APPS ON ACCESSING AND DEMULTIPLEXING ILLUMINA DATA

PELLIGRINI LAB

Overview

A modular set of tools for easily accessing and demultiplexing Illumina HiSeq data from the BSCRC sequencing core. They can be used individually, or sequentially as needed. They are:

1) downloader

> makes it quick and easy to copy the reads (as gziped qseq files) from the BSCRC servers to your local machine or to the cluster

2) qseq2fastq

> will uncompress the read files (if needed) and convert them from qseq format to fastq format

3) demultiplexer

> use to segregate pooled reads based on their index sequence (i.e. sort multiplexed reads)

These tools are written in PERL, and should work on any standard *nix computer, such as Mac OSX, and Linux; including the hoffman2 cluster. They are made publically available at bitbucket and are released under the GNU General Public License.

Installation

All three tools are available in the htSeqTools bitbucket repository at:

> https://bitbucket.org/gallaher/htseqtools/downloads

To install on hoffman2, login to the cluster and type the following:

```
curl https://bitbucket.org/gallaher/htseqtools/get/master.zip –o htSeqTools.zip
unzip htSeqTools.zip
mv gallaher* htSeqTools/
rm htSeqTools.zip
```

Next, you may want to add the tools to your $PATH. To do this, use your preferred text editor to edit your ~/.bash_profile. Add the following line:

```
export PATH=$HOME/htSeqTools/bin/:$PATH
```

Save the file, and then type:

```
source ~/.bash_profile
```

Sample Usage

In the following hypothetical example, you have generated 10 different RNAseq libraries with the Illumina TruSeq single index kit, pooled them, and sequenced them on two lanes of the HiSeq 2000 sequencers at the BSCRC sequencing core. When the run was complete, you were given the following lane credentials in an email from the BSCRC:

```
SxaQSEQsYA007L6:s23NtC92kkQ6

SxaQSEQsYA007L5:u7xOr74QM9v5
```

You wish to copy the files to your account on the hoffman2 cluster, and demultiplex / sort the pooled reads back into 10 different sets.

The following example assumes that you are logged into the hoffman2 cluster, that you have installed htSeqTools, and that you have initiated an interactive session on one of the computing nodes. Be aware, that the demultiplexer can take a day or more to run, so you will probably want submit these to the queue. See script below for details.

Step 1: downloader

This tool makes it easy to transfer your data from BSCRC to the cluster.

1) Create a directory in your user folder with the name of your experiment and cd into it:

```
mkdir ~/myRNAseqExp/
cd ~/myRNAseqExp/
```

2) Run the downloader with your lane credential as an argument:

```
downloader SxaQSEQsYA007L5:u7xOr74QM9i5
```

The tool will create a directory in myRNAseqExp called 01_qseq/lane_5/, and then it will transfer your data to it from the BSCRC server (pan.pellegrini.mcdb.ucla.edu).

3) Repeat for your other lane(s) from the same run of the sequencer:

```
downloader SxaQSEQsYA007L6:s23NtC92kkQ6
```

This will put the reads from the second lane in 01_qseq/lane_6/.

Step 2: qseq2fastq

This tool will expand the files (if needed) and convert them from qseq to fastq format. It expects files of the format:

s_5_1_1101_qseq.txt.gz

which is the default format of reads from the BSCRC.

1) cd into the directory containing the raw read files:

   ```
   cd ~/myRNAseqExp/01_qseq/lane_5
   ```

2) run qseq2fastq:

   ```
   qseq2fastq
   ```

   The tool will create a directory in myRNAseqExp called 02_fastq/lane_5/, and copy the fastq files to it.

3) Repeat as needed for the other lanes from the same run of the sequencer:

   ```
   cd ~/myRNAseqExp/01_qseq/lane_6
   qseq2fastq
   ```

Running qseq2fastq on one lane of 50 NT single end reads generated by a HiSeq 4000 (>400 million reads total) takes approximately 2 hours to complete.

Step 3: demultiplexer

This tool will sort the reads from a lane of sequencing based on their index sequence. It expects matched sets of fastq sequence files formatted like this:

s_5_1_1101.fastq	the first sequencing read (required)
s_5_2_1101.fastq	the index read (required)
s_5_3_1101.fastq	the second sequencing read if pair-end sequencing was used (optional)

To run the demultiplexer with default settings and the standard Illumina TruSeq-compatible indices:

1) cd into the directory containing the fastq files:

```
cd ~/myRNAseqExp/02_fastq/lane_5
```

2) run the demultiplexer

```
demultiplexer
```

The tool will create a series of directories in myRNAseqExp called:

03_demultiplexed/lane_5/Index_01

03_demultiplexed/lane_5/Index_02

et cetera, for each index that was observed, and one called:

03_demultiplexed/lane_5/Unmatched

In addition, it will create an index report that indicates how many reads were matched to each index.

Running demultiplexer on one lane of 50 NT single end reads off of the HiSeq 4000 (>400 million reads total) takes approximately 34 hours to complete.

Advanced Usage

The demulitplexer has some additional features, such as the ability to specify your own index sequences, and to adjust its sequence matching stringency.

User-specified Indices

By default, the demultiplexer uses the following 24 Illumina TruSeq index sequences:

ATCACG	Index01	GATCAG	Index09	GTCCGC	Index18
CGATGT	Index02	TAGCTT	Index10	GTGAAA	Index19
TTAGGC	Index03	GGCTAC	Index11	GTGGCC	Index20
TGACCA	Index04	CTTGTA	Index12	GTTTCG	Index21
ACAGTG	Index05	AGTCAA	Index13	CGTACG	Index22
GCCAAT	Index06	AGTTCC	Index14	GAGTGG	Index23
CAGATC	Index07	ATGTCA	Index15	ACTGAT	Index25
ACTTGA	Index08	CCGTCC	Index16	ATTCCT	Index27

Alternatively, you can specify your own set of index sequences. Simply place them in a tab-delimited text file, one index per line, with the index sequence followed by a unique identifier, separated by a tab. The index sequence should consist of some combination of the letters [ACGT].

For example:

```
GCGTAC myFirstIndex
TTCGAA mySecondIndex
ATGCAT myThirdIndex
```

When you run the demultiplexer, specify your file of user-supplied indices as an argument to the -i flag:

```
demultiplexer –i myIndexFile.txt
```

Stringency

The demulitplexer uses a scoring system to match the index read to each index sequence, and you can adjust the threshold score that is considered a match. The score is calculated by taking the length of the index sequence, and subtracting 1 for each mismatch or 0.5 for each N, and then dividing that number by the length.

By default, the demultiplexer uses the moderate setting. Since each Illumina index sequence has at least two nucleotides distance from the others, the probability of getting a false match at this level is low, but not zero.

Setting	**Threshold Score**	**Allows in a 6 NT Index**
strict	1.00	perfect 6 / 6 matches
moderate (default)	0.80	1 mismatch or 2 Ns
loose	0.65	2 mismatches or 4 Ns

If two or more indices receive the same score, the read is placed in an “ambiguous” directory.

You can increase or decrease the stringency with the -s flag:

```
demultiplexer –s strict
```

or

```
demultiplexer –s loose
```

One-step qsub Script

You can run all three steps automatically on hoffman2 with the following script. Just replace the <add lane credentials here> with your lane credentials and run on the queue with the qsub command.

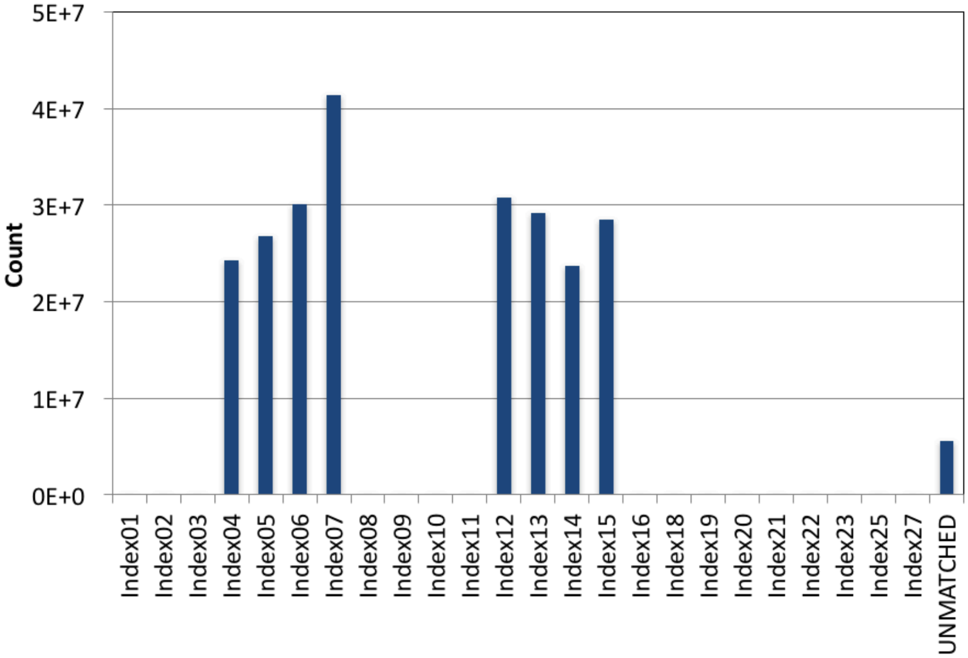
```
#!/bin/bash
#$ -cwd
#$ -V
#$ -l h_data=8g,h_rt=48:00:00,highp

CRED=<add lane credentials here>
LANE=$(echo $CRED | sed -E 's/SxaQSEQs.{5}L(.):.{12}/lane_\1/')

downloader $CRED
cd 01_qseq/$LANE/
qseq2fastq
cd ../../02_fastq/$LANE/
demultiplexer
```

Quality Control

When the demultiplexer is done, it generates a text file summarizing the results. This report specifies the number of reads matched to each index, plus the number of unmatched, plus the number of ambiguous reads, if any, and the total number of reads.

Review the report. There should be significantly more reads matched to the indices that you expect to be present relative to the unmatched and unused indices. You may wish to plot the resulting report data to visualize them.

Here is a real world example of a lane of sequencing containing 8 pooled samples, that was run by the demultiplexer with default settings:



A comparable number of reads was matched to the 8 indices that were used. There were negligible matches to the other 16 indices. Approximately 2% of the reads went unmatched.