FAQ of Dr.seq

Dr.seq is a QC and analysis pipeline for Drop-seq data. By applying this pipeline, Dr.seq takes 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 answers for frequently asked question, if this section doesn't satisfy you, feel free to write email to 2xp10tarela@tongji.edu.cn. Your question will be daily replied.

Q1: In the Manual section you use STAR as default parameter and generate example results with STAR, but you use bowtie2 for Quick start. What's the difference?
A: We suggest using STAR as aligner mainly because of its speed. Drop-seq data requires huge amount reads to support thousand of single cell transcriptome to be detected, so the speed of alignment becomes a major factor we concern. In the quick start section we use bowtie2 because STAR requires > 30G memory for human genome, which is impossible for mac computers and small servers, while bowtie2 requires only 2~3G memory, which is suitable for almost all machines. Another reason is the mapping index for bowtie2 is much smaller than STAR (4G for bowtie2 index and 20G for STAR index), which is much easier to download.

Q2: I download your testing data but the result is not as clear as your example output. What's the difference between testing data and the data you use for generating example?
A: We provide testing is only for users to get familiar with and see the flexibility of Dr.seq. The Drop-seq data we use to generate example contains 500 million total reads, and the total file size of 2 FASTQ files is greater than 200 gigabyte (200G), which is too big and not suitable for users to get started in a short time. To get users quickly familiar with Dr.seq, we provide a small dataset, which could be download and process very quickly. 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).

Q3: What should I do if I do want to generate your example output myself?
A: If you want to generate our example output yourself, you can download the corresponded Drop-seq data (accession ID mentioned in Quick Start section) and run Dr.seq with all default parameter (you can use both simple mode and standard mode with STAR as aligner, mm10 as genome version, see other default parameter from template configure file).

Q4: The published Drop-seq data I download is not FASTQ format, but .sra format. Also I only get a single .sra file. Where is my barcode and reads FASTQ file?
A: If you download the Drop-seq data we mentioned in Quick Start section from SRA, what you get is a single .sra file. You can transform and split the .sra file you download to barcode file and reads file in FASTQ format with free software called fastq-dump (from package SRA toolkit http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software). After you successfully installed SRA toolkit, Type command line:

```
$ fastq-dump.2.2.2 --split-files SRR1853178.sra
```

Then the .sra file will be split to 2 FASTQ files, named as SRR1853178_1.fastq and SRR1853178_2.fastq. SRR1853178_1.fastq is barcode file and SRR1853178_2.fastq is reads file.
For other published Drop-seq data, you can read their description of data transform their published data to barcode.fastq and read.fastq for Dr.seq input. Note that sequences in barcode file should be composed with cell barcodes and UMIs. For each sequence, cell barcode part should locate before UMI part. For example, if the barcode in your fastq is AAAAAAAAAAAACCCCCCCC, then the cell_barcode of this reads is A*12 and the UMI is C*8. The length of cell barcode and UMI can be specified by two parameters, "cell_barcode_length" and "umi_length".

Q5: I use Dr.seq to process my Drop-seq data with all default parameters, but the default parameter seems not suitable for my data. What should I do now?
A: To get initial sense of a new Drop-seq data, we suggest using Dr.seq with all default parameter (for example, simple mode). And after you get first feedback from Dr.seq output, you can specify the

parameter and make Dr.seq suitable for your data. This time you can use aligned reads file in SAM format from the result of last run. To do this you need to input the location of aligned SAM file for "reads_file" in configure file (eg. "reads_file = /home/user/drseq1/mapping/defaultrun.sam"). The aligned SAM file is in the mapping/ folder in Dr.seq output and named as outname.sam. The original barcode file in FASTQ format should also be inputted. Use aligned reads file for Dr.seq can save you more than 80% of total time of last run.

Q6: There are too many parameters in standard mode, its too complicated for me to handle them.
A: Every parameter has its unique function and is required, but only some of them have great influence on the performance of Dr.seq. For example, you can change "select_cell_measure = 2" if you don't like covered gene number to be the cutoff of STAMP barcodes. You can set "remove_non_dup_cell = 0" to include all cells if you don't believe low duplicate rate barcodes comes from bias. You can use DBSCAN for clustering if you don't like our novel method for clustering.
If you don't like change parameter in a configure file and do want to get feedback with "clicking a bottom", we provide simple mode for you guys.

Q7: I know there are 2000 cells in my Drop-seq sample, but from the result of default Dr.seq, you only gave me 200 cells. How can I get my cells back?
A: By default we select STAMP barcodes with more than 1000 genes have reads covered. If your Drop-seq sample doesn't have enough reads coverage, this cutoff may be too stringent for your sample. You can change the parameter "covergncluster" to get more or less STAMP barcodes. For those who want to get every cell back, we provide another STAMP selection method, that is, select STAMPs with topN highest reads count (UMI count). You can set "select_cell_measure = 2" and "topumicellnumber = 2000" if you put approximately 2000 real cells in your Drop-seq sample.

Q8: Why I get so many clusters from Dr.seq clustering result?
A: We designed a method to use kmeans + Gap statistics to cluster selected cells. To make there is enough cluster for kmeans to assign to small cell types, our method will generate slightly more cluster number than it suppose to be. We test the performance of our method on mouse retina cell Drop-seq data and simulated data, both gives satisfying result (see our paper). Users can combine nearby cluster from our provided clustering result, using optional kmeans+Gap statistics+maxSE to decide the number of cluster or using DBSCAN as clustering method if you don't like the performance of our method.

Q9: I saw "error" in the analysis step, but Dr.seq still processed and "say" It's DONE.

```
Epoch: Iteration #100 error is: 16.5179010356764
Epoch: Iteration #200 error is: 1.47606465146746
Epoch: Iteration #300 error is: 0.969390101026092
Epoch: Iteration #400 error is: 0.80049049634396
Epoch: Iteration #500 error is: 0.6741274422159
Epoch: Iteration #600 error is: 0.621842871094754
Epoch: Iteration #700 error is: 0.600730020015712
Epoch: Iteration #800 error is: 0.583418889593835
Epoch: Iteration #900 error is: 0.569868811419969
Epoch: Iteration #1000 error is: 0.559695999785545
Step4 analysis QC DONE
```

A: Here error is doesn't means bug or crash down. It's mathematics error calculated by our dimensional reduction method: t-SNE.
There's an easy way for you to check whether Dr.seq successfully generate the result. You can check summary/plots/ and summary/results/ folder for every results mentioned in our output list (see manual section for output list).

Q10: I don't have pdflatex, and I don't have root permission to install that. How can I get Dr.seq result?
A: Dr.seq will still process and generate all results if lack of pdflatex, you can check summary/plots/ folder and summary/results/ folder for everything you want. Also, if you do want the summary report, you can copy the whole summary/plots/ folder to another machine with pdflatex installed and type:

```
$ pdflatex outname_summary.tex
```

In the plots/ folder to generate the summary report.

Q11: My sample doesn't show clear bimodal about duplicate rate and I don't believe low duplicate rate represent bias.

A: In the STAMP selection step in the original Drop-seq paper (Macosko, et al., 2015), they use topN cell barcodes with highest reads count, which contains almost NO cells with low duplicate rate. If you don't like our method of removing low duplicate rate cells, you can turn off this function by set "remove_non_dup_cell = 0".

Q12: Dr.seq also provide PCA table and paired wise correlation table for all selected STAMP barcodes. What's the usage of these two optional outputs?

A: We provide these two table incase some users don't like our default workflow (PCA + t-SNE + kmeans + Gap stat) and do prefer PCA analysis only or distance correlation. So we provided these two optional outputs for those users to conduct further analysis based on their preferred intermediate results.