新規ゲノムアセンブリとアノテーション

野口　英樹

次世代型DNAシーケンサー（NGS）の登場とその後の急速な発展により，全ゲノム情報の取得にかかる時間と費用は大幅に削減された．2014年にNatureに掲載された「1,000ドルゲノム」の記事[1]によれば，NGSが普及し始めた2007年以降，シーケンシングにかかる費用は実に2,000分の1にまで下がってきている．NGSのスループット向上は今なお継続中であり，それに伴い生命科学の幅広い分野で，塩基配列データに基づいた研究が行われるようになっている．NGSの応用範囲は多岐にわたっているが，解析にはリファレンスとなるゲノム配列が存在することを前提としたものが多い．そのため，すでにゲノムが決定されているモデル生物などを扱っている研究者以外の研究者がNGSを利用したいと考えた場合には，新規に対象生物種のゲノムを決定するところから始める必要がある．

本稿では，NGSを用いた新規ゲノムアセンブリの手順と，ゲノム配列決定後のゲノムアノテーションの手法について主要な解析ツールとともに紹介する．

新規ゲノムアセンブリ

NGSの配列データから新規（*de novo*）にゲノムアセンブリを構築する手順は，使用するシーケンサーにより大きく異なる．ここでは，ショートリード（リード長が数百bp）のシーケンサーとしてIllumina Hiseq/MiSeqを，ロングリード（平均10 Kbp以上）ではPacBio RS II/Sequelを想定したアセンブリ手法について解説する．

ショートリードアセンブリ　Illumina HiSeq/MiSeqから得られるリードは長さが最長で300 bpと短いが，その分リード数は多く，シーケンシングエラーも比較的少ない．シーケンサーの登場初期はリード長もより短く，DNAフラグメントの両端配列を決定するペアードエンドの手法も確立されていなかったため，*de novo*ゲノムアセンブリへの適用は微生物のようなゲノムサイズが小さく反復配列の少ない生物に限られていた．その後リード長も伸び，ペアードエンド（数百bpの短いDNAフラグメント）やメイトペア（数Kbpの長いフラグメント）の手法も確立されたことで，ジャイアントパンダのゲノム決定[2]を皮切りに現在では大型真核生物ゲノムの*de novo*アセンブリにもショートリードのNGSが広く用いられるようになっている．

ショートリードの場合，十分なデータ量（ゲノムサイズの60～100倍）を確保しようとすると，その分配列リード数は必然的に多くなる．NGS登場以前に一般的に用いられてきたOverlap-Layout-Consensus（OLC）というアセンブリ手法では，はじめにリードどうしの重なりを調べる必要があるが，その計算量はリード数の2乗に比例する．そのため，ショートリードのアセンブリにOLCを適用することは難しく，もっぱら*k*-merグラフを用いたアセンブリ手法が採用されている（図1）．この手法の場合，リードはより短い（長さ*k* bpの）*k*-merに分解される．*k*-merの種類数は全部で4^k個存在するが，リード全体の中でどの*k*-merが何回出現するかをカウントしておけば，*k*-merどうしのつながり（*k*-1 bpオーバーラップする*k*-mer）を短時間で調べることができる．このような*k*-merグラフに基づいたアセンブリツールは，Velvet[3]やAllpaths-LG[4]，SOAPdenovo2[5]，Platanus[6]など数多くが開発され，利用されている．

k-merグラフを利用したアセンブリは，①生データの

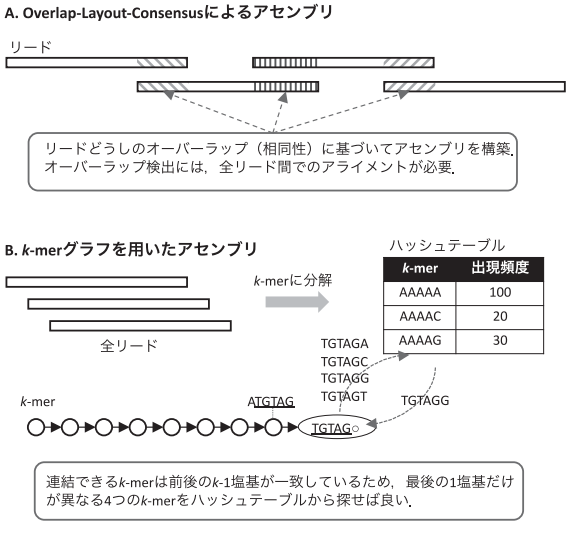

図1．ショートリードのアセンブリ手法

著者紹介　情報・システム研究機構データサイエンス共同利用基盤施設ゲノムデータ解析支援センター（センター長，特任教授）
E-mail: hnoguchi@nig.ac.jp

前処理，②*k*-merグラフを用いたアセンブリ，③ペアードエンドとメイトペアを用いたスキャフォールディング，④仕上げ（ギャップ埋め）処理，という手順で行われる．

得られたNGSデータはFastQC[7]などのツールで品質を評価できるが，生データの場合，特にリードの3'側の品質が低かったり，アダプターの配列が混入していたりすることがある．またメイトペアの場合には，リンカー配列の混入も考えられる．これらの余分な配列をリードデータから除去するのが前処理の目的であり，その後の計算の効率化や高品質なアセンブリ構築には欠かすことができない．もっとも，前処理自体は*de novo*アセンブリだけでなく，ゲノムリシーケンシングやRNA-seq解析などNGSを用いた解析全般で必要な処理であり，そのためのツールにはFASTX-Toolkit[8]やCutadapt[9]などがある．

前処理を行った後は，上述のゲノムアセンブラを用いることで最終的なアセンブリ構築までが行える．図2に，Platanusを例にした実際のツールの使い方を示す．この例では，ペアードエンドのライブラリが一組（PE），メイトペアのライブラリが3 Kbpと5 Kbpの二組（MP3K, MP5K）あるものと仮定している．また，前処理で生じる片エンドのみの配列（SE）も使用している．

k-merグラフを用いたアセンブリにおいて，最終結果に影響する重要なパラメータの一つが*k*をいくつに設定するかであろう．最適な*k*の値はリード長やデータ量に依存するだけでなく，ターゲットとなる生物種のヘテロ接合度や反復配列，遺伝子重複の度合いなどによっても変わってくる．そのため，最適値を決定するには結局のところ実際にいろいろな値を試して調べるしかない．データ量とのバランスが良い一つの目安は，リード長の半分の値に*k*を設定することである．Platanusの場合は，複数の*k*-merサイズを用いて自動的に最適なアセンブリを構築してくれる．

アセンブリの品質（ここではアセンブリの長さ）にもっとも影響を与えるのが，スキャフォールディングのプロセス，とりわけペアードエンド・メイトペアライブラリ作製時のDNAフラグメントの長さ（以下ライブラリ長）である．*k*-merベースのアセンブリではリードをより短い*k*-merに分解してグラフを構築するが，多くの場合*k*-merグラフだけではアセンブリの問題は解けない．図3にアセンブリ構築で問題となる典型的な*k*-merグラフ構造とその主な原因を示した．図3のB，Cはヘテロ接合やシーケンシングエラーにより生じる構造である．エラーが原因の場合，それにより生じる*k*-merの出現頻度は極端に低いため，比較的簡単に正解のパスが分かる．ヘテロ接合が原因の場合はどちらのパスも正しいため，任意のパスを選択する必要がある．最近はフェージング（アリルの組合せを解いて両方の配列を出力）も行われる．これらはいずれも解決にペアの情報は必要としない．しかし図3のAのような，反復配列や遺伝子重複など，ゲノム中に何度も現れる（*k*-merより長い）配列が原因

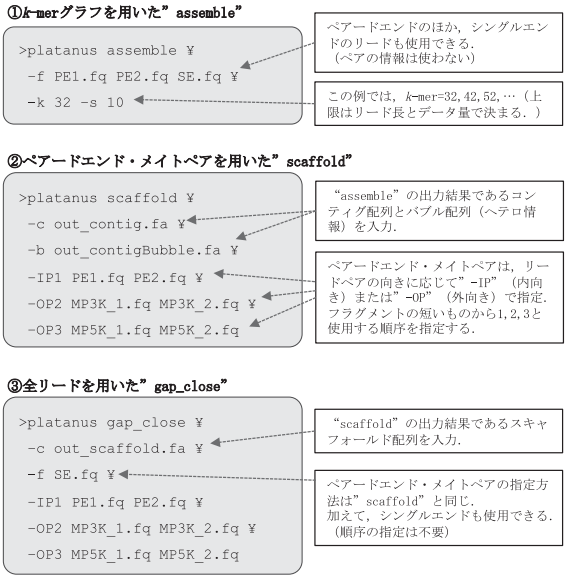

図2．Platanusを用いたショートリードアセンブリ

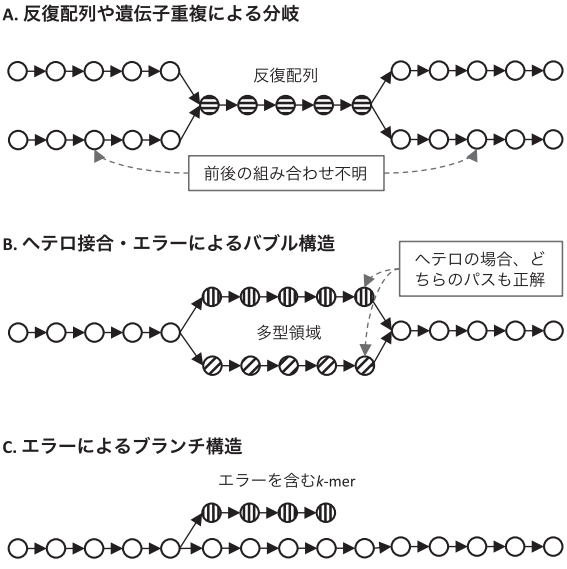

図3．複雑なグラフ構造とその要因

の枝分かれ構造を解くには，反復配列より長いペアードエンド・メイトペアの情報が必要となる．反復配列や遺伝子重複は高等真核生物のゲノム中には頻繁に出現するため，この解決なしに高品質なアセンブリを得ることはできない．ライブラリ長は長ければ長いほどより長い反復配列にまで対応できるが，逆に長すぎるとリード間に未解決の分岐を抱えることになる．これを避けるには，さまざまなサイズのライブラリを複数用意する必要がある．実際，初期のショートリードアセンブリであるジャイアントパンダゲノムの場合も，150 bp，500 bpのペアードエンドライブラリと，2 Kbp，5 Kbp，10 Kbpのメイトペアライブラリを用いてアセンブリを行っている．ペアードエンドだけでアセンブルした段階でのパンダゲノムのコンティグN50長は1.5 Kbpしかないが，さまざまなサイズのライブラリをすべて用いてスキャフォールディングを行うことで，最終アセンブリのコンティグN50長は40 Kbp，スキャフォールドN50長は1.3 Mbpまで伸長している．このように，単純にデータ量を増やすのではなく，対象の生物種に応じてさまざまなサイズのライブラリを用意することで，ショートリードであっても高品質なゲノムアセンブリを構築することができる．

ロングリードアセンブリ　PacBio RS II/Sequelの特徴は何と言ってもその配列リードの長さである．登場初期でも平均長が4～5 Kbp，現在では平均20 Kbp近くまで読めるようになっている．リード長そのものが長いため，ショートリードの時のように複数のライブラリを用意するまでもなく，さまざまなサイズの反復配列に対応できる．その一方で配列の精度は低く，エラーを15%前後も含んでいるのが最大の欠点であり，これをいかに克服するかがPacBioアセンブリの肝となる．PacBioのアセンブリでは，①リードのエラー補正，②OLCによるアセンブリ，③構築されたアセンブリを再度修正，というのが基本的な流れとなる．PacBio用のアセンブラにもCanu[10]やFalcon[11]などさまざまなものが開発されているが，PacBioが提供するSMRT Analysisの中にはHGAP[12]というゲノムアセンブリパイプラインも含まれており，これを用いることで上記の一連の作業を自動で行ってくれる．

PacBioリードのエラー補正プロセスは，HGAPではプレアセンブリと呼ばれている．ここではまず全リードの中から一定以上の長さを持つリード（シードリード）を選び出し，シードリードに対して他のリードをマッピングしてコンセンサス配列を作製することでリードの補正を行う．また，コンセンサス配列両端のクオリティの低い配列をトリムすることで，その後のアセンブリステップの精度向上を図っている．アセンブリステップでは補正後のシードリードだけを用いることになる．そのため，シードリードだけで全ゲノムをカバーできるデータ量となるようにシードリード長の下限を設定する必要がある．

ロングリードのアセンブリ手法には前述のようにOLCが用いられることが多い．HGAPの場合，HGAP3までのバージョンではCelera Assembler[13]が，またHGAP4ではFalconがアセンブリエンジンとして用いられている．Falconは入力データとして2倍体のゲノムを想定しており，メインとなるアセンブリ（primary contig）とともに別のアリルを含むコンティグ（associate contig）が出力される．これらの情報をFalcon-unzipに与えることで，2対の染色体を区別してアセンブルを行うこともできる．

補正後のリードを用いているとはいえ，出来上がったアセンブリには依然シーケンシングエラーが多く含まれている．特に挿入や欠失は，タンパクのコード領域に存在するとフレームシフトを引き起こすため，PacBioアセンブリでは最後にもう一度アセンブリの修正を行うことが望ましい．HGAP3まではQuiver，HGAP4ではArrowというツールがこの修正ステップ（アセンブリポリッシング）を担っている．

最終的に得られるPacBioアセンブリのスキャフォールドN50長は，適切に構築されたショートリードのアセンブリと大きく変わらないことが多い．しかしながら，ショートリードのアセンブリには配列が不定（N）の領域が多く存在しておりコンティグN50長が短くなる傾向にあるのに対し，PacBioのアセンブリではNをほとんど含まない連続した配列が得られる点が魅力である．

ハイブリッドアセンブリ　ショートリードとロングリードを組み合わせ，ショートリードの精度の高さとロングリードの長いリードという利点を生かしたアセンブリ手法はハイブリッドアセンブリと呼ばれている．ハイブリッドアセンブリの手法は，ロングリードのエラー補正にショートリードを用いる方法と，ショートリードで構築したコンティグのスキャフォールディングやギャップ埋めにロングリードを用いる方法とに大別できる．前者の手法は反復配列への対応が難しいことから主に反復配列の少ない小規模なゲノムのアセンブリに用いられ

る．一方，大型真核生物ゲノムのアセンブリには後者の手法の方が向いており，うまく組み合わせることで精度が高く連続した長い（Nを含まない）コンティグを構築することができる．このようなツールにはPBjelly[14]がある．

ゲノムアノテーション

ゲノムアセンブリが構築できても，当然のことながらその配列上には遺伝子の位置や構造などの情報は載っていない．ただACGTの文字が並んでいるだけである．せっかく構築したアセンブリを有効に利用するためには，次にゲノムアノテーションの作業が必要となる（図4）．原核生物とは異なり真核生物の場合は一つの遺伝子がエキソンとイントロンとに分断されているためアノテーションの難度も相対的に高くなる．ここでは，NGSによるRNA-seqデータを用いたアノテーションを中心に，公共データベース中の既知遺伝子配列を用いた相同性検索，ゲノム配列と遺伝子統計量に基づいた*ab-initio*遺伝子予測についても紹介する．

RNA-seq解析　転写産物の配列を網羅的に決定するRNA-seq解析でもNGSが活躍する．ロングリードでは遺伝子の全長配列や選択的スプライシングによるバリアントの配列を容易に決定できるという利点があるが，発現量の少ない遺伝子まで押さえようとするとまだデータ量的に不足がある．ショートリードの場合は得られるリードの多くが遺伝子の部分配列であるため，ゲノムの場合と同様にアセンブリ作業が必要となるが，得られるデータ量が多いことと，現在はストランドの方向まで評価できるようになっていることから，ゲノムアノテーションにも広く用いられている．

RNA-seqによるアノテーションは，RNA-seqデータだけを用いた*de novo*アセンブリと，参照ゲノム配列へのマッピングを通したアノテーションとに分けられる．*de novo*アセンブリの場合は，ゲノムアセンブリ同様*k-mer*グラフを用いたアセンブリ手法が広く利用されており，Trinity[15]をはじめ，Oases[16]やTrans-ABySS[17]，SOAPdenovo-Trans[18]といったツールが開発されている．RNA-seqアセンブリでは選択的スプライシングにより複数の「正解パス」がグラフ中に存在している点がゲノムアセンブリとは異なる．また，発現量の違いにより遺伝子ごと（アイソフォームごと）にデータの厚みが異なる点も問題を難しくしており，最終的なアセンブリにはミスアセンブリも多く含まれる．

図4．ゲノムアノテーションの例

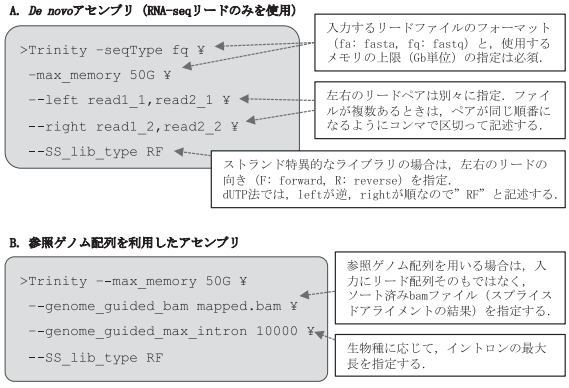

図5．Trinityを用いたRNA-seqアセンブリ

参照ゲノム配列へのマッピング情報を利用したアセンブリでは，短いリードを，スプライシングを考慮してマッピングするという難しさはあるが，シーケンシングエラーやアリルによる配列の違いを参照ゲノム配列側で吸収できるため，比較的ミスアセンブリは少なくてすむ．また，既知遺伝子のアノテーションを利用できるツールもあり，新規アイソフォームの同定とアイソフォームごとの転写量推定を同時に行ってくれるものも多く開発されている．代表的なツールとしては，Cufflinks[19]，iReckon[20]，StringTie[21]などのほか，前出のTrinityでも参照ゲノム配列を利用したアセンブリが行える．

図5に，Trinityを用いたde novoアセンブリ（A）および参照ゲノム配列を利用したアセンブリ（B）のそれぞれについて実際の使用方法を示す．この例では，ストランド特異的なRNA-seqのリードが二組（read1，read2）あるものと仮定している．参照ゲノム配列を用

いたアセンブリでは，事前にRNA-seqリードを参照ゲノムにマッピングしておく必要がある．このとき，bwa[22]やbowtie2[23]のような一般的な（リードの全域をマッピングする）アライメントツールではなく，スプライシングを考慮してアライメントを行ってくれるツールを使用しなければならない．このようなツールには，TopHat[24]やGSNAP[25], STAR[26]などがある．Trinityでは，Cufflinksなどの他の参照ゲノムを利用するアセンブリ手法と異なり，参照ゲノムはリードをグルーピングする（ゲノム上でオーバーラップする，同一遺伝子由来と思われるリードを集める）ためにだけ使用し，その後集められたリードのみで*de novo*アセンブリを行う．そうすることで，参照ゲノム配列上の変異やエラーによる影響を回避している．

相同性検索および*ab initio*予測　対象生物種やその近縁種の既知遺伝子配列が利用できる場合は，これらの配列とゲノム配列とをアライメントすることで遺伝子領域を特定できる．特に，近縁種の全ゲノム配列が決定され全遺伝子配列が明らかになっている場合はきわめて有効な手段である．近縁種の遺伝子は分岐年代に応じて配列が異なるため，一般的には塩基配列よりもアミノ酸配列を用いた方が予測の感度は高くなる．スプライシングを考慮しながらアミノ酸配列をゲノム配列にマッピングできるツールには，Exonerate[27]やGenewise[28]がある．Genewiseでは，マルチプルアライメントから作製した遺伝子のプロファイルHMMとゲノムとのアライメントを行うこともでき，より感度の高い予測が行える．

RNA-seqではサンプル中で発現している遺伝子を，また相同性検索では進化の過程で高度に保存されている遺伝子を同定することはできるが，発現量が低い（または発現していない）遺伝子や保存度の低い遺伝子，その生物種に特異的な遺伝子などは検出できない．そのような遺伝子まで含めて網羅的に遺伝子予測を行いたい場合には*ab initio*遺伝子予測が有効である．*ab initio*遺伝子予測ではコード領域に見られる塩基組成の偏りやスプライス部位などの機能領域における出現塩基の偏りを利用して遺伝子を予測する．これらの遺伝子統計量の学習から遺伝子予測までを行えるツールには，Augustus[29]やSNAP[30]などがある．Augustusでは，RNA-seqのデータや相同性検索の結果を予測のためのヒントとして用いることも可能で，予測精度の向上が期待できる．

最後に

NGSの普及によりゲノムデータの利用は身近なものとなり，データ解析のためのツールも日々新しいものが開発され，公開されている．これらを用いれば，誰もが手軽に最先端の解析を行うことができるが，精度の高い解析結果を得るには手法の特性を理解することはもちろん，解析に応じたデータセットを用意することも重要である．本稿が，NGSという強力なゲノム解析ツールを使いこなすための一助となれば幸いである．

文　献

1) Hayden, E. C.: *Nature*, **507**, 294 (2014).
2) Li, R. *et al.*: *Nature*, **463**, 311 (2010).
3) Zerbino, D. R. and Birney, E.: *Genome Res.*, **18**, 821 (2008).
4) Gnerre, S. *et al.*: *Proc. Natl. Acad. Sci. USA*, **108**, 1513 (2011).
5) Luo, R. *et al.*: *GigaScience*, **1**, 18 (2012).
6) Kajitani, R. *et al.*: *Genome Res.*, **24**, 1384 (2014).
7) Andrews, S.: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2017/0523).
8) FASTX-Toolkit: http://hannonlab.cshl.edu/fastx_toolkit/ (2017/0523).
9) Martin, M.: *EMBnet.journal*, **17**, 10 (2011).
10) Koren, S. *et al.*: *Genome Res.*, **27**, 722 (2017).
11) Chin, C-S. *et al.*: *Nat. Methods*, **13**, 1050 (2016).
12) Chin, C-S. *et al.*: *Nat. Methods*, **10**, 563 (2013).
13) Miller, J. R. *et al.*: *Bioinformatics*, **24**, 2818 (2008).
14) English, A. C. *et al.*: *PLoS ONE*, **7**, e47768 (2012).
15) Grabherr, M. G. *et al.*: *Nat. Biotechnol.*, **29**, 644 (2011).
16) Schulz, M. H. *et al.*: *Bioinformatics*, **28**, 1086 (2012).
17) Robertson, G. *et al.*: *Nat. Methods*, **7**, 909 (2010).
18) Xie, Y. *et al.*: *Bioinformatics*, **30**, 1660 (2014).
19) Roberts, A. *et al.*: *Bioinformatics*, **27**, 2325 (2011).
20) Mezlini, A. M. *et al.*: *Genome Res.*, **23**, 519 (20313).
21) Pertea, M. *et al.*: *Nat. Biotechnol.*, **33**, 290 (2015).
22) Li, H. and Durbin, R.: *Bioinformatics*, **25**, 1754 (2009).
23) Langmead, B. and Saizberg, S.: *Nat. Methods*, **9**, 357 (2012).
24) Kim, D. *et al.*: *Genome Biol.*, **14**, R36 (2013).
25) Wu, T. D. and Nacu, S.: *Bioinformatics*, **26**, 873 (2010).
26) Dobin, A. *et al.*: *Bioinformatics*, **29**, 15 (2013).
27) Slater, G. S. and Birney, E: *BMC Bioinformatics*, **6**, 31 (2005).
28) Birney, E. *et al.*: *Genome Res.*, **14**, 988 (2004).
29) Stanke, M. *et al.*: *BMC Bioinformatics*, **7**, 62 (2006).
30) Korf, I.: *BMC Bioinformatics*, **5**, 59 (2004).