QAA notebook

Thursday, August 31, 2023 3:56 PM

```
 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 as 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. If --delete is
            also specified then the zip file will be removed after the
            contents are unzipped.

-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.

--noextract    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.

--nogroup      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!

--min_length   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 comaparable statistics from 
                    datasets with somewhat variable read lengths.

--dup_length    Sets a length to which the sequences will be truncated when
                    defining them to be duplicates, affecting the duplication and
                    overrepresented sequences plot.  This can be useful if you have
                    long reads with higher levels of miscalls, or contamination with
                    adapter dimers containing UMI sequences.


-f --format     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


--memory        Sets the base amount of memory, in Megabytes, used to process 
                    each file.  Defaults to 512MB.  You may need to increase this if
                    you have a file with very long sequences in it.

--svg           Save the graphs in the report in SVG format.

-t --threads    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

-c              Specifies a non-default file which contains the list of
--contaminants  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.

-a              Specifies a non-default file which contains the list of
--adapters      adapter sequences which will be explicity searched against
                    the library. The file must contain sets of named adapters
```

```
                the library. The file must contain sets of named adapters
                in the form name[tab]sequence.  Lines prefixed with a hash
                will be ignored.

    -l          Specifies a non-default file which contains a set of criteria
    --limits       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.

   -k --kmers      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.

   -q --quiet      Suppress all progress messages on stdout and only report errors.

   -d --dir        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
    or in www.bioinformatics.babraham.ac.uk/bugzilla/
```

Both R1 time info:

```
Command being timed: "fastqc /projects/bgmp/shared/2017
_sequencing/demultiplexed/34_4H_both_S24_L008_R1_
001.fastq.gz --outdir /projects/bgmp/leylacuf/bioinfo/Bi623"
        User time (seconds): 47.37
        System time (seconds): 2.13
        Percent of CPU this job got: 92%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 0:53.49
        Average shared text size (kbytes): 0
```

```
		Average unshared data size (kbytes): 0
		Average stack size (kbytes): 0
		Average total size (kbytes): 0
		Maximum resident set size (kbytes): 189324
		Average resident set size (kbytes): 0
		Major (requiring I/O) page faults: 320
		Minor (reclaiming a frame) page faults: 136219
		Voluntary context switches: 5612
		Involuntary context switches: 1964
		Swaps: 0
		File system inputs: 62456
		File system outputs: 2520
		Socket messages sent: 0
		Socket messages received: 0
		Signals delivered: 0
		Page size (bytes): 4096
		Exit status: 0
------------------------------------------
```

Both R2 time info:

```
Command being timed: "fastqc /projects/bgmp/shared/2017
_sequencing/demultiplexed/34_4H_both_S24_L008_R2_001.fastq.gz --
outdir /projects/bgmp/leylacuf/bioinfo/Bi623/fastqc_output"
	User time (seconds): 46.84
	System time (seconds): 2.24
	Percent of CPU this job got: 87%
	Elapsed (wall clock) time (h:mm:ss or m:ss): 0:56.25
	Average shared text size (kbytes): 0
	Average unshared data size (kbytes): 0
	Average stack size (kbytes): 0
	Average total size (kbytes): 0
	Maximum resident set size (kbytes): 183148
	Average resident set size (kbytes): 0
	Major (requiring I/O) page faults: 0
	Minor (reclaiming a frame) page faults: 133501
	Voluntary context switches: 4701
	Involuntary context switches: 1726
```

```
        Involuntary context switches: 1720
        Swaps: 0
        File system inputs: 0
        File system outputs: 2560
        Socket messages sent: 0
        Socket messages received: 0
        Signals delivered: 0
        Page size (bytes): 4096
        Exit status: 0
```

Mbnl R1 time info:

```
Command being timed: "fastqc /projects/bgmp/shared/2017
_sequencing/demultiplexed/6_2D_mbnl_S5_L008_R1_001.fastq.gz --
outdir /projects/bgmp/leylacuf/bioinfo/Bi623/fastqc_output"
        User time (seconds): 57.35
        System time (seconds): 2.52
        Percent of CPU this job got: 98%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 1:00.63
        Average shared text size (kbytes): 0
        Average unshared data size (kbytes): 0
        Average stack size (kbytes): 0
        Average total size (kbytes): 0
        Maximum resident set size (kbytes): 196184
        Average resident set size (kbytes): 0
        Major (requiring I/O) page faults: 1
        Minor (reclaiming a frame) page faults: 119115
        Voluntary context switches: 5571
        Involuntary context switches: 2278
        Swaps: 0
        File system inputs: 24
        File system outputs: 2544
        Socket messages sent: 0
        Socket messages received: 0
        Signals delivered: 0
        Page size (bytes): 4096
        Exit status: 0
```

Mbnl R2 time info:

```
Command being timed: "fastqc /projects/bgmp/shared/2017
_sequencing/demultiplexed/6_2D_mbnl_S5_L008_R2_001.fastq.gz --
outdir /projects/bgmp/leylacuf/bioinfo/Bi623/fastqc_output"
	User time (seconds): 56.89
	System time (seconds): 2.86
	Percent of CPU this job got: 98%
	Elapsed (wall clock) time (h:mm:ss or m:ss): 1:00.72
	Average shared text size (kbytes): 0
	Average unshared data size (kbytes): 0
	Average stack size (kbytes): 0
	Average total size (kbytes): 0
	Maximum resident set size (kbytes): 165744
	Average resident set size (kbytes): 0
	Major (requiring I/O) page faults: 0
	Minor (reclaiming a frame) page faults: 106143
	Voluntary context switches: 5315
	Involuntary context switches: 2005
	Swaps: 0
	File system inputs: 0
	File system outputs: 2528
	Socket messages sent: 0
	Socket messages received: 0
	Signals delivered: 0
	Page size (bytes): 4096
	Exit status: 0
```

```
Command being timed: "./Demultiplex_quality_score_plotting -f
/projects/bgmp/shared/2017_sequencing/demultiplexed/34_
4H_both_S24_L008_R1_001.fastq.gz -l 102 -o
/projects/bgmp/leylacuf/bioinfo/Bi623/both_R1_plotting_output"
	User time (seconds): 177.96
	System time (seconds): 0.22
	Percent of CPU this job got: 95%
	Percent of CPU this job got: 25%
	Elapsed (wall clock) time (h:mm:ss or m:ss): 11:33.67
	Average shared text size (kbytes): 0
	Average unshared data size (kbytes): 0
	Average stack size (kbytes): 0
	Average total size (kbytes): 0
	Maximum resident set size (kbytes): 58320
	Average resident set size (kbytes): 0
	Major (requiring I/O) page faults: 0
	Minor (reclaiming a frame) page faults: 14462
	Voluntary context switches: 479
	Involuntary context switches: 16492
	Swaps: 0
	File system inputs: 0
	File system outputs: 0
	Socket messages sent: 0
	Socket messages received: 0
	Signals delivered: 0
	Page size (bytes): 4096
	Exit status: 0
Command being timed: "./Demultiplex_quality_score_plotting -f
/projects/bgmp/shared/2017_sequencing/demultiplexed/34_
4H_both_S24_L008_R2_001.fastq.gz -l 102 -o
/projects/bgmp/leylacuf/bioinfo/Bi623/both_R2_plotting_output"
	User time (seconds): 184.07
	System time (seconds): 0.23
	Percent of CPU this job got: 26%
	Elapsed (wall clock) time (h:mm:ss or m:ss): 11:34.58
	Average shared text size (kbytes): 0
	Average unshared data size (kbytes): 0
	Average stack size (kbytes): 0
	Average total size (kbytes): 0
	Maximum resident set size (kbytes): 58504
	Average resident set size (kbytes): 0
	Major (requiring I/O) page faults: 0
	Minor (reclaiming a frame) page faults: 14464
	Voluntary context switches: 488
	Involuntary context switches: 16806
	Swaps: 0
```