

生物信息编程基础 其他部分综合终测

题目卷 | 最后一次考前模拟

覆盖范围	Extended 1-3: PyPI/Git/GitHub、Scanpy 单细胞 RNA-seq、Spatial Transcriptomics / Squidpy
命题方式	参考已提供的 S1-S3 综合终测与历年/模拟题风格：概念简答、选择/填空、输出追踪、代码填空、Debug、综合问答。
总分 / 建议用时	130 分 / 120-140 分钟
使用建议	按真实考试计时完成；不会完整手写代码时，写清伪代码或编程思路。答案文件单独使用。

说明：所有题目均重新设计，不照搬讲义例题；重点覆盖老师明确提示的基础概念、函数/类、代码 Debug/填空，以及单细胞相关算法原理。

最后冲刺优先级

优先级	内容	复习策略
最高	Scanpy 流程与算法原理、AnnData	必须能从 QC 一路讲到 annotation，并解释每步为什么
高	空间转录组、Squidpy neighborhood/ligrec	分清转录相似图与空间邻居图，理解置换检验
中高	Git/PyPI	掌握核心术语、命令方向与包发布流程

第一部分 PyPI、Python 包与发布（15 分）

重点：pip 常用命令、package/module 区别、setup.py、__init__.py、本地安装、构建与上传。

1. PyPI 与 pip 概念（5 分）

简答：

- （1）PyPI 是什么？
- （2）pip 的作用是什么？
- （3）分别写出安装、卸载、列出已安装包、查看某包信息的命令。

2. Python 包结构填空（5 分）

一个最小可导入的课程示例包结构如下，填空并说明关键文件作用。

```
MyPackage/  
├── MyPackage/  
│   ├──   
│   └── tool.py  
└──   

```

3. 本地安装与发布流程（5 分）

把下列步骤按合理顺序排列，并写出相应关键命令：

- A. 上传 dist 中的发行包

- B. 在项目目录本地安装测试
 - C. 生成 source distribution
 - D. 编写 setup.py / 包结构
-
-

第二部分 Git / GitHub 与版本控制（20 分）

重点：version control 的意义；commit/repository/branch；clone/pull/push；staging；README；好 commit message。

4. 核心概念简答（6 分）

分别解释 version control、repository、commit、branch、clone、pull、push。

5. 提交工作流代码填空（6 分）

你修改了 analysis.py，准备提交。补全典型工作流。

```
git ____ analysis.py
git ____ -m "Update analysis workflow"
git ____
```

6. 协作情境判断（4 分）

判断应使用 clone、pull 还是 push：

- （1）第一次把 GitHub 上整个项目复制到本机。
 - （2）队友刚上传了两个新 commit，你本地要同步。
 - （3）你本地完成提交，要更新 GitHub 上的远程仓库。
-
-

7. README 与 commit message（4 分）

简答：README.md 为什么重要？一个好的 commit message 至少应满足哪两点？

第三部分 单细胞 RNA-seq 基础与 AnnData（20 分）

重点：Bulk vs scRNA-seq、SMART/液滴式、barcode/UMI、稀疏/高维/技术噪声、表达矩阵与 metadata、AnnData 结构。

8. Bulk RNA-seq 与 scRNA-seq 比较（5 分）

从输入、分辨率、输出、异质性、稀有细胞检测五方面比较 Bulk RNA-seq 与 scRNA-seq。

9. 测序技术与 barcode/UMI（5 分）

简答：

- （1）SMART-based scRNA-seq 的课程特点至少写两点。
 - （2）droplet-based scRNA-seq 中 cell barcode 与 UMI 分别标记什么？
-
-

10. scRNA-seq 数据三大特征（4 分）

解释 sparsity、high dimensionality、technical noise，并各举一个课程例子。

11. AnnData 结构匹配（6 分）

把对象与内容匹配：.X、.obs、.var、.obsm、.uns、.layers、.obsp。

可选内容：表达矩阵；细胞/spot 元数据；基因元数据；PCA/UMAP/空间坐标；非结构化分析信息；额外表达矩阵版本；细胞/spot 两两矩阵。

第四部分 Scanpy 标准分析流程与算法原理（35 分）

老师明确提示要熟悉单细胞相关算法原理。本部分是最后冲刺最高优先级：QC→标准化→HVG→PCA→neighbors→UMAP→Leiden→marker/annotation。

12. 流程排序（7 分）

把下列步骤按常见 Scanpy 分析顺序排序，并写出每一步的目的：

- A. Leiden clustering
 - B. QC/filter
 - C. UMAP
 - D. normalize_total + log1p
 - E. highly variable genes
 - F. PCA
 - G. neighbors
-
-

13. QC 指标与代码（7 分）

回答：

- （1）低质量细胞常表现出哪三类 QC 异常？
 - （2）n_genes_by_counts、total_counts、pct_counts_mt 分别表示什么？
 - （3）写出识别线粒体基因并计算其比例的关键代码思路。
-
-

14. 标准化与 HVG（5 分）

解释 normalize_total、log1p、highly_variable_genes 三步各解决什么问题，并写出对应的典型函数。

15. PCA、t-SNE、UMAP 原理比较（6 分）

简答：

- （1）PCA 是线性还是非线性降维？核心目标是什么？课程中提到它可通过什么分解实现？
 - （2）UMAP 为什么适合单细胞可视化？
 - （3）课程如何描述 t-SNE 的局部/全局距离性质？
-
-

16. Neighbor graph 与 Leiden (6 分)

回答:

- (1) `sc.pp.neighbors` 的核心产物是什么? `n_neighbors` 与 `n_pcs` 分别控制什么?
- (2) Leiden/Louvain 在什么对象上做聚类?
- (3) Leiden 的 `resolution` 参数增大通常会怎样?

17. 细胞类型注释 (4 分)

简答 marker-based annotation 与 automatic annotation 的基本思路, 并各写一个课程示例工具/函数。

第五部分 空间转录组与 Squidpy (25 分)

重点: 空间表达 + 组织位置; Visium/成像型概念; spatial 坐标与 `scalefactor`; 空间邻居图、`nhood enrichment`; `ligand-receptor permutation test`。

18. 空间转录组核心概念 (5 分)

比较 Bulk RNA-seq、scRNA-seq、Spatial Transcriptomics 的主要分辨率和信息差异; 说明空间转录组相比 scRNA-seq 多保留了什么关键维度。

19. 空间 AnnData (5 分)

回答:

- (1) 空间坐标通常存在哪?
- (2) 组织图像与 `scalefactors` 通常存在哪?
- (3) `spatial_connectivities` / `spatial_distances` 通常存在哪?
- (4) 为什么 Visium 坐标可能需要 `scalefactor` 才能与组织图对齐?

20. Visium/空间分析流程代码理解 (5 分)

把代码对应到“QC、标准化、HVG、转录相似性降维/聚类、空间可视化”五个阶段。

```
sc.pp.filter_cells(adata, min_counts=5000)
sc.pp.normalize_total(adata)
sc.pp.log1p(adata)
sc.pp.highly_variable_genes(adata, n_top_genes=2000)
sc.pp.pca(adata)
sc.pp.neighbors(adata)
sc.tl.umap(adata)
sc.tl.leiden(adata, key_added='clusters')
sc.pl.spatial(adata, img_key='hires', color='clusters')
```

21. Neighborhood enrichment (4 分)

说明 Squidpy 的 `spatial_neighbors` 与 `nhood_enrichment` 各做什么; `nhood_enrichment` 的核心问题是什么?

22. Ligand-receptor 分析（6 分）

回答：

- （1）课程中 Squidpy ligrec/CellPhoneDB 核心显著性算法是什么？
 - （2）n_perms 的含义？
 - （3）means 与 pvalues 两个结果矩阵分别代表什么？
 - （4）OmniPath 在流程中提供什么？
-
-

第六部分 算法问答与 Debug（15 分）

历年风格会给一个单细胞问题，让你讲“背景→现有局限→改进思路”；不会写代码时，把算法流程和验证指标写清楚即可。

23. 算法设计问答（三选一）（10 分）

任选其一作答，按“问题背景→为什么难→算法流程→如何验证”四段写：

- A. Batch effect: 不同批次按批次而不是细胞类型分开，如何校正同时保留生物信号？
 - B. Sparsity/dropout: 如何区分真实 0 与技术性 dropout，并合理插补？
 - C. Trajectory inference: 如何从静态单细胞快照推断分化顺序与分支？
-
-

24. Scanpy Workflow Debug（5 分）

下列流程有明显逻辑问题。指出至少 3 点并给出合理顺序。

```
sc.tl.umap(adata)
sc.tl.leiden(adata)
sc.pp.normalize_total(adata)
sc.pp.filter_cells(adata, min_genes=200)
sc.pp.neighbors(adata)
sc.tl.pca(adata)
```

交卷前自检

- PyPI 是平台，pip 是包管理；setup.py、__init__.py、pip install .、sdist、twine upload。
- commit/repo/branch/clone/pull/push/staging 各是什么意思，能否按协作场景判断。
- Bulk vs scRNA vs spatial: 分辨率、异质性、空间信息。
- cell barcode 标细胞，UMI 标分子；scRNA 的 sparsity/high dimensionality/technical noise。
- AnnData: X/obs/var/obsm/uns/layers/obsp 一次背清。
- Scanpy 主链: QC→normalize/log1p→HVG→PCA→neighbors→UMAP→Leiden→annotation。
- PCA 是线性降维；UMAP 非线性可视化；neighbors 构建细胞相似图；Leiden 是图社区聚类。
- 空间分析中转录相似图 ≠ 空间邻居图；obsm['spatial'] 与 uns['spatial'] 分工。
- Squidpy: spatial_neighbors→nhood_enrichment；ligrec 用 permutation test，mean 与 p-value 要分开看。
- 算法问答按“背景→难点→流程→验证”写，伪代码/流程图也能拿分。