Quick Start for Dr.seq

Dr.seq is a QC and analysis pipeline for Drop-seq data. By applying this pipeline, Dr.seq take two sequencing file as input (data_1.fastq for barcode information, data_2.fastq for reads information, see our testing data and Manual section for more information) and provides four groups of QC measurements for given Drop-seq data, including reads level, bulk-cell level, individual-cell level and cell-clustering level QC.
Here we provide an example to get you easily started on a linux/MacOS system with only python and R installed. To run Dr.seq with options specific to your data, you need to see Manual section for detailed usage.

Step1.Install pipeline

1. Make sure you have python2.7 and R(version >= 2.14.1) on linux or MAC OSX environment.
2. Get Dr.seq from https://Tarela@bitbucket.org/tarela/drseq.
3. Install Dr.seq on your server/computer (Please contact the administrator of that machine if you want their help)

```
$ unzip Dr.seq.1.0.zip
$ cd Dr.seq.1.0 #find your Dr.seq.1.0 folder and change directory to it
```

for root users:

```
$ sudo python setup.py install
```

if you are not root users, you can install Dr.seq at specific location with write permission

```
$ python setup.py install --prefix /home/drseq      # here you can
replace "/home/drseq" with any location you want

$ export PATH=/home/drseq/bin:$PATH    # setup PATH, so that system
knows where to find executable file

$ export PYTHONPATH=/home/drseq/lib/python2.7/site-packages:$PYTHONPATH
# setup PYTHONPATH, so that Dr.seq knows where to import modules
```

NOTE: To install Dr.seq on MAC OSX, users need to download and install Xcode beforehand
type:

```
$ Drseq.py --help
```

If you see help manual, you have successfully installed Dr.seq

Step2. Prepare annotation and mapping software

1. Get gene annotation according to the species of your sample.
   We provide convenient link for human (hg38) and mouse (mm10) at our webpage.
   You can download full annotation table from UCSC for other species (see Manual section)
   (Skip 2 and 3 if you already have bowtie2 and bowtie2 index)
2. Prepare mapping software (We use bowtie2 (version 2.2.6) for users to get quick start. By default we use STAR (version >= 2.5.0) as aligner, see Manual).
   We provide convenient link of executable bowtie2 for linux and MacOS user at out webpage. (otherwise you have to download full bowtie2 package and compile yourself, see Manual section):
   For root user, you can just copy executable bowtie2 to any default PATH, (for example, /usr/local/bin)

```
$ unzip bowtie2–2.2.6-linux-x86_64.zip
$ cd bowtie2-2.2.6-linux-x86_64
$ sudo cp bowtie2* /usr/local/bin     # don't forget to type * mark here
```

If you are not root user, you can copy bowtie2 to the PATH (/home/drseq/bin) you setup in step1

```
$ unzip bowtie2—2.2.6-linux-x86_64.zip
$ cd bowtie2-2.2.6-linux-x86_64
$ cp bowtie2* /home/drseq/bin     # don't forget to type * mark here
```

type

```
$ echo $PATH
```

to check the list of your default PATH

3. Prepare bowtie2 index
   We provide pre-built bowtie2 index at our webpage for quick start, it takes some time for downloading (You can build bowtie2 index yourself if the genome version is not hg38 nor mm10, see Manual section).

Step3. Run Dr.seq

1. Before you start running, Dr.seq will check your computer for pdflatex. If you have already installed pdflatex, Dr.seq will generate a summary QC report in addition to QC and analysis reports (see Manual section for the installation of pdflatex).
2. Now you can run Dr.seq pipeline to generate QC and analysis reports of your Dropseq data with two parts of sequencing data input (barcode file and reads file).
3. Here we provide an example of our simple mode on published Drop-seq data (GSM1626793) and display Dr.seq output in the following panel.
   Below is an example command for Dr.seq:

   ```
   $ Drseq.py simple -b SRR1853178_1.fastq -r SRR1853178_2.fastq -n
   GSM1626793_mouse_retinal -g /home/user/annotation/mm10_refgenes.txt
   --maptool bowtie2 --mapindex /home/user/bowtie2_index/mm10
   ```

   # Brief description of major parameter, see Manual section for more information
   # -b SRR1853178_1.fastq: FASTQ file containing barcode information, each barcode is composed by 12bp cell barcode and 8bp UMI
   # -r SRR1853178_2.fastq: FASTQ file containing reads (RNA sequence) information, this file will be aligned to genome with given parameter
   # -n name of your output results and output directory, note that no "/" should be appeared here
   # -g /home/user/annotation/mm10_refgenes.txt: absolute path of the genome annotation file
   # --mapindex /home/user/bowtie2_index/mm10: absolute path of bowtie2 index. Note that the genome version of the index should corresponded to the annotation file. If you have an index file looks like: /home/user/bowtie2_index/mm10.1.bt2, the red part is the one you should input here.
   # --maptool bowtie2: name of your mapping tools, choose from bowtie2 and STAR, by default we use STAR because of the speed. We use bowtie2 for quick start because bowtie2 consume less memory and suitable for both linux server and Mac computer.

Step4. Output and testing data

1. We provide Dr.seq output result for published Drop-seq data (GSM1626793) on our webpage
2. We also provide small testing data on our webpage for users to quickly test the flexibility of Dr.seq. Note that the testing data is only for users to get familiar with Dr.seq in a very short time, so we don't except this testing data to generate any meaningful results (For example, the duplicate rate distribution may not show different pattern because of low reads count, and there is not enough cells to provide a clear clustering pattern).