猫も杓子も

Geneious Prime でシークエンス解析

第 9 回　アンプリコンメタゲノミクス（その 1）

メタゲノミクスは、環境サンプルから直接回収された遺伝物質の研究です。今回から数回にわたり、発酵プロセスに関連する細菌群をプロファイルするため、自然発酵したザワークラウトから PCR 増幅された 16S rRNA 遺伝子配列を解析する手法をご紹介します。

今回はメタゲノムアンプリコンデータの前処理についてです。解析前の準備段階ですが、解析の最後にまで大きく影響する可能性のある重要な部分になります。解析例として Short Read Archive (SRA)の SRR7140083 データセットを用います。このデータセットは、16S rRNA アンプリコンの V4 領域（約 260bp）について、Illumina MiSeq の 2 x 250 bp ペアリードで得られたものです。

アンプリコンデータから信頼性の高い分類を行うための鍵の一つは、適切にキュレートされたリファレンス配列のデータベースを使用することです。今回の例では、まず配列のトリミング、フィルタリング、de novo アセンブラを用いた OTU へのクラスタリングを行います。次に代表的な配列を NCBI の 16S Microbial データベース（バクテリアとアーキアの 16S 配列のキュレーションセット）に BLAST します。最終的に BLAST 結果は、Sequence Classifier プラグインでリードセットを分類するためのターゲットデータベースとして使用されます。

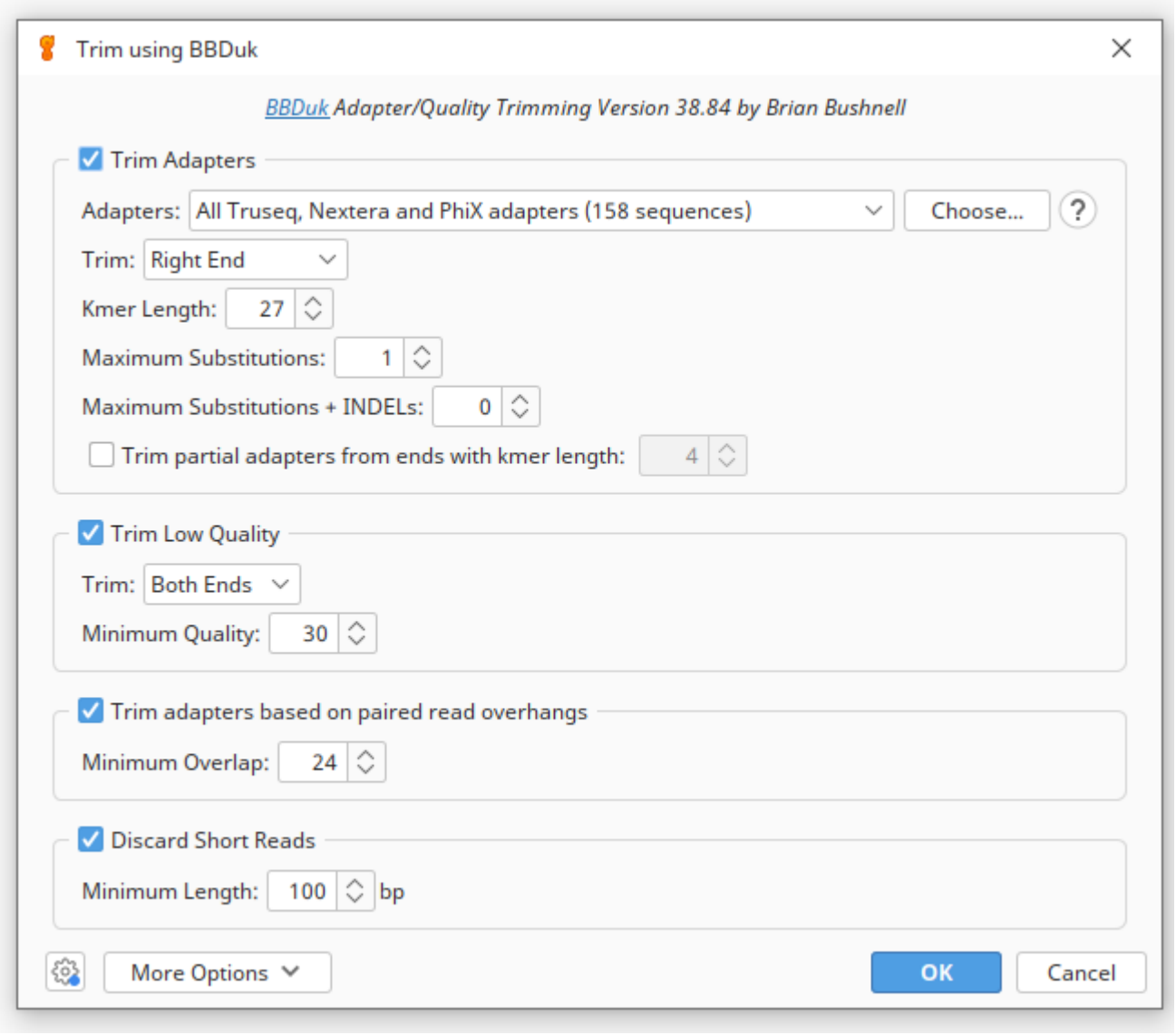
また、ご紹介する例では 16S を対象としていますが、BLAST 用に適切にキュレートされたデータベースがあれば、18S、ITS、CO1 など、他のメタゲノムマーカーにも適用することができます。

Tutorials → Metagenomics Analysis フォルダにある SRR7140083_50000 ファイルを練習用に用いることもできます。これは SRR7140083 全データからのサブセットで、50,000 の 16S アンプリコンペアリードが含まれています。

イルミナのペアリードでは通常、フォワードとリバースのリードが別々に fastq 形式のリストとして提供されます。これらを同時にインポートした場合、Geneious はそれらをペアリングし、1 つのペアードリードリストを作成することができます。また、別々にインポートしたリストも **Sequence → Set Paired Reads** でペアリングすることができます。SRR7140083_50000 ファイルはすでにペアリング済みのリードリストです。ペアリードであることは各リード名の左側に記号で表されています。

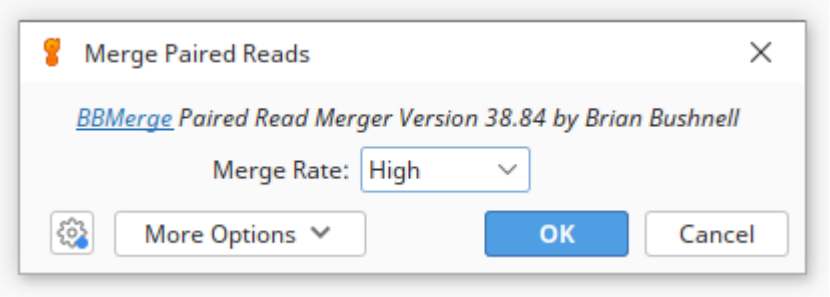
アンプリコンメタゲノムリードでは、PCR やシーケンシングエラーによる配列のわずかな違いを実際の変異と間違えないために、クオリティによるトリミングが非常に重要です。トリミングには、Geneious デフォルトのトリマー（Trim Ends）よりも多くの機能を持つ BBDuk プラグインを使用することをお勧めしています。

はじめに、**Annotate and Predict → Trim using BBDuk** で、以下のスクリーンショットのようにパラメータを設定します。これはリード中に残っている Illumina アダプター配列、クオリティスコア 30 以下の塩基をトリミングし、トリミング後に 100 bp 以下となるリードを除去するものです。

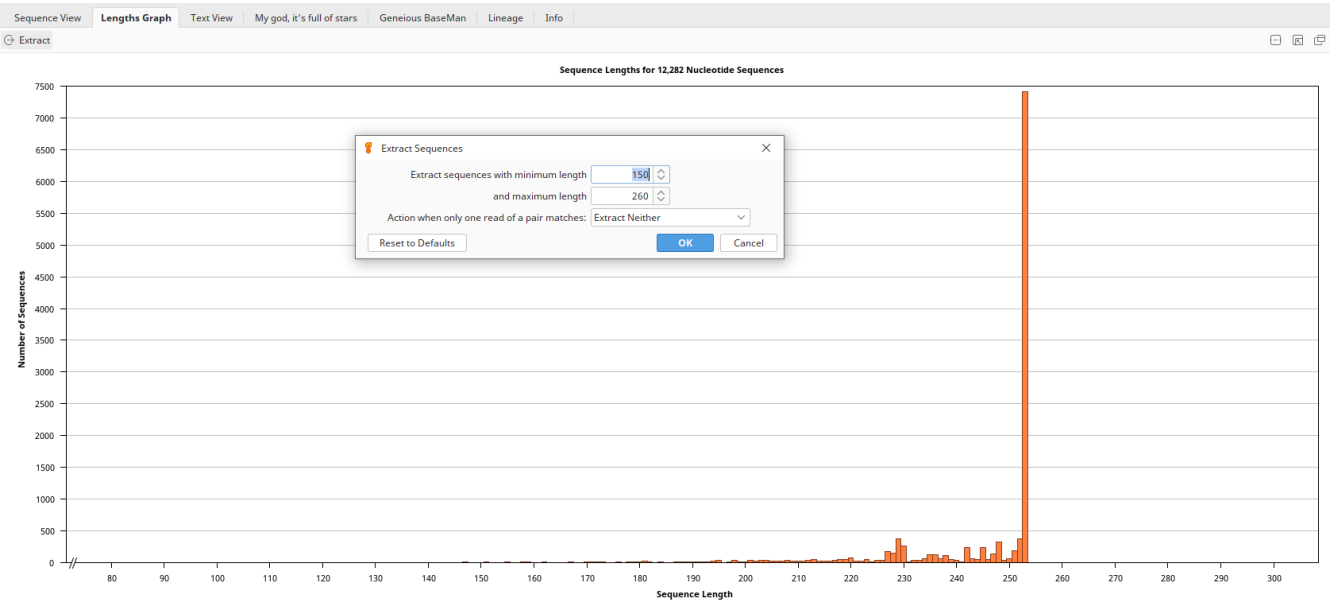


トリミングが完了すると、**SRR7140083_50000(trimmed)**というファイルがドキュメントテーブルに表示されます。このファイルには約 25,000 の配列が含まれており、クオリティの低いデータが相当量削除されたことがわかります。もし Trim Low Quality のしきい値を 30 とした設定ではデータが削除されすぎてしまう場合は 20 に下げ、代わりに Discard Short Reads のしきい値を 150 bp に上げて、クラスタリングと BLAST が 16S 配列のなるべく長い部分に基づいて行われるようにすることをお勧めします。

次にこのデータの場合、増幅された 16s rRNA の領域はプライマーとアダプター配列を除くと約 250 bp ですが、F と R のリード長も 250 bp であるためオーバーラップしており、それぞれのリードペアで単一のコンセンサス配列を作成するためにマージすることができます。リードのマージには、BBMerge ツールを使用します。トリミングしたリードセットを選択し、**Sequence → Merge Paired Reads** で、下図の設定(Merge Rate: High)で **OK** をクリックします。



マージできなかったリード SRR7140083_50000 (trimmed) (couldn't be merged)と、マージされたリード **SRR7140083_50000 (trimmed) (merged)**が作成されます。このファイルをクリックし、ビューアーの上側にある Lengths Graph タブを見ると、マージ後のいくつかの配列が、予想されるプロダクトのサイズ(約 250 bp)よりもずっと短いか長いことがわかります。長い配列はコンタミネーションか、間違ってマージされた配列の可能性がありますので、これらを削除します。また、非常に短い配列は、正しく分類するために十分な長さの配列が含まれていないため、削除します。残したいリードを抽出するためには、Lengths Graph で Extract ボタンをクリックし、150～260bp の配列を抽出します。このデータの場合には約 12,000 のリードが含まれているはずです。



また、必須ではありませんが、データセットによってはこの段階でキメラリードを削除することで、より良い解析結果となることがあります。Sequence メニューにある Remove Chimeric Reads オプションでは、ご自身でデータベース(例:RDP-Gold)をご用意いただいて、リファレンスベースの UCHIIME を実行することができます(de novo モードはサポートしていません)。また、より高速な解析のために USEARCH を使用することも可能です。Geneious 内のツールではありませんが、de novo またはリファレンスアプローチで動作する VSEARCH も代替となります。

次回は de novo アセンブラを使用してリードを OTU (Operational Taxonomic Unit)にクラスタリングするステップをご紹介する予定です。

Geneious 製品概要についてはこちら

TDB News 4. 2022
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