

猫も杓子も

Geneious Prime でシーケンス解析

第 27 回 リファレンスマッピングと SNP 検出 (その 2)



前回:[第 26 回 リファレンスマッピングと SNP 検出 \(その 1\)](#)より、NGS リードをリファレンス配列にマッピングし、SNP を検出するための一般的なワークフローについてご紹介しています。2 回目となる今回は、マッピングで作成されたコンティグの操作についてです。

今回もチュートリアル用データが用意されていますので、実際に操作して試してみたい方は、[こちらからダウンロード](#)し、ZIP ファイルを解凍せずにそのまま Geneious Prime にドラッグ&ドロップしてインポートしてください。

マッピングで作成されたコンティグドキュメント(チュートリアルデータの場合は **yghJ paired Illumina reads (trimmed) assembled to yghJ CDS (divergent reference)**)をクリックして開き、リードがリファレンス配列にどのようにマッピングされているかを確認します。

シーケンスビューアの右側にある **Advanced settings** タブ(歯車マーク)で、**Vertically compress contig** にチェックすることでリードが横一列に表示されます。

General タブ(家マーク)で、**Colors** を Paired Distance に設定することで、ペアエンドリードをセットアップする際に指定したインサートサイズに基づいて、ペアエンドリードが予想通りの距離でマッピングされているかどうかを色で確認することができます。チュートリアルデータのコンティグでは、ほとんどのリードが緑色になっており、ほぼ予想通りのインサートサイズでマッピングされていることがわかります。

またシーケンスビューアの上にある **Insert sizes** タブをクリックすると、実際のインサートサイズの分布を確認することができます。チュートリアルデータのコンティグでは、ほとんどのペアが 450-500 bp のインサートでマップされ、予想される 500 bp のサイズに近いことがわかります。

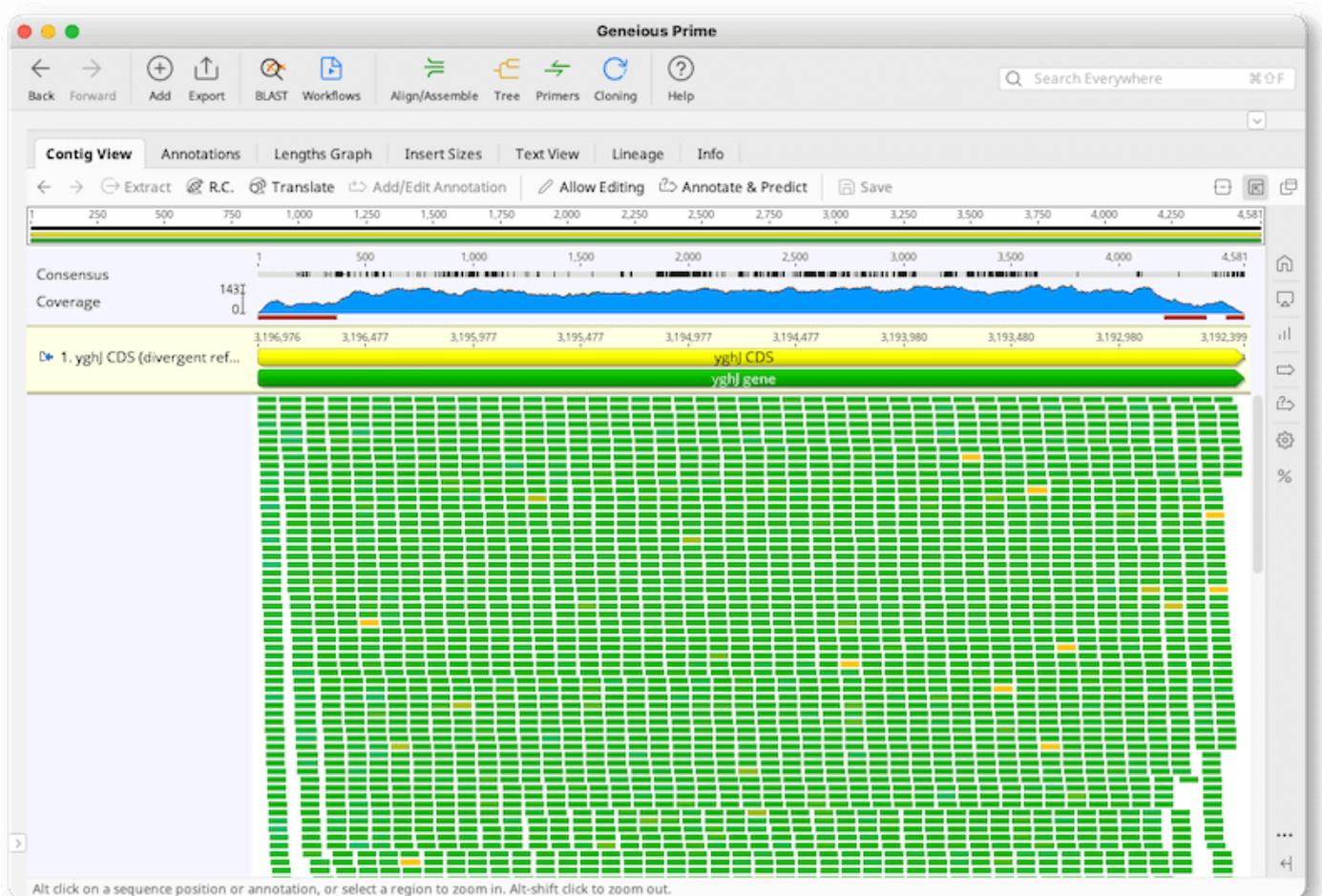
Contig View タブに戻すと、コンティグの上部に **Consensus** が表示されています。はじめは最大にズームアウトされているため配列を確認することができませんが、シーケンスビューアの右にあるサイドパネルの上に表示されているズームコントロールを使ってズームインすることで、塩基配列を確認することができます。

表示されている塩基配列はマッピングされたリードから作成されたコンセンサス配列であり、リファレンスの配列ではありません。このコンセンサス配列のコールに使用する設定は、シーケンスビューアの右側にある **Display** タブ(モニタマーク)で設定します。通常、NGS リードにはクオリティスコアが

ついているので、コンセンサス配列をコールするための **Threshold** として、デフォルトでは **Highest Quality** が選択されています。この設定では、その位置にある各塩基の相対的なクオリティスコアを考慮して多数決でコンセンサスを算出します。Threshold の設定に関する詳細は[マニュアル](#)をご参照ください。

Threshold を 100% – Identical に変更すると、配列の中に混合塩基が表示されます。この設定では、例えば 1000 リードのうち 1 リードにだけ異なる塩基が含まれていたとしても、塩基が混在しているリードには混合塩基が挿入されます。ただし NGS リードのマッピングの場合には、これは実際の多型を表すものではなく、シーケンシングエラーの結果であることが多いため、通常は使用しないでください。もしリードにクオリティスコアがない場合に、実際の多型のみをコンセンサス配列に混合塩基として表示したい場合には、90-99%の Threshold が最も適切です。

また **Consensus** の下には、下のスクリーンショットのように青い **Coverage** グラフが表示されます(もし表示されていない場合は、シーケンスビューアの右側にあるグラフタブをクリックし、Show graphs と Coverage にチェックして有効にしてください)。Coverage グラフは、各塩基に何本のリードがマッピングされたかを示すもので、マッピングの質を評価するのに便利です。



Geneious Prime では、カバレッジの高い領域や低い領域を素早く特定するために 2 つのツールをご用意しています:

1. **Graphs** タブで、特定のカバレッジ以上/未満の領域をハイライトすることができます。例えば **Highlight above** にチェックし、50 に設定すると、Coverage グラフの下に黄色いバーが表示され、カバレッジが 50 を超える領域をカバーするようになります。

2. **Annotate and Predict** → **Find Low/High Coverage** で、カバレッジの高い領域や低い領域にアノテーションを付けることができます。

SNP を検出する時にこれらの領域を除外できるよう、このオプションを使用してカバレッジの低い領域にアノテーションを付けてみます。**Find regions with coverage below** をチェックし、**Standard deviations from the mean** は 2 を設定します。**Merge regions** オプションは両方チェックし、**High Coverage** オプションはチェックを外して **OK** をクリックします。これで、リファレンス配列に Coverage アノテーショントラックが表示され、低カバレッジ領域にアノテーションが表示されます。**Save** をクリックし、元の配列に変更を適用するため、**Yes** を選択します。このアノテーションは次々回に使用する予定です。

Find High/Low Coverage

Only Find In

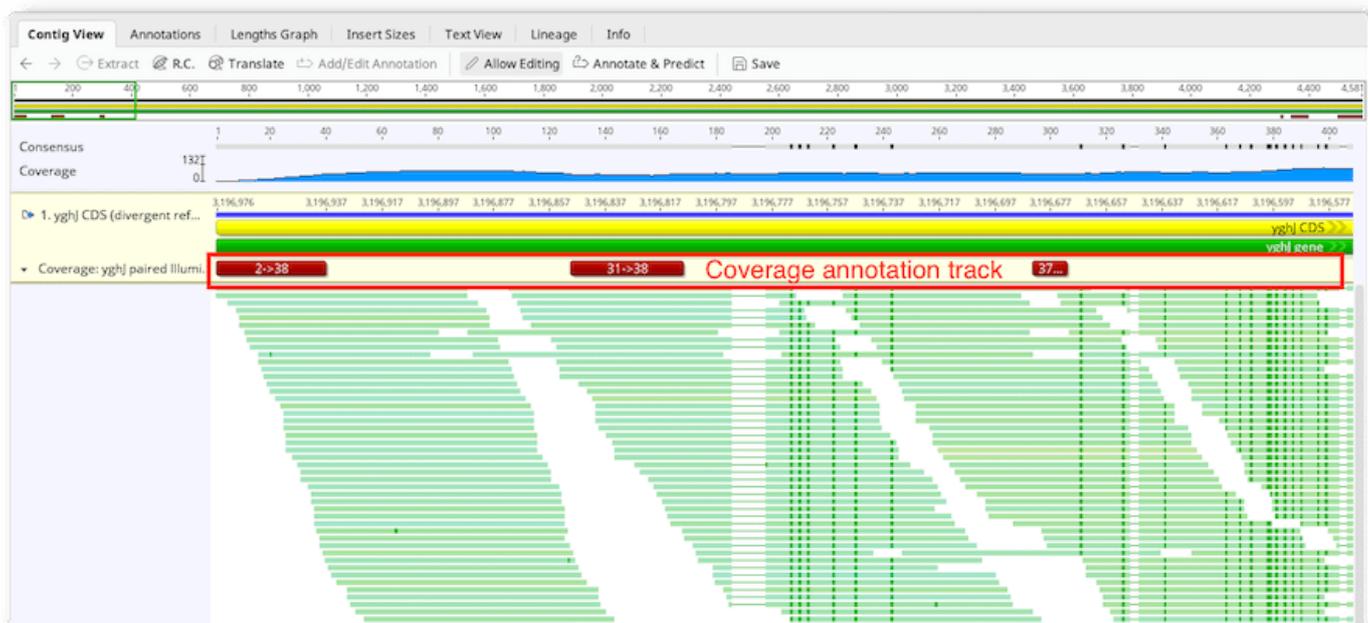
- Selected region
- Annotations in reference sequence of type
- Create coverage annotations on reference sequence
- Create same type of annotations on consensus sequence

Low Coverage

- Find regions with coverage below
 - Standard deviations from mean:
 - Number of sequences:
 - Merge regions if coverage between them is within of low coverage
 - Merge low coverage regions less than bases apart

High Coverage

- Find regions with coverage above
 - Standard deviations from mean:
 - Number of sequences:
 - Merge regions if coverage between them is within of high coverage
 - Merge high coverage regions less than bases apart



今回はコンティグから SNP を検出する手法をご紹介します。

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!](#)

TDB News 1. 2024
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