【報告様式Ａ】

課題管理番号: 23bm0404073h0003　作成／更新日:令和 6 年 5 月 28 日

日本医療研究開発機構
再生・細胞医療・遺伝子治療実現加速化プログラム事業
事後評価報告書

公開

Ⅰ 基本情報

研究開発課題名：　（日本語）高出力マルチオミクスによる細胞特性計測の深化
（英　語）Development of Massively Multi-Sample Omics Analysis

研究開発実施期間：令和/平成　年 4 月 1 日～令和　年 3 月 31 日(予定)

研究開発代表者　氏名：(日本語) 二階堂愛
（英　語）Itoshi Nikaido

研究開発代表者　所属機関・部署・役職：
（日本語）東京医科歯科大学難治疾患研究所バイオデータ科学部門ゲノム機能情報分野・教授
（英　語）Department of Functional Genome Informatics, Division of Biological Data Science, Medical Research Institute. Tokyo Medical and Dental University (TMDU)

II 研究開発の概要

【研究開発の成果】

本課題では、ヒト由来多能性幹細胞(hiPS 細胞)やその分化細胞の細胞機能と遺伝的多型を同時に数千検体から調べる手法を開発した。さらに数百検体から細胞機能と遺伝的多型の不均一性を 1 細胞計測する手法を開発した。この技術を用いて本プログラム内外の研究グループと連携し、培地環境や化合物、細胞ロット・株が hiPS 細胞や分化細胞の特性に与える影響を解析する。この結果を臨床研究や治験にフィードバックし、細胞・遺伝子治療の早期実現を加速への貢献を目指した。

細胞・遺伝子治療に用いられる細胞や hiPS 細胞、その分化細胞は、ロットや株ごとに細胞特性が異なり、移植の安全性・有効性に影響を与える。細胞特性に影響する因子は、培地や容器、継代数、接着因子、遺伝的背景などがあり、その組み合わせは膨大で調べ尽せない。

我々は、これまで細胞機能を数千検体の RNA シーケンス法によって計測する技術を開発した。一方、細胞の安全性や有効性の評価にとって遺伝子変異の計測も重要である。しかし、依存の手法は細胞集団に含まれるわずかな細胞を調べられず、スループットが低くコストも非常に高い。またこれらの手法は細胞機能の情報は得られないため、多型が細胞特性へどのように影響するか評価できない。

そこで本課題では、細胞機能と遺伝的多型の多検体同時計測法の開発を行った。我々が開発した世界最高性能の 1 細胞/多検体 RNA-seq 法 Quartz-Seq2 をベースとし、RNA 全長をシーケンスできるようにし遺伝子構造変化を捉えられるようにした。そのために RNA 全長ライブラリ作製法や長鎖 DNA シーケンサーへの対応を行った。さらに処理できる検体数を向上させるため、Quartz-Seq2 法をマイクロ流体デバイスや複数検体の混合反応法に対応させた。これらのデータより遺伝子機能と遺伝的多型情報を引き出す人工知能技術を開発した。これらの技術をプロジェクト内外の研究チームと連携し、細胞・遺伝子治療の開発に貢献を行った。さらに企業に技術導出を行い技術の普及に努めた。

3 つの研究開発項目からなる。研究開発項目 1 は「細胞機能と遺伝的多型の多検体同時計測法の開発」である。研究開発項目 1 はさらに研究開発項目 1-1「数千検体の全長 RNA シーケンスの達成」と研究開発項目 1-2「細胞機能と遺伝的多型の情報を抽出する AI の実装」に分けられる。研究開発項目 2 では 1-1 で開発した技術を 1 細胞レベルで数百検体に対して利用できるようにする。研究開発項目 3 では、研究開発項目 1 と 2 で開発した技術を用いて、細胞機能と遺伝的多型の多検体同時計測法の応用を行う。

研究開発項目 1 の細胞機能と遺伝的多型の多検体同時計測法の開発は現在の達成度は 100%である。世界最高精度の 1 細胞 RNA シーケンスや大規模検体向け RNA シーケンス法である Quartz-Seq2 をベースに Oxford Nanopore 社のロングリード DNA シーケンサーに応用した Long-Read Quartz-Seq2 の開発に成功した。最終的な性能としては、検出遺伝子性能ではショートリード Quartz-Seq2 と同等で既報の手法を超える性能を示した。バーコード認識性能でも既報の手法を大幅に上回った。また遺伝子の完全長の決定率も高い水準を示した。

Long-Read Quartz-Seq2 で得られたデータからの遺伝子機能と遺伝的多型情報を取り出す人工知能技術の開発を行った。シーケンスリードから遺伝子発現行列やゲノム上のリードの位置などの情報を自動的に計算するデータ解析ワークフローQ2-pipeline を開発した完成させた。また、ゲノム多型情報を引き出すワークフローを構築し、多型解析が行えることをミトコンドリアゲノムの多型を例に示した。さらに、大規模生成 AI モデルを利用して、遺伝子配列への摂動が遺伝子発現変動に及ぼす影響を予測するモデルを構築した。このモデルは予測された遺伝子発現と実測値の相関係数を用いて性能評価を行った。またこのモデルを用いて遺伝子発現に影響を及ぼすゲノム領域も同定できることを示した。

研究開発項目 2 では、開発した技術を 1 細胞レベルで数百検体に対して利用できるようにするために実験のスループット向上、低コスト化を目指した。細胞採取速度はセルソーターから液滴生成流路に変更することで、300-600 倍のスループット向上を達成した。シーケンスコスト低下については両端に異なるバーコ

ードを付与するシーケンスライブラリ作製法を導入し最新のシーケンサーで配列決定することで、1/20 のコストダウンを達成した。さらにシーケンスコストを低下させるために cDNA 収量改善を進めた結果、３つの新たな改善点を発見できた。現在、これらの技術を統合し、最終的なスループットと性能を確認した

研究開発項目 3 については、研究開発項目 1 と 2 で開発してきた技術を用いて、細胞機能と遺伝的多型の多検体同時計測法の応用を行った。これらの技術を用いて 4 件の拠点との連携、3 社への技術導出を行った。

我々は本研究で開発された技術の普及のため、ナレッジパレット社へ技術移転している。創薬や再生医療等製品の開発を目指すスタートアップ、ナレッジパレット社は、我々のラボに所属し本事業で雇用されていたポスドクが 2018 年度に設立した。NEDO, 川崎市などの支援、2021 年 3 月にシリーズ A ラウンド、2021 年 8 月に追加投資を受けた。ジャパン・ヘルスケアベンチャー・サミット 2020 の審査員特別賞、Startup Pitch@CIC Deep Tech で優勝を獲得した。複数の国内外の製薬企業(田辺三菱製薬株式会社、小野薬品工業株式会社、マルホ、Axcelead Drug Discovery Partners 株式会社など)と共同開発をスタートしている。また国内外の再生医療研究を行っているアカデミアのグループとの共同研究も多数行っている。さらに 2023 年にシリーズ B ラウンドの大型資金調達に成功した。

【研究開発の意義】

本技術は、細胞・遺伝子治療の開発のみならず、創薬における細胞スクリーニングやコホート研究、バイオバンクなど大量の検体が得られるプロジェクトに応用できる。我々のラボでは、数千から数万を数名のスタッフで従来の数桁以下のコストで実施できるシステムを構築できた。遺伝子発現を収集する国際プロジェクトや国家プロジェクトは複数あるが、数十年、数十億年、数千人の研究者が集まって実施してきたものだが、我々は同等の規模を数ヶ月で得ることができる。

このような大規模データは、大規模な生成 AI の学習データとして利用でき、細胞・遺伝子治療をサポートする AI の開発に大きく貢献すると期待できる。近年、ChatGPT のような大規模言語モデル(Large Laungage Model: LLM)を実装したサービスが様々な分野で活躍しつつある。現在、自然言語ではなく DNA 配列を学習した生成 AI モデルの開発が激化している。このようなモデルはひとつの AI で遺伝子発現予測や多型解析、細胞機能の予測・設計など様々なタスクに応用できるため、多様なデータ解析技術はたったひとつの AI モデルに集約されているとされており、官民学での開発競争が起きている。

現在、大規模生成 AI 構築のため GPU 獲得競争が起きているが、本質的には新しいデータの枯渇のほうが問題である。AI を賢くするためには、GPU だけでなく新規の大規模データが必須であるからである。これをエネルギー資源に例えれば、GPU は原油を精製する石油プラントに相当し、新規データは原油そのものに相当する。日本の研究競争力を保つためには、原油たる新規大規模データを生み出す「データ油田」の確保が必須である。

本研究で開発された技術は他に類を見ない規模と精度で大規模データを取得できるため、AI が必要とする規模のデータを独自に得られるため、細胞・遺伝子治療を支援する AI を生み出すためのデータ油田となりうる。事実、本課題で開発した大規模生成 AI では、大規模なオミクスデータを学習することで、実験することなく、細胞状態や遺伝子発現の変化、転写領域の同定などを１つの AI モデルで実現できている。今後は、より多様な細胞状態のデータを生産し学習することで、細胞・遺伝子治療法の開発に貢献すると期待できる。

In this project, we developed a method to simultaneously investigate the cellular functions and genetic polymorphisms of human-induced pluripotent stem cells (hiPS cells) and their differentiated cells from thousands of samples. Additionally, we developed a method to measure these parameters at the single-cell level from hundreds of samples. Using this technology, we collaborated with research groups to analyze how culture conditions, compounds, and cell lots/strains affect the characteristics of hiPS cells and differentiated cells. The goal is to feedback these results into clinical research and trials to accelerate the early realization of cell and gene therapy.

Cells used in cell and gene therapy have different characteristics depending on the lot and strain, affecting transplantation safety and efficacy. Factors influencing cell characteristics include culture media, containers, passage numbers, adhesion factors, genetic background, etc. We previously developed a method to measure cellular functions from thousands of samples using RNA sequencing. However, current methods for measuring genetic mutations cannot examine a few cells within a cell population, have low throughput, and are very costly. These methods also do not provide information on cellular functions, making it impossible to evaluate how polymorphisms affect cell characteristics.

To address those technical issues, we developed a novel method for simultaneous measurement of cellular functions and genetic polymorphisms from many samples. Based on Quartz-Seq2, the world's highest-performing single-cell/multi-sample RNA-seq method, we developed Long-Read Quartz-Seq2 by adapting it to sequence entire RNA molecule, capturing gene structure changes. We introduced full-length sequencing library preparation methods and adapted long-read DNA sequencers. To increase the number of samples processed, we adapted Quartz-Seq2 to microfluidic devices and mixed-reaction methods for multiple samples. We also developed artificial intelligence technology to extract genetic function and polymorphism information from these data.

The project consists of three research and development (R&D) items. R&D Item 1 involved developing a method for simultaneous measurement of cellular functions and genetic polymorphisms from many samples. This was divided into achieving full-length RNA sequencing of thousands of samples and implementing AI to extract information on cellular functions and genetic polymorphisms. R&D Item 2 aimed to apply this technology at the single-cell level for hundreds of samples by improving throughput and reducing costs. R&D Item 3 applied the technology developed in Items 1 and 2 for practical use.

The development of Long-Read Quartz-Seq2 was successful, achieving performance comparable to short-read Quartz-Seq2 and exceeding other reported methods in gene detection and cell/sample barcode recognition. We developed the Q2-pipeline, a data analysis workflow for automatically calculating gene expression matrices and the positions of reads on the genome. We demonstrated variant analysis using mitochondrial genome polymorphisms. Using a large-scale generative AI model, we built a model to predict the impact of gene sequence perturbations on gene expression changes, evaluated by the correlation between predicted and measured expressions.

To reduce costs, we improved the throughput of cell collection and sequencing processes. This included changing the single-cell collection method and introducing a new library preparation method, significantly reducing sequencing costs. These improvements were integrated and confirmed for their final throughput and performance.

We transferred the developed technology to Knowledge Palette Inc., a startup established by a former postdoctoral researcher from our lab. The company has collaborated with multiple pharmaceutical companies and research groups, successfully raising significant funds in the Series B round in 2023.

This technology can be applied to drug discovery, cohort studies, and large-scale omics projects. Our lab has built a system to execute such projects at significantly lower costs. The large-scale data generated can be used for training large-scale generative AI, contributing to the development of AI supporting cell and gene therapy.

In conclusion, the technology developed in this project provides a unique and highly accurate method for obtaining large-scale data, essential for advancing cell and gene therapy and supporting AI development in this field.