# Scvelocity 结果说明

1. **背景介绍**

单细胞 RNA-seq 只能提供测量时细胞状态的静态快照，而无法随时间推移跟踪细胞。RNA速度的概念（La Manno等，2018）不仅可以获取细胞的当前状态，还可以获取其在转录组空间中的运动方向和速度，从而开启了研究细胞动力学的新方法，实现细胞动力学的预测模型。

RNA Velo 基于基因剪接和非剪接信使 RNA (mRNA) 的比率，描述了其在给定时间点的表达变化率。然而，如果违反了关于一般剪接速率的中心假设，以及观察到的稳态 mRNA 水平的完整剪接动力学，RNA速度估计可能会出现误差。ScVelo 通过使用基于似然的动力学模型来求解完整的剪接动力学，克服了这些限制。这将 RNA Velo 推广到具有瞬时细胞状态的系统，这种状态在发育和对扰动的反应中是常见的。

ScVelo 能够推断基因特定的转录、剪接和降解速率，恢复每个细胞在潜在分化过程中的位置，并检测可能的驱动基因。ScVelo共提出了3个模型，其中稳态模型（La Manno等，[2018](https://www.embopress.org/doi/full/10.15252/msb.202110282" \l "msb202110282-bib-0009" \t "_blank))将速度估计为观察到的未剪接与剪接 mRNA 的比率与推断的稳态比率的残差。在稳态模型中，RNA速率通过对在预期达到稳态表达水平的下分位数和上分位数中发现的细胞进行线性回归来近似估计。但该方法的假定前提是每个细胞的降解率、剪切率等参数都是固定不变的，但这显而易见是不可能的，因此ScVelo推出了动态模型，该模型引入了[期望最大化算法](https://links.jianshu.com/go?to=https://towardsdatascience.com/expectation-maximization-explained-c82f5ed438e5" \t "_blank)(Expectation Maximization，EM)来估计参数，该算法通过最大似然法迭代逼近转录速率、未剪接到剪接的剪接速率和剪接后 mRNA 产物的降解速率，并学习给定基因的剪接/未剪接轨迹。通过这种方式能够恢复每个细胞在潜在分化过程中的位置，并检测可能的关键驱动基因。与稳态模型相比，动态模型通常能在相邻细胞之间产生更一致的速度估计，并更准确地识别转录状态。

**2.背景介绍**

#### 2.1.数据状态概览

#### 该部分结果见1.spliced\_unspliced\_Statistics文件夹

图1 基于celltype的剪切/未剪切比例展示

该图显示的是每个细胞类型（或者其他数据标签）的剪切/未剪切count的比例，理论上比例变化与细胞的发育水平相关。

**2.2.细胞速度矢量可视化**

该部分结果见2.scvelo\_stream文件夹

此文件夹结果展示了软件推断的细胞可能的状态变化(从一个细胞到另一个细胞的过渡)，并将速度矢量嵌入umap图(或tsne)中。

该文件夹包含三种单细胞速度矢量嵌入降维聚类结果的方法。

分别是1.单个细胞水平：single-cell\_level\_dynamical\_velocity\_by\*.pdf；

2.网格线：scvelo\_embedding\_grid\_by\*.pdf；

3.流线型：velocity\_embedding\_stream\_by\*.pdf

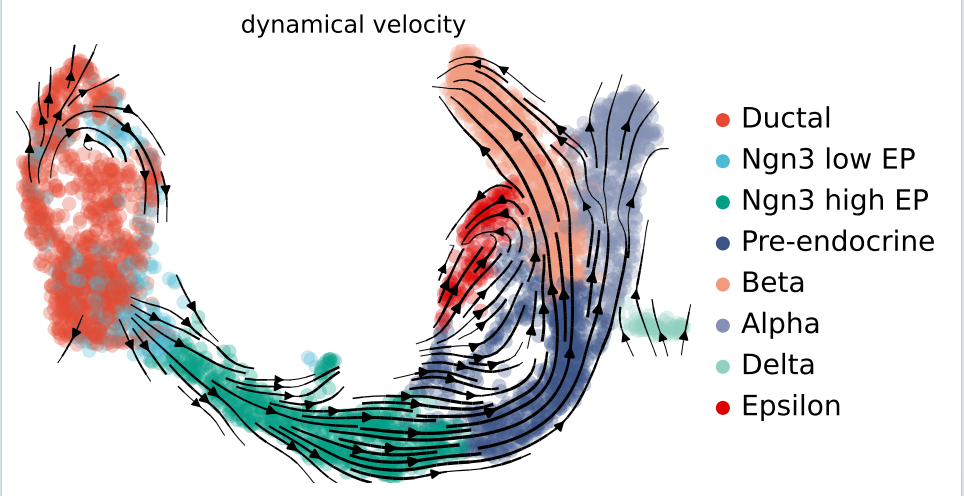


图2：velocity\_embedding\_stream\_by\*.pdf



图3：single-cell\_level\_dynamical\_velocity\_by\*.pdf

该图反映的是由箭头长短和方向指代单个细胞水平的细胞分化速度和方向。细胞的着色可以替代成任何细胞的标签，如样本名，分组名等。

**2.3.潜伏时间**

该部分结果见3.latent\_time文件夹

潜在时间（Latent Time）是由SCVelo推断的，代表了细胞的内时钟，能够准确描述细胞在潜在生物过程中的位置。与现有的基于相似性的伪时间方法不同，这种潜在时间仅基于转录动力学，并且考虑了基因表达变化的速度和方向。因此，它比基于相似性的伪时间方法更好地捕捉实际时间的各个方面。



图4：latent\_time\_by\_dynamical.pdf

该图是每个细胞的分化潜在时间展示图，数值越低的细胞，表示该细胞分化程度越低，而数值越高的细胞，表示该细胞的分化接近终末状态。

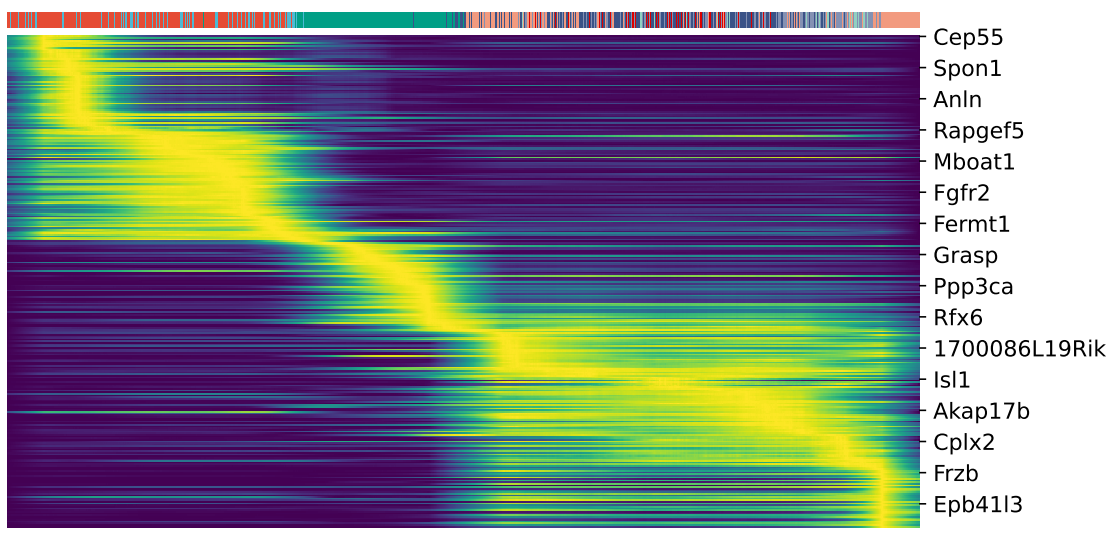


图5：latent\_time\_heatmap\_by\*.pdf

该图的横坐标是按潜在时间（latent time）排序的细胞，纵坐标是基因。热图数据展示了前300个高似然基因在这些细胞中的表达量。通过将这300个高似然基因沿着潜在时间进行动态分析，可以观察到一个清晰的转录级联过程。

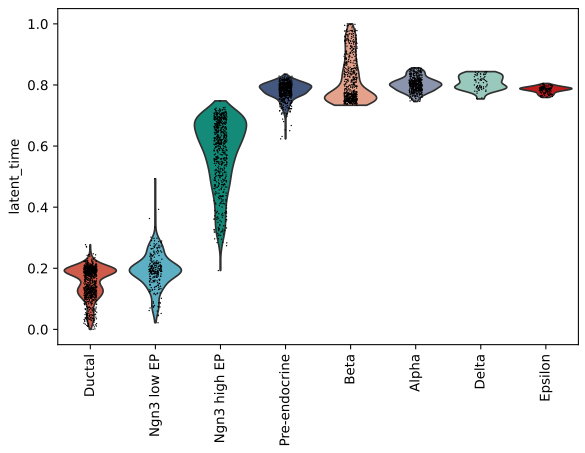
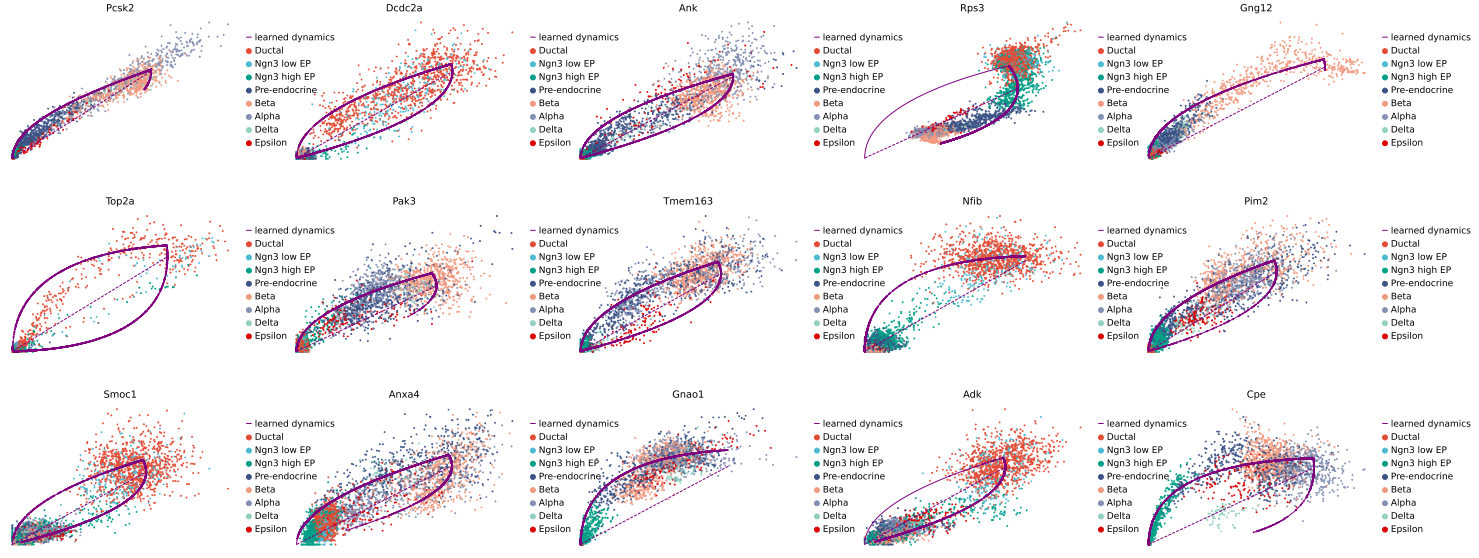


图6：scVelo-violin-latent\_time\_by\*.pdf

**2.4.驱动基因计算与展示**

该部分结果见4.likelihood\_gene文件夹。

高似然基因在动态模型中可以被认为是驱动基因。这些基因展示出显著的动态特征，并且其变化可以明显地被检测到。

图7：Top15-likelihood\_genes\_by\_\*.pdf

该图为前15个高似然基因绘制的基因相图，横坐标为剪切数，纵坐标为未剪切数，紫色实线为推断的轨迹，虚线为稳态比率。

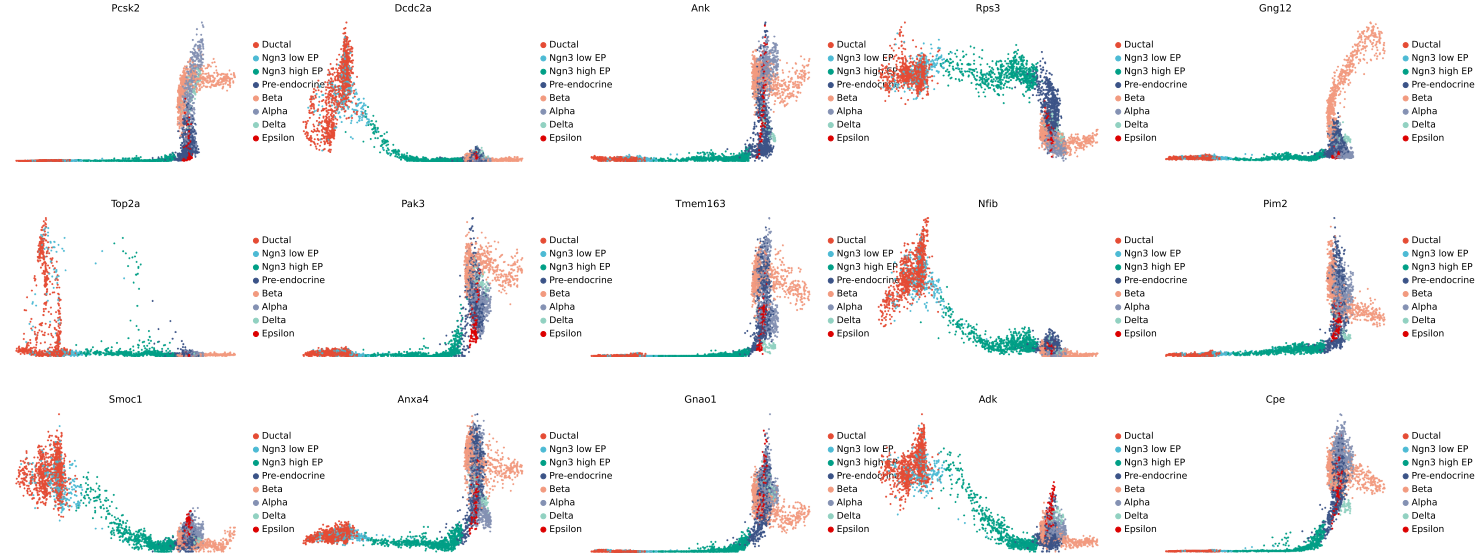


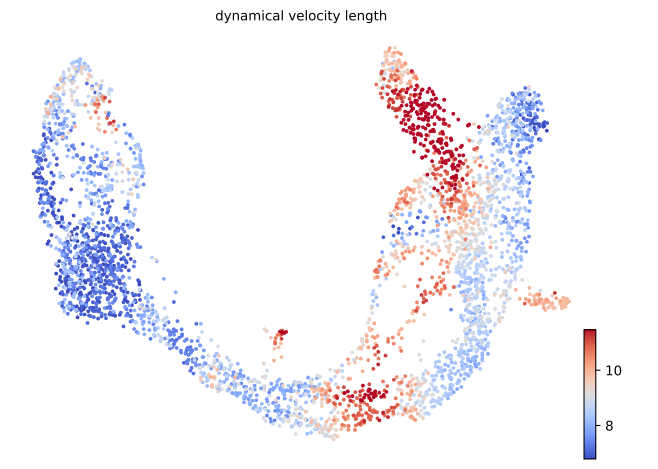
图8：Top15-likelihood\_genes\_latent\_time\_by\_\*.pdf

该图为选取前15个高似然基因绘制的基因随latent time的表达轨迹图，横坐标代表latent time,纵坐标代表基因表达量，可以看出高似然基因的表达量随着潜在时间的变化，在某个点发生了较为明显的上升或者下降。

Top10\_specific\_likelihood\_gene\_by\*文件夹为每个cluster（或其他细胞标签）的前10个高似然基因相图。

使用Welch t检验的差异表达测试的方法，检测在某个群中的基因，它们显示出与所有其他群相比在转录调控上存在不同的动态变化（例如，在该群中表现为诱导，而在其他群体中则保持稳态），并选取top10的基因绘制基因相图。结果保存在Top10\_specific\_velocity\_genes\_by\*文件夹中。

**2.5.速度与相干性计算**

图9：dynamical\_velocity\_length.pdf

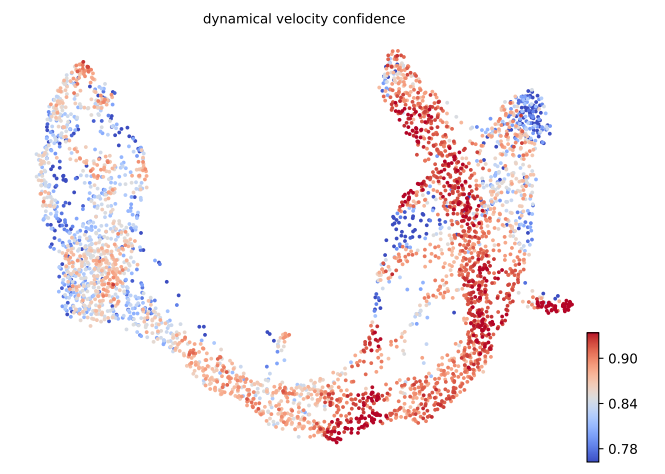
该图展示的是细胞分化速率，颜色由蓝到红表示分化速率由低到高。

图10:dynamical\_velocity\_confidence.pdf

该图为置信度打分，颜色由蓝到红表示置信度由低到高。

每个细胞都有一个长度为 g（表示基因数量）的速度向量，这个向量的长度可以用来表示该细胞分化的速率。然而，这个速率是无方向的。当计算细胞间的速度向量的余弦距离后，可以判断这些细胞是否在速度上共表达。用余弦距离定义的共表达（correlates）称为细胞间的相干性（coherence），这种相干性即为置信度（理论上，同簇的近邻细胞应具有相干性）。