Working with Circular Reference Sequences in NextGENe® Software

John McGuigan, Ni Shouyong, CS Jonathan Liu

Introduction

Alignment of reads to circular reference genomes poses a particular challenge because most alignment algorithms treat sequences as a single linear string of bases. NextGENe v2.4.0 introduces new features that allow easier analysis with circular references like mitochondrial and bacterial genomes. GenBank (.gbk) files may be annotated as circular sequences within the file. These can be used to build a new amplicon reference or they can be used directly for whole (circular) genome sequencing.

Procedure

Amplicon Sequencing

The "Build Preloaded Reference" tool in the NextGENe Tools menu may be used to build an amplicon reference (figure 1).

1. Enter a name for the reference.
2. Click "Add Files" and select a circular GenBank reference file
3. Select "Create index based on BED file(s)"
4. Click "Add BEDs" and select one (or multiple) BED file(s) that define amplicon positions
5. Choose "Merge BED Overlaps" to combine overlapping regions, or uncheck it to treat each amplicon individually.
   - When overlaps are not merged it is possible for the same position to occur in multiple amplicons. These will be treated as separate positions in the mutation report, so the same variant may be listed multiple times.
6. Click "Build Index". The new reference will now appear in the list of preloaded references. Figure 2 shows alignment results against one such reference (with unmerged overlapping amplicons).

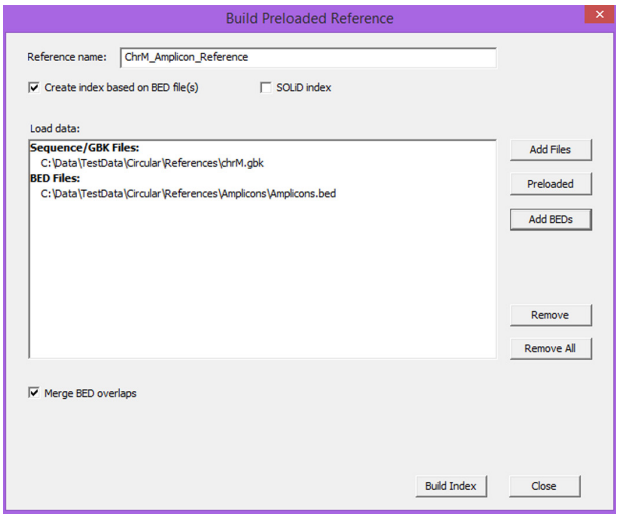

Figure 1: The Build Preloaded Reference dialog

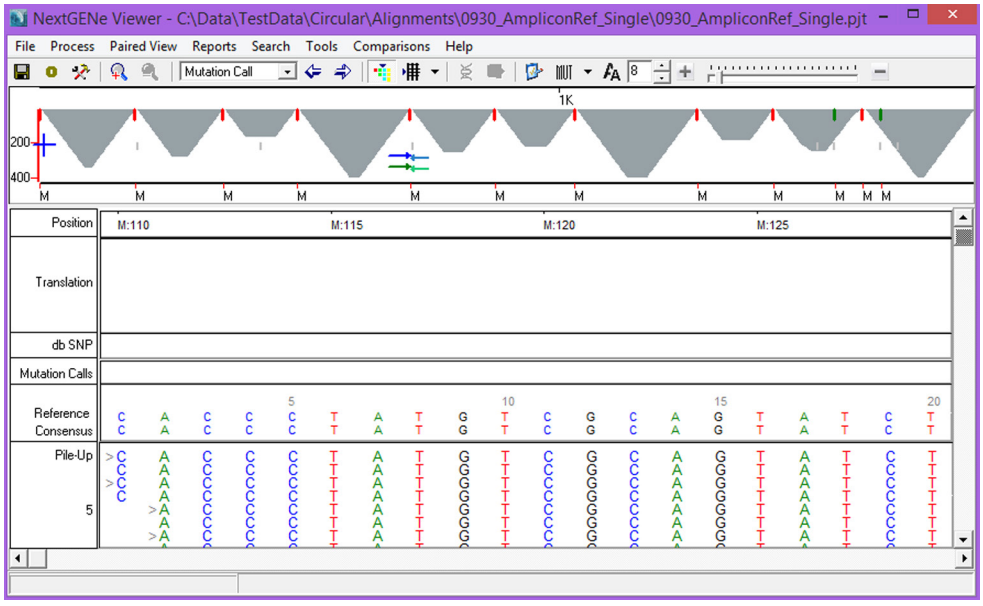

Figure 2: Alignment of simulated data to an amplicon reference. Red lines and green lines can be seen in the top panel which also shows depth of coverage in gray. Red lines separate the contigs (amplicons) while green lines show the position of the reference origin within amplicons.

Procedure (cont.)

Whole Genome Sequencing

The same circular genome GenBank files can be loaded as a reference during alignment without creating a preloaded reference. When this happens, NextGENe will automatically add 1,000 bp to each end of the reference before alignment. After alignment is completed it will then correctly position reads that overlap the ends of the reference sequence. Figure 3 shows alignment results for the R23-81 dataset found on the Ion Community website. This sample is E. coli DH10B sequenced using a 314 chip on the Ion Torrent PGM.

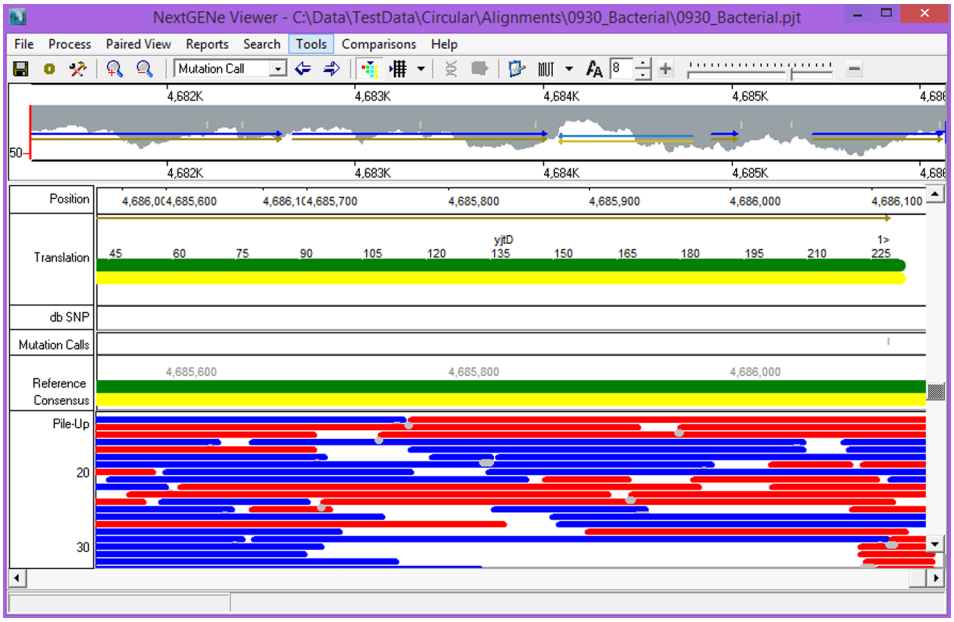

Figure 3: Alignment at one end of the reference E. coli DH10B genome. Some reads can be seen aligned to the end of the sequence - these continue at the beginning of the sequence because it is a circular genome.

Acknowledgements

We would like to thank Ion Torrent for providing sample data on the Ion Community Website.