T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

題目(和文)	ハイスループ　ットDNA シークエンスデ　ータによる全ゲ　ノム配列の構築と解析に関する研究
Title(English)	
著者(和文)	梶谷嶺
Author(English)	Rei Kajitani
出典(和文)	学位:博士（理学）, 学位授与機関:東京工業大学, 報告番号:甲第9728号, 授与年月日:2015年3月26日, 学位の種別:課程博士, 審査員:伊藤　武彦,工藤　明,徳永　万喜洋,黒川　顕,山口　雄輝
Citation(English)	Degree:, Conferring organization: Tokyo Institute of Technology, Report number:甲第9728号, Conferred date:2015/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Powered by T2R2 (Tokyo Institute Research Repository)

平成 26 年度　博士論文

ハイスループット DNA シークエンスデータによる全ゲノム配列の構築と解析に関する研究

指導教官　伊藤　武彦　教授

東京工業大学
大学院生命理工学研究科
生命情報専攻

梶谷　嶺

第 1 章　序論 4

第 2 章　真核生物ドラフトゲノム配列の構築 8

2.1　背景・目的 8

2.2　真核生物用 *de novo* アセンブラ Platanus の開発 9

2.2.1　Platanus のアルゴリズムの概要 10

2.2.2　Contig-assembly のアルゴリズム 13

・*k*-mer 出現回数の分布の算出 13

・de Bruijn グラフの構築 13

・エラー由来の枝構造（tip）の除去 15

・*k*-mer 伸長と de Bruijn グラフの再構築 15

・$k_{\max}$ と c_i の計算方法 18

・バブル構造の除去 20

・リードの再マッピングによる contig の修正 21

2.2.3　Scaffolding のアルゴリズム 22

・paired-end（mate-pair）ライブラリのマッピング 22

・Contig-assembly のバブル配列を利用したリードのマッピング 23

・インサートサイズの推定 23

・scaffold グラフ 24

・scaffold グラフの節点（contig）の衝突 25

・scaffold 配列の構築 26

・エラー由来の辺の除去 26

・scaffold グラフ中のバブル構造の除去（bubble removal） 27

・ヘテロ領域に由来する枝構造の除去（branch cut） 29

・インサートサイズの大きいライブラリによる修正 30

2.2.4　Gap-close のアルゴリズム 31

・ギャップをカバーするリードの収集 31

・ローカルな *de novo* アセンブリによるギャップ部分の配列構築 32

2.3　真核生物データに対する Platanus の有用性の検証 34

2.3.1　概要 34

・ベンチマーク対象の生物種と 17-mer 出現回数の解析 34

・ベンチマーク対象の *de novo* アセンブラ 39
・ベンチマークデータの前処理 39
2.3.2　ヘテロ接合度をシミュレートしたデータによるベンチマークと考察 41
・*C. elegans* 実データの取得およびヘテロ接合度をシミュレートしたデータ作成 41
・ベンチマーク結果および考察 42
・ベンチマーク結果の詳細データ 49
2.3.3　高ヘテロ接合性線虫によるベンチマークと考察 51
・*S. venezuelensis* データの取得および特徴 51
・ベンチマーク結果および考察 51
2.3.4　高ヘテロ接合性かつ高リピート率のサンプルによるベンチマークと考察 61
・*C. gigas* データの取得および特徴 61
・ベンチマーク結果および考察 61
2.3.5　Assemblathon2 のデータによるベンチマーク結果と考察 65
・Assemblathon2 データ(bird、snake、fish)の取得および特徴 65
・ベンチマーク結果および考察 67
2.3.6　ダウンサンプリングテストによる最適 coverage に関する考察 70
2.3.7　実行時間、メモリ使用量についてのベンチマーク結果と考察 72

2.4　Platanus の実データに対する適用例：シーラカンスゲノム解読 74
2.4.1　シーラカンスゲノムのアセンブリ 74
・シーラカンスシークエンスデータの取得およびその特徴 74
・アセンブリ結果および考察 76
2.4.2　変異解析、5 個体の比較解析 84

2.5　考察 88

第 3 章　原核生物の全ゲノム、環境ドラフトゲノム配列の構築 92

3.1　背景・目的 92

3.2　原核生物データに対する Platanus の有用性の検証 94
3.2.1　大腸菌によるベンチマークと考察 94
・*E. coli* 2 株のデータの取得および特徴 94
・ベンチマーク結果および考察 96

3.2.2　大腸菌の完全ゲノム構築の条件検討 103

3.3 メタゲノム用 *de novo* アセンブラ MetaPlatanus の開発 108

3.3.1　MetaPlatanus のアルゴリズムの概要 108

3.3.2　MetaPlatanus の Contig-assembly のアルゴリズム 109

3.3.3　MetaPlatanus の Scaffolding のアルゴリズム 111

3.3.4　Scaffolding、Gap-close の反復のアルゴリズム 112

3.3.5　scaffold 配列のクラスタリングと re-scaffolding のアルゴリズム 114

3.4　メタゲノムデータに対する MetaPlatanus の有用性の検証 116

3.4.1　仮想メタゲノムデータによるベンチマークと考察 116

・仮想メタゲノムデータの取得および特徴 116

・ベンチマーク結果および考察 122

3.4.2　環境メタゲノム実データのアセンブリ 132

・Cow rumen メタゲノムデータの取得および特徴 132

・アセンブリ結果および考察 132

3.5　考察 133

第 4 章　総括 135

参考文献 138

謝辞 147

第1章　序論

1980 年代、大腸菌（*Escherichia coli*）等の全ゲノム配列解読計画が提唱され始めた。当時の主流な方法はゲノムの一部分をクローニングして配列を決定し、遺伝地図や物理地図に従ってそれらを並べていくことでより長い配列を構築（*de novo* アセンブリ）するというものである。大腸菌の全物理地図は 1987 年に完成しているが（Kohara et al. 1987）、全ゲノム配列解読の発表（Blattner et al. 1997）はそれから 10 年後であり、配列決定は長期的な計画を要する状況であった。しかしながら 1990 年代に入ると、サンガー法 DNA シークエンサの性能が向上したことや、国際的なヒトゲノム解読計画が立ち上げられたことで全ゲノム配列の解読計画が活発化していくこととなる。1996 年の酵母の全ゲノム解読（Goffeau et al. 1996）に続き、1998 年には線虫の全ゲノムが解読（The *C. elegans* Sequencing Consortium. 1998）された。これらはクローニングと物理地図を用いる点は従来と同一であるが、多国籍のチームが数十台のシークエンサを用いて配列を決定していくという点で規模に大きな差が存在した。一方でゲノム DNA をランダムに断片化してそれぞれをシークエンスし、計算機上で一度に元のゲノム配列を構築する全ゲノムショットガン法も考案されていた。その有効性はインフルエンザ桿菌ゲノムの決定（Fleischmann et al. 1995）で示され、その後もショウジョウバエなどの真核生物ゲノムに適用されている（Myers et al. 2000）。全ゲノムショットガン法は遺伝地図や物理地図を用いないため時間と予算の効率が高いという利点を持つが、アセンブリ時の計算量が莫大になるため、高速なアルゴリズムが要求されるのみならず、繰り返し配列などゲノム上で類似している配列を正しく区別しミスアセンブルを起こさず正確な配列を得るための、より複雑なアルゴリズムが要求される。シークエンサのスループットの向上や全ゲノムショットガン法の導入はヒトゲノム計画にも影響を強く与えており、公的機関からなる国際コンソーシアムと私企業のセレラ社が 2001 年に同時にドラフトゲノム配列の論文を発表する結果となったが（Lander et al. 2001; Venter et al. 2001）、これは両者の競争によるもので、計画開始時の予定より早い解読宣言であった。前者は精度を重視し、保守的な物理地図構築に基づく手法が、後者は速度を重視し、全ゲノムショットガン法が適用されている。セレラ社はヒトゲノム全ゲノムショットガン法のために Cerela Assembler という新たなアセンブル用ソフトウェア（アセンブラ）を開発し、この手順が正しく機能することを確認するた

めにショウジョウバエのゲノム解読をプロトタイプとして実施している（Myers et al. 2000）。以上が、1980–90 年代にかけてシークエンサとショットガン法の進歩により全ゲノム配列の研究が加速されてきた経緯である。

ヒトの全ゲノム配列の解読宣言以前より Expressed Sequence Tag（EST）による網羅的な発現遺伝子の解析などは行われていたが（Adams et al. 1992）、全ゲノム配列が利用可能になったことで非コード領域も含めた解析が推進されるようになってきた。ヒトゲノムを用いた代表的なプロジェクトとしては ENCODE 計画（Bernstein et al. 2012）が挙げられる。この計画は 2003 年に開始され、ゲノム中で機能を持った要素（element）を全て注釈付けることが目的となっており、解析結果からはゲノム配列上の 80%は何らかの機能を持つという仮説が提唱されている。疾患関連の変異が非コード領域で発見されることもあり、コード領域以外を"junk DNA"と呼ぶことはもはや不適切である。ヒト以外でも、ショウジョウバエや線虫の modENCODE 計画（modENCODE consortium. 2010）など類似の計画は進められており、全ゲノム配列からは多数の知見が得られることが期待されている。

2000 年代に入り、従来のサンガー法とは原理の異なるハイスループット DNA シークエンサと呼ばれる機器の登場により、シークエンサのスループットは劇的に上昇した。代表的なものは 454 社（現 Roche 社）製、Illumina 社製のシークエンサである。2014 年時点で最も 1 塩基あたりのコストが低く、1 運転でのスループットが大きいシークエンサは Illumina 社製であると考えられるが、機器が発表された当時はシークエンス可能な断片（リード）長が短いことから *de novo* アセンブリへの適用例は少なかった。リード長が伸びるにつれ、バクテリアなどゲノムサイズの小さな生物種の新規ゲノム決定に用いられるようになり、ジャイアントパンダのドラフトゲノム（2.25 Gbp）の決定（Li et al. 2010）に用いられた以後は真核生物ゲノム決定にも多く適用されているようになっている。ここで、長い DNA 断片の両端をシークエンスする paired-end 法、mate-pair 法と呼ばれる手法もアセンブリ結果の長さ向上に寄与しており、本論文でもこれらの手法で得られたデータを多く使用している。これ以降数多くのゲノムプロジェクトが Illumina 社等のハイスループット DNA シークエンサの利用で進められており、1990 年代のように必ずしも生物種毎に国際コンソーシアムを立ち上げることなく、多数の生物種の全ゲノム配列を決定することが可能となっている。しかしながら、ハイスループットシークエンサを用いても *de novo* アセンブリの

プロトコルは完全に自動化されているとは言い難く、サンプルの特徴によってはアセンブリ結果の長さが不十分である場合や、配列の精度が悪いという状況が起こりえる。アセンブリの障害としてはゲノム中のリピート配列やシークエンスエラーの存在など数多くの要因があるが、非モデル生物ゲノム特有の問題としては高いヘテロ接合性（相同染色体間の差異）がアセンブリ結果に悪影響を及ぼすケースが報告されている（Zhang et al. 2012）。ヘテロ接合性は集団内の多様性を反映しており、研究室内で近交系が確立されているモデル生物と比較して野生型の生物では値が大きくなる。DNA シークエンサのスループットは 2014 年時点でも増加し続けているが、ヘテロ接合性やリピート配列などサンプルの性質に基づく問題は単にデータ量を増やせば解決するものではない。非モデル生物の全ゲノム配列を効率的に決定し解析を行うためには、それらの問題に対応した *de novo* アセンブラの開発が必要である。

一方、解析対象を原核生物に目を向けても、Illumina データからの新規ゲノム配列決定において決定的な手法は未だ存在しておらず、*de novo* アセンブリのアルゴリズムを改善することにより、ハイスループット DNA シークエンサを用いて、完成ゲノムに近づいたギャップ領域が残らない配列の決定が期待されている。さらに、環境 DNA サンプルにショットガン法を適用することで、複数の難培養性生物種のドラフトゲノムを同時に構築するメタゲノム研究も盛んに行われており（Nielsen et al. 2014）、この分野でもより高精度で長い配列を構築できるアセンブラは大いに期待されている。

これらの背景を踏まえ、本研究では上記問題を解決すべく *de novo* アセンブラ Platanus を新たに開発した（Kajitani et al. 2014）。最大の特徴としては高ヘテロ接合性の２倍体データからも高精度な配列をアセンブルすることができるという点が挙げられるが、ヘテロ接合性以外の障害に対しても対策は施されており、バクテリアや複数種を含むメタゲノムデータへの適用も考慮し開発された。

次章以降で Platanus アセンブラのアルゴリズムおよび性能評価などを述べるが、本論文の主要部分は以下に示す 2 章からなっている。まず第 2 章では真核生物データのアセンブルを取扱い、次のようなデータに対して Platanus を適用することでその有用性を示す。

・　ヘテロ接合性のシミュレーションデータ

低ヘテロ接合性の線虫 *C. elegans* の実データに対して、計算機上で 0.1–2.0%

のヘテロ接合度をシミュレートし、アセンブリ結果の精度評価を行なった。既存ツールと比較して、ヘテロ接合度の増加による影響を Platanus は受けにくいことが示された。

- 高ヘテロ接合性サンプルの実データ

  高ヘテロ接合性の２サンプル、線虫 *S. venezuelensis* と牡蠣 *C. gigas* の実データに Platanus を適用した。シミュレーションデータだけでなく、高ヘテロ接合性の実データに対しても Platanus が優れた精度を持つことが示された。

- **Assemblathon2 のデータ**

  de novo アセンブリの国際コンテストである Assemblathon2（Bradnam et al. 2013）のデータに Platanus を適用し、ゲノムサイズが 1 Gbp 前後のサンプルに対しても他のツールより優れた結果が得られることを確認した。

- ゲノム解読計画での実用例

  Platanus の初の実用はシーラカンス（*L. chalumnae, L. menadoensis*）のゲノム解読計画である（Nikaido et al. 2013）。Platanus のアセンブリ結果を基に解析が行われた例として記す。

引き続いて第 3 章では次のような原核生物のデータを取扱い、Platanus アルゴリズムの応用の可能性を示す。

- バクテリアの単一種データ

  真核生物と比較してゲノムサイズが小さいバクテリアに Platanus を適用すると、生物種によってはギャップが少なく完成に近いゲノム配列が得られることが示された。

- メタゲノムデータ

  Platanus にメタゲノムデータ用のアルゴリズム変更を施し、複数のバクテリアを含む DNA サンプルのデータに適用した。環境 DNA サンプルから培養を経ずにドラフトゲノムが構築可能であることが示唆された。

また、最後に第 4 章として以上の総括を述べることで全体のまとめを実施した。

第2章　真核生物ドラフトゲノム配列の構築

2.1　背景・目的

Illumina 社製 HiSeq に代表されるハイスループット DNA シークエンサのデータを *de novo* アセンブリする際、一番大きな問題が計算量の多さである。サンガー法データに用いられていた overlap-layout-consensus アルゴリズム（Myers et al. 2000; Batzoglor et al. 2002）は、リード間のオーバーラップを検出する段階やグラフ中の経路を探索する段階で計算量やメモリ使用量が増大しやすいという欠点を持つ。データ量が多い場合は計算時間が計画の進行に支障をきたすことや、メモリ使用量過多による異常終了が起こる可能性が存在する。そこで注目されたアルゴリズムは、Euler アセンブラ（Pevzner et al. 2001）に採用されていた de Bruijn グラフを用いたものである。de Bruijn グラフとは長さ k の部分文字列を節点とし、$k-1$ のオーバーラップを辺としたグラフである。グラフ中の経路がアセンブリ結果（contig）の配列に対応する。このアルゴリズムではオーバーラップ検出のためのアライメントを行わないため、計算量が抑えられるという利点を持つ。Euler はサンガー法データ用に開発されたツールであったが、ハイスループットシークエンスデータに対応した Velvet（Zerbino and Birney. 2008）が実用化され、その有効性が示された。しかし、Velvet もゲノムサイズの大きなサンプルではメモリ使用量が急増するというケースが存在しており（Salzberg et al. 2012）、ゲノムサイズが大きい真核生物向けのアセンブラが新規開発された（Li et al. 2010; Gnerre et al 2011）。その中でも SOAPdenovo はジャイアントパンダゲノムを Illumina データのみでアセンブリし（Li et al. 2010）、ゲノムサイズ 2 Gbp 以上のサンプルで Illumina データと de Bruijn グラフの有効性を示した点で重要である。

序論で示したが、真核生物の *de novo* アセンブルにおける大きな問題の一つが、ヘテロ接合性の高さに起因したものである。この問題点はサンガー法、Illumina 両方のデータで報告されており（Vinson et al. 2005; Sodergren et al. 2006; Velasco et al. 2007; The Potato Genome Sequencing Consortium 2011; Star et al. 2011; Takeuchi et al. 2012; Zhang et al. 2012; Nystedt et al. 2013; You et al. 2013; Zheng et al. 2013）、このことからヘテロ接合性の問題はデータやアルゴリズムに特有のものではなく、*de novo* アセンブリにおける普遍的な問題であることが分かる。この問題に対し、様々な解決策が取られてきた。例えばジャガイモゲノムのアセ

ンブルでは、元来の異質 4 倍体に対して交雑実験を行い、double monoploid クローンと呼ばれる同質 2 倍体を作成することで解決している（The Potato Genome Sequencing Consortium 2011）。また、牡蠣（Zhang et al. 2012）、コナガ（You et al. 2013）、オウシュウトウヒ（Nystedt et al. 2013）のゲノム解読では、fosmid pooling と呼ばれる手法でゲノム全体をカバーする fosmid 配列のライブラリを構築し、各配列をショットガン法でアセンブリした後、更に overlap-layout-consensus アルゴリズムで長い配列を構築している。これは階層的ショットガン法の 1 種と言える。各 fosmid はハプロイド由来であるため、その配列をアセンブリする段階まではヘテロ接合性の問題を避けることができる。その後の overlap-layout -consensus を行う際にはヘテロ領域に対応する必要があり、それぞれの解読計画で独自のパイプラインを用いている。以上見てきたような手法でヘテロ接合性の高い生物種のゲノム決定は試みられてきているが、double monoploid の構築は植物など限られた生物種に対してのみの適用可能である上に、fosmid ライブラリの構築と同様に時間と金額のコストがかかるため、全ゲノムショットガン法によるアセンブリで同等以上の結果が得られるならば、効率的に全ゲノム配列を決定する上で極めて効果的であるといえる。Illumina 社製シークエンサの適用により、ゲノム配列の決定コストが大幅に引き下げられたことで、非モデル生物のゲノム決定が試みられる機会が増え、ヘテロ接合性が問題となるケースが増えてきていることは注目すべき点である。

2.2　真核生物用 *de novo* アセンブラ Platanus の開発

非モデル真核生物のゲノムを構築することを目的として、本研究では *de novo* アセンブラ、Platanus を新たに開発した（Kajitani et al. 2014）。特徴としては高ヘテロ接合性の 2 倍体データからも高精度な配列をアセンブルすることができるという点が挙げられるが、ヘテロ接合性以外の障害に対しても対策は施されており、低ヘテロ接合性サンプルやゲノムサイズが 1 Gbp を超える生物種のデータにも汎用的に使用可能となるよう設計されている。

ソースコードについては、バージョン 1.0.0 から 1.1.4 を筆者が C 言語で実装し、その後コードの可読性向上の目的で 1.1.4 と同様の内容のものが C++でバージョン 1.2.0 として実装された。その後機能を追加し、バージョン 1.2.1 とした。以降の節で説明するアルゴリズムはバージョン 1.2.1 のものである。

2.2.1　Platanus のアルゴリズムの概要

Platanus は以下の様な３つのサブプログラムからなっている。

1. Contig-assembly

   ハイスループット DNA シークエンサのデータから de Bruijn グラフというデータ構造を用いて contig（アセンブルされた配列）を出力する。小規模な変異によって生じるグラフ上の構造を単純化する機能を持つ。

2. Scaffolding

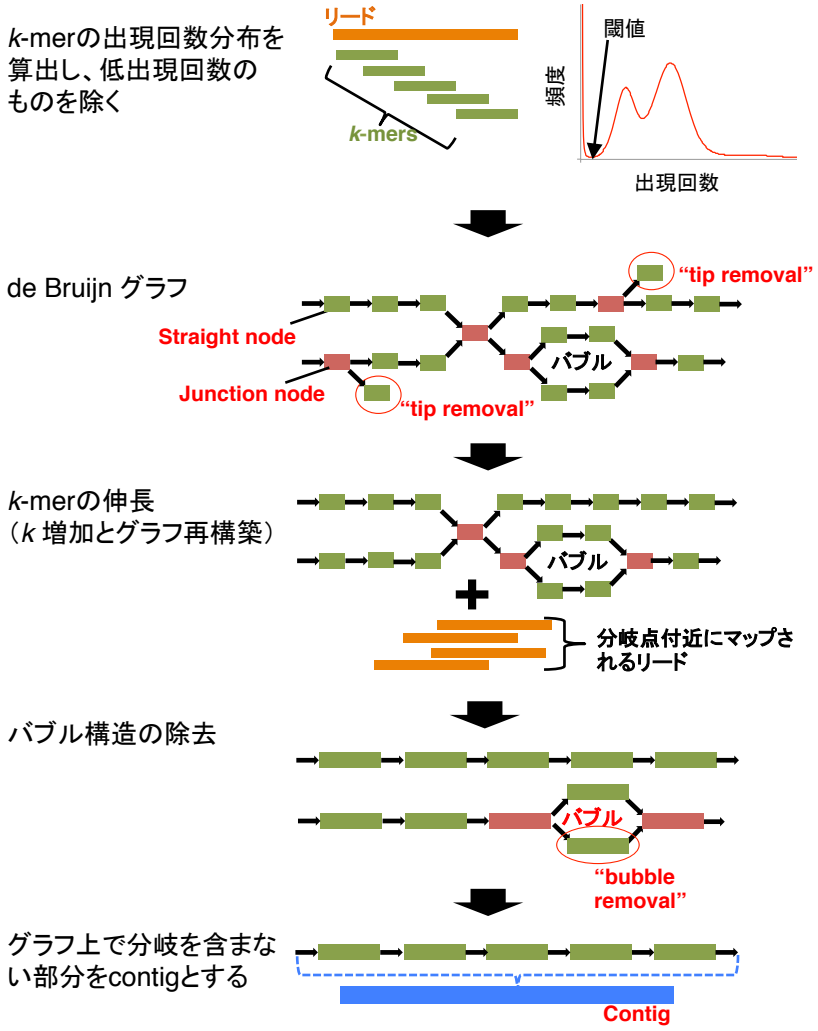
   Contig 上に paired-ends（長い断片の両端の配列）reads をマッピングしてレイアウトを決定し、ギャップを含む scaffold 配列を出力する。構造変異を含む高ヘテロ接合性の領域を統合する機能を持つ。

3. Gap-close

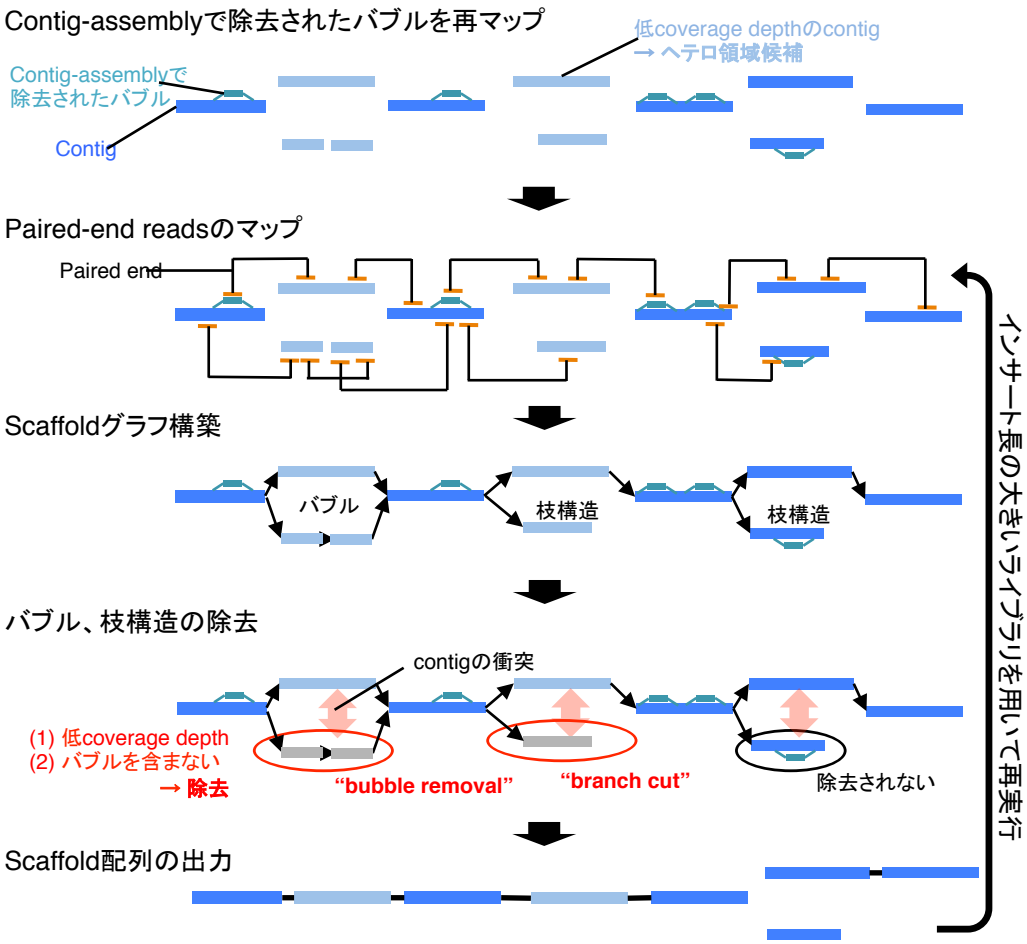
   Scaffold 配列中のギャップ部分を paired-ends reads を用いて構築する。

次節以降でこれらのアルゴリズムを詳細に述べることとする。全体像の模式図を図 2-1 に示す。

(A) Contig-assembly



(B) Scaffolding



(C) Gap-close

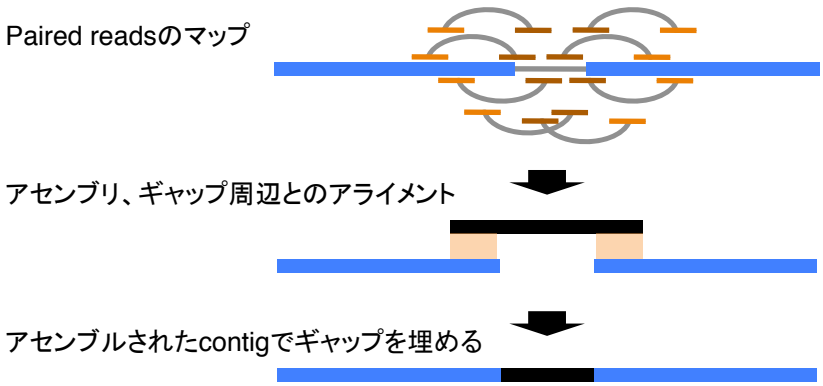


図 2-1 Platanus の全体像

2.2.2　Contig-assembly のアルゴリズム

Contig-assembly の模式図を図 2-1A に示す。

・*k*-mer 出現回数の分布の算出

最初に、入力断片配列（リード）中の長さ *k* の部分文字列（*k*-mer）をカウントする。*k* の初期値はデフォルトでは $k_0 = 32$ であり、A、T，G，C 以外の文字を含む *k*-mer は無視される。カウント後、各 *k*-mer の出現回数（coverage depth）はハッシュテーブルとして記録され、分布も算出される。その後、coverage depth の小さい *k*-mer はエラーを含んでいるという仮定より、エラー由来とゲノム配列由来の *k*-mer を区別するための閾値が決定される。ここでは、分布左側の極小値をウィンドウ（サイズ 7）を用いて検出して閾値（c_0）とする（図 2-2）。ここで、ウィンドウサイズの 7 は *k*-mer coverage depth の分布のピーク値が比較的小さい（50 以下）の場合を考慮し、また奇数である方が極小値の計算が容易であることから選択された。

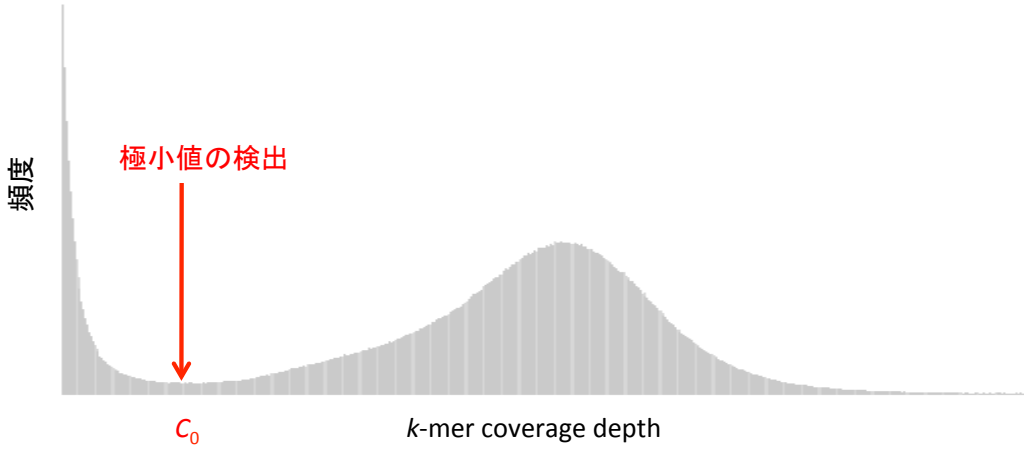


図 2-2　*k*-mer coverage depth の分布と c_0 の決定

・de Bruijn グラフの構築

de Bruijn グラフは、ハイスループットシークエンスデータの *de novo* アセンブリに対して多くのツールで採用されている（Zerbino and Birney 2008; Li et al. 2010b; Bnerre et al. 2011）。サンガー法データに用いられていた overlap–layout -consensus アルゴリズム（Myers et al. 2000; Batzoglou et al. 2002）と比較して高速

であるという利点があり、Platanus でも用いられている。Contig-assembly においては、coverage depth が c_0 以上の 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。Platanus での実装上は厳密な意味での de Bruijn グラフとは異なり、次のような性質を持つ。

(1) ある k-mer とその相補配列は同一に扱われる。

(2) グラフは有向グラフである。

(3) 各節点は coverage depth の値を持つ。

また、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）として扱われる。つまり、straight node は複数の k-mer を含む。最終的には、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 と呼ばれる（図 2-3）。これ以後、ある節点を v としたとき、入次数、出次数、coverage depth、含まれる k-mer 数をそれぞれ $in(v)$、$out(v)$、$c(v)$、$|v|$と表記する。

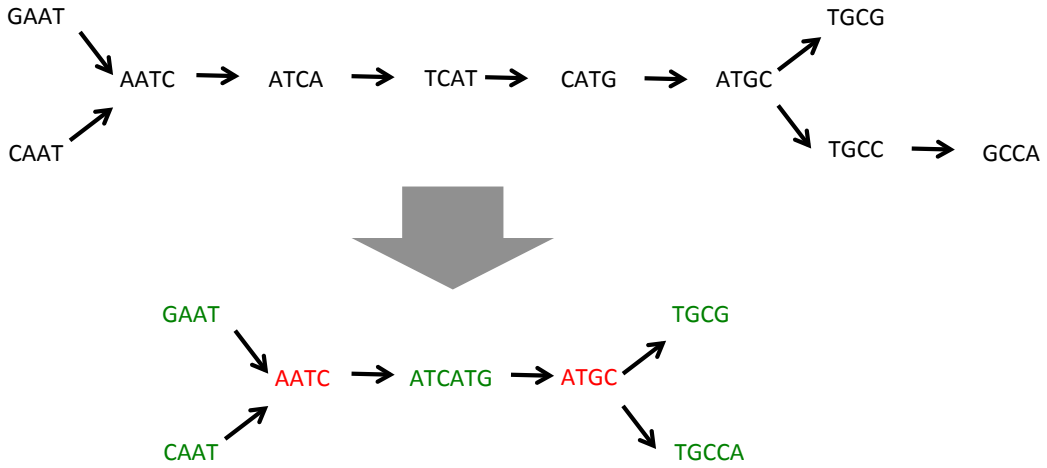


図 2-3 de Bruijn グラフ（$k = 4$）

上のグラフは k-mer を節点とした de Bruijn グラフ。下のグラフでは、緑色の節点は straight node、赤色の節点は junction node を表す。

straight node の組において両側の隣接節点が共有されているとき、バブル構造と呼ばれる（本節で後述）。この構造は single nucleotide variant（SNV）、small indel、エラー等に対応する。グラフの構築後、バブル構造に含まれない節点から

coverage depth の平均が算出され、この値はホモ領域（両方の相同染色体に存在する領域）の coverage detph としてその後のステップで扱われる。

・エラー由来の枝構造（tip）の除去

coverage depth が c_0 以上の k-mer 中にもエラーを含み、元のゲノム配列には存在しないものが一部含まれる。エラーは枝構造（tip）として表れることが報告されており（Zerbino & Birney, 2008）Platanus もそれを除去する機能を持つ。ここでは、ある straight node の coverage depth が低く、長さ（含まれる k-mer 数）が短いとき tip として除去される。具体的には、次の 3 つの条件が満たされたとき除かれる（図 2-4）。C_{tip} は定数で、デフォルト値は 0.5 である。

$$|v| < 2k$$

$$out(v) + in(v) = 1$$

$$c(v) \leq C_{\mathrm{tip}} \cdot \max_{w \in W} c(w)$$

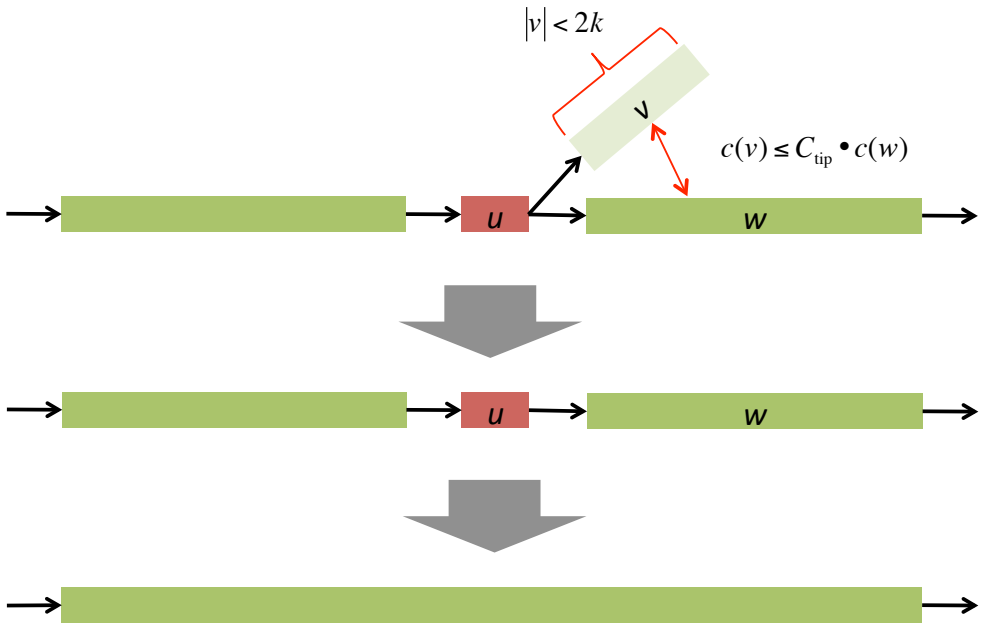


図 2-4　エラー由来の枝構造の除去

・k-mer 伸長と de Bruijn グラフの再構築

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 の値の利点を活用するため、k_0（デフォルト 32）で de Bruijn グラフを構築した後、k_{step}（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（図 2-5）。k_0 のデフォルト値は、リード長が 75 bp 以上の場合に十分な数のリード間オーバーラップを確保し、さらに 64 bit 整数としてプログラム中で保持されることで実行時間やメモリを削減できるという点から 32 という値が選択された。

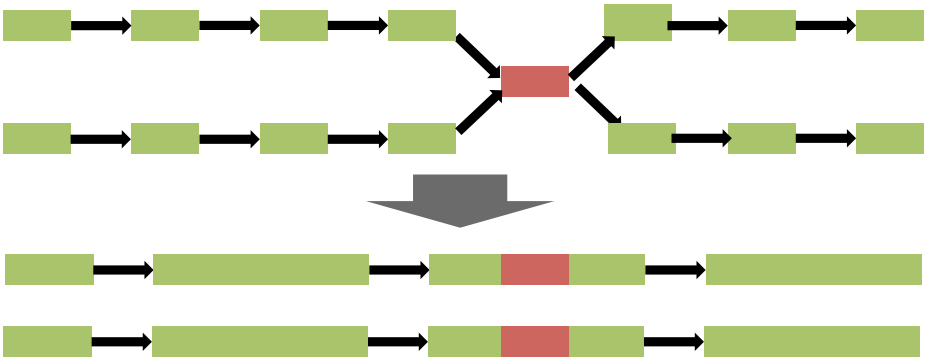


図 2-5　*k*-mer 伸長によるリピートの解決

緑色の四角形は *k*-mer、赤色の四角形はリピート配列を表す。

k の最大値 k_{max} は k_0-mer の平均 coverage depth と平均リード長より自動的に計算される（本節で後述）。以下に、k_{pre}-mer のグラフから k-mer のグラフを再構築する手順を示す（$k_{pre} < k$）。

(1) k_{pre}-mer のグラフにおいて、各 straight node の配列を分岐点に到達するまで延長する。ただし、延長する長さは最大$(k - k_{pre})$文字までである（図 2-6）。

(2) 延長された straight node 配列中の k-mer を抽出する。各 k-mer の coverage depth は次のように再計算する。

$$c(v) \cdot \frac{r - k + 1}{r - k_{\mathrm{pre}} + 1}$$

(3) k_{pre}-mer のグラフ中で、分岐点（junction node）から $(k - k_{pre})$ 以内の距離にある k_{pre}-mer を記録し、それを含むリードを収集する（図 2-7）。

(4) (2)で得られた k-mer と(3)で収集されたリード中の k-mer から de Bruijn グラ

フを構築する。coverage detph が c_i より小さい k-mer は使われない（c_i の算出方法は本節で後述）。

この方法を用いることで、k_{pre}-mer グラフでのオーバーラップ情報を次のグラフに活用しつつ、k を増加させることができる。なお、全てのグラフでエラー由来の枝の除去は行われる。

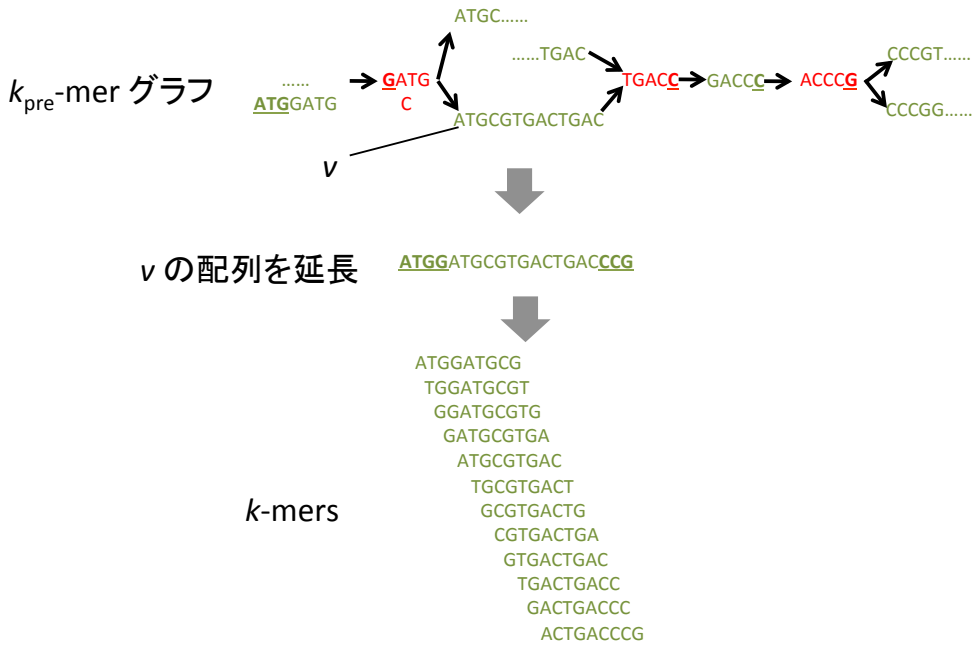


図 2-6　グラフ再構築時の straight node 配列の延長

$k_{pre} = 5$、$k = 9$ の場合を例として示す。5-mer のグラフから 9-mer のグラフを再構築する場合に対応する。

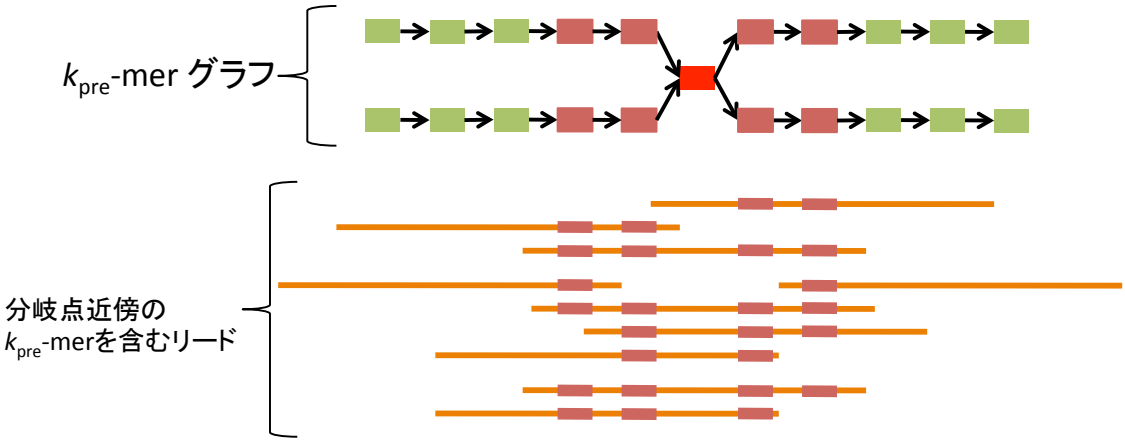


図 2-7　グラフ再構築時の、分岐点付近にマップされるリードの収集

$k - k_{pre} = 2$ の場合を例として示す。

・k_{max} と c_i の計算方法

リピート配列を介してミスアセンブリが起こる場合が考えられるが、coverage depth が十分に高くない場合はその数が増加する可能性がある（図 2-8）。ここでは、coverage depth がポアソン分布に従うと仮定し、本来あるべき辺が coverage depth のゆらぎにより構築されない確率を反映したスコアを以下のように定義する。a を平均 coverage depth、c を coverage depth の下限として、

$$s_{\mathrm{split}}(k, k_{\mathrm{pre}}, a, c) = 1 - (1 - e^{-a} \sum_{j=0}^{c-1} \frac{a^j}{j!})^{k-k_{\mathrm{pre}}}$$

とする。s_{split} の値が低い場合、coverage depth が十分でなくミスアセンブリが起こる可能性が多いと想定する。k_{pre}-mer グラフの coverage depth の平均、下限をそれぞれ a_{pre}、c_{pre} とし、平均リード長を r とすると、k-mer グラフの平均 coverage depth a は

$$a = a_{\mathrm{pre}} \frac{r - k + 1}{r - k_{\mathrm{pre}} + 1}$$

と計算される。k-mer グラフの新たな coverage depth の下限 c_i は次の条件をともに満たす最大の c の値である。

(1) $c_{\min} \leq c \leq c_{\mathrm{pre}}$

(2) $s_{\mathrm{split}}(k, k_{\mathrm{pre}}, a, c) \leq s_{\max}$

ここで $c_{\min}$、$s_{\max}$ は定数でデフォルトではそれぞれ 2、10^{-10} である。条件を満たす c_i が存在するような k の値のうち、最大のものを $k_{\max}$ とする。つまり、c_i が算出できないとき、k を増加させることを中止する。これらの計算により、ミスアセンブリの可能性を高める coverage depth の下限、k-mer 長の適用が防がれると期待される。

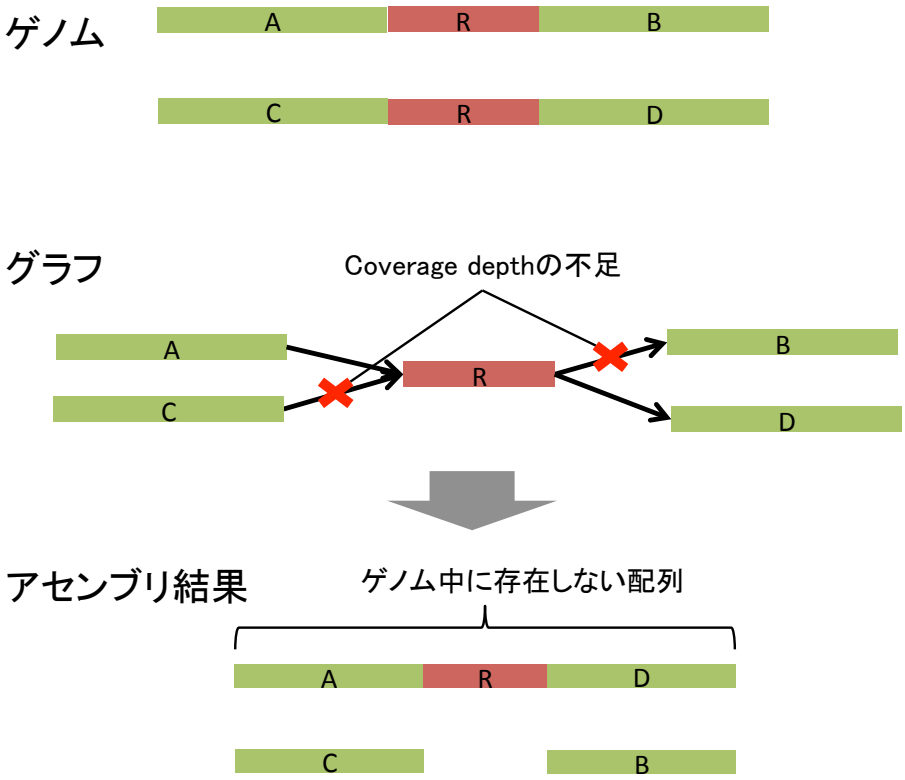


図 2-8　リピート配列を介したミスアセンブリの模式図

R（赤色の四角形）はリピート配列を表す。

・バブル構造の除去

k_{max}-mer のグラフを構築し、エラー由来の枝構造（tip）の除去を行なった後、変異またはエラーに由来するバブル構造の除去を行う。このようなアルゴリズムは他の de Bruijn グラフに基づく *de novo* アセンブラにも実装されているが（Zerbino and Birney 2008; Li et al. 2010b; Bnerre et al. 2011）、Platanus での具体的な除去基準についてここでは述べる。バブル構造は２つの staright node と２つの junction node から成り、straight node はそれぞれ同一の junction node に接続されている（図 2-9）。ヘテロ領域は相同染色体の片方のみにしか配列が存在しないため、coverage depth はホモ領域と比較して低くなる。Platanus はバブル内の straight node の配列の相同性が高く、coverage depth は低い場合に、片方の straight node を除くことでバブル構造を除去する。straight node を v、u とし、対応する配列の組でアライメントを行なったときの編集距離（ミスマッチとギャップの合計）を $edit(v, u)$とする。C_{bubble} を定数とし（デフォルト　0.1）以下の条件がともに満たされるとき、coverage depth の高い方の straight node（v か u）をグラフから除く。

(1) $c(v) + c(u) < 1.5a$　(a: 平均 k_{max}-mer coverage)

(2) $edit(v, u) \leq C_{bubble} \cdot \max(|v|, |u|)$

除去された配列は別のファイルに保存され、後の解析に用いることもできる。

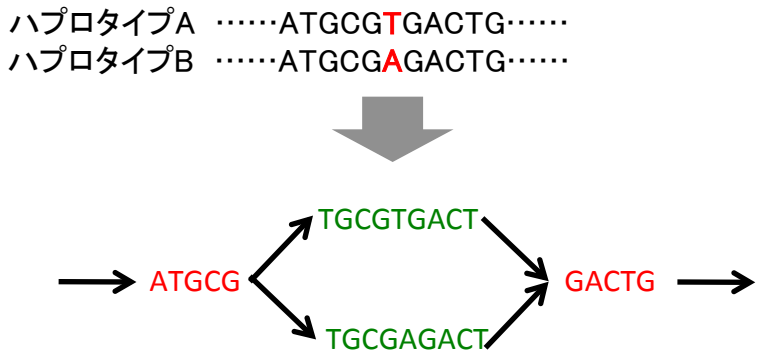


図 2-9　de Bruijn グラフにおけるバブル構造

SNV により生じた構造の例を示している（$k = 5$）

・リードの再マッピングによる contig の修正

以下にアルゴリズムを記す。バブル構造の除去後、straight node 上にリードを再マップしミスアセンブリが疑われる箇所を検出する。具体的には、straight node 上に完全一致のみでリードをマップし、リード間で長さ k_0（デフォルト 32）以上のオーバーラップが存在しない位置で straight node を分断する（図 2-10）。分断後の straight node がそれぞれ最終的な cotnig に対応する。

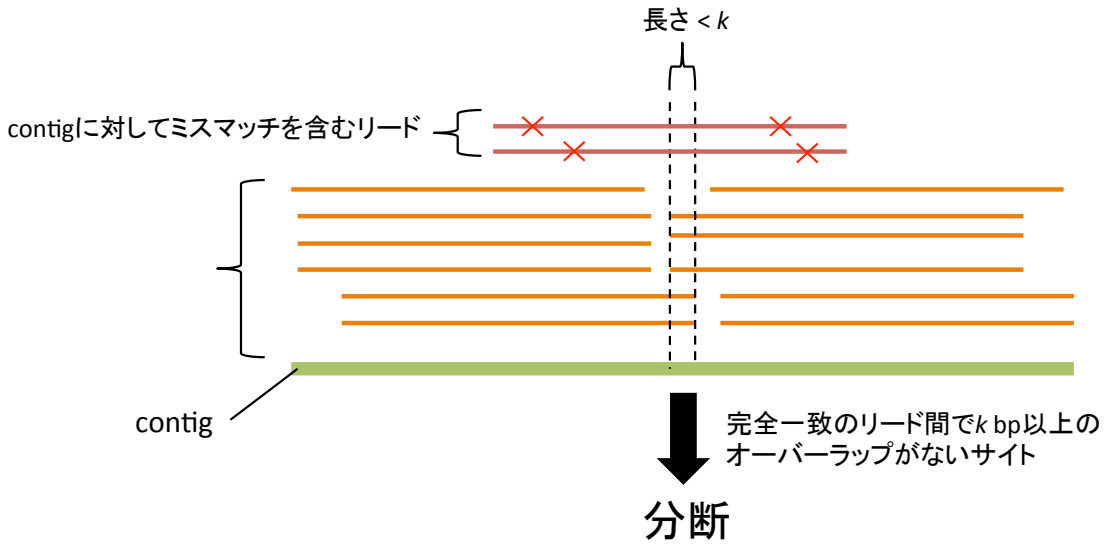


図 2-10　リードの再マップによる contig 配列の修正（分断）

2.2.3　Scaffolding のアルゴリズム

Scaffolding の模式図を図 2-1B に示す。

・paired-end（mate-pair）ライブラリのマッピング

Scaffolding では最初に contig 上に paired-end または mate-pair のリードをマップする。それらのライブラリは長い配列の両端をシークエンスしたものであり、ゲノム上での contig の順序を決定し、scaffold 配列を構築するために用いることができる。リードのマップは contig 中のユニークな *k*-mer（デフォルト　$k = 32$）をキー、contig 中の位置を値としたハッシュテーブルを用いて行う。ユニークな *k*-mer とは、全 contig 中に 1 つだけ含まれる *k*-mer である。リードのマップの際には、リード配列から重複なしで *k*-mer を取り出し contig にマップする。リードがマップされる contig は多数決で決め、位置は平均により求める（図 2-11）。この方法はアライメントを用いるより高速であり、ローカルアライメントのようにリードの一部のみがターゲット配列と一致する場合もマップされるため、contig の端に位置するリードもマップ可能である。さらに、ユニークな *k*-mer のみを用いることでリピート配列によるミスマッピングを防ぐ機能も持つ。

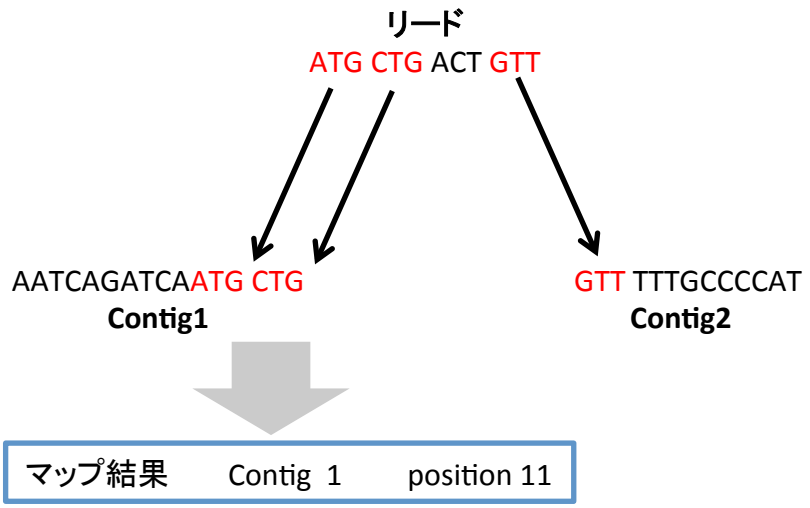


図 2-11　Scaffolding での contig へのリードのマッピング

3-mer によるマッピングの例。リード中の 4 つの 3-mer のうち 2 つが contig1 にマップされるため、リードは contig1 にマップされる。

・Contig-assembly のバブル配列を利用したリードのマッピング

ヘテロ領域については、contig 配列は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、リードをより多くマップするため、Contig-assembly の際に除去されたバブルの配列を利用する。そのために、リードのマップ前に Contig-assembly 時のバブル配列を contig 上にマップしておく。次に、contig にマップされないリードに対してバブルの配列へのマップを試み、マップされたならばバブル配列上の位置を contig 上の位置に変換する（図 2-12）。

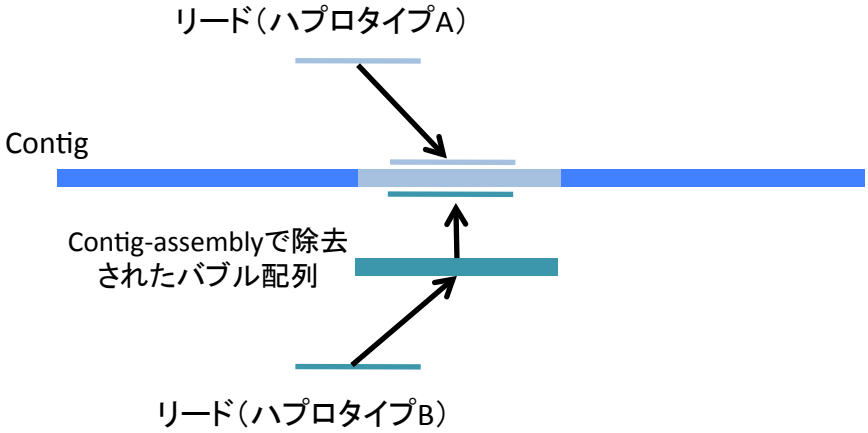


図 2-12　バブル配列を利用したリードのマッピング

・インサートサイズの推定

ライブラリ毎のインサートサイズの平均、標準偏差は、同一配列上にマップされたリードの組から推定する。ここで、インサートサイズはペアリード間の距離に対応する。Platanus はインサートサイズが小さいライブラリから順に scaffolding に用いるよう設計されているため、あるライブラリの値は、一段階手前のライブラリで構築された scaffold へのマッピング結果より推定される。これにより、マッピング対象の配列（contig または scaffold）が短いことにより、インサートサイズの大きいライブラリについての推定値が狂う可能性を減らすことができる。平均と標準偏差の算出時には外れ値の影響を防ぐため、インサートサイズの分布中でピークを検出し、[0.25×ピーク値, 1.75×ピーク値] の範囲内の値のみを用いる。ピークはサイズ 101 のウィンドウを用いて検出する。

・scaffold グラフ

リードのペアが異なる congtig 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する。 各 contig を節点とし、 閾値 n 以上の数のリンクを持つ contig の組を辺で結び、 scaffold グラフを構築する。 ここで、辺はリンクの数と contig 間の距離情報を持つものとする（図 2-13）。

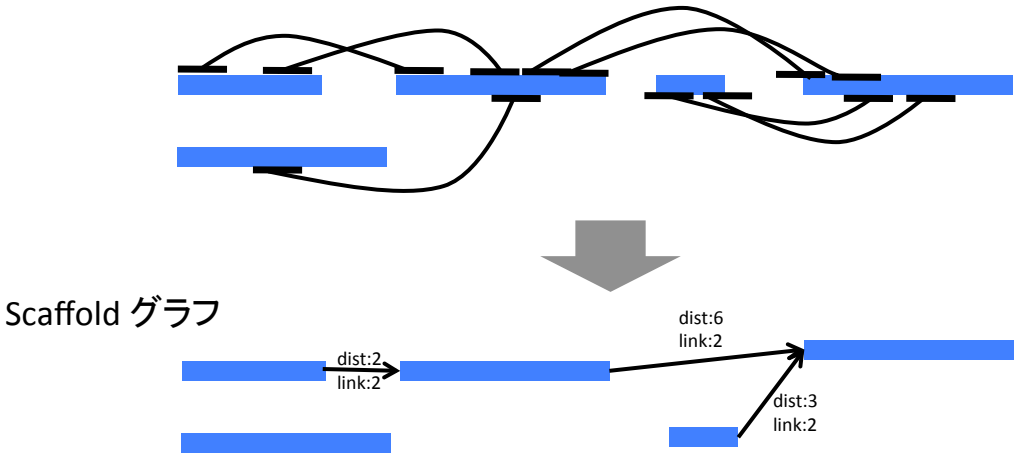


図 2-13　scaffold グラフ

リンクのためのペアリード数の下限（n）が 2 の場合の例。"link"はリンクに含まれるペアリード数、"dist"は contig 間の距離。

contig 間距離の推定とリンク数の下限 n の決定方法を記すため、以下の変数を定義する。

r: 平均リード長

c: 平均 coverage depth (マップされたリードの合計長 / contig の合計長)

μ: 平均インサートサイズ

σ: インサートサイズの標準偏差

l_1, l_2: contig 長

g: 実際の contig 間の距離

インサートサイズ分布は正規分布（平均 μ、分散 σ^2）に従うと仮定し、最尤法により contig 間距離を推定する。contig 間距離を d としたとき、リンクに含まれるペアリードのそれぞれのインサートサイズを $s_i(d)$ (i = 1, 2, ..., ペア数) とする。

推定される距離は以下のようになる。

$$\underset{d}{\operatorname{argmax}} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(s_i(d)-\mu)^2}{2\sigma^2}\right)$$

リンクのペア数の下限 n は２通りの方法で決定される。１つは定数 $n_{\min}$（デフォルト　3）であり、もう１つの $n_{\exp}$ は $g=\mu-2r$、 $l_1=l_2=$ 平均 contig 長、としたときのペア数の期待値である。その期待値は以下のように計算される。

$$\int_0^{l_1-r} \int_{l_1+g+r-y}^{l_1+g+l_2-y} \frac{c}{2r} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dxdy$$

Scaffolding ステップでは、$n_{\exp}$ を適用して scaffold を構築した後、$n_{\min}$ を適用してさらに配列の延長を行う。なお、$n_{\exp}$ はライブラリ毎に計算される。

・scaffold グラフの節点（contig）の衝突

scaffold グラフ中で節点 u、w が節点 v に接続している場合を考える。v を基準にした u と w の位置より、それらが長さ o 以上オーバーラップしている場合、v と w は「衝突」していると定義する（図 2-14）。閾値 o は２通り適用される。最初に $o=2\sigma$（σ：インサートサイズ標準偏差）で scaffold を構築後、$o=3\sigma$ として再構築を行う。

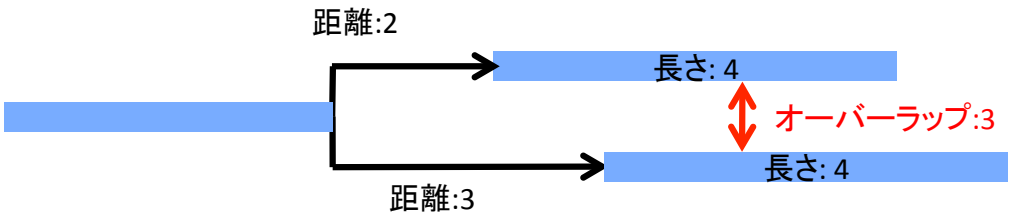


図 2-14　scaffold グラフにおける節点の衝突

・scaffold 配列の構築

V_{repeat} を衝突する隣接節点の組を持つ節点の集合、 V_{used} を scaffold 構築に既に使われた節点の集合とする。以下に scaffold 配列構築の手順を述べる。リピート配列候補（V_{repeat}）を介して節点をつながないようアルゴリズムは設計されている。

(1) $(v_0 \notin V_{\mathrm{repeat}}$ and $v_0 \notin V_{\mathrm{used}})$ を満たす節点をランダムに v_0 とし、節点集合 V_{scaffold} を $V_{\mathrm{scaffold}} = \{v_0\}$ として初期化する。

(2) 次の条件 $(u \in V_{\mathrm{scaffold}}$ and $u \notin V_{\mathrm{repeat}})$ を満たす辺(u, w) を探す。節点 w が V_{scaffold} の要素と衝突しないとき、w は V_{scaffold} と V_{used} に加えられる。w の候補が複数あるときは、v_0 からのグラフ上の距離が小さいものが選ばれる。

(3) V_{scaffold} に加えられる候補の節点が無くなるまで(2)を繰り返す。

(4) V_{scaffold} を scaffold 配列として保存する。

(5) ステップ(1)–(4) を v_0 となる節点の候補が無くなるまで繰り返す。

・エラー由来の辺の除去

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より生じたものとみなして除去する（図 2-15）。具体的な手順は次の通りである。節点 v, u, w について、u と w が衝突している場合を考える。v, u 間で $n(v, u)$をペアリードの数、$e(v, u)$をペアリード数の期待値とする。C_{cut}（デフォルト　0.5）を定数として、下の条件が満たされるとき、辺 (v, u)を除く。

$$\frac{n(v,u)}{e(v,u)} < C_{\mathrm{cut}} \cdot \frac{n(v,w)}{e(v,w)}$$

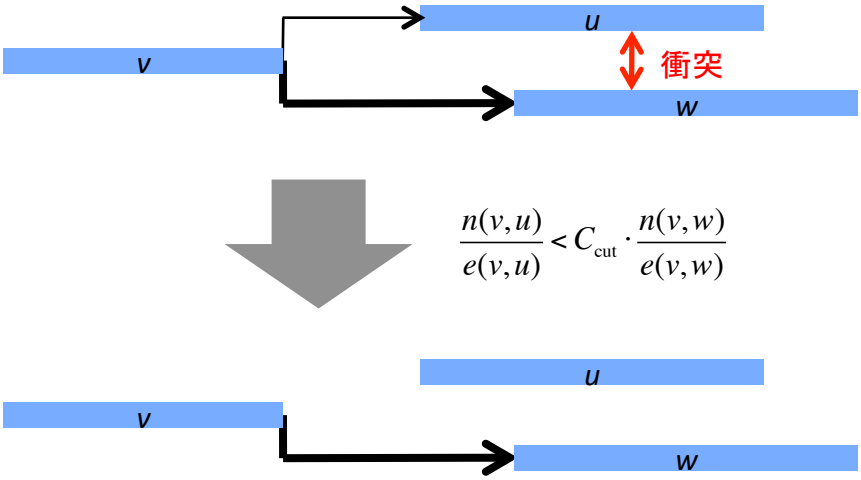


図 2-15　エラー由来の辺の除去

・scaffold グラフ中のバブル構造の除去（bubble removal）

de Bruijn グラフのバブル構造と同様に、変異はバブル構造として表れる。ここでは、SNV や small indel のみならず構造変異なども検出できる可能性がある。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変異からは起こりえない。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（図 2-16）。以下に具体的な方法を述べる。u と w はともに v に隣接しており、 かつ衝突しているとする。 u と w を起点 v_0 として 2 つの scaffold V_u、 V_w を構築する。 V_u、V_w の両端で contig が共有されているならば、その contig を除いた部分を V'_u 、V'_w とする。V'_u 、V'_w を配列としてアライメントし、編集距離を計算する。 表記について以下のように定義する。

$|V'|$： scaffold V'の長さ

$c(V')$： scaffold V'の coverage depth

$bub(V')$： scaffold V'上にマップされるバブル（Contig-assembly 由来）の数

$edit(V'_u, V'_w)$：V'_u と V'_w の編集距離

C_{sim}: 定数（デフォルト　0.1）

C_{homo}: 定数（デフォルト　1.5）

C_{hetero}: 定数（デフォルト　0.75）

次の２条件のうちいずれかが満たされたとき、バブルを除去する。

(1)

$c(V'_u) + c(V'_w) \leq 2a$ and

$edit(V'_u, V'_w) \leq C_{\mathrm{sim}} \cdot \max(|V'_u|, |V'_w|)$ and

$(bub(V'_u) = 0 \text{ or } bub(V'_w) = 0)$

(2)

$c(V'_u) \leq C_{\mathrm{hetero}} \cdot a$ and

$c(V'_w) \leq C_{\mathrm{hetero}} \cdot a$ and

$c(v_{\mathrm{right}}) \leq C_{\mathrm{homo}} \cdot a$ and

$c(v_{\mathrm{left}}) \leq C_{\mathrm{homo}} \cdot a$ and

$(bub(V'_u) = 0 \text{ or } bub(V'_w) = 0)$

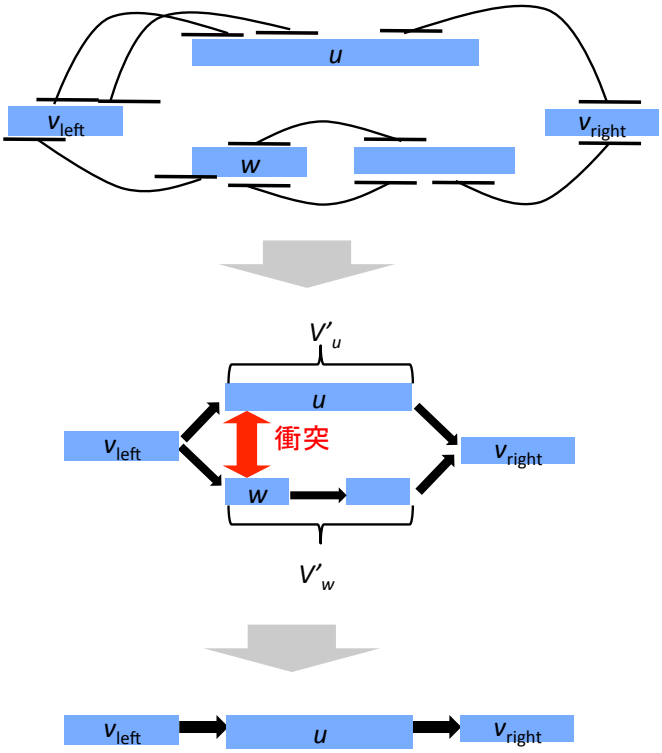


図 2-16　scaffold グラフ中のバブル除去

・ヘテロ領域に由来する枝構造の除去（branch cut）

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。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（図 2-17）。節点 V_{root}、V_u 、V_w について、V_u と V_w が衝突しているとする。次の条件が満たされるとき、接続している辺が少ない節点（V_u か V_w）をグラフから除く。

C_{homo}: 定数（デフォルト　1.5）

C_{hetero}: 定数（デフォルト　0.75）

$c(V_{\mathrm{root}}) \leq C_{\mathrm{homo}} \cdot a$ and

$\max\left(c(V_u), c(V_w)\right) < C_{\mathrm{hetero}} \cdot a$ and

$bub(V_u) = 0$ and

$bub(V_w) = 0$

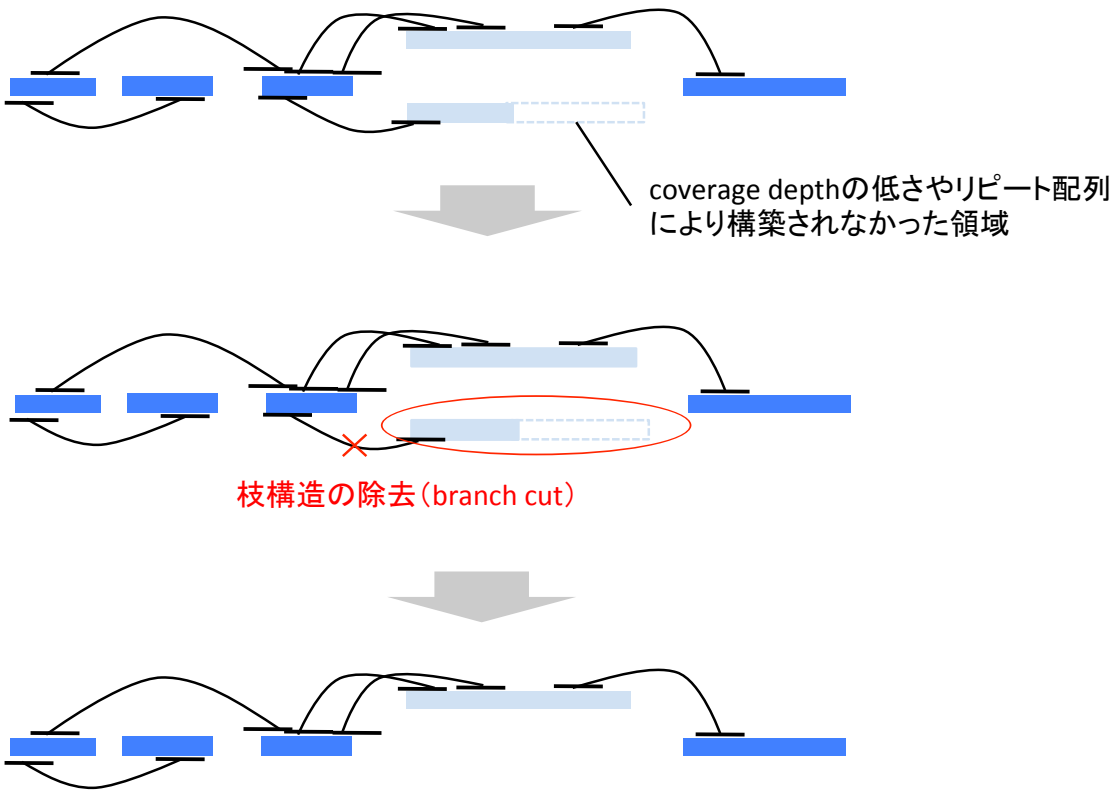


図 2-17　ヘテロ領域由来の枝構造の除去（branch-cut）

・インサートサイズの大きいライブラリによる修正

インサートサイズが小さいライブラリで構築した scaffold の配列を、よりインサートサイズが大きいライブラリで修正できる場合がある。ライブラリの physical coverage を（平均インサートサイズ × マップされたリード数 / contig 合計長）と定義する。インサートサイズと physical coverage がともに大きいライブラリで scaffolding を行う前に、各ギャップをまたぐリードペアが存在するか確認し、存在しない場合はミスアセンブリとみなして分断する（図 2-18）。具体的には、次の条件(1)–(4)が満たされたとき分断を行う。

s：ギャップをまたぐリードペア数

s_{exp}：ギャップをまたぐリードペア数の期待値

n_{min}：定数（デフォルト　3）

(1) $s < n_{\mathrm{min}}$

(2) $s/s_{\mathrm{exp}} < 0.1$

(3) $s_{\mathrm{exp}} > 1$

(4) 分断候補のギャップとは別にリンクが存在する（リードペア数 $\geq n_{\mathrm{min}}$）

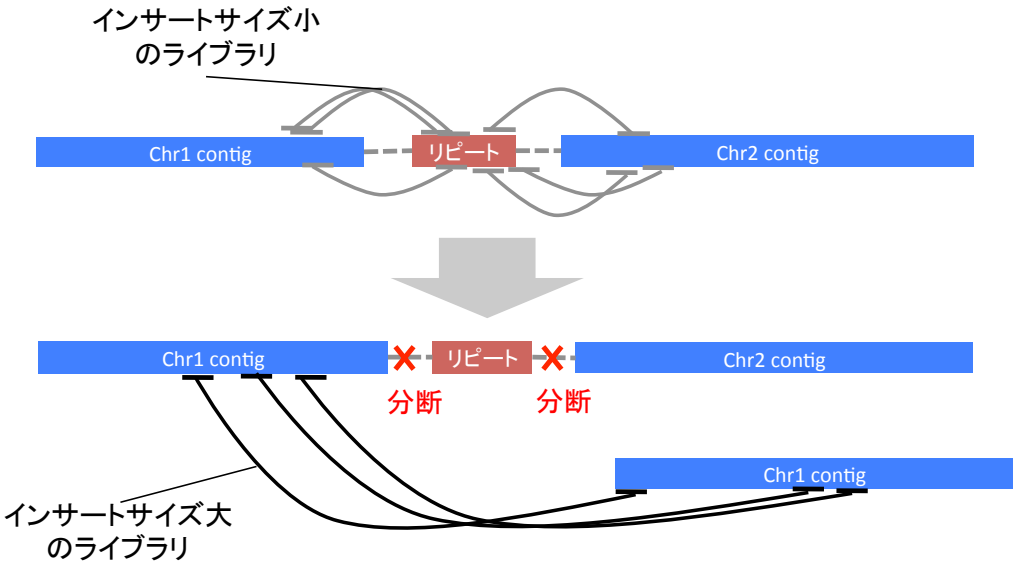


図 2-18　インサートサイズの大きいライブラリによるギャップの分断

2.2.4　Gap-close のアルゴリズム

Gap-close の模式図を図 2-1C に示す。

・ギャップをカバーするリードの収集

Gap-close では scaffold 配列上に paired-end（mate-pair）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を試みる。Scaffolding と同様にリードを scaffold 配列にマップし、 各ライブラリのインサートサイズを推定し、インサートサイズの小さなライブラリから順にギャップを閉じるステップを実行していく。最後に全てのライブラリのリードを一度に用いる。

ギャップ部分をカバーするリードの収集方法を述べる。あるペアの片方のリードがマップされたとき、もう片方のリードの位置をインサートサイズの情報から推定できる。

h_{left}: マップされなかったリードの左端位置
h_{right}: マップされなかったリードの右端位置
g_{left}: ギャップの左端位置
g_{right}: ギャップの右端位置
σ: インサートサイズの標準偏差

次の 2 条件のいずれかが満たされるとき、マップされなかったリードはギャップに対応付けられる（図 2-19）。

(1) $g_{\mathrm{left}} - 3\sigma \leq h_{\mathrm{right}}$
(2) $h_{\mathrm{left}} \leq g_{\mathrm{right}}$

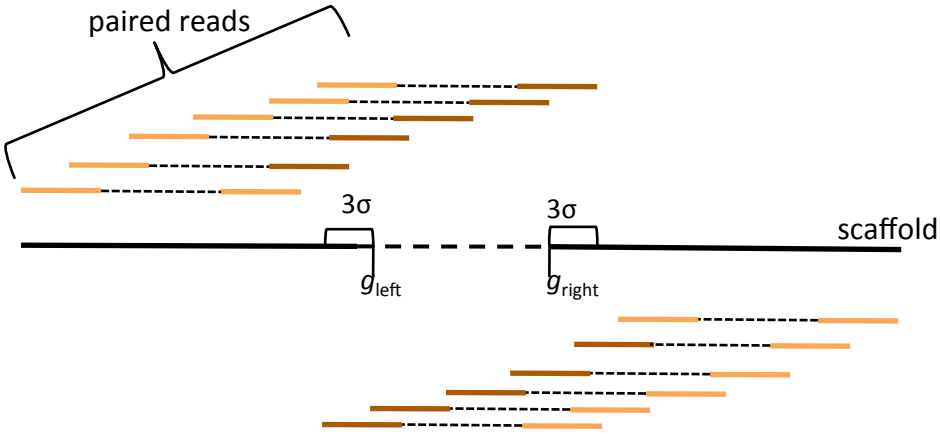


図 2-19　ギャップをカバーするリードの収集

・ローカルな *de novo* アセンブリによるギャップ部分の配列構築

ギャップ毎にリードを集めた後は、 そのリードをアセンブルし contig を構築する。 de Bruijn グラフを用いたアセンブリでは、*k* と coverage depth の下限を動的に変更しない。 *k* の値は 24、 72 の２通りのみとし、coverage depth の下限は 3 とし、*k*-mer 伸長のステップ数を減らすことで高速化が図られる。 ゲノムサイズが数ギガ bp のサンプルではギャップの個数がメガ単位になることがあるため、速度を重視している。また、リードのマップ結果よりギャップをカバーするリードのストランド方向が定まるため、相補的な *k*-mer の組を区別する。de Bruijn グラフでギャップを閉じる contig が構築されない場合、overlap-layout -consensus によるアセンブリが実行される。最小オーバーラップ長は 32 である。アセンブリ後、各 contig についてギャップの周囲の配列とオーバーラップしているかをアライメントにより調べる。 デフォルトでは、 長さ 32 以上でミスマッチの割合が 0.05 以下のアライメントをオーバーラップと判定する。contig の両端がギャップ周辺の配列とオーバーラップしているとき、 その contig でギャップを閉じる（図 2-20）。

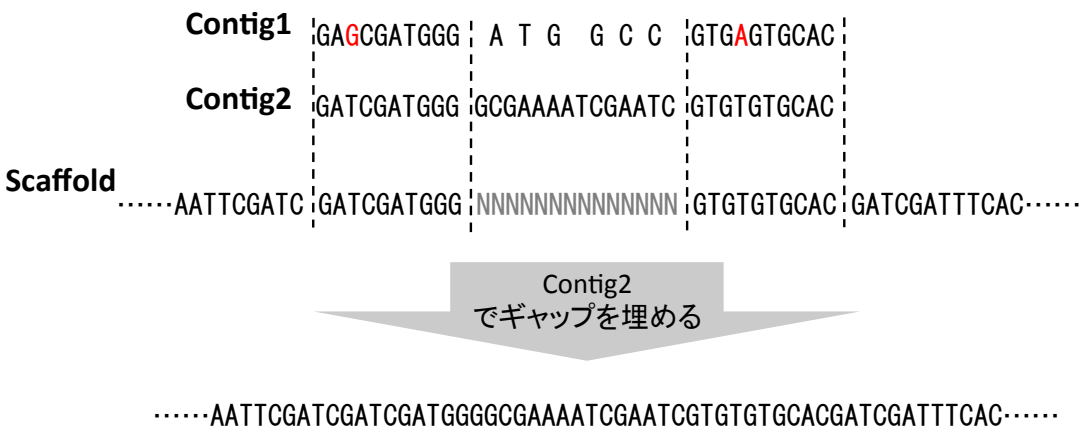


図 **2-20** **contig** によるギャップ部分の配列構築

2.3　真核生物データに対する Platanus の有用性の検証

本節では、前節で説明したアルゴリズムに基づいて開発された Platanus の性能評価をするため、様々な値のヘテロ接合性を持つシミュレーションおよび実データでベンチマークを行なった結果を示す。

2.3.1　概要

ベンチマークの概要を以下に示す。アセンブリ結果の長さを評価するには NG50 という値を用いた。アセンブリ結果において、ゲノムサイズの 50%がこの値以上の長さの配列に含まれることを意味し、長さの平均値や中央値より、エラーやリピート由来の短い配列の影響を受けにくい性質がある。NG50 は scaffold と contig に対してそれぞれ計算することができる。contig の定義は、ベンチマークツール GAGE に従い、scaffold を 3 bp より長い'N'が存在する箇所で分断した配列とする。精度評価としては、リファレンス配列が存在する *C. elegans*（The *C. elegans* Sequencing Consortium 1998）については評価ツールの GAGE（Salzberg et al. 2012）を用いた。その他の生物種については、fosmid または BAC により構築されたゲノムの部分配列とアセンブリ結果を比較することで評価を行なった。

・ベンチマーク対象の生物種と 17-mer 出現回数の解析

用いた生物種は以下の通りである。

・*Caenorhabditis elegans*

線虫のモデル生物であり、ヘテロ接合性は低い。実データと、それを基にヘテロ接合性をシミュレートしたデータを生成しベンチマークに用いた。

・*Strongyloides venezuelensis*

高ヘテロ接合性の線虫。ゲノムデータは未発表である。

・*Crassostrea gigas*

高ヘテロ接合性の牡蠣であり、ドラフトゲノムが発表されている（Zhang et al. 2012）。

・*Melopsittacus undulatus* ("bird")
de novo アセンブリコンテストである Assemblathon 2（Bradnam et al. 2013）にて使用された鳥の一種。

・*Boa constrictor constrictor* ("snake")
Assemblathon2 で使用された蛇の一種。

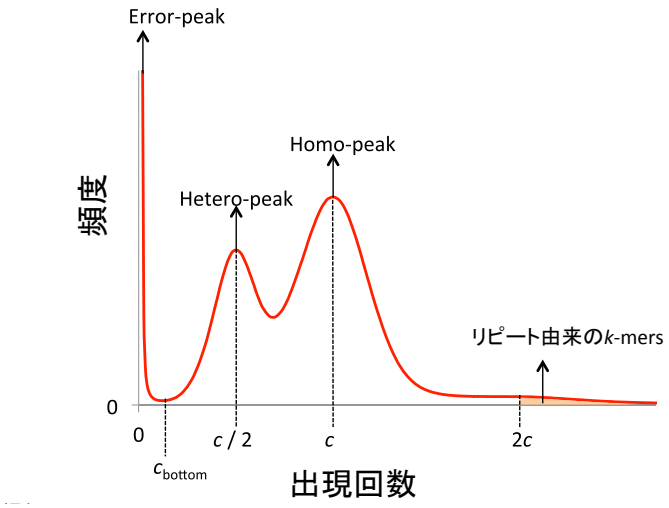
・*Maylandia zebra* ("fish")
Assemblathon2 で使用された魚の一種。

M. undulatus、*B. constrictor constrictor*、*M. zebra* は Assemblathon2 ではそれぞれ "bird"、"snake"、"fish" と呼ばれており、本論文でもそのように表記する場合がある。

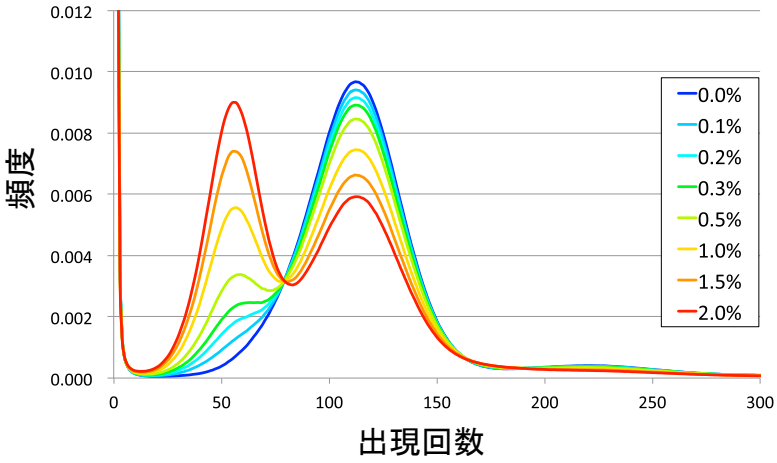
各生物種のゲノムの特徴を把握するため、17-mer の出現回数分布による解析を行なった（図 2-21）。これにより、ヘテロ接合性の大小、リピート配列の割合の大小、ゲノムサイズを推定することができる。paired-end リード中の全ての 17-mer をカウントし、出現回数の分布を算出している。ゲノムサイズに対して十分なデータ量がある場合、ホモ領域に対応するピークが見られるが、高ヘテロ接合性のサンプルの場合はヘテロ領域から産出されるピークの存在により２峰性の形をとる（図 2-21B）。さらに、リピート配列の割合が高いゲノムの場合はそれに対応するピークも見られる。またゲノムサイズは、出現回数が左側の極小値（図 2-21 A の c_{bottom}）より小さいエラー由来の 17-mer を除いて、（全 17-mer の数 / ホモ領域のピーク値）を求めることで推定できる。17-mer 出現回数分布解析結果を表 2-1 シークエンスデータの詳細を表 2-2 に示す。これらの解析結果より以下の知見がベンチマークデータに対して得られた。

・*C. elegans* と *S. venezuelensis* はリピート配列の割合が近い。
・*S. venezuelensis* と *C. gigas* は高ヘテロ接合性。
・*C. gigas* はリピート配列の割合が比較的高い。
・Assemblathon2 の３種はゲノムサイズが 1 Gbp 前後。

(A) 模式図



(B) *C. elegans* のヘテロ接合性（0.0–2.0%）シミュレーションデータ



(C) ベンチマーク対象の生物種

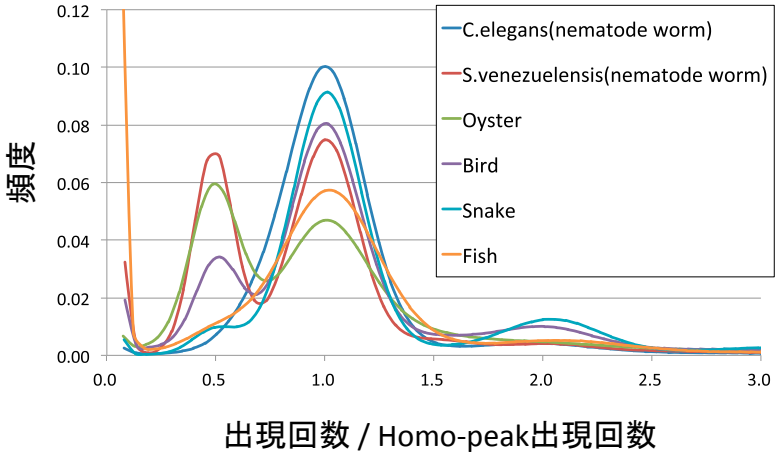


図 2-21　17-mer の出現回数の分布

(C)はホモ領域のピーク値で正規化されている。

表 2-1　17-mer 出現回数分布解析結果

生物種	ゲノムサイズ (Mbp)	ホモ領域の出現回数 (ピーク値)	Hetero-peak-height / Homo-peak-height	リピート由来 17-merの割合
C. elegans (nematode worm)	100.3	113	0.0704	0.236
S. venezuelensis (nematode worm)	57.7	111	0.955	0.289
Crassostrea gigas (oyster)	565.7	98	1.27	0.471
Melopsittacus undulates (bird)	1085.2	91	0.424	0.313
Boa constrictor constrictor (snake)	1431.5	77	0.108	0.436
Maylandia zebra (fish)	915.0	41	0.194	0.441

C. elegans 以外のゲノムサイズは 17-mer の出現回数分布から推定した値。Homo-peak、Hetero-peak、リピート由来の 17-mer については図 2-21A の模式図に準ずる。

表 2-2　シークエンスデータの詳細

生物種	*Caenorhabditis elegans* (nematode worm)		
ゲノムサイズ (bp)	100.3 M		
インサートサイズ (bp)	230	420	4,660
リード長 (raw) (bp)	110	110	100
合計サイズ (raw) (bp)	7.5G	7.2G	13.9G
リード長 (preprocessed) (bp)	107	106	87
合計サイズ (preprocessed) (bp)	7.2G	6.8G	2.8G

生物種	*Strongyloides venezuelensis* (nematode worm)		
ゲノムサイズ (bp)	57.7M		
インサートサイズ (bp)	200	450	3400
リード長 (raw) (bp)	110	100	100
合計サイズ (raw) (bp)	2.9G	5.2G	5.3G
リード長 (preprocessed) (bp)	104	96	69
合計サイズ (preprocessed) (bp)	2.7G	5.0G	3.6G

生物種	*Crassostrea gigas* (oyster)						
ゲノムサイズ (bp)	565.7M						
インサートサイズ (bp)	170	500	800	2000	5000	10,000	20,000
リード長 (raw) (bp)	90	90	90	90	90	90	90
合計サイズ (raw) (bp)	36.3G	18.7G	18.5G	50.7G	16.6G	18.8G	25.7G
リード長 (preprocessed) (bp)	86	84	82	71	82	66	60
合計サイズ (preprocessed) (bp)	34.7G	17.6G	17.0G	29.7G	4.2G	2.0G	2.6G

生物種	*Melopsittacus undulates* (bird)							
ゲノムサイズ (bp)	1085.2M							
インサートサイズ (bp)	220	500	800	2,000	5,000	10,000	20,000	40,000
リード長 (raw) (bp)	150	150	150	90	90	90	90	90
合計サイズ (raw) (bp)	48.4G	47.2G	43.1G	47.6G	35.0G	17.0G	16.1G	15.7G
リード長 (preprocessed) (bp)	135	128	115	82	81	79	78	65
合計サイズ (preprocessed) (bp)	43.6G	40.5G	33.0G	26.0G	15.7G	7.3G	3.7G	1.7G

生物種	*Boa constrictor constrictor* (snake)			
ゲノムサイズ (bp)	1431.5M			
インサートサイズ (bp)	400	2,000	4,000	10,000
リード長 (raw) (bp)	121	101	101	101
合計サイズ (raw) (bp)	136.0G	25.1G	17.4G	20.5G
リード長 (preprocessed) (bp)	118	65	68	75
合計サイズ (preprocessed) (bp)	132.1G	14.3G	1.0G	6.5G

生物種	*Maylandia zebra* (fish)						
ゲノムサイズ (bp)	915.0M						
インサートサイズ (bp)	180	2,500	5,000	7,000	9,000	11,000	40,000
リード長 (raw) (bp)	101	101	101	101	101	101	101
合計サイズ (raw) (bp)	60.4G	71.7G	14.5G	16.0G	14.9G	11.6G	3.9G
リード長 (preprocessed) (bp)	80	62	62	58	52	58	50
合計サイズ (preprocessed) (bp)	48.0G	27.7G	3.8G	3.8G	5.1G	2.0G	1.0G

"preprocessed"はアダプタ配列、低クオリティ領域の除去後を指す。mate-pair に対しては PCR-duplicate とインサートサイズが極端に小さいペアも除去される。

・ベンチマーク対象の *de novo* アセンブラ

Platanus（version 1.2.1）に加えて、ALLPATHS-LG（Gnerre et al. 2011）（version 44837）、MaSuRCA（Zimin et al. 2013）（version 2.0.4）、Velvet（Zerbino and Birney 2008）（version 1.2.07）、SOAPdenovo2（Luo et al. 2012）（version 2.04）をベンチマークした。様々な *de novo* アセンブラを比較した GAGE の論文（Salzberg et al. 2012）では、ヒト 14 番染色体データにおけるテストにおいて、これらのツールは scaffold NG50 で上位 1–4 位の順位を達成している。ALLPATHS-LG、Velvet、SOAPdenovo2 のアルゴリズムは de Bruijn グラフに基づく。Velvet は比較的ゲノムサイズの小さな生物向けに設計されているが、ALLPATHS-LG と SOAPdenovo2 はギガ単位のゲノムサイズにも対応している。MaSuRCA は overlap-layout-consensus アルゴリズムを用いる Celera assembler（Myers et al. 2000）をベースに、ハイスループットデータへ対応するための改良を加えたツールである。overlap-layout-consensus は一般に多くの計算時間を必要とするが、リピート配列や coverage depth の低い領域に有効である可能性も有する。

パラメータ設定については、Platanus はデフォルト設定で実行された。Velvet と SOAPdenovo2 については、重要なパラメータである *k*-mer 長を最適化することを試みている。ここでは、21 から 91 の *k*-mer 長を 10 刻みで与えてそれぞれ実行し、scaffold NG50 が最大となるものを選んでいる。SOAPdenovo2 については、ヘテロ領域の解決に重要と思われる"mergeLevel"（1–3 の範囲で、大きいほど多くの相同な配列の組をマージする）を 1 または 3 に設定し、同様に scaffold NG50 が大きい方を選択しており、さらに付属の GapCloser プログラムも実行し、ギャップ数を減らすようにした。ALLPATHS-LG は２倍体モードで実行し、その他はデフォルトの設定で実行した。MaSuRCA はメモリ消費量に関する設定以外はデフォルトで実行した。

・ベンチマークデータの前処理

全てのサンプルに共通するデータの前処理方法について述べる。まず、全てのライブラリのリードに対して、アダプタ配列と低クオリティ領域の除去を行なった。アダプタ配列は、11 bp の完全一致をシードとして、30 bp 以上のアライメントが検出された場合、その箇所より外側を除去する。ペアリードの場合は、リードのオーバーラップ情報を用いて 30 bp より短いアダプタ配列の除去も試みられる。低クオリティ領域に対しては、リードの両端から順に

quality value ≥ 15 の塩基が 11 bp 以上連続している領域を検出し、それより外側の領域を除去する。mate-pair ライブラリについては、PCR-duplicate とインサートサイズが極端に小さいペアの除去を行なった。これらを検出するためにはリードをリファレンス配列にマップする必要があるが、*de novo* アセンブリの場合はそれが存在しないため、paired-end のみから Platanus で構築した scaffold 配列（gap close 済み）にリードをマップした。マッピングツールは Bowtie2（Langmead and Salzberg 2012）である。両リードの 5'端のマップ位置が一致するペアを集め、その中から最も mapping quality（Bowtie2 が算出）が高いペアのみを残すことで PCR-duplicate を除去する。また、リード間の距離が nominal インサートサイズ（ライブラリ調整時に定められたインサートサイズ）の半分以下のペアも除いた。

2.3.2　ヘテロ接合度をシミュレートしたデータによるベンチマークと考察

・*C. elegans* 実データの取得およびヘテロ接合度をシミュレートしたデータ作成

C. elegans（ゲノムサイズ 100 Mbp：国立遺伝学研究所 小原雄治教授より提供）のゲノム DNA を Illumina HiSeq 2000 によりシークエンスして頂いた（表 2-2：国立遺伝学研究所 豊田敦准教授によるシークエンス）。ライブラリは 2 paired-ends と 1 mate-pair（インサートサイズ　230 bp、420 bp、4,660 bp）で構成される。リードの前処理後、リファレンスゲノムに paired-end をマップし SNV と small indel 変異を以下の手順で検出した。

(1)　リードのマッピング

Bowtie2 の single-end モードでリードをそれぞれマップする。許容する編集距離（ミスマッチ + ギャップ）は 5 以下。

(2)　マッピング結果のフィルタリング

同一の配列に両リードがマップされるペアのうち、インサートサイズが、全体の平均±2×全体の標準偏差 の範囲内におさまるもののみを残す。PCR-duplicate も除く（SAMtools rmdup コマンド）。

(3)　パイルアップ

SAMtools（Li et al. 2009）の mpileup コマンドで行う。塩基の quality value の下限は 30 とする（-Q 30）。

(4)　変異のコール

coverage depth が[20, 全体平均×2]の範囲内にあるサイトで、変異（SNV か small indel）を示すリードの割合が[0.25, 0.75]の範囲にあるとき、変異をコールする。ここで、変異は全てヘテロであると仮定しているので 0.75 という上限を設けた。また、リファレンス上のギャップまたは配列端から 100 bp 以内のサイトは対象外とする。

検出した変異の数からサンプルのヘテロ接合性を推定すると、1.85×10^{-3} %と極めて低い値になった。このように元のヘテロ接合性が低いことを確認した上で、計算機上で人為的に 0.1–2.0%のヘテロ接合度になるような変異を加えることでシミュレーションデータを構築した。作成手順は以下の通りである。

(1) リファレンスと対になるハプロタイプ配列の構築
リファレンス配列上にランダムに SNV または indel を導入する。SNV と indel の数の比は 9:1。変異密度が 0.1–2.0%となる配列をそれぞれ構築する。

(2) リードのマッピング
全てのライブラリのリードをリファレンスにマップし位置を決定する。

(3) リードへの変異の導入
各リードのハプロタイプ（リファレンス配列または構築したハプロタイプ配列）を決定し、リファレンスと異なるタイプのリードは配列を変換する。その際、各サイトで比が 1:1 になるようにする。また、この比は正規分布に従ってばらつく。この方法では変異間の連鎖も正しく再現される。

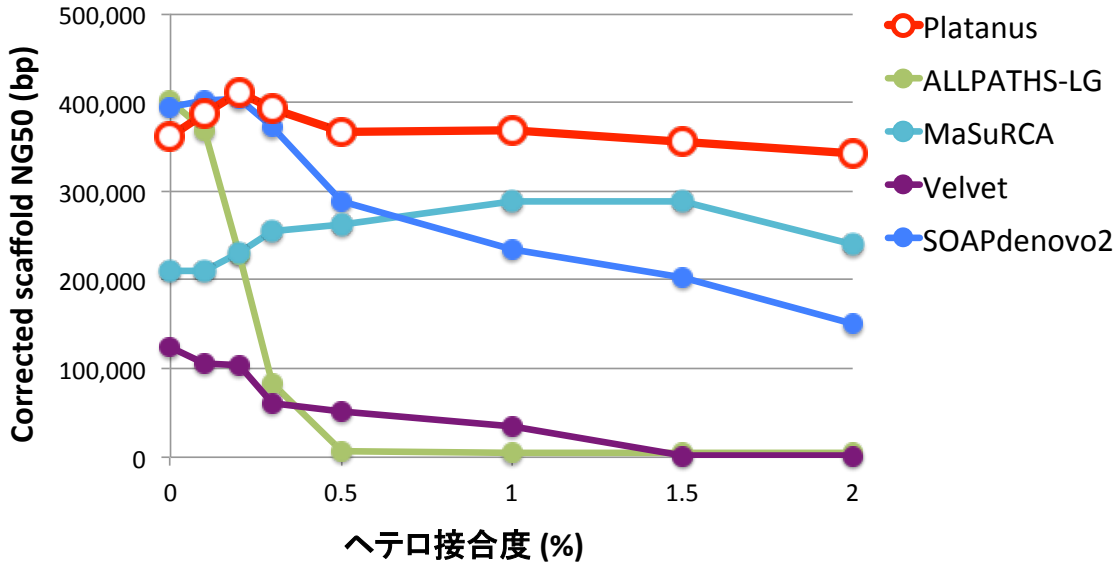
シミュレートしたヘテロ接合度の値は 0.1、0.2、0.3、0.5、1.0、1.5、2.0%の 7 通りであり、それに対応した 7 つのデータセットが作られた。

・ベンチマーク結果および考察

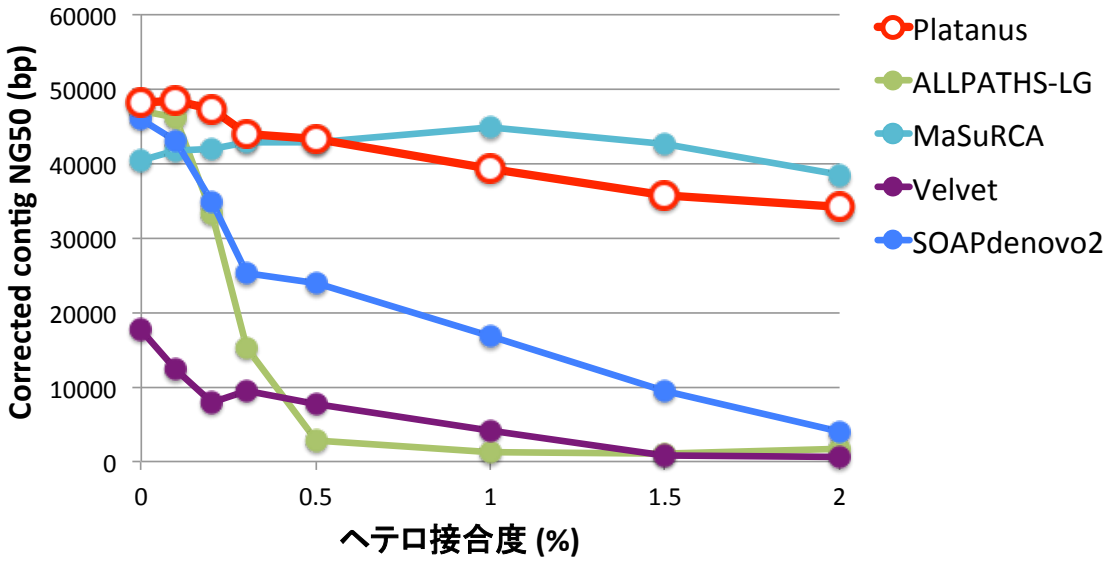
2.3 節で述べた方法で各 *de novo* アセンブラを実行し、評価ツール GAGE を用いて評価を行なった。GAGE ではリファレンスゲノム配列にアセンブリ結果配列（scaffold 配列）をアライメントし、ミスアセンブリ数（inversion、translocation、relocation の合計数）等が算出される。また、scaffold、contig をミスアセンブリ箇所で分断して NG50 を再計算し、corrected NG50 という値として報告される。この指標は長さと精度の両方を考慮した値として用いることができる。

まず、精度の情報を含む指標（corrected NG50、ミスアセンブリ数）について、各アセンブラの値がヘテロ接合性によってどのように変化するかを図 2-22 に示す。

(A) Corrected scaffold NG50



(B) Corrected contig NG50



(C) ミスアセンブリ数

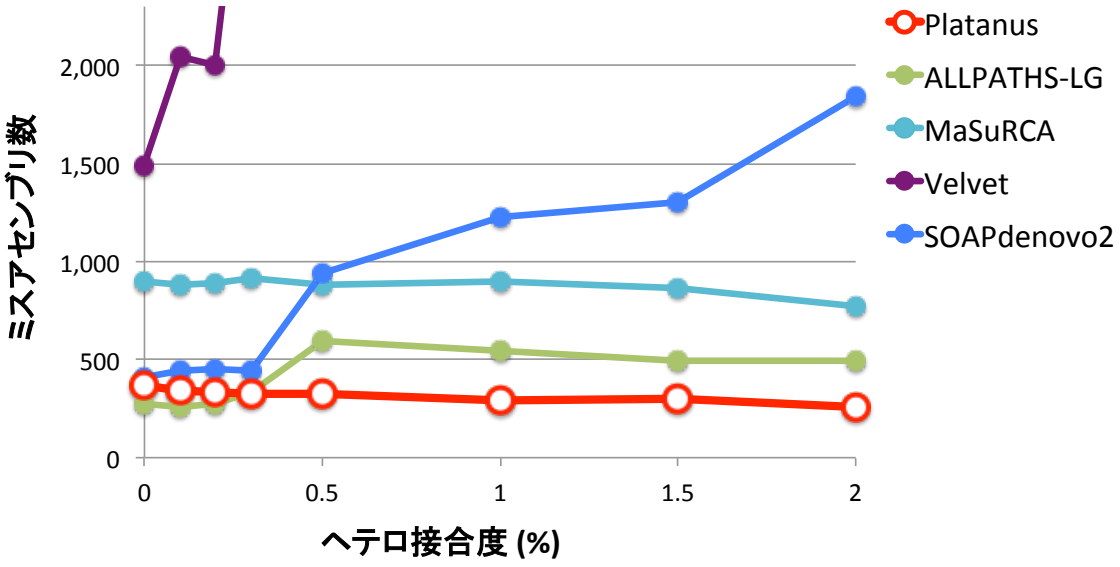


図 2-22　ヘテロ接合性シミュレーションデータでのベンチマーク結果（精度情報を含む指標）

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度の増加に従い corrected NG50 が急激に減少し、ミスアセンブリ数も急激に増加している。ヘテロ接合性がアセンブリの障害となることはゲノム解読計画で個別に報告されていたが（Sodergren et al. 2006; Xu et al. 2011; Zhang et al. 2012）本研究のように複数のヘテロ接合度の値毎にその影響を調べたものは新規である。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。特にヘテロ接合度が高い（≥ 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、塩基レベルでの精度をリファレンスと scaffold 間のミスマッチ数と indel（< 5bp）数で評価した（表 2-3）。ここでは、ヘテロ接合性をシミュレートしていないデータを対象としている。

表 2-3　*C. elegans*（ヘテロ接合度 0.0%）データのアセンブリにおける塩基レベルの精度評価

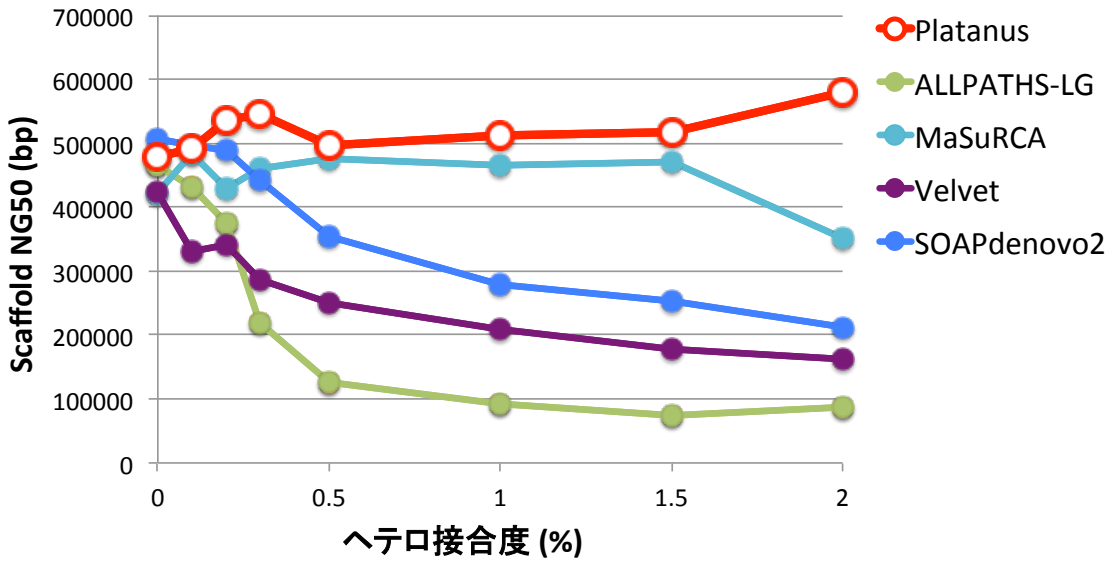
	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
ミスマッチ数	4,534	5,762	15,521	16,650	16,941
Indel数 (< 5 bp)	3,352	5,125	9,142	5,236	5,102
'N'の割合 (%)	1.40	2.63	0.77	0.43	3.33

リファレンス配列と scaffold 配列間の値（GAGE が算出）。

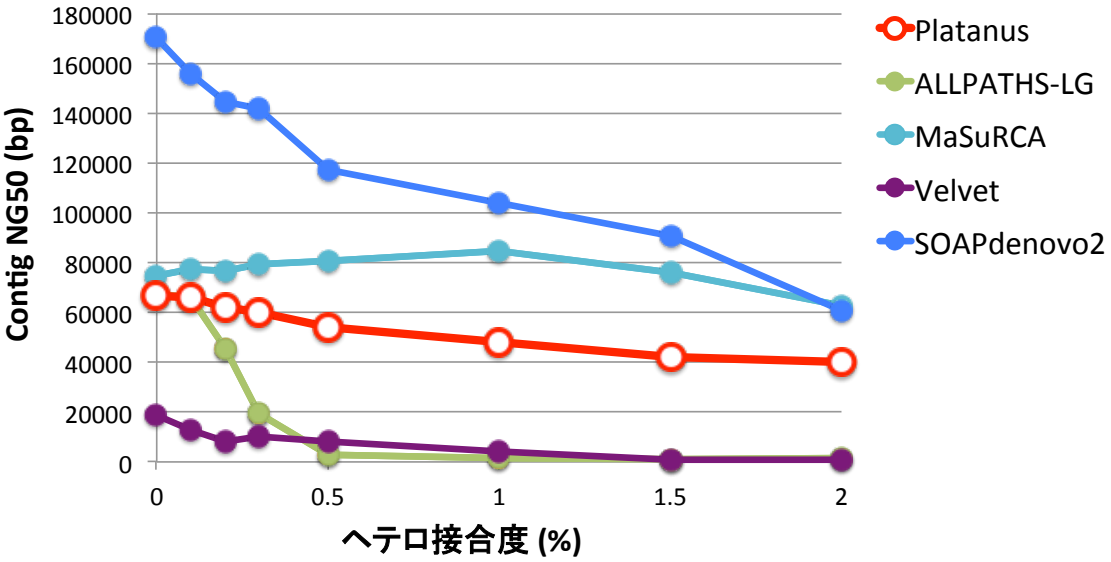
C. elegans サンプルの元のヘテロ接合度は 1.85×10^{-3} %と推定されたため、ミスマッチ数と indel 数の合計が 1850 個を上回ったとき、少なくともその分はエラーであると考えられる。Platanus のミスマッチと indel の合計は最小であり、塩基レベルでの精度も高いことが示唆される。ミスマッチ、indel 数は、アセンブリ結果に'N'を多く含めることで減らせる可能性があるが、Platnus の'N'の割合は 3 番目であり、中間の順位である。つまり、Platanus は'N'を多く報告することでこれらの数を下げているということはなく、実用にも適した精度を有していると考えられる。

精度に関する情報を含まない他の指標（NG50、配列合計長）を図 2-23 に示す。scaffold NG50 については、corrected scaffold NG50 と比較して大きな傾向の変化はない。contig NG50 は、SOAPdenovo2 の値がヘテロ接合度 2.0%の場合以外で最大となっている。しかし、corrected contig NG50 については SOAPdenovo2 では多くのケースで Platanus や MaSuRCA より小さい値になっているため、ミスアセンブリで分断される配列が多いことが示唆される。配列の合計長は、ALLPATHS-LG 以外は全てのケースで *C. elegans* のゲノムサイズ（100 Mbp）に近い値となっている。特に Platanus はヘテロ接合度 2.0%でもゲノムサイズとの差が小さく、相同領域のマージが機能していることを示している。

(A) Scaffold NG50



(B) Contig NG50



(C) Scaffold 合計長

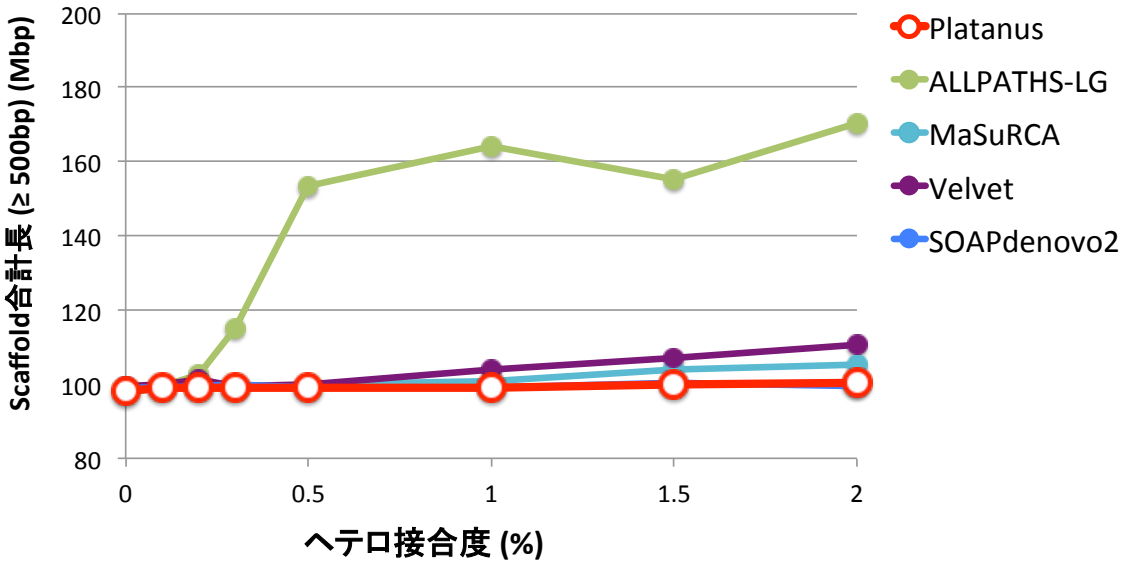


図 2-23　ヘテロ接合性シミュレーションデータでのベンチマーク結果（精度情報を含まない指標）

ALLPATHS-LG のアセンブル結果の合計長はヘテロ接合度 ≥0.5%のとき 150 Mbp 以上となり、ゲノムサイズの 2 倍（200 Mbp）に近い値となっている。これについては、ALLPATHS-LG がハプロタイプの両方の配列を構築している可能性が存在するため詳細を調べた。まず、ヘテロ接合度 2.0%のデータについて、リファレンスゲノム配列と、シミュレーション時に作られた仮想的なハプロタイプ配列を合わせて、2 倍体のリファレンス配列とする。次に各アセンブラの scaffold を 200 bp に区切り、それぞれを 2 倍体リファレンス配列に Bowtie2 でマッピングする。片方のハプロタイプ配列にユニークにベストヒットする場合、その 200 bp 領域のハプロタイプを決定する。変異がない領域やギャップなど、ユニークなベストヒットが存在しない場合はハプロタイプ不明とする。ある scaffold 内で、片方のハプロタイプからもう片方へ切り替わる点が存在するとき、その点を"haplotype junction"とし各アセンブラでの数を調べた（表 2-4）。なお、あるハプロタイプから不明領域への切り替わりはカウントしない。

表 2-4　ハプロタイプの切り替え（haplotype junction）に関する統計

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
Haplotype junction数	143,359	92,940	170,172	184,989	149,030
Haplotype junctionを含むscaffoldの割合 (%)	17.88	85.84	11.66	43.67	32.96
Haplotype junction間の平均距離 (bp)	1398.9	2157.7	1178.5	1084.1	1345.7

その結果、ALLPATHS-LG にも多くの haplotype junction が含まれることが分かった。Platanus を含めどのアセンブラも相同領域をマージする設計になっているため、haplotype junction が存在することは予想外ではない。これらのアセンブラの結果を用いて解析を行う際には、たとえ合計長がゲノムサイズより大きくなっていても、ハプロタイプが正確に構築できている可能性は低いことに注意する必要がある。ある 10 kbp の領域における各アセンブラの haplotype junction の例を図 2-24 に示す。

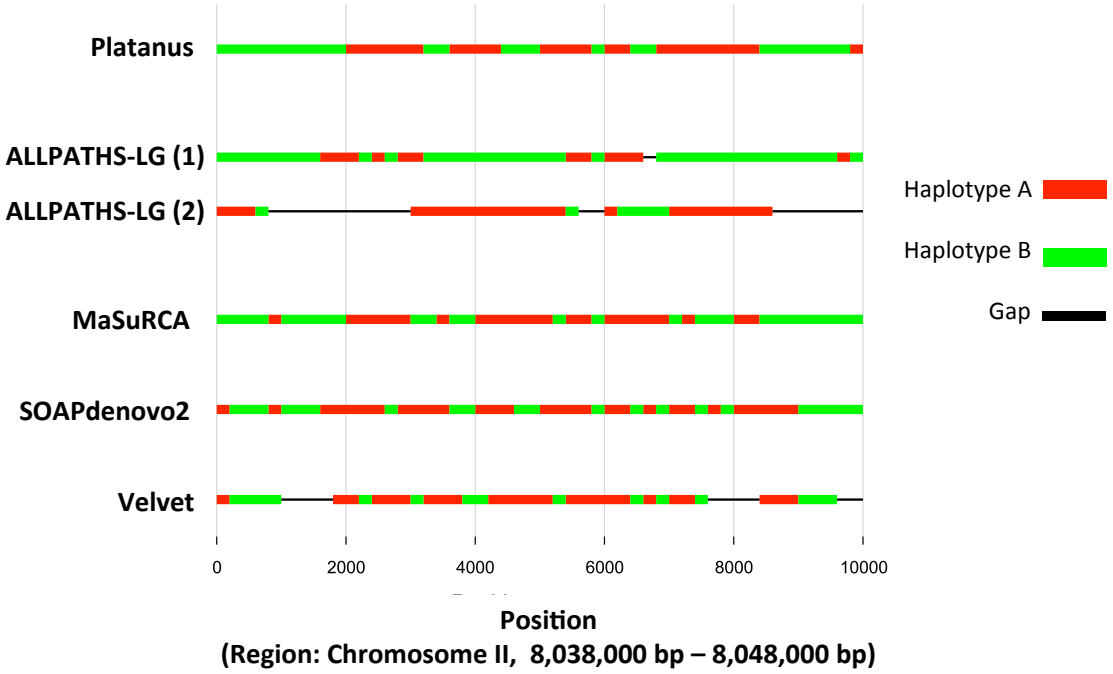


図 2-24　scaffold 内でのハプロタイプ切り替え（haplotype junction）の例
C. elegans のヘテロ接合度 2.0%のケース。Haplotype A はリファレンス配列、Haplotype B はシミュレーション時の仮想ハプロタイプ。

・ベンチマーク結果の詳細データ

ベンチマーク結果についての詳細な数値を表 2-5 にまとめる。

表 2-5　*C. elegans* データを用いたベンチマーク結果の詳細

(A) Corrected scaffold NG50

ヘテロ接合度 (%)	Corrected scaffold NG50 (bp)				
	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
0.0	361,608	402,062	210,683	394,556	123,438
0.1	388,028	368,779	209,917	403,115	104,756
0.2	411,618	228,033	229,808	403,701	104,352
0.3	392,137	82,457	255,210	372,430	60,253
0.5	366,498	5,994	263,146	288,181	51,097
1.0	368,857	4,928	288,752	234,395	34,491
1.5	355,057	4,999	287,646	202,140	2,099
2.0	341,914	5,178	240,778	150,059	1,932

(B) Scaffold NG50

ヘテロ接合度 (%)	Scaffold NG50 (bp)				
	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
0.0	478,744	466,658	420,694	507,513	424,862
0.1	490,975	431,770	482,920	497,363	332,019
0.2	535,328	375,904	430,011	489,092	340,229
0.3	545,914	219,404	460,620	441,950	286,218
0.5	497,387	127,365	475,513	353,955	251,000
1.0	511,190	91,413	466,806	280,050	209,807
1.5	516,958	73,543	472,079	252,105	178,132
2.0	580,832	86,979	351,406	212,590	162,062

(C) Corrected contig NG50

ヘテロ接合度 (%)	Corrected contig NG50 (bp)				
	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
0.0	48,054	46,361	40,248	45,927	17,739
0.1	48,250	45,792	41,645	43,026	12,263
0.2	47,179	33,146	41,819	34,822	7,831
0.3	43,969	15,120	42,900	25,294	9,474
0.5	43,196	2,746	42,700	23,831	7,652
1.0	39,291	1,316	44,746	16,784	4,231
1.5	35,786	1,057	42,593	9,419	707
2.0	34,030	1,648	38,465	4,013	610

(D) Contig NG50

ヘテロ接合度 (%)	Contig NG50 (bp)				
	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
0.0	66,446	67,596	420,694	74,686	19,084
0.1	65,717	66,289	482,920	77,050	12,988
0.2	61,969	45,370	430,011	76,439	8,135
0.3	59,795	19,629	460,620	79,194	10,174
0.5	53,873	2,807	475,513	80,813	8,142
1.0	48,090	1,358	466,806	84,827	4,445
1.5	42,013	1,103	472,079	75,918	766
2.0	39,915	1,810	351,406	62,656	666

(E) Scaffold 合計長（≥ 500 bp）

ヘテロ接合度 (%)	Scaffold合計長 (≥ 500 bp) (bp)				
	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
0.0	97,938,690	98,864,571	97,742,565	99,132,377	99,382,943
0.1	98,876,811	99,521,025	99,171,512	99,215,054	99,851,591
0.2	98,876,741	102,601,583	98,996,688	99,504,741	101,204,194
0.3	98,920,320	114,954,930	99,210,505	99,969,847	99,482,862
0.5	98,916,151	153,369,255	99,925,896	99,196,053	100,074,381
1.0	99,223,085	163,953,538	100,913,677	99,086,653	103,773,061
1.5	99,706,135	155,053,323	103,830,661	100,438,152	107,239,890
2.0	100,454,422	170,519,463	105,340,084	99,546,387	110,798,448

(F) ミスアセンブリ数（#inversion + #translocation + #relocation）

ヘテロ接合度 (%)	ミスアセンブリ数				
	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
0.0	367	272	900	407	1,490
0.1	346	259	878	441	2,044
0.2	333	276	892	450	1,999
0.3	324	335	911	442	3,575
0.5	325	593	881	941	4,206
1.0	290	542	895	1,226	6,285
1.5	300	498	861	1,304	35,440
2.0	256	496	774	1,839	36,840

2.3.3　高ヘテロ接合性線虫によるベンチマークと考察

・*S. venezuelensis* データの取得および特徴

高ヘテロ接合性線虫（*S. venezuelensis*）の DNA を宮崎大学 丸山治彦教授より提供頂き、国立遺伝学研究所 豊田敦准教授により HiSeq 2000 にてシークエンスして頂いた。シークエンスされたライブラリは 2 paired-ends と 1 mate-pair（インサートサイズ　200 bp、450 bp、3,400 bp）で構成される。

17-mer の出現回数分布の解析より（図 2-21、表 2-1）、線虫 *S. venezuelensis* のゲノムサイズは 57.7 Mbp と推定される。リピート由来 17-mer の割合は *S. venezuelensis*: 0.289、*C. elegans*: 0.239、となり、リピート配列の割合は *C. elegans* と比較的近い値と考えられる。よって、ヘテロ接合性の影響について、シミュレーション（*C. elegans*）と実データで傾向を比較することができる。

全ゲノムリードとは別に、ゲノムの一部分に対応する 8 本の完成 fosmid 配列（合計長　272,981 bp）をヘテロ接合度の推定やアセンブリの精度評価に用いた *S. venezuelensis* に関する fosmid 配列は、宮崎大学 丸山治彦教授に提供して頂いた DNA よりライブラリ調整したものであり、シークエンスは東京大学 白髭克彦教授により実施されたものである。ヘテロ接合度は、fosmid に paired-end リードをマップし *C. elegans* と同じ手法（2.3.2）で 0.93 と推定した。

・ベンチマーク結果および考察

NG50 は推定ゲノムサイズ 57.7 Mbp を用いて求めた。精度は、リファレンスゲノム配列が存在しないため、fosmid 配列と scaffold 配列を比較して評価した。具体的な手順は次の通りで、アライメントツールの MUMmer パッケージ（Kurtz et al. 2004）中のプログラム nucmer、delta-filter を用いる。

(1) fosmid 配列を nucmer で scaffold 配列にアライン
(2) アライメント結果の delta-filter -g（"one-to-one global alignment, not allowing for rearrangements"）でフィルタリング
(3) fosmid 配列毎に最も長いアライメントを選び、トップヒットとする。
(4) トップヒットのアライメント長の合計を Top-hits-length とし、相同性（identity）の平均も求める。

Top-hits-length は、fosmid 配列と対応する scaffold が短かったり、ミスアセンブリやギャップを含んでいたりすると小さくなるため、アセンブリ結果の長さと精度の両方を反映した指標となる。トップヒットのアライメント長が fosmid 配列長の 90%以上の場合は、"contained"としてカウントした。identity は変異の存在によっても低下するが、アセンブリのエラーがない場合は 99.53%（100－0.93／2）程度になると予想される。結果を表 2-6 に示す。

表 2-6　*S. venezuelensis* アセンブリ結果

		Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
アセンブリ結果	合計長 (≥500 bp) (bp)	58,503,663	61,205,926	66,053,722	52,677,856	63,982,183
	#Scaffolds	2,560	9,608	4,876	3,383	11,696
	Scaffold NG50 (bp)	274,622	16,765	176,206	87,219	17,006
	Contig NG50 (bp)	71,357	2,008	84,739	48,010	1,946
Fosmidによる評価	Top-hits-lengths (bp)	272,164	69,792	256,848	270,392	78,159
	Average Identity (%)	99.42	99.31	99.39	98.72	99.31
	#Contained fosmids	8	0	7	8	0

500 bp 以上の配列が対象。

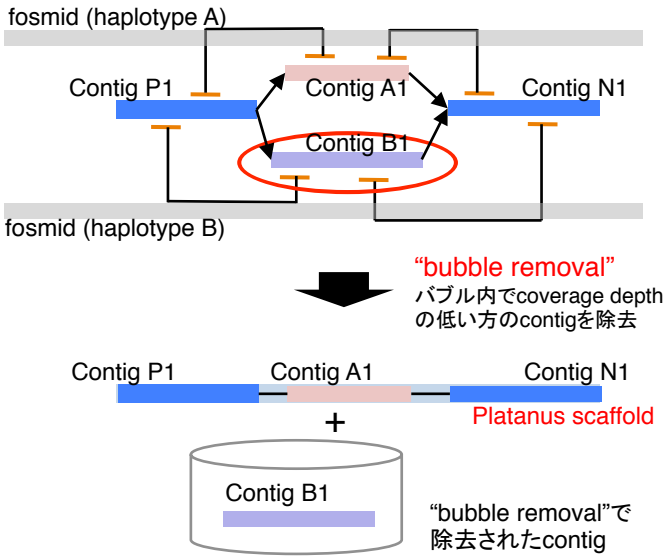
Platanus のアセンブリ結果については、scaffold NG50、fosmid 配列に対する Top-hits-length、identity はともに最大となっている。"Contained fosmids"の数が 8 であることは、8 本の fosmid 配列がいずれも全長が構築されたことを意味し、大きなミスアセンブリも検出されなかったことを意味する。scaffold NG50 について、各アセンブラと Platanus の値の比を表 2-7 に示す。ここでは *C. elegans*（2.3.2）と *S. venezuelensis* の両方について算出した。ヘテロ接合度 1.0%の *C. elegans* と比較すると、*S. venezuelensis* の方が全アセンブラで値が低くなっている。これは、実データではシミュレーションより Platanus と他のアセンブラの差が広がったことを意味する。シミュレーションでは NG50 へのヘテロ接合度の影響が小さかった MaSuRCA についても値の低下は見られる。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。

表 2-7　scaffold-NG50 / Platanus-scaffold-NG50

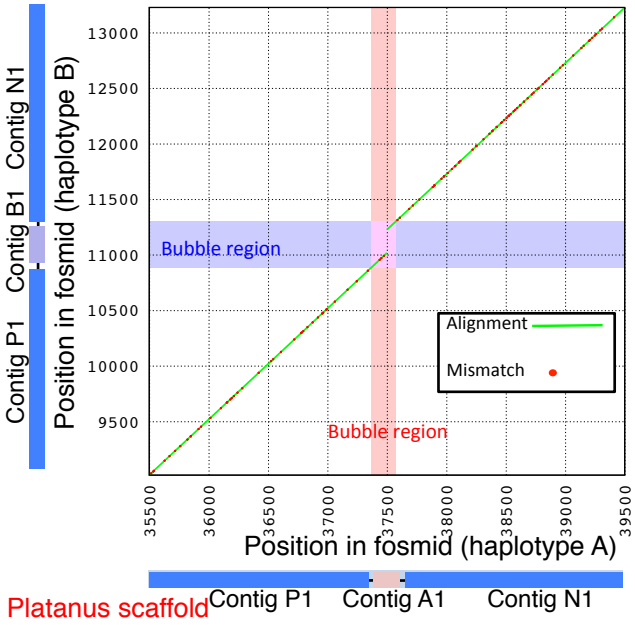
	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
C. elegans (ヘテロ接合度 0.0%)	0.975	0.879	1.060	0.887
C. elegans (ヘテロ接合度 0.1%)	0.879	0.984	1.013	0.676
C. elegans (ヘテロ接合度 0.2%)	0.702	0.803	0.914	0.636
C. elegans (ヘテロ接合度 0.3%)	0.402	0.844	0.810	0.524
C. elegans (ヘテロ接合度 0.5%)	0.256	0.956	0.712	0.505
C. elegans (ヘテロ接合度 1.0%)	0.179	0.913	0.548	0.410
C. elegans (ヘテロ接合度 1.5%)	0.142	0.913	0.488	0.345
C. elegans (ヘテロ接合度 2.0%)	0.150	0.605	0.366	0.279
S. venezuelensis	0.061	0.642	0.318	0.062

シミュレーションと実データで異なる点としては、構造変異の有無などが挙げられる。シミュレーションではSNVとsmall indelのみを導入している。Platanusは構造変異等への対策をScaffoldingステップで備えており、それらが他のアセンブラより大きなNG50に寄与している可能性が存在する。Platanusが Scaffolding時のバブル構造除去（bubble removal）、枝構造除去（branch cut）アルゴリズムで解決した領域について、それぞれ1箇所ずつfosmid配列の組を用いて検証した。各領域でfosmid配列の組は両方のハプロタイプに1本ずつ対応している。模式図とfosmid組のアライメント結果を図 2-25に示す。bubble removalを行なった領域では209 bpのindelが存在し、アライメントが得られた領域のidentityは97.9%であった（図 2-25 AB）。branch cutを行なった領域では更に規模の大きい126 bp、715 bp、1206 bpの3つのindelが存在し、identityは98.13%であった（図 2-25 CD）。両方のfosmid組について、ハプロタイプの片方でも全長を構築できていたのはPlatanusのみであった。注目すべき点としては、これらの領域では構造変異が存在するのみではなく、SNVとsmall indelの密度も高いということが挙げられる。今回の例では、タイプの異なる変異が実際には組み合わさって存在しアセンブリをより困難にしていること、Platanusがそのような領域を解決することで長いscaffoldを構築していることが示唆された。

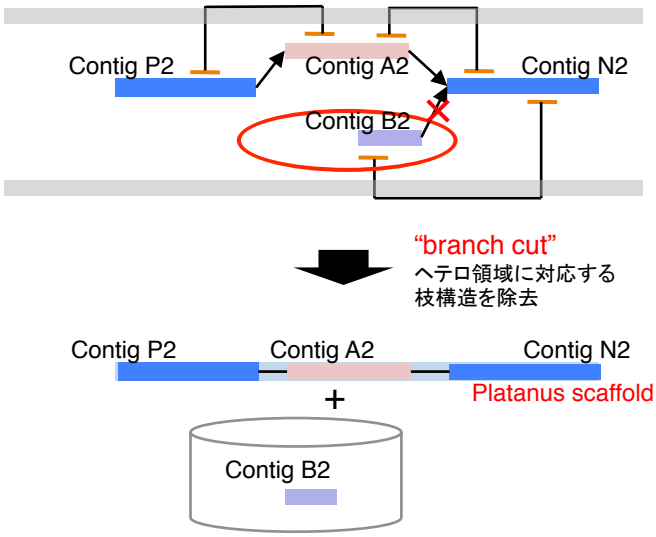
(A)


fosmid (haplotype A)
Contig P1
Contig A1
Contig N1
Contig B1
fosmid (haplotype B)
"bubble removal"
バブル内でcoverage depth
の低い方のcontigを除去
Contig P1
Contig A1
Contig N1
Platanus scaffold
+
Contig B1
"bubble removal"で
除去されたcontig

(B)


Contig N1
Contig B1
Contig P1
Position in fosmid (haplotype B)
13000
12500
12000
11500
11000
10500
10000
9500
Bubble region
Alignment
Mismatch
Bubble region
35500
36000
36500
37000
37500
38000
38500
39000
39500
Position in fosmid (haplotype A)
Platanus scaffold
Contig P1
Contig A1
Contig N1

(C)



(D)

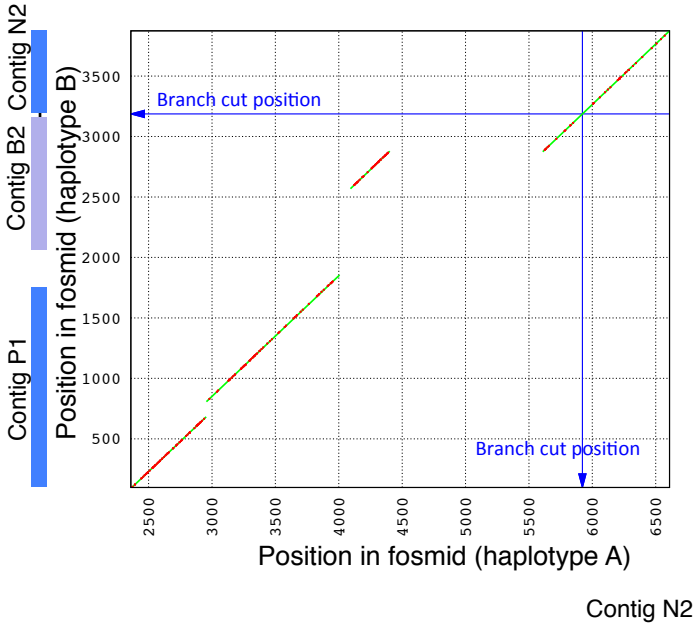
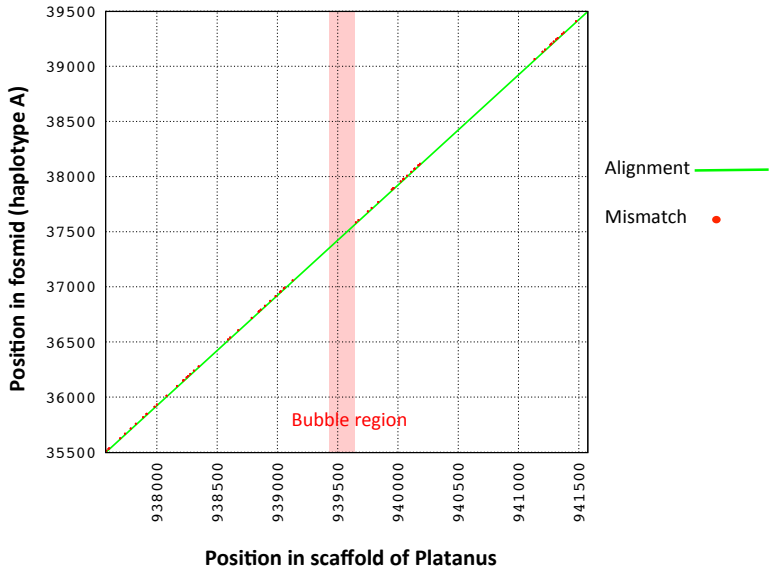


図 2-25　Scaffolding で解決された構造変異の例

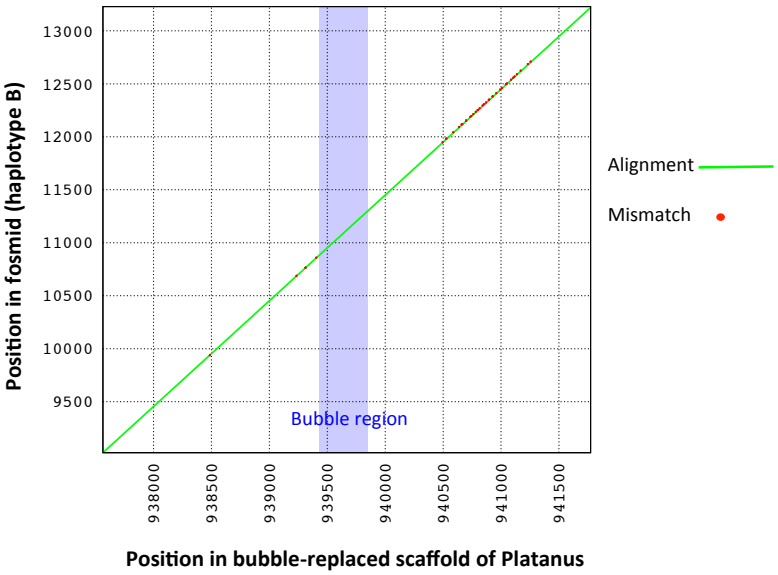
(A) バブル構造除去（bubble removal）が行われた領域の解析の模式図。(B) A に対応する fosmid 間のドットプロット図。(C) 枝構造除去（branch cut）が行われた領域の解析の模式図。(D) C に対応する fosmid 間のドットプロット図。

Platanus の scaffold と fosmid 配列をアラインし、ミスアセンブリの有無についても調べることとした（図 2-26）。バブル領域に関しては、scaffold に残った配列（haplotype A）と除去された配列（haplotype B）がそれぞれ片方のハプロタイプの fosmid にアラインされるかどうかを確かめている。結果、ミスアセンブリは検出されず、Platanus が構造変異を正しく解決できていることが分かった。

(A)



(B)



(C)

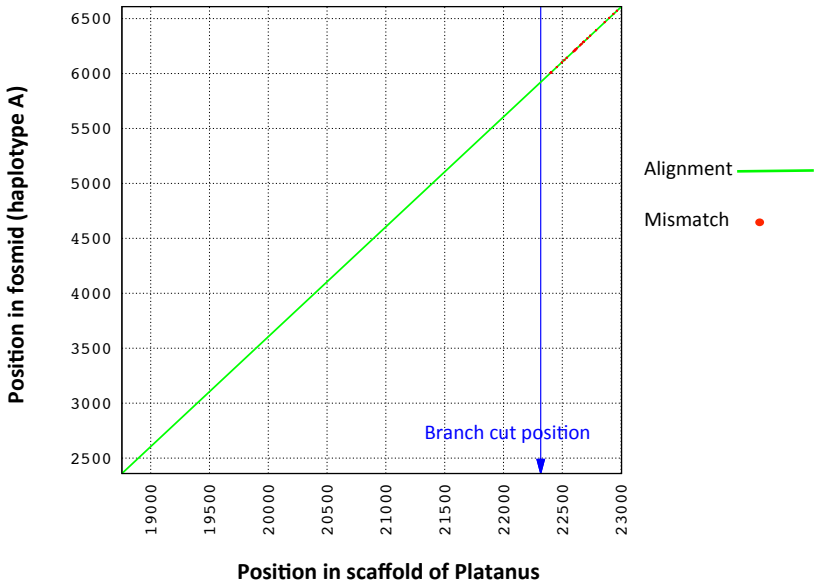


図 2-26 Platanus の scaffold と fosmid 配列のアライメント

(A)(B) scaffolding 中に bubble removal が行われた領域（図 2-25AB に対応）。両方のハプロタイプについてそれぞれアライメントを示す。 (C) scaffolding 中に branch cut が行われた領域（図 2-25CD に対応）。

ここまでの fosmid 配列を用いた解析は、あくまでゲノムの一部分のみを対象としている。続いて、ゲノム全体で実データとシミュレーションデータのヘテロ接合度の分布の比較を試みた。*S. venezuelensis* のリファレンスゲノム配列が存在しないため、アセンブリ結果の中で NG50 が最大となる Platanus の scaffold 配列にリードをマップし、変異を検出した（方法 : 2.3.2）。全体の変異密度は 0.95%で、fosmid 配列上で推定したヘテロ接合度 0.93%と近い値となっている。1 kbp のウィンドウを用いて求めた分布を図 2-27 に示す。

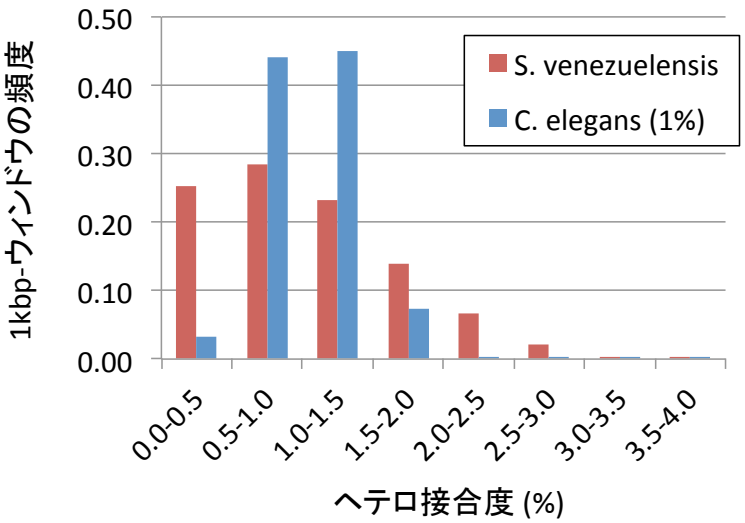


図 2-27　実データ（*S. venezuelensis*）とシミュレーションデータ（*C. elegans*）のヘテロ接合度の分布

Platanus の scaffold 配列に paired-end をマップして変異を検出後、1 kbp-ウィンドウを用いて算出した。

ゲノム全域に均一に変異を導入したシミュレーションデータと比較して、実データの方はばらつきが大きくなっている。このような分布の違いも、実データで Platanus と他のアセンブラの NG50 の差が広がっている理由の 1 つである可能性がある。しかしながら、実データでは低ヘテロ接合性の領域も多く存在することも意味し、ヘテロ接合度のばらつきが大きいとき必ずしも NG50 が低下するとは限らない。そこで、高ヘテロ接合度（>1%）を示す 1 kbp-ウィンドウの間隔の平均を調べると、1,930 bp という値になった。これは約 2 kbp おきに高ヘテ

ロ接合性領域が現れることを意味し、ヘテロ接合性に関してモザイク状の構造をとっていることが分かる。他アセンブラの scaffold は、短い間隔で現れる高ヘテロ接合性領域で分断されることが多いため、NG50 の低下が引き起こされるという可能性が示唆される。

Platanus に特徴的な機能が実際に NG50 等に寄与しているかを知るため、それぞれの機能を無効化したときのアセンブリ結果を表 2-8 に示す。比較のため、*C. elegans*（ヘテロ接合度 0.0%、1.0%、2.0%）での結果も含めた。ここで対象とした機能は、Contig-assembly の *k*-mer 伸長、Contig-assembly のバブル構造の除去、Scaffolding のバブルおよび枝構造の除去である。scaffold と contig の NG50 に注目すると、*k*-mer の伸長の無効化の際にはヘテロ接合性に関わらず全てのケースで値が低下している。Contig-assembly のバブル除去を無効化すると、ヘテロ接合度 0.0%の *C. elegans* 以外のケースで NG50 が低下しており、変異の解決にこの機能が寄与していることが分かる。Scaffolding のヘテロ領域対策（バブル、枝構造除去）についても傾向は同じで、高ヘテロ接合性のサンプルに対して効果を発揮しているが、特に *S. venezuelensis* で無効化時の NG50 の減少度合いが大きくなっている。この結果は、これらの機能が実データに特有の構造変異の解決に寄与しているという仮説と一致する。

表 2-8　Platanus の特徴な機能を無効化した際のアセンブリ結果

(A) *C. elegans*（ヘテロ接合度　0.0%）

	Default	− (*k*-mer-extension)	− (bubble-removal in contig-assembly)	− (hetero-removal in scaffolding)
合計長 (≥500 bp)	97,938,690	96,528,882	98,022,430	98,715,700
Scaffold数 (≥500 bp)	942	1,953	852	1,214
Scaffold NG50 (bp)	478,744	306,733	481,025	415,878
Corrected scaffold NG50 (bp)	361,608	226,959	375,389	334,052
Contig NG50 (bp)	66,446	35,712	67,574	66,447
Corrected contig NG50 (bp)	48,054	30,899	49,046	48,800
ミスアセンブリ数	367	257	342	274

(B) *C. elegans*（ヘテロ接合度　1.0%）

	Default	− (*k*-mer-extension)	− (bubble-removal in contig-assembly)	− (hetero-removal in scaffolding)
合計長 (≥500 bp)	99,223,085	96,211,233	140,267,578	100,641,997
Scaffold数 (≥500 bp)	1,289	2,193	24,301	4,579
Scaffold NG50 (bp)	511,190	291,355	45,673	381,587
Corrected scaffold NG50 (bp)	368,857	204,518	20,728	195,751
Contig NG50 (bp)	48,090	8,887	6,268	14,472
Corrected contig NG50 (bp)	39,291	8,547	5,970	13,665
ミスアセンブリ数	290	160	203	228

(C) *C. elegans*（ヘテロ接合度　2.0%）

	Default	− (*k*-mer-extension)	− (bubble-removal in contig-assembly)	− (hetero-removal in scaffolding)
合計長 (≥500 bp)	100,454,422	96,157,873	183,642,064	108,879,145
Scaffold数 (≥500 bp)	2,264	2,434	10,902	17,986
Scaffold NG50 (bp)	580,832	270,290	81,927	207,088
Corrected scaffold NG50 (bp)	341,914	204,425	26,451	100,787
Contig NG50 (bp)	39,915	6,704	22,094	7,545
Corrected contig NG50 (bp)	34,030	6,517	16,994	7,316
ミスアセンブリ数	256	129	146	170

(D) *S. venezuelensis*

	Default	− (*k*-mer-extension)	− (bubble-removal in contig-assembly)	− (hetero-removal in scaffolding)
合計長 (≥500 bp)	58,503,689	45,493,271	74,553,871	60,793,871
Scaffold数 (≥500 bp)	2,560	1,660	9,304	7,026
Scaffold NG50 (bp)	274,622	127,238	52,920	108,844
Contig NG50 (bp)	71,357	15,708	7,754	15,983

"−(*k*-mer-extension)" は Contig-assembly の *k*-mer 伸長、"−(bubble-removal in contig-assembly)"は Contig-assembly のバブル構造の除去、"−(hetero-removal in scaffolding)"は Scaffolding のバブルおよび枝構造の除去を無効化した際の結果を示す。

2.3.4　高ヘテロ接合性かつ高リピート率のサンプルによるベンチマークと考察

・*C. gigas* データの取得および特徴

C. gigas（牡蠣、oyster）のドラフトゲノムは Oyster Genome Project（Zhang et al. 2012）により発表されている。本節で用いたシークエンスリード、BAC 配列、RNA-seq データは全て公開データである（表 2-2）。17-mer の出現回数分布の解析より、*C. gigas* のゲノムサイズは 565.7 Mbp と推定される。リピート由来 17-mer の割合は 0.471 となり、線虫（*S. venezuelensis*: 0.289、*C. elegans*: 0.239）よりリピート配列の割合が多いことが分かる。ゲノムの一部分に対応する 8 本の完成 BAC 配列（合計長　1,081,613 bp）をヘテロ接合度の推定やアセンブリの精度評価に用いた。ヘテロ接合度は、BAC に paired-end のリードをマップし *C. elegans* と同じ手法（2.3.2）で 0.923 %と推定した。*C. gigas* はヘテロ接合度が高いだけでなく、線虫と比較してゲノムサイズやリピート配列の割合が大きく、*de novo* アセンブリがより困難なサンプルであると考えられる。

・ベンチマーク結果および考察

各アセンブラの結果と Oyster Genome Project 発表のドラフトゲノム（fosmid-based reference）とを比較した。ドラフトゲノムは、ゲノム全体をカバーする fosmid 配列のセットと全ゲノムショットガンリードを組み合わせ、独自のパイプラインにより構築されたものである（Zhang et al. 2012）。ここで、Velvet と MaSuRCA は実行途中に異常終了したため比較対象としていない。その際の実行環境は、RAM: 512 GB、プロセッサ数: 32 である。Velvet はゲノムサイズの大きいサンプルではメモリ使用量が増大することが GAGE ベンチマークなどで報告されており（Salzberg et al. 2012）、今回もそれが原因と考えられる。MaSuRCA は 1 ヶ月以上計算を行なった後異常終了した。明確な原因は不明であるが、MaSuRCA の論文（Zimin et al. 2014）では 500 Mbp 以上のゲノムサイズを持つサンプルでのテストは行なっておらず、ゲノムサイズの大きいケースでの問題が顕在化した可能性がある。

長さの評価基準である NG50 を算出する際のゲノムサイズは、17-mer 出現回数分布からの推定値（565.7 Mbp）を用いた。精度評価としては、8 本の BAC 配列とアセンブリ結果を、*S. venezuelensis* ベンチマーク時の fosmid と同じ手順

（2.3.3）で比較した。また、遺伝子配列のカバー率を評価するため、RNA-seq データを *de novo* アセンブルすることで RNA-contig を構築し、ゲノムのアセンブリ結果へのマップ率も調べた。その際、RNA-seq データのアセンブリには Trinity（Grabherr et al. 2011）を用い、その結果（RNA-contig）のうち 500 bp 以上の配列を blat（Kent 2002）によりアラインした。各 RNA-contig において identity ≥ 90 %かつ coverage ≥ 90 %のアライメントが得られたとき、マップされたと判定する。マップされた RAN-contig について、アライメントに'N'が含まれないものの割合も調べた。アセンブリ結果の評価を表 2-9 にまとめる。

表 2-9　*C. gigas* アセンブリ結果

		Platanus	ALLPATHS-LG	SOAPdenovo2	Fosmid-based Reference
アセンブリ結果	合計長 (≥500 bp)	684,614,954	655,152,639	859,413,081	557,340,816
	Scaffold数 (≥500 bp)	36,091	18,238	67,846	6,432
	Scaffold NG50 (bp)	381,943	154,144	116,321	392,835
	Contig NG50 (bp)	9,011	12,025	11,719	26,430
BACによる評価	Top-hits-length (bp)	864,992	752,977	851,083	750,984
	平均identity (%)	96.48	96.41	96.28	96.92
	Contained BAC数	3	2	2	1
RNA-seqによる評価	平均identity (%)	42,801,107	38,060,320	40,846,500	42,241,208
	Average Identity (%)	98.48	98.34	98.47	98.52
	マップされた RNA-contig数	30,700	28,152	30,230	30,150
	マップされた RNA-contigs数 ('N' を含まないアライメント)	28,452	25,914	27,092	28,520

Platanus の scaffold NG50 は他のアセンブラの値より大きく、fosmid-based reference の値に近くなっている。BAC の Top-hits-length は Platanus が最大の値を示し、アセンブリ結果の精度と長さが両立していることが示唆される。fosmid-based reference でこの値が最小となっている明確な理由は不明であるが、ゲノム中で fosmid ではカバーされていない領域が存在している可能性がある。contig NG50 については Platanus の値が最小となっているが、RNA-contig のマップ率は最大となっており、'N'をふくむマッピング結果を除いても fosmid-based reference に近い値であり他のアセンブラの結果よりは大きい。このことから Platanus の結果について、遺伝子領域ではギャップが少ないことが示されている。Platanus の scaffold が fosmid-based reference の代替となり得るかということを調べるため、Platanus と fosmid-based reference のそれぞれに特異的にマップされる RNA-contig の数をカウントした（表 2-10）。

表 2-10　Platanus の scaffold、fosmid-based reference に特異的にマップされる RNA-contig

	合計長 (bp)	Contig数	平均長 (bp)	nrにヒットを持つcontig数
Platanus-specific (Reference: nohit)	1,680,407	1,357	1238.3	928
Reference-specific (Platanus: nohit)	540,263	366	1476.1	271

"nr にヒットを持つ contig 数"は、NCBI nr データベースに e-value $<10^{-5}$ のヒットを持つ RNA-contig の数。

その結果、Platanus の scaffold にのみマップされる RNA-contig 数の方が多く、さらにその内 68 %（928 / 1357）は NCBI nr データベースにヒット（e-value $<10^{-5}$）を示す。Platanus の scaffold は、ゲノム配列解析を行うにあたって fosmid-based reference の代替となり得る精度を有していることが示唆されている。

Platanus のアセンブリが成功している個別の例として、BAC のうち fosmid-based reference が 3 つに分断されているものを示す（図 2-28）。

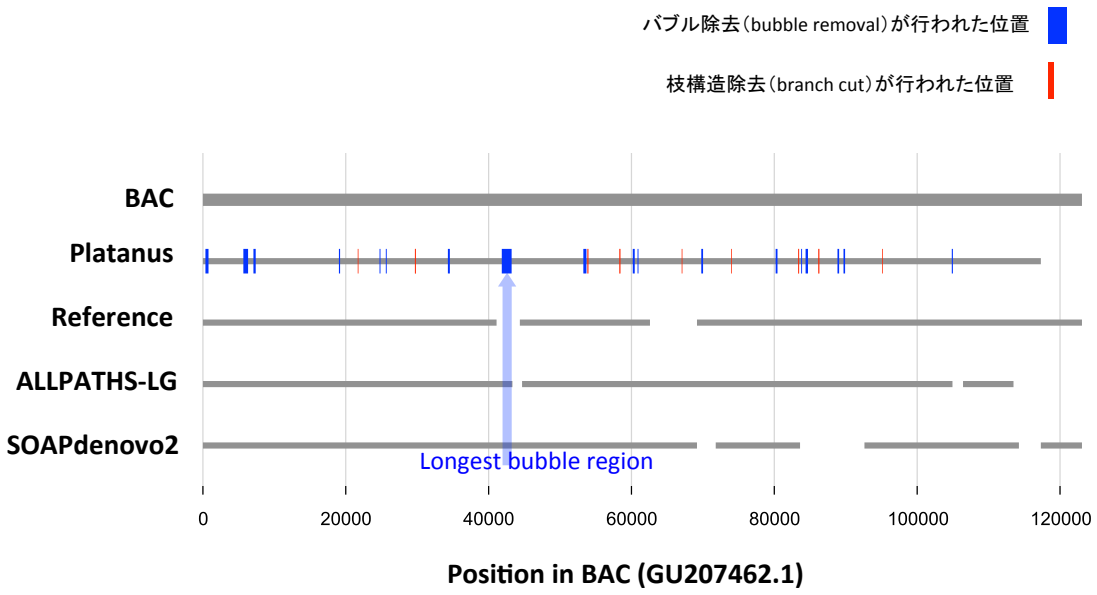


図 2-28　BAC と scaffold の比較の例

Platanus の配列上では、scaffolding 中にバブル構造除去（bubble removal）、枝構造除去（branch cut）が行われた箇所をそれぞれ青色と赤色で示す。

fosmid-based reference が分断されている箇所の 1 つは、Platanus が scaffolding におけるバブル除去で解決した領域である。除去されたバブル配列の大きさは 1 kbp 強であり、約 800 bp の indel を含んでいた。この箇所の近傍では ALLPATHS-LG の scaffold も分断されており、*C. gigas* においても *S. venezuelensis* と同様に構造変異を解決することで、Platanus は scaffold の長さに関して優位を得ていることが示唆されている。

2.3.5　Assemblathon2 のデータによるベンチマーク結果と考察

・Assemblathon2 データ(bird、snake、fish)の取得および特徴

Assemblathon2 は *de novo* アセンブリの国際コンテストである（Bradnam et al. 2013）。*Melopsittacus undulatus* ("bird")、*Boa constrictor constrictor* ("snake")、*Maylandia zebra* ("fish")の 3 種の脊椎動物のシークエンスデータが事前に公開されており、参加者は各自アセンブリを行い結果の配列を運営側へアップロードし、登録期間終了後に評価が行われる。アセンブリに用いるツール、計算機に制限はなく、多くのチームが複数のツールを組み合わせてパイプラインを作っている。参加者名（Team name）、使用ツールを表 2-11 に示す。同一チームが複数のアセンブリを登録している場合は Team name に'*'をつけて区別している。また、bird のみは Illumina データだけでなく Roche-454、PacBio（Pacific Biosciences 社製 1 分子 DNA シークエンサ）データが公開されており、それらを使用したチームも含まれる。なお、Platanus は登録期間中に開発が完了していなかったため、Assemblathon2 に参加はしていない。

3 種の推定ゲノムサイズは、17-mer 出現回数分布より bird: 1.43 Gbp、snake: 1.08 Gbp、fish: 0.915 Gbp である（図 2-21、表 2-1）。ヘテロ接合度については、bird と snake については公開されている fosmid 配列、fish は Platanus の scaffold 上にリードをマップして求めた。その推定値は bird: 0.465 %、snake: 0.165%、fish: 0.147%となった。3 種のヘテロ接合度は *S. venezuelensis* や *C. gigas* と比較すると高くはないが、1 Gbp 前後のゲノムサイズを持つサンプルへ Platanus が適用可能であるかを調べることができる。また、ベンチマークの際は、あるツールの実行方法が最適化されていない可能性が存在するという問題があるが、Assemblathon 2 では各ツールの開発者も多く参加しており、十分に最適化されたツールの性能を比較できるという利点がある。長さの指標である NG50 の算出時のゲノムサイズは 17-mer 出現回数からの推定値を用いた。精度評価としては、bird と snake は Assemblathon 2 で用いられ公開されている fosmid 配列（bird: 1,035,129 bp、snake: 378,186 bp）をアセンブリ結果にアラインし、Top-hits-length（方法: 2.3.3）などを算出した。

表 2-11　Assemblathon2 の参加者情報

(A) Bird

Team name	主要な使用ツール	使用データ
ABL	HyDA	Illumina, Roche-454
Allpaths	ALLPATHS-LG	Illumina
BCM-HGSC	SeqPrep, KmerFreq, Quake, BWA, Newbler, ALLPATHS-LG, Atlas-Link, Atlas-GapFill, Phrap	Illumina, Roche-454, PacBio
BCM-HGSC*	SeqPrep, KmerFreq, Quake, BWA, Newbler, ALLPATHS-LG, Atlas-Link, Atlas-GapFill, Phrap	Illumina, Roche-454, PacBio
CBCB	Celera assembler and PacBio Corrected Reads (PBcR)	Illumina, Roche-454, PacBio
CoBig2	4Pipe4 pipeline, Seqclean, Mira, Bambus2	Roche-454
MLK-Group	ABySS	Illumina
Meraculous	meraculous	Illumina
Newbler-454	Newbler	Roche-454
Phusion	Phusion2, SOAPdenovo, SSPACE	Illumina
Ray	Ray	Illumina
SGA	SGA	Illumina
SOAPdenovo	SOAPdenovo	Illumina
SOAPdenovo*	SOAPdenovo	Illumina
SOAPdenovo**	SOAPdenovo	Illumina

(B) Snake

Team name	主要な使用ツール
ABySS	ABySS
BCM-HGSC	SeqPrep, KmerFreq, Quake, BWA, Newbler, ALLP A THS-LG, Atlas-Link, Atlas-GapFill, Phrap
CRACS	ABySS, SSPACE, Bowtie, and FASTX
Curtain	SOAPdenovo, fastx_toolkit, bwa, samtools, velvet, curtain
GAM	GAM, CLC and ABySS
Meraculous	meraculous
PRICE	PRICE
Phusion	Phusion2, SOAPdenovo, SSPACE
Ray	Ray
SGA	SGA
SOAPdenovo	SOAPdenovo
Symbiose	Monument, SSPACE, SuperScaffolder, GapCloser

(C) Fish

Team name	主要な使用ツール
ABySS	ABySS
Allpaths	ALLPATHS-LG
BCM-HGSC	SeqPrep, KmerFreq, Quake, BWA, Newbler, ALLPATHS-LG, Atlas-Link, Atlas-GapFill, Phrap
CSHL	Metassembler, ALLPATHS, SOAPdenovo
CSHL*	Metassembler, ALLPATHS, SOAPdenovo
CSHL**	Metassembler, ALLPATHS, SOAPdenovo
CTD	Unspecified
CTD*	Unspecified
CTD**	Unspecified
IOBUGA	ALLPATHS-LG, SOAPdenovo
IOBUGA*	ALLPATHS-LG, SOAPdenovo
Meraculous	meraculous
Ray	Ray
SGA	SGA
SOAPdenovo	SOAPdenovo
Symbiose	Monument, SSPACE, SuperScaffolder, GapCloser

・ベンチマーク結果および考察

Assemblathon2 の参加者全てに Platanus を加えた結果を表 2-12 に示す。注目すべきことに、bird と snake に関しては、Platanus の scaffold NG50、fosmid の Top-hits-length が最大となっており、1 Gbp 以上のゲノムサイズを持つ生物種に対しても Platanus の scaffold が長さと精度を両立していることが示唆された。他の参加者の多くは複数のツールを組み合わせているのに対して、Platanus は単体で使用されており、デフォルトのパラメータを用いている。今回の結果の傾向が他のサンプルにも当てはまるならば、Platanus の利用により *de novo* アセンブリプロトコルの簡略化が期待される。fish については、Platanus の scaffold NG50 の値は 17 の参加者中 5 位であり、単体のツールを用いた参加者の内では Allpaths（ALLPATHS-LG）に次いで 2 位であった。fish のデータの特徴としては、paired-end の coverage depth が 52.5 と低めであることが挙げられる。Platanus の NG50 が低 coverage depth のデータで低下することは *C. elegans* データのダウンサンプリングテスト（後述 2.3.6）でも示されており、fish データでもその影響を強く受けたと考えられる。

表 2-12　Assemblathon 2 (Bird, Snake, Fish)アセンブリ結果

(A) Bird

Team name	アセンブリ結果			Fosmidによる評価		
	合計長 (≥ 500bp)	Scaffold NG50 (bp)	Contig NG50 (bp)	Top-hits-length (bp)	Identity (%)	Contained fosmids数
Platanus	**1,129,507,736**	**21,684,294**	**56,973**	**1,026,396**	**99.40**	**84**
ABL	962,501,883	3,003	3,003	415,799	99.63	18
Allpaths	1,148,183,934	17,716,398	57,617	1,001,521	99.22	77
BCM-HGSC	1,322,103,957	17,484,024	184,968	995,482	99.40	80
BCM-HGSC*	1,323,179,685	17,488,428	115,689	992,609	99.44	80
CBCB	1,219,131,251	2,030,397	116,375	1,021,740	99.34	81
CoBig2	11,169,988	0	0	30,036	89.25	0
MLK-Group	1,871,173,099	150,983	36,470	876,624	99.45	58
Meraculous	1,081,601,988	9,834,474	37,061	975,863	99.43	71
Newbler-454	1,117,368,293	11,495,203	66,030	995,738	99.38	74
Phusion	1,334,459,170	1,479,065	74,321	931,879	99.26	66
Ray	1,266,700,501	673,924	44,663	987,974	99.41	78
SGA	1,152,568,402	3,584,181	17,785	912,595	99.49	59
SOAPdenovo	1,150,787,380	14,644,723	40,684	1,009,988	99.24	81
SOAPdenovo*	1,147,593,642	14,617,523	54,932	1,010,089	99.29	81
SOAPdenovo**	1,147,651,023	14,617,693	53,557	1,010,137	99.29	81

(B) Snake

Team name	アセンブリ結果			Fosmidによる評価		
	合計長 (≥ 500bp)	Scaffold NG50 (bp)	Contig NG50 (bp)	Top-hits-length (bp)	Identity (%)	Contained fosmids数
Platanus	**1,441,357,474**	**17,165,953**	**48,614**	**368,203**	**99.80**	**53**
ABySS	1,513,411,700	508,539	28,651	356,343	99.80	47
BCM-HGSC	1,439,656,907	1,563,800	12,727	333,109	99.51	39
CRACS	1,514,104,169	738,101	16,852	344,776	99.76	39
Curtain	1,497,321,002	58,954	4,476	279,969	99.34	25
GAM	1,367,480,854	19,350	4,499	306,517	99.22	29
Meraculous	1,426,938,655	1,247,790	34,344	356,103	99.85	47
PRICE	312,669,723	0	0	60,537	94.92	2
Phusion	1,522,735,963	4,494,848	77,302	365,061	99.71	53
Ray	1,527,313,311	143,603	18,158	340,844	99.76	42
SGA	1,442,930,293	4,536,273	28,397	357,395	99.71	51
SOAPdenovo	1,608,796,330	2,004,523	18,848	365,687	99.77	51
Symbiose	1,989,598,057	1,794,214	82,841	339,360	99.58	45

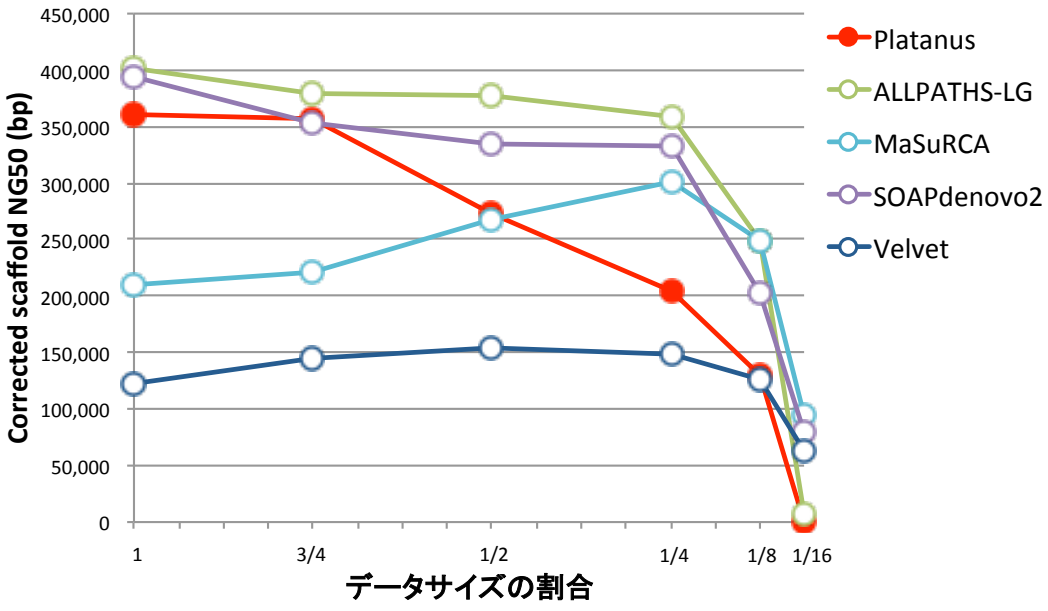
(C) Fish

Team name	アセンブリ結果		
	合計長 (≥ 500bp)	Scaffold NG50 (bp)	Contig NG50 (bp)
Platanus	**825,154,699**	**2,371,946**	**6,587**
ABySS	849,812,973	1,013,854	4,030
Allpaths	845,883,785	3,677,909	14,105
BCM-HGSC	868,269,785	4,850,564	17,895
CSHL	845,089,232	3,418,986	16,740
CSHL*	844,657,792	3,418,986	16,517
CSHL**	545,734,294	1,560	1,550
CTD	1,325,939,605	3,139	935
CTD*	933,447,632	935	3,139
CTD**	989,896,657	1,285	1,285
IOBUGA	825,949,698	261,156	1,472
IOBUGA*	2,193,012,109	51,717	556
Meraculous	801,481,081	764,900	3,962
Ray	787,647,067	37,280	6,829
SGA	812,239,698	88,883	4,739
SOAPdenovo	1,063,706,964	1,665,791	7,001
Symbiose	1,101,461,395	1,731,822	30,444

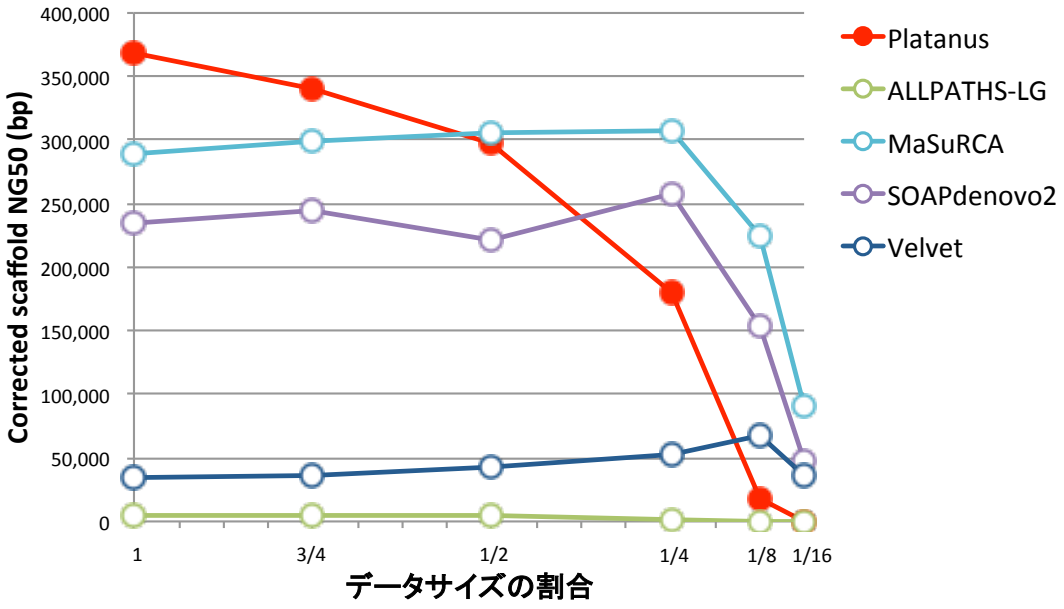
2.3.6　ダウンサンプリングテストによる最適 coverage に関する考察

アセンブラ毎に最適な coverage depth が異なる可能性があるため、シークエンスデータをダウンサンプルしてアセンブリを行い、coverage depth と corrected scaffold NG50 の関係を調べた。具体的には、*C. elegans* のヘテロ接合度 0.0%、1.0%、2.0%の各ケースで、全ライブラリから 3/4、1/2、1/4、1/8、1/16 の 5 通りの割合でリードを抽出した。ベンチマーク結果を図 2-29 に示す。データサイズの割合 3/4 では、Platanus の corrected scaffold NG50 は大きく減少せず、ヘテロ接合度が 1.0%以上のケースにおいて他のアセンブラより大きな値を示す。しかし、データサイズの割合がそれより下がるにつれて corrected scaffold NG50 も急激に減少する。データサイズが 3/4 のとき、paired-end の coverage depth は 104 であり、Platanus が最も性能を発揮する条件は 100 以上の coverage depth であると考えられる。また、coverage depth が低いデータに対応できているアセンブラは overlap-layout-consensus アルゴリズムを採用している MaSuRCA であることが示唆される。ここで、他のアセンブラは coverage depth を上げると corrected scaffold NG50 の値がプラトーに達するかむしろ減少していると解釈することもできるが、Platanus の値は増加し続ける傾向があり、この性質はシークエンサのスループットが増加し続けている状況に適しているという可能性もある。

(A) *C. elegans* ヘテロ接合度 0.0%



(B) *C. elegans* ヘテロ接合度 1.0%



(C) *C. elegans* ヘテロ接合度 2.0%

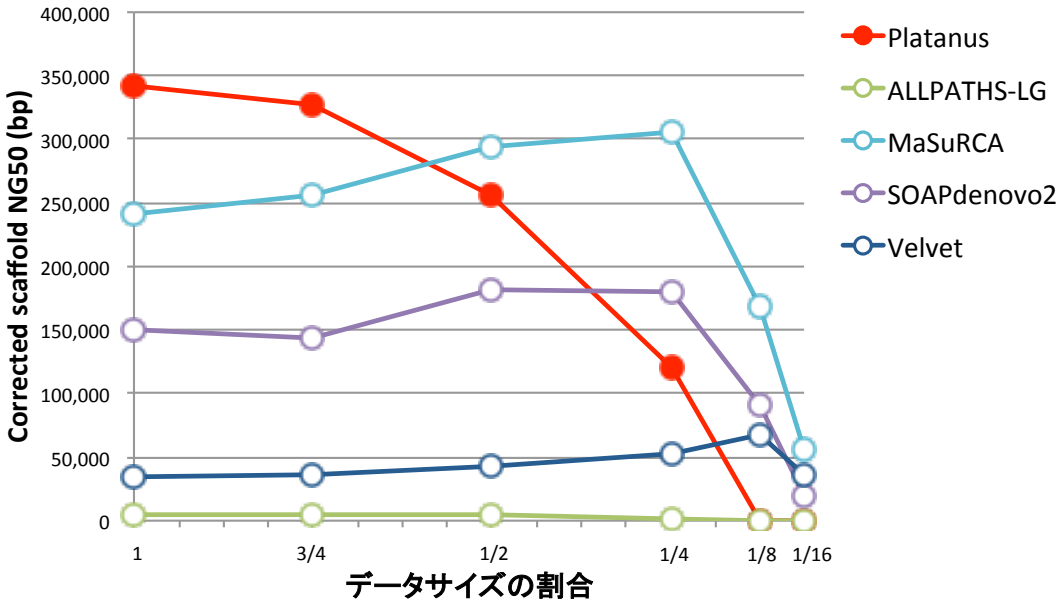


図 2-29 ダウンサンプリングテスト（corrected scaffold NG50）

2.3.7　実行時間、メモリ使用量についてのベンチマーク結果と考察

C. elegans（ヘテロ接合度 0.0%）、*S. venezuelensis*、*C. gigas* データについて、各アセンブラの実行時間とメモリ使用量ピーク値を表 2-13 に示す。実行環境は次の通りである。

・プロセッサ: Intel(R) Xeon(R) CPU X7560 2.27 GHz
・プロセッサ数: 32
・RAM: 512 GB

全てのアセンブラのスレッド数は 32 と指定した。SOAPdenovo2 については、付属プログラムである GapCloser の実行時間、メモリ使用量についても測定対象としている。*C. gigas* のアセンブリでは、MaSuRCA と Velvet は異常終了したため値を記していない。全体的な傾向としては、overlap-layout-consensus アルゴリズムを採用している MaSuRCA の実行時間は大きい。他の de Bruijn グラフを用いたアセンブラについては、線虫（*C. elegans*、*S. venezuelensis*）のデータでは SOAPdenovo2 が実時間、メモリ使用量ピーク値ともに最小であるが、データ量の大きい *C. gigas* では Platanus の実時間、メモリ使用量ピーク値の方が小さくなっている。ここで、表 2-13 には、SOAPdenovo2 と Velvet のパラメータ調整時（*k*-mer 長などを複数試すステップ）についての値は入っていないため、それらのアセンブラについては、実際にはより多くの実行時間が必要になると予想される。それを考慮に入れると、特に大きなサイズのデータを扱う際には、Platanus は時間とメモリの両方に関して、今回テストしたアセンブラの中では効率が良いと考えられる。

表 2-13　*de novo* アセンブリの実行時間、メモリ使用量ピーク値

(A) *C. elegans* ヘテロ接合度 0.0%

	CPU time	Real time	Peak memory (GB)
Platanus	588,408 s (163 h)	23,966 s (7 h)	20.0
ALLPATHS-LG	648,721 s (180 h)	62,844 s (17 h)	129.6
MaSuRCA	802,214 s (223 h)	64,055 s (18 h)	72.9
SOAPdenovo2	86,605 s (24 h)	6,873 s (2 h)	36.1
Velvet	23,191 s (6 h)	4,727 s (1 h)	35.0

(B) *S. venezuelensis*

	CPU time	Real time	Peak memory (GB)
Platanus	238,767 s (66 h)	10,431 s (3 h)	19.8
ALLPATHS-LG	424,661 s (118 h)	26,515 s (7 h)	73.1
MaSuRCA	748,571 s (208 h)	118,230 s (33 h)	70.1
SOAPdenovo2	53,453 s (15 h)	5,449 s (2 h)	16.6
Velvet	19,442 s (5 h)	3,639 s (1 h)	38.2

(C) *C. gigas*

	CPU time	Real time	Peak memory (GB)
Platanus	2,485,919 s (691 h)	114,107 s (32 h)	98.2
ALLPATHS-LG	3,860,440 s (1,072 h)	306,899 s (85 h)	322.7
MaSuRCA		異常終了	
SOAPdenovo2	2,254,545 s (626 h)	248,160 s (69 h)	148.4
Velvet		異常終了	

2.4　Platanus の実データに対する適用例：シーラカンスゲノム解読

Platanus が活用された初の例は、シーラカンスゲノム解読計画である（Nikaido et al. 2013）。本計画では、Platanus でドラフトゲノムを構築しただけでなく、アノテーション、変異解析、2 種（*L. chalumnae*、*L. menadoensis*）の配列比較など様々な解析を行なった上で発表を行なっている。本節では、シーラカンスゲノム解読における Platanus の適用例と個体の比較解析について記すこととする。

2.4.1　シーラカンスゲノムのアセンブリ

・シーラカンスシークエンスデータの取得およびその特徴

シークエンスされたシークエンス個体は 5 個体で、生息地がアフリカ南東沖である *L. chalumnae* が 4 個体、インドネシア沖である *L. menadoensis* が 1 個体（"Indonesia"）である。*L. chalumnae* は更にタンザニア沖の 3 個体（TCC041-004、S2、TCC25）とコモロ諸島沖の 1 個体（"Comoro"）に分けられる。これらは漁業で偶然に混獲された個体であり、ワシントン条約に基づき許可を得て輸入されている。標本の輸入経路は以下の通りである。

・タンザニア沖の 3 個体

タンザニア水産研究所から東京工業大学

・Comoro

Centre National de Documentation et de Recherche Scientifique, Musee National des Comores からアクアマリンふくしま

・Indonesia

サムラトゥランギ大学から東京大学

標本の移送後は、東京工業大学 岡田典弘教授から国立遺伝学研究所 藤山秋佐夫教授に提供された稚魚個体（TCC041-004）から抽出された DNA、および岡田研究室にて凍結保存サンプル(S2, TCC25)より抽出された DNA、アクアマリンふくしまより提供された DNA(Comoro)、さらに東京大学 菅野純夫教授より提供された DNA(Indonesia)を国立遺伝学研究所 豊田敦准教授がシークエンスし、配列

データが得られた。ドラフトゲノムの構築においては、タンザニア沖産の 1 個体（TCC041-004）の全ゲノムショットガンデータが用いられた。シークエンスデータを表 2-14 に示す。DNA シークエンサは Illumina HiSeq 2000 である。データは 3 つの paired-end と 2 つの mate-pair ライブラリから構成され、インサートサイズは 300–5,000 bp である。また、アダプタ配列、低クオリティ領域の除去はベンチマーク時と同様に行なった。

表 2-14　*L. chalumnae*　TCC041-004 のシークエンスデータ

インサートサイズ (bp)		300	500	1,000	2,500	5,000
ライブラリ		paired-end	paired-end	paired-end	mate-pair	mate-pair
リード長 (bp)	**Raw**	100	100	100	100	100
	Preprocessed	97.3	92.3	93.2	96.7	94.8
合計長 (Gbp)	**Raw**	242 G	299 G	168 G	72.8 G	103 G
	Preprocessed	236 G	276 G	156 G	70.8 G	98.8 G

"Preprocessed"はアダプタ配列、低クオリティ領域の除去後の値。

インサートサイズ 300 bp、500 bp の paired-end ライブラリより 32-mer の出現回数の分布を求めた（図 2-30）。90（横軸）付近のピークはホモ領域に対応し、ヘテロ領域のピークは明確には見られない。180 付近のピークはリピート配列由来である。この分布の形からヘテロ接合度は低いと予想され、更にゲノムサイズは 2.67 Gbp と推定された。

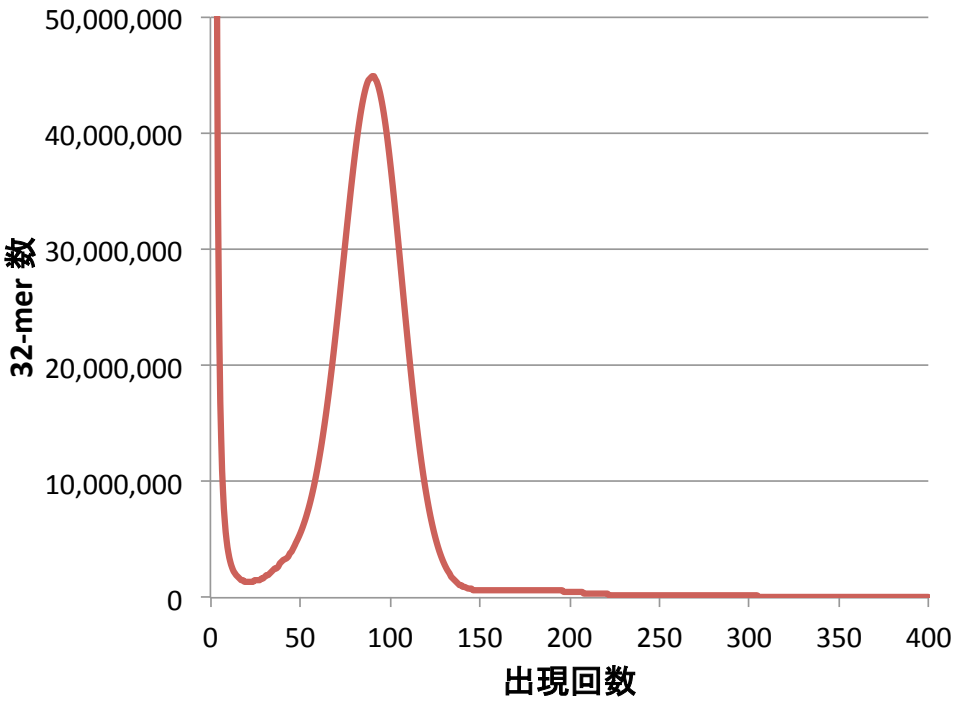


図 2-30　*L. chalumnae* paired-end データの 32-mer 出現回数分布

・アセンブリ結果および考察

シーラカンスのゲノム解析においては、まずタンザニア沖産の 1 個体（TCC041-004）の全ゲノムショットガンデータを *de novo* アセンブリした。続いて、そのドラフトゲノムに他の個体の全ゲノムリードをマップすることで全ての個体のゲノム配列を構築した。

de novo アセンブリには Platanus（version 1.0.0）が用いられた。インサートサイズが 1 kbp 以上のライブラリは PCR-duplicate や短いインサートサイズのペアを比較的多く含むため、Contig-assembly はインサートサイズ 300 bp、500 bp の paired-end を入力として行われた。1 kbp よりインサートサイズが大きいライブラリを用いる際には、それより小さなインサートサイズのライブラリで Scaffolding、Gap-close まで行なって scaffold を構築し、その配列にリードをマップ後、PCR-duplicate と短いインサートサイズのペアを除くという作業を繰り返した（方法: 2.3.1・ベンチマークデータの前処理）。アセンブリ結果を表 2-15 に示す。比較のため、SOAPdenovo（Li et al. 2010）の結果も含めている。NG50 を求める際のゲノムサイズは、32-mer 出現回数分布からの推定値（2.67 Gbp）を用

いた。SOAPdenovo の *k*-mer 長は 31 から 71 まで 10 刻みで入力し、scaffold NG50 が最大となる *k* = 61 を採用した。ここで、付属プログラムの GapCloser は異常終了するため実行していない（実行環境　プロセッサ: Intel(R) Xeon(R) CPU E7-8837, 2.67GHz, プロセッサ数: 32, RAM: 1TB）。Platanus と SOAPdenovo の結果を比較すると、scaffold NG50、contig NG50、'N'の割合は全て Platanus が優っている。

表 2-15　*L. chalumnae* アセンブリ結果

	Scaffold数 (≥ 300 bp)	合計長 (≥300 bp) (bp)	Scaffold NG50 (bp)	Contig NG50 (bp)	'N'の割合 (%)
Initial contig	1,630,173	2,360,333,583	-	1,990	0
Pre gap-close	42,503	2,745,636,434	342,148	3,217	7.5
Final	37,861	2,736,338,780	340,559	8,816	4.5
SOAPdenovo	46,439	2,727,167,611	195,285	1,751	16.4

"Initial contig"は Contig-assembly の結果、"Pre gap-close"は Scaffolding の結果、"Final"は gap-close を行なった最終結果。

アセンブリ結果の評価は、fosmid-end ライブラリ（HiSeq 2000 およびサンガー法）と完成 fosmid 配列（4 本、合計長 147,605 bp）を用いて行なった。HiSeq 2000 で fosmid-end をシークエンスする際には、NxSeq 40 kb Mate-Pair Cloning kit（Wu et al. 2012）を用いられており、全てのシークエンスは国立遺伝学研究所 豊田敦准教授に実施して頂いた。fosmid-end 配列は Bowtie2 のローカルアライメントモード（--local）で scaffold 配列にアラインされ、両エンド配列が identity ≥ 95%、アライメント長 ≥ 50 bp でマップされたペアについてインサートサイズの分布を調べた。fosmid 断片は約 40 kbp と予想されるため、インサートサイズが 40 kbp 付近を示すペアが多ければ、scaffold にミスアセンブリが少ない、またはギャップが少ないことが示唆されるため、インサートサイズ 30–50 kbp を示すペア数もカウントした。Platanus と SOAPdenovo の scaffold に対して、マッピングの結果を表 2-16、に示す。2 つアセンブラの結果を比較すると、Platanus の方が 30–50 kbp のインサートサイズ（"Normal insert size"）を示すペア数が多いことが分かる。SOAPdenovo の scaffold には GapCloser を適用していないことを考慮し、Platanus の方でも Gap-close 適用前の scaffold（"pre-gap-close"）へのマップ結果を示して

いるが、それでも SOAPdenovo より 30–50 kbp のインサートサイズを示すペアは多くなっている。

表 2-16　Fosmid-end マッピング結果

(A) HiSeq 2000 (ペア数: 5,777,574)

	#Normal insert size	#Abnormal insert size	#Different scaffolds	#Multi best hits	#Unmapped
Platanus (pre-gap-close)	1,525,731 (26.41%)	94,809 (1.64%)	2,497,141 (43.22%)	433,689 (7.51%)	1,226,204 (21.22%)
Platanus (post-gap-close)	**1,639,620 (28.38%)**	102,510 (1.77%)	2,142,983 (37.09%)	856,106 (14.82%)	**1,036,355 (17.94%)**
SOAPdenovo	1,113,386 (19.27%)	68,388 (1.18%)	2,626,459 (45.46%)	463,960 (8.03%)	1,505,381 (26.06%)

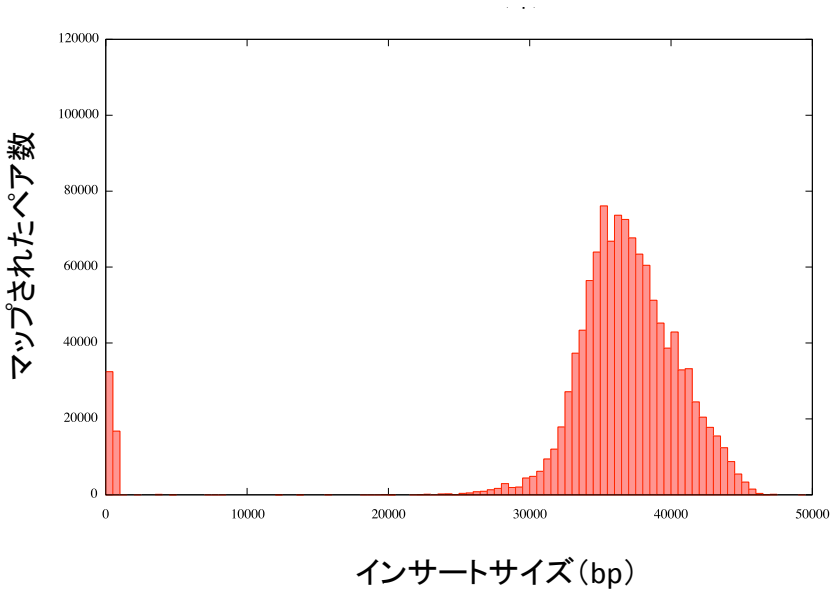
(B) サンガー法 (ペア数: 55,111)

	#Normal insert size	#Abnormal insert size	#Different scaffolds	#Multi best hits	#Unmapped
Platanus (pre-gap-close)	22,149 (40.19%)	317 (0.58%)	9,466 (17.18%)	140 (0.25%)	23,039 (41.80%)
Platanus (post-gap-close)	**28,137 (51.06%)**	497 (0.90%)	16,639 (30.19%)	475 (0.86%)	**9363 (16.99%)**
SOAPdenovo	14,561 (26.42%)	188 (0.34%)	10,604 (19.24%)	1,055 (1.91%)	28,703 (52.08%)

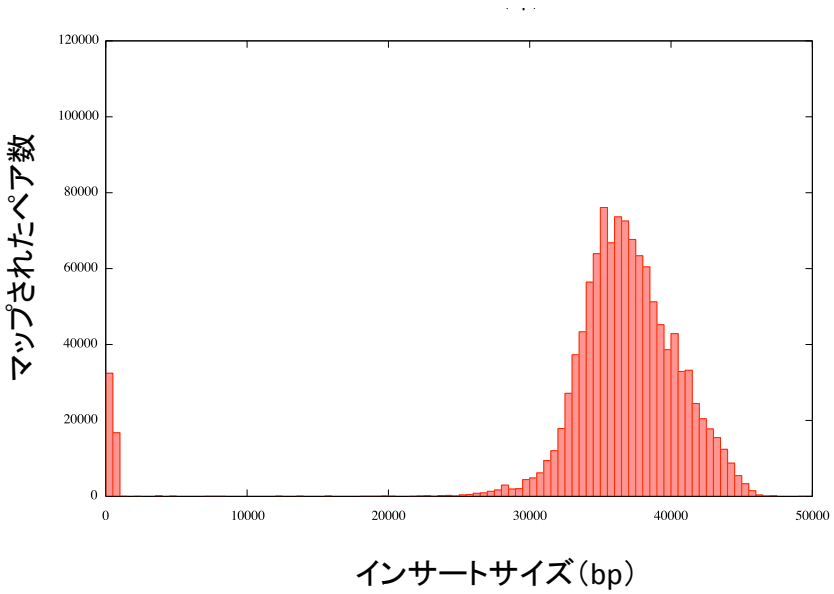
"Normal insert size": 30–50 kbp のインサートサイズを示すペア
"Abnormal insert size": インサートサイズが 30–50 kbp の範囲外のペア
"Different scaffolds": 異なる scaffold へマップされるペア
"Multi best hits": 少なくとも片方がユニークなベストヒットを持たないペア
"Unmapped": 少なくとも片方がマップされないペア

インサートサイズの分布を図 2-31 に示す。HiSeq と Sanger のデータで傾向は変わらず、Platanus と SOAPdenovo の両方でピークは 40 kbp 付近であるが、ピークの高さ（ペア数）は Platanus の方が大きくなっている。以上のことから、Platanus の scaffold の方がマップ率は高く、ギャップまたはミスアセンブリが少ないことが示唆された。

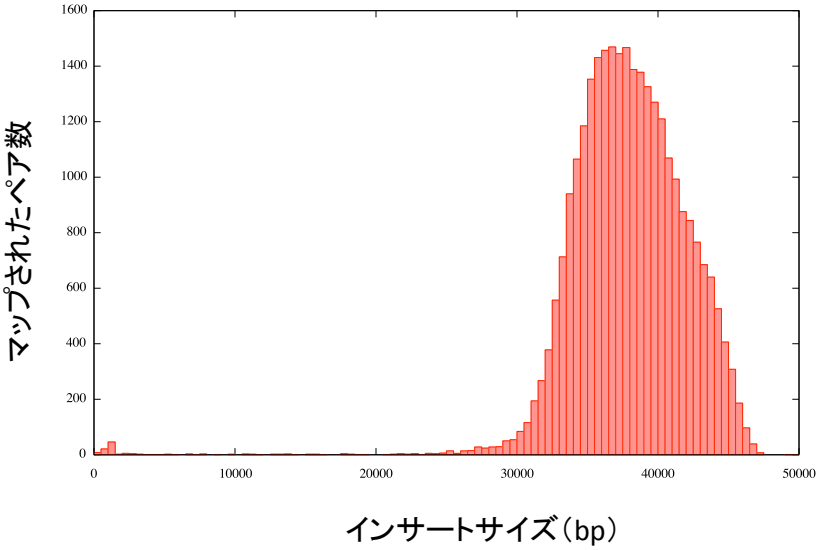
(A) HiSeq 2000-Platanus



(B) HiSeq 2000-SOAPdenovo



(C) サンガー法-Platanus



(D) サンガー法-SOAPdenovo

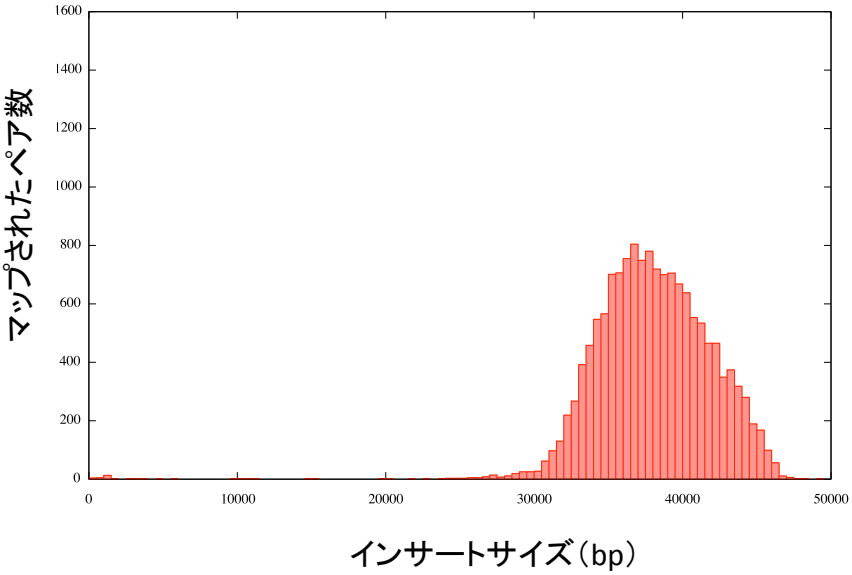


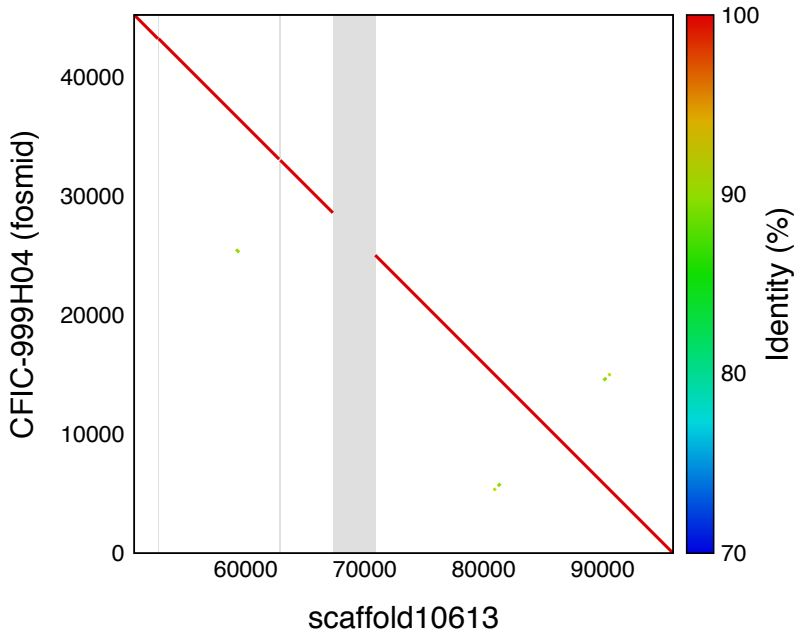
図 2-31　Scaffold 配列にマップされた fosmid-end のインサートサイズ分布

続いて 4 本の完成 fosmid 配列（合計長 147,605 bp）を用いて Platanus の scaffold とアラインすることで塩基レベルでの精度を評価した。用いたツールは MUMmer パッケージに含まれるプログラム（nucmer、delta-filter、dnadiff）である。具体的な手順を以下に記す。

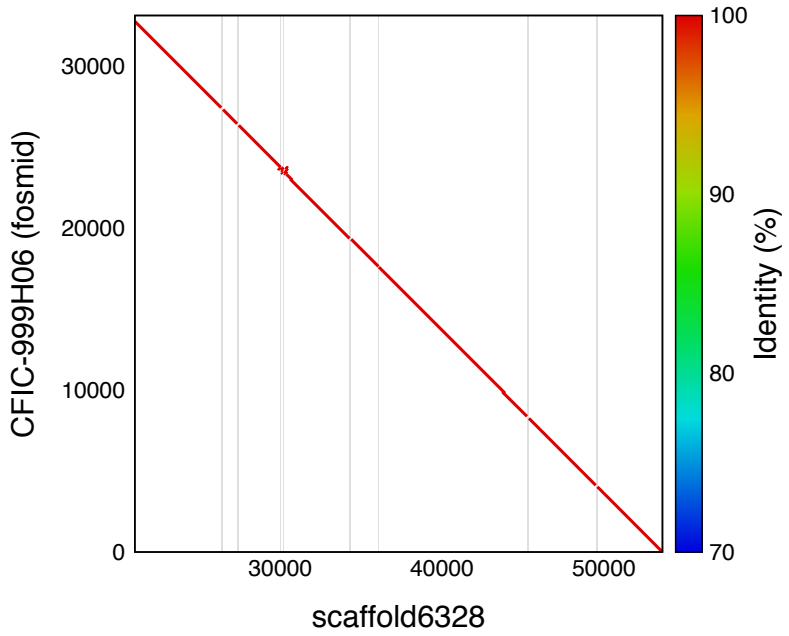
(1) fosmid をクエリ、scaffold をリファレンスとし、nucmer によりアライメントする。その際 -noextend オプション（"do not execute the cluster extension step"）を指定した。
(2) delta-filter を -g オプション（" 1-to-1 global alignment not allowing rearrangements"）とともに使用し、アライメント結果をフィルタリングする。
(3) dnadiff で SNV（ミスマッチ）をコールする。

その結果、アライメントの合計長は 138,070 bp、コールされた SNV は 2 個であった。その SNV は変異解析（後述）でも検出されており、アセンブリのエラーではないと考えられる。アライメントのドットプロットを図 2-32 に示す。この図では、規模の大きいミスアセンブリも見当たらない。以上の結果から、Platanus の scaffold の精度も高いことが示唆される。

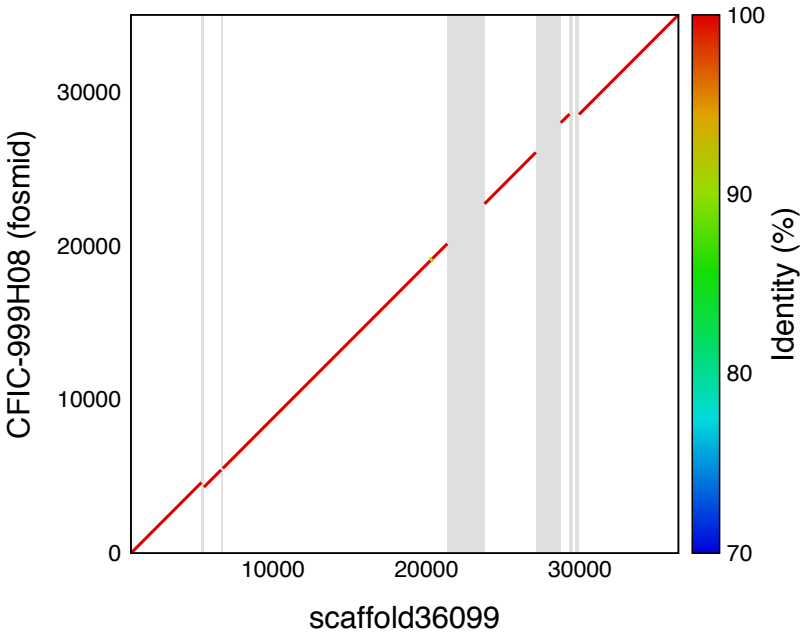
(A) CFIC-999H04



(B) CFIC-999H06



(C) CFIC-999H08



(D) CFIC-999H10

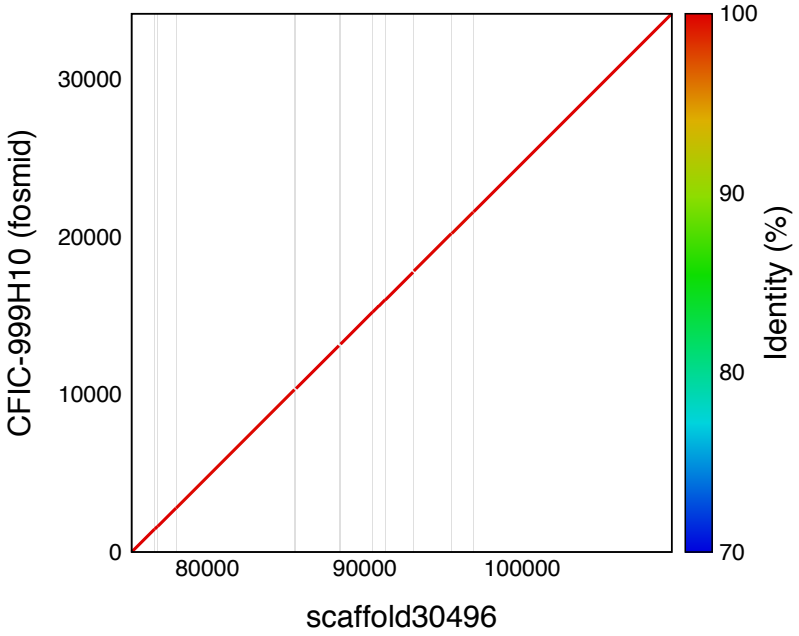


図 2-32　Platanus の scaffold と完成 fosmid 配列のドットプロット

横軸が scaffold、縦軸が fosmid のポジション。灰色の四角形はギャップを示す。

2.4.2　変異解析、5 個体の比較解析

Platanus で構築した scaffold 配列の活用例として、変異解析について述べる。scaffold を *L. chalumnae* のドラフトゲノムとして、*L. menadoensis* も含めた他の 4 個体のゲノム配列の決定を試みた。方法としては、scaffold に各サンプルの paired-end をマップし、ホモおよびヘテロの変異を検出する。検出の手順は 2.3.2 で記した手順と類似であるが、変異をコールする条件について、以下の点が異なる。

・coverage depth の下限は、平均 coverage depth×0.5
・各ストランドで、coverage depth が平均 coverage depth×0.2 以上
・各ストランドで、変異を示すリードの割合が 0.25 以上

これらは主に偽陽性を減らすための追加条件である。Illumina リードの系統的エラーは、各ストランドで変異がコールされているかを考慮することで一部は除くことができる（Nakamura et al. 2011）。シーラカンスの変異密度は極めて低く、真の変異に対する偽陽性の割合が大きくなる恐れがあるため、このように条件を厳しくしている。偽陽性が増加する可能性も存在するが、完成 fosmid と scaffold の比較によって検出された SNV（2.4.1）と、リードをマッピングして検出した SNV は一致したので、変異検出はこの方法で行うこととした。

各個体の paired-end ライブラリについての情報を表 2-17 に示す。また、マッピング結果における各リードの編集距離の分布を図 2-33 に示す。注目すべきことに、"Indonesia"は *L. chalumnae* とは別種の *L. menadoensis* の個体であるが、リードは十分にマップされており（coverage depth ≈ 70×）、編集距離 0 でマッピングされるリードも約 70%を占める。このことから、2 種のゲノム配列の相同性は高く、*L. menadoensis* のゲノム配列もマッピング結果から構築できると考えられる。

表 2-17　シーラカンスの各個体の paired-end ライブラリ

	TCC041-004		S2	TCC25	Comoro	Indonesia
	300bp	500bp				
合計長 (bp)	236 G	299 G	165 G	172 G	196 G	256 G
リード長 (bp)	97.3	92.3	94.5	93.9	89.4	93.9
平均インサートサイズ (bp)	253	495	425	449	476	431
インサートサイズ標準偏差 (bp)	46	46	117	82	42	48
PCR-duplicateの割合 (%)	0.260	4.030	1.160	0.703	0.311	0.677
Coverage depth	130.4		44.4	48	56.2	69.9

"Removed duplicates"は samtools rmdup コマンドで除去された PCR-duplicate の割合。"Coverage depth"は実際にマップされたリードによる coverage depth。TCC041-004 は 2 ライブラリ（インサートサイズ：300、500 bp）の値を示す。

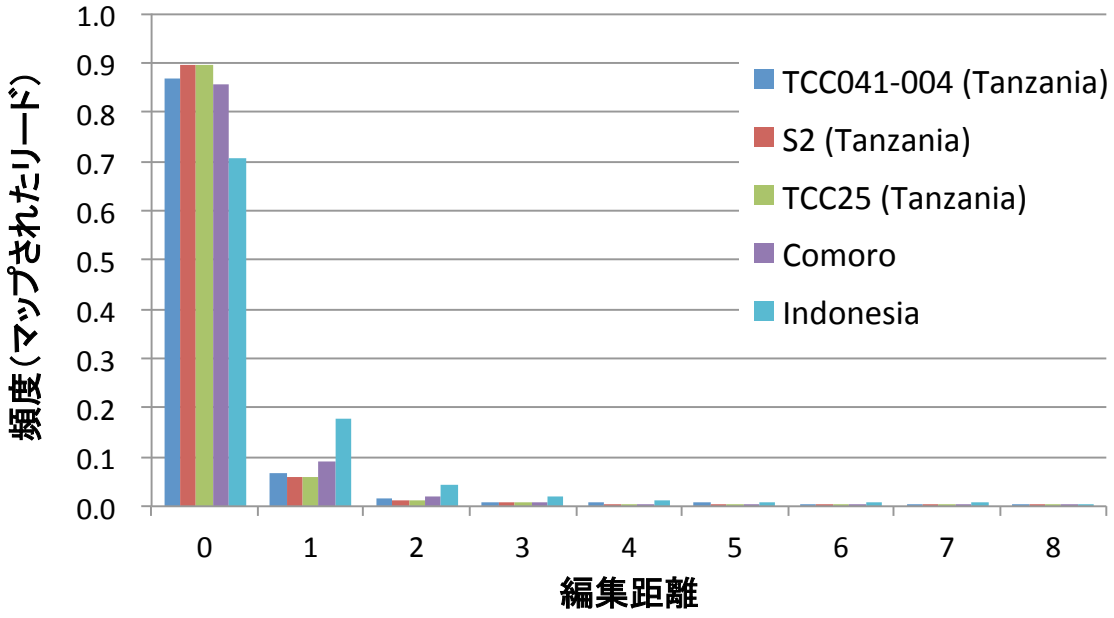


図 2-33　マップされた各リードの編集距離の分布

全ての個体で変異検出の対象となる（coverage depth が十分な）領域は 1,673,302,134 bp であった。この領域内で検出されたヘテロな SNV の密度を表 2-18 に示す。ヘテロな SNV の密度をヘテロ接合度の推定値とすると、これらの値（0.00188%–0.00611%）はヒト（0.069%)（Wang et al. 2008）やゴリラ（0.076%–0.189%）（Scally et al. 2012）よりかなり低い値となる。ヘテロ接合性は集団内の多様性と言い換えることができ、進化速度と集団のサイズを反映している。それらの寄与率を推定するためには追加の情報が必要となるが、シーラカンスの集団サイズが極めて小さい可能性があるという、重要な示唆を得ることができる。

表 2-18　ヘテロな SNV の密度

個体	ヘテロSNV密度
TCC041-004	0.00234%
S2	0.00235%
TCC025	0.00233%
Comoro	0.00188%
Indonesia	0.00611%

全ての個体で変異検出が可能な 1,673,302,134 bp のサイト上の SNV が対象

続いて個体間で塩基が異なる SNV サイトの割合を調べると表 2-19 の値となる。アフリカ南東沖の個体（*L. chalumnae*）とインドネシアの個体（*L. menadoensis*）のゲノムの差異は 0.18%程度と解釈することができる。先行研究では、シーラカンスのミトコンドリアゲノムの解析から、*L. chalumnae* と *L. menadoensis* の分岐年代は 20–30 Ma（million years ago）と推定されている（Inoue et al. 2005; Saitoh et al. 2011）。この分岐年代を仮定すると、シーラカンスゲノムの置換速度は 0.03–0.045×10^{-9} per year となる。これは、ヒトとチンパンジーゲノムの置換速度の推定値 1.2×10^{-9} （Watanabe et al. 2004）よりかなり低い値となっている。

表 2-19　シーラカンス個体間で塩基が異なる SNV サイトの密度

	TCC041-004	S2	TCC25	Comoro
TCC041-004				
S2	0.00326%　(10.0%)			
TCC25	0.00318%　(10.7%)	0.00311%　(11.4%)		
Comoro	0.00343%　(18.7%)	0.00336%　(16.4%)	0.00342%　(17.1%)	
Indonesia	0.18263%　(95.5%)	0.18268%　(95.5%)	0.18267%　(95.5%)	0.18248%　(95.7%)

括弧内はホモな SNV サイトの割合。

全ての個体で変異検出が可能な 1,673,302,134 bp のサイト上の SNV が対象。

2.5　考察

Platanus のアルゴリズムの内、特に新規性が高い部分は Scaffolding 中のヘテロ領域対策（バブル、枝構造の除去）である。de Bruijn グラフにおけるバブル除去は今回ベンチマーク対象としたアセンブラ（ALLPATHS-LG、SOAPdenovo2、Velvet）においても実装されているが、構造変異を含む実データにおいては、それだけでは不十分である可能性がある。Contig-assembly に関しては、複数の *k*-mer 長を活用する点が Platanus の特徴の 1 つであり、低ヘテロ接合性のデータ（*C. elegans*、Assemblathon2 のデータ、シーラカンス *L. chalumnae*）での良好な性能に寄与していると考えられる。SOAPdenovo2 や ALLPATHS-LG も内部で複数の *k*-mer 長を用いるため、その発想自体は新規ではないが、*k* の値の決定方法、複数の de Bruijn グラフのマージ方法はアセンブラ毎に異なっている。グラフの分岐点付近にリードをマップしてグラフを再構築する点、*k* の範囲を coverage depth を考慮して自動で決定する点は Platanus のみが持つ機能である。Platanus は *k* を手動で調整せずとも安定した性能を示すことから、パラメータ調整の時間を節約できるという利点も持つ。Gap-close についても、アルゴリズムの大枠に新規性はないものの、配列の精度が犠牲になると報告されている（Simpson and Durbin 2012）SOAPdenovo2 付属の GapCloser とは異なり、Platanus はギャップ部分構築後も高い精度を保つ（表 2-3）。これには、グリーディーにギャップ周辺を延長する GapCloser と、ギャップ周辺が構築できたときのみギャップを閉じる（図 2-20）Platanus との差が表れている可能性が存在する。

C. elegans のヘテロ接合性シミュレーションテストでは、ヘテロ接合度が高くなるにつれて Velvet、SOAPdenovo、ALLPATHS-LG の性能（NG50、精度）が悪化することが示された。ヘテロ接合性がアセンブリ結果に影響を与えることは報告されていたものの、系統的にテストを行なった研究は 2014 年時点では本研究が最初であると考えられる。高ヘテロ接合性の実データ（*S. venezuelensis*、*C. gigas*）では、シミュレーションの高ヘテロ接合性データで性能が低下しなかった MaSuRCA を含めて、Platanus と他のアセンブラの scaffold NG50 の差が広がった。*S. venezuelensis* での fosmid 配列を用いた解析からは、比較的大きな indel を含む、相同染色体間の配列の違いが大きい領域の存在が示され、それを Scaffolding で解決できる Platanus がより有利となっていることが示唆された。このような領域の存在は *C. gigas* ゲノム解読の元論文（Zhang et al. 2012）でも BAC 配列の比較から示されており、2 種のホヤ *Ciona savignyi* と *Ciona intestinalis* のゲ

ノム中にも検出されている（Vinson et al. 2005; Kim et al. 2007;）。報告された領域では、局所的に相同染色体間の identity が 95%以下になる。このような多様化した領域は多くの生物種のゲノム中で普遍的に存在する可能性があり、野生型サンプルのゲノムを *de novo* アセンブリする際には Platanus のように対策を立てることが重要と考えられる。

Platanus はヘテロ領域をマージする方法を採用しているが、各ハプロタイプ配列を別々に構築する方法も考えられる。もしハプロタイプ配列が十分に長く構築できたならば、変異の連鎖情報などが得られ解析においてはより有効である。そのような方法でアセンブリを行なった例としてはホヤ *C. savignyi* と *C. intestinalis* のゲノム解析（Kim et al 2007; Small et al. 2007）が挙げられる。ここで、*C. scvignyi* のヘテロ接合度は 4.6%、ハプロタイプ配列の N50 は 496 kbp であり、*C. intestinalis* はそれぞれ 1.2%、37. 9 kbp である。ハプロタイプ配列を構築する際には、ヘテロ接合度が高いほど変異間の連鎖情報が得られやすくなるため、配列が長くなると考えられる。言い換えると、ヘテロ接合度が低い場合は変異が存在しない長い領域が多くなるため、リード配列からは連鎖を解くことができない変異の組が増加し、アセンブリ結果は分断される。*C. scvignyi* のように極めて高いヘテロ接合度（4.6%）を持つサンプルの場合はハプロタイプのアセンブリが有効であるが、*C. intestinalis* のように、ヘテロ接合度が 2%以下の場合はアセンブリ結果の N50 が小さくなるため（37.9 kbp）、Platanus のマージ戦略が有効である可能性が高い。

C. gigas のテストにおいては、Platanus のアセンブリ結果と fosmid ベースで構築したリファレンス配列（Zhang et al. 2012）を比較した。NG50 の値は Platanus が劣るが、RNA-seq データや BAC による評価では Platanus が優っている指標が多く、Platanus が fosmid ベースのアセンブリ手法の代替となり得ることが示された。fosmid ベースの手法は *C. gigas* のゲノム解読で導入され、"cost effective" な方法であると表現された。続いてコナガ（*Plutella xylostella*）のゲノム解読においても適用されている（You et al. 2013）。しかしながら、各 fosmid 配列をショットガン法で決定する際に coverage depth が必要となる他、構築された fosmid 配列をアセンブルする際に fosmid 配列がオーバーラップしている必要が生じる。各 fosmid を 100×の depth でシークエンスし、fosmid 配列の合計長がゲノムサイズの 10×であると仮定した場合、総データ量はゲノムサイズの 1,000×となり、全ゲノムショットガン法と比較すると大きいサイズである。実際、*C. gigas*、*P.*

xylostella のケースではそれぞれ 390 Gbp（690×）、855 Gbp（2,170×）の paired-end リードが用いられている。加えて、これらのゲノム解読計画では fosmid のアセンブリ結果を全ゲノムショットガンデータで scaffolding している点にも注意する必要がある。つまり、Platanus により全ゲノムショットガンデータのみからドラフトゲノムが得られれば、fosmid ベースの方法より確実にコストは下げることができると考えられる。

Platanus の適用例としてはシーラカンスゲノム解読計画を扱っている。シーラカンスゲノムはリピート配列の割合が 60%と、哺乳類と比較して多く、このことは *de novo* アセンブリの障害となる。また、推定ゲノムサイズは 2.7 Gbp と本研究で扱った生物種の中では最大のサイズを持つ。リードのマッピングによる別個体との比較では、個体間の配列の差異が極めて小さいことが示されている。もしアセンブリ結果の配列の精度が悪いならば、アセンブリのエラーが変異と誤認識され差異が過大評価される可能性があるが、別個体や別種（*L. menadoensis*）とも配列が高い類似度を持つことは、アセンブリ結果の精度の高さを示唆している。完成 fosmid 配列と scaffold 配列とのアライメント結果に関しても、精度の高さを支持する結果となっている。よって、Platanus は 2.5 Gbp 以上のサイズを持ち、リピート配列率が比較的高い生物種に対しても有効であることが示唆された。アセンブリ後の解析の例としては、構築されたドラフトゲノムを元に行なった、タンザニア沖、コモロ諸島、インドネシア（*L. menadoensis*）の個体間の変異解析を扱った。1 個体の全ゲノムを代表として決定してしまえば、集団内の他のゲノムはリードのマッピングを用いてより安価に決定できる。全ゲノムの *de novo* アセンブリは集団遺伝学的な知見を得る上で有効な手段であり、ドラフトゲノム配列の数多い活用例の 1 つとして今回は示すこととした。シーラカンスゲノムについては、米国の Broad Institute を中心とするチームも 2013 年にドラフトゲノムおよび解析結果を発表している（Amemiya et al. 2013）。この論文発表の方が Nikaido et al.による発表（オンライン）より約 3 ヶ月早いが、2 つの解読計画は独立に進められており、ドラフトゲノムのアセンブリに用いたデータも全て異なっている。Broad Institute のドラフトゲノムは ALLPATHS-LG によってアセンブリされた。その scaffold NG50 を算出すると 1,012,495 bp となり、Platanus の値（340,559 bp）より大きいが、この差はライブラリ構成に起因する可能性が高い。Platanus に入力された mate-pair のインサートサイズは 5 kbp までであるが、Broad Institute では 40 kbp の fosmid-ends ライブラリを用いている。

paired-end の coverage depth については Broad Institute の方が少なく（61）、アセンブリ結果にも'N'の割合が多い（23.7 %）という特徴がある。Platanus の値は 4.5 % である。2 つのドラフトゲノムの優劣を決めることはできないが、同じ生物種でも解読計画の違いにより、アセンブリ結果が異なる例と言える。

第3章　原核生物の全ゲノム、環境ドラフトゲノム配列の構築

3.1　背景・目的

序論で触れたように、1980–90 年代にかけては、大腸菌の物理地図決定（Kohara et al. 1987）から全ゲノム配列発表（Blattner et al. 1997）まで 10 年を要していたが、原核生物のゲノム決定もハイスループットシークエンサの利用により効率的に行われることが期待される。バクテリアのゲノムサイズは 10 Mbp 以下と小さいため、シークエンサの 1 回の運転で十分な coverage depth を持つデータを得る事ができる。真核生物と比較してリピート配列の割合も小さいため、*de novo* アセンブリの工程を自動化できれば、シークエンサの運転から全ゲノム決定まで一ヶ月未満で行うことができる。しかしながら、リピート配列が少ないとはいえ、リボソーム RNA オペロンなどは 5 kbp 以上の長さでゲノム中に複数コピーが存在し、paired-end ライブラリのみでは解決が困難である。バクテリアの Illumina データに対して複数の *de novo* アセンブラのベンチマーク（"GAGE-B"）が行われたが（Magoc et al. 2013）その際は paired-end のみを用いていたため、どのアセンブラも完全ゲノムを構築することはできていない。本章の前半では、Platanus を含めた複数のアセンブラに mate-pair（インサートサイズ　4–12 kbp）を含むデータを入力してベンチマークを行い、バクテリア完全ゲノム構築が可能な条件を調べている。

バクテリアのゲノム決定のためには、シークエンス前に培養のプロセスが必要であるという問題も生じる。難培養性バクテリアの種類の方が自然界には多いと考えられるため、培養を経ないで全ゲノムを決定する方法は新種のゲノム決定を効率化する上で重要となる。方法の 1 つとしては、1 細胞の DNA を Multiple displacement amplification（MDA）等で増幅してシークエンスを行うものがあり、共生菌の完全ゲノム決定（Hongoh et al. 2008）に成功した例も報告された。しかし、MDA で得られたデータの欠点としては、ゲノム上の領域毎に増幅効率が異なりシークエンスデータの coverage depth のばらつきが大きいことや、キメラリードの割合が比較的多いということが挙げられる（Chitsaz et al. 2011）。また、1 細胞 DNA サンプルから mate-pair ライブラリを構築した例はないと考えられ、リピート領域の解決が難しいという問題点もある。難培養性バクテリアのゲノム配列を得るための別の手段としては、環境 DNA サンプルを直接シークエンスし、メタゲノムデータとして *de novo* アセンブリを行う方法が存在する。メタゲ

ノム解析はハイスループットシークエンサが広まる以前からサンガー法により行われていたが（Kurokawa et al. 2007）、サンガー法リードでは各菌種の全ゲノム配列を構築するのに十分な coverage depth を得ることが困難であった。しかし、ハイスループットシークエンサが活用されたメタゲノム解析計画ではドラフトゲノムの構築まで行われるケースが存在する（Hess et al. 2011; Albertsen et al. 2013; Sharon et al. 2013; Nielsen et al. 2014）。また、適用例は少ないものの mate-pair を活用してアセンブリを行なった研究もある（Hess et al. 2011）。本章の後半では、目標として環境 DNA サンプルから完全ゲノムに近いドラフトゲノムを得ることを挙げて、Platanus を基に開発したメタゲノムデータ用アセンブラ MetaPlatanus のアルゴリズムと性能評価の結果を示す。

3.2　原核生物データに対する Platanus の有用性の検証

本節では、真核生物データ用に設計された Platanus が原核生物データに応用可能であるかを検証するため、バクテリア（大腸菌）の実データでベンチマークを行なった結果を示す。

3.2.1　大腸菌によるベンチマークと考察

・*E. coli* 2 株のデータの取得および特徴

大腸菌 *E. coli* の K-12 MG1655 株（以下 MG1655）と O157 Sakai 株（以下 O157）の DNA サンプル（いずれも宮崎大学 林教授提供）を Illumina MiSeq でシークエンスしたデータを用いて、バクテリアゲノムについてのベンチマークを行なった。生データの合計サイズはゲノムサイズの 500 倍以上とサイズが大きいため、paired-end のデータサイズがゲノムサイズの 100 倍となるケースを仮定し、リードをランダムに抽出してサイズを縮小した。ベンチマークに用いたデータのライブラリ構成を表 3-1 に示す。リードの前処理を施した後の値なので、paired-end のサイズはゲノムサイズ 100 倍より小さい値となっている。比較のため、1 分子 DNA シークエンサである PacBio RS のデータのアセンブリと評価も行なっており、そのデータの値も記している。PacBio RS のシークエンスデータおよびアセンブル結果は国立遺伝学研究所 豊田敦准教授より提供頂いた。

表 **3-1** ***E. coli* 2** 株のシークエンスデータ

(A) *E. coli* K-12 MG1655 （ゲノムサイズ 4.64 Mbp）

シークエンサ	Illumina MiSeq				PacBio RS
ライブラリ	paired-end	mate-pair	mate-pair	mate-pair	
インサートサイズ **(bp)**	650	4,000	8,000	12,000	
リード長 **(bp)**	263.2	71.0	71.1	71.0	3,973
合計長 **(bp)**	394.1 M	87.4 M	86.9 M	87.4 M	649.4 M

(B) *E. coli* O157 Sakai （ゲノムサイズ 5.59 Mbp）

シークエンサ	Illumina MiSeq				PacBio RS
ライブラリ	paired-end	mate-pair	mate-pair	mate-pair	
インサートサイズ **(bp)**	650	4,000	8,000	12,000	
リード長 **(bp)**	264.6	70.9	71.3	70.6	7,497
合計長 **(bp)**	479.1 M	103.6 M	106.8 M	101.5 M	774.6 M

Illumna MiSeq データの値は前処理（アダプタ配列、低クオリティ領域除去）後の値。

・ベンチマーク結果および考察

ベンチマーク対象としたアセンブラは、Platanus に加えて第 2 章で用いた MaSuRCA、SOAPdenovo2、Velvet である。ALLPATHS-LG は実行時に短いインサートサイズ（< 2×リード長）の paired-end ライブラリを必要とし、今回はそのようなライブラリを欠くため対象外とした。PacBio データは国立遺伝学研究所豊田敦准教授により、HGAP3（Chin et al 2013）を用いてアセンブリを行なって頂いた。精度評価は、完成リファレンスゲノム配列と評価ツールである QUAST（Gurevich et al. 2013）を用いて行なった。QUAST はリファレンスゲノムが環状であることを考慮して、GAGE と同様にミスアセンブリ数等を報告する。その際、アライメント領域の NG50 である"NGA50"という値も算出する。この評価指標は GAGE の"corrected NG50"に対応し、長さと精度を統合した評価指標である。scaffold によりカバーされたゲノム配列の割合は、QUAST が報告する値を用いる（"Covered rate"）。scaffold 配列上の遺伝子数はその後の解析で重要となるため、ミスマッチ、ギャップ無しで全長が構築された ORF 数も調べた（"#Complete ORF"）。正解となる ORF 配列は、アノテーション方法を統一するため、リファレンスゲノム上に予測ツール MetaGeneAnnotator（Noguchi et al. 2008）でアノテートされた完全 ORF 配列とする。

MG1655 のベンチマーク結果を表 3-2 に示す。

表 3-2　*E. coli* K-12 MG1655 データのアセンブリ結果

(A) 合計長・Scaffold 数・NG50

	合計長 (≥1 kbp)	Scaffold数 (≥1 kbp)	Contig NG50 (bp)	Scaffold NG50 (bp)
Platanus	4,643,205	2	4,637,996	4,637,996
MaSuRCA	5,655,655	44	505,082	701,670
SOAPdenovo2	4,674,534	3	3,763,305	4,667,984
Velvet	4,653,183	14	403,377	4,628,977
HGAP3 (PacBio)	4,652,257	1	4,652,257	4,652,257

(B) NGA50

	Contig NGA50 (bp)	Scaffold NGA50 (bp)
Platanus	4,546,375	4,546,375
MaSuRCA	504,935	578,477
SOAPdenovo2	2,649,161	2,649,161
Velvet	292,346	2,072,353
HGAP3 (PacBio)	4,536,261	4,536,261

(C) アセンブリエラーに関する情報

	#Misassemblies	#Local-misassemblies	#Mismatches /100kbp	#Indels/100kbp
Platanus	2	5	0.19	0.11
MaSuRCA	15	36	3.51	0.82
SOAPdenovo2	11	50	6.75	0.93
Velvet	6	44	10.67	1.80
HGAP3 (PacBio)	5	2	0.26	0.73

"#Local-misassemblies"は 1 kbp 以下の規模のミスアセンブリ

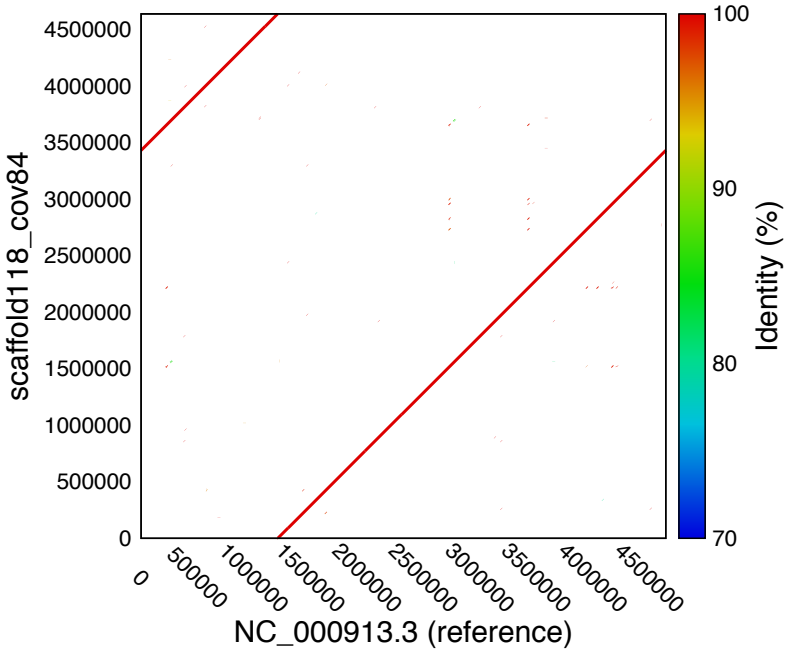
(D) ゲノムの再現度に関する情報

	Covered rate (%)	#Complete ORF
Platanus	99.94	4,314
MaSuRCA	99.92	4,304
SOAPdenovo2	99.89	4,282
Velvet	99.50	4,251
HGAP3 (PacBio)	99.88	4,311

全 ORF 数は 4,319

Platanus の contig NG50 はゲノムサイズの 99.92%であり、これはゲノムのほぼ全長が 1 本の contig として構築されたことを意味する。SOAPdenovo2、Velvet の scaffold NG50 もゲノムサイズの 99%以上となっているが、アセンブリのエラー数を示す指標（#Misassemblies、#Local-misassemblies、#Mismatches/100kbp、#Indels/100kbp）は全て Platanus の値より悪化しており、精度は Platanus が優れていることが分かる。全体的に Platanus に近い値を示すアセンブラは HGAP3（PacBio）である。精度の情報を含む指標（表 3-2 BCD）は、#Local-misassemblies 以外は Platanus が優っているが、各値の差は小さく、条件次第で大小関係は逆転する可能性があることは注意して評価する必要がある。Platanus、HGAP3 の最長 contig とリファレンスゲノムとのドットプロットを図 3-1 に示す。QUAST によりミスアセンブリは報告されるものの、この図では大規模な構造の変化は両方で認められず、局所的なミスアセンブリのみが存在することが分かる。

(A) Platanus（MG1655）



(B) HGAP3（PacBio）（MG1655）

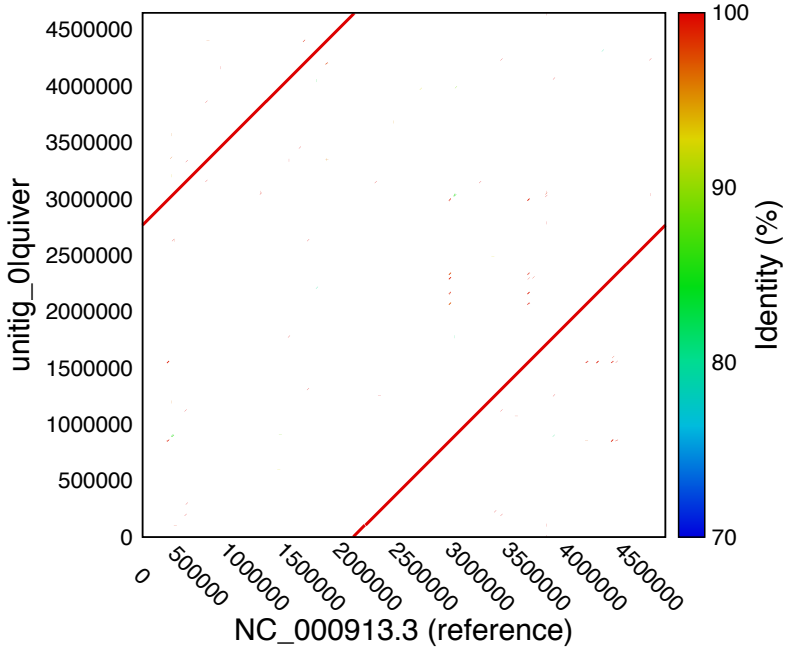


図 3-1　*E. coli* K-12 MG1655 のアセンブリ結果とリファレンスゲノム配列の比較

縦軸は scaffold、横軸はリファレンスゲノム上のポジション。

O157 のゲノムは、*E. coli* の株間で保存されている基本骨格の配列にファージ由来等の外来配列が多く挿入されることにより、比較的大きいサイズを持つに至っている（Hayashi et al. 2001）。外来配列には互いに相同性が高い組もあり、*de novo* アセンブリの障害となることが予想される。O157 のベンチマーク結果を表 3-3 に示す。他の Illumina データ用アセンブラと比較して、Platanus の NG50、NGA50 は大きく、エラー数が少ないという傾向は MG1655 と同様である。構築された完全 ORF についても Platanus の方が多い。しかし、HGAP3 と Platanus の結果を比べると、精度情報を含んだ指標（表 3-3 BCD）で Platanus が優っているのは Scaffold NGA50、#Misassemblies、#Indels/100kbp、のみであり、8 項目中 3 項目である。Platanus と HGAP3 の配列（≥10 kbp）と、O157 の染色体配列とのドットプロットを図 3-2 に示す。ゲノムのほぼ全長が 1 本の contig になっていた MG1655 のケースに対して、今回は複数配列に分断されていることが分かる。また、HGAP3 の結果ではリファレンスの 2 Mbp 付近の位置で比較的大規模なミスアセンブリが起こっている。

表 3-3　*E. coli* O157 Sakai データのアセンブリ結果

(A) 合計長・Scaffold 数・NG50

	合計長 (≥1 kbp)	Scaffold数 (≥1 kbp)	Contig NG50 (bp)	Scaffold NG50 (bp)
Platanus	5,539,690	16	2,898,084	4,580,494
MaSuRCA	5,778,443	122	184,383	877,026
SOAPdenovo2	5,650,247	26	433,248	3,734,657
Velvet	5,613,781	38	181,865	4,483,701
HGAP3 (PacBio)	5,679,991	5	5,345,079	5,345,079

(B) NGA50

	Contig NGA50 (bp)	Scaffold NGA50 (bp)
Platanus	2,898,080	3,987,011
MaSuRCA	184,383	331,744
SOAPdenovo2	270,946	275,944
Velvet	181,865	993,232

(C) アセンブリエラーに関する情報

	#Misassemblies	#Local-misassemblies	#Mismatches /100kbp	#Indels/100kbp
Platanus	2	20	1.13	0.25
MaSuRCA	55	116	21.63	2.40
SOAPdenovo2	52	49	27.74	2.10
Velvet	22	94	9.59	1.39
HGAP3 (PacBio)	6	0	0.13	1.22

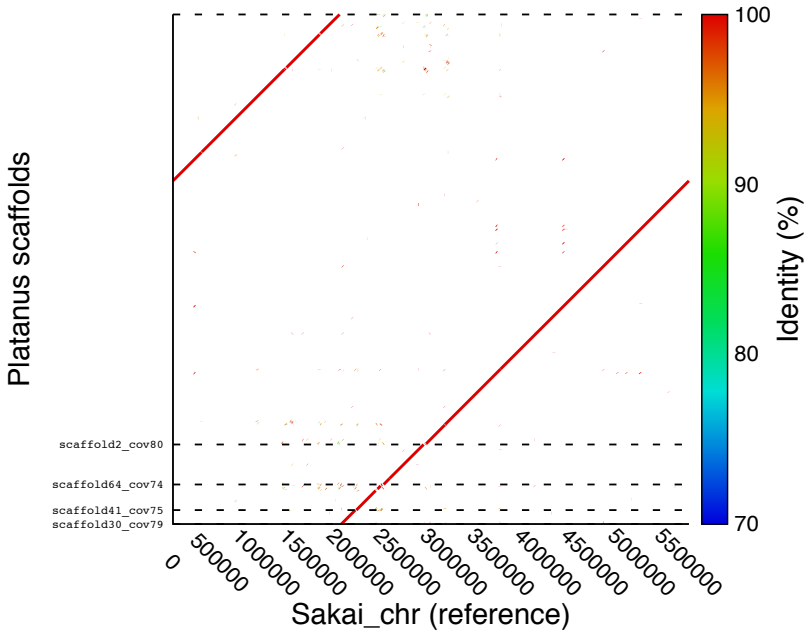
"#Local-misassemblies"は 1 kbp 以下の規模のミスアセンブリ

(D) ゲノムの再現度に関する情報

	Covered rate (%)	#Complete ORF
Platanus	98.23	5,230
MaSuRCA	96.69	5,073
SOAPdenovo2	98.80	5,139
Velvet	98.06	5,078
HGAP3 (PacBio)	99.94	5,327

全 ORF 数は 5,354

(A) Platanus（O157）



(B) HGAP3（PacBio）（O157）

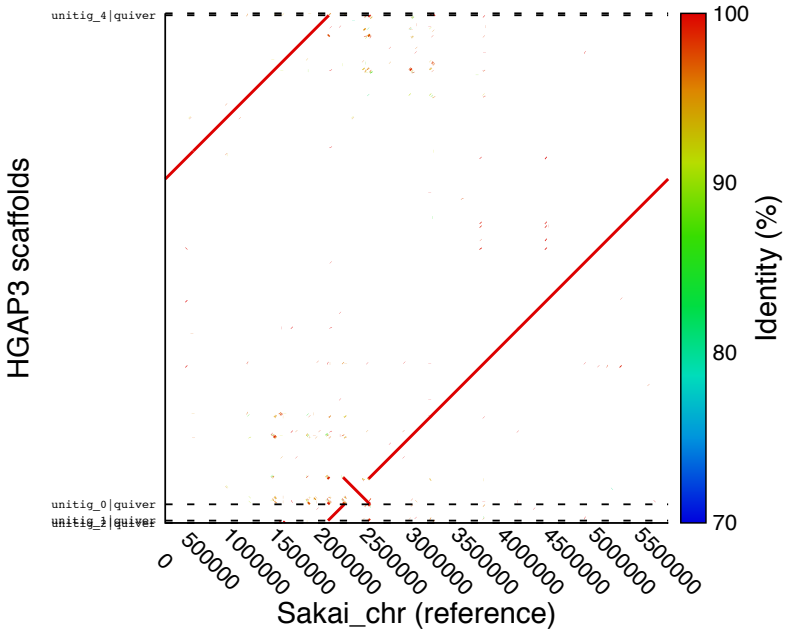


図 3-2　*E. coli* O157 のアセンブリ結果とリファレンス染色体配列の比較
縦軸は scaffold、横軸はリファレンスゲノム上のポジション。横方向の点線は scaffold 端を表す。scaffold については、10 kbp 未満またはプラスミドに対応するものは除いた。

3.2.2　大腸菌の完全ゲノム構築の条件検討

最初に、paired-end のサンプル調整方法を変更した場合の Platanus のアセンブリ結果を比較した。前節で用いた paired-end は Illumina TruSeq DNA Sample Prep Kit（以下 TruSeq）で調整したものである。TruSeq より安価で調整時間が短く、DNA 量が少ない場合もシークエンス可能な Illumina Nextera XT DNA Sample Prep Kit（以下 Nextera XT）という調整キットも存在するが、シークエンスのエラー率がより高くなるという可能性も存在する。ここでは、データサイズは統一し、paired-end ライブラリのサンプル調整方法のみを変えて Platanus のアセンブリをそれぞれ行なった（表 3-4）。mate-pair ライブラリは両方のケースで共通である。評価指標は、長さと精度の情報を含んだ contig NGA50、scaffold NGA50、#Complete ORF を選んでいる。MG1655 については、contig NGA50 と scaffold NGA50 が Nextera XT を用いた場合に半減しているが、これは最長の配列にミスアセンブリが検出されたことを示している。O157 の NGA50 も減少するが、その差は小さい。ただし、今回のように最長の配列がゲノムサイズの 50%を超えるような場合では、1 箇所のミスアセンブリにより NGA50 の値は大きくばらつくことに注意して結果を解釈する必要がある。#Complete ORF についてもやはり TruSeq の方が大きい値となるが、差は 10–27 個と小さい。これらの結果から、完全ゲノム配列の構築を目指す場合は精度の良い TruSeq が好ましいが、Nextera XT でもゲノムサイズの 50%以上をカバーする 1 本の contig を構築し、全 ORF 配列の 99%以上を得る程の結果を達成できることが示唆された。

表 3-4　Paired-end サンプル調整方法についての Platanus アセンブリ結果の比較

(A) MG1655

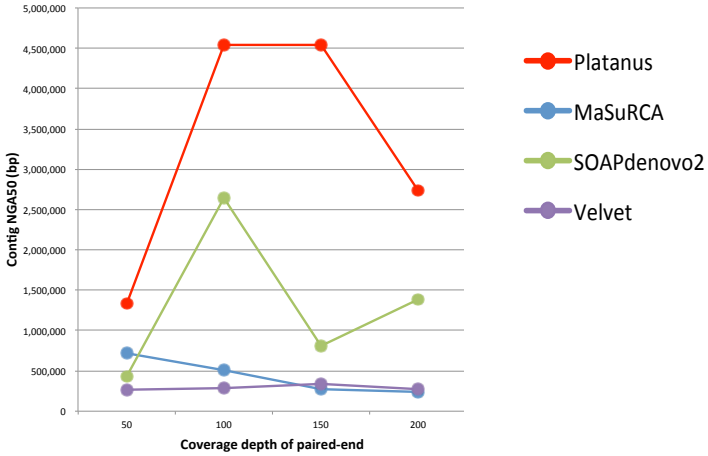
	Contig NGA50 (bp)	Scaffold NGA50 (bp)	#Complete ORF
TruSeq	4,637,996	4,637,996	4,314
Nextera XT	2,459,671	2,728,106	4,304

(B) O157

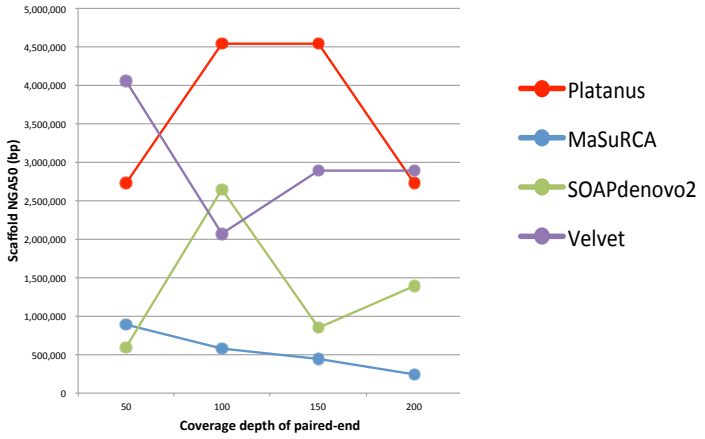
	Contig NGA50 (bp)	Scaffold NGA50 (bp)	#Complete ORF
TruSeq	2,898,080	3,987,011	5,230
Nextera XT	2,894,009	3,773,498	5,203

次に、データ量を 0.5×、1.5×、2.0×の倍率で変化させ、同様にアセンブリ結果を比較した（図 3-3）。それぞれ paired-end の coverage depth が 50、150、200 の場合に相当する。Platanus については、coverage depth が 50 から 100 へ増加した際に全ての指標が改善しているが、150 以上に増加させた場合は顕著な改善は見られない。他のアセンブラは coverage depth を 100 以上にしても明確な改善傾向は認められない。この結果から、Platanus によるアセンブリでは paired-end の coverage depth は 100 以上が望ましいことが示された。

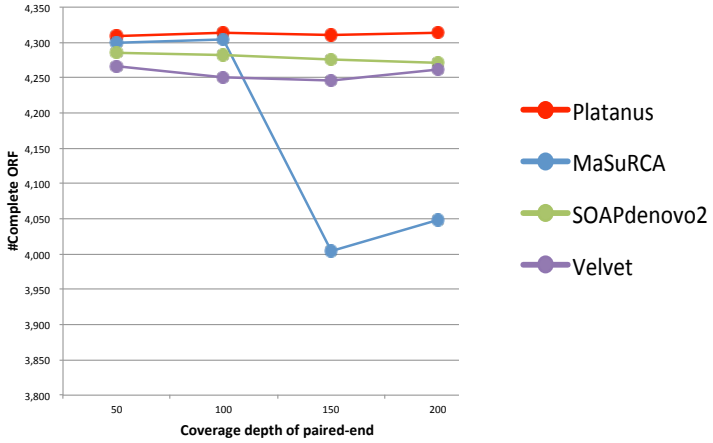
(A) Contig NGA50（MG1655）



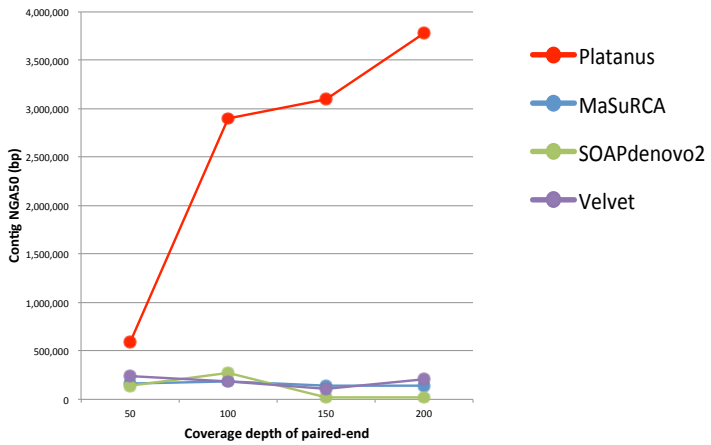
(B) Scaffold NGA50（MG1655）



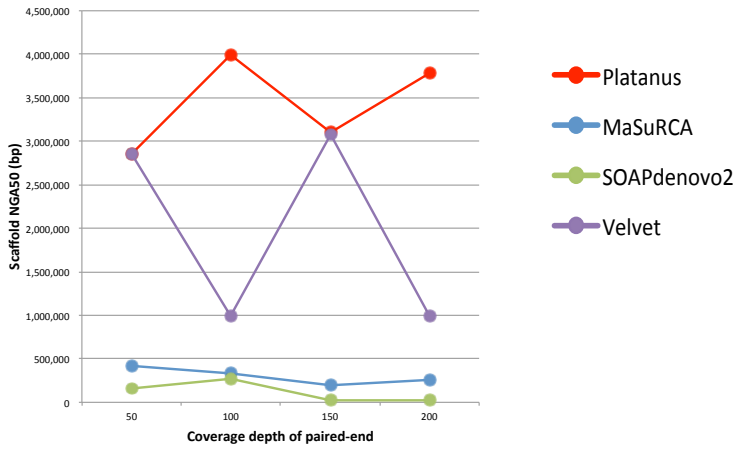
(C) #Complete ORF（MG1655）



(D) Contig NGA50（O157）



(E) Scaffold NGA50（O157）



(F) #Complete ORF（O157）

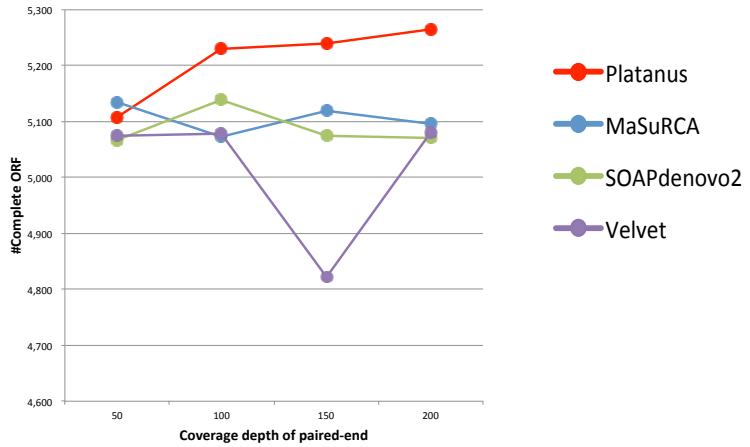


図 3-3　Coverage depth についての Platanus アセンブリ結果の比較

最後に、ライブラリ構成を変更した際の結果を比較した（表 3-5）。scaffold NGA50 はインサートサイズ 8 kbp の mate-pair を加えると大きく増加する。これについては、インサートサイズがゲノム中に複数存在するリボソーム RNA オペロンの長さ（5–6 kbp）を超えるかどうかが重要な要因となっていると考えられる。contig NGA50 に関しても、入力するライブラリ数が増えるほど値が増加しているが、特に mate-pair が 1 つの場合（"PE, MP4k"、"PE, MP8k"）と 2 つの場合（"PE, MP4k, MP8k"）で大きく改善が見られる。4 kbp-mate-pair はリボソーム RNA オペロンを解決できないが、ギャップを減らし contig NGA50 を増加させることには寄与していることが分かる。インサートサイズ 12 kbp の mate-pair については、MG1655 では加えても指標の改善が見られず、8kbp-mate-pair までで十分であることが示されているが、リピート配列の多い O157 では改善が見られる。ただし、その差は比較的小さい。

表 3-5　ライブラリ構成についての Platanus アセンブリ結果の比較

(A) MG1655

	Contig NGA50 (bp)	Scaffold NGA50 (bp)	#Complete ORF
PE	176,239	176,239	4,249
PE, MP4k	977,952	977,952	4,307
PE, MP8k	1,807,406	4,540,635	4,314
PE, MP4k, MP8k	4,546,375	4,546,375	4,314
PE, MP4k, MP8k, MP12k	4,546,375	4,546,375	4,314

(B) O157

	Contig NGA50 (bp)	Scaffold NGA50 (bp)	#Complete ORF
PE	148,315	160,636	5,051
PE, MP4k	297,224	538,926	5,173
PE, MP8k	541,385	3,735,793	5,164
PE, MP4k, MP8k	2,892,505	3,775,755	5,211
PE, MP4k, MP8k, MP12k	2,898,080	3,987,011	5,230

PE は paired-end、MP は mate-pair を表す。

3.3 メタゲノム用 *de novo* アセンブラ MetaPlatanus の開発

3.3.1 MetaPlatanus のアルゴリズムの概要

MetaPlatanus（バージョン 1.0.1）は、Platanus を基にメタゲノムデータに対応するよう開発された *de novo* アセンブラである（図 3-4）。Contig-assembly、Scaffolding、Gap-close の各モジュールは Platanus から継承されているが、Scaffolding と Gap-close の反復を行う機能、Scaffolding において coverage depth と 4-mer 頻度情報を基に contig 組が同種由来か判定する機能、ダイコドン（2 連コドン）頻度情報を用いて scaffold 配列のクラスタリングを行う機能が追加されている。これらはそれぞれ、生物種毎の coverage detph のばらつきへの対応、異種間のミスアセンブリの防止、scaffold 配列を由来する生物種毎に分類することを目的としている。最終結果は scaffold 配列のクラスタで、クラスタと生物種が 1 対 1 対応することを理想としている。

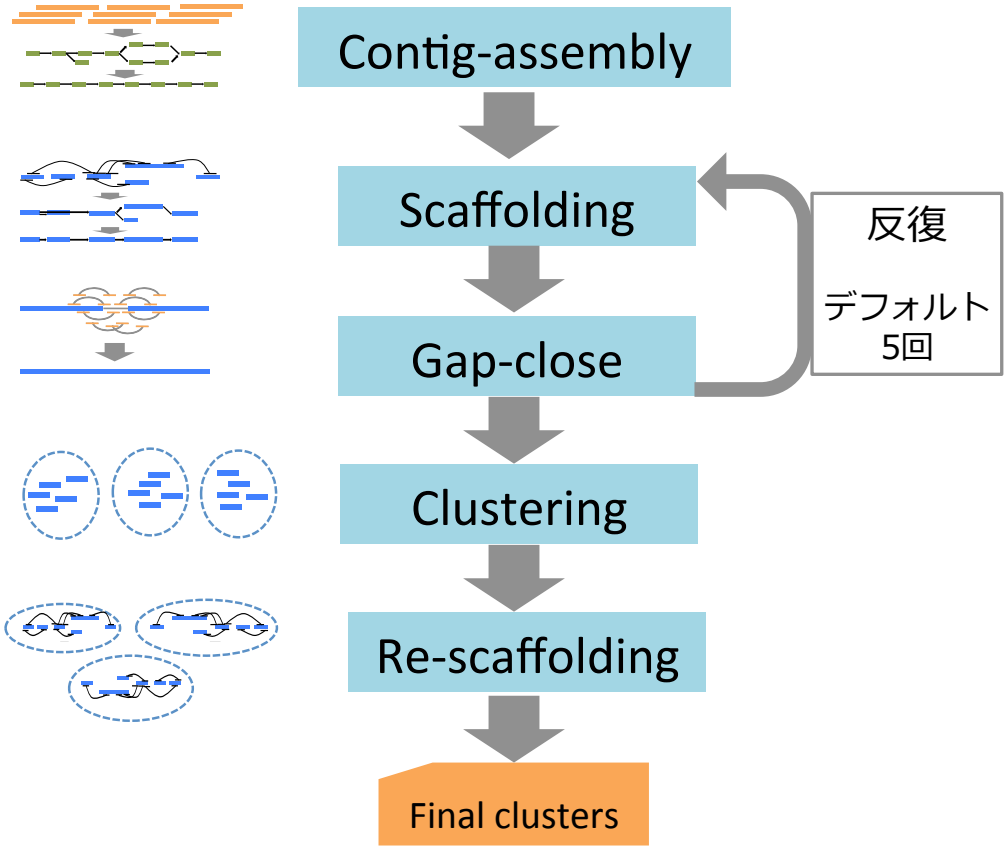


図 3-4　MetaPlatanus の全体像

3.3.2　MetaPlatanus の Contig-assembly のアルゴリズム

メタゲノムデータは複数種のゲノムを含み、サンプル中の組成を反映してそれぞれ異なる coverage depth を持つ。coverage depth のばらつきに対応するため、Contig-assembly を行う際の *k*-mer 長と coverage depth の下限（coverage cutoff）の決定方法を変更している。具体的な方法は以下の通りである。

・*k*

k_0: 定数（デフォルト　25）、k_{max}: 平均リード長×0.9

Platanus では k_0: 32 とし、k_{max} は coverage depth が均一であると仮定して算出していたが、その仮定は適切でないため変更を施した。k_0 から k_{max} まで小さい値から順に *k* の値を適用する。用いる *k* の数（ステップ数）はデフォルトで 3 である。

・coverage cutoff

各 *k* の値に対し 2 種類の定数の coverage cutoff 値（c_{lower}, c_{upper}）を適用する。デフォルトでは c_{lower}: 3、c_{upper}: 6 である。

MetaPlatanus の Contig-assembly の模式図を図 3-5 に示す。各 *k* 値のステップでは、最初に coverage cutoff: c_{upper}（初回のみは c_{lower}）で de Bruijn グラフを構築後、straight node の端の *k*-mer を含むリードを収集する。続いて、coverage cutoff を c_{lower} とし、前のグラフの straight node と収集したリードから de Bruijn グラフを再構築する。グラフを再構築後、長さが短く coverage depth の低い straight node（長さ < 2×平均リード長、coverage depth < c_{upper}）を除去する。次に、Platanus と同様に *k* を増加させ、coverage cut を c_{uppper} てグラフを再構築し、同じ手順を繰り返す。なお、バブル構造の除去は 2 倍体サンプルを想定した機能であるため、MetaPlatanus ではその機能は無効化している。

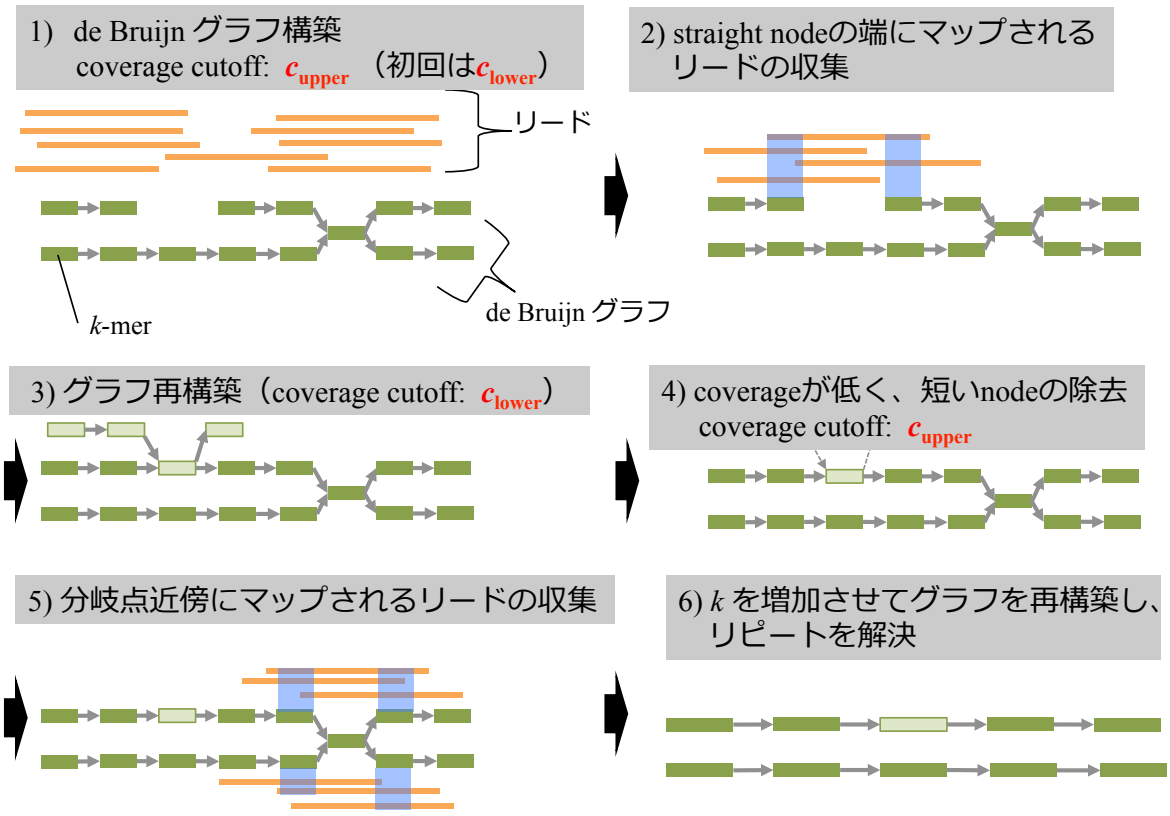


図 3-5 MetaPlatanus の Contig-assembly の模式図

3.3.3　MetaPlatanus の Scaffolding のアルゴリズム

異なる生物種由来の配列を誤って接続した場合、後の解析に重大な悪影響を及ぼすと考えられるため、そのようなミスアセンブリを防ぐ機能を Scaffolding に追加している。用いる情報は contig の coverage depth と 4-mer 頻度である。coverage depth の比率、各 4-mer の頻度が大きく異なる contig の組は異種由来であると判定されリンクされない。具体的な手順は次の通りである。

(1)　同種由来配列間の coverage depth 比率分布の推定

1 kbp 以上の長さを持つ各 contig を 2 等分し、それぞれの領域で平均 coverage depth を算出する。coverage depth は *k*-mer（デフォルト 25-mer）の出現回数から推定する。同一 contig 上の領域間での coverage depth の比率（大きい値 / 小さい値）を集計しその分布を求める。ここで、その分布の相補累積分布関数（1 − 累積分布関数）を F_{coverage} とする。

(2)　同種由来配列間の 4-mer 頻度距離の分布の推定

4 文字{A, T, G, C}からなる 4-mer を s_i（$i = 1, 2, ..., 4^4$）、s_i の接頭辞の 3-mer を $s_i[1, 3]$とする。ある配列の部分文字列 s の出現回数を $n(s)$と表し、s_i の頻度を

$$f_i = \frac{n(s_i)}{n(s_i[1,3])}$$

とする。これは $s_i[1, 3]$が出現した時に s_i が出現する条件付き確率に対応する。ある配列の 4-mer 頻度はベクトル$(f_1, f_2, \ldots, f_{4^4})$と表し、2 つ頻度間の距離はベクトル間のユークリッド距離とする。

(1)と同様に 1 kbp 以上の contig を 2 等分し、各組で 4-mer 頻度距離を計算し、その分布を求め、相補累積分布関数を $F_{\text{4-mer}}$ とする。

(3)　scaffold グラフの構築（Platanus と共通）

(4)　異種間のリンク候補を削除

scaffold グラフの辺のそれぞれについて、接続している節点（contig）間で coverage depth の比率 r と 4-mer 頻度距離 d を計算する。次の条件が成り立つとき、辺を削除する。

$$F_{\mathrm{coverage}}(r) \times F_{4-\mathrm{mer}}(d) \leq 0.05$$

ここで、左辺は条件付き確率

P(coverage depth の比率 $> r$　かつ　4-mer 頻度距離 $>$ d | contig 組が同種由来)

に対応する。

3.3.4　Scaffolding、Gap-close の反復のアルゴリズム

scaffold（contig）の長さを伸ばすため、Scaffolding と Gap-close の反復を行う。反復の回数はデフォルトでは 5 回である。その際、Platanus とは異なり Gap-close で配列端の延長を行う。端が延長された scaffold 配列のセットはオーバーラップを含んでいるので、de Bruijn グラフを構築して contig 配列へ変換することでマージをする（図 3-6）。更に、contig 内のミスアセンブリ部分を修正するため、contig の端領域の coverage depth が配列内の中央値から大きく異なる場合、その部分を分断するという機能も追加されている。具体的な手順は次のようになる。

(1)　前段階の結果の contig の修正

contig 内の各 *k*-mer（デフォルト $k = 25$）の coverage depth から、各 contig で中央値を算出する。*k*-mer の coverage depth を $c_{k\text{-mer}}$、contig の coverage depth 中央値を c_{median} とし、

$$1.5 \times c_{k\text{-mer}} < c_{\text{median}} \quad \text{または} \quad 1.5 \times c_{\text{median}} < c_{k\text{-mer}}$$

が成り立つとき、その *k*-mer は異種由来の候補とする。contig 端から、異種由来の *k*-mer 候補がなくなるまで、contig 配列をトリムする。

(2)　Scaffolding

前節で述べたように、異種間のミスアセンブリを防ぐ機能を持った Scaffolding を実行する。

(3)　Gap-close

ギャップ部分の配列構築に加えて、scaffold 配列端の延長を行う。また、ギャップまたは scaffold 端の構築に用いられなかった contig も次の段階で用いるため保存される。

(4)　de Bruijn グラフを用いた contig 配列の再構築

延長された scaffold 配列の重複部分をマージするため、scaffold（contig）配列上の *k*-mer（デフォルト　*k* = 1.5×平均リード長）から de Bruijn グラフを構築する。ここで、(3)で新たに保存された contig 配列上の *k*-mer もグラフに取り込まれる。エラー由来の枝構造の除去を行なった後、straight node を contig 配列として次の反復の入力とする（(1)に戻る）。

最後の反復では、(4)は実行せず、さらに(3)で scaffold 端の延長を行わないで構築した scaffold 配列を出力とする。

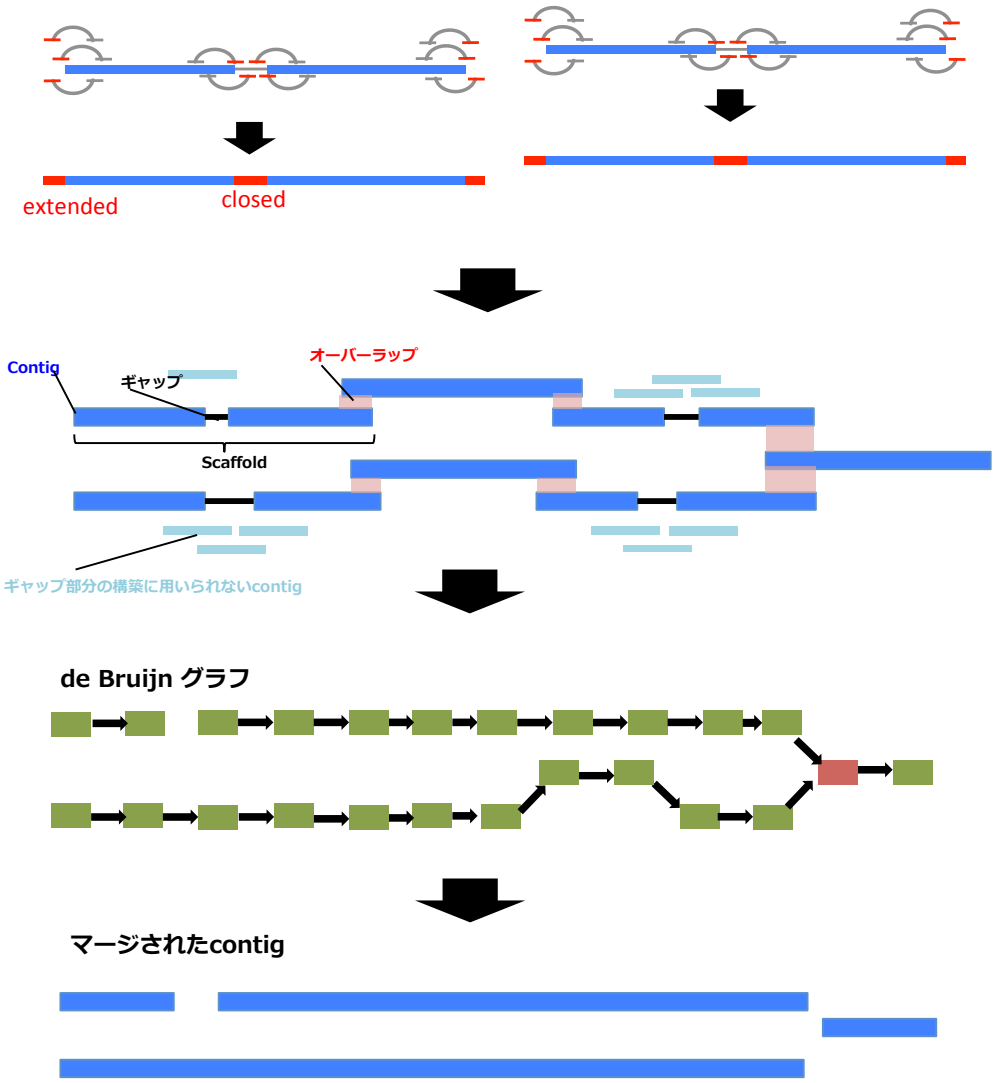


図 3-6　Gap-close 結果のマージの模式図

3.3.5　scaffold 配列のクラスタリングと re-scaffolding のアルゴリズム

国立遺伝学研究所 野口英樹准教授開発のクラスタリングプログラムMGC（未発表）のアルゴリズムを採用している。MGC は各 scaffold 配列の ORF 上のダイコドン（2 連コドン）頻度情報を基に、*k*-means 法によるクラスタリングを行う。クラスタリング後、同一クラスタに属する配列に再度 Scaffolding を実行する（re-scaffolding）。具体的な手順を以下に示す。

(1)　各 scaffold 配列上に ORF 位置を定める。

MetaGeneAnnotator（Noguchi et al. 2008）により予測される。

(2)　各 scaffold のダイコドン頻度を求める

ここでダイコドン頻度は、コドンの出現頻度を直前のコドン毎にまとめたものであり、条件付き確率に対応する。

(3)　scaffold セットを *k*-means 法でクラスタリング

クラスタリングの対象とする scaffold は、予測された ORF の合計長が 1,200 bp 以上のものとする。*k*-means 法を行う際、クラスタ中心と要素（scaffold）の距離は、次のように定義される確率ベースの類似度 S を用いる。$x = (x_1, x_2, \ldots, x_l)$ は l 個のコドンからなる ORF 配列、P はクラスタ中心 c におけるダイコドンまたはモノコドン頻度とするとき、

$$S(c, x) = \log_2 P(x_1) + \sum_{i=2}^{l} \log_2 P(x_i \mid x_{i-1})$$

k-means 法を実行する際には、クラスタ数を 1 から指定された値（デフォルト scaffold 合計長 / 3 Mbp）に至るまで 1 つずつ分割により増加させる。分割の対象となるクラスタは、クラスタ中心と要素間の類似度の総和が最大となるものとする。また、2 分割のための代表属性の決定には FastMap（Faloutos et al. 1995）と呼ばれる高速な次元圧縮アルゴリズムを用いる。

(4)　Re-scaffolding

ORF の合計長が 1,200 bp 以上の scaffold はクラスタに属し、クラスタ番号を持つ。そうでない scaffold は"unclassified"として扱われ、クラスタ番号を持た

ない。Scaffolding を再実行する際、同一クラスタ番号を持つ scaffold 間、クラスタ番号を持つ scaffold と unclassified の scaffold 間のリンクは有効とし、別のクラスタ番号を持つ scaffold の組のリンクは無効とする。また、unclassified 配列を介して別のクラスタ番号を持つ scaffold の組が接続されることも無効となる。これらのリンクの制限により、異種由来の配列が接続されることを防ぎつつ配列が延長されることが期待される。

3.4　メタゲノムデータに対する MetaPlatanus の有用性の検証

本節では、前節で説明したアルゴリズムに基づいて開発された MetaPlatanus の性能評価をするため、複数のバクテリア DNA を混合して作成した仮想メタゲノムデータおよび実データでベンチマークを行なった。

3.4.1　仮想メタゲノムデータによるベンチマークと考察

・仮想メタゲノムデータの取得および特徴

リファレンスゲノムが登録されている 20 種のバクテリアのゲノム DNA を混合し、仮想的な環境メタゲノムサンプルとした。DNA サンプルは宮崎大学 林哲也教授、香川大学 桑原知己教授に提供して頂いた。生物種名とリファレンスゲノムの構築状況を表 3-6 に示す。含まれるバクテリアは主にヒト腸内に生息する種であり、11 種は完成ゲノム、9 種はドラフトゲノムが公開されている。ゲノムの GC 含量は 28–67%の範囲に渡っている。また、10 種が *Bacteroides* と *Parabacteroides* 属に、7 種が *Clostridium* 属に含まれ、同属中では互いにゲノム配列の相同性が比較的高くなっている。系統的に近い種の配列間では相同領域を介してミスアセンブリが起こる危険性が存在し、今回はそのような *de novo* アセンブリの障害も再現するよう種を選んでいる。

表 3-6　20 種のバクテリアサンプルの詳細

生物種	株		リファレンスゲノム			
			Status	サイズ（bp）	配列数	**N50 (bp)**
Bacteroides caccae	ATCC43185	JCM9498	Draft	4,564,814	21	500,031
Parabacteroides distasonis	ATCC8503		Finished	4,811,379	1	4,811,379
Bacteroides eggerthii	ATCC27754	DSM20697	Draft	4,197,635	20	1,194,706
Bacteroides fragilis	YCH46		Finished	5,310,990	2	5,277,274
Parabacteroides merdae	ATCC43184	JCM9497	Draft	4,434,377	93	334,494
Bacteroides ovatus	ATCC8483		Draft	6,465,369	32	507,553
Bacteroides stercoris	ATCC43183	JCM9496	Draft	4,009,829	17	476,026
Bacteroides thetaiotaomicron	VPI-5482		Finished	6,293,399	2	6,260,361
Bacteroides uniformis	ATCC8492		Draft	4,719,097	33	287,799
Bacteroides vulgatus	ATCC8482		Finished	5,163,189	1	5,163,189
Clostridium acetobutylicum	ATCC824		Finished	4,132,880	2	3,940,880
Clostridium cellulolyticum	ATCC35319	H10	Finished	4,068,724	1	4,068,724
Clostridium difficile	630		Finished	4,298,133	2	4,290,252
Clostridium hylemonae	DSM15053	JCM10539	Draft	3,889,859	167	2,898,417
Clostridium pasteurianum	ATCC6013	DSM525	Draft	4,420,100	12	859,467
Clostridium perfringens	ATCC13124		Finished	3,256,683	1	3,256,683
Clostridium ramosum	DSM1402	JCM1298	Draft	3,235,195	12	512,969
Escherichia coli	K12	MG1655	Finished	4,641,652	1	4,641,652
Pseudomonas aeruginosa	PAO1		Finished	6,264,404	1	6,264,404
Serratia marcescens	Db11		Finished	5,113,802	1	5,113,802

"Status"列はリファレンスゲノムが完成（Finished）かドラフト（Draft）かを表す。

混合する DNA 量の比が異なる 3 ケース（Case 1–Case3）のサンプルがそれぞれを Illumina MiSeq でシークエンスされ、ベンチマーク用データとして使用された。ライブラリ構成は共通で、1 種類の paired-end（インサートサイズ 550 bp）と 2 種類の mate-pair（インサートサイズ　4 kbp、8 kbp）をそれぞれのケースで用意している。

リード長とデータサイズを表 3-7 に示す。前処理としては、全てのライブラリでアダプタ配列と低クオリティ領域の除去を行い、mate-pair に関しては FastUniq（Xu et al. 2012）を用いて PCR-duplicate の除去を施した。FastUniq はリードをリファレンスにマップすることなく PCR-duplicate を除くツールであり、今回の mate-pair は短いインサートサイズのペアの混入が比較的少なくそれらの除去が不要と考えられるため採用した。*de novo* アセンブリのベンチマークにおいて、前処理の段階でリファレンスあるいは scaffold 配列の情報を用いることを避けられるという利点がある。paired-end を Bowtie2（編集距離の上限 5）でリファレンスゲノムにマップして算出した sequence coverage depth（マップされたリー

ド合計長 / ゲノムサイズ）を図 3-7 に示す。sequence coverage depth のばらつきは Case 1 から Case 3 の順で大きくなる。Case 1 は *S. mar.*以外は 29–65 となっており、Case 2 は 41–651、Case 3 は 9–2811 の範囲を持つ。mate-pair を同様にリファレンスにマップして算出した physical coverage depth（マップされたペア数×平均インサートサイズ / ゲノムサイズ）を図 3-8 に示す。physical coverage depth は、両端の配列が得られている DNA 断片で各サイトが平均で何本カバーされているかを表している。菌種間の coverage depth の比率に関しては、paired-end と mate-pair は必ずしも一致しない。この原因については、特に mate-pair 用の長い DNA 断片を調整する際に、菌種毎に抽出効率が異なるからだと考えられる。菌種によっては 8kbp-mate-pair の physical coverage depth が 10 以下と低くなってしまうことがあるが（例　Case 1 の *C. ram.*: 0.50）、4 kbp と 8 kbp の mate-pair の両方で depth が 10 以下になることはない。つまり、2 つの mate-pair ライブラリを組み合わせれば、少なくともどちらかで 10×以上カバーされていることになり、複数ライブラリを用いることの有効性が示唆されている。

表 **3-7**　仮想メタゲノムデータ

(A) Case 1

インサートサイズ **(bp)**		550	4,000	8,000
ライブラリ		paired-end	mate-pair	mate-pair
リード長 **(bp)**	**Raw**	150.0	75.4	75.5
	Preprocessed	141.9	70.9	71.0
合計長 **(bp)**	**Raw**	5.1 G	532.8 M	284.4 M
	Preprocessed	4.8 G	359.5 M	197.6 M

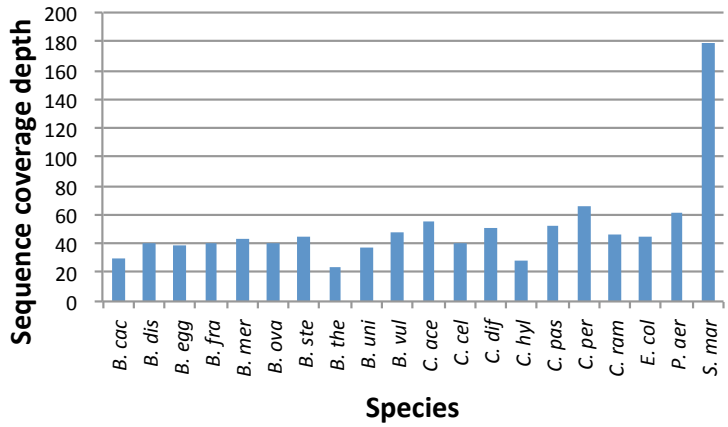
(B) Case 2

インサートサイズ **(bp)**		550	4,000	8,000
ライブラリ		paired-end	mate-pair	mate-pair
リード長 **(bp)**	**Raw**	150.0	75.4	75.4
	Preprocessed	143.8	70.4	70.7
合計長 **(bp)**	**Raw**	21.4 G	2.1 G	1.0 G
	Preprocessed	20.5 G	891.4 M	545.2 M

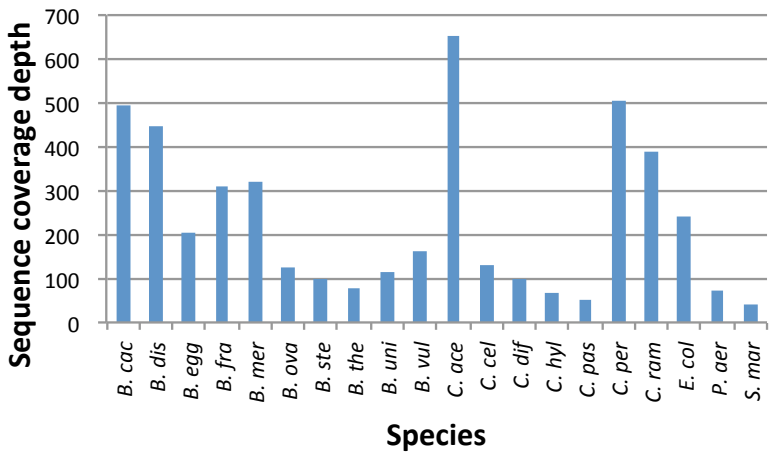
(C) Case 3

インサートサイズ **(bp)**		550	4,000	8,000
ライブラリ		paired-end	mate-pair	mate-pair
リード長 **(bp)**	**Raw**	150.0	75.4	75.4
	Preprocessed	142.7	70.4	70.2
合計長 **(bp)**	**Raw**	35.7 G	4.4 G	2.1 G
	Preprocessed	34.0 G	1.3 G	608.5 M

(A) Case 1



(B) Case 2



(C) Case 3

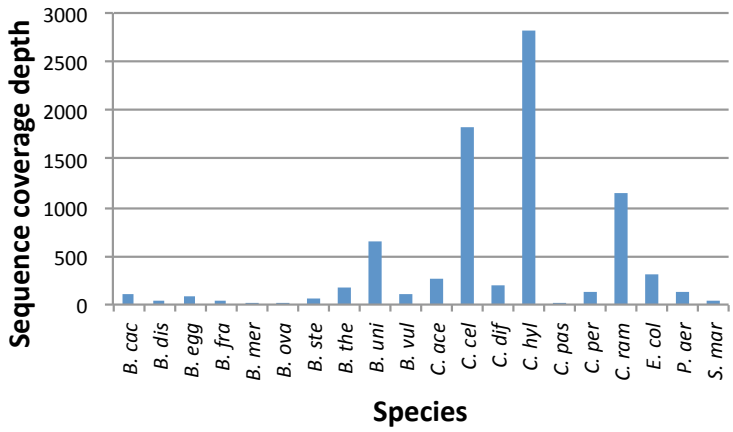
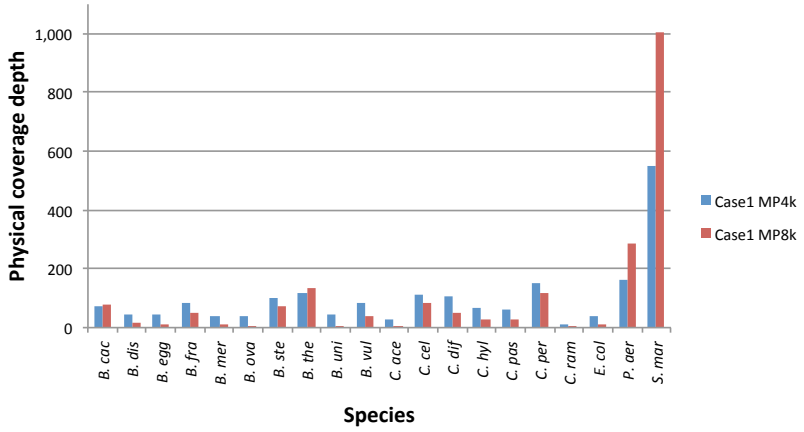
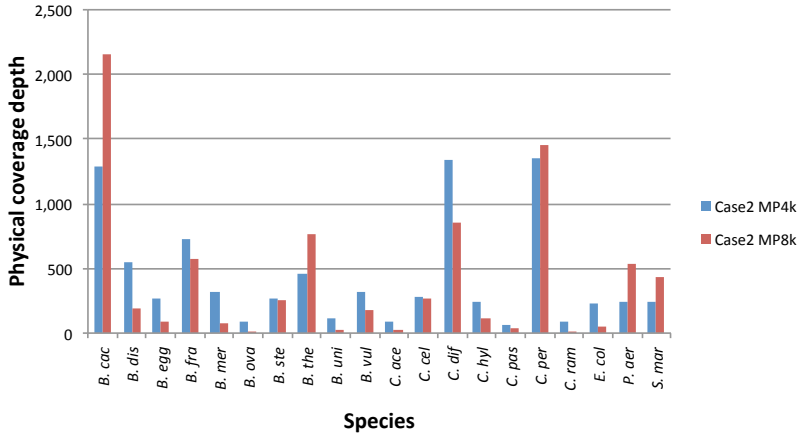


図 3-7　菌種毎の paired-end の sequence coverage depth

(A) Case 1



(B) Case 2



(C) Case 3

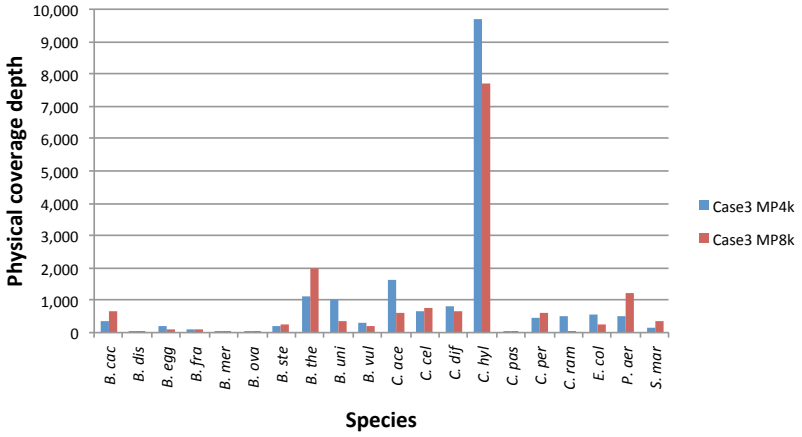


図 3-8　菌種毎の mate-pair（MP）の physical coverage depth

・ベンチマーク結果および考察

ベンチマーク対象は、いずれもメタゲノム用 *de novo* アセンブラの MetaPlatanus、IDBA_UD（Peng et al. 2012）、Ray_Meta（Boisvert et al. 2012）、MetaVelvet（Namiki et al. 2012）、Omega（Haider et al. 2014）である。バージョンとパラメータ調整方法を以下に示す。

・MetaPlatanus（version 1.0.1）
デフォルト設定。

・IDBA_UD（version 1.1.1、de Bruijn グラフ、複数の *k*-mer 長を用いる）
デフォルト設定。

・Ray_Meta（version 2.3.0、de Bruijn グラフ）
k について 31–91 の範囲を 20 刻みで入力し、scaffold NG50 が最大となる値を採用。

・MetaVelvet（version 1.2.02、de Bruijn グラフ）
k について 31–91 の範囲を 20 刻みで入力し、scaffold NG50 が最大となる値を採用。coverage depth に関するパラメータは全て"auto"。

・Omega（version 1.0.2、overlap-layout-consensus に類似したアルゴリズム）
複数ライブラリを入力できないため paired-end のみを入力。また、オーバーラップ長（-l）は 40–100 を 20 刻みで入力し、scaffold NG50 が最大の結果を採用。

精度の評価にはリファレンスゲノムと評価ツールの QUAST を用いた。異種間のミスアセンブリ（"chimera"）は、QUAST の報告する"translocation"のうち異種間であることを示すものである。加えて、scaffold を 10 kbp のブロックに区切り独自に検出したものもカウントした。QUAST の報告するものは、リファレンスのドラフトゲノムでギャップになっている部分が scaffold 配列側で構築されているとき、その部分を他種のゲノムにアラインして誤検出する場合がある。また、リピート配列の一単位だけ scaffold 配列側で間違えている場合など、比較的小規模なミスアセンブリを含む。10 kbp のブロックを用いる場合は、Blastn でリファ

レンス配列にアラインしたとき、5 kbp 以上のアライメント長、98%以上の identity を示す 10 kbp ブロックを対象とし、同一 scaffold 内のブロックが複数種にアラインされるときミスアセンブリを認める。これは、解析に重大な影響を及ぼす大規模なミスアセンブリに対応しており、"#Chimera (10 kbp)"としてカウントした。この値が 0 になるようなアセンブリ結果が解析には望ましいと考えられる。全ゲノム配列の再現度の指標としては、大腸菌（*E. coli*）でのテストと同様に QUAST が報告するカバー率（"covered rate"）と、完全に（ミスマッチ、ギャップ無し）構築された ORF の割合を算出した（"complete ORF"）。ここで、正解の ORF は菌種間でアノテーション方法を統一するため、MetaGeneAnnotator で予測した ORF とした。

全菌種のリファレンスゲノムを用いたベンチマーク結果を表 3-8 に示す。リファレンスがドラフトゲノムである種も含むことにより、NGA50（アライメント領域長についての NG50）や QUAST の報告するミスアセンブリ数（"#Misassemblies"）は正しく算出されないため記していない。scaffold NG50 については Case 2 を除いて MetaPlatanus の値が最大となっており、Case 1, 3 では他のアセンブラの 2 倍以上の値となっている。Case 2 で最大の scaffold NG50 を記録したのは Ray_Meta であるが、異種間のミスアセンブリ数（"#Chimera"）を示す指標は MetaPlatanus を上回っており、精度では劣る可能性がある。アセンブリのエラー数（表 3-8 B）に関しては、規模の大きい異種間ミスアセンブリ数（"#Chimera (10 kbp)"）が全て 0 となっているのは MetaPlatanus だけであり、ミスアセンブリの対策が機能していることを示唆している。IDBA_UD に対しては "#Mismatch / 100 kbp"の値が全ケースで劣っているが、"Indels / 100 kbp"や "#Local-misassemblies"は逆に全ケースで優っているため、単純に精度の優劣を決めることはできない。ゲノム再現度の指標（表 3-8 C）では、6 項目中 3 項目で MetaPlatanus の値が最大となっており、残りの項目でも 2–3 位である。Case 3 で IDBA_UD に劣っているのは、paired-end の coverage depth が 10 以下の 2 種（*B. mer.* と *B. ova.*）の配列を MetaPlatanus が構築できていないからであると考えられる。

表 3-8　全ての菌種のアセンブリ結果

(A) 合計長・Scaffold 数・NG50

		合計長 (≥1 kbp)	Scaffold数 (≥1 kbp)	Contig NG50 (bp)	Scaffold NG50 (bp)
Case 1	MetaPlatanus	92,113,993	617	256,322	1,237,875
	IDBA_UD	91,065,334	716	137,837	435,318
	Ray_Meta	91,854,120	1,487	183,265	591,843
	MetaVelvet	82,570,471	3,068	58,922	59,162
	Omega	89,950,753	2,288	241,431	241,431
Case 2	MetaPlatanus	92,209,090	483	354,379	1,666,007
	IDBA_UD	91,231,500	890	141,588	616,402
	Ray_Meta	95,403,771	403	276,793	1,775,371
	MetaVelvet	88,331,270	5,783	38,576	38,833
	Omega	98,132,034	1,902	227,450	227,450
Case 3	MetaPlatanus	86,032,644	1,345	291,832	1,492,916
	IDBA_UD	90,398,721	2,781	108,952	525,936
	Ray_Meta	92,139,274	2,382	206,620	449,631
	MetaVelvet	79,171,340	3,977	40,240	40,274
	Omega	91,478,963	4,191	143,186	143,186

(B) アセンブリエラーに関する情報

		#Mismatches /100kbp	#Indels/100kbp	#Local-misassemblies	#Chimera (QUAST)	#Chimera (≥10 kbp)
Case 1	MetaPlatanus	3.90	1.07	1,006	20	0
	IDBA_UD	2.62	1.70	1,309	46	1
	Ray_Meta	7.03	1.20	943	36	6
	MetaVelvet	6.68	0.97	59	39	4
	Omega	11.70	2.83	1,049	103	15
Case 2	MetaPlatanus	3.41	1.00	477	32	0
	IDBA_UD	2.09	1.62	1,038	27	2
	Ray_Meta	5.60	1.30	393	35	5
	MetaVelvet	2.63	0.90	29	1	0
	Omega	17.44	2.32	938	149	14
Case 3	MetaPlatanus	3.72	1.08	789	25	0
	IDBA_UD	3.11	1.18	884	19	0
	Ray_Meta	11.23	1.82	2,791	30	2
	MetaVelvet	1.97	0.86	31	18	0
	Omega	22.87	4.77	13,892	179	9

(C) ゲノムの再現度に関する情報

		Covered rate (%)	Complete ORF (%)
Case 1	MetaPlatanus	96.90	96.18
	IDBA_UD	96.87	95.63
	Ray_Meta	95.88	93.42
	MetaVelvet	88.91	87.80
	Omega	96.36	92.70
Case 2	MetaPlatanus	97.58	97.47
	IDBA_UD	96.92	95.89
	Ray_Meta	97.88	97.15
	MetaVelvet	95.53	90.96
	Omega	98.02	92.89
Case 3	MetaPlatanus	91.73	90.50
	IDBA_UD	96.48	93.56
	Ray_Meta	87.12	83.21
	MetaVelvet	85.37	83.83
	Omega	90.71	79.10

次に、完成ゲノムが公開されている 11 種に関してベンチマークを行なった。事前に各 scaffold を全菌種のリファレンスゲノムに Blastn でアラインし、トップヒットを示す菌種に scaffold を割り当て、完成ゲノムを持つ 11 種に対応するもののみを抽出して QUAST を実行した。リファレンスゲノムが完成していることから、NGA50 の値や QUAST の報告する"#Misassemblies"の値を正しく評価することができる。結果を表 3-9 に示す。scaffold NGA50、contig NGA50 は全ケースで MetaPlatanus の値が最大となっている。NGA50 はアセンブリ結果の長さと精度を統合した指標であり、MetaPlatanus がそれらを両立していることを示している。

表 **3-9**　完成リファレンスゲノムが存在する種のアセンブリ結果

(A) 合計長・Scaffold 数・NG50

		合計長 (≥1 kbp)	Scaffold数 (≥1 kbp)	Contig NG50 (bp)	Scaffold NG50 (bp)
Case 1	MetaPlatanus	52,944,894	427	301,186	1,951,008
	IDBA_UD	52,532,825	294	168,141	724,393
	Ray_Meta	52,114,041	1,151	208,993	699,352
	MetaVelvet	47,538,481	1,635	80,159	81,188
	Omega	50,623,969	1,847	255,659	255,659
Case 2	MetaPlatanus	52,987,047	302	400,615	1,951,826
	IDBA_UD	52,466,076	273	178,297	1,003,417
	Ray_Meta	56,503,307	295	314,624	2,021,549
	MetaVelvet	51,002,766	2,689	51,139	51,278
	Omega	56,950,193	1,412	232,535	232,535
Case 3	MetaPlatanus	52,936,380	222	362,171	1,883,390
	IDBA_UD	52,455,722	486	160,206	825,026
	Ray_Meta	56,971,592	245	228,195	530,793
	MetaVelvet	51,605,182	1,849	66,441	66,659
	Omega	53,799,649	980	227,183	227,183

(B) NGA50

		Contig NGA50 (bp)	Scaffold NGA50 (bp)
Case 1	MetaPlatanus	252,344	730,905
	IDBA_UD	164,911	456,977
	Ray_Meta	186,350	347,715
	MetaVelvet	78,637	78,637
	Omega	221,658	221,658
Case 2	MetaPlatanus	338,295	929,095
	IDBA_UD	170,506	647,683
	Ray_Meta	255,282	293,722
	MetaVelvet	51,056	51,235
	Omega	200,087	200,087
Case 3	MetaPlatanus	320,088	922,434
	IDBA_UD	155,965	656,312
	Ray_Meta	184,506	218,857
	MetaVelvet	65,555	66,307
	Omega	203,375	203,375

(C) アセンブリエラーに関する情報

		#Mismatches /100kbp	#Indels/100kbp	#Local-misassemblies	#Misassemblies
Case 1	MetaPlatanus	4.20	0.65	823	118
	IDBA_UD	3.89	1.23	770	99
	Ray_Meta	6.11	0.67	494	165
	MetaVelvet	3.56	0.57	35	87
	Omega	8.38	2.72	769	212
Case 2	MetaPlatanus	3.14	0.58	323	89
	IDBA_UD	2.77	1.02	677	75
	Ray_Meta	5.24	0.80	272	340
	MetaVelvet	2.67	0.52	14	36
	Omega	13.18	2.33	724	454
Case 3	MetaPlatanus	3.82	0.62	288	94
	IDBA_UD	2.02	0.77	567	68
	Ray_Meta	12.48	1.54	1,983	516
	MetaVelvet	2.00	0.50	21	56
	Omega	13.95	2.55	948	340

coverage depth と scaffold の長さの関係を調べるため、Case 3 で菌種毎の scaffold NG50 を算出した。今回もリファレンスゲノムに対し scaffold をクエリとして Blastn を実行し、トップヒットを基に各 scaffold を菌種に対応付ける。各菌種でリファレンスゲノム配列の合計長をゲノムサイズとして scaffold NG50 を計算した結果が図 3-9 である。MetaPlatanus は paired-end の coverage depth が 10 以下の 2 種以外では scaffold NG50 は 400 kbp 以上となっている。また、coverage depth が 23 と 2,811 の 2 種（*C. pas.*、*C. hyl.*）で 1 Mbp 以上の scaffold NG50 を達成しており、23–2811 の広範囲の coverage depth に対応できていることが示唆されている。

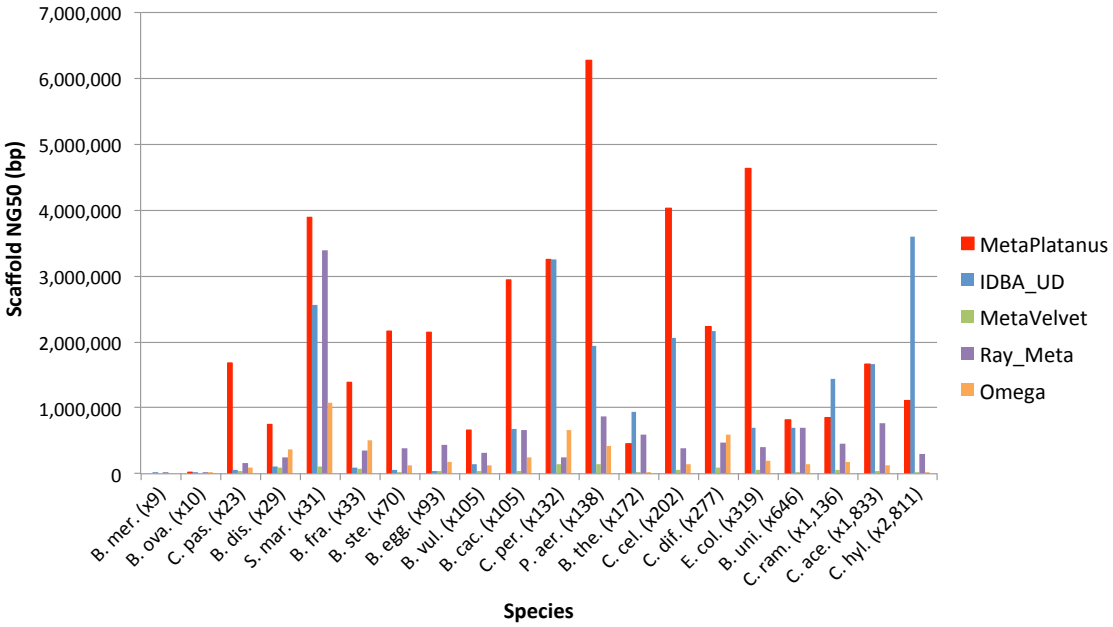


図 3-9　Case 3 での菌種毎の scaffold NG50

横軸の括弧内は paired-end の sequence coverage depth。

続いて、MGC によるクラスタリングの精度を調べた。事前に各 scaffold をリファレンスゲノムに対して Blastn でアラインし、トップヒットを基に対応する菌種を決定しておき、purity と recall の 2 つの指標を算出した。purity は各クラスタの精度を表し、あるクラスタに属する scaffold が全て同じ菌種由来ならば値は 1 となる。

$$purity = \frac{\text{クラスタ内で最大の菌種の scaffold 合計長}}{\text{クラスタ内の scaffold 合計長}}$$

recall は各菌種のゲノムの再現度を表し、ある菌種由来の scaffold が全て 1 クラスタに属しているとき 1 となる。

$$recall = \frac{\text{菌種内で最大のクラスタの scaffold 合計長}}{\text{菌種内の scaffold 合計長}}$$

purity はクラスタについて、recall は菌種についての平均値を求め、表 3-10 に示す。クラスタ数はデフォルト（scaffold 合計長 /3 Mbp）と 20 を指定した場合を

試した。デフォルトではクラスタ数は 29–31 個と多めに推定され、クラスタ数 20 指定の場合と比較して recall は下がるが、Case1, 3 では purity は上がっている。どちらの設定でも purity、recall とも 0.83 以上という結果を得ることができた。

表 3-10　クラスタリング結果の評価

(A) クラスタ数：デフォルト

	平均 purity	平均 recall
Case1	0.883	0.854
Case2	0.835	0.873
Case3	0.866	0.835

(B) クラスタ数：20 を指定

	平均 purity	平均 recall
Case1	0.873	0.952
Case2	0.918	0.926
Case3	0.844	0.859

最後に、MetaPlatanus の幾つかの機能を無効化した場合のアセンブリ結果を比較する。対象とした機能は、配列の組が異種由来かの判定に関するものであり、次の 5 通りの試行を各ケースで行なった。

・"simple"
Scaffolding 時に contig の組が異種由来かの判定を行わない。

・"coverage info."
Scaffolding 時に contig の組が異種由来かの判定を、coverage depth の情報のみでおこなう。3.3.3 節における $F_{4\text{-mer}}(d) = 1$ （一定）とした状態に対応する。

・"coverage + 4-mer info."
Scaffolding 時に coverage depth と 4-mer 頻度の情報を用いる。デフォルトの設定でクラスタリング以降を行わない状態に対応する。

・"MGC"
デフォルトの設定で MGC によるクラスタリングと re-scaffolding を行う。表 3-8、表 3-9 の"MetaPlatanus"と同一。

・"MGC n20"
MGC のクラスタ数を理想的な 20 に設定し、クラスタリングと re-scaffolding を行う。

規模の大きい（≥ 10 kbp）異種間のミスアセンブリ数（"#Chimera"）に加えて、アセンブリ結果の長さと精度を統合した指標である contig NGA50 と scaffold NGA50 を指標の代表として表 3-11 に示す。なお、NGA50 は完成リファレンスゲノムを用いて求めている。"simple"では Scaffolding の際にリンクの制限を行わないので、アセンブリ結果の長さ（NGA50）は最大となるが、Case 1, 2 で計 3 つの異種間ミスアセンブリが起こっており、精度に問題があることが分かる。今回のテストでは coverage depth の情報を活用するだけでも異種間ミスアセンブリ数を 0 にすることができるが、scaffold、contig NGA50 に注目すると、4-mer 頻度情報、MGC のクラスタリングを活用することで、全ケースで値を上げることができる。ここで、クラスタ数を理想的な値（20）に設定しても NGA50 は変化しない。表 3-10 で示されているように、クラスタリングの評価値である recall はクラスタ数を適切に指定すると増加するが、配列レベルで評価を行う場合はデフォルトの設定でも劣らない値が得られることが示された。

表 **3-11　MetaPlatanus** の各機能の効果の比較

		#Chimera (全菌種) (≥10 kbp)	Contig NGA50 (完成ゲノムの種) (bp)	Scaffold NGA50 (完成ゲノムの種) (bp)
Case 1	simple	2	301,186	878,579
	coverage info.	0	253,244	691,551
	coverage + 4-mer info.	0	252,344	695,556
	MGC	0	252,344	730,905
	MGC n20	0	252,344	730,905
Case 2	simple	1	446,047	1,146,864
	coverage info.	0	327,200	897,606
	coverage + 4-mer info.	0	338,295	899,095
	MGC	0	338,295	929,095
	MGC n20	0	338,295	929,095
Case 3	simple	0	428,293	955,055
	coverage info.	0	308,566	811,153
	coverage + 4-mer info.	0	320,088	888,950
	MGC	0	320,088	922,434
	MGC n20	0	320,088	922,434

3.4.2　環境メタゲノム実データのアセンブリ

・Cow rumen メタゲノムデータの取得および特徴

環境 DNA サンプルの実データとして、公開されている Cow rumen（牛の反芻胃）のメタゲノムデータ（Hess et al. 2011）を用いて MetaPlatanus のテストを行なった。このデータには paired-end に加えて 2 種類の mate-pair が含まれているという特徴があり、Scaffolding に新たな機能が追加された MetaPlatanus のテストには適していると判断した。また、paired-end の合計サイズが生データで 200 Gbp 以上と大きいという点も特徴である。

リードの前処理としては、全てのライブラリでアダプタ配列、低クオリティ領域、コンタミネーション由来のペアの除去を行なっている。コンタミネーションの候補としては、牛（*Bos taurus*）に加えて、胃の内容物である switch grass（*Panicum virgatum*）が考えられるため、全リードを Bowtie2（--local）でそれら 2 種のゲノム配列にアラインし、どちらか一方のリードで 20 bp 以上のアライメントが検出された場合ペアを除去した。mate-pair については、PCR-duplicate および短いインサートサイズを持つペア（インサートサイズ ＜nominl-インサートサイズ×0.5）を除いた。生データと前処理後のデータサイズを表 3-12 に示す。

表 3-12　Cow rumen サンプルシークエンスデータ

インサート長 (bp)		200	300	3,000	5,000
ライブラリ		paired-end	paired-end	mate-pair	mate-pair
リード長 (bp)	**Raw**	125.0	101.0	65.7	65.2
	Preprocessed	98.2	85.0	55.9	58.0
合計長 (bp)	**Raw**	17.3 G	158.1 G	31.8 G	26.8 G
	Preprocessed	13.6 G	82.3 G	20.8 G	17.5 G

・アセンブリ結果および考察

元の論文（Hess et al. 2011）では、Velvet を基に独自のパイプラインを構築して *de novo* アセンブリおよび 4-mer 頻度情報によるクラスタリングを行い、複数のバクテリアのドラフトゲノムをアセンブリしている。ただし、そのドラフトゲノムは公開されていないため、今回は配列比較は行なっていない。なお、

MetaVelvet、IDBA_UD、Ray_Meta、Omega も実行を試みたが異常終了という結果に終わった。その際の実行環境は以下の通りである。

・プロセッサ: Intel(R) Xeon(R) CPU X7560 2.27 GHz

・プロセッサ数: 32

・RAM: 512 GB

論文に記載されたアセンブリ結果（"Publication-assembly"）と MetaPlatanus のアセンブリ結果を表 3-13 に示す。MetaPlatanus については、paired-end のみの場合（"PE"）と mate-pair まで加えた場合（"PE+MP"）の両方を記す。

表 3-13　Cow rumen データのアセンブリ結果

	Scaffold長の下限	合計長 (bp)	Scaffold数	NG50 (bp)	最大長 (bp)
MetaPlatanus (PE)	1 kbp	2,421,427,830	710,754	6,939	737,230
MetaPlatanus (PE+MP)		2,427,103,784	460,268	49,625	1,999,196
Publication-assembly		1,930,000,000	179,092	34,338	1,529,637
MetaPlatanus (PE+MP)	1 Mbp	31,000,698	23	1,622,938	1,999,196
Publication-assembly		9,500,000	8	1,985,061	1,529,637

NG50 は Publication-assembly の合計長をゲノムサイズとして算出した。

1 kbp 以上の配列（"Scaffold 長の下限": 1 kbp）に関しては MetaPlatanus の NG50 は mate-pair を加える事で約 7 倍となり、メタゲノムの実データにおいても mate-pair の有効性が示されている。1 kbp 以上の配列について元論文のアセンブリ結果と比較すると、NG50 と合計長の両方で MetaPlatanus の値がより大きくなっており、長さとゲノムのカバー率に関して MetaPlatanus が優位であることが示唆された。1 Mbp 以上の配列の数と合計長も MetaPlatanus の方が多く、MetaPlatanus が個別に最適化されたパイプラインに優る性能を持つ可能性が高いといえる。また、他のアセンブラが異常終了するようなサイズのデータに関しても実行可能であり、大容量データへの汎用性があることも示された。

3.5　考察

大腸菌 MG1655 株を用いたベンチマークでは、coverage depth ≈ 100 の

paired-end とインサートサイズ 8 kbp までの mate-pair を Platanus に入力することで、染色体全体を 1 本でカバーする contig が得られることが確認された。このときギャップ（'N'）は contig 上に存在せず、データベース登録時などには「完成ゲノム」として扱うことも可能な完成度を示している。必要なデータサイズは Illumina シークエンサの 1 運転毎のスループットより小さい。マルチプレックスと呼ばれる方法で、シークエンスの 1 単位であるレーン中に複数サンプルを入力することができ、それを駆使して運転 1 回で複数の完成ゲノムを決定する計画の実現可能性が示唆されている。O157 株では 10–20 kbp のリピート配列が存在しているため、インサートサイズ 12 kbp の mate-pair を用いても完成ゲノムには至らない。このケースではリード長の長い（～20 kbp）PacBio データでもミスアセンブリが起きており、PacBio シークエンサを用いれば単純に解決する問題ではないことが分かる。ただし、O157 は大腸菌株の中でも外来のリピート配列の多さで特徴付けられており、バクテリア全体でも特殊なケースである可能性がある。複数のバクテリアの完成ゲノムを構築していくには、最初に Illumina シークエンサと Platanus を用いてアセンブリを行い、完成に至らない菌種のみについて追加のシークエンスや手作業によるパラメータ調整などを行なっていくことが、コストの点では優れている場合もあると考えられる。

仮想メタゲノムデータによるベンチマークにおいても、インサートサイズが 8 kbp の mate-pair を含むライブラリ構成では MetaPlatanus によりドラフトゲノムが多数得られることが示された。Case 3 においては、paired-end の coverage depth が 23×以上のサンプル（18 種）は全て 400 kbp 以上の NG50 を達成している。今回用いた菌種のリファレンスゲノムの内、最も N50 が小さいものは *Bacteroides uniformis* であり、その値は 287 kbp である。つまり、大半が「ドラフトゲノム」と呼ぶことが可能な長さのアセンブリ結果であると言える。実データ（Cow rumen メタゲノム）においても MetaPlatanus が有効であることは示唆されたが、やはり scaffold 配列を延長するにあたって mate-pair の効果は大きい。メタゲノムサンプルにおける mate-pair の活用は一般的ではないが、*de novo* アセンブリを行う際には検討する価値があると考えられる。

第 4 章　総括

本研究で開発された *de novo* アセンブラ：Platanus が最も既存ツールに対して優位な性能を発揮するのは、高ヘテロ接合性サンプルを対象とした場合である。ゲノム解読計画の対象がモデル生物から非モデル生物へ移り変わるに従って、ヘテロ接合性が全ゲノムショットガン法へ悪影響を与えることが顕在化し、fosmid 等を用いた階層的ショットガン法を用いた研究が再び増加していた。本研究は高ヘテロ接合性データに対しても全ゲノムショットガン法が有効であることを実証した。一方、第 1 章、第 2 章を通して、Platanus は汎用的に用いることができるという側面も明らかとなった。シーラカンスのような低ヘテロ接合性の真核生物や、1 倍体であるバクテリアに対しても他のツールと比較して優れた精度を示している。これは、ヘテロ領域解決のための機能が正しく働き、過不足なくヘテロ領域を認識して統合していることを示唆する。加えて、Platanus の派生プログラムである MetaPlatanus は環境 DNA のメタゲノムデータ用に開発され、仮想メタゲノムデータによるテストではバクテリアのドラフトゲノムが構築可能であることを示している。また、全てのサンプルで Platanus のパラメータはデフォルト値であり、手作業によるパラメータ調整が不要という点も注目に値する。Platanus の実際の適用例としては、第 1 章で取り上げたシーラカンスの他、ネムリユスリカのゲノム解読計画（Gusev et al. 2014）があり、今後も Platanus が広く活用され全ゲノム配列の決定がより効率化されることが期待される。

多数の生物種の全ゲノム配列解読を目的とした計画は、2014 年時点で複数が進行中である。真核生物を対象とした例としては、脊椎動物 10,000 種を対象とした Genome 10K（Genome 10K Community of Scientists. 2009）、昆虫 5,000 種を対象とした i5K（i5K Consortium. 2013）が挙げられる。これらの計画で発表されるゲノム配列の精度評価は未だ行われていないが、*de novo* アセンブリの効率化が求められる典型的なケースと言える。対象となる生物種の多くは研究室内での近交系が確立されておらず、ヘテロ接合度の高いサンプルも含まれると予想されるため、Platanus の長所が発揮されることが見込まれる。原核生物を対象とした類似の計画としては、100 以上のバクテリア、アーキアの全ゲノム解読を目的とした Genomic Encyclopedia of Bacteria and Archea（GEBA）（Wu et al. 2009）などが実施されている。GEBA では決定された全ゲノム配列が既にデータベースに登録されており、多数の原核生物のゲノム解読計画は実現可能であること

が示されている。さらに、難培養性の種を対象とした計画である GEBA-Microbial Dark Matter（GEBA-MDM）も発表されている（Rinke et al. 2013）。この計画ではメタゲノムや 1 細胞ゲノムのシークエンスが用いられていた。GEBA や GEBA-MDM に類似する計画は今後増加する可能性があり、原核生物データにも有効である Platanus、MetaPlatanus が活用されることにより、精度の高い全ゲノムデータが多く得られることが期待される。

本研究で主に扱った Illumina シークエンサの他に注目すべきシークエンス技術としては、1 分子シークエンス法が存在する。2014 年時点では Pacific Biosciences（PacBio）社の DNA シークエンサ（Eid et al. 2009）が実用化されている。そのリード長は最大 20 kbp に達し、coverage depth のバイアスが小さいなどの利点がある一方、エラー率が約 15%と高い欠点も有する。1 塩基あたりのシークエンスコストは、2014 年時点では Illumina シークエンサの方が小さいが、複数の mate-pair ライブラリを調整する場合、その差は縮まることに注意する必要がある。*de novo* アセンブリに PacBio データを用いる場合、当初は Illumina データを用いて PacBio データのエラーを修正する手法（Koren et al. 2012）が考案されていたが、その後 PacBio データ単独でアセンブリを行う HGAP などのアセンブラも開発されている（Chin et al. 2013）。第 3 章での大腸菌データによるベンチマークでは、Illumina と PacBio のデータをそれぞれ Platanus と HGAP でアセンブリした結果を比較した場合、明確な優劣を決めることはできなかった。Assemblathon2 の bird サンプルについては、PacBio データが用意されており、それを用いた参加者も存在している中、Platanus は NG50 と fosmid 配列による評価で最高の性能を示した。ここでの PacBio データは coverage depth が小さく（約 10）、Illumina データが *de novo* アセンブリに適していると結論付けることはできないが、現時点で Illumina データと Platanus を用いることが選択肢として考えられるという点は確かである。その他の新規な技術としては、より小型な機器で 1 分子シークエンスを行う Oxford Nanopore Technologies 社製のシークエンサ（Laszro et al. 2014; Ashton et al. 2014）や、10 kbp の DNA 断片にタグ配列情報を付加して Illumina シークエンサを適用し、局所的アセンブリにより擬似的なロングリードを得る Moleculo という手法（Voskoboynik et al. 2013; Kuleshov et al. 2014）などが存在する。これらを *de novo* アセンブリに用いた際の詳細な精度評価は未だ行われていないが、今後も注目すべき技術であると言える。

アセンブリ結果が断片化されている（contig、scaffold が短い）場合や、精度

が低い場合は、遺伝子予測を行う場合に悪影響が起こることが報告されている（Denton et al. 2014）。遺伝子数の推定時には、contig 長が短いときは遺伝子領域が分断されて過大になり、ギャップの割合が大きいときは過小となる。遺伝子ファミリーの多様化、欠失は進化について考察する際に重要であり、*de novo* アセンブラの性能がその後の解析にも深く関わる例となる。遺伝子領域のみに注目するならば、転写産物のみをシークエンスする RNA-seq を実施し、transcriptome 解析を行うという選択も考えられる。しかし、ゲノム配列を用いない場合は遺伝子のクラスタ構造や他種とのシンテニー構造について解析することはできない。また、ENCODE 計画（Bernstein et al. 2012）で提唱されているように、遺伝子以外のゲノム領域も多くは機能を持っていると考えられる。第 2 章で扱ったシーラカンスゲノムの解析（Nikaido et al. 2013）では、非コード保存領域（conserved noncoding element, CNE）が鰭と四肢の進化に関わっているという興味深い仮説が立てられた。また、リピート配列の一種である転移因子の解析では、転移因子の多様化と種分化の時期に関して考察が試みられている。これらは全ゲノム配列が構築されていなければ行えない解析の例である。バクテリアについても第 3 章で扱った *E. coli* O157 Sakai 株のゲノム解析（Hayashi et al. 2001）では、リピートの一種となる外来配列が表現形に大きな影響を及ぼしている可能性が示された。全ゲノム配列はゲノムサイズを問わず解析に重要であることが分かる。本研究で開発、評価を行なった *de novo* アセンブリ手法が活用されることで高精度な配列が構築され、ゲノミクスの知見の蓄積に貢献することが望まれる。

参考文献

Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C, Venter JC. 1992. Sequence identification of 2,375 human brain genes. *Nature* **355**: 632-634.

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**: 533–538.

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, Maccallum I, Braasch I, Manousaki T, Schneider I, Rohner N, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**: 311–316.

Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J. 2014. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol*. ePub ahead of print Dec. 8.

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Eric S. 2002. ARACHNE: A Whole-Genome Shotgun Assembler. *Genome Res.* **12**: 177–189.

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-vides J, Glasner JD, Rode CK, Mayhew GF*, et al.* 1997. The Complete Genome Sequence of *Escherichia coli* K-12. *Science* **277**: 1453-1462.

Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* **13**: R122.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman J a, Chapuis G, Chikhi R, *et al*. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**: 10.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**: 2012–2018.

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, *et al.* 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Meth.* **10**: 563–569.

Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo M-J, Dupont CL, Badger JH, Novotny M, Rusch DB, Fraser LJ, Gormley N, *et al.* 2011. Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nat. Biotechnol.* **29**: 915–921.

Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. 2014. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLoS Comput. Biol.* **10**: e1003998.

Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. 2014. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLoS Comput. Biol.* **10**: e1003998.

Faloutsos C and Lin KI. 1995. FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. *SIGMOD* **24**: 163-174.

Fleischmann RD, Adams MD, White O, Clayton RA, Ewen F, Kerlavage AR, Bult CJ, Tomb J, Dougherty BA, Merrick JM, *et al*. 1995. Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd. *Science* **269:** 496–512.

Genome 10K Community of Scientists. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.* **100**: 659–674.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, *et al.* 1996. Life with 6000 Genes. *Science* **274**: 546-567.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall

G, Shea TP, Sykes S, *et al*. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**: 1513–1518.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al*. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**: 644–652.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072–1075.

Gusev O, Suetsugu Y, Cornette R, Kawashima T, Logacheva MD, Kondrashov AS, Penin A a, Hatanaka R, Kikuta S, Shimura S, *et al*. 2014. Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge. *Nat. Commun.* **5**: 4784.

Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han C, Ohtsubo E, Nakayama K, Murata T, *et al*. 2001. Complete Genome Sequence of Enterohemorrhagic Eschelichia coli O157 : H7 and Genomic Comparison with a Laboratory Strain K-12. *DNA Res.* **22**: 11–22.

Haider B, Ahn T-H, Bushnell B, Chai J, Copeland A, Pan C. 2014. Omega: an Overlap-graph de novo Assembler for Metagenomics. *Bioinformatics* **30**: 2717–2722.

Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, *et al*. 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**: 463–467.

Hongoh Y, Sharma VK, Prakash T, Noda S, Toh H, Taylor TD, Kudo T, Sakaki Y. 2008. Genome of an Endosymbiont Within Protist Cells in Termite Gut. *Science* **322**: 1108–1109.

i5K Consortium. 2013. The i5K Initiative : Advancing Arthropod Genomics for Knowledge , Human Health , Agriculture , and the Environment. *J. Hered.* **104**: 595–600.

Inoue JG, Miya M, Venkatesh B, Nishida M. 2005. The mitochondrial genome of Indonesian coelacanth Latimeria menadoensis (Sarcopterygii: Coelacanthiformes) and divergence time estimation between the two coelacanths. *Gene* **349**: 227–235.

Kim JH, Waterman MS, Li LM. 2007. Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savigny*i. *Genome Res.* **17**: 1101–1110.

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, *et al*. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**: 1384–1395.

Kent WJ. 2002. Blat—the BLAST-like alignment tool. *Genome Res*. **12**: 656– 664.

Kohara Y, Akiyama K, Isono K.1987. The Physical Map of the Whole E. coli Chromosome: Application of a New Strategy for Rapid Analysis and Sorting of a Large Genomic Library. *Cell* **50**: 495–508.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko D a, McCombie WR, Jarvis ED, et al. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**: 693–700.

Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, Kertesz M, Snyder M. 2014. Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotech*. **32**: 261–266.

Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP, *et al*. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* **14**: 169–181.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol*. **5**: R12.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. 2001. Initial sequencing and analysis of the human

genome. *Nature* **409**: 860-921

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**: 357–359.

Laszlo AH, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC, Craig JM, Langford KW, Samson JM, Daza R, *et al*. 2014. Decoding long nanopore sequencing reads of natural DNA. *Nat. Biotechnol.* **32**: 829–834.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, *et al.* 2010. The sequence and de novo assembly of the giant panda genome. *Nature* **463**: 311–317.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, *et al.* 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**: 265–272.

Magoc T, Pabinger S, Salzberg SL, Canzar S, Liu X, Su Q, Puiu D, Tal- LJ. 2013. GAGE-B : An Evaluation of Genome Assemblers for Bacterial Organisms. *Bioinformatics* **28**: 1718–1725.

modENCODE Consortium. 2010. Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science* **330**: 1787–1797.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, *et al.* 2000. A Whole-Genome Assembly of Drosophila. *Science* **287**: 2196–2204.

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, *et al*. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39**: e90.

Namiki T, Hachiya T, Tanaka H, Sakakibara Y. 2012. MetaVelvet: an extension of

Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* **40**: e155.

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, *et al.* 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**: 822-828.

Nikaido M, Noguchi H, Nishihara H, Toyoda A, Suzuki Y, Kajitani R, Suzuki H, Okuno M, Aibara M, Ngatunga BP, *et al.* 2013. Coelacanth genomes reveal signatures for evolutionary transition from water to land. *Genome Res.* **23**: 1740–1748.

Noguchi H, Taniguchi T, Itoh T. 2008. MetaGeneAnnotator : Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. *DNA Res.* **15**: 387–396.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, *et al.* 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**: 579–584.

Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420–1428.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA* **98**: 9748-9753.

The Potato Genome Sequencing Consortium. 2011. Genome sequence and analysis of the tuber crop potato. *Nature* **475**: 189–195.

Laszlo AH, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC, Craig JM, Langford KW, Samson JM, Daza R, *et al*. 2014. Decoding long nanopore sequencing reads of natural DNA. *Nat. Biotechnol.* **32**: 829–834.

Saitoh K, Sado T, Doosey MH, Bart HLJ, Inoue JG, Nishida M, Mayden RL, Nishida

M, Miya M. 2011. Evidence from mitochondrial genomics supports the lower Mesozoic of South Asia as the time and place of basal divergence of cypriniform fishes (Actinopterygii: Ostariophysi). *Zool. J. Linn. Soc.* **161**: 633–662.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, *et al.* 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **22**: 557–567.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, *et al.* 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**: 169– 175.

Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman D a, Banfield JF. 2013. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* **23**: 111–120.

Simpson JT, Durbin R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* **22**: 549–556.

Small KS, Brudno M, Hill MM, Sidow A. 2007. Extreme genomic variation in a natural population. *Proc. Natl. Acad. Sci. USA* **104**: 5698–5703.

Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs R a, Angerer RC, Angerer LM, Arnone MI, Burgess DR, Burke RD, *et al*. 2006. The genome of the sea urchin Strongylocentrotus purpuratus. *Science* **314**: 941–952.

Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A, et al. 2011. The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477**: 207–210.

Takeuchi T, Kawashima T, Koyanagi R, Gyoja F, Tanaka M, Ikuta T, Shoguchi E, Fujiwara M, Shinzato C, Hisata K, et al. 2012. Draft genome of the pearl oyster Pinctada fucata: a platform for understanding bivalve biology. *DNA Res.* **19**: 117–130.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M,

Fitzgerald LM, Vezzulli S, Reid J, *et al*. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**: e1326.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans C a, Holt R a*, et al.* 2001. The sequence of the human genome. *Science* **291**: 1304–1351.

Vinson JP, Jaffe DB, O'Neill K, Karlsson EK, Stange-Thomann N, Anderson S, Mesirov JP, Satoh N, Satou Y, Nusbaum C, *et al.* 2005. Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. *Genome Res.* **15**: 1127–1135.

Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ, *et al*. 2013. The genome sequence of the colonial chordate, Botryllus schlosseri. *Elife* **2**: e00569.

WangJ,WangW,LiR,LiY,TianG,GoodmanL,FanW,ZhangJ,LiJ,GuoY, *et al*. 2008. The diploid genome sequence of an Asian individual. *Nature* **456**: 60–65.

Watanabe H, Fujiyama A, Hattori M, Taylor TD, Toyoda A, Kuroki Y, Noguchi H, BenKahla A, Lehrach H, Sudbrak R, *et al.* 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429**: 382– 388.

Wu C, Ye R, Jasinovica S, Godiska R, Tong AH, Lok S, Krerowicz A, Knox C, Mead D. 2012. Long-span , mate-pair scaffolding and other methods for faster next-generation sequencing library creation. *Nat. Methods* **9**: i–ii.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, *et al*. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**: 1056–1060.

Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S. 2012. FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One* **7**: e52249.

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, *et al*. 2011. Genome sequence and analysis of the tuber crop potato. *Nature* **475**: 189–195.

You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, et al. 2013. A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.* **45**: 220–225.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**: 821–829.

Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, Yang P, Zhang L, Wang X, Qi H, *et al*. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**: 49–54.

Zimin A, Marçais G, Puiu D, Roberts M, Salzberg L, Yorke JA. 2013. The MaSuRCA genome Assembler. *Bioinformatics* **29:** 2669-2677

謝辞

本研究は東京工業大学 伊藤武彦教授の御指導のもとで行われました。伊藤教授には心より感謝申し上げます。

以下に本論文で扱ったテーマの順番に従い、研究に御協力頂いた方々への謝辞を述べます。Platanus のソースコード改良に御協力を頂きました理化学研究所 年本広太研究員に御礼申し上げます。線虫（*C. elegans*）の DNA サンプルを御提供頂きました国立遺伝学研究所 小原雄治教授に感謝申し上げます。線虫（*S. venezuelensis*）の DNA サンプルを御提供頂きました宮崎大学 丸山治彦教授に御礼申し上げます。*S. venezuelensis* の fosmid 配列のシークエンスを行なって頂きました東京大学 白髭克彦教授に御礼申し上げます。シーラカンス（*L. chalumnae*、*L. menadoensis*）の DNA サンプルを御提供頂きました東京工業大学 岡田典弘名誉教授、東京大学 菅野純夫教授、国立遺伝学研究所 藤山秋佐夫教授、アクアマリンふくしまの方々に御礼申し上げます。*C. elegans*、*S. venezuelensis*、*L. chalumnae*、*L. menadoensis* のシークエンスを行なって頂きました国立遺伝学研究所 豊田敦准教授に御礼申し上げます。第 3 章の多くのバクテリア DNA サンプルの御提供および研究についての御助言を頂きました宮崎大学 林哲也教授、小椋義俊助教、後藤恭弘助教に御礼申し上げます。大腸菌と仮想メタゲノムのシークエンスを行なって頂きました東京工業大学 流水利恵技術員に御礼申し上げます。クラスタリングプログラムのソースコードおよび研究への御助言を頂きました国立遺伝学研究所 野口英樹准教授に御礼申し上げます。仮想メタゲノム用 DNA サンプルを御提供頂きました香川大学 桑原知己教授に御礼申し上げます。

博士課程在学中は博士課程教育リーディングプログラム　情報生命博士教育院より経済的、教育的な御支援を頂いたことにつきまして、御礼申し上げます。

同研究室の方々、卒業生の方々には日頃より有意義な議論をさせて頂き感謝致します。特に伊藤研究室秘書 坂東由衣さんには在学中の多くの手続きで御世話になりましたことに御礼申し上げます。