SEQUENCHER®

Tutorial for Windows and Macintosh

Next-Generation Sequence Alignment – Advanced Analysis

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

Next-Generation Sequence Alignment – Advanced Analysis

Next-Generation Sequence Alignment – Advanced Analysis

This is the second in a series of tutorials dealing with Next-Generation Sequencing. In this tutorial, you will learn to use either **GSNAP** or **Maq** to look for SNPs. You will also learn to use **GSNAP** to analyze bisulfite-treated DNA in methylation studies and to analyze DNA for evidence of changes caused by the ADAR gene product (A to I changes). Finally, you will learn how to analyze the resultant SAM/BAM file to create VCF files that can be visualized in **Tablet** or your favorite Genome browser.

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

File Formats

In this tutorial, you will be provided with Next-Gen reads in **FastQ** and **FastA** formats. 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** format, Sanger and Illumina. You will need to know which type your FastQ files are.

If you are using your own data and it is in its original 454 format (SFF), you will need to extract the data in **FastQ** format before proceeding. You can read more about this in the "Next-Generation Sequence Alignment" tutorial, which can be found in the Tutorials folder.

Getting Started

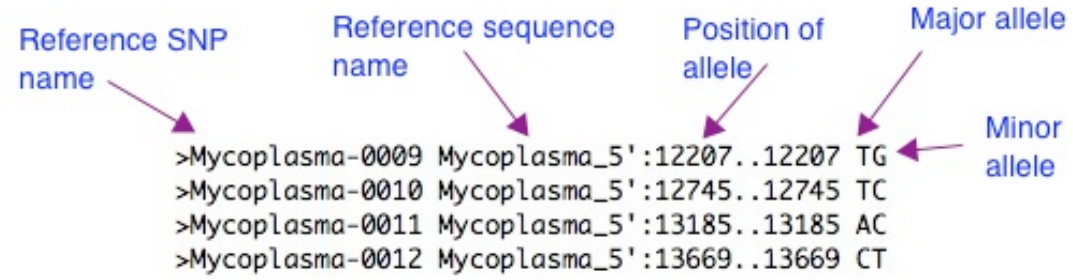
In this tutorial, you will use the Next-Gen algorithms to align your Next-Gen reads to a reference sequence and then analyze them. You will first need to open a project. The project provided contains a reference sequence for use with the sample Next-Gen data. If you want to learn how to build GSNAP databases or BWA indexes, refer to the "Preparing Your Data for NGS Alignment" tutorial on our website at http://genecodes.com/training/tutorials.

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

You can use the **External Data Browser** to annotate each run with important information such as the samples used, the source of the samples, and settings. It will open automatically whenever you choose to build a reference database or index, do a reference-guided alignment, or a de novo assembly.

SNP Analysis with GSNAP

GSNAP performs a SNP analysis by aligning reads to the reference in an SNP-tolerant fashion. The results indicate whether the SNP is a known SNP or an unknown mismatch. You will need to prepare and supply a file that contains a list of known SNPs. This file has to be in a specific format that is described below.

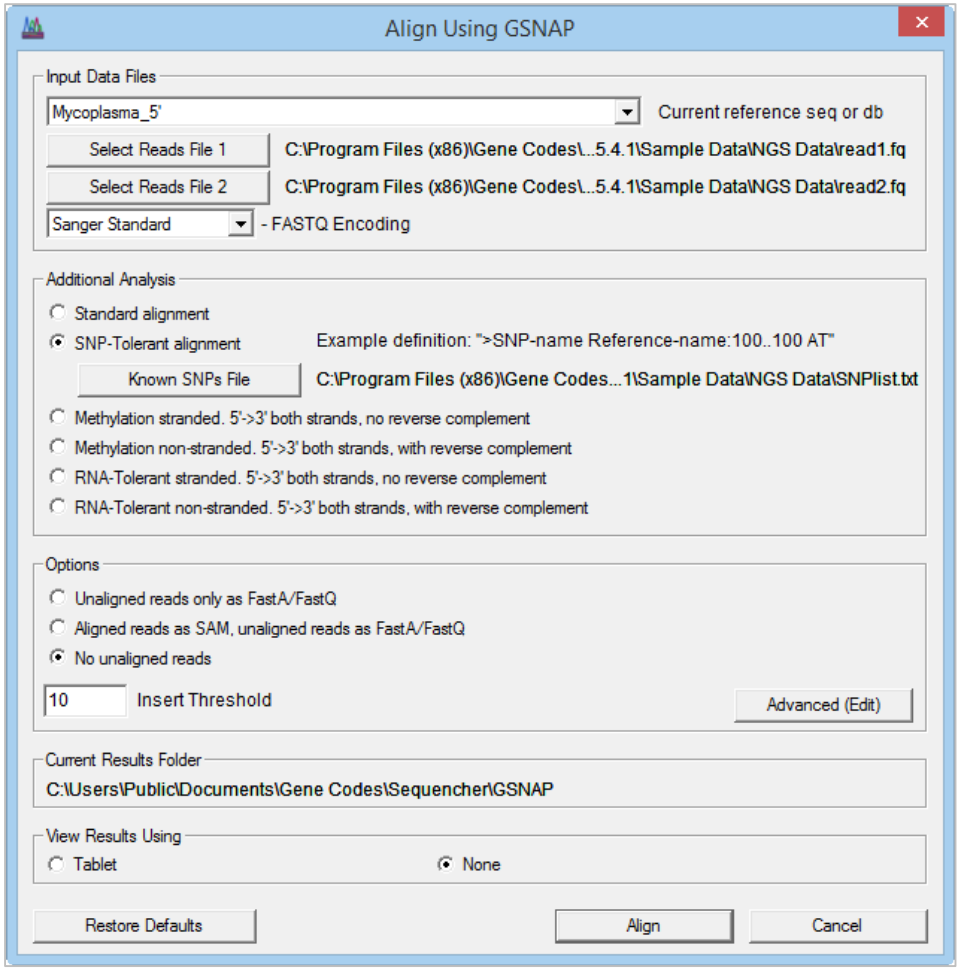


The file is a text file and each SNP is placed on a separate line. The lines begin with a > (angle bracket). This is followed by the reference name of the SNP. The next column contains the name of the reference sequence. In this case, we are using Mycoplasma_5'. It is very important that this name matches the name of the chosen reference sequence in the **Project Window**. Otherwise, the analysis will fail. The next columns (following the colon) indicate the position of the SNP. Note that these SNPs are single bases so the position is written from 12207 to 12207 so it is written as 12207:12207. The final column contains the major and minor alleles for the SNP. The file itself can be created in any text editor (not a word processor).

Note that the position information in this file always assumes that the first base of the reference sequence in **Sequencher** is 1, no matter its actual numbering relative to its chromosomal or contig position.

To perform the analysis, you need to select your reference sequence from the **Project Window**.

- Back in the **Project Window**, select the **Mycoplasma_5'** sequence.
- Choose **Assemble > Align Data Files to Ref Using > GSNAP...**
- The **External Data Browser** and the **Align Using GSNAP** dialogs will appear. From the **Align Using GSNAP** dialog, click on the **Select Reads File 1** button and browse to the first reads file you want to use. We have supplied one called **read1.fq** in **Sequencher's Sample Data/NGS Data** folder.
- Select that file and click on the **Open** button.
- Click on the **Select Reads File 2** button if you are using paired-end data and select the second reads file you want to use. We have supplied **read2.fq** in **Sequencher**'s **Sample Data/NGS Data** folder.
- Click on the **Open** button.
- Now choose **SNP-Tolerant alignment** in the **Additional Analysis** groupbox.
- The **Known SNPs File** button is now enabled. Click on that button and browse to the file containing your list of known SNPs. We have supplied one in the **NGS Data** folder that goes with the supplied reads files, read1.fq and read2.fq. Select file **SNPlist.txt** and click on the **Open** button.
- Now choose **No unaligned reads** in the **Options** groupbox.

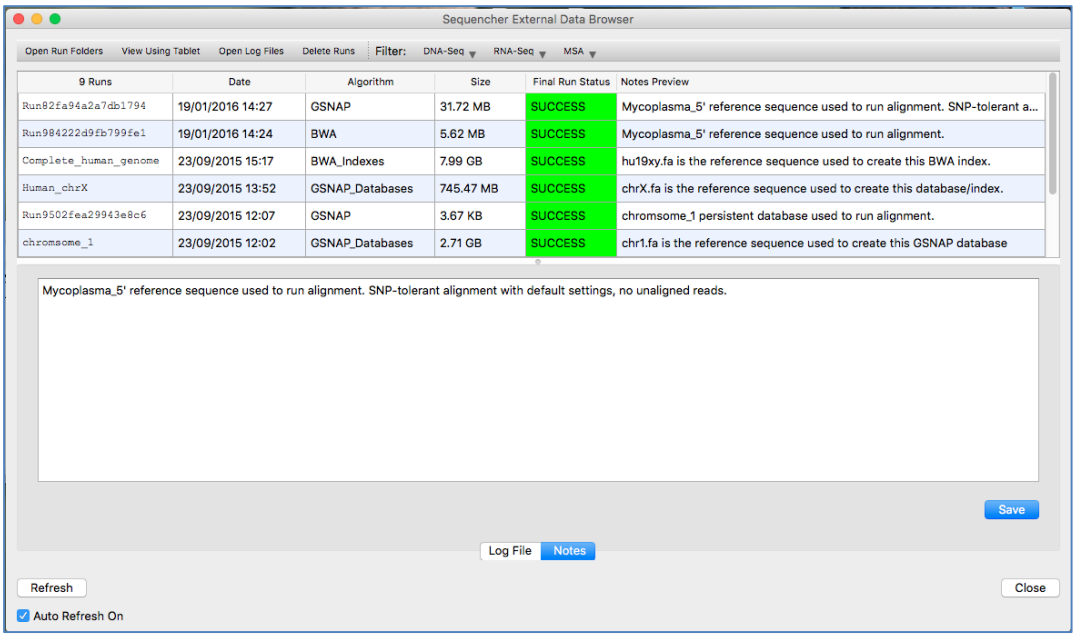


Also choose whether you want to see the results file in **Tablet** immediately after the alignment has finished. In the previous image, **None** has been selected. Note that if you do not choose a viewer before submitting the reads for alignment, you will not be able to invoke it later if using the Viewer version of **Sequencher**. You will get a message to that effect if you try. You will still see a new contig in the **Project Window** though.

- Now click on the **Align** button to initiate the analysis.

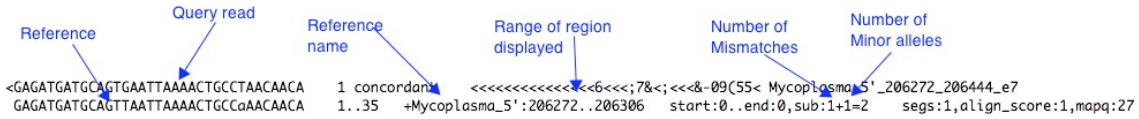
The analysis is complete when a new contig appears in the **Project Window**.

- Back in the **External Data Browser** window, click on the **Refresh** button and click on the most recent **GSNAP** run.
- Then click on the **Notes** tab and in the **Notes** window enter additional details for your analyses similar to "**SNP-tolerant alignment with default settings, no unaligned reads**".
- Click on the **Save** button to save your note. You will see the note in the **Notes Preview** cell.



- To review the results, choose the contig of interest in the **Project Window** if it is not already selected.
- Go to the **Sequence** menu.
- Choose the **Analyses > GSNAP SNP Analysis** command.
- The SNP report opens in your default browser.

The report you see presents the results as a series of genomic segments with reads aligned to the segment.



Note that if you are working with a Viewer version of **Sequencher**, you will only see a sample report. Reading across the report you will see information which shows how many known SNPs and how many mismatches are in each segment. The image above explains the kind of result you might see in more detail.

The first line is always a query read with the line below being an aligned segment from the reference sequence. Next comes the range of the alignment. In the aligned segment example, it ranges from base 1 to 35. Characters + or – in front of the region information indicates stranded-ness. Note that, if the region is from the – strand, the range information will be of the form 10..1 rather than 1..10. The word concordant refers to a paired-end read if both reads of the pair are on the same reference sequence and in the expected direction.

In the result above, there is a mismatch, sub:1, and a minor allele, +1, leading to a total of 2 mismatches relative to the reference sequence (1+1=2). Only one segment was found (segs:1) with a mapping quality of 27. If the report shows a mismatch, then this is likely to be a new SNP. If the report shows a minor allele, then this is a known SNP.

- Use your browser's **Find** function to search for sub:1 or +1 to find other SNPs.

- When you are finished, close out of your browser and return to **Sequencher**.

Methylation Analysis with GSNAP

The study of epigenetics deals with heritable changes in gene expression and usually involves modifications to the genome but do not involve DNA base changes. One such example is DNA methylation. Sequencing studies of methylated parts of the genome involve using bisulfite treatment of the DNA followed by sequencing of the regions of interest. The bisulfite treatment chemically converts unmethylated Cs to Ts. **GSNAP** is capable of aligning the sequenced reads to the genomic (untreated) sequence.

To perform the analysis, you need to supply a reference sequence and one or two (if paired-end) FastQ files containing reads from bisulfite treated DNA.

You have 2 options for this type of analysis, stranded or non-stranded. Stranded data allows only 5′ to 3′ reads on each strand of the genome. Non-stranded data allows both the 5′ to 3′ reads and their reverse complements on each strand.

In this example, you will be using a reference sequence called methylation_reference and analyzing a set of single-ends sequencing reads using the stranded analysis option.

- Click on the sequence called **methylation_reference**.
- Choose **Assemble > Align Data Files to Ref Using > GSNAP...**
- The **External Data Browser** and the **Align Using GSNAP** dialogs will appear. From the **Align Using GSNAP** dialog, click on the **Select Reads File 1** button.
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder.
- Within the **Sample Data** folder you will find a folder called **NGS Data**. Select **bisulfite_read.fa** and click on the **Open** button.
- Now choose **Methylation stranded. 5′->3′ both strands, no reverse complement** in the **Additional Analysis** groupbox and **No unaligned reads** in the **Options** groupbox.
- Choose the **View Results None** radio button.
- Click on the **Align** button.
- Don't forget to add an annotation for this run in the **External Data Browser** window after clicking on the **Refresh** button.

Although you chose to not view the results at the time of alignment, you will still be able to invoke a viewer later to review the results if using the full version of **Sequencher**. After the alignment is completed, you will see a new contig in the **Project Window**.

In order to see the results of the analysis, do the following:

- Go to the **Sequence** menu.
- Choose **Analyses > GSNAP Methylation Analysis**.

If you create an alignment while running in Viewer Mode, any new report generated from that alignment can be displayed in your browser but will be an example of what it would look like and will indicate it is a demo report. If you are looking at reports that were generated when you were not in Viewer Mode, you can view those in their entirety.

The Methylation report comes up in your default browser. This report will contain a great deal of information. As you browse through the contents, you will see periods below some of the Ts. These characters represent cytosines in the genomic sequence and are paired against some thymidines in the query reads. They indicate that the read contains a converted C > T, where the C was unmethylated and therefore unprotected from bisulfite treatment.

In the segment of the report shown below, the read data follows the > character on the first line. The next column on the first line lists the number of matching genomic locations to which the read aligned. The line below it (containing the periods) is the genomic sequence followed by the range of the aligned read. In the example below, all 63 bases of the read sequence aligns to the genomic reference–1 to 63. The polarity of the genomic strand is indicated by a + or – character preceding the name of the reference (genomic) sequence.

```
>AATAATAATAATATTAAAAGTTAAAATTAAATTTATTAAATTAATTAAGGTTAAAGTTAAAAT    1   Frag[0001]_0-63_0
 AA.AA.AATAATA..AAAAG..AAAATTAAA...ATTAAAT.AATTAAGGTTAAAG..AAAAT    1..63   -
methylation_reference:63..1   start:0..end:0,sub:0

>ATTTTGGTTTTAATTTTAATTGATTTAATGGGTTTAATTTTGGTTTTTGGTATTATTGTTGTT    1   Frag[0001]_0-63_1
 ATTTTGG.TTTAA..TTAATTGATTTAATGGGTTTAATTTTGG.TTTTGGTATTATTGTTGTT    1..63
   +methylation_reference:1..63  start:0..end:0,sub:0
```

Any mismatches reported will be in lower case and the number of mismatches is indicated by the word sub. Thus, sub:2 means there are two mismatches in the reference vs. query alignment. In the segment of the report shown in the image above, there are no mismatches and this is indicated by sub:0.

- If you so desire, save your work and then **Quit** (Mac) or **Exit** (Windows) from **Sequencher**.

RNA Analysis with GSNAP

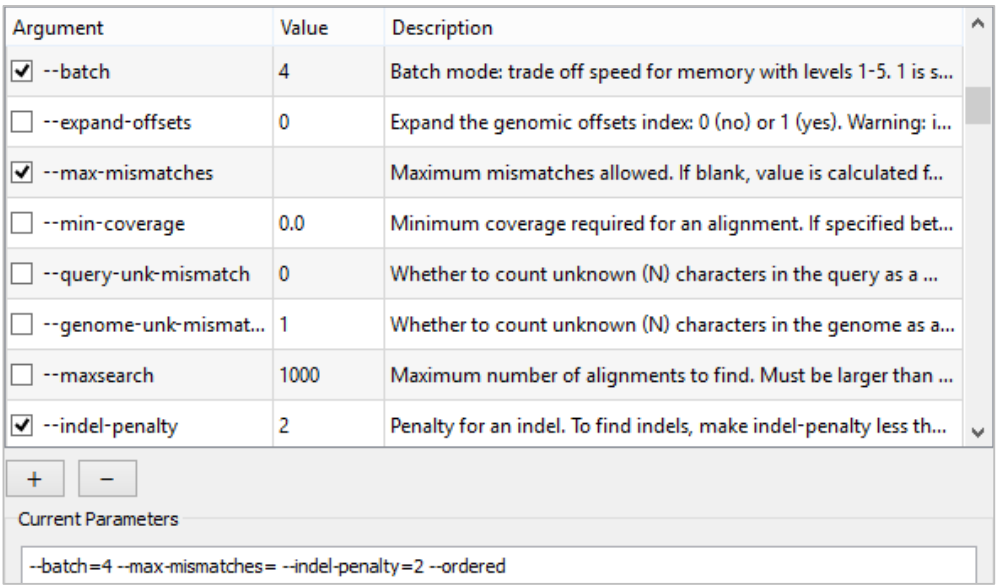
The **GSNAP** aligner has an alignment mode that is tolerant to A to G changes. These changes occur when RNA post-transcriptional editing causes A's to be converted to I's.

To perform the analysis, you need to supply a reference sequence and one or two (if paired-end) FastQ files containing data you suspect of having undergone this modification. You have 2 options for this type of analysis, stranded or non-stranded. Stranded data allows only 5′ to 3′ reads on each strand of the genome. Non-stranded data allows both the 5′ to 3′ reads and their reverse complements on each strand.

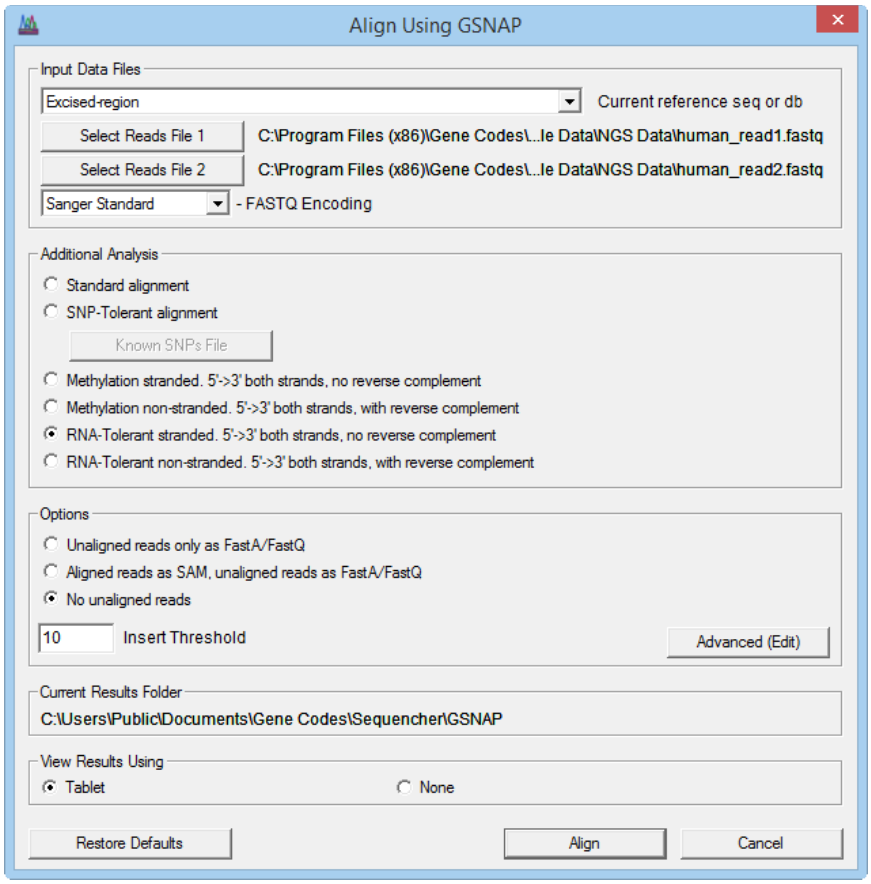
In this example, you will be using a reference sequence called Excised-region and analyzing a set of paired-ends sequencing reads.

- In the **Project Window**, click on the sequence called **Excised-region** (green label).
- Choose **Assemble > Align Data Files To Ref Using > GSNAP….** The **External Data Browser** will open automatically if it is not already open.
- Click on the **Select Reads File 1** button, browse to the **NGS Data** folder, choose **human_read1.fastq**, and click on the **Open** button.

- Click on the **Select Reads File 2** button, browse to the **NGS Data** folder, choose **human_read2.fastq**, and click on the **Open** button.
- Click on the **RNA-Tolerant stranded. 5′->3′ both strands, no reverse complement** radio button.
- Click on the **Advanced (Edit)** button and in the **GSNAP Advanced Options** dialog and change your settings to match the following image:



- Click on the **OK** button.
- Ensure that the **Tablet** radio button has been selected.



- Click on the **Align** button.

The alignment starts and the progress dialog is displayed. Please be patient. This will take some time to run (depending on the amount of RAM you have installed). The progress dialog will be automatically dismissed once the alignment has finished.

- Back in the **External Data Browser** window, click on the **Refresh** button and note the name of the results folder.
- Click on the **Notes** tab for the current Run, click in the **Notes** window, and add additional text to the note like "**RNA-Tolerant Excised-region run and Variant Calling**", and then click on the **Save** button. Your note should appear in the **Notes Preview** cell.

You can monitor the progress of the alignment using the following steps:

- Go back to the **External Data Browser** window and click on the **Log File** tab. During a long alignment, the Log File view will refresh periodically.
- Using the scrollbar, scroll to the bottom of the window to see the progress of the alignment.
- Click on the **Refresh** button again to get the latest progress. The status in the **Final Run Status** column will be a green **SUCCESS**. If the alignment failed, you'll see a red **FAILED** status.

Tablet will automatically launch and load the results file when the alignment is completed.

- Leave **Tablet** open.

Viewing RNA results in your default browser

If you chose to not view the results at the time of alignment, you will still be able to invoke a viewer later to review the results if using the full version of **Sequencher**. After the alignment is completed, you will see a new contig in the **Project Window**.

In order to see the results of the analysis, do the following:

- Go to the **Sequence** menu.
- Choose **Analyses > GSNAP RNA-Tolerant Analysis**.

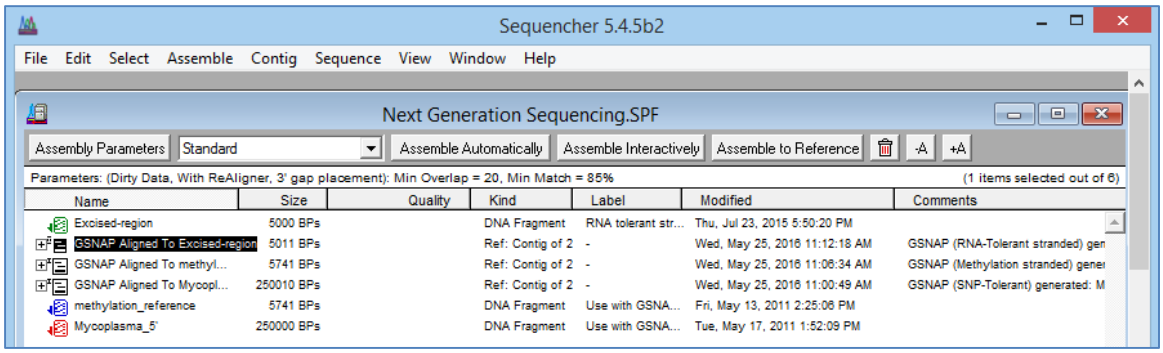
The RNA-Tolerant Analysis report comes up in your default browser. This report will contain a great deal of information. Any mismatches reported will be in lower case and the number of mismatches is indicated by the word sub.

- A good tip is to reset the Default settings in **GSNAP** before you use it again.
- Select a sequence and choose **Assemble > Align Data Files To Ref Using > GSNAP….**
- Click on the **Advanced (Edit)** button, then click on the **Restore Defaults** button.
- Click on the **OK** button to dismiss the **Advanced (Edit)** dialog. Then click on the **Cancel** button to dismiss the **Align Using GSNAP** dialog.
- If you so desire, save your work and then **Quit** (Mac) or **Exit** (Windows) from **Sequencher**.

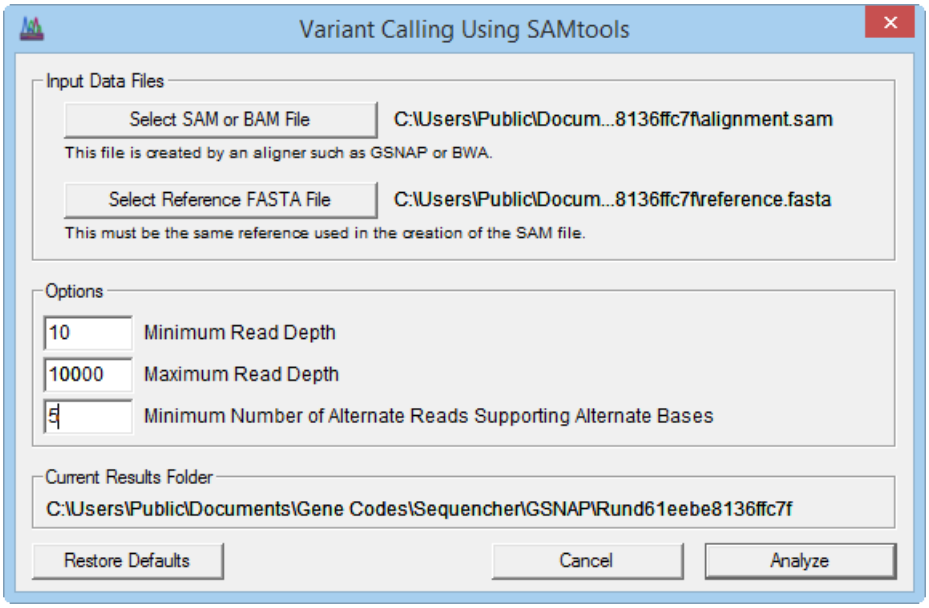
Variant Calling with SAMtools

Once you have performed an alignment, you can analyze the results file (SAM/BAM) for the presence of variants. This is done using SAMTools, which takes the SAM/BAM file and analyzes it for the presence of insertions and deletions as well as single base changes. The SAM/BAM file is converted to an mpileup file. An association test is carried out on the file. This is followed by a step that estimates the allele frequency spectrum and finally a filter step is performed. The results are written to a VCF (Variant Call Format) file, which can be viewed in Tablet or most genome browsers.

- Click on the contig created in the last step.



- Choose **Sequence > Analyses > SAMtools Variant Calling**...
- Enter **10** as the value for **Minimum Read Depth**.
- Enter **10000** as the value for **Maximum Read Depth**.
- Enter **5** as the value for **Minimum Number of Alternate Reads Supporting Alternate Bases**.



- Click on the **Analyze** button.

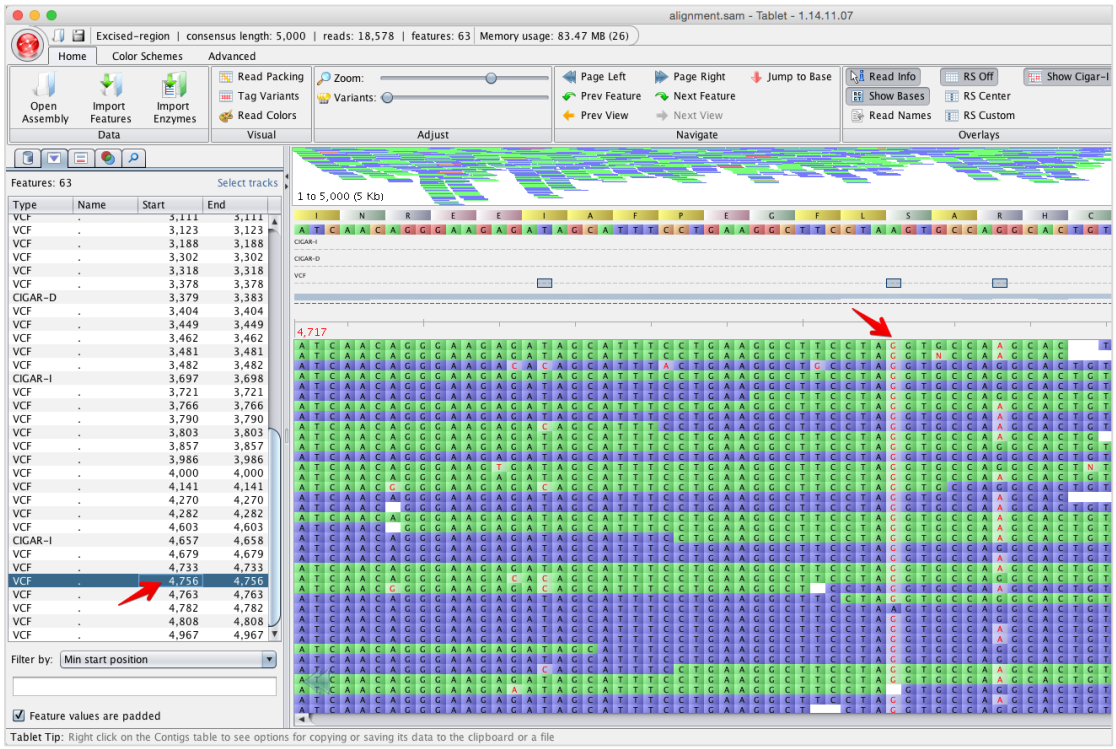
The variant calling starts and the progress dialog is displayed. This will be automatically dismissed once the analysis is complete. The VCF results file is stored in the same folder as the parent SAM/BAM file.

Viewing your Variants Using Tablet

The results of your alignment and variant analysis can be viewed using the **Tablet** genome browser as follows.

- Click on the **Tablet** window to bring its window to the front.
- Click on the contig listed in the left-hand pane to load it.
- Click on the **Import Features** button.
- Navigate to the run folder whose name you noted above, the one for the **Note** you added above.
- Select the **alignment.vcf** file in the folder and click on the **Open** button.
- Close the confirmation dialog.
- Click on the second tab (Features) in **Tablet** to display the list of variants (all Type VCF) found by **SAMtools**.
- Locate the entry for position 3766 in the list on the left of the **Tablet** window and click on it.
- Scroll the alignment window on the right-hand side of the **Tablet** window up and down and note how many Gs there are.
- Locate the entry for position 4756 and click on it.
- Explore this column of data in the right-hand side of the **Tablet** window.

In both instances, you will see that there are very few As present. These could be good candidates for RNA editing by the ADAR gene product compared to other processes. You can continue to explore your contig further or quit Tablet and **Sequencher** to end the tutorial.



Conclusion

In this tutorial, you have worked with two programs, one for aligning Next-Generation sequences to a reference sequence and one for calling variants on the aligned results. You have learned how to use these programs to hunt for SNPs, study methylation patterns in DNA, which has been sequenced following bisulfite treatment and analyse DNA for post-transcriptional RNA editing. You have been introduced to the **External Data Browser** and used it for annotating your runs. You have also learned how to use **SAMtools for Variant Calling**. Finally you have learned how to view your results using **Tablet**.

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.