

CEGX-QC

User Documentation v0.2

CEGX Bioinformatics Team

Cambridge Epigenetix (CEGX), Babraham Research Campus, Cambridge, CB22 3AT

Contact email: technical@cegx.co.uk

www.cegx.co.uk

cegQC: <https://bitbucket.org/cegx-bfx/cegqc>

11/02/2015



Contents

1	What is cegxQC?	2
2	Download cegxQC	2
3	Installation Instructions	2
4	Running cegxQC	2
5	CEGX Specific Analysis Modules	3
5.1	CEGX Control Coverage	3
5.2	CEGX Conversion Summary	4
5.3	CEGX Per Control Conversion	4
6	Feature Requests and Bug Reports	5
7	Other Recommended Packages	5
7.1	FastQC	5
7.2	bsExpress	5
7.3	Bismark	6
8	GPLv3 License	6

1 What is cegxC?

CEGX has developed a post-sequencing program, cegxC, a customised version of FastQC, to output a set of summary documents and QC reports based on the conversion performance of the sequencing spike-in controls.

cegxC is designed to perform quality control analysis of fastq files from bisulfite (BS-Seq) and oxidised bisulfite (oxBS-Seq) sequencing.

2 Download cegxC

cegxC is freely available as a pre-compiled binary

<https://bitbucket.org/ceg-bfx/cegxc/downloads>

Or as source code

<https://bitbucket.org/ceg-bfx/cegxc/src>

3 Installation Instructions

An INSTALL.txt file is provided in the cegxC software package. Below is a brief summary of the steps required to install and run cegxC.

1. cegxC, like its parent program FastQC, is a java application. In order to run it needs your system to have a suitable Java Runtime Environment (JRE) installed. A JRE is available from <http://www.java.com>
2. Unzip the downloaded cegxC package

As cegxC is a customised version of FastQC the installation and running of the program are almost identical. We therefore recommend reading the FastQC documentation

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.

4 Running cegxC

cegxC can be run in either an interactive graphical mode or as part of a pipeline using the command line interface.

Running cegxC Interactively

Windows: Double click on the `run_cegxc` bat file.

MacOSX/Linux: A wrapper script, `cegxc`, is the simplest way to launch the program.

Running cegxC as part of a pipeline

To run cegxC non-interactively use the `cegxC` wrapper script and specify a list of files to process on the command line

```
cegxC afaftqfile.afaftq anotherfaftqfile.fq
```

As many faftq files can be specified on the command line as required. We recommend only launching two or three to avoid running out of memory. If no faftq files are specified cegxC will launch in graphical mode, which could cause errors if the display isn't set on e.g. a cluster.

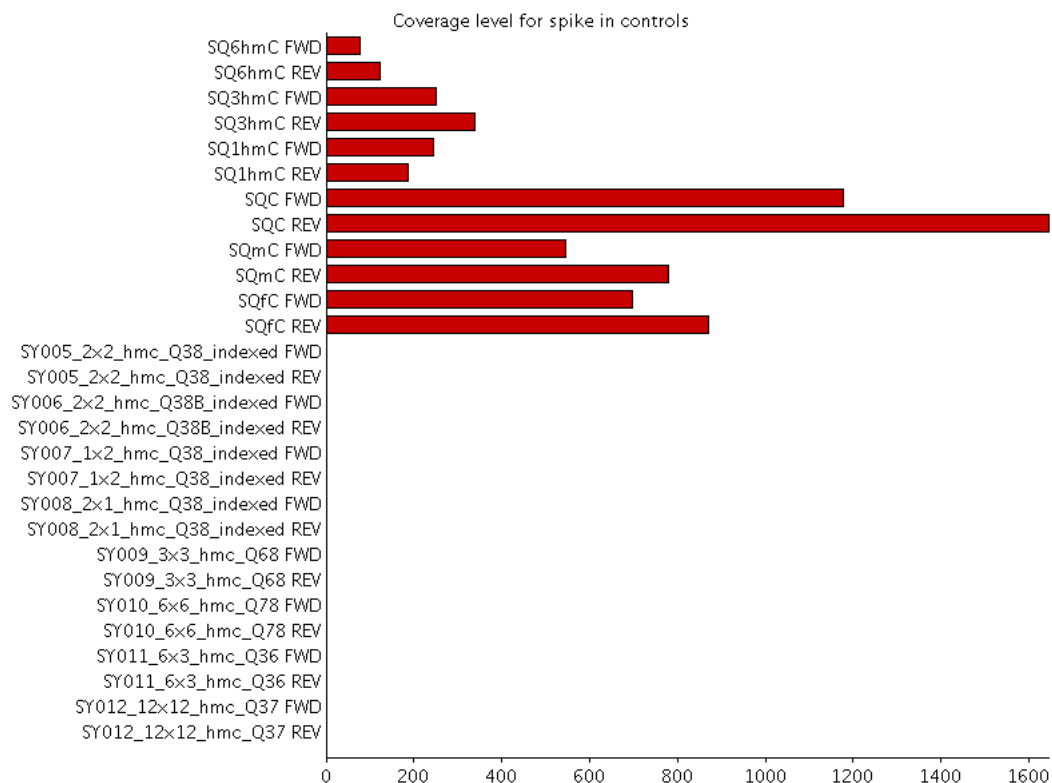
5 CEGX Specific Analysis Modules

As cegxC is a customised version of FastQC we recommend reading the FastQC documentation for the more general non-control sequence features

<http://www.bioinformatics.babraham.ac.uk/projects/afaftqc>.

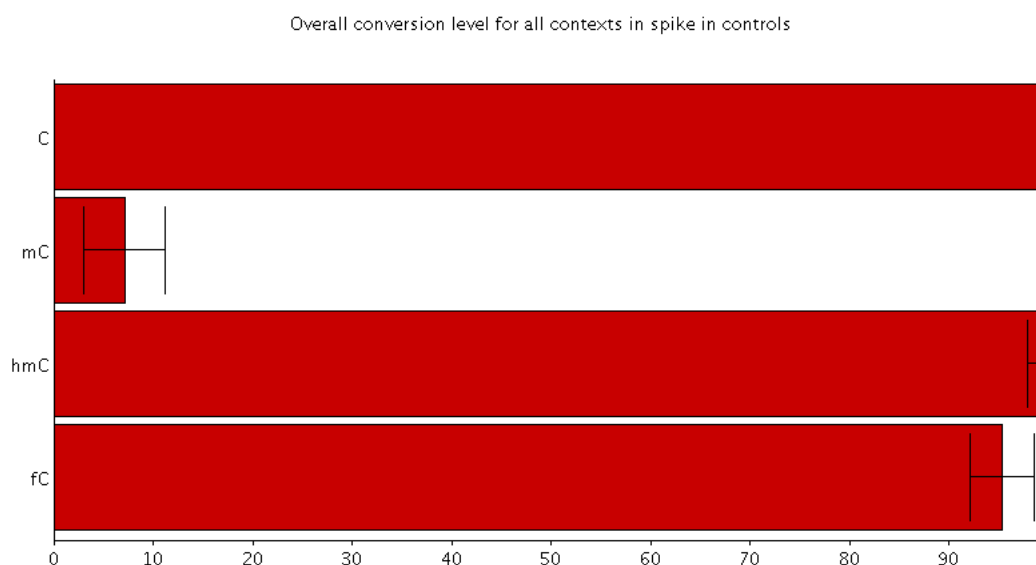
5.1 CEGX Control Coverage

Coverage levels are plotted for each of the spike in controls in both forward and reverse orientations.



5.2 CEGX Conversion Summary

The conversion levels are summarised and plotted across all of the spike-in-controls for cytosines, methyl cytosines, hydroxymethyl cytosines and formyl cytosine.



5.3 CEGX Per Control Conversion

The modified cytosine positions in each of the controls are displayed as coloured blocks. The height of the blocks are proportional to the conversion rate of the specific cytosine position.

SO6hmC FWD (76)	mC C hmC C hmC hmC mC C mC C mC hmC hmC mC C hmC
SQ6hmC REV (123)	mC hmC mC C C hmC C hmC mC mC hmC C mC hmC hmC
SQ3hmC FWD (250)	mC C mC C mC hmC C mC mC mC hmC C mC hmC hmC
SQ3hmC REV (338)	mC C mC C mC hmC C mC mC mC hmC C mC hmC hmC
SQ1hmC FWD (245)	mC C mC C mC C mC C mC C mC hmC C mC C mC
SQ1hmC REV (188)	mC C mC C C C mC C mC mC mC mC hmC C mC C mC
SQC FWD (1177)	C C C C C C C C mC mC C C C C C C C C
SQC REV (1644)	C C C C C C C C mC C C C C C C C C C
SQmC FWD (546)	mC mC mC mC mC mC mC mC mC mC mC mC mC mC mC
SQmC REV (779)	mC mC mC mC mC mC mC mC mC mC mC mC mC mC mC
SQfC FWD (698)	C C fC C C C C C mC C C C C C C C C C
SQfC REV (871)	C C fC C C C C C mC mC C C C C C C C C
SY005 2x2 hmc Q38 indexed FWD (0)	C mC C mC C C C mC C C hmC hmC
SY005 2x2 hmc Q38 indexed REV (0)	C C mC mC C C mC C C mC C hmC hmC
SY006 2x2 hmc Q38B indexed FWD (0)	C mC C mC C C C mC C C mC C C C C
SY006 2x2 hmc Q38B indexed REV (0)	C C C mC C C C mC C C mC C C C C
SY007 1x2 hmc Q38 indexed FWD (0)	C mC C mC C C C C mC C C C hmC
SY007 1x2 hmc Q38 indexed REV (0)	C C mC C C C mC C C C hmC hmC
SY008 2x1 hmc Q38 indexed FWD (0)	C mC C mC C C C mC C C C hmC hmC
SY008 2x1 hmc Q38 indexed REV (0)	C C C mC C C C mC C C C hmC
SY009 3x3 hmc Q68 FWD (0)	C mC C mC C C C mC C C hmC hmC hmC
SY009 3x3 hmc Q68 REV (0)	C C mC C C C mC C C mC C hmC hmC hmC
SY010 6x6 hmc Q78 FWD (0)	C mC C mC mC C C C hmC hmC hmC hmC hmC hmC
SY010 6x6 hmc Q78 REV (0)	C C mC C C C mC C C hmC hmC hmC hmC hmC hmC
SY011 6x3 hmc Q36 FWD (0)	C mC C mC C C C mC C C hmC hmC hmC hmC hmC
SY011 6x3 hmc Q36 REV (0)	C mC C mC C C C mC C C hmC hmC hmC
SY012 12x12 hmc Q37 FWD (0)	C mC C mC C C C mC C C hmC hmC hmC hmC hmC hmC
SY012 12x12 hmc Q37 REV (0)	C mC C mC mC C C C hmC hmC hmC hmC hmC hmC

Note: This section is undergoing development and we hope to release an updated version soon.

6 Feature Requests and Bug Reports

We are always very happy to hear about any features you think should be added to cegxC. We are actively developing the program and plan to add new features.

Bug reports or any issues using the software can be reported on our software repository site in the issue tracking section

<https://bitbucket.org/ceg-x-bfx/cegxc/issues>

7 Other Recommended Packages

7.1 FastQC

The parent program for cegxC

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

7.2 bsExpress

bsExpress is designed to perform quality control bisulfite (BS-Seq) and oxidised bisulfite (oxBS-Seq) libraries using ad hoc control sequences where cytosine modification

are known. However, the pipeline is not limited to control sequences but is also suitable for processing (ox)BS-Seq data from raw fastq files to genome-wide methylation calls.

<https://code.google.com/p/oxbs-sequencing-qc/wiki/bsExpressDoc>

7.3 Bismark

Short read aligner for bisulfite (BS-Seq) and oxidised bisulfite (oxBS-Seq) sequencing data

<http://www.bioinformatics.babraham.ac.uk/projects/bismark>

8 GPLv3 License

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <http://www.gnu.org/licenses/>.