

SAMap 简介与概述

SAMap (Self-Assembling Manifold Alignment for Projection) 是一个用于跨物种单细胞转录组数据整合的工具，主要通过构建同源基因图谱和嵌入空间来映射不同物种的单细胞数据。它特别适用于跨物种比较，包括植物物种间的整合。该工具基于Python实现，通常与Scanpy等库结合使用。

根据提供的Jupyter Notebook (SAMap_vignette.ipynb) 的分析，SAMap在进行植物跨物种单细胞数据整合时，对输入文件有严格要求。这些要求旨在确保数据的一致性、同源关系的准确性和计算效率。植物数据整合的特殊性在于植物基因组往往涉及全基因组重复 (Whole Genome Duplication, WGD) 、基因家族扩张 (如MADS-box、MYB家族) 和物种特异性注释，因此输入文件需考虑这些因素。下面我将详细分析输入文件的要求，包括文件类型、格式、预处理步骤、植物特异性考虑，以及SAMap函数的相关参数。

1. 所需输入文件类型

SAMap的主要输入包括三个核心文件类型：计数矩阵 (expression data) 、基因注释 (gene annotations) 和同源/映射表 (homology/mapping table) 。这些文件需针对每个物种准备，并在整合时指定物种代码 (通常为两个字符，如'ar'代表Arabidopsis) 。

文件类型	描述	必须包含的内容	植物特异性考虑
计数矩阵 (Count Matrix)	单细胞表达数据，通常为原始UMI计数或归一化计数。	<ul style="list-style-type: none"> - 细胞作为列（或行，在TSV中）。 - 基因作为行（或列，在TSV中）。 - 稀疏矩阵格式（如 <code>scipy.sparse.csr_matrix</code>）。 	<ul style="list-style-type: none"> - 使用稳定基因标识符（如 Ensembl、TAIR、MSU、Araport、Phytozome）。 - 对于多倍体基因组（如小麦、棉花），需区分基因拷贝。 - 过滤低表达基因（例如，在 >10% 细胞中计数<1的基因）。
基因注释 (Gene Annotations)	基因元数据，用于匹配计数矩阵中的基因 ID。	<ul style="list-style-type: none"> - 基因ID、符号、描述。 - 可选：功能注释（如GO、KEGG）。 	<ul style="list-style-type: none"> - 包含基因家族、旁系同源、ortholog信息。 - 使用植物特异数据库（如 Plaza、Phytozome）进行注释。
同源/映射表 (Homology Table)	基因间同源关系，用于构建无向同源图。	<ul style="list-style-type: none"> - 成对基因映射 (<code>query_gene, reference_gene</code>)。 - 列：E-value ($\leq 1e-6$)、bit-score (初始边权重)。 - 互惠命中 (reciprocal hits) 组合成无向图。 	<ul style="list-style-type: none"> - 考虑WGD和基因家族扩张。 - 包含旁系/ortholog标志，可选WGD校正标志。 - 使用植物资源（如Plaza、Phytozome）推断 orthology。

2. 文件格式规范

SAMap支持多种格式，但推荐使用AnnData格式以简化工作流。备选格式适用于从R或其他来源导入数据。

	文件类型		主要格式
计数	AnnData (.h5ad) – 原始计数存储在 adata.X中，必须为稀疏矩阵。	- TSV/CSV (基因×细胞 , 带表头) 。 - Matrix Market (.mtx)。 - R matrix/data.frame。	- 无需额外处理 , 仅移除低质量细胞。 - 基因ID必须匹配注释和同源表。
基因	嵌入AnnData的 adatavar中，或独立GFF/GTF文件。	CSV/TSV (列 : gene_id, symbol, description) 。 R data.frame。	- 标准GFF3/GTF字段 , 至少包含 gene_id 和 gene_name。 - 与计数矩阵基因 ID 匹配。
同源	目录中的成对 BLAST结果文件 (TSV , 如 pl_to_sc.txt) , 或单一TSV文件。	OrthoFinder/OrthoMCL 输出。 Ensembl Compara文件。 自定义TSV (query_gene, reference_gene, homology_type, confidence) 。	chr1 Araport11 gene 1000 2000 . + . gene_id "AT1G01010"; gene_name "MYB113"; AT1G01010 0s01g01000 0.98 或 BLAST: dd_Smed_v4_10001_0_1 Smp_042920 30.594 219 ... 4.21000e-30 112. - 过滤高置信边 (bit-score >50, E-value <1e-6) 。
表			

3. 预处理步骤

输入文件需经过预处理，以确保数据质量和兼容性。植物数据常需额外处理WGD和基因长度变异。

步

骤 操作

工具/命令

植物特异性提示

过滤	移除低质量细胞 (<500 总计数) 和基因 (在 <10% 细胞中表达)。	Scanpy: sc.pp.filter_cells , sc.pp.filter_genes 。	对于多倍体，保留参考等位基因或使用等位特异计数。
归一化	Log1p、TPM 或 scTransform , 但保留原始计数 (若 save_processed=False) 。	Scanpy: sc.pp.normalize_total , sc.pp.log1p ; scvi for scTransform 。	TPM 优先，因为它考虑植物基因长度变异。
基因ID统一	映射到共同命名空间 (如 Ensembl、Araport) 。	biomaRt、AnnotationHub、自定义表。	Arabidopsis: TAIR → Araport; Rice: MSU → Gramene。
生成同源图	对所有物种对运行 BLAST/DIAMOND , 使用蛋白质组/转录组。互惠最佳命中 → 无向图。	blastp -query pl.faa -db sc.faa -outfmt 6 -evalue 1e-6 或 vignette 中的 map_genes.sh 。	使用Plaza/Phytozome 蛋白质组；过滤低复杂度重复。
过滤同源	保留高置信边。	pandas/numpy 过滤；可选 homology_filter.py 。	保留旁系组 (如 MADS-box 家族) 作为单独节点或家族节点。
注释对齐	确保计数矩阵中的每个基因出现在 GFF/GTF 和同源表中。	内连接 (inner join) 基因 ID 。	移除非编码位点 (如假基因) , 若未在同源图中表示。
物种标识	分配两字符代码 (如 'ar' 、 'os') 。	字典: { 'Arabidopsis': 'ar' , 'Rice': 'os' } 。	在所有文件 (矩阵、BLAST、注释) 中保持一致代码。

4. SAMap 函数参数与输入相关性

SAMap的核心是 SAMAP 类初始化，它直接使用上述输入文件。以下是与输入文件相关的关键参数（基于Notebook示例）：

参数	类型	描述	默认/示例	植物特性提示
filenames	dict {species_code: path_to_h5ad}	每个物种的 AnnData 路径。	{'ar': 'arabidopsis.h5ad', 'os': 'rice.h5ad'}	包含所有物种。
f_maps	str	同源文件目录路径。	'data/maps/'	包含植物和植物-射（若合）。
save_processed	bool	是否保存过滤后文件 (*_pr.h5ad)。	True	对于长用；False I/O。
pairwise	bool	True: 每个细胞映射到每个其他物种的 k 最近邻 (>2 物种推荐)。 False: 跨所有其他物种。	True	对于两种 (如 Arabid vs. Rice) False 用比较。
resolutions	dict {species_code: float}	每个物种的 Leiden 聚类分辨率，用于定义邻域。	{'ar': 1.8, 'os': 1.5}	低值用性高的组织 (如 mesophyll, xylem)。
keys	dict {species_code: obs_column}	adata.obs 中的预注释细胞类型列，用于邻域构建。	{'ar': 'cell_type'}	植物细胞如 “mesophyll” “xylem”
neigh_from_keys	dict {species_code: bool}	True: 使用 keys 列 定义邻域 (覆盖 resolutions)。	{'ar': True}	用于已有的植物集。

reference_species	str (可选)	参考物种，用于联合流形锚定。	'ar' (Arabidopsis)	选择注整的物种。
wgd_correction	bool (植物特异)	True: 将WGD旁系分组为单一节点。	True	对于植物开启；对于非植物关闭。
gene_family	list or dict (可选)	显式基因家族分组。	None	用于主转录组基因家族。
batch_key	str (可选)	adata.obs中的批次列。	'batch'	若物种不同测序。
n_pcs	int	用于邻域构建的PC数。	50	大植物 (>100 胞) 增加。
homology_type	str or list (可选)	限制同源类型 (如['1:1'])。	None (所有类型)	['1:1'] 避免旁系噪音。

示例代码 (基于Notebook) :

Python

5. 植物跨物种整合的整体工作流建议

1. 获取/生成物种特异表达矩阵（.h5ad，原始计数）。
2. 过滤和归一化（优先TPM）。
3. 统一基因ID（使用植物数据库）。
4. 生成同源（BLAST，使用Plaza/Phytozome蛋白质组；E-value $\leq 1e-6$, bit-score > 50 ）。
5. 构建同源表（包含homology_type和confidence；标记WGD）。
6. 准备注释（GFF/GTF匹配基因ID）。
7. 初始化SAMAP并运行。

这些要求确保整合准确，尤其在处理植物的基因组复杂性时。若数据中存在批次效应或混合王国（植物-动物），需额外调整参数。Notebook强调，植物整合成功依赖高质量同源表，因此优先使用专用植物资源。