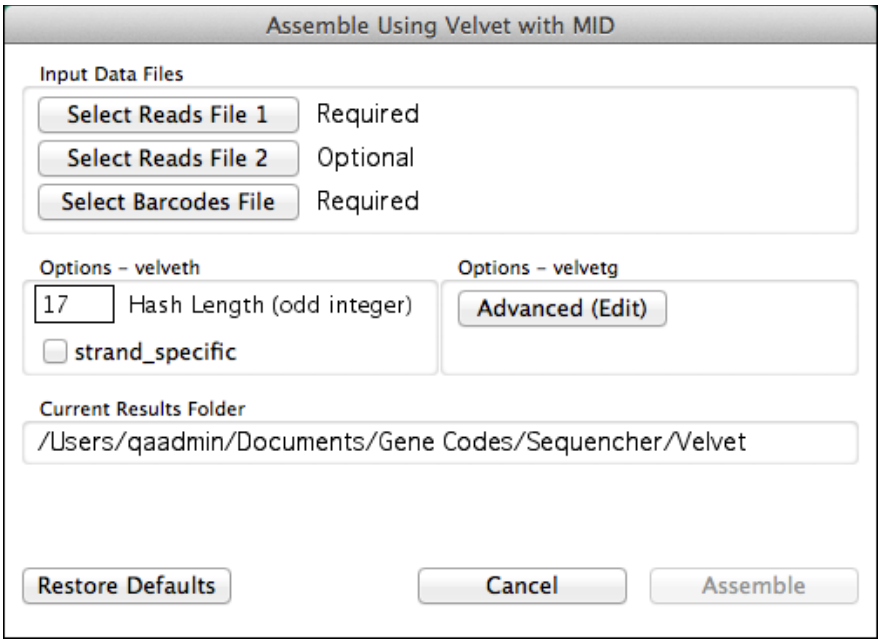
- Launch **Sequencher** if it is not already open.
- Set the **Assembly Mode** to **Multiplex ID** if it is not already.




- Go to the **Assemble** menu and select **Assemble Data Files Using>Velvet...**

The **Assemble Using Velvet with MID** dialog will appear along with the **External Data Browser** dialog if it is not already open. Now make changes to **Velvet's** settings as follows:

- Change the **Hash Length** to **17**.
- Click on the button called **Advanced (Edit)**.
- Click on the checkbox for **–ins_length** (insert length between paired-end reads) and set its value to **675**.
- Click on the **OK** button.

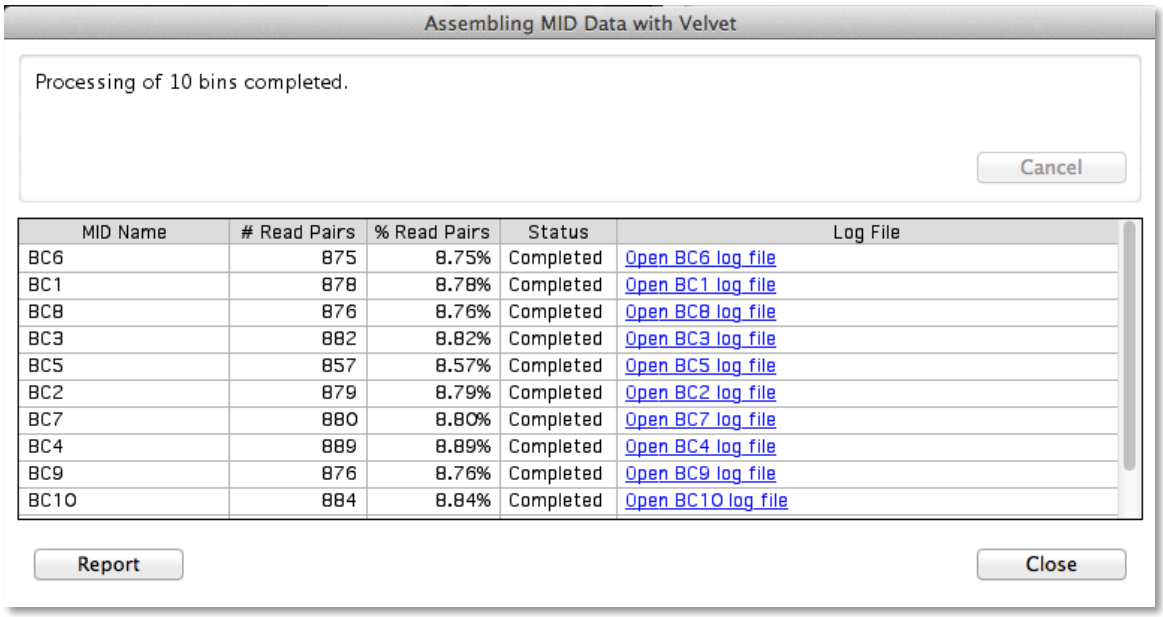



- Click on the **Select Reads File 1** button.
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder, then to the **Multiplex ID Data** folder.
- Choose **MID-Reads1.fasta** and click on the **Open** button.
- Click on the **Select Reads File 2** button. The file picker will automatically take you to the last location you chose.
- Choose **MID-Reads2.fasta** and click on the **Open** button.
- Now click on the **Select Barcodes File** button.
- The file picker will automatically take you to the last location you chose.
- Choose **MID-list.txt** and click on the **Open** button.

Finally, start the assembly process.

- Click on the **Assemble** button.

You can follow the progress of your assembly on the **Assembling MID Data with Velvet** dialog. The top part of the dialog shows you the progress of separating your reads into their different files. The lower part of the dialog shows the progress of the assembly for each barcode. It also shows how many reads are in each file and what this represents in terms of the overall number of reads. A status column will show whether a file is in the process of being assembled or whether assembly has been completed. Once the run is complete, the **Report** and **Close** buttons will become enabled.




## Reviewing Log File

The final column in the status lists the log files for each file.

- Click on a link to open a log file.

The log file contains details on the assembly run. Note that the top of the file contains details of the run including the sequence of the barcode for this particular run.

```
Start time: 2015-Nov-18 12:45:20

Using "Velvet" version "(1.2.10)".
Original reads file(s):
Reads file 1: /Applications/Sequencher 5.4.1/Sample Data/Multiplex ID Data/MID-Reads1.fasta
Reads file 2: /Applications/Sequencher 5.4.1/Sample Data/Multiplex ID Data/MID-Reads2.fasta
Intermediate reads file(s):
Reads file 1: /var/folders/xg/h3t9pwjs43179wmxgqr6wjmr0000gn/T/Bins609d5d7db4b91fa1/1/BC6.fa
Reads file 2: /var/folders/xg/h3t9pwjs43179wmxgqr6wjmr0000gn/T/Bins609d5d7db4b91fa1/2/BC6.fa
Reads file 1 is paired: No
Hash length: 35
Strand specific: No
MID: BC6
Barcode: TCAGATATCGCGAG
```

- Close the log file when done looking at it.

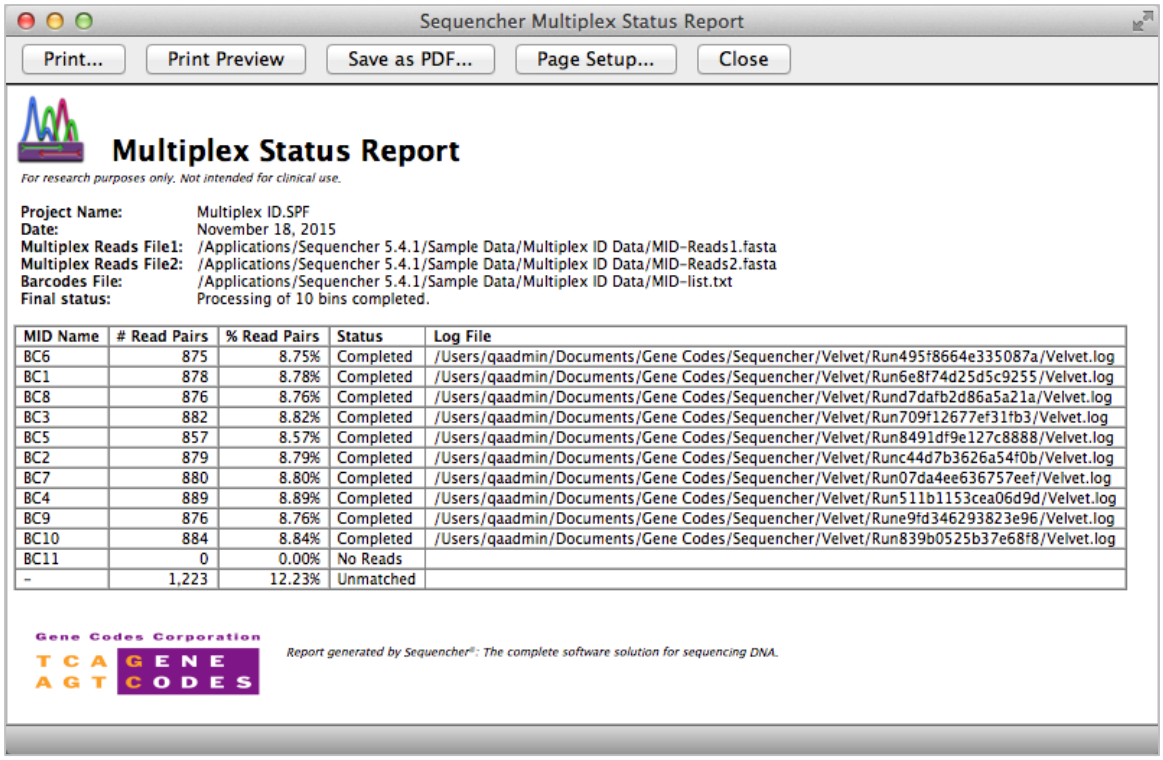
You can also monitor the progress of your assembly using the **External Data Browser** dialog. You can watch as the Log File pane updates. Notice that, once your alignment is complete, the status field updates to SUCCESS.

## Printing Status Report

To write out the status information contained in the **Assembling MID Data with Velvet** dialog, do the following:

- Click on the **Report** button at the bottom of the dialog.

This creates a preview window called **Sequencher Multiplex Status Report** that will contain the same information as the status table. This window has a button bar across the top. From this button bar, you can save the report as PDF, change the Page Setup settings, and Print the status report.




- Close the status report window.
- Close the **Assembling MID Data with Velvet** dialog.

## Reviewing Your Results in Tablet

The results of the assembly are returned to your **Project Window** as a series of consensus sequences (if any are created). You can sort the results by clicking the header of any column. One column that is particularly useful to sort on is the MID column that will group together the results from each barcode file.

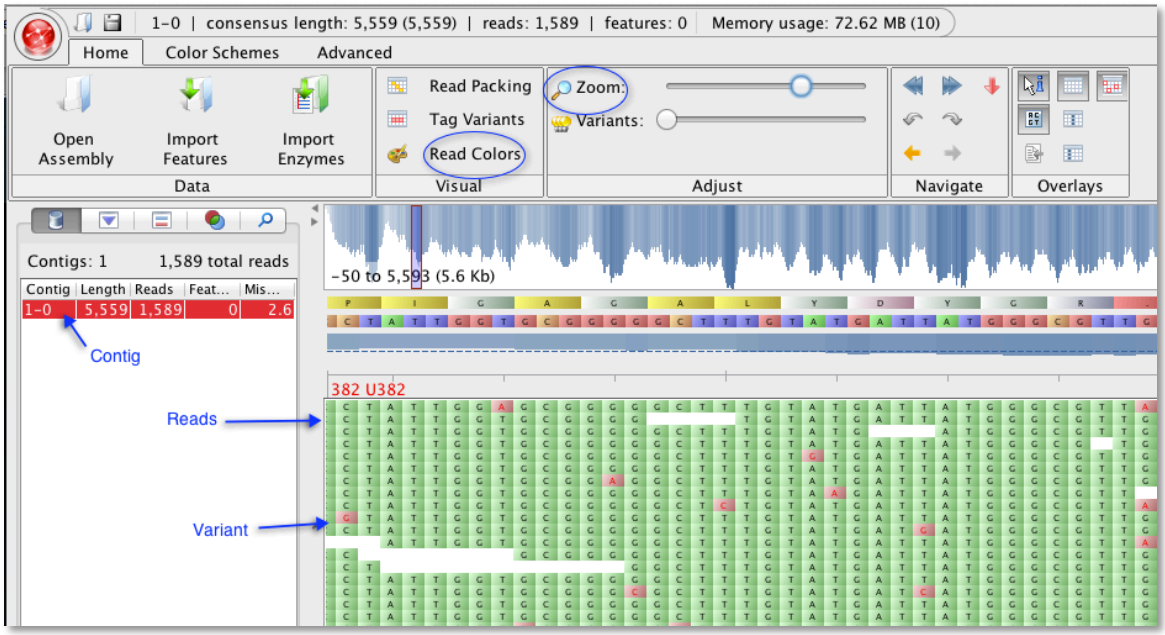
If you want to review your data in **Tablet** for a consensus sequence belonging to a barcode group you are interested in, then do the following:

- Click on a consensus sequence belonging to the barcode group you are interested in.

- Go to the **Contig** menu.
- Choose **Show NGS Data Using>Tablet**.

**Tablet** will launch and load all the data belonging to the barcode group. On the left-hand side of the **Tablet** window, you will see a list of contig(s).

- Click once on the contig of choice.
- The reads will load and appear on the right-hand side of the window.
- Zoom in and out using the slider at the top of the window.




- Click on **Read Colors**.
- Choose a class of **Read Colors** from the drop-down menu.

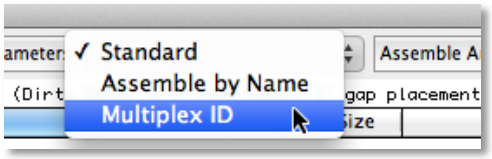
You will see a list of read classes in the drop-down menu such as **Nucleotide** which colors the individual bases such that all A's will be in one color and all T's will be in a different color and so on. Choosing **Direction** shows the direction of each read with coloring of the forward reads in one color and coloring of the reverse reads in a different color. **Read Group** refers to paired-end reads and **Variants** highlights any variants. In more recent versions of Tablet, you can customize the colors for these different classes. The view in the image above includes customized colors (pink backgrounds indicate variant bases).

- Close the **Tablet viewer** when done.
- Close the project without saving and exit **Sequencher** if desired.

## Aligning Multiplex ID Data with BWA-MEM

- Launch **Sequencher**.

- Set the **Assembly Mode** to **Multiplex ID** if it is not already.

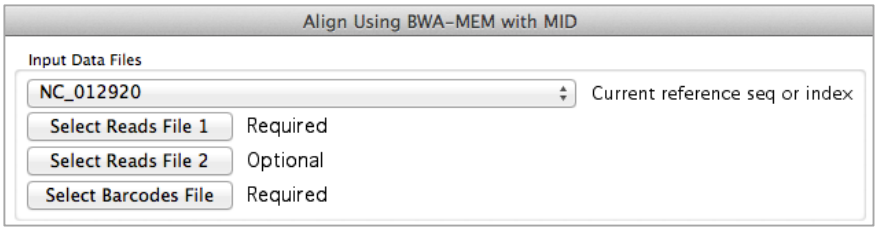



- Go to the **File** menu and select **Open Project…**
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder and choose the project called **Multiplex ID.SPF**.
- Click on the **Open** button.

The **Sequencher** project opens and contains a single reference sequence called NC_012920. This sequence is the complete human mitochondrion genome.

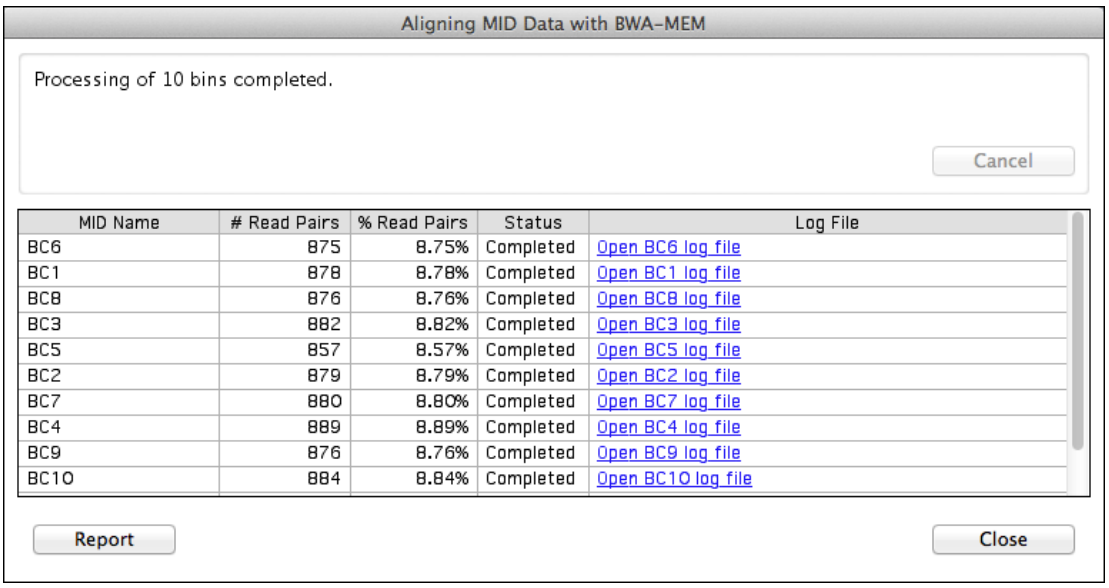
- Select the **NC_012920** sequence in the **Project Window**.
- Go to the **Assemble** menu and select **Align Data Files To Ref Using>BWA-MEM…**

The following dialog appears along with the **External Data Browser** dialog if it is not already opened:




- Click on the **Select Reads File 1** button.
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder, then to the **Multiplex ID Data** folder.
- Choose **MID-Reads1.fasta** and click on the **Open** button.
- Click on the **Select Reads File 2** button.
- The file picker will automatically take you to the last location you chose.
- Choose **MID-Reads2.fasta** and click on the **Open** button.
- Now click on the **Select Barcodes File** button.
- The file picker will automatically take you to the last location you chose.
- Choose **MID-list.txt** and click on the **Open** button.
- Now click on the **Align** button.

You can follow the progress of your analysis on the **Aligning MID Data with BWA-MEM** dialog. The top part of the dialog shows you the progress of separating your reads into their different files. The lower part of the dialog contains a status table that shows the progress of alignment for each sample. It also shows how many reads are in each file and what this represents in terms of the overall number of reads. A status column will show whether a file is in the process of being aligned or whether alignment has been completed. Once the run is complete, the **Report** and **Close** buttons will become enabled.




## Reviewing Log File

The final column in the status table lists the log files for each file.

- Click on one of the log file links to open the log file.

The log file contains details on the alignment run. Note that the top of the file contains details of the run including the sequence of the barcode for this particular run.

```
Start time: 2015-Nov-18 13:07:09

Using BWA version "0.7.12-r1039".
Reference Sequence: NC_012920
Original reads file(s):
Reads file 1: /Applications/Sequencher 5.4.1/Sample Data/Multiplex ID Data/MID-Reads1.fasta
Reads file 2: /Applications/Sequencher 5.4.1/Sample Data/Multiplex ID Data/MID-Reads2.fasta
Intermediate reads file(s):
Reads file 1: /var/folders/xg/h3t9pwjs43179wmxgqr6wjmr0000gn/T/Bins756c65831c8b1c7d/1/BC6.fa
Reads file 2: /var/folders/xg/h3t9pwjs43179wmxgqr6wjmr0000gn/T/Bins756c65831c8b1c7d/2/BC6.fa
Match score: 1
Mismatch penalty: 4
Gap open penalty: 6
Gap extension penalty: 1
Insert filter: 10
MID: BC6
Barcode: TCAGATATCGCGAG
```

- Close the log file when done looking at it.

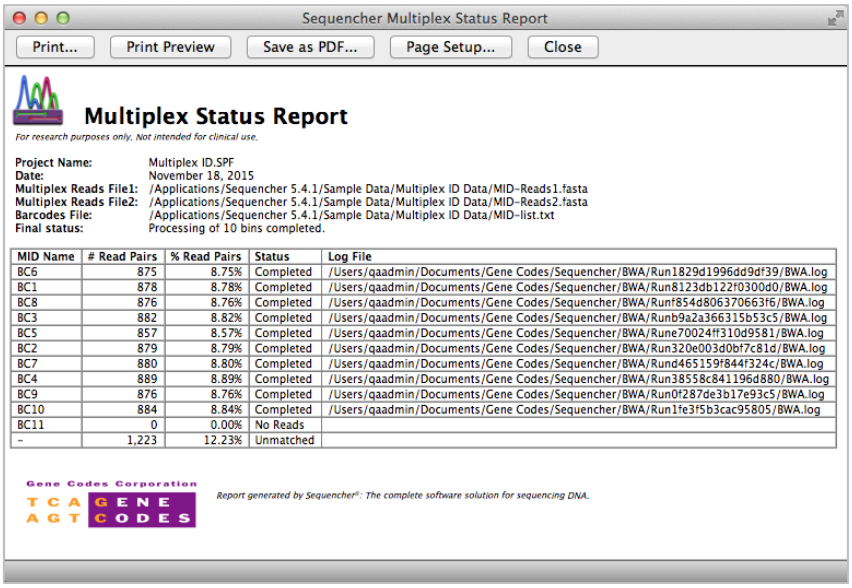
You can also monitor the progress of your assembly using the **External Data Browser** dialog. You can watch as the Log File pane updates. Notice that, once your alignment is complete, the status field updates to SUCCESS.

## Printing Multiplex Status Report

To write out the status information contained in the **Aligning MID Data with BWA-MEM** dialog, do the following:

- Click on the **Report** button at the bottom of the dialog.

This creates a preview window called **Sequencher Multiplex Status Report** that will contain the same information as the status table but also includes added information such as the date of the run and the location of the data files. This window has a button bar across the top. From this button bar, you can save the report as PDF, change the Page Setup settings, and Print the status report.




- Close the status report window.
- Close the **Aligning MID Data with BWA-MEM** dialog.

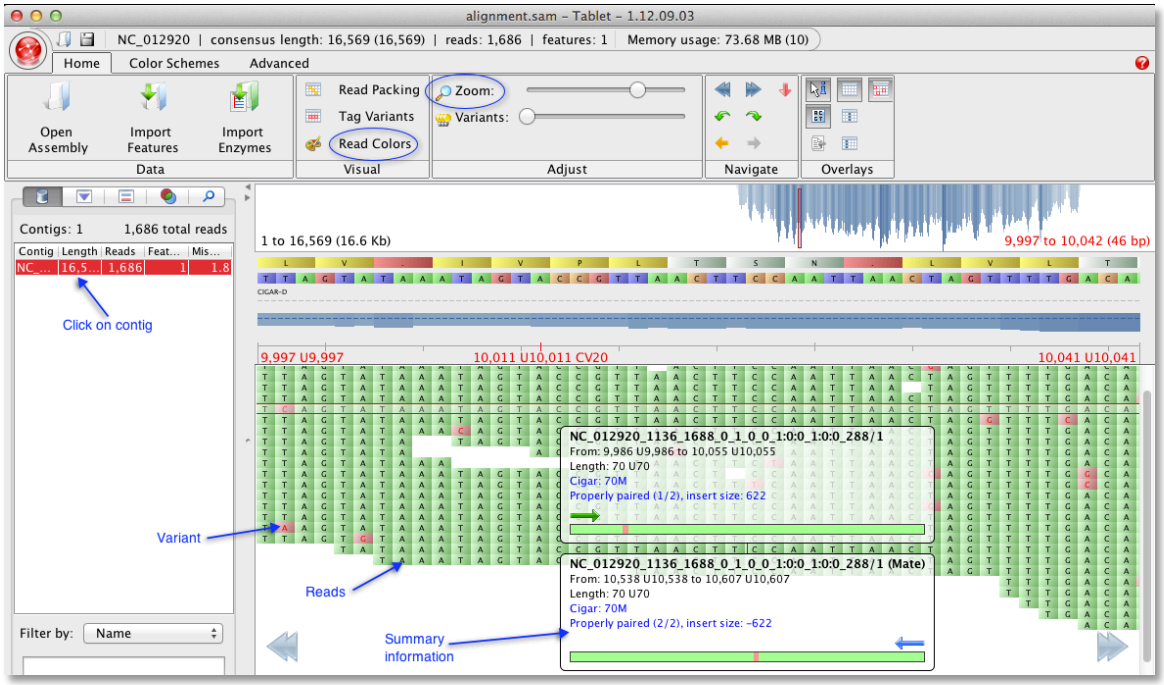
## Reviewing Your Results in Tablet

If you want to review your data in **Tablet**, click on an item of interest in the **Project Window**.

- Go to the **Contig** menu.
- Choose **Show NGS Data Using>Tablet**.

**Tablet** will launch and load your data and reference sequence automatically. On the left-hand side of the **Tablet** window, you will see a list of contig(s).

- Click once on the contig of choice (there may only be one).
- The reads will load and appear on the right-hand side of the window.
- Zoom in and out using the slider at the top of the window (top blue oval).




- Click on **Read Colors** (lower blue oval).
- Choose a class of **Read Colors** from the drop-down menu.

You will see a list of read classes in the drop-down menu such as **Nucleotide** which colors the individual bases such that all A's will be in one color and all T's will be in a different color and so on. Choosing **Direction** shows the direction of each read with coloring of the forward reads in one color and coloring of the reverse reads in a different color. **Read Group** refers to paired-end reads and **Variants** highlights any variants. In more recent versions of **Tablet**, you can customize the colors for these different classes. The view in the image above includes customized colors (pink backgrounds indicate variant bases).

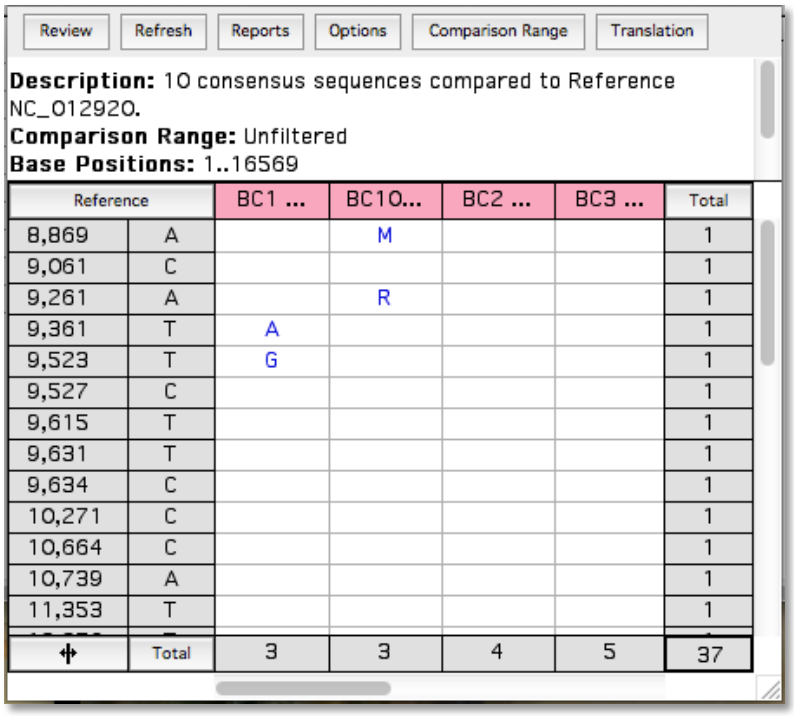
- Close the **Tablet viewer**.

## Reviewing in Sequencher

Since you have aligned your data against the same reference sequence, you can use the **Variance Table** to review the number of differences between samples.

- Select the contigs you wish to review.
- Choose the **Contig** menu and select **Compare Consensus to Reference**.

A **Variance Table** is generated and with it you can compare the position and number of variants between each contig and the reference sequence.




| Reference | | BC1 ... | BC10... | BC2 ... | BC3 ... | Total |
|---|---|---|---|---|---|---|
| 8,869 | A | | M | | | 1 |
| 9,061 | C | | | | | 1 |
| 9,261 | A | | R | | | 1 |
| 9,361 | T | A | | | | 1 |
| 9,523 | T | G | | | | 1 |
| 9,527 | C | | | | | 1 |
| 9,615 | T | | | | | 1 |
| 9,631 | T | | | | | 1 |
| 9,634 | C | | | | | 1 |
| 10,271 | C | | | | | 1 |
| 10,664 | C | | | | | 1 |
| 10,739 | A | | | | | 1 |
| 11,353 | T | | | | | 1 |
| | Total | 3 | 3 | 4 | 5 | 37 |

- Close the **Variance Table** window.

## Conclusion

In this tutorial, you have learned how to work with three separate programs, **BWA-MEM**, **GSNAP**, and **Velvet**, and Multiplex ID data. You have learned how to change settings, review your data in **Tablet**, and examine log files. You have also learned how to understand and print out the Multiplex Report.

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