NGS QC Toolkit (v2.3) Manual

TABLE OF CONTENTS

1. Essential requirements

- **Operating system**:
  - Windows (PC)
  - Linux
- **Software**:
  - Perl (ActivePerl for Windows)
- **Additional Perl modules required**:
  - GD::Graph (optional, used to prepare graphs)
  - String::Approx (required to speed up the string matching for primer/adapter)

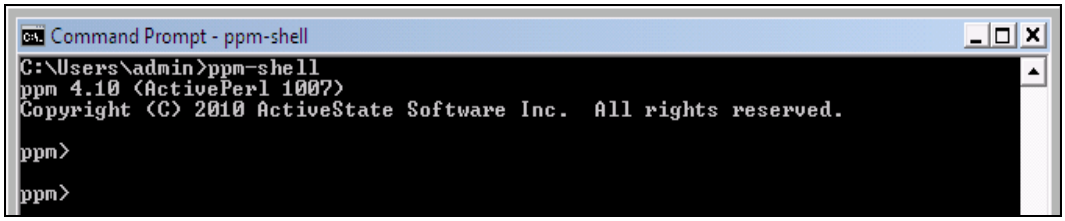
2. How to install additional perl modules

On Windows XP/Vista/7:

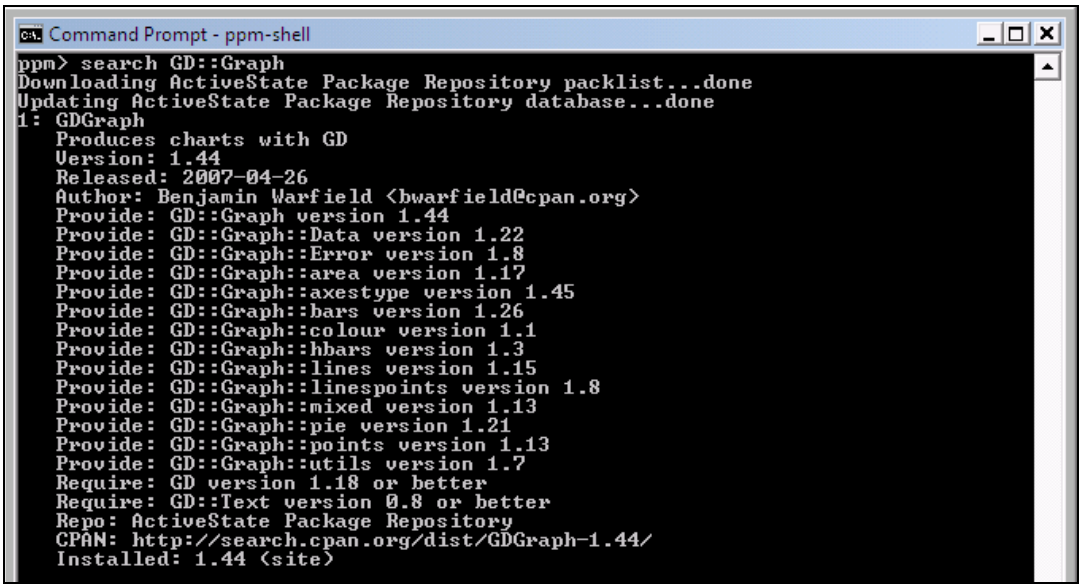
Note: Activeperl (http://www.activestate.com/activeperl/downloads) has to be installed on windows.

- Install perl modules using DOS command prompt. Following are the steps for installing GD::Graph

  (1) Open DOS command prompt and type "ppm-shell" and press enter. The "ppm>" prompt will come.



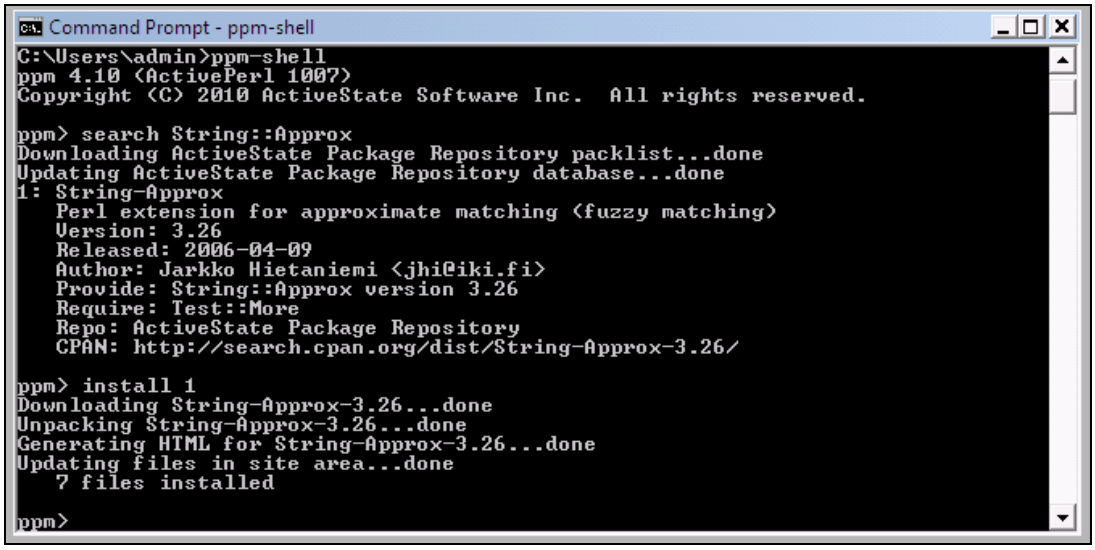
  (2) On "ppm" command prompt, the modules have to be searched using "search <module name>". This will show all available modules for the given keyword with index number on left (In this case a single module is available for GD::Graph).



(3) Install the module using “install <index number>” command.

(4) And the module is installed.

- Following is the screenshot of String::Approx installation



On Linux:

Note: Check that perl has been installed on your system.

Download modules from http://search.cpan.org

Following are the web links for required modules:

(1) GD::Graph
(http://search.cpan.org/~bwarfield/GDGraph-1.44/Graph.pm)
This module requires two dependencies:
- GD
(http://search.cpan.org/~lds/GD-2.45/GD.pm)
- GD::Text
(http://search.cpan.org/~mverb/GDTextUtil-0.86/Text.pm)

(2) String::Approx
(http://search.cpan.org/~jhi/String-Approx-3.26/Approx.pm)

Install these modules according to the "README" information provided with the module.

3. How to use NGS QC Toolkit

- Download and install required software and perl modules
- Download source code from the home page (http://www.nipgr.res.in/ngsqctoolkit.html)
- Extract the compressed file
- Tools are available in the extracted folder
- Run these perl script using command "perl <tool name>"

4. Sample data

Sample input and output data is provided to download from the homepage of the NGS QC toolkit. Sample input data includes Illumina paired-end (~0.13 million reads), Illumina single-end (~0.13 million reads), 454 paired-end (~0.12 million reads) and 454 single-end (more than 90 thousand reads) sequencing data. Sample output presents output of QC, trimming and statistics tools for above mentioned Illumina and 454 sample data. QC output contains high-quality filtered data, text file and graphs for QC statistics and consolidated HTML report file.

5. Tools in toolkit

- **QC**
  - IlluQC.pl: Tool for quality control of sequencing data generated using Illumina platform (FASTQ format)
  - IlluQC_PRLL.pl: This tool has the same functionality as IlluQC.pl. However, it provides an additional option to use multiple CPUs to speed up the analysis
  - 454QC.pl: Tool for quality control of sequencing data generated using 454 platform (read and quality in FASTA format)
  - 454QC_PRLL.pl: Tool performs same quality control analysis as 454QC.pl and helps to analyze data using multiple CPUs
  - 454QC_PE.pl: Tool for quality control of paired-end sequencing data generated using 454 platform (read and quality in FASTA format)
- **Format-converter**
  - SangerFastqToIlluFastq.pl: To convert fastq-sanger variant to fastq-illumina variant of FASTQ format
  - SolexaFastqToIlluFastq.pl: To convert fastq-solexa variant to fastq-illumina variant of FASTQ format
  - FastqTo454.pl: To convert FASTQ format (any variant) to 454 format (two files in FASTA format: one for reads/sequences (.fna) and another for quality (.qual))
  - FastqToFasta.pl: To convert FASTQ format file to FASTA format file for reads/sequences
- **Trimming**
  - TrimmingReads.pl: Tool for trimming reads from 5' and/or 3' end of the read (FASTQ or FASTA format)
  - HomoPolymerTrimming.pl: Tool for trimming 3' end of the reads from the first base of homopolymer of given length
  - AmbiguityFiltering.pl: Tool for filtering reads containing ambiguous bases or trimming flanking ambiguous bases
- **Statistics**
  - AvgQuality.pl: Tool to calculate average quality score for each read and overall quality score for the given FASTA quality file
  - N50Stat.pl: Tool to generate statistics for read/sequence data given in FASTA format

6. Detailed help information

Following is the detailed help for each tool provided in the NGS QC Toolkit.

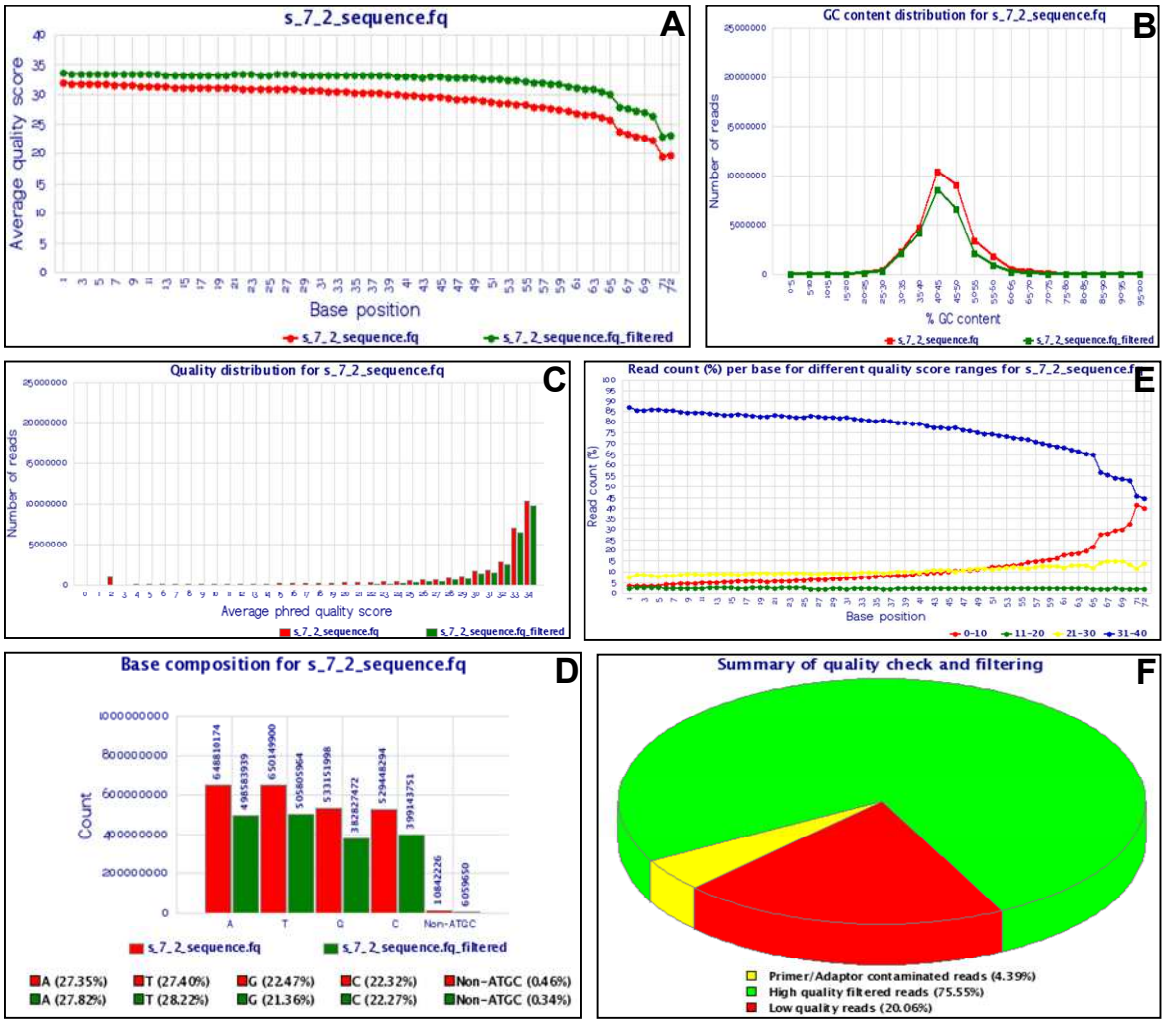
(A) QC Tools

(1) IlluQC.pl: This tool performs quality check and filtering of the sequencing data generated using Illumina technology. Input to this tool is FASTQ files (any variant) containing read and corresponding quality scores. Following are the options available with IlluQC.pl.

```
Usage: perl IlluQC.pl <options>

IlluQC.pl options:

### Input reads (FASTQ) options (Atleast one option is required)
  -pe <Forward reads file> <Reverse reads file> <Primer/Adaptor
library> <FASTQ variant>
    Paired-end read files (FASTQ) with primer/adaptor library and
FASTQ variant
    User may choose from the provided primer/adaptor library or can
give a file containing primer/adaptor sequences, one per line
    Multiple libraries can be given using multiple '-pe' options
      For eg.: -pe r1.fq r2.fq 3 1 -pe t1.fq t2.fq 2 A

  -se <Reads file> <Primer/Adaptor library> <FASTQ variant>
    Single-end read file (FASTQ) with primer/adaptor library and
FASTQ variant
    Multiple libraries can be given using multiple '-se' options
      For eg.: -se r1.fq 3 2 -se t2.fq 2 2

    Primer/Adaptor libraries:
      1 = Genomic DNA/Chip-seq Library
      2 = Paired End DNA Library
      3 = DpnII gene expression Library
      4 = NlaIII gene expression Library
      5 = Small RNA Library
      6 = Multiplexing DNA Library
      N = Do not filter for Primer/Adaptor
      <File> = File for user defined primer/adaptor sequences, one
per line

    FASTQ variants:
      1 = Sanger (Phred+33, 33 to 73)
      2 = Solexa (Phred+64, 59 to 104)
      3 = Illumina (1.3+) (Phred+64, 64 to 104)
      4 = Illumina (1.5+) (Phred+64, 66 to 104)
      5 = Illumina (1.8+) (Phred+33, 33 to 74)
      A = Automatic detection of FASTQ variant

### Other options [Optional]
  -h | -help
    Prints this help
---------------------------- QC Options -----------------------------
```

```
  -l | -cutOffReadLen4HQ <Real number, 0 to 100>
    The cut-off value for percentage of read length that should be
of given quality
    default: 70
  -s | -cutOffQualScore <Integer, 0 to 40>
    The cut-off value for PHRED quality score for high-quality
filtering
    default: 20
------------------------ Processing Options ------------------------
  -p | -processes <Integer>
    Number of processes to be used
    default: 1
  -onlyStat
    Outputs only statistics without filtered data output
-------------------------- Output Options --------------------------
  -t | -statOutFmt <Integer>
    Output format for statistics
    Formats:
      1 = formatted text
      2 = tab delimited
    default: 1
  -o | -outputFolder <Output folder name/path>
    Output will be stored in the given folder
    default: By default, output folder (IlluQC_Filtered_files) will
be generated where the input files are
  -z | -outputDataCompression <Character>
    Output format for HQ filtered data
    Formats:
      t = text FASTQ files
      g = gzip compressed files
    default: t
```

Output: IlluQC.pl generates statistics for quality check and filtering steps along with quality of input and high-quality filtered data in the form of text files and graphs. Following are the sample output graphs showing the average quality score at each base position (A), GC content distribution (B), average quality distribution (C) and base composition (D) for input and filtered reads. (E) shows the percentage of reads for different quality score ranges at each base position. The pie chart shows the summary of quality control analysis (F).

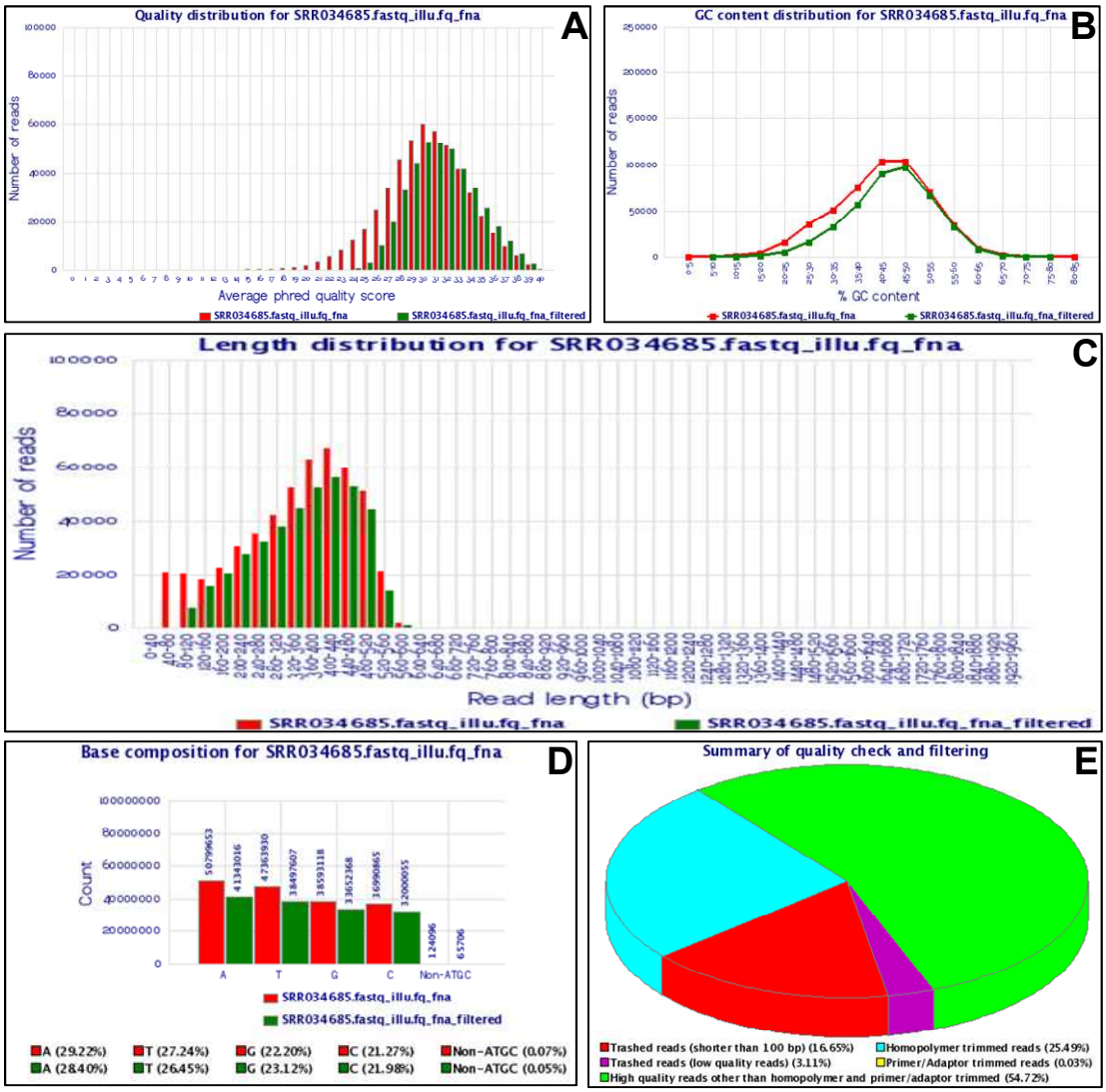


(2) 454QC.pl: This tool requires two files as an input: 1) .fna file containing reads/sequences in FASTA format and 2) .qual file containing quality score in FASTA format. On the basis of quality provided in the second file the quality check is performed and reads are filtered. Following is the detailed help for 454QC.pl.

```
Usage: perl 454QC.pl <options>

454QC.pl options:

### Input reads (FASTA format; .fna and .qual files) (Required)
  -i <Read file> <Quality file> <Primer/Adaptor library>
    Read and quality file in FASTA format with primer/adaptor
library
    User may choose from the provided primer/adaptor library or can
give a file containing primer/adaptor sequences, one per line
    Multiple libraries can be given using multiple '-i' options
      For eg.: -i read1.fna read1.qual 3 -i read2.fna read2.qual 2

    Primer/Adaptor libraries:
      1 = Rapid Library (Standard)
```

```
      2 = Paired End Library
      3 = Amplicon PE Library
      4 = Small RNA Library
      N = Do not filter for Primer/Adaptor
      <File> = File for user defined primer/adaptor sequences, one
per line

### Other options [Optional]
  -h | -help
    Prints this help
---------------------------- QC Options -----------------------------
  -l | -cutOffReadLen4HQ <Real number, 0 to 100>
    The cut-off value for percentage of read length that should be
of given quality
    default: 70
  -s | -cutOffQualScore <Integer, 0 to 40>
    The cut-off value for PHRED quality score for high-quality
filtering
    default: 20
  -n | -homoPolyLen <Integer>
    Minimum length of the homopolymer to be trimmed (0: to skip the
homopolymer trimming)
      For eg.: -n 8, will trim the right end of read from the
homopolymer of at least 8 bases long
    default: 0 (homopolymer trimming is off)
  -m | -minLen <Integer>
    Filter sequences shorter than the given minimum length
    default: 100
  -f | -lenFilter <Y/N>
    Are sequences to be filtered on the basis of length: (Y)es or
(N)o
    default: Y
----------------------- Processing Options -----------------------
  -p | -processes <Integer>
    Number of processes to be used
    default: 1
  -onlyStat
    Outputs only statistics without filtered data output
-------------------------- Output Options --------------------------
  -t | -statOutFmt <Integer>
    Output format for statistics
    Formats:
      1 = formatted text
      2 = tab delimited
    default: 1
  -o | -outputFolder <Output folder name/path>
    Output will be stored in the given folder
    default: By default, output folder (454QC_Filtered_files) will
be generated where the input files are
  -z | -outputDataCompression <Character>
    Output format for HQ filtered data
    Formats:
      t = text FASTA files
      g = gzip compressed files
    default: t
```

Output: 454QC.pl generates statistics after each step of analysis (number and percentage of trimmed, trashed and high-quality reads) and statistics for both input and filtered data (minimum, maximum, mean, N25, N50, N75, N90 and N95 read length) in the form of text files and graphs including average quality distribution (A), GC content distribution (B) and average length distribution (C) and base composition (D) for input and filtered data. The pie charts summarize the quality control analysis (E).



(3) IlluQC_PRLL.pl: This tool provides a utility to process input data parallely on multiple CPUs (using '-c' option) to speed-up the quality control analysis. Otherwise, it is identical to the IlluQC.pl in context of algorithm for processing Illumina data.

(4) 454QC_PRLL.pl: This tool can use multiple CPUs (using '-c' option) and process the large amount of data very fast. It performs same function as 454QC.pl for quality control of 454 sequencing data.

(5) 454QC_PE.pl: This tool performs quality control of paired-end sequencing data generated using 454 platform. It identifies the linker sequence to separate the PE reads as first step. The following steps of QC on these PE reads and unpaired reads (where linker sequence could not be identified) are same as other 454QC tools.

(B) Format-converter Tools

(1) SangerFastqToIlluFastq.pl: A tool for the conversion of FASTQ file containing quality score encoded in fastq-sanger format (score: 33 to 73) to the FASTQ file having quality score encoded in fastq-illumina format (score: 64 to 104). Following are the options provided with this tool.

```
Usage: perl SangerFastqToIlluFastq.pl <options>

SangerFastqToIlluFastq.pl options:

### Input reads (FASTQ) (Required)
  -i <Sanger FASTQ read file>
    Read file in Sanger FASTQ format

### Other options [Optional]
  -h | -help
    Prints this help
  -o | -outputFile <Output file name>
    Output will be stored in the given file
    default: By default, file will be stored where the input file is
```

(2) SolexaFastqToIlluFastq.pl: A tool for the conversion of FASTQ file containing quality score encoded in fastq-solexa format (score: 59 to 73) to the FASTQ file having quality score encoded in fastq-illumina format (score: 64 to 104). Following are the options provided with this tool.

```
Usage: perl SolexaFastqToIlluFastq.pl <options>

SolexaFastqToIlluFastq.pl options:

### Input reads (FASTQ) (Required)
  -i <Solexa FASTQ read file>
    Read file in Solexa FASTQ format

### Other options [Optional]
  -h | -help
    Prints this help
  -o | -outputFile <Output file name>
    Output will be stored in the given file
```

```
    default: By default, file will be stored where the input file is
```

(3) FastqTo454.pl: This tool converts FASTQ format file (any variant) to 454 format, i.e. separates reads/sequences and quality in different FASTA files (.fna and .qual). Options provided with the tool:

```
Usage: perl FastqTo454.pl <options>

FastqTo454.pl options:

### Input reads (FASTQ) (Required)
  -i <Illumina FASTQ read file>
    Read file in Illumina FASTQ format

### Other options [Optional]
  -h | -help
    Prints this help
  -o | -outputFolder <Output folder name>
    Output will be stored in the given folder
    default: By default, files will be stored where the input file
is
  -v | -fastqVariant <FASTQ variant>
    FASTQ variants:
      1 = Sanger (Phred+33, 33 to 73)
      2 = Solexa (Phred+64, 59 to 104)
      3 = Illumina (1.3+) (Phred+64, 64 to 104)
      4 = Illumina (1.5+) (Phred+64, 66 to 104)
      5 = Illumina (1.8+) (Phred+33, 33 to 74)
      A = Automatic detection of FASTQ variant
    default: "A"
```

(4) FastqToFasta.pl: It exports reads/sequences from the FASTQ file to the FASTA format file. Following are the options available with the tool.

```
Usage: perl FastqToFasta.pl <options>

FastqToFasta.pl options:

### Input reads (FASTQ) (Required)
  -i <FASTQ read file>
    Read file in FASTQ format

### Other options [Optional]
  -h | -help
    Prints this help
  -o | -outputFile <Output file name>
    Output will be stored in the given file
    default: By default, file will be stored where the input file is
```

(C) Trimming Tools

(1) TrimmingReads.pl: This tool trims the reads/sequences and their quality scores (in case of FASTQ file) in two ways. First, it trims fixed (user-specified) number of bases from 5' and/or 3' end of the reads and corresponding qualities from the input FASTQ file. Second, it trims low quality bases from 3' end of the read using user-defined threshold value of quality score. Input to this tool is either FASTQ or FASTA format file. Options are provided to specify the number of bases to be trimmed and the quality threshold for quality based trimming.

```
Usage: perl ..\TrimmingReads.pl <options>

..\TrimmingReads.pl options:

### Input reads/sequences (FASTQ/FASTA) (Required)
  -i <Forward read/sequence file>
    File containing reads/sequences in either FASTQ or FASTA format

### Input reads/sequences (FASTQ) [Optional]
  -irev <Reverse read/sequence file of paired-end data>
    File containing reverse reads/sequences of paired-end data in
FASTQ format

### Other options [Optional]
  -h | -help
    Prints this help
--------------------------------- Trimming Options -----------------
----------------
  -l | -leftTrimBases <Integer>
    Number of bases to be trimmed from left end (5' end)
    default: 0
  -r | -rightTrimBases <Integer>
    Number of bases to be trimmed from right end (3' end)
    default: 0
  -q | -qualCutOff <Integer> (Only for FASTQ files)
    Cut-off PHRED quality score for trimming reads from right end
(3' end)
      For eg.: -q 20, will trim bases having PHRED quality score
less than 20 at 3' end of the read
      Note: Quality trimming can be performed only if -l and -r are
not used
    default: 0 (i.e. quality trimming is OFF)
  -n | -lenCutOff <Integer>
    Read length cut-off
    Reads shorter than given length will be discarded
    default: -1 (i.e. length filtering is OFF)
--------------------------------- Output Options -------------------
--------------
  -o | -outputFile <Output file name>
    Output will be stored in the given file
    default: By default, output file will be stored where the input
file is
```

(2) HomoPolymerTrimming.pl: The tool finds homopolymer of the given length and trims the 3' end of reads/sequences from the first base of the homopolymer from the data in FASTA format.

```
Usage: perl HomopolymerTrimming.pl <options>

HomopolymerTrimming.pl options:

### Input reads/sequences (FASTA format; .fna and .qual files)
(Required)
  -i <Read/Sequence file> [Quality file (optional)]
    Read/Sequence and quality file in FASTA format

### Other options [Optional]
  -h | -help
    Prints this help
  -l | -minReadLen <Integer>
    Minimum length of a read/sequence to be retained in output
    default: 100
  -n | -homoPolyLen <Integer>
    Minimum length of the homopolymer to be trimmed
      For eg.: -n 8, will trim the right end of read/sequence from
the homopolymer of at least 8 bases long
      Note:- use -n 0 to skip homopolymer trimming (for only length
filtering)
    default: 8
  -o | -outputFolder <Output folder name/path>
    Output will be stored in the given folder
    default: By default, files will be stored where the input files
are
```

(3) AmbiguityFiltering.pl: The tool helps filtering ambiguous base content in two ways: 1) Trimming 5' and/or 3' ambiguous bases, and 2) Filtering reads based on user defined cut-off values for maximum number/percentage of allowed ambiguous bases.

```
Usage: perl ..\AmbiguityFiltering.pl <options>

..\AmbiguityFiltering.pl options:

### Input reads/sequences (FASTQ/FASTA) (Required)
  -i <Forward read/sequence file>
    File containing reads/sequences in either FASTQ or FASTA format

### Input reads/sequences (FASTQ) [Optional]
  -irev <Reverse read/sequence file of paired-end data>
    File containing reverse reads/sequences of paired-end data in
FASTQ format

### Other options [Optional]
  -h | -help
    Prints this help
--------------------------------- Trimming Options -----------------
---------------
  -c | -countN <Integer>
```

```
    Maximum number of allowed ambiguous bases
    default: 0
  -p | -percentN <Integer>
    Maximum percentage of allowed ambiguous bases
    default: 0
  -t5 | -trim5EndN
    Trim ambiguous bases from 5' end of the sequence
    default: off
  -t3 | -trim3EndN
    Trim ambiguous bases from 3' end of the sequence
    default: off
  -n | -lenCutOff <Integer>
    Sequence length cut-off
    Sequences shorter than given length will be discarded
    default: -1 (i.e. length filtering is OFF)
 NOTE: filtering can be performed using any one of (-c), (-p) and (-
t5 and/or -t3) switches at a time
--------------------------------- Output Options -------------------
--------------
  -o | -outputFile <Output file name>
    Output will be stored in the given file
    default: By default, output file will be stored where the input
file is
```

(D) Statistics Tools

(1) AvgQuality.pl: Tool calculates average quality score for each read and overall average quality score for the given file. This tool takes a quality file in FASTA format as an input.

```
Usage: perl AvgQuality.pl <options>

AvgQuality.pl options:

### Input quality (FASTA) (Required)
  -i <Quality file>
    Quality file in FASTA format

### Other options [Optional]
  -h | -help
    Prints this help
  -o | -outputFile <Output file name>
    Output will be stored in the given file
    default: By default, quality statistics file will be stored
where the input file is
```

(2) N50Stat.pl: This tool calculates different statistics for read file given in FASTA format. It calculates total number of reads/sequences, total and individual (A,T,C,G and N) number of bases, G+C and A+T counts, and minimum, maximum, average, median, N25, N50, N75, N90 and N95 read/sequence length. Following are the options provided for this tool.

```
Usage: perl N50Stat.pl <options>
```

```
N50Stat.pl options:

### Input reads/sequences (FASTA) (Required)
  -i <Read/Sequence file>
    Read/Sequence in fasta format

### Other options [Optional]
  -h | -help
    Prints this help
  -o | -outputFile <Output file name>
    Output will be stored in the given file
    default: By default, N50 statistics file will be stored where
the input file is
```

7. Sample commands

- For quality control analysis of two paired-end sequencing data (pe11.fq-pe12.fq and pe21.fq-pe22.fq) and two single-end sequencing data (se1.fq and se2.fq) from Illumina platform using default parameters except statistics in tab-delimited file format:

  perl IlluQC.pl -pe pe11.fq pe12.fq 2 A -pe pe21.fq pe22.fq 2 A -se se1.fq 1 1 -se se2.fq 1 A -statOutFmt 2 -p 4

  Options used:

  -pe pe11.fq pe12.fq 2 A: The first paired read data is inputted using *-pe* switch followed by file names for both the ends followed by the primer/adaptor library used in sequencing (`2 = Paired-End DNA Library`) and FASTQ variant (`A = Automatic detection of FASTQ variant`)

  -se se1.fq 1 1: For single-end sequencing data, *-se* switch is used followed by name of the read file followed by the primer/adaptor library used (`1 = Genomic DNA/Chip-seq Library`) and FASTQ variant (`1 = Sanger (Phred+33, 33 to 73)`)

  -p 4: This option state the number of processes to be used. All four read data are processed simultaneously. For instance, the value "2" for this switch will process two datasets concurrently.

  -statOutFmr 2: The statistics will be printed in a tab-delimited file format.

This command will perform quality control and filtering of two paired-end and two single-end sequencing datasets simultaneously (-p 4, four libraries are processed all together). The filtered files are generated where the input files are.

- For quality control analysis without primer/adaptor contamination removal for two 454 datasets (lib1.fna-lib1.qual and lib2.fna-lib2.qual):

  perl 454QC.pl -i lib1.fna lib1.qual n -i lib2.fna lib2.qual n -n 7 -o FilteredFiles

  Options used:

  -i lib1.fna lib1.qual n: The switch *-i* is used to input read and quality files (FASTA format) of 454 sequencing data followed by the primer/adaptor library used for sequencing (n, skips the primer/adaptor filtering step)

  -n 7: The switch *-n* states the minimum length of the homopolymer to be trimmed. This will trim reads from the first base of homopolymer (at least 7 base pairs in length).

  -o FilteredFiles: A folder will be created in the current directory and the resulting filtered and statistics files are stored in it.

8. Contact details

Ravi Patel (ravipatel@nipgr.res.in)
Mukesh Jain (mjain@nipgr.res.in)

9. Citation

If toolkit has been used for any publication, please cite as below

Patel RK, Jain M (2012). NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE*, **7(2)**: e30619. (http://www.nipgr.res.in/ngsqctoolkit.html)

10. Terms of Use