様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １０ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)
研究期間：2012～2014
課題番号：２４３１０１４２

研究課題名（和文）ヘテロ接合性を考慮した次世代シークエンサ用ゲノムアセンブラ／遺伝子予測手法の開発

研究課題名（英文）Development of denovo genome/gene assembler for highly heterozygous samples from NGS sequence data

研究代表者
伊藤　武彦（ITOH, TAKEHIKO）

東京工業大学・生命理工学研究科・教授

研究者番号：９０５０１１０６

交付決定額（研究期間全体）：（直接経費）　　16,100,000円

研究成果の概要（和文）：高ヘテロ接合性を持つ二倍体ゲノムのアセンブルはde bruijnグラフが複雑化するため極めて困難な課題であるが、非モデル生物ゲノムの配列決定への要求は極めて高い。そこで、contig作成時、scaffold作成時の双方でヘテロ接合性に起因したグラフ構造の単純化を図る事によりこの問題をクリアした新規Platanusアセンブラを開発した。
Platanusアセンブラはシミュレーションデータ、実高ヘテロ接合性サンプルであるベネズエラ糞線虫データ双方に対して、既存アセンブラと比べて高い精度で長く繋がった結果を得る事に成功した。

研究成果の概要（英文）：Assembling the highly heterozygous diploid genomes is a big scientific challenge due to the increased complexity of the de Bruijn graph structure. To deal with an increasing demand for sequencing of non-model and/or wild-type sample, we developed a novel de novo assembler, Platanus, which can effectively manage high-throughput data from heterozygous samples. Platanus assembles DNA fragments into contigs by constructing de Bruijn graphs, followed by scaffolding of contigs based on paired-end information. The complicated graph structures that result from the heterozygosity are simplified during not only the contig assembly step but also the scaffolding step.
We evaluated the assembly results on eukaryotic samples with various levels of heterozygosity. Compared with other assemblers, the Platanus assembly results have a larger NG50 length without any accompanying loss of accuracy in both simulated data and real data including highly heterozygous Strongyloides samples.

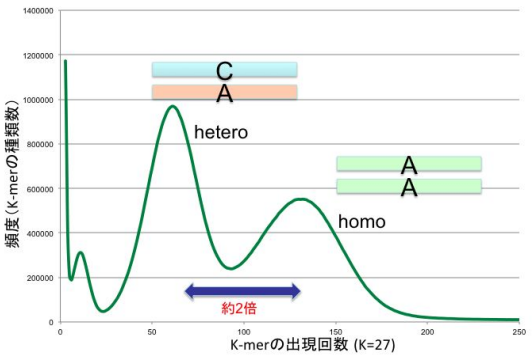
研究分野：ゲノム情報

キーワード：ゲノム情報　バイオインフォマティクス

１．研究開始当初の背景

Roche 社 454、Illumina 社 GA などのいわゆる「次世代シークエンサ」は 2005-6 年頃に登場した。これらのシークエンサは、それまで主流であった Sanger 法シークエンサと比べ大規模な並列化が可能であり、2012 年当時で一度に 650Mb(454 Titanium+)や、300GB(Illumina 社 Hiseq 2000)もの塩基配列が得られるようになっていた。特に Illumina 社プラットフォームでは塩基単価が極めて低いため、本来の目的であるゲノム既知生物に対するマッピング解析に留まらず、新規ゲノム配列決定への試みがなされ始めており、Velvet、ABySS などのアセンブラ使用により、バクテリアなど小型ゲノムへの適用が、SOAPdenovo、Allpath-LG などの登場により、パンダ、ヒト、マウスといった Gb オーダーのゲノムサイズを持つ真核生物の新規解読への適用が報告されるようになってきていた。

このような現状を踏まえ、申請者らも分担研究者らと共にベネズエラ糞線虫やヒトデゲノムの解読を、Illumina プラットフォームによるシークエンス、既存アセンブラの利用により試みた。しかし N50 は 200-800bp にとどまり、ほとんど繋がらない結果しか得ることができなかった。様々な角度から原因究明を試みたところ、扱ったゲノムの「ヘテロ接合性」の高さが主な原因であることが判明してきた。下図は、得られたシークエンス全 read を 27bp に分割し、ある 27bp が全 read 中に何回出てくるかを調べ、横軸に出現回数、縦軸に頻度(27bp の種類数)を取ったものである。



一般的なゲノム解読時には、平均 redundancy を中心とした正規分布様を示す。しかし、ベネズエラ糞線虫のゲノムでは下図のように二山が認められ、その peak はほぼ２倍の出現回数位置に存在する。これは図中の左の山がゲノム中ヘテロな領域、右の山がホモな領域由来の 27mer であることに起因すると考えられる。実際左の山に含まれる 27mer を精査すると 1bp 違いのペアが非常に多数見つかることもこの説を強力にサポートする。

上で示したアセンブラはいずれも計算時間、使用メモリ量の観点から de bruijn グラフアルゴリズムを利用することでアセンブルを可能にしている。しかし、その場合１塩基の多型が存在してもグラフ構造が分岐を持ったバブル構造となり、分岐の前後でアセンブルを切ってしまう。これを既存のプログラムの利用で解決することはほぼ不可能であり、新規プログラムの開発が必須である。ジャガイモゲノム解読に関する論文においても、栽培に広く用いられている株にはヘテロ接合性があるため、そのまま解読することは現時点の技術では無理であり、ホモ接合性倍加一倍体を作成、そのゲノムを解読／アセンブルし、その結果上にヘテロ接合性二倍体クローン由来のデータをマッピングすることで解析がようやく可能になったと述べられている。また同様にカキゲノムの解読では、一旦 Fosmid ライブラリを作成し、クローン毎のシークエンスを実施している。

以上示したように、Illumina プラットフォームの利用による新規ゲノム決定が成功するケースはモデル生物や近交系が確立された生物をターゲットにした稀なケースであり、野生種など多くの生物を対象としたゲノム決定においては、ヘテロ接合性の高さに起因して極めて困難な状況であった。

２．研究の目的

研究開始当初の背景で示した通り、ヘテロ接合性の高いゲノムを既存の手法でアセンブルすることは現時点ではほぼ不可能である。まず、アセンブルがうまくいっていないデータの特徴／理由をさらに精査するとともに、その解析結果を踏まえ、de bruijn グラフに分岐構造を組み込むことにより、ヘテロ接合性の高いゲノムに適応した short-read 用新規 denovo ゲノムアセンブラを開発することを第一の目的とする。当然このアセンブラは接合性の低いゲノムにも適応可能である。

次に、分担研究者である丸山協力の元、本研究開始の動機に繋がったベネズエラ糞線虫ゲノムの解読、アセンブル、アノテーションを行い、ヘテロ接合性の高い実サンプルでの解析を実施することを第二の目的とする。

本研究の最大の特徴は、シークエンス費用の低下に伴って今後益々増加するであろう、近交系が確立されていない非モデル生物や野生種など、ヘテロ接合性の高いゲノムに対応したゲノムアセンブラを広く研究者に提供できることにある。このようなヘテロ接合性の高いゲノムに対して有効なアセンブル手段を提供する手法は世界的に見ても皆無であり、非常に独創的である。さらに本研究では、理想的なシミュレーションデータではなく、実験研究者と組む事でベネズエラ糞線虫の実データから如何に正

しい解析結果を導くかに重点を置いて計画を立案しているため、幅広く実験研究者の研究推進に資することが可能であると考えられる。

3．研究の方法

ヘテロ接合性が高いゲノム対応のルーチンを組み込むために、まずその基礎となるshort-readを入力としたゲノムアセンブラの研究開発を実施した。アセンブラの基本アルゴリズムには、overlap-layout-consensus, greedy-graph, de bruijn graphなどが存在するが、本研究では最終的なターゲットとして、ヒトなど数Gbのゲノムを持つ生物種を想定しているため、計算時間、使用メモリ量の観点から、de bruijn graphを採用した。本アルゴリズムは”An Eulerian path approach to DNA fragment assembly”として発表されたのが最初であり、全readをk-merに分割後、全k-mer(node)を通るEulerパス問題へとゲノムアセンブル問題を置き換えることでアライメントが不必要となり、効率的にcontig構築が可能となる。

また、contig構築後short-readのペア情報に基づいたscaffolding、gap-closeを実現するルーチンも合わせて開発し、contig構築、scaffold構築、gap-closeの4モジュールからなるde bruijnアルゴリズムに基づいたゲノムアセンブラ開発を行った。

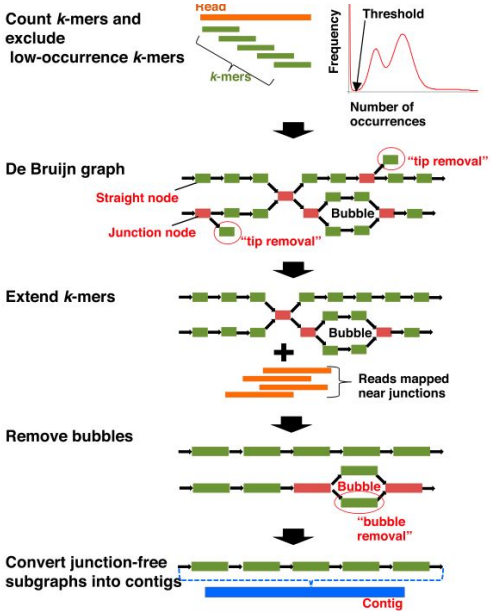
一方、ヘテロ接合性が高いことが先行研究にて示されているベネズエラ糞線虫ゲノムについて、その特徴を事前に取得したデータを基により精査した。具体的には、illuminaによるWGSデータとは別に、Fosmidを作成しそれをシークエンスする事でハプロタイプ別の配列を取得し、Fosmid由来の配列に対して両染色体由来のWGSデータをマッピングし、k-merグラフが分岐構造になる箇所を緻密に調べ上げることで、アセンブラに組み込むべきグラフの形状／特徴を調査した。調査の結果を踏まえ、あらかじめ開発したde bruijnグラフベースのゲノムアセンブラに高ヘテロ接合性対策サブルーチンを組み込む事で目的のアセンブラ開発を実施した。

アセンブラ開発後は、シミュレーションデータなど解がわかっているデータに対してのベンチマークテストを繰り返し行う事でアルゴリズムの改良を続け、最終的にPla公開可能なプログラム開発を実現した。

4．研究成果

最終的に、コンティグアセンブル、スキャッフォールディング、ギャップクローズの3つの工程からなるPlatanusアセンブラの開発に成功した。以下にその概要を説明する。

コンティグアセンブルでは、入力リードからk-merを節点、k－1のオーバーラップを辺としたde Bruijnグラフを構築する。

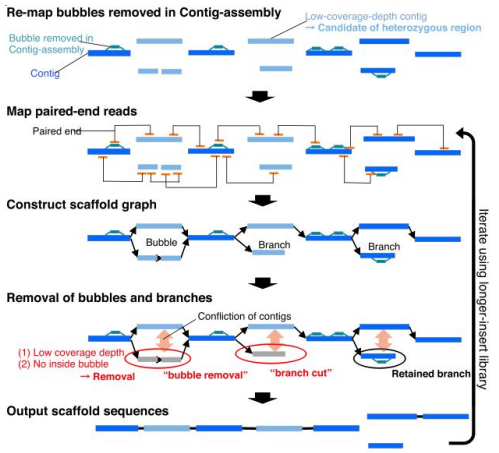


メモリ使用量を減らすためグラフ中で分岐のない領域の節点は1つに圧縮され、1つの節点（straight node）として扱われ、最終的に、straight nodeがアセンブルされた配列（contig）に対応する。分岐を持つ節点はjunction nodeと呼ばれ、straight nodeの組において両側の隣接節点が共有されている場合、バブル構造と呼ぶ。この構造はSNV、small indel、エラー等に起因して発生する。グラフの構築後、バブル構造に含まれない節点からcoverage depthの平均を算出し、この値をホモ領域（両方の相同染色体に存在する領域）のcoverage detphと判断する。その値よりも出現回数の低いk-merはシークエンスエラーに起因した枝構造と判断して除去、さらにはバブル構造領域がホモ領域の約半分の出現頻度から構成されている場合には、相同染色体の違いによるものと判断する等に用い、バブル構造を形成するstraight nodeの片側を一旦除去する事でcontigを長く伸長することに成功した。

また、straight nodeをcontigとして出力する場合、kより長いリピート配列については解決できないという問題が生じる。そのため、kが大きい方がゲノム中のリピート配列に対応するには適しているが、データ量が少なくcoverage depthが低い場合にはギャップが多くなる。これは、リード間のkより短いオーバーラップを検出できなくなるためである。Platanusは複数のkの値の利点を活用し、k0（デフォルト32）でde Bruijnグラフを構築した後、kstep（デフォルト10）ずつ増加させながらグラフを再構築していくことでこの問題にも対処している。

スキャッフォールディングでは、最初に

contig 上に paired-end または mate-pair データをマッピングする。これにより contig の位置情報を特定し、scaffold 配列を構築するために用いることができる。ヘテロ領域については、contig 配列が片方のハプロタイプに対応するので、もう片方のハプロタイプ由来のリードは適切にマッピングされない可能性が存在する。そこで、Contig-assembly の際に除去されたバブルの配列も合わせて利用する。



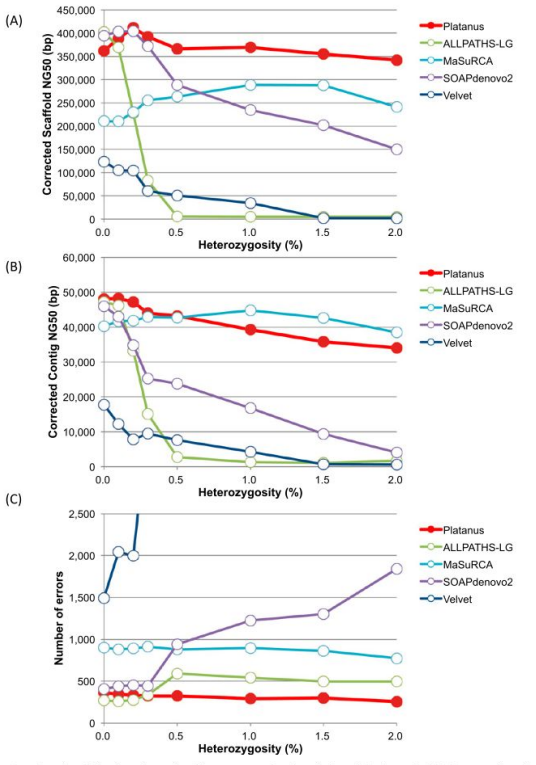
マッピング結果により、リードのペアが異なる contig の組にそれぞれマップされるとき、contig の組がリンクされると判断する。各 contig を節点とし、閾値 n 以上のリンクを持つ contig の組を辺で結び、scaffold グラフを構築する。この際に、ある contig から繋がる複数の contig がインサートサイズを考慮しても衝突する場合が存在する。節点の衝突が存在したとき、片方の辺のペアリード数が閾値より少ない場合、辺はマッピングのミス等に生じたもので真ではないとみなして除去する。

一方、de Bruijn グラフのバブル構造と同様に、SNP/small indel さらには構造変異はバブル構造として表れ、この場合も contig は衝突するように見える。このようなヘテロ領域では coverage depth が低いと考えられ、また、２倍体ゲノムを想定しているため、バブルの中にバブルが存在するような構造は起こりえない。これらの仮定と、ハプロタイプ間の相同性の情報を考慮した基準を考え、scaffold グラフ中のバブルがヘテロ領域に対応すると判定された場合、coverage depth の低い方のパスを除去することで解決を図る。

また、ヘテロ領域中にギャップやリピート配列が存在する場合、バブル構造をとらず枝構造をとる可能性がある。この場合には節点の衝突が起こったとき、衝突している組がともに coverage depth が低く、Contig-assembly のバブルもマップされないとき、片方を除去することで解決を図る。

ギャップクローズは、scaffold 配列上に paired-end をマップし、ギャップ付近にマップされるリード配列から de Bruijn グラフまたは overlap-layout-consensus アルゴリズムをもちいてギャップ部分の配列を構築することで実現している。

続いて、上記アルゴリズム従って開発した Platnus アセンブラの性能を他アセンブラと比較したベンチマーク結果を示す。まず始めに、ヘテロ接合性の低い *C.elegans* ゲノムを illumina Hiseq2000 でシークエンス(230bp-PE, 420bp-PE, 4660bp-MP)し、それに対して、計算機上で変異をヘテロ接合性が 0.1～2.0%になるようにランダムに挿入したデータに対する結果を示す。比較対象は、Allpaths-LG, SOAPdenovo2, Velvet, MaSuRCA である。MaSuRCA は overlap-layout-consensus アルゴリズムを、その他のアセンブラは de bruijn アルゴリズムを採用している。各ヘテロ接合性データに対して、アセンブルを実施し得られた scaffold, contig の Corrected NG50（ミスアセンブル箇所でアセンブル結果を切断したもの）の値を以下のグラフに示す。



グラフより Platanus 以外の de Bruijn グラフに基づくアセンブラ（ALLPATHS-LG、Velvet、SOAPdenovo2）においてはヘテロ接合性の増加に従い corrected NG50 が急激に減少していることが確認できる。overlap-layout-consensus アルゴリズムを採用している MaSuRCA は比較的 corrected NG50 の減少が緩やかであるが、ミスアセンブリ数はどのデータでも Platanus の 2 倍以上という結果になってい

る。特にヘテロ接合性が高い（≥ 1.0%）データにおいては、Platanus の scaffold corrected NG50 は最大、ミスアセンブリ数は最小となっており、精度と長さを両立できていると考えられる。

続いて、高ヘテロ接合性ベネズエラ糞線虫ゲノム(58Mb)に対して Hiseq2000 でシークエンスしたデータ(200bp-PE, 450bp-PE, 3.4kb-MP)をアセンブルした結果を以下に示す。また同時に個別に配列決定した 8 本の fosmid 配列(合計 273kb)を用いたベンチマーク結果も示す。Fosmid 配列を用いた評価では、各 fosmid 配列と得られたアセンブル結果を nunmer によりアラインし、各 fosmid に対して最も長くアライメントされた配列を選び、その合計長と相同性の平均を用いている。

Assembly statistics

	Total (≥500 bp)	Number of scaffolds (≥ 500 bp)	NG50 (bp)
Platanus	58,503,663	2,560	274,622
ALLPATHS-LG	61,205,926	9,608	16,765
MaSuRCA	66,053,722	4,876	176,206
SOAPdenovo2	52,677,856	3,383	87,219
Velvet	63,982,183	11,696	17,006

Fosmid validation

	Top-hits-lengths (bp)	Average Identity (%)	Number of contained fosmids
Platanus	272,164	99.42	8
ALLPATHS-LG	69,792	99.31	0
MaSuRCA	256,848	99.39	7
SOAPdenovo2	270,392	98.72	8
Velvet	78,159	99.31	0

Platanus のアセンブリ結果については、scaffold NG50、fosmid 配列に対する Top-hits-length、identity はともに最大となっている。"Contained fosmids"の数が 8 であることは、8 本の fosmid 配列がいずれも全長が構築されたことを意味し、大きなミスアセンブリも検出されなかったことを意味する。これらの結果は、Platanus はシミュレートされた高ヘテロ接合性データだけでなく実データに対しても有効であることのみならず、他のアセンブラに対してより大きな scaffold NG50 の優位性を得ることを示唆している。

5．主な発表論文等

〔雑誌論文〕（計 9 件）

① Nishikawa H, Iijima T, Kajitani R, 他 13 名, Itoh T, Fujiwara H, A genetic mechanism for female-limited Batesian mimicry in Papilio butterfly., Nat Genet. 47:405-9 (2015) 査読有 doi: 10.1038/ng.3241

② Kajitani R, Toshimoto K, 他 11 名, Itoh T, Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads., Genome Res. 24:1384-95 (2014) 査読有 doi: 10.1101/gr.170720.113.

③ Kawai M, Futagami T, 他 8 名, Ito T, Fujiyama A, Inagaki F, Takami H, High frequency of phylogenetically diverse reductive dehalogenase-homologous genes in deep subseafloor sedimentary metagenomes., Front Microbiol. 3; 80 (2014) 査読有 doi: 10.3389/fmicb.2014.00080.

④ Suda N, Itoh T, 他 7 名, Tanaka M, Dimeric combinations of MafB, cFos and cJun control the apoptosis-survival balance in limb morphogenesis., Development. 141:2885-94 (2014) 査読有 doi: 10.1242/dev.099150.

⑤ Nikaido M, Noguchi H, 他 21 名, Itoh T, Sugano S, Kohara Y, Fujiyama A, Okada N, Coelacanth genomes reveal signatures for evolutionary transition from water to land., Genome Res. 23:1740-8 (2013) 査読有 doi: 10.1101/gr.158105.113.

⑥ Nagayasu E, Ogura Y, Itoh T, Yoshida A, Chakraborty G, Hayashi T, Maruyama H, Transcriptomic analysis of four developmental stages of Strongyloides venezuelensis. Parasitol Int. 62:57-65 (2013) 査読有 doi: 10.1016/j.parint.2012.09.006

⑦ Deardorff MA, Bando M, Nakato R, Watrin E, Itoh T, 他 35 名, Shirahige K. HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle., Nature. 489: 313-7 (2012) 査読有 doi: 10.1038/nature11316.

⑧ De Piccoli G, Katou Y, Itoh T, Nakato R, Shirahige K, Labib K, Replisome stability at defective DNA replication forks is independent of S phase checkpoint kinases., Mol Cell. 45:696-704 (2012) 査読有 doi: 10.1016/j.molcel.2012.01.007.

⑨ Takami H, Noguchi H, Takaki Y, Uchiyama I, Toyoda A, Nishi S, Chee GJ, Arai W, Nunoura T, Itoh T, Hattori M, Takai K, A deeply branching thermophilic bacterium with an ancient acetyl-CoA pathway dominates a subsurface ecosystem., PLoS One. 7:e30559 (2012) 査読有 doi: 10.1371/journal.pone.0030559.

〔学会発表〕（計 2 件）

① 吉村大，後藤恭宏，小椋義俊，林哲也，

伊藤武彦, Whole Genome Shotgun を用いた病原菌に関する疫学研究のための解析手法の開発, ゲノム微生物学会, 2014/3/7, 東京農業大学, 東京都・世田谷区

② Miki Okuno, Yukiko Kodama, Takehiko Itoh, A whole genome comparison between lager brewing yeast Weihenstephan 34/70 and its ancestral strains., Yeast 2013, 2013/8/29, Westend of Goethe University, フランクフルト(ドイツ)

６．研究組織

(1)研究代表者

伊藤　武彦（ITOH TAKEHIKO）
東京工業大学・大学院生命理工学研究科・教授
研究者番号：90501106

(2)研究分担者

野口　英樹（NOGUCHI HIDEKI）
国立遺伝学研究所・先端ゲノミクス推進センター・特任准教授
研究者番号：50333349

丸山　治彦（MARUYAMA HARUHIKO）
宮崎大学・医学部・教授
研究者番号：90229625