猫も杓子も

Geneious Prime でシークエンス解析

第 25 回 De novo アセンブリ（その 5：Tips）

Geneious Prime で行うことができる解析のうち、de novo アセンブリは最もマシンパワーが要求されるものの 1 つです。de novo アセンブリに必要となるハードウェアの詳細については、こちらのナレッジベースの情報をご参照ください。

アセンブリの品質を向上させ、アセンブリに必要な時間と RAM を削減するためには、常に次のことを行う必要があります。

1. BBDuk を使用して低クオリティな配列をトリミングします。Q 値 20 以上で厳密にトリミングを行います。

2. アセンブリのカバレッジは 50～100 倍が目安です。これ以上のカバレッジはアセンブリ結果を改善せずに、アセンブリに必要な時間と RAM だけが増加してしまうことが多いです。推定ゲノムサイズ、平均リード長、リード数から推定カバレッジを計算する方法については、こちらのリンクをご参照ください。

リードのサブセットを de novo アセンブリするためには、次のオプションのいずれかを使用します。

a) De novo アセンブリの設定ウィンドウで、**Use X% of data** オプションをチェックし、カバレッジ計算に基づいて適切な値を設定します。これでリスト内の最初から X%のリードが使用されます。

b) **Workflows → Randomly sample sequences** を使用して、リードリストからランダムにサンプリングされたサブセット（ペア）を作成します。

c) Normalization を使用してデータセットのサイズを小さくします。

de novo アセンブリで完全な連続ゲノムは得られますか？

de novo アセンブリアルゴリズムは、その手法にかかわらず、ゲノム上のリピート領域がシークエンスリード長やペアリードのインサートサイズより長い場合、完全なリピート配列を明確にアセンブリすることができません。これは現実的には、イルミナショートリードデータによるゲノムアセンブリでは、ほとんどの場合、オーバーラップしないコンティグが複数作成されることを意味します。

例えば、微生物ゲノムは通常ほぼ完全なリピート配列を含む SSU(16S)、LSU(23S)や、5S rRNA 遺伝子といった rRNA 遺伝子クラスターのコピーを複数含んでいます。これらのほぼ同一配列の rRNA クラスターは、単一の連続したコンセンサス配列のアセンブリの障害となります。ほとんどの場合、他のリピートユニット(重複遺伝子、トランスポゾンなど)も、de novo アセンブリにさらなる「切れ目」を生じさせることになります。

しかしリピート配列に由来するすべてのリードは、どれか 1 つのコンティグの末尾で一緒にアセンブリされることになります。このようにアセンブリされたリピート領域のカバレッジは、結果としてゲノムのユニーク部分の平均カバレッジよりも高くなるために見分けることが可能です。

このような問題を解決する手段の一つが、PacBio や Oxford Nanopore のロングリードテクノロジーを組み合わせることです。特に PacBio の HiFi リードは 99.99%の精度で最大 25 kb のリード長を得ることができるため、リピート配列を含むゲノムの de novo アセンブリには最適な選択肢の一つです。Geneious Prime では Flye アルゴリズムによるロングリードのアセンブリ、SPAdes アルゴリズムによるロング/ショートリードのハイブリッドアセンブリに対応しています。

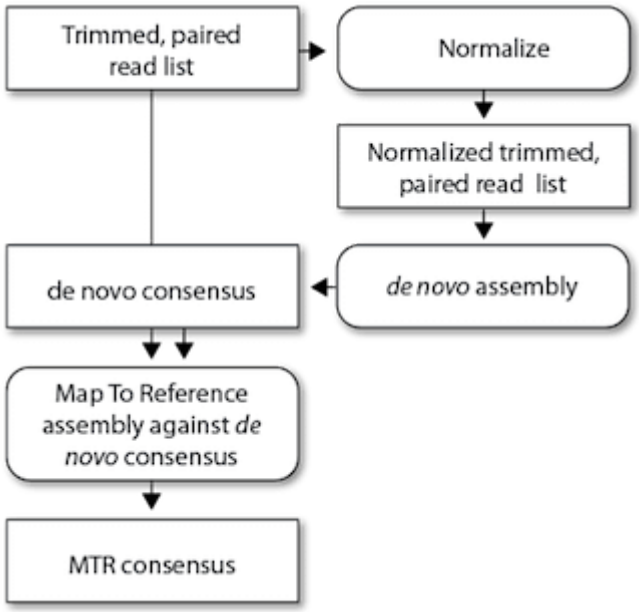
リードのエラー補正と正規化

Sequence メニュー → Error correct and Normalize reads から行うことができます。

Error correct and Normalize read ツールは BBNorm を使用しています。BBNorm は、ゲノムの高デプス領域のリードをダウンサンプリングしてアセンブリのカバレッジを正規化し、カバレッジ分布をより均一にするよう設計されています。この正規化ではカバレッジの低い領域のリードは削除されないというのが重要です。

正規化により、データセットのサイズを大幅に縮小することができ、その後の de novo アセンブリでは、必要な RAM とアセンブリ時間を削減することができます。

正規化によって、シークエンシングが困難な領域のエラーが増幅され、エラーが大きく見えるようになってしまう可能性があることにご注意ください。
そのため、正規化を使用する場合は、右のフロー図に示すようなストラテジを採用されることをお勧めします。 この正規化/de novo アセンブリ/リファレンスへのマップの組み合わせ手法は、通常、全データセットの de novo アセンブリを試みるよりもはるかに速いです。



ペアリードのマージ

Sequence メニュー → Merge paired reads から使用することができます。
このツールは BBMerge を使用し、オーバーラップする 2 つのペアリードを 1 つのリードにマージするよう設計されています。アンプリコンシーケンスで生成されたオーバーラップリードからコンセンサスを生成するのに便利なツールです。

重複したリードの削除

Sequence メニュー → Remove duplicate reads から使用することができます。
Dedupe を利用してデータセットに含まれる重複する配列をすべて見つけ出し、削除するツールです。

キメラの除去

Sequence メニュー → Remove chimeric reads から使用することができます。
リファレンスデータベースと比較することで、シークエンスデータからキメラリードをフィルタリングするツールです。パブリックドメインの UCHIME アルゴリズムか、より高速な USEARCH 8 をダウンロードして使用することができます。無料版の USEARCH 8 は使用できる RAM が 4GB までに制限されており、大きな NGS データセットを扱うことができないことにご注意ください。

バーコードによる分割

Sequence → Separate by barcodes メニューから使用することができます。
このツールはカスタムバーコードデータを別々のリストにデマルチプレックスします。このツールには 454 MID バーコードのプリセットがあり、また、独自のカスタムバーコードセットを定義して使用することもできます。
注: デマルチプレックスする場合は、BBduk を使用してトリミングする前に必ず実行する必要があります。

NGS リードの de novo アセンブリについては、今回がひとまずの最終回となります。
次回からはリファレンス配列がある場合のマッピングをご紹介する予定です。

Geneious 製品概要・フリートライアルリクエストについてはこちら
『Geneious Prime で猫も杓子もシークエンス解析』 過去の記事はこちらでチェック！

トミーデジタルバイオロジー株式会社
Phone 03-6240-0843 Fax 03-6240-0461