

2025 年度年次報告書

ライフサイエンスデータベース統合推進事業 統合化推進プログラム

【研究課題名】「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」

研究代表者：

尾崎 遼（理化学研究所 生命機能科学研究センター チームディレクター）

研究分担者：

尾崎 遼（筑波大学 医学医療系 生命医科学域 客員准教授）



©2025 尾崎 遼(筑波大学) licensed under CC表示4.0国際

1. 研究進捗状況と成果

【研究開発項目 A】 トランスクリプトーム計測ベースの細胞機能・表現型に係る公開用データの作成

NCBI Short Read Archive に収載された神経科学および免疫学分野の Chromium 系 1 細胞 RNA-seq データを収集した。トランスクリプトームデータ処理パイプラインにより、遺伝子発現量行列、マーカー遺伝子リスト、細胞型アノテーションラベル、次元圧縮結果等を得た。さらに、メタデータ抽出パイプラインにより、サンプルの表現型や実験条件についてのラベルデータを整理した。これらにより、 β 版 DB 公開用データセットとして、約 3,500 実験（約 1,190 万細胞）、741 細胞型、3 生物種（ヒト、マウス、ラット）、42 組織種について整備した。加えて、超大規模 scRNA-seq データの急増に対応するため、既設サーバーに GPU を増設し、パイプラインの GPU 化による処理時間の短縮効果を評価するとともに、シングルセル基盤モデルを用いたリファレンスマッピングによる細胞型アノテーションの性能検証を行った。

【研究開発項目 B】 文献ベースの細胞機能・表現型に係る公開用データの作成

文献から細胞機能・表現型を抽出する大規模言語モデルベースのパイプラインを構築・改良した。このパイプラインでは、細胞型を「主語」、細胞の機能・表現型を「述語」「目的語」とした組を抽出する。これを PubMed Central の Open Access Subset に収載された神経科学および免疫学分野の文献に適用し、 β 版 DB 公開用データセット（約 134,100 組）を構築した。

【研究開発項目 C】 細胞型の名寄せに係るツールの開発

細胞型名、由来サンプル名等を Cell Ontology 等のオントロジーにマッピングする名寄せツールを開発・改良した。トランスクリプトーム計測ベースのデータでは、細胞型は Cell Ontology をあらかじめ紐付けたリファレンスデータでアノテーションし、メタデータは機械学習ベースのオントロジーマッピングで名寄せした。文献ベースのデータはより記述自由度が高いため、汎用 LLM を使用して様々なオントロジーへマッピングしたのち、LLM-as-a-judge にて出力された信頼度によってマッピングを最終決定するパイプラインとした。これらのツールを、研究開発項目 A および B で作成した β 版 DB 公開用データセットに適用し、細胞型名の名寄せを完了した。

【研究開発項目 D】 データ公開システムの開発

β 版 DB を 2026 年 3 月末に評価用として公開した。細胞型名やオントロジー ID での細胞検索機能、キーワードによる文献ベースの細胞機能情報の検索機能のほか、各サンプルでの細胞型構成比、次元圧縮の結果なども確認できる。

2. 代表的な成果事例

1. 尾崎 遼，研究自動化による、遺伝子発現データと文献データからの細胞レベルの情報統合，シンポジウム「オープンライフサイエンスの新時代」，第 48 回日本分子生物学会年会，2025-12-05，国内，招待，口頭。

概要：細胞レベルの機能・表現型と遺伝子発現を関連付けるウェブデータベースの開発についてアピールするとともに、その中でのワークフロー言語を用いたデータ解析の自動化や、大規模言語モデル（LLM）によるテキストマイニングに基づく文献知識抽出の自動化等の研究自動化事例を紹介した。

以上