**Figure 1-figure supplement 1A, B**

Codes of analysis RiboD-PETRI and PETRI-seq scRNA-seq sequencing data

1.Since both reads1 and reads2 files contain information about barcode and genome, in order to improve the data utilisation efficiency of reads1 and reads2 files, we pre-processed the reads1 and reads2 files respectively. We used codes to combine the reads1 and reads2 files of scRNA-seq clean data (such as: "{sample}\_R1.fq.gz" and "{sample}\_R2.fq.gz") to mention barcode sequences and genome sequences, respectively, and processed them into four fastq files for downstream splitting of barcode and comparison processes. Make sure all fastq files have names in the form: {sample name}\_{S#}\_L00{#}\_R1\_001.fastq.gz , {sample name}\_{S#}\_L00{#}\_R2\_001.fastq.gz , {sample name}B\_{S#}\_L00{#}\_R1\_001.fastq.gz and {sample name}B\_{S#}\_L00{#}\_R2\_001.fastq.gz.

The codes are:

$cutadapt -g NNNNNNNNAGAATACACGACGCTCTTCCGATCT -o r1.fastq {sample}\_R1.fq.gz

$cutadapt -g NNNNNNNNAGAATACACGACGCTCTTCCGATCT -o r2.fastq {sample}\_R2.fq.gz

$cat r1.fastq | seqkit subseq -r 1:85 > {sample}\_S5\_L001\_R1\_001.fastq

$cat r2.fastq | seqkit subseq -r 80:100 > {sample}\_S5\_L001\_R2\_001.fastq

$cat r2.fastq | seqkit subseq -r 1:85 > {sample}B\_S5\_L001\_R1\_001.fastq

$cat r1.fastq | seqkit subseq -r 80:100 > {sample}B\_S5\_L001\_R2\_001.fastq

$gzip {sample}\_S5\_L001\_R1\_001.fastq

$gzip {sample}\_S5\_L001\_R2\_001.fastq

$gzip {sample}B\_S5\_L001\_R1\_001.fastq

$gzip {sample}B\_S5\_L001\_R2\_001.fastq

2.Create a folder named demo and then in the demo folder create the folders with the sample names : {sample} and {sample}B. Put {sample}\_S5\_L001\_R1\_001.fastq and {sample}\_S5\_L001\_R2\_001.fastq in the folder named {sample}; Put {sample}B\_S5\_L001\_R1\_001.fastq and {sample}B\_S5\_L001\_R2\_001.fastq in the folder named {sample}B.

3.Then we performed single-cell sequencing data analyses according to the processes and scripts of previous articles "Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing". The scripts folder was also been put in "scripts" folder which we made some modifications owing to differences in python versions. The data are all in one lane, so we removed the step of merging the lanes. Then in the "demo" folder, run the following codes:

$python [path]/scripts/sc\_pipeline\_11.py {sample}\_S5 {n\_lanes}

$python [path]/scripts/sc\_pipeline\_11.py {sample}B\_S5 {n\_lanes}

# “sample” is sample name and S number (eg first1000\_S5)

# n\_lanes is the number of sequencing lanes for analysis - if lanes are merged, then the single file should be names with suffix \_L001\_R1\_001.fastq.gz and n\_lanes set to 1. The script will count lanes from 1 to n\_lanes so always start numbering from 1.

# sc\_pipeline\_11 runs fastqc, quality filter, and barcode demultiplexing

For example:

$python path\_to/scripts/sc\_pipeline\_11.py 1214-7\_S5 1

$python path\_to/scripts/sc\_pipeline\_11.py 1214\_7B\_S5 1

4.Look at {sample}\_bc1\_ReadsPerBC.eps, {sample}\_bc1\_kneePlot.eps, {sample}B\_bc1\_ReadsPerBC.eps and {sample}B\_bc1\_kneePlot.eps to determine number of BCs to include in further analysis. In the following analysis step, we only need to generate statistical tables for UMI, genes and cells, so we remove the step of synthesising the matrix. This script generates two files named as ("{sample}\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt" and "{sample}B\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt").

The codes were:

SCRIPT: [path]/scripts/pipeline.sh {sample} {n\_BCs} {fasta} {gtf} {custom\_name}

# n\_BCs is number of BCs to include in further analysis (typically 10000-80000)

# fasta is location and name of fasta for alignment

# gtf is location and name of gtf for feature calling - see example.gtf for example format (specifically, gene names should be indicated by 'name=')

# Custom name is a new name for the sample, corresponding to maybe the gtf used or other specific input of the pipeline. For example, we might analyze the same cells by CDS or by exon and would indicate that in the custom name. Custom name can be the same as sample name if desired.

# pipeline.sh includes a number of cleanup commands at the end. If interested in intermediate files, these can be easily commented out.

For example:

$sh ./scripts/pipeline.sh 1214\_7 10000 Escherichia\_coli\_str\_k\_12\_substr\_mg1655.ASM584v2.dna.toplevel.fa.gz Escherichia\_coli\_str\_k\_12\_substr\_mg1655.ASM584v2.46.gtf 1214\_7\_CDS

$sh ./scripts/pipeline.sh 1214\_7B 10000 Escherichia\_coli\_str\_k\_12\_substr\_mg1655.ASM584v2.dna.toplevel.fa.gz Escherichia\_coli\_str\_k\_12\_substr\_mg1655.ASM584v2.46.gtf 1214\_7B\_CDS

5.The files ("{sample}\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt" and "{sample}B\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt") obtained in step 4 are combined into a matrix containing UMI, gene and barcode information.

The codes were:

$cat {sample}\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt {sample}B\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt > {sample}.txt

$sed -i "s/{sample}B/{sample}/g" {sample}.txt

$sed -i "s/Chromosome://g" {sample}.txt

$awk 'BEGIN{ FS=" ";OFS="+" }{ print $1,$2,$3 }' {sample}.txt > {sample}.2.txt

$cat {sample}.2.txt | sort | uniq > {sample}.3.txt

$awk 'BEGIN{ FS="+";OFS=" " }{ print $1,$3 }' {sample}.3.txt > {sample}.4.txt

$awk 'BEGIN{ FS=" ";OFS="+" }{ print $1,$2 }' {sample}.4.txt > {sample}.5.txt

$cat {sample}.5.txt | sort | uniq -c > {sample}.6.txt

$awk 'BEGIN{ FS=" ";OFS="\t" }{ print $2,$3,$1 }' {sample}.6.txt > {sample}.7.txt

$awk 'BEGIN{ FS="+";OFS="\t" }{ print $1,$2,$3 }' {sample}.7.txt > {sample}.8.txt

$awk 'BEGIN{ FS="\t\t";OFS="\t" }{ print $1,$2,$3 }' {sample}.8.txt > {sample}.9.txt

$awk 'BEGIN{ FS="\t";OFS="+" }{ print $1,$2,$3 }' {sample}.9.txt > {sample}.10.txt

$sed -i '1i\cell+gene+count' {sample}.10.txt

$awk 'BEGIN{ FS="+";OFS="\t" }{ print $1,$2,$3 }' {sample}.10.txt > {sample}.11.txt

$awk 'BEGIN{ FS="\t";OFS="," }{ print $1,$2,$3 }' {sample}.11.txt > {sample}.12.csv

>>>python

>>>import pandas as pd

>>>import numpy as np

>>>table = pd.read\_csv("{sample}.12.csv")

>>>table=pd.pivot\_table(table,index=["cell"],columns=["gene"],values=["count"],fill\_value=0)

>>>df = pd.DataFrame(table)

>>>df.to\_csv('{sample}.csv')

1. Download "{sample}.csv", removed rows and columns that clearly not belong to cells and genes (such as "cell", "16" and so on.) , and count the number of UMI in each cell in the gene expression matrix, and select cells with a UMI count greater than 15 to calculate the median UMI count.

**Figure 1-figure supplement 1C**

1. For UMI/cell analysis, we used {sample}\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt and {sample}B\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt to counted the number of UMIs in each cell using the following codes:

$cat {sample}\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt {sample}B\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt > {sample}.txt

$sed -i "s/{sample}B/{sample}/g" {sample}.txt

$sed -i "s/Chromosome://g" {sample}.txt

$awk 'BEGIN{ FS=" ";OFS="+" }{ print $1,$2,$3}' {sample}.txt > {sample}.2.txt

$cat {sample}.2.txt | sort | uniq > {sample}.3.txt

$sed -i "/+VN+1.6/d" {sample}.3.txt

$sed -i "/+SN+Chro:LN/d" {sample}.3.txt

$sed -i "/+ID+bwa/d" {sample}.3.txt

$sed -i "/ambiguous/d" {sample}.3.txt

$awk 'BEGIN{ FS="+";OFS=" " }{ print $1}' {sample}.3.txt > {sample}.4.txt

$cat {sample}.4.txt | sort | uniq -c > {sample}.5.txt

$mv {sample}.5.txt {sample}.UMI-cell.txt

1. For reads/cell analysis, we used {sample}\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt and {sample}B\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt to counted the number of reads in each cell using the following codes:

$cat {sample}\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt {sample}B\_CDS\_v11\_threshold\_0\_filtered\_mapped\_UMIs.txt > {sample}.txt

$sed -i "s/{sample}B/{sample}/g" {sample}.txt

$sed -i "s/Chromosome://g" {sample}.txt

$awk 'BEGIN{ FS=" ";OFS="+" }{ print $1,$2,$3,$4}' {sample}.txt > {sample}.2.txt

$sed -i "/+VN+1.6/d" {sample}.2.txt

$sed -i "/+SN+Chro:LN/d" {sample}.2.txt

$sed -i "/+ID+bwa/d" {sample}.2.txt

$sed -i "/ambiguous/d" {sample}.2.txt

$awk 'BEGIN{ FS="+";OFS=" " }{ print $1,$4}' {sample}.2.txt > {sample}.3.txt

$awk '{sum[$1]+=$2}END{for(i in sum)print i" "sum[i]}' {sample}.3.txt > {sample}.4.txt

$mv {sample}.4.txt {sample}.reads-cell.txt

1. Calculate the correlation coefficient between the UMI count and reads count of each counted cell.

The codes were :

>f1<-read.table("{sample}1 vs {sample}2.txt")

>f2<-f1 + 1

>f3<-log2(f2)

>plot(f3[,1],f3[,2],xlab='{sample}1',ylab='{sample}2',las=1,mgp=c(1.5,0.5,0),cex.axis=0.8,mai=c(1,1,1,1))

>cor.test(f1.1.1[,1],f1.1.1[,2],method = "pearson")

**Figure 1-figure supplement 1E, F**

1. Firstly, divide the sequencing data into five equal parts, and then select 20%, 40%, 60%, 80%, and 100% of the data for single-cell analysis.
2. Next, we will use the gene expression matrices detected from five sets of data to count the number of UMIs and genes in each cell, and sort the cells from high to low.
3. Select different numbers of cells such 1,000, 5,000 and 10,000 cells to calculate the median UMI or gene count.