

SEQUENCHER[®]

Version 5.0

Release Notes

© 2011 Gene Codes Corporation

Gene Codes Corporation



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

Sequencher 5.0 Release Notes

Rely on Results from Sequencher 5.0[™]. Easy... Fast... Powerful... DNA Sequence Assembly Software.

What's New

- **Align Next Generation Sequences**

Sequencher's Next Generation Sequencing (NGS) data alignment using GSNAP[®] or Maq[†] takes data input files, performs the alignment to the selected reference sequence and adds the resulting contig's consensus to your Sequencher Project with links to detailed data results. These data results may often be very large files. You can choose where these files are saved by setting the Gene Codes Home Directory on the new **External Data** User Preference pane. To perform an alignment with either algorithm, select a single sequence in your project and choose the **Assemble > Align Data Files to Ref Using** command.

- **New Assemble Menu**

There is a new Assemble menu that consolidates the commands for the NGS alignment algorithms with Sequencher's existing Assemble Contigs commands that were previously on the Contig menu.

- **View Next Gen Sequence Alignment Data**

You may immediately view the NGS alignment results in Maqview[†] or Tablet[§] viewers after alignment with Maq or GSNAP has completed. Regardless of whether you choose to view the aligned reads immediately, the consensus results aligned to the Reference Sequence will be added to your current project allowing you to view the contig later in the appropriate viewer by selecting the consensus sequence or contig in your project and choosing the **Contig > Show NGS Data Using** command.

- **SNP Analysis with Maq**

Maq allows you to search for SNPs using a two stage filtering process, the results of which are presented in a report. To open the report, choose the **Sequence > Analyses > Maq SNP Report** command.

- **SNP Analysis with GSNAP**

You may perform a SNP-tolerant alignment with GSNAP. You provide the reference sequence, reads and a list of known SNPs. GSNAP reports back on the found known SNPs and any new mismatches. To view the SNP-tolerant alignment results, choose the **Sequence > Analyses > GSNAP SNP Analysis** command.

- **Methylation Analysis**

GSNAP performs a mismatch tolerant alignment that allows Cs in the reference sequence to match Ts in the sequencing reads. The Ts will have been converted by the bisulfite treatment prior to sequencing and, therefore, were unmethylated. To view the Methylation tolerant alignment results, choose the **Sequence > Analyses > GSNAP Methylation Analysis** command.

- **Simple NCBI BLAST Search**

We've added the ability to perform Blast searches from within Sequencher. The search mode is designed for quickly checking regions of your data. Select a range of bases or 1 or more sequences and contigs in your project and choose the **Sequence > NCBI Blast Search** command.

[™] Sequencher 5.0 Requirements: Mac OS X 10.5 or 10.6, Windows XP or later. Hardware requirements depend upon your project's needs but at minimum should be 512MB RAM and 175MB hard disk space.

What's Improved

- **Performance Improvements for Large Sequences and Contigs**

Working with sequences and contigs millions of bases long is now up to hundreds of times faster. This performance improvement is most noticeable in calculating and displaying the Overview (including features, motifs, and codon map), the Restriction Map and Sequence inset map elements.

- **Display of Restriction Sites, Features, Motifs and Start/Stop Codons**

Depending on the size of the sequence and the window, there were certain situations when Restriction maps, Features and Motifs, and Start/Stop maps would display incorrectly, truncating or even wrapping in the middle. This has now been corrected. All of the maps draw correctly now no matter how many millions of bases are being represented.

- **Variance Table Changes**

Positions where one sample has an insert result in a Variance Table row with a gap listed in the Reference Sequence's column. Samples that match the Reference Sequence and have no insert display a blank (matching) cell in the table. These cells were incorrectly colored dark blue to indicate low confidence. That has been corrected.

Variance Table Reports may be saved as PDF files. The default name for these files has been standardized to include a timestamp and to save to the last used save location (will save to your Desktop by default).

- **Mac OS X Only**

When projects are saved on Mac, they will now automatically have an .SPF extension added.

The obsolete File Export User Preference pane has been removed.

- **New Menu for Clustal**

Alignment with Clustal was added in Sequencher 4.9 under the **Contig > Assemble Contigs** menu. This functionality is now available from the **Assemble** menu. Select sequences in your project you wish to align and choose the **Assemble > Align Using > Clustal** command. For instructions on installing the clustalw2 alignment algorithm, please see [Using External Tools with Sequencher](http://www.genecodes.com/training/tutorials) available at www.genecodes.com/training/tutorials.

- **Trim to Reference Sequence Improvements**

The Trim to Reference Sequence command is available from the **Contig** menu. It trims every sequence in a contig that extends beyond the boundaries of the contig's Reference Sequence. The command no longer leaves behind extra gaps in sequences, is consistently available in any open contig window as well as the Project Window, and updates the display of contigs that are significantly shortened after trimming.

- **Fixes**

Get Info on a project correctly reports the number of contigs • Changing the orientation of a sequence no longer updates the modified date • It's now possible to save to a disk larger than 1TB without getting a disk space error • You may set the default Assembly Parameters for new blank projects from the **New Project User Preferences** pane.

What's New

ALIGNING NEXT GEN SEQUENCES

You can now add the alignment programs, Maq and GSNAP to Sequencer yourself and use those algorithms to align your Next Generation sequences to a reference sequence. Please see [Using External Tools with Sequencer](#) for detailed help in setting up your machine to use Maq and GSNAP as well as the associated viewers, Tablet and Maqview.

Sequencer accepts many sequence formats. For Next Generation sequencing, the most relevant formats are FastA and FastQ. Although most sequencers have their own native formats, as long as the data can be converted into FastA or FastQ format, Sequencer will be able to align the reads.

You may align Single or Paired-end data to the reference sequence using either algorithm. Maq is best suited to reads data up to 127 bases in length while GSNAP supports reads from 14 to 500 bases (and may be configured for even longer read lengths). The sequences should be in FastQ format for Maq and FastA or FastQ format for GSNAP. For more detailed information on how to prepare NGS data for use with Sequencer, please see [Using External Tools with Sequencer](#).

Alignment with both Maq and GSNAP follows the same basic pattern:

1. Import the genomic reference sequence into the project.
2. Select the reference sequence and choose Align Data Files to Ref Using either Maq or GSNAP.
3. Select Input Data Files.
4. Choose whether or not to include any Additional Analysis option.
5. Decide whether or not to immediately open the aligned reads in a viewer.
6. Click the Align button.

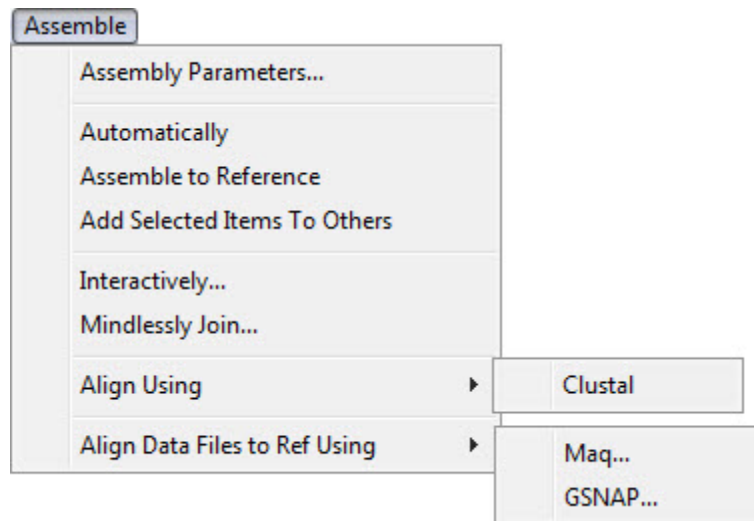
The screenshot shows the 'Align Data Files' dialog box in Sequencer. It is divided into several sections: 'Input Data Files' with 'Select File 1' (B:\Genomic Alignment Data\Ecoli_PEREads\EcoliReads1.FQ) and 'Select File 2' (Optional), a 'FASTQ Encoding' dropdown set to 'Sanger Standard', 'Additional Analysis' with a 'None' dropdown and a 'Known SNPs File' button, 'Current Results Folder' set to 'M:\NGS Data Results\Gene Codes\Sequencer\GSNAP', and 'View Results Using' with radio buttons for 'Tablet' (selected) and 'None'. At the bottom are 'Cancel' and 'Align' buttons.

Regardless of whether you choose to view the aligned reads immediately, the consensus results aligned to the Reference Sequence will be added to your current project. The generated data files are saved, and may be viewed later by selecting the consensus sequence or contig in your project and choosing the **Contig > Show NGS Data Using** command.

For more information on how to use all of the new NGS features of Sequencer please see the tutorial [Next Gen Sequence Alignment](#). Both the NGS algorithm installation guide—Using External Tools with Sequencer—and the Next Gen Sequence Alignment tutorial are available for download from the Gene Codes website at www.genecodes.com/training/tutorials. They are also available in the Tutorials folder on the Sequencer 5.0 CD and installed with Sequencer on your computer.

NEW ASSEMBLE MENU

There have been some changes to Sequencer's menus. A new menu named **Assemble** has been added.



All of the Assembly-related commands that were previously listed on the **Contig** menu (**Assembly Parameters** and the **Assemble Contigs** submenu) have been re-located to the new **Assemble** menu.

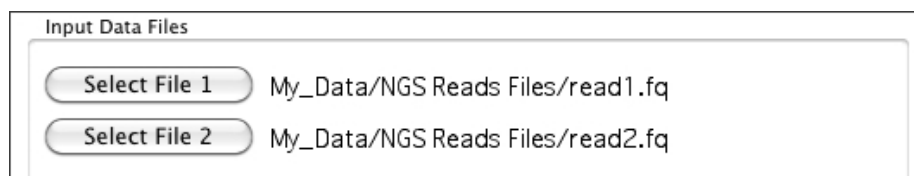
There are two **Align** items on the menu. **Align Using** has a submenu for **Clustal** (which was introduced in Sequencer 4.9). The most exciting additions to this menu are those relating to Next Generation Sequence alignment. **Maq...** and **GSNAP...** are the two current items on the new **Align Data Files to Ref Using** submenu.

VIEW NEXT GEN SEQUENCE ALIGNMENT RESULTS

To align your Next Gen sequences in Sequencer using the Maq alignment algorithm, you must first have a sequence in the project to align with. You can use any sequence in your Sequencer project to act as the reference. It does not have to be marked as a Reference Sequence using **Sequence > Reference Sequence** first.

Launch Sequencer and import your reference sequence. Note that you could use any sequence in your Sequencer project to act as the reference in this instance. It does not have to be marked as a Reference Sequence using **Sequence > Reference Sequence** first.

Select the reference sequence and choose **Assemble > Align Data Files to Ref Using > Maq...** or **GSNAP...**



Choose your first reads file by clicking on the **Select File 1** button. Browse to the file you want to use, select the file, and then click on the **Open** button. Next, optionally choose the second reads file by clicking on the **Select File 2** button and open your choice in

the same manner. Click the **Align** button. By default the aligned reads will open in Tablet immediately upon completion of the alignment.

View Results Using

☒ Tablet
 ☐ Maqview
 ☐ None

Regardless of whether you choose to view the aligned reads immediately, the consensus results aligned to the Reference Sequence will be added to your current project and may be viewed later.

Once the alignment is complete, the resulting consensus aligned to your selected Reference Sequence will be added to your current project. The contig will be named based on the alignment algorithm used. Maq generated contig consensi are calculated by Maq as part of the alignment process. GSNAP generated contig consensi are calculated by Sequencher using Plurality rules.

Name	Size	Quality	Kind	Comments
Mycoplasma_5'	250000 BPs		Ref: DNA Fragment	
Maq Aligned To Mycoplasma_5'	250000 BPs	100.0%	Ref: Contig of 2	Maq generated
Maq Consensus To Mycoplasma_5'	249997 BPs	100.0%	DNA Fragment	Maq generated
Mycoplasma_5'	250000 BPs		Ref: DNA Fragment	

Information contained in the comments for any GSNAP or Maq generated contigs and consensus sequences include the names of the read files used for the alignment, the paths to those reads files and the location of the results file. For the contig in the above example, you have the following information:

Maq Aligned To Mycoplasma_5'

Kind : Ref: Contig of 2

Size : 250000 BPs

Where : In project window.

Original : My_Data:NGS Data Alignment Results:Gene Codes :Sequencher :Maq:Runa47f47a43bf42d2

Created : Mon, May 16, 2011, 8:54 PM

Modified : Mon, May 16, 2011, 8:54 PM

Version : Modifiable.

Comments :

Maq generated: May 16, 2011 20:54:47 from My_Data:NGS Reads Files:read1.fq and My_Data:NGS Reads Files:read2.fq, External Data Location :Maq/Runa47f47a43bf42d2

Base Count : 87079 As, 80572 Ts

39010 Cs, 43189 Gs

147 Ambigs, 110 Disagreements, 3 Gaps

You may open the contig in a Contig Editor or compare the NGS aligned Consensus to the Reference Sequence or any other Contig operation you choose.

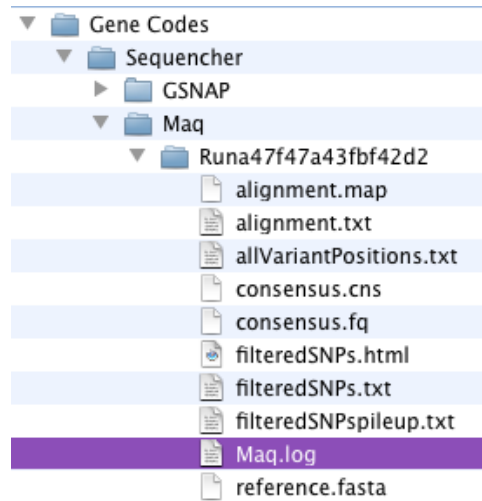
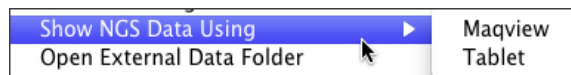
Maq Consensus To Mycoplasma_5'	AACAAACCTATCAAGTAGTGTATCCGTAACCTATCTAATTAATAATGAACCTTGAAAAAGCTAACCAACTTTTAGTGGA
Mycoplasma_5'	AACAAACCTATCAAGTAGTGTATCCGTAACCTATCTAATTAATAATGAACCTTGAAAAAGCTAACCAACTTTTAGTGGA
2 frag bases selected at consensus position 184,447	184410 184420 184430 184440 184450 184460 184470 184480
Select Next Ambiguous Base = spacebar	AACAAACCTATCAAGTAGTGTATCCGTAACCTATCTAATTAATAATGAACCTTGAAAAAGCTAACCAACTTTTAGTGGA
CDS riboflavin biosynthesis protein RibF.....

In addition to the standard contig operations, there are several additional features available for a Maq or GSNAP generated contig and consensus sequence.

The supporting data result files as well as a log listing the history of the alignment are saved in a unique location for every run. This location is saved with the new config in your project, and the data may be accessed at any point.

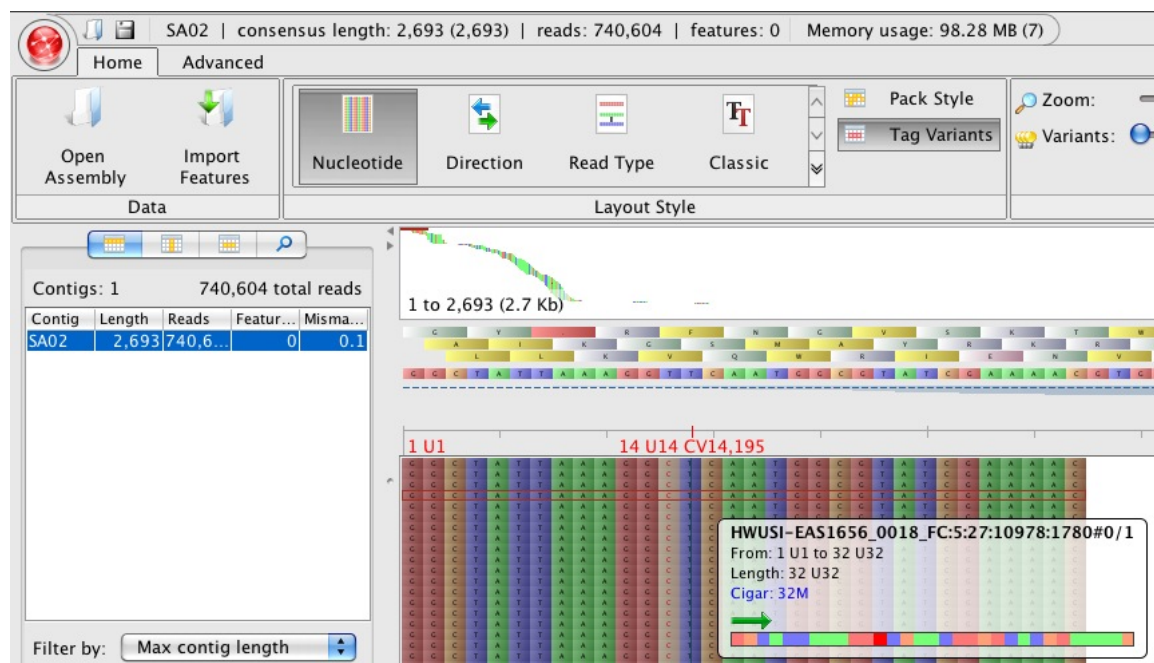
To review the log, select the Contig or Consensus and choose the **Open External Data Folder** from the right-click menu or **Windows** menu. The associated Run folder will open.

To view the detailed aligned reads associated with the Maq or GSNAP generated data in your Sequencer project, select the contig and choose the **Show NGS Data Using** command from the right-click or **Contig** menu.



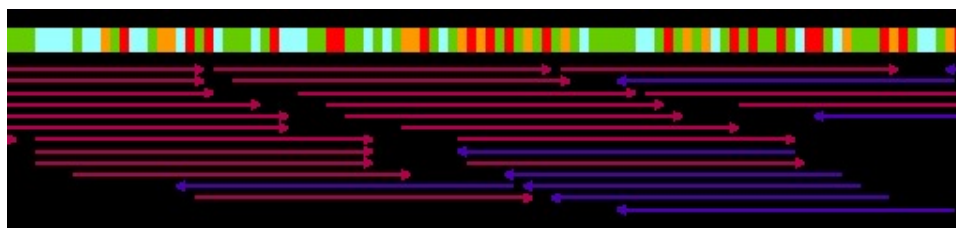
Maq alignments may be viewed in either Maqview or Tablet. GSNAP alignments may be viewed in Tablet.

If you choose Tablet as your viewer option, your aligned reads and Reference Sequence will be displayed. Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments available from The James Hutton Institute. By default, all Maq and GSNAP alignments will open in Tablet as soon as the alignments are complete. Once Tablet is open, click on the contig in the list at the left hand side of the window and explore the displayed alignment.



You can move about and change the look of the view. Using the “Zoom” slider increases and decreases the size of the reads bases. Using the “Variants” slider increases or decreases the background grey on which the bases are displayed so that moving the slider to the extreme right will effectively hide normal bases while potential SNPs will be highlighted. Click on the “Direction” button changes the color of the bases from their normal individual base color to dark blue or khaki depending on the direction of the read. The “Classic” view shows the bases without any background coloration.

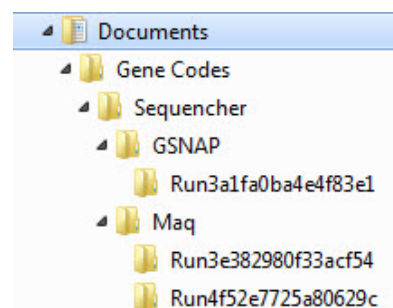
If, instead, you choose the Maqview program, then a console window will open and display the reads. Maqview has different views; to move between these views use the function keys on your keyboard. F1 switches to the bases mode, F2 switches to the colored blocks mode, and F3 switches to the overview mode. In the F2 and F3 views you can zoom in and out using + and – keys.



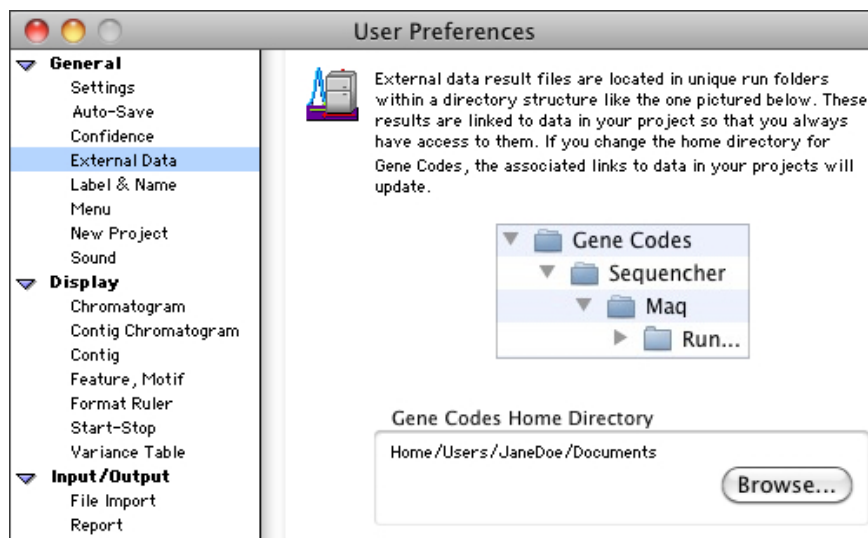
See the instructions in [Using External Tools with Sequencer](#) for information on how to download and install Tablet and Maqview.

CHOOSING THE EXTERNAL DATA HOME DIRECTORY

Each invocation of a Maq or GSNAP alignment within Sequencer will create several data and log files written to a unique Run folder located in the Gene Codes Home Directory. Links to these unique locations are stored with Maq and GSNAP results in your project file. Although you may not change the Gene Codes directory structure without destroying these links, you may choose to change the Gene Codes Home Directory location to save this data wherever you wish. By default, the Gene Codes Home directory is in your user Documents folder.



After a few runs, you may decide to change the location of the Gene Codes Home Directory. Do this by going to the **External Data** pane in Sequencer's **User Preferences** and clicking on the **Browse...** button to choose a different location.

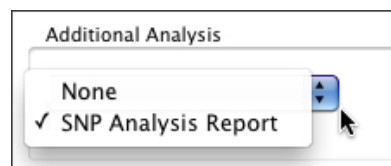


The value set in the **External Data User Preferences** pane for the Gene Codes Home Directory will be remembered and all new alignment results will be written to this new location. Note that if you do decide to move the Home Directory elsewhere, you will need to move any existing Run folders and place them in the Home Directory in the new location. This will maintain referential integrity between your projects, their contigs, and the associated Run folders that contain the data results.

SNP ANALYSIS WITH MAQ

You may perform a SNP analysis in addition to alignment when choosing **Align Data Files to Ref Using > Maq**. Choose your Input Data Files as you would normally and then select **SNP Analysis Report** from the **Additional Analysis** drop-down menu.

If you chose to generate a SNP Analysis Report as part of the Maq alignment, then, in addition to viewing the aligned reads in Tablet or Maqview, you may open the Maq SNP Report. Select the Maq generated contig or consensus sequence and choose the **Sequence > Analyses > Maq SNP Report** command. The results will be displayed in your default browser.



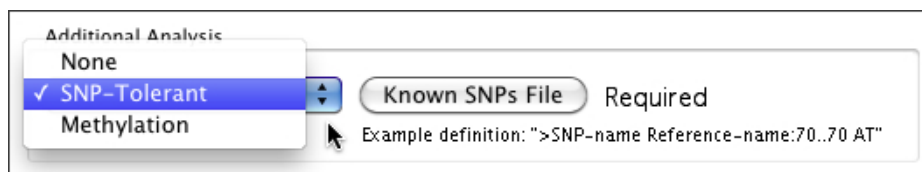
Sequencher Maq SNP Report											
Gene Codes Corporation											
T C A G G E N E A G T C O D E S											
Mycoplasma_5'	348	T	A	81	18	1.00	63	62	N	176	N
Mycoplasma_5'	392	T	A	99	24	1.00	63	62	N	173	N
Mycoplasma_5'	587	T	A	57	10	1.00	63	54	N	104	N

The first column is the name of the reference sequence, the second column is the position of the reference base, next is the reference base itself, followed by variant base. Column 5 contains a Phred-like quality value that will assist in determining the reliability of the SNP. Column 6 gives you the read depth. When the table reports 1.00 in column 7, this means that the region is near enough unique. The final columns relate to the second-best and third-best base calls at this position.

SNP ANALYSIS WITH GSNAP

Similarly, you may perform additional analyses when choosing **Align Data Files to Ref Using > GSNAP**. GSNAP performs SNP hunting in a different way than Maq. You must supply a file that contains a list of known SNPs.

To perform a SNP-Tolerant alignment using GSNAP, choose **SNP-Tolerant** from the **Additional Analysis** drop-down menu.



The **Known SNPs File** button will then be enabled. Click on that button and browse to a file containing the list of known SNPs. This text file has to list each SNP in a specific format—one SNP per line.

Name	Size	
My Hflu Ref	1985832 BPs	>rs004341 My_Hflu_Ref:65..65 AT >rs004342 My_Hflu_Ref:154..154 AT >rs004343 My_Hflu_Ref:227..227 CG >rs004344 My_Hflu_Ref:632..632 AC >rs004345 My_Hflu_Ref:1396..1396 AG >rs004346 My_Hflu_Ref:1413..1413 GT

Each line must begin with the > character followed by a SNP identifier. In the example above, it is an rs number. The next pieces of data are reference and positional information in the format RefName:#..# followed by a major and minor allele. In real data, the reference name to use before the colon is the name of the sequence you select in the Project Window. If there are spaces in the name, these should be replaced with underscores. In the example above, 'My Hflu Ref' is selected for a SNP-tolerant GSNAP alignment. The reference identifier used in the known SNP file is therefore My_Hflu_Ref. The position information in this file always assumes that the first base of the reference sequence in Sequencher is 1, no matter what its actual numbering relative to its chromosomal or contig position is.

If you chose to perform a GSNAP SNP-Tolerant analysis as part of the GSNAP alignment, you can view the results of the generated SNP-Tolerant analysis by selecting the contig of interest in the Project Window and choosing the **Sequence > Analyses > GSNAP SNP Analysis** command. The results will be displayed in your default browser. For more details regarding the GSNAP SNP-tolerant alignment results, please see the tutorial, [Next Gen Sequence Alignment](#).

METHYLATION ANALYSIS

To generate a Methylation analysis report using GSNAP instead, choose **Methylation** from the **Additional Analysis** drop-down menu. Studies of methylated parts of the genome involve using bisulfite treatment of the DNA following sequencing of the regions of interest. The bisulfite treatment chemically converts unmethylated Cs to Ts. GSNAP is capable of aligning the sequenced reads to genomic (untreated) sequence.

If you chose to perform a GSNAP Methylation analysis as part of the GSNAP alignment, the resulting GSNAP generated contig will be added to the Sequencer project with an Inclusively calculated consensus. You may view the results of the generated Methylation analysis by selecting the contig of interest in the Project Window and choosing the **Sequence > Analyses > GSNAP Methylation Analysis** command. The results will be displayed in your default browser. The top line of each block is the aligned read sequence. Note the periods in the sequence underneath. They represent Cs in the corresponding genomic reference. For more details regarding these Methylation results, please see the tutorial [Next Gen Sequence Alignment](#).

Sequencher GSNAP Methylation Analysis			Gene Codes Corporation	
			T C A	GENE
			A G T	CODES
>AATAATAATAATATTTAAAGTTAAATTAATTTATTAATTAATTAAGGTTAAAGTTAAAT	1	Frag[0001]_0-63_0		
AA.AA.AATAATA..AAAAG..AAAATTA..ATTAAAT.AATTAAGGTTAAAG..AAAAT	1..63	-methylation_reference:63..1		start:0..
>ATTTGGTTTAAATTTAATTGATTTAATGGGTTTAAATTTGGTTTTGGTATTATTGTTGT	1	Frag[0001]_0-63_1		
ATTTGG.TTTAA..TTAATTGATTTAATGGGTTTAAATTTGG.TTTGGTATTATTGTTGT	1..63	+methylation_reference:1..63		start:0..
>TTTTGGTTTAAATTTAATTGATTTAATGGGTTTAAATTTGGTTTTGGTATTATTGTTGTTA	1	Frag[0001]_1-64_1		
TTTTGG.TTTAA..TTAATTGATTTAATGGGTTTAAATTTGG.TTTGGTATTATTGTTGTTA	1..63	+methylation_reference:2..64		start:0..
>TTGTTAAAGGTTGTTATAGAAATTATTAATAAATGAATTTAAATTAATGTTAAGGAAT	1	Frag[0001]_100-163_0		
.TGTTAAAGG.TG.TATAGAA.TA..AAAATAAATGAA.TTAAAAATTAATG.TAAGGAAT	1..63	-methylation_reference:163..101		start:0..

SIMPLE NCBI BLAST SEARCH

The United States Centers for Disease Control asked us to put in a simple NCBI Blast search for rapid checking of short viral fragments.

With this feature, you may choose any sequence or contig in your Sequencher project and send it to NCBI's Blast.

Select a single sequence or contig in the project window, or select a range of bases from any window and choose **Sequence > NCBI Blast Search** to initiate the search. The search is limited to the first 7800bp (Mac OS X) or 1600bp (Windows) of the selected sequence or consensus sequence. Click on the View Report button in the opened web browser to see the actual results.

Depending on the speed of your Internet connection, you will be able to query and verify short sequences rapidly and locate features or find related sequences while still working on your project.

* Thomas D. Wu and Colin K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 2005 21: 1859-1875. And Thomas D. Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010 26: 873-881.

† Li, et al, *Genome Research*, 2008 and Thompson, et al, *Nucleic Acids Res.*, 1994 Nov 11.

‡ Li, H., Ruan, J. & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18, 1851-1858.

§ Milne, I., et al. (2010) Tablet—next generation sequence assembly visualization. *Bioinformatics*, 26, 401-402.