CEGX
Cambridge-Epigenetix

Document: cegx_bsexpress_docker
Version: v0.5
Website: www.cegx.co.uk
Contact: bfx@cegx.co.uk

CEGX_BsExpress_Docker

Introduction

CEGX has developed a custom post-processing script, bsExpress, that interfaces with Bismark to output a set of summary documents and QC reports based on the conversion performance of the sequencing spike-in controls. bsExpress is a program to perform quality control analysis of bisulfite (BS-Seq) and oxidised bisulfite (oxBS-Seq) sequencing libraries. 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.

Previous versions of bsExpress were problematic to install due to the long list of prerequisite programs that also needed to be installed. To simplify the installation and running there is now a dockerised version. Docker wraps up bsExpress and all of its required programs and configurations into a single package.

The instructions below assume you are using a Mac or Linux (Windows instructions to be added)

If you aren't familiar with the command line, how to change directories and list the files within a directory then this simple guide may help: Linux Guide

Instructions for installing cegx_bsexpress_0.5

1. Install the Docker Toolbox

This installs the docker environment on your computer allowing dockerised applications to be run. Docker ToolBox

To start the Docker engine running: in Finder, navigate to Applications, open "Docker QuickStart Terminal". You should see an image of a whale appear (rendered in text).

2. Download the cegx_bsexpress_0.5 image

If you are familiar with git, use:

```
cd /User/joebloggs/Desktop

git clone https://russellshamilton@bitbucket.org/cegx-bfx/cegx_bsexpress_docker.git
```

Otherwise you can download `cegx_bsexpress_0.5.tar.gz` from [https://bitbucket.org/cegx-bfx/cegx_bsexpress_docker/src]

Note: Don't double click the file when it appears in the bottom bar of your browser. This will uncompress it into an incorrect format.

3. Import the cegx_bsexpress_0.5 docker image into Docker

Open a terminal (In Finder Applications/Utilities/Terminal.app)
Change to the directory where you downloaded the cegx_bsexpress_0.5 docker image

```
cd /User/joebloggs/Desktop
```

Import the image

```
docker load < cegx_bsexpress_0.5.tar.gz
```

Check the docker image was loaded

```
docker images
```

You should see something very similar to this:

```
REPOSITORY          TAG       IMAGE ID       CREATED       VIRTUAL SIZE
cegx_bsexpress_0.5  latest    0884de2e3bc4   2 days ago    1.274 GB
```

The docker version of bsExpress should now be ready to use!

Running cegx_bsexpress_0.5

1. Start the Docker Engine

In Finder, navigate to Applications and open "Docker QuickStart Terminal". You can also do this from Launchpad (little rocket icon in the status bar), look for the Docker QuickStart Terminal icon and double click to launch.

You should see a whale picture appear in the terminal. It takes a few minutes to load, so be patient! Once open, do not close this terminal window until you are finished, this keeps the Docker instance live.

2. Change to the directory containing your fastq.gz files

At the prompt, change directory to the location where you saved the fastq file.

```
cd /User/joebloggs/MyData/
```

3. Run bsExpress

- **Option A ("easy, but hard to remember"):**

```
docker run -v=/Users/joebloggs/Desktop/MyData/:/Data -it cegx_bsexpress_0.5
auto_bsExpress
```

or if you are already in the directory with the fastq.gz.files

```
docker run -v=`pwd`:/Data -it cegx_bsexpress_0.5 auto_bsExpress
```

TIP: *fastq.gz downloaded from BaseSpace are buried deep in the folder structure from the sequencer. Change directory down to the level of the folder structure where the fastq are located and then execute the docker run command. If you execute the command in the top level of the run folder the analysis won't run.*

- **Option B ("easier, but involves extra set up step"):**

The command above isn't easy to remember, so there is a script available which should be installed into /usr/local/bin on your computer

Install (note you will be asked for your computers password)

Download local_auto_bsExpress from https://bitbucket.org/cegx-bfx/cegx_bsexpress_docker/src

Or use git

```
git clone https://bitbucket.org/cegx-bfx/cegx_bsexpress_docker/src
```

```
sudo cp local_auto_bsExpress /usr/local/bin/

sudo chmod 755 /usr/local/bin/local_auto_bsExpress
```

Change into the directory where your fastq.gz file is:

```
cd /User/joebloggs/Desktop/MyData/
```

Run the script with

```
local_auto_bsExpress
```

This will automatically run bsExpress on all fastq.gz R1 files in the directory. It will create two directories one for the DC (digestion control) and one for the SC (Spike in control).

- **Option C ("difficult": traditional version supplying command line arguments):**

```
docker run -v=/Users/joebloggs/Documents/CEGX-Projects/cegx-controls/:/Data -it
cegx_bsexpress_0.5 bsExpress
```

Analysis results

In Finder, navigate to the analysis folder. Alongside the original fastq.gz data files there will be the following assortment of analysis result files:

<fastq>SQ.bsExpressSummary.txt

This is a high level conversion metric summary. It provides the C2U, mC2T, hmC2U and fc2U conversion rates as an average of all modified bases per control, averaged across all controls.

```
SQ controls
Num Reads:       875199       After Trim:       53381
chrom        mod        pct.met        tot_reads
SQ           5fC        6.76           1063
SQ           5hmC       4.43           1063
SQ           5mC        94.99          1063
SQ           C          0.43           1063
```

<fastq>runqc_SQ sequencing control results folder

This contains a set of useful files relating to the sequencing control analysis.

<fastq.gz>trimming_report.txt

Summary of high level run metadata.

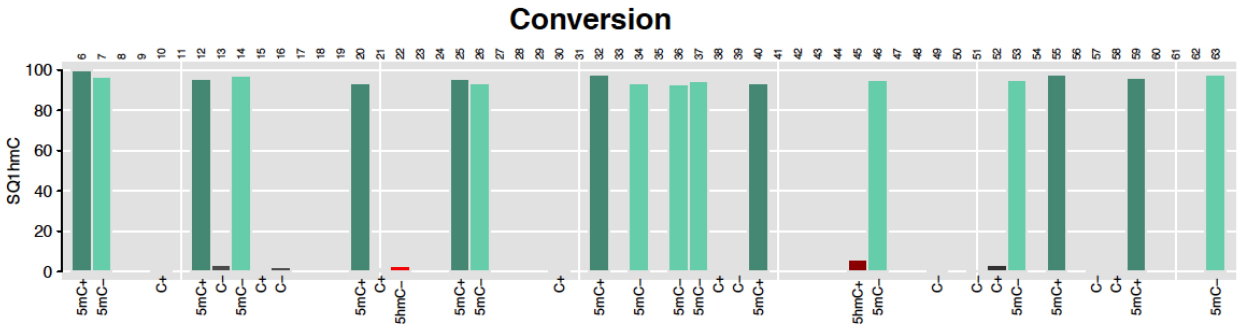
```
SUMMARISING RUN PARAMETERS
Input filename: data.fastq.gz
Trimming mode: single-end
Trim Galore version: 0.3.7
Quality Phred score cutoff: 20
Quality encoding type selected: ASCII+33
Adapter sequence: 'AGATCGGAAGAGC'
Maximum trimming error rate: 0.1 (default)
Minimum required adapter overlap (stringency): 13 bp
Minimum required sequence length before a sequence gets removed: 20 bp
All Read 1 sequences will be trimmed by 50 bp from their 3' end to avoid poor qualities
or biases
Output file will be GZIP compressed
```

<fastq.gz>runqc_SQ.bam and <fastq.gz>runqc_SQ.bam.bai

Binary alignment and index files for the sequencing control reads. Non-human readable files.

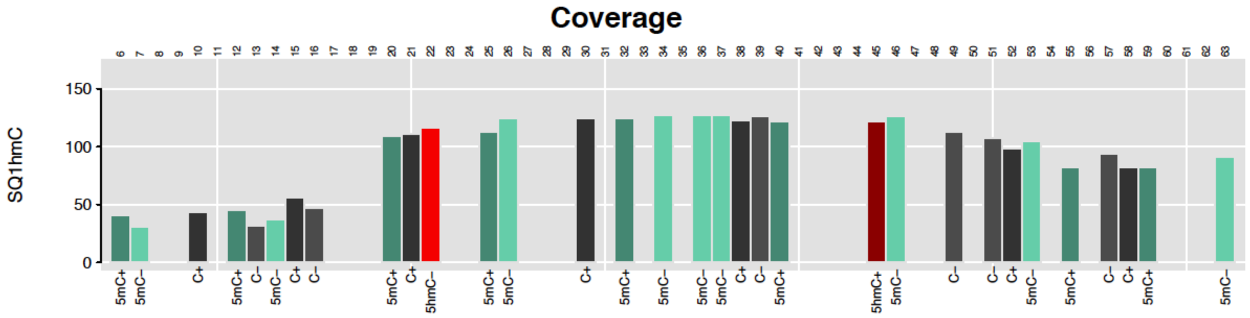
<fastq.gz>runqc_SQ.conversion.pdf

Conversion plot split out per control. The level of conversion can be visualized per base for each control. Use this to diagnose whether there is any systematic positional bias in conversion within the control reads.



<fastq.gz>runqc_SQ.coverage.pdf

Coverage plot split out per control. The level of coverage can be visualized per base for each control. Use this to diagnose whether there is any systematic positional bias in coverage within the control reads.



<fastq.gz>runqc_SQ.mcall.bdg.gz

A bed/bedgraph style format file describing the methylation status of each cytosine position

The columns with example data:

Control sequence name	Cytosine position 0-based	Cytosine position 1-based.	Percentage unconverted C (% methylated)	Number of unconverted C (count methylated)	Total number of converted and unconverted C.	Strand: + for C on forward strand, - for C on reverse strand
chr1	424	425	0.0	0	54	+

<fastq.gz>runqc_SQ.oxqc_summary.txt

Per-control conversion averages summary.

chrom	mod	pct.met	cnt.met	tot_reads
SQ1hmC	5hmC	4.2	10	238
SQ1hmC	5mC	95.23	1538	1615
SQ1hmC	C	0.78	9	1157
SQ3hmC	5hmC	3.5	14	400

SQ3hmC	5mC	93.77	873	931
SQ3hmC	C	0.66	5	759
SQ6hmC	5hmC	5.45	23	422
SQ6hmC	5mC	88.38	350	396
SQ6hmC	C	0.3	1	336
SQC	5mC	97.94	95	97
SQC	C	0.15	1	660
SQfC	5fC	6.76	5	74
SQfC	5mC	96.15	175	182
SQfC	C	0.29	7	2444
SQmC	5mC	96.22	2242	2330
all	5fC	6.76	5	74
all	5hmC	4.43	47	1060
all	5mC	94.99	5273	5551
all	C	0.43	23	5356

<fastq.gz>runqc_SQ.oxqc.txt

Per-position, per-control conversion summary.

chrom	pos	pct.met	cnt.met	tot_reads	strand	base_iupac	short_description
SQ1hmC	6	100	41	41	+	C	5mC+
SQ1hmC	7	96.77	30	31	-	G	5mC-
SQ1hmC	10	0	0	43	+	C	C+
SQ1hmC	12	95.56	43	45	+	C	5mC+
SQ1hmC	13	3.12	1	32	-	G	C-
SQ1hmC	14	97.3	36	37	-	G	5mC-
SQ1hmC	15	0	0	56	+	C	C+
SQ1hmC	16	2.13	1	47	-	G	C-
SQ1hmC	20	93.58	102	109	+	C	5mC+
SQ1hmC	21	0.9	1	111	+	C	C+

<fastq.gz>runqc_SQ.R1.short.fq.gz_bismark_bt2_SE_report.txt

Bismark alignment report, showing analysis parameters, alignment metadata and high level CPG, CPH, CHH methylation information.

```
Bismark report for: data.runqc_SQ/data.runqc_SQ.R1.short.fq.gz (version: v0.14.0)
Option '--directional' specified (default mode): alignments to complementary strands
(CTOT, CTOB) were ignored (i.e. not performed)
Bismark was run with Bowtie 2 against the bisulfite genome of
/cegx_bsexpress/control_reference/ with the specified options: -q --phred33 --score-min
L,0,-0.2 --ignore-qualsFinal
```

```
Alignment report
======================
Sequences analysed in total: 53381
Number of alignments with a unique best hit from the different alignments: 1063
Mapping efficiency: 2.0%
Sequences with no alignments under any condition: 45992
Sequences did not map uniquely: 6326
Sequences which were discarded because genomic sequence could not be extracted: 0
Number of sequences with unique best (first) alignment came from the bowtie output:
CT/CT:        604        ((converted) top strand)
CT/GA:        459        ((converted) bottom strand)
GA/CT:          0        (complementary to (converted) top strand)
GA/GA:          0        (complementary to (converted) bottom strand)...
```

<fastq.gz>fq.gz

Compressed trimmed fastq files. No need to do anything with these. They are intermediate fastqs generated during the analysis.