SOLiD Filtering Framework

This script is written and Perl and was designed to overcome memory constraints imposed by uploading all of the reads and their corresponding quality values (QVs) into memory. The program sacrificed runtime in order to have a small memory footprint. Uploading the data into memory for analysis will only prove more difficult as the quantities of reads sequenced grows with the new generations of the SOLiD platform. Below is a list of switches and how they are used to run the filter.

```
[-f csfasta file for the forward strand
      [-g quality file for the forward strand
       -i input type
       -r csfasta file for the reverse strand
       -s quality file for the reverse strand
       -p Mean (Avg QV for the read) anlysis on/off
       -q Mean minimum QV score
       -n Removal of all reads containing a missing colorcall on/off
       -t truncation on/off
       -u truncation length
       -a quality analysis on/off]
 -o output file name
      [-v Output matching QV files on/off]
]
```

Here is an example
$ ./ SOLiD_preprocess_filter_v1.pl -i mp -f a_F3.csfasta -g a_F3_QV.qual -r b_R3.csfasta -s b_R3_QV.qual -p y -q 20 –n y –t y –u 25 –a y -o filter_ouput –v n

Input Name	Required Options and Defaults	Description
-i or –input_type	No – default is set for a fragment run	The type of data - identifies mate pair or fragment analysis. If the -i option is left blank or excluded a Fragment analysis will be preformed, and if -r and -s options are used in such a case, they will be ignored. For mate pair analysis, the mate pair option must be selected. [F, Frag, Fragment, f, frag, fragment, mp, MP, mate-pair, Mate-pair]
-f or –f3	Yes	csfasta file for analysis - this the name of the file coming from the SOLiD primary analysis. If this is analysis for mate pairs the F3 file set must be passed in here and the mates under the –r option.
-g or –f3QV	No	Corresponding quality file for the –f option. If this is left blank, the name of the corresponding QV file must match and be in the same location as the csfasta file except the ending will replace the .csfasta with _QV.qual.
-r or –r3	No	fasta file for mate pair analysis - this is the name of the mates' csfasta file to the f3 option. If mate pair analysis is turned on then this field becomes required.
-s or –r3QV	No	Corresponding quality file for the –r option. If this is left blank, the name of the corresponding QV file must match and be in the same location as the csfasta file except the ending will replace the .csfasta with _QV.qual.
-p or –mean_analysis	No – default is on	Average anlysis on/off - Average analysis filter looks at the average of the all the color calls. It asks that the average of all the calls must exceed the qv score of interest (p|mean_qv). [on,yes,y,off,no,n]
-q or –mean_qv	No – default is a quality score of 20	Average analysis min QV score - This is the minimum score for the average analysis. This is currently limited to a number between 0-34.
-n or –neg_qv	No – default is off	Removal of all reads containing a missed color call on/off – If removal of all reads containing missing color calls (identified by negative quality scores), this flag must be turned on. [on,yes,y,off,no,n]
-t or –trunk	No – default is off	truncation of reads on/off - If truncation of reads is desired, this flag must be turned on. If turned on, option u must be used as well. [on,yes,y,off,no,n]
-u or –tr_len	No	length of desired read after truncation – This is the length of the sequence desired, any color calls after this length are removed. This option must be filled in if truncation is turned on and be an integer greater than 0.
-a or –qv_analysis	No	Analysis of the quality values for all of the inputted reads and the passing reads. Analysis returns a file with a matrix of a count of scores by position.
-o or –output	Yes	output file name - this is the beginning of the name for the output information. The endings are filled in as needed.
-v or –ouput_qv	No – default is on	Output matching QV files on/off – this will print the matching QV files to the outputted csfasta files. [on,off]