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）
研究開発実施報告書　様式

2024 年度 研究開発実施報告

概要

研究開発課題名	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
開発対象データベースの名称（URL）	Cell IO (https://未定)
研究代表者氏名	尾崎 遼　（10743346）
所属・役職	筑波大学 医学医療系 准教授（2025 年 3 月時点）

□目次

§1. 研究実施体制

グループ名	研究代表者・研究分担者氏名	所属機関・役職名	研究題目
尾崎グループ	尾崎 遥	筑波大学・准教授	Cell IOデータベースの開発

§2. 研究開発対象とするデータベース・ツール等

(1) データベース一覧

【主なデータベース】

No.	名称	別称・略称	URL
1	Cell IO		未定

【その他のデータベース】

No.	名称	別称・略称	URL
1			

(2) ツール等一覧

No.	名称	別称・略称	URL
1			

§3. 実施内容

(1) 本年度に計画されていた研究開発項目・タスク

【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

【項目 A-1】パイプラインの構築

- 収集するデータの実験手法に応じたデータ処理方法の選定
- 1次解析のためのデータ処理パイプラインの構築
- データセットの作成者が利用するレポジトリとファイル形式に基づいたデータ収集方法の洗い出し
- 細胞表面抗原とトランスクリプトームの同時計測データセットを収集
- 異なる実験手法由来のデータセットを複数のレポジトリから選定と・選定したデータセットの収集と処理の実施
- 収集・処理方法の最適化を行い、データセットを効率的に収集する方法の洗い出し

【項目 A-2】公開用データの作成

- 項目 A-1 で構築されたパイプラインの Gene Expression Omnibus などのデータベースへの適用
- パイプラインにて処理後のデータの評価

【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

【項目 B-1】パイプラインの構築

- 文献からの細胞機能・表現型抽出パイプラインの評価データの作成
- パイプラインの構築
- パイプラインの評価

【項目 C】細胞型の名寄せに係るツールの開発

【項目 C-1】細胞型の名寄せに係るツールの開発

- 評価用データセットの作成
- 名寄せツールの構築
- 名寄せツールの評価

(2) 進捗状況

【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

【項目 A-1】パイプラインの構築

本年度は、最も多くのサンプルが公開されている 10x Chromium 手法からパイロットとして取り組みを開始し、合計 22,985 サンプルの収集と処理を実施した。パイプラインの構築に際しては、これまで人手によって個々のデータセットごとに設定されていた閾値やパラメータを自動化することに成功した。特にデータ品質コントロール（QC）ステップおよび細胞型アノテーションステップについては、完全に自動化を達成し（図1）、これにより従来必要とされていた主観的な判断の排除に成功し、処理の安定性が向上した。また、この自動化により、処理時間および研究者の労力が大幅に削減され、今後さらに多様な手法由来のデータへの適用拡張が容易になった。

当初計画では複数の実験手法由来のデータを対象としてパイプライン構築を行う予定であったが、10x Chromium データの収集と処理を先行して集中実施したため、他の手法への適用は次年度以降に後ろ倒しとなった。

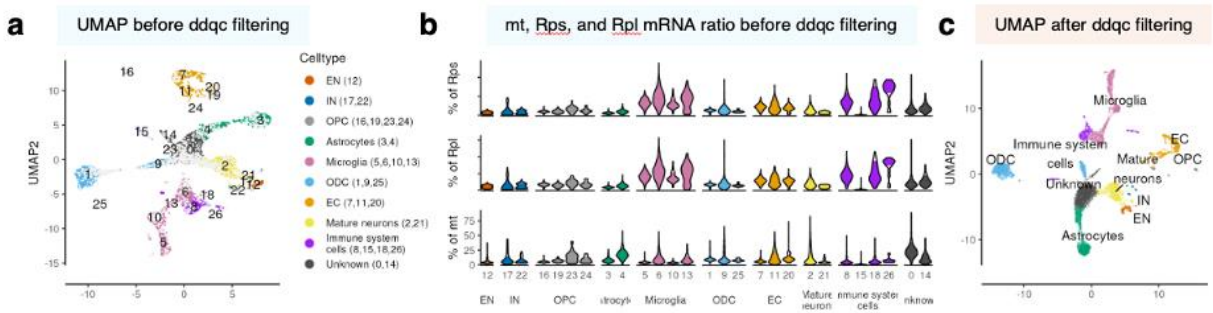

図１　１細胞 RNA-seq のデータ QC ステップおよび細胞型アノテーションステップの自動化

【項目 A-2】公開用データの作成

構築したパイプラインを Gene Expression Omnibus（GEO）をはじめとする公開データベースから取得したデータに適用し、公開用データの作成を進めた。しかしながら、収集したデータセットのメタデータの欠如、特にサンプルの性別などの重要な情報が記載されていないケースが多く認められ、原著論文を直接参照しなければならないなどの問題が明らかになった。そのため、これらのメタデータの抽出パイプラインの改善を次年度に引き続き取り組む。

【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

【項目 B-1】パイプラインの構築

文献ベースの細胞機能・表現型抽出パイプラインの評価用データセットの構築においては、専門家（Wet 系研究者）計 10 名（神経科学 5 名、免疫学 5 名）の協力を得てマニュアルキュレーションを実施した。研究者一人当たり 200 件の文献アブストラクトを対象に、細胞機能・表現型に関する記載がある文献をそれぞれ 20 件抽出し、詳細なアノテーション（細胞型、細胞型が影響を与える生命現象、影響の種類、影響の判断根拠を示す文章表現に対するアノテーション）を行った。最終的に神経科学分野 90 件、免疫学分野 103 件の計 193 件の文献において評価データセットを整備した。

この評価データを用いて GPT-4.1、Claude、Gemini、GPT-o3 の4種類の言語モデル、3 種類のプロンプトにて抽出パイプラインの性能評価を行った（図2）。細胞型（cell_score）、生命現象スコア（phenomenon_score）、影響の種類（type_score）それぞれの一致度スコアにおいて安定した高性能を示した。影響の判断根拠の一致度スコア（basis_score）については、言語モデル間で差異が見られ、GPT-4.1 を除く他の GPT モデルは比較的高精度を維持したものの、プロンプト変更による著しい精度改善は確認されなかった。これらの結果を踏まえ、次年度は特に影響の判断根拠の抽出性能の改善に取り組む。

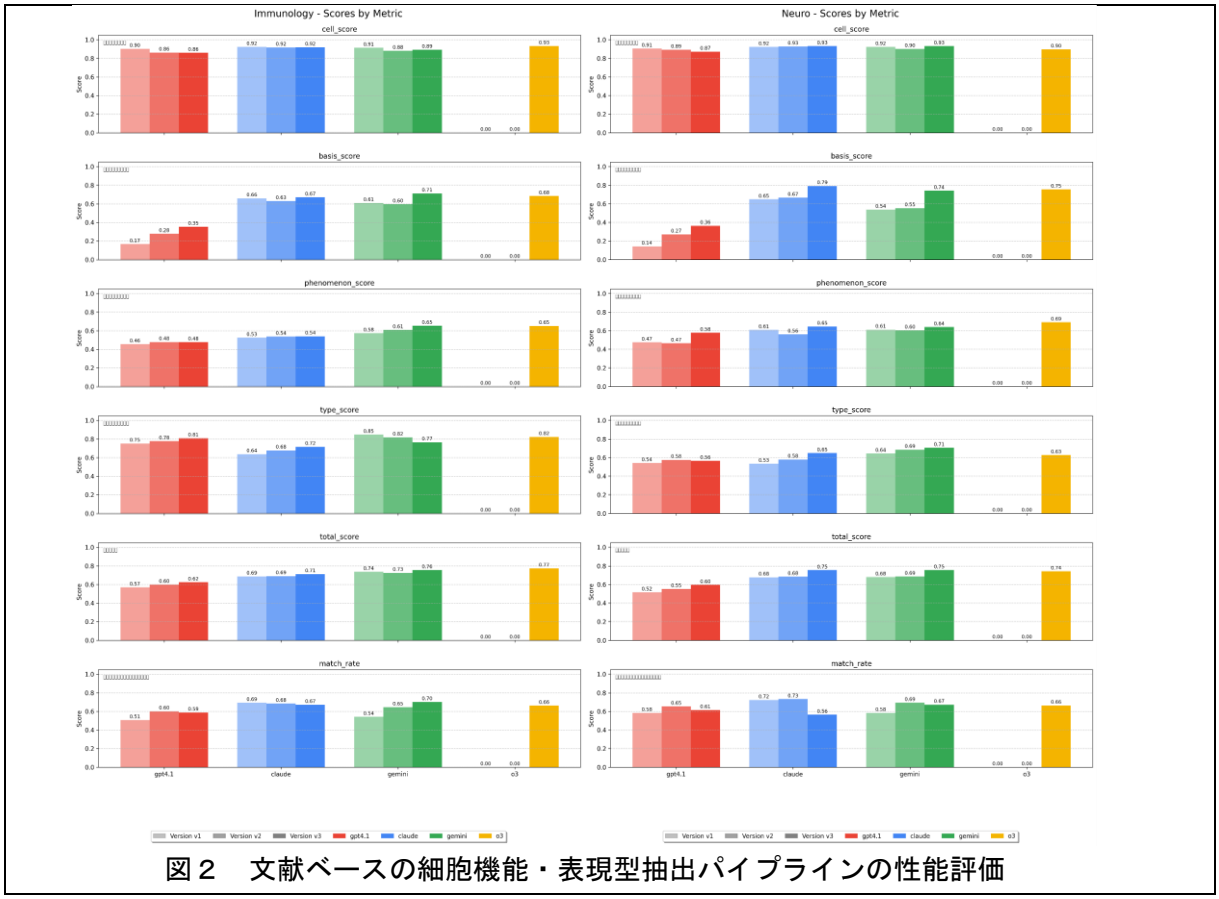

図２　文献ベースの細胞機能・表現型抽出パイプラインの性能評価

【項目 C】細胞型の名寄せに係るツールの開発

【項目 C-1】細胞型名寄せツールの開発

細胞型名寄せツールの評価については、専門家のマニュアルキュレーションに基づいて 26 件の細胞型名および由来サンプル名の評価用データセットを作成し、精度約 80%という実用レベルの性能を達成した。当初計画ではさらに多くのサンプルを評価する予定であったが、ツール構築および評価に想定よりも時間を要したため、評価対象数を絞って実施した。次年度はサンプル数を増やし、ツールの更なる精度向上を目指す。

§4. 成果発表等

(1) 原著論文発表

① 論文数概要

種別	国内外	件数
発行済論文	国内(和文)	0 件
	国際(欧文)	0 件
未発行論文 (accepted, in press 等)	国内(和文)	0 件
	国際(欧文)	0 件

② 論文詳細情報

なし

(2) その他の著作物(総説、書籍など)

なし

(3) 国際学会および国内学会発表

① 概要

種別	国内外	件数
招待講演	国内	1 件
	国際	0 件
口頭発表	国内	0 件
	国際	1 件
ポスター発表	国内	0 件
	国際	1 件

② 招待講演

〈国内〉

1. 尾崎 遼、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、トーゴーの日シンポジウム 2024、東京, 港, 品川ザ・グランドホール、2024-10-05

③ 口頭講演

〈国際〉

1. Haruka Ozaki, Ryota Yamada, Shinya Nakata, Kazuya Miyanishi, Ami Kaneko, Haruto Ijiri, Yoshihiko Sakaguchi, Extraction of cellular function knowledge from literature using large language models, 1st Asia & Pacific Bioinformatics Joint Conference (APBJC 2024), NAHA CULTURAL ARTS THEATER NAHArt, Naha, Okinawa, Japan, 2024-10-25,

④ ポスター発表

〈国際〉

1. Haruka Ozaki, Ryota Yamada, Shinya Nakata, Kazuya Miyanishi, Ami Kaneko, Yoshihiko Sakaguchi, Haruto Ijiri, Extraction of cellular function knowledge from literature using large language models, 1st Asia & Pacific Bioinformatics Joint Conference (APBJC 2024), NAHA CULTURAL ARTS THEATER NAHArt, Naha, Okinawa, Japan, 2024-10-25

(4) 知的財産権の出願 （国内の出願件数のみ公開）

① 出願件数

種別		件数
特許出願	国内	0 件

② 一覧

1) 国内出願

なし

(5) 受賞・報道等

① 受賞

なし

② メディア報道

なし

③ その他の成果発表

なし

§5. 主要なデータベースの利活用状況

1. アクセス数

① 実績

表 1 研究開発対象の主要なデータベースの利用状況

名称	種別	2024 年度(月間平均値)
Cell IO(公開前)	訪問者数	
	訪問数	
	ページ数	

② 分析

2. データベースの利用状況を示すアクセス数以外の指標

なし。

3. データベースの利活用により得られた研究成果(生命科学研究への波及効果)

なし。

4. データベースの利活用によりもたらされた産業への波及効果や科学技術のイノベーション(産業や科学技術への波及効果)

なし。

§6. 研究開発期間中に主催した活動（ワークショップ等）

(1) 進捗ミーティング

年月日	名称	場所	参加人数	目的・概要
2024 年 4 月 1 日～202 5 年 3 月 31 日 （隔週開催）	チーム内ミーティング （非公開）	筑波大学　健康医科学イノベーション棟（ハイブリッド）	7 人	研究進捗報告のためのミーティング

(2) 主催したワークショップ、シンポジウム、アウトリーチ活動等

なし。

以上

別紙１　既公開のデータベース・ウェブツール等

No.	正式名称	別称・略称	概要	URL	公開日	状態	分類	関連論文
1								