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
       -x Polyclonal anlysis on/off
       -p Polyclonal analysis count required
       -q Polyclonal analysis min QV score
       -y error analysis on/off
       -e error analysis max error count allowed
       -d error analysis max error 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 –x y -p 3 -q 25 -y y -e 2 -d 10 –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.
-x or –poly_analysis	No – default is on	Polyclonal analysis on/off - Polyclonal analysis looks only at the first 10 color calls. It asks that within those first 10 calls there must be a certain number (p|p_cnt) of qv scores that exceed the qv score of interest (q|p_qv). [on,yes,y,off,no,n]
-p or – p_cnt	No – default is 1	Polyclonal analysis count required - This is the count required for the polyclonal analysis. This number must be between [0-10]. Zero is equivalent to having the polyclonal analysis turned off.
-q or –p_qv	No – default is a quality score of 25	Polyclonal analysis minimum QV score - This is the minimum score for the polyclonal analysis. This must be a number between 0-50. Please be aware that scores above 34 are exceedingly rare.
-y or –error_analysis	No – default is on	error analysis on/off - Error analysis looks at the entire read for quality scores that fall below the error score passed (e_sc). These calls are counted, and the total number of these erroneous calls must be under the err_cnt passed. [on,yes,y,off,no,n]
-e or –e_cnt	No – default is 3	error analysis maximum error count allowed - This is the maximum number of errors allowed per read. If the number is greater then the read length it is equivalent to having the error analysis turned off.
-d or –e_sc	No – default is a quality score of 10	error analysis maximum QV score - This is the maximum score for error analysis. This must be a number between 0-50. Please be aware that scores above 34 are exceedingly rare.
-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]