

## Technical Note

# Creating SNPtrees in Geneious R10

# Creating SNPtrees in Geneious Prime R10



## Introduction

In epidemiological studies, isolates usually have very little genetic diversity and require full genomic analysis to accurately differentiate and classify.

Reducing a whole genome assemblies to a subset of statistically-validated variant sequence positions provides a condensed dataset that can be used for phylogenetic analysis and classification of isolates (1,2).

Geneious R10 includes a new Workflow that allows you to easily generate a whole-genome alignment suitable for generating phylogenetic trees based on single nucleotide variant information (SNPtrees).

This technical note provides an overview of how to generate a SNPtree in Geneious Prime. The data used in the overview is next-generation sequencing (NGS) data generated by the U.S. Army Medical Research Institute of Infectious Diseases (USAMRIID) from Ebola virus isolated during the 2014 West African Ebola outbreak. A published Ebola genome, NC\_002549, is used as the reference for assembly.

The data used in this example can be downloaded using the links given below in Table 1. Note that Ebola genome NGS datasets were chosen due to their relatively small (i.e. quick to download) sizes. However, the procedure described here will work equally well for larger genomes. The reference sequence used in this example, NC\_002549, can be downloaded directly from within Geneious Prime via the NCBI search tool in the Geneious Prime Sources panel.

**Table 1:** USAMRIID Ebola NGS datasets used in this example

Isolate	Source
LIBR0284	<a href="http://www.ebi.ac.uk/ena/data/view/SRR3152967">http://www.ebi.ac.uk/ena/data/view/SRR3152967</a>
LIBR0380	<a href="http://www.ebi.ac.uk/ena/data/view/SRR3152978">http://www.ebi.ac.uk/ena/data/view/SRR3152978</a>
LIBR10279	<a href="http://www.ebi.ac.uk/ena/data/view/SRR3152980">http://www.ebi.ac.uk/ena/data/view/SRR3152980</a>
LIBR11079	<a href="http://www.ebi.ac.uk/ena/data/view/SRR3152981">http://www.ebi.ac.uk/ena/data/view/SRR3152981</a>
LIBR11177	<a href="http://www.ebi.ac.uk/ena/data/view/SRR3152982">http://www.ebi.ac.uk/ena/data/view/SRR3152982</a>
LIBR0622	<a href="http://www.ebi.ac.uk/ena/data/view/SRR3152984">http://www.ebi.ac.uk/ena/data/view/SRR3152984</a>
LIBR0707	<a href="http://www.ebi.ac.uk/ena/data/view/SRR3152985">http://www.ebi.ac.uk/ena/data/view/SRR3152985</a>
LIBR010180	<a href="http://www.ebi.ac.uk/ena/data/view/SRR3153068">http://www.ebi.ac.uk/ena/data/view/SRR3153068</a>

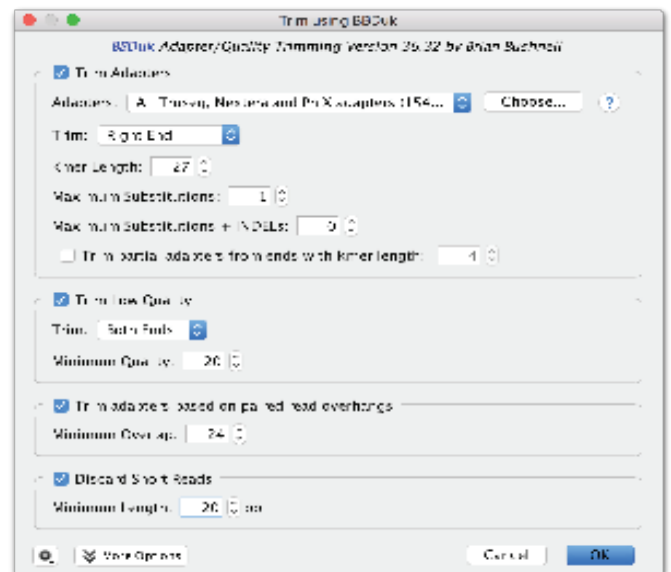
## Method

### Preprocessing of NGS Data

NGS read datasets for each isolate should be paired, trimmed and filtered appropriately prior to mapping to the reference sequence as follows:

1. Download the Illumina MiSeq FASTQ data using the links in Table 1 (download Files 2 & 3 only. File 1 contains upaired reads and can be ignored), then import the fastq files into Geneious Prime by dragging them into an appropriate folder in the Local data folder.

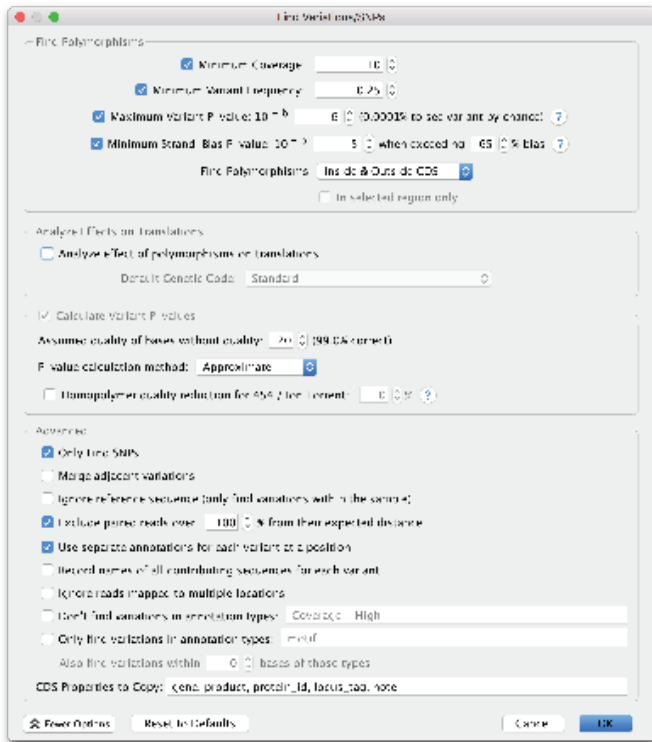
2. **Set Paired reads:** Select all pairs of read lists and use menu *Sequence* → *Set paired reads*, set an expected distance of 300 bp. This will create a new paired list for each data set.
3. **Trimming & Filtering:** Select all paired read lists and trim using menu *Annotate & Predict* → *Trim using BBDuk* (download the BBDuk plugin first if required) with the settings shown in Fig. 1.



**Figure 1.** BBDuk example settings for trimming paired read data.

## Assembly of NGS data sets

1. Perform a Map to Reference assembly for each paired/trimmed NGS list. You can assemble all lists to your reference in one step if you wish. To do this, select all paired-lists and your reference, go *Align/Assemble* → *Map to Reference* and check the option to *Assemble each sequence list separately*. In this example the default map to reference assembly settings were used.
  - Exclude repeat regions and potential incorrect mapping by setting "Map multiple best matches: To None" (Advanced option)
2. Select the contigs generated by map to reference assembly and go *Annotate & Predict* → *Find Variants/SNP's*. See Fig. 2. for the settings used in this example. Again, this step can be performed on all contigs in one step if you wish.
  - Ensure the option to *Analyse effect of polymorphisms on Translation* is unchecked (otherwise Overlapping CDS annotations will result in multiple variant annotations at single positions)
  - Ensure the option to *Use separate annotations for each variant at a position* is checked (required for the *Apply Variants to Reference Sequence* Workflow)



**Figure 2.** An example of *Find Variations/SNPs* settings used for identification of variants in each assembly.

Other settings to consider using to ensure you are identifying “real” variants.

- Only find SNPs (SNVs)
  - Exclude reads that don’t meet expected paired read distances
  - Exclude regions below a defined Minimum coverage threshold
  - Exclude polymorphisms with high strand bias
  - Exclude variants with low p-value
  - Turn on Homopolymer quality reduction for 454/Ion Torrent if analysing either of these “gappy” data types
3. Save the assembly with the new variant call annotations. Choose Yes when asked if you want to apply changes to the original reference sequence. This will apply the variant

call annotation tracks to the original reference sequence file. Variant annotations from each assembly will each be grouped as separate tracks on your reference sequence (see Fig. 3).

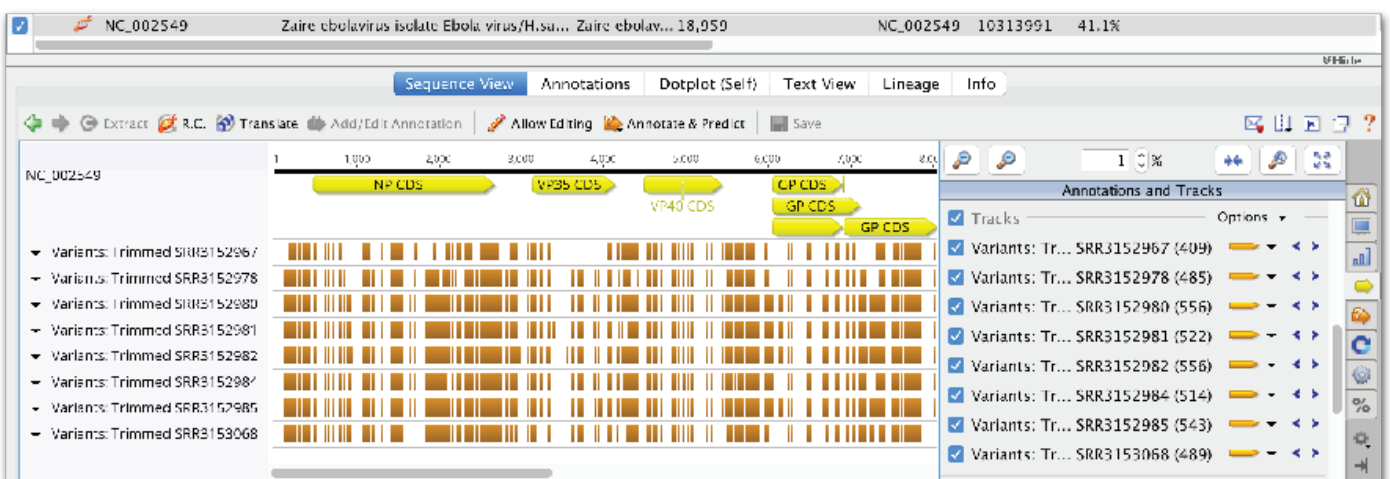
## Generating an alignment of Pseudogenomes

The reference sequence with its associated variant tracks can now be used to generate a “pseudogenome” alignment using the *Apply Variants to Reference Sequence* Workflow.

Each “pseudogenome” sequence is based on the reference sequence, altered to match the variant call at each variant position in each track. Each sequence will be named based on the original NGS dataset name. The final output of this Workflow is an alignment of “pseudogenome” sequences that can be for phylogenetic analysis.

To create a pseudogenome alignment:

1. Select your Reference sequence. Go *Workflow* → *Apply Variants to Reference Sequence*.
  - This Workflow provides a number of options, Check the option to Create alignment
  - If your assembly contains allelic variation, or is derived from a mixed population, then consider using the *Create ambiguities* option
2. Optional: If you wish you can then strip invariant or ambiguous positions from your alignment. To do this, select the alignment from the above operation and go menu *Tools* → *Mask Alignment* and select the option to *Save a copy with sites stripped* and set the option for *Sites containing: Identical bases*. This creates an alignment of all sequences comprising only those positions that differed in one or more sequences from the reference.
3. Repeat the stripping step with other criteria if required. For example, if the reference sequence contains ambiguities or N’s then these positions will be called as variants. The final stripped alignment will comprise only variant positions that differ from the reference in one or more sequences.
  - Note that stripping of invariant sites can adversely affect maximum likelihood, Bayesian, and distance-based tree builders, as the tree topologies inferred by these



**Figure 3.** Reference sequence with associated SNP tracks after assembly and variant calling.

methods are dependent on estimated branch lengths and observed substitution frequencies. Maximum parsimony methods (PAUP\*) do not consider branch length so can be used with stripped alignments

## Phylogenetic analysis

The initial BBDuk trimming step used in this example appended the word "Trimmed" to the read list names which then carried over to assemblies and also to our alignment of pseudogenomes sequences. Before performing a phylogenetic analysis we will rename the sequences in our stripped alignment so that sensible names are displayed on the tree.

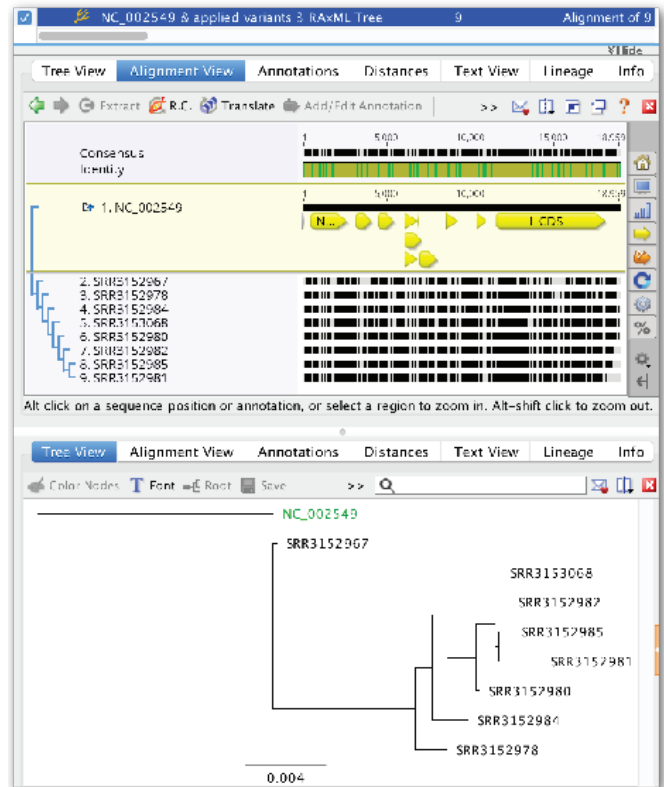
Select your alignment go menu *Edit -> Batch/Rename*. Set *Aspect to rename*: to *Sequences in alignment*. Set *Property to rename*: to *Name of Sequence*. Select *More Options*. Check *Advanced*. Set *Replace parts matching*: to "Trimmed" (include a trailing space) and leave the *Set With*: field empty. This will trim the pseudosequence file names.

Finally, select your alignment and analyse with the tree building algorithm of your choice (See Fig. 4.).

You can learn about the assembly and tree-building operations used in this note in more detail by downloading and running through the tutorials found at [www.geneious.com/tutorials](http://www.geneious.com/tutorials).

## References:

1. Leekitcharoenphon *et al.* (2014) PLOS One 9:e87991
2. Kaas *et al.* (2014) PLOS One 9:e104984



**Figure 4.** Split view of RaxML bootstrap tree (200 replicates) generated using an unstripped pseudogenome alignment containing a total of 610 unique SNV positions.