

## FastQC - A high throughput sequence QC analysis tool

### SYNOPSIS

```
fastqc seqfile1 seqfile2 .. seqfileN
```

```
fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam]
      [-c contaminant file] seqfile1 .. seqfileN
```

### DESCRIPTION

FastQC reads a set of sequence files and produces from each one a quality control report consisting of a number of different modules, each one of which will help to identify a different potential type of problem in your data.

If no files to process are specified on the command line then the program will start as an interactive graphical application. If files are provided on the command line then the program will run with no user interaction required. In this mode it is suitable for inclusion into a standardised analysis pipeline.

The options for the program are as follows:

- h --help            Print this help file and exit
- v --version        Print the version of the program and exit
- o --outdir        Create all output files in the specified output directory. Please note that this directory must exist as the program will not create it. If this option is not set then the output file for each sequence file is created in the same directory as the sequence file which was processed.
- casava            Files come from raw casava output. Files in the same sample group (differing only by the group number) will be analysed as a set rather than individually. Sequences with the filter flag set in the header will be excluded from the analysis. Files must have the same names given to them by casava (including being gzipped and ending with .gz) otherwise they won't be grouped together correctly.
- nano             Files come from nanopore sequences and are in fast5 format. In this mode you can pass in directories to process and the program will take in all fast5 files within those directories and produce a single output file from the sequences found in all files.
- nofilter          If running with --casava then don't remove read flagged by casava as poor quality when performing the QC analysis.
- extract          If set then the zipped output file will be uncompressed in the same directory after it has been created. By default this option will be set if fastqc is run in non-interactive mode.
- j --java           Provides the full path to the java binary you want to use to launch fastqc. If not supplied then java is assumed to be in your path.

<code>--noextract</code>	Do not uncompress the output file after creating it. You should set this option if you do not wish to uncompress the output when running in non-interactive mode.
<code>--nogroup</code>	Disable grouping of bases for reads >50bp. All reports will show data for every base in the read. WARNING: Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned!
<code>--min_length</code>	Sets an artificial lower limit on the length of the sequence to be shown in the report. As long as you set this to a value greater or equal to your longest read length then this will be the sequence length used to create your read groups. This can be useful for making directly comparable statistics from datasets with somewhat variable read lengths.
<code>-f --format</code>	Bypasses the normal sequence file format detection and forces the program to use the specified format. Valid formats are bam,sam,bam_mapped,sam_mapped and fastq
<code>-t --threads</code>	Specifies the number of files which can be processed simultaneously. Each thread will be allocated 250MB of memory so you shouldn't run more threads than your available memory will cope with, and not more than 6 threads on a 32 bit machine
<code>-c</code> <code>--contaminants</code>	Specifies a non-default file which contains the list of contaminants to screen overrepresented sequences against. The file must contain sets of named contaminants in the form name[tab]sequence. Lines prefixed with a hash will be ignored.
<code>-a</code> <code>--adapters</code>	Specifies a non-default file which contains the list of adapter sequences which will be explicitly searched against the library. The file must contain sets of named adapters in the form name[tab]sequence. Lines prefixed with a hash will be ignored.
<code>-l</code> <code>--limits</code>	Specifies a non-default file which contains a set of criteria which will be used to determine the warn/error limits for the various modules. This file can also be used to selectively remove some modules from the output all together. The format needs to mirror the default limits.txt file found in the Configuration folder.
<code>-k --kmers</code>	Specifies the length of Kmer to look for in the Kmer content module. Specified Kmer length must be between 2 and 10. Default length is 7 if not specified.
<code>-q --quiet</code>	Suppress all progress messages on stdout and only report errors.
<code>-d --dir</code>	Selects a directory to be used for temporary files written when generating report images. Defaults to system temp directory if not specified.

## BUGS

Any bugs in fastqc should be reported either to [simon.andrews@babraham.ac.uk](mailto:simon.andrews@babraham.ac.uk)

or in [www.bioinformatics.babraham.ac.uk/bugzilla/](http://www.bioinformatics.babraham.ac.uk/bugzilla/)