SEQUENCHER®

Tutorial for Windows and Macintosh

Preparing Your Data for NGS Alignment

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

Preparing Your Data for NGS Alignment

Preparing Your Data for NGS Alignment

Sequencher accepts many sequence formats. For Next-Generation sequencing, your reads should be in **FastA** or **FastQ** format. Although most sequencers have their own native formats, as long as the data can be converted into **FastA** or **FastQ** format, **Sequencher** will be able to align the reads.

When aligning sequences using **BWA-MEM**, **Maq**, or **GSNAP**, you will be using a reference sequence that has been imported into **Sequencher**. The advantage of using a GenBank sequence is that it will usually carry annotations. If you are performing a de novo assembly with the **Velvet** algorithms, a reference sequence is not required.

All the algorithms for Next-Generation sequence analysis will work with single-end or paired-end data. Each algorithm has its own requirements for data input files that may require some modification to your files in advance of performing the alignment in order to succeed.

One thing to note is that, although it is possible to get **FastQ** format files from both Illumina and 454, the formats differ in how they encode confidence scores. Both represent confidence scores in single ASCII characters, but the key to decode them back into scores is different. 454 data conforms to what is popularly known as Sanger standard format. Illumina represents scores using a different range of ASCII characters, unless your pipeline is Casava 1.8, in which case it is equivalent to Sanger standard format. You will need to specify which FastQ variant you'll be using when aligning with **GSNAP** using the **FASTQ Encoding** drop-down menu.

Data File Requirements for Multiplex ID (MID) Files

Valid entries for a barcodes file consist of a barcode name followed by a tab character followed by the barcode sequence itself. Barcode sequences must all be of the same length.

MID1 TCAGATATCGCGAG

Data File Requirements for GSNAP

- **GSNAP** supports both **FastA** and **FastQ** file formats (both Sanger standard and Illumina variants).
- Read lengths may vary in size and fall within the range of 14bp to 1500bp. **GSNAP** may be configured for even longer read lengths though.
- 2 FastQ files will be treated as paired-end reads and 1 FastQ file will be treated as single-ends reads. Paired-end data in **FastQ** format must list the reads in the same order in both files. Here is an example:

```
File 1:
       @NC_014230.1|_1831264_1831446_0/1
       TCTCCATAAGTTGAGATAAGTTAGAAACCAAGTGTT
       +
       &IIIIIII+IIIIIIIBBII)$&IIII>.&+%E*I0
       @NC_014230.1|_1066261_1066432_1/1
       GCTGAACTTGCATAATAGTGGACCAATCATAAGAAT
       +
       D!"IIIIIII$!!*(&%.$IIIIIIIIIII3&III
File 2:
       @NC_014230.1|_1831264_1831446_0/2
       ATAGGATTCAAGGCAGATTTAAAATTGACGGCGCGC
       +
       III<IIIIIIIIIIIIIIII'%IICII3/II>+=
       @NC_014230.1|_1066261_1066432_1/2
       AATCCTGGTAACAAAATGTTTTTACATTATAGCCTA
       +
       IIIIICIIIIIIIADIIIIIIIIIIIIIIII0II
```

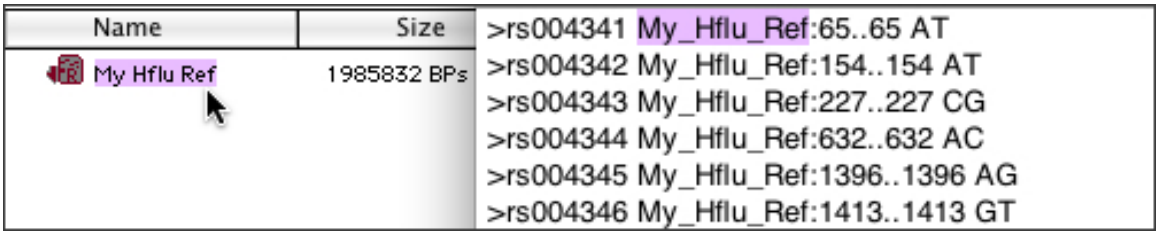
- FastA files may also represent both single-ends and paired-ends reads. **GSNAP** has specific requirements for modifying the basic **FastA** format for alignment.
    - The entire read must be on a single line–no line breaks in the DNA sections.
    - If you have paired-ends data, the second read must be on the next line. For example:

```
>name sequence header information
ATGAACAGGCGCGATCTTCTTTTACAAGAAATGGGCATTTCCCAGTGGGA
GAATGTAAGCAGCCTATTCGTTATTGGTTACTATCAGAAAATAGCGACCA
>next-sequence
CACTTTGCCATTTTGCAAGCAGGCTGAGCAGGTTTATCGC
TATCGCCCCGAGGTACTGCAAGGTTCAGTAGGAATTAGTG
```

  In the above example, there are 4 DNA sequences–2 pairs, not 2 sequences.

Preparing a Known SNPs Data File For GSNAP

To perform an **SNP-Tolerant** alignment using **GSNAP**, you must provide a file containing the list of known SNPs. **GSNAP** performs SNP hunting in a different way than **Maq**. This text file has to list each SNP in a specific format–one SNP per line.



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 an SNP-tolerant **GSNAP** alignment. The reference identifier used in

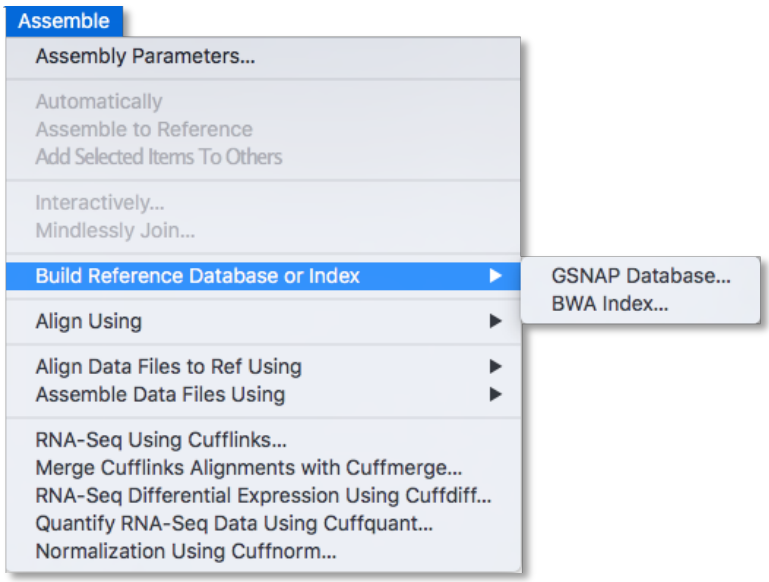
the known SNP file is therefore My_Hflu_Ref. The position information in this file always assumes that the first base of the reference sequence in **Sequencher** is 1, no matter what its actual numbering relative to its chromosomal or contig position is.

Building Your GSNAP Reference Sequence Database

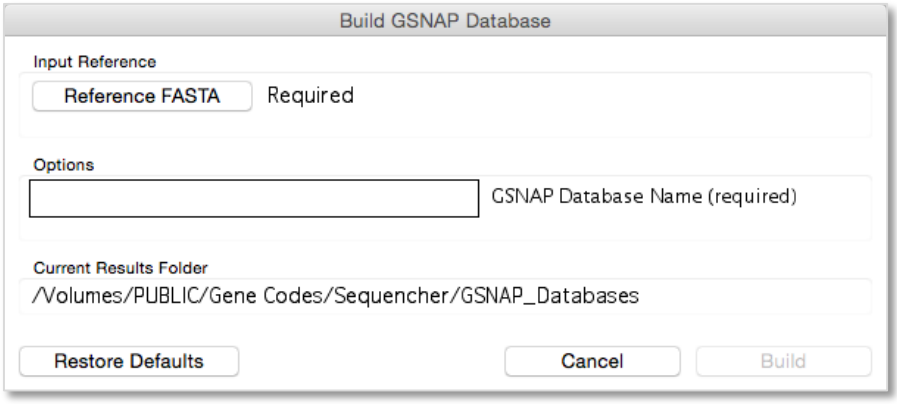
You can use a single sequence in **FastA** format or a series of sequences in concatenated **FastA** format (for example a series of chromosomes) on your file system or you can use a sequence in your **Sequencher** project. If you plan to perform an in silico experiment such as comparing the **BWA-MEM** algorithm to the **GSNAP** algorithm, you will need to build a database/index for both aligners.

The files for the database are placed in the **External Data Home** (the default location for this is your **Documents** folder but you may change this through a **User Preference**). You will be able to give the database a name of your own choosing. In addition, information about the database run will be visible in the **External Data Browser**. You can launch this by choosing **Window>Open External Data Browser**. The **External Data Browser** will also open automatically when you select either of the **Build Reference Database or Index** commands.

Choose **Assemble>Build Reference Database or Index>GSNAP Database…** to use the reference sequence with the **GSNAP** algorithm.

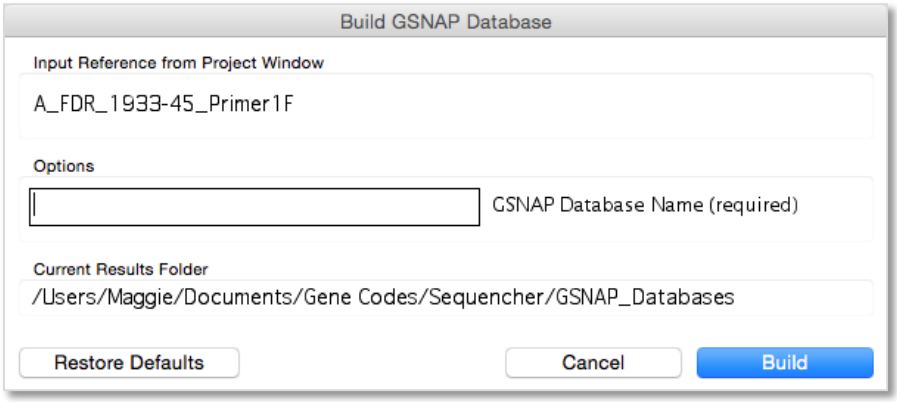


If you have no selection in the **Project Window**, the following **Build GSNAP Database** dialog appears.



Click on the **Reference FASTA** button and browse to the FastA file containing your reference sequence(s). Click on the **Open** button. Give your database a name and click on the **Build** button.

If instead you have a selection in the **Project Window**, the following **Build GSNAP Database** dialog appears.



Note that the name of your **Input Reference from Project Window** sequence is the same as the name of the sequence you have selected in the **Project Window**. Give your database a name and click on the **Build** button.

If you are using the **External Data Browser** to monitor the progress of the build (especially useful with large genomes), click on the **Refresh** button from time to time and note the progress in the log file tab. If the **Auto Refresh On** control is on, refreshes will occur periodically.

Once **Sequencher** has finished the creation of the database, you will see a green **SUCCESS** status in the Final Run Status column. If there was a problem with the creation of the database, you may see a red **FAILED** status. Consulting the log file will often give you a clue as to the reason for the failure. Very rarely there will be no status message at all. This is also an indicator of problems with the creation/build run.

The **Notes** field allows you to add pertinent information about your reference sequence. This is a good place to record information such as the build version for complete or partial genomes. **Sequencher** will automatically add the name of the sequence file used to create the database to the **Notes** field.

Once the database is built, you will see it listed in a drop-down menu the next time you use **GSNAP**.

Data File Requirements for BWA-MEM

- **BWA-MEM** supports both **FastA** and **FastQ** file formats (both Sanger standard and Illumina variants).
- 2 FastQ files will be treated as paired-end reads and 1 FastQ file will be treated as single-ends reads. Paired-end data in **FastQ** format must list the reads in the same order in both files. Here is an example:

```
File 1:
       @NC_014230.1|_1831264_1831446_0/1
       TCTCCATAAGTTGAGATAAGTTAGAAACCAAGTGTT
       +
       &IIIIIII+IIIIIIIBBII)$&IIII>.&+%E*I0
       @NC_014230.1|_1066261_1066432_1/1
       GCTGAACTTGCATAATAGTGGACCAATCATAAGAAT
       +
       D!"IIIIIII$!!*(&%.$IIIIIIIIIIII3&III


File 2:
       @NC_014230.1|_1831264_1831446_0/2
       ATAGGATTCAAGGCAGATTTAAAATTGACGGCGCGC
       +
       III<IIIIIIIIIIIIIIIII'%IICII3/II>+=
       @NC_014230.1|_1066261_1066432_1/2
       AATCCTGGTAACAAAATGTTTTTACATTATAGCCTA
       +
       IIIIICIIIIIIIIADIIIIIIIIIIIIIIIII0II
```

- Paired-end data in **FastA** format must list the reads in the same order in both files. **BWA-MEM** has specific requirements for modifying the basic **FastA** format for assembly.
- The entire read must be on a single line–no line breaks in the DNA sections.

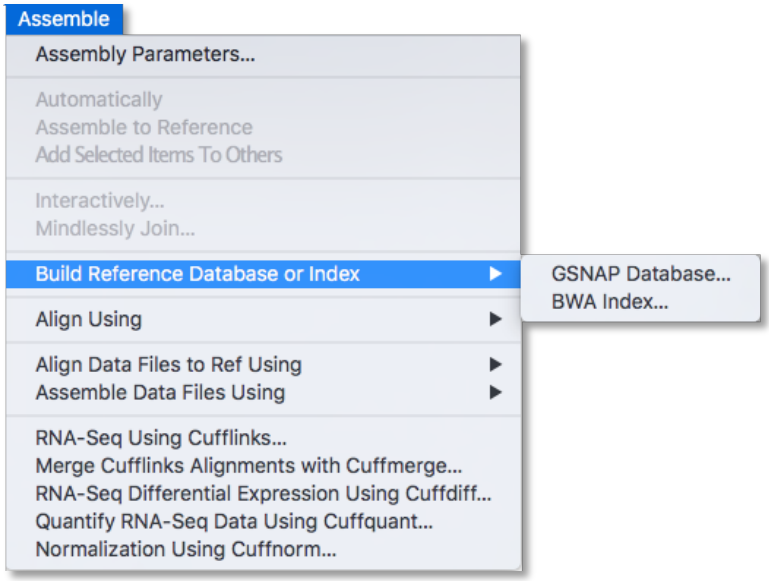
Building Your BWA-MEM Reference Sequence Index

You can use a single sequence in **FastA** format or a series of sequences in concatenated **FastA** format (for example a series of chromosomes) on your file system or you can use a sequence in your **Sequencher** project. If you plan to perform an in silico experiment such as comparing the **BWA-MEM** algorithm to the **GSNAP** algorithm, you will need to build a database/index for both aligners.

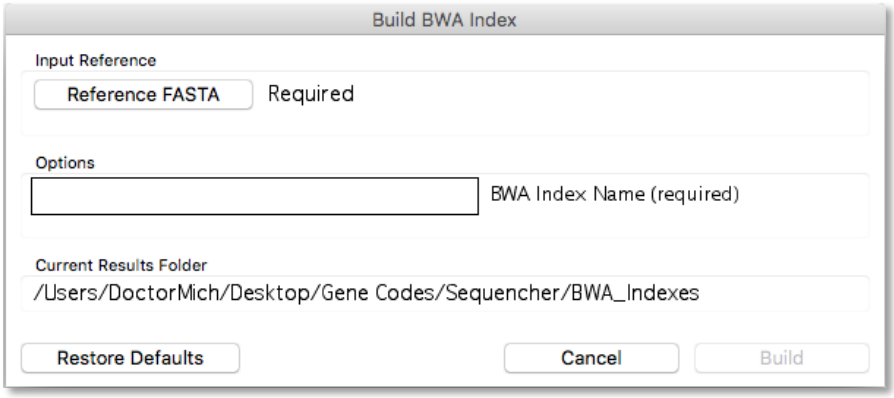
The files for the index are placed in the **External Data Home** (the default location for this is your **Documents** folder but you may change this through a **User Preference**). You will be able to give the

index a name of your own choosing. In addition, information about the index run will be visible in the **External Data Browser**. You can launch this by choosing **Window>Open External Data Browser**. The **External Data Browser** will also open automatically when you select either of the **Build Reference Database or Index** commands.

Choose **Assemble>Build Reference Database or Index>BWA Index…** to use the reference sequence with the **BWA-MEM** algorithm.

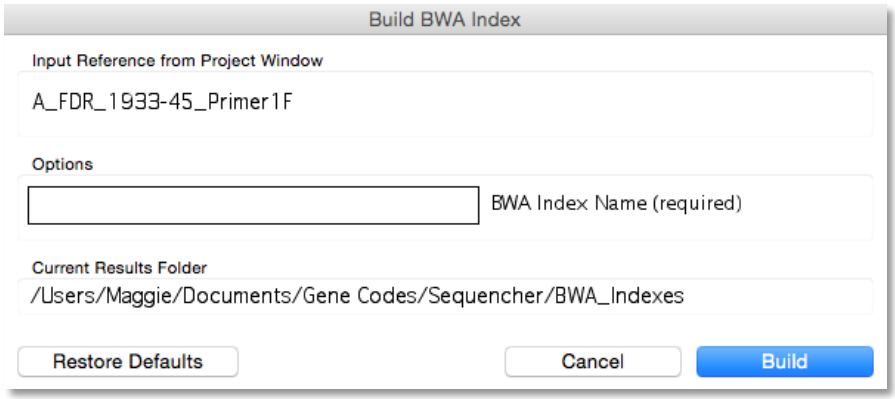


If you have no selection in the **Project Window**, the following **Build BWA Index** dialog appears.



Click on the **Reference FASTA** button and browse to the FastA file containing your reference sequence(s). Click on the **Open** button. Give your index a name and click on the **Build** button.

If instead you have a selection in the **Project Window**, the following **Build BWA Index** dialog appears.



Note that the name of your Input Reference from **Project Window** sequence is the same as the name of the sequence you have selected in the **Project Window**. Give your index a name and click on the **Build** button.

If you are using the **External Data Browser** to monitor the progress of the build (especially useful with large genomes), click on the **Refresh** button from time to time and note the progress in the log file tab. If the **Auto Refresh On** control is on, refreshes will occur periodically.

Once **Sequencher** has finished the creation of the index, you will see a green **SUCCESS** status in the Final Run Status column. If there was a problem with the creation of the index, you may see a red **FAILED** status. Consulting the log file will often give you a clue as to the reason for the failure. Very rarely there will be no status message at all. This is also an indicator of problems with the creation/build run.

The **Notes** field allows you to add pertinent information about your reference sequence. This is a good place to record information such as the build version for complete or partial genomes. **Sequencher** will automatically add the name of the sequence file used to create the index to the **Notes** field.

Once the index is built, you will see it listed in a drop-down menu the next time you use **BWA-MEM**.

Data File Requirements for Velvet

- **Velvet** supports both **FastA** and **FastQ** file formats (both Sanger standard and Illumina variants).
- **Velvet** may be configured for longer k-mers.
- 2 FastQ files will be treated as paired-end reads and 1 FastQ file will be treated as single-ends reads. Paired-end data in **FastQ** format must list the reads in the same order in both files. Here is an example:

```
File 1:
       @NC_014230.1|_1831264_1831446_0/1
       TCTCCATAAGTTGAGATAAGTTAGAAACCAAGTGTT
       +
```

```
&IIIIIII+IIIIIIIBBII)$&IIII>.&+%E*I0
@NC_014230.1|_1066261_1066432_1/1
GCTGAACTTGCATAATAGTGGACCAATCATAAGAAT
+
D!"IIIIIII$!!*(&%.$IIIIIIIIIII3&III


File 2:
@NC_014230.1|_1831264_1831446_0/2
ATAGGATTCAAGGCAGATTTAAAATTGACGGCGCGC
+
III<IIIIIIIIIIIIIIIII'%IICII3/II>+=
@NC_014230.1|_1066261_1066432_1/2
AATCCTGGTAACAAAATGTTTTTACATTATAGCCTA
+
IIIIICIIIIIIIIADIIIIIIIIIIIIIIIII0II
```

- Paired-end data in **FastA** format must list the reads in the same order in both files. **Velvet** has specific requirements for modifying the basic **FastA** format for assembly.
- The entire read must be on a single line–no line breaks in the DNA sections.
- If you have paired-ends data, the second read must be on the next line. For example:

```
>name sequence header information
ATGAACAGGCGCGATCTTCTTTTACAAGAAATGGGCATTTCCCAGTGGGA
GAATGTAAGCAGCCTATTCGTTATTGGTTACTATCAGAAAATAGCGACCA
>next-sequence
CACTTTGCCATTTTGCAAGCAGGCTGAGCAGGTTTATCGC
TATCGCCCCGAGGTACTGCAAGGTTCAGTAGGAATTAGTG
```

In the above example, there are 4 DNA sequences in 2 pairs, not 2 sequences.

Data File Requirements for Maq

- Must be in **FastQ** format. **Maq** expects Sanger standard encoded quality scores.
- Read lengths can be no greater than 127 bases and must be the same size for every read.
- Paired-ends data (2 FastQ files) and single-ends data (1 FastQ file) are supported.