scBIRD User Interface Manual

Yueyi Li, Weiqiang Zhou, Runzhe Li, Hongkai Ji

April 20, 2020

1 Install scBIRD

scBIRD package can be installed via Github by running the following commands in R:

```
devtools::install_github("liyueyikate/scBIRD")
```

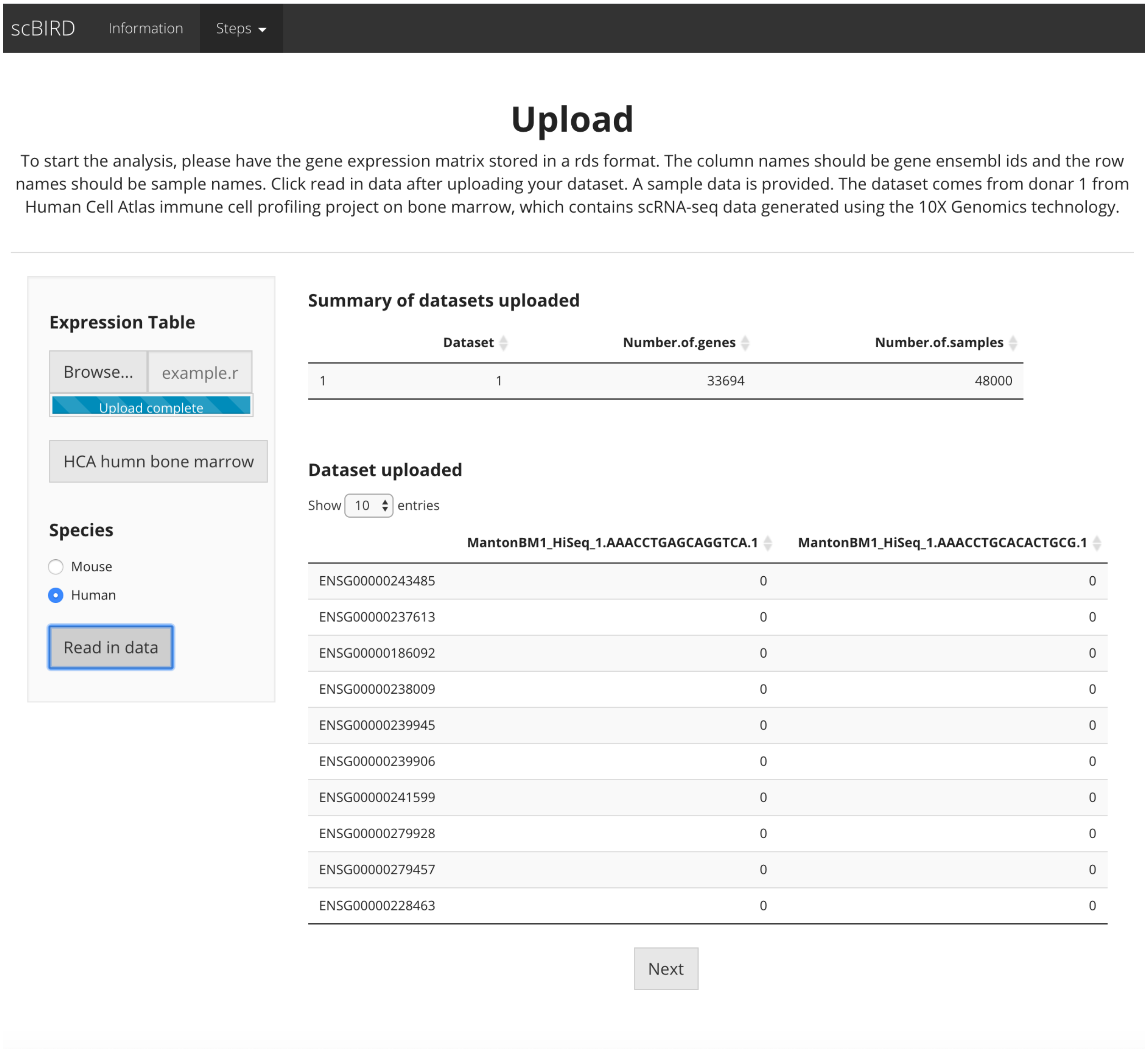
The user interface can be run by the following commands:

```
library(scBIRD)
scBIRDui()
```

A complete scBIRD analysis consists 6 steps: 1.Data uploadidng; 2.Quality control; 3.Normalization; 4.Feature Selection; 5.Dimension reduction and clustering 6.Bird prediction. User can switch between each step using the slider bar on the top or the "Next" button on the bottom.

2 Data uploadidng

The first step is to input a gene expression count matrix saved in a rds format. Click browse to upload a file from local and then choose species. An example data is provided. The dataset comes from donar 1 from Human Cell Atlas immune cell profiling project on bone marrow, which contains scRNA-seq data generated using the 10X Genomics technology. A simple summary describes the number of genes and cells in the dataset. Entries for the first 50 genes and the first two cells are printed out.



Upload

To start the analysis, please have the gene expression matrix stored in a rds format. The column names should be gene ensembl ids and the row names should be sample names. Click read in data after uploading your dataset. A sample data is provided. The dataset comes from donar 1 from Human Cell Atlas immune cell profiling project on bone marrow, which contains scRNA-seq data generated using the 10X Genomics technology.

Expression Table

Species

○ Mouse
● Human

Summary of datasets uploaded

	Dataset	Number.of.genes	Number.of.samples
1	1	33694	48000

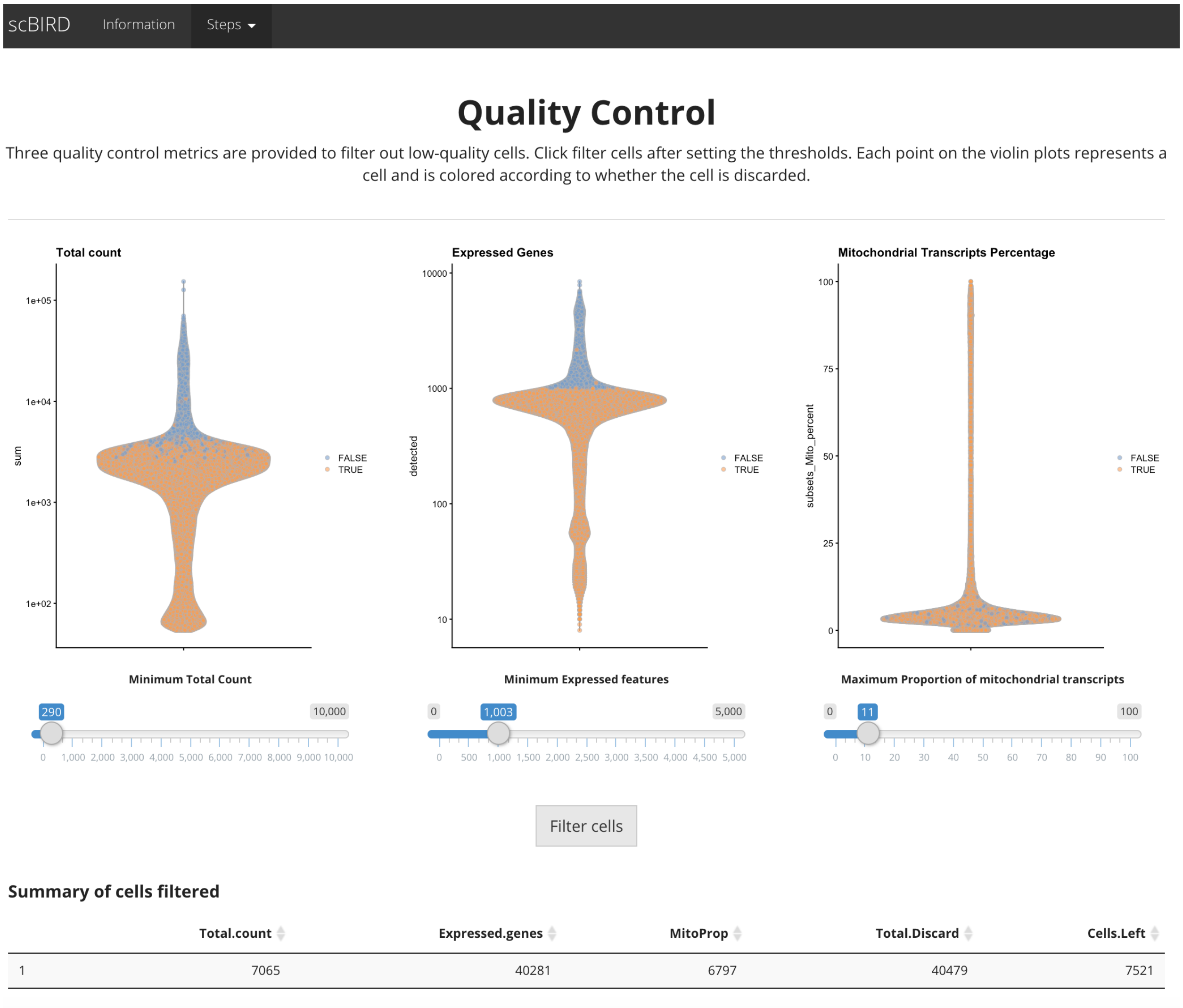
Dataset uploaded

Show 10 entries

	MantonBM1_HiSeq_1.AAACCTGAGCAGGTCA.1	MantonBM1_HiSeq_1.AAACCTGCACACTGCG.1
ENSG00000243485	0	0
ENSG00000237613	0	0
ENSG00000186092	0	0
ENSG00000238009	0	0
ENSG00000239945	0	0
ENSG00000239906	0	0
ENSG00000241599	0	0
ENSG00000279928	0	0
ENSG00000279457	0	0
ENSG00000228463	0	0

3 Quality control

After uploading the data into scBIRD, the quality control metrics are calculated for each cell. scBIRD provides three metric options to identify low-quality cells: the total counts across all genes of a cell, the number of expressed genes, the fraction of counts mapped to mitochondrial genes. Users could choose the thresholds of metrics to filter cells. The default thresholds are three median absolute deviation away from the the median value of the metrics, after being log transformed. Click “Filter cells” after setting the threshold. The distributions of the metrics are visualized using violin plots (Figure 1). Each point represents a cell corresponding to its metric value on the y-axis. The points are coloured orange if they are of low-quality. Based on the thresholds of the metrics, low-quality cells will be removed. A summary table is provided.



Quality Control

Three quality control metrics are provided to filter out low-quality cells. Click filter cells after setting the thresholds. Each point on the violin plots represents a cell and is colored according to whether the cell is discarded.

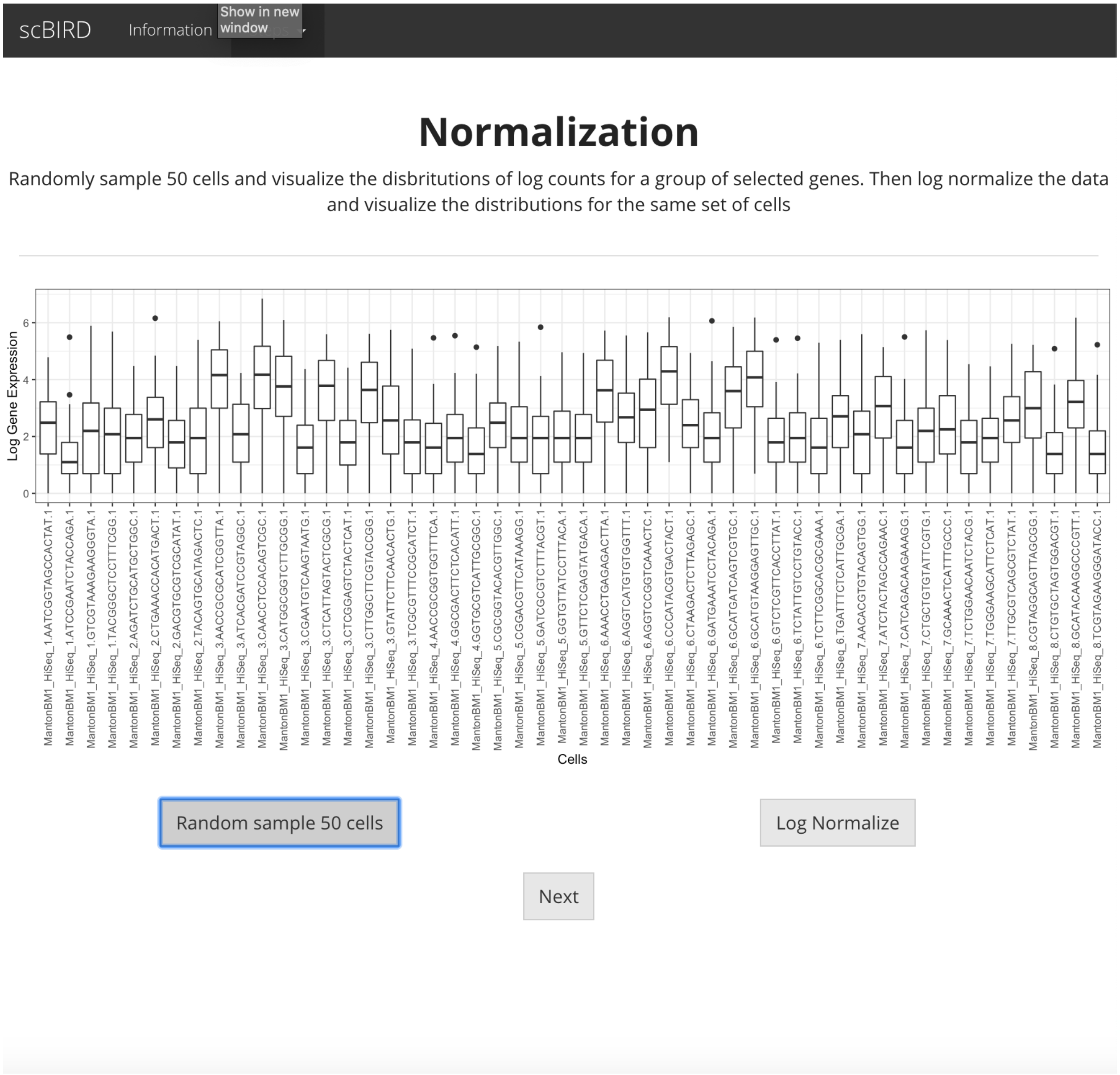
Summary of cells filtered

	Total.count	Expressed.genes	MitoProp	Total.Discard	Cells.Left
1	7065	40281	6797	40479	7521

4 Normalization

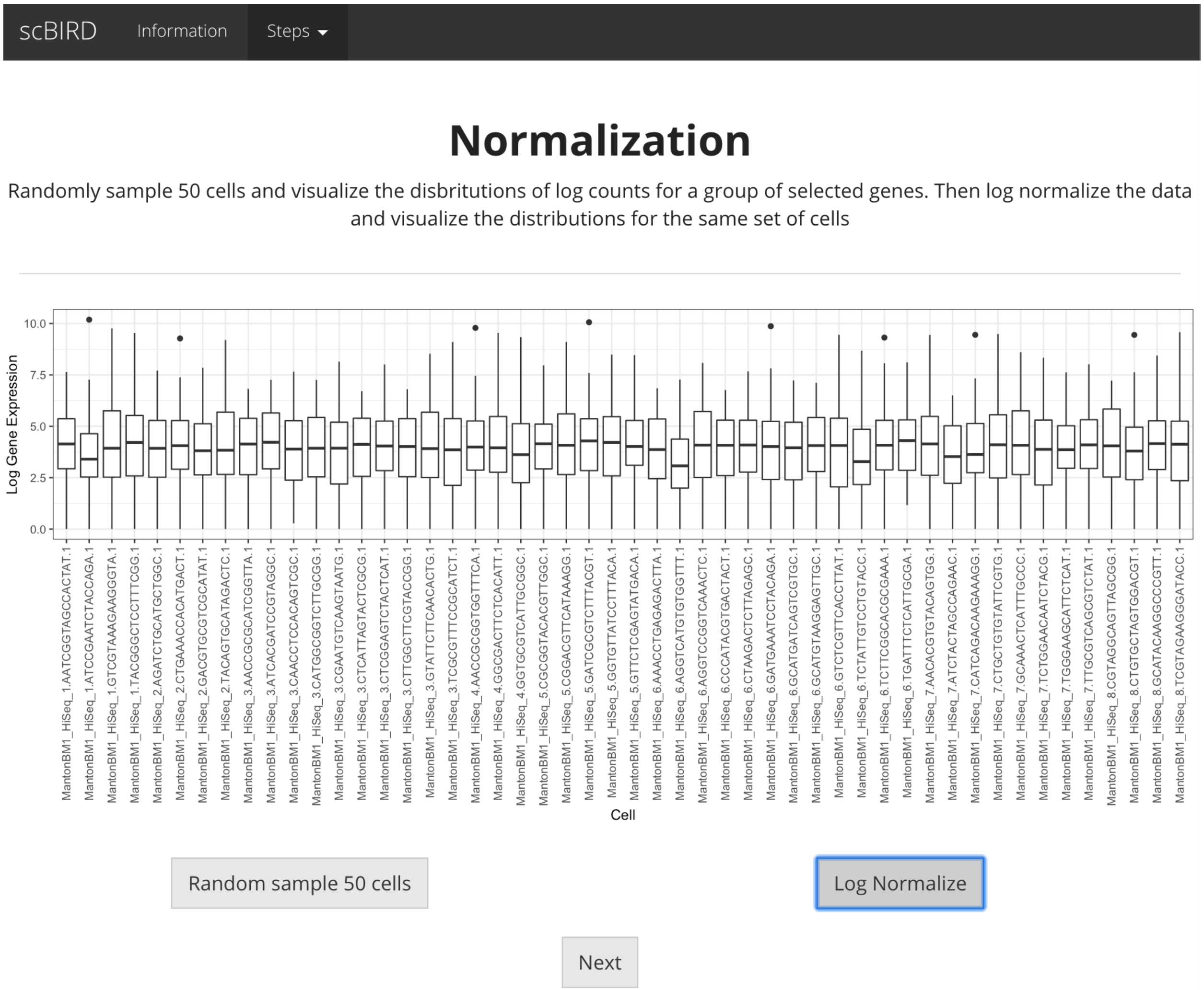
scBIRD uses library size normalization. The expression values are normalized by dividing the count for each gene of a cell with a size factor for that cell. The size factor for a cell is proportional to the cell's library size and the mean of the size factors across all cells is 1. Then the normalized expression values are log-transformed.

Press “Random sample 50 cells” to generate boxplots (Figure 2) that show the gene expression distributions. The x-axis of the boxplots are 50 randomly generated cells and the y-axis are the log-transformed expression values of a set of selected genes, which need to be expressed in over 80 % of the cells. Click “Log normalize” to perform library size normalization. Normalized gene expression distribuions of the 50 cells will be displayed (Figure 3).


scBIRD
Information
Show in new window
Log Gene Expression
0
2
4
6
MantonBM1_HiSeq_1.AATCGGTAGCCACTAT.1
MantonBM1_HiSeq_1.ATCCGAATCTACCAGA.1
MantonBM1_HiSeq_1.GTCGTAAAGAAGGGTA.1
MantonBM1_HiSeq_1.TACGGGCTCCTTTCGG.1
MantonBM1_HiSeq_2.AGATCTGCATGCTGGC.1
MantonBM1_HiSeq_2.CTGAAACCACATGACT.1
MantonBM1_HiSeq_2.GACGTGCGTCGCATAT.1
MantonBM1_HiSeq_2.TACAGTGCATAGACTC.1
MantonBM1_HiSeq_3.AACCGCGCATCGGTTA.1
MantonBM1_HiSeq_3.ATCACGATCCGTAGGC.1
MantonBM1_HiSeq_3.CAACCTCCACAGTCGC.1
MantonBM1_HiSeq_3.CATGGCGGTCTTGCGG.1
MantonBM1_HiSeq_3.CGAATGTCAAGTAATG.1
MantonBM1_HiSeq_3.CTCATTAGTACTCGCG.1
MantonBM1_HiSeq_3.CTCGGAGTCTACTCAT.1
MantonBM1_HiSeq_3.CTTGGCTTCGTACCGG.1
MantonBM1_HiSeq_3.GTATTCTTCAACACTG.1
MantonBM1_HiSeq_3.TCGCGTTTCCGCATCT.1
MantonBM1_HiSeq_4.AACCGCGGTGGTTTCA.1
MantonBM1_HiSeq_4.GGCGACTTCTCACATT.1
MantonBM1_HiSeq_4.GGTGCGTCATTGCGGC.1
MantonBM1_HiSeq_5.CGCGGTACACGTTGGC.1
MantonBM1_HiSeq_5.CGGACGTTCATAAAGG.1
MantonBM1_HiSeq_5.GATCGCGTCTTTACGT.1
MantonBM1_HiSeq_5.GGTGTTATCCTTTACA.1
MantonBM1_HiSeq_5.GTTCTCGAGTATGACA.1
MantonBM1_HiSeq_6.AAACCTGAGAGACTTA.1
MantonBM1_HiSeq_6.AGGTCATGTGTGGTTT.1
MantonBM1_HiSeq_6.AGGTCCGGTCAAACTC.1
MantonBM1_HiSeq_6.CCCATACGTGACTACT.1
MantonBM1_HiSeq_6.CTAAGACTCTTAGAGC.1
MantonBM1_HiSeq_6.GATGAAATCCTACAGA.1
MantonBM1_HiSeq_6.GCATGATCAGTCGTGC.1
MantonBM1_HiSeq_6.GCATGTAAGGAGTTGC.1
MantonBM1_HiSeq_6.GTCTCGTTCACCTTAT.1
MantonBM1_HiSeq_6.TCTATTGTCCTGTACC.1
MantonBM1_HiSeq_6.TCTTCGGCACGCGAAA.1
MantonBM1_HiSeq_6.TGATTTCTCATTGCGA.1
MantonBM1_HiSeq_7.AACACGTGTACAGTGG.1
MantonBM1_HiSeq_7.ATCTACTAGCCAGAAC.1
MantonBM1_HiSeq_7.CATCAGACAAGAAAGG.1
MantonBM1_HiSeq_7.CTGCTGTGTATTCGTG.1
MantonBM1_HiSeq_7.GCAAACTCATTTGCCC.1
MantonBM1_HiSeq_7.TCTGGAACAATCTACG.1
MantonBM1_HiSeq_7.TGGGAAGCATTCTCAT.1
MantonBM1_HiSeq_7.TTGCGTCAGCGTCTAT.1
MantonBM1_HiSeq_8.CGTAGGCAGTTAGCGG.1
MantonBM1_HiSeq_8.CTGTGCTAGTGGACGT.1
MantonBM1_HiSeq_8.GCATACAAGGCCCGTT.1
MantonBM1_HiSeq_8.TCGTAGAAGGGATACC.1
Cells
Random sample 50 cells
Log Normalize
Next

Normalization

Randomly sample 50 cells and visualize the disbritutions of log counts for a group of selected genes. Then log normalize the data and visualize the distributions for the same set of cells


scBIRD
Information
Steps ▾

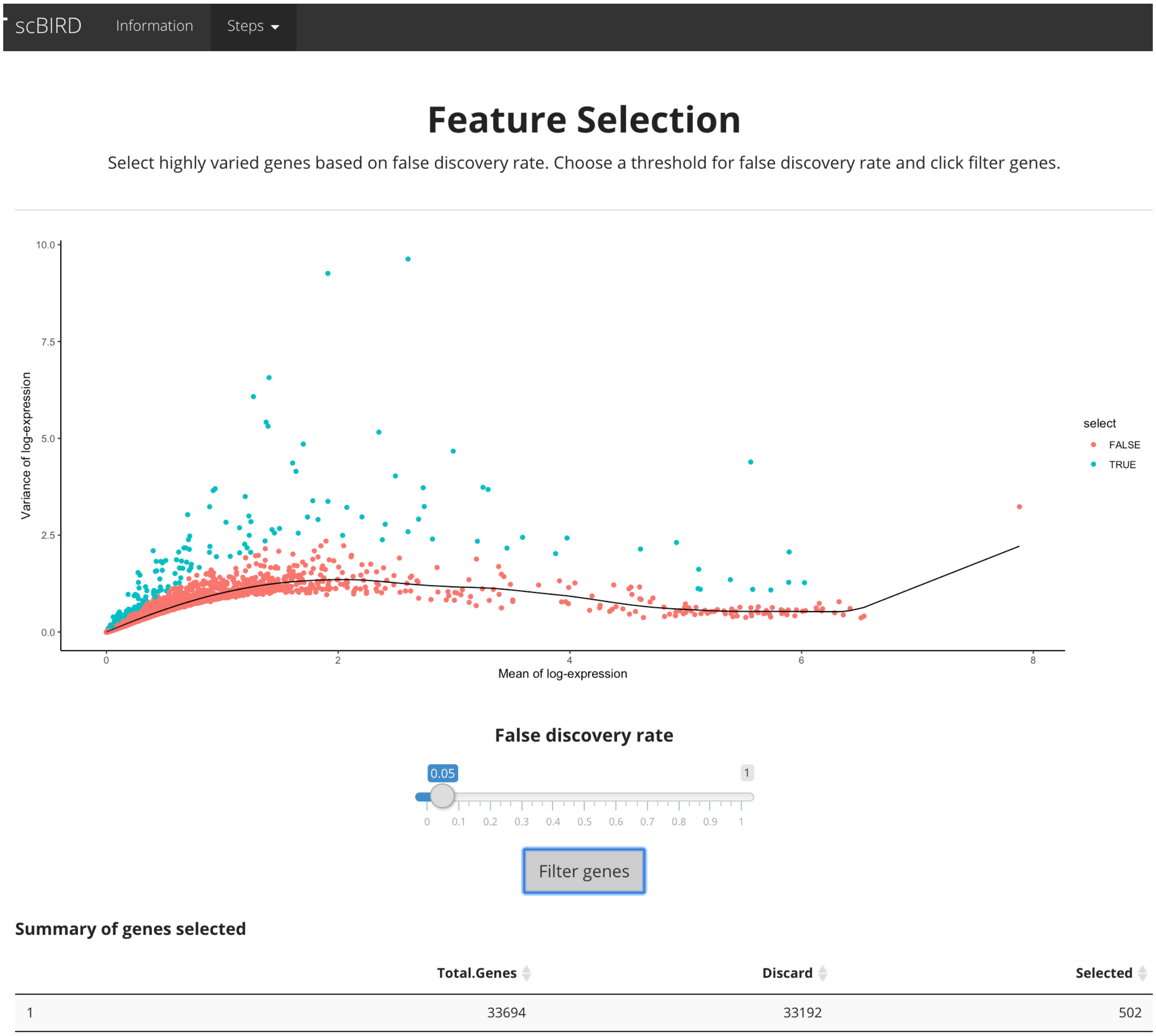
Normalization

Randomly sample 50 cells and visualize the disbritutions of log counts for a group of selected genes. Then log normalize the data and visualize the distributions for the same set of cells

Log Gene Expression
0.0
2.5
5.0
7.5
10.0
Cell
MantonBM1_HiSeq_1.AATCGGTAGCCACTAT.1
MantonBM1_HiSeq_1.ATCCGAATCTACCAGA.1
MantonBM1_HiSeq_1.GTCGTAAAGAAGGGTA.1
MantonBM1_HiSeq_1.TACGGGCTCCTTTCGG.1
MantonBM1_HiSeq_2.AGATCTGCATGCTGGC.1
MantonBM1_HiSeq_2.CTGAAACCACATGACT.1
MantonBM1_HiSeq_2.GACGTGCGTCGCATAT.1
MantonBM1_HiSeq_2.TACAGTGCATAGACTC.1
MantonBM1_HiSeq_3.AACCGCGCATCGGTTA.1
MantonBM1_HiSeq_3.ATCACGATCCGTAGGC.1
MantonBM1_HiSeq_3.CAACCTCCACAGTCGC.1
MantonBM1_HiSeq_3.CATGGCGGTCTTGCGG.1
MantonBM1_HiSeq_3.CGAATGTCAAGTAATG.1
MantonBM1_HiSeq_3.CTCATTAGTACTCGCG.1
MantonBM1_HiSeq_3.CTCGGAGTCTACTCAT.1
MantonBM1_HiSeq_3.CTTGGCTTCGTACCGG.1
MantonBM1_HiSeq_3.GTATTCTTCAACACTG.1
MantonBM1_HiSeq_3.TCGCGTTTCCGCATCT.1
MantonBM1_HiSeq_4.AACCGCGGTGGTTTCA.1
MantonBM1_HiSeq_4.GGCGACTTCTCACATT.1
MantonBM1_HiSeq_4.GGTGCGTCATTGCGGC.1
MantonBM1_HiSeq_5.CGCGGTACACGTTGGC.1
MantonBM1_HiSeq_5.CGGACGTTCATAAAGG.1
MantonBM1_HiSeq_5.GATCGCGTCTTTACGT.1
MantonBM1_HiSeq_5.GGTGTTATCCTTTACA.1
MantonBM1_HiSeq_5.GTTCTCGAGTATGACA.1
MantonBM1_HiSeq_6.AAACCTGAGAGACTTA.1
MantonBM1_HiSeq_6.AGGTCATGTGTGGTTT.1
MantonBM1_HiSeq_6.AGGTCCGGTCAAACTC.1
MantonBM1_HiSeq_6.CCCATACGTGACTACT.1
MantonBM1_HiSeq_6.CTAAGACTCTTAGAGC.1
MantonBM1_HiSeq_6.GATGAAATCCTACAGA.1
MantonBM1_HiSeq_6.GCATGATCAGTCGTGC.1
MantonBM1_HiSeq_6.GCATGTAAGGAGTTGC.1
MantonBM1_HiSeq_6.GTCTCGTTCACCTTAT.1
MantonBM1_HiSeq_6.TCTATTGTCCTGTACC.1
MantonBM1_HiSeq_6.TCTTCGGCACGCGAAA.1
MantonBM1_HiSeq_6.TGATTTCTCATTGCGA.1
MantonBM1_HiSeq_7.AACACGTGTACAGTGG.1
MantonBM1_HiSeq_7.ATCTACTAGCCAGAAC.1
MantonBM1_HiSeq_7.CATCAGACAAGAAAGG.1
MantonBM1_HiSeq_7.CTGCTGTGTATTCGTG.1
MantonBM1_HiSeq_7.GCAAACTCATTTGCCC.1
MantonBM1_HiSeq_7.TCTGGAACAATCTACG.1
MantonBM1_HiSeq_7.TGGGAAGCATTCTCAT.1
MantonBM1_HiSeq_7.TTGCGTCAGCGTCTAT.1
MantonBM1_HiSeq_8.CGTAGGCAGTTAGCGG.1
MantonBM1_HiSeq_8.CTGTGCTAGTGGACGT.1
MantonBM1_HiSeq_8.GCATACAAGGCCCGTT.1
MantonBM1_HiSeq_8.TCGTAGAAGGGATACC.1
Random sample 50 cells
Log Normalize
Next

5 Feature Selection

scBIRD identifies highly variable genes (HVGs) by calculating the variances of the log counts for each gene and decomposing the variances into biological and technical components. A trend (black line in Figure 5) is fit to the variance of log counts with respect to the mean across all genes using loess. Assuming that the majority of genes are not differentially expressed and that the technical component makes up the most of the total variance, the technical component for the genes are estimated by the fitted values of the trend. The biological component is the difference between the total variance and the technical component. Users could control the number of HVGs by false discovery rate (FDR), calculated by testing against the null hypothesis that the variance is equal to the trend. Press the “Filter genes” button. Genes below the FDR threshold will be selected as HVGs (blue). A summary table will indicate how many genes are selected.



Feature Selection

Select highly varied genes based on false discovery rate. Choose a threshold for false discovery rate and click filter genes.

False discovery rate

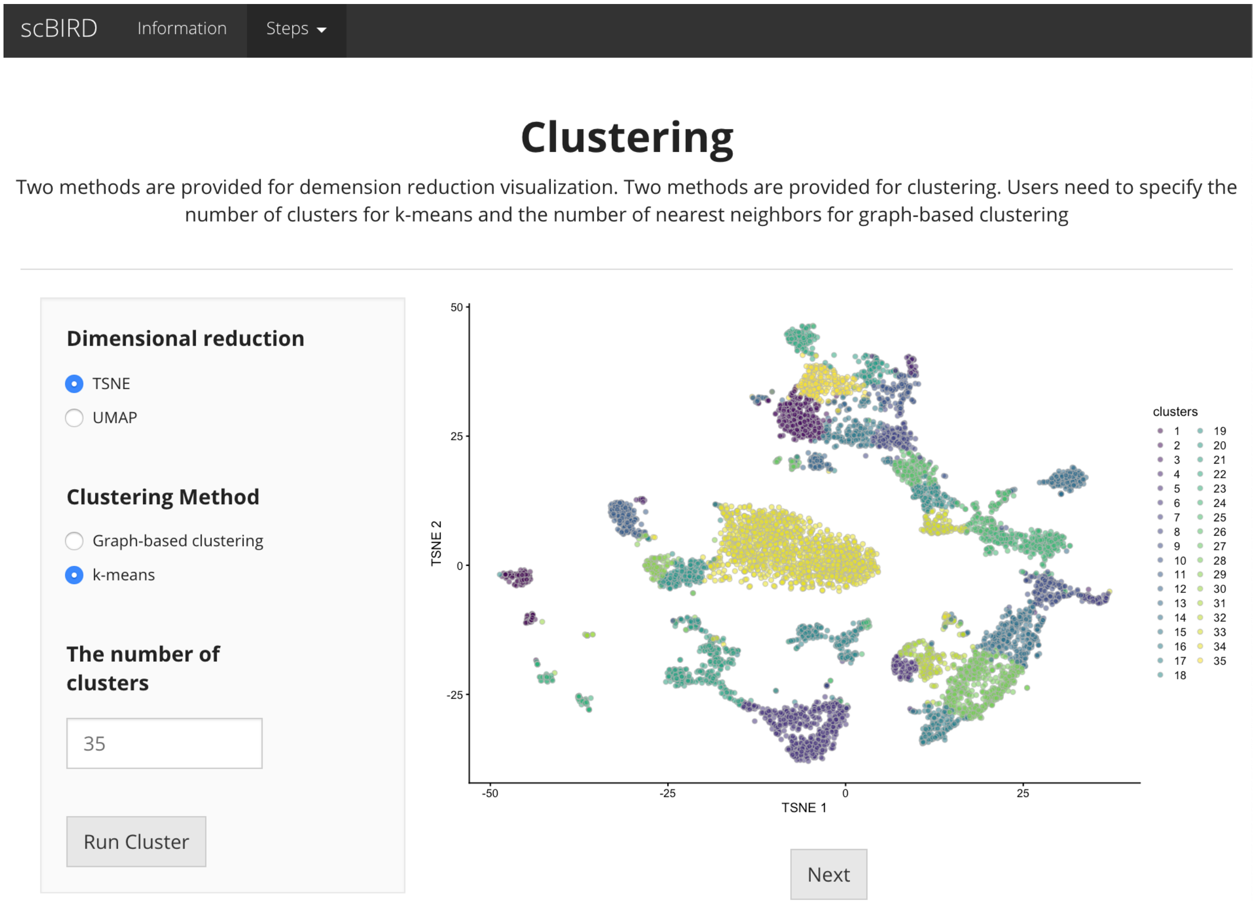
Summary of genes selected

	Total.Genes	Discard	Selected
1	33694	33192	502

6 Dimension reduction and clustering

Principal components analysis is first performed to reduce the data into 30 dimensions using the genes selected in the feature selection step. Then the data is further compressed into 2 dimensions using t-stochastic neighbor embedding (T-SNE) or uniform manifold approximation and projection (UMAP) by the user's choice (Dimension reduction in Figure 6) .

Two clustering methods are provided: graph-based clustering and k-means. Users need to set the number of nearest neighbors for the graph-based clustering method. A larger number of nearest neighbors will lead to a more connected graph and smaller number of clusters. For k-means, users need to choose the number of clusters. A T-SNE or UMAP plot by the user's choice will be generated where each point represents a cells and is colored according to its cluster.



Clustering

Two methods are provided for demension reduction visualization. Two methods are provided for clustering. Users need to specify the number of clusters for k-means and the number of nearest neighbors for graph-based clustering

Dimensional reduction

● TSNE

○ UMAP

Clustering Method

○ Graph-based clustering

● k-means

The number of clusters

35

## 7 Bird prediction

Before performing BIRD prediction, one need to have a training model. Four models prebuilt by Weiqiang Zhou are available for BIRD:

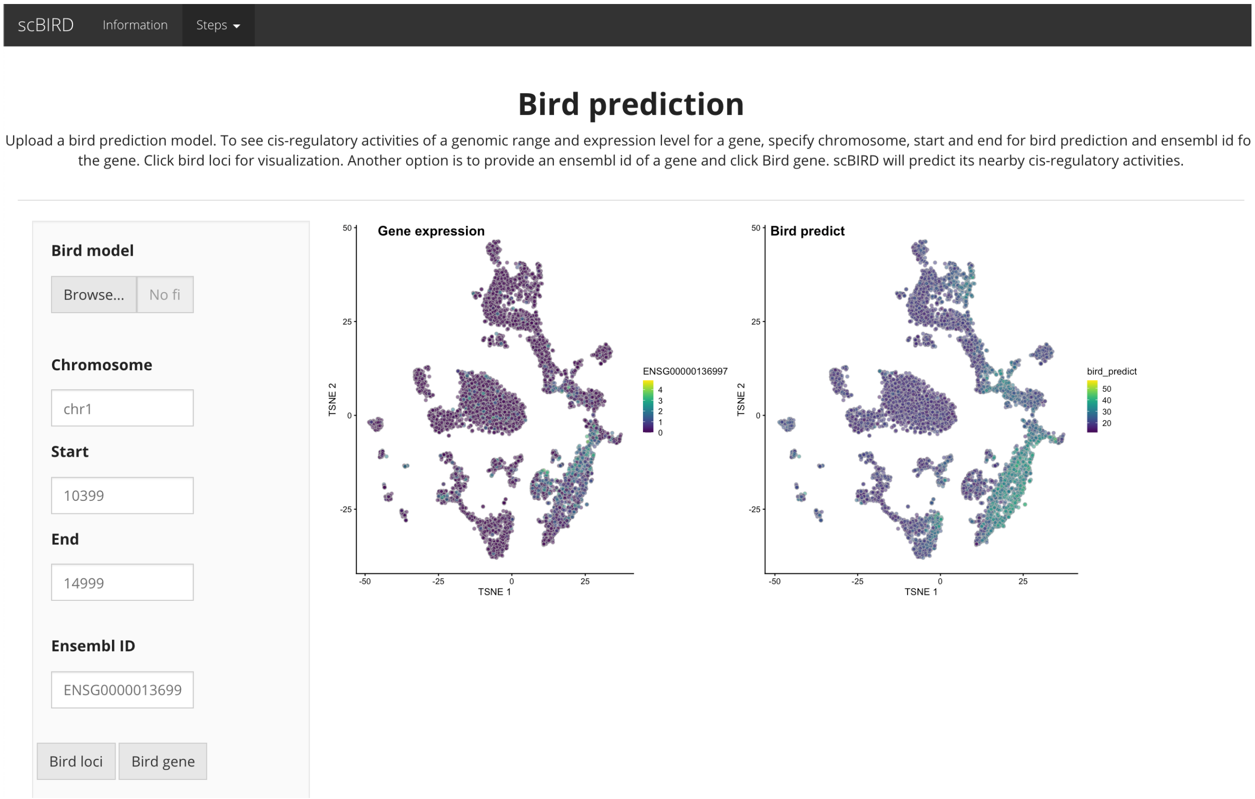
RNA-seq model, current release (trained with 167 ENCODE samples)

RNA-seq model, previous release (trained with 70 Epigenome Roadmap samples)

RNA-seq model for 2 million loci, previous release (trained with 70 Epigenome Roadmap samples)

Exon Array model

Upload the model bin file from local. The genomic loci are selected based on their signal in the training DNase-seq samples. scBIRD could also output a vector of aggregate signals for a set of user-selected genomics regions. The genomic regions could be specified by chromosome, start and end. User could also choose a gene and the aggregate signals for the nearby loci will be predicted. The loci are match to nearest genes using HOMER. Predictions are performed on all the loci matched to the selected gene.



Bird prediction

Upload a bird prediction model. To see cis-regulatory activities of a genomic range and expression level for a gene, specify chromosome, start and end for bird prediction and ensembl id fo the gene. Click bird loci for visualization. Another option is to provide an ensembl id of a gene and click Bird gene. scBIRD will predict its nearby cis-regulatory activities.

Bird model

Chromosome

chr1

Start

10399

End

14999

Ensembl ID

ENSG0000013699