様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　７　月　２８　日現在

機関番号：　２２３０３
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６４３０１９９

研究課題名（和文）機器の特性を考慮した次世代シーケンサー解析技術の高度化

研究課題名（英文）Development of NGS Data Analysis Softwares Considering the Data Profile of Each Platform

研究代表者

中村　建介（Nakamura, Kensuke）

前橋工科大学・工学部・教授

研究者番号：２０２１２０９５

交付決定額（研究期間全体）：（直接経費）　　4,000,000円

研究成果の概要（和文）：次世代シーケンサーリードマッピングプログラム maps の開発をすすめることで、以下の成果をあげることが出来た。(1) ChIP-seqを高精度化したGef-seqデータ解析の開発に寄与することが出来た。(2) イルミナ社シーケンサーのエラープロファイルの解析を進めることができた。(3) 植物オルガネラゲノムの一部に起きる相同、あるいは非相同組換変異の解析をすすめることが出来た。(4)酵母のゲノム比較研究を行った。(1)については論文発表とプログラムの公開をおこない、プロトコルの投稿を準備している。(2)については学会発表を行った。(3,4)については、論文の作成を進めている。

研究成果の概要（英文）：I have been developing a mapping program "maps" suite for next generation sequencing. I was able to achieve the following accomplishment during the 3 years period. (1) I have contributed the establishment of Gef-seq (Genome footprint sequence) analysis method, which is an improved method of ChIP-seq analysis, by preparing the computational analysis pipeline. (2) I continued the error-profile analysis of Illumina Next generation sequencer. (3) I have contributed the analysis of homologous/non-homologous rearrangements occuring in a part of organera genomes. (4) I contributed a comparative genome analysis of several yeast strains by genome assembling. For (1), we published a apaper and made our program public, and preparing another manuscript to describe the protocol. For (2), I made a couple of presentations. For (3) and (4), we are preparing for a publication.

研究分野：　生命情報学

キーワード：　次世代シーケンサー　イルミナ　Gef-seq　相同組換変異解析　転写因子　解析プログラム開発　エラープロファイル　ゲノムアセンブル

１．研究開始当初の背景

10 年ほど前から広く使われるようになったいわゆる「次世代シーケンサー」は 15 年程前には 3000 億円もの予算と何年もの歳月をかけて行われたヒトゲノムプロジェクトに匹敵する質と量のデータを数十万円と数日のコストで得ることを可能とし、分子生物学の研究に大きなインパクトを与えた。中でも米国イルミナ社のシーケンサーは得られるデータ量の大きさと、その冗長性により保証される高い信頼性から次世代シーケンサーのシェアの 75%を占めると言われ、現在、世界的に最も広く普及している。申請者は独自に maps と呼ぶマッピングプログラムを開発することで、このイルミナ社のシーケンサーの持つエラープロファイルを解析した。BWA や bowtie など通常よく使われているマッピングプログラムでは、リード配列を参照配列上にマッピングする際にギャップを含めたアラインメントを行う。これに対し、maps ではギャップアラインメントを行わずに、参照配列に対して一致する塩基の数が最も多くなるようにリードの対応する位置のみを決定する。その結果、イルミナ社のシーケンサーによるリードエラーの大きな要因であるフェーズシフトを含むエラープロファイルを特定することが出来た。この解析結果について、Nucleic Acids Resarch 誌に投稿し、現在までに被引用件数 385 件と、大きな反響を得ることができた。こうした、エラープロファイルの特定を進めることで、エラーが起こりうる箇所の予測を行うことができる。さらにはエラーの補正を行うことで、高性能な機器の有効性を更に高めることが可能になる。そこで、より詳細なエラープロファイルの特定を目指すこととした。

次世代シーケンサーの利用方法としては、その高性能な配列読み取りの力を活用して、本来の用途である、アセンブルやリシーケンシングなどのゲノム配列の決定という目的にとどまらず、メッセンジャーRNA の配列を読み取る RNA-seq や、転写因子等の DNA 結合タンパク質の認識配列を読み取る、ChIP-seq など、様々な、いわゆる-seq メソッドと呼ばれる新しい利用方法が次々と生みだされつつある。こうした、新しい次世代シーケンサーの利用方法の幾つかについて我々の開発するマッピング手法の特性を活かしたシーケンシングデータ解析手法の開発と発展を試みることとした。

２．研究の目的

本申請研究では先の研究で得た知見をもとに、次世代シーケンサーにより得られたデータを解析する為のプログラムの開発を継続しておこない、分子生物学研究の発展に寄与することを主な目的とする。

具体的な方向性としてはまず、先に述べたイルミナシーケンサーエラープロファイルの原因となる要素を解明し、エラーの発生箇所を予測することで、データの補正および個々のデータについて信頼性の評価を行うことを目指した。

また、実験を行う研究者との共同により、次世代シーケンサーを利用する新しい手法の開発に取り組むこととした。こうした手法の開発には、実験的な工夫だけでなく、データ解析を行うためのプログラムの整備が必要不可欠となる。そこで、我々の開発してきた maps プログラムの改良を進め、さらにその特性を活かすことで、新しい実験技術の開発・発展に貢献していくことを目指した。

３．研究の方法

マッピングプログラム maps の開発を中心に進めた。Maps は当初バクテリアゲノムへのイルミナシーケンサーリード（固定長リード）のマッピングを目的として開発したが、多染色体への対応、長さの異なるリードの混ざったデータへの対応、等の改良を進めより多様な解析への利用を可能とした。またマッピング結果の解析を行うプログラムの改良を進めることで、次世代シーケンサーデータを利用した新しい実験手法への対応を進めた。また、同時に解析結果を可視化するプログラムの開発を並行して進めた。

使用するデータとしては、SRA/DRA 等の公開データベースからダウンロードしたシーケンシングデータを用いたほか、それぞれのデータ解析手法の開発プロジェクトについては共同研究者により提供されるデータを使用して解析を行った。

プログラムの開発および解析計算の実行には、本研究費で購入した MacOS および Linux(Ubuntu)の動作するワークステーションを利用した。プログラム開発は主に C 言語(gcc)を使用し、一部 GUI の開発では Object-C を併用した。

４．研究成果

(1) DNA 結合タンパク質の認識配列を特定する方法として ChIP-seq という手法が知られている。タンパク質が DNA に結合した状態でクロスリンクを形成し、DNA 二本鎖を超音波などによって破砕し、免疫沈降によりタンパク質を回収し、同時に得られる DNA の断片の配列を読むことでタンパク質が認識して結合していた塩基配列を知ることができる。この手法では超音波による切断位置がまちまちであり、認識配列のみを含む正確な断片を得ることが困難であったため、DNAseI による DNA の切断を行うことで、タンパク質の認識配列を高い精度で特定する Gef-seq と呼ぶ手法が、奈良先端大の大島拓博士（現：富山大学）と石

川周博士（現：神戸大学）等により開発された（図１）。この Gef-seq により得られたデータについて、ドライデータ解析を行うためのプログラム開発をおこない、コンセンサス配列を抽出するためのプロシジャを確立した。Gef-seq 実験で得られる DNA 断片は一般にリード長より短く、残りの部分には特定のアダプタ配列が挿入される、maps プログラムにより多くのミスマッチを許容したマッピングを行うことで、一つのリードのマップ開始位置とアダプタ配列に変化して参照配列とのミスマッチが起きる位置との間がこのタンパク質の認識していた配列であると推定することができる。もう一つの手法としては、得られた配列断片の両端からペアエンド法によりシーケンシングを行うことで、ペアとなる向かい合うリード配列のそれぞれの初めの塩基位置に挟まれた領域として DNA 認識配列が特定できる。これにより１塩基単位の精度で DNA 認識配列を読み取ることが可能になった。この解析により、枯草菌の転写因子 ResD, NsrR, Fur についての認識配列の解析をおこない、それぞれの認識配列の配列モチーフに関する結果をまとめた論文が掲載された（研究成果１）。また、計算手順を含めたプロトコルの Methods in Molecular Biology 誌への投稿を準備している。

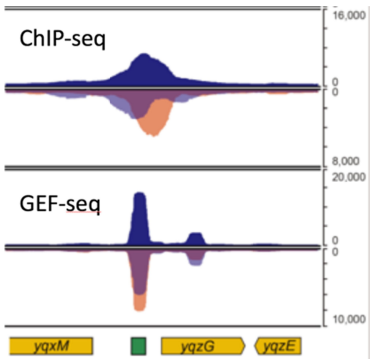

図１． ChIP-seq と Gef-seq の解像度の比較

(2) イルミナ社シーケンサーのエラープロファイル解析を継続して行った。先に発表した論文では、フェーズシフトによるエラーが発生する要因として、（１）原核生物のターミネーター配列として見られることの多い、長い（10 数塩基）インバーテッドリピート、（２）参照配列とのミスマッチが起きる箇所の上流に高い頻度で GCC トリプレットが観測されること、を報告した。一方で、比較的長いインバーテッドリピート配列の周辺でもミスマッチがほとんど起きないケースも観測されることや、GCC トリプレットは確率的には 64 塩基に一度は存在するが、実際にミスマッチが観測される頻度はずっと低いことから、これらの配列パターンは読み取りエラーを起こす要因の一つではあるが、十分条件とはなっていないことが示唆されていた。この状況では、エラーの予測ならびに補正を行うためには情報が不十分なため、継続してエラープロファイルの解析を進めた。一方向に配列特異的エラーが誘起される箇所（図２(a)）の収集を自動的におこない、これらの部位の上流にある配列に共通するモチーフを MEME により抽出することで以前より詳細な配列モチーフを得ることが出来た（図２(b)）。この結果については学会発表で報告したが、読み取りエラーの予測には依然不充分であるため、配列特異的エラープロファイルの特定へ向けた解析を継続して行っている。

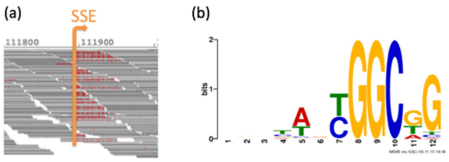

図２．(a) 配列特異的エラーと(b) 誘起パターン

(3) 次世代シーケンサーを利用して、参照ゲノム配列の存在する生物種について、株間等の小さな変異の有無を調べたい場合には、次世代シーケンサーによりゲノムシーケンシングを行ったリードデータをレファレンス配列上にマッピングすることで変異箇所を特定するリシーケンシングという手法が取られる。通常のリシーケンシングでは SNP や短い INDEL 等の微小な変異の特定が行われるが、相同組換え等、離れた位置での DNA 配列の組換による転座や繰り返しの解析は容易ではない。一般にペアエンドリードの対応するリードのマップ位置からの推定が行われるが、この手法では組換点の特定を簡便に行うことが出来ない。組換点を含むリードは、リードの前半と後半で参照配列上の対応する位置が異なり、この様なリードをキメラリード、あるいはジャンクションリード、と呼ぶ。通常のマッピングソフトウェアによるデフォルトのマッピングではこのようなリードデータは排除されることが多いが、われわれの maps を用いて多くのミスマッチを許容するマッピングを行うことで、キメラリードについても一致配列が長い側の部位へのマッピングが行われる。こうしたリードのミスマッチ配列についてのコンセンサスを取り、(2)で述べたシーケンシングエラーなどの可能性を排除するための一定のクライテリアを満たす配列として、参照ゲノム配列上に探索することで、離れた位置での組換を同定することができる。この手法により、サンプル中の全

ててはなく、ごく一部分の配列に起きている組換をも検出する事ができる。現在、共同研究者による実験が進行中であり、我々の開発したプログラムによるデータ解析を行った上で論文の投稿をおこなう予定である。

(4) これまで、次世代シーケンサーデータのマッピングデータを扱う上で蓄積してきたエラープロファル等に関する知見を活用して、類似した参照配列が存在しない場合に行う配列アセンブルについても解析手法の開発を進めている。特に倍数体の様に類似した配列が複数系統混在する場合の解析は困難であるが、まず de Bruijn 法によりアセンブルしたコンティグおよびスキャッフォルドへのペアエンドリードデータのマッピングを行うことで、組み換えによる遺伝子位置の相違の特定や、繰り返し回数の特定等の解析を行うためのプログラム開発を行っている。開発手法の具体的な適用対象として、学内の研究者と協力して雑種２倍体を含む酵母のゲノム解析を行っており、その結果について論文の投稿準備を進めている。

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１件）

Onuma Chumsakul, Divya P. Anantsri, Tai Quirke, Taku Oshima, Kensuke Nakamura, Shu Ishikawa and Michiko M. Nakano, “Genome-Wide Analysis of ResD, NsrR, and Fur Binding in Bacillus subtilis during Anaerobic Fermentative Growth by In Vivo Footprinting”, Journal of Bacteriology, 199(13), e00086-17, 2017.

〔学会発表〕（計３件）

中村建介・松本秀太　“Gef-seq に対応したマッピングプログラム及び GUI の開発” 第９回ゲノム微生物学会年会 2015 年 3 月 6 日 神戸大学

松本秀太・中村建介　“次世代シーケンサーのベースコール精度の検証” 第 38 回分子生物学会年会　2015 年 12 月 1 日　神戸ポートアイランド

中村建介・竹内敬一郎・松本秀太　“シーケンサー配列解析ソフトウェア maps の開発とシーケンシングエラープロファイル解析” 第 39 回分子生物学会年会 2016 年 12 月 2 日 パシフィコ横浜

〔図書〕（計　１件）

大島拓・石川周・Onuma Chumsakul・中村建介
「高精度で結合領域を決定する Gef-seq」
次世代シークエンス解析スタンダード～NGS のポテンシャルを活かしきる WET&DRY pp.131-142, 羊土社　2014 年

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ウェブサイト：
http://metalmine.mydns.jp/maps/gefseq/

６．研究組織
(1)研究代表者
中村　建介　（Kensuke Nakamura）
前橋工科大学・工学部・教授
研究者番号：２０２１２０９５

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）