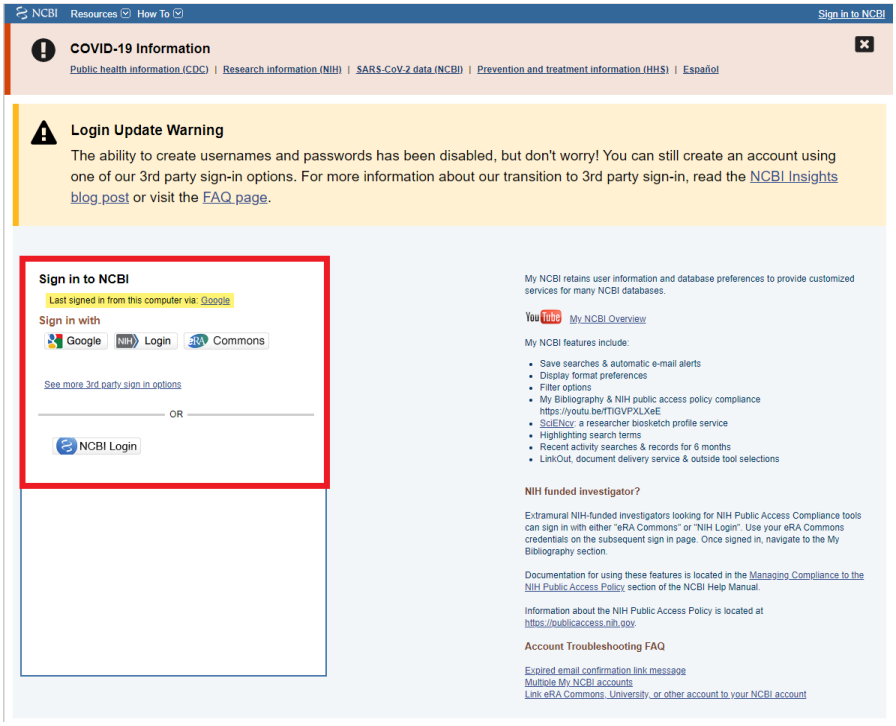
Being a **BLAST**ED Geneious

Table of Contents

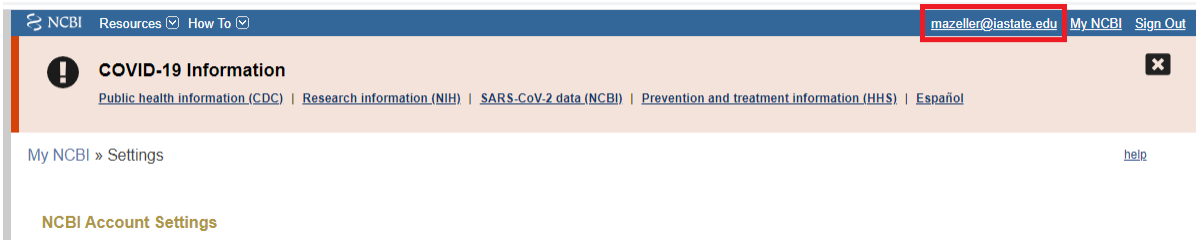
Setting up an NCBI account/API Key

Keys only work for Geneious 11 and above. R7 and R10 users will not have access to this. Currently, API Key Holders can make 10 requests to NCBI per second. Non-API key holders can only make 3 requests per second.

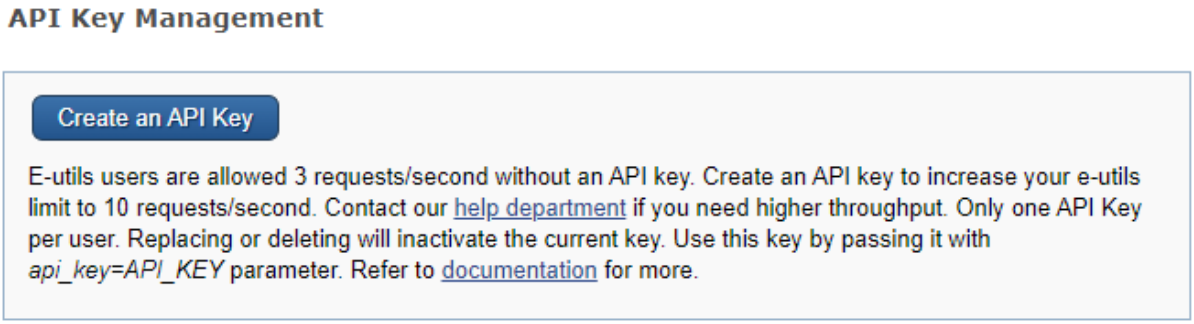
1. Create an NCBI account.



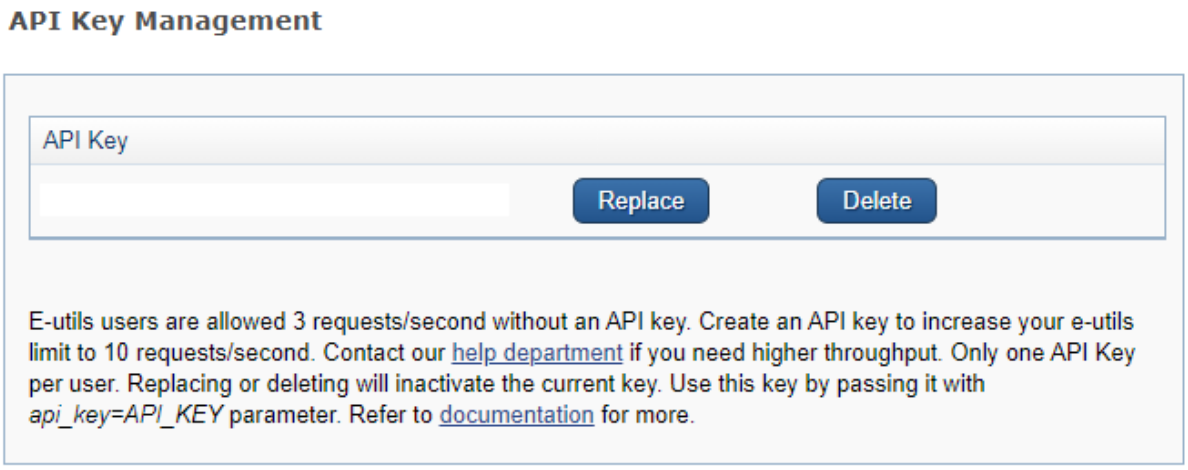
2. Click your account name to access settings



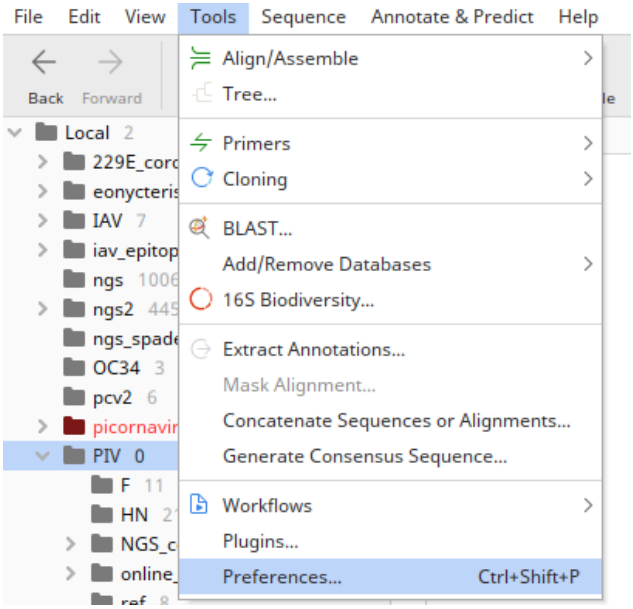
3. Create an API key



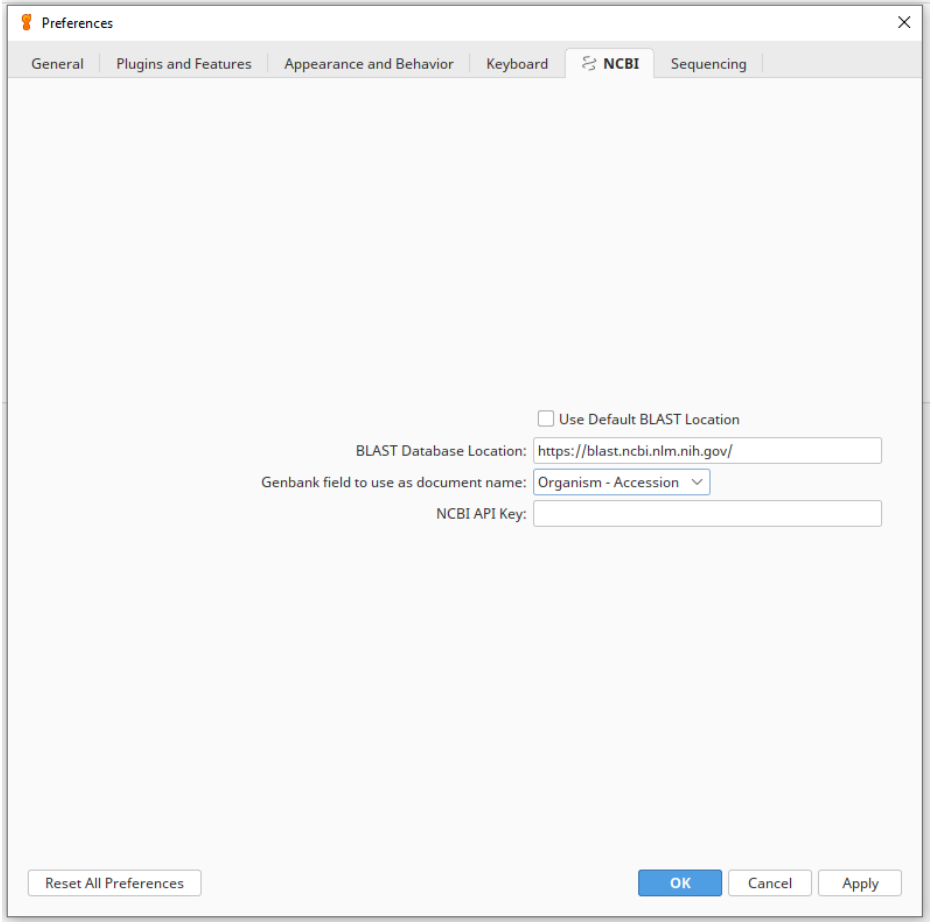
4. Create and copy the key



5. Open the preferences of Geneious



6. Paste it on the NCBI tab



Exercise 1 – Phylogenetic reconstruction of WGS Data using single genes

Goal – From assembled WGS data, find the phylogenetic relationship

Old school method – Initially one would pick a reference strain and mark where the gene of interest was. After running an alignment, the entire aligned block with the gene would be pulled out. If there were indels present, the extracted genes might need to be realigned or re-examined. This method can be quiet long, as alignment algorithms do not have linear time complexity.

Q1: Why do we not make trees from whole genome alignments typically

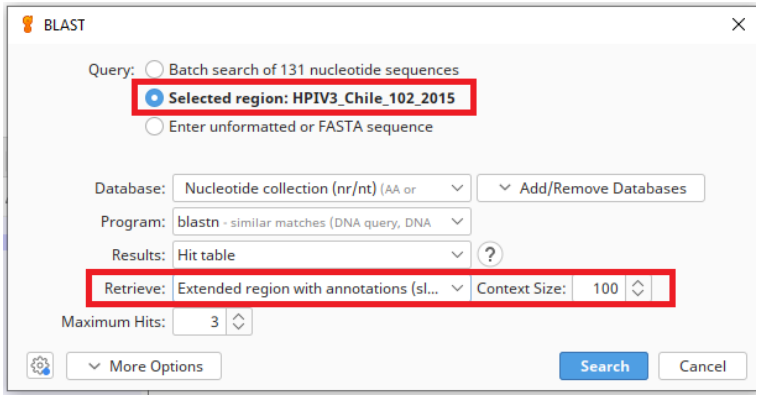
1. *Time complexity Trying to align and then create trees from a multitude of large sequences can be computationally demanding.*
2. *Recombination can disrupt signal and mess up topology*

Load in PIV3 NGS Data

Double click the attached Geneious file. If asked about where to import, create a specific folder for this project. Organization is important.

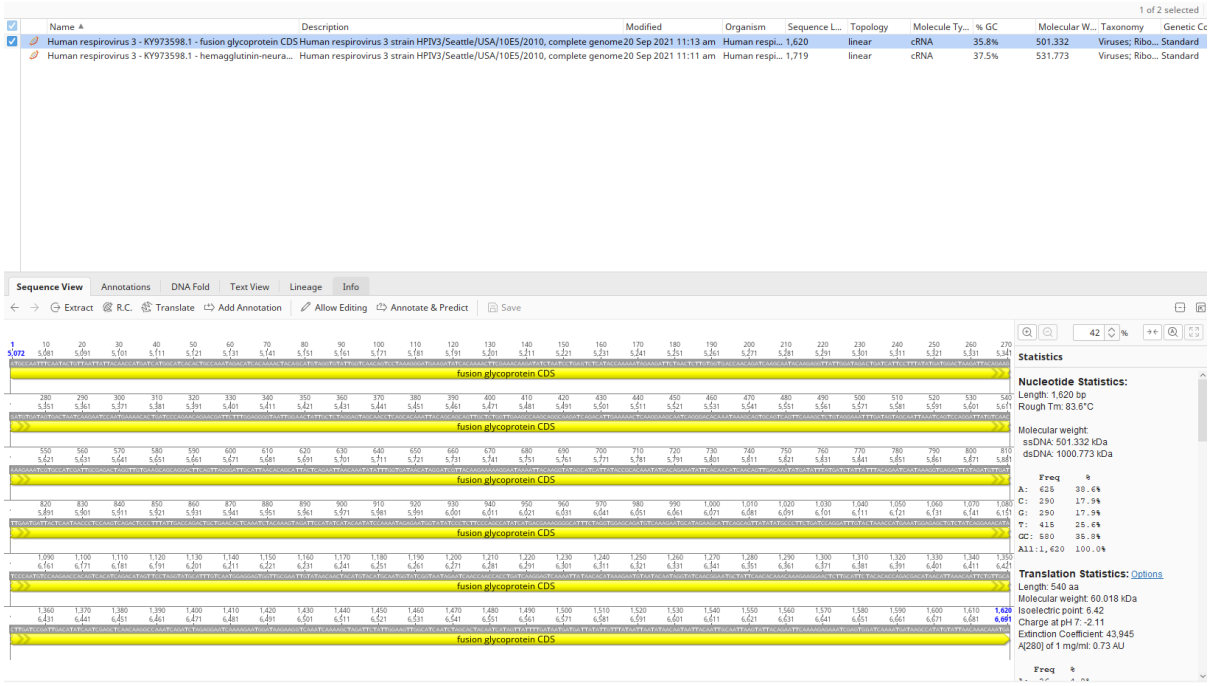
Find a reference sequence

Select the first sequence in the sequence list and choose BLAST. Make sure to minimize the hits to improve performance, WHGS with annotations is slow 1-3 (options). Because this is a previously well studied virus, there is a good chance we can find a well annotated sequence on GenBank.



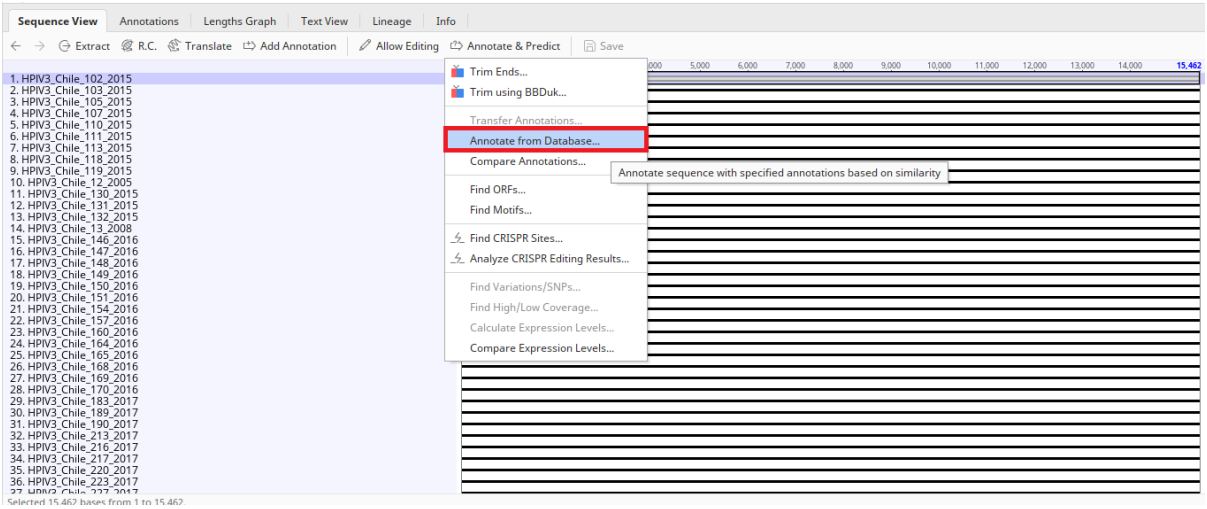
Picking annotations

1. Create a reference folder named “ref” inside your current working directory.
2. Click the HN CDS annotation in the first BLAST hit and copy it. Paste it into the ref directory
3. Delete the source and BLAST HIT annotations
4. Repeat 1-2 for the Fusion protein

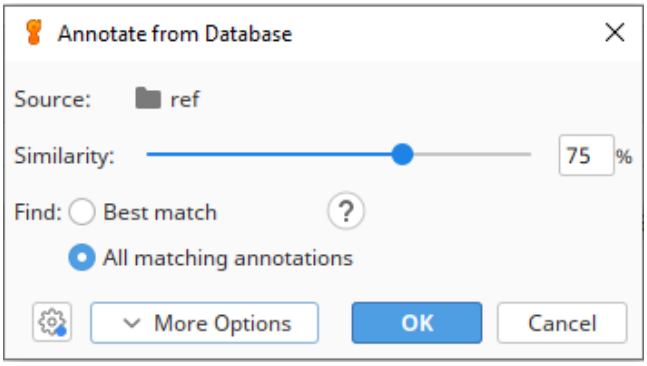


Predicting the annotations

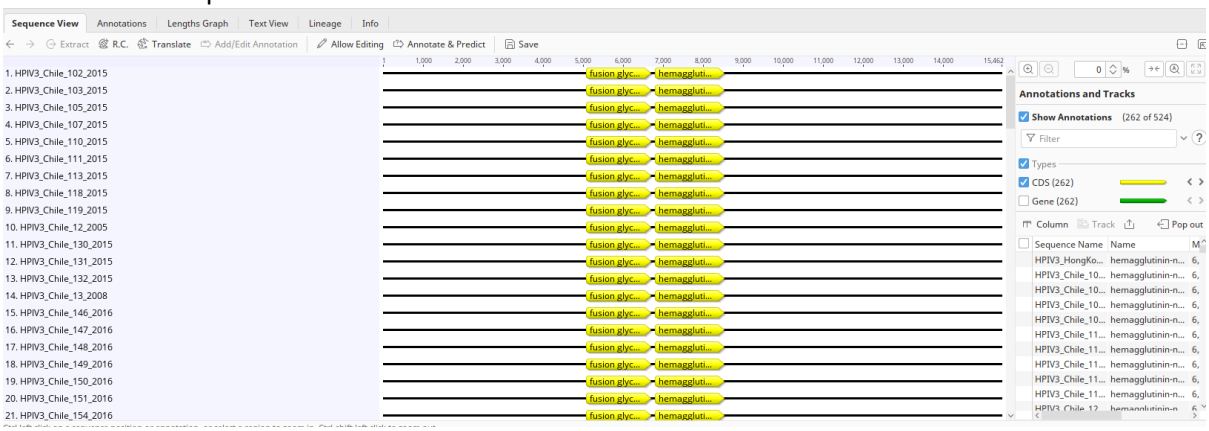
1. Click Annotate and predict - > Annotate from Database



2. Select your 'ref' folder as the source, and set similarity to a feasible percent (between 75% - 85%). In more complex applications, you may need to expand the "More Options" section.

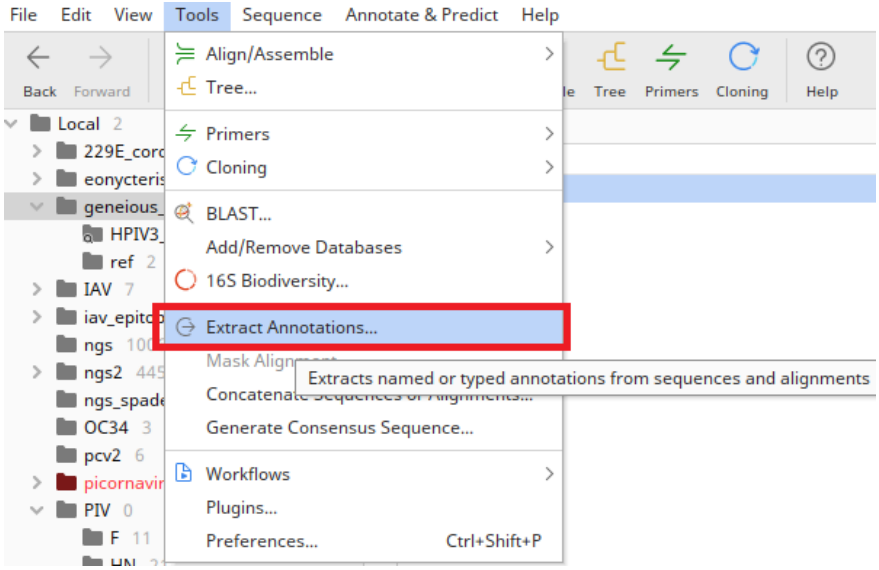


3. Check that all sequences are annotated.

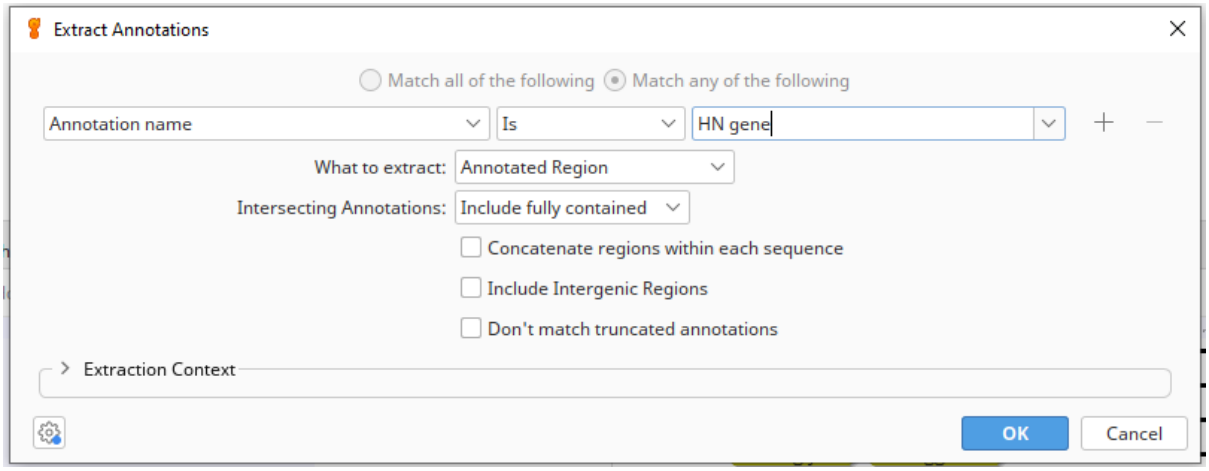


Extract Genes of interest

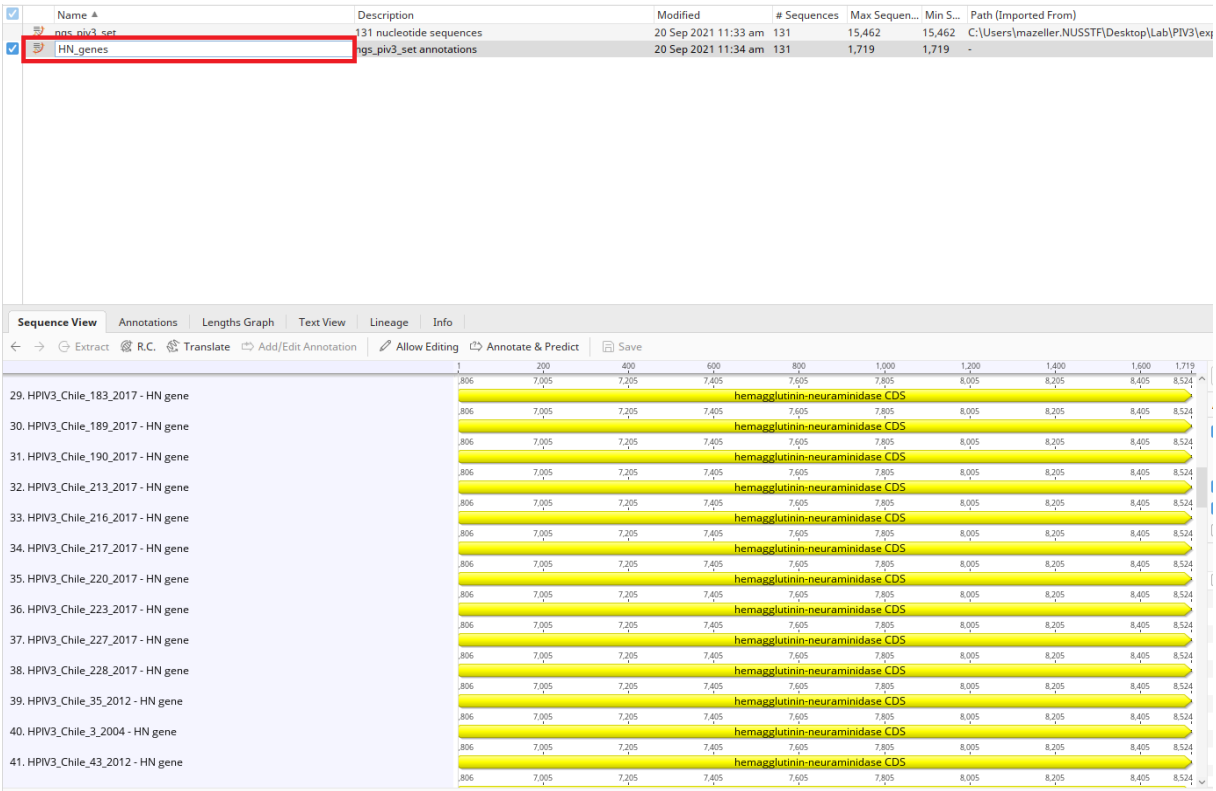
1. Select "Extract Annotations"



2. Complete selection logic for the HN gene



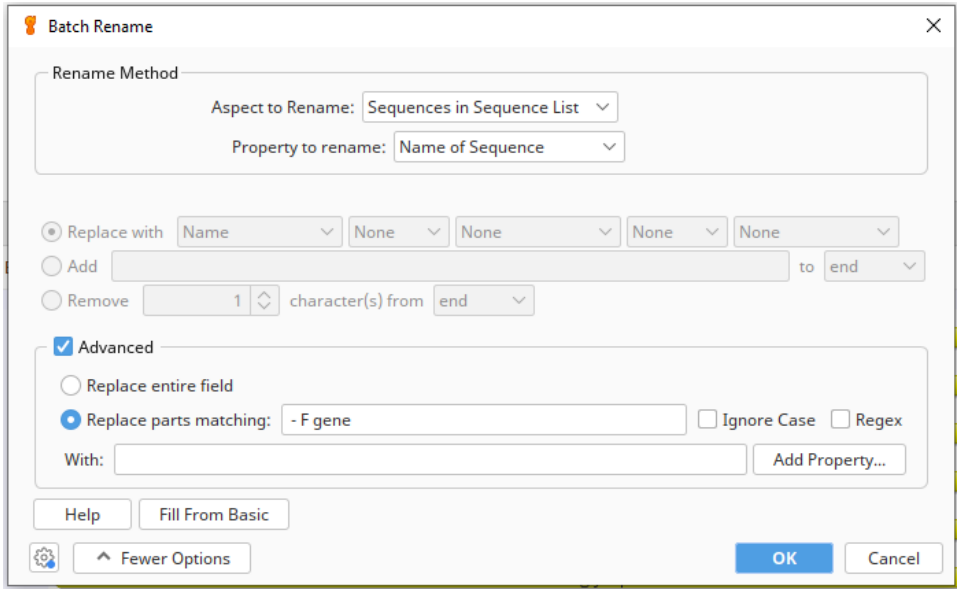
3. Rename the document to something sensible



4. Repeat for F gene

Batch renaming example

1. Select the F_genes document
2. Select Edit->Batch Rename
3. Find " - F gene" and replace with blank



4. More complex renaming requires either extra fields or advanced regex knowledge. It is typically easier to do with a bash script.

Constructing alignments and building trees

1. Align the F gene and the HN gene separately using either the Geneious, MUSCLE, or MAFFT alignment

2. Construct a neighbor joining tree using the Geneious Tree builder
   NJ trees as based solely on the metric of distance between sequences in the tree. This allows it to be much faster than maximum-likelihood trees and Bayesian trees. The consequence is that genetic information is lost There is also the property that if an exact solution exists, the NJ algorithm will find it. Issues are that often the exact solution does not exist, and this can sometimes result in trees with negative branch lengths. In terms of computational complexity, the NJ algorithm is not very efficient and runs at $O(n^3)$, or cubic runtime under the original formulation by Masotoshi Nei and Koichiro Tamura. Common heuristics implemented in programs runs at better time complexities.

Q2: Do the topologies match between the HN and F trees?

Q3: Does one gene appear to be more conserved than the other?

Q4: Are genes located towards the 5' or the 3' have a slower or faster mutation compared to each other?

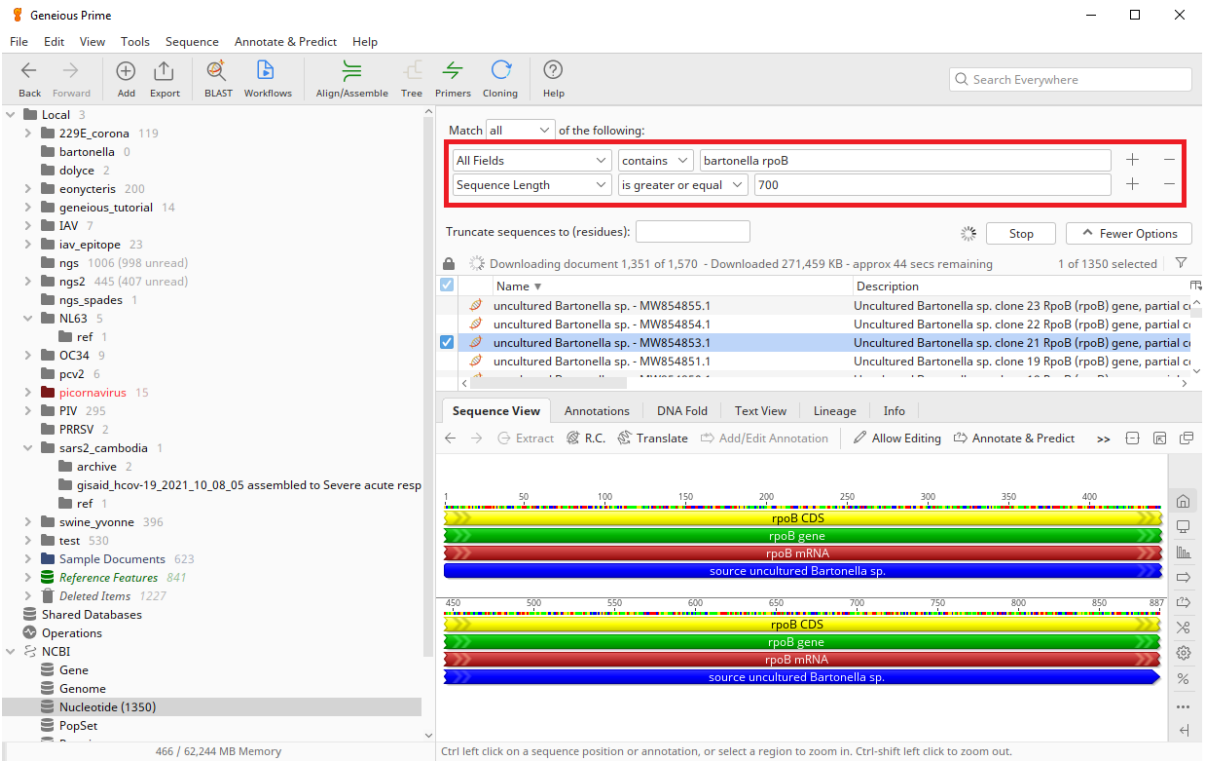
Q5: Are genes located closer to the 5' end replicated more or less than genes closer to the 3' end?

Exercise 2 Building a bartonella reference set

Problem: You have a pleathora of bartonella rpoB genes and you would like to capture the general diversity present within your sample set. To do this, an adequate dataset needs to be created to encapsulate the range of genetic diversity present in the population. The goal of this exercise is to use ad hoc measures to generate a reference set for bartonella based on the rpoB gene, while giving some consideration to the underlying genetics.

Downloading bartonella from NCBI

Based on prior knowledge, rpoB is approx. 4120nt in length. Many instances of the rpoB available in genbank are shorter than this length, coming in around 800nt approx. To get our initial datatset, we will selectively pull documents by length.

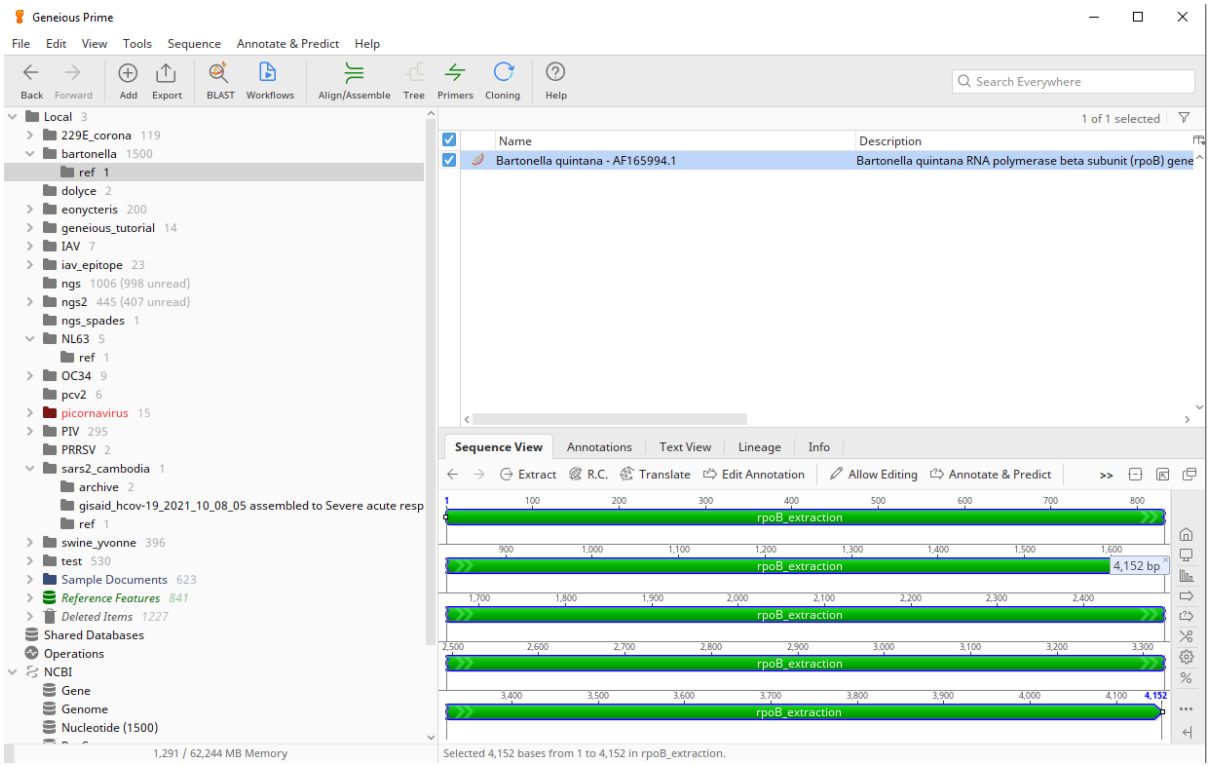


*Adding a less than 5000 is a good idea for this instance.

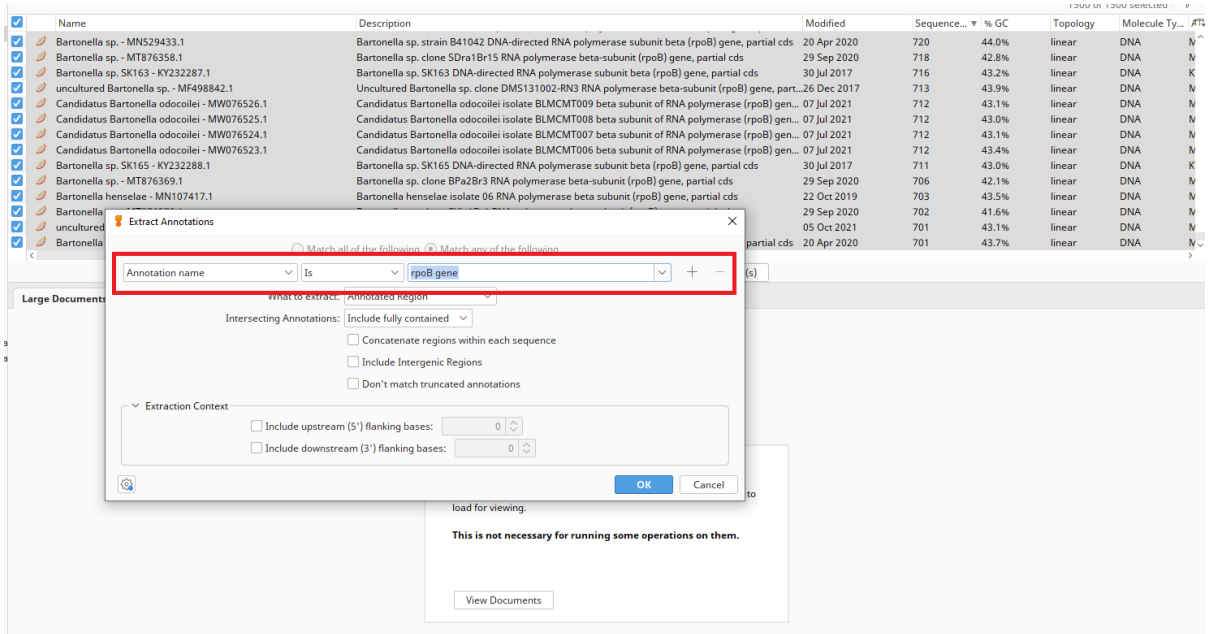
If the run is taking too long, you can stop it and maintain the current results. Move results to a new folder.

Annotating and extracting genes of interest

Create a reference 'ref' folder. Find a complete rpoB gene and copy the annotation over to the ref folder. Rename the annotation something unique for easy extraction, and remove superfluous annotations.

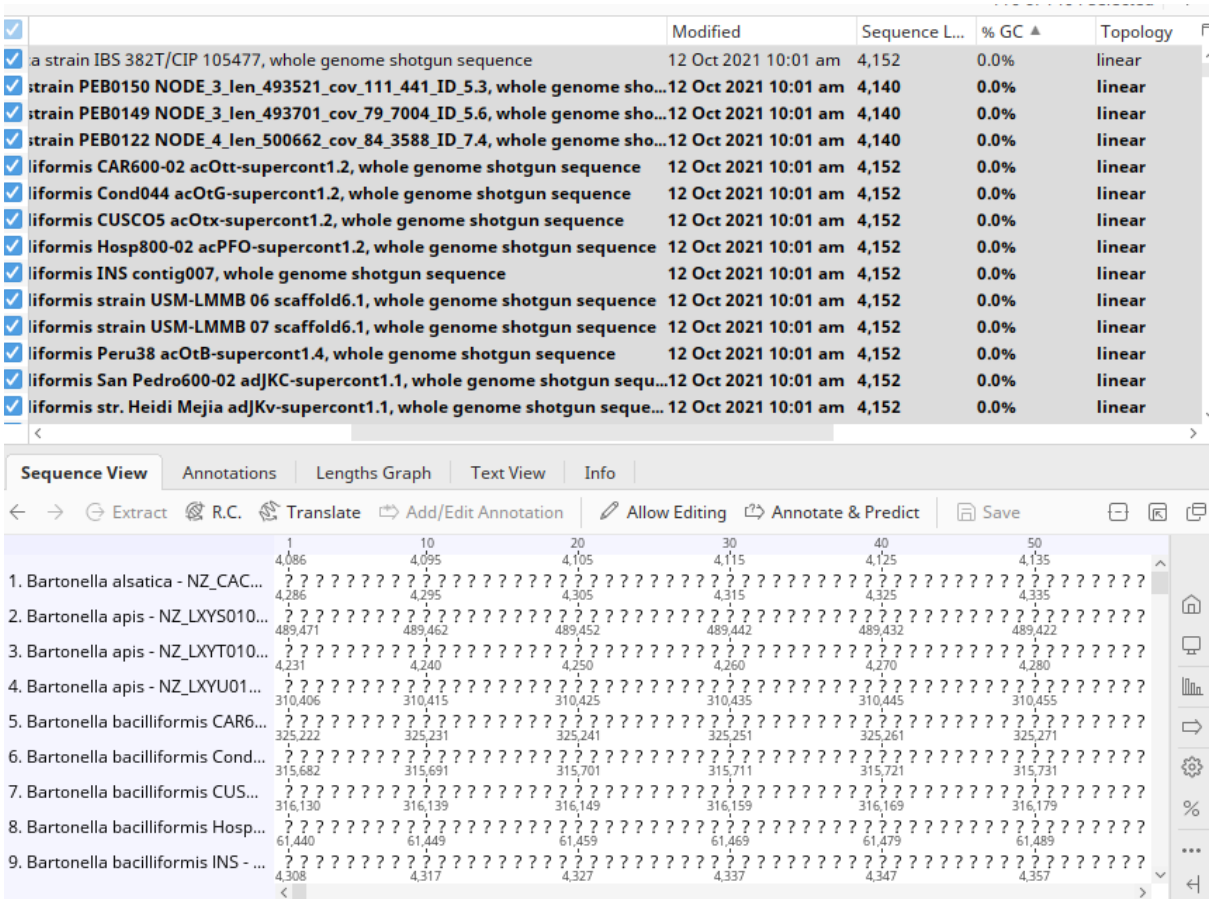


Select all Bartonella sequences of interest for processing. Using the Tools > Annotate Extractions tool, select the rpoB genes for extraction.

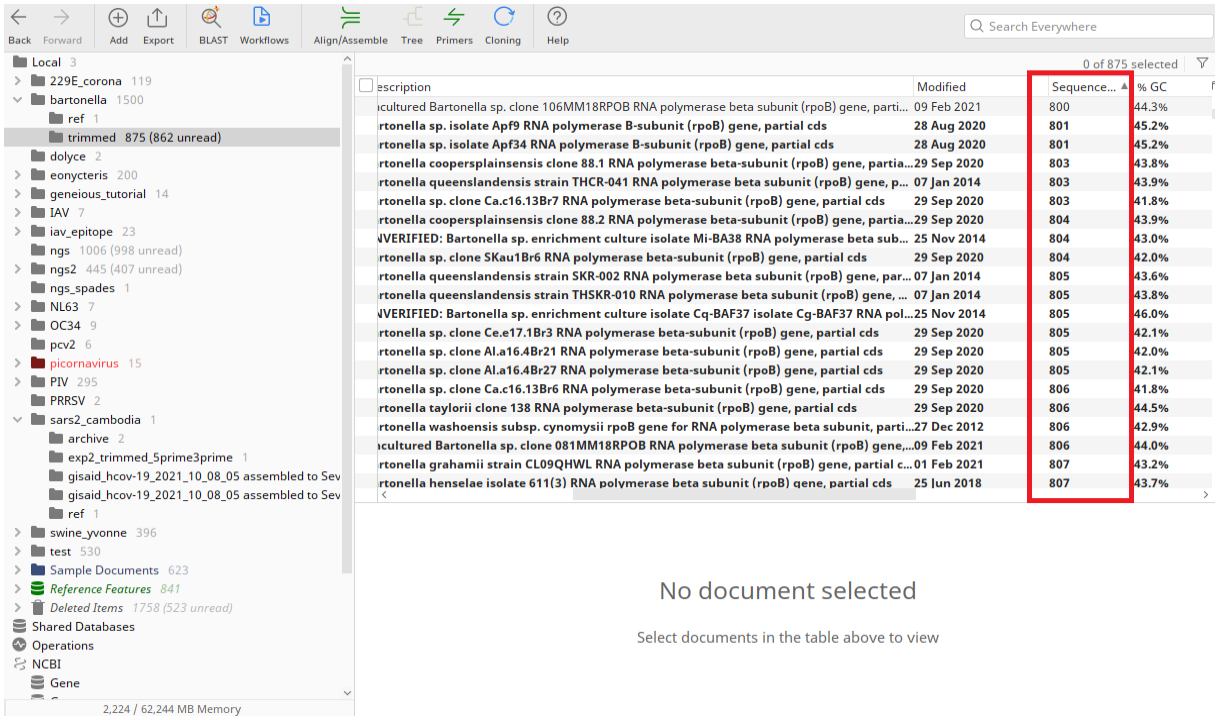


Select correct lengths and other cleaning measures

Not all sequences returned by NCBI will be useful. Some sequences may be too short, while other extractions may be unusable. First, useless sequences will be identified by having %GC content as 0.0% and removed. These sequences were introduced into NCBI from people submitting weak WGS.



Next, sort by sequence length and remove sequence under a threshold in length. If you only take sequences > 1k in length, you end up with 61 sequences. While this is a convenient size for a reference set, this may not capture the full diversity of Bartonella rpoB. For this reason, this example will only delete sequences with length less than 800nt.



At this point, approximately 800 sequences remain. To gain an idea of what information is available, the remaining sequences are aligned and pushed into a tree to check the diversity. The fastest way to do this would be to do a fast alignment paired with a neighbour-joining tree. Make sure to check the alignment for irregularities.

Subsampling larger datasets

Assuming you have a priori knowledge on the grouping and number of clades you want in your reference set, you can make an ad hoc assessment of the number of taxa that would be appropriate for the reference set. A general probability of selection from sampling without replacement can be made as a heuristic. Given that that there may be 15 distinct clades *equally* represented in a group of 800 sequences of bartonella, we can reasonably select 4 or 5 from each group and have a good probability of selection of at least 1 object from each clade. Please refer to prior smof/smot tutorial on how to perform random sampling, or sampling proportionally from the tree topology.