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　３０　年　　６　月　１４　日現在

機関番号：　８２４０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６４３０２００

研究課題名（和文）未知ゲノムの解読：ドラフトレベルのアセンブリ配列を改築するシステムの開発

研究課題名（英文）Development of system to improve the quality of draft assemblies

研究代表者
小杉　俊一（Kosugi, Shunichi）

国立研究開発法人理化学研究所・統合生命医科学研究センター・研究員

研究者番号：３０３６５４５７

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究で取得したイネデータ（イルミナショートリード、PacBioロングリード）を含めた全ゲノムシークエンスリードを用いて、酵母、線虫、イネのゲノムアセンブリを構築し、そのエラー特性等の統計値を計測した。コンティグやリードのアライメント特性についての真・偽アライメント間の尤度比を利用することにより、アセンブリ中のギャップ領域を高精度で修復するツール（GMcloser）を開発した。そのほかにアセンブリ中のミスアセンブリを計測するツール（GMvalue）やアセンブリを伸張・再アセンブルするツール（Exterm）の開発を行った。

研究成果の概要（英文）：I constructed genome assemblies with the whole genome sequencing reads from budding yeast, nematoda, and rice, and determined the statistics for the metrics and error properties of the assemblies. Using the likelihood ratio of the alignment properties (i.e., alignment length, alignment identity, and mapping rate of paired-end short reads) between the true and false alignment data, I developed a gap-closing tool, GMcloser, that closes gaps in assemblies with high accuracy. In addition, I developed another too tools, GMvalue and Extrerm. GMvalue determines the metrics of misassemblies including the number of misassembles. Extreme reassembles a preassembled config set by extending the termini of the contigs followed by assembling the extended contigs with overlap-layout-consensus algorithm.

研究分野：　ゲノム科学

キーワード：　アセンブリ　ゲノム

１．研究開始当初の背景

新規ゲノム配列の構築は、近年の次世代シークエンシング技術の発達によって、より高速かつ安価に進めることが可能となっている。しかし、従来のサンガー法によるシークエンシング手法に比べて、次世代シーケンシングリードの de novo アセンブリングによるゲノム構築は、アセンブリのエラーが多く、繋がりも悪く、ギャップも多い。特に、スキャッフォールドにおけるコンティグ連結エラーは、全スキャッフォールド中の数十%に及ぶことがある。この理由は、次世代シークエンシング技術では相対的にシークエンシングエラー率が高くなるためであるが、現在の de novo アセンブリング手法の特性から、ゲノムのリピート配列およびヘテロ接合度合いが高くなるほどエラー率は高くなる。逆にこのような配列特性は、de novo アセンブリングの配列伸長を停止させ、アセンブリの連結性を低下させる。このため、複雑度が高くゲノムサイズの大きなゲノムほどアセンブリのエラー率は高くなり、連結性も低くなる。ミスアセンブリングやヘテロな遺伝子型の誤配置が遺伝子領域内に生じると、実際には存在しない翻訳産物やキメラ遺伝子が生じるなど、その後の解析に大きな影響を及ぼす結果となる。

このため、解析に資するゲノム配列を得る為には、アセンブリ配列に存在するエラーとギャップを修復することが最重要となる。これらの過程に関わる現在の手法について以下に解説する。

① ミスアセンブルの検出・修復　ミスアセンブリの正確な特定は、ミスアセンブリの殆どないリファレンスゲノムが構築された生物種でのみ可能である。リファレンスの無い生物種では、リードをマッピングすることによってミスアセンブリの箇所を類推することしかできない。この類推法（ALE, CGAL, Reapr 等）では、アセンブリ配列にアライメントされたリードのカバレッジが特に低くなった領域をミスアセンブリ箇所の候補とする。しかしこの手法は、リードがゲノム全域に渡って均一に読まれていることを前提としており、実際にはエラーのないリファレンスゲノムにリードをアライメントした場合にもカバレッジの極端に低くなる領域が多く存在する。さらに、これらのツールをエラー修復したイネアセンブリ配列に適用すると、多くの擬陽性（ミスアセンブリでない箇所をミスアセンブリとして検出する数）を検出する結果を得ている。

② ギャップの修復　現在用いられているギャップを埋めるツール（GapCloser, GapFiller 等）は、アセンブリ配列にペアードエンドリードをアライメントし、ギャップ周辺にアライメントされたリードのアセンブリングによってギャップの修復を行う。この手法は効率的であるがエラーが多い。申請者が現在開発を進めているギャップ修復ツール GMcloser は、コンティグセットとペアードエンドリードを用いることで、精度の高いギャップ修復を可能にしている。

ギャップ修復のこの工程は、エラー検出とカップルしており、ギャップ修復の過程で得られるアライメントデータから、ミスアセンブリ箇所を推定することができる。また、①②の両工程において最も重要な作業は、paired-end リードの正確なアライメントデータを得ることである。申請者は、リードのアライメントデータの質を向上させ DNA 多型を精度良く検出する手法を開発しており（文献 1, 2）、本手法を取り入れることによって精度の高いアライメントデータを得ることができる。

＜引用文献＞

(1) Abe A. and Kosugi S. et al., Nat. Biotechnol. 30, 174-178 (2012)
(2) Kosugi S. et., PLoS One 8, e75402 (2013)

２．研究の目的

本研究では、高等真核生物を含むドラフトレベルのゲノム配列について、エラーの検出と修復、ギャップの修復、再アセンブリングの 3 工程を互いに連携させながら一連の作業を繰り返すことで配列を改築する-アセンブリ配列フィニッシャーパイプライン-を構築することを目的とする。これらの各工程では、実データの統計解析を基に計算した尤度スコアを用いて正確性の判別を行うアルゴリズムが確立される。このようなシステムはこれまでに無く、実データの統計解析データを基にした尤度判定手法を用いて、正確で高効率のツールを作製することを目標にする。

３．研究の方法

構築するアセンブリフィニッシャーパイプラインでの各 3 工程（ミスアセンブル箇所の修復、ギャップの修復、再アセンブル）を効率良く、精度高く実行するため、尤度に基づいた正確性を判定するアルゴリズムを確立する。このアルゴリズムは、種々のデータセット（HiSeq 及び PacBio リード、RNA-seq データ、cDNA 配列、アセンブリ配列セット等）についてのアライメントデータを統計解析し、各データ項目について計算した尤度比を基にする。各工程での判別に用いられる尤度値や関連するデータは、他の工程でも利用され、複数のデータを統合した判別子を確立することによって、より精度高く実行するプログラムを開発する。さらに、アセンブリの質をさらに向上させるため、これら一連の作業を繰り返すプログラムを構築する。

構築するパイプラインの各 3 工程についてのアルゴリズムを確立するため、線虫とイネをモデルとするデータセットの作製を進める。

① 線虫およびイネアセンブリ配列の調製
ミスアセンブリについての統計データを得るため、リファレンスゲノムが構築されている線虫（ゲノムサイズ：100 Mb）とイネ（ゲノムサイズ：380 Mb）の次世代シークエンシングリードを取得し、複数のアセンブリングツールを用いて異なるアセンブリセットを調製する。イネゲノムのデータは、そのゲノムサイズとリピート含有率の高さから、多くの高等真核生物ゲノムのアセンブリ構築のモデルとなるため、必須のデータとなる。SOAPdenovo、ALLPATHS-LG、Newbler等を用いて、取得したリードセット（PacBioリードを除く）のアセンブリングを行い、線虫およびイネの各種アセンブリセットを調製する。

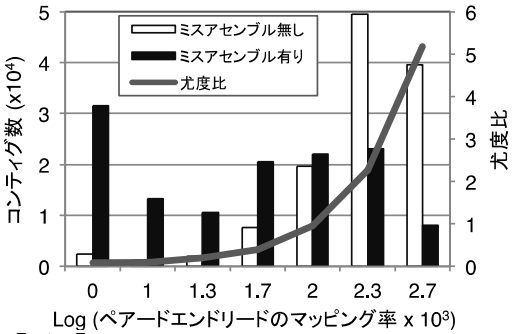
② 線虫およびイネアセンブリ配列中に存在するミスアセンブル箇所の特定
各アセンブリセットを線虫またはイネリファレンスゲノムにアライメントし、断片的に異なるリファレンス領域にアライメントされたミスアセンリ配列とそのミスアセンブル部位を特定する。さらに、ミスアセンブル部位の周辺配列の特性を調査し、リファレンスと異なるSNPやindelの部位も特定する。SNP/indelの特定にはCoval（研究背景・文献2）を用いる。

③ アセンブリ配列へ各種リードおよびcDNA配列のマッピング
アセンブリングに用いたリードセットをアセンブリ配列にアライメントし、リードカバレッジやリードペアの整合性についてのデータを得る。PacBioリードは、HiSeqリードを用いてエラー補正後使用する。また、cDNA配列や近縁種で同定されているタンパク質配列についてもアセンブリ配列へアライメントを行い、遺伝子構造を推定したアライメントデータを取得する。同時にそれぞれのリードセットはリファレンスにもアライメントを行い、リードのカバレッジ特性を取得する。アライメントデータはCovalを用いて精度の改善を行う。

④ アセンブリデータセット間でのアライメント
アセンブリングに用いるリードセットやツールの違いによって、ミスアセンブリングの起こる部位が異なる場合がある。このことを利用して、アセンブリデータセット間でアセンブリ配列のマルチプルアライメントを行い、アライメントの比較からミスアセンブルが存在する部位の候補を収集する。

取得した種々のアライメントデータについて、尤度に基づいた正確性判定アルゴリズムを確立し、フィニッシャーツールの作製を行う。
⑤ 各種アライメントデータの統計解析
②~④の工程で取得したデータを基に統計解析を行う。例えば、ゲノムのGC含量やリピート配列の存在、マップされたリードのカバレッジ、ペアードエンドリード、RNA-seqリード、PacBio長鎖リードのマッピング率などの頻度分布、またはこれら因子を組み合わせた頻度分布をミスアセンブルの存在の有無に分けて解析を行う。

⑥ 尤度に基づいた正確性判定アルゴリズムの確立
上記の頻度分布がミスアセンブルの有無で差が生じた場合、ミスアセンブルか否かを判定するための尤度比を得ることができる。申請者がこれまでに行った１つの例を【図 1】に挙げる。この例は、シミュレーションリードで作製したイネアセンブリセットにペアードエンドリードをマッピングしたとき、特定の位置を挟んでマップされるペアードエンドリードの頻度分布をグラフにしたものである。この特定位置にあたる部位がミスアセンブル部位に相当するとき（黒棒）とそうでないとき（白棒）では頻度分布に明らかな差があり、計算した尤度比（灰色線）を基にミスアセンブルがその特定部位に存在するか否かを確率的に判定出来る。このような尤度比を複数の因子について取得し、それらの積算をとって融合することにより、より判定の精度を高めたアルゴリズムを確立する。



【図 1】ミスアセンブルの有無によるイネコンティグのペアードエンドリードマッピング率の頻度分布と尤度比
ミスアセンブル部位を挟んでマッピングされるペアードエンドリードのマッピング率の頻度分布（黒棒）とミスアセンブルを含まないものの頻度分布（白棒）から計算された尤度比（灰色線）は、目的の部位がミスアセンブル部位であるか否かを判定する指標となる（この例の場合、尤度比が低いほど、ミスアセンブル部位の可能性が高くなる）。

⑦ アセンブリフィニッシャーパイプラインの作製
確立したアルゴリズムを基に、ミスアセンブル箇所の特定・修復、ギャップ修復、処理後のアセンブリセットの再アセンブルを実行するツールを作製する。これらの過程を単独で処理出来るツールを作製するとともに、各工程での判別に用いられる尤度値や関連するデータを他の工程でも利用できようにし

た、一連の工程を連携して繰り返し処理するパイプラインを作製する。

４．研究成果

① イネアセンブリ配列セットを構築するために、イネ次世代シークエンシングリードを取得した。リードは、イルミナ HiSeq ペアードエンドリード（インサート長：500 bp)、メイトペアリード（インサート長：3 Kb、10 Kb)、および PacBio リードを受託シークエンシング解析サービスを利用して取得した。イルミナリードに関しては、得られたリードを用いて SOAPdenovo2、Platanus、FERMI 等を用いてデノボアセンブリを行った。また、イネのオーバーラップ paired-end リードについては、公共データベース（Sequence Read archive）から取得できため、そのリードを用いて ALLPATHS-LG によるアセンブリを行った。また、イネ PacBio リードについては、種々のエラー修正ツール（PacBioToCA、Proovread、ECtools 等）を用いてエラー修正を行い、エラー修正効率、精度等を計測した。線虫に関しては、イルミナ HiSeq ペアードエンドリード（インサート長：500 bp)、メイトペアリード（インサート長：2 Kb）を Sequence Read Archive から取得し、イネと同様にデノボアセンブリを行った。

② 得られたそれぞれのアセンブリ配列セットについて、配列全長や N50 値などの統計値を計測し、イネのリファレンス配列を用いてミスアセンブリ数およびミスアセンブリ部位の特性やそれらのカテゴリー集計を行った。さらに、各アセンブリセットにイルミナペアードエンドリードをアライメントし、アライメントカバレッジ特性を解析するとともに、既存のミスアセンブリ検出ツール（Reapr、CGAL 等）を用いて、ミスアセンブリ部位の推定を行い、ミスアセンブリ部位についての情報を収集した。

③ 構築したイネゲノムアセンブリデータセットを含め、線虫や酵母等のゲノムアセンブリデータを用いて、リファレンス配列に アライメントを行うことにより、ミスアセンブリ部位の同定を行った。ミスアセンブルを含むアセンブリ配列セット（True contig set）とミスアセンブリを含まないアセンブリ配列セット（False contig set）に分割し、それぞれのデータセットにおけるコンティグ 間のアライメントの (1) オーバーラップ配列長、(2) オーバーラップ配列相同性、および (3) paired-end リードのマッピンクグ率（forward read, reverse read の 2 コンティグ間サポートマッピンクグ率）を計測し、統計解析を行った。これらの 3 つの統計量について 、True contig set と False contig set 間で尤度比を算出し、積算したスコアを用いることで、ミスアセンブリを含むコンティグアライメントとミスアセンブリ含まないコンティグアライメントを精度高く識別することができた。この尤度を利用したアルゴリズムを用いて、ゲノムアセンブリに存在するギャップをコンティグや長鎖リードを用いて精度高く修復するツール（GMcloser）を開発した。GMcloser は、既存のギャップ修復ツールと比較して同等なギャップ修復効率を維持しながら、5~100 倍高い修復精度を示した。 さらに、ゲノムアセンブリ中のミスアセンブリの数、種類、部位について計測するツール（GMvalue）を開発した。本ツールは、アセンブリ配列をリファレンスにアライメントすることでミスアセンブリ部位を同定する手法を取り、既存の同様のツールでは困難な 300 Mb 以上の大きなサイズのゲノムにも適用できることや、ミスアセンブリを定義、同定するための細かな設定変更が出来ることが特徴である。

④ これまでに構築したイネ、線虫、酵母のゲノムアセンブリデータを用いて、ペアードエンド短鎖リードを bwa を用いてアライメントし、コンティグアセンブリ配列中のミスアセンブリ部位に存在する特性を調査した。各タイプ (local misassembly, translocation, inversion 等)のミスアセンブリ領域とミスアセンブリの無い領域間でのリードの (1) カバレッジ、(2) discordant リード(元のインサートサイズと大きく異なるリードペアやアライメントの方向が forward-reverse になっていないリードペア)の割合、(3) split（soft-clipped）リードの割合について統計的調査を行なった。その結果、(1), (2), (3) いずれの因子においてもミスアセンブリ領域とミスアセンブリの無い領域間で差が認められた。 近年第 3 世代シークエンシング技術により得られる長鎖リードを用いたゲノムアセンブリの構築が盛んに進められており、短鎖リードを用いたアセンブリに比べて格段に長いコンティグ配列が取得されている。しかし、長鎖リードは 15%程度の高いエラーを含み、このエラーが十分に修正されずにアセンブリ中に残ってしまうことが多い。長鎖リードのエラー修正効率、修正精度を調べるため、イネ 、線虫、酵母の長鎖リード（PacBio リード）を用いて、既存のエラー修正ツールのエラー修正精度を調査した。その結果、調べた全てのツール（特に短鎖リードを用いたハイブリッドエラー修正ツール)では、リード中の多くのエラーを十分に修正しきれておらず、ゲノムサイズが大きいほどエラー修正率は低下し、ツールの実行時間が極度に長くなることが観察された。

⑤ この他に、ドラフトレベルのアセンブリを再アセンブルするツール（Exterm）の開発を行った。Exterm は、各コンティグ末端領域に paired-end short reads のアライメントを行い、得られたアライメント情報を基にした local assembly によってコンティグ末端

を伸張させ、末端を伸張させたコンティグ同士をoverlap-layout-consensusアルゴリズムを用いて連結する。本過程でミスアセンブリを極力低減させるため、以下の処理を行った(1) コンティグリピート領域のマスク、(2) アライメントしたリードのフィルタリング、(3) 伸張させたコンティグ配列が他のコンティグと end-to-end でアライメントされない場合、伸張させた配列の除去、(4) コンティグ同士を連結させたアセンブリ配列に再度paired-end short readsのアライメントを行い、不合理なリードペアのアライメントが認められた際、コンティグの連結を解除する。これらの処理を行わなかった時と比較すると、約95%のミスアセンブリを除去することができた。酵母、線虫、イネの種々のヘテロ接合度を導入した人工アセンブリセットおよびリアルアセンブリセットを用いてExterm の性能を計測したところ、ミスアセンブリの生成を抑えつつ、最大で13倍のN50増加を生じた。特に、ヘテロ接合性の高いゲノムのアセンブリはより断片化する傾向にあるため、本ツールはヘテロ接合性の高いゲノムの構築に有用性を発揮すると期待される。

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 8 件）

① Akira Abe, Hiroki Takagi, Satoshi Natsume, Hiroki Yaegashi, Hideko Kikuchi, Kentaro Yoshida, Shunichi Kosugi, Hiromasa Saitoh, Naoya Urasaki, Hideo Matsumura, Hiroyuki Kanzaki, and Ryohei Terauchi, Rice breeding based on whole genome sequencing using next-generation sequencer. Seikagaku, **査読有**, vol88, 2016, 44-53
DOI: 10.14952/SEIKAGAKU.2016.880044

② Kentaro Yoshida, Diane G. O. Saunders, Chikako Mitsuoka, Satoshi Natsume, Shunichi Kosugi, Hiromasa Saitoh, Yoshihiro Inoue, Izumi Chuma, Yukio Tosa, Liliana M. Cano, Sophien Kamoun, Ryohei Terauchi, Host specialization of the blast fungus Magnaporthe oryzae is associated with dynamic gain and loss of genes linked to transposable elements. BMC Genomics, **査読有**, vol17, 2016, 370
DOI: 10.1186/s12864-016-2690-6

③ Hidenori Tanaka, Hideki Hirakawa, Shunichi Kosugi, Shinobu Nakayama, Akiko Ono, Akiko Watanabe, Masatsugu Hashiguchi, Takahiro Gondo, Genki Ishigaki, Melody Muguerza, Katsuya Shimizu, Noriko Sawamura, Takayasu Inoue, Yuichi Shigeki, Naoki Ohno, Satoshi Tabata, Ryo Akashi, Shusei Sato, Sequencing and comparative analyses of the genomes of zoysiagrasses. DNA Research, **査読有**, vol23, 2016, 171-180
DOI: 10.1093/dnares/dsw006

④ **阿部 陽, 高木 宏樹, 夏目 俊, 八重樫 弘樹, 菊池 秀子, 吉田 健太 郎, 小杉 俊二, 齋藤 宏昌, 浦崎 直也, 松村 英生, 神崎 洋之, 寺 内 良平、次世代シークエンサーを活用した全ゲノム解析によるイネ育種、生化学、査読有、88 巻、2016、44-53**
DOI: 10.14952/SEIKAGAKU

⑤ Shunichi Kosugi, Hideki Hirakawa, and Satoshi Tabata, GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long- read alignments. Bioinformatics, **査読有**, vol33, 2015, 445-449
DOI: 10.1093/bioinformatics/btv465

⑥ Hideki Nishikawa, Takuro Iijima, Rei Kajitani, Junichi Yamaguchi, Toshiya Ando, Yutaka Suzuki, Sumio Sugano, Asao Fujiyama, Shunichi Kosugi, Hideki Hirakawa, Satoshi Tabata, Katsuhisa Ozaki, Hiroya Morimoto, Kunio Ihara, Madoka Obara, Hiroshi Hori, Takehiko Itoh, and Haruhiko Fujiwara, A genetic mechanism for female-limited Batesian mimicry in Papilio butterfly. Nature Genetics, **査読有**, vol47, 2015, 405-409
DOI: DOI: 10.1038/ng.3241

⑦ Shunichi Kosugi, Hiroshi Yanagawa, Ryohei Terauchi, Satoshi Tabata, NESmapper: accurate prediction of leucine-rich nuclear export signals using activity-based profiles. Plos Computational Biology, **査読有**, vol18, 2014, e1003841
DOI: 10.1371/journal.pcbi.1003841

⑧ Hiroki Takagi,Muluneh Tamiru, Akira Abe, Kentaro Yoshida, Aiko Uemura, Hiroki Yaegashi, Tsutomu Obara, Kaori Oikawa, Hiroe Utsushi, Eiko Kanzaki, Chikako Mitsuoka, Satoshi Natsume, Shunichi Kosugi, Hiroyuki Kanzaki, Hideo Matsumura, Naoya Urasaki, Sophien Kamoun, and Ryohei Terauchi, MutMap accelerates breeding of a salt-tolerant rice cultivar. Nature Biotechnology, **査読有**, vol33, 2014, 445-449
DOI: 10.1038/nbt.3188

〔学会発表〕（計 1 件）

① Shunichi Kosugi, Yoichiro Kamatani,

Comprehensive evaluation of strucutral variation calling tools and development of an integrated pipeline for merging calling results. **第 5 回生命医薬情報連合大会**、2016

〔図書〕（計　0 件）

〔産業財産権〕

○出願状況（計　0 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　0 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〕
ホームページ等
GMcloser ダウンロードサイト：
https://sourceforge.net/projects/gmcloser/

６．研究組織
(1)研究代表者
小杉　俊一（KOSUGI Shunichi）
国立研究開発法人理化学研究所・統合生命医科学研究センター・研究員
研究者番号：30365457

(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
（　　）