Manual of Dr.seq

Dr.seq is a QC and analysis pipeline for Drop-seq data. By applying this pipeline, Dr.seq takes two sequencing files 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 detailed manual about installation and usage of Dr.seq. See FAQ section if this section doesn't make you satisfied.

Changelog

2015.12.18 V1.0.2: The first released version, which generates the results of the paper.

Installation of Dr.seq

Prerequisite

1. python2.7
2. R (version >= 2.14.1)
3. Dr.seq will generate a summary QC report if you have pdflatex installed, otherwise you only get a package of QC plots and analysis results in the summary.

Install Dr.seq
Dr.seq uses Python's distutils tools for source installations. To install a source distribution of Dr.seq, unpack the distribution zip file and open up a command terminal. Go to the directory where you unpacked Dr.seq, and simply run the install script (We provide an example for installation in quick start section for users who feel confusing about the following description.):

$ python setup.py install
By default, the script will install python library and executable codes globally, which means you need to be root or administrator of the machine so as to complete the installation. Please contact the administrator of that machine if you want their help. If you need to provide a nonstandard install prefix, or any other nonstandard options, you can provide many command line options to the install script. Use the --help option to see a brief list of available options.

$ python setup.py install --prefix /home/drseq
If you are not root user, you will be mentioned "permission denied" when you tried to install Dr.seq globally. Then you can use --prefix parameter to install Dr.seq to any directory you have write permission, you might need to add the install location to your PYTHONPATH and PATH environment variables. The process for doing this varies on each platform, but the general concept is the same across platforms.

Setup environment variable
PYTHONPATH
To set up your PYTHONPATH environment variable, you'll need to add the value PREFIX/lib/pythonX.Y/site-packages to your existing PYTHONPATH. In this value, X.Y stands for the major–minor version of Python you are using (in Dr.seq it should be 2.7; you can find this with sys.version[:3] from a Python command line). PREFIX is the install prefix where you installed Dr.seq. If you did not specify a prefix on the command line, Dr.seq will be installed using Python's sys.prefix value.
For example, if you specify the parameter "--prefix /home/drseq", you have to add /home/drseq/lib/python2.7/site-packages to your PYTHONPATH (below is an example, DON'T add space near the equals sign).
$ export PYTHONPATH=/home/drseq/lib/python2.7/site-packages:$PYTHONPATH

PATH

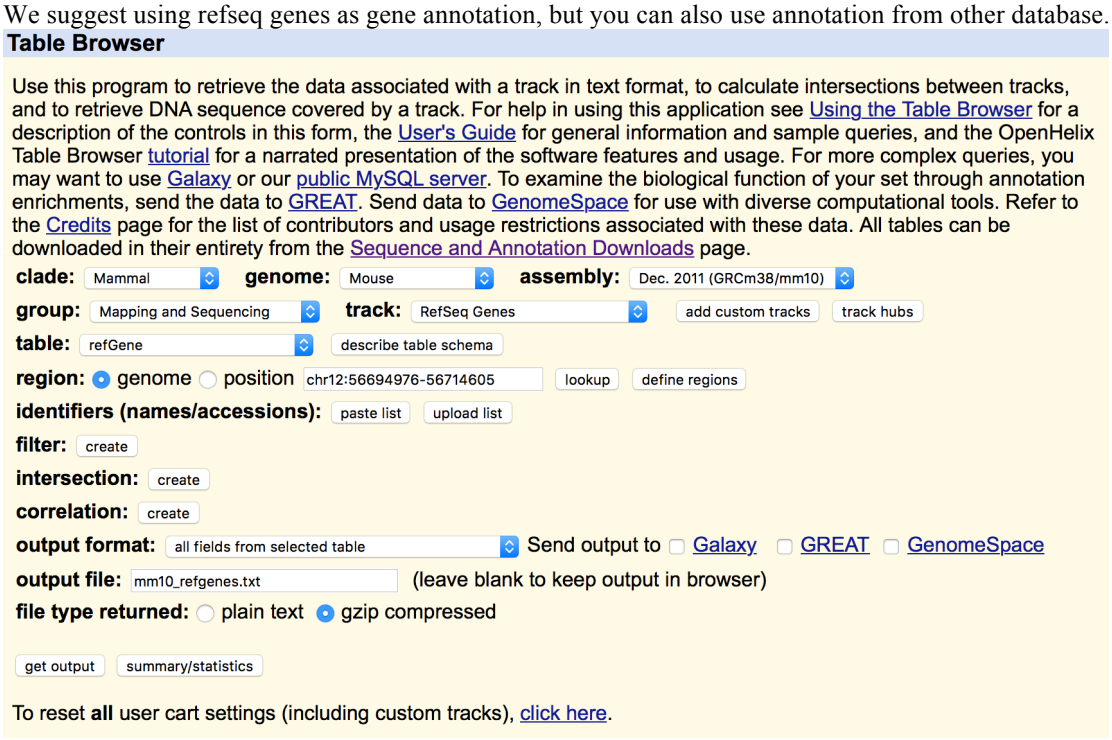
Just like your PYTHONPATH, you'll also need to add a new value to your PATH environment variable so that you can use the Dr.seq command line directly. Unlike the PYTHONPATH value, however, this time you'll need to add PREFIX/bin to your PATH environment variable. The process for updating this is the same as described above for the PYTHONPATH variable:
$ export PATH=/home/drseq/bin:$PATH

To check your default PATH and PYTHONPATH, you can type:
echo $PATH
echo $PYTHONPATH

Automatically export PATH and PYTHONPATH
Feel bothering export environment variable every time?
On Linux, using bash, you can include the new value in my PYTHONPATH by adding these lines
export PATH=/home/drseq/bin:$PATH
export PYTHONPATH=/home/drseq/lib/python2.7/site-packages:$PYTHONPATH
to your ~/.bashrc (or ~/.bash_profile). Then you export environment automatically and don't need to export environment next time.

Now you have Dr.seq installed on your computer, type "Dr.seq --help" to check your installation.

Prepare gene annotation file
To conduct Dr.seq, a gene annotation file is required for the information of the location, name and composition of each gene. Dr.seq requires gene annotation file in full information format downloaded from UCSC genome browser http://genome.ucsc.edu/cgi-bin/hgTables. Below is an example for downloading gene annotation file from UCSC genome browser in mm10 genome version.
We suggest using refseq genes as gene annotation, but you can also use annotation from other database.



Prepare alignment software and build index

After you installed Dr.seq on your computer, you can handle Drop-seq data with barcode file in FASTQ format and reads file in SAM format (reads file is pre-aligned to corresponded genome by any aligner). If you want to use the full function of Dr.seq (including alignment step, that is, users can just input raw barcode file and raw reads file in FASTQ format and get QC and analysis results back), you should install alignment software in your default PATH and build index yourself.

In Dr.seq we accept STAR (default) and bowtie2 as aligner.

We provide download link for executable alignment software and mapping index to save time and effort, see the following description and download section.

1. Install STAR and build STAR index

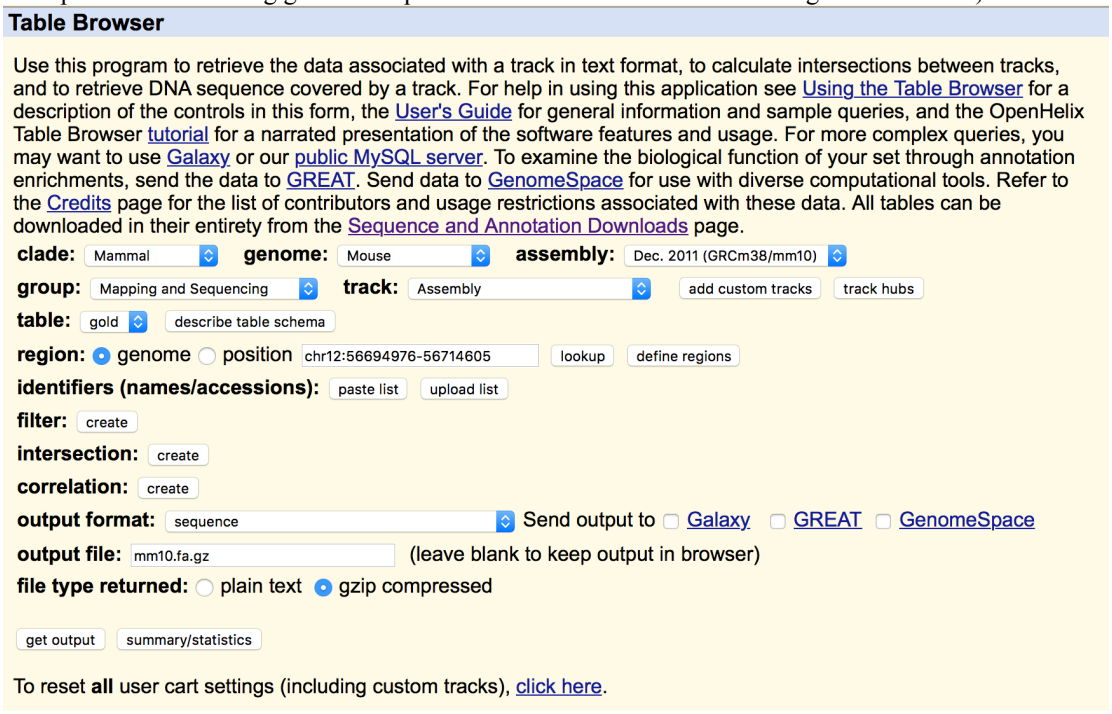
By default we use STAR (version >= 2.5.0) as alignment software because of its speed and performance, but STAR also consume great memory size. To use STAR for alignment in Dr.seq, first you should make sure the memory of your server is greater than 40G. Dr.seq will check the total memory of your server before alignment step if you chose STAR as alignment software (see the parameter description section for corresponded parameter: checkmem). We don't suggest users to run STAR on a Mac.

To install STAR on you server, first you should download STAR package and install STAR according to their manual (github for STAR: https://github.com/alexdobin/STAR)

If you don't want to or have trouble install STAR, we provide download link for executable STAR compiled in linux_x86_64 and macos_x86_64 system. See the download section of our webpage.

After you successfully install STAR or download executable STAR, the next step is to build STAR index. If you don't want to or have trouble build STAR index, we provide download link for STAR index in different species and genome versions, including hg38, hg19, mm10, mm9. See the download section of our webpage.

To build STAR index first you have to download genome assembly file (regularly its .fa file, you can easily download it from UCSC genome browser http://genome.ucsc.edu/cgi-bin/hgTables, below is an example for downloading genome sequence in FASTA/fa format in mm10 genome version)



After you download genome .fa file, what you need is to run a STAR build command to transform your .fa file to an index package (note that to you have to create the directory yourself before you build STAR index).

```
$ mkdir /home/user/STAR_index
$ STAR --runMode genomeGenerate --genomeDir /home/user/STAR_index
--genomeFastaFiles /home/data/mm10.fa --runThreadN 8
```

--runMode genomeGenerate: choose STAR build-index mode
--genomeDir /home/user/STAR_index: name of directory you build STAR index in,
--genomeFastaFiles /home/data/mm10.fa: the genome .fa file you just download from UCSC

--runThreadN 8: how many threads you want to use to generate STAR index, it's a speed up parameter

After you finish index building, you can use STAR as aligner in Dr.seq simple mode (see the example in Quick Start section for the usage of simple mode)

```
$ Drseq.py simple -b SRR1853178_1.fastq -r SRR1853178_2.fastq -n
GSM1626793_mouse_retina1 -g /home/user/annotation/
mm10_refgenes.txt --maptool STAR --mapindex /home/user/STAR_index
```

Note that "STAR" is case sensitive in --maptool parameter. In another word, Dr.seq will exit and let you input correct alignment software if you want to conduct alignment step and input "--maptool star" as a parameter. The directory in which you generate STAR index is just the parameter of --mapindex Here the parameter -g, --mapindex should be inputed with absolute path (type pwd in command line to get the absolute path of your current directory).

2. Install bowtie2 and build bowtie2 index

If your server/computer don't have more than 40G memory (for example, you want to run Dr.seq on a Mac), you can use bowtie2 (version 2.2.6) as an aligner instead, because bowtie2 consume less memory (< 3G memory for human genome) though slower. And that's also the reason we use bowtie2 for Quick Start (another reason is that the index of bowtie2 is smaller and take less time to download).

You may already get familiar with in the Quick Start section. Similar like STAR, you can download and compile bowtie2 according to the manual from bowtie2 website (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml).

If you don't want to or have trouble install bowtie2 and build bowtie2 index, we provide download link for executable bowtie2 compiled in linux_x86_64 and macos_x86_64 system and bowtie2 index in different species and genome version. See the download section of our webpage.

Bowtie2 has its own index. After you finish installing bowtie2, you can use bowtie2-build command to build bowtie2 index.

```
$ mkdir /home/user/bowtie2_index
$ cd  /home/user/bowtie2_index
$ bowtie2-build mm10.fa mm10
```

After you finish index building, you can use bowtie2 as aligner in Dr.seq simple mode (see the example in Quick Start section for the usage of simple mode)

```
$ Drseq.py simple -b SRR1853178_1.fastq -r SRR1853178_2.fastq -n
GSM1626793_mouse_retina1 -g /home/user/annotation/
mm10_refgenes.txt --maptool STAR --mapindex /home/user/
bowtie2_index/mm10
```

--mapindex /home/user/bowtie2_index/mm10: absolute path of bowtie2 index. Note that the genome version of the index should correspond 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.

Install pdflatex

For mac user, you can get download "MacTex" from http://www.tug.org/mactex/. After downloading the package MacTex.pkg from http://tug.org/cgi-bin/mactex-download/MacTeX.pkg, you just double click to install MacTex and get the pdflatex.

For linux user, you can type

```
$ apt-get install texlive-all
```

to install pdflatex on your server/computer.

Note that the installation of pdflatex on both mac and linux requires root privilege.

Usage of Dr.seq

Standard mode and simple mode

Dr.seq provides two modes, you can edit and design a parameter complex suitable for your data with standard mode but you need to generate a configure file and edit it before you start Dr.seq. Simple mode is designed for convenience and quick start. With this mode, you can run Dr.seq with a simple command line, but you can only edit major parameters with this mode.

1. Simple mode

You may get familiar with simple mode in the quick start section. In simple mode you only need to specify input data, output name, mapping tool, annotation and mapping index with a command line and Dr.seq will generate QC and analysis reports with all default parameters.

All parameters in simple mode are also in standard mode and will be described below

```
usage: Drseq.py simple [-h] -b BARCODE -r READS -n NAME
                       [--cellbarcodelength CBL] [--umilength UMIL] [-g GA]
                       [--maptool {STAR,bowtie2}] [--checkmem {0,1}]
                       [--mapindex MAPINDEX] [--thread P] [-f] [--clean]
                       [--select_cell_measure {1,2}]
                       [--remove_low_dup_cell {0,1}]
```

Type

```
$ Drseq.py simple -h
```

for detailed description of all major parameters of simple mode.

2. Standard mode

In standard mode you can edit all parameters you want and make Dr.seq suitable for your data.
First you should generate a configure file according to the template configure file (we will describe how to find and edit template configure file below) by typing:

```
$ Drseq.py gen your_config_name.conf
```

Now you have generated a configure file (named your_config_name.conf in the example) according to the template configure file. If you open the configure file you just generated you will find all changeable parameters listed with the pattern “parameter = value” (for example, barcode_file = SRR1853178_1.fastq) and corresponded description in the bottom of each panel (starts with “#”).

We use a python package called “ConfigParser” to parse configure file, which required specific pattern of input configure file. So user have certain caution when editing configure file:

1. Don't edit the description lines (the lines start with “#”), keep these lines start with “#”
2. Keep the pattern “parameter = value” for parameter lines. Don’t remove value for any parameter line if you don’t have specific value for it. And keep the space next to the equal sign.

Now you can edit your configure file and make it suitable for your Drop-seq data (If you don’t know the suitable parameter at first time, you can run Dr.seq simple mode with only major parameters changed and edit other parameters according to the performance of Dr.seq output, then you can re-run Dr.seq with updated parameters, the re-run of Dr.seq only take 20% of total time, see FAQ section).

After you finish configure file editing, you can use standard mode to run Dr.seq with your edited configure file input.

```
$ Drseq.py run -c your_edited_config.conf -f --clean
```

-c your_edited_config.conf: name of your input configure file, the configure file should be generated from Drseq.py gen mode or copy from other Drseq run. The major parameters (parameter with “[required]” in the description line) including input file location, gene annotation file location, alignment software and mapping index corresponded to the software should be specified by user.
-f: force_overwrite, Dr.seq will overwrite output result if the output folder already exist when -f is added, otherwise Dr.seq will exit.
--clean: Dr.seq will remove intermediate result if --clean is added.

3. DIY template configure file to save your specific parameters

If it's too complicated for you to edit configure file for same parameters for each Dr.seq run. You can edit the template configure file in the Dr.seq package before you install Dr.seq, to save your changes for some major parameters (for example, edit gene annotation file, alignment software, mapping index and other parameters) in the template configure file. If you have already installed Dr.seq, you can edit the template configure file and install again.

First you should find the template configure file. It locates in the "lib/Config/" folder of Dr.seq package and named as "Drseq_template.conf" (do NOT change the file name of the template configure file).

After you find the template configure file, you can edit it to change some commonly used parameters. For example if you change parameter "gene_annotation = /data/mm10_refgenes.txt" and then re-install Dr.seq, the configure file you generate next time will always has "gene_annotation = /data/mm10_refgenes.txt" and you don't need to edit it again. Generally we suggest editing gene_annotation, mapping_software_main and mapindex. You can also change cell_barcode_length if your barcode looks different from the published Drop-seq data (12bp for cell barcode, 8bp for UMI). The way you edit template configure file is similar like editing generated configure file.

Both **standard mode** and **simple mode** can inherit your editing on the template configure file.

4. Full parameter description

Below is the description of all changeable parameters. The value after equal sign is the default (suggested) value of the parameter. [required] means this parameter should be specified for different samples.

A more convenient way is to read the same version in configure file while you are editing configure file and specify parameters.

[General]
In General panel we describe major parameters of Dr.seq
barcode_file =
# [required]barcode_file : Fastq file only, by default, every barcode in fastq file consist of 20 bp including 12bp cell barcode(1-12bp) and 8bp UMI(13-20bp). Cell barcode should locate before UMI, the length of cell barcode and UMI can be defined below
cell_barcode_length = 12
umi_length = 8
# cell_barcode_length/umi_length: Length of cell barcode and UMI, (see the annotation of barcoe_fastq). By default they are 12,8, respectively
reads_file =
[required]reads_file: Accept raw sequencing file (fastq) or aligned file(sam), file type is specified by extension. (regard as raw file and add mapping step if .fastq, regard as aligned file and skip alignment step if .sam), sam file should be with header
outputdirectory =
[required]outputdirectory: (absolute path) Directory for all result. Default is current dir "." if user left it blank , but not recommended
outname =
# [required]outname: Name of all your output results, your results will looks like outname.pdf, outname.txt
gene_annotation = /home/user/annotation/mm10_refgenes.txt
# [required]gene_annotation: (absolute path) Gene model in full text format, download from UCSC genome browser, eg: /yourfolder/mm10_refgenes.txt (absolute path, refseq version recommended)
png_for_dot = 0
# png_for_dot: Only works for dotplots in individual cell QC. Set 1 to plot dotplots in individual cell QC in png format, otherwise they will be plotted in pdf format. The format of other plots is fixed to pdf.

```
#     Make sure your Rscript is able to generate PNG plots if you turn on this function.

[Step1_Mapping]
# In Mapping panel we describe parameters related to the alignment of reads file
mapping_software_main = STAR
# [required if reads file is FASTQ format]mapping_software_main: name of your mapping software, choose from STAR and bowtie2 (case sensitive), STAR (bowtie2) should be installed in your default PATH
checkmem = 1
# checkmem: (only take effect when mapping_software_main = STAR) Dr.seq will check your total memory to make sure its greater 40G if you choose STAR as mapping tool. We don't suggest running STAR on Mac. You can turn off (set 0) this function if you prefer STAR regardless of your memory (which may cause crash down of your computer)
mapping_p = 8
# mapping_p: Number of alignment threads to launch
mapindex = /home/user/STAR_index
# [required if reads file is FASTQ format]mapindex: Mapping index of your alignment tool, absolute path,
#     Mapping index should be built before you run this pipeline, note that STAR and bowtie2 use different index type (see STAR/bowtie2 document for more details.).
#     For STAR, mapindex should be the absolute path of the folder you built STAR index, this parameter will be directly used as the mapping index parameter of STAR
#     eg: /mnt/Storage3/mapping_idnex/mm10.star
#     For bowtie2, mapindex should be absolute path of index filename prefix (minus trailing .X.bt2).this parameter will be directly used as the mapping index parameter of bowtie2
#     eg: /mnt/Storage3/mapping_index/mm10.bowtie2/mm10  (then under your folder /mnt/Storage3/mapping_index/mm10.bowtie2/  there should be mm10.1.bt2, mm10.rev.1.bt2 ....  ),
q30filter = 1
# q30filter: Use q30 criteria (Phred quality scores) to filter reads in samfile, set this parameter to 1(default) to turn on this option, set 0 to turn off

[Step2_ExpMat]
# In ExpMat panel we describe parameters related to generating expression matrix
filterttsdistance = 0
# filterttsdistance: Whether discard reads far away from transcript terminal site (TTS), set 1 to turn on this function, default is 0 (not use).
ttsdistance = 400
# ttsdistance: Default filter distance is 400bp (discard reads 400bp away from TTS), only take effect when filterttsdistance = 1 is set
covergncutoff = 100
# covergncutoff: Only plot cell barcodes with more than 100 (default) genes covered in the dotplots in individual-cell QC step.
#     Note that this parameter is only for visualization of QC report and convenience of following analysis, because too many cell_barcodes will influence users' interpretation of the QC report.
#     Users can change this parameter when the drop-seq sample doesn’t have enough coverage/read depth
#     To determine "STAMP barcodes" from cell barcodes for clustering, users can change parameter in [Step4_Analysis] "covergncluster"
duplicate_measure = 1
# duplicate_measure: method to consider duplicate reads in each cell barcodes when generate expression matrix,
#     1(default): combine duplicate reads with same UMI and same genomic location (position and strand)
#     2: only consider UMI (combine duplicate reads with same UMI)
#     3: only consider genomic location only (combine duplicate reads with same genomic location)
#     0: do not remove (combine) duplicate reads (keep all duplicate reads, skip combination step)
umidis1 = 0
# umidis1: only take effect when duplicate_measure = 1 or 2, ignore this parameter if duplicate_measure = 0 or 3
```

```
#     Set 1: if two reads from same cell barcode have same genomic location (chrom,position,strand) and their UMI distance <= 1, regard them as duplicate reads and combine them
#     Set 0 (default): regard them as duplicate reads only when their UMI distance = 0 and have genomic location
#     By default this function is turned off

[Step3_QC]
# In QC panel we describe parameters related to quality control and STAMP barcodes selection
select_cell_measure = 1
# select_cell_measure: Method to select real cells from cell_barcodes, choose from 1 or 2
#     1 (default): Cell_barcodes with more than 1000 genes covered are selected as real cells for following analysis including dimentional reduction and clustering, cutoff of covered gene number (1000, default) is determined in parameter "covergncluster"
#     2: Top 1000 cell_harcodes with highest UMI count will be selected as real cells, number of highest UMI cell (1000, default) is determined in parameter "topumicellnumber". Suitable for Drop-seq sample with known cell number
covergncluster = 1000
# covergncluster: cell_barcodes with more than 1000 (default) genes covered will be selected as STAMP barcodes for clustering
#     This parameter takes effect only when "select_cell_measure" is set to 1
topumicellnumber = 1000
# topumicellnumber: Top 1000(default) cell_barcodes with highest UMI count will be selected as STAMP barcodes for clustering
#     This parameter takes effect only when "select_cell_measure" is set to 2
remove_non_dup_cell = 1
# remove_non_dup_cell: A group of cell barcodes have almost no duplicate reads, which show clearly different pattern from STAMP barcodes, we set an option to remove these cell_barcodes for analysis step.
#     Choose from 0 or 1, set 1 (default) to turn on this function (remove low duplicate rate cell_barcodes), set 0 to turn off (keep low duplicate rate cell_barcodes)
#     Removing low duplicate cells maybe not so effective when sequencing depth is not enough.
non_dup_cutoff = 0.1
# non_dup_cutoff: Only take effect when remove_non_dup_cell is turned on. Cutoff of low-duplicate cell_barcode (default is 0.1), a cell_barcode will be defined as low-duplicate cell_barcode if its duplicate rate < 0.1. Duplicate rate for each cell_barcode is defined as (#total_reads - #UMI)/#total_reads


[Step4_Analysis]
# In analysis panel we describe parameters related to dimensional reduction, highly variable gene selection and clustering
highvarz = 1.64
# highvarz: Cutoff for high variance selection. In the first step we divide all genes to 20 groups based on average expression level across all individual cells. Then in each group we select genes whose z-normalized CV (var/mean) >= 1.64 (default, corresponding to pvalue =0.05)
selectpccumvar = 0.5
# selectpccumvar: Select topN PC until they explain 50% (default is 0.5,  eg. 0.3 for 30% variance) of total variance. Then conduct 2 dimensional t-SNE based on selected PCs
pctable = 1
# pctable: choose from 0(default) and 1, you can turn on this function(set 1) to output a 2 column table for PC1 v.s PC2
#     This paramter is designed for users who prefer and conduct following analysis based on PCA output
#     0(default): Do not output PCA result, only output PCA + 2 dimensional t-SNE result (see document for more details)
#     1: Output PC1 v.s. PC2 table in addition to t-SNE result.
cortable = 1
# cortable: Choose from 0(default) and 1, you can turn on this function(set 1) to output a cell to cell correlation matrix
#     Correlation is based on log scale TPM (transcript per million reads).
```

# User can turn on this function to generate cell to cell correlation matrix for custom analysis
clustering_method = 1
# clustering_method: Method for cluster cells based on t-SNE output
# Choose from 1(default), 2, 3 and 4.
# 1(default): k-means, use Gap statistics followed by our "first stable Gap" method to determine k
# 2: k-means, but use Gap statistics followed by traditional "Tibs2001SEmax" method to determine k (Tibshirani et al (2001))
# 3: k-means with custom determined k, k value is defined in following parameter "custom_k", this option is designed for users who know the number of subgroup of the drop-seq sample
# 4: (make sure your R enviroment has library "fpc" installed) Use dbscan as clustering method, the (eps) parameter is defined in following parameter "custom_d"
maxknum = 100
# maxknum: Maximum k number for gap statistics, only take effect when clustering_method = 1 or 2
custom_k = 5
# custom_k: Only take effect when clustering_method = 3, cells will be clustered to N group based on t-SNE result according to user determined k
custom_d = 2
# custom_d: Only take effect when clustering_method = 4, refer to the parameter "eps"(Reachability distance, see Ester et al. (1996)) of dbscan. By default we set it to 2 according to the orginal Drop-seq paper, but it varies a lot between different datasets.
rdnumber = 1007
# rdnumber: Set initial random number to keep your result reproducible

5. Output files

All output files will be generated in summary/ folder. The QC plots will be generated in summary/plots/ folder and results will be generated in summary/results/ folder

1.QC report (pdf format, see demo.pdf in the package) [outname_summary.pdf]
The QC report will be generated if you have pdflatex installed on your computer, otherwise all QC plots with ordered name (Figure 1-11) will be generated in the summary/plots/ folder.

2.Expression matrix for STAMP barcodes. [outname_expmat_clustercell.txt]
In Drop-seq experiment there are millions of cell barcodes generated, but most of them don't contain transcriptome though they still contain some garbage RNA sequenced. Only a very little number of cell barcodes correspond with single cell transcriptome. In the expression matrix we only report transcriptome of these cell barcodes (call STAMP barcode). Expression matrix is a big table in plain text format, whose rows represent each gene and columns represent each selected STAMP barcodes.

3.QC measurements for STAMP barcodes [outname_qcmat_clustercell.txt]
Similar like 3 but in this table we only show QC measurements (column) of STAMPs (row). Measurements include mappable reads, uniq reads, covered gene number, intron rate and other information.

4.Dimensional reduction and clustering result for STAMP barcodes [outname_cluster.txt]
In the analysis step we conduct a t-SNE dimensional reduction and clustering with automatically determined cluster number. We print the dimensional reduction result and cluster assignment of each STAMP barcodes to the table for further analysis.

5.Other optional outputs. [outname_pctable.txt, outname_correlation_table.txt]
Dr.seq can also output PCA dimensional result and paired wise correlation matrix of selected STMAPs in case not all users prefer the default t-SNE method.