

Demonstration of SCRAT analysis

In this example, we demonstrated the major functions of SCRAT using a dataset which contains 230 HEK293T cells and 20 GM12878 cells. The aligned bam files (aligned to hg19) for this example are available at https://github.com/zji90/SCRATdata/tree/master/SCRAT_example_data_bam. Users can also use the “Load example data” function in Step 1 to load this dataset. For readers’ convenience, we also saved the SCRAT summarized features (obtained after performing Step 1 and Step 2 below) of this dataset and provide them at the following web link: https://github.com/zji90/SCRATdata/blob/master/SCRAT_summarized_features_GM12878_HEK293T.txt. If users start the SCRAT analysis from these summarized features, they can skip the procedure described below in Step 1 and Step 2, and use instead the “Upload Summarization Table” function in Step 2 to read in the summarized features (**Fig. 1a**). For instance, one can upload the data in the “Input Summary Table” section using the “Choose Files” button, and read in the data using the “Read in” button after the upload is completed (**Fig. 1b**). Users can also use the “Load example data” function in Step 2 to load the summary table (**Fig. 1c**). The summarized features can be viewed in the “Results” panel (**Fig. 1d**). Then, one can proceed to Step 3 and Step 4.

The screenshot shows the SCRAT web interface at Step 2: Feature summarization. The sidebar on the left has three main sections: 'Previous Step' and 'Next Step' buttons at the top; a section for uploading a summary table with a 'Choose File' button and a 'Read in' button; and a 'Load example data' button. The main panel displays a table of summarized features. The table has columns for 'Feature', 'CV', and several cell lines (GM12878_14.bam, GM12878_17.bam, GM12878_42.bam, GM12878_45.bam, GM12878_53.bam, GM12878_90.bam, GM12878_91.bam). The table shows read counts for various genes across different cell lines. The table is paginated, showing 1 to 10 of 5,650 entries.

Feature	CV	GM12878_14.bam	GM12878_17.bam	GM12878_42.bam	GM12878_45.bam	GM12878_53.bam	GM12878_90.bam	GM12878_91.bam
GENE: ENSG00000199347.1:RNU5E-1	2.34058098383612	0	2.12514679142593	3.65699989402624	0	0	3.98721543530858	0
GENE: ENSG00000207005.1:RNU1-2	2.50189910773427	0	0	0	0	0	0	0
GENE: ENSG00000272426.1:RP11-108M9.6	2.94909443944779	3.82795314147536	0	0	0	0	0	0
GENE: ENSG00000201405.1:Y_RNA	3.0835062669181	0	0	4.59865126918869	2.94699456218395	0	0	0
GENE: ENSG00000117713.13:ARID1A	3.03634388730825	0	0	0	0	0	0	0
GENE: ENSG00000009780.11:FAM76A	2.97234531722972	0	0	0	0	0	0	2.833386
GENE: ENSG00000117758.9:STX12	2.96984342090524	0	0	0	0	4.02596074006307	3.98721543530858	4.276671
GENE: ENSG00000269971.1:RP3-426I6.5	3.02874458975156	0	0	0	0	4.02596074006307	3.98721543530858	4.276671
GENE: ENSG00000158161.11:EYA3	2.35366552870056	0	0	0	0	0	0	0
GENE: ENSG00000126698.6:DNAJC8	3.16664955088	0	0	0	0	0	0	0

Figure 1. User can upload the previously saved summarized features into SCRAT for analysis.

Step 1: Data input and preprocessing

The first step is to input the single-cell data (aligned bam files) into SCRAT. First, one has to select the corresponding reference genome from the “Select Genome” section (**Fig. 2a**). Then, one can upload bam files in the “Input Bam Files” section using the “Choose Files” button, and read in the data using the “Read in” button after the upload is completed (**Fig. 2b**). For this example, one can simply use the “Load example data” function to input the bam files (**Fig. 2c**). By default, SCRAT will filter blacklist regions and exclude samples with total number of reads less than 500 (adjustable by the user). Information about the input data will be shown after they have been read in (**Fig. 2d**). One can proceed to feature summarization using the “Next step” button (**Fig. 2e**).

SCRAT

Step 1: Data input and preprocessing

Step 2: Feature summarization

Step 3: Cell heterogeneity analysis

Step 4: Differential feature analysis

Next Step

To start the analysis, users should have the already aligned bam files. Select the corresponding genome used for alignment and upload the bam files.

Select Genome

hg19 (Human)

To upload a summary table from previous SCRAT session, skip this step and go directly to step 2.

Input Bam Files

Choose File

Browse... 250 files

Upload complete

☒ Filter blacklist

Read in

Load example data

Filter Bam Files

Exclude samples with reads less than

500

Exclude specific samples

250 bam files read in

250 bam files retained

Reads for each bam file:

Show 10 entries

Search:

BAM	Reads	Type
GM12878_14.bam	1515	paired-end
GM12878_17.bam	2974	paired-end
GM12878_42.bam	861	paired-end
GM12878_45.bam	1490	paired-end
GM12878_53.bam	654	paired-end
GM12878_90.bam	673	paired-end
GM12878_100.bam	1632	paired-end
GM12878_119.bam	1058	paired-end
GM12878_132.bam	988	paired-end
GM12878_143.bam	604	paired-end

Showing 1 to 10 of 250 entries

Previous 1 2 3 4 5 ... 25 Next

Figure 2. SCRAT analysis step 1 -- Data input and preprocessing.

Step 2: Feature summarization

The second step is to summarize the input data into different features according to the feature definitions. To demonstrate, we summarized signals in this test dataset using the pre-defined features in SCRAT. In the “Choose Summarizing Method” section, we selected all pre-defined SCRAT features for our analysis (i.e., *Motif*, *ENCODE Cluster*, *Gene* and *Gene set*; **Fig. 3a**).

SCRAT provides rich tuning options for each feature type. The parameters of each feature types can be adjusted in the “Method Details” section of the user interface (**Fig. 3d**). When the example dataset was summarized based on *Motif*, we asked SCRAT to aggregate reads within 100 base pair (bp) flanking region from both sides of the motif sites. When the data were summarized based on *ENCODE Cluster*, we set the cluster number to be 2000. When the data were summarized based on *Gene*, we asked SCRAT to aggregate reads within the 3000 bp upstream and 1000 bp downstream region from the transcription start site (TSS) of each gene. When the data were summarized based on *Gene Set*, we chose to include only Gene Ontology (GO) gene sets for analysis. For each gene set, we asked SCRAT to aggregate reads within the 3000 bp upstream and 1000 bp downstream regions from TSSs of all genes.

SCRAT allows one to normalize the features and filter them based on the user-provided parameters (**Fig. 3b**). Once all parameters are set, one can start the summarization process using the “Run Summarization” button (**Fig. 3c**). After the summarization is done, the summarized features can be viewed and downloaded from the “Results” panel (**Fig. 3e**). Then, one can proceed to cell heterogeneity analysis using the “Next step” button (**Fig. 3f**).

2

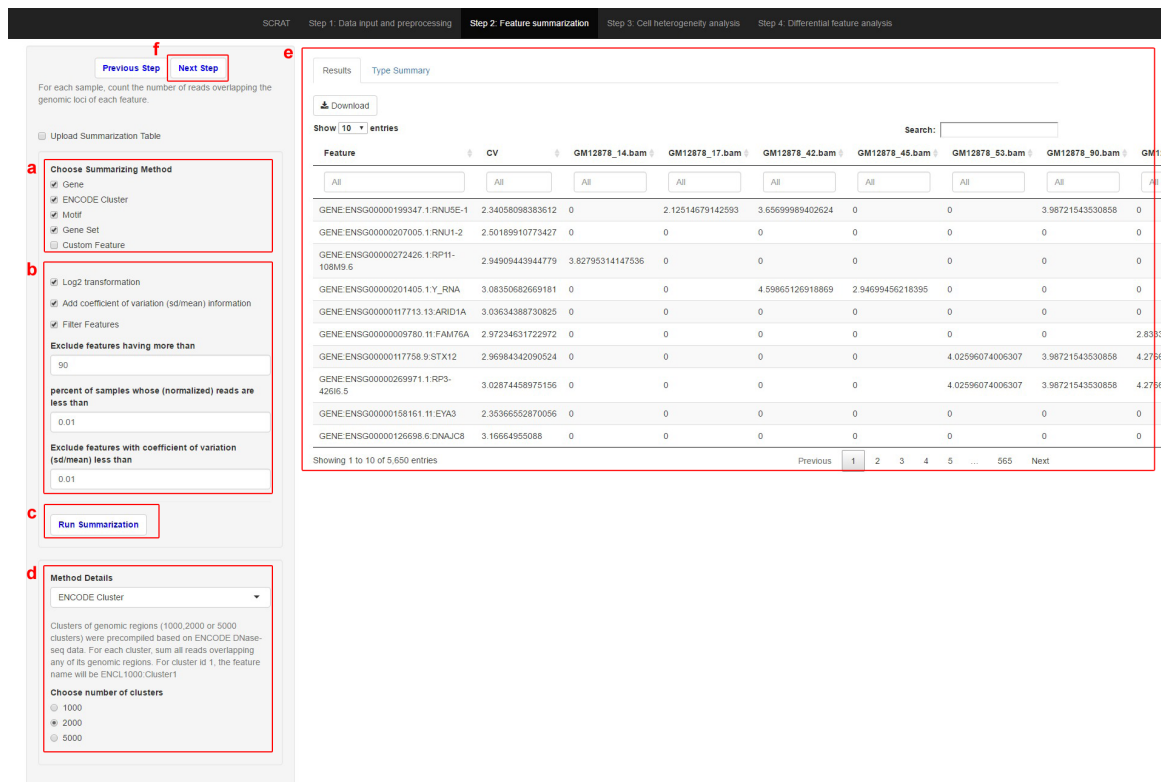


Figure 3. SCRAT analysis step 2 -- Feature summarization.

Step 3: Cell heterogeneity analysis

The third step is to dissect the cell heterogeneity by clustering the cells. First, one can select different types of features for clustering in the “*Select Feature Type*” section (Fig. 4a). Second, one can choose what type of methods to use to reduce the dimension of the features in the “*Dimension reduction method*” section (Fig. 4b). Third, one can choose the clustering method in the “*Clustering method*” section (Fig. 4c). By default, SCRAT selects the *ENCODE Cluster* features and uses the principal components of these features to cluster cells based on model-based clustering. One can start the clustering process using the “*Perform Clustering*” button (Fig. 4d). Then, the result will be shown in the “*Clustering Result*” panel (Fig. 4e). Applying this procedure to the example data yields two clusters (Fig. 4e), corresponding to GM12878 and HEK293T respectively.

After obtaining the cell clustering results, one can further explore the cell identities by comparing the individual cells with the existing cell types in our pre-compiled bulk DNase-seq database. First, one can use the “*Include existing cell types*” function (Fig. 4f) to select samples from the existing cell types in the database and project them to the principal component space of the single cells. Second, one can also evaluate the similarity between each cell and the existing cell types using the “*Similarity to existing cell types*” function (Fig. 4g) based on the selected features (Fig. 5a). One can start the analysis using the “*Calculate Correlations*” button (Fig. 5b). The results will be

visualized as a heatmap (Fig. 5c). Then, one can proceed to differential feature analysis using the “Next step” button (Fig. 4h).

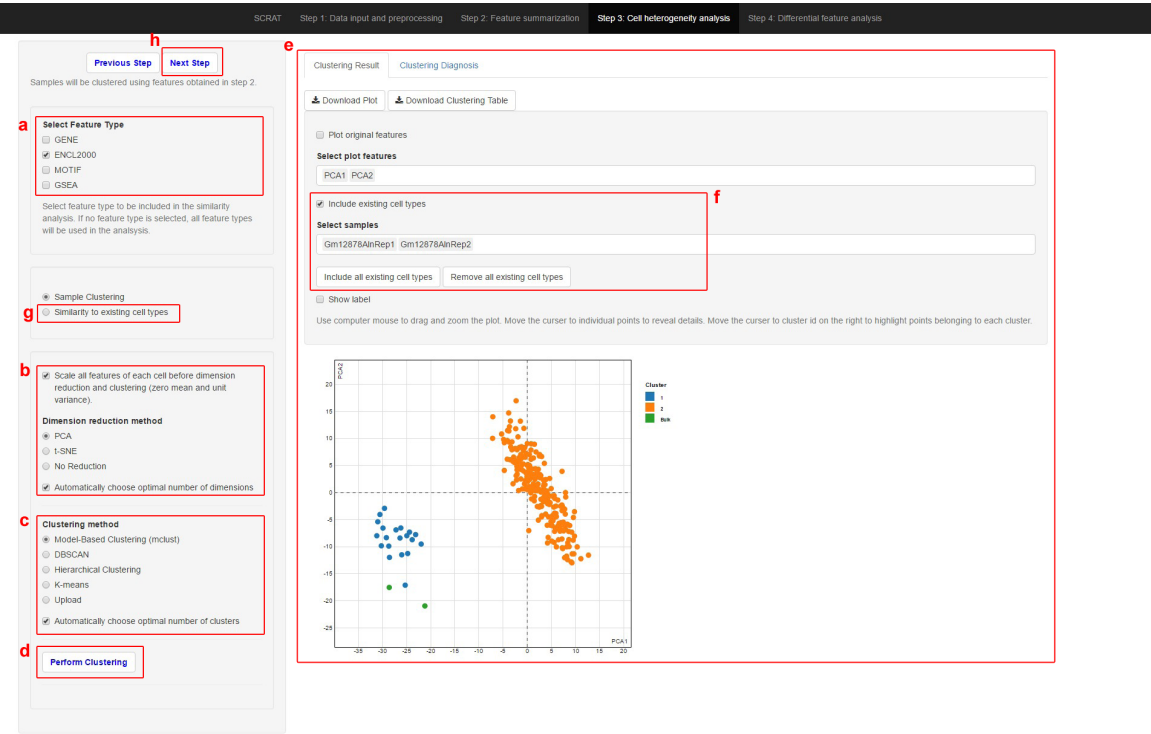


Figure 4. SCRAT analysis step 3 -- Cell heterogeneity analysis.

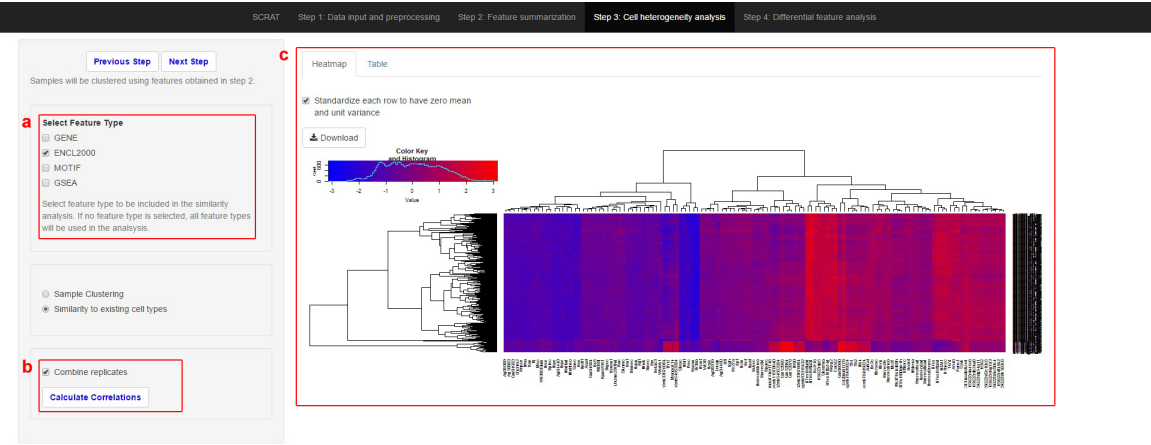


Figure 5. SCRAT analysis step 3 -- Cell heterogeneity analysis (cont'd). Evaluating similarity to existing cell types.

Step 4: Differential feature analysis

The last step is to identify the differential features among different cell subpopulations. One can perform analysis to all cell clusters obtained from Step 3 or a subset of selected cell clusters (**Fig. 6a**). Then, one can choose what type of statistical test to be used (**Fig. 6b**). If more than two cell clusters are selected, ANOVA F-test, Kruskal-Wallis test, or permutation test can be used to perform analysis to each feature. If only two cell clusters are selected, t-test, Wilcoxon rank-sum test, or permutation test can be used to perform analysis to each feature. One can start the analysis using the “*Perform Test*” button (**Fig. 6c**). The results including the name of the feature, the test statistics and the adjusted p-value (FDR) will be shown in the “*Results*” panel (**Fig. 6d**).

Previous Step

Perform ANOVA or t tests to identify key features that mostly explains the between cluster variance. The sample clusters are obtained in step 3.

Select Feature Type

☒ GENE
☒ ENCL2000
☒ MOTIF
☒ GSEA

Select feature type to be included in the differential feature analysis. If no feature type is selected, all feature types will be used in the analysis.

☐ Perform tests for all clusters

Select clusters where ANOVA or t test will be performed (at least two should be selected)

1 2

Select test method

☒ t test
☐ wilcoxon test (nonparametric)
☐ Permutation test

Select alternative hypothesis type

☒ two-sided
☐ less
☐ greater

Cluster 1 will be compared with cluster 2. The alternative hypothesis is that Cluster 1 is not equal to cluster 2.

Perform Test

Results Summary

Download

Show 10 entries

Search:

Feature	statistics	FDR
GSEA	All	All
GSEA.DEFENSE_RESPONSE_TO_VIRUS	5.76525016862885	3.01715329252574e-7
GSEA.CELLULAR_DEFENSE_RESPONSE	4.57264960563719	0.000632303539580002
GSEA.LOCOMOTORY_BEHAVIOR	4.48484988281781	0.000090369202395809
GSEA.MANNOSYLTRANSFERASE_ACTIVITY	4.40259624258474	0.000126779748372507
GSEA.OXIDOREDUCTASE_ACTIVITY_ACTING_ON_THE_CH_NH_GROUP_OF_DONORS	4.19651899547787	0.000279065826727942
GSEA.SULFUR_METABOLIC_PROCESS	-4.12182314293521	0.000369621547460509
GSEA.RESPONSE_TO_VIRUS	3.99094155863226	0.000590237504721182
GSEA.VASCULATURE_DEVELOPMENT	-3.9835975110619	0.000606916162829049
GSEA.ORGAN_MORPHOGENESIS	-3.92341209810456	0.00075000830240635
GSEA.DEFENSE_RESPONSE	3.92267019426233	0.000751304353943667

Showing 1 to 10 of 1,435 entries (filtered from 5,650 total entries)

Previous 1 2 3 4 5 ... 144 Next

Figure 6. SCRAT analysis step 4 -- Differential feature analysis.