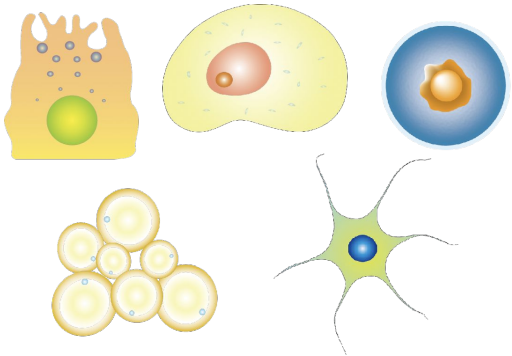
2024/10/05 トーゴーの日シンポジウム2024

細胞レベルの機能・表現型と遺伝子発現を関連付ける「Cell IO」データベースの開発

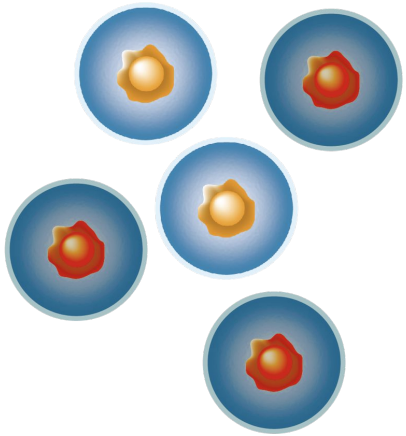
尾崎　遼
筑波大 医学医療系 バイオインフォマティクス研究室
筑波大 人工知能科学センター
理研BDR バイオインフォマティクス研究開発チーム

細胞レベルの機能・表現型は多様である

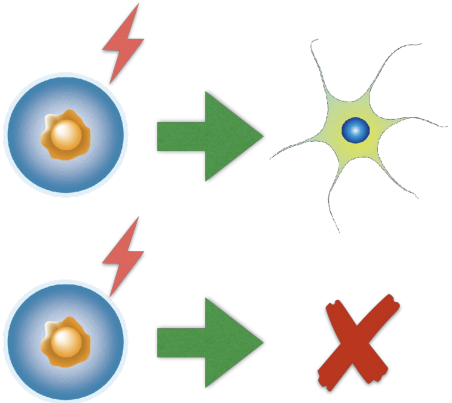
細胞型
Cell type



細胞状態
Cell state



分化ポテンシャル
Differentiation potential



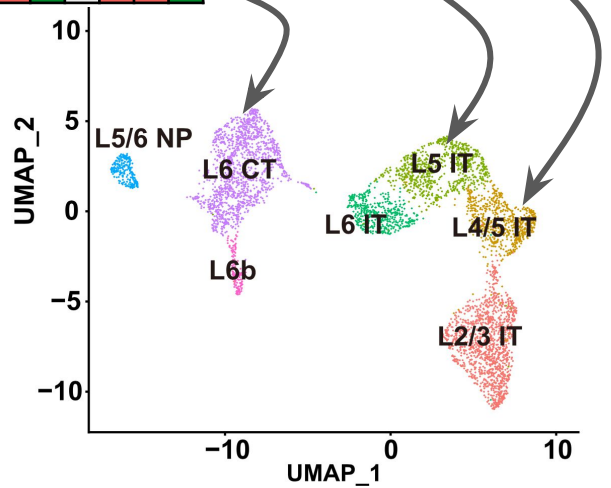
1細胞RNA-seq → 細胞の分類の細分化・再定義

単一の細胞

1細胞RNA-seq

遺伝子発現プロファイル

遺伝子 1
遺伝子 2
遺伝子 3
…



細胞地図
（細胞アトラス）

Nakata et al., *Neuroscience Research*, 2024

「細胞レベルの機能・表現型」の知識へのアクセス

1細胞(空間)トランスクリプトームデータ

・細胞型↔遺伝子発現
・細胞型↔遺伝子発現↔摂動
・細胞型↔遺伝子発現↔細胞機能・表現型

例:遺伝子ノックダウン、化合物

例:CD8陽性、炎症活性化、細胞サイズ

文献データ

・細胞型↔摂動
・細胞型↔細胞機能・表現型

1細胞RNA-seqベースの知識:アクセスしづらい

レポジトリ・2次DB	・遺伝子リスト等はみられる ・カタログ → 詳しくは自分で解析
データを報告した原著論文	・部分的な報告 ・データと紐づかない
データ解析	・遺伝子発現がわかっても どんな細胞かわからない ・GO解析でも解釈しづらい

文献ベースの知識:アクセスしづらい

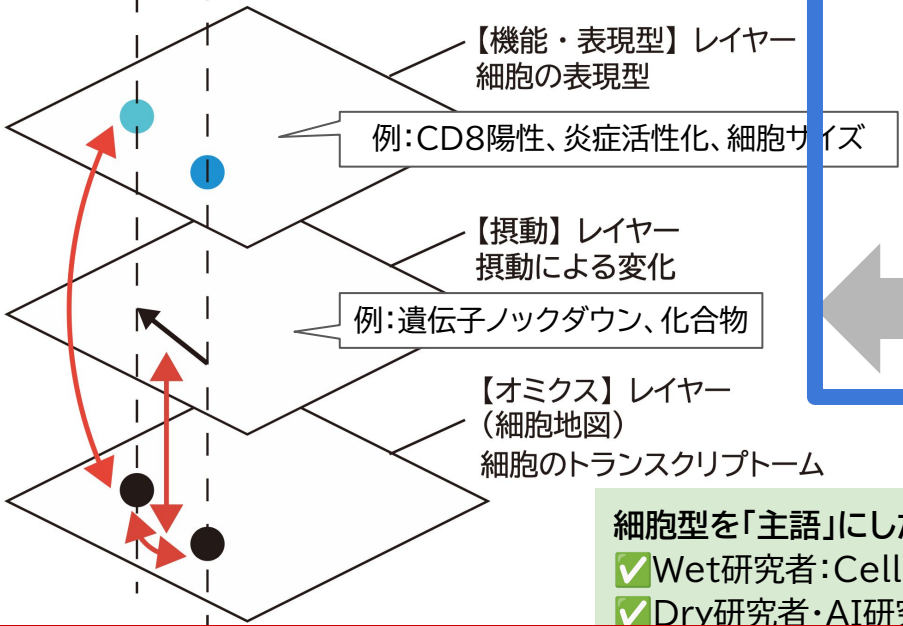
PubMed等	・細胞型名で検索しづらい ・MeSHが完璧ではない ・読み込むにはドメイン知識が必要
LLMサービス	・ハルシネーション ・ドメイン知識がないと合っているかわからない

Cell IO（Cell Input/Output）

1細胞RNA-seqデータ

専用データ処理パイプライン

・遺伝子発現
・遺伝子発現↔摂動
・遺伝子発現↔細胞機能・表現型



文献データ

LLMによる情報抽出

例：遺伝子ノックダウン、化合物

摂動 → 細胞型

細胞型 → 機能・表現型

例：CD8陽性、炎症活性化、細胞サイズ

細胞型名、解剖学的位置等によるマッピング

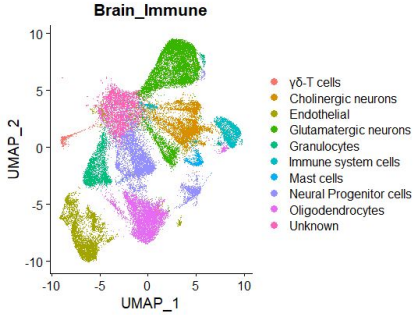
細胞型を「主語」にした生物学的知識の整理

✅Wet研究者：Cell IOで遺伝子と機能・表現型の探索を加速

✅Dry研究者・AI研究者：Cell IOで細胞ベースのDry・AI研究を加速

Cell IO：提供機能

・トランスクリプトームに基づく
細胞地図の表示



細胞のメタデータや遺伝子発現を表示
サブセットのダウンロードも可能

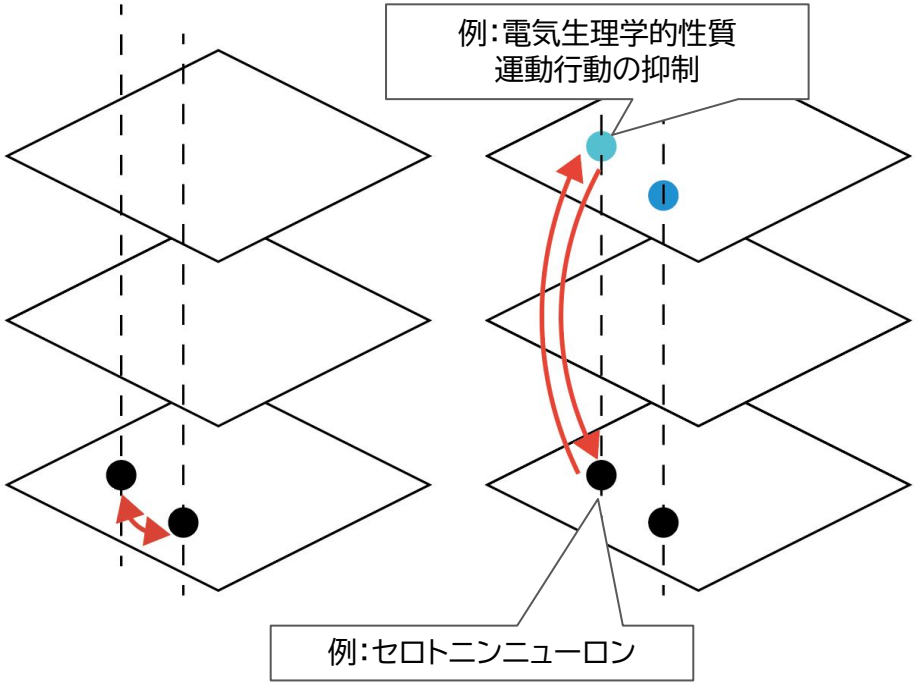
・文献に基づく
細胞レベルの機能・表現型

摂動 → 細胞型

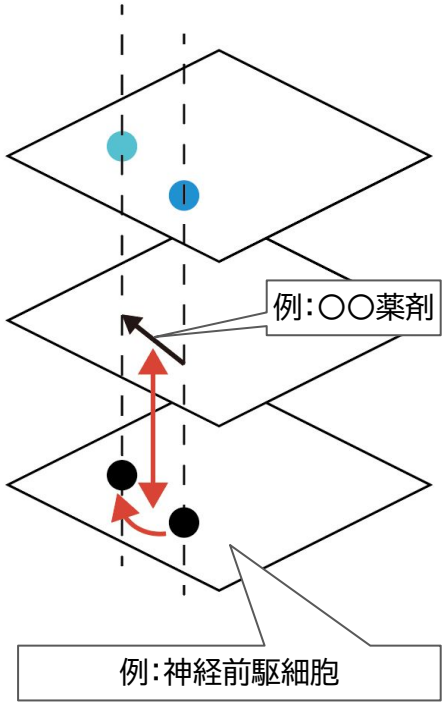
細胞型 → 機能・表現型

・細胞型名や遺伝子群による
類似細胞の検索

・機能・表現型による検索
・機能・表現型の予測



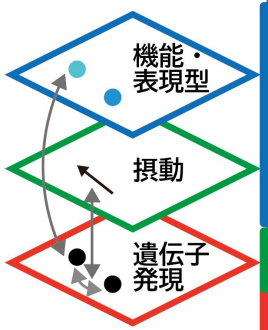
・2種類の遺伝子群による細胞変動
・摂動の予測



Cell IOデータベースに収載予定のデータの収集

初期段階でのスコープ：ヒト・マウス・ラット、免疫学および神経科学

1細胞（空間）トランスクリプトームデータ　遺伝子発現、細胞レベルの機能・表現型の情報



解像度	摂動	細胞の機能・表現型	主な計測手法
1細胞	-	細胞の形態等	Quartz-seq2, RamDA-seq（FACS-based）
1細胞	-	細胞表面抗原	CITE-seq, scGR-seq
1細胞	-	電気生理学的性質	Patch-seq
1細胞	-	細胞の空間座標	MERFISH, 10x Xenium（※1）
1細胞	CRISPR、化合物	-	Perturb-seq, scifi-RNA-seq
1細胞	-	-	10x Chromium, Smart-seq

※1　空間トランスクリプトームデータにおける、各細胞の空間座標も細胞の表現型として取り扱いできる。

文献ベースの細胞型の機能・表現型情報　細胞型 - 機能・表現型 - 目的型

摂動 → 細胞型

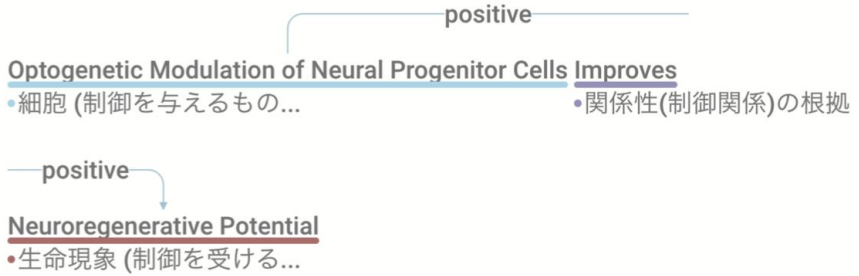
細胞型 → 機能・表現型

- PubMed Central等の文献から抽出した機能・表現型のグラフ
  - 免疫学・神経科学の細胞関連 かつ PMCのテキストマイニング可能なサブセット
    - 免疫学：約15万報（全種類）、約2.2万報（総説）
    - 神経科学：約11万報（全種類）、約1.2万報（総説）

LLMによる細胞レベルの機能・表現型の文献からの抽出

- POC：神経科学分野で摂動と細胞型に関連した論文の要旨
- 精度検証用アノテーションセットの構築：専門家5名・100件
- グラフ：細胞＋関係性（positve, neutral, negative）＋生命現象

■細胞型の機能・表現型のアノテーションの例

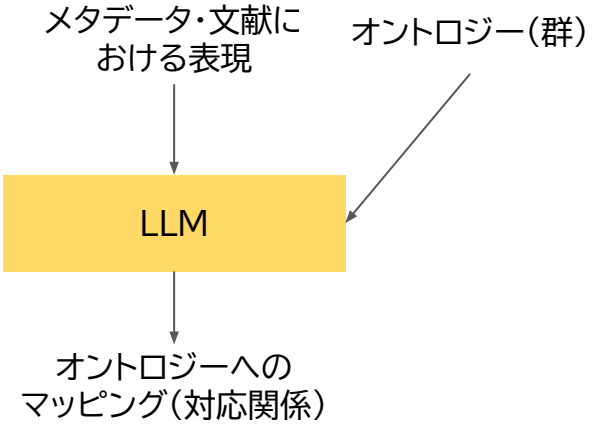


■抽出された細胞型の機能・表現型

PMID: 28193320	細胞	関係性の根拠	生命現象
Answer (annotator)	dorsal raphe nucleus (DRN) 5-HT neurons	phasic optogenetic activation	suppression of spontaneous locomotor behavior
Zero-shot	5-HT neurons	phasic optogenetic activation	suppression of spontaneous locomotor behavior
One-shot	5-HT neurons	phasic optogenetic activation	suppression of spontaneous locomotor behavior
Two-shot	5-HT neurons	phasic optogenetic activation	suppression of spontaneous locomotor behavior
Three-shot	activating dorsal raphe nucleus (DRN) 5-HT neurons	induced	suppression of spontaneous locomotor behavior

細胞の呼称、機能・表現型、サンプルメタデータの名寄せに係るツールの開発

細胞型の名寄せに係るツールの開発

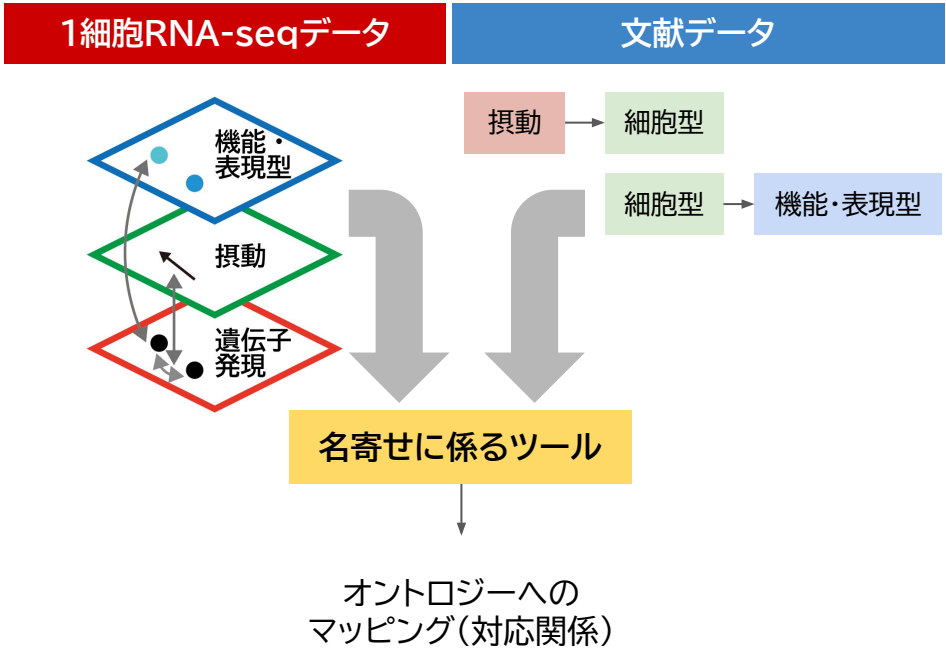


細胞型の表現：Cell Ontology、Cell Line Ontology、Uberon

摂動の表現：Experimental Factor Ontology、MONDO

表現型の表現：HPO、MPO、GOなど

細胞型の名寄せに係るツールの運用



利用者像 (1/2)

ウェット研究者

特徴
- オミクス解析の馴染が薄い
- データ解析できる経験がない、人材がいない

課題
- データ解析がハードル

Cell IOの機能
- 遺伝子からの検索
- 機能・表現型からの探索

↓

Cell IOがもたらす効果
- 人材、研究資源が乏しい研究室での研究を加速

✅Cell IOで遺伝子と機能・表現型の探索を加速

- 例1：遺伝子の探索
  - 注目する、細胞レベルの機能・表現型がある
  - → Cell IOでどんな遺伝子を発現しているかわかる
  - → 遺伝子に着目した研究へ
- 例2：機能・表現型の探索
  - 注目する遺伝子、細胞型がある
  - → Cell IOでこんな機能・表現型を示すかわかる
  - → 機能解析のための仮説導出へ

例： 膵臓がんの研究者
リスク遺伝子 → T細胞サブタイプ → 表現型X

それぞれの立場の利用者のボトルネックを解消することで研究を加速

利用者像 (2/2)

データ解析者

特徴
- いろんなプロジェクトのデータ解析に従事
- ドメイン知識が比較的浅い

課題
- 文献検索がハードル

Cell IOの機能
- 遺伝子発現→細胞型名→細胞機能・表現型を検索

↓

Cell IOがもたらす効果
- データの解釈と深い解析を高速に回せる

AI研究者

特徴
- データセットがあればAI研究をしたい
- データキュレーションに興味はない

課題
- データ入手がハードル

Cell IOの機能
- 遺伝子発現と細胞機能・表現型の学習データを得られる

↓

Cell IOがもたらす効果
- AI研究を加速

✅Cell IOで細胞ベースのDry・AI研究を加速

- 例1：1細胞RNA-seqデータ解析結果の解釈
  - 条件間で変動する細胞サブタイプがある
  - → Cell IOで、遺伝子発現から機能・表現型がわかる
  - → GO解析等では到達が難しい解釈へ
- 例2：細胞制御AIのためのデータセット提供
  - 所望の細胞表現型を示す細胞を作るための摂動を生成したい
  - → Cell IOで細胞の遺伝子発現とラベルの紐づいたデータセットを得る
  - → AIモデル開発が捗る

例：疾患Xのモデルマウスでマイクログリアのサブタイプが増加
→ Cell IO で慢性陣痛を活性化 → 疾患Xの病態と関連？

それぞれの立場の利用者のボトルネックを解消することで研究を加速

Acknowledgements

We thank Shinya Nakata, Kazuya Miyanishi, Ami Kaneko, Yoshihiko Sakaguchi, and Haruto Ijiri for text annotation.

We also thank Ryota Yamada (fuku Inc.) for designing, implementing, and evaluating the text mining system.

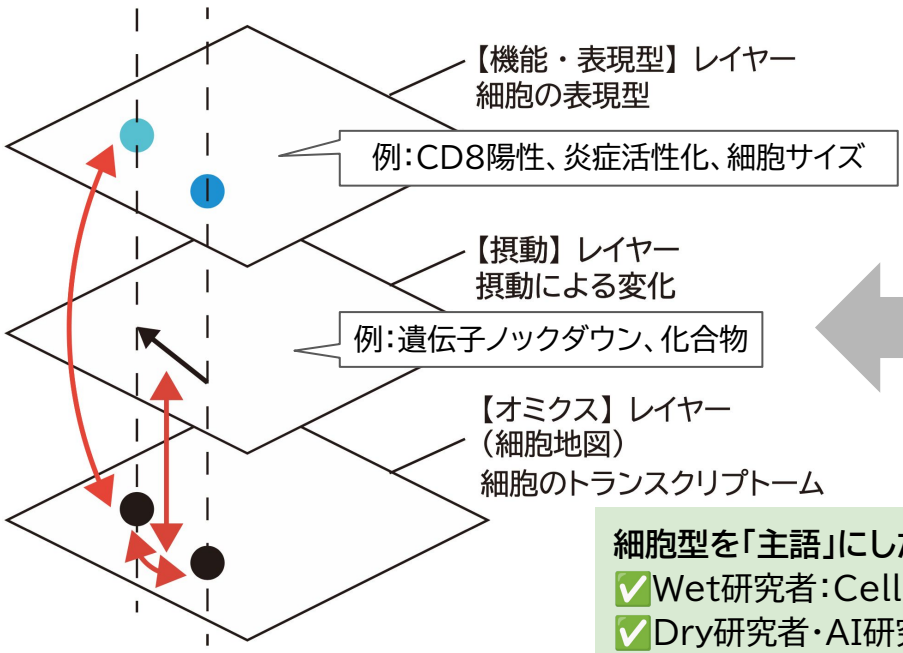
This work is supported by JST-NBDC DICP (JPMJND2402).

Cell IO（Cell Input/Output）

1細胞RNA-seqデータ

専用データ処理パイプライン

・遺伝子発現
・遺伝子発現↔摂動
・遺伝子発現↔細胞機能・表現型



文献データ

LLMによる情報抽出

例：遺伝子ノックダウン、化合物

摂動 → 細胞型

細胞型 → 機能・表現型

例：CD8陽性、炎症活性化、細胞サイズ

細胞型名、解剖学的位置等によるマッピング

細胞型を「主語」にした生物学的知識の整理

✅Wet研究者：Cell IOで遺伝子と機能・表現型の探索を加速

✅Dry研究者・AI研究者：Cell IOで細胞ベースのDry・AI研究を加速