**AGEseq (Analysis of Genome Editing by Sequencing)**

Email Liangjiao Xue (lxue@uga.edu ) for help and bug reports.

AGEseq compares amplicon sequences with designed targets and finds the insertion/deletion sites in the amplicon sequences. It is written in Perl and calls BLAT software from the working directory or environment PATH.

AGEseq is available at AspenDB(<http://aspendb.uga.edu/downloads>).

Download the “AGEseq.zip” files and unzip it on your system, the AGEseq directory is the **working directory** during the analysis.

**1 Prepare dependent software**

1. Perl

a) Windows user:  
Download and install ActivePerl ( <http://www.activestate.com/activeperl/downloads>) .

Try the command line if there is difficulty with installation on Windows 8  
msiexec /i C:\dirctory\to\ActivePerl-XXX-MSWin32XXX.msi TARGETDIR="c:\perl" PERL\_PATH="Yes"

b) Other systems:

Perl is functional by default.

1. BLAT

a) Windows user:  
Download blat\_windows.zip from <http://aspendb.uga.edu/downloads>

Unzip the directory and copy the two files (blat.exe and cygwin1.dll) into the working directory.

b) Other systems:  
Go to download page: <http://hgdownload.cse.ucsc.edu/admin/exe/>  
Select platform version  
Scroll down application list and select “blat/” under the Name column  
Download “blat” and copy the **executable file** to the working directory or put blat holding directory in the PATH  
 export PATH=$PATH:/usr/blat\_dir  
Change the permission of blat by typing : chmod 770 blat

To run software with test data, go to **step 4** directly.

**2 Prepare read files**

The amplicon reads need to be put into the directory “reads”, which can be find in the working directory. Two testing files are there, which need to be deleted or moved before running real jobs.

Put read files of following types into directory “reads”:

1. fastq (end with .fastq or .fq), fq.gz files need to be unzipped at first
2. fasta with multiple sequences (end with .fasta, .fas, .fa, .seq or .txt)
3. fasta with single sequences or raw .seq files from sanger sequences (ending same as 2)

**3 Prepare design file**

Open “targets.txt” file in the working directory using Excel-like software. (On your system, this file may be shown as “targets” as the extension name is hidden. )

The contents look like this:

|  |  |
| --- | --- |
| targets | sequences |
| Amp1 | CAGTGCATCGATCGA(demo\_only) |
| Amp2 | CAGTGAATAGTTCAA(demo\_only) |

Edit the “targets” and “sequences”, and the sequences usually expand 30-40 bp at each side of the predicted editing site. Shorter sequences result in more precise annotation. Sequences of potential off-targeting sites can also be provided. Save the file.

**4 Run AGEseq**

Windows user:

**Double click** the **Run\_AGEseq.bat** file.

Mac user:

Change the permission of “**run\_AGEseq.sh**” by typing “ chmod 770 run\_AGEseq.sh “ in the working directory.

**Double click** the **Run\_AGEseq.sh** file.

All users:

The command to run AGEseq is like this:

cd /dir/to/AGEseq

**perl AGEseq.pl reads targets.txt All\_output.txt**

# 1 read directory # 2 design file # 3 output file

The first two inputs are required and output file (the third augment) is optional, All\_output.txt will be generated by default if the output name is not provided. The Perl script calls BLAT inside it; please make sure BLAT is in the current working directory or PATH environment.

**5. Understanding the output file**

The output needs to be opened with Excel-like software. There are 8 columns in the file.

Column A: Names of input read files. Raw read files from Sanger sequences are merged into one file.

Columns B to G:

Targets: Names of amplicon references in design file

Sequence of target: Sequences of amplicon regions in design file

Read hit: Sequences mapped to “Sequence of target” with best score.

Read Number: Numbers of “Read hit”

Alignment(Target): Alignment result of Sequence of target

Alignment(Read): Alignment result of input read sequences

Indel Sites: Containing indel or not

SNP Sites: SNPs in the alignment if any.

One summary line is also generated for each amplicon “Type” of each input sequences file.

Total Reads: Number of reads in each file.  
 Total Hits: Total number of reads matching sequences in the design file  
 Sub Hits: Numbers of reads matching sequences of each amplicon “Type”  
 Indel Hits: Numbers of reads with indel comparing Ref\_Type

Only top 20 non-indel reads and top 50 indel reads, if any, are shown for each amplicon “Targets”.

Tip: change the font into “**Courier**“ for these two columns, Alignment (query) and Alignment (read), when saving the file as an Excel file.