Directions for Python program:

NOTE: This program will ONLY analyze deletion repair junctions. If there are SNPs or InDels, you must use the R script.

To run the python script, unzip the folder and proceed to **Python26->Scripts->SD-MMEJ** and double click on “data\_ui\_newest\_11\_03\_2014.py”

There will be three options in the menu. If you select 1, then you will see directions for how to prepare your sequences for the program. I have pasted them here.

“Creating A Sequence Data File:

The file should have one sequence per line. Do not leave blank lines.

The end-joining repair junctions should be shown as if they were

aligned with the original sequence. Use dashes (-) to indicate

deleted bases. example: ATGATG---GCGCTATGCG

VERY IMPORTANT: For the analysis to work correctly, all of the junction

sequences in the data file MUST comprise EXACTLY the same sequence

interval as the sequence that you enter when prompted for the original

sequence. In other words,

\* All junction sequences MUST be the same length as the original sequence

AND

\* There must be an exact one to one correspondence of the remaining

non-deleted bases in the junction sequences with bases in the original

sequence.

example: if you provide the original sequence AAAATTTTGGGGCCCC,

all of the deletions in your data file must be able to

align exactly with that sequence (but do NOT include the

original sequence in the data file):

AAAATTTTGGGGCCCC (original)

AAAAT----GGGCCCC (OK)

AAAATTTTGG-GCCCC (OK)

CAAAT--TGGGGCCCC (NOT OK - mismatch with the original sequence.)

CCAAAAT---GGGGCCCC (NOT OK - extra bases in the junction sequence)

AATTTTG--CCCCC (NOT OK - too few bases in the junction sequence)

ALSO IMPORTANT: To be able to compare the results generated by

this program with results generated by the program that analyzes

the original sequence, the original sequence that you provide to this

program must be the exact same as the original sequence you provide to

the other program.”

To run the program, select option 2.

1. You will be prompted to input the DNA sequence to the left of the left nick (you can copy and paste the sequence)
2. You will then be prompted to input the DNA sequence between the nicks (leave blank if blunt DSB).
3. Input the DNA sequence to the right of the right nick
4. Then you will be asked to input the number of bases to the left and right of the repair junction to search for SD-MMEJ consistent repeats. We always use 30.
5. Then you must insert the minimum size of an SD-MMEJ consistent repeat. We use 4.
6. You will then be prompted to enter the data file. I always just drag the fil into the window. It will automatically input the path and file name
7. Then you must choose a name for the output file, and for the output spreadsheet with a breakdown of all of the parameters which may be important for SD-MMEJ
8. The program will run, but it will not save the output files until you quit the python program.

The output files can be opened in any text editor. I usually open the spreadsheet in excel.

Directions for R script:

NOTE: This script has not been thoroughly vetted for all types of cuts or junctions.

This script analyzes insertions only. It probably could be modified to analyze deletions as well but is unnecessary because we use the Python program for that.

Instructions for running the script are written in the script.

The script can only analyze repair junctions from one original (reference) DNA sequence at once. If you have junctions from multiple different DNA sequences, they must be separated based on their original cut site and analyzed separately.

If there are SNPs surrounding the repair junctions, the script will view it as the start of an indel and attempt to determine if it is SD-MMEJ consistent. The script will likely determine that it is not, and continue on. Therefore, if you have common SNPs surrounding the repair junctions, it is likely best to analyze them separately.

In the script you must provide the DNA sequence Left and Right of the break, written 5’->3’ for both. If there is an overhang, include it in both.

*Formatting sequences for the script*: The sequences should not be aligned, however they all must start at the same nucleotide, and end at the same nucleotide (see mut\_m5\_ins\_for\_R\_script.csv). The sequences should be saved as .csv.

This is a very crude script that requires manual curation of the output to ensure it’s fidelity. If certain junctions are not SD-MMEJ consistent, you can make changes to the output file, reload it, and rerun it for the correct outcome. There is a little bit of script at the end to help you through this.