SCRAT User Interface Manual

Zhicheng Ji, Weiqiang Zhou, Hongkai Ji

August 22, 2016

1 SCRAT Online GUI

SCRAT Online GUI can be launched directly online: https://zhiji.shinyapps.io/scrat/. The only needed software is the web browser and no additional software is required to be installed. All currently supported genomes are available for the online GUI (hg19, hg38, mm9, mm10).

2 Installation SCRAT software

SCRAT software can be installed via Github. Users should have R installed on their computer before installing SCRAT. R can be downloaded here: http://www.r-project.org/. Users should first install the SCRAT data packages by running following commands in R. Note that one does not need to install all data packages. For example if the reads are aligned to hg19 genome, then only SCRATdatahg19 package should be installed.

```
if (!require("devtools"))
  install.packages("devtools")
devtools::install_github("SCRATdatahg19","zji90")
devtools::install_github("SCRATdatahg38","zji90")
devtools::install_github("SCRATdatamm10","zji90")
devtools::install_github("SCRATdatamm9","zji90")
```

After that, one can install the latest version of SCRAT package via Github by running the following commands in R (For Windows users, if the installation fails please use R version 3.2.x):

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("GenomicAlignments")
devtools::install_github("SCRAT","zji90")
```

After that, one can launch the user interface by running the following commands in R:

```
library(SCRAT)
SCRATui()
```

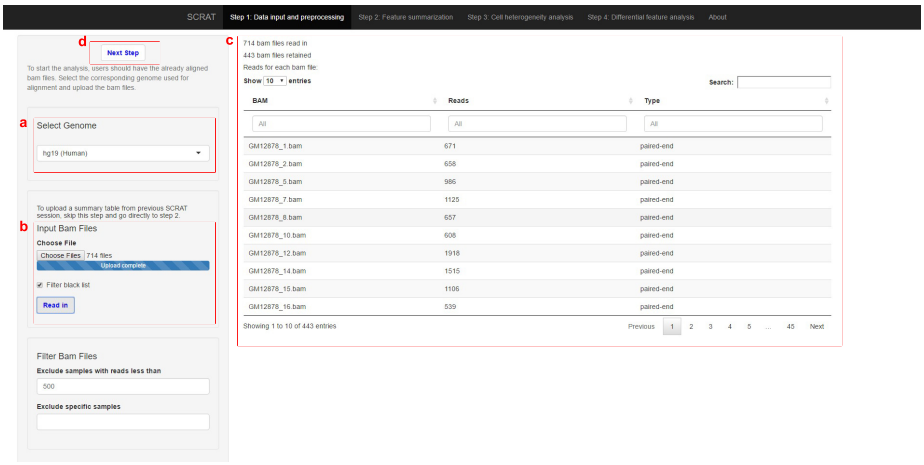

Figure 1: Screenshot demonstrating step 1: Data input and preprocessing

3 SCRAT GUI Overview

The SCRAT GUI consists of three main parts: the main menu on the top of the GUI where users can switch between analysis steps, the sidebar panel on the left side where users can specify options, and the main panel where plots and results will appear.

A typical SCRAT analysis consists of four major analysis steps: 1.Data input and preprocessing; 2.Feature summarization; 3.Cell heterogeneity analysis and 4.Differential feature analysis. The first two steps are necessary if users want to perform cell heterogeneity or differential feature analyses. Users can switch between each step using the main menu on the top or the "Previsous Step" (Figure 1d) and "Next Step" (Figure 2f) buttons on the top-left corner. All key buttons in the analysis steps will be highlighted with blue color.

4 Data input and preprocessing

The first step is to input bam files into SCRAT and select the corresponding genome. First, one should select the correct genome for the bam files to be uploaded in "Select Genome" (Figure 1a). Note that if launched from one's local computer, SCRAT GUI will only show genomes that has been installed on the computer (refer to "Installation SCRAT software"). Then, one can input the bam files by clicking "Choose Files" button (Figure 1b). Note that multiple files can be chosen from the current folder. To choose multiple files hold "Control" key and keep clicking the mouse. To select all files in the folder hold "Control" key and press "a". Only files ending with ".bam" can be chosen. For the online GUI, it may take time to upload all the bam files to the system.

After selecting the bam files, one can choose whether to exclude the reads overlapping with the ENOCDE black list regions [1] by selecting "Filter black list" (Figure 1b).

[1] https://sites.google.com/site/anshulkundaje/projects/blacklists

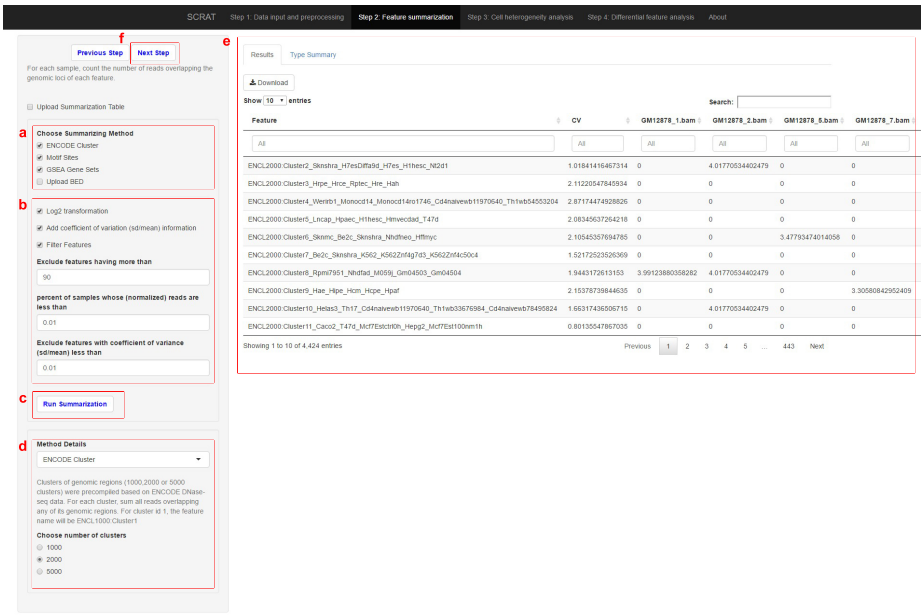

Figure 2: Screenshot demonstrating step 2: Feature summarization

Click "Read in" button to read in the bam files (Figure 1b). A summary table will appear showing the total number of bam files read in and number of samples kept for analysis (Figure 1c). One can choose to exclude samples with too few reads or certain samples from the analysis in "Filter Bam Files". After that, one can click "Next Step" or use the menu bar to go to step 2.

5 Feature summarization

SCRAT can summarize reads counts into four different kinds of features (Figure 2a). Press "Run Summarization" button to run summarization with all default parameters (Figure 2c). Each feature corresponds to one or multiple genomic loci and SCRAT will calculate and report the total number of reads intersecting with all the loci for each feature. The four categories are as follows:

1. ENCODE cluster: using all the ENCODE DNase-seq bulk samples, we split the whole genome into 200 bp windows, retain windows that are likely to have DNase-seq signals and clustered them into 1000, 2000 or 5000 clusters.

2. Motif sites: we mapped all the jaspar core vertebrata motifs [2] and some transfac motifs [3] to the whole genome. We retained the mapped motif sites which has non-zero DNase-seq counts in at least one of the ENCODE DNase-seq bulk samples. By default the 100bp flanking region (both sides) of the motif site will be taken.

3. GSEA: we compiled all GSEA [4] gene sets where each gene set contains a

[2] http://jaspar.genereg.net/
[3] http://www.gene-regulation.com/pub/databases.html
[4] http://software.broadinstitute.org/gsea/index.jsp

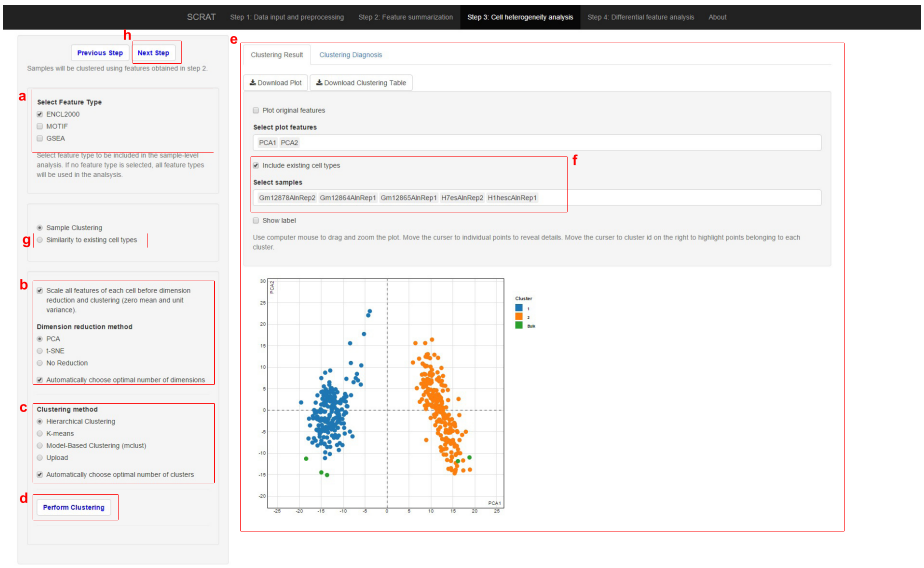

Figure 3: Screenshot demonstrating clustering function in step 3: Cell heterogeneity analysis

list of genes. For each gene set we calculate the total read counts intersecting with the promoter regions of all its genes.

4. upload: one can input genomic regions by uploading bed file.

After obtaining the read counts for each feature and for each sample, SCRAT will normalize the read counts accounting for library size (total read counts for each bam file). Specifically, for each sample the read counts will be divided by the library size of this sample and multiplied by 10000 for all samples. Users can also choose to take log2 transformation of the normalized read counts and filter out features that are low across all samples (Figure 2b). The coefficient of variation can also be calculated for each feature (Figure 2b).

Users also have the option to adjust the definition of the four features categories (Figure 2d). For example, they can choose different number of ENCODE cluster.

It is worth mentioning that one can save the feature table as txt files (Figure 2e). These files can also be load back into SCRAT to continue or reproduce the results from previous SCRAT sessions. Note that by doing so one does not need to upload bam files but still need to select the appropriate genome in step 1.

6 Cell heterogeneity analysis

First users should choose feature types to be included in the analysis (Figure 3a). Note that changing the feature types requires users to manually rerun all analyses since the clustering results will not be updated automatically. Two types of analyses are available: sample clustering and similarity to existing cell types (Figure 3g).

Before sample clustering, users can choose to perform dimension reduction

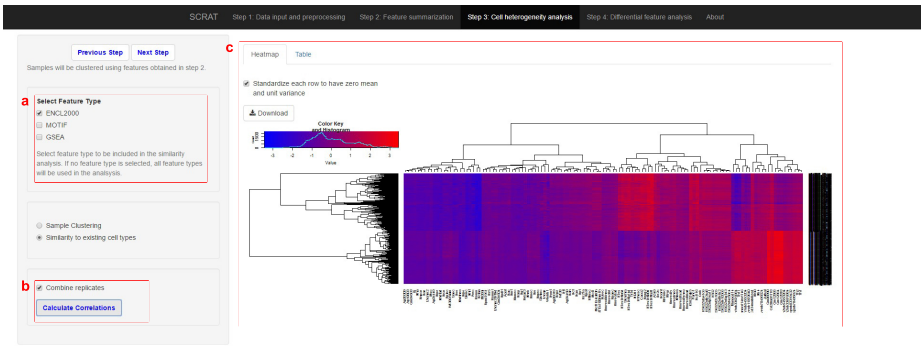

Figure 4: Screenshot demonstrating similarity function in step 3: Cell heterogeneity analysis

using PCA or t-SNE (Figure 3b). For PCA users can manually specify the number of PCs or let SCRAT to automatically choose the optimal number of PCs. For t-SNE users need to manually specify the number of dimensions. Users can also choose to omit dimension reduction and perform sample clustering based on the original data. SCRAT provides three methods to perform the sample clustering: hierarchical clustering, k-means clustering and model-based clustering (Figure 3c). Users also have the option to manually specify the number of clusters or let SCRAT to automatically determine the optimal number of clusters. Click "Perform Clustering" button to do the clustering (Figure 3d). By default the clustering results will show up as an interactive scatterplot in the main panel (Figure 3e). One can move the mouse curser to the individual points to show the details of the points, move the curser to the cluster id to highlight all samples that belong to this cluster or zoom in/out using the mouse. Users can also specify what kind of features are to be shown on the plot (Figure 3e). SCRAT will show a boxplot for one feature, interactive scatterplot for two features and heatmap for more than two features. For PCA users can also add existing bulk cell types obtained from ENCODE to the plot (Figure 3f). These bulk cell types are preprocessed using exactly the same SCRAT procedures and then projected to the space of principal components.

One can also compare the single-cell samples to ENCODE bulk samples by calculating the pearson correlations (Figure 4). All ENCODE bulk DNase-seq samples are downloaded and processed from UCSC ENCODE [5]. Then we calculated all the features for the ENCODE bulk samples using exactly the same SCRAT procedures. When calculating the correlations between single-cell and bulk samples, we take the intersect of all features selected in Figure 4a. Click "Calculate Correlations" button to perform the analysis (Figure 4b). The results will be shown using heatmap or table (Figure 4c). Note that one can choose to combine bulk replicates with the same cell line and treatment (Figure

[5]https://genome.ucsc.edu/ENCODE/

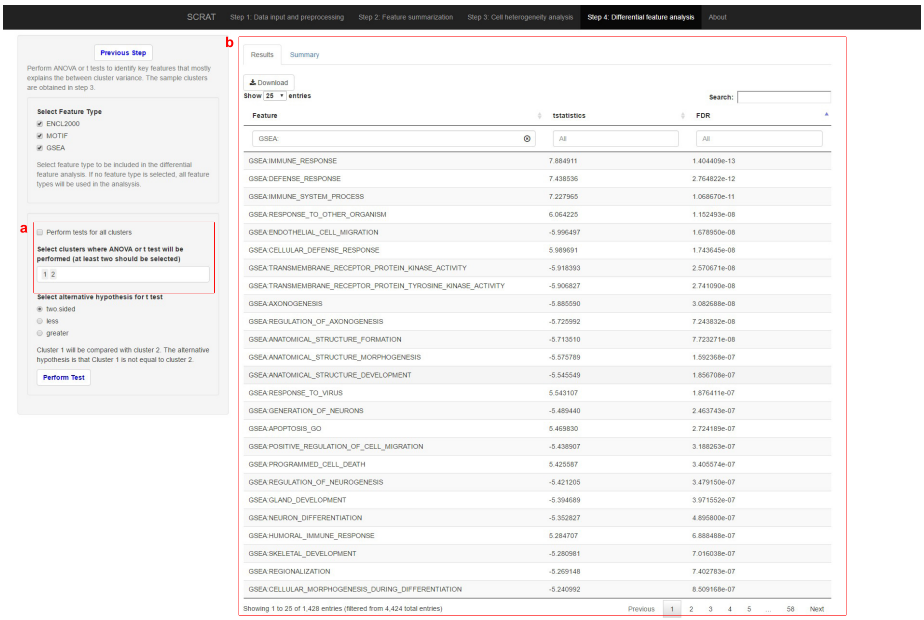

Figure 5: Screenshot demonstrating step 4: Differential feature analysis

5b). In this case the average of features for bulk replicates will be used.

7 Differential feature analysis

After sample clustering, one can perform differential feature analysis by identifying the features that drive the differneces between sample clusters. Users can specify what clusters are to be included in the analysis (Figure 5a). For two samples the t tests will be performed and for more than two samples ANOVA tests will be performed. The pvalues are further adjusted using the BH procedure to obtain FDR. The results will show up in Figure 5b.

8 Contact

To report bugs and provide suggestions for the SCRAT GUI as well as the SCRAT package, please contact the maintainer Zhicheng Ji (zji4@jhu.edu).