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

令和　元 年　 ６ 月 １７ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)（一般）
研究期間：2016～2018
課題番号：１６Ｈ０４７１９

研究課題名（和文）ハプロタイプを区別する新規ゲノムアセンブラの開発および超多様化ゲノム領域の解析

研究課題名（英文）Development of haplotype phased genome assembler

研究代表者
伊藤　武彦（Itoh, Takehiko）

東京工業大学・生命理工学院・教授

研究者番号：９０５０１１０６

交付決定額（研究期間全体）：（直接経費）　　12,700,000円

研究成果の概要（和文）：二倍体ゲノムの新規構築にあたって究極のゴールは、相同染色体を区別した染色体配列を構築することであるが、高ヘテロ接合ゲノムにも適用可能なプログラムは現存しない。そこで我々は、Platanus-alleeと呼ばれる新規アセンブラを開発した。このアセンブラでは、まず相同染色体を別々に構築することで、高ヘテロ接合ゲノムや、局所的に相同染色体間で大きな配列の隔たりを持つ領域（既存研究でしばしば重要性が指摘されている箇所）の構築に成功した。また、ベンチマークテストを通じ、既存手法よりも高精度・高感度で長く構築できることも明らかとなった。

研究成果の学術的意義や社会的意義
本研究を通じて開発されたPlatanua-alleeアセンブラは、そのアルゴリズムを論文にて公開する一方で、ホームページより利用可能な形のプログラムとして公開している。このプログラムを用いることで、従来はfosmid構築など特別なコストのかかる方法で配列決定をしないとアクセスできなかったようなゲノム領域に対し、安価なIlluminaデータの使用のみで迫ることができるようになったという点で、学術的な意義は極めて大きいと考えられる。

研究成果の概要（英文）：The ultimate goal for diploid genome determination is to completely decode homologous chromosomes independently, and several phasing programs from consensus sequences have been developed. These methods work well for lowly heterozygous genomes, but the manifold species have high heterozygosity. Additionally, there are highly divergent regions, where the haplotype sequences differ considerably and are likely to relate to various interesting biological phenomena. However, they cannot be accessed by existing phasing methods, and we have to adopt costly traditional methods. Here, we develop a de novo haplotype assembler, Platanus-allee, which initially constructs each haplotype sequence and then untangles the assembly graphs utilizing sequence links and synteny information. A comprehensive benchmark analysis reveals that Platanus-allee exhibits high recall and precision, particularly for highly divergent regions.

研究分野：ゲノム情報

キーワード：ゲノムアセンブラ　ヘテロ接合性

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

NGS とりわけ Hiseq の登場により、100-250bp 程度の短い配列ではあるものの、一度の稼働で数百 Gb にも及ぶ塩基配列が安価に産出される時代になり、様々な生物種の新規ゲノム配列決定にも適用されている。これに伴い、近交化されていない生物のドラフトゲノムや野生集団のリシークエンスデータが急速に蓄積されており、解析結果から配列多様性がゲノム中の領域によって大きく変動することが示されている(*Science*. 2010:330,512-4)。中でも多様性の高い領域は、種分化に関係する可能性が示唆されている(*Genome Res.*2015:Sep9) のみならず、形態の表現型に決定的な影響を及ぼす場合も報告されている (*Nature* 2011:477, 203-6; Nature 2014:507,229-32)。例えば、シロオビアゲハ擬態の責任領域において supergene を形成している約 130 kbp の逆位変異が存在し、ハプロタイプ間の配列相同性が 90%以下と極めて低いことなども報告され(*Nature Genetics* 2015:47,405-409)、このような構造変異とハプロタイプ間の低相同性を併せ持つ「超多様化ゲノム領域」は進化の駆動力として重要であることが示唆されている。また、差異の大きい 2 型を持つ常染色体領域が性決定機能を持つ例はトゲウオ科の生物種などで報告されており(*Mol Biol. Evol.* 2013: 30,1131-44)、組み換え頻度の低下に配列分化が伴う点は蝶の supergene と類似性を有する。

また、2013 年 *Nature* 誌に、HeLa 細胞をハプロタイプに分けてゲノム解析を実施した（Haplotype phasing）論文が発表された（*Nature*. 2013:500,207-211）のを皮切りに、ヒトゲノムのハプロタイプ別アセンブルに関する論文が複数報告されている。がん等疾患のゲノム的な理解にはハプロタイプ毎に入った変異を詳細に解析する必要性があるため、様々な手法が提案されている。Fosmid ライブラリの構築、10-40 kbp のゲノム領域を局所的に Illumina シークエンサで読み擬似的にロングリードを得る方法(*Nat. Biotech.* 2014:32,261-6)、Pacific Biosciences (PacBio)社製 1 分子 DNA シークエンサ、Irys システムを Illumina データに組み合わせた方法(*Nat. Methods* 2015:12,780-6)などにより長いハプロタイプ配列が構築されており、解析結果からは挿入・欠失・逆位・タンデムリピートが隣接する複雑な構造変異の存在が明らかになっている。しかし、いずれも極めてコストがかかる方法であり、一般的な手法とはなり得ない。

研究開始当初における一般的な解析では、シークエンスされた断片配列をリファレンスゲノムにマッピングする手法が用いられている。しかし、解析対象に超多様化ゲノム領域が含まれるあるいはハプロタイプ間で構造変異がある場合、マッピングにより解決する事は極めて困難である。これら領域を解析するためには配列を新規構築(アセンブリ)する必要があり、アセンブリベースの変異検出用ツールとして Cortex、Discovar 等が存在するが、これらは大規模な構造変異は検出対象としていないなど、ゲノム全体でのハプロタイプ別の解析には適しておらず、新規手法の開発が望まれていた。

２．研究の目的

上記背景で述べたように、次世代シークエンサ(NGS)の普及に伴い、様々な生物種の新規ゲノム配列決定や個人ゲノムの決定が比較的容易に行われるようになっている。近年ではそれに伴った次なる研究対象として「相同染色体の違い」への注目が高まってきている。非モデル生物においてハプロタイプ間の差異が著しく大きい「超多様化ゲノム領域」が存在し、種分化や表現形の多様化に深く関連している事例や、がんを始めとした疾患原因の究明を対象とした「ハプロタイプ別ゲノム配列」解析事例等が報告されている。しかし、広く行われているリファレンスゲノムへのマッピングに基づいたハプロタイプ解析では、相同染色体間での差異が大きいあるいは構造多型がある場合への対応が困難であり、fosmid などに基づいたハプロタイプ構築手法は効果的であるが、コストの高さがネックである。そこで本研究ではハイスループットショートリードシークエンサおよびロングリードシークエンサを用いて、既存手法より大幅に低いコストで数 100 kbp のハプロタイプを構築する手法を開発し、ハプロタイプを区別したゲノム解析研究の基盤を築くことを目的とする。

３．研究の方法

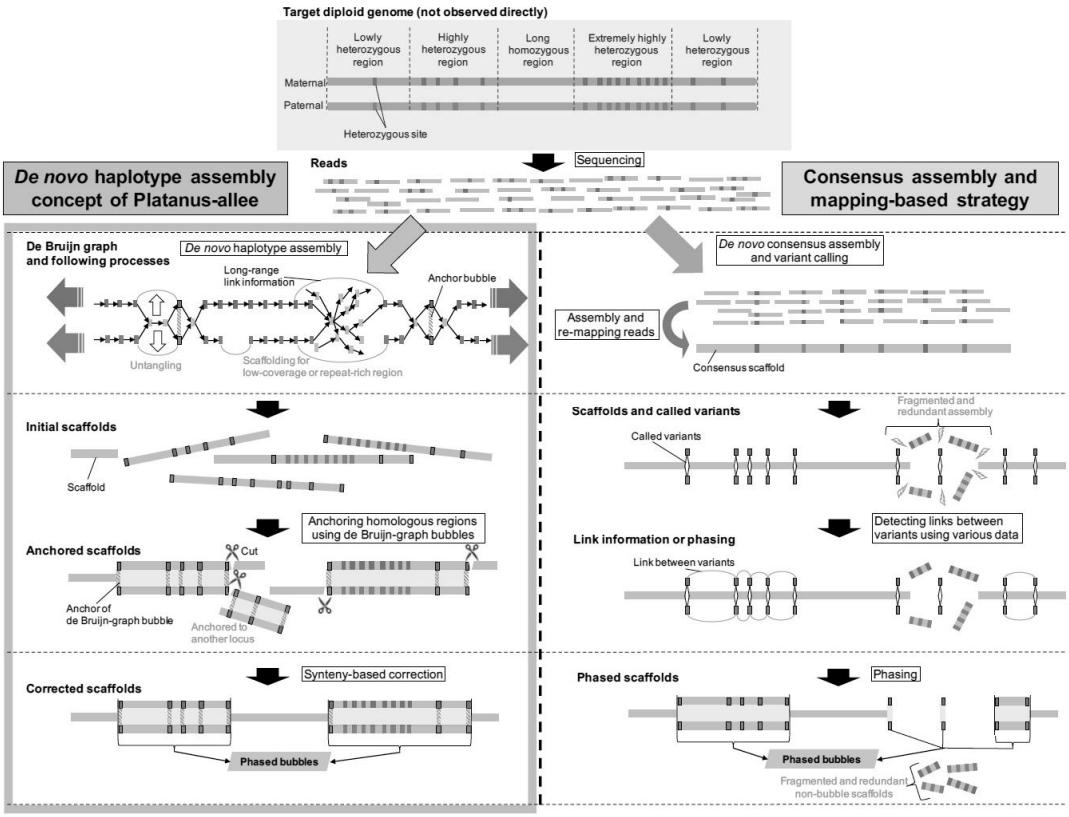
本研究ではまず、代表者および分担者らが開発した Platanus アセンブラをベースとし、Illumina リード向けハプロタイプ別アセンブラの開発を実施する。Platanus には相同染色体間の配列の違いが大きな(ヘテロ接合性が高い)ゲノムに対しても、良好な成績を収められるという最大の特徴があり多くの成果を挙げているため、この手法を元として相同染色体間の差異により de Bruijn グラフに生じるバブル構造間の連鎖関係を、Pair 情報を用いて解決することで実現を目指す。Platanus アセンブラ開発時には、想定していなかったほどの相同染色体間における差異が大きい領域がゲノム中には存在している場合も多いため、従前の相同染色体のコンセンサスを作成後に、SNV 情報などを用いて相同染色体を「分ける」手法のみならず、独立に相同染色体配列の構築も行うこととする。また、並行して、PacBio データの活用に向けた OLC をベースとしたアルゴリズムの実装を試みる。最終的に両アルゴリズムを組み合わせることで、ハイブリッドデータを入力としたハプロタイプ別アセンブラおよび超多様化領域抽出アルゴリズムの開発を実施する。

開発されたアルゴリズムに関しては、ゲノム既知な生物を使ったベンチマークを行うことで、その有用性と、他アセンブラに対する優位性の検証を実施する。

４．研究成果

・新規アセンブラ(Platanus-allee)の開発

上記目的を達成するために、新規ゲノムアセンブラ Platanus-allee を開発した。既存の各種アセンブラがハプロタイプ配列間のコンセンサス配列をまず構築し、相同染色体由来配列へと「分ける」（下図右）のとは異なり、Platanus-allee では、各ハプロタイプ由来の配列を独立に２本構築するという基本的なストラテジーを採用した。（下図左）

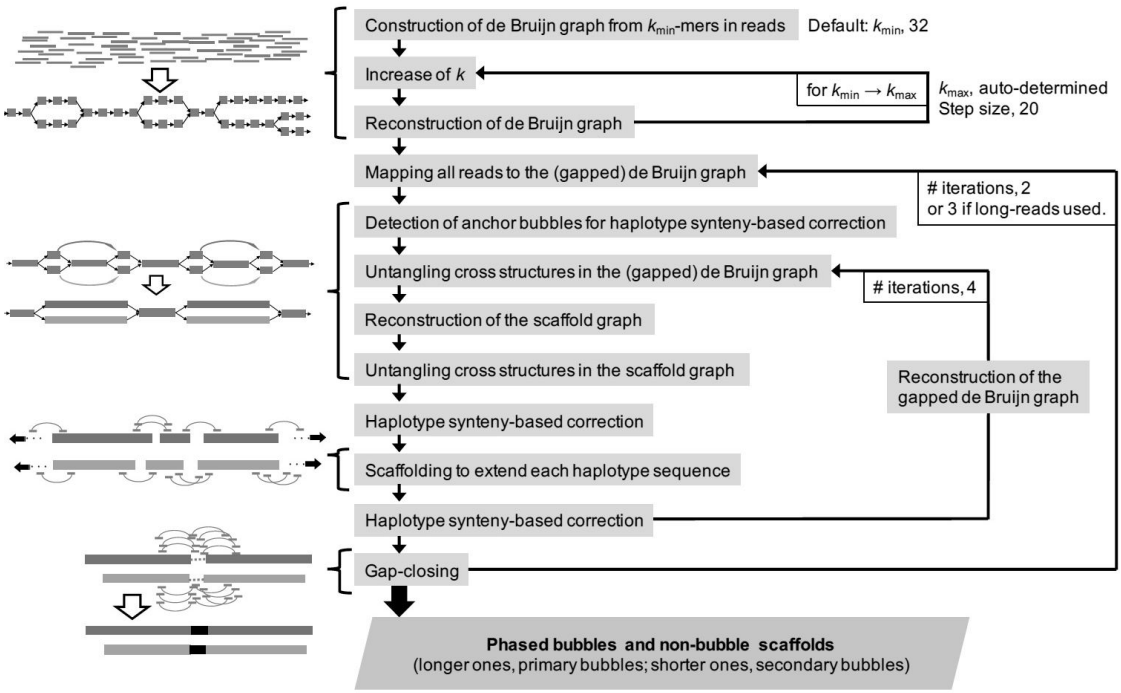


このストラテジーは、主に de Bruijn グラフ構築後の scaffolding，gap_close モジュールにおいて実現されており、具体的には得られた contig 配列同士をアライメントすることで、相同染色体間の対応付けを行っている。この際の contig の作成には、*k*-mer 同士の１塩基の違いも厳密に区別しており、実現には Illumina データが持つ高い配列精度が欠かせない。一方 Illumina short リードでは、繰り返し配列など graph 構造が複雑となる箇所の解決はリード長の関係より困難である。そのような箇所は、scaffolding 後の gap_close において解決を試みることとし、これらのルーチンを繰り返し実行することで、より長く正確なアセンブルを目指した。

Platanus-allee においては、グラフ構造における×の字構造を、シングルリード、ペアエンド、メイトペアのリンク情報を de Bruijn グラフにマップすることで解決する。×の字構造は、一つの中央ノードと、4 つのそれ以外のノードから構成されるが、前者が相同染色体間におけるホモな配列、後者がヘテロな配列に相当する。このヘテロなノード間の連鎖関係が解決した場合、中央のホモな配列は両アレルにおいて二回使われることで×の字から＝の字構造へとグラフが解決される。この工程は、de Bruijn グラフ上および contig 間の並びを解決する scaffolding 時の双方で行われ、繰り返し実施される。また、×の字構造の解決および、scaffolding においては、de Bruijn グラフ上に PacBio などのデータをマッピングすることで、Longread も活用できるようになっている。

相同染色体を分けた(phased)配列をより長く正確に得るために、Platanus-allee では単に graph を解決するのみならず、相同性に基づくシンテニー候補を考慮するアルゴリズムを採用している。これは例え局所的には相同染色体間で大きく異なった配列が存在しても、染色体スケールで見た場合には、大域的なシンテニーは保存されているであろうという仮設に基づいている。このアルゴリズムの採用により、完全にグラフがバブルとして閉じない場合にも、相同性によりシンテニーと判断されると繋いでいくことが可能となり、よいアセンブル結果が期待される。

上記アルゴリズムの概要を表したのが、下記フローである。



・ベンチマークテスト

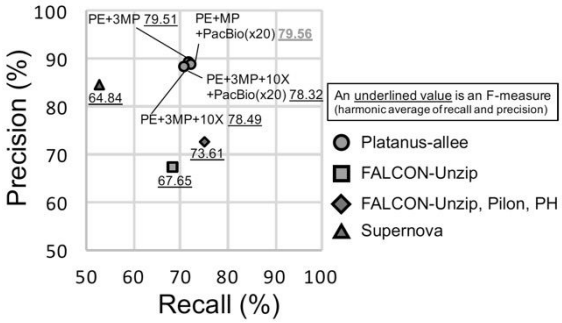
開発した Platanus-allee について、ヘテロ接合度が高い生物種に関するベンチマークテストを実施した。比較対象は、ハプロタイプ毎のアセンブルが可能とされている FALCON-Unzip, Supernova とし、ヘテロ接合度の高いゲノムを持つシロオビアゲハ、ナメクジウオ、さらには精度のベンチマークを行うために、線虫 N2 株および CB4856 株由来のゲノムを混ぜた疑似 diploid ゲノム、ヘテロ接合度の低いヒトゲノムをアセンブルした結果に対する phased block の各種統計値を以下の表に示す。

Sample (name; diploid genome size; heterozygosity)	Assembler	Input data	Total (bp)	Bubble total (bp)	Bubble-total/genome-size	Scaffold NG50 (bp)	Scaffold LG50 (#)
P. polytes (butterfly; 480 Mbp; 1.52%)	Platanus-allee	PE + 4 MP(≤15 kb)	441,668,515	392,697,455	0.818	403,792	344
		PE + 4 MP(≤15 kb) + PacBio(x20)	473,444,046	455,992,202	0.950	3,225,089	51
		PE + 4 MP(≤15 kb) + 10X	449,125,510	407,085,972	0.848	697,870	207
		PE + 4 MP(≤15 kb) + PacBio(x20) + 10X	475,717,493	460,173,985	0.959	2,391,666	65
	FALCON-Unzip	PacBio(x99)	480,851,586	404,432,220	0.843	413,454	352
	FALCON-Unzip, Pilon, PH	PacBio(x99) + PE	470,795,219	422,370,064	0.880	420,842	353
	Supernova	10X	313,260,798	121,664,227	0.253	78,567	789
B. japonicum (amphioxus; 780 Mbp; 3.48%)	Platanus-allee	PE + 3 MP(≤10 kb)	719,710,097	686,385,470	0.880	1,090,309	194
		PE + 3 MP(≤10 kb) + PacBio(x20)	739,278,992	714,636,843	0.916	1,513,530	142
		PE + 3 MP(≤10 kb) + 10X	731,512,578	693,785,345	0.889	1,154,834	172
		PE + 3 MP(≤10 kb) + PacBio(x20) + 10X	749,591,461	719,559,447	0.923	1,515,647	140
	FALCON-Unzip	PacBio(x156)	917,509,827	377,851,536	0.484	171,861	1,179
	FALCON-Unzip, Pilon, PH	PE + PacBio(x156)	978,100,476	852,042,162	1.092	1,074,877	162
	Supernova	10X	696,582,435	177,104,030	0.227	18,368	8,779
C. elegans (worm, synthetic diploid; 200 Mbp; 0.38%)	Platanus-allee	PE + 3 MP(≤16 kb)	195,025,359	179,373,871	0.897	469,838	112
		PE + 3 MP(≤16 kb) + PacBio(x20)	204,646,867	197,622,584	0.988	902,153	64
	FALCON-Unzip	PacBio(x192)	242,644,110	224,221,346	1.121	511,326	105
	FALCON-Unzip, Pilon, PH	PacBio(x192) + PE	232,095,471	221,854,009	1.109	512,209	105
H. sapiens (human, NA1287; 6.2 Gbp; 0.13%)	Platanus-allee	PE + 4 MP(≤15 kb)	3,898,167,507	2,017,742,240	0.325	4,093	194,055
		PE + 4 MP(≤15 kb) + PacBio(x20)	5,684,002,821	5,405,978,105	0.872	306,349	5,294
		PE + 4 MP(≤15 kb) + 10X	4,918,237,808	4,025,214,794	0.649	59,172	23,612
		PE + 4 MP(≤15 kb) + PacBio(x20) + 10X	5,672,647,688	5,459,715,622	0.881	658,165	2,584
	FALCON-Unzip	PacBio(x77)	4,721,106,192	3,690,539,270	0.595	108,668	13,508
	FALCON-Unzip, Pilon, PH	PacBio(x77) + PE	4,851,423,579	3,783,489,783	0.610	123,947	11,817
	Supernova	10X	5,404,712,080	5,027,768,096	0.811	2,489,267	675
	Mostovoy et al. 2016	PE + 1 MP(2 kb)+ 10X + Bionano	5,534,675,788	5,352,992,348	0.863	3,998,058	423

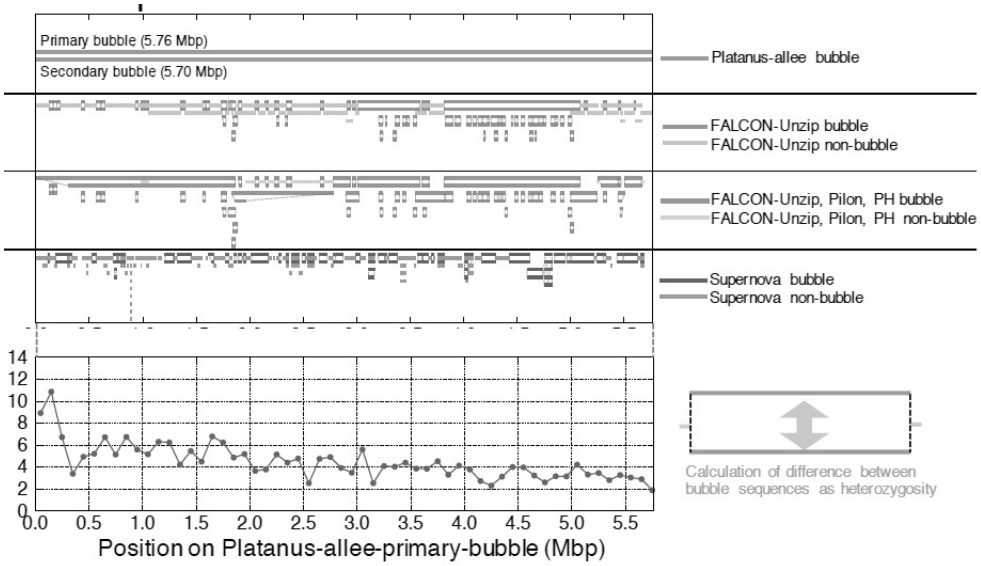
結果より、ヘテロ接合度の高いシロオビアゲハ、ナメクジウオにおいて、他アセンブラより圧倒的によくつながった結果が得られていることがわかる。Phased されたゲノム領域の割合もいずれも 90%を超えており、推定ゲノムサイズとの乖離差も小さく良好な結果である。一方、ヘテロ接合度が 0.13%と低いヒトゲノムに関しては、phased block のサイズは小さくなってしまう。これは相同染色体間での SNV が低頻度であり、short read では複数の SNV にまたがる情報を得ることが困難なことが主な要因と考えられる。事実、Platanus-allee においても Longread を加えることで phased black の N50 が 5kb から 658kb まで飛躍的に大きくなっていることとからも想定される。

次にナメクジウオゲノムを用い、構築された phased block の精度・感度の評価を実施した。ナメクジウオゲノムでは、Moleculo を用いたシークエンスも実施しており、10kb と短いものの、完全にハプロタイプを区別した合成ロングリードデータが得られている。これらを用いることで各アセンブル結果の精度と感度を測定した結果が、下記のグラフである。

結果より、Platanus-allee の結果が精度においても感度においても他アセンブラ結果を上回っており、長くつながるのみならず、高い精度･感度を持ってつながっていることが明らかとなった。同様の結果は、疑似 diploid を構築した線虫の結果からも示されており､ゲノムを 1kb の区画に分けて 100% identity で構築できた割合を調べると、Platanus-allee ではゲノム全体の約 85%が構築できているのに対し、FALCON-Unzip ではポリッシュを行っても約 60%, Supernova では約 20%しか構築できていないことが明らかとなった。



最後に、ナメクジウオゲノムにおいて、Platanus-allee ではうまく phased block が構築できた領域について、他のアセンブル結果がどのようになっているかを調べ、図示化したものを以下に示す。



上段の二本の線が Platanus-allee の bubble 配列であり、二段目が FALCON-unzip, 四段目が Supernova の結果である。他のアセンブラでは、断片化されているのみならず、3 本目、4 本目などの配列が冗長に出力されている結果が見て取れる。この傾向は、表で示した FALCON-unzip, Supernova のアセンブル長合計が、推定ゲノムサイズ長を大きく超えていることとも一致する。

以上、4 種のゲノムデータを用いたベンチマークを通し、本研究を通じて開発した Platanus-allee アセンブラが高いヘテロ接合度を有するゲノムに関しては、高い精度・感度を持って長く phased したブロック配列を構築できていることが明らかとなった。

５．主な発表論文等

〔雑誌論文〕（計１件）

① Kajitani R, Yoshimura D, Okuno M, Minakuchi Y, Kagoshima H, Fujiyama A, Kubokawa K, Kohara Y, Toyoda A, Itoh T., Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. ***Nat Commun.*** 2019 Apr 12;10(1):1702. doi: 10.1038/s41467-019-09575-2.

〔学会発表〕（計１件）

① 梶谷嶺, 吉村大, 奥野未来, 豊田敦, 伊藤武彦 Platanus2: a de novo haplotype assembler enabling comprehensive accesses to divergent heterozygous region. 第 6 回生命医薬情報学連合大会(IIBMP2017) 2017 年

〔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：

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：
国内外の別：

〔その他〕
ホームページ等
http://platanus.bio.titech.ac.jp/platanus2/

６．研究組織

(1)研究分担者
研究分担者氏名：豊田　敦
ローマ字氏名：Toyoda Atsushi
所属研究機関名：国立遺伝学研究所
部局名：ゲノム・進化研究系
職名：特任教授
研究者番号（8 桁）：１０２６７４９５

研究分担者氏名：梶谷　嶺
ローマ字氏名：Kajitani Rei
所属研究機関名：東京工業大学
部局名：生命理工学院
職名：助教
研究者番号（8 桁）：４０７５６７０６

(2)研究協力者
研究協力者氏名：
ローマ字氏名：