QC and analysis reports for Drop-seq data : mouse_retina_cells

December 18, 2015

Contents

1 Data description

Table 1 mainly describe the input file and mapping and analysis parameters.

Table 1: Data description

parameter	value
output name	mouse_retina_cell
barcode file(file name only)	SRR1853178_1.fastq
reads file(file name only)	SRR1853178_2.sam
reads file format	SAM
genome version	mm10
cell barcode length	12
UMI length	8
mapping software	STAR
Q30 filter mapped reads	True
remove reads away TTS	False
duplicate rate in each cell	UMI + location
merge UMI ED = 1	False
select STAMPs	1000 covered gene
remove low duplicate rate cell	True
low duplicate rate cutoff	0.1
z-score for highly variable gene	1.64
cumulative variance for selecting PC	50.0%
cluster method	k-means (Gap statistics, first stable)

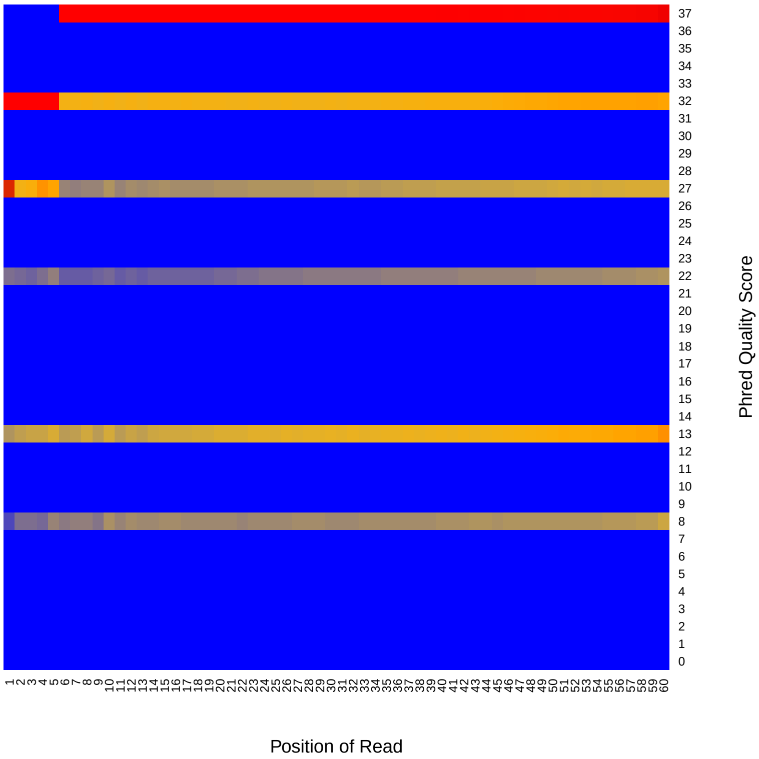
2 Reads level QC

In the reads QC step we measured the quality of sequencing reads, including nucleotide quality and composition. In the reads level QC step and Bulk-cell level QC step we randomly sampled down total aligned reads to 5 million and used a published package called "RseQC" for reference.(Wang, L., Wang, S. and Li, W. (2012))

2.1 Reads quality

Reads quality is one of the basic reads level quality control methods. We plotted the distribution of a widely used Phred Quality Score at every position of sequence. Phred Quality Score is calculate by a python function $ord(Q) - 33$. Color in the heatmap represent frequency of this quality score observed at this position. Red represents higher frequency while blue is lower frequency.

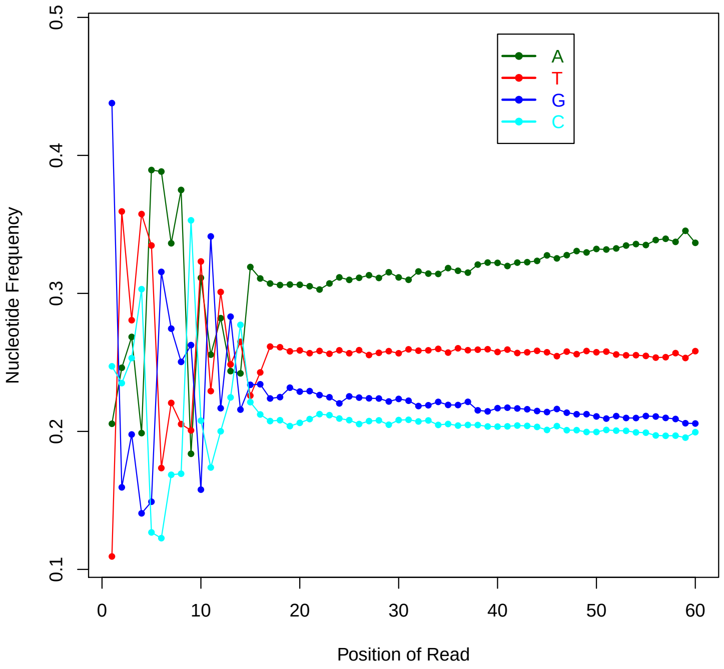
Figure 1: Reads quality



2.2 Reads nucleotide composition

We assessed the nucleotide composition bias of a sample. The proportion of four different nucleotides is calculated at each position of reads. Theoretically four nucleotides have similar proportion at each position of reads. For Drop-seq sample we observed higher A count at the 3'end of reads, because of the 3'end polyA tail generated in sequencing cDNA libaray.

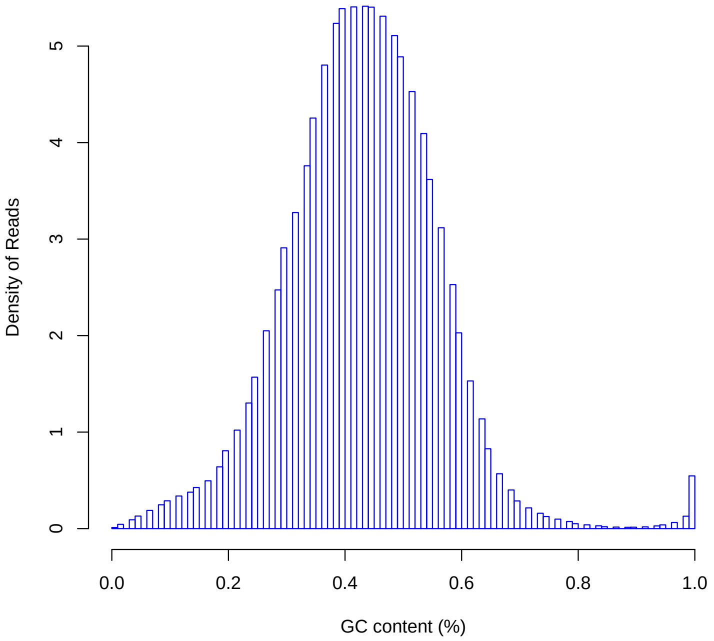
Figure 2: Reads nucleotide composition



2.3 Reads GC content

Distribution of GC content of each read.

Figure 3: Reads GC content



3 Bulk-cell level QC

In the bulk-cell level QC step we measured the performance of total Drop-seq reads. In this step we did't separate cell or remove "empty" cell barcodes, just like treated the sample as bulk RNA-seq sample.

3.1 Reads alignment summary

The following table shows mappability and distribution of total Drop-seq reads. Note that UMI number was calculated by removing duplicate reads (which have identical genomic location, cell barcode and UMI sequences). Mappable reads was after Q30 filtering if Q30 filter function was turned on.
* the percentage was calculated by dividing total reads number
** the percentage was calculated by divding total UMI number

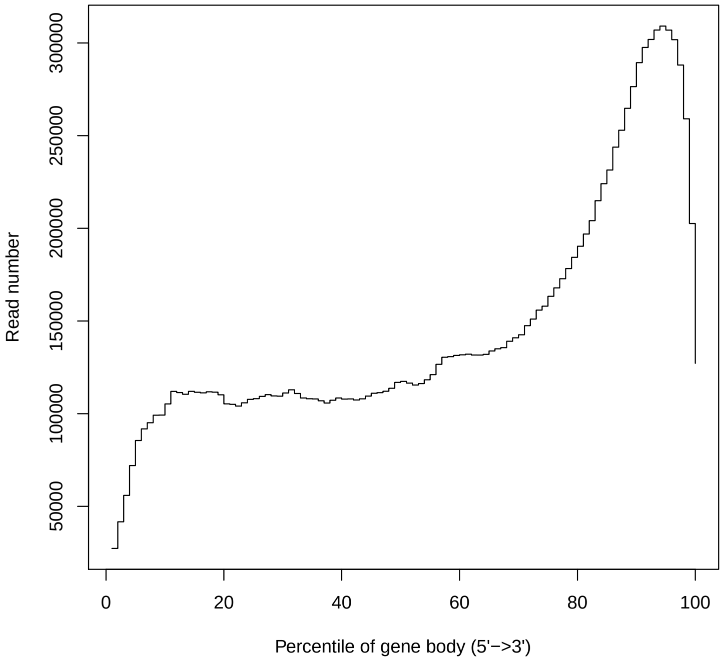
Table 2: Reads alignment summary

genomic region(Category)	reads number
total reads	510,210,716
mappble reads	278,424,189 (54.57%)*
total UMI count	186,631,071 (36.58%)*
CDS exon UMI count	56,970,626 (30.53%)**
3'UTR UMI count	27,735,552 (14.86%)**
5'UTR UMI count	32,383,752 (17.35%)**
intron UMI count	41,007,430 (21.97%)**
intergenic UMI count	28,533,711 (15.28%)**

3.2 Gene body coverage

Aggregate plot of reads coverage on all genes. Theoretically we observe a unimodal(single bell) distribution, but for Drop-seq sample we observed an enrichment at 3'end because of the CEL-seq like protocol used in sequencing cDNA library. (Klein, A.M., et al. (2015))

Figure 4: Gene body coverage



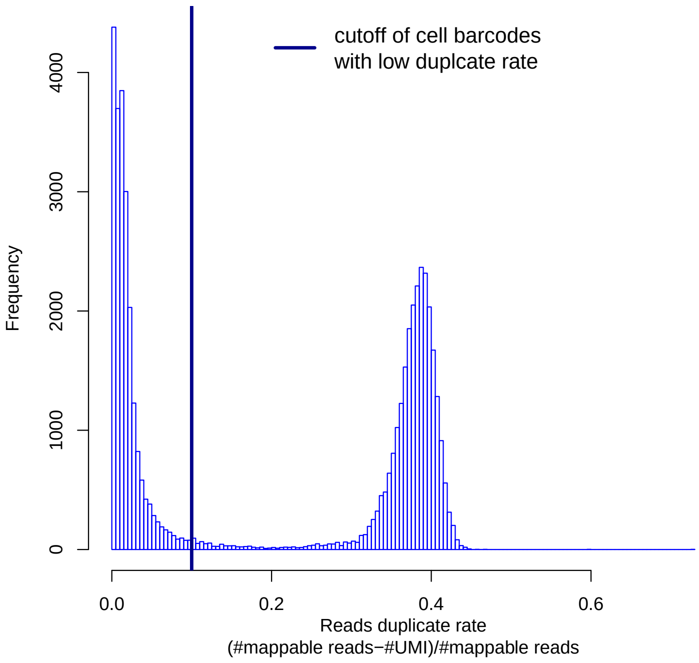
4 Individual-cell level QC

In this step we focused on the quality of individual cell and distinguishing cell barcodes from STAMPs (single-cell transcriptomes attached to microparticles)

4.1 Reads duplicate rate distribution

Drop-seq technology has an innate advantage of detect duplicate reads and amplification bias because of the barcode and UMI information. Here we plotted the distribution of duplicate rate in each cell barcode (though most of cell barcodes don't contain cells, they still have RNA) and observed a bimodal distribution of duplicate rate. We set an option for users to discard cell barcodes with low duplicate rate in following steps. The vertical line represented the cutoff (duplicate rate >= 0.1) of discarding cell barcodes with low duplicate rate.

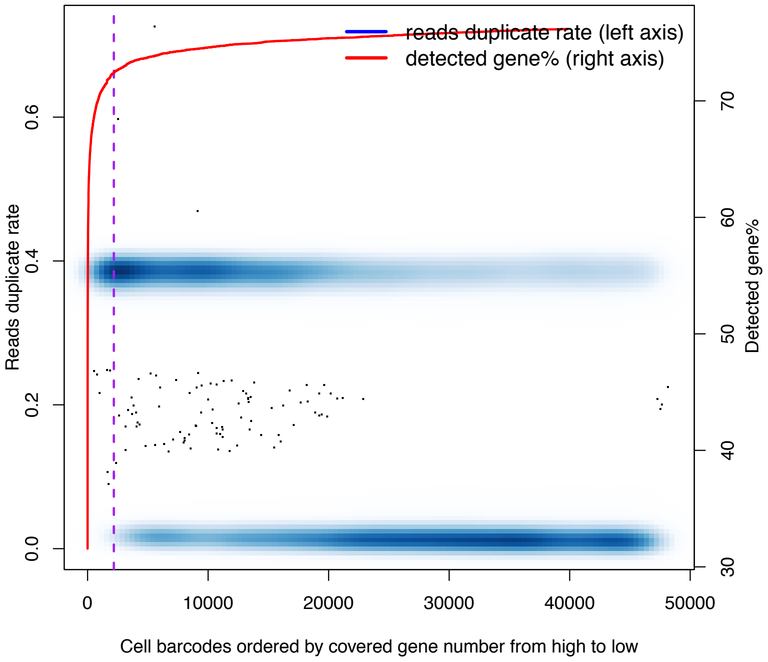
Figure 5: Reads dupliate rate distribution



4.2 Reads duplicate rate vs. cumulative covered gene number

Reads duplicate rate versus cumulative covered gene numbers. Cell barcodes were ranked by the number of covered genes. The duplicate rate (y-axis, left side) was plotted as a function of ranked cell barcode. Red curve represented the number of genes covered by top N cell barcodes (y-axis, right side). N was showed by x-axis.

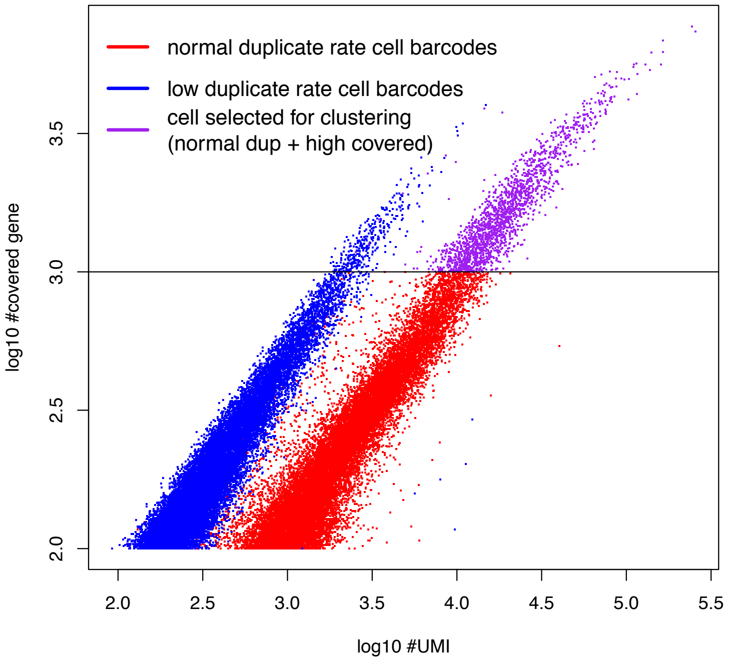
Figure 6: Reads duplicate rate vs. cumulative covered gene number



4.3 UMI vs. covered gene number

Covered gene number was plotted as a function of the number of UMI (i.e. unique read). We observed a clearly different pattern for two groups of cell barcodes with different reads duplicate rate (blue dots versus red and purple dots). Purple dots represented the selected STAMPs for the cell-clustering analysis.Note that we used only STAMPs selected in this step for following analysis. The other cell barcodes were discarded.

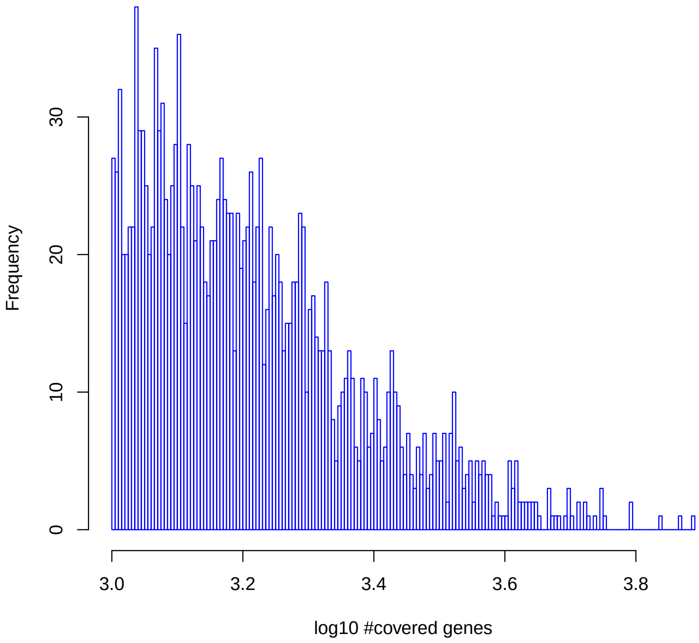
Figure 7: UMI v.s. covered gene number



4.4 Covered gene number distribution

Histogram of covered gene number of selected STAMPs

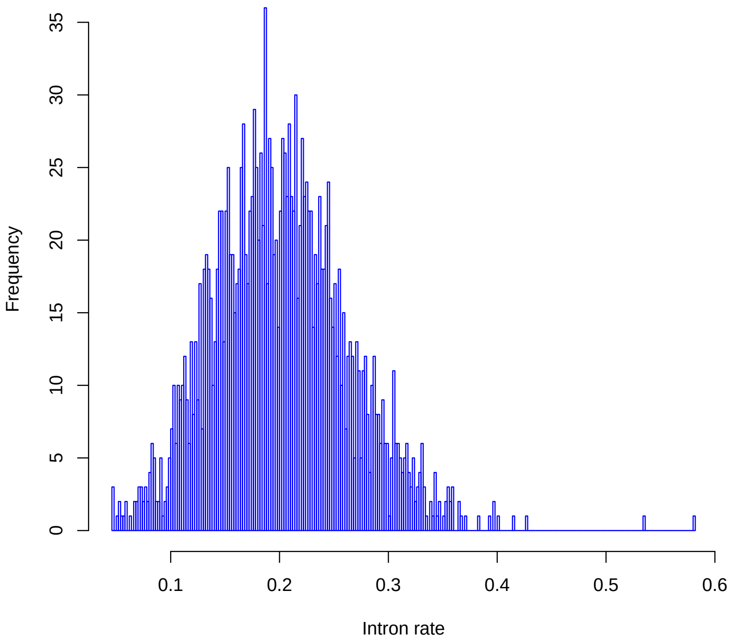
Figure 8: Covered gene number



4.5 Intron rate distribution

Intron rate is a effective method to measure the quality of a RNA-seq sample. We plotted a histogram of intron rate of every STAMP barcodes. Intron rate was defined as $\frac{intron\ reads\ number}{intron+exon\ reads\ number}$

Figure 9: Intron rate distribution



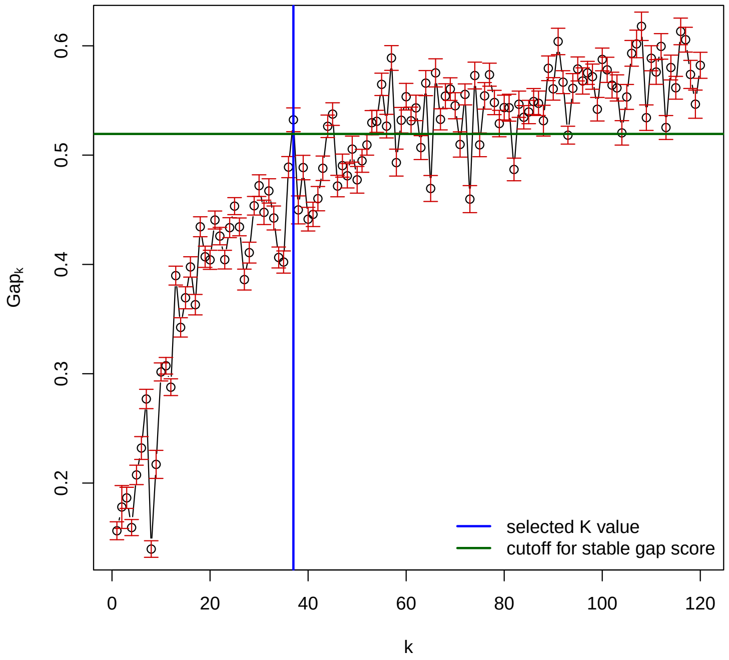
5 Cell-clustering level QC

This step composed by k-means clustering based on t-SNE dimentional reduction result and Gap statistics to determine best k.

5.1 Gap statistics

We conducted a k-means clustering based on t-SNE dimentional reduction output. Gap statistics followed by the "first stable gap" method was performed to determine the best k in k-means clustering (to determine how many groups the data should have).

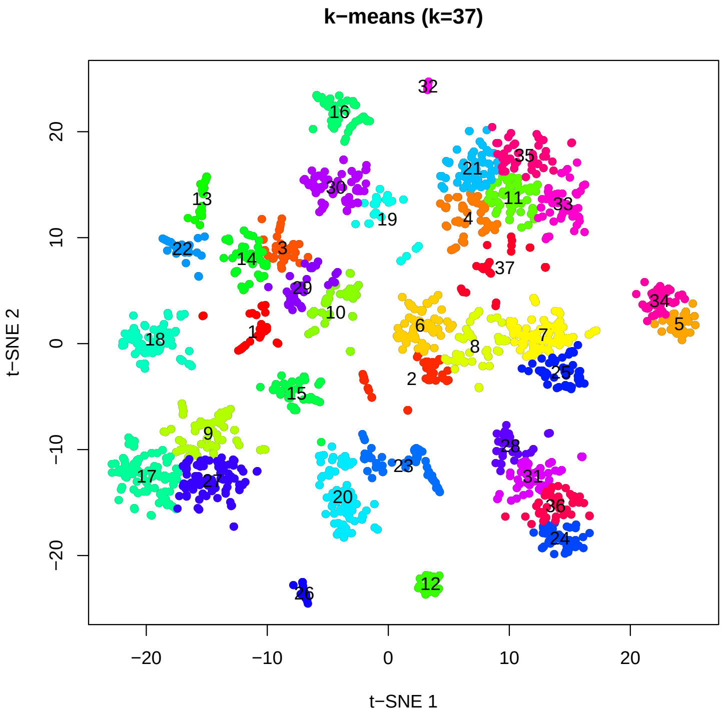
Figure 10: Gap statistics



5.2 Clustering plot

Scatter plot represented visualization of t-SNE dimensional reduction output of selected STAMP barcodes. STAMP barcodes were colored according to the clustering result and cluster numbers were printed in the center of each cluster.

Figure 11: Clustering plot



6 Output list

All output files were described in the following table

Table 3: output list

description	filename
expression matrix for selected STAMPs	mouse_retina_cell_expmat_clustercell.txt
QC measurements for selected STAMPs	mouse_retina_cell_qcmat_clustercell.txt
top2 components of PCA dimentional reduction result	mouse_retina_cell_pctable.txt
pairwise correlation matrix	mouse_retina_cell_correlation_table.txt
t-SNE dimentional reduction and clustering result	mouse_retina_cell_cluster.txt
summary QC report	mouse_retina_cell_summary.pdf