様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成 25　年　5　月　20　日現在

機関番号：１１３０１
研究種目：挑戦的萌芽研究
研究期間：2011～2012
課題番号：23650150
研究課題名（和文）ヒト数千人規模 ncRNA の網羅的探索と機能予測をスパコンで一気通貫に実現する技術

研究課題名（英文）RNA-Seq data analysis for ncRNA function prediction to thousands public RNA-Seq data on super computer

研究代表者
長崎　正朗（NAGASAKI MASAO）
東北大学・東北メディカル・メガバンク機構・教授
研究者番号：90396862

研究成果の概要（和文）：次世代シークエンサではさまざまな転写産物を同時に見ることができるが同じ領域から出る転写産物について実際にどの転写産物がどの程度出ているかを区別することが課題である。特に機能を行っているかどうか未知である ncRNA や microRNA の網羅的探索においてはより正確な転写産物の推定が重要である。本研究ではこれらの特性を生かした転写量推定アルゴリズムを開発した。また、スーパーコンピュータを用いて lincRNA および mRNA を抽出できる環境整備を本研究成果により実現することができた。

研究成果の概要（英文）：High performance sequencer (Hit-Seq) generates massive amounts of short reads. Hit-Seq is used for the analysis of transcriptome expressions. Especially, it is important to detect fusion transcripts and isoforms expression from overlapped region in a chromosome. This research developed a new method to detect isoform expressions from RNA-Seq data. By applying this method on supercomputer environment, more accurate expressions to unkown lincRNAs and mRNAs from thousands RNA-Seq data can be inferred.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	1,400,000	420,000	1,820,000

研究分野：総合領域
科研費の分科・細目：情報学・生体生命情報学
キーワード：次世代シーケンサ

１．研究開始当初の背景

ヒトの全ゲノムの 97%は非コード領域である。その内 microRNA については研究が進んでいるが、他の非コード領域にもまだまだ機能領域が隠されているといわれている。実際に、DNA 複製の開始点として定義される

複製起点、あるいはプロモーターといった遺伝子の発現を制御する領域が推定や確認がされている。

一方、計測機器として、次世代、次々世代シーケンサの開発が着々と進み、現在、数百万程度必要とするデータを､数年後には 1000 ドル以下で数分以内にヒトゲノムの配列情報が得られるという発表さえされている。この装置により、DNA だけでなく、RNA を計測することで、非コード領域の詳細な発現情報(≒ncRNA)についても非常に低コストで取得できるようになる。

しかし、取得されるシーケンスのデータは現在の Solexa で 1 検体あたり 300GB 以上であり、100 検体の同時解析を考えると、生データだけで 30T のデータ処理を行いつつ解析を行う必要がある。

また次世代シーケンサは 1TetaByte のデータを出力することが想定され、スーパーコンピュータのリソースを利用し効率よくデータを処理する手法の開発、またそのような大量データに適した解析技術開発が急務である。

２．研究の目的

ヒト数千人規模 ncRNA の網羅的探索と機能予測をスパコンで一気通貫に実現するための技術「次世代シーケンサがもたらす肝がんなどのヒト数千人規模の超大規模な RNA シーケンス情報から、大型計算機の計算リソースとストレージを最大限に活用し、生体内の機能が不明瞭であるヒトの全ゲノムの９７％の非コード領域の中から、機能をもつ領域とその役割を発見・推定する情報科学・統計数理科学の技術・手法を開発すること」を目的とする。

３．研究の方法

(1) 概要

平成 23 年度は、内部のプロジェクトで取得される RNASeq のデータと公共データベースから取得できるあらゆる RNASeq のデータを取得･整理する(120 件程度)。さらに RNASeq の一次解析パイプラインを構築する。このパイプラインを、ヒトゲノム解析センターのスーパーコンピュータ上で実行し効率よく一次解析を行う。さらに、これらの情報に対して、有意に機能をもつ microRNA、lincRNA の抽出を数理統計の技術を用いて二次解析を行う。

平成 24 年度においては、数千件程度のデータ解析を行うとともに、ChipSeq のデータについても一次解析を行い、さらに RNASeq の２次解析につなげることでより信頼度の高い機能 microRNA、lincRNA を抽出する。また､肝がんで特異的な機能 microRNA､lincRNA を抽出する。

(2) 23 年度に予定していた研究計画

東京大学医科学研究所ヒトゲノム解析センターには 2009 年 1 月に導入した，ピーク性能値で 75TFLOPS の PC クラスター型スーパーコンピュータ及び 2TB の共有メモリ型スーパーコンピュータ があり、1PB 容量のディスクが整備されている。すでに、ネットワークの推定、シミュレーション、大規模データ解析などの大規模計算この設備を用いた実績がある。

この計算資源を用いて非コード領域の解析のための準備を進める。平成 23 年度研究開始時点において、肝がんデータについて数十検体のデータが取得され利用できる状況にあると想定されるが、その他のデータと同時に、公開されている次世代シーケンサ(NGS)データを積極的に用いて解析準備を整える。具体的には、NCBI の Sequence Read Archive に登録されている NGS の RNAseq のデータと国立遺伝研の NGS の Archive データで利用可能なデータを対象に解析を開始する。SRA には RNASeq のデータは 2010 年 10 月時点で８９件登録されており、2011 年 4 月時点では 120 件程度登録されていることが想定され、研究を開始するにあたって十分な数のデータがそろっていることが保障されている。このデータの中には、National Human Genome Research Institute (NHGRI)の ENCODE Project (ENCyclopedia of DNA Elements; http://www.genome.gov/10005107) のデータも含んでおり、主要なデータセットを網羅していると考えている。

NGS のデータの解析には、まずリファレンスゲノムへのマッピングが必要である。そのために、高速な short read のマッピングが可能なソフトウェア BWA を用いて行う。このソフトウェアを用いても、上記の 120 件のデータのマッピング処理には、4core のマシン 1 台を用いて 180 日程度必要であると見積もられる。しかし、本スーパーコンピュータを用いることで準備する解析パイプラインさえ安定して動作すれば、2 週間程度で終了する（本マシンは占有ではなく共有で利用するためマシンの込み具合によって多少前後することが想定される。なお 2 週間については、過去の利用率の経験から見積もっている。）

マッピング後はさらに、プロトコル上避けることができない、PCR-duplicate の除去 SamTools もしくは Picard などの利用を想定)、マッピングのカバレッジを利用したシーケンスデータから出力される Quality

Score の更新など(GATK などの利用を想定)のアルゴリズムを適用し、解析に必要となるデータの品質を高めつつデータ整理を行う。なお、この時点で後に必要となる中間ファイルを含めて、10T 程度のストレージが必要となる。通常のパソコンでは通常扱えない量のデータかつ十分な I/O 性能が要求されるが、本スーパーコンピュータの Luster を用いた超高速なストレージ環境を利用することで円滑に研究を推進できることが保障できる。

非コード領域の RNA といっても、microRNA や lincRNA など複数のクラスが存在し、それぞれの生物学的な理解レベルには差がある。microRNA については、まずバイオインフォマティクスの分野で開発されている microRNA 推定ソフトウェア群の解析結果を用いてデータを整備する。

なお、利用する microRNA の推定ソフトウェアに依存して、microRNA として機能するかどうかの判定結果が大きく変わることを経験している。そこで、１つのソフトウェアには絞らず、複数の推定ソフトウェアで１つでも機能するとされる領域については候補対象とすることで、よりロバストなデータセットを用意する。それら機能が推定される部位についての 120 件のデータについての発現パターンを解析することで、有意に機能を持っていると考えられる microRNA を同定する。一方、lincRNA のクラスについては、2009 年に Guttman ら (2009, Nature) により新たな ncRNA のクラスとして lincRNA の存在が明らかになり、さらに、その lincRNA の１つHOTAIR 遺伝子に関してエピジェネティクスの制御、癌化との関連が示され(Rinn et al. 2007, Cell; Gupta et al. 2010, Nature) ncRNA の中に重要な新規機能因子が含まれることが再確認されている。

しかし、依然単発的な研究にとどまり、その他大多数の lincRNA の機能、生理的意義についてはほとんど情報がないのが現状である。そのため、microRNA のように予測ツールは開発されていない。lincRNA のクラスは 1000base 以上の非常に長い ncRNA であることは知られている。そこで、本研究では、これらの 120 件のデータからパターンをもって発現している有意な長い領域を統計数理科学の技術を用いることで抽出を行うことで解析を行う。

(3)24 年度に予定していた研究計画

平成 23 年度において大量の NGS のデータが SRA などに登録されることが想定される。見積もりでは、前年度の数倍以上 1000 件以上のデータが登録されるものと推定している。本研究では前年度で開発したマッピングと後処理を行うパイプラインを、スパコンを用いて実行することで更新追加するとともに、情報量がふえることで、より信頼度の高い機能をもつ microRNA・lincRNA の推定を行う。

各 NGS の計測機器には GC リッチな領域やリピートを含む領域についてはシーケンス結果が不正確になるという特徴などがあり、RNA-Seq の結果のみから機能 microRNA・lincRNA を推定するのは不十分である。そこで、24 年度においては、NGS によって取得が進む Chip-Seq などの他の種類のデータについても合わせて解析に加えることでより信頼の高い機能 ncRNA を抽出する。そのために、前年度の RNA-Seq 処理するためのパイプラインを改良することで Chip-Seq のための解析パイプラインを構築する。特に、PolII などの転写にかかわる汎用的な因子を使った Chip-Seq のデータを利用することで、実際に転写がおこなわれている領域を同定することができ、場合によっては格段に信頼度の高い機能 ncRNA を抽出できること考えらえる。なお、新しい計測機器による１実験あたりのデータ量の増加と解析対象のサンプルの増加による、解析に必要なストレージと計算リソースが増加することが想定されるが、平成２４年１月から、所属する東京大学医科学研究所ヒトゲノム解析センターでは 225TFLOPS、4PB 容量のディスクに増強することが計画されている。そのため、データの増加においても問題なく円滑に研究を推進できることが保障できる。

また、この時点において、肝がんゲノムプロジェクトで取得されるデータ量についても数百件以上となる。このデータを用いることで、肝がんにおいて機能発現している ncRNA を上記で開発したパイプラインを用いることで推定を行う。

４．研究成果

平成 23 年度は、東京大学医科学研究所ヒトゲノム解析センターには 2009 年 1 月に導入した，ピーク性能値で 75TFLOPS の PC クラスター型スーパーコンピュータ及び 2TB の共有メモリ型スーパーコンピュータ (http://supcom.hgc.jp/japanese/) があり、1PB 容量のディスクが整備されている。すでに、ネットワークの推定、シミュレーション、大規模データ解析などの大規模計算にこの設備を用いた実績がある。この計算資源を用いてコード、非コード領域トランスクリプトーム解析のための準備を進めた。公開されている次世代シーケンサ(NGS)データを積極的に用いて解析準備を整えた。具体的には、NCBI の Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) に登録されている NGS の RNAseq のデータと国立

遺伝研の NGS の Archive データで利用可能なデータを対象に解析を開始した (200 件程度)。このデータの中には、National Human Genome Research Institute (NHGRI) の ENCODE Project (ENCyclopedia of DNA Elements; http://www.genome.gov/10005107) のデータも含んでおり、主要なデータセットを網羅していると考えている。

このデータに対してスーパーコンピュータ上で動作するパイプラインの実装を行った。また、高機能シークエンサのデータに適した数理統計モデルの基本コンセプトの検討を行い次年度の実装準備を行った。

次世代シークエンサではさまざまな転写産物を同時に見ることができるが同じ領域から出る転写産物について実際にどの転写産物がどの程度出ているかを区別することが課題である。特に機能を行っているかどうか未知である ncRNA や microRNA の網羅的探索においてはより正確な転写産物の推定が重要である。また、各 NGS の計測機器には GC リッチな領域やリピートを含む領域についてはシーケンス結果が不正確になるという特徴などがある。また、ペアエンド法によってシークエンスを行う方法や転写方向を意識したシークエンスの手法などが考案されている。さらに、シークエンサの種類によってシークエンスされるデータのエラー率やエラーの入り方のパターンが異なることが報告されている。

平成 24 年度は、前年度の検討に基づき、これらの特性を生かした転写量推定アルゴリズムを開発した。また、前年度で取得した大規模な SRA のデータに対して、本成果をスーパーコンピュータおよびの上で適用することで信頼度の高い機能している lincRNA および mRNA を抽出できる環境整備を本研究成果により実現することができた。

これらのデータを参考にすることでマイクロアレイのデータ解析などのデータ解析研究の因子の絞り込みのヒントとして利用することができた。今後は、これらのデータの整理を行うとともに新しく取得される公開データについて同様の手法を適用し共通して発現している転写産物の抽出などを今後の研究として発展させていく予定である。

この情報科学・統計数理科学をスーパーコンピュータ上で最大限に活かした研究成果に基づき、予想された機能 ncRNA と予測された機能を道しるべとし、安心して何千人もの分子生物学者・医学系研究者が得意とする生物学的な知識と in vivo/vitro の解析技術を用いて、より詳細に機能同定を行う領域に踏み込むことができると考えている。さらに、データ取得の部分については、次々世代シーケンサが数年後に控えており、本研究の最終年度の終わりあたりには非常に安価・高速かつ高品質で取得できるようになることが想定される。その時点において、肝がんプロジェクトでは非常に高額の公的資金を投入して行われたデータ取得と同じレベルのデータ取得、それ以上の高品質かつ大規模のデータ取得が行われはじめることは明らかであり、この研究の成果は機能 ncRNA を推定・抽出するために必要不可欠な解析インフラにもなりうる成果であると考えている。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

(1) Nagasaki M, Fujita A, Sekiya Y, Saito A, Ikeda E, Li C and Miyano S, XiP: a computational environment to create, extend, and share workflows, Bioinformatics, 2012, 28:21 doi: 10.1093/bioinformatics/bts630 （査読有）

(2) Koso H, Takeda H, Yew CC, Ward JM, Nariai N, Ueno K, Nagasaki M, Watanabe S, Rust AG, Adams DJ, Copeland NG, Jenkins NA, Transposon mutagenesis identifies genes that transform neural stem cells into glioma-initiating cells, Proc Natl Acad Sci U S A, 2012, 109(44):E2998-3007 doi:10.1073/pnas.1217039110 (査読有)

〔学会発表〕（計４件）

① Data Management and Bioinformatics of High Throughput Sequencing Data on the Massive Parallel Supercomputer Environment and Future, Bilateral Workshop between Tohoku University and National Tsing Hua University（招待講演）,2012/12/11,仙台

② ゲノム超ビックデータ時代における次世代シークエンスデータ解析環境の構築と解析，化学工学会第 44 回秋季大会秋季大会シンポジウム(招待講演),2012/9/21,仙台

③ Data Managenoment and Bioinformatics of High Throughput Sequencing Data on the Massive Parallel Supercomputer Environment, CSI-GCOE Joint workshop

on Genome Science(招待講演), 2012/8/21, 仙台

④ スーパーコンピュータを用いた超快適高速次世代シークエンスデータ解析環境の構築と解析, 東北大学研究科シンポジウム（招待講演）, 2012/4/26,仙台

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

〔その他〕
特になし

６．研究組織
(1)研究代表者
長崎　正朗（NAGASAKI MASAO）
東北大学・東北メディカル・メガバンク機構・教授
研究者番号：90396862