SEQUENCHER®

Installing External Tools for Sequencher

Gene Codes Corporation

TCA GENE
AGT CODES

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

Installing External Tools for Sequencher

Quick and Easy Set Up for NGS Aligners and Assemblers 3
Installing External Applications on Windows 3
Installing the Tablet Viewer on Windows 3
Installing Clustal on Windows 3
Installing External Applications on Mac OS X 4
Installing the Tablet Viewer on Mac OS X 4
Installing Clustal on Mac OS X 5
Preparing Your Data for Sequence Alignment 6
Data File Requirements for Multiplex Identifier (MID) Files For Use with BWA-MEM, GSNAP, and Velvet: 6
Data File Requirements for GSNAP: 6
Preparing a Known SNPs Data File For GSNAP 7
Data File Requirements for BWA-MEM: 7
Data File Requirements for Velvet: 8
Data File Requirements for Maq: 9
More Information 9

Introduction

With the advent of Next-Generation sequencing, several new external alignment algorithms have been integrated into Sequencher to align the large quantity of data produced by Next-Generation sequencing machines. These algorithms, BWA-MEM, GSNAP, Maq, and Velvet, join Clustal (the very first external algorithm to be added to Sequencher). The MUSCLE external algorithm, while not an NGS algorithm, has also been added to Sequencher. The BWA-MEM, GSNAP, Maq, MUSCLE, and Velvet algorithms were written to run on Unix-based systems. Mac OS X is a Unix-based operating system. We have built native binaries for Windows, so that there is no longer any need to have Cygwin, a Unix emulation environment, installed in order to run the alignment tools. To take full advantage of these tools, you will also need to install Tablet—a third-party graphical viewer.

The instructions which follow show you how to use the all-in-one installers which will place almost everything you need to run these algorithms on your computer. The Tablet viewer is a separate installer but only takes moments to install.

Quick and Easy Set Up for NGS Aligners and Assemblers

Installing External Applications on Windows

The Windows installer will place BWA-MEM, GSNAP, Maq, Maqview, MUSCLE, and Velvet and their related files where they need to be. That location is Programs Files/Gene Codes/Sequencher External Tools.

1. Download the External Tools Installer zip archive from the Gene Codes website at www.genecodes.com/download/external-tools-download.
2. Extract/unzip the archive and open the installer folder. It will contain ExternalToolsInstaller.exe, Tablet, src, links, and ClustalW2 folders.
3. Double-click on *ExternalToolsInstaller.exe.* Accept the installation defaults including the license agreement terms.
4. A new window appears which shows the progress of the installer. When the installation is complete, BWA-MEM, GSNAP, Maq, Maqview, MUSCLE, and Velvet are ready for use in Sequencher.

Installing the Tablet Viewer on Windows

To install the Tablet viewer on your computer, follow these steps:

1. Open the Tablet folder. There are two installers. The tablet_windows_x64 installer is for 64-bit systems and the tablet_windows_x86 installer is for 32-bit systems.
2. Tablet requires administrator privileges in order to install. If you are an Administrator user, double-click on the correct Tablet installer. If you are not an Administrator user, get assistance with getting administrator privileges so you can install.
3. Accept all of the installation defaults. You will be prompted for a destination location for the installation. The default location is Program Files\Tablet. Don't change this location.

Installing Clustal on Windows

To install Clustal on your computer, follow these steps:

1. Open the ClustalW2 folder.

2. Clustal requires administrator privileges in order to install. If you are an Administrator user, double-click on the Clustal installer. If you are not an Administrator user, get assistance with getting administrator privileges so you can install.
3. Accept all of the installation defaults. You will be prompted for a destination location for the installation. The default location is Program Files\ClustalW2 for 32-bit systems or Program Files (x86)\ClustalW2 for 64-bit systems. Don't change this location.

INSTALLING EXTERNAL APPLICATIONS ON MAC OS X

You must be an Administrator or have an administrator's password in order to install the external tools on Mac. If you are installing on a computer with OS X 10.8 (Mountain Lion) or above, change your security setting in User Preferences to Allow applications downlaoded from: Anywhere. The Mac OS X installer will place BWA-MEM, GSNAP, Maq, Maqview, MUSCLE, Velvet, and their related files where they need to be.

1. Download the External Tools installer disk image file (.dmg) from the Gene Codes website at www.genecodes.com/download/external-tools-download.
2. Double-click on the downloaded dmg file.
3. Double-click on the new volume icon named 'External Tools Installers' to open it if it doesn't open automatically.
4. Double-click on the ExternalTools package to begin installation.
5. Accept installation defaults including the license agreement.
6. The installer starts. You will see various status messages as the files are written to your system. Finally, a screen appears telling you that everything was installed successfully. Click on the **Close** button. The external tools, BWA-MEM, GSNAP, Maq, Maqview, MUSCLE, and Velvet are now ready for use in Sequencher.
7. Eject the disk image icon from your desktop unless you are planning to install Clustal or the Tablet viewer next.

INSTALLING THE TABLET VIEWER ON MAC OS X

To install the Tablet viewer on your computer, follow these steps:

1. By default, Tablet will install in the Applications folder and that is where Sequencher expects to find it. Because installing in the Applications folder requires administrator privileges, you must be an administrator to install.
2. Double-click on the Other Installers folder icon to open the folder.
3. If you are on Mountain Lion or above, change the security setting for 'Allow applications downloaded from:' to Anywhere.
4. Double-click on the tablet.dmg file to open the installer.
5. Double-click on the Tablet Installer to start the installation.
6. Accept all of the installation defaults. You are prompted for a destination location for the installation. The default location is Applications. Don't change this location.
7. Eject the disk image icon from your desktop if you are done.
8. If you are on Mountain Lion or above and you changed the 'Allow applications downloaded from:' security setting to Anywhere, change it back to 'Mac App Store and identified developers'.

INSTALLING CLUSTAL ON MAC OS X

To install Clustal on your computer, follow these steps:

1. **By default, Clustalw2 needs to be in the Applications folder and that is where Sequencher expects to find it.**
2. **Double-click on the Other Installers folder icon to open the folder.**
3. **Double-click on the clustalw2.dmg file.**
4. **Double-click on the clustalw-2.1-macosx folder to open the folder.**
5. **Drag the clustalw2 application to the Applications folder on your Macintosh Hard Drive.**
6. **If you are on Mountain Lion or above, double-click on the clustalw2 application to run it once.**
7. **Eject the disk image icon from your desktop if you are done.**

Preparing Your Data for Sequence 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 Identifier (MID) Files For Use with BWA-MEM, GSNAP, and Velvet:

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)$&IIII>.&+%E*I0
       @NC_014230.1|_1066261_1066432_1/1
       GCTGAACTTGCATAATAGTGGACCAATCATAAGAAT
       +
       D!"IIIIIII$!!*(&%.$IIIIIIIIIIII3&III
File 2:
       @NC_014230.1|_1831264_1831446_0/2
       ATAGGATTCAAGGCAGATTTAAAATTGACGGCGCGC
       +
       III<IIIIIIIIIIIIIIIIII'%IICII3/II>+=
       @NC_014230.1|_1066261_1066432_1/2
       AATCCTGGTAACAAAATGTTTTTACATTATAGCCTA
```

```
        +
        IIIIICIIIIIIIIADIIIIIIIIIIIIIIIII0II
```

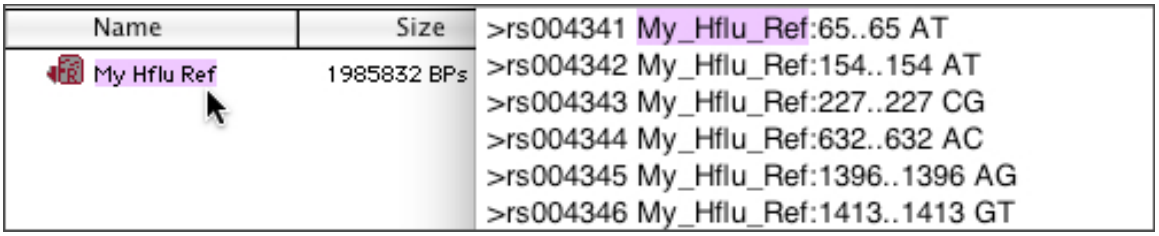
- FastA files may also represent both single-ends and paired-ends reads. GSNAP has specific requirements for modifying the basic FastA format for alignment.
    - ○ The entire read must be on a single line—no line breaks in the DNA sections.
    - ○ If you have paired-ends data, the second read must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



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 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.

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<IIIIIIIIIIIIIIIIII'%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.

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


File 2:
       @NC_014230.1|_1831264_1831446_0/2
       ATAGGATTCAAGGCAGATTTAAAATTGACGGCGCGC
       +
       III<IIIIIIIIIIIIIIIIII'%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—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.

MORE INFORMATION

If you would like more information on BWA-MEM, Maq, GSNAP, Velvet, or working with MIDs, refer to the NGS for DNA and RNA-Seq chapter of the Sequencher User Manual, chapter 16. You will find this manual in the Sequencher installation folder. You can also find additional information in the following tutorials : Next Gen Sequence Alignment, Advanced Next Gen Sequence Alignment, De Novo Sequence Assembly, and Multiplex IDs with BWA, GSNAP, and Velvet.