celloracle 0.1.0 documentation

[celloracle](index.html#document-index)

* [Welcome to celloracle’s documentation!](index.html#document-index)
* [Installation](index.html#document-installation/index)
  + [Docker image](index.html#docker-image)
  + [System Requirements](index.html#system-requirements)
  + [Python Requirements](index.html#python-requirements)
    - [0. (Optional) Make a new environment](index.html#optional-make-a-new-environment)
    - [1. Add conda channels](index.html#add-conda-channels)
    - [2. Install velocyto](index.html#install-velocyto)
    - [3. Install scanpy](index.html#install-scanpy)
    - [4. Install gimmemotifs](index.html#install-gimmemotifs)
    - [5. Install other python libraries](index.html#install-other-python-libraries)
    - [6. install celloracle from github](index.html#install-celloracle-from-github)
  + [R requirements](index.html#r-requirements)
    - [Seurat](index.html#id3)
    - [Cicero](index.html#id5)
    - [igraph](index.html#id7)
    - [linkcomm](index.html#id9)
    - [rnetcarto](index.html#id11)
    - [Check installation](index.html#check-installation)
* [Tutorial](index.html#document-tutorials/index)
  + [1. ATAC-seq data preprocessing](index.html#document-tutorials/atac)
    - [A. Extract TF binding information from scATAC-seq data](index.html#a-extract-tf-binding-information-from-scatac-seq-data)
      * [0. Import library](index.html#document-notebooks/01_ATAC-seq_data_processing/option1_scATAC-seq_data_analysis_with_cicero/01_atacdata_to_cicero)
      * [1. Prepare data](index.html#1.-Prepare-data)
      * [2. Load data and make Cell Data Set (CDS) object](index.html#2.-Load-data-and-make-Cell-Data-Set-(CDS)-object)
        + [2.1. Process data to make CDS object](index.html#2.1.-Process-data-to-make-CDS-object)
      * [3. Qauality check and Filtering](index.html#3.-Qauality-check-and-Filtering)
      * [4. Process cicero-CDS object](index.html#4.-Process-cicero-CDS-object)
      * [5. Run cicero to get cis-regulatory connection scores](index.html#5.-Run-cicero-to-get-cis-regulatory-connection-scores)
      * [6. Save results for next step](index.html#6.-Save-results-for-next-step)
      * [0. Import libraries](index.html#document-notebooks/01_ATAC-seq_data_processing/option1_scATAC-seq_data_analysis_with_cicero/02_preprocess_peak_data)
      * [1. Load data made with cicero](index.html#1.-Load-data-made-with-cicero)
      * [2. Make TSS annotation](index.html#2.-Make-TSS-annotation)
        + [IMPORTANT: Please make sure that you are setting correct reference genoms.](index.html#IMPORTANT:-Please-make-sure-that-you-are-setting-correct-reference-genoms.)
      * [3. Integrate TSS info and cicero connections](index.html#3.-Integrate-TSS-info-and-cicero-connections)
      * [4. Filter peaks](index.html#4.-Filter-peaks)
      * [5. Save data](index.html#5.-Save-data)
    - [B. Extract TF binding information from bulk ATAC-seq data or Chip-seq data](index.html#b-extract-tf-binding-information-from-bulk-atac-seq-data-or-chip-seq-data)
      * [0. Import libraries](index.html#document-notebooks/01_ATAC-seq_data_processing/option2_Bulk_ATAC-seq_data/01_preprocess_Bulk_ATAC_seq_peak_data)
      * [1. Load bed file](index.html#1.-Load-bed-file)
      * [2. Make TSS annotation](index.html#2.-Make-TSS-annotation)
      * [3. Save data](index.html#3.-Save-data)
  + [2. Transcription factor binding motif scan](index.html#document-tutorials/motifscan)
    - [0. Import libraries](index.html#document-notebooks/02_motif_scan/02_atac_peaks_to_TFinfo_with_celloracle_190901)
    - [1. Load data](index.html#1.-Load-data)
    - [2. Check data](index.html#2.-Check-data)
      * [2.1. Remove short peaks](index.html#2.1.-Remove-short-peaks)
    - [3. Instantiate TFinfo object and search for TF binding motifs](index.html#3.-Instantiate-TFinfo-object-and-search-for-TF-binding-motifs)
    - [3.1 check reference genome installation](index.html#3.1-check-reference-genome-installation)
      * [3.2. Install reference genome (if refgenome is not installed)](index.html#3.2.-Install-reference-genome-(if-refgenome-is-not-installed))
    - [4. Scan motifs and save object](index.html#4.-Scan-motifs-and-save-object)
    - [5. Filtering motifs](index.html#5.-Filtering-motifs)
    - [6. Get Final results](index.html#6.-Get-Final-results)
      * [6.1. Get resutls as a dictionary](index.html#6.1.-Get-resutls-as-a-dictionary)
      * [6.2. Get results as a dataframe](index.html#6.2.-Get-results-as-a-dataframe)
    - [7. Save TFinfo as dictionary or dataframe](index.html#7.-Save-TFinfo-as-dictionary-or-dataframe)
  + [3. Single-cell RNA-seq data preprocessing](index.html#document-tutorials/scrnaprocess)
    - [A. scRNA-seq data preprocessing with scanpy](index.html#a-scrna-seq-data-preprocessing-with-scanpy)
      * [0. Import libraries](index.html#document-notebooks/03_scRNA-seq_data_preprocessing/scanpy_preprocessing_with_Paul_etal_2015_data)
      * [1. Load data](index.html#1.-Load-data)
      * [2. Filtering](index.html#2.-Filtering)
      * [3. Normalization](index.html#3.-Normalization)
      * [4. Identification of highly variable genes](index.html#4.-Identification-of-highly-variable-genes)
      * [5. Log transformation](index.html#5.-Log-transformation)
      * [6. Dimensional reduction](index.html#6.-Dimensional-reduction)
      * [7. Clustering](index.html#7.-Clustering)
      * [(Optional) Re-calculate Dimensional reduction graph](index.html#(Optional)-Re-calculate-Dimensional-reduction-graph)
      * [8. Check data](index.html#8.-Check-data)
        + [8.1. Visualize marker gene expression](index.html#8.1.-Visualize-marker-gene-expression)
      * [8. Make annotation for cluster](index.html#8.-Make-annotation-for-cluster)
        + [8.1. Make annotation (1)](index.html#8.1.-Make-annotation-(1))

[!! Please change the dictionary below depending on the clustering results. The results may change depending on the execution environment.](index.html#!!-Please-change-the-dictionary-below-depending-on-the-clustering-results.-The-results-may-change-depending-on-the-execution-environment.)

* + - * + [8.2. Make annotation (2)](index.html#8.2.-Make-annotation-(2))

[!! Please change the dictionary below depending on the clustering results. The results may change depending on the execution environment.](index.html#id1)

* + - * [9. (Option) Subset cells](index.html#9.-(Option)-Subset-cells)
      * [10. Save data](index.html#10.-Save-data)
    - [B. scRNA-seq data preprocessing with Seurat](index.html#b-scrna-seq-data-preprocessing-with-seurat)
  + [4. Network analysis](index.html#document-tutorials/networkanalysis)
    - [0. Import libraries](index.html#document-notebooks/04_Network_analysis/Network_analysis_with_with_Paul_etal_2015_data)
      * [0.1. Check installation](index.html#0.1.-Check-installation)
      * [0.2. Make a folder to save graph](index.html#0.2.-Make-a-folder-to-save-graph)
    - [1. Load data](index.html#1.-Load-data)
      * [1.1. Load processed gene expression data (anndata)](index.html#1.1.-Load-processed-gene-expression-data-(anndata))
      * [1.2. Load TF data.](index.html#1.2.-Load-TF-data.)
    - [2. Initiate Oracle object](index.html#2.-Initiate-Oracle-object)
      * [2.1. load gene expression data into oracle object.](index.html#2.1.-load-gene-expression-data-into-oracle-object.)
      * [2.2. Load TFinfo into oracle object](index.html#2.2.-Load-TFinfo-into-oracle-object)
      * [2.3. (Optional) Add TF info manually](index.html#2.3.-(Optional)-Add-TF-info-manually)
        + [2.3.1. Make TF info dictionary manually](index.html#2.3.1.-Make-TF-info-dictionary-manually)
        + [2.3.2. Add TF informatio dictionary into the oracle object](index.html#2.3.2.-Add-TF-informatio-dictionary-into-the-oracle-object)
    - [3. Knn imputation](index.html#3.-Knn-imputation)
      * [3.1. PCA](index.html#3.1.-PCA)
      * [3.2. KNN imputation](index.html#3.2.-KNN-imputation)
    - [4. Save and Load.](index.html#4.-Save-and-Load.)
    - [4. GRN calculation](index.html#4.-GRN-calculation)
      * [4.1. Get GRNs](index.html#4.1.-Get-GRNs)
      * [4.2. (Optional) Export GRNs](index.html#4.2.-(Optional)-Export-GRNs)
      * [4.3. (Optional) Change order](index.html#4.3.-(Optional)-Change-order)
    - [5. Network preprocessing](index.html#5.-Network-preprocessing)
      * [5.1. Filter network edges](index.html#5.1.-Filter-network-edges)
      * [5.2. Degree distribution](index.html#5.2.-Degree-distribution)
      * [5.3. Calculate netowrk score](index.html#5.3.-Calculate-netowrk-score)
      * [5.4. Save](index.html#5.4.-Save)
    - [6. Network analysis; Network score for each gene](index.html#6.-Network-analysis;-Network-score-for-each-gene)
      * [6.1. Network score in each cluster](index.html#6.1.-Network-score-in-each-cluster)
      * [6.2. Network score comparison between two clusters](index.html#6.2.-Network-score-comparison-between-two-clusters)
      * [6.3. Network score dynamics](index.html#6.3.-Network-score-dynamics)
      * [6.4. Gene cartography analysis](index.html#6.4.-Gene-cartography-analysis)
    - [7. Network analysis; network score distribution](index.html#7.-Network-analysis;-network-score-distribution)
      * [7.1. Distribution of network degree](index.html#7.1.-Distribution-of-network-degree)
      * [7.2. Distribution of netowrk entolopy](index.html#7.2.-Distribution-of-netowrk-entolopy)
  + [5. Simulation with GRNs](index.html#document-tutorials/simulation)
    - [0. Import libraries](index.html#document-notebooks/05_simulation/Gata1_KO_simulation_with_with_Paul_etal_2015_data)
      * [0.1. Import public libraries](index.html#0.1.-Import-public-libraries)
      * [0.1. Make a folder to save graph](index.html#0.1.-Make-a-folder-to-save-graph)
    - [1. Load data](index.html#1.-Load-data)
      * [1.1. Load processed oracle object](index.html#1.1.-Load-processed-oracle-object)
      * [1.2. Load inferred GRNs](index.html#1.2.-Load-inferred-GRNs)
    - [3. Make predictive models for simulation](index.html#3.-Make-predictive-models-for-simulation)
    - [4. in silico Perturbation-simulation](index.html#4.-in-silico-Perturbation-simulation)
      * [4.1. Check gene expression pattern.](index.html#4.1.-Check-gene-expression-pattern.)
      * [4.1. calculate future gene expression after perturbation.](index.html#4.1.-calculate-future-gene-expression-after-perturbation.)
      * [4.2. calculate transition probability between cells](index.html#4.2.-calculate-transition-probability-between-cells)
      * [4.3. Visualization](index.html#4.3.-Visualization)
        + [4.3.1. Detailed directed trajectory graph](index.html#4.3.1.-Detailed-directed-trajectory-graph)
        + [4.3.2. Grid graph](index.html#4.3.2.-Grid-graph)
      * [4.4. Markov simulation to analyze the effects of perturbation on cell fate transition](index.html#4.4.-Markov-simulation-to-analyze-the-effects-of-perturbation-on-cell-fate-transition)
        + [4.4.1. Do simulation](index.html#4.4.1.-Do-simulation)
        + [4.4.2. Check the results of the simulation for specific cells](index.html#4.4.2.-Check-the-results-of-the-simulation-for-specific-cells)
        + [4.4.3. Summarize the results of simulation by plotting sankey diagram](index.html#4.4.3.-Summarize-the-results-of-simulation-by-plotting-sankey-diagram)
* [API](index.html#document-modules/index)
  + [Command Line API](index.html#command-line-api)
  + [Python API](index.html#python-api)
    - [Custom class in celloracle](index.html#document-modules/celloracle)
    - [Modules for ATAC-seq analysis](index.html#modules-for-atac-seq-analysis)
      * [celloracle.motif\_analysis module](index.html#document-modules/celloracle.motif_analysis)
    - [Modules for Network analysis](index.html#modules-for-network-analysis)
      * [celloracle.network\_analysis module](index.html#document-modules/celloracle.network_analysis)
    - [Other modules](index.html#other-modules)
      * [celloracle.go\_analysis module](index.html#document-modules/celloracle.go_analysis)
      * [celloracle.utility module](index.html#document-modules/celloracle.utility)
      * [celloracle.data module](index.html#document-modules/celloracle.data)
      * [celloracle.data\_conversion module](index.html#document-modules/celloracle.data_conversion)
* [Changelog](index.html#document-changelog)
* [License](index.html#document-license/index)
* [Authors and citations](index.html#document-citation/index)
  + [Cite celloracle](index.html#cite-celloracle)
  + [celloracle software development](index.html#celloracle-software-development)

[celloracle](index.html#document-index)

* [Docs](index.html#document-index) »
* celloracle 0.1.0 documentation

# Welcome to celloracle’s documentation![¶](#welcome-to-celloracle-s-documentation)

CellOracle is a python library for the analysis of Gene Regulatory Network with single cell data.

Source code are available at [celloracle git hub repository](https://github.com/morris-lab/CellOracle)

Note

**For colleagues in Morris Lab**

**Thank you for your kind help!!**

Documentation is also available as a word and pdf file.

[pdf documentation](_downloads/2b68115d706e6ca9c7d63e468f820981/celloracle.pdf)

[word documentation](_downloads/be48a717cb6d826b0fff88ea52479901/documentation_microsoft_word.docx)

The word documentation does not include figures.

**Could you please add comments on the file and send it back to Kenji?**

I really appreciate your support!

Thanks,

Kenji

(This message will be deleted when we launch celloracle.)

Warning

CellOracle is still under development. It is alpha version and functions in this package may change in the future release.

# Contents[¶](#contents)

## Installation[¶](#installation)

celloracle uses several python libraries and R library. Please follow this guide below to install the dependent software of celloracle.

### Docker image[¶](#docker-image)

* Not available now. Comming soon.

### System Requirements[¶](#system-requirements)

* Operating system: macOS or linux are highly recommended. celloracle was developed and tested in Linux and macOS.
* We found that the celloracle calculation may be EXTREMELY SLOW under an environment of Windows Subsystem for Linux (WSL). We do not recommend using WSL.
* While you can install celloracle in Windows OS, please do so at your own risk and responsibility. We DO NOT provide any support for the use in the Windows OS.
* Memory: 8 G byte or more. Memory usage also depends on your scRNA-seq data. Especially in silico simulation requires large amount of memory.
* CPU: Core i5 or better processor. GRN inference supports multicore calculation. Higer number of CPU cores enables faster calculation.

### Python Requirements[¶](#python-requirements)

* celloracle was developed with python 3.6. We do not support python 2.7x or python <=3.5.
* Please install all dependent libraries before installing celloracle according to the instructions below.
* celloracle is still beta version and it is not available through PyPI or anaconda distribution yet. Please install celloracle from GitHub repository according to the instruction below.

#### 0. (Optional) Make a new environment[¶](#optional-make-a-new-environment)

This step is optional. Please make a new python environment for celloracle and install dependent libraries in it if you get some software conflicts.

conda create -n celloracle\_env python=3.6  
conda activate celloracle\_env

#### 1. Add conda channels[¶](#add-conda-channels)

Installation of some libraries requires non-default anaconda channels. Please add the channels below. Instead, you can explicitly enter the channel when you install a library.

conda config --add channels defaults  
conda config --add channels bioconda  
conda config --add channels conda-forge

#### 2. Install [velocyto](http://velocyto.org/velocyto.py/install/index.html)[¶](#install-velocyto)

Please install velocyto with the following commands or [the author’s instruction](http://velocyto.org/velocyto.py/install/index.html) . On Mac OS, you may have a compile error during velocyto installation. I recommend installing [Xcode](https://developer.apple.com/xcode/) in that case.

conda install numpy scipy cython numba matplotlib scikit-learn h5py click

Then

pip install velocyto

#### 3. Install [scanpy](https://scanpy.readthedocs.io/en/stable/installation.html)[¶](#install-scanpy)

Please install scanpy with the following commands or [the author’s instruction](https://scanpy.readthedocs.io/en/stable/installation.html) .

conda install seaborn statsmodels numba pytables python-igraph louvain

Then

pip install scanpy

#### 4. Install [gimmemotifs](https://gimmemotifs.readthedocs.io/en/master/installation.html)[¶](#install-gimmemotifs)

Please install gimmemotifs with the following commands or [the author’s instruction](https://gimmemotifs.readthedocs.io/en/master/installation.html) .

conda install gimmemotifs genomepy=0.5.5

#### 5. Install other python libraries[¶](#install-other-python-libraries)

Please install other python libraries below with the following commands.

conda install goatools pyarrow tqdm joblib jupyter

#### 6. install celloracle from github[¶](#install-celloracle-from-github)

pip install git+https://github.com/morris-lab/CellOracle.git

### R requirements[¶](#r-requirements)

celloracle use R libraries for the network analysis and scATAC-seq analysis. Please install [R](https://www.r-project.org) (>=3.5) and R libraries below according to the author’s instruction.

#### [Seurat](https://satijalab.org/seurat/install.html)[¶](#id3)

Please install Seurat with the following r-script or [the author’s instruction](https://satijalab.org/seurat/install.html) . celloracle is compatible with both Seurat V2 and V3. If you use only scanpy for the scRNA-seq preprocessing and do not use Seurat , you can skip installation of Seurat.

In R console,

install.packages('Seurat')

#### [Cicero](https://cole-trapnell-lab.github.io/cicero-release/docs/#installing-cicero)[¶](#id5)

Please install Cicero with the following r-script or [the author’s instruction](https://cole-trapnell-lab.github.io/cicero-release/docs/#installing-cicero) . If you do not have scATAC-seq data and plan to use celloracle’s base GRN, you do not need to install Cicero.

In R console,

if (!requireNamespace("BiocManager", quietly = TRUE))  
install.packages("BiocManager")  
BiocManager::install("cicero")

#### [igraph](https://igraph.org/r/)[¶](#id7)

Please install igraph with the following r-script or [the author’s instruction](https://igraph.org/r/) .

In R console,

install.packages("igraph")

#### [linkcomm](https://cran.r-project.org/web/packages/linkcomm/index.html)[¶](#id9)

Please install linkcomm with the following r-script or [the author’s instruction](https://cran.r-project.org/web/packages/linkcomm/index.html) .

In R console,

install.packages("linkcomm")

#### [rnetcarto](https://github.com/cran/rnetcarto/blob/master/src/rgraph/README.md)[¶](#id11)

Please install rnetcarto with the following r-script or [the author’s instruction](https://github.com/cran/rnetcarto/blob/master/src/rgraph/README.md) .

In R console,

install.packages("rnetcarto")

#### Check installation[¶](#check-installation)

These R libraries above are necessary for the network analysis in celloracle. You can check installation using celloracle’s function.

In python console,

import celloracle as co  
co.network\_analysis.test\_R\_libraries\_installation()

Please make sure that all R libraries are installed. The following message will be shown when all R libraries are appropriately installed.

checking R library installation: igraph -> OK

checking R library installation: linkcomm -> OK

checking R library installation: rnetcarto -> OK

## Tutorial[¶](#tutorial)

The analysis proceeds through multiple steps. Please run the notebooks sequentially. If you do not have ATAC-seq data and want to use the default TF binding information, you can skip the first and second step and start from the third step.

Please refer to the celloracle paper for scientific premise and the detail of the algorithm of celloracle.

The jupyter notebook files in this tutorial are available [here](https://github.com/morris-lab/CellOracle/tree/master/docs/notebooks) .

### 1. ATAC-seq data preprocessing[¶](#atac-seq-data-preprocessing)

In this step, we process scATAC-seq data (or bulk ATAC-seq data) to obtain the accessible promoter/enhancer DNA sequence. We can get the active proximal promoter/enhancer genome sequences by picking up the ATAC-seq peaks that exist around the transcription starting site (TSS). Distal cis-regulatory elements can be picked up using [Cicero](https://cole-trapnell-lab.github.io/cicero-release/docs/#installing-cicero) . Cicero analyzes scATAC-seq data to calculate a co-accessible score between peaks. We can identify cis-regulatory elements using Cicero’s co-access score and TSS information.

If you have bulk ATAC-seq data instead of scATAC-data, we’ll get only the proximal promoter/enhancer genome sequences.

#### A. Extract TF binding information from scATAC-seq data[¶](#a-extract-tf-binding-information-from-scatac-seq-data)

If you have scATAC-seq data, you can get information on the distal cis-regulatory elements. This step uses Cicero and does not use celloracle. Please refer to [the documentation of Cicero](https://cole-trapnell-lab.github.io/cicero-release/) for the detailed usage.

R notebook

##### 0. Import library[¶](#0.-Import-library)

[2]:

library(cicero)

##### 1. Prepare data[¶](#1.-Prepare-data)

In this tutorial we’ll use acATAC-seq data from the 10x genomics database. You do not need to download these data if you analyze your own scATAC-seq data.

[4]:

# Create folder to store data  
dir.create("data")  
  
# Download demo dataset from 10x genomics  
system("wget -O data/matrix.tar.gz http://cf.10xgenomics.com/samples/cell-atac/1.1.0/atac\_v1\_E18\_brain\_fresh\_5k/atac\_v1\_E18\_brain\_fresh\_5k\_filtered\_peak\_bc\_matrix.tar.gz")  
  
# Unzip data  
system("tar -xvf data/matrix.tar.gz -C data")

[6]:

# You can substitute the data path below with the data path of your scATAC data.  
data\_folder <- "data/filtered\_peak\_bc\_matrix"  
  
# Create a folder to save results  
output\_folder <- "cicero\_output"  
dir.create(output\_folder)

##### 2. Load data and make Cell Data Set (CDS) object[¶](#2.-Load-data-and-make-Cell-Data-Set-(CDS)-object)

###### 2.1. Process data to make CDS object[¶](#2.1.-Process-data-to-make-CDS-object)

[7]:

# Read in matrix data using the Matrix package  
indata <- Matrix::readMM(paste0(data\_folder, "/matrix.mtx"))  
# binarize the matrix  
indata@x[indata@x > 0] <- 1  
  
# Format cell info  
cellinfo <- read.table(paste0(data\_folder, "/barcodes.tsv"))  
row.names(cellinfo) <- cellinfo$V1  
names(cellinfo) <- "cells"  
  
# Format peak info  
peakinfo <- read.table(paste0(data\_folder, "/peaks.bed"))  
names(peakinfo) <- c("chr", "bp1", "bp2")  
peakinfo$site\_name <- paste(peakinfo$chr, peakinfo$bp1, peakinfo$bp2, sep="\_")  
row.names(peakinfo) <- peakinfo$site\_name  
  
row.names(indata) <- row.names(peakinfo)  
colnames(indata) <- row.names(cellinfo)  
  
# Make CDS  
input\_cds <- suppressWarnings(newCellDataSet(indata,  
 phenoData = methods::new("AnnotatedDataFrame", data = cellinfo),  
 featureData = methods::new("AnnotatedDataFrame", data = peakinfo),  
 expressionFamily=VGAM::binomialff(),  
 lowerDetectionLimit=0))  
input\_cds@expressionFamily@vfamily <- "binomialff"  
input\_cds <- monocle::detectGenes(input\_cds)  
  
#Ensure there are no peaks included with zero reads  
input\_cds <- input\_cds[Matrix::rowSums(exprs(input\_cds)) >= 100,]

##### 3. Qauality check and Filtering[¶](#3.-Qauality-check-and-Filtering)

[8]:

# Visualize peak\_count\_per\_cell  
hist(Matrix::colSums(exprs(input\_cds)))

<_images/notebooks_01_ATAC-seq_data_processing_option1_scATAC-seq_data_analysis_with_cicero_01_atacdata_to_cicero_8_0.png>

[9]:

# Filter cells by peak\_count  
max\_count <- 15000 # Please change the threshold value according to the distribution of the peak\_count of your data  
min\_count <- 2000 # Please change the threshold value according to the distribution of the peak\_count of your data  
input\_cds <- input\_cds[,Matrix::colSums(exprs(input\_cds)) >= min\_count]  
input\_cds <- input\_cds[,Matrix::colSums(exprs(input\_cds)) <= max\_count]

##### 4. Process cicero-CDS object[¶](#4.-Process-cicero-CDS-object)

[10]:

# Run cicero to get cis-regulatory networks  
set.seed(2017)  
input\_cds <- detectGenes(input\_cds)  
input\_cds <- estimateSizeFactors(input\_cds)  
  
input\_cds <- reduceDimension(input\_cds, max\_components = 2, verbose=T,scaling = FALSE,relative\_expr=FALSE,  
 reduction\_method = 'tSNE', norm\_method = "none")  
  
tsne\_coords <- t(reducedDimA(input\_cds))  
row.names(tsne\_coords) <- row.names(pData(input\_cds))  
cicero\_cds <- make\_cicero\_cds(input\_cds, reduced\_coordinates = tsne\_coords)  
  
# Save cicero-CDS object if you want.  
#saveRDS(cicero\_cds, paste0(output\_folder, "/cicero\_cds.Rds"))

Remove noise by PCA ...  
  
Reduce dimension by tSNE ...  
  
Overlap QC metrics:  
Cells per bin: 50  
Maximum shared cells bin-bin: 44  
Mean shared cells bin-bin: 0.76256263875674  
Median shared cells bin-bin: 0

##### 5. Run cicero to get cis-regulatory connection scores[¶](#5.-Run-cicero-to-get-cis-regulatory-connection-scores)

[11]:

# Import genome length, which is needed for the function, run\_cicero  
mm10\_chromosome\_length <- read.table("./mm10\_chromosome\_length.txt")  
  
# Run the main function  
conns <- run\_cicero(cicero\_cds, mm10\_chromosome\_length) # Takes a few minutes to run  
  
# Check results  
head(conns)

[1] "Starting Cicero"  
[1] "Calculating distance\_parameter value"  
[1] "Running models"  
[1] "Assembling connections"  
[1] "Done"

A data.frame: 6 × 3

Peak1

Peak2

coaccess

<fct>

<fct>

<dbl>

2

chr1\_3094484\_3095479

chr1\_3113499\_3113979

-0.316289004

3

chr1\_3094484\_3095479

chr1\_3119478\_3121690

-0.419240532

4

chr1\_3094484\_3095479

chr1\_3399730\_3400368

-0.050867246

5

chr1\_3113499\_3113979

chr1\_3094484\_3095479

-0.316289004

7

chr1\_3113499\_3113979

chr1\_3119478\_3121690

0.370342744

8

chr1\_3113499\_3113979

chr1\_3399730\_3400368

-0.009276026

##### 6. Save results for next step[¶](#6.-Save-results-for-next-step)

[ ]:

all\_peaks <- row.names(exprs(input\_cds))  
write.csv(x = all\_peaks, file = paste0(output\_folder, "/all\_peaks.csv"))  
write.csv(x = conns, file = paste0(output\_folder, "/cicero\_connections.csv"))

Next, the results of Cicero analysis will be processed to make TSS annotations.

Python notebook

In this notebook, we process the results of cicero analysis to get active promoter/enhancer DNA peaks. First, we pick up peaks around the transcription starting site (TSS). Second, we merge cicero data with the peaks around TSS. Then we remove peaks that have a weak connection to TSS peak so that the final product includes TSS peaks and peaks that have a strong connection with the TSS peaks. We use this information as an active promoter/enhancer elements.

##### 0. Import libraries[¶](#0.-Import-libraries)

[1]:

import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
%matplotlib inline  
  
import seaborn as sns  
  
  
import os, sys, shutil, importlib, glob  
from tqdm import tqdm\_notebook as tqdm  
  
from celloracle import motif\_analysis as ma

[2]:

%config InlineBackend.figure\_format = 'retina'  
  
plt.rcParams['figure.figsize'] = [6, 4.5]  
plt.rcParams["savefig.dpi"] = 300

##### 1. Load data made with cicero[¶](#1.-Load-data-made-with-cicero)

[3]:

# Load all peaks  
peaks = pd.read\_csv("cicero\_output/all\_peaks.csv", index\_col=0)  
peaks = peaks.x.values  
peaks

[3]:

array(['chr1\_3094484\_3095479', 'chr1\_3113499\_3113979',  
 'chr1\_3119478\_3121690', ..., 'chrY\_90804622\_90805450',  
 'chrY\_90808626\_90809117', 'chrY\_90810560\_90811167'], dtype=object)

[4]:

# Load cicero results  
cicero\_connections = pd.read\_csv("cicero\_output/cicero\_connections.csv", index\_col=0)  
cicero\_connections.head()

/home/k/anaconda3/envs/test/lib/python3.6/site-packages/numpy/lib/arraysetops.py:568: FutureWarning: elementwise comparison failed; returning scalar instead, but in the future will perform elementwise comparison  
 mask |= (ar1 == a)

[4]:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Peak1 | Peak2 | coaccess |
| 2 | chr1\_3094484\_3095479 | chr1\_3113499\_3113979 | -0.316289 |
| 3 | chr1\_3094484\_3095479 | chr1\_3119478\_3121690 | -0.419241 |
| 4 | chr1\_3094484\_3095479 | chr1\_3399730\_3400368 | -0.050867 |
| 5 | chr1\_3113499\_3113979 | chr1\_3094484\_3095479 | -0.316289 |
| 7 | chr1\_3113499\_3113979 | chr1\_3119478\_3121690 | 0.370343 |

##### 2. Make TSS annotation[¶](#2.-Make-TSS-annotation)

###### IMPORTANT: Please make sure that you are setting correct reference genoms.[¶](#IMPORTANT:-Please-make-sure-that-you-are-setting-correct-reference-genoms.)

[5]:

tss\_annotated = ma.get\_tss\_info(peak\_str\_list=peaks, ref\_genome="mm10")  
  
# Check results  
tss\_annotated.tail()

que bed peaks: 72402  
tss peaks in que: 16987

[5]:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | chr | start | end | gene\_short\_name | strand |
| 16982 | chr1 | 55130650 | 55132118 | Mob4 | + |
| 16983 | chr6 | 94499875 | 94500767 | Slc25a26 | + |
| 16984 | chr19 | 45659222 | 45660823 | Fbxw4 | - |
| 16985 | chr12 | 100898848 | 100899597 | Gpr68 | - |
| 16986 | chr4 | 129491262 | 129492047 | Fam229a | - |

##### 3. Integrate TSS info and cicero connections[¶](#3.-Integrate-TSS-info-and-cicero-connections)

The output file after the integration process has three columns; “peak\_id”, “gene\_short\_name”, and “coaccess”. “peak\_id” is either the TSS peak or the peaks that have a connection with the TSS peak. “gene\_short\_name” is the gene name that associated with the TSS site. “coaccess” is the co-access score between a peak and TSS peak. Note, the TSS peak is indicated by a score of 1.

[8]:

integrated = ma.integrate\_tss\_peak\_with\_cicero(tss\_peak=tss\_annotated,  
 cicero\_connections=cicero\_connections)  
print(integrated.shape)  
integrated.head()

(263279, 3)

[8]:

|  |  |  |  |
| --- | --- | --- | --- |
|  | peak\_id | gene\_short\_name | coaccess |
| 0 | chr10\_100015291\_100017830 | Kitl | 1.000000 |
| 1 | chr10\_100018677\_100020384 | Kitl | 0.086299 |
| 2 | chr10\_100050858\_100051762 | Kitl | 0.034558 |
| 3 | chr10\_100052829\_100053395 | Kitl | 0.167188 |
| 4 | chr10\_100128086\_100128882 | Tmtc3 | 0.022341 |

##### 4. Filter peaks[¶](#4.-Filter-peaks)

Remove peaks that have weak coaccess score.

[9]:

peak = integrated[integrated.coaccess >= 0.8]  
peak = peak[["peak\_id", "gene\_short\_name"]].reset\_index(drop=True)

[10]:

print(peak.shape)  
peak.head()

(15680, 2)

[10]:

|  |  |  |
| --- | --- | --- |
|  | peak\_id | gene\_short\_name |
| 0 | chr10\_100015291\_100017830 | Kitl |
| 1 | chr10\_100486534\_100488209 | Tmtc3 |
| 2 | chr10\_100588641\_100589556 | 4930430F08Rik |
| 3 | chr10\_100741247\_100742505 | Gm35722 |
| 4 | chr10\_101681379\_101682124 | Mgat4c |

##### 5. Save data[¶](#5.-Save-data)

Save the promoter/enhancer peak.

[11]:

peak.to\_parquet("peak\_file.parquet")

-> go to next notebook

#### B. Extract TF binding information from bulk ATAC-seq data or Chip-seq data[¶](#b-extract-tf-binding-information-from-bulk-atac-seq-data-or-chip-seq-data)

Bulk DNA-seq data can be used to get the accessible promoter/enhancer sequences.

Python notebook

##### 0. Import libraries[¶](#0.-Import-libraries)

[1]:

import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
%matplotlib inline  
  
import seaborn as sns  
  
import os, sys, shutil, importlib, glob  
from tqdm import tqdm\_notebook as tqdm  
  
%config InlineBackend.figure\_format = 'retina'  
  
plt.rcParams['figure.figsize'] = [6, 4.5]  
plt.rcParams["savefig.dpi"] = 300

[2]:

# Import celloracle function  
from celloracle import motif\_analysis as ma

##### 1. Load bed file[¶](#1.-Load-bed-file)

Import ATAC-seq bed file. This script can also be used with DNase-seq or Chip-seq data.

[3]:

file\_path\_of\_bed\_file = "data/all\_peaks.bed"

[4]:

# Load bed\_file  
bed = ma.read\_bed(file\_path\_of\_bed\_file)  
print(bed.shape)  
bed.head()

(436206, 4)

[4]:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | chrom | start | end | seqname |
| 0 | chr1 | 3002478 | 3002968 | chr1\_3002478\_3002968 |
| 1 | chr1 | 3084739 | 3085712 | chr1\_3084739\_3085712 |
| 2 | chr1 | 3103576 | 3104022 | chr1\_3103576\_3104022 |
| 3 | chr1 | 3106871 | 3107210 | chr1\_3106871\_3107210 |
| 4 | chr1 | 3108932 | 3109158 | chr1\_3108932\_3109158 |

[6]:

# Convert bed file into peak name list  
peaks = ma.process\_bed\_file.df\_to\_list\_peakstr(bed)  
peaks

[6]:

array(['chr1\_3002478\_3002968', 'chr1\_3084739\_3085712',  
 'chr1\_3103576\_3104022', ..., 'chrY\_631222\_631480',  
 'chrY\_795887\_796426', 'chrY\_2397419\_2397628'], dtype=object)

##### 2. Make TSS annotation[¶](#2.-Make-TSS-annotation)

IMPORTANT: Please make sure that you are setting the correct ref genome!

[7]:

tss\_annotated = ma.get\_tss\_info(peak\_str\_list=peaks, ref\_genome="mm9")  
  
# Check results  
tss\_annotated.tail()

que bed peaks: 436206  
tss peaks in que: 24822

[7]:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | chr | start | end | gene\_short\_name | strand |
| 24817 | chr2 | 60560211 | 60561602 | Itgb6 | - |
| 24818 | chr15 | 3975177 | 3978654 | BC037032 | - |
| 24819 | chr14 | 67690701 | 67692101 | Ppp2r2a | - |
| 24820 | chr17 | 48455247 | 48455773 | B430306N03Rik | + |
| 24821 | chr10 | 59861192 | 59861608 | Gm17455 | + |

[9]:

# Change format  
peak\_id\_tss = ma.process\_bed\_file.df\_to\_list\_peakstr(tss\_annotated)  
tss\_annotated = pd.DataFrame({"peak\_id": peak\_id\_tss,  
 "gene\_short\_name": tss\_annotated.gene\_short\_name.values})  
tss\_annotated = tss\_annotated.reset\_index(drop=True)  
print(tss\_annotated.shape)  
tss\_annotated.head()

(24822, 2)

[9]:

|  |  |  |
| --- | --- | --- |
|  | peak\_id | gene\_short\_name |
| 0 | chr7\_50691730\_50692032 | Nkg7 |
| 1 | chr7\_50692077\_50692785 | Nkg7 |
| 2 | chr13\_93564413\_93564836 | Thbs4 |
| 3 | chr13\_14613429\_14615645 | Hecw1 |
| 4 | chr3\_99688753\_99689665 | Spag17 |

##### 3. Save data[¶](#3.-Save-data)

[10]:

tss\_annotated.to\_parquet("peak\_file.parquet")

-> go to next notebook

### 2. Transcription factor binding motif scan[¶](#transcription-factor-binding-motif-scan)

We identified accessible Promoter/enhancer DNA regions using ATAC-seq data. Next, we will obtain a list of TFs for each target gene by scanning the regulatory genomic sequences for TF-binding motifs. In the later GRN inference process, this list will be used to define potential regulatory connections.

Python notebook

#### 0. Import libraries[¶](#0.-Import-libraries)

[2]:

import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
%matplotlib inline  
  
import seaborn as sns  
  
import os, sys, shutil, importlib, glob  
from tqdm import tqdm\_notebook as tqdm  
  
%config InlineBackend.figure\_format = 'retina'  
  
plt.rcParams['figure.figsize'] = (15,7)  
plt.rcParams["savefig.dpi"] = 600

[3]:

from celloracle import motif\_analysis as ma  
from celloracle.utility import save\_as\_pickled\_object

#### 1. Load data[¶](#1.-Load-data)

[4]:

# Load annotated peak data.  
peaks = pd.read\_parquet("../01\_ATAC-seq\_data\_processing/option1\_scATAC-seq\_data\_analysis\_with\_cicero/peak\_file.parquet")  
peaks.head()

[4]:

|  |  |  |
| --- | --- | --- |
|  | peak\_id | gene\_short\_name |
| 0 | chr10\_100015291\_100017830 | Kitl |
| 1 | chr10\_100486534\_100488209 | Tmtc3 |
| 2 | chr10\_100588641\_100589556 | 4930430F08Rik |
| 3 | chr10\_100741247\_100742505 | Gm35722 |
| 4 | chr10\_101681379\_101682124 | Mgat4c |

#### 2. Check data[¶](#2.-Check-data)

[5]:

# Check data  
print(f"number of peak: {len(peaks.peak\_id.unique())}")  
  
def getLength(x):  
 a, b, c = x["peak\_id"].split("\_")  
 return int(c) - int(b)  
  
df = peaks.apply(lambda x: getLength(x), axis=1)  
print(f"mean peak length: {df.values.mean()}")

number of peak: 13919  
mean peak length: 1756.1744260204082

##### 2.1. Remove short peaks[¶](#2.1.-Remove-short-peaks)

Short DNA fragment that are less than 5 bases, cannot be used for motif scanning. Therefore, we will remove the short DNA fragments.

[6]:

peaks = peaks[df>=5]

#### 3. Instantiate TFinfo object and search for TF binding motifs[¶](#3.-Instantiate-TFinfo-object-and-search-for-TF-binding-motifs)

The motif analysis module has a custom class; TFinfo. The TFinfo object converts a peak data into a DNA sequences and scans the DNA sequences searching for TF binding motifs. Then, the results of motif scan will be filtered and converted into either a python dictionary or a depending on your preference. This TF information is necessary for GRN inference.

#### 3.1 check reference genome installation[¶](#3.1-check-reference-genome-installation)

[7]:

# PLEASE make sure that you are setting correct ref genome.  
ref\_genome = "mm10"  
  
ma.is\_genome\_installed(ref\_genome=ref\_genome)

genome mm10 is not installed in this environment.  
Please install genome using genomepy.  
e.g.  
 >>> import genomepy  
 >>> genomepy.install\_genome("mm9", "UCSC")

[7]:

False

##### 3.2. Install reference genome (if refgenome is not installed)[¶](#3.2.-Install-reference-genome-(if-refgenome-is-not-installed))

[9]:

import genomepy  
genomepy.install\_genome(ref\_genome, "UCSC")

downloading from http://hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/chromFa.tar.gz...  
done...  
name: mm10  
local name: mm10  
fasta: /home/k/.local/share/genomes/mm10/mm10.fa

[9]:

# check again  
ma.is\_genome\_installed(ref\_genome=ref\_genome)

[9]:

True

[14]:

# Instantiate TFinfo object  
tfi = ma.TFinfo(peak\_data\_frame=peaks, # peak info calculated from ATAC-seq data  
 ref\_genome=ref\_genome)

#### 4. Scan motifs and save object[¶](#4.-Scan-motifs-and-save-object)

This step may take long time

[15]:

%%time  
# Scan motifs  
tfi.scan(fpr=0.02, verbose=True)  
  
# Save tfinfo object  
tfi.to\_hdf5(file\_path="test.celloracle.tfinfo")

initiating scanner ...

2019-09-22 23:00:18,604 - INFO - Using background: genome mm10 with length 200  
2019-09-22 23:00:18,986 - INFO - Determining FPR-based threshold

getting DNA sequences ...  
scanning motifs ...

CPU times: user 52min 23s, sys: 36.8 s, total: 53min  
Wall time: 52min 58s

[16]:

# Check motif scan results  
tfi.scanned\_df.head()

[16]:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | seqname | motif\_id | factors\_direct | factors\_indirect | score | pos | strand |
| 0 | chr10\_100015291\_100017830 | GM.5.0.Homeodomain.0001 | TGIF1 | ENSG00000234254, TGIF1 | 10.311002 | 1003 | 1 |
| 1 | chr10\_100015291\_100017830 | GM.5.0.Mixed.0001 |  | SRF, EGR1 | 7.925873 | 481 | 1 |
| 2 | chr10\_100015291\_100017830 | GM.5.0.Mixed.0001 |  | SRF, EGR1 | 7.321375 | 911 | -1 |
| 3 | chr10\_100015291\_100017830 | GM.5.0.Mixed.0001 |  | SRF, EGR1 | 7.276585 | 811 | -1 |
| 4 | chr10\_100015291\_100017830 | GM.5.0.Nuclear\_receptor.0002 | NR2C2 | NR2C2, Nr2c2 | 9.067331 | 449 | -1 |

We have the score for each sequence and motif\_id pair. In the next step we will filter the motifs with low score.

#### 5. Filtering motifs[¶](#5.-Filtering-motifs)

[17]:

# Reset filtering  
tfi.reset\_filtering()  
  
# Do filtering  
tfi.filter\_motifs\_by\_score(threshold=10.5)  
  
# Do post filtering process. Convert results into several file format.  
tfi.make\_TFinfo\_dataframe\_and\_dictionary(verbose=True)

peaks were filtered: 12934005 -> 2285279  
1. converting scanned results into one-hot encoded dataframe.

2. converting results into dictionaries.  
converting scan results into dictionaries...

#### 6. Get Final results[¶](#6.-Get-Final-results)

##### 6.1. Get resutls as a dictionary[¶](#6.1.-Get-resutls-as-a-dictionary)

[18]:

td = tfi.to\_dictionary(dictionary\_type="targetgene2TFs")

##### 6.2. Get results as a dataframe[¶](#6.2.-Get-results-as-a-dataframe)

[20]:

df = tfi.to\_dataframe()  
df.head()

[20]:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | peak\_id | gene\_short\_name | 9430076c15rik | Ac002126.6 | Ac012531.1 | Ac226150.2 | Afp | Ahr | Ahrr | Aire | ... | Znf784 | Znf8 | Znf816 | Znf85 | Zscan10 | Zscan16 | Zscan22 | Zscan26 | Zscan31 | Zscan4 |
| 0 | chr10\_100015291\_100017830 | Kitl | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 1 | chr10\_100486534\_100488209 | Tmtc3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 2 | chr10\_100588641\_100589556 | 4930430F08Rik | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | chr10\_100741247\_100742505 | Gm35722 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | chr10\_101681379\_101682124 | Mgat4c | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

5 rows × 1092 columns

#### 7. Save TFinfo as dictionary or dataframe[¶](#7.-Save-TFinfo-as-dictionary-or-dataframe)

We’ll use this information when making the GRNs. Save the results.

[21]:

folder = "TFinfo\_outputs"  
os.makedirs(folder, exist\_ok=True)  
  
# save TFinfo as a dictionary  
td = tfi.to\_dictionary(dictionary\_type="targetgene2TFs")  
save\_as\_pickled\_object(td, os.path.join(folder, "TFinfo\_targetgene2TFs.pickled"))  
  
# save TFinfo as a dataframe  
df = tfi.to\_dataframe()  
df.to\_parquet(os.path.join(folder, "TFinfo\_dataframe.parquet"))

### 3. Single-cell RNA-seq data preprocessing[¶](#single-cell-rna-seq-data-preprocessing)

Network analysis and simulation in celloracle will be performed using scRNA-seq data. The scRNA-seq data should include the components below.

* Gene expression matrix; mRNA counts before scaling and transformation.
* Clustering results.
* Dimensional reduction results.

In addition to these minimum requirements, we highly recommend doing these analyses below in the preprocessing step.

* Data quality check and cell/gene filtering.
* Normalization
* Identification of highly variable genes

We recommend processing scRNA-seq data using either Scanpy or Seurat. If you are not familiar with the general workflow of scRNA-seq data processing, please go to [the documentation for scanpy](https://scanpy.readthedocs.io/en/stable/) and [the documentation for Seurat](https://satijalab.org/seurat/vignettes.html) before celloracle analysis.

If you already have preprocessed scRNA-seq data, which includes the necessary information above, you can skip this part.

#### A. scRNA-seq data preprocessing with scanpy[¶](#a-scrna-seq-data-preprocessing-with-scanpy)

scanpy is a python library for the analysis of scRNA-seq data.

In this tutorial, we introduce an example of scRNA-seq preprocessing for celloracle with scanpy. We wrote the notebook based on [one of scanpy’s tutorials](https://scanpy-tutorials.readthedocs.io/en/latest/paga-paul15.html) with some modifications.

Python notebook

##### 0. Import libraries[¶](#0.-Import-libraries)

[1]:

import os  
import matplotlib.pyplot as plt  
import numpy as np  
import pandas as pd  
import scanpy as sc

[2]:

%matplotlib inline  
%config InlineBackend.figure\_format = 'retina'  
plt.rcParams["savefig.dpi"] = 300  
plt.rcParams["figure.figsize"] = [6, 4.5]

##### 1. Load data[¶](#1.-Load-data)

In this notebook, we will show an example of how to process scRNA-seq data using a scRNA-seq data of hematopoiesis (Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., et al. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell, 163(7), 1663–1677. <http://doi.org/10.1016/j.cell.2015.11.013>). You can easily download this scRNA-seq data with a scanpy function.

Please change the code below if you want to use your data.

[3]:

# Download dataset. You can change the code blow if you use another data.  
adata = sc.datasets.paul15()

WARNING: In Scanpy 0.\*, this returned logarithmized data. Now it returns non-logarithmized data.

... storing 'paul15\_clusters' as categorical  
Trying to set attribute `.uns` of view, making a copy.

##### 2. Filtering[¶](#2.-Filtering)

[4]:

# Only consider genes with more than 1 count  
sc.pp.filter\_genes(adata, min\_counts=1)

##### 3. Normalization[¶](#3.-Normalization)

[5]:

# Normalize gene expression matrix with total UMI count per cell  
sc.pp.normalize\_per\_cell(adata, key\_n\_counts='n\_counts\_all')

##### 4. Identification of highly variable genes[¶](#4.-Identification-of-highly-variable-genes)

Removing non-variable genes not only reduces the calculation time during the GRN reconstruction and simulation, but also improve the accuracy of GRN inference. We recommend using the top 2000~3000 variable genes.

[6]:

# Select top 2000 highly-variable genes  
filter\_result = sc.pp.filter\_genes\_dispersion(adata.X,  
 flavor='cell\_ranger',  
 n\_top\_genes=2000,  
 log=False)  
  
# Subset the genes  
adata = adata[:, filter\_result.gene\_subset]  
  
# Renormalize after filtering  
sc.pp.normalize\_per\_cell(adata)

Trying to set attribute `.obs` of view, making a copy.

##### 5. Log transformation[¶](#5.-Log-transformation)

We will do log transformation scaling because these are necessary for PCA, clustering, and differential gene calculations. However, we also need non-transformed gene expression data in the celloracle analysis. Thus we keep raw count in anndata using the following command before the log transformation.

[7]:

# keep raw cont data before log transformation  
adata.raw = adata  
  
# Log transformation and scaling  
sc.pp.log1p(adata)  
sc.pp.scale(adata)

##### 6. Dimensional reduction[¶](#6.-Dimensional-reduction)

Dimensional reduction is one of the most important parts of the scRNA-seq analysis. Celloracle needs dimensional reduction embeddings to simulate cell transition.

Please choose a proper algorithm for dimensional reduction so that the embedding appropriately represents the data structure. We recommend using one of these dimensional reduction algorithms (or trajectory inference algorithms); UMAP, tSNE, diffusion map, force-directed graph drawing or PAGA.

In this example, we use a combination of four algorithms; diffusion map, force-directed graph drawing, and PAGA.

[9]:

# PCA  
sc.tl.pca(adata, svd\_solver='arpack')

[10]:

# Diffusion map  
sc.pp.neighbors(adata, n\_neighbors=4, n\_pcs=20)  
  
sc.tl.diffmap(adata)  
# Calculate neihbors again based on diffusionmap  
sc.pp.neighbors(adata, n\_neighbors=10, use\_rep='X\_diffmap')

##### 7. Clustering[¶](#7.-Clustering)

[11]:

sc.tl.louvain(adata, resolution=0.8)

##### (Optional) Re-calculate Dimensional reduction graph[¶](#(Optional)-Re-calculate-Dimensional-reduction-graph)

[12]:

# PAGA graph construction  
sc.tl.paga(adata, groups='louvain')

[13]:

# Check current cluster name  
cluster\_list = adata.obs.louvain.unique()  
cluster\_list

[13]:

[5, 2, 12, 13, 0, ..., 6, 20, 14, 15, 21]  
Length: 23  
Categories (23, object): [5, 2, 12, 13, ..., 20, 14, 15, 21]

[14]:

plt.rcParams["figure.figsize"] = [6, 4.5]

[15]:

sc.pl.paga(adata)

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_22_0.png>

[16]:

sc.tl.draw\_graph(adata, init\_pos='paga', random\_state=123)

[17]:

sc.pl.draw\_graph(adata, color='louvain', legend\_loc='on data')

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_24_0.png>

##### 8. Check data[¶](#8.-Check-data)

###### 8.1. Visualize marker gene expression[¶](#8.1.-Visualize-marker-gene-expression)

[18]:

plt.rcParams["figure.figsize"] = [4.5, 4.5]

[19]:

markers = {"Erythroids":["Gata1", "Klf1", "Gypa", "Hba-a2"],  
 "Megakaryocytes":["Itga2b", "Pbx1", "Sdpr", "Vwf"],  
 "Granulocytes":["Elane", "Cebpe", "Ctsg", "Mpo", "Gfi1"],  
 "Monocytes":["Irf8", "Csf1r", "Ctsg", "Mpo"],  
 "Mast\_cells":["Cma1", "Gzmb", "Kit"],  
 "Basophils":["Mcpt8", "Prss34"]  
 }  
  
for cell\_type, genes in markers.items():  
 print(f"marker gene of {cell\_type}")  
 sc.pl.draw\_graph(adata, color=genes, use\_raw=False, ncols=2)  
 plt.show()

marker gene of Erythroids

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_27_1.png>

marker gene of Megakaryocytes

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_27_3.png>

marker gene of Granulocytes

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_27_5.png>

marker gene of Monocytes

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_27_7.png>

marker gene of Mast\_cells

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_27_9.png>

marker gene of Basophils

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_27_11.png>

##### 8. Make annotation for cluster[¶](#8.-Make-annotation-for-cluster)

Based on the marker gene expression and previous reports, we will manually annotate each cluster. When using your own data, you will need to annotate the clusters appropriately.

###### 8.1. Make annotation (1)[¶](#8.1.-Make-annotation-(1))

[20]:

sc.pl.draw\_graph(adata, color=['louvain', 'paul15\_clusters'],  
 legend\_loc='on data')

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_31_0.png>

[21]:

# Check current cluster name  
cluster\_list = adata.obs.louvain.unique()  
cluster\_list

[21]:

[5, 2, 12, 13, 0, ..., 6, 20, 14, 15, 21]  
Length: 23  
Categories (23, object): [5, 2, 12, 13, ..., 20, 14, 15, 21]

!! Please change the dictionary below depending on the clustering results. The results may change depending on the execution environment.[¶](#!!-Please-change-the-dictionary-below-depending-on-the-clustering-results.-The-results-may-change-depending-on-the-execution-environment.)

[22]:

# Make anottation dictionary  
annotation = {"MEP":[5],  
 "Erythroids": [15, 10, 16, 9, 8, 14, 19, 3, 12, 18],  
 "Megakaryocytes":[17, 22],  
 "GMP":[11, 1],  
 "late\_GMP" :[0],  
 "Granulocytes":[7, 13, 4],  
 "Monocytes":[6, 2],  
 "DC":[21],  
 "Lymphoid":[20]}  
  
# change dictionary format  
annotation\_rev = {}  
for i in cluster\_list:  
 for k in annotation:  
 if int(i) in annotation[k]:  
 annotation\_rev[i] = k  
  
# check dictionary  
annotation\_rev

[22]:

{'5': 'MEP',  
 '2': 'Monocytes',  
 '12': 'Erythroids',  
 '13': 'Granulocytes',  
 '0': 'late\_GMP',  
 '10': 'Erythroids',  
 '3': 'Erythroids',  
 '18': 'Erythroids',  
 '11': 'GMP',  
 '7': 'Granulocytes',  
 '8': 'Erythroids',  
 '22': 'Megakaryocytes',  
 '16': 'Erythroids',  
 '1': 'GMP',  
 '17': 'Megakaryocytes',  
 '4': 'Granulocytes',  
 '19': 'Erythroids',  
 '9': 'Erythroids',  
 '6': 'Monocytes',  
 '20': 'Lymphoid',  
 '14': 'Erythroids',  
 '15': 'Erythroids',  
 '21': 'DC'}

[23]:

adata.obs["cell\_type"] = [annotation\_rev[i] for i in adata.obs.louvain]

[24]:

# check results  
sc.pl.draw\_graph(adata, color=['cell\_type', 'paul15\_clusters'],  
 legend\_loc='on data')

... storing 'cell\_type' as categorical

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_36_1.png>

###### 8.2. Make annotation (2)[¶](#8.2.-Make-annotation-(2))

We’ll make another annotation manually for each Louvain clusters.

[25]:

sc.pl.draw\_graph(adata, color=['louvain', 'cell\_type'],  
 legend\_loc='on data')

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_38_0.png>

!! Please change the dictionary below depending on the clustering results. The results may change depending on the execution environment.[¶](#id1)

[26]:

annotation\_2 = {'5': 'MEP\_0',  
 '15': 'Ery\_0',  
 '10': 'Ery\_1',  
 '16': 'Ery\_2',  
 '14': 'Ery\_3',  
 '9': 'Ery\_4',  
 '8': 'Ery\_5',  
 '19': 'Ery\_6',  
 '3': 'Ery\_7',  
 '12': 'Ery\_8',  
 '18': 'Ery\_9',  
 '17': 'Mk\_0',  
 '22': 'Mk\_0',  
 '11': 'GMP\_0',  
 '1': 'GMP\_1',  
 '0': 'GMPl\_0',  
 '7': 'Gran\_0',  
 '13': 'Gran\_1',  
 '4': 'Gran\_2',  
 '6': 'Mo\_0',  
 '2': 'Mo\_1',  
 '21': 'DC\_0',  
 '20': 'Lym\_0'}

[27]:

adata.obs["louvain\_annot"] = [annotation\_2[i] for i in adata.obs.louvain]

[28]:

# Check result  
sc.pl.draw\_graph(adata, color=['louvain\_annot', 'cell\_type'],  
 legend\_loc='on data')

... storing 'louvain\_annot' as categorical

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_42_1.png>

We’ve done secveral scRNA-preprocessing steps; filtering, normalization, clustering, and dimensional reduction. In the next step, we’ll do the GRN inference, network analysis, and in silico simulation based on this information.

##### 9. (Option) Subset cells[¶](#9.-(Option)-Subset-cells)

In this tutorial, we are using scRNA-seq data of hematopoiesis. In the latter part, we will focus on the cell fate decision in the myeloid lineage. So we will remove non-myeloid cell cluster; DC and Lymphoid cell cluster.

[29]:

adata.obs.cell\_type.unique()

[29]:

[MEP, Monocytes, Erythroids, Granulocytes, late\_GMP, GMP, Megakaryocytes, Lymphoid, DC]  
Categories (9, object): [MEP, Monocytes, Erythroids, Granulocytes, ..., GMP, Megakaryocytes, Lymphoid, DC]

[30]:

cell\_of\_interest = adata.obs.index[~adata.obs.cell\_type.isin(["Lymphoid", "DC"])]  
adata = adata[cell\_of\_interest, :]

[31]:

# check result  
sc.pl.draw\_graph(adata, color=['louvain\_annot', 'cell\_type'],  
 legend\_loc='on data')

<_images/notebooks_03_scRNA-seq_data_preprocessing_scanpy_preprocessing_with_Paul_etal_2015_data_47_0.png>

##### 10. Save data[¶](#10.-Save-data)

[32]:

adata.write\_h5ad("data/Paul\_etal\_15.h5ad")

#### B. scRNA-seq data preprocessing with Seurat[¶](#b-scrna-seq-data-preprocessing-with-seurat)

R notebook … comming in the future update.

Note

If you use Seurat for preprocessing, you need to convert the scRNA-seq data (Seurat object) into anndata to analyze the data with celloracle. celloracle has a python API and command-line API to convert a Seurat object into an anndata. Please go to the documentation of celloracle’s API documentation for more information.

### 4. Network analysis[¶](#network-analysis)

celloracle imports the scRNA-seq dataset and TF binding information to find active regulatory connections for all genes, generating sample-specific GRNs.

The inferred GRN is analyzed with several network algorithms to get various network scores. The network score is useful to identify key regulatory genes.

Celloracle reconstructs a GRN for each cluster, enabling us to compare GRNs to each other. It is also possible to analyze how the GRN changes over differentiation. The dynamics of the GRN structure can provide us insight into the context-dependent regulatory mechanisms.

Python notebook

#### 0. Import libraries[¶](#0.-Import-libraries)

[1]:

# 0. Import  
  
import os  
import sys  
  
import matplotlib.pyplot as plt  
import numpy as np  
import pandas as pd  
import scanpy as sc  
import seaborn as sns

[2]:

import celloracle as co

[3]:

# visualization settings  
%config InlineBackend.figure\_format = 'retina'  
%matplotlib inline  
  
plt.rcParams['figure.figsize'] = [6, 4.5]  
plt.rcParams["savefig.dpi"] = 300

##### 0.1. Check installation[¶](#0.1.-Check-installation)

Celloracle uses some R libraries in network analysis. Please make sure that all dependent R libraries are installed on your computer. You can test the installation with the following command.

[4]:

co.network\_analysis.test\_R\_libraries\_installation()

checking R library installation: igraph -> OK  
checking R library installation: linkcomm -> OK  
checking R library installation: rnetcarto -> OK

##### 0.2. Make a folder to save graph[¶](#0.2.-Make-a-folder-to-save-graph)

[5]:

save\_folder = "figures"  
os.makedirs(save\_folder, exist\_ok=True)

#### 1. Load data[¶](#1.-Load-data)

##### 1.1. Load processed gene expression data (anndata)[¶](#1.1.-Load-processed-gene-expression-data-(anndata))

Please refer to the previous notebook in the tutorial for an example of how to process scRNA-seq data.

[6]:

# Load data. !!Replace the data path below when you use another data.  
adata = sc.read\_h5ad("../03\_scRNA-seq\_data\_preprocessing/data/Paul\_etal\_15.h5ad")

##### 1.2. Load TF data.[¶](#1.2.-Load-TF-data.)

For the GRN inference, celloracle needs TF information, which contains lists of the regulatory candidate genes. There are several ways to make such TF information. We can generate TF information from scATAC-seq data or bulk ATAC-seq data. Please refer to the first step of the tutorial for the details of this process.

If you do not have your scATAC-seq data, you can use some built-in data in celloracle. The built-in TFinfo wqs made using various tissue/cell-types from the mouse ATAC-seq atlas dataset (<http://atlas.gs.washington.edu/mouse-atac/>).

You can load and use the data with the following command.

[7]:

# Load TF info which was made from mouse cell atlas dataset.  
TFinfo\_df = co.data.load\_TFinfo\_df\_mm9\_mouse\_atac\_atlas()  
  
# Check data  
TFinfo\_df.head()

[7]:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | peak\_id | gene\_short\_name | 9430076c15rik | Ac002126.6 | Ac012531.1 | Ac226150.2 | Afp | Ahr | Ahrr | Aire | ... | Znf784 | Znf8 | Znf816 | Znf85 | Zscan10 | Zscan16 | Zscan22 | Zscan26 | Zscan31 | Zscan4 |
| 0 | chr10\_100050979\_100052296 | 4930430F08Rik | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | chr10\_101006922\_101007748 | SNORA17 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 2 | chr10\_101144061\_101145000 | Mgat4c | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 3 | chr10\_10148873\_10149183 | 9130014G24Rik | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | chr10\_10149425\_10149815 | 9130014G24Rik | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

5 rows × 1095 columns

#### 2. Initiate Oracle object[¶](#2.-Initiate-Oracle-object)

Celloracle has a custom called Oracle. We can use Oracle for the data preprocessing and GRN inference steps. The Oracle object stores all of necessary information and does the calculations with its internal functions. We instantiate an Oracle object, then input the gene expression data (anndata) and a TFinfo into the Oracle object.

[8]:

# Instantiate Oracle object  
oracle = co.Oracle()

##### 2.1. load gene expression data into oracle object.[¶](#2.1.-load-gene-expression-data-into-oracle-object.)

When you load a scRNA-seq data, please enter the name of clustering data and dimensional reduction data. The clustering data should be to be stored in the attribute of “obs” in the anndata. Dimensional reduction data suppose to be stored in the attribute of “obsm” in the anndata. You can check these data by the following command.

If you are not familiar with anndata, please look at the documentation of annata (<https://anndata.readthedocs.io/en/stable/>) or Scanpy (<https://scanpy.readthedocs.io/en/stable/>).

For the celloracle analysis, the anndata shoud include (1) gene expression count, (2) clustering information, (3) trajectory (dimensional reduction embeddings) data. Please refer to another notebook for more information on anndata preprocessing.

[9]:

# show data name in anndata  
print("metadata columns :", list(adata.obs.columns))  
print("dimensional reduction: ", list(adata.obsm.keys()))

metadata columns : ['paul15\_clusters', 'n\_counts\_all', 'n\_counts', 'louvain', 'cell\_type', 'louvain\_annot']  
dimensional reduction: ['X\_diffmap', 'X\_draw\_graph\_fa', 'X\_pca']

[ ]:

# In this notebook, we use raw mRNA count as an input of Oracle object.  
adata.X = adata.raw.X.copy()  
  
# Instantiate Oracle object.  
oracle.import\_anndata\_as\_raw\_count(adata=adata,  
 cluster\_column\_name="louvain\_annot",  
 embedding\_name="X\_draw\_graph\_fa")

##### 2.2. Load TFinfo into oracle object[¶](#2.2.-Load-TFinfo-into-oracle-object)

[ ]:

# You can load TF info dataframe with the following code.  
oracle.import\_TF\_data(TF\_info\_matrix=TFinfo\_df)  
  
# Alternatively, if you saved the informmation as a dictionary, you can use the code below.  
# oracle.import\_TF\_data(TFdict=TFinfo\_dictionary)

##### 2.3. (Optional) Add TF info manually[¶](#2.3.-(Optional)-Add-TF-info-manually)

While we mainly use TF info data made from scATAC-seq data, we can also add additional information about the TF-target gene pair manually.

For example, if there is a study or database that includes specific TF-target pairs, you can use such information in the following way.

###### 2.3.1. Make TF info dictionary manually[¶](#2.3.1.-Make-TF-info-dictionary-manually)

Here, we will introduce how to add TF binding information.

We will start with TF binding data from supplemental table 4 in (<http://doi.org/10.1016/j.cell.2015.11.013>).

In order to import TF data into the Oracle object, we need to convert them into a python dictionary. The dictionary keys will be the target genes, and the values will be the regulatory candidate TFs.

[12]:

# We have TF and its target gene information. This is from a supplemental Fig of Paul et. al, (2015).  
Paul\_15\_data = pd.read\_csv("TF\_data\_in\_Paul15.csv")  
Paul\_15\_data

[12]:

|  |  |  |
| --- | --- | --- |
|  | TF | Target\_genes |
| 0 | Cebpa | Abcb1b, Acot1, C3, Cnpy3, Dhrs7, Dtx4, Edem2, ... |
| 1 | Irf8 | Abcd1, Aif1, BC017643, Cbl, Ccdc109b, Ccl6, d6... |
| 2 | Irf8 | 1100001G20Rik, 4732418C07Rik, 9230105E10Rik, A... |
| 3 | Klf1 | 2010011I20Rik, 5730469M10Rik, Acsl6, Add2, Ank... |
| 4 | Sfpi1 | 0910001L09Rik, 2310014H01Rik, 4632428N05Rik, A... |

[13]:

# Make dictionary: dictionary Key is TF, dictionary Value is list of target genes  
TF\_to\_TG\_dictionary = {}  
  
for TF, TGs in zip(Paul\_15\_data.TF, Paul\_15\_data.Target\_genes):  
 # convert target gene to list  
 TG\_list = TGs.replace(" ", "").split(",")  
 # store target gene list in a dictionary  
 TF\_to\_TG\_dictionary[TF] = TG\_list  
  
# We have to make a dictionary, in which a Key is Target gene and value is TF.  
# We invert the dictionary above using a utility function in celloracle.  
TG\_to\_TF\_dictionary = co.utility.inverse\_dictionary(TF\_to\_TG\_dictionary)

###### 2.3.2. Add TF informatio dictionary into the oracle object[¶](#2.3.2.-Add-TF-informatio-dictionary-into-the-oracle-object)

[14]:

# Add TF information  
oracle.addTFinfo\_dictionary(TG\_to\_TF\_dictionary)

#### 3. Knn imputation[¶](#3.-Knn-imputation)

Celloracle uses almost the same strategy as velocyto for visualizing cell transitions. This process requires KNN imputation in advance.

For the KNN imputation, we need PCA and PC selection first.

##### 3.1. PCA[¶](#3.1.-PCA)

[15]:

# Perform PCA  
oracle.perform\_PCA()  
  
# Select important PCs  
plt.plot(np.cumsum(oracle.pca.explained\_variance\_ratio\_)[:100])  
n\_comps = np.where(np.diff(np.diff(np.cumsum(oracle.pca.explained\_variance\_ratio\_))>0.002))[0][0]  
plt.axvline(n\_comps, c="k")  
print(n\_comps)  
n\_comps = min(n\_comps, 50)

45

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_27_1.png>

##### 3.2. KNN imputation[¶](#3.2.-KNN-imputation)

Estimate the optimal number of nearest neighbors for KNN imputation.

[16]:

n\_cell = oracle.adata.shape[0]  
print(f"cell number is :{n\_cell}")

cell number is :2671

[17]:

k = int(0.025\*n\_cell)  
print(f"Auto-selected k is :{k}")

Auto-selected k is :66

[18]:

oracle.knn\_imputation(n\_pca\_dims=n\_comps, k=k, balanced=True, b\_sight=k\*8,  
 b\_maxl=k\*4, n\_jobs=4)

#### 4. Save and Load.[¶](#4.-Save-and-Load.)

Celloracle has some custom-classes: Links, Oracle and TFinfo. You can save such an object using “to\_hdf5”.

Pleasae use “load\_hdf5” function to load the file.

[19]:

# Save oracle object.  
oracle.to\_hdf5("Paul\_15\_data.celloracle.oracle")

[4]:

# Load file.  
# oracle = co.load\_hdf5("Paul\_15\_data.celloracle.oracle")

#### 4. GRN calculation[¶](#4.-GRN-calculation)

The next step is constructing a cluster-specific GRN for all clusters.

You can calculate GRNs with the “get\_links” function, and the function returns GRNs as a Links object. The Links object stores inferred GRNs and the corresponding metadata. You can do network analysis with the Links object.

The GRN will be calculated for each cluster/sub-group. In the example below, we construct GRN for each unit of the “louvain\_annot” clustering.

The GRNs can be calculated at any arbitrary unit as long as the clustering information is stored in anndata.

[21]:

# check data  
sc.pl.draw\_graph(oracle.adata, color="louvain\_annot")

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_37_0.png>

##### 4.1. Get GRNs[¶](#4.1.-Get-GRNs)

[25]:

%%time  
# Calculate GRN for each population in "louvain\_annot" clustering unit.  
# This step may take long time.  
links = oracle.get\_links(cluster\_name\_for\_GRN\_unit="louvain\_annot", alpha=10,  
 verbose\_level=10, test\_mode=False)

inferring GRN for Ery\_0...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Ery\_1...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Ery\_2...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Ery\_3...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Ery\_4...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Ery\_5...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Ery\_6...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Ery\_7...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Ery\_8...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Ery\_9...  
method: bagging\_ridge  
alpha: 10

inferring GRN for GMP\_0...  
method: bagging\_ridge  
alpha: 10

inferring GRN for GMP\_1...  
method: bagging\_ridge  
alpha: 10

inferring GRN for GMPl\_0...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Gran\_0...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Gran\_1...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Gran\_2...  
method: bagging\_ridge  
alpha: 10

inferring GRN for MEP\_0...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Mk\_0...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Mo\_0...  
method: bagging\_ridge  
alpha: 10

inferring GRN for Mo\_1...  
method: bagging\_ridge  
alpha: 10

CPU times: user 22min 20s, sys: 11.5 s, total: 22min 31s  
Wall time: 47min 55s

##### 4.2. (Optional) Export GRNs[¶](#4.2.-(Optional)-Export-GRNs)

Although celloracle has many functions for network analysis, you can analyze GRNs by hand if you choose. The raw GRN data is stored in the attribute of “links\_dict”.

For example, you can get the GRN for the “Ery\_0” cluster with the following commands.

[11]:

links.links\_dict["Ery\_0"]

[11]:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | source | target | coef\_mean | coef\_abs | p | -logp |
| 0 | Id2 | 0610007L01Rik | 0.000552 | 0.000552 | 5.724488e-01 | 0.242263 |
| 1 | Klf2 | 0610007L01Rik | 0.000000 | 0.000000 | NaN | -0.000000 |
| 2 | Stat5a | 0610007L01Rik | -0.005179 | 0.005179 | 5.294497e-04 | 3.276175 |
| 3 | Elf1 | 0610007L01Rik | 0.002107 | 0.002107 | 1.498884e-01 | 0.824232 |
| 4 | Gata1 | 0610007L01Rik | -0.000700 | 0.000700 | 5.731063e-01 | 0.241765 |
| ... | ... | ... | ... | ... | ... | ... |
| 74460 | Cxxc1 | Zyx | -0.004999 | 0.004999 | 1.578852e-02 | 1.801659 |
| 74461 | Mef2c | Zyx | 0.017708 | 0.017708 | 3.011616e-07 | 6.521200 |
| 74462 | Nfe2 | Zyx | 0.034433 | 0.034433 | 2.548244e-12 | 11.593759 |
| 74463 | Nr3c1 | Zyx | -0.022663 | 0.022663 | 2.265408e-08 | 7.644854 |
| 74464 | Ets1 | Zyx | 0.012826 | 0.012826 | 4.813285e-09 | 8.317558 |

74465 rows × 6 columns

You can export the file as follows.

[ ]:

# Set cluster name  
cluster = "Ery\_0"  
  
# Save as csv  
links.links\_dict[cluster].to\_csv(f"raw\_GRN\_for\_{cluster}.csv")

##### 4.3. (Optional) Change order[¶](#4.3.-(Optional)-Change-order)

The links object has a color information in an attribute, “palette”. This information is used for the visualization

The sample will be visualized in that order. Here we can change the order.

[16]:

# Show the contents of pallete  
links.palette

[16]:

|  |  |
| --- | --- |
|  | palette |
| Ery\_0 | #7D87B9 |
| Ery\_1 | #BEC1D4 |
| Ery\_2 | #D6BCC0 |
| Ery\_3 | #BB7784 |
| Ery\_4 | #8E063B |
| Ery\_5 | #4A6FE3 |
| Ery\_6 | #8595E1 |
| Ery\_7 | #B5BBE3 |
| Ery\_8 | #E6AFB9 |
| Ery\_9 | #E07B91 |
| GMP\_0 | #D33F6A |
| GMP\_1 | #11C638 |
| GMPl\_0 | #8DD593 |
| Gran\_0 | #C6DEC7 |
| Gran\_1 | #EAD3C6 |
| Gran\_2 | #F0B98D |
| MEP\_0 | #0FCFC0 |
| Mk\_0 | #9CDED6 |
| Mo\_0 | #D5EAE7 |
| Mo\_1 | #F3E1EB |

[17]:

# Change the order of pallete  
order = ['MEP\_0', 'Mk\_0','Ery\_0', 'Ery\_1', 'Ery\_2', 'Ery\_3', 'Ery\_4', 'Ery\_5',  
 'Ery\_6', 'Ery\_7', 'Ery\_8', 'Ery\_9','GMP\_0', 'GMP\_1',  
 'GMPl\_0', 'Mo\_0', 'Mo\_1', 'Gran\_0', 'Gran\_1', 'Gran\_2']  
links.palette = links.palette.loc[order]  
links.palette

[17]:

|  |  |
| --- | --- |
|  | palette |
| MEP\_0 | #0FCFC0 |
| Mk\_0 | #9CDED6 |
| Ery\_0 | #7D87B9 |
| Ery\_1 | #BEC1D4 |
| Ery\_2 | #D6BCC0 |
| Ery\_3 | #BB7784 |
| Ery\_4 | #8E063B |
| Ery\_5 | #4A6FE3 |
| Ery\_6 | #8595E1 |
| Ery\_7 | #B5BBE3 |
| Ery\_8 | #E6AFB9 |
| Ery\_9 | #E07B91 |
| GMP\_0 | #D33F6A |
| GMP\_1 | #11C638 |
| GMPl\_0 | #8DD593 |
| Mo\_0 | #D5EAE7 |
| Mo\_1 | #F3E1EB |
| Gran\_0 | #C6DEC7 |
| Gran\_1 | #EAD3C6 |
| Gran\_2 | #F0B98D |

#### 5. Network preprocessing[¶](#5.-Network-preprocessing)

##### 5.1. Filter network edges[¶](#5.1.-Filter-network-edges)

Celloracle utilizes bagging ridge or Bayesian ridge regression to infer gene regulatory networks. These methods provide a network edge strength as a distribution rather than a point value. We can use the distribution to know the certainness of the connection.

We filter the network edges as follows.

1. Remove uncertain network edges based on the p-value.
2. Remove weak network edge. In this tutorial, we pick up the top 2000 edges in terms of network strength.

The raw network data is stored as an attribute, “links\_dict,” while filtered network data is stored in “filtered\_links.” Thus the filtering function keeps raw network information rather than overwriting the data. You can come back to the filtering process to filter the data with different parameters if you want.

[32]:

links.filter\_links(p=0.001, weight="coef\_abs", thread\_number=2000)

##### 5.2. Degree distribution[¶](#5.2.-Degree-distribution)

In the first step, we examine the network degree distribution. Network degree, which is the number of edges for each node, is one of the important metrics used to investigate the network structure (<https://en.wikipedia.org/wiki/Degree_distribution>).

Please keep in mind that the degree distribution may change depending on the filtering threshold.

[50]:

plt.rcParams["figure.figsize"] = [9, 4.5]

[51]:

links.plot\_degree\_distributions(plot\_model=True, save=f"{save\_folder}/degree\_distribution/")

Ery\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_1.png>

Ery\_1

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_3.png>

Ery\_2

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_5.png>

Ery\_3

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_7.png>

Ery\_4

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_9.png>

Ery\_5

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_11.png>

Ery\_6

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_13.png>

Ery\_7

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_15.png>

Ery\_8

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_17.png>

Ery\_9

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_19.png>

GMP\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_21.png>

GMP\_1

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_23.png>

GMPl\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_25.png>

Gran\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_27.png>

Gran\_1

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_29.png>

Gran\_2

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_31.png>

MEP\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_33.png>

Mk\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_35.png>

Mo\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_37.png>

Mo\_1

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_52_39.png>

[52]:

plt.rcParams["figure.figsize"] = [6, 4.5]

##### 5.3. Calculate netowrk score[¶](#5.3.-Calculate-netowrk-score)

Next, we calculate several network score using some R libraries. Please make sure that R libraries are installed in your PC before running the command below.

[39]:

# Calculate network scores. It takes several minutes.  
links.get\_score()

processing... batch 1/5  
Ery\_0: finished.  
Ery\_1: finished.  
Ery\_2: finished.  
Ery\_3: finished.  
processing... batch 2/5  
Ery\_4: finished.  
Ery\_5: finished.  
Ery\_6: finished.  
Ery\_7: finished.  
processing... batch 3/5  
Ery\_8: finished.  
Ery\_9: finished.  
GMP\_0: finished.  
GMP\_1: finished.  
processing... batch 4/5  
GMPl\_0: finished.  
Gran\_0: finished.  
Gran\_1: finished.  
Gran\_2: finished.  
processing... batch 5/5  
MEP\_0: finished.  
Mk\_0: finished.  
Mo\_0: finished.  
Mo\_1: finished.  
the scores are saved in ./louvain\_annot/

The score is stored as a attribute called “merged\_score”, and the score will also be saved in a folder in your computer.

[57]:

links.merged\_score.head()

[57]:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | degree\_all | degree\_in | degree\_out | clustering\_coefficient | clustering\_coefficient\_weighted | degree\_centrality\_all | degree\_centrality\_in | degree\_centrality\_out | betweenness\_centrality | closeness\_centrality | ... | assortative\_coefficient | average\_path\_length | community\_edge\_betweenness | community\_random\_walk | community\_eigenvector | module | connectivity | participation | role | cluster |
| Stat3 | 91 | 0 | 91 | 0.019780 | 0.020122 | 0.167279 | 0.000000 | 0.167279 | 0 | 0.000013 | ... | -0.172207 | 2.475244 | 1 | 1 | 1 | 0 | 4.052023 | 0.632291 | Connector Hub | Ery\_0 |
| Mycn | 28 | 0 | 28 | 0.002646 | 0.001828 | 0.051471 | 0.000000 | 0.051471 | 0 | 0.000004 | ... | -0.172207 | 2.475244 | 2 | 1 | 5 | 2 | 2.347146 | 0.625000 | Connector | Ery\_0 |
| Zbtb1 | 27 | 0 | 27 | 0.005698 | 0.008576 | 0.049632 | 0.000000 | 0.049632 | 0 | 0.000011 | ... | -0.172207 | 2.475244 | 3 | 1 | 5 | 1 | 2.275431 | 0.666667 | Connector | Ery\_0 |
| E2f4 | 186 | 3 | 183 | 0.009474 | 0.011623 | 0.341912 | 0.005515 | 0.336397 | 2788 | 0.000010 | ... | -0.172207 | 2.475244 | 4 | 1 | 3 | 0 | 8.496354 | 0.637760 | Connector Hub | Ery\_0 |
| Ybx1 | 71 | 9 | 62 | 0.027364 | 0.027485 | 0.130515 | 0.016544 | 0.113971 | 1153 | 0.000004 | ... | -0.172207 | 2.475244 | 5 | 11 | 4 | 3 | 5.433857 | 0.564967 | Connector Hub | Ery\_0 |

5 rows × 22 columns

##### 5.4. Save[¶](#5.4.-Save)

Save processed GRN. We use this file in the next notebook; “in silico perturbation with GRNs”.

[42]:

# Save Links object.  
links.to\_hdf5(file\_path="links.celloracle.links")

[9]:

# You can load files with the following command.  
links = co.load\_hdf5(file\_path="links.celloracle.links")

#### 6. Network analysis; Network score for each gene[¶](#6.-Network-analysis;-Network-score-for-each-gene)

The Links class has many functions to visualize network score. See the documentation for the details of the functions.

##### 6.1. Network score in each cluster[¶](#6.1.-Network-score-in-each-cluster)

We have calculated several network scores using different centrality metrics. We can use the centrality score to identify key regulatory genes because centrality is one of the important indicators of network structure (<https://en.wikipedia.org/wiki/Centrality>).

Let’s visualize genes with high network centrality.

[ ]:

# Check cluster name  
links.cluster

[53]:

# Visualize top n-th genes that have high scores.  
links.plot\_scores\_as\_rank(cluster="MEP\_0", n\_gene=30, save=f"{save\_folder}/ranked\_score")

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_64_0.png>

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_64_1.png>

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_64_2.png>

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_64_3.png>

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_64_4.png>

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_64_5.png>

##### 6.2. Network score comparison between two clusters[¶](#6.2.-Network-score-comparison-between-two-clusters)

By comparing network scores between two clusters, we can analyze differences in GRN structure.

[54]:

plt.ticklabel\_format(style='sci',axis='y',scilimits=(0,0))  
links.plot\_score\_comparison\_2D(value="eigenvector\_centrality",  
 cluster1="MEP\_0", cluster2="GMPl\_0",  
 percentile=98, save=f"{save\_folder}/score\_comparison")

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_67_0.png>

[55]:

plt.ticklabel\_format(style='sci',axis='y',scilimits=(0,0))  
links.plot\_score\_comparison\_2D(value="betweenness\_centrality",  
 cluster1="MEP\_0", cluster2="GMPl\_0",  
 percentile=98, save=f"{save\_folder}/score\_comparison")

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_68_0.png>

[56]:

plt.ticklabel\_format(style='sci',axis='y',scilimits=(0,0))  
links.plot\_score\_comparison\_2D(value="degree\_centrality\_all",  
 cluster1="MEP\_0", cluster2="GMPl\_0",  
 percentile=98, save=f"{save\_folder}/score\_comparison")

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_69_0.png>

##### 6.3. Network score dynamics[¶](#6.3.-Network-score-dynamics)

In the following session, we focus on how a gene’s network score changes during the differentiation.

Using Gata2, we will demonstrate how you can visualize networks scores for a single gene.

Gata2 is known to play an essential role in the early MEP and GMP populations. .

[57]:

# Visualize Gata2 network score dynamics  
links.plot\_score\_per\_cluster(goi="Gata2", save=f"{save\_folder}/network\_score\_per\_gene/")

Gata2

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_71_1.png>

##### 6.4. Gene cartography analysis[¶](#6.4.-Gene-cartography-analysis)

Gene cartography is a method for gene network analysis. The method classifies gene into several groups using the network module structure and connections. It provides us an insight about the role and regulatory mechanism for each gene. For more information on gene cartography, please refer to the following paper (<https://www.nature.com/articles/nature03288>).

The gene cartography will be calculated for the GRN in each cluster. Thus we can know how the gene cartography change by comparing the the score between clusters.

[58]:

# Plot cartography as a scatter plot  
links.plot\_cartography\_scatter\_per\_cluster(scatter=True,  
 kde=False,  
 gois=["Gata1", "Gata2", "Sfpi1"],  
 auto\_gene\_annot=False,  
 args\_dot={"n\_levels": 105},  
 args\_line={"c":"gray"}, save=f"{save\_folder}/cartography")

Ery\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_1.png>

Ery\_1

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_3.png>

Ery\_2

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_5.png>

Ery\_3

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_7.png>

Ery\_4

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_9.png>

Ery\_5

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_11.png>

Ery\_6

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_13.png>

Ery\_7

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_15.png>

Ery\_8

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_17.png>

Ery\_9

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_19.png>

GMP\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_21.png>

GMP\_1

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_23.png>

GMPl\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_25.png>

Gran\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_27.png>

Gran\_1

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_29.png>

Gran\_2

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_31.png>

MEP\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_33.png>

Mk\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_35.png>

Mo\_0

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_37.png>

Mo\_1

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_73_39.png>

[66]:

# Plot the summary of cartography analysis  
links.plot\_cartography\_term(goi="Gata2", save=f"{save\_folder}/cartography")

Gata2

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_74_1.png>

#### 7. Network analysis; network score distribution[¶](#7.-Network-analysis;-network-score-distribution)

Next, we visualize the distribution of network score to get insight into the global trend of the GRNs.

##### 7.1. Distribution of network degree[¶](#7.1.-Distribution-of-network-degree)

[60]:

plt.subplots\_adjust(left=0.15, bottom=0.3)  
plt.ylim([0,0.040])  
links.plot\_score\_discributions(values=["degree\_centrality\_all"], method="boxplot", save=f"{save\_folder}")

degree\_centrality\_all

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_77_1.png>

[61]:

plt.subplots\_adjust(left=0.15, bottom=0.3)  
plt.ylim([0, 0.40])  
links.plot\_score\_discributions(values=["eigenvector\_centrality"], method="boxplot", save=f"{save\_folder}")

eigenvector\_centrality

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_78_1.png>

##### 7.2. Distribution of netowrk entolopy[¶](#7.2.-Distribution-of-netowrk-entolopy)

[62]:

plt.subplots\_adjust(left=0.15, bottom=0.3)  
links.plot\_network\_entropy\_distributions(save=f"{save\_folder}")

/home/k/anaconda3/envs/test/lib/python3.6/site-packages/scipy/stats/\_distn\_infrastructure.py:2614: RuntimeWarning: invalid value encountered in true\_divide  
 pk = 1.0\*pk / np.sum(pk, axis=0)  
/home/k/anaconda3/envs/test/lib/python3.6/site-packages/celloracle/network\_analysis/links\_object.py:345: RuntimeWarning: divide by zero encountered in log  
 ent\_norm.append(en/np.log(k[i]))  
/home/k/anaconda3/envs/test/lib/python3.6/site-packages/celloracle/network\_analysis/links\_object.py:345: RuntimeWarning: invalid value encountered in double\_scalars  
 ent\_norm.append(en/np.log(k[i]))

<_images/notebooks_04_Network_analysis_Network_analysis_with_with_Paul_etal_2015_data_80_1.png>

Using the network scores, we could pick up cluster-specific key TFs. Gata2, Gata1, Klf1, E2f1, for example, are known to play an essential role in MEP, and these TFs showed high network score in our GRN.

However, it is important to note that network analysis alone cannot shed light on the specific functions or roles these TFs play in cell fate determination.

In the next section, we will begin to investigate each TF’s contribution to cell fate by running GRN simulations

[ ]:

### 5. Simulation with GRNs[¶](#simulation-with-grns)

celloracle leverage GRNs to simulate signal propagation inside a cell. We can estimate the effect of gene perturbation by the simulation with GRNs.

Additonally, we will combine the signal propagation simulation with a cell state transition simulation. The latter simulation is performed by a python library for RNA-velocity analysis, called velocyto . This analysis may provide an insight into a complex system how TF controls enormous target genes to determines cell fate.

Python notebook

#### 0. Import libraries[¶](#0.-Import-libraries)

##### 0.1. Import public libraries[¶](#0.1.-Import-public-libraries)

[1]:

import os  
import sys  
  
import matplotlib.colors as colors  
import matplotlib.pyplot as plt  
import numpy as np  
import pandas as pd  
import scanpy as sc  
import seaborn as sns

[2]:

import celloracle as co

[3]:

plt.rcParams["font.family"] = "arial"  
plt.rcParams["figure.figsize"] = [9,6]  
%config InlineBackend.figure\_format = 'retina'  
plt.rcParams["savefig.dpi"] = 600  
  
%matplotlib inline

##### 0.1. Make a folder to save graph[¶](#0.1.-Make-a-folder-to-save-graph)

[5]:

# Make folder to save plots  
save\_folder = "figures"  
os.makedirs(save\_folder, exist\_ok=True)

#### 1. Load data[¶](#1.-Load-data)

##### 1.1. Load processed oracle object[¶](#1.1.-Load-processed-oracle-object)

Load the oracle object. See the previous notebook for the notes on how to prepare the oracle object.

[7]:

oracle = co.load\_hdf5("../04\_Network\_analysis/Paul\_15\_data.celloracle.oracle")

##### 1.2. Load inferred GRNs[¶](#1.2.-Load-inferred-GRNs)

In the previous notebook, we calculated GRNs. Now, we will use these GRNs for simulation. We import GRNs which were saved in the Links object.

[8]:

links = co.load\_hdf5("../04\_Network\_analysis/links.celloracle.links")

#### 3. Make predictive models for simulation[¶](#3.-Make-predictive-models-for-simulation)

We will fit ridge regression models again. This process takes less time than the GRN inference in the previous notebook because we only use significant TFs to predict target gene instead of all regulatory candidate TFs.

[12]:

links.filter\_links()  
oracle.get\_cluster\_specific\_TFdict\_from\_Links(links\_object=links)  
oracle.fit\_GRN\_for\_simulation(alpha=10, use\_cluster\_specific\_TFdict=True)

calculating GRN using cluster specicif TF dict...  
calculating GRN in Ery\_0

genes\_in\_gem: 1999  
models made for 1074 genes  
calculating GRN in Ery\_1

genes\_in\_gem: 1999  
models made for 1092 genes  
calculating GRN in Ery\_2

genes\_in\_gem: 1999  
models made for 1064 genes  
calculating GRN in Ery\_3

genes\_in\_gem: 1999  
models made for 1105 genes  
calculating GRN in Ery\_4

genes\_in\_gem: 1999  
models made for 1102 genes  
calculating GRN in Ery\_5

genes\_in\_gem: 1999  
models made for 1116 genes  
calculating GRN in Ery\_6

genes\_in\_gem: 1999  
models made for 1097 genes  
calculating GRN in Ery\_7

genes\_in\_gem: 1999  
models made for 1062 genes  
calculating GRN in Ery\_8

genes\_in\_gem: 1999  
models made for 1117 genes  
calculating GRN in Ery\_9

genes\_in\_gem: 1999  
models made for 1121 genes  
calculating GRN in GMP\_0

genes\_in\_gem: 1999  
models made for 1107 genes  
calculating GRN in GMP\_1

genes\_in\_gem: 1999  
models made for 1104 genes  
calculating GRN in GMPl\_0

genes\_in\_gem: 1999  
models made for 1089 genes  
calculating GRN in Gran\_0

genes\_in\_gem: 1999  
models made for 1067 genes  
calculating GRN in Gran\_1

genes\_in\_gem: 1999  
models made for 1076 genes  
calculating GRN in Gran\_2

genes\_in\_gem: 1999  
models made for 1105 genes  
calculating GRN in MEP\_0

genes\_in\_gem: 1999  
models made for 1152 genes  
calculating GRN in Mk\_0

genes\_in\_gem: 1999  
models made for 1114 genes  
calculating GRN in Mo\_0

genes\_in\_gem: 1999  
models made for 1085 genes  
calculating GRN in Mo\_1

genes\_in\_gem: 1999  
models made for 1074 genes

#### 4. in silico Perturbation-simulation[¶](#4.-in-silico-Perturbation-simulation)

Next, we will simulate the effects of perturbing a single TF to investigate its function and regulatory mechanism. See the celloracle paper for the details and scientific premise on the algorithm.

In this notebook, we’ll show an example of the simulation; we’ll simulate knock-out of Gata1 gene in the hematopoiesis.

Previous studies have shown that Gata1 is one of the TFs that regulates cell fate decisions in myeloid progenitors. Additionally, Gata1 has been shown to affect erythroid cell differentiation.

Here, we will analyze Gata1 for the demonstration of celloracle; Celloracle try to recapitulate the previous findings of Gata1 gene above.

##### 4.1. Check gene expression pattern.[¶](#4.1.-Check-gene-expression-pattern.)

[26]:

# Check gene expression  
goi = "Gata1"  
sc.pl.draw\_graph(oracle.adata, color=[goi, oracle.cluster\_column\_name],  
 layer="imputed\_count", use\_raw=False, cmap="viridis")

<_images/notebooks_05_simulation_Gata1_KO_simulation_with_with_Paul_etal_2015_data_15_0.png>

[33]:

# Plot gene expression in histogram  
sc.get.obs\_df(oracle.adata, keys=[goi], layer="imputed\_count").hist()  
plt.show()

<_images/notebooks_05_simulation_Gata1_KO_simulation_with_with_Paul_etal_2015_data_16_0.png>

##### 4.1. calculate future gene expression after perturbation.[¶](#4.1.-calculate-future-gene-expression-after-perturbation.)

Although you can use any gene expression value for the input of in silico perturbation, we recommend avoiding extreme values which are far from natural gene expression ranges. If you set Gata1 gene expression to 100, for example, it may lead to biologically infeasible results.

Here we simulate Gata1 KO; we predict what happens to the cells if Gata1 gene expression changed into 0.

[34]:

# Enter perturbation conditions to simulate signal propagation after the perturbation.  
oracle.simulate\_shift(perturb\_condition={goi: 0.0},  
 n\_propagation=3)

##### 4.2. calculate transition probability between cells[¶](#4.2.-calculate-transition-probability-between-cells)

In the step above, we simulated simulated future gene expression values after perturbation. This prediction is based on itelative calculations of signal propagations within the GRN.

Next step, we will calculate the probabilityof a cell state transition based on the simulated data. Using the transition probability between cells, we can predict how a cell changes after perturbation.

This transition probability will be used in two ways.

1. Visualization of directed trjectory graph.
2. Markof simulation.

In Step 4.2 and 4.3, we usefunctions imported from the velocytoloom class in velocyto.py. Please see the documentation of VelocytoLoom for more information. <http://velocyto.org/velocyto.py/fullapi/api_analysis.html>

[35]:

# Get transition probability  
oracle.estimate\_transition\_prob(n\_neighbors=200, knn\_random=True, sampled\_fraction=0.5)  
  
# Calculate embedding  
oracle.calculate\_embedding\_shift(sigma\_corr = 0.05)  
  
# Calculate global trend of cell transition  
oracle.calculate\_grid\_arrows(smooth=0.8, steps=(40, 40), n\_neighbors=300)

/home/k/anaconda3/envs/test/lib/python3.6/site-packages/IPython/core/interactiveshell.py:3326: FutureWarning: arrays to stack must be passed as a "sequence" type such as list or tuple. Support for non-sequence iterables such as generators is deprecated as of NumPy 1.16 and will raise an error in the future.  
 exec(code\_obj, self.user\_global\_ns, self.user\_ns)  
WARNING:root:Nans encountered in corrcoef and corrected to 1s. If not identical cells were present it is probably a small isolated cluster converging after imputation.

##### 4.3. Visualization[¶](#4.3.-Visualization)

###### 4.3.1. Detailed directed trajectory graph[¶](#4.3.1.-Detailed-directed-trajectory-graph)

[36]:

plt.figure(None,(6,6))  
quiver\_scale = 40  
  
  
ix\_choice = np.random.choice(oracle.adata.shape[0], size=int(oracle.adata.shape[0]/1.), replace=False)  
  
embedding = oracle.adata.obsm[oracle.embedding\_name]  
  
plt.scatter(embedding[ix\_choice, 0], embedding[ix\_choice, 1],  
 c="0.8", alpha=0.2, s=38, edgecolor=(0,0,0,1), lw=0.3, rasterized=True)  
  
quiver\_kwargs=dict(headaxislength=7, headlength=11, headwidth=8,  
 linewidths=0.35, width=0.0045,edgecolors="k",  
 color=oracle.colorandum[ix\_choice], alpha=1)  
plt.quiver(embedding[ix\_choice, 0], embedding[ix\_choice, 1],  
 oracle.delta\_embedding[ix\_choice, 0], oracle.delta\_embedding[ix\_choice, 1],  
 scale=quiver\_scale, \*\*quiver\_kwargs)  
  
plt.axis("off")  
#plt.savefig(f"{save\_folder}/full\_arrows{goi}.png", transparent=True)

[36]:

(-10815.27020913708, 10950.84121716522, -10711.36365432337, 10949.477199695968)

<_images/notebooks_05_simulation_Gata1_KO_simulation_with_with_Paul_etal_2015_data_22_1.png>

###### 4.3.2. Grid graph[¶](#4.3.2.-Grid-graph)

[37]:

# Plot whole graph  
plt.figure(None,(10,10))  
oracle.plot\_grid\_arrows(quiver\_scale=2.0,  
 scatter\_kwargs\_dict={"alpha":0.35, "lw":0.35,  
 "edgecolor":"0.4", "s":38,  
 "rasterized":True},  
 min\_mass=0.015, angles='xy', scale\_units='xy',  
 headaxislength=2.75,  
 headlength=5, headwidth=4.8, minlength=1.5,  
 plot\_random=False, scale\_type="relative")  
#plt.savefig(f"{save\_folder}/vectorfield\_{goi}.png", transparent=True)

<_images/notebooks_05_simulation_Gata1_KO_simulation_with_with_Paul_etal_2015_data_24_0.png>

##### 4.4. Markov simulation to analyze the effects of perturbation on cell fate transition[¶](#4.4.-Markov-simulation-to-analyze-the-effects-of-perturbation-on-cell-fate-transition)

We can also simulate cell state transition using Markof simulation.

###### 4.4.1. Do simulation[¶](#4.4.1.-Do-simulation)

We will simulate using the parameters, “n\_steps=200” and “n\_duplication=5” in the following example.

To elaborate, this means:

1. We will do 200 times of iterative simulations to predict how the cell changes over time
2. We will repeat 5 rounds of simulations

[83]:

%%time  
# n\_steps is the number of steps in markov simulation.  
# n\_duplication is the number of technical duplication for the simulation  
oracle.run\_markov\_chain\_simulation(n\_steps=200, n\_duplication=5)

CPU times: user 1.33 s, sys: 0 ns, total: 1.33 s  
Wall time: 1.33 s

###### 4.4.2. Check the results of the simulation for specific cells[¶](#4.4.2.-Check-the-results-of-the-simulation-for-specific-cells)

Check the results of simulation. Pick up some cells and visualize their transition trajectory.

[88]:

# Randomly pick up 3 cells  
np.random.seed(12)  
cells = oracle.adata.obs.index.values[np.random.choice(oracle.ixs\_mcmc, 3)]  
  
# Visualize the simulated results of cell transition after perturbation  
for k in cells:  
 print(f"cell {k}")  
 plt.figure(figsize=[9, 3])  
 for j, i in enumerate([0, 20, 50]): # time points  
 plt.subplot(1, 3, (j+1))  
 oracle.plot\_mc\_result\_as\_trajectory(k, range(0, i))  
 plt.title(f"simulation step: 0~{i}")  
 plt.axis("off")  
 plt.show()

cell 1961

<_images/notebooks_05_simulation_Gata1_KO_simulation_with_with_Paul_etal_2015_data_32_1.png>

cell 43

<_images/notebooks_05_simulation_Gata1_KO_simulation_with_with_Paul_etal_2015_data_32_3.png>

cell 1567

<_images/notebooks_05_simulation_Gata1_KO_simulation_with_with_Paul_etal_2015_data_32_5.png>

###### 4.4.3. Summarize the results of simulation by plotting sankey diagram[¶](#4.4.3.-Summarize-the-results-of-simulation-by-plotting-sankey-diagram)

Sankey diagrams are useful when you want to visualize proportional cell transitions between some groups.

For the grouping of cells, you can use arbitrary cluster unit.

[89]:

# Plot sankey diagram  
plt.figure(figsize=[5,6])  
cl = "louvain\_annot"  
oracle.plot\_mc\_resutls\_as\_sankey(cluster\_use=cl, start=0, end=100)

<_images/notebooks_05_simulation_Gata1_KO_simulation_with_with_Paul_etal_2015_data_35_0.png>

The Sankey diagram above looks messy because the cluster order is random.

Let’s change the cluster order and make the plot again

[90]:

cl = "louvain\_annot"  
order = ['MEP\_0', 'Mk\_0','Ery\_0', 'Ery\_1', 'Ery\_2', 'Ery\_3', 'Ery\_4',  
 'Ery\_5', 'Ery\_6', 'Ery\_7', 'Ery\_8', 'Ery\_9',  
 'GMP\_0', 'GMP\_1', 'GMP\_2', 'GMPl\_0', 'GMPl\_1',  
 'Mo\_0', 'Mo\_1', 'Mo\_2', 'Gran\_0', 'Gran\_1', 'Gran\_2', 'Gran\_3']  
  
plt.figure(figsize=[5,6])  
plt.subplots\_adjust(left=0.3, right=0.7)  
oracle.plot\_mc\_resutls\_as\_sankey(cluster\_use=cl, start=0, end=100, order=order)  
#plt.savefig(f"{save\_folder}/mcmc\_{cl}.png")

<_images/notebooks_05_simulation_Gata1_KO_simulation_with_with_Paul_etal_2015_data_37_0.png>

Make another Saneky diagram with different cluster units.

[92]:

order = ['Megakaryocytes', 'MEP', 'Erythroids', 'GMP', 'late\_GMP', 'Monocytes', 'Granulocytes']  
cl = "cell\_type"  
  
plt.figure(figsize=[5,6])  
plt.subplots\_adjust(left=0.35, right=0.65)  
oracle.plot\_mc\_resutls\_as\_sankey(cluster\_use=cl, start=0, end=100, order=order, font\_size=14)  
#plt.savefig(f"{save\_folder}/mcmc\_{cl}{goi}.png", transparent=True)

<_images/notebooks_05_simulation_Gata1_KO_simulation_with_with_Paul_etal_2015_data_39_0.png>

Based on the results, we may conclude several things as follows.

Gata1 KO induced both cell state transitions from Erythroids to MEP, and from MEP to GMP.

1. These results suggest that Gata1 may play a role in the progression of Erythroid differentiation and cell state determination between the MEP and GMP lineages.
2. Gata1 KO also induced cell state transitions from granulocytes to late GMP, suggesting Gata1’s involvement in Granulocytes differentiation.

These results agree with previous reports about Gata1 and recapitulate Gata1’s cell-type-specific function regarding the cell fate decisions in hematopoiesis.

## API[¶](#api)

### Command Line API[¶](#command-line-api)

CellOracle has a command line API. This command can be used to convert scRNA-seq data. If you have a scRNA-seq data which was processed with Seurat and saved as Rds file, you can use the following command to make anndata from Seurat object. The anndata object produced by this command can be used for input of celloracle.

seuratToAnndata YOUR\_SEURAT\_OBJECT.Rds OUTPUT\_PATH

### Python API[¶](#python-api)

#### Custom class in celloracle[¶](#custom-class-in-celloracle)

We define some custom classes in celloracle.

*class* celloracle.Oracle[¶](#celloracle.Oracle)

Bases: celloracle.trajectory.modified\_VelocytoLoom\_class.modified\_VelocytoLoom

Oracle is the main class in CellOracle. Oracle object imports scRNA-seq data (anndata) and TF information to infer cluster-specific GRNs. It can predict the future gene expression patterns and cell state transitions in response to the perturbation of TFs. Please see the CellOracle paper for details. The code of the Oracle class was made of the three components below.

1. Anndata: Gene expression matrix and metadata from single-cell RNA-seq are stored in the anndata object. Processed values, such as normalized counts and simulated values, are stored as layers of anndata. Metadata (i.e., Cluster info) are saved in anndata.obs. Refer to scanpy/anndata documentation for detail.
2. Net: Net is a custom class in celloracle. Net object processes several data to infer GRN. See the Net class documentation for details.
3. VelycytoLoom: Calculation of transition probability and visualization of directed trajectory graph will be performed in the same way as velocytoloom. VelocytoLoom is class from Velocyto, a python library for RNA-velocity analysis. In celloracle, we use some functions in velocytoloom for the visualization.

adata[¶](#celloracle.Oracle.adata)

Imported anndata object

Type

anndata

cluster\_column\_name[¶](#celloracle.Oracle.cluster_column_name)

The column name in adata.obs containing cluster info

Type

str

embedding\_name[¶](#celloracle.Oracle.embedding_name)

The key name in adata.obsm containing dimensional reduction cordinates

Type

str

addTFinfo\_dictionary(*TFdict*)[¶](#celloracle.Oracle.addTFinfo_dictionary)

Add new TF info to pre-existing TFdict. Values in the old TF dictionary will remain.

Parameters

**TFdict** (*dictionary*) – Python dictionary of TF info.

copy()[¶](#celloracle.Oracle.copy)

Deepcopy itself.

fit\_GRN\_for\_simulation(*GRN\_unit='cluster'*, *alpha=1*, *use\_cluster\_specific\_TFdict=False*)[¶](#celloracle.Oracle.fit_GRN_for_simulation)

Do GRN inference. Please see the paper of CellOracle paper for details.

GRN can be constructed for the entire population or each clusters. If you want to infer cluster-specific GRN, please set [GRN\_unit=”cluster”]. You can select cluster information when you import data.

If you set [GRN\_unit=”whole”], GRN will be made using all cells.

Parameters

* **GRN\_unit** (*str*) – Select “cluster” or “whole”
* **alpha** (*float* *or* *int*) – The strength of regularization. If you set a lower value, the sensitivity increases, and you can detect weaker network connections. However, there may be more noise. If you select a higher value, it will reduce the chance of overfitting.

get\_cluster\_specific\_TFdict\_from\_Links(*links\_object*)[¶](#celloracle.Oracle.get_cluster_specific_TFdict_from_Links)

Extract TF and its target gene information from Links object. This function can be used to reconstruct GRNs based on pre-existing GRNs saved in Links object.

Parameters

**links\_object** ([*Links*](index.html#celloracle.Links)) – Please see the explanation of Links class.

get\_links(*cluster\_name\_for\_GRN\_unit=None*, *alpha=10*, *bagging\_number=20*, *verbose\_level=1*, *test\_mode=False*)[¶](#celloracle.Oracle.get_links)

Makes GRN for each cluster and returns results as a Links object. Several preprocessing should be done before using this function.

Parameters

* **cluster\_name\_for\_GRN\_unit** (*str*) – Cluster name for GRN calculation. The cluster information should be stored in Oracle.adata.obs.
* **alpha** (*float* *or* *int*) – The strength of regularization. If you set a lower value, the sensitivity increases, and you can detect weaker network connections. However, there may be more noise. If you select a higher value, it will reduce the chance of overfitting.
* **bagging\_number** (*int*) – The number used in bagging calculation.
* **verbose\_level** (*int*) – if [verbose\_level>1], most detailed progress information will be shown. if [verbose\_level > 0], one progress bar will be shown. if [verbose\_level == 0], no progress bar will be shown.
* **test\_mode** (*bool*) – If test\_mode is True, GRN calculation will be done for only one cluster rather than all clusters.

import\_TF\_data(*TF\_info\_matrix=None*, *TF\_info\_matrix\_path=None*, *TFdict=None*)[¶](#celloracle.Oracle.import_TF_data)

Load data about potential-regulatory TFs. You can import either TF\_info\_matrix or TFdict. For more information on how to make these files, please see the motif analysis module within the celloracle tutorial.

Parameters

* **TF\_info\_matrix** (*pandas.DataFrame*) – TF\_info\_matrix.
* **TF\_info\_matrix\_path** (*str*) – File path for TF\_info\_matrix (pandas.DataFrame).
* **TFdict** (*dictionary*) – Python dictionary of TF info.

import\_anndata\_as\_normalized\_count(*adata*, *cluster\_column\_name=None*, *embedding\_name=None*)[¶](#celloracle.Oracle.import_anndata_as_normalized_count)

Load scRNA-seq data. scRNA-seq data should be prepared as an anndata object. Preprocessing (cell and gene filtering, dimensional reduction, clustering, etc.) should be done before loading data. The method will import NORMALIZED and LOG TRANSFORMED data but NOT SCALED and NOT CENTERED data. See the tutorial for more details on how to process scRNA-seq data.

Parameters

* **adata** (*anndata*) – anndata object containing scRNA-seq data.
* **cluster\_column\_name** (*str*) – the name of column containing cluster information in anndata.obs. Clustering data should be in anndata.obs.
* **embedding\_name** (*str*) – the key name for dimensional reduction information in anndata.obsm. Dimensional reduction (or 2D trajectory graph) should be in anndata.obsm.
* **transform** (*str*) – The method for log-transformation. Chose one from “natural\_log” or “log2”.

import\_anndata\_as\_raw\_count(*adata*, *cluster\_column\_name=None*, *embedding\_name=None*, *transform='natural\_log'*)[¶](#celloracle.Oracle.import_anndata_as_raw_count)

Load scRNA-seq data. scRNA-seq data should be prepared as an anndata object. Preprocessing (cell and gene filtering, dimensional reduction, clustering, etc.) should be done before loading data. The method imports RAW GENE COUNTS because unscaled and uncentered gene expression data are required for the GRN inference and simulation. See tutorial notebook for the details about how to process scRNA-seq data.

Parameters

* **adata** (*anndata*) – anndata object that stores scRNA-seq data.
* **cluster\_column\_name** (*str*) – the name of column containing cluster information in anndata.obs. Clustering data should be in anndata.obs.
* **embedding\_name** (*str*) – the key name for dimensional reduction information in anndata.obsm. Dimensional reduction (or 2D trajectory graph) should be in anndata.obsm.
* **transform** (*str*) – The method for log-transformation. Chose one from “natural\_log” or “log2”.

plot\_mc\_result\_as\_kde(*n\_time*, *args={}*)[¶](#celloracle.Oracle.plot_mc_result_as_kde)

Pick up one timepoint in the cell state-transition simulation and plot as a kde plot.

Parameters

* **n\_time** (*int*) – the number in Markov simulation
* **args** (*dictionary*) – An argument for seaborn.kdeplot. See seaborn documentation for details (<https://seaborn.pydata.org/generated/seaborn.kdeplot.html#seaborn.kdeplot>).

plot\_mc\_result\_as\_trajectory(*cell\_name*, *time\_range*, *args={}*)[¶](#celloracle.Oracle.plot_mc_result_as_trajectory)

Pick up several timepoints in the cell state-transition simulation and plot as a line plot. This function can be used to visualize how cell-state changes after perturbation focusing on a specific cell.

Parameters

* **cell\_name** (*str*) – cell name. chose from adata.obs.index
* **time\_range** (*list of int*) – the list of index in Markov simulation
* **args** (*dictionary*) – dictionary for the arguments for matplotlib.pyplit.plot. See matplotlib documentation for details (<https://matplotlib.org/api/_as_gen/matplotlib.pyplot.plot.html#matplotlib.pyplot.plot>).

plot\_mc\_resutls\_as\_sankey(*cluster\_use*, *start=0*, *end=-1*, *order=None*, *font\_size=10*)[¶](#celloracle.Oracle.plot_mc_resutls_as_sankey)

Plot the simulated cell state-transition as a Sankey-diagram after groping by the cluster.

Parameters

* **cluster\_use** (*str*) – cluster information name in anndata.obs. You can use any cluster information in anndata.obs.
* **start** (*int*) – The starting point of Sankey-diagram. Please select a step in the Markov simulation.
* **end** (*int*) – The end point of Sankey-diagram. Please select a step in the Markov simulation. if you set [end=-1], the final step of Markov simulation will be used.
* **order** (*list of str*) – The order of cluster name in the Sankey-diagram.
* **font\_size** (*int*) – Font size for cluster name label in the Sankey diagram.

prepare\_markov\_simulation(*verbose=False*)[¶](#celloracle.Oracle.prepare_markov_simulation)

Pick up cells for Markov simulation.

Parameters

**verbose** (*bool*) – If True, it plots selected cells.

run\_markov\_chain\_simulation(*n\_steps=500*, *n\_duplication=5*, *seed=123*)[¶](#celloracle.Oracle.run_markov_chain_simulation)

Do Markov simlations to predict cell transition after perturbation. The transition probability between cells has been calculated based on simulated gene expression values in the signal propagation process. The cell state transition will be simulated based on the probability. You can simulate the process multiple times to get a robust outcome.

Parameters

* **n\_steps** (*int*) – steps for Markov simulation. This value is equivalent to the amount of time after perturbation.
* **n\_duplication** (*int*) – the number for multiple calculations.

simulate\_shift(*perturb\_condition=None*, *GRN\_unit='cluster'*, *n\_propagation=3*, *ignore\_warning=False*)[¶](#celloracle.Oracle.simulate_shift)

Simulate signal propagation with GRNs. Please see the CellOracle paper for details. This function simulates a gene expression pattern in the near future. Simulated values will be stored in anndata.layers: [“simulated\_count”]

The simulation use three types of data. (1) GRN inference results (coef\_matrix). (2) Perturb\_condition: You can set arbitrary perturbation condition. (3) Gene expression matrix: The simulation starts from imputed gene expression data.

Parameters

* **perturb\_condition** (*dictionary*) – condition for perturbation. if you want to simulate knockout for GeneX, please set [perturb\_condition={“GeneX”: 0.0}] Although you can set any non-negative values for the gene condition, avoid setting biologically infeasible values for the perturb condition. It is strongly recommended to check gene expression values in your data before selecting the perturb condition.
* **GRN\_unit** (*str*) – GRN type. Please select either “whole” or “cluster”. See the documentation of “fit\_GRN\_for\_simulation” for the detailed explanation.
* **n\_propagation** (*int*) – Calculation will be performed iteratively to simulate signal propagation in GRN. You can set the number of steps for this calculation. With a higher number, the results may recapitulate signal propagation for many genes. However, a higher number of propagation may cause more error/noise.

summarize\_mc\_results\_by\_cluster(*cluster\_use*)[¶](#celloracle.Oracle.summarize_mc_results_by_cluster)

This function summarizes the simulated cell state-transition by groping the results into each cluster. It returns sumarized results as a pandas.DataFrame.

Parameters

**cluster\_use** (*str*) – cluster information name in anndata.obs. You can use any arbitrary cluster information in anndata.obs.

to\_hdf5(*file\_path*)[¶](#celloracle.Oracle.to_hdf5)

Save object as hdf5.

Parameters

**file\_path** (*str*) – file path to save file. Filename needs to end with ‘.celloracle.oracle’

updateTFinfo\_dictionary(*TFdict*)[¶](#celloracle.Oracle.updateTFinfo_dictionary)

Update a TF dictionary. If a key in the new TF dictionary already exists in the old TF dictionary, old values will be replaced with a new one.

Parameters

**TFdict** (*dictionary*) – Python dictionary of TF info.

*class* celloracle.Links(*name*, *links\_dict={}*)[¶](#celloracle.Links)

Bases: object

This is a class for the processing and visualization of GRNs. Links object stores cluster-specific GRNs and metadata. Please use “get\_links” function in Oracle object to generate Links object.

links\_dict[¶](#celloracle.Links.links_dict)

Dictionary that store unprocessed network data.

Type

dictionary

filtered\_links[¶](#celloracle.Links.filtered_links)

Dictionary that store filtered network data.

Type

dictionary

merged\_score[¶](#celloracle.Links.merged_score)

Network scores.

Type

pandas.dataframe

cluster[¶](#celloracle.Links.cluster)

List of cluster name.

Type

list of str

name[¶](#celloracle.Links.name)

Name of clustering unit.

Type

str

palette[¶](#celloracle.Links.palette)

DataFrame that store color information.

Type

pandas.dataframe

filter\_links(*p=0.001*, *weight='coef\_abs'*, *thread\_number=10000*, *genelist\_source=None*, *genelist\_target=None*)[¶](#celloracle.Links.filter_links)

Filter network edges. In most cases, inferred GRN has non-significant random edges. We have to remove these edges before analyzing the network structure. You can do the filtering in any of the following ways.

1. Filter based on the p-value of the network edge. Please enter p-value for thresholding.
2. Filter based on network edge number. If you set the number, network edges will be filtered based on the order of a network score. The top n-th network edges with network weight will remain, and the other edges will be removed. The network data has several types of network weight, so you have to select which network weight do you want to use.
3. Filter based on an arbitrary gene list. You can set a gene list for source nodes or target nodes.

Parameters

* **p** (*float*) – threshold for p-value of the network edge.
* **weight** (*str*) – Please select network weight name for the filtering
* **genelist\_source** (*list of str*) – gene list to remain in regulatory gene nodes. Default is None.
* **genelist\_target** (*list of str*) – gene list to remain in target gene nodes. Default is None.

get\_network\_entropy(*value='coef\_abs'*)[¶](#celloracle.Links.get_network_entropy)

Calculate network entropy scores.

Parameters

**value** (*str*) – Default is “coef\_abs”.

get\_score(*test\_mode=False*)[¶](#celloracle.Links.get_score)

Get several network sores using R libraries. Make sure all dependent R libraries are installed in your environment before running this function. You can check the installation for the R libraries by running test\_installation() in network\_analysis module.

plot\_cartography\_scatter\_per\_cluster(*gois=None*, *clusters=None*, *scatter=True*, *kde=False*, *auto\_gene\_annot=False*, *percentile=98*, *args\_dot={'n\_levels': 105}*, *args\_line={'c': 'gray'}*, *args\_annot={}*, *save=None*)[¶](#celloracle.Links.plot_cartography_scatter_per_cluster)

Make a gene network cartography plot. Please read the original paper describing gene network cartography for more information. <https://www.nature.com/articles/nature03288>

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See network\_analysis.Links class for detail.
* **gois** (*list of srt*) – List of gene name to highlight.
* **clusters** (*list of str*) – List of cluster name to analyze. If None, all clusters in Links object will be analyzed.
* **scatter** (*bool*) – Whether to make a scatter plot.
* **auto\_gene\_annot** (*bool*) – Whether to pick up genes to make an annotation.
* **percentile** (*float*) – Genes with a network score above the percentile will be shown with annotation. Default is 98.
* **args\_dot** (*dictionary*) – Arguments for scatter plot.
* **args\_line** (*dictionary*) – Arguments for lines in cartography plot.
* **args\_annot** (*dictionary*) – Arguments for annotation in plots.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_cartography\_term(*goi*, *save=None*)[¶](#celloracle.Links.plot_cartography_term)

Plot the gene network cartography term like a heatmap. Please read the original paper of gene network cartography for the principle of gene network cartography. <https://www.nature.com/articles/nature03288>

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See network\_analysis.Links class for detail.
* **gois** (*list of srt*) – List of gene name to highlight.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_degree\_distributions(*plot\_model=False*, *save=None*)[¶](#celloracle.Links.plot_degree_distributions)

Plot the network degree distributions (the number of edge per gene). The network degree will be visualized in both linear scale and log scale.

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See network\_analysis.Links class for detail.
* **plot\_model** (*bool*) – Whether to plot linear approximation line.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_network\_entropy\_distributions(*update\_network\_entropy=False*, *save=None*)[¶](#celloracle.Links.plot_network_entropy_distributions)

Plot the distribution for network entropy. See the CellOracle paper for more detail.

Parameters

* **links** (*Links object*) – See network\_analysis.Links class for detail.
* **values** (*list of str*) – The list of score to visualize. If it is None, all network score (listed above) will be used.
* **update\_network\_entropy** (*bool*) – Whether to recalculate network entropy.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_score\_comparison\_2D(*value*, *cluster1*, *cluster2*, *percentile=99*, *annot\_shifts=None*, *save=None*)[¶](#celloracle.Links.plot_score_comparison_2D)

Make a scatter plot that compares specific network scores in two groups.

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See network\_analysis.Links class for detail.
* **value** (*srt*) – The network score type.
* **cluster1** (*str*) – Cluster name. Network scores in cluster1 will be visualized in the x-axis.
* **cluster2** (*str*) – Cluster name. Network scores in cluster2 will be visualized in the y-axis.
* **percentile** (*float*) – Genes with a network score above the percentile will be shown with annotation. Default is 99.
* **annot\_shifts** (*(float,* *float)*) – Annotation visualization setting.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_score\_discributions(*values=None*, *method='boxplot'*, *save=None*)[¶](#celloracle.Links.plot_score_discributions)

Plot the distribution of network scores. An individual data point is a network edge (gene).

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See Links class for details.
* **values** (*list of str*) – The list of score to visualize. If it is None, all of the network score will be used.
* **method** (*str*) – Plotting method. Select either “boxplot” or “barplot”.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_score\_per\_cluster(*goi*, *save=None*)[¶](#celloracle.Links.plot_score_per_cluster)

Plot network score for a gene. This function visualizes the network score for a specific gene between clusters to get an insight into the dynamics of the gene.

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See network\_analysis.Links class for detail.
* **goi** (*srt*) – Gene name.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_scores\_as\_rank(*cluster*, *n\_gene=50*, *save=None*)[¶](#celloracle.Links.plot_scores_as_rank)

Pick up top n-th genes wich high-network scores and make plots.

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See network\_analysis.Links class for detail.
* **cluster** (*str*) – Cluster name to analyze.
* **n\_gene** (*int*) – Number of genes to plot. Default is 50.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

to\_hdf5(*file\_path*)[¶](#celloracle.Links.to_hdf5)

Save object as hdf5.

Parameters

**file\_path** (*str*) – file path to save file. Filename needs to end with ‘.celloracle.links’

*class* celloracle.Net(*gene\_expression\_matrix*, *gem\_standerdized=None*, *TFinfo\_matrix=None*, *cellstate=None*, *TFinfo\_dic=None*, *annotation=None*, *verbose=True*)[¶](#celloracle.Net)

Bases: object

Net is a custom class for inferring sample-specific GRN from scRNA-seq data. This class is used inside the Oracle class for GRN inference. This class requires two types of information below.

1. Single-cell RNA-seq data: The Net class needs processed scRNA-seq data. Gene and cell filtering, quality check, normalization, log-transformation (but not scaling and centering) have to be done before starting the GRN calculation with this class. You can also use any arbitrary metadata (i.e., mRNA count, cell-cycle phase) for GRN input.
2. Potential regulatory connection (or base GRN): This method uses the list of potential regulatory TFs as input. This information can be calculated from ATAC-seq data using the motif-analysis module. If sample-specific ATAC-seq data is not available, you can use general TF-binding info derived from public ATAC-seq dataset of various tissue/cell type.

linkList[¶](#celloracle.Net.linkList)

The results of the GRN inference.

Type

pandas.DataFrame

all\_genes[¶](#celloracle.Net.all_genes)

An array of all genes that exist in the input gene expression matrix

Type

numpy.array

embedding\_name[¶](#celloracle.Net.embedding_name)

The key name name in adata.obsm containing dimensional reduction coordinates

Type

str

annotation[¶](#celloracle.Net.annotation)

Annotation. you can add custom annotation.

Type

dictionary

coefs\_dict[¶](#celloracle.Net.coefs_dict)

Coefs of linear regression.

Type

dictionary

stats\_dict[¶](#celloracle.Net.stats_dict)

Statistic values about coefs.

Type

dictionary

fitted\_genes[¶](#celloracle.Net.fitted_genes)

List of genes where the regression model was successfully calculated.

Type

list of str

failed\_genes[¶](#celloracle.Net.failed_genes)

List of genes that were not assigned coefs

Type

list of str

cellstate[¶](#celloracle.Net.cellstate)

A metadata for GRN input

Type

pandas.DataFrame

TFinfo[¶](#celloracle.Net.TFinfo)

Information about potential regulatory TFs.

Type

pandas.DataFrame

gem[¶](#celloracle.Net.gem)

Merged matrix made with gene\_expression\_matrix and cellstate matrix.

Type

pandas.DataFrame

gem\_standerdized[¶](#celloracle.Net.gem_standerdized)

Almost the same as gem, but the gene\_expression\_matrix was standardized.

Type

pandas.DataFrame

library\_last\_update\_date[¶](#celloracle.Net.library_last_update_date)

Last update date of this code. This info is for code development. It can be deprecated in the future

Type

str

object\_initiation\_date[¶](#celloracle.Net.object_initiation_date)

The date when this object was made.

Type

str

addAnnotation(*annotation\_dictionary*)[¶](#celloracle.Net.addAnnotation)

Add a new annotation.

Parameters

**annotation\_dictionary** (*dictionary*) – e.g. {“sample\_name”: “NIH 3T3 cell”}

addTFinfo\_dictionary(*TFdict*)[¶](#celloracle.Net.addTFinfo_dictionary)

Add a new TF info to pre-exiting TFdict.

Parameters

**TFdict** (*dictionary*) – python dictionary of TF info.

addTFinfo\_matrix(*TFinfo\_matrix*)[¶](#celloracle.Net.addTFinfo_matrix)

Load TF info dataframe.

Parameters

**TFinfo** (*pandas.DataFrame*) – information about potential regulatory TFs.

copy()[¶](#celloracle.Net.copy)

Deepcopy itself

fit\_All\_genes(*bagging\_number=200*, *scaling=True*, *model\_method='bagging\_ridge'*, *command\_line\_mode=False*, *log=None*, *alpha=1*, *verbose=True*)[¶](#celloracle.Net.fit_All_genes)

Make ML models for all genes. The calculation will be performed in parallel using scikit-learn bagging function. You can select a modeling method (bagging\_ridge or bayesian\_ridge). This calculation usually takes a long time.

Parameters

* **bagging\_number** (*int*) – The number of estimators for bagging.
* **scaling** (*bool*) – Whether or not to scale regulatory gene expression values.
* **model\_method** (*str*) – ML model name. Please select either “bagging\_ridge” or “bayesian\_ridge”
* **command\_line\_mode** (*bool*) – Please select False if the calculation is performed on jupyter notebook.
* **log** (*logging object*) – log object to output log
* **alpha** (*int*) – Strength of regularization.
* **verbose** (*bool*) – Whether or not to show a progress bar.

fit\_All\_genes\_parallel(*bagging\_number=200*, *scaling=True*, *log=None*, *verbose=10*)[¶](#celloracle.Net.fit_All_genes_parallel)

IMPORTANT: this function being debugged and is currently unavailable.

Make ML models for all genes. The calculation will be performed in parallel using joblib parallel module.

Parameters

* **bagging\_number** (*int*) – The number of estimators for bagging.
* **scaling** (*bool*) – Whether or not to scale regulatory gene expression values.
* **log** (*logging object*) – log object to output log
* **verbose** (*int*) – verbose for joblib parallel

fit\_genes(*target\_genes*, *bagging\_number=200*, *scaling=True*, *model\_method='bagging\_ridge'*, *save\_coefs=False*, *command\_line\_mode=False*, *log=None*, *alpha=1*, *verbose=True*)[¶](#celloracle.Net.fit_genes)

Make ML models for genes of interest. This calculation will be performed in parallel using scikit-learn’s bagging function. You can select a modeling method; Please chose either bagging\_ridge or bayesian\_ridge.

Parameters

* **target\_genes** (*list of str*) – gene list
* **bagging\_number** (*int*) – The number of estimators for bagging.
* **scaling** (*bool*) – Whether or not to scale regulatory gene expression values.
* **model\_method** (*str*) – ML model name. Please select either “bagging\_ridge” or “bayesian\_ridge”
* **save\_coefs** (*bool*) – Whether or not to store details of coef values in bagging model.
* **command\_line\_mode** (*bool*) – Please select False if the calculation is performed on jupyter notebook.
* **log** (*logging object*) – log object to output log
* **alpha** (*int*) – Strength of regularization.
* **verbose** (*bool*) – Whether or not to show a progress bar.

plotCoefs(*target\_gene*, *sort=True*, *threshold\_p=None*)[¶](#celloracle.Net.plotCoefs)

Plot the distribution of Coef values (network edge weights).

Parameters

* **target\_gene** (*str*) – gene name
* **sort** (*bool*) – Whether or not to sort genes by its strength
* **bagging\_number** (*int*) – The number of estimators for bagging.
* **threshold\_p** (*float*) – the threshold for p-values. TFs will be filtered based on the p-value. if None, no filtering is applied.

to\_hdf5(*file\_path*)[¶](#celloracle.Net.to_hdf5)

Save object as hdf5.

Parameters

**file\_path** (*str*) – file path to save file. Filename needs to end with ‘.celloracle.net’

updateLinkList(*verbose=True*)[¶](#celloracle.Net.updateLinkList)

Update LinkList. LinkList is a data frame that store information about inferred GRNs.

Parameters

**verbose** (*bool*) – Whether or not to show a progress bar

updateTFinfo\_dictionary(*TFdict*)[¶](#celloracle.Net.updateTFinfo_dictionary)

Update TF info matrix

Parameters

**TFdict** (*dictionary*) – A python dictionary in which a key is Target gene, value are potential regulatory genes for the target gene.

celloracle.load\_hdf5(*file\_path*, *object\_class\_name=None*)[¶](#celloracle.load_hdf5)

Load an object of celloracle’s custom class that was saved as hdf5.

Parameters

* **file\_path** (*str*) – file\_path.
* **object\_class\_name** (*str*) – Types of object. If it is None, object class will be identified from the extension of file\_name. Default is None.

#### Modules for ATAC-seq analysis[¶](#modules-for-atac-seq-analysis)

##### celloracle.motif\_analysis module[¶](#module-celloracle.motif_analysis)

The [motif\_analysis](#module-celloracle.motif_analysis) module implements transcription factor motif scan.

Genomic activity information (peak of ATAC-seq or Chip-seq) is extracted first. Then the peak DNA sequence will be subjected to TF motif scan. Finally we will get list of TFs that potentially binds to a specific gene.

celloracle.motif\_analysis.is\_genome\_installed(*ref\_genome*)[¶](#celloracle.motif_analysis.is_genome_installed)

Celloracle motif\_analysis module uses gimmemotifs and genomepy internally. Reference genome files should be installed in the PC to use gimmemotifs and genomepy. This function checks the installation status of the reference genome.

Parameters

**ref\_genome** (*str*) – names of reference genome. i.e., “mm10”, “hg19”

celloracle.motif\_analysis.peak2fasta(*peak\_ids*, *ref\_genome*)[¶](#celloracle.motif_analysis.peak2fasta)

Convert peak\_id into fasta object.

Parameters

* **peak\_id** (*str* *or* *list of str*) – Peak\_id. e.g. “chr5\_0930303\_9499409” or it can be a list of peak\_id. e.g. [“chr5\_0930303\_9499409”, “chr11\_123445555\_123445577”]
* **ref\_genome** (*str*) – Reference genome name. e.g. “mm9”, “mm10”, “hg19” etc

Returns

DNA sequence in fasta format

Return type

gimmemotifs fasta object

celloracle.motif\_analysis.read\_bed(*bed\_path*)[¶](#celloracle.motif_analysis.read_bed)

Load bed file and return as dataframe.

Parameters

**bed\_path** (*str*) – File path.

Returns

bed file in dataframe.

Return type

pandas.dataframe

celloracle.motif\_analysis.load\_TFinfo\_from\_parquets(*folder\_path*)[¶](#celloracle.motif_analysis.load_TFinfo_from_parquets)

Load TFinfo object which was saved with the function; “save\_as\_parquet”.

Parameters

**folder\_path** (*str*) – folder path

Returns

Loaded TFinfo object.

Return type

[TFinfo](index.html#celloracle.motif_analysis.TFinfo)

celloracle.motif\_analysis.make\_TFinfo\_from\_scanned\_file(*path\_to\_raw\_bed*, *path\_to\_scanned\_result\_bed*, *ref\_genome*)[¶](#celloracle.motif_analysis.make_TFinfo_from_scanned_file)

This function is currently an available.

*class* celloracle.motif\_analysis.TFinfo(*peak\_data\_frame*, *ref\_genome*)[¶](#celloracle.motif_analysis.TFinfo)

Bases: object

This is a custom class for motif analysis in celloracle. TFinfo object performs motif scan using the TF motif database in gimmemotifs and several functions of genomepy. Analysis results can be exported as a python dictionary or dataframe. These files; python dictionary of dataframe of TF binding information, are needed during GRN inference.

peak\_df[¶](#celloracle.motif_analysis.TFinfo.peak_df)

dataframe about DNA peak and target gene data.

Type

pandas.dataframe

all\_target\_gene[¶](#celloracle.motif_analysis.TFinfo.all_target_gene)

target genes.

Type

array of str

ref\_genome[¶](#celloracle.motif_analysis.TFinfo.ref_genome)

reference genome name that was used in DNA peak generation.

Type

str

scanned\_df[¶](#celloracle.motif_analysis.TFinfo.scanned_df)

Results of motif scan. Key is a peak name. Value is a dataframe of motif scan.

Type

dictionary

dic\_targetgene2TFs[¶](#celloracle.motif_analysis.TFinfo.dic_targetgene2TFs)

Final product of motif scan. Key is a target gene. Value is a list of regulatory candidate genes.

Type

dictionary

dic\_peak2Targetgene[¶](#celloracle.motif_analysis.TFinfo.dic_peak2Targetgene)

Dictionary. Key is a peak name. Value is a list of the target gene.

Type

dictionary

dic\_TF2targetgenes[¶](#celloracle.motif_analysis.TFinfo.dic_TF2targetgenes)

Final product of motif scan. Key is a TF. Value is a list of potential target genes of the TF.

Type

dictionary

copy()[¶](#celloracle.motif_analysis.TFinfo.copy)

Deepcoty itself.

filter\_motifs\_by\_score(*threshold*, *method='cumlative\_score'*)[¶](#celloracle.motif_analysis.TFinfo.filter_motifs_by_score)

Remove motifs with low binding scores.

Parameters

**method** (*str*) – thresholding method. Select either of [“indivisual\_score”, “cumlative\_score”]

filter\_peaks(*peaks\_to\_be\_remained*)[¶](#celloracle.motif_analysis.TFinfo.filter_peaks)

Filter peaks.

Parameters

**peaks\_to\_be\_remained** (*array of str*) – list of peaks. Peaks that are NOT in the list will be removed.

make\_TFinfo\_dataframe\_and\_dictionary(*verbose=True*)[¶](#celloracle.motif_analysis.TFinfo.make_TFinfo_dataframe_and_dictionary)

This is the final step of motif\_analysis. Convert scanned results into a data frame and dictionaries.

Parameters

**verbose** (*bool*) – Whether to show a progress bar.

reset\_dictionary\_and\_df()[¶](#celloracle.motif_analysis.TFinfo.reset_dictionary_and_df)

Reset TF dictionary and TF dataframe. The following attributes will be erased: TF\_onehot, dic\_targetgene2TFs, dic\_peak2Targetgene, dic\_TF2targetgenes.

reset\_filtering()[¶](#celloracle.motif_analysis.TFinfo.reset_filtering)

Reset filtering information. You can use this function to stat over the filtering step with new conditions. The following attributes will be erased: TF\_onehot, dic\_targetgene2TFs, dic\_peak2Targetgene, dic\_TF2targetgenes.

save\_as\_parquet(*folder\_path=None*)[¶](#celloracle.motif_analysis.TFinfo.save_as_parquet)

Save itself. Some attributes are saved as parquet file.

Parameters

**folder\_path** (*str*) – folder path

scan(*background\_length=200*, *fpr=0.02*, *n\_cpus=-1*, *verbose=True*)[¶](#celloracle.motif_analysis.TFinfo.scan)

Scan DNA sequences searching for TF binding motifs.

Parameters

* **background\_length** (*int*) – background length. This is used for the calculation of the binding score.
* **fpr** (*float*) – False positive rate for motif identification.
* **n\_cpus** (*int*) – number of CPUs for parallel calculation.
* **verbose** (*bool*) – Whether to show a progress bar.

to\_dataframe(*verbose=True*)[¶](#celloracle.motif_analysis.TFinfo.to_dataframe)

Return results as a dataframe. Rows are peak\_id, and columns are TFs.

Parameters

**verbose** (*bool*) – Whether to show a progress bar.

Returns

TFinfo matrix.

Return type

pandas.dataframe

to\_dictionary(*dictionary\_type='targetgene2TFs'*, *verbose=True*)[¶](#celloracle.motif_analysis.TFinfo.to_dictionary)

Return TF information as a python dictionary.

Parameters

**dictionary\_type** (*str*) – Type of dictionary. Select from [“targetgene2TFs”, “TF2targetgenes”]. If you chose “targetgene2TFs”, it returns a dictionary in which a key is a target gene, and a value is a list of regulatory candidate genes (TFs) of the target. If you chose “TF2targetgenes”, it returns a dictionary in which a key is a TF and a value is a list of potential target genes of the TF.

Returns

dictionary.

Return type

dictionary

to\_hdf5(*file\_path*)[¶](#celloracle.motif_analysis.TFinfo.to_hdf5)

Save object as hdf5.

Parameters

**file\_path** (*str*) – file path to save file. Filename needs to end with ‘.celloracle.tfinfo’

celloracle.motif\_analysis.get\_tss\_info(*peak\_str\_list*, *ref\_genome*, *verbose=True*)[¶](#celloracle.motif_analysis.get_tss_info)

Get annotation about Transcription Starting Site (TSS).

Parameters

* **peak\_str\_list** (*list of str*) – list of peak\_id. e.g., [“chr5\_0930303\_9499409”, “chr11\_123445555\_123445577”]
* **ref\_genome** (*str*) – reference genome name.
* **verbose** (*bool*) – verbosity.

celloracle.motif\_analysis.integrate\_tss\_peak\_with\_cicero(*tss\_peak*, *cicero\_connections*)[¶](#celloracle.motif_analysis.integrate_tss_peak_with_cicero)

Process output of cicero data and returns DNA peak information for motif analysis in celloracle. Please see the celloracle tutorial for more information.

Parameters

* **tss\_peak** (*pandas.dataframe*) – dataframe about TSS information. Please use the function, “get\_tss\_info” to get this dataframe.
* **cicero\_connections** (*dataframe*) – dataframe that stores the results of cicero analysis.

Returns

DNA peak about promoter/enhancer and its annotation about target gene.

Return type

pandas.dataframe

#### Modules for Network analysis[¶](#modules-for-network-analysis)

##### celloracle.network\_analysis module[¶](#module-celloracle.network_analysis)

The [network\_analysis](#module-celloracle.network_analysis) module implements Network analysis.

celloracle.network\_analysis.get\_links(*oracle\_object*, *cluster\_name\_for\_GRN\_unit=None*, *alpha=10*, *bagging\_number=20*, *verbose\_level=1*, *test\_mode=False*)[¶](#celloracle.network_analysis.get_links)

Make GRN for each cluster and returns results as a Links object. Several preprocessing should be done before using this function.

Parameters

* **oracle\_object** ([*Oracle*](index.html#celloracle.Oracle)) – See Oracle module for detail.
* **cluster\_name\_for\_GRN\_unit** (*str*) – Cluster name for GRN calculation. The cluster information should be stored in Oracle.adata.obs.
* **alpha** (*float* *or* *int*) – The strength of regularization. If you set a lower value, the sensitivity increases, and you can detect weaker network connections. However, there may be more noise. If you select a higher value, it will reduce the chance of overfitting.
* **bagging\_number** (*int*) – The number used in bagging calculation.
* **verbose\_level** (*int*) – if [verbose\_level>1], most detailed progress information will be shown. if [verbose\_level > 0], one progress bar will be shown. if [verbose\_level == 0], no progress bar will be shown.
* **test\_mode** (*bool*) – If test\_mode is True, GRN calculation will be done for only one cluster rather than all clusters.

celloracle.network\_analysis.test\_R\_libraries\_installation()[¶](#celloracle.network_analysis.test_R_libraries_installation)

CellOracle.network\_analysis use several R libraries for network analysis. This is a test function to check for instalation of the necessary R libraries.

celloracle.network\_analysis.load\_links(*file\_path*)[¶](#celloracle.network_analysis.load_links)

Load links object saved as a hdf5 file.

Parameters

**file\_path** (*str*) – file path.

Returns

loaded links object.

Return type

[Links](index.html#celloracle.Links)

*class* celloracle.network\_analysis.Links(*name*, *links\_dict={}*)[¶](#celloracle.network_analysis.Links)

Bases: object

This is a class for the processing and visualization of GRNs. Links object stores cluster-specific GRNs and metadata. Please use “get\_links” function in Oracle object to generate Links object.

links\_dict[¶](#celloracle.network_analysis.Links.links_dict)

Dictionary that store unprocessed network data.

Type

dictionary

filtered\_links[¶](#celloracle.network_analysis.Links.filtered_links)

Dictionary that store filtered network data.

Type

dictionary

merged\_score[¶](#celloracle.network_analysis.Links.merged_score)

Network scores.

Type

pandas.dataframe

cluster[¶](#celloracle.network_analysis.Links.cluster)

List of cluster name.

Type

list of str

name[¶](#celloracle.network_analysis.Links.name)

Name of clustering unit.

Type

str

palette[¶](#celloracle.network_analysis.Links.palette)

DataFrame that store color information.

Type

pandas.dataframe

filter\_links(*p=0.001*, *weight='coef\_abs'*, *thread\_number=10000*, *genelist\_source=None*, *genelist\_target=None*)[¶](#celloracle.network_analysis.Links.filter_links)

Filter network edges. In most cases, inferred GRN has non-significant random edges. We have to remove these edges before analyzing the network structure. You can do the filtering in any of the following ways.

1. Filter based on the p-value of the network edge. Please enter p-value for thresholding.
2. Filter based on network edge number. If you set the number, network edges will be filtered based on the order of a network score. The top n-th network edges with network weight will remain, and the other edges will be removed. The network data has several types of network weight, so you have to select which network weight do you want to use.
3. Filter based on an arbitrary gene list. You can set a gene list for source nodes or target nodes.

Parameters

* **p** (*float*) – threshold for p-value of the network edge.
* **weight** (*str*) – Please select network weight name for the filtering
* **genelist\_source** (*list of str*) – gene list to remain in regulatory gene nodes. Default is None.
* **genelist\_target** (*list of str*) – gene list to remain in target gene nodes. Default is None.

get\_network\_entropy(*value='coef\_abs'*)[¶](#celloracle.network_analysis.Links.get_network_entropy)

Calculate network entropy scores.

Parameters

**value** (*str*) – Default is “coef\_abs”.

get\_score(*test\_mode=False*)[¶](#celloracle.network_analysis.Links.get_score)

Get several network sores using R libraries. Make sure all dependent R libraries are installed in your environment before running this function. You can check the installation for the R libraries by running test\_installation() in network\_analysis module.

plot\_cartography\_scatter\_per\_cluster(*gois=None*, *clusters=None*, *scatter=True*, *kde=False*, *auto\_gene\_annot=False*, *percentile=98*, *args\_dot={'n\_levels': 105}*, *args\_line={'c': 'gray'}*, *args\_annot={}*, *save=None*)[¶](#celloracle.network_analysis.Links.plot_cartography_scatter_per_cluster)

Make a gene network cartography plot. Please read the original paper describing gene network cartography for more information. <https://www.nature.com/articles/nature03288>

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See network\_analysis.Links class for detail.
* **gois** (*list of srt*) – List of gene name to highlight.
* **clusters** (*list of str*) – List of cluster name to analyze. If None, all clusters in Links object will be analyzed.
* **scatter** (*bool*) – Whether to make a scatter plot.
* **auto\_gene\_annot** (*bool*) – Whether to pick up genes to make an annotation.
* **percentile** (*float*) – Genes with a network score above the percentile will be shown with annotation. Default is 98.
* **args\_dot** (*dictionary*) – Arguments for scatter plot.
* **args\_line** (*dictionary*) – Arguments for lines in cartography plot.
* **args\_annot** (*dictionary*) – Arguments for annotation in plots.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_cartography\_term(*goi*, *save=None*)[¶](#celloracle.network_analysis.Links.plot_cartography_term)

Plot the gene network cartography term like a heatmap. Please read the original paper of gene network cartography for the principle of gene network cartography. <https://www.nature.com/articles/nature03288>

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See network\_analysis.Links class for detail.
* **gois** (*list of srt*) – List of gene name to highlight.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_degree\_distributions(*plot\_model=False*, *save=None*)[¶](#celloracle.network_analysis.Links.plot_degree_distributions)

Plot the network degree distributions (the number of edge per gene). The network degree will be visualized in both linear scale and log scale.

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See network\_analysis.Links class for detail.
* **plot\_model** (*bool*) – Whether to plot linear approximation line.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_network\_entropy\_distributions(*update\_network\_entropy=False*, *save=None*)[¶](#celloracle.network_analysis.Links.plot_network_entropy_distributions)

Plot the distribution for network entropy. See the CellOracle paper for more detail.

Parameters

* **links** (*Links object*) – See network\_analysis.Links class for detail.
* **values** (*list of str*) – The list of score to visualize. If it is None, all network score (listed above) will be used.
* **update\_network\_entropy** (*bool*) – Whether to recalculate network entropy.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_score\_comparison\_2D(*value*, *cluster1*, *cluster2*, *percentile=99*, *annot\_shifts=None*, *save=None*)[¶](#celloracle.network_analysis.Links.plot_score_comparison_2D)

Make a scatter plot that compares specific network scores in two groups.

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See network\_analysis.Links class for detail.
* **value** (*srt*) – The network score type.
* **cluster1** (*str*) – Cluster name. Network scores in cluster1 will be visualized in the x-axis.
* **cluster2** (*str*) – Cluster name. Network scores in cluster2 will be visualized in the y-axis.
* **percentile** (*float*) – Genes with a network score above the percentile will be shown with annotation. Default is 99.
* **annot\_shifts** (*(float,* *float)*) – Annotation visualization setting.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_score\_discributions(*values=None*, *method='boxplot'*, *save=None*)[¶](#celloracle.network_analysis.Links.plot_score_discributions)

Plot the distribution of network scores. An individual data point is a network edge (gene).

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See Links class for details.
* **values** (*list of str*) – The list of score to visualize. If it is None, all of the network score will be used.
* **method** (*str*) – Plotting method. Select either “boxplot” or “barplot”.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_score\_per\_cluster(*goi*, *save=None*)[¶](#celloracle.network_analysis.Links.plot_score_per_cluster)

Plot network score for a gene. This function visualizes the network score for a specific gene between clusters to get an insight into the dynamics of the gene.

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See network\_analysis.Links class for detail.
* **goi** (*srt*) – Gene name.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

plot\_scores\_as\_rank(*cluster*, *n\_gene=50*, *save=None*)[¶](#celloracle.network_analysis.Links.plot_scores_as_rank)

Pick up top n-th genes wich high-network scores and make plots.

Parameters

* **links** ([*Links*](index.html#celloracle.Links)) – See network\_analysis.Links class for detail.
* **cluster** (*str*) – Cluster name to analyze.
* **n\_gene** (*int*) – Number of genes to plot. Default is 50.
* **save** (*str*) – Folder path to save plots. If the folder does not exist in the path, the function creates the folder. Plots will not be saved if [save=None]. Default is None.

to\_hdf5(*file\_path*)[¶](#celloracle.network_analysis.Links.to_hdf5)

Save object as hdf5.

Parameters

**file\_path** (*str*) – file path to save file. Filename needs to end with ‘.celloracle.links’

celloracle.network\_analysis.transfer\_scores\_from\_links\_to\_adata(*adata*, *links*, *method='median'*)[¶](#celloracle.network_analysis.transfer_scores_from_links_to_adata)

Transfer the summary of network scores (median or mean) per group from Links object into adata.

Parameters

* **adata** (*anndata*) – anndata
* **links** ([*Links*](index.html#celloracle.Links)) – Likns object
* **method** (*str*) – The method to summarize data.

celloracle.network\_analysis.linkList\_to\_networkgraph(*filteredlinkList*)[¶](#celloracle.network_analysis.linkList_to_networkgraph)

Convert linkList into Graph object in NetworkX.

Parameters

**filteredlinkList** (*pandas.DataFrame*) – GRN saved as linkList.

Returns

Network X graph objenct.

Return type

Graph object

celloracle.network\_analysis.draw\_network(*linkList*, *return\_graph=False*)[¶](#celloracle.network_analysis.draw_network)

Plot network graph.

Parameters

* **linkList** (*pandas.DataFrame*) – GRN saved as linkList.
* **return\_graph** (*bool*) – Whether to return graph object.

Returns

Network X graph objenct.

Return type

Graph object

#### Other modules[¶](#other-modules)

##### celloracle.go\_analysis module[¶](#celloracle-go-analysis-module)

The [go\_analysis](#module-celloracle.go_analysis) module implements Gene Ontology analysis. This module use goatools internally.

celloracle.go\_analysis.geneSymbol2ID(*symbols*, *species='mouse'*)[¶](#celloracle.go_analysis.geneSymbol2ID)

Convert gene symbol into Entrez gene id.

Parameters

* **symbols** (*array of str*) – gene symbol
* **species** (*str*) – Select species. Either “mouse” or “human”

Returns

Entrez gene id

Return type

list of str

celloracle.go\_analysis.geneID2Symbol(*IDs*, *species='mouse'*)[¶](#celloracle.go_analysis.geneID2Symbol)

Convert Entrez gene id into gene symbol.

Parameters

* **IDs** (*array of str*) – Entrez gene id.
* **species** (*str*) – Select species. Either “mouse” or “human”.

Returns

Gene symbol

Return type

list of str

celloracle.go\_analysis.get\_GO(*gene\_query*, *species='mouse'*)[¶](#celloracle.go_analysis.get_GO)

Get Gene Ontologies (GOs).

Parameters

* **gene\_query** (*array of str*) – gene list.
* **species** (*str*) – Select species. Either “mouse” or “human”

Returns

GO analysis results as dataframe.

Return type

pandas.dataframe

##### celloracle.utility module[¶](#celloracle-utility-module)

The [utility](#module-celloracle.utility) module has several functions that support celloracle.

*class* celloracle.utility.makelog(*file\_name=None*, *directory=None*)[¶](#celloracle.utility.makelog)

Bases: object

This is a class for making log.

info(*comment*)[¶](#celloracle.utility.makelog.info)

Add comment into the log file.

Parameters

**comment** (*str*) – comment.

celloracle.utility.save\_as\_pickled\_object(*obj*, *filepath*)[¶](#celloracle.utility.save_as_pickled_object)

Save any object using pickle.

Parameters

* **obj** (*any python object*) – python object.
* **filepath** (*str*) – file path.

celloracle.utility.load\_pickled\_object(*filepath*)[¶](#celloracle.utility.load_pickled_object)

Load pickled object.

Parameters

**filepath** (*str*) – file path.

Returns

loaded object.

Return type

python object

celloracle.utility.intersect(*list1*, *list2*)[¶](#celloracle.utility.intersect)

Intersect two list and get components that exists in both list.

Parameters

* **list1** (*list*) – input list.
* **list2** (*list*) – input list.

Returns

intersected list.

Return type

list

celloracle.utility.exec\_process(*commands*, *message=True*, *wait\_finished=True*, *return\_process=True*)[¶](#celloracle.utility.exec_process)

Excute a command. This is a wrapper of “subprocess.Popen”

Parameters

* **commands** (*str*) – command.
* **message** (*bool*) – Whether to return a message or not.
* **wait\_finished** (*bool*) – Whether or not to wait for the process to finish. If false, the process will be perfomed in background and the function will finish immediately
* **return\_process** (*bool*) – Whether to return “process”.

celloracle.utility.standard(*df*)[¶](#celloracle.utility.standard)

Standardize value.

Parameters

**df** (*padas.dataframe*) – dataframe.

Returns

Data after standardization.

Return type

pandas.dataframe

celloracle.utility.load\_hdf5(*file\_path*, *object\_class\_name=None*)[¶](#celloracle.utility.load_hdf5)

Load an object of celloracle’s custom class that was saved as hdf5.

Parameters

* **file\_path** (*str*) – file\_path.
* **object\_class\_name** (*str*) – Types of object. If it is None, object class will be identified from the extension of file\_name. Default is None.

celloracle.utility.inverse\_dictionary(*dictionary*, *verbose=True*, *return\_value\_as\_numpy=False*)[¶](#celloracle.utility.inverse_dictionary)

Make inverse dictionary. See examples below for detail.

Parameters

* **dictionary** (*dict*) – python dictionary
* **verbose** (*bool*) – Whether to show progress bar.
* **return\_value\_as\_numpy** (*bool*) – Whether to convert values into numpy array.

Returns

Python dictionary.

Return type

dict

Examples

>>> dic = {"a": [1, 2, 3], "b": [2, 3, 4]}  
>>> inverse\_dictionary(dic)  
{1: ['a'], 2: ['a', 'b'], 3: ['a', 'b'], 4: ['b']}

>>> dic = {"a": [1, 2, 3], "b": [2, 3, 4]}  
>>> inverse\_dictionary(dic, return\_value\_as\_numpy=True)  
{1: array(['a'], dtype='<U1'),  
 2: array(['a', 'b'], dtype='<U1'),  
 3: array(['a', 'b'], dtype='<U1'),  
 4: array(['b'], dtype='<U1')}

##### celloracle.data module[¶](#celloracle-data-module)

The [data](#module-celloracle.data) module implements data download and loading.

celloracle.data.load\_TFinfo\_df\_mm9\_mouse\_atac\_atlas()[¶](#celloracle.data.load_TFinfo_df_mm9_mouse_atac_atlas)

Load Transcription factor binding information made from mouse scATAC-seq atlas dataset. mm9 genome was used for the reference genome.

Args:

Returns

TF binding info.

Return type

pandas.dataframe

##### celloracle.data\_conversion module[¶](#celloracle-data-conversion-module)

The [data\_conversion](#module-celloracle.data_conversion) module implements data conversion between different platform.

celloracle.data\_conversion.seurat\_object\_to\_anndata(*file\_path\_seurat\_object*, *delete\_tmp\_file=True*)[¶](#celloracle.data_conversion.seurat_object_to_anndata)

Convert seurat object into anndata.

Parameters

* **file\_path\_seurat\_object** (*str*) – File path of seurat object. Seurat object should be saved as Rds format.
* **delete\_tmp\_file** (*bool*) – Whether to delete temporary file.

Returns

anndata object.

Return type

anndata

## Changelog[¶](#changelog)

## License[¶](#license)

The software is provided under Apache License Version 2.0.

Copyright 2020 Kenji Kamimoto, Christy Hoffmann, Samantha Morris

Apache License Version 2.0, January 2004 <http://www.apache.org/licenses/>

TERMS AND CONDITIONS FOR USE, REPRODUCTION, AND DISTRIBUTION

1. Definitions.

“License” shall mean the terms and conditions for use, reproduction, and distribution as defined by Sections 1 through 9 of this document.

“Licensor” shall mean the copyright owner or entity authorized by the copyright owner that is granting the License.

“Legal Entity” shall mean the union of the acting entity and all other entities that control, are controlled by, or are under common control with that entity. For the purposes of this definition, “control” means (i) the power, direct or indirect, to cause the direction or management of such entity, whether by contract or otherwise, or (ii) ownership of fifty percent (50%) or more of the outstanding shares, or (iii) beneficial ownership of such entity.

“You” (or “Your”) shall mean an individual or Legal Entity exercising permissions granted by this License.

“Source” form shall mean the preferred form for making modifications, including but not limited to software source code, documentation source, and configuration files.

“Object” form shall mean any form resulting from mechanical transformation or translation of a Source form, including but not limited to compiled object code, generated documentation, and conversions to other media types.

“Work” shall mean the work of authorship, whether in Source or Object form, made available under the License, as indicated by a copyright notice that is included in or attached to the work (an example is provided in the Appendix below).

“Derivative Works” shall mean any work, whether in Source or Object form, that is based on (or derived from) the Work and for which the editorial revisions, annotations, elaborations, or other modifications represent, as a whole, an original work of authorship. For the purposes of this License, Derivative Works shall not include works that remain separable from, or merely link (or bind by name) to the interfaces of, the Work and Derivative Works thereof.

“Contribution” shall mean any work of authorship, including the original version of the Work and any modifications or additions to that Work or Derivative Works thereof, that is intentionally submitted to Licensor for inclusion in the Work by the copyright owner or by an individual or Legal Entity authorized to submit on behalf of the copyright owner. For the purposes of this definition, “submitted” means any form of electronic, verbal, or written communication sent to the Licensor or its representatives, including but not limited to communication on electronic mailing lists, source code control systems, and issue tracking systems that are managed by, or on behalf of, the Licensor for the purpose of discussing and improving the Work, but excluding communication that is conspicuously marked or otherwise designated in writing by the copyright owner as “Not a Contribution.”

“Contributor” shall mean Licensor and any individual or Legal Entity on behalf of whom a Contribution has been received by Licensor and subsequently incorporated within the Work.

2. Grant of Copyright License. Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute the Work and such Derivative Works in Source or Object form.

3. Grant of Patent License. Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable (except as stated in this section) patent license to make, have made, use, offer to sell, sell, import, and otherwise transfer the Work, where such license applies only to those patent claims licensable by such Contributor that are necessarily infringed by their Contribution(s) alone or by combination of their Contribution(s) with the Work to which such Contribution(s) was submitted. If You institute patent litigation against any entity (including a cross-claim or counterclaim in a lawsuit) alleging that the Work or a Contribution incorporated within the Work constitutes direct or contributory patent infringement, then any patent licenses granted to You under this License for that Work shall terminate as of the date such litigation is filed.

4. Redistribution. You may reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that You meet the following conditions:

(a) You must give any other recipients of the Work or Derivative Works a copy of this License; and

(b) You must cause any modified files to carry prominent notices stating that You changed the files; and

(c) You must retain, in the Source form of any Derivative Works that You distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work, excluding those notices that do not pertain to any part of the Derivative Works; and

(d) If the Work includes a “NOTICE” text file as part of its distribution, then any Derivative Works that You distribute must include a readable copy of the attribution notices contained within such NOTICE file, excluding those notices that do not pertain to any part of the Derivative Works, in at least one of the following places: within a NOTICE text file distributed as part of the Derivative Works; within the Source form or documentation, if provided along with the Derivative Works; or, within a display generated by the Derivative Works, if and wherever such third-party notices normally appear. The contents of the NOTICE file are for informational purposes only and do not modify the License. You may add Your own attribution notices within Derivative Works that You distribute, alongside or as an addendum to the NOTICE text from the Work, provided that such additional attribution notices cannot be construed as modifying the License.

You may add Your own copyright statement to Your modifications and may provide additional or different license terms and conditions for use, reproduction, or distribution of Your modifications, or for any such Derivative Works as a whole, provided Your use, reproduction, and distribution of the Work otherwise complies with the conditions stated in this License.

5. Submission of Contributions. Unless You explicitly state otherwise, any Contribution intentionally submitted for inclusion in the Work by You to the Licensor shall be under the terms and conditions of this License, without any additional terms or conditions. Notwithstanding the above, nothing herein shall supersede or modify the terms of any separate license agreement you may have executed with Licensor regarding such Contributions.

6. Trademarks. This License does not grant permission to use the trade names, trademarks, service marks, or product names of the Licensor, except as required for reasonable and customary use in describing the origin of the Work and reproducing the content of the NOTICE file.

7. Disclaimer of Warranty. Unless required by applicable law or agreed to in writing, Licensor provides the Work (and each Contributor provides its Contributions) on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied, including, without limitation, any warranties or conditions of TITLE, NON-INFRINGEMENT, MERCHANTABILITY, or FITNESS FOR A PARTICULAR PURPOSE. You are solely responsible for determining the appropriateness of using or redistributing the Work and assume any risks associated with Your exercise of permissions under this License.

8. Limitation of Liability. In no event and under no legal theory, whether in tort (including negligence), contract, or otherwise, unless required by applicable law (such as deliberate and grossly negligent acts) or agreed to in writing, shall any Contributor be liable to You for damages, including any direct, indirect, special, incidental, or consequential damages of any character arising as a result of this License or out of the use or inability to use the Work (including but not limited to damages for loss of goodwill, work stoppage, computer failure or malfunction, or any and all other commercial damages or losses), even if such Contributor has been advised of the possibility of such damages.

9. Accepting Warranty or Additional Liability. While redistributing the Work or Derivative Works thereof, You may choose to offer, and charge a fee for, acceptance of support, warranty, indemnity, or other liability obligations and/or rights consistent with this License. However, in accepting such obligations, You may act only on Your own behalf and on Your sole responsibility, not on behalf of any other Contributor, and only if You agree to indemnify, defend, and hold each Contributor harmless for any liability incurred by, or claims asserted against, such Contributor by reason of your accepting any such warranty or additional liability.

END OF TERMS AND CONDITIONS

APPENDIX: How to apply the Apache License to your work.

To apply the Apache License to your work, attach the following boilerplate notice, with the fields enclosed by brackets “[]” replaced with your own identifying information. (Don’t include the brackets!) The text should be enclosed in the appropriate comment syntax for the file format. We also recommend that a file or class name and description of purpose be included on the same “printed page” as the copyright notice for easier identification within third-party archives.

Copyright 2020 Kenji Kamimoto, Christy Hoffmann, Samantha Morris

Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at

<http://www.apache.org/licenses/LICENSE-2.0>

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

## Authors and citations[¶](#authors-and-citations)

### Cite celloracle[¶](#cite-celloracle)

If you use celloracle please cite our bioarxiv preprint TITLE.

### celloracle software development[¶](#celloracle-software-development)

celloracle is developed and maintained by the effort of [the members of Samantha Morris Lab](http://morrislab.wustl.edu/lab-members/). Please post troubles or questions on [the Github repository](https://github.com/KenjiKamimoto-wustl122/CellOracle/issues).

# Indices and tables[¶](#indices-and-tables)

* [Index](genindex.html)
* [Module Index](py-modindex.html)
* [Search Page](search.html)

© Copyright 2019, Samantha Morris Lab

Built with [Sphinx](http://sphinx-doc.org/) using a [theme](https://github.com/rtfd/sphinx_rtd_theme) provided by [Read the Docs](https://readthedocs.org).