

时空数据分析软件 (SDAS) 使用手册

版本: SDAS_beta.2

作者: 三箭齐发

更新日期: 2025-04-30

目录

- 一、软件简介
- 二、安装指南
- 三、快速入门
- 四、各功能模块使用说明
- 五、各功能模块性能测试
- 六、流程示例
- 七、常见问题 (FAQ)

附录 SDAS日志和退出码规范

附录 SAW aggr(alpha ver.)流程介绍

一、软件简介

StereoDataAnalysisSolution (SDAS)软件：一套时空组学数据分析的生信工具，囊括了空间组学数据解读的关键步骤。

- 功能模块： 9个功能模块，共14个算法；具体有数据处理，细胞类型注释，空间结构域识别，CNV分析，差异基因识别，基因集富集分析，空间基因共表达，细胞通讯分析，轨迹分析。
- 输入： h5ad文件（如SAW count，SAW convert gef2h5ad生成） ， SAW aggr生成的h5mu文件， rds文件
- 输出： h5ad, rds, csv, txt, png和pdf文件

	功能模块	算法/方法	环境	核心包版本	必须输入文件	输出文件形式
1	数据处理	-	python 3.10; R 4.3.3	schard_0.0.1 seurat_5.1.0 scanpy 1.10.4	h5ad/h5mu/csv	h5ad; rds; txt
2	细胞类型注释	Spotlight	R 4.3.3	spotlight 1.6.3	h5ad	h5ad; csv; png; pdf
		cell2location	python 3.10	cell2location 0.1.4	h5ad	
		RCTD	R 4.3.3	r-spacexr 2.2.1	h5ad	
3	空间结构域识别	GraphST	python 3.8	GraphST 1.1.1	h5ad	h5ad; png; pdf
4	CNV分析	inferCNV	R 4.2.3	infercnv 1.14.2	rds和h5ad	rds; csv; png; pdf
5	空间基因共表达	hotspot	python 3.10	nest-analysis 1.0.4	h5ad	h5ad; csv; png; pdf
		nest		hotspotsc 1.1.1	h5ad	
6	差异基因分析	scanpy	python 3.10	scanpy 1.10.4	h5ad 和 csv	csv; png; pdf
7	基因集富集分析	gsea	python 3.10	gseapy 1.1.4	h5ad 和 csv	csv; png; pdf
		enrichr			csv	csv; png; pdf
		prerank			csv	csv; png; pdf
		gsva			h5ad 和 csv	csv; png; pdf
8	细胞通讯分析	cellchat v2	R 4.3.3	CellChat 2.1.2	rds	rds; csv; png; pdf
9	轨迹分析	monocle3	R 4.3.3	monocle3 1.3.1	rds	rds; csv; png; pdf

1. 系统配置要求

SDAS分析软件包在Linux系统上被解压和安装，计算环境需满足下列基本要求：

- 8-core Intel or AMD processor (>24 cores recommended)
- 128GB RAM (>256GB recommended)
- 200G free disk space or higher
- 64-bit CentOS/RedHat 7.8 or Ubuntu 20.04

2. 软件下载

华大网盘：<https://bgipan.genomics.cn/>（链接和提取密码通过其他方式获取）

Github: `git clone https://github.com/STOmics/SDAS.git`（无tar.gz和测试数据）

3. 解压

```
tar -xzf SDAS_beta.2.tar.gz
cd SDAS_beta_v2504/anno
./conda-unpack
```

4. 测试运行

```
./SDAS -h
```

其他说明：

1. 软件压缩包共14G，解压后34G

2. 关于GPU使用的说明：请用户自行安装CUDA >= 11.7，参考<https://docs.nvidia.com/cuda/cuda-installation-guide-linux/>

1. 获取SAW输出的文件:

方式一: 直接下载SAW count输出的h5ad文件

方式二: 从gef转h5ad, `saw convert gef2h5ad --gef=sample.tissue.gef --bin-size=100 --h5ad=sample_bin100.h5ad`

方式三: 直接下载SAW aggr(alpha ver.)输出的h5mu文件

2. 进行细胞类型注释 (cell2location、RCTD) :

```
SDAS_beta_v2504/SDAS cellAnnotation cell2locationMakeRef --reference Test_data/single_slice/sample_ref.h5ad -o  
output/cell2location_ref --label_key annotation2 --filter_rare_cell 0 --cell_percentage_cutoff2 0.05 --nonz_mean_cutoff 1.45  
--gpu_id 0
```

```
SDAS_beta_v2504/SDAS cellAnnotation cell2location -i Test_data/single_slice/sample.h5ad -o output/cell2location --  
reference_csv output/cell2location_ref/sample_ref_inf_aver.csv --input_gene_symbol_key _index --bin_size 100 --gpu_id 0
```

```
SDAS_beta_v2504/SDAS cellAnnotation rctd -i Test_data/single_slice/sample.h5ad -o output/rctd --reference  
Test_data/single_slice/sample_ref.h5ad --label_key annotation2 --input_gene_symbol_key _index --filter_rare_cell 0 --bin_size  
100
```

3. 进行空间结构域识别:

```
SDAS_beta_v2504/SDAS spatialDomain graphst -i Test_data/single_slice/sample.h5ad -o output/graphST --gpu_id 0 --tool mclust -  
-n_clusters 10 --n_hvg 3000 --bin_size 100
```

4. CNV分析

- 使用细胞注释后的h5ad, 准备rds文件:

```
SDAS_beta_v2504/SDAS dataProcess h5ad2rds -i output/rctd/sample_anno_rctd.h5ad -o output/rctd
```

- 运行inferCNV:

```
SDAS_beta_v2504/SDAS infercnv -i output/rctd/sample_anno_rctd.rds --h5ad output/rctd/sample_anno_rctd.h5ad -o output/inferCNV  
--bin_size 100 --label_key anno_rctd --gene_symbol_key _index --species human --cutoff 0.02 --ref_group_names CAF_CXCL14
```

5. 空间基因共表达分析 (hotspot) :

```
SDAS_beta_v2504/SDAS coexpress hotspot -i Test_data/single_slice/sample.h5ad -o output/hotspot --bin_size 100 --selected_genes top5000 --n_cpus 8
```

6. 差异基因分析:

```
SDAS_beta_v2504/SDAS DEG -i output/graphST/sample_graphst.h5ad -o output/DEG --diff_plan Test_data/single_slice/deg_plan.csv
```

7. 基因集富集分析 (GSEA、Enrichr) :

```
SDAS_beta_v2504/SDAS geneSetEnrichment gsea -i output/graphST/sample_graphst.h5ad -o output/gsea --gsea_plan Test_data/single_slice/gsea_plan.csv --species human
```

```
SDAS_beta_v2504/SDAS geneSetEnrichment enrichr -i output/DEG/3.vs.8.deg_filtered.csv -o output/enrichr --species human --cut_off 0.05
```

8. 细胞通讯分析:

• 使用注释后的h5ad, 准备rds文件:

```
SDAS_beta_v2504/SDAS dataProcess h5ad2rds -i output/cell2location/sample_anno_cell2location.h5ad -o output/cell2location
```

• 运行cellchat:

```
SDAS_beta_v2504/SDAS CCI cellchat -i output/cell2location/sample_anno_cell2location.rds -o output/cellchat_nospatial --bin_size 100 --label_key anno_cell2location --gene_symbol_key _index --species human --method truncatedMean
```

9. 轨迹分析:

• 使用注释后h5ad, 准备rds文件:

```
SDAS_beta_v2504/SDAS dataProcess h5ad2rds -i output/rctd/sample_anno_rctd.h5ad -o output/rctd
```

• 运行monocle3:

```
SDAS_beta_v2504/SDAS trajectory monocle3 -i output/rctd/sample_anno_rctd.rds -o output/monocle3 --gene_symbol_key _index --root_key anno_rctd --root CAF_CXCL14 --gene_color_label pseudotime
```

四、各功能模块使用说明

- 对9个功能模块进行详细介绍，包括用途，运行方式，输入参数说明（包含参数推荐）和输出结果说明；
- 9个功能模块介绍的顺序为：数据处理，细胞类型注释，空间结构域识别，CNV分析，空间基因共表达，差异基因识别，基因集富集分析，细胞通讯分析，轨迹分析

	功能模块	算法/方法	是否支持GPU	是否支持CPU并行运算	进程/线程是否可控	并行运算开放参数	是否支持多片	是否支持单细胞数据分析
1	数据处理	-	否	否	/	/	支持	支持
2	细胞类型注释	Spotlight	否	支持	线程可控	n_threads	支持	否
		cell2location	支持	支持	线程可控	n_threads	支持	否
		RCTD	否	支持	进程可控，每个进程使用1个线程	n_cpus	支持	否
3	空间结构域识别	GraphST	支持	支持	不可控	/	支持	否
4	CNV分析	inferCNV	否	支持	/	/	支持	支持
5	空间基因共表达	hotspot	否	支持	进程可控，每个进程固定1个线程	n_cpus	否	否
		nest	否	支持	进程可控，每个进程固定1个线程	n_cpus	否	否
6	差异基因分析	scanpy	否	否	/	/	支持	支持
7	基因集富集分析	gsea	否	否	/	/	支持	支持
		enrichr	否	否	/	/	否	支持
		prerank	否	否	/	/	否	支持
		gsva	否	否	/	/	支持	支持
8	细胞通讯分析	cellchat v2	否	否	/	/	否	支持
9	轨迹分析	monocle3	否	支持	进程可控，每个进程固定1个线程	n_cpus	支持	支持

h5ad文件：为anndata对象

1. 原始表达矩阵: 可存在adata.layers (在SDAS中通过--layer参数来指定key) , 或者adata.raw.X, adata.X
2. 基因名 (即gene symbol) : 可存在adata.var["real_gene_name"], 或adata.var的其他列 (在SDAS中通过--gene_symbol_key来指定key)
3. 空间坐标 (x,y) : 存在adata.obsm["spatial"], 用于scanpy.pl.spatial()画空间图或者其他用途

rds文件：为Seurat v4 assay对象

1. 原始表达矩阵: 存在默认assay的counts中(如@assays\$RNA@counts), 或者在SDAS中通过--assay指定assay名
2. 基因名 (即gene symbol) : 可存在meta.features["real_gene_name"], 或meta.features的其他列 (在SDAS中通过--gene_symbol_key来指定key)
3. 空间坐标 (x,y) : 存在@reductions\$spatial中, 用于FeaturePlot(reduction= "spatial") , DimPlot(reduction = "spatial") 画空间图

用户可以按照如上h5ad和rds格式说明自行准备文件，作为SDAS的输入

用途：对SAW的h5ad进行多片h5ad合并，h5ad转rds， SAW aggr输出的h5mu转h5ad， 打印h5ad信息，提取h5ad的子集
运行方式：

- mergeAdata: SDAS dataProcess mergeAdata -i mult.csv -o outdir
- h5ad2rds: SDAS dataProcess h5ad2rds -i st.h5ad -o outdir
- h5mu2h5ad: SDAS dataProcess h5mu2h5ad -i st.h5mu -o outdir
- printAdataInfo: SDAS dataProcess printAdataInfo -i st.h5ad -o outdir
- subsetAdata: SDAS dataProcess subsetAdata -i st.h5ad --subset_key total_counts --min 100 --max 5000 -o outdir
- subsetAdata: SDAS dataProcess subsetAdata -i st.h5ad --subset_key anno_spotlight --list_include B,Fibroblast -o outdir

输入参数说明

dataProcess参数	是否必须	描述
-i / --input	是	输入为SAW生成的h5ad， SAW aggr输出的h5mu或者csv文件（mergeAdata进行多片合并时，输入为csv文件（第一行#开头为表头））
-o / --output	否	输出文件夹， printAdataInfo不加-o时，将adata信息输出到shell
--label_key	是	在subsetAdata使用，提取adata子集使用的obs或者var的列名
--run_mode	否	在h5ad2rds使用，输入数据类型，stRNA或scRNA，默认为stRNA
--gene_symbol_key	否	在mergeAdata使用，h5ad.var中表示基因名(symbol)的列的名称（_index 表示使用h5ad.var.index）
--layer	否	在h5ad2rds和subsetAdata使用，指定h5ad存放raw counts的layer层
--list_include	否	在subsetAdata使用，当label_key是列表时，需要提取的列表元素，比如Fibroblast,B,NK
--list_exclude	否	在subsetAdata使用，当label_key是列表时，不需要提取的列表元素，比如Fibroblast,B,NK
--min	否	在subsetAdata使用，当label_key是数值时，提取的最小值
--max	否	在subsetAdata使用，当label_key是数值时，提取的最大值

多片输入为csv文件示例，包含样本名，分组信息，样本的h5ad绝对路径，以逗号分隔（第一行#开头为表头）

```
#sample,group,h5adPath
sample1,control,sample1.h5ad.path
sample2,control,sample2.h5ad.path
sample3,test,sample3.h5ad.path
sample4,test,sample4.h5ad.path
```

输出结果说明

结果文件	描述
<input_name>.h5ad	h5mu转换的h5ad
<input_name>_subset.h5ad	subsetAdata的h5ad
combine.h5ad	多片合并的h5ad
<input_name>_rds	由h5ad转换的rds
<input_name>_adata_info.txt	adata的详细信息

<input_name>_adata_info.txt实例

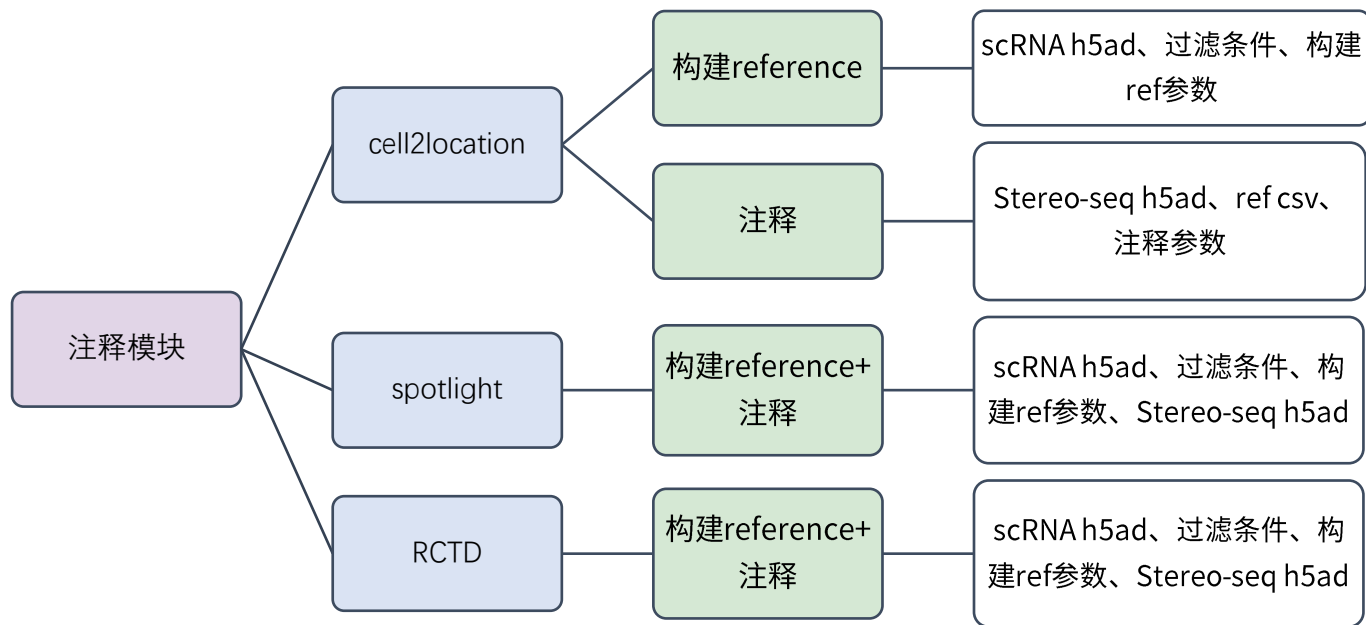
```
AnnData object with n_obs × n_vars = 320 × 32757
  obs: 'total_counts', 'n_genes_by_counts', 'pct_counts_mt', 'leiden', 'orig.ident', 'x', 'y'
  var: 'read_gene_name', 'n_cells', 'n_counts', 'mean_umt', 'mean', 'dispersions', 'dispersions_norm', 'highly_variable'
  uns: 'bin_size', 'bin_type', 'gene_exp_leiden', 'hug', 'leiden_resolution', 'neighbors', 'omics', 'pca_variance_ratio',
  'rank_genes_groups', 'resolution', 'sn'
  obsm: 'X_pca', 'X_umap', 'spatial'
  layers: 'raw_counts'
  obsm: 'connectivities', 'distances'

The 'obs' attribute of the AnnData contains 7 columns.
The 'var' attribute of the AnnData contains 8 columns.

Number of unique values in each column of 'obs'(except 'total_counts', 'n_genes_by_counts', 'pct_counts_mt', 'x', 'y'):
leiden: 30 unique values
orig.ident: 1 unique values

Unique values in each column of 'obs':
-----
leiden: Index(['1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '13',
              '14', '15', '16', '17', '18', '19', '20', '21', '22', '23', '24', '25', '26',
              '27', '28', '29', '30'],
              dtype='object')
-----
orig.ident: Index(['sample'], dtype='object')
```


细胞注释模块：整体介绍



```
(# ^.^ #)SDAS cellAnnotation -h
usage: main.pyc [-h] [--version] {cell2location,cell2locationMakeRef,spotlight,rctd} ...

annotate celltype of spatial transcriptomics data

positional arguments:
  {cell2location,cell2locationMakeRef,spotlight,rctd}
    cell2location          Select annotation method
    cell2locationMakeRef   cell2location
    spotlight              make cell2location reference
    rctd                   SPOTlight
                           RCTD

options:
  -h, --help            show this help message and exit
  --version              show program's version number and exit
```

输入h5ad格式说明:

1. 空转和单细胞数据都是raw counts; 可以通过--layer参数指定原始表达矩阵, 没有指定时会用raw.X的矩阵, raw.X也没有时会使用.X的矩阵
2. 空转和单细胞数据的var中都要有基因名(gene symbol), 存在自定义的gene_symbol_key或者var.index中
3. 空转数据要有空间坐标, 存在obs里的'x'和'y'列中; 或存在obsm里的'spatial'中
4. 多片空转数据可以提供自定义的slice_key, 默认为'sampleID', 用于画图(在cell2location中也用于提供批次信息)。如果不提供, 则当做单片处理
5. 单细胞obs中要有注释信息, 存在自定义的label_key中; **rctd算法不支持注释信息有 '/' 的情况**
6. 单细胞obs中可以有批次信息, 存在自定义的batch_key中; 如果不提供, 则会使用所有样本

输出h5ad格式说明:

1. 输入h5ad+注释结果: 每个细胞类型的分数存在obsm['anno_score_cell2location']中, 分数最高的类型存在obs['anno_cell2location']中
2. 统一未被注释的细胞的类型为NA

细胞注释模块： cell2location 构建单细胞参考csv

用途：构建cell2location的单细胞参考inf_aver.csv文件

运行方式：SDAS cellAnnotation cell2locationMakeRef -o ./ref --reference sc.h5ad --label_key annotation --batch_key id --nonz_mean_cutoff 1.45 --gpu_id 3

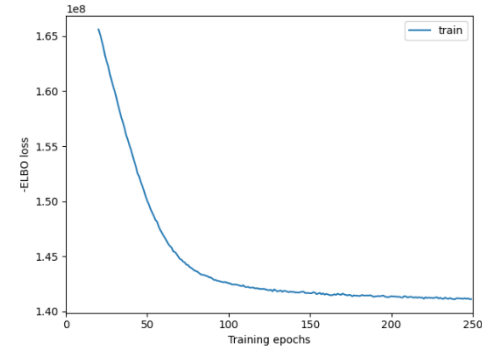
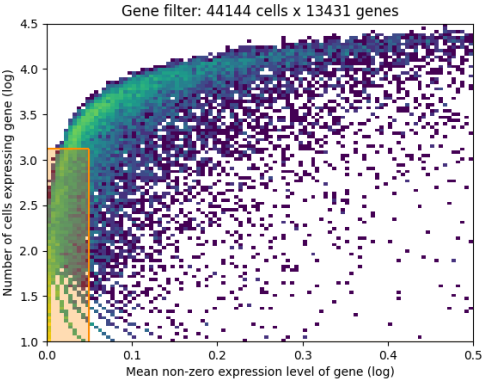
输入参数说明

cell2locationMakeRef参数	是否必须	默认值	描述
-o / --output	是		输出文件夹
--reference	是		单细胞ref h5ad, 要求有原始表达矩阵
--label_key	是		单细胞ref h5ad.obs中表示细胞类型的列的名称
--ref_layer	否		单细胞ref h5ad存放raw counts的layer
--ref_gene_symbol_key	否	_index	单细胞ref h5ad.var中表示基因名(symbol)的列的名称 (_index 表示使用h5ad.var.index)
--batch_key	否		单细胞ref h5ad.obs中表示批次的列的名称, 不输入则不考虑批次
--filter_rare_cell	否	100	如果某些细胞类型在单细胞ref中细胞数小于此值, 则过滤掉这些细胞类型
--check_filter_genes	否		如果设置此参数, 则只输出筛选基因的结果图filter_genes.png
--cell_count_cutoff	否	5	控制cell2location筛选基因的参数, 一般不调整
--cell_percentage_cutoff2	否	0.03	控制cell2location筛选基因的参数, 值越大筛选出的基因越少, 基因数推荐控制在8k-16k
--nonz_mean_cutoff	否	1.12	控制cell2location筛选基因的参数, 值越大筛选出的基因越少, 基因数推荐控制在8k-16k
--max_epochs	否	250	模型训练epoch数
--seed	否	42	指定随机种子
--gpu_id	否	-1	使用的GPU的编号, 如果为-1, 则使用CPU。此参数只指定主要使用的GPU, 其他GPU也会被占用, 但占用量很低。如果需要严格指定GPU, 请在运行前设置环境变量, 如: export CUDA_VISIBLE_DEVICES=2, 此时再设置--gpu_id 0, 则会只使用2号GPU
--n_threads	否		CPU模式下使用的线程数, 默认为全部CPU

输出结果说明

cell2locationMakeRef结果文件	描述
<reference_name>_filter_genes.png	cell2location筛选基因的结果图 (<reference_name>为单细胞ref h5ad文件名)
<reference_name>_train_history.png	训练Loss下降图
<reference_name>_inf_aver.csv	cell2location构建的单细胞ref csv

*输出结果展示:



	Fibroblast	B cells	Lymphatic	Squamous	Neutrophil
AL627309.1	0.000721	0.000315	0.002332	0.018779	0.003534
AP006222.1	0.047347	0.005848	0.057651	0.076984	0.005694
SAMD11	0.166152	0.003667	0.002706	0.000453	0.001062
NOC2L	0.060385	0.027762	0.056389	0.20657	0.004468
PLEKHN1	0.000367	0.000175	0.002783	0.036837	0.001424
HES4	0.28838	0.013985	0.885376	0.419497	0.023717
ISG15	0.746277	0.112931	1.634906	0.729634	0.485852
AGR1	0.044908	0.001752	0.162751	0.064732	0.002179
RNF223	0.000225	0.000142	0.003744	0.022773	0.001292
C1orf159	0.008118	0.000874	0.008737	0.023208	0.001796
TNFRSF18	0.010459	0.005404	0.018063	0.164901	0.00889

细胞注释模块：cell2location解卷积

用途：使用cell2location做解卷积细胞注释

运行方式：SDAS cellAnnotation cell2location -i st.h5ad -o outdir --reference_csv ./ref/inf_aver.csv --bin_size 20 --input_gene_symbol_key_index --gpu_id 3

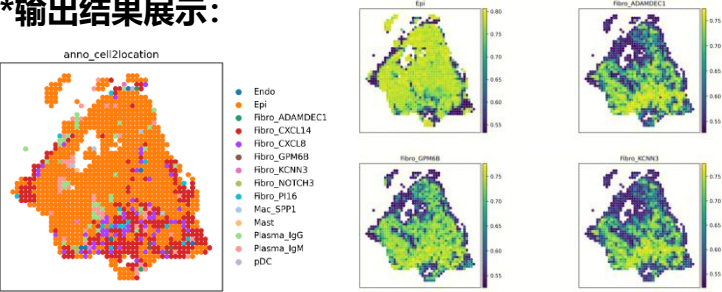
输入参数说明

cell2location参数	是否必须	默认值	描述
-i / --input	是		Stereo-seq h5ad, 要求有原始表达矩阵
-o / --output	是		输出文件夹
--reference_csv	是		单细胞ref csv文件
--bin_size	是		Bin大小, 用于控制每个bin的细胞数和图中点的大小; 如20, 50, 100, cellbin (等效于20)
--input_layer	否		Stereo-seq h5ad存放raw counts的layer
--input_gene_symbol_key	否	real_gene_name	Stereo-seq h5ad.var中表示基因名(symbol)的列的名称 (_index 表示使用 h5ad.var.index)
--slice_key	否	sampleID	多片h5ad.obs中表示片编号的列的名称, 提供批次信息和用于画图
--detection_alpha	否	20	规则化参数。空转数据的技术性变异越大, 适合的detection_alpha越小, 一般不调整
--data_split_strategy	否	chunk	当bin数量太大时, 对空转数据进行拆分, 此参数为数据拆分策略。chunk 表示先随机拆分再运行cell2location, batch表示在算法内部进行拆分
--data_split_size	否	10000	当bin数量太大时, 对空转数据进行拆分, 此参数为拆分的数据大小。越大运行得越快, 但所占显存也越大。如果为-1, 则不进行拆分
--max_epochs	否	5000	模型训练epoch数
--seed	否	42	指定随机种子
--gpu_id	否	-1	使用的GPU的编号, 如果为-1, 则使用CPU。此参数只指定主要使用的GPU, 其他GPU也会被占用, 但占用量很低。如果需要严格指定GPU, 请在运行前设置环境变量, 如: export CUDA_VISIBLE_DEVICES=2, 此时再设置--gpu_id 0, 则会只使用2号GPU
--n_threads	否		CPU模式下使用的线程数, 默认为全部CPU

输出结果说明

cell2location结果文件	描述
<input_name>_anno_cell2location.csv	每个spot的注释结果, 包括每种细胞类型的分数
<input_name>_anno_cell2location.h5ad	输入h5ad+注释结果。每个细胞类型的分数存在 obsm['anno_score_cell2location']中, 分数最高的类型存在 obs['anno_cell2location']中
<input_name>_anno_cell2location.png/pdf	总体注释结果图, 多片情况下每片画一张图, 同时输出png和pdf
<input_name>_anno_score_cell2location.png/pdf	每个细胞类型的分数图, 多片情况下每片画一张图, 同时输出png和pdf

*输出结果展示:



不同细胞类型分数和总体结果的表格

index	B_act	B_naive	CD4_CXCL	CD4_Tcm	CD4_Treg	CD4_act	CD8_Cyto
CRCP99_T	0.537523	0.536948	0.533196	0.534602	0.534567	0.540949	0.530005
CRCP99_T	0.534857	0.540527	0.552721	0.540647	0.54286	0.536034	0.537996
CRCP99_T	0.53337	0.538663	0.535225	0.532017	0.534083	0.528293	0.538153
CRCP99_T	0.538592	0.533116	0.533197	0.530321	0.526087	0.526096	0.534302
CRCP99_T	0.531412	0.530635	0.530761	0.530957	0.533807	0.533881	0.523617
CRCP99_T	0.538245	0.534404	0.549497	0.539566	0.539161	0.541039	0.530749
CRCP99_T	0.536277	0.531047	0.542305	0.539332	0.535449	0.535668	0.533795
CRCP99_T	0.531966	0.534707	0.530665	0.535111	0.531214	0.533513	0.537625
CRCP99_T	0.538371	0.529251	0.536385	0.529158	0.53626	0.532688	0.528587
CRCP99_T	0.534723	0.525861	0.537191	0.5304	0.531984	0.534209	0.527593
CRCP99_T	0.533587	0.530908	0.540009	0.536223	0.5392	0.534765	0.529211
CRCP99_T	0.538523	0.527799	0.53069	0.527776	0.534062	0.533603	0.529701
CRCP99_T	0.535306	0.535971	0.533804	0.528555	0.535749	0.53728	0.526181

细胞注释模块：spotlight解卷积

用途：使用SPOTlight做解卷积细胞注释

运行方式：SDAS cellAnnotation spotlight -i st.h5ad -o outdir --reference sc.h5ad --bin_size 20 --label_key annotation2 --input_gene_symbol_key_index

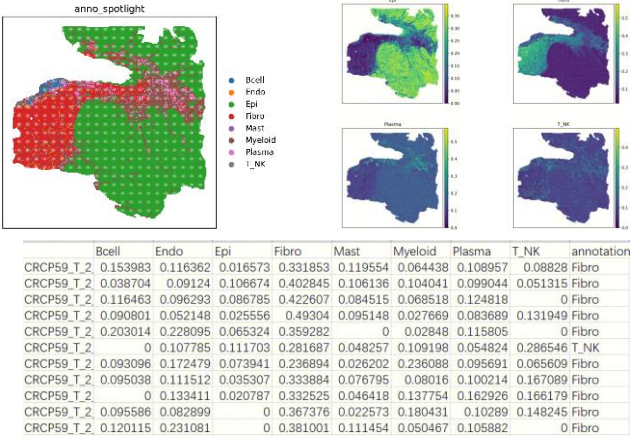
输入参数说明

spotlight参数	是否必须	默认值	描述
-i / --input	是		Stereo-seq h5ad, 要求有原始表达矩阵
-o / --output	是		输出文件夹
--reference	是		单细胞ref h5ad, 要求有原始表达矩阵
--label_key	是		单细胞ref h5ad.obs中表示细胞类型的列的名称
--bin_size	是		Bin大小, 用于控制图中点的大小, 不用于计算,比如20,50,100, cellbin (等效于20)
--input_layer	否		Stereo-seq h5ad存放raw counts的layer
--ref_layer	否		单细胞ref h5ad存放raw counts的layer
--input_gene_symbol_key	否	real_gene_name	Stereo-seq h5ad.var中表示基因名(symbol)的列的名称
--ref_gene_symbol_key	否	_index	单细胞ref h5ad.var中表示基因名(symbol)的列的名称 (_index 表示使用h5ad.var.index)
--batch_key	否		单细胞ref h5ad.obs中表示批次的列的名称, 不输入则不考虑批次
--batch	否		如果指定, 则只使用该批次的细胞做ref, 否则使用所有细胞
--slice_key	否	sampleID	多片h5ad.obs中表示片编号的列的名称, 用于画图
--filter_rare_cell	否	100	如果某些细胞类型在单细胞ref中细胞数小于此值, 则过滤掉这些细胞类型
--n_cells	否	100	单细胞ref中每个细胞类型随机选取的细胞数, 用于训练spotlight模型
--n_hvg	否	3000	单细胞ref中的高变基因个数, 高变基因和marker基因共同作为基因集
--auc_threshold	否	0.5	用于过滤单细胞ref中每个细胞类型marker基因的AUC阈值, 高变基因和marker基因共同作为基因集
--norm_sc	否	False	是否用logNormCounts函数处理单细胞ref数据
--norm_sp	否	False	是否用logNormCounts函数处理Stereo-seq数据
--seed	否	42	指定随机种子
--n_threads	否	8	使用的线程数

输出结果说明

spotlight结果文件	描述
<input_name>_anno_spotlight.csv	每个spot的注释结果, 包括每种细胞类型的分数
<input_name>_anno_spotlight.h5ad	输入h5ad+注释结果。每个细胞类型的分数存在 obsm['anno_score_spotlight']中, 分数最高的类型存在 obs['anno_spotlight']中
<input_name>_anno_spotlight	总体注释结果图, 多片情况下每片画一张图, 同时输出png和pdf
<input_name>_anno_score_spotlight	每个细胞类型的分数图, 多片情况下每片画一张图, 同时输出png和pdf

*输出结果展示:



细胞注释模块：RCTD解卷积

用途：使用RCTD做解卷积细胞注释

运行方式：SDAS cellAnnotation rctd -i st.h5ad -o outdir --reference sc.h5ad --bin_size 20 --label_key annotation2 --filter_rare_cell 0

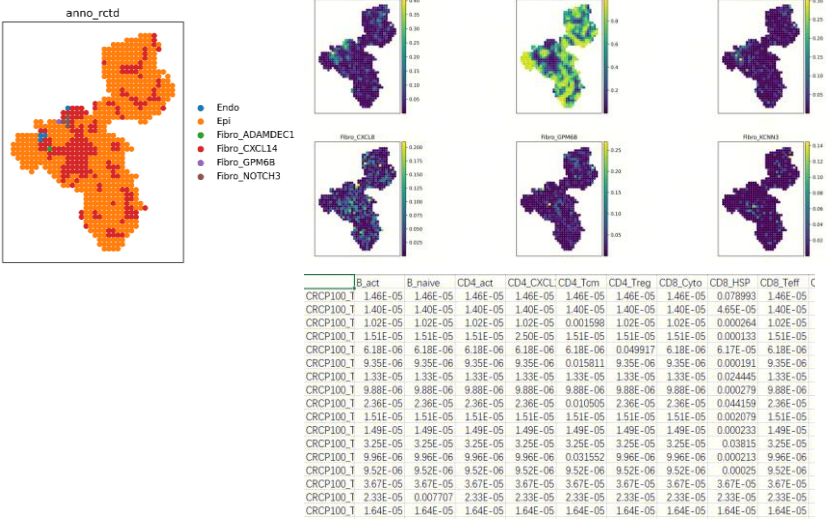
输入参数说明

rctd参数	是否必须	默认值	描述
-i / --input	是		Stereo-seq h5ad, 要求有原始表达矩阵
-o / --output	是		输出文件夹
--reference	是		单细胞ref h5ad, 要求有原始表达矩阵
--label_key	是		单细胞ref h5ad.obs中表示细胞类型的列的名称
--bin_size	是		Bin大小, 用于控制图中点的大小, 不用于计算,比如20,50,100, cellbin (等效于20)
--input_layer	否		Stereo-seq h5ad存放raw counts的layer
--ref_layer	否		单细胞ref h5ad存放raw counts的layer
--input_gene_symbol_key	否	real_gene_name	Stereo-seq h5ad.var中表示基因名(symbol)的列的名称
--ref_gene_symbol_key	否	_index	单细胞ref h5ad.var中表示基因名(symbol)的列的名称 (_index 表示使用h5ad.var.index)
--batch_key	否		单细胞ref h5ad.obs中表示批次的列的名称, 不输入则不考虑批次
--batch	否		如果指定, 则只使用该批次的细胞做ref, 否则使用所有细胞
--slice_key	否	sampleID	多片h5ad.obs中表示片编号的列的名称, 用于画图
--filter_rare_cell	否	100	如果某些细胞类型在单细胞ref中细胞数小于此值, 则过滤掉这些细胞类型
--mode	否	full	RCTD模式。选项: doublet, multi, full
--seed	否	42	指定随机种子
--n_cpus	否	8	使用的进程数

输出结果说明

rctd结果文件	描述
<input_name>_anno_rctd.csv	每个spot的注释结果, 包括每种细胞类型的分数
<input_name>_anno_rctd.h5ad	输入h5ad+注释结果。每个细胞类型的分数存在obs['anno_score_rctd']中, 分数最高的类型存在obs['anno_rctd']中
<input_name>_anno_rctd	总体注释结果图, 多片情况下每片画一张图, 同时输出png和pdf
<input_name>_anno_score_rctd	每个细胞类型的分数图, 多片情况下每片画一张图, 同时输出png和pdf

*输出结果展示:



用途：使用graphST算法进行空间结构域识别

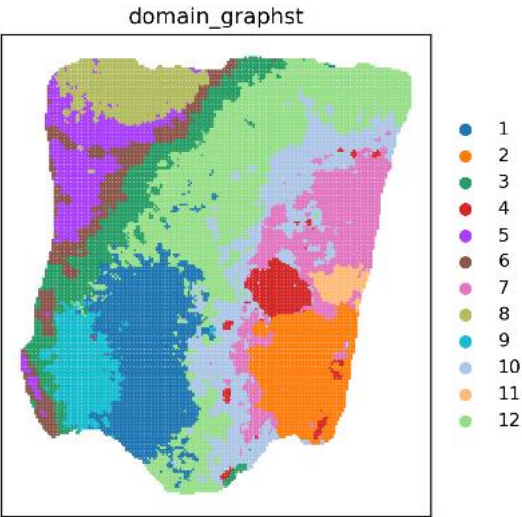
运行方式：

- **CPU模式：** SDAS spatialDomain graphst -i st.h5ad -o outdir --gpu_id -1 --tool ‘mclust’ --n_clusters 10 --n_hvg 3000
- **GPU模式：** SDAS spatialDomain graphst -i st.h5ad -o outdir --gpu_id 1 --tool ‘mclust’ --n_clusters 10 --n_hvg 3000

输入参数说明：

graphst参数	是否必须	默认值	描述
-i / --input	是		Stereo-seq h5ad, 要求有原始表达矩阵
-o / --output	是		输出文件夹
--n_cluster	是		聚类数目
--tool	否	‘mclust’	Graphst所选的聚类方法, 可选 ‘mclust’ , ‘leiden’ , ‘louvian’
--bin_size	否	20	Bin大小, 用于控制图中点的大小, 不用于计算,比如20,50,100
--gpu_id	否	-1	使用的GPU的编号, 如果为-1, 则使用CPU
--n_hvg	否	3000	当输入数据中有' highly_variable'字段时, 会直接调用现有高变基因集, 并选取--n_hvg参数设置的高变基因数量。反之, 则GraphST会进行预处理并计算高变基因, 并选取 --n_hvg参数设置的高变基因数量
--slice_key	否	None	多片h5ad.obs中表示片编号的列的名称, 如果是多片, 必须指定
--seed	否	42	随机种子设置
--layer	否	None	指定Stereo-seq h5ad存放raw counts的layer

*输出结果展示:



输出结果说明

cell2location结果文件	描述
<input_name>_graphst.h5ad	输入h5ad+空间域聚类结果, 聚类结果保存在obs[‘graphst_domain’]中
graphst_domain.png	空间域cluster类在组织上分布图

*符合下列情况，可以进行多片分析：

- 1、邻片，组织纵向切片，做空间域分析前，需要进行邻片配准，参数slice_key设置为h5ad的obs中记录多片的key。
 - 2、同一个组织切片，在水平方向上分成若干连续多片，分开测序，需要多片在水平方向上配准。
- 具体可参考：
https://deepst-tutorials.readthedocs.io/en/latest/Tutorial%204_Horizontal%20Integration.html （邻片）
https://deepst-tutorials.readthedocs.io/en/latest/Tutorial%205_Vertical%20Integration.html （水平方向）

用途：使用infercnv做空间转录组CNV分析

运行方式：SDAS inferCNV -i st.rds --h5ad st.h5ad --bin_size 50 --slice_key batch -o outdir --label_key anno_cell2location --species human --ref_group_names B,T --min_counts_per_cell 200

输入参数说明

InferCNV参数	是否必须	默认值	描述
-i / --input	是		rds文件，要求有原始表达矩阵
-o / --output	是		所有文件的输出目录
--h5ad	是		h5ad格式的sample.h5ad，用于画CNV score的空间热图，以单细胞模式运行时，该值无需设置
--bin_size	是		Bin大小，用于控制图中点的大小，不用于计算,比如20,50,100，以单细胞模式运行时，该值无需设置
--label_key	是		rds对象中存在metadata中的注释信息，cluster信息或样本信息字段
--ref_group_names	否	None	推断infercnv时当作reference的正常 cell或normal sample的分组，默认为用所有cell当作reference(不推荐)
--gene_order_file	否	None	指定用户自定义的所有基因的染色体位置信息的文本文件
--cluster_heatmap	否	False	绘制CNV heatmap时是否对cell进行聚类展示，可选True或者False
--species	否	human	指定预先构建好的物种的*_pos.txt， 'human' 或者 'mouse' ，默认 'human' ，当指定--gene_order_file参数时，该参数不起作用
--slice_key	否	batch	多片h5ad.obs中表示片编号的列的名称，用于画图
--gene_symbol_key	否	real_gene_name	rds文件中中表示基因名(symbol)的列(meta.features)的名称（_index 表示使用矩阵的列名）
--assay	否	None	rds文件中指定用于计算cnv的assay的名称，若不设置，则使用默认的assay
--min_counts_per_cell	否	100	细胞过滤参数，每个细胞/bins中包含的最小total_counts数
--cutoff	否	0.02	基因过滤参数，参考细胞/bins中每个基因的最小平均counts数
--run_mode	否	stRNA	选择以空间转录组模式（stRNA）或是单细胞模式（scRNA）运行

gene_order_file文件格式示例：
Gene_symbol chr start end

MIR1302-2HG	1	29554	31109
FAM138A	1	34554	36081
OR4F5	1	65419	71585
AL627309.1	1	89295	133723
AL627309.3	1	89551	91105
AL627309.2	1	139790	140339
AL627309.4	1	160446	161525
AL732372.1	1	358857	366052
OR4F29	1	450703	451697
AC114498.1	1	587629	594768
OR4F16	1	685679	686673
AL669831.2	1	760911	761989
AL669831.5	1	778770	810060
FAM87B	1	817371	819837

可由gtf_to_position_file.py生成，脚本链接如下：
https://github.com/broadinstitute/infercnv/blob/master/scripts/gtf_to_position_file.py

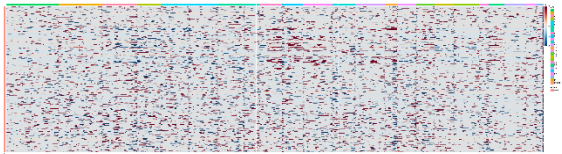
注：--species参数指定的human的gene_order_file来源于GRCh38.p12- NCBI:GCA_000001405.27。mouse的gene_order_file来源于GRCm38.p6- NCBI:GCA_000001635.8。行名均为genesymbol

输出结果说明

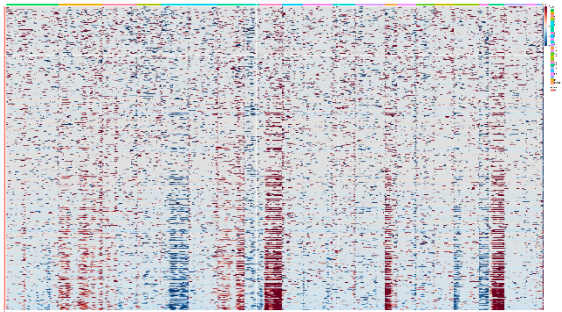
InferCNV结果文件	描述
<input_name>_run.final.infercnv_obj.rds	包含所有基因和spot的cnv矩阵的rds对象
<input_name>_CNV_score.csv	包含每个spot的cnv score
<input_name>_CNV_ref.png/pdf	参考细胞的cnv的表达热图
<input_name>_CNV_obs.png/pdf	用于观测的细胞的cnv的表达热图，当--ref_group_names为None时，该图不会输出
<input_name>_CNV_score.png/pdf	cnv score的空间热图，多片情况下每片画一张图，当以单细胞模式运行时，该图不会输出

*输出结果展示:

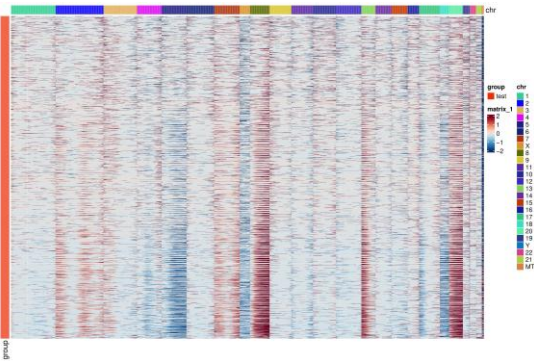
CNV_ref.png



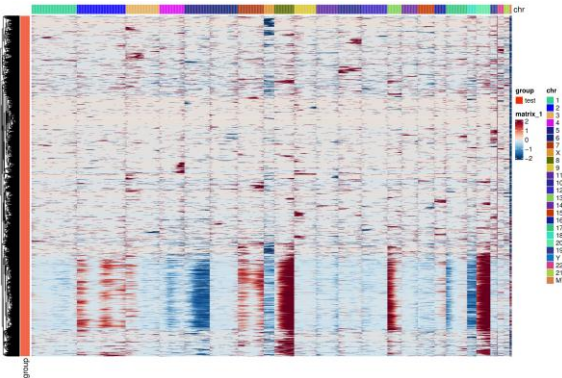
CNV_obs.png



--cluster_heatmap False



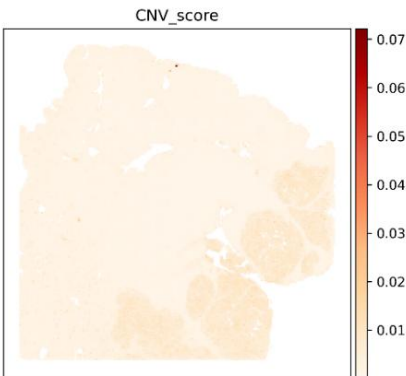
--cluster_heatmap True



CNV_score.csv

CNV_score	
429496737600_D03663C6	0.0018216
429496737700_D03663C6	0.0018385
429496738900_D03663C6	0.0009589
858993460800_D03663C6	0.0015203
858993460900_D03663C6	0.0031701
858993461000_D03663C6	0.0015783
858993461100_D03663C6	0.002563
858993461200_D03663C6	0.00436
858993461300_D03663C6	0.0094898
858993461400_D03663C6	0.0036031
858993461500_D03663C6	0.0034979
858993461600_D03663C6	0.00312
858993461700_D03663C6	0.0039391
858993461800_D03663C6	0.002287
858993461900_D03663C6	0.0019334
858993462000_D03663C6	0.0031127
858993462100_D03663C6	0.0025705
858993462200_D03663C6	0.0016642

CNV_score.png



空间基因共表达： NeST和Hotspot算法

用途：通过识别在空间上具有相似表达模式的基因模块，了解基因间相互作用脉络（高度协同变化的基因集），从而挖掘基因功能以及寻找核心基因的一类分析方法

运行方式：

- NeST: `SDAS coexpress nest -i st.h5ad -o outdir --bin_size 100`
- Hotspot: `SDAS coexpress hotspot -i st.h5ad -o outdir --bin_size 100`

输入参数说明：

coexpress参数	是否必须	默认值	描述
-i / --input	是		Stereo-seq h5ad, 要求有原始表达矩阵
-o / --output	是		输出文件夹
--bin_size	是	50	分辨率Bin大小（从20, 50, 100, 200, cellbin中选择, 与输入的h5ad一致）画图与计算均需要
--input_layer	否	None	指定h5ad中存有原始表达矩阵的layer层（例如layers['raw_counts']）
--selected_genes	否	top5000	基因列表（topn 高变基因, full 全部基因）
--moran_path	否	None	已经计算好的基因莫兰指数列表路径(包含了squidpy分析的Moran 's I 指数分析的结果, 前2列为必需列: 与h5ad一致的基因名称, 和moranI)
--n_cpus	否	8	并行计算使用的进程数
--seed	否	42	随机种子
--fdr_cutoff	否	0.05	hotspot算法中使用, 统计检验空间高变基因与共表达基因集的FDR矫正阈值
--model	否	normal	hotspot算法中使用, 统计检验假设 (normal, bernoulli, danb, none)
--hotspot_min_size	否	30	nest算法中使用, 空间高变单基因覆盖的最少spot/细胞个数
--hotspot_min_samples	否	4	nest算法中使用, 识别空间高变单基因时DBSCAN算法覆盖的最少邻域spot/细胞个数 (k neighbor)
--min_cells	否	100或者30	nest算法中使用, 共表达基因集覆盖的最少spot/细胞个数, 默认: cellbin/bin20/bin50时为100; bin100/bin200时为30

* **NeST算法调参建议：**若样本bin20/50基因数低于200，或其他特殊样本

- 识别的空间高变基因较少，建议降低hotspot_min_size到10
- 识别的空间共表达基因集较少，建议降低min_cells到10
- 空间共表达基因集识别pattern过于精细，出现 “NumPy Unable to allocate X GiB array ”报错，建议升高hotspot_min_size, hotspot_min_samples

moran_path示例:
./moran.csv

```
,moranI,pval_norm,var_norm,pval_norm_fdr_bh,real_gene_name
FABP5,0.8273024881707159,0.0,0.00010227497014174042,0.0,FABP5
STX3,0.8203151346356835,0.0,0.00010227497014174042,0.0,STX3
HSPB1,0.7993499075159156,0.0,0.00010227497014174042,0.0,HSPB1
S100A9,0.7927948719212896,0.0,0.00010227497014174042,0.0,S100A9
NTS,0.7868677451739036,0.0,0.00010227497014174042,0.0,NTS
```

selected_genes_file 示例:
./genelist

```
$head MP.genelist
A2ML1
ANXA1
ASF1B
ASPM
ATAD2
ATF3
```

空间基因共表达： NeST和Hotspot算法

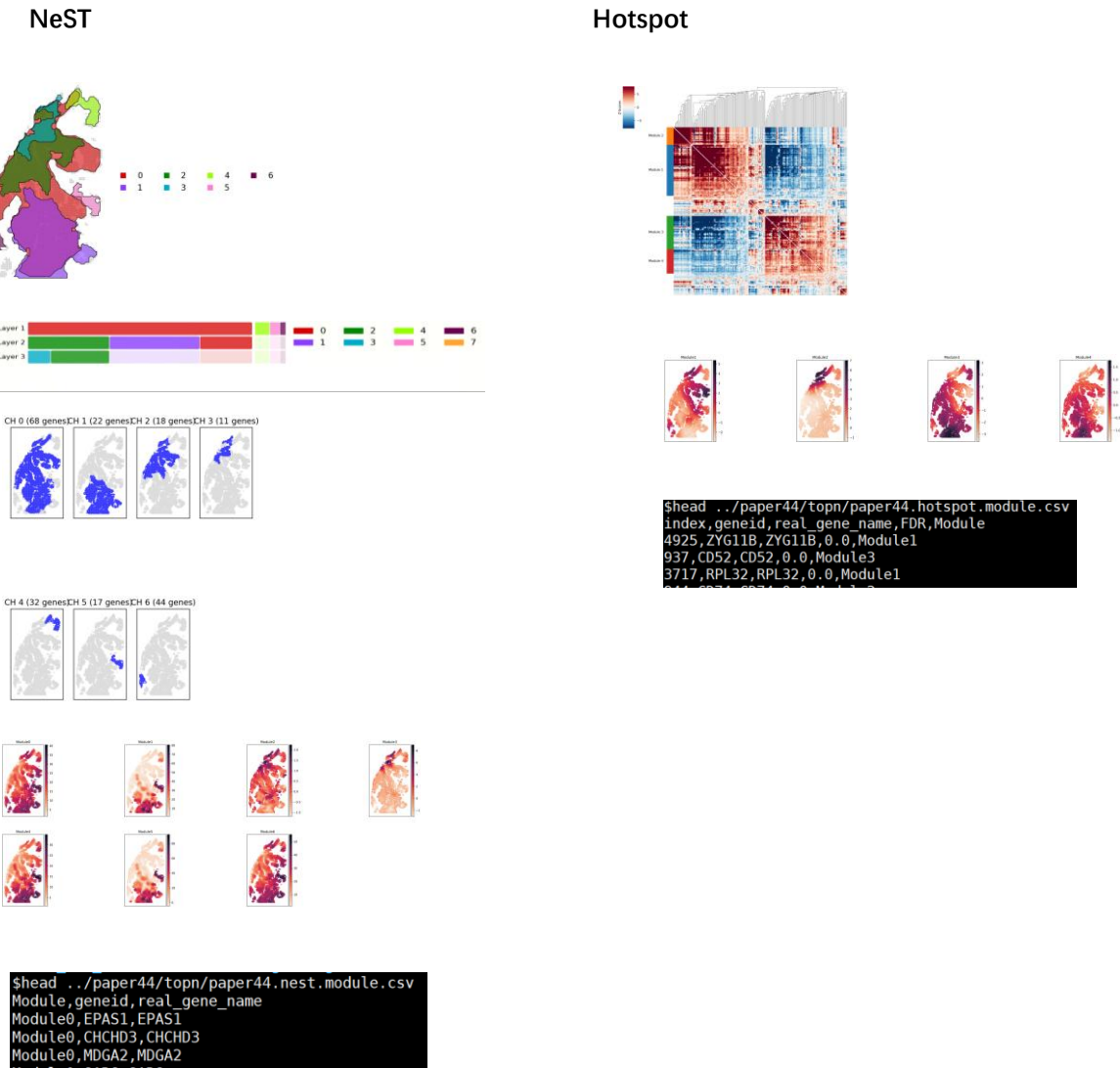
输出结果说明：

结果文件	描述
module.csv	空间高变基因（gene symbol+gene id）对应的共表达基因集（module）
h5ad	含有共表达基因集结果的h5ad文件（adata.obsm['module_score_nest'] /adata.obsm['module_score_hotspot'] ）
png/pdf	共表达基因集之间的关系图 共表达基因集单独modle score热图
moran.csv	如果使用topn计算，输出全部基因的莫兰指数以及P值

输出结果解读：

- NeST算法：共表达基因集从Module0开始， 没有Module为不符合共表达基因集聚类要求的基因
- Hotspot算法：共表达基因集从Module1开始， Module-1/没有Module为不符合共表达基因集聚类要求的基因

*输出结果展示：



差异基因分析

用途：进行基因差异表达分析

运行方式：SDAS DEG -i st.h5ad --diff_plan diff_plan.csv -o outdir

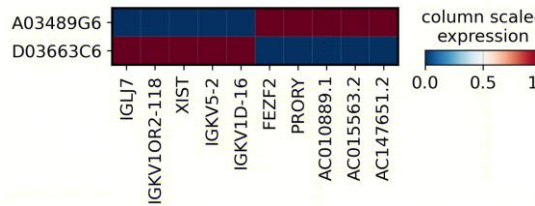
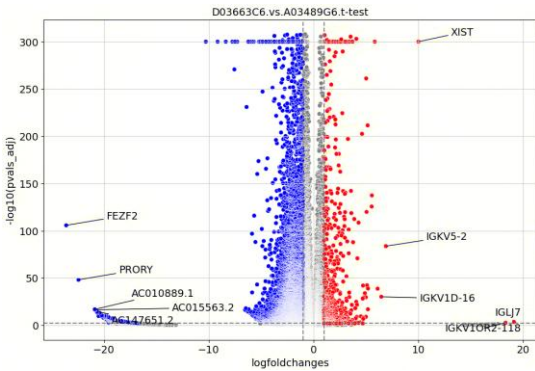
输入参数说明

DEG参数	是否必须	默认值	描述
-i / --input	是		Stereo-seq h5ad，要求有原始表达矩阵
--diff_plan	是		差异分析方案，csv格式，每个方案一行，至少包含4列信息：obs对象,处理组,对照组,差异方法
--layer	否	None	指定h5ad中存有原始表达矩阵的layer层
--gene_symbol_key	否	real_gene_name	指定用于进行分析的基因名称所在列
--padj_threshold	否	0.01	padj值的阈值，用于筛选显著差异基因
--logfc_threshold	否	1	log2FC的绝对值阈值，用于筛选显著差异基因
--top_gene	否	10	指定用于画图的基因数，只对all vs rest方案有效
--genelist	否	5	在图中标出感兴趣的基因，多个基因时用逗号“,”分割，默认上下调最显著的5个基因，对all vs rest的方案无效
--add_label	否		对h5ad的obs对象增加感兴趣的标签，再基于增加的标签信息进行差异分析
--min_gene	否	0	某个细胞或spot允许的最少基因数
--max_gene	否	20000	某个细胞或spot允许的最多基因数
--min_cell	否	0	某个基因存在的最少细胞或spot数
-o / --output	否	当前文件夹	分析结果存放文件夹

输出结果说明

DEG结果文件	描述
{处理组}.vs.{对照组}.deg.csv	处理组/对照组所有差异基因
{处理组}.vs.{对照组}.deg_filtered.csv	处理组/对照组显著差异基因
matrixplot_{处理组}.vs.{对照组}.pdf/png	显著差异基因表达量热图
volcano_plot. {处理组}.vs.{对照组}.pdf/png	差异基因火山图（all vs rest方案不做火山图）

*输出结果展示：



gene_ids	scores	logfoldch	pvals	pvals_adj	real_gene	treat_group
MTRNR2L	237.6234	5.849801	0	0	MTRNR2L	D03663C6
XIST	230.4059	10.01164	0	0	XIST	D03663C6
JCHAIN	115.5075	2.657836	0	0	JCHAIN	D03663C6
IGHA1	101.331	3.162789	0	0	IGHA1	D03663C6
EEF1A1	98.69592	2.471968	0	0	EEF1A1	D03663C6
MT-CO1	90.51968	1.116584	0	0	MT-CO1	D03663C6
IGKC	87.89241	2.113462	0	0	IGKC	D03663C6
MT-ND5	84.20027	1.048821	0	0	MT-ND5	D03663C6
PARP8	82.23981	1.817068	0	0	PARP8	D03663C6
B4GALNT2	82.01656	1.519611	0	0	B4GALNT2	D03663C6
FP671120	80.31432	1.592818	0	0	FP671120	D03663C6
RPL7	77.76428	3.313637	0	0	RPL7	D03663C6
AAMDC	74.7632	1.250964	0	0	AAMDC	D03663C6
WDR74	70.67052	2.254388	0	0	WDR74	D03663C6
MUC12	68.16598	1.189166	0	0	MUC12	D03663C6
IGHA2	65.82799	1.911193	0	0	IGHA2	D03663C6

流程说明：

- 1. 只支持h5ad格式的文件输入，且h5ad文件必须包含原始表达矩阵，可以通过--layer参数指定表达矩阵，没有指定时会用raw.X的矩阵，raw.X也没有时会使用.X的矩阵
- 2. 基因名称通过--gene_symbol_name参数指定，没有指定时使用h5ad文件var的real_gene_name，否则用var的index
- 3. 做差异前，会对基因名称进行make_unique操作，但分析结果中会输出make_unique前后的基因名称
- 4. 作图都使用make_unique后的基因名称
- 5. 流程允许对细胞和gene进行过滤，如果输入的h5ad文件已经经过过滤，可以不用再设置相关过滤参数
- 6. 流程目前只支持sc.tl.rank_genes_groups函数进行差异分析，差异方法只支持't-test', 't-test_overestim_var', 'wilcoxon', 'logreg'
- 7. 差异分析方案:
 - 以csv格式的文件提供，方案中需要指定obs的一个具体对象，以及该对象下的2个元素做为处理组和对照组，第4列信息用于指定差异分析方法
 - 支持对obs key指定的特定对象进行差异分析（如T cells），可以在差异方案的5、6列分别指定obs key及特定对象来进行细胞/bins的筛选
 - 支持对obs key的所有类别进行subset后再做差异，例如对cell_type中所有细胞类型分别进行subset再做差异

diff_plan的csv文件示例：

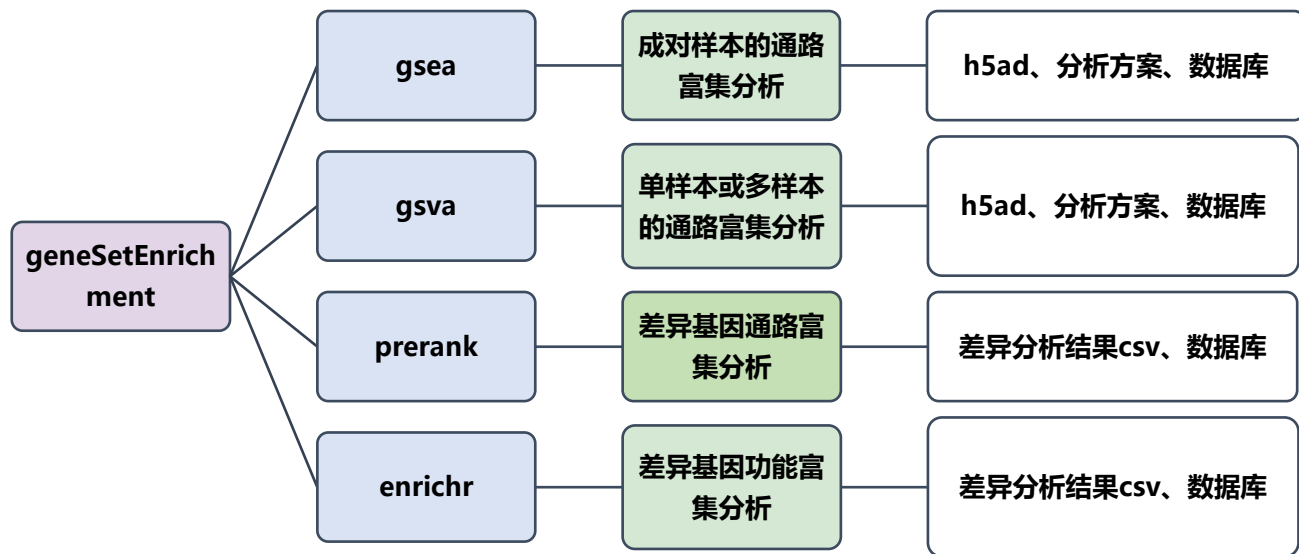
#object,treatment,control,method,1stObject,1stClass
leiden_cluster,all,rest,t-test
tumor_subregion,warm-up,cold,t-test
Sample,Treatment,control,t-test,bayes_cluster,1
Sample,Treatment,control,t-test,anno_cell2location

- 单片差异基因示例
- 单片差异基因示例
- 多片差异基因示例
- 多片差异基因示例

diff-plan.csv文件说明：

- 1、跳过'#'开头的行；
- 2、对h5ad文件obs中'leiden_cluster'列的每个cluster进行 all vs rest的差异基因分析；
- 3、对h5ad文件obs中'tumor_subregion'列指定的warm-up组和cold组进行差异基因分析；
- 4、对h5ad文件obs中'bayes_cluster'列指定的类别1，在obs中'Sample'指定的Treatment组和控制组进行差异基因分析
- 5、对h5ad文件obs中'anno_cell2location'列指定的所有细胞类型，分别在obs中'Sample'指定的Treatment组和控制组进行差异基因分析

基因集富集分析：整体介绍



流程说明：

1. 关于输入文件：

- `gsea`和`gsva` 直接基于h5ad文件的原始表达量对样本进行通路富集分析，通过`--layer`参数指定原始表达矩阵，没有指定时会用`adata.raw.X`的矩阵，`adata.raw.X`也没有时会使用`adata.X`的矩阵；
- 基因信息通过`--gene_symbol_name`指定，没有指定时使用h5ad文件var的`real_gene_name`，否则用var的index，分析中会对基因信息进行`make_unique`
- `prerank`要求输入文件有`deg_gene_name`和`logfoldchange`两列，分析结果与`gsea`类似，建议输入所有基因的结果列表
- `enrichr`要求输入文件中有`deg_gene_name`这一列，建议输入显著差异基因列表

2. 关于数据库：

- 4个富集分析功能都支持gmt格式的数据库文件，且数据库中的gene name都为大写，SDAS中有预先构建好的人和小鼠数据库
- 输入自定义的gmt数据库时，gene信息必须为大写，因为流程会将gene信息都改为大写再与数据库进行匹配
- 软件预下载了human和mouse的Msigdb数据库，如果想选择性使用某个gmt，可以在SDAS软件路径`sdas_deg_enrichment/lib/GSEADB`下根据需求使用`--gmt`进行指定

```
( # ^.^ # )SDAS geneSetEnrichment -h
usage: Gene_enrichment.pyc [-h] [--version] {gsea,prerank,gsva,enrichr} ...

Gene_enrichment.pyc -- Gene Set Enrichment Analysis in Python

positional arguments:
  {gsea,prerank,gsva,enrichr}
    gsea                Main GSEapy Function: run GSEapy instead of GSEA on spatial h5ad file.
    prerank             Run GSEapy Prerank tool on deg genes.
    gsva                Run GSVA on gene expression.
    enrichr            Run Enrichr on deg genes.

options:
  -h, --help            show this help message and exit
  --version             show program's version number and exit

For command line options of each command, type: COMMAND -h
```


基因集富集分析：gsea

用途：使用GSEAPY中的gsea模块进行基因通路富集分析

运行方式：SDAS geneSetEnrichment gsea -i st.h5ad --gsea_plan gsea_plan.csv --gmt h.all.v2024.1.Hs.symbols.gmt -o outdir --permutation_type gene_set

输入参数说明

gsea参数	是否必须	默认值	描述
-i / --input	是		Stereo-seq h5ad, 要求有原始表达矩阵
--gsea_plan	是		csv格式的分析方案, 每个方案一行, 至少包含3列信息: obs对象,处理组,对照组
-o / --output	是		结果存放文件夹路径
--layer	否	None	指定h5ad中存有原始表达矩阵的layer层
--gene_symbol_key	否	real_gene_name	指定用于进行分析的基因名称所在列
--species	否	human	指定预先构建好的物种的数据库, 'human' 或者 'mouse', 默认 'human'; 当指定--gmt参数时, 该参数不起作用。
--gmt	否		gmt格式的数据库文件其中gene_name信息必须转为大写, 多个文件时用 ',' 隔开, 不提供时使用--species参数指定的物种数据库。
--graph	否	5	通路数量, 用得分最高的top通路进行画图, 默认' 5 '。用了--pathways参数时, 该参数不起作用
--pathways	否		通过txt文件指定1到多个感兴趣的通路进行画图, 这些通路必须在gmt数据库中, 不指定时使用--graph参数。
--permutation_type	否	gene_set	小于1000个样本时用 'gene_set', 大于1000个样本时用 'phenotype'

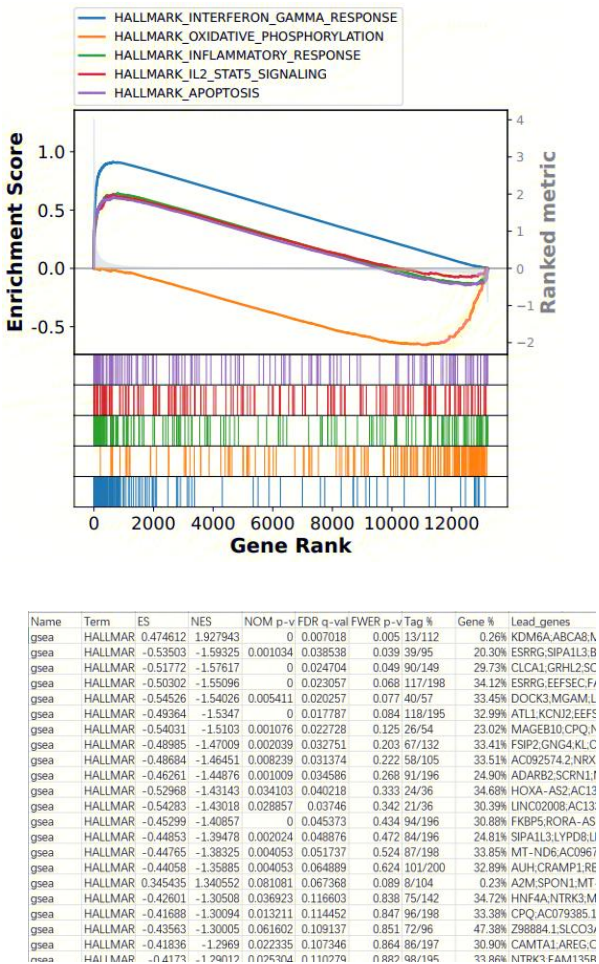
gsea_plan.csv示例:

#obs1,Treat_group,Control_group,obs2,obs2_info
stim,STIM,CTRL,seurat_annotations,CD14 Mono
stim,STIM,CTRL

输出结果说明

gsea结果文件	描述
GSEA.{database}.csv	csv格式的结果文件
GSEA.{database}.top5.pdf/png	pdf和png格式的图像文件

*输出结果展示:



基因集富集分析： gsva

用途：使用GSEAPY中的gsva模块进行基因通路富集分析

运行方式： SDAS geneSetEnrichment gsva -i st.h5ad --gsva_plan gsva_plan.csv -gmt h.all.v2024.1.Hs.symbols.gmt
-o outdir --kernel_cdf Poisson

输入参数说明

gsva参数	是否必须	默认值	描述
-i / --input	是		Stereo-seq h5ad, 要求有原始表达矩阵
--gsva_plan	是		csv格式的分析方案, 至少包含2列信息, 多个样本用 “;” 隔开: obs对象,样本1;样本2;样本3
-o / --output	是		结果存放文件夹路径
--layer	否	None	指定h5ad中存有原始表达矩阵的layer层
--gene_symbol_key	否	real_gene_name	指定用于进行分析的基因名称所在列
--species	否	human	指定预先构建好的物种的数据库, ‘human’ 或者 ‘mouse’, 默认 ‘human’ ; 当指定--gmt参数时, 该参数不起作用。
--gmt	否		gmt格式的数据库文件其中gene_name信息必须转为大写, 多个文件时用 ‘,’ 隔开, 不提供时使用--species参数指定的物种数据库。
--kernel_cdf	否	Poisson	当使用的表达值为原始表达量时选 ‘Poisson’, 其他选 ‘Gaussian’

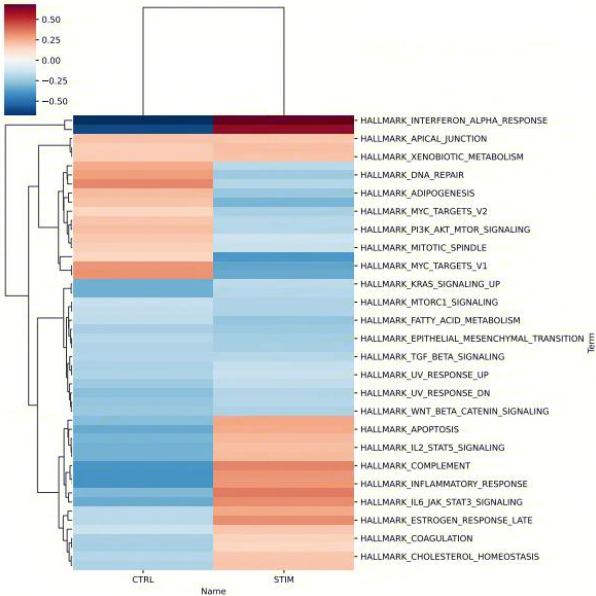
gsva_plan.csv示例：

#obs1,obs1_infos,obs2,obs2_info
stim,STIM;CTRL,seurat_annotations,CD14 Mono
stim,STIM

输出结果说明

gsva结果文件	描述
GSVA.{database}.csv	csv格式的结果文件
GSVA.{database}.pdf/png	pdf和png格式的图像文件

*输出结果展示：



Term	A03489G6	D03663C6
HALLMARK_ADIPOGENESIS	-0.585921595	-0.58010288
HALLMARK_ALLOGRAFT_REJECTION	-0.456441325	-0.44022777
HALLMARK_ANDROGEN_RESPONSE	-0.639092435	-0.671114335
HALLMARK_ANGIOGENESIS	-0.499586231	-0.539279722
HALLMARK_APICAL_JUNCTION	-0.526470667	-0.546878344
HALLMARK_APICAL_SURFACE	-0.477814236	-0.504528357
HALLMARK_APOPTOSIS	-0.565404478	-0.573781697
HALLMARK_BILE_ACID_METABOLISM	-0.466687203	-0.478519795
HALLMARK_CHOLESTEROL_HOMEOSTASIS	-0.59573537	-0.590166228
HALLMARK_COAGULATION	-0.391725864	-0.384757076
HALLMARK_COMPLEMENT	-0.51070563	-0.526563746
HALLMARK_DNA_REPAIR	-0.565980565	-0.561181619
HALLMARK_E2F_TARGETS	-0.596284547	-0.600694738
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	-0.556293578	-0.555338388
HALLMARK_ESTROGEN_RESPONSE_EARLY	-0.558211932	-0.582859387
HALLMARK_ESTROGEN_RESPONSE_LATE	-0.521162417	-0.546434554
HALLMARK_FATTY_ACID_METABOLISM	-0.551373342	-0.564614081
HALLMARK_G2M_CHECKPOINT	-0.593441974	-0.615553283
HALLMARK_GLYCOLYSIS	-0.544694403	-0.557160524
HALLMARK_HEDGEHOG_SIGNALING	-0.555757612	-0.606071982
HALLMARK_HEME_METABOLISM	-0.531717139	-0.527456344
HALLMARK_HYPOXIA	-0.550757298	-0.55392465
HALLMARK_IL2_STAT5_SIGNALING	-0.574251487	-0.579891378
HALLMARK_IL6_JAK_STAT3_SIGNALING	-0.521303183	-0.496991027

基因集富集分析：enrichr

用途：使用GSEAPY中的enrichr模块进行基因通路富集分析

运行方式：SDAS geneSetEnrichment enrichr -i A.vs.B.deg_filtered.csv --gmt h.all.v2024.1.Hs.symbols.gmt -o outdir

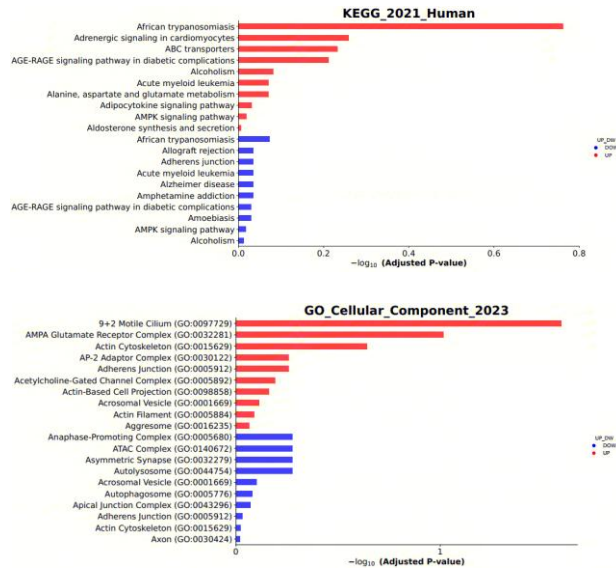
输入参数说明

enrichr参数	是否必须	默认值	描述
-i / --input	是		SDAS软件输出的显著差异分析结果文件，格式为csv。(注意all vs rest方案的差异基因会作为一个列表进行富集，建议用户自行处理，如分别提取每个cluster的列表)
-o / --output	是		结果存放文件夹路径
--species	否		指定预先构建好的物种的数据库，‘human’ 或者 ‘mouse’ ，默认 ‘human’ ，当指定--gmt参数时，该参数不起作用。
--gmt	否		gmt格式的数据库文件其中gene_name信息必须转为大写，多个文件时用 ‘，’ 隔开，不提供时使用--species参数指定的物种数据库。
--cut_off	否	1	富集结果作图时过滤的pvalue阈值，默认值为1，设太小的阈值时可能会由于没有显著富集结果导致无法作图。
--background	否		设定富集分析时使用的background，默认为所用数据库的gene数。
--top_term	否	10	指定最富集的top通路进行作图，默认为10。

输出结果说明

enrichr结果文件	描述
enrichment_{database}.UP.csv	上调基因的富集分析结果
enrichment_{database}.DOWN.csv	下调基因的富集分析结果
enrichment_{database}.pdf/png	上调和下调基因显著富集通路图

*输出结果展示：



Gene_set	Term	Overlap	P-value	Adjusted f	Odds Rati	Combined Genes
KEGG_202	ABC trans	3/45	0.118343	0.585341	3.016586	6.437902 ABCA3;ABCC9;ABCC9
KEGG_202	AGE-RAGI	5/100	0.129335	0.613144	2.115097	4.326118 SELE;IL6;COL4A4;IL1B
KEGG_202	AMPK sigi	3/120	0.627185	0.957882	1.080642	0.504134 PPARGC1A;PPP2R2C;
KEGG_202	Acute myt	3/67	0.266066	0.848717	1.982082	2.624298 SPI1;BCL2A1;FLT3
KEGG_202	Adipocyto	2/69	0.554799	0.93119	1.346024	0.793009 PPARGC1A;ACSBG1
KEGG_202	Adrenergic	7/150	0.107005	0.55031	1.925815	4.30397 CREB3L3;PIK3R5;CAC
KEGG_202	African try	4/37	0.016463	0.172861	4.949289	20.32497 IL1B;IL6;SELE;IDO1
KEGG_202	Alanine, a	2/37	0.260216	0.848717	2.569848	3.459642 GAD1;CPS1
KEGG_202	Alcoholism	7/186	0.230093	0.828335	1.532399	2.251509 CREB3L3;GRIN3A;SHC
KEGG_202	Aldosterol	2/98	0.742585	0.984483	0.938017	0.279171 CREB3L3;CACNA1I
KEGG_202	Allograft r	1/38	0.643822	0.971515	1.452529	0.639597 HLA-DRA
KEGG_202	Alzheimer	8/369	0.777429	0.985024	0.845966	0.212983 SNCA;MAPT;WNT11
KEGG_202	Amoebias	6/102	0.055257	0.386801	2.485188	7.196497 GNA14;LAMA2;IL6;CC
KEGG_202	Amphetar	5/69	0.036844	0.299503	3.144153	10.37908 CREB3L3;GRIN3A;GRI
KEGG_202	Amyotro	11/364	0.38226	0.93119	1.188797	1.14321 HAP1;NUP210L;NRG
KEGG_202	Antigen p	2/78	0.621844	0.957882	1.186297	0.563569 KLRD1;HLA-DRA
KEGG_202	Apelin sigi	5/137	0.304193	0.881111	1.517204	1.805614 PIK3R5;PPARGC1A;RY
KEGG_202	Apoptosis	1/142	0.979414	0.997924	0.37983	0.007901 BCL2A1
KEGG_202	Arachidoni	1/61	0.809777	0.989582	0.897682	0.189408 PLA2G4A
KEGG_202	Arginine a	1/50	0.743172	0.984483	1.098713	0.326129 DAO
KEGG_202	Arginine t	1月22日	0.449572	0.93119	2.538661	2.029555 CPS1
KEGG_202	Arrhythmic	4/77	0.150403	0.631693	2.244274	4.251633 CACNB4;ACTN2;CAC

基因集富集分析： prerank

用途：使用GSEAPY中的prerank模块进行基因通路富集分析

运行方式： SDAS geneSetEnrichment prerank -i A.vs.B.deg.csv --gmt h.all.v2024.1.Hs.symbols.gmt -o outdir

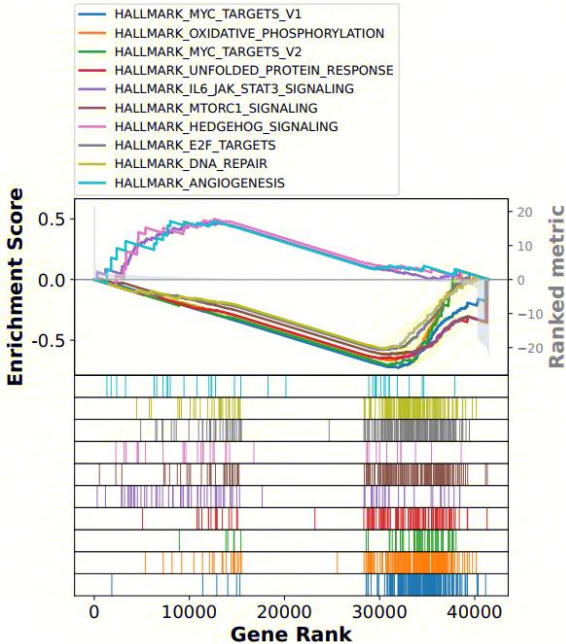
输入参数说明

prerank参数	是否必须	默认值	描述
-i / --input	是		SDAS软件输出的所有差异分析结果文件，格式为csv。(注意all vs rest方案的差异基因会作为一个列表进行富集，建议用户自行处理，如分别提取每个cluster的列表)
-o / --output	是		结果存放文件夹路径
--species	否		指定预先构建好的物种的数据库， 'human' 或者 'mouse' ，默认 'human' ，当指定--gmt参数时，该参数不起作用。
--gmt	否		Gmt格式的数据库文件其中gene_name信息必须转为大写，多个文件时用 '，' 隔开，不提供时使用--species参数指定的物种数据库。
--graph	否	5	通路数量，用得分最高的top通路进行画图，默认' 5 '。用了--pathways参数时，该参数不起作用。
--pathways	否		通过txt文件指定1到多个感兴趣的通路进行画图，这些通路必须在gmt数据库中，不指定时使用--graph参数。

输出结果说明

prerank结果文件	描述
prerank_{database}.csv	csv格式的结果文件
prerank_{database}.top10.pdf/png	pdf和png格式的图像文件

*输出结果展示：



Name	Term	ES	NES	NOM p-v	FDR q-val	FWER p-v	Tag %	Gene %	Lead_genes
prerank	Phospholi	0.485233	1.434222	0.018868	1	0.815	79/147	31.17%	GRM6,DGKK,CX
prerank	Ribosome	-0.74349	-1.418	0.007519	1	0.695	120/136	21.70%	RNA559,RPS18I
prerank	Cholinergi	0.503666	1.383584	0.014706	1	0.911	56/112	27.47%	KCNQ2,CAMK2
prerank	Olfactory	0.459646	1.379304	0.052632	1	0.918	42/267	16.05%	OR4D1,OR10R2
prerank	Serotoner	0.5158	1.348361	0.013699	1	0.956	50/112	25.41%	HTR4,CYP4X1,H
prerank	Circadian	0.520843	1.336239	0.085106	1	0.971	51/95	27.47%	MTNR1B,GRIA4
prerank	Proteasom	-0.7542	-1.29054	0.072306	1	1	32/45	19.82%	PSMC4,PSME2I
prerank	Spliceosor	-0.65647	-1.27222	0.079484	1	1	105/143	25.73%	RNVU1-1,PCBP
prerank	Fatty acid	-0.77883	-1.26461	0.109635	1	1	2月27日	1.81%	ACOT1L,THEM5
prerank	alpha-Lini	-0.80802	-1.26308	0.096429	1	1	3月25日	1.41%	PLA2G3,PLA2GJ
prerank	Arrhythm	0.49947	1.244309	0.144	1	0.999	34/77	26.10%	ACTN3,CACNGI
prerank	ECM-rece	0.480068	1.228121	0.134615	1	1	27/87	23.76%	IBSP,TNX8,COL
prerank	Salivary se	-0.65891	-1.22775	0.141563	1	1	9/87	5.54%	CST2,ADRB1,AD
prerank	Linoleic ac	-0.76817	-1.22733	0.147011	1	1	3月29日	1.41%	PLA2G3,PLA2GJ
prerank	Starch anx	-0.76264	-1.22526	0.146084	1	1	2/32	1.38%	GY52,MGMAM2
prerank	Carbohydri	-0.73774	-1.22324	0.136691	1	1	4/42	3.27%	SLC37A4,MGMAM
prerank	Allograft r	-0.74605	-1.22308	0.143079	1	1	5/33	4.99%	IL12A,HLA-DOA
prerank	Antigen p	-0.66785	-1.21757	0.186667	1	1	13/65	9.29%	PSME2,HLA-DC
prerank	Maturity c	-0.77366	-1.21424	0.155979	1	1	2月26日	2.14%	MAFA,NEUROG
prerank	Arachidon	-0.68932	-1.20979	0.165432	1	1	3/59	1.41%	PLA2G3,PLA2GJ
prerank	Ether lipid	-0.70125	-1.20473	0.177181	1	1	3/47	1.41%	PLA2G3,PLA2GJ
prerank	One carbx	-0.77263	-1.19894	0.23475	1	1	1月20日	0.99%	FTCD
prerank	Nucleoti	-0.69018	-1.18247	0.218878	1	1	26/46	22.64%	GTF2H4,RPA4,R
prerank	Fat digest	-0.69203	-1.17106	0.252454	1	1	3/43	1.41%	PLA2G3,PLA2GJ

用途：使用cellchat v2进行细胞通讯分析

运行方式：SDAS CCI cellchat -i st.rds --bin_size 100 --label_key anno_spotlight --add_spatial -o outdir

输入参数说明：

cellchat参数	是否必须	默认值	描述
-i / --input	是		rds文件，要求有原始表达矩阵
-o / --output	是		输出文件夹
--label_key	是		细胞类群的列名称
--assay	否	None	rds文件中用于分析的assay的名称，若不设置，则使用默认的assay
--gene_symbol_key	否	real_gene_name	rds文件中表示基因名(symbol)的列(meta.features)的名称（_index表示使用矩阵的列名）
--run_mode	否	stRNA	软件运行模式，选择以空间转录组模式（stRNA）或是单细胞模式（scRNA）运行
--add_spatial	否		当--run_mode为stRNA时，如果计算时需考虑空间信息，则使用“--add_spatial”
--bin_size	否		如果使用“--add_spatial”，则需要提供bin_size, 如cellbin,10, 20, 50,100等
--species	否	human	cellchat内置的数据库，human或mouse，默认‘human’；当指定--database参数时，该参数不起作用。
--database	否		用户自定义的数据库
--method	否	triMean	computeCommunProb的计算方法，triMean或truncatedMean
--trim	否	0.1	当method为truncatedMean时，可调整trim，trim越小找到的交互越多，0.1表示截断上下各 10% 的数据
--scale_distance	否	2	用于计算空间距离的缩放因子
--pathway_file	否		用户自定义的pathway列表路径，用于生成<input_name>_pathway_communication_network.png/pdf
--seed	否	42	指定随机种子

pathway_file示例
./pathway.file（无表头）

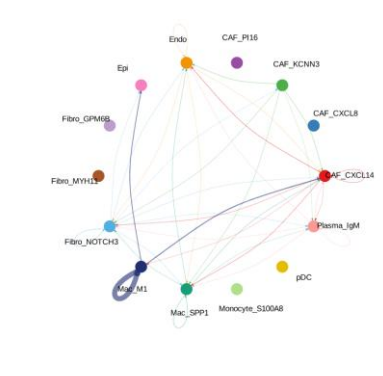
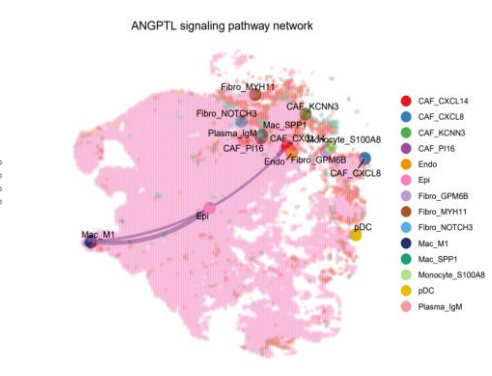
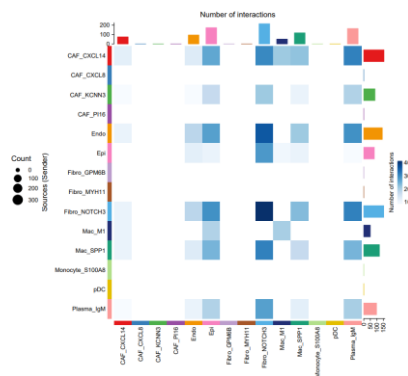
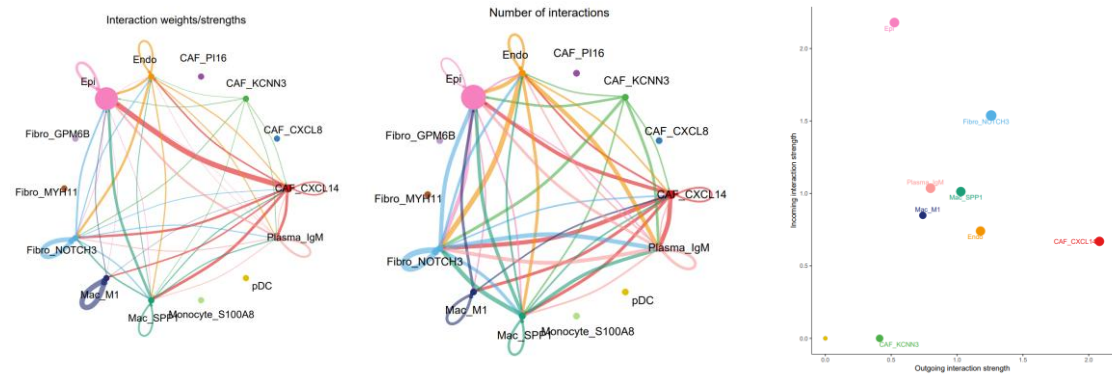
```
$ head pathway.file
COLLAGEN
FN1
LAMININ
JAM
```

***method参数建议：**预测较强的细胞间通信，推荐使用默认的 triMean 方法。如果希望获得更多的交互，可以尝试使用 truncatedMean 方法，并根据需要调整 trim 参数，trim设置越小，找到的交互越多

输出结果说明

结果文件	描述
<input_name>_cellchat_LR.csv	互作的配受体结果(没有找到显著配受体时不输出)
<input_name>_cellchat_LR_pathway.csv	配受体富集的通路结果(没有找到显著配受体时不输出)
<input_name>_interaction_strength_circle.png/pdf <input_name>_interaction_strength_heatmap.png/pdf	细胞互作强度的circle图和heatmap图，同时输出pdf和png
<input_name>_number_of_interactions_circle.png/pdf <input_name>_number_of_interactions_heatmap.png/pdf	细胞互作数目的circle图和heatmap图，同时输出pdf和png
<input_name>_signalingRole_scatter.png/pdf	细胞互作点图，同时输出pdf和png
<input_name>_pathway_communication_network.png/pdf	指定通路的细胞网络图，如果加空间信息分析，输出的是空间网络图，同时输出pdf和png（如果指定通路没有显著富集则不输出）
<input_name>_cellchat.rds	包含细胞互作结果的rds文件

*输出结果展示：



sample_cellchat_LR.csv

source	target	ligand	receptor	prob	pval	interaction_name	interaction_name 2	pathway_name	annotation	evidence
<fct>	<fct>	<chr>	<chr>	<dbl>	<dbl>	<fct>	<chr>	<chr>	<chr>	<chr>
MP	AT1	Spp1	Cd44	0.2031065	0	SPP1_CD44	Spp1 - Cd44	SPP1	Secreted Signaling	PMID: 21907283
MP	AT2	Spp1	Cd44	0.2017045	0	SPP1_CD44	Spp1 - Cd44	SPP1	Secreted Signaling	PMID: 21907283
MP	B	Spp1	Cd44	0.1982072	0	SPP1_CD44	Spp1 - Cd44	SPP1	Secreted Signaling	PMID: 21907283
MP	Club1	Spp1	Cd44	0.1875791	0	SPP1_CD44	Spp1 - Cd44	SPP1	Secreted Signaling	PMID: 21907283
MP	Club2	Spp1	Cd44	0.1754759	0	SPP1_CD44	Spp1 - Cd44	SPP1	Secreted Signaling	PMID: 21907283

sample_cellchat_LR_pathway.csv

source	target	pathway_name	prob	pval
<chr>	<chr>	<chr>	<dbl>	<dbl>
AT1_AT2	AT1_AT2	PTPRM	0.038414675	0.000
Deuterosomal	IM	NRXN	0.015143084	0.000
IM	IM	NRXN	0.018496709	0.000
Mesothelial	IM	NRXN	0.008612413	0.010
Mesothelial	Mesothelial	NCAM	0.026620328	0.005
VEC	IM	NRXN	0.055055605	0.000

轨迹分析：monocle3

用途: 使用monocle3进行细胞轨迹分析

运行方式: SDAS trajectory monocle3 -i st.rds -o outdir --root_key annotation --root CAF_DES

输入参数说明:

monocle3参数	是否必须	默认值	描述
-i / --input	是		rds文件, 要求有原始表达矩阵
-o / --output	是		输出文件夹
--root_key	是		meta.data中根节点所在的列名称
--root	是		设置为根节点的名称
--assay	否	None	rds文件中用于分析的assay的名称, 若不设置, 则使用默认的assay
--gene_symbol_key	否	real_gene_name	rds文件中表示基因名(symbol)的列(meta.features)的名称 (_index 表示使用矩阵的列名)
--run_mode	否	stRNA	软件运行模式, 选择以空间转录组模式 (stRNA) 或是单细胞模式 (scRNA) 运行
--batch_key	否		进行批次校正的meta.data的列名称, 不输入则不做去批次
--resolution	否	NULL	Leiden聚类时的resolution参数; 默认为NULL, 函数会自行决定参数
--use_existing_umap_cluster	否	False	使用输入rds的umap和cluster信息
--umap_key	否	umap	输入rds的存储umap信息的名称
--cluster_key	否	leiden	输入rds的存储cluster信息的名称
--deg	否	False	要分析随拟时序变化的差异基因则输入" --deg"
--n_cpus	否	8	多线程的线程数, 默认8线程
--top_gene_num	否	5	输出随时间变化的top数目的差异基因的表达图
--gene_file	否		输出指定gene的随时间变化的表达图
--gene_color_label	否	pseudotime	基因图展示的列的名称
--pval_cutoff	否	0.05	差异基因筛选p值
--qval_cutoff	否	0.05	差异基因筛选q值
--seed	否	42	指定随机种子

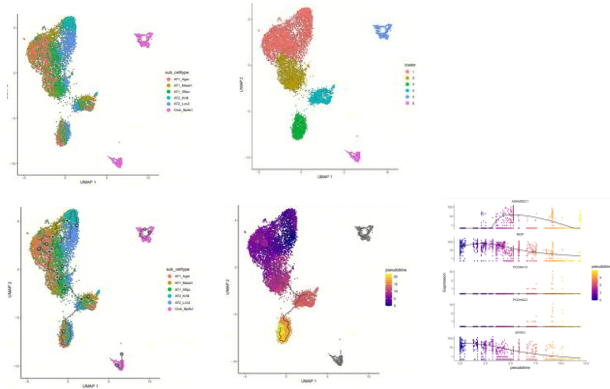
输出结果说明

结果文件	描述
<input_name>_dimension.png/pdf	降维图
<input_name>_dimension_color_by_batch.png/pdf	以批次信息展示降维图 (有做批次校正时输出)
<input_name>_cluster.png/pdf	聚类图
<input_name>_roots.png/pdf	root图
<input_name>_pseudotime.png/pdf	拟时序图
<input_name>_top_genes_in_pseudotime.png/pdf	随时间变化的top gene表达图
<input_name>_custom_genes_in_pseudotime.png/pdf	随时间变化的指定gene表达图
<input_name>_monocle3.rds	包含轨迹分析结果的rds文件
<input_name>_pseudotime.csv	细胞对应的拟时序结果
<input_name>_deg_trajectory.xls	随拟时序变化的差异基因结果

*输出结果展示:

gene_file示例
./genelist (无表头)

```
$ head genelist
ABLIM1
FAU
CLMN
RBBP8
HCG14
```



五、各功能模块性能测试

对功能模块进行性能测试，包括运行时间和使用内存峰值。

- 测试环境：研发内部集群（实际生产环境），CPU节点计算资源为112个CPU、1T内存，GPU节点为112个CPU, 2T内存，16G显存。
- 测试数据：研发内部数据，包括小鼠样本和病理样本，Stereo-seq V1.3和FFPE各5例，每个模块测试使用样本为10例中的子集。

Stereo-Seq FFPE数据				Stereo-Seq V1.3数据			
	Bin20	Bin50	Bin100	Cellbin	Bin20	Bin50	Bin100
Bins数	50w-70w	8w-12w	2w-3w	30w-40w	50w-70w	8w-12w	2w-3w
基因数	100-200	600-1k	2k-3k	300~1k	400-900	2k-4k	5k+

描述：

- 时间： bin50/bin100时三个算法都在10h以内， bin20时， cell2location-GPU时间显著增加。
- 内存： RCTD的内存在bin20/50/100， 都比较一致30-40G。 Spotlight， cell2location随着bin数增加， 内存会显著增加， 300G+。

性能表现（内存和时间）：

Stereo-Seq FFPE数据					Stereo-Seq V1.3数据			
算法	性能(平均)	Bin20	Bin50	Bin100	Cellbin	Bin20	Bin50	Bin100
	bins数	50w-70w	8w-12w	2w-3w	30w-40w	50w-70w	8w-12w	2w-3w
	基因数	100-200	600-1k	2k-3k	300~1k	400-900	2k-4k	5k+
Cell2location-GPU模式	运行时间	43.5h	8.2h	2.7h	14.3h	35.5h	10.8h	2.0h
	内存	377.1G	64.8G	21.2G	95.9G	426.5G	72.8G	23.5G
	显存	10G	8.8G	8.8G	9.5G	10.1G	9G	8.6G
Spotlight	运行时间	4.7h	1.7h	0.83h	11.0h	14.2h	8.9h	6.5h
	内存	490.1G	87.0G	33.8G	261.3G	433G	79.4G	27.3G
RCTD	运行时间	7.4h	1.0h	0.48h	10.9h	24.2h	2.9h	0.9h
	内存	46.8G	38.9G	37.3G	37.2G	48.0G	33.1G	30.9G

描述：

- CPU模式： GraphST算法支持并行运算，但无法通过参数控制线程数，该性能测试是在测试节点计算资源112个CPU、1T内存的条件下进行的测试。运行时间和内存与细胞数量呈现正相关
- GPU模式：测试服务器显存为16G，若cell数量不超过4.9万个时，能够正常完成测试，且运行时间与细胞数量呈现正相关。
- Cellbin/Bin20时，CPU或者GPU模式下均不能完成测试，因需求内存过大，导致程序运行失败。

性能表现（CPU模式）（内存和时间）：

	Stereo-Seq FFPE数据				Stereo-Seq V1.3数据			
算法	性能(平均)	Bin20	Bin50	Bin100	Cellbin	Bin20	Bin50	Bin100
	bins数	50w-70w	8w-12w	2w-3w	30w-40w	50w-70w	8w-12w	2w-3w
	基因数	100-200	600-1k	2k-3k	300~1k	400-900	2k-4k	5k+
GraphST	运行时间	/	4h	1h	/	/	2h	1h
	内存	/	213G	15G	/	/	206G	17G

性能表现（GPU模式）（内存和时间）

	Stereo-Seq FFPE数据				Stereo-Seq V1.3数据			
算法	性能(平均)	Bin20	Bin50 (cell <= 4.9w)	Bin100	Cellbin	Bin20	Bin50 (cell <= 4.9w)	Bin100
GraphST	运行时间	/	12m	6m	/	/	11m	6m
	内存	/	63G	24G	/	/	56G	23G
	显存	/	11275M	6160M	/	/	11392M	5516M

备注：GPU模式下，16G显存时，细胞数量大于4.6w且小于等于4.9w时，可以通过设置export PYTORCH_CUDA_ALLOC_CONF=max_split_size_mb:512能够跑通，小于4.6w时，无需设置即能跑通

描述：

- NeSt/hotspot算法使用n_cpus控制并行线程数，该性能测试使用n_cpus=8。
- 运行时间与细胞数和基因数有关，细胞数越多、基因数越多所需时间越长，内存越大。
- 时间：hotspot > nest
- 内存：nest > hotspot

性能表现（内存和时间）：

Stereo-Seq FFPE数据					Stereo-Seq V1.3数据			
算法	性能（平均）	Bin20	Bin50	Bin100	Cellbin	Bin20	Bin50	Bin100
	bins数	50w-70w	8w-12w	2w-3w	30w-40w	50w-70w	8w-12w	2w-3w
	基因数	100-200	600-1k	2k-3k	300~1k	400-900	2k-4k	5k+
NeST	运行时间	4h	2h	1h	3.5h	6h	3h	3h
	内存	33G	20G	16G	57G	117G	68G	64G
Hotspot	运行时间	106h	32.5h	23.5h	94h	133h	35h	27h
	内存	34G	10G	6G	22G	38G	14G	9G

描述:

- 运行时间和内存与细胞数和细胞类群个数有关，细胞数和细胞类群个数越多所需时间越长，内存越大

性能表现（内存和时间）：

样本	细胞数目	细胞类群(个)	运行时间(h)	内存(G)
sample1_bin100	20629	34	0.3	13.2
sample2_bin100	21500	11	0.3	16.4
sample1_bin50	81843	34	0.6	38.9
sample2_bin50	84510	12	0.5	48.5
sample2_bin50_2	84510	25	0.9	49
sample1_cellbin	321856	35	1.6	137.1
sample2_cellbin	369313	12	2	218.4
sample2_cellbin_2	369313	17	2.7	173.9
sample1_bin20	508859	35	3.9	214.6
sample2_bin20	522051	12	3.6	240.8
sample2_bin20_2	522051	18	5.1	242.1

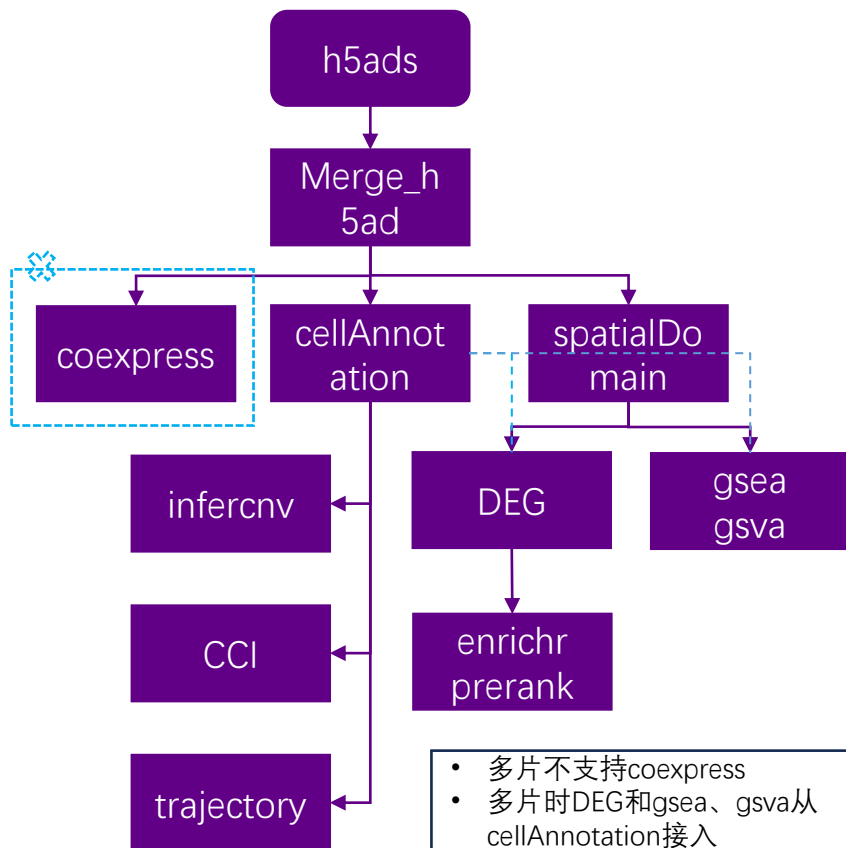
描述：

- monocle3使用n_cpus控制并行进程数，每个进程固定1个线程，该性能测试使用n_cpus=8，即8*1=8线程。
- 运行时间和内存与细胞数和细胞类群个数有关，细胞数和细胞类群个数越多所需时间越长，内存越大

性能表现（内存和时间）：

样本	细胞数目	细胞类群(个)	运行时间(h)	内存(G)
sample1_bin100	20629	34	0.4	8.5
sample2_bin100	21500	11	0.8	11.1
sample1_bin50	81843	34	0.9	13.3
sample2_bin50	84510	12	1.6	24.1
sample2_bin50_2	84510	25	2.2	21.8
sample1_cellbin	321856	35	3.6	208.1
sample2_cellbin	369313	12	4.8	252.6
sample2_cellbin_2	369313	17	5.6	236.1
sample2_bin20	522051	12	12.1	579.2

六、流程示例



Step1.按需求修改conf文件

```
#####
#          SDAS pipeline          #
#####

SDAS_software = /path/SDAS_beta_v2504/SDAS

##### input h5ad #####
## input h5ad files, use "," to split more than 1 files, example:
# h5ad_files = A.h5ad
# h5ad_files = S1.group1,A.h5ad;S2.group1,B.h5ad;S3.group2,C.h5ad
# or
# h5ad_files = S1,,A.h5ad;S2,,B.h5ad
# notice:

h5ad_files = /path/input.h5ad

## analysis process, chose 1 or more process, example:
# process = coexpress,spatialDomain,cellAnnotation,infercnv,CCI,trajectory
# notice: 1. infercnv or CCI or trajectory depend on cellAnnotation result
#         2. DEG and gene enrichment analysis depend on spatialDomain and input h5ad

process = coexpress,spatialDomain,cellAnnotation,infercnv,CCI,trajectory

##### coexpress #####
```

Step2.运行SDAS_pipeline.py

usage: python3 SDAS_pipeline.py -c pipeline_input.conf -o /

positional arguments:
{} Commands help.

options:
-h, --help show this help message and exit
-c CONF, --conf CONF input conf file
-o OUTDIR, --outdir OUTDIR output directory

Step3.按照all_shell.conf文件运行任务

```
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/input_file/shell/merge_h5ads.sh:cpu:3:mem:10G /storeData/
USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/coexpress_hotspot.sh:cpu:8:mem:30G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/merge_h5ads.sh:cpu:3:mem:10G /storeData/
USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/cellAnnotation/shell/cellAnnotation_rcd.sh:cpu:4:mem:20G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/input_file/shell/merge_h5ads.sh:cpu:3:mem:10G /storeData/
USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/cellAnnotation/shell/cellAnnotation_rcd.sh:cpu:3:mem:20G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/infercnv/shell/infercnv.sh:cpu:3:mem:20G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/cellAnnotation/shell/cellAnnotation_rcd.sh:cpu:4:mem:20G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/cellAnnotation/shell/cellAnnotation_rcd.sh:cpu:4:mem:20G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/cellAnnotation/shell/cellAnnotation_rcd.sh:cpu:4:mem:20G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/cellAnnotation/shell/cellAnnotation_rcd.sh:cpu:4:mem:20G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/spatialDomain/shell/spatialDomain_graphst.sh:cpu:3:mem:20G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/deg/shell/deg.sh:cpu:1:mem:2G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/deg/shell/deg.sh:cpu:1:mem:2G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/geneEnrichment/shell/deg_enrich.sh:cpu:1:mem:2G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/spatialDomain/shell/spatialDomain_graphst.sh:cpu:3:mem:20G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/geneEnrichment/shell/gsea.sh:cpu:1:mem:2G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/spatialDomain/shell/spatialDomain_graphst.sh:cpu:3:mem:20G
/storeData/USER/data/02.Bioinformatics_for_STOmics/01.user/xiehongqing/analysis_work/work_ST_DEG_and_Enrichment/test_pipeline/single_slice/geneEnrichment/shell/gsva.sh:cpu:1:mem:2G
```

测试数据: ./Test_data/single_slice 和 ./Test_data/multiple_slices

脚本: ./Scripts/pipelines

七、常见问题 (FAQ)

1. 输入文件的格式是什么样的？

答：SDAS支持SAW count, SAW gef2h5ad (\geq SAW V8), SAW aggr(alpha ver.)的输出文件；同时支持用户按照使用手册第7页 [\(SDAS输入数据格式描述\)](#) 自行准备的文件。

2. 细胞注释模块有3种算法，均需要对应的scRNA-Seq数据，SDAS是否提供，如果不提供，用户应该如何准备？

答：SDAS软件暂时不提供细胞注释的单细胞数据，用户需要按照使用手册中的格式说明来准备h5ad文件。

3. SDAS中的各模块支持scRNA-Seq数据分析吗，如果支持，用户应该如何准备单细胞数据？

答：SDAS中的dataProcess, DEG, geneSetEnrichment, inferCNV, CCI, trajectory模块均支持scRNA-Seq数据分析，用户按照使用手册准备输入文件（h5ad或rds）即可。

4. SDAS各模块输出结果，进行可视化作图，是否有支持方案？

答：大部分模块环境中（除spatialDomain, DEG, geneSetEnrichment外）安装jupyter kernel (ipykernel/IRkernel) ，支持用户直接调用该环境使用Jupyter Notebook等交互式环境画图。

1. 日志：SDAS将标准输出 (stdout) 与错误信息 (stderr) 进行分离，方便用户快速定位问题

- 标准输出分类：打印正常结果，程序运行状态（如进度条，成功消息等）
- 错误信息分类：错误信息，警告信息，异常信息

```
example_bash_script > stdout.log 2> stderr.log
```

- stdout.log文件将包含正常输出
- stderr.log文件将包含错误信息

2. 退出码：退出码通常是由程序在运行结束时返回给操作系统的状态码，用于表示程序的运行结果是否成功。通常返回值 0 表示程序正常退出，非 0 表示程序因错误退出，方便后续在脚本中(如WDL、snakemake等)通过退出码来判断程序完成状态。SDAS退出码定义如下：

- 正常退出：0
- 核心算法main function错误：1
- 命令行参数错误：2
- 输入文件格式或者内容错误：3

附录 SAW aggr(alpha ver.)流程介绍

1. **SAW aggr是什么**：包含数据整合、去批次、聚类、差异基因的多片分析和可视化流程
2. **如何与SDAS进行衔接**：SAW aggr生成的h5mu文件，通过SDAS dataProcess h5mu2h5ad即可转成SDAS的标准输入格式
3. **如何获取SAW aggr (alpha ver.)**：包括软件和使用手册，测试数据和测试结果报告

华大网盘 <https://bgipan.genomics.cn/#/link/vfaGuYJLpxgrpGzZ7fdn> 提取密码：YUM0

SDAS beta版本材料包的内容:

1. SDAS_beta.2.tar.gz
2. Documents: SDAS_beta.2_user_manual.pdf
3. Test_data: single_slice (单片)、multiple_slices (多片)
4. Scripts: quick_start (单模块运行脚本), pipelines (串流程脚本)
5. Application_cases:
 - NC2024_paper_bioinformatics_analysis_withSDAS_beta.2.pdf
 - Application_test_data: 包括单细胞和空转数据
 - Application_scripts: 包括R, python代码和notebook

Test data来源:

1. 单片和应用案例: Feng, Y., Ma, W., Zang, Y. et al. Spatially organized tumor-stroma boundary determines the efficacy of immunotherapy in colorectal cancer patients. *Nat Commun* 15, 10259 (2024). <https://doi.org/10.1038/s41467-024-54710-3>
2. 多片: Zhang R, Feng Y, Ma W, et al. Spatial transcriptome unveils a discontinuous inflammatory pattern in proficient mismatch repair colorectal adenocarcinoma. *Fundam Res.* 2022;3(4):640-646 (2022). <https://doi.org/10.1016/j.fmre.2022.01.036>