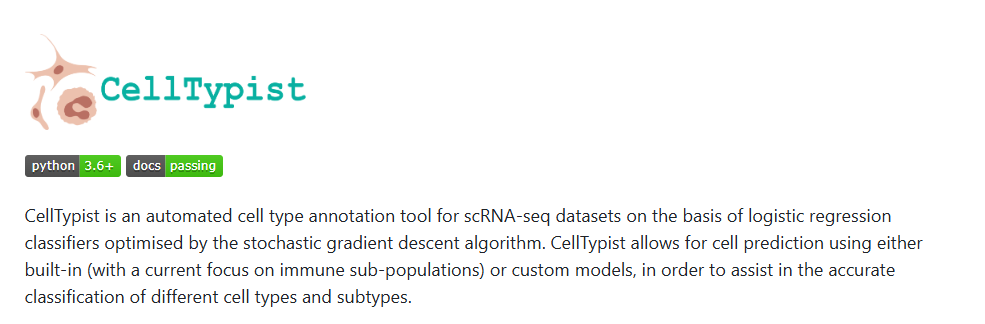
Celltype annotation

【文献与方法分享】三：使用CellTypist完成细胞类型的自动注释

<https://mp.weixin.qq.com/s/p4qGBv9JHCGg-0WCwXqWwA>



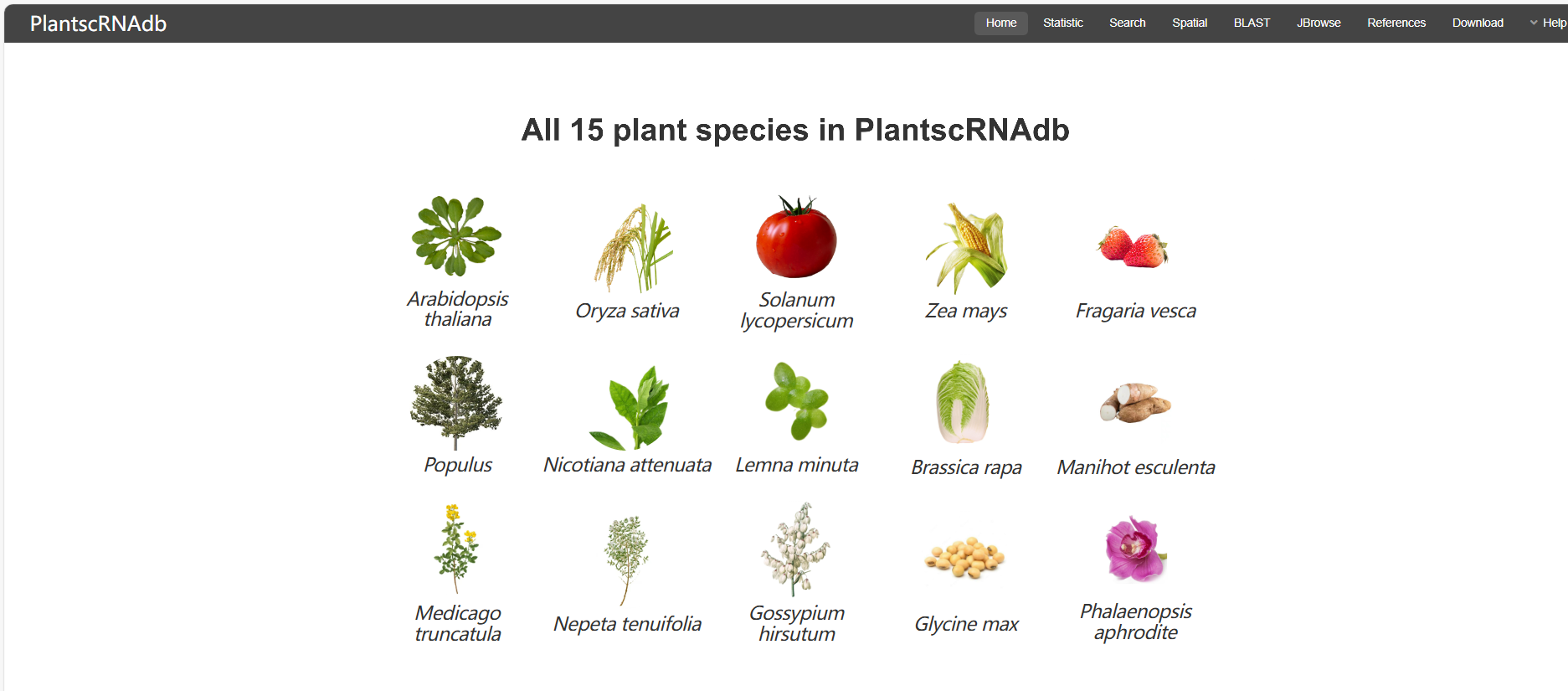
<https://www.celltypist.org/>

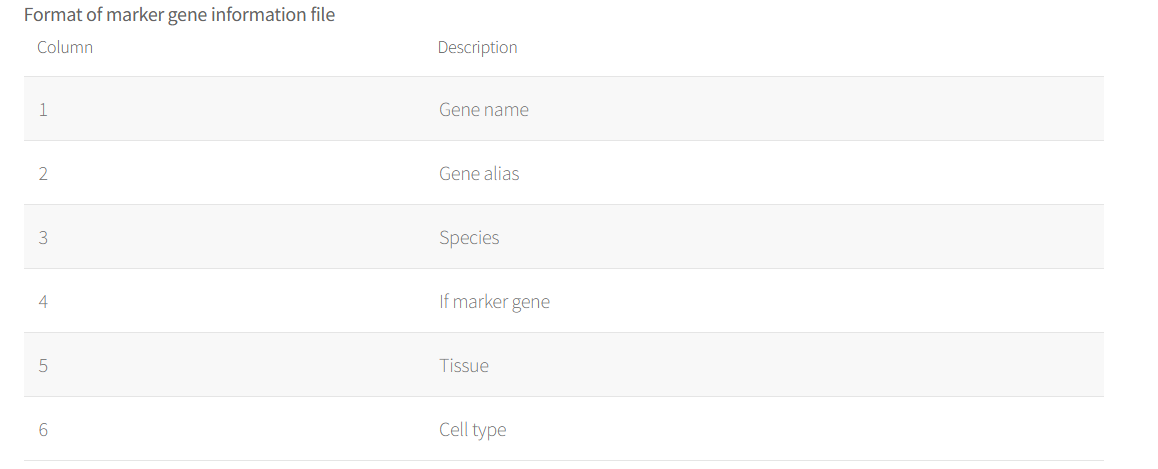
Celltypist似乎只有人的细胞模型，植物的还是要看其它的自动注释package

数据库提供植物单转数据的marker，比如PlantCellmarker、PlantscRNAdb、Plant Single Cell hub等。

|  |  |  |
| --- | --- | --- |
| Cell Taxonomy | 由北京基因组所（国家生物信息中心）创建于2023年。其基于4,299篇文献以及15个相关数据资源的单细胞测序数据分析，目前共收录了34个物种、387种组织的3,143种细胞类型和26,613个细胞标志物（cell marker）。是目前收录物种，组织最多，细胞类型最丰富，在线功能最实用的的数据库之一。 | [https://ngdc.cncb.ac.cn/celltaxonomy/](https://ngdc.cncb.ac.cn/celltaxonomy/" \o "https://ngdc.cncb.ac.cn/celltaxonomy/) |
| PlantscRNAdb | 由浙江大学创建于2021年，该数据库更新速度比较高，至今已更新了5版。由2021年3月V1.0的4个物种更新到2023年8月的V3.0的15个物种。目前V3.0收录了15个物种的38个组织的451种细胞类型和114,770个细胞标志物（cell marker）。Marker基因也是来源于大量文献整理，其中包含实验验证的，特定细胞类型的bulk RNA-seq以及scRNA-seq数据的，针对scRNA-seq作者又使用了统一的方法和共同的参数进行重新分析。 | <http://ibi.zju.edu.cn/plantscrnadb/index.php> |
| PCMDB | plant cell marker database, 由郑州烟草研究院也是创建于2021年，与PlantscRNAdb相比，耗时更长，从最初的12.5万篇文献中通过人工一步一步的筛选获得，marker同样来自三个途径，实验相关的，scRNA-seq发现的，以及一些组织/细胞的bulkRNA-seq的显著差异marker基因。最终获得了6种植物的22个组织的263种细胞类型的81117个marker基因，其中3119来源于实验研究，平均每种celltype有17个实验相关marker可用。 | <https://www.tobaccodb.org/pcmdb/homePage> |

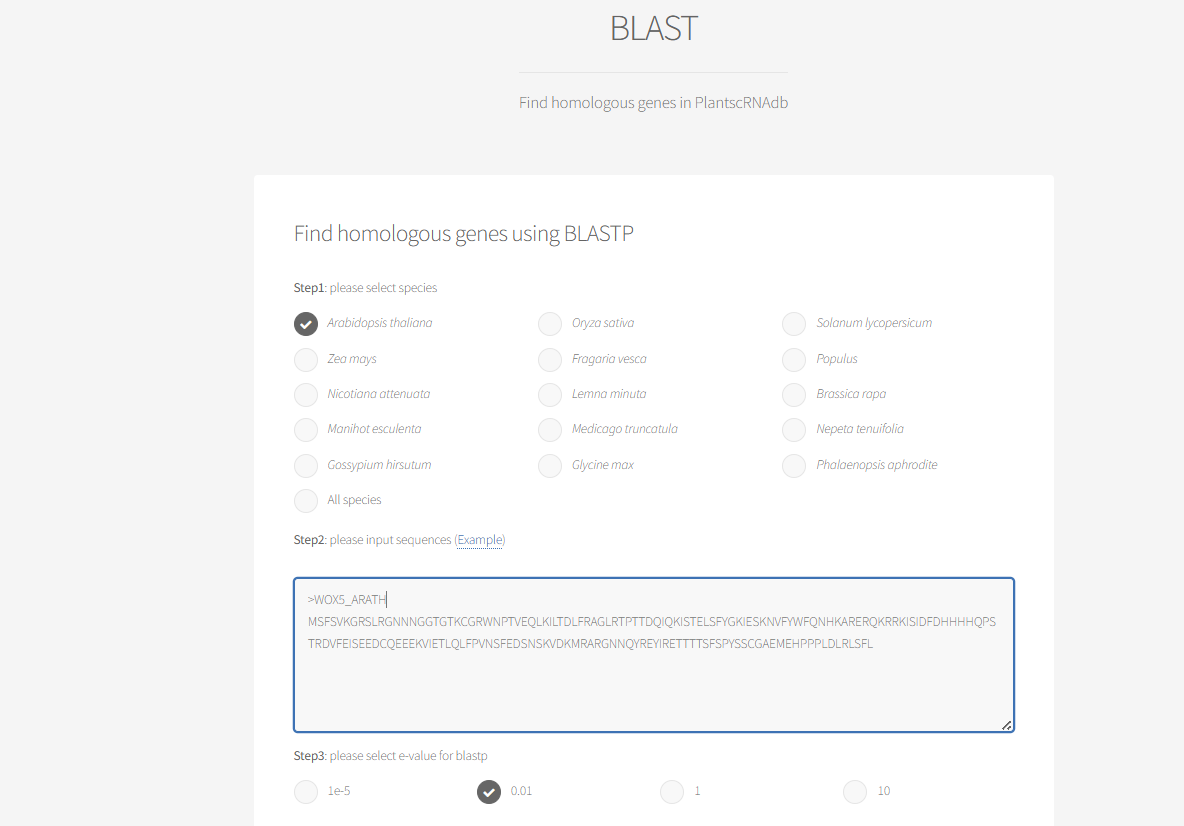
PlantscRNAdb 感觉是植物物种比较全的，主要的marker资源是拟兰芥。





首先水蕨Cer这里面是没有的，要将marker基因与细胞类型对应起来，那么肯定也需要一种对应表，A基因→细胞类型1；B基因→细胞类型2；注意到，对于拟兰芥而言，它应该是Ara0001→细胞类型1；水蕨是Cer0001→细胞类型？；现在问题在于拿到水蕨每个cluster的差异表达基因后，它的基因命名是按照水蕨来的，你要借助拟兰芥的基因和细胞类型对应表来的推测水蕨的细胞类型，那么就得在Ara0001和Cer0001之间建立联系，它们具有同源性，那么蛋白质产物相似，那么形状相似，从数据来看就是序列相似，所以应该就要用同源比对。这时候两个方案：①运用这两个的核酸序列来进行比对；②基因的蛋白质来比对；蛋白质比对可能更加可靠，但是要比对蛋白质就必须对物种基因组要求，基因组注释的质量要求，对于非模式生物的化怀疑可能主要是进行核苷酸序列的比对。

问题：获得单细胞的数据Cer的基因名列和核酸序列，获得拟兰芥的基因名和核酸序列表。将其进行比对，获得一个列表，列表应该未Cer基因名，Ara基因名

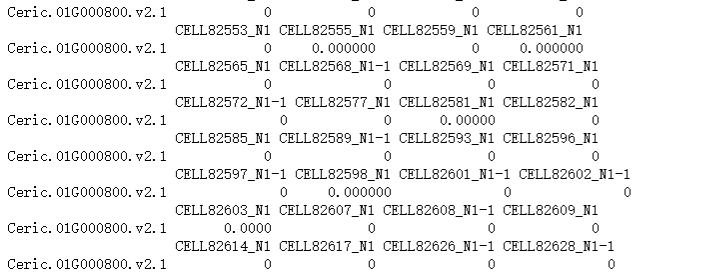


用的是blastp，用的蛋白质比对，也很正常，由基因组到蛋白质组本身就不是简单的翻译，蛋白质组才是实在的表型实现。哦，看来要从转录本的mRNA得到对应的翻译产物protein，然后将protein和marker基因列表获得同源，得到对应列表，然后通过这样猜测细胞类型

Ara的marker注释是基因名-组织-细胞类型-评分/等级（？）

Cer通过FindMarker可以找到cluster的显著基因，通过dotplot可以看那些基因是在其cluster里面显著表达的，然后通过Cer的基因通过蛋白质比对检索到Ara的marker基因列表，获得其细胞类型名称，进而给自己的细胞进行注释

难点：Cer的cluster显著基因的获得、Cer的基因蛋白质的准确性、能够准确检索到Ara的marker列表，也考研其数据的准确全面性，遇到两个cluster有很多相近表达还真是会比较难注释



Rds格式的数据中并没有保留原始数据的探针和测到的核酸序列，只保留对应的名称，当然这也没关系，在下机的10X格式里面有。10X标准文件包含cellranger上游比对分析产生的barcodes.tsv.gz，features.tsv.gz和matrix.mtx.gz 3个文件，分别代表细胞标签(barcode)、基因ID(feature)、表达数据（matrix）。那么features.tsv.gz肯定是有这个对应关系的，用这个去做比对也是ok的。