猫も杓子も

Geneious Prime でシークエンス解析

第 21 回 De novo アセンブリ（その 1：前処理の概要）

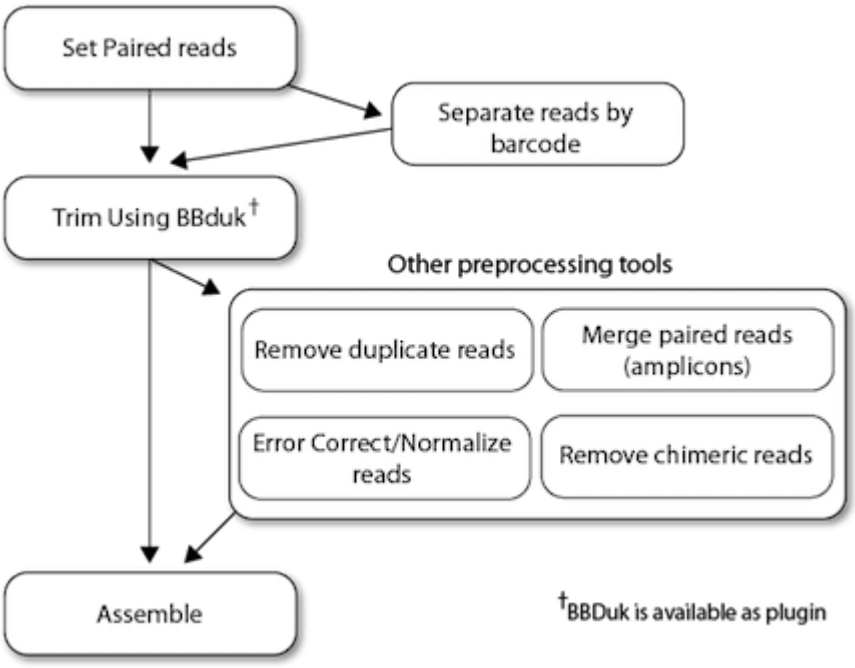
今回から数回にわたり、NGS リードを適切に前処理し、de novo アセンブルする一般的な流れについてご紹介します。

Geneious Prime には、独自の Geneious de novo アセンブラが搭載されていますが、SPAdes, Tadpole, Velvet, MIRA などのサードパーティ製 de novo アセンブラもプラグインで導入することができます。他のアセンブラの詳細については、Which *de novo* assembler is best for my data をご参照ください。

初回となる今回は、アセンブリの前に行うべき NGS リードのペアリング、トリミング、フィルタリングなどの前処理ステップの概要についてです。

NGS リードの前処理を適切に行うことは非常に大切で、アセンブリに必要な計算/時間コストを大幅に削減し、アセンブリの成功率と精度を向上することに繋がります。

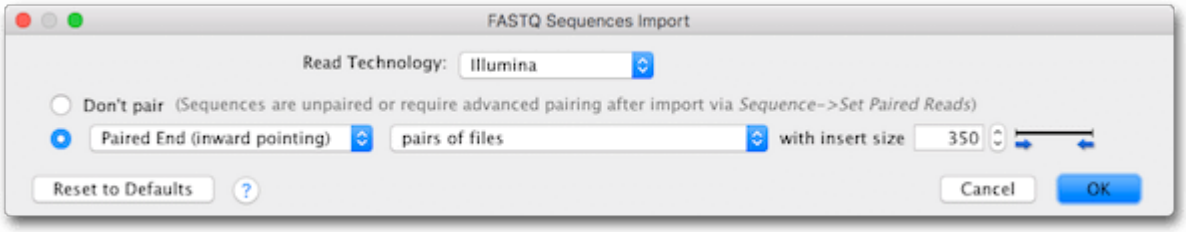
ペアリングされたデータがある場合、最初のステップは **Set paired reads** です。次にトリミングを行い、必要に応じて次のフローに示すような他の前処理ステップを行う必要があります。



通常、NGS のサービスプロバイダーは illumina のペアリードデータを 2 つの別々（フォワードリードとリバースリード）の fastq フォーマットで提供します。また多くの場合、この.fastq は.gzip(.gz)により圧縮されていますが、Geneious Prime は圧縮または非圧縮のどちらでもインポートできます。

File → From Multiple files でフォワードリードとリバースリードのファイルを一緒にインポートすると、Geneious Prime は 2 つのファイルをペアリングして 1 つのペアリードリストを作成することを提案します。同様に、ペアリードのファイルを Geneious Prime のウィンドウにドラッグ&ドロップした時も、インポート中にリードをペアリングするオプションが表示されます。

リードテクノロジーは Geneious Prime が自動的に判別しますので、予想インサートサイズ(アダプターを除いた予想平均インサートサイズ)を設定し、**OK** を押すだけです。



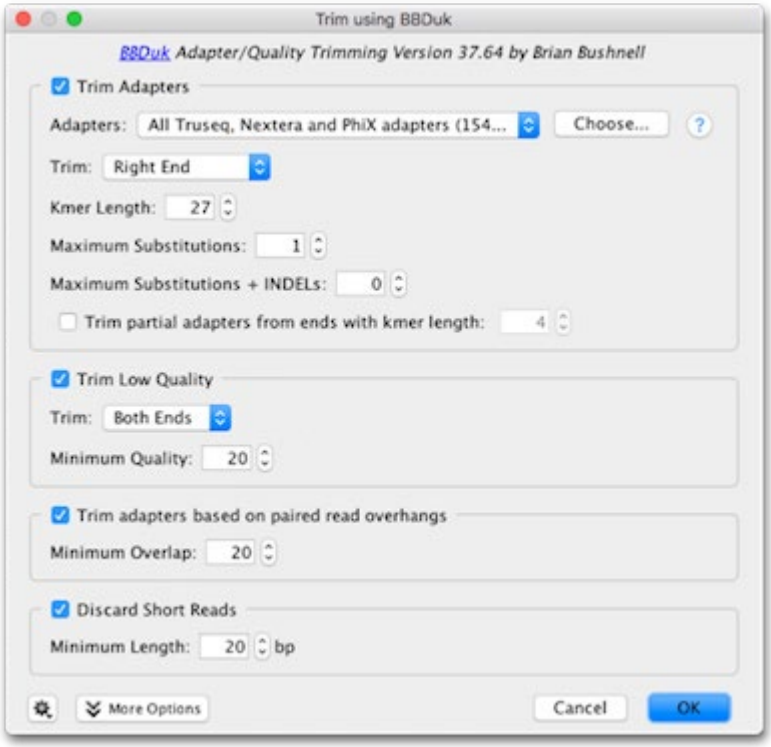
ペアリングの出力は、インターレースされたフォワードリードとリバースリードのリストになります。

もし、フォワードリードとリバースリードをすでに別々のリストとしてインポートしている場合は、それらのリストを選択し、**Sequence → Set paired reads** でペアリングすることができます。

次のステップはトリミングです。illumina リードの末端にある低クオリティなコールは、正しいアセンブリを阻害し、アセンブリ実行に必要な計算/時間コストを増加させる可能性があるため、アセンブリの前にリードをトリミングすることが重要です。

Geneious Prime では BBDuk をトリマーとして使用することを推奨しています。BBDuk (**D**econtamination **U**sing **K**mers) は、NGS リードのトリミングとフィルタリングのための高速かつ正確なツールで、illumina アダプター用のプリセットを使用したアダプターのトリミング、クオリティによるトリミング、ペアリードのオーバーハングに基づくアダプターのトリミング、設定した長さ以下にトリミングされたリードの除去ができます。

Geneious Prime 2023 以降のバージョンでは BBDuk は標準でバンドルされており、**Annotate & Predict → Trim using BBDuk** からアクセス可能です。Geneious Prime 2022 以前のバージョンをお使いの場合は、**Tools → Plugins** からプラグインをインストールする必要があります。



多くの場合、標準的な illumina アダプターの配列はサービスプロバイダーによってすでにトリミングされていますが、もし必要な場合は **Trim Adapters** で設定することができます。

Minimum Quality で設定する Q 値は Phred スコア(modified Mott アルゴリズム)です。
以下の表は Q 値と%尤度(Likelihood)の相関例を示しています。 適切な設定値はデータの全体的なクオリティによって異なりますが、一般的に設定値を高くしてトリミングを行うと、リードのかなりの部分がトリミングされてしまわない限りはアセンブリの速度と品質が向上します。 illumina リードの場合、20～30 に値を設定することをお勧めします。(Nanopore リードの場合は 6 が推奨値です)

Q value	% Likelihood call will be correct
6	75
10	90
13	95
20	99
30	99.9

次回はサンプルデータを用いた前処理の実例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 でシークエンス解析』 の過去の記事はこちらでチェック！

TDB News 7. 2023
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