お詫びと注意

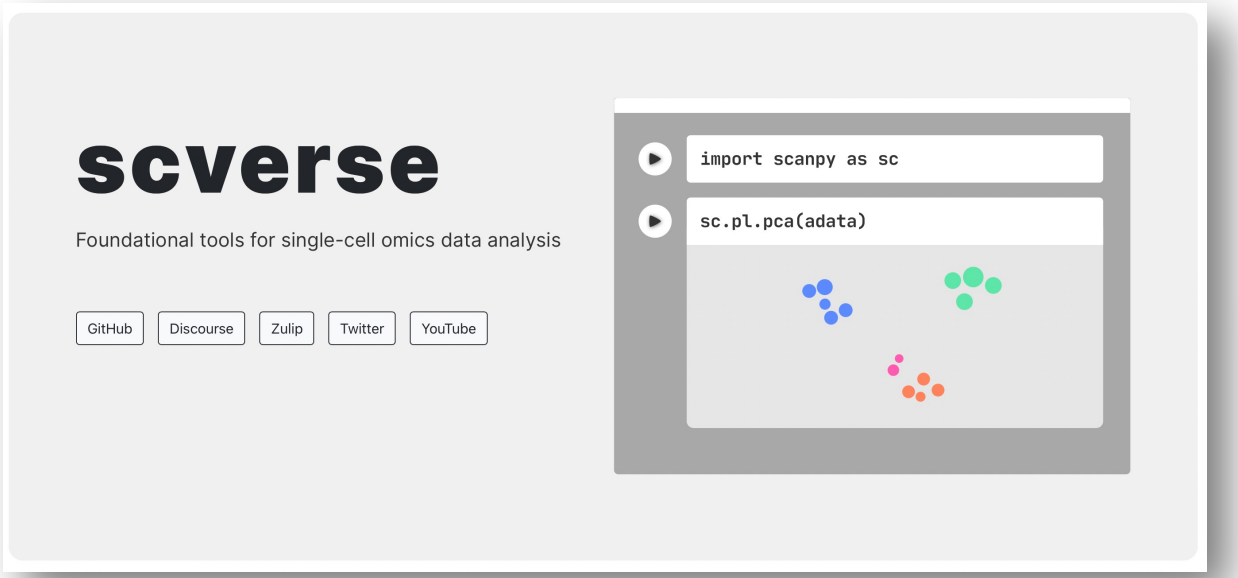
- シングルセル解析のみ。Pythonプログラミングの講習というよりシングルセルRNA-seq解析の講習。

- 「ベストプラクティス」というわけではありません。発展的なツールの紹介もあって、中には有効性がまだじゅうぶんに検証されていないものもある。

- M1チップのMacなど、必要パッケージがインストールできなかった場合はGithubページのHTML版か、Google Colabのバージョンで実行。
  インストールできなかったのがscikit-miscだけの場合は講習中に変更点がひとつあるだけなのでそのまま実行していただいて大丈夫です。

PythonによるシングルセルRNA-seq解析

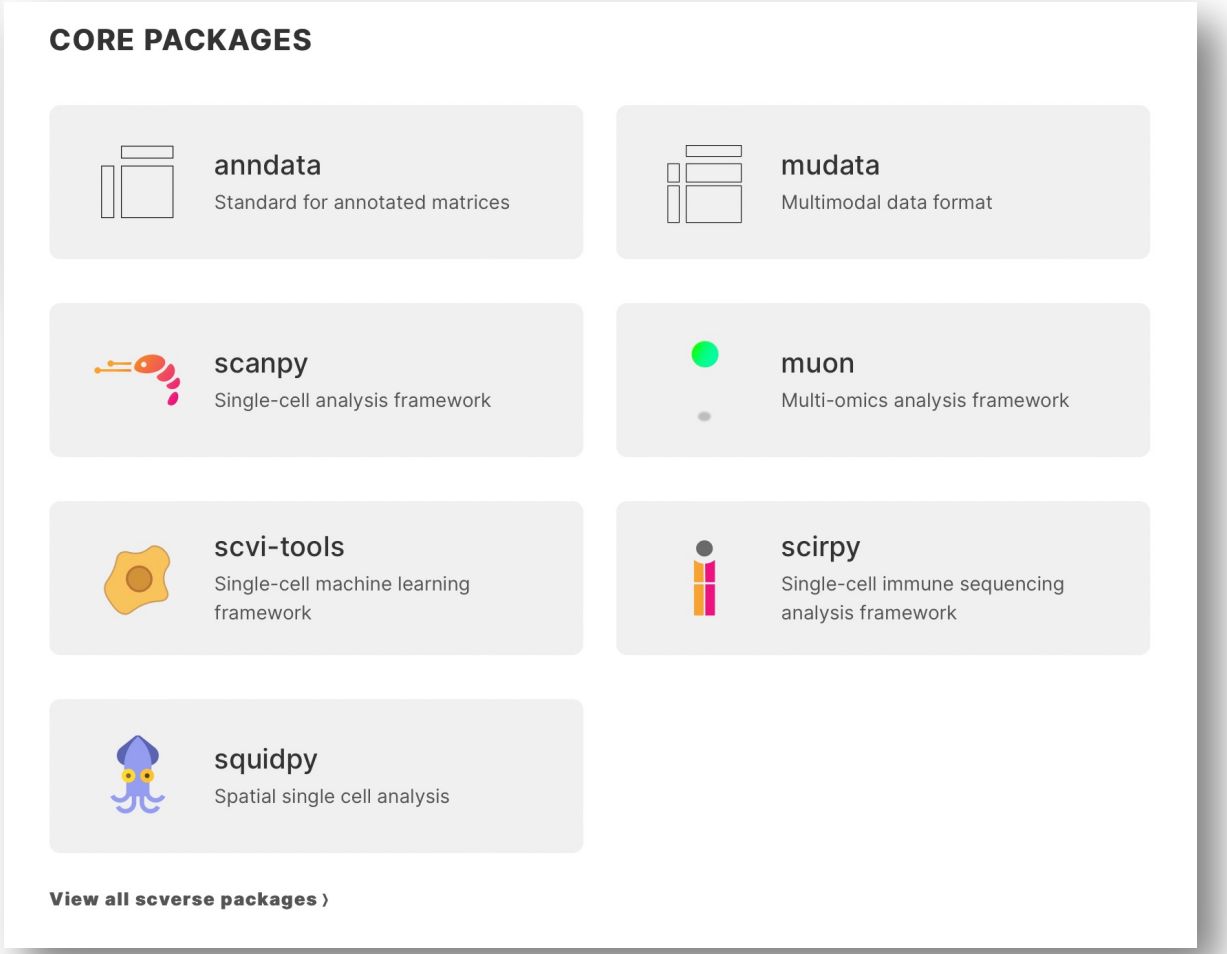
scRNA-tools (https://www.scrna-tools.org)が収録しているツールの開発言語
Zappia, L., Theis, F.J. *Genome Biol* **22**, 301 (2021).



https://scverse.org

単一細胞オミクス解析に関連するPythonツールの開発・維持を目的に2022年に組織されたコンソーシアム。

AnnDataとScanpyをコア技術とする。
マルチモーダルデータ（scRNA-seq + scATAC-seq）の解析に対する拡張として MuData, Muon の開発、
空間トランスクリプトーム解析のためのSquidpyの開発など。
それぞれの相互運用性の改善やファイルフォーマットの統一など、
一体として扱いやすいツール群の開発を目指していくコミュニティ。



AnnData

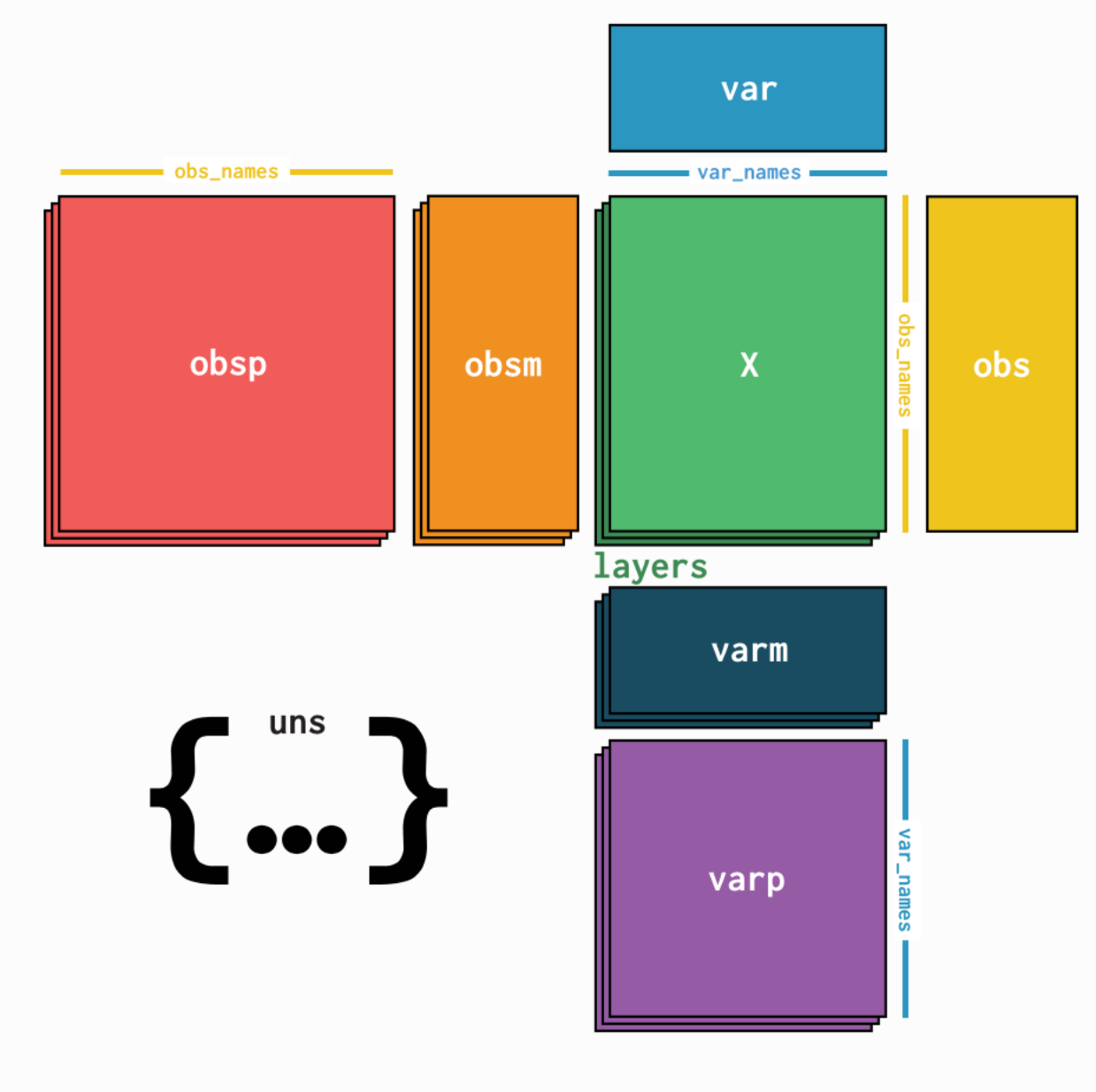
“Annotated Data”（アノテーションされたデータ）の略。

オミクスデータ格納のためにPandasのDataFrameを拡張したデータ構造。
シングルセル解析のためのPythonパッケージの多くが、このオブジェクトに対する計算として実装されている。

オミクスデータは実験で測定された数値テーブルのほかに、
観測値（obs）、変数（var）それぞれが多様な情報を持つ。
たとえばRNA-seqの場合、観測値であるサンプルは実験条件・性別・年齢など
様々なメタデータを持つ。変数である遺伝子も、遺伝子IDやシンボルだけでなく、
機能カテゴリや、DEGか否かなどのメタデータを持つ。

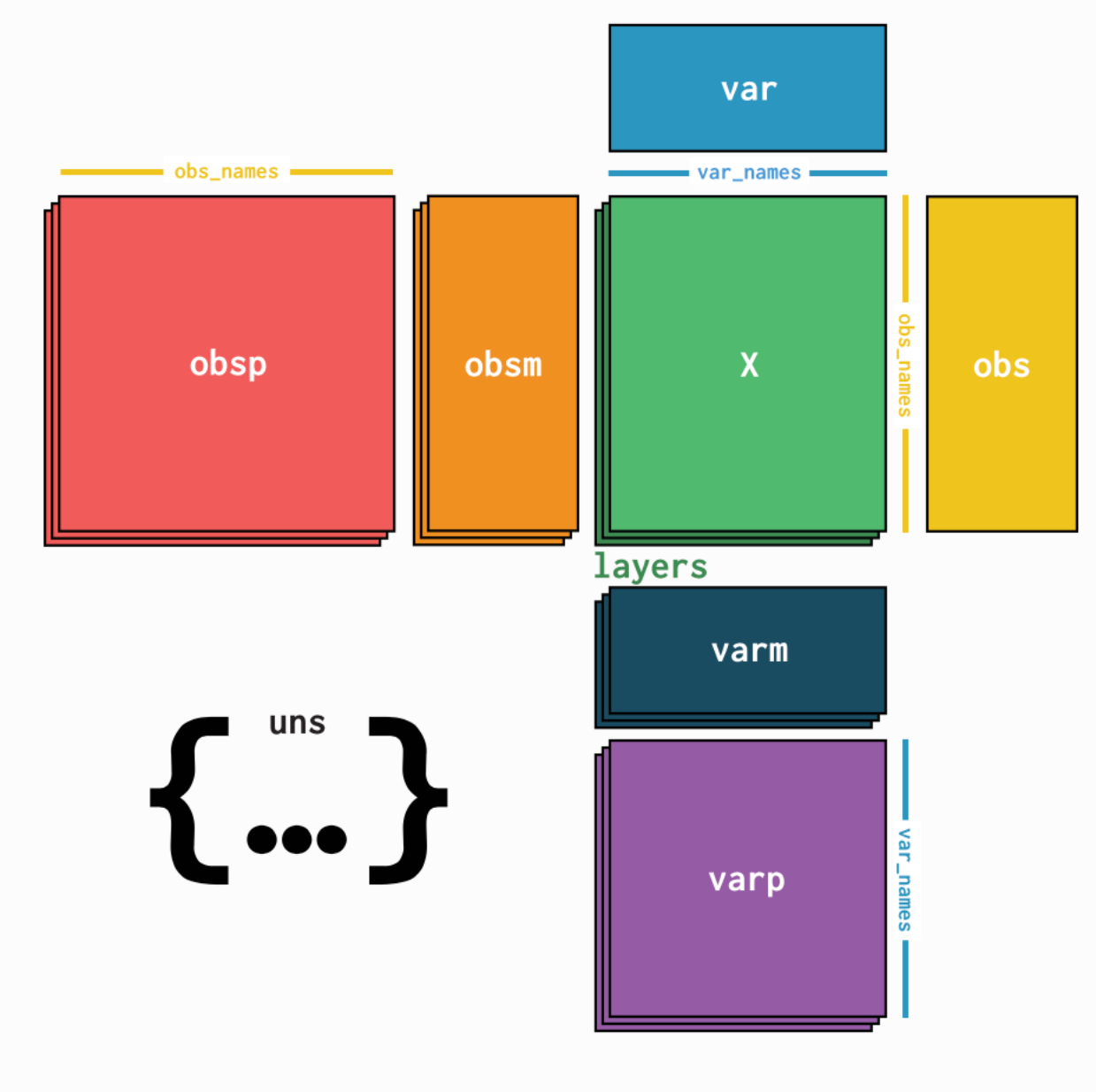
それぞれを個別のオブジェクトとして管理するのはとても面倒。
数値テーブルになんらかの操作を施した結果が、観測値や変数のメタデータに即座に反映されない。
なので、複数のオブジェクトをいちいち行ったり来たりしなきゃならない。
テーブルに対する計算の結果わかったことを観測値のメタデータに入れて、
その結果に基づいて観測値をセレクションしたから今度は数値テーブルを同じようにスライスして、、、みたいな。

そういった面倒を避けるために、すべての観測と計算結果をひとつのオブジェクトに詰め込んで管理しやすくしたのが、AnnData というオブジェクトの特徴。



AnnData

- **.X**
  n_obs × n_varsの数値テーブル。numpy.ndarrayやscipyのスパースマトリックス。scRNA-seqのカウントマトリックスなど、実験の根幹となるデータ。**layers** に、同じshapeの複数のマトリックスを保持しておける。たとえば全体をノーマライズしたけど元々のカウントデータも残しておきたいときは別のレイヤーに入れておく。スライスの影響はすべてのlayerに作用する。
- **.obs**
  observationsの略。観測値に関するメタデータ。PandasのDataFrameなのでPandasの操作は全部実行できる。長さは必ず n_obs
- **.var**
  variablesの略。変数（遺伝子など）に関するメタデータ。PandasのDataFrame。長さは必ず n_var
- **.obsm**
  multi-dimensional annotations for obs. 複数の数値のまとまりでそれぞれの観測値を表現したいときに使う。各観測値の低次元空間座標など。次元サイズは任意。n_obs × 次元サイズの numpy.ndarray.
- **.varm**
  multi-dimensional annotations for var. n_var × 次元サイズのnumpy.ndarray
- **.obsp**
  Pairwise annotation of obs. 観測値のペアに関する情報。距離行列など。n_obs × n_obs のnumpy.ndarray
- **.varp**
  Pairwise annotation of var. 変数のペアに関する情報。距離行列など。n_var × n_var のnumpy.ndarray
- **.uns**
  それ以外のデータ。とくに構造の制限はない。その他の関連データをひとまとめにしておきたいときに辞書型で放り込んでおく。クラスタの色指定とか。



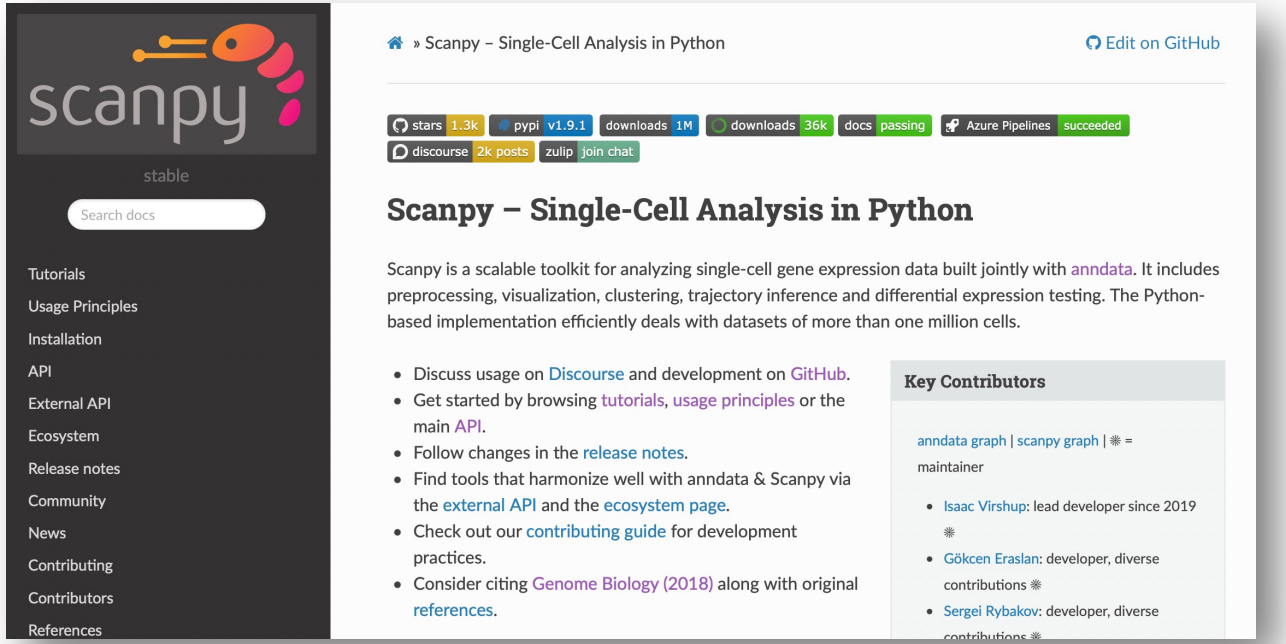
Scanpy

Pythonでシングルセル解析をする際のコアパッケージ。

データの前処理や、近傍グラフ構築、t-SNEなど、
標準的な解析を実行できる。

基本的に、
AnnDataオブジェクトを入力して関数を実行すると、
結果が同じAnnDataオブジェクトに追加されていく。
新しいAnnDataを返すのではなく、
inplaceで（＝破壊的に）AnnDataが変換されていくのが特徴。

一見どこにどんな変化が生じたのかわかりにくい。
観測値や変数のデータフレームにいつのまにか勝手にカラムが追加されていることがある。



https://scanpy.readthedocs.io

Scanpyの関数

- **scanpy.pp.XXX**
  前処理（**pre**processing）に関連する関数がある。
  細胞や遺伝子のフィルタリング、対数変換や、近傍グラフの構築など

- **scanpy.tl.XXX**
  さまざまなツール（**to**o**ls**）のセット。
  PCA, t-SNE, UMAPなどの次元削減や、Louvain/Leidenクラスタリングなど。

- **scanpy.pl.XXX**
  プロット（**pl**otting）用の関数。
  PCA用のプロット、UMAP用のプロットなど、それぞれの可視化に適した関数が用意されている。
  複雑な処理を書かなくても、anndataに含まれるメタデータから自動的に、遺伝子発現量による色のグラデーションや、クラスタごとの色分けなどをやってくれる。

注：scanpyはたいてい “sc” の短縮名で呼び出すことが多いので、以上の関数は、sc.pp.XXX, sc.tl.XXXなどと呼び出す

scVI : 深層生成モデルを利用したシングルセルデータの確率的解析

Gayoso, Adam, et al. "A Python library for probabilistic analysis of single-cell omics data." *Nature Biotechnology* 40.2 (2022): 163-166.

SOLO : ダブレット検出のための半教師付き深層学習

scVIでモデリングした変分オートエンコーダの構造を流用。
エンコーダ（カウントデータから潜在表現への変換）の出力部分に、
single/doubletの二分類を予測するニューラルネットワークを接続。
シミュレーションデータ（適当なふたつの細胞の平均発現パターン）でニューラルネットワークを学習してから、実際のデータのダブレットを予測する。

Bernstein, Nicholas J., et al. *Cell systems* 11.1 (2020): 95-101.

CellAssign : Cell typeの自動推定

事前に定義されたマーカー遺伝子の情報を活用して、それぞれの細胞を既知のCell type、あるいは“unassigned”に割り当てる手法。

マーカーの事前知識は、閾値設定などが必要なく、「マーカーか否か」の1/0の情報だけあればいいことが特徴。

Zhang, Allen W., et al. *Nature methods* 16.10 (2019): 1007-1015.

scANVI : アトラスとの統合、ラベル転移

基本的にはscVIと同じ、
エンコーダ / デコーダのVAEモデル。
ただし、一部の細胞にはCell typeの
ラベルがついていて、潜在表現 z が
このラベルに依存して決まるように
モデリング。

Xu, Chenling, et al. *Molecular systems biology* 17.1 (2021): e9620.

RNA velocity --- Velocytoの場合

La Manno, Gioele, et al. "RNA velocity of single cells." *Nature*560.7719 (2018): 494-498.

$$\frac{du}{dt} = \alpha(t) - \beta(t)\,u(t)$$

$$\frac{ds}{dt} = \beta(t)\,u(t) - \gamma(t)s(t)$$

$$\frac{du}{dt} = \alpha - u(t)$$

$$\frac{ds}{dt} = u(t) - \gamma s(t)$$

転写のパラメータαは一定、
スプライシング 効率のパラメータβは
全遺伝子共通（1）と仮定

RNA velocity --- scVeloの場合

定常状態が観測できなかった場合
（ spliced/unsplicedのkinetics全体像の一部しか観測できなかった場合）、速度の推定は難しくなる。

Bergen, Volker, et al. *Nature biotechnology* 38.12 (2020): 1408-1414.

RNA velocityの注意点：二次元表現の矢印はいったいなんなのか

本来のRNA速度は高次元空間（全遺伝子の空間）
で定義されているはず。
ならばなぜ、非線形変換したt-SNEやUMAPなどの
二次元平面に矢印なんて描けるのか。

いろいろと自明ではない計算を重ねて無理やり二次元表現に
落とし込んでいる

細胞ごとに発現パターンの差分と速度ベクトルの相関を計算

$$\pi_{ij} = \cos\angle\left(\delta_{ij}, \mathbf{v}_i\right) = \frac{\delta_{ij}^T v_i}{\left\|\delta_{ij}\right\|\|v_i\|},$$

その値を確率値に変換

$$\tilde{\pi}_{ij} = \frac{1}{z_i}\exp\left(\frac{\cos\angle\left(x_j - x_i, v_i\right)}{\sigma_i^2}\right),$$

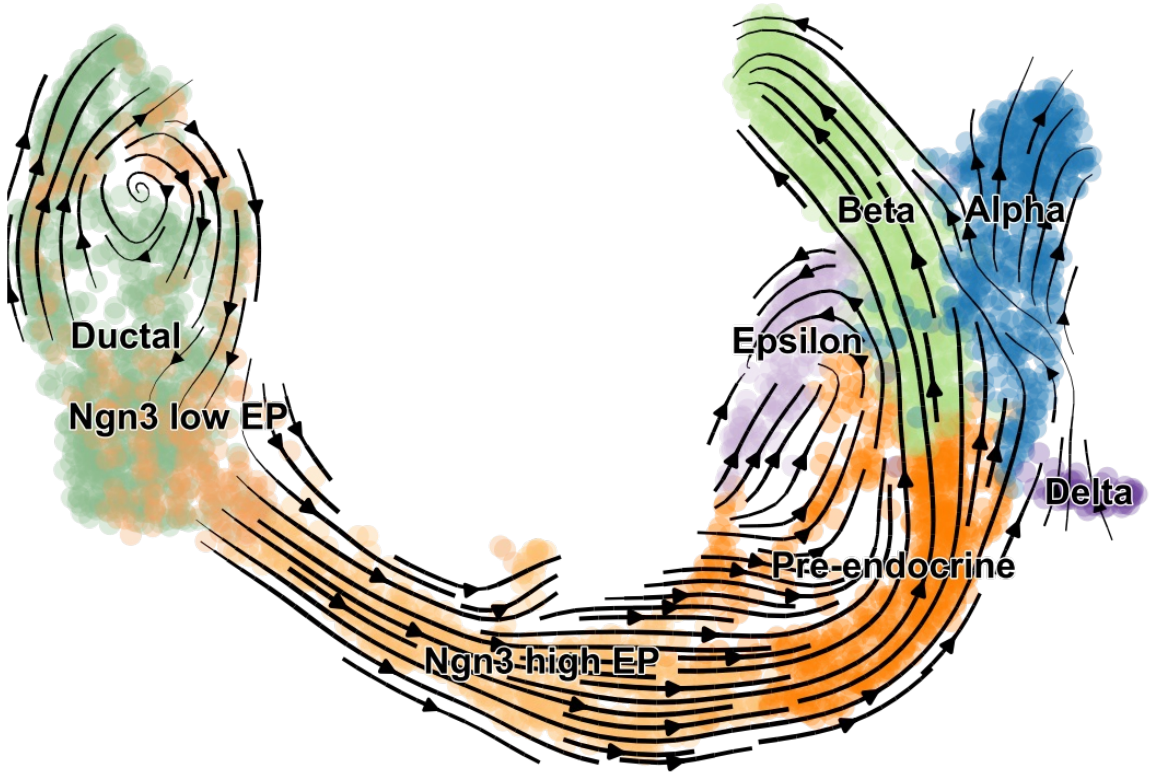
埋め込み空間における座標で細胞間の方向ベクトルを定義

$$\tilde{\delta}_{ij} = \frac{\tilde{s}_j - \tilde{s}_i}{\left\|\tilde{s}_j - \tilde{s}_i\right\|},$$

その方向ベクトルを前述の確率値で重み付け

$$\mathbf{v}_i = \mathbb{E}_{\tilde{\pi}_{i\cdot}}\left[\tilde{\delta}_{i\cdot}\right] = \sum_{j \neq i}\left(\tilde{\pi}_{ij} - \frac{1}{n}\right)\tilde{\delta}_{ij},$$

1/nは埋め込み空間で細胞が均等に分布していない点を補正しているらしい
細胞が密集しているエリアに方向ベクトルが引っ張られやすいため



CellRank

細胞をランク付けしたりオーダリングするツール、ではなく、名前の由来は初期のGoogle検索に用いられたアルゴリズムの “PageRank”

近傍グラフ上で、細胞間の「発現パターンの差分」とRNA速度ベクトルとの相関を基にして、細胞間の遷移確率行列を構成する。

粗視化した遷移確率行列から「初期状態」「終端状態」に対応する細胞集団を特定する。

すべての細胞について、どの「終端状態」にたどり着くか、その運命確率を計算できる。

Lange, Marius, et al. *Nature methods* 19.2 (2022): 159-170.