AGEseq (Analysis of Genome Editing by Sequencing) User Guide v2

Liang-Jiao Xue and Chung-Jui Tsai, University of Georgia

January 8, 2021
(this version includes updated instructions for Mac users, provided by Chen Hsieh)

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

AGEseq is available at AspenDB (http://aspendb.org/downloads) and GitHub (https://github.com/liangjiaoxue/AGEseq).

Download the "AGEseq.zip" file and unzip it on your local system, the AGEseq directory is the **working directory** for the analysis.

1. Download necessary software

1) Perl

a) Windows:

Download and install ActivePerl (http://www.activestate.com/activeperl/downloads) .

If you have difficulty installing the program on Windows 8 by double clicking the downloaded file, try the following command line:

msiexec /i C:\dirctory\to\ActivePerl-XXX-MSWin32XXX.msi TARGETDIR="c:\perl" PERL_PATH="Yes"

Hint: search for "cmd" APP to open the command line console. Type or copy the command line to the console, making sure the full address to the downloaded file name ".msi" is correct.

For more information, see: http://docs.activestate.com/activeperl/5.20/install.html

b) Other systems:

Perl is functional by default.

2) BLAT

a) Windows:

Download blat_windows.zip from http://aspendb.uga.edu/downloads

Unzip the directory and copy the blat.exe and cygwin1.dll files into the AGEseq working directory.

b) Other systems:

Select an appropriate platform from: http://hgdownload.cse.ucsc.edu/admin/exe/

Scroll down the application list, select "blat/", and download "blat"

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

For newer macOS version, the following steps are also needed:

In Finder, right click the icon of blat, click 'open' and grant permission.

In Finder, right click the icon of blat, click 'Get Info'. In the 'Open with:' column, change the default app to 'Terminal'. The icon of blat should appear as executable (a black window with green 'exec' characters).

To run AGEseq with the test data provided with the package, go to **step 4** directly.

2. Prepare the "reads" folder(s)
The amplicon sequences need to be put into the "reads" folder, which can be found in the working directory. Two test files are provided, which need to be deleted or moved before running real jobs. Only a small number of reads is provided in the test data. A complete set of data from one transgenic line is available at http://aspendb.uga.edu/downloads.

The following file types are supported:

1) fastq (.fastq or .fq). The fq.gz files need to be unzipped first.
2) fasta (.fasta, .fas, .fa, .seq or .txt): either single or multiple sequences per file.

3. Prepare the "design file"
Open the text-delimited file named "targets" in the working directory using Excel or similar software.

The contents look like this:

target	Sequence
4CL1_1	CTAAGTCACCTGATCTTGACAAGCATGACTTGTCTTCTTTGAGGATGATAAAATCTGGAGGGGCTCCATTG
4CL1_2	CTAGGTCACCTGATCTTGACAAGCATGACTTGTCTTCTTTGAGGATGATAAAATCTGGAGGGGCTCCATTG
4CL5_1	CCAAGTCACCCGATCTTGATAAACATGACTTGTCTTCGTTGAGGATGTTGAAGTCTGGAGGGTCGCCATTG
4CL5_2	CCAAGTCACCTGATCTTGATAAACATGACTTGTCTTCGTTGAGGATGTTGAAGTCTGGAGGGTCGCCGTTG

Replace the target (gene/allele) names and sequences with your own. The sequence usually spans 30-40 bp at each end of the predicted editing site. It is not necessary to use the whole amplicon sequences if they are longer than the sequencing length. We routinely use 100-150 nt sequences with satisfactory results. If allele sequences are included, they must be of the same length (see examples above), or the BLAT engine may favor the longer allele for read assignment. Sequences of potential off-target sites can also be included, if your amplicon primers were designed to amplify homologous genes. Save the file.

4. Run AGEseq

1) Default settings:
Windows:
Double click the **Run_AGEseq.bat** file.

Mac:
Change the permission of "**run_AGEseq.command**" by typing "chmod 777 run_AGEseq.command" in the working directory. Then **Double click** the **Run_AGEseq.command** file.

Mac and other systems:
Change the permission of "**run_AGEseq.sh**" by typing "chmod 770 run_AGEseq.sh" in the working directory. Then run the program by typing:
./run_AGEseq.sh

2) Advanced options:
The command to run AGEseq is as follows:
cd /dir/to/AGEseq

perl	**AGEseq.pl**	**reads**	**targets.txt**	**AGE_output.txt**
		# 1 read directory	# 2 design file	# 3 output file

The first two inputs are required and the output file (the third augment) is optional. A text-delimited output file named "AGE_output.txt" will be generated by default if the output name is not provided. The filename extension ".txt" may be not shown on your system. The Perl script calls BLAT within the same directory to execute the analysis; please make sure BLAT is copied in the working directory or PATH environment. BLAT parameters can be customized to adjust the read matching sensitivity.

5. User-configurable options:

These parameters can be changed to customize the analysis stringency and reporting format.

Open the file "AGEseq.pl" with text editor like Notepad, and find the following lines (top 20 lines).

```
## setting for reports
my $mismatch_cutoff  = 0.1 ;  # mismatch rate to filter low quality alignment, default = 0.1 (10%)
my $min_cutoff       = 0   ;  # cutoff to filter reads with low abundance, default = 0
my $wt_like_report   = 20  ;  # report top xx WT like records, default = 20
my $indel_report     = 50  ;  # report top xx records with indels, default = 50
my $remove_files     =  1;  # keep (0) or delete (1) intermediate files, default = 1
```

Modify the numbers shown in red color to re-set the values as appropriate.

Reads with mismatches larger than $mismatch_cutoff are dropped before read counting.

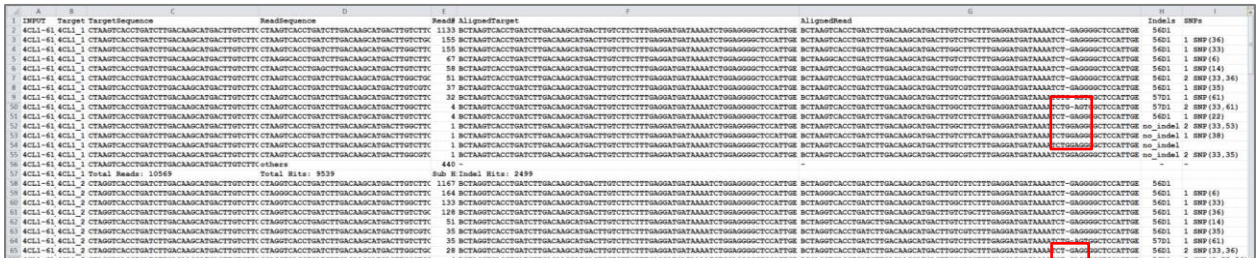
Reads with abundance less than $min_cutoff are dropped before read counting. $wt_like_report and $indel_report are used to set total number of cases shown in the output file. The "dropped" reads are still counted in the summary section.

Users can set $remove_files to 0 to keep intermediate (.psl) files, which can be mined for follow-up analysis for cases with unusual genome editing events.

6. Understanding the output file

AGEseq results are summarized in a text-delimited output file which can be opened in Excel. Select a monospaced font (such as Courier or Courier New) to display sequence data.

There are 9 columns (A through I) in the file (see the screenshot below).



A. INPUT: Name of the input read file. For Sanger data, individual read files are merged into one file and displayed with one name.
B. Target: Gene or allele name as provided in the design file.
C. TargetSequence: Target sequence as provided in the design file that has the best match to D.
D. ReadSequence: Amplicon sequence mapped to the target sequence in C.

E. Read#: Number of reads shown in D.
F. AlignedTarget: Target sequence displayed according to its alignment with the read sequence.
G. AlignedRead: Amplicon sequence displayed according to its alignment with the target sequence. Dashes are introduced to denote indels between the alignment pair (F and G).
H. Indels: Pattern of indels, if any, denoted by position (1st integer) of insertion (I) or deletion (D), followed by the number of indels (2nd integer). For example, 56D1 means 1-nt deletion at position 56. Unusual editing patterns with large insertions or deletions are flagged as "strange editing" and manual inspection for those events is necessary. The raw sequences can be extracted from the "reads" folder using "ReadSequence" for the given event from the AGEseq output (it may be necessary to use reverse complementary sequence). In some cases, additional bench experiments involving cloning and sequencing may be necessary to identify the exact editing patterns.
I. SNPs: Patterns of SNPs, if any, denoted by the nucleotide position(s).

When multiple targets and/or samples are included in the analysis, each group (e.g., one target from one transgenic line) is separated by read statistics of the group. Within each group, only the top 20 non-indel reads and the top 50 indel reads, if present, are reported by default.

A summary is provided for each analysis group following the detailed output. For each target gene/allele, two separate entries are provided for edited as well as WT-like patterns. The representative sequence, their detection frequency and relevant information are summarized in 9 columns.

	A	B	C	D	E	F	G	H	I
167	Sum:INPUT	Target	AlignedTarget	AlignedRead	Total Hits	Sub Hits	Indel or WT Hits	Indel or WT rate%	Pattern
168	Sum:4CL1-61	4CL1_1	BCTAAGTCACCTGATCTTGACAAGCATGACTTGTCTTCTTTGAGGATGATAAAATCTGGAGGGGCTCCATTGE	BCTAAGTCACCTGATCTTGACAAGCATGACTTGTCTTCTTTGAGGATGATAAAATCT-GAGGGGCTCCATTGE	9592	2531	2494	98.53	Indel:-1
169	Sum:4CL1-61	4CL1_1	BCTAAGTCACCTGATCTTGACAAGCATGACTTGTCTTCTTTGAGGATGATAAAATCTGGAGGGGCTCCATTGE	BCTAAGTCACCTGATCTTGACAAGCATGACTTGTCTTCGATTGAGGATGATAAAATCTGGAGGGGCTCCATTGE	9592	2531	37	1.47	WT_like
170	Sum:4CL1-61	4CL1_2	BCTAGGTCACCTGATCTTGACAAGCATGACTTGTCTTCTTTGAGGATGATAAAATCTGGAGGGGCTCCATTGE	BCTAGGTCACCTGATCTTGACAAGCATGACTTGTCTTCTTTGAGGATGATAAAATCT-GAGGGGCTCCATTGE	9592	2580	2561	99.26	Indel:-1
171	Sum:4CL1-61	4CL1_2	BCTAGGTCACCTGATCTTGACAAGCATGACTTGTCTTCTTTGAGGATGATAAAATCTGGAGGGGCTCCATTGE	BCTAGGTCACCTGATCTTGACAAGCATGACTTGTCTTCTTTGAGGATGATAAAATCTGGAGGGGCTCCATTGE	9592	2580	19	0.74	WT_like
172	Sum:4CL1-61	4CL5_1	BCCAAGTCACCCGATCTTGATAAACATGACTTGTCTTCGTTGAGGATGTTGAAGTCTGGAGGGTCGCCATTGE	BCCAAGTCACCCGATCTTGATAAACATGACTT-TCTTCGTTGAGGATGTTGAAGTCTGGAGGGTCGCCATTGE	9592	2102	2	0.09	Indel:-1
173	Sum:4CL1-61	4CL5_1	BCCAAGTCACCCGATCTTGATAAACATGACTTGTCTTCGTTGAGGATGTTGAAGTCTGGAGGGTCGCCATTGE	BCCAAGTCACCCGATCTTGATAAACATGACTTGTCTTCGTTGAGGATGTTGAAGTCTGGAGGGTCGCCATTGE	9592	2102	2100	99.91	WT_like
174	Sum:4CL1-61	4CL5_2	BCCAAGTCACCTGATCTTGATAAACATGACTTGTCTTCGTTGAGGATGTTGAAGTCTGGAGGGTCGCCGTTGE	BCCAAGTCACCTGATCTTGAT-AACATGACTTGTCTTCGTTGAGGATGGTGAAGTCTGGAGGGGCGCCGTTGE	9592	2379	1	0.04	Indel:-1
175	Sum:4CL1-61	4CL5_2	BCCAAGTCACCTGATCTTGATAAACATGACTTGTCTTCGTTGAGGATGTTGAAGTCTGGAGGGTCGCCGTTGE	BCCAAGTCACCTGATCTTGATAAACATGACTTGTCTTCGTTGAGGATGTTGAAGTCTGGAGGGTCGCCGTTGE	9592	2379	2378	99.96	WT_like

A. "Sum:INPUT": Name of the input read file with a "Sum:" prefix. Users can use the "sort" function in excel to group the summary section of individual files/samples together.
B. Target: Gene or allele name as provided in the design file.
C. AlignedTarget: Target sequence displayed according to its alignment with the read sequence.
D. AlignedRead: Edited sequence displayed according to its alignment with the target sequence. By default, sequence from the predominant editing event is shown. Dashes are introduced to denote indels between the alignment pair (C and D).
E. Total hits: Total number of reads matching all target sequences in the file.
F. Sub hits: Number of reads matching the specific target/allele sequence.
G. Indel or WT hits: Number of reads with indels or WT-like sequence.
H. Indel or WT rate%: Fraction of indel or WT hits (G) over the gene/allele-specific sub-hits (F).
I. Pattern: Patterns of indels, if any, denoted by + (insertion) or – (deletion) and the number of nucleotides affected. Unusual editing patterns are noted as "strange editing" as described above.

Known limitations and suggestions:

1) Due to sequence alignment artifacts with BLAT or other similar programs concerning inconsistent handling of gaps in the presence of homonucleotides, it is highly recommended that the user manually inspects the AGEseq output and makes necessary adjustments. The first screenshot on the previous page shows an example of inconsistent indel calls due to alignment artifacts (56D1 and 57D1 for both 4CL1_1 and 4CL1_2).

2) Sequence errors can be introduced during amplicon library preparation and sequencing that involve PCR, or by base-calling algorithms. These errors will appear as SNPs/indels in the AGEseq report. We suggest only indel reads with a coverage of 3-5 be considered further in the analysis. By default, AGEseq computes the "indel %" or "WT-like %" using the total number of indel (or WT-like) reads regardless of the indel/sequence pattern. In the example above for 4CL1_1, we detected 2531 hits matching this allele, 2494 with a 1-nt deletion and 37 with no indels. Of the 2494 indel hits, the predominant event (1133 hits) has perfect sequence match with the target sequence, except for the 1-nt deletion. The remaining indel events contained distinct SNPs ranging from 1 to 3 nt, likely derived from sequence errors. User inspection and adjustment is therefore highly recommended. Alternatively, sequence matching stringency can be customized, or only reads without SNPs be considered.

3) Unusual genome editing events may not be detected using the default settings, since sequence alignment quality is usually lower for those cases. To improve their detection, users may relax the mismatch cutoff ($mismatch_cutoff= 0.5) in conjunction with a higher read coverage ($min_cutoff= 10). Using a longer target sequence may improve detection of events with large deletions.

4) For reporting purpose, the data can be condensed to include only WT_like (no editing) reads for control samples, and indel-containing reads for edited samples (retaining one row per sample per allele). See the table below as an example.

Plant	Allele	Sequence	Read#	Indel/WT#	Indel/WT%	Pattern
WT	4CL1a	TTGAGGATGATAAAATCTGGAGGGGCTCCATTGGGCAAGGAACTTGAAGATACTGTCAGAG	6387	6289	98.5	WT_like
	4CL1t	TTGAGGATGATAAAATCTGGAGGGGCTCCATTGGGCAAGGAACTTGAAGAAACTGTCAGAG	6325	6265	99.1	WT_like
	4CL5a	CGTTGAGGATGTTGAAGTCTGGAGGGTCGCCATTGGGGAAGGAGCTTGAAGATACTGTCAGAG	4745	4682	98.7	WT_like
	4CL5t	CGTTGAGGATGTTGAAGTCTGGAGGGTCGCCGTTGGGGAAGGAGCTTGAAGATACTGTCAGAG	6591	6507	98.7	WT_like
Cas9	4CL1a	TTGAGGATGATAAAATCTGGAGGGGCTCCATTGGGCAAGGAACTTGAAGATACTGTCAGAG	6314	6247	98.9	WT_like
	4CL1t	TTGAGGATGATAAAATCTGGAGGGGCTCCATTGGGCAAGGAACTTGAAGAAACTGTCAGAG	6406	6337	98.9	WT_like
	4CL5a	CGTTGAGGATGTTGAAGTCTGGAGGGTCGCCATTGGGGAAGGAGCTTGAAGATACTGTCAGAG	6039	5952	98.6	WT_like
	4CL5t	CGTTGAGGATGTTGAAGTCTGGAGGGTCGCCGTTGGGGAAGGAGCTTGAAGATACTGTCAGAG	8964	8879	99.1	WT_like
1-40	4CL1a	TCTTTGAGGATGATAAAATCT-GAGGGGCTCCATTGGGCAAGGAACTTGAAGATACTGTCAGAG	9583	9578	99.9	Indel:-1
	4CL1t	TCTTTGAGGATGATAAAATCT-GAGGGGCTCCATTGGGCAAGGAACTTGAAGAAACTGTCAGAG	9623	9620	100.0	Indel:-1
	4CL5a	CGTTGAGGATGTTGAAGTCTGGAGGGTCGCCATTGGGGAAGGAGCTTGAAGATACTGTCAGAG	9217	9082	98.5	WT_like
	4CL5t	CGTTGAGGATGTTGAAGTCTGGAGGGTCGCCGTTGGGGAAGGAGCTTGAAGATACTGTCAGAG	12217	11970	98.0	WT_like