【報告様式Ａ】

作成／更新日：令和４年 ５月 30 日
課題管理番号：21am0101102j0005

日本医療研究開発機構　創薬等ライフサイエンス研究支援基盤事業
事後評価報告書

公開

Ⅰ 基本情報

補助事業課題名：　（日本語）創薬等ライフサイエンス研究支援基盤事業
（プログラム名）　（英　語）Platform Project for Supporting Drug Discovery and Life Science Research

実施期間：平成２９年４月１日～令和４年３月 31 日

補助事業担当者　氏名：（日本語）カルニンチ　ピエロ
（英　語）Piero Carninci

補助事業担当者　所属機関・部署・役職：
（日本語）国立研究開発法人理化学研究所・生命医科学研究センター・副センター長
（英　語）RIKEN・Center for Integrative Medical Sciences・Deputy Director

Ⅱ 補助事業の概要

補助事業の成果およびその意義等

支援の成果

理化学研究所（代表機関）は国立遺伝学研究所（分担機関）と協力連携して、支援業務が円滑に運営されるよう課題全体を管理した。ゲノミクス解析のコンサルティング支援申請窓口として、支援申請・審査により採択された支援課題について、機能ゲノミクス解析パイプライン（サンプル受入、核酸抽出、ライブラリ作製、シーケンス、データ解析）を実行し、プロセス毎に統合的な支援を実施した。これまでの 5 年間に採択された 39 課題に連携・共同研究 4 課題及び追加支援 2 課題を加えた合計 45 課題の支援を実施した。また、令和 3 年 8 月より開始したヒト免疫系解析支援について、1 課題を実施した。生物種分類ではヒト 26 課題、マウス 18 課題、サル、メダカ、各 1 課題である。これら課題の研究対象は、感染・免疫疾患、癌、神経疾患等、全て疾患関連研究（20 課題は臨床検体を使用）で占められており、疾患メカニズムの理解や疾患バイオマーカーや治療標的の探索を目指す研究である。これらの課題で依頼された解析仕様は、RNA-seq 32 件（11 件は微量 RNA-seq）、CAGE-seq 10 件、small RNA-seq 5 件、完全長 cDNA-seq 3 件、WGBS 1 件（合計 51 解析案件 995 サンプル）に加えて、Single Cell（sc）解析（scRNA-seq 及び scATAC-seq、計 11 件 46 サンプル）で、トランスクリプトーム解析が主である。理化学研究所で支援を実施して産出されたシーケンスデータについて、国立遺伝学研究所でマルチゲノミクス解析を実施して統合的に解析を行い、解析結果を依頼者に還元した。BINDS 事業での研究成果として、前事業（PDIS）から共同研究として継続した 10 課題で 16 報、BINDS 支援課題で 7 報の論文発表がなされた。

高度化シーケンス技術開発の成果

支援に必要とされる要素技術の開発（高度化）として、理化学研究所の独自技術である CAGE 法をさらに発展させ、技術開発とハイスループット化を軸とした高度化として、以下の 8 課題を設定した。

(1) 自動微量非増幅 CAGE ライブラリ調製システムの構築

フルレングス cDNA を安定的かつハイスループットに合成してライブラリ作製を行うために、ssCAGE ライブラリ作製の工程を Bravo 自動化 NGS システムに適用した。ssCAGE ライブラリ作製の自動化について検証し、アダプター混入率の低減化に改良を加えてデータ量の増加とコストの低減化を実現した（2019 年度より支援メニュー化した）。これにより、2 週間以上の複雑な工程におけるエラーの発生を抑制し、より効率的にライブラリ作製が行えるようになり、かつ、事業期間後半で行ったプロトコル改良によりアダプター混入率の低減化を達成し、最終データに含まれる有効データ量を増やすことが出来た（データの質の改善を得た）。また共同研究にてヒト疾患の病態解明に資する研究プロジェクトに本 ssCAGE 法により貢献した（達成度 100%）。

(2) NET-CAGE 法のメニュー化に向けた取り組み

新生 RNA 鎖を濃縮して検出することで転写開始地点を鋭敏・正確に決定し、かつエンハンサー領域から発現する eRNA を検出することでゲノム上のエンハンサー領域を同定するのに有用である NET-CAGE 法を導入した。培養細胞株及び複数のヒト患者由来初代培養細胞を用いた NET-CAGE 解析を共同研究として実施して実績を蓄積し、さらにライブラリ作製を自動化することによりハイスループット化に成功した。後半年度では共同研究としてヒト希少難治性疾患の病態評価に応用した取り組みを行なった（達成度 100%）。

(3) 完全長 cDNA-seq 開発

CAGE 法を利用して合成した完全長の cDNA をロングリードシーケンサーであるナノポア MinION を用いて 1 本鎖レベルで long-read シーケンスすることにより、これまで short-read シーケンスに必要であった 2 本鎖 cDNA 合成と断片化の工程が省かれ、完全長 cDNA を 1 本の連続した配列情報として得られる技術を開発することに成功した。これにより遺伝子発現のバリアントフォームの高精度解析技術が達成された。 2019 年度より共同研究レベルで、本手法の実用化に向けた取組みを実施している。共同研究として既に 1 報の論文を報告し、さらに他にも支援として共同研究を令和 3 年度に実施した（達成度 100%）。

(4) 極微量（pg レベル）RNA-seq の支援導入

Total RNA（100 ng～10 ng）を出発原料とするトランスクリプトーム解析を行うとともに、単純な遺伝子発現解析に加えて、パスウェー解析等の高次解析も支援することを目標とした。市販キットの比較検証により、Ovation SoLo RNA-Seq System（NuGEN Technologies Inc.）がライブラリ作製法として再現性に優れていることを見出した。これにより Total RNA 100pg というごく微量サンプルからの RNA-seq が可能となった。また、別課題である極微量 RNA の品質検査については、ヒト限定ではあるが、豊富に含まれる rRNA を指標とした qPCR 法をサンプル定量法として構築した。2018 年度からメニュー化して技術開発を終了し、支援を開始した（達成度 100%）。

(5) Chromium 及び ONT MinION long read による genome phasing 解析技術の支援導入

Chromium System（10X Genomics 社）を用いて、ゲノムフェーシングライブラリを作製し、シーケンスするまでをメニューとして提供することを目標とした。培養細胞由来 Genome DNA を用いた Chromium system によるライブラリ調製―シーケンス―データ解析パイプラインの構築が完了し、メニュー化して技術開発を終了した。（達成度 100%）。MinION を用いた genome phasing 解析技術については、現時点ではスループットおよびコスト面で課題が多く、Chromium System による技術開発の完成により支援・連携への応用が可能となったため、MinION については課題(3)への活用を優先して技術開発を中止とした。

(6) Direct RNA-seq の技術評価

実サンプルを用い、MinION による RNA 分子の直接シーケンスが可能であるが、創薬及び医療への実用化

に資する技術として確立するには、RNA 量の低減化が必須である。本手法の主目的である完全長 RNA の配列決定については、完全長 cDNA-seq がメニュー化で補完が可能であることから、完全長 cDNA-seq 法に一本化して推進した。

(7) Large-scale low-cost directional single cell RNA-seq の技術開発

1 細胞から ①（非 poly-A RNA を含む）すべてのトランスクリプトを②ストランド情報を保ったままキャプチャーして、③UMI (Unique Molecular Identifier) を付加することで cDNA 増幅バイアスの影響を低減し、さらに Total RNA の全長シーケンスを可能とする RNA 解析技術を開発することを目標とした。独自に開発した total RNA を解析対象とした新規 sc RNA-seq 法（新法）と polyA+ RNA を解析対象とした最新の高感度低コスト型 Smart-seq3 法との比較評価を行い、遺伝子検出数ではやや劣るものの non-codingRNA の検出に成功した。超微量組織を対象とした sc 解析技術の実用化への取り組みをヒト・マウスの組織検体で実施し、共同研究を交えて着実に研究支援と研究高度化の成果を挙げている（達成度 80%）。

(8) 長鎖塩基配列のターゲットシーケンス

長鎖塩基配列のターゲットシーケンスを目的として CRISPR/Cas9 システムにより関心領域を切断し、MinION ロングリードシーケンサーで解析するプロトコルの改良を行った。長さや Read 深度などのデータ改善のため、長鎖 DNA の最終回収率を上げるための、ライブラリ精製法の改良を行った。また技術開発と並行して、令和 3 年度は共同研究にて神経筋疾患の遺伝子変異の同定への応用（ターゲットロングリードゲノム phasing を用いた複合ヘテロ変異の確認など）を行ない高度化の成果を挙げた（論文投稿準備中）（達成度 80%）。

高度化データ解析技術開発の成果

創薬支援の一環として、NGS データを中心にしたデータマイニングを通じてターゲットを絞り込み、疾患分子機構モデルの構築を進めるために、ゲノム多様性、タンパク相互作用などの関連情報や統計情報学的手法を有効利用することを目的として、以下の高度化 3 課題を実施した。加えて、令和 2 年度より (4) Single Cell データ解析パイプラインへの対応を進めた。

(1) 疾患マーカーおよび創薬ターゲット探索のためのヒト以外の生物種を含めたゲノムリファレンス情報の充実、遺伝子発現、相互作用データ利用の促進

公開情報をデータベース化して利用するための環境整備を継続し、支援に提供した。さらに、モデル生物ゲノムのリファレンス情報に加えて、独自に決定された各種生物ゲノム情報の利用に向けたゲノム配列アノテーションパイプラインを整備した。未発表ゲノム配列や異なるバージョンのゲノム配列を簡便に利用出来る解析パイプラインと解析結果ビューワーの整備や利用を開始した。これらのパイプラインについては、支援における活用に加え、本事業で開発を進めている大規模データ解析システム(Maser)を通じて、一般登録者による利用も可能とした。加えて、ウイルス等の感染症を念頭に、宿主ゲノム及びウイルス等のゲノム配列を同時にリファレンスとして利用出来るパイプラインも開発し、宿主ゲノム上でのウイルス配列の探索、宿主とウイルス融合遺伝子の探索を可能とするアルゴリズムやパイプラインの開発を進め、支援に提供した。（達成度 100％）

(2) 完全長 cDNA-seq および微量サンプル解析パイプラインの開発

従来の RNA-seq 解析パイプラインを改良拡張し、完全長 cDNA 解析にも対応を可能とする開発を進めた。特に、既存データと微量 RNA-seq の解析結果について、比較検証を行うとともに、それぞれの解析結果にもとづいたマップ率、発現変動遺伝子の比較検討や個々の遺伝子リード数の比較検証を可能とする解析フローの開発とパイプライン化を進め、支援への提供を開始した。（達成度 100％）具体的には nanopore シーケンサで得られる long-read cDNA 配列をマッピングし、StringTie を用いて遺伝子構造予測・リードカウントを行うとともに、（differential splicing 検索のために）GffCompare で構造比較を実施するパイプラインの

開発と利用を開始した。また、重複配列データを用いて遺伝子発現変動解析も可能とした。【1439】との共同研究では、このパイプラインを用いてマッピングしたデータについて、DEXSeq による differential splicing 解析を行うとともに、exon 単位で発現変動解析を加えて、exon usage が異なる遺伝子バリアントを多数検出した。

(3) Chromium 及び ONT MinION long read による genome phasing 解析技術のデータ解析支援技術の導入

long-read コンティグ作成ツールとフローを整備し、さらに SNP 検出によるハプロタイプの同定、phasing に有用な解析フローの構築を行った。また、新規に出現した変異を含め疾患責任変異の同定を可能とするための解析技術の開発を行っており、自動解析パイプラインとして解析支援での利用が開始されている。この際、家系データがある場合には、より詳細な解析を可能とする統合解析パイプラインの開発も同時に行った。また、得られた SNPs の機能への影響を予測するためのアルゴリズムを開発し、それぞれの SNP の疾患への影響の度合いを予測することが可能なパイプラインを構築して支援への提供を可能とした。（達成度 100%）

(4) Single Cell データ解析への対応

sc データ解析への対応として、ゲノムマッピングベースでデータ処理パイプラインの実装を完了した。また、sc 解析で大規模データ処理の高速化を図り、コンテナ技術を用いて、従来サーバー上で実施していた解析処理をスパコン上に移行して本格運用を開始した。この結果、1 ジョブ当たり 10 億リード/日の処理能力を実現した（スパコンの空きスロット状況により、複数ジョブ並行処理が可能）。また、AI を利用した細胞タイプアノテーションのデータフロー解析により、機械学習を用いた遺伝子発現プロファイルの細胞種アノテーションが可能となり、未知細胞の推定を可能とするアルゴリズムの開発に至った。また、trajectory 解析に加えて、得られた配列の変異パターンを抽出して利用することにより、細胞系譜を再構築するアルゴリズムを開発し、疾患関連遺伝子の変異や癌のドライバー変異を細胞レベルで検出可能とする解析を実現した（達成度 100%）

ユニット内連携の成果

BINDS で前事業（PDIS）からの支援継続により共同研究に発展した 1 課題及び BINDS 支援 1 課題について、ユニット内連携を実施した。連携内容として、九州大学及び東京大学がエピゲノム解析（BS-seq、ChIP-seq）を担当し、理研がトランスクリプトーム解析（RNA-seq）を実施した。これら 2 課題についてはそれぞれ論文化がなされた。

AMED 連携の成果

ジャパン・キャンサーリサーチ・プロジェクト次世代がん医療創生研究事業（AMED P-CREATE）との連携課題として、順天堂大学と共同研究を実施した。本研究では、リンパ節転移診断バイオマーカーの探索を目的として、子宮体癌組織について CAGE 法による転写開始点解析及び遺伝子発現解析を実施した。リンパ節転移の有無に最も相関する 2 遺伝子、TACC2 新規アイソフォームと SEMA3D を同定し、これら 2 遺伝子発現量の組み合わせにより、高精度にリンパ節転移の識別が可能であることを明らかにした。さらに高度化課題「(2) 完全長 cDNA-seq 開発」を応用した共同研究として、MinION を用いた TACC2 新規アイソフォームの網羅的解析を行った結果、子宮体がん細胞においては、800 以上のアイソフォームが見出され、生体内ではこれまで考えられてきたよりも遥かに多くのアイソフォームが機能していることが示唆された。

Support

RIKEN（the representative organization) cooperated with National Institute of Genetics (NIG; the co-representative organization) to manage all the research projects and promoted the smooth operation of business support. We executed the functional genomics analysis pipeline (sample acceptance, nucleic acid extraction, library preparation, sequencing and data analysis) for the research projects, which ware adopted by the support application and subsequent acceptance via a one-stop contact for consulting with a wide range of life science researchers regarding functional genomics analysis. We provided integrated support for each process of the pipeline. We provided support for a total of 45 projects, including 39 projects adopted over the past five years, 4 collaborative research projects and 2 additional support projects. We carried out one research project for human immunological system analysis support started in August 2021. The sequence analysis specifications requested for these research projects were 32 RNA-seq including 11 low quantity RNA-seq, 10 CAGE-seq, 5 small RNA-seq, 3 full-length cDNA-seq, and 1 WGBS, respectively. In addition to the analysis for the total of 51 requests, single cell analysis for the total of 11 requests of scRNA-seq and scATAC-seq was also carried out. Most of the requests were occupied by transcriptome analysis. The sequence data produced with the support of RIKEN was analyzed in an integrated manner by multigenomic analysis at NIG, and the analysis results were returned to the client. As a results of BINDS achievement, 16 papers were published in 10 projects that continued as collaborative research from the previous project (PDIS), and 7 papers were published in the BINDS support project.

Improvements（Sequence Technologies）

We customized the CAGE (Cap Analysis of Gene Expression) method, which was originally developed in our institute, for developing individual technologies needed for above described “Supports”, and conducted the following eight “Improvements (Sequence Technologies)” subprojects.

(1) Development of automated low input non-amplified CAGE library preparation system

We applied Bravo automation system (Agilent) for preparation of ssCAGE library, which is very complicated and requiring much time and effort if manually conducted. We succeeded in automation as well as improving the quality of prepared libraries, especially in reducing the portion of adaptor carry-over, which is the most major reason for low throughput.

(2) Introduction of NET-CAGE (native elongating transcript–cap analysis of gene expression) method

NET-CAGE, a derivative method of CAGE, can condense and detect nascent RNA species, and accurately and sensitively detect the transcription start site. NET-CAGE also can detects eRNAs, which are bilaterally transcribed from the putative enhancer element, and thus is presumed to be useful for detection of enhancer elements. We introduced this novel technologies to add to the menu list of “support”.

(3) Development of the full length cDNA-sequencing technology

We modified the CAGE method to sequence the full length cDNA, by applying CAGE-derived full length cDNA to long read sequencer nanopore MINION (Oxford Nanopore Technologies). We succeeded in consolidating the merge of these two technologies to achieve the full length long read sequencing. Further, we successfully applied this developed technology to the analysis of cancer related isoform variations, which is already published.

(4) Ultra-low input (order of ～pg) RNA sequencing

We surveyed commercially available kits of low input RNA sequencing protocols. In parallel, we conducted ordinary low input (100-10ng of total RNA) RNA sequencing and high-order data analyses as a “support”. By evaluating commercial kits, we found that Ovation SoLo RNA-seq System (NuGEN Technologies Inc.) has the highest reproducibility. By applying this kit, we developed RNA sequencing protocol which can start with ultra-low input (100pg of total RNA).

(5) Large-scale low-cost directional single cell RNA sequencing

We developed a new method to sequence all the transcripts including non-coding RNAs, conserving strand information, and counting UMI (unique molecular identifier).

(6) Target long read sequencing

By combination of CRISPR/Cas9 enrichment of region of interests and ONT MinION long read sequencer, we developed a novel protocol for more efficient sequencing. We applied the technology to resolve phases in compound heterozygotes with neuromuscular diseases.

Results of Advanced Data Analysis Technology Development

As part of drug discovery support, the following four upgrading tasks were carried out.

(1) Enhancement of genome reference information including non-human species for the search of disease markers and drug targets, and promotion of the use of gene expression and interaction data

To the utilization of these pipelines in the support, they were made available to the general public through the data analysis system (Maser). Host genomes and genomes of viruses and other organisms were made available for simultaneous analysis, and the search for viral sequences on host genomes and host-virus fusion genes were provided for support. (100% achievement)

(2) Development of full-length cDNA-seq and trace sample analysis pipelines

We started using a pipeline that maps long-read cDNA sequences, performs gene structure prediction and differential splicing searches.

(3) Introduction of data analysis support technology for genome phasing analysis technology using Chromium and ONT MinION long read.

We developed analysis technology to enable identification of disease responsible mutations including newly emerged mutations and more detailed analysis when genealogical data are available. (100% achievement)

(4) Support for Single Cell data analysis

Using container technology, the analysis process was moved to a supercomputer, achieving a processing capacity of 10 million reads/day per job. AI was also used to estimate the function of unknown cells. Developed an algorithm to reconstruct cellular lineage by using mutation patterns. (100% achievement)

Intra-unit collaboration

Intra-unit collaboration was carried out for one research project that developed into joint research that continued to be supported from the previous project (PDIS) and one research project that was supported by BINDS. In these collaborations, Kyushu University and the University of Tokyo were in charge of epigenome analyses (BS-seq, ChIP-seq), and RIKEN conducted transcriptome analysis (RNA-seq). Each of these two projects was published as a paper.

AMED collaboration

As a collaborative project with AMED P-CREATE, we conducted joint research with Juntendo University. In this study, comprehensive analysis for transcription initiation site analysis and gene expression by CAGE-seq was performed on endometrial cancer tissue for the purpose of searching for biomarkers for diagnosing lymph node metastasis. We identified two genes, TACC2 novel isoform and SEMA3D, that most correlate with the presence or absence of lymph node metastasis and clarified that the combination of these two gene expression levels enables highly accurate identification of lymph node metastasis. As a result of in detail analysis of the new TACC2 isoform using MinION in a collaboration applying "(3) Development of full-length cDNA-seq", more than 800 isoforms were found in endometrial cancer cells, suggesting that far more isoforms are functioning in vivo than previously thought.