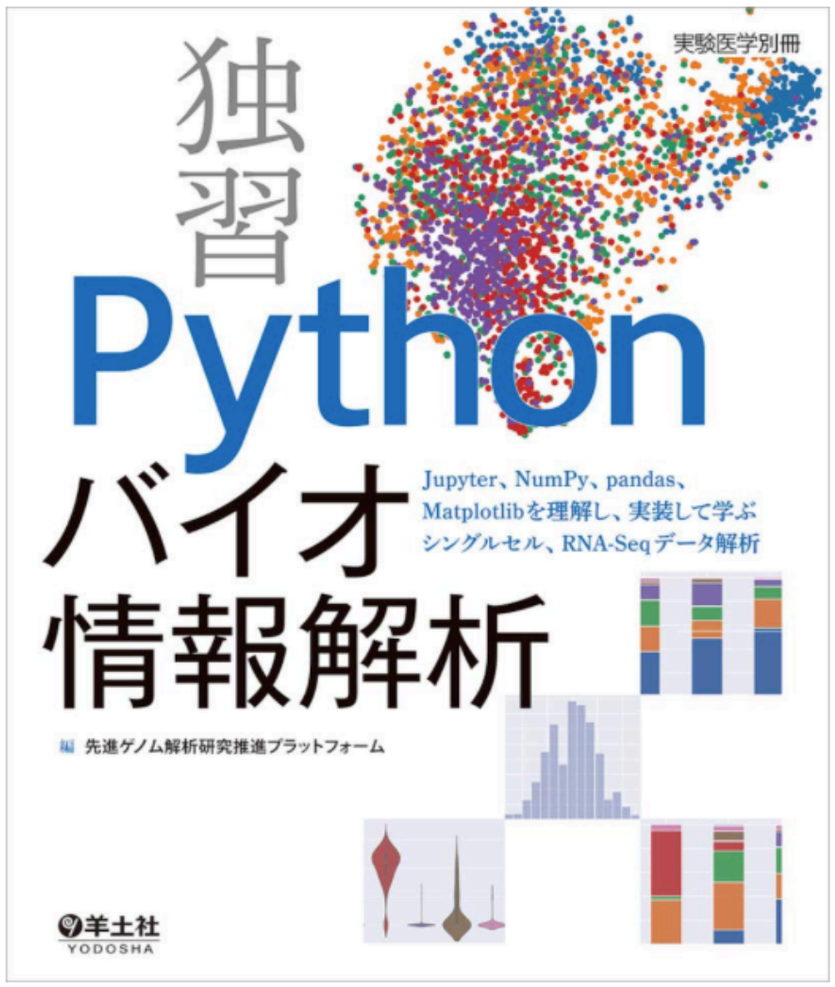
Pythonを用いたシングルセルRNA-seq解析（基礎）

2024/02/14

遺伝研 望月孝子

謝辞

本講習の内容は、2019年度、2023年度の先進ゲノム支援の情報解析講習会 遺伝研・東光一先生の資料、ご発表の内容をベースにさせて頂いております。



独習 Python バイオ情報解析の内容でも勉強させて頂きました。

scRNA-seq の解析のステップにおいて、なぜその処理が必要なのか？そのアルゴリズムはどうなっているのか？を説明して下さっていて大変参考になりました。

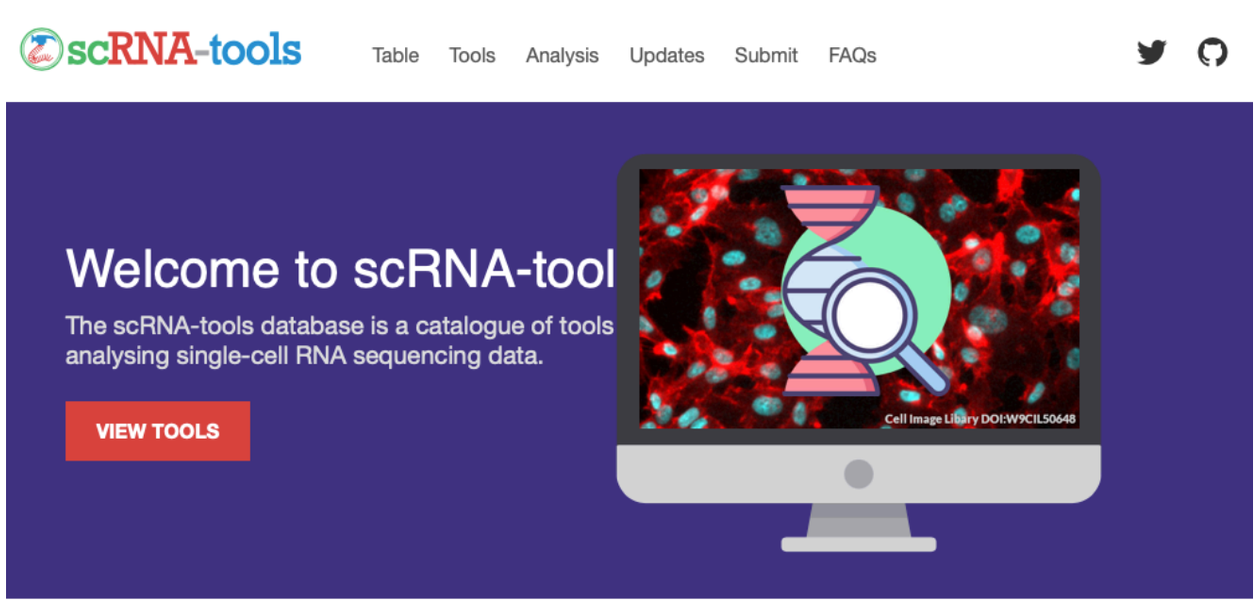
東先生、ありがとうございました。

ご注意ください。

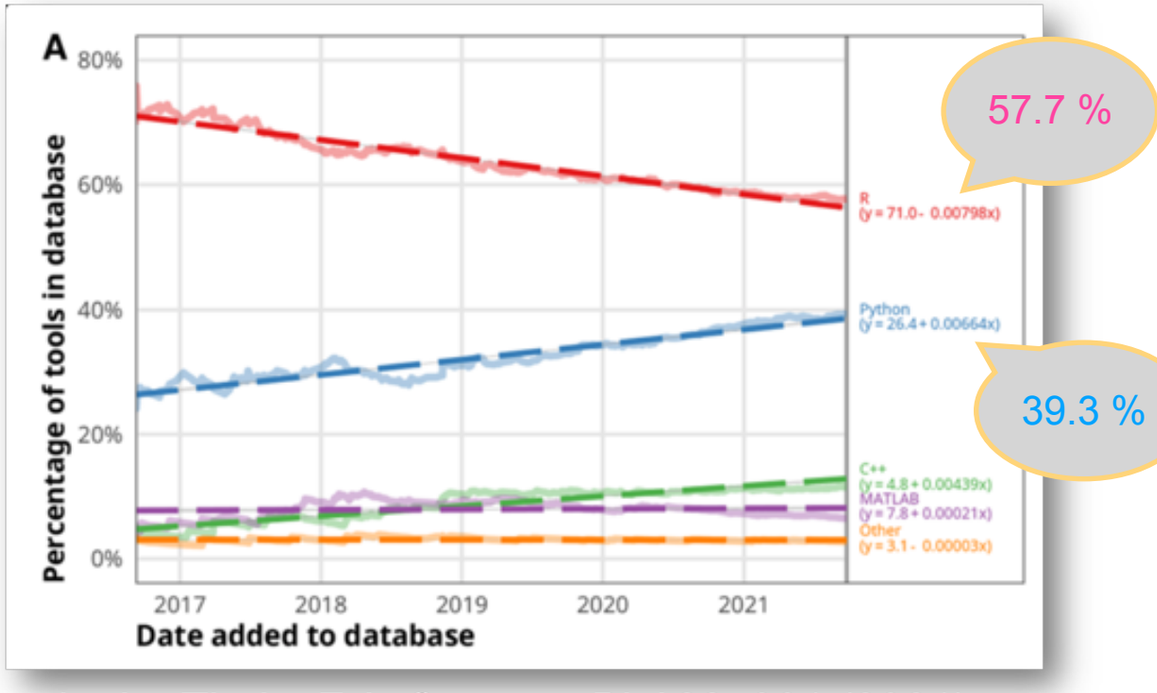
- 「ベストプラクティス」というわけではありません。発展的なツールの紹介もあって、中には有効性がまだじゅうぶんに検証されていないものもある。

- 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 Docker か、Google Colabのバージョンで実行。

Pythonによるシングルセル RNA-seq解析

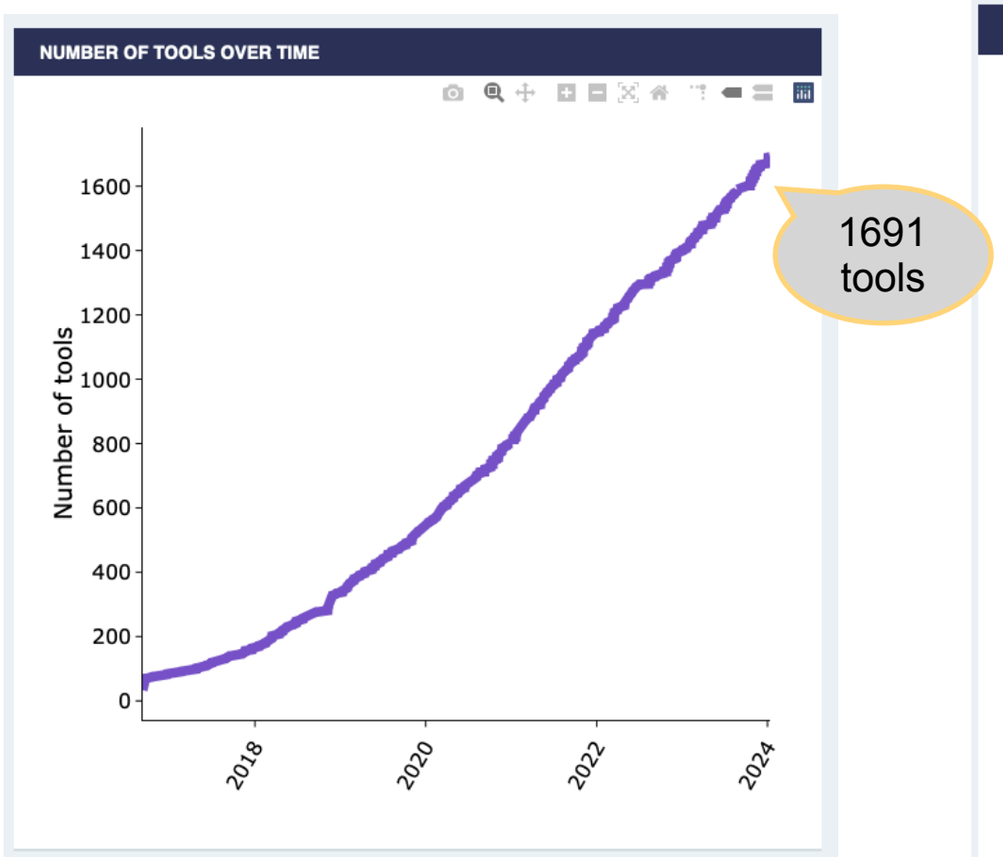


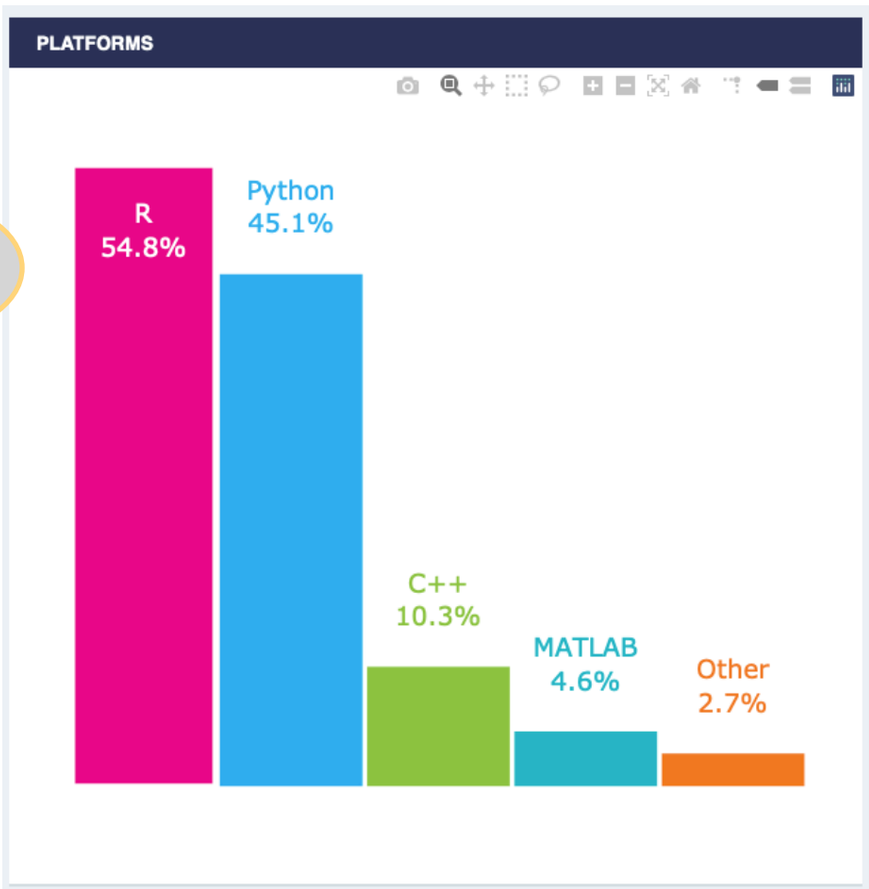
https://www.scrna-tools.org/



Zappia, L., Theis, F.J. *Genome Biol* **22**, 301 (2021).

2024年1月16日現在

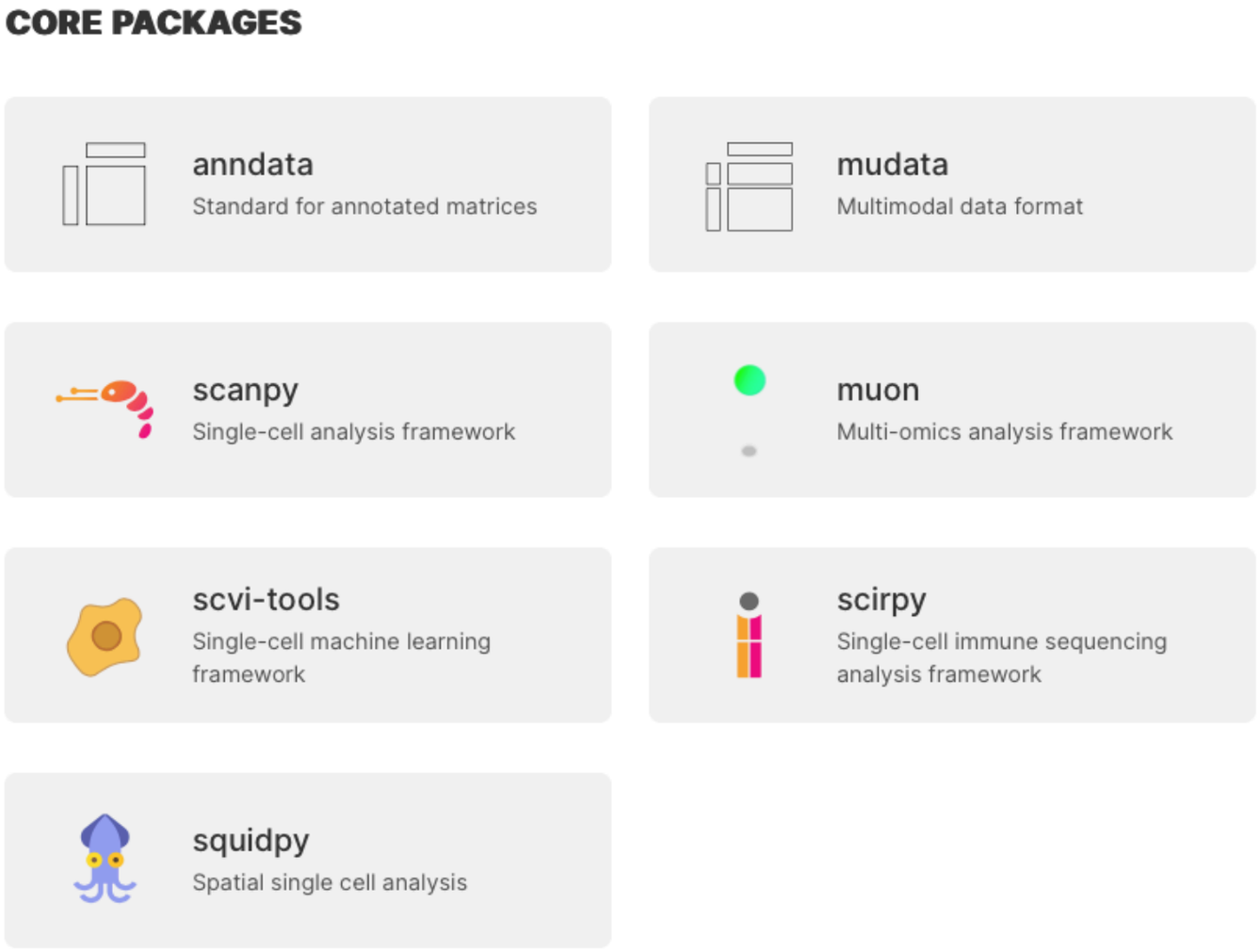




scverse　単一細胞オミクス解析に関連するPythonツールの開発・維持のためのコンソーシアム

https://scverse.org

Virshup I. et al. *Nature Biotechnology* **41**, 604 (2023)



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。

本日の講習

Cell ranger にて遺伝子毎のリードカウントを計算する

↓

前処理: 細胞、遺伝子のフィルタリング

↓

正規化

↓

特徴量選択（発現量の変動が大きい遺伝子）

↓

主成分分析 (PCA)

↓

次元削減 (UMAP, tSNE)

↓

Leiden クラスタリング

Scanpy

計算
効率のために PCA
を行ってから UMAP,
tSNE を行う。

scRNA-seq データの統合 (バッチ補正) scVI

↓

バッチ補正後の UMAP～ Leiden クラスタリング

Scanpy

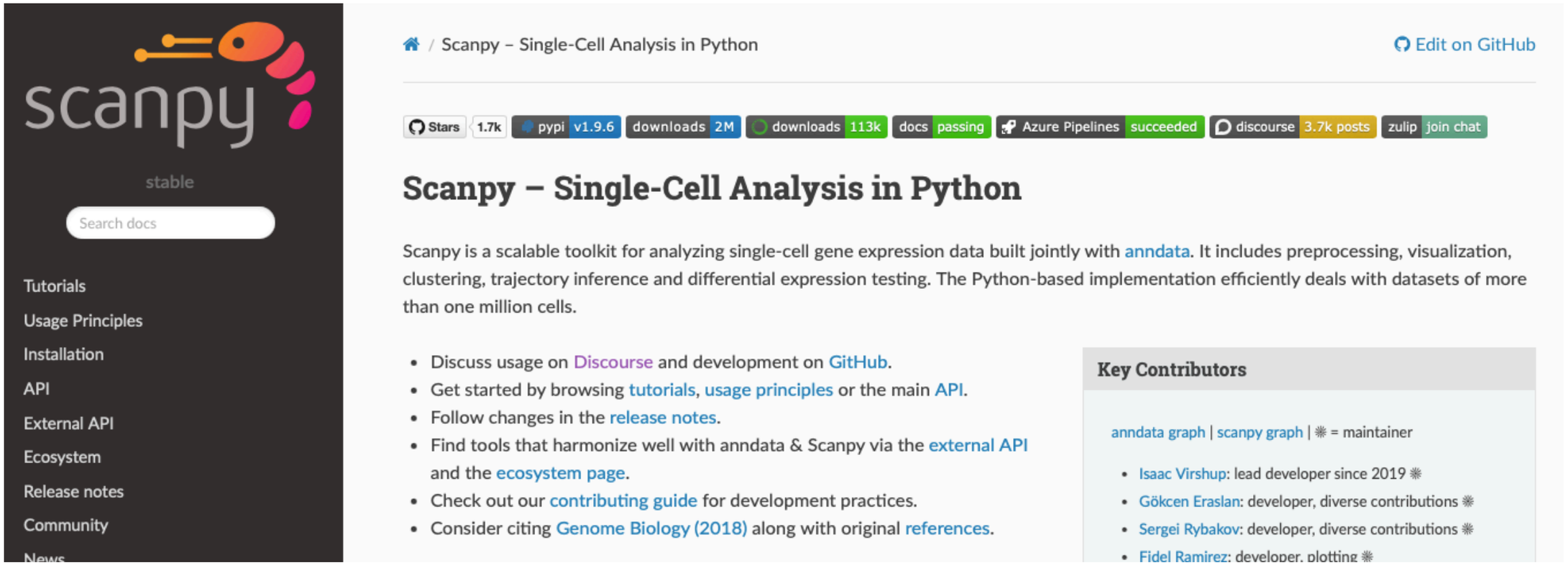
↓

Doublet 検出 Solo

↓

DEG解析 scVI

Scanpy: Pythonでシングルセル解析をする際のコアパッケージ



10x のデータの場合、リード処理と定量化を Cell Ranger で行い、その後の処理を Scanpy で行う。

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。

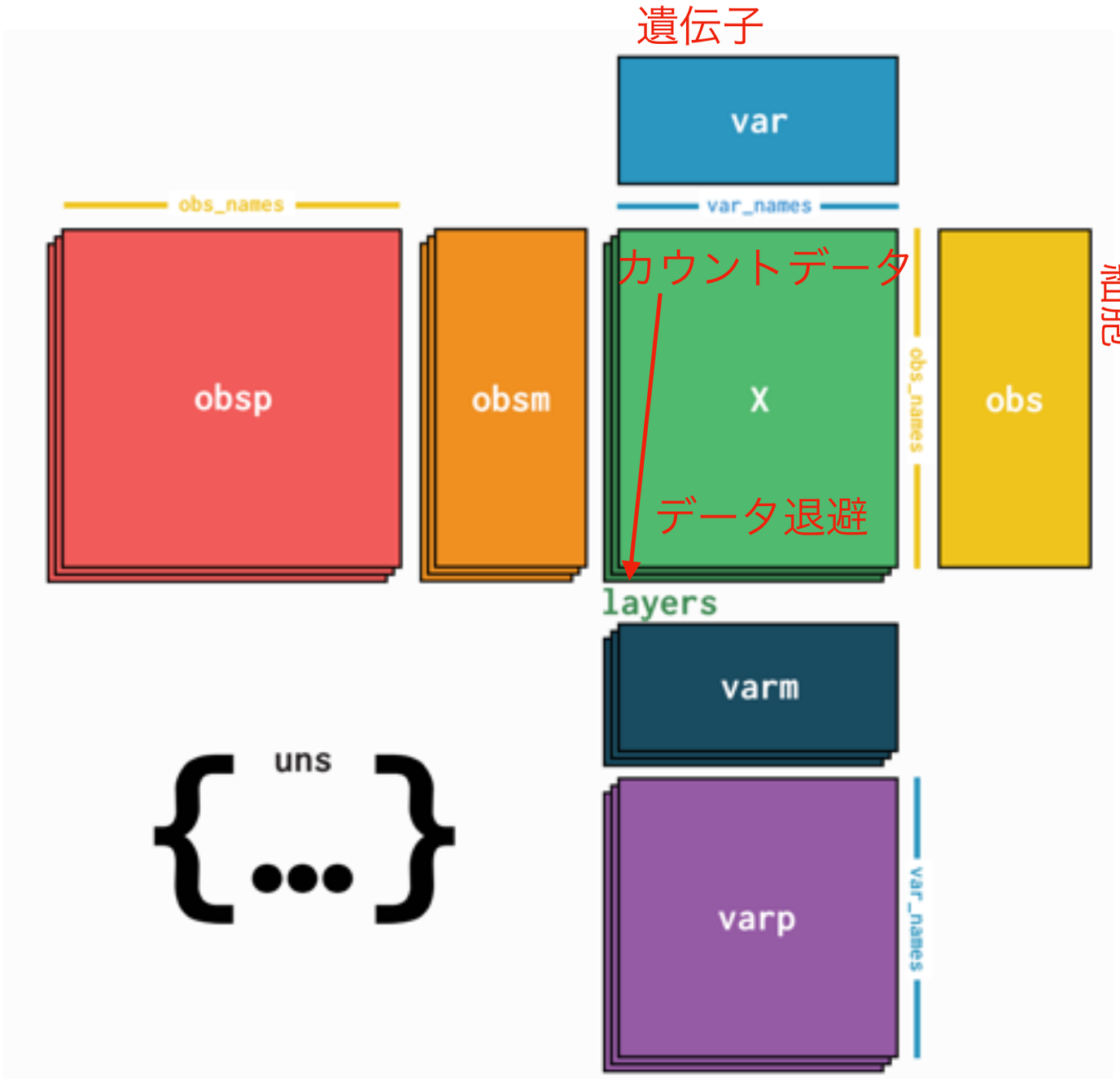
AnnData

(“Annotated Data” の略)

シングルセル解析のためのPythonパッケージの多くが、このオブジェクトを使用

PandasのDataFrameを拡張したデータ構造

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。



ひとつのオブジェクトに遺伝子発現量のデータ、サンプルや細胞のアノテーション、遺伝子の情報などをまとめて格納できる。

anndataを使うことで、実験の情報が詰まったひとつのオブジェクトに対して処理を次々に実行し、さらに処理結果をそこに蓄積していくことができる。

Virshup I. et al. *bioRxiv* 2021 Dec 19.

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。PCA で次元削減したデータもここに入る。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の関数

•scanpy.pp.XXX

前処理（preprocessing）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（tools）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（plotting）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-tools: 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

複雑な確率モデルを
ニューラルネット
ワークを使って表現

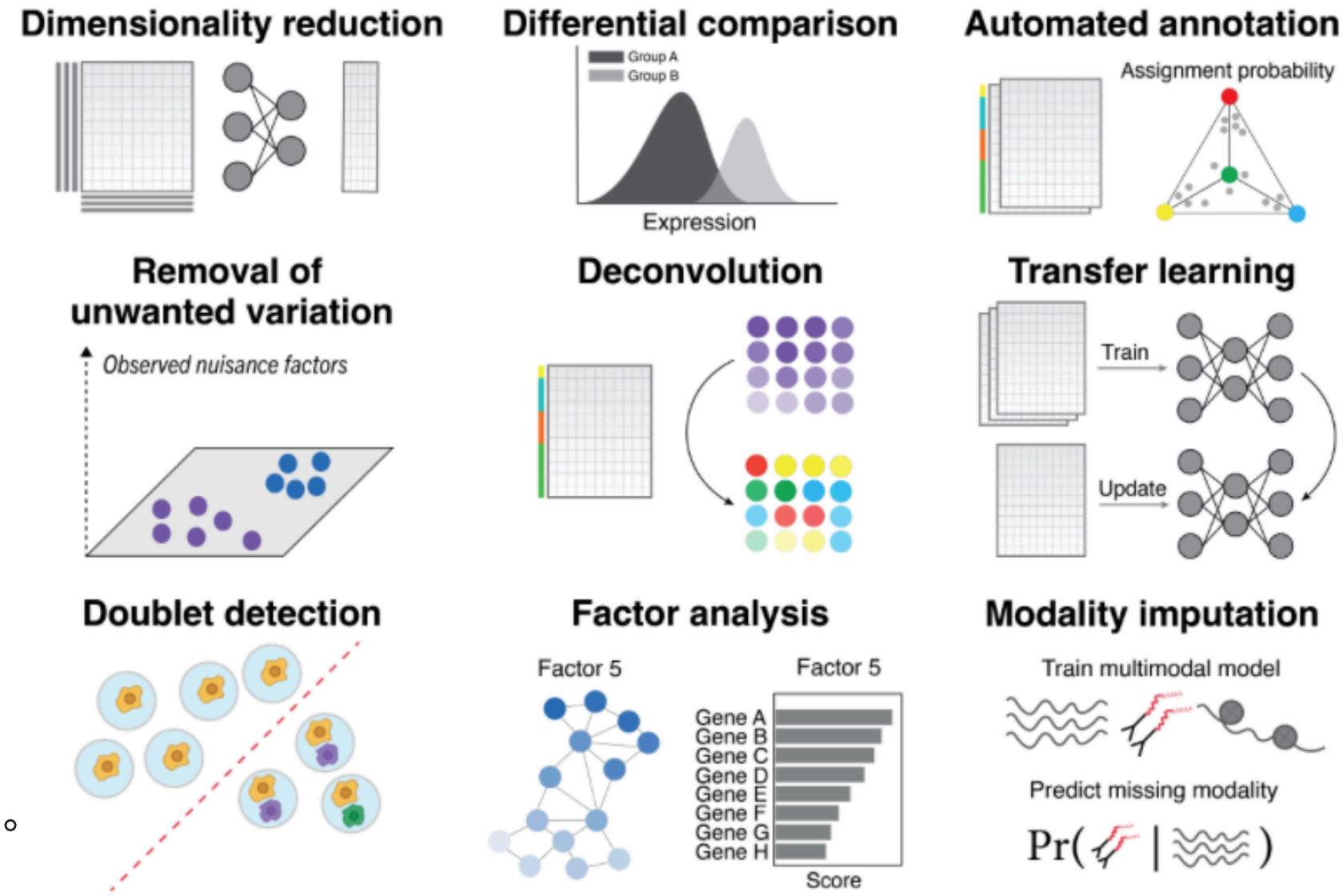
次元削減

正規化

Replicate のデータ統合

インプテーション

などの機能が実装されている。



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。

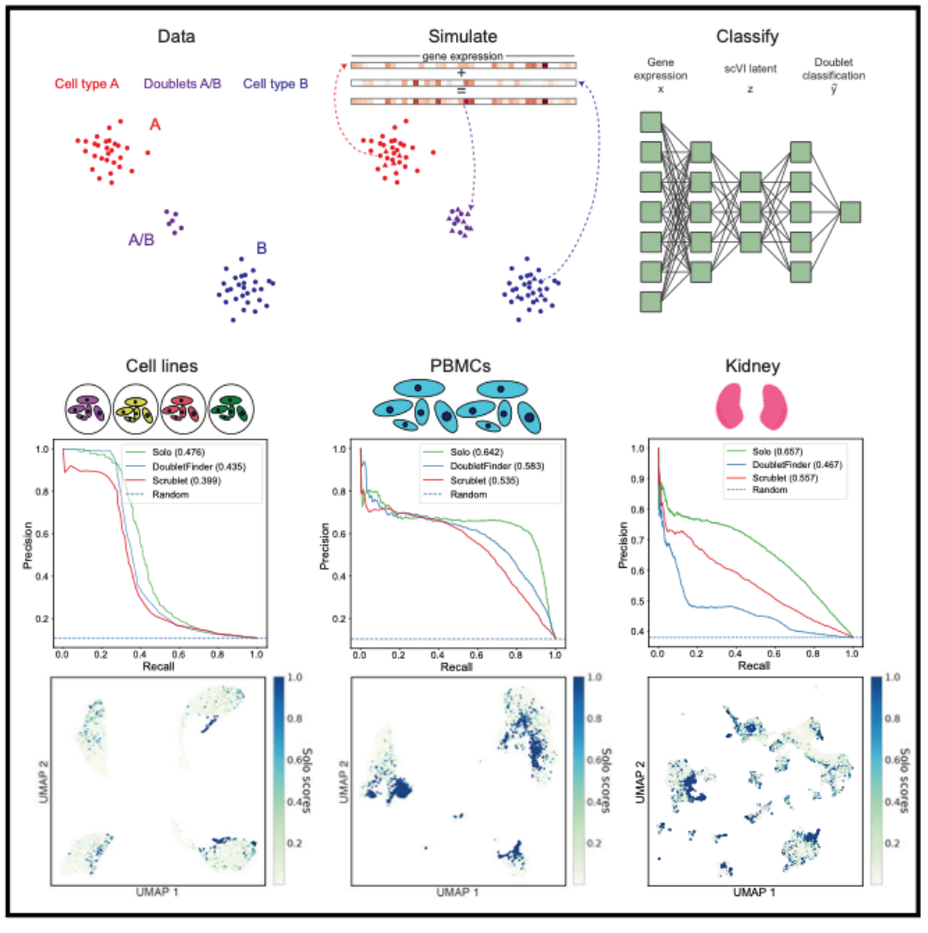
シミュレーションデータ（適当なふたつの細胞の平均発現パターン）でニューラルネットワークを学習してから、実際のデータのダブレットを予測する。

Brief Report

Cell Systems

Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning

Graphical Abstract



Authors

Nicholas J. Bernstein, Nicole L. Fong, Irene Lam, Margaret A. Roy, David G. Hendrickson, David R. Kelley

Correspondence

dgh@calicolabs.com (D.G.H.), drk@calicolabs.com (D.R.K.)

In Brief

Current single-cell RNA sequencing technologies occasionally allow multiple cells to be combined into a single profile, which challenges downstream analyses. Bernstein et al. introduce a semi-supervised deep learning method called Solo that identifies these “doublet” cells with greater accuracy than existing methods.

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