

Being a **BLASTED** Geneious

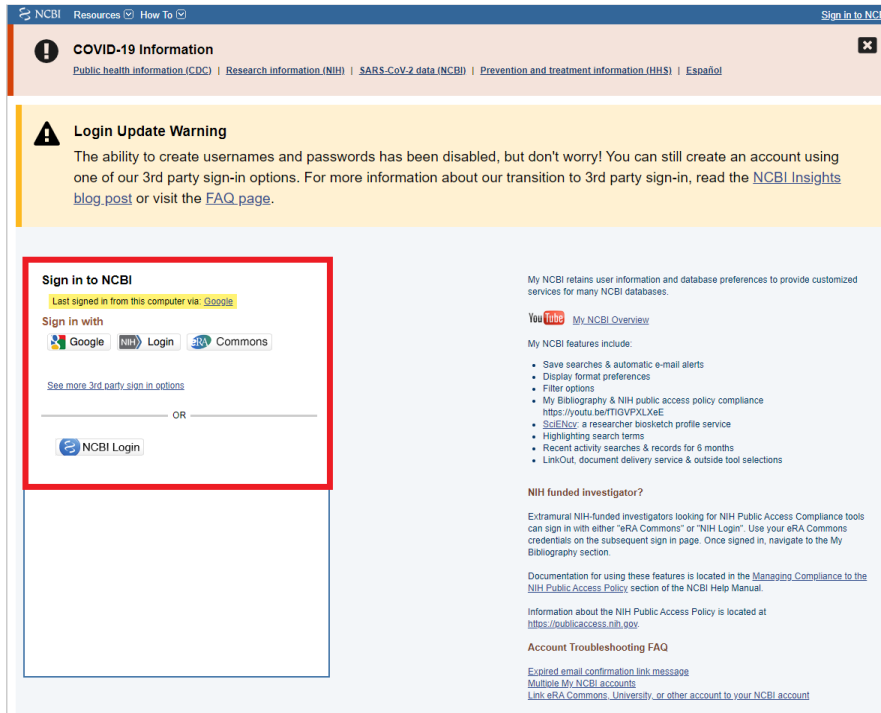
Table of Contents

| | |
|---|----|
| Setting up an NCBI account/API Key | 1 |
| Exercise 1 – Phylogenetic reconstruction of WGS Data using single genes | 4 |
| Load in PIV3 NGS Data | 4 |
| Find a reference sequence | 4 |
| Picking annotations | 5 |
| Predicting the annotations..... | 5 |
| Extract Genes of interest | 6 |
| Batch renaming example | 8 |
| Constructing alignments and building trees | 8 |
| Exercise 2 Building a bartonella reference set..... | 9 |
| Downloading bartonella from NCBI..... | 9 |
| Annotating and extracting genes of interest | 10 |
| Select correct lengths and other cleaning measures | 11 |
| Subsampling larger datasets | 13 |

Setting up an NCBI account/API Key

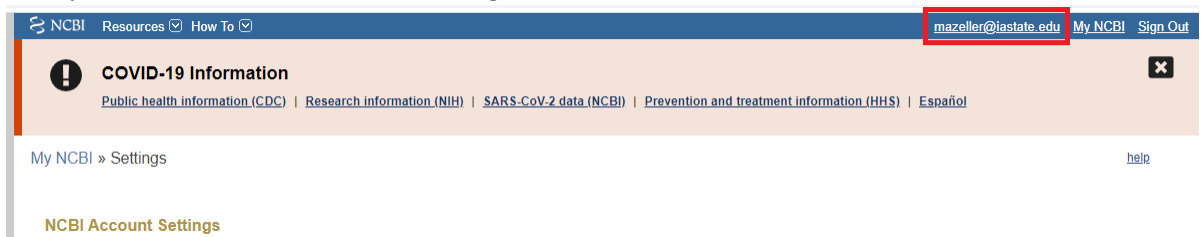
Keys only work for Geneious 11 and above. R7 and R10 users will not have access to this. Currently, API Key Holders can make 10 requests to NCBI per second. Non-API key holders can only make 3 requests per second.

1. Create an NCBI account.



The screenshot shows the NCBI homepage with a blue header bar containing 'NCBI', 'Resources', 'How To', and a 'Sign in to NCBI' link. A red box highlights the 'Sign in to NCBI' section on the left. This section includes a 'Last signed in from this computer via' dropdown set to 'Google', followed by 'Sign in with' buttons for Google, NIH, Login, and Commons. Below these is a link to 'See more 3rd party sign in options', an 'OR' separator, and an 'NCBI Login' button. To the right of the sign-in section, there is a 'My NCBI' section with a 'My NCBI Overview' link, a list of features (Save searches, Display format preferences, Filter options, My Bibliography, SciENcy, Highlighting search terms, Recent activity, LinkOut), and a 'NIH funded investigator?' section with links to 'Managing Compliance to the NIH Public Access Policy' and 'Account Troubleshooting FAQ'.

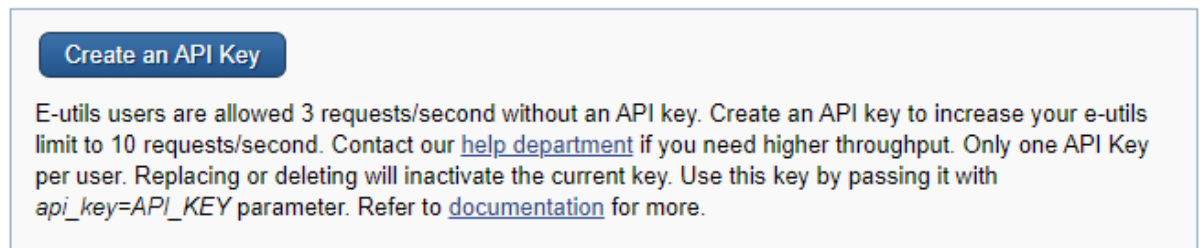
2. Click your account name to access settings



The screenshot shows the 'My NCBI' account settings page. The header bar now includes the email address 'mazeller@iastate.edu' next to the 'My NCBI' and 'Sign Out' links. A red box highlights the email address. Below the header, there is a 'COVID-19 Information' section with links to 'Public health information (CDC)', 'Research information (NIH)', 'SARS-CoV-2 data (NCBI)', 'Prevention and treatment information (HHS)', and 'Español'. The main content area shows 'My NCBI » Settings' and a 'help' link. Below this is a section titled 'NCBI Account Settings'.

3. Create an API key

API Key Management



The screenshot shows the 'API Key Management' page. It features a blue button labeled 'Create an API Key'. Below the button, there is a paragraph of text: 'E-utils users are allowed 3 requests/second without an API key. Create an API key to increase your e-utils limit to 10 requests/second. Contact our [help department](#) if you need higher throughput. Only one API Key per user. Replacing or deleting will inactivate the current key. Use this key by passing it with `api_key=API_KEY` parameter. Refer to [documentation](#) for more.'

4. Create and copy the key

API Key Management

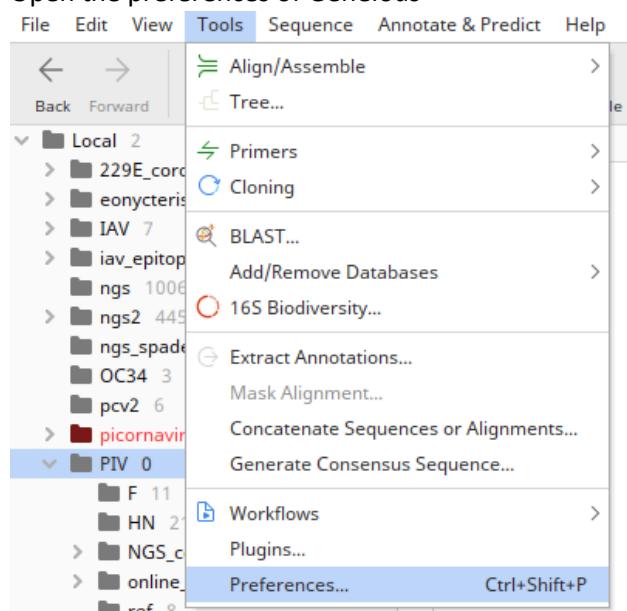
API Key

Replace

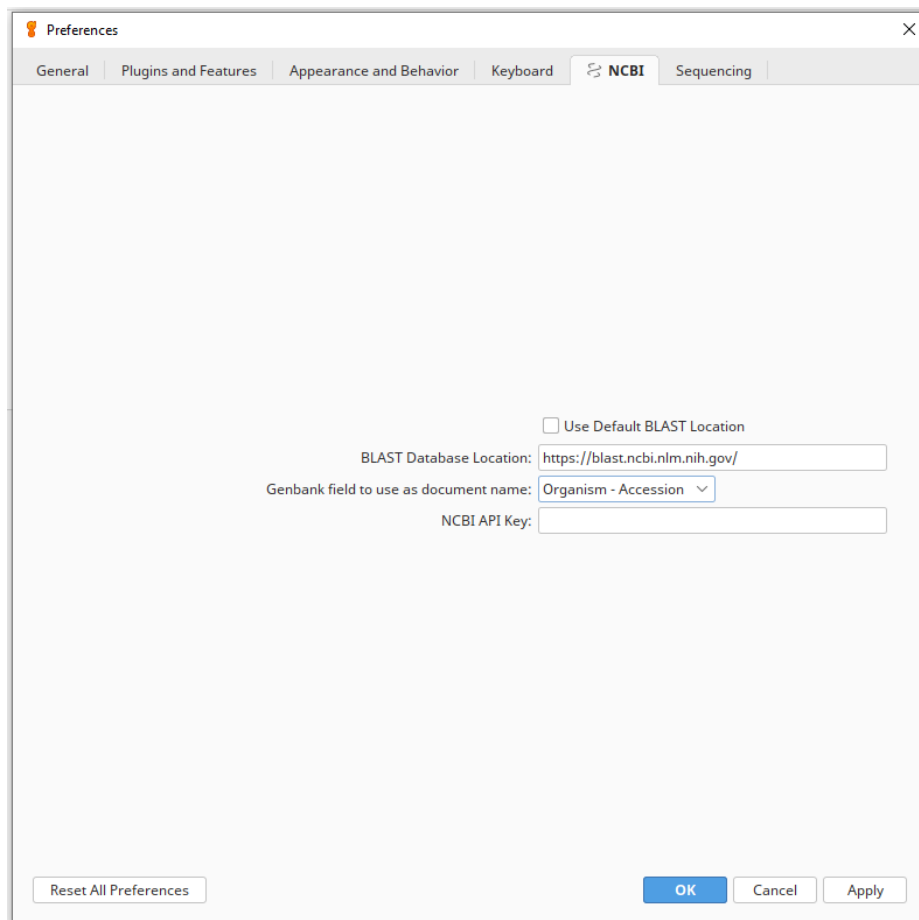
Delete

E-utils users are allowed 3 requests/second without an API key. Create an API key to increase your e-utils limit to 10 requests/second. Contact our [help department](#) if you need higher throughput. Only one API Key per user. Replacing or deleting will inactivate the current key. Use this key by passing it with `api_key=API_KEY` parameter. Refer to [documentation](#) for more.

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 quite 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.

BLAST

Query: ☐ Batch search of 131 nucleotide sequences
☒ **Selected region: HPIV3_Chile_102_2015**
☐ Enter unformatted or FASTA sequence

Database: Nucleotide collection (nr/nt) (AA or) Add/Remove Databases

Program: blastn - similar matches (DNA query, DNA)

Results: Hit table (?)

Retrieve: Extended region with annotations (sl...) Context Size: 100

Maximum Hits: 3

More Options Search Cancel

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

Human coronavirus HKU-1

| Name * | Description | Modified | Organism | Sequence L. | Topology | Molecule Ty... | % GC | Molecular W... | Taxonomy | Genetic Co |
|---|--|----------------|----------|-------------|----------|----------------|---------|------------------|----------|------------|
| Human coronavirus HKU-1 - fusion glycoprotein CDS | Human coronavirus HKU-1 strain HKU-1/USA/10E5/2010, complete genome 20 Sep 2011 11:13 am | Human respi... | 1,620 | linear | cRNA | 35.8% | 501.332 | Viruses; Ribo... | Standard | |
| Human coronavirus HKU-1 - hemagglutinin-neura... | Human coronavirus HKU-1 strain HKU-1/USA/10E5/2010, complete genome 20 Sep 2011 11:11 am | Human respi... | 1,719 | linear | cRNA | 37.5% | 531.773 | Viruses; Ribo... | Standard | |

Sequence View
Annotations
DNA Fold
Text View
Lineage
Info

Extract
R.C.
Translate
Add Annotation
Allow Editing
Annotate & Predict
Save

Statistics

Nucleotide Statistics:
Length: 1,620 bp
Rough Tm: 83.6°C

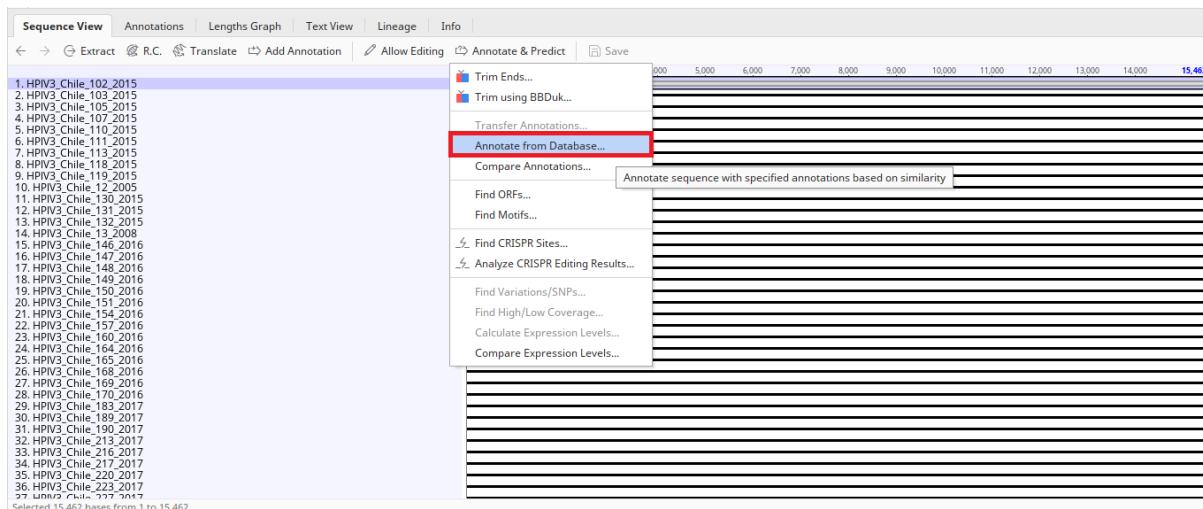
Molecular weight
ssDNA: 501.332 kDa
sdNA: 1000.773 kDa

Freq %
A: 29.6%
C: 17.9%
G: 25.6%
T: 26.9%

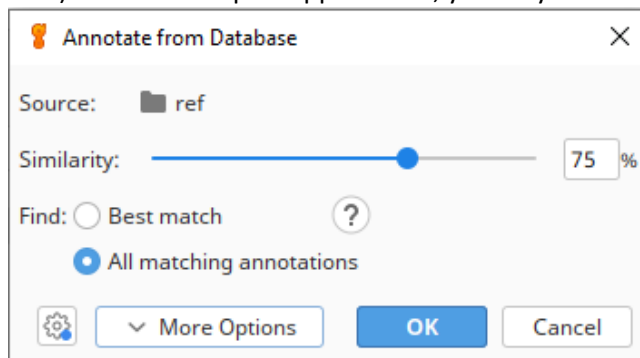
Translation Statistics:
Options
Length: 540 aa
Molecular weight: 60.018 kDa
Isoelectric point: 6.42
Charge at pH 7: -2.11
Extinction Coefficient: 43,945 AU(280) of 1 mg/ml: 0.73 AU

Predicting the annotations

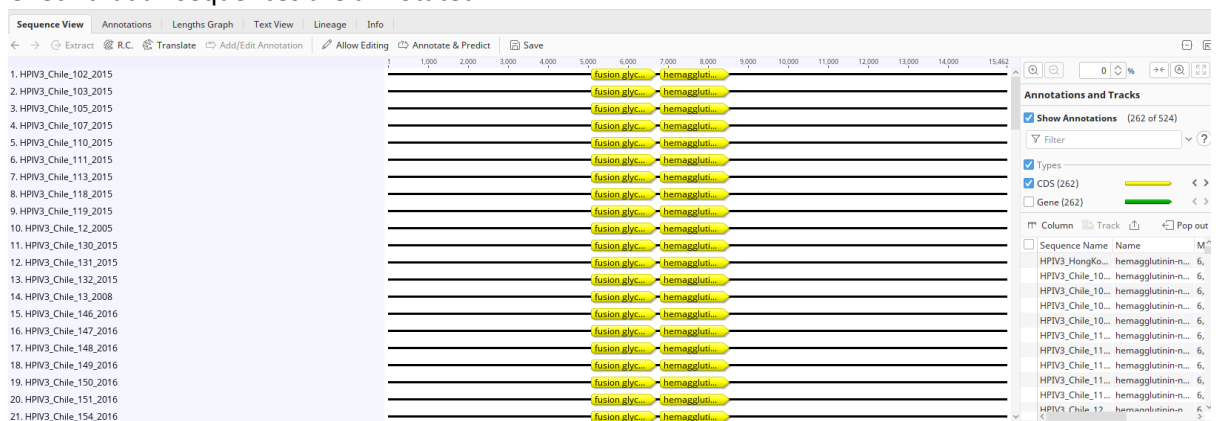
1. Click Annotate and predict -> Annotate from Database



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.

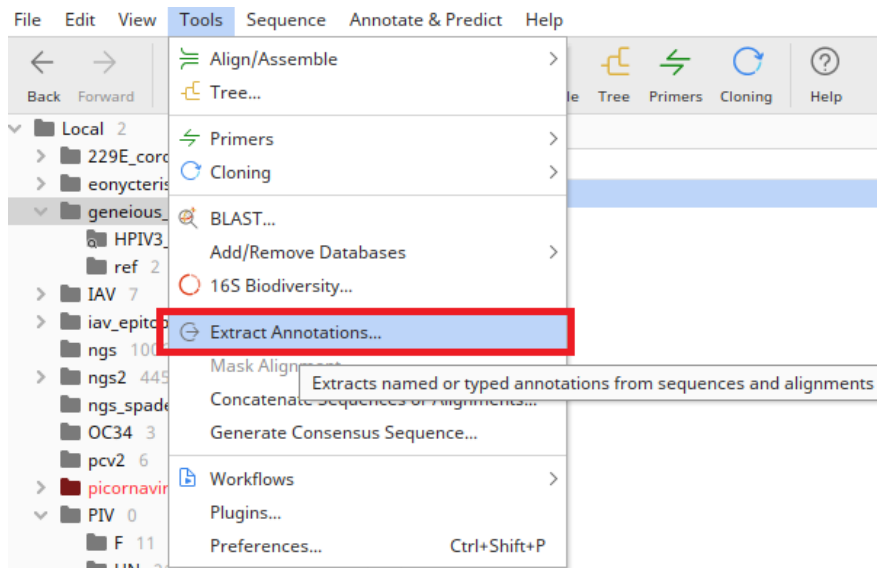


3. Check that all sequences are annotated.

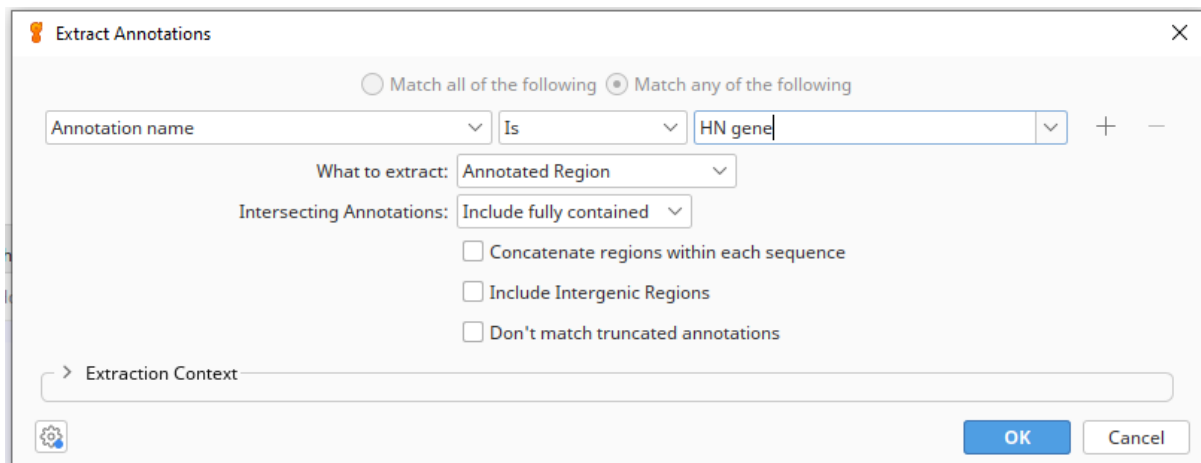


Extract Genes of interest

1. Select "Extract Annotations"



2. Complete selection logic for the HN gene



3. Rename the document to something sensible

| Name | Description | Modified | # Sequences | Max Sequen... | Min S... | Path (Imported From) |
|--------------|--------------------------|----------------------|-------------|---------------|----------|--|
| ngs_piv3_set | 131 nucleotide sequences | 20 Sep 2021 11:33 am | 131 | 15,462 | 15,462 | C:\Users\mazeller.NUSSTF\Desktop\Lab\PIV3\ex |
| HN_genes | ngs_piv3_set annotations | 20 Sep 2021 11:34 am | 131 | 1,719 | 1,719 | - |

| Sequence View | Annotations | Lengths Graph | Text View | Lineage | Info |
|---|---------------------------------|---------------|-----------|---------|------|
| <div> ← → ↺ Extract R.C. Translate Add/Edit Annotation Allow Editing Annotate & Predict Save </div> | | | | | |
| 29. HPIV3_Chile_183_2017 - HN gene | hemagglutinin-neuraminidase CDS | | | | |
| 30. HPIV3_Chile_189_2017 - HN gene | hemagglutinin-neuraminidase CDS | | | | |
| 31. HPIV3_Chile_190_2017 - HN gene | hemagglutinin-neuraminidase CDS | | | | |
| 32. HPIV3_Chile_213_2017 - HN gene | hemagglutinin-neuraminidase CDS | | | | |
| 33. HPIV3_Chile_216_2017 - HN gene | hemagglutinin-neuraminidase CDS | | | | |
| 34. HPIV3_Chile_217_2017 - HN gene | hemagglutinin-neuraminidase CDS | | | | |
| 35. HPIV3_Chile_220_2017 - HN gene | hemagglutinin-neuraminidase CDS | | | | |
| 36. HPIV3_Chile_223_2017 - HN gene | hemagglutinin-neuraminidase CDS | | | | |
| 37. HPIV3_Chile_227_2017 - HN gene | hemagglutinin-neuraminidase CDS | | | | |
| 38. HPIV3_Chile_228_2017 - HN gene | hemagglutinin-neuraminidase CDS | | | | |
| 39. HPIV3_Chile_35_2012 - HN gene | hemagglutinin-neuraminidase CDS | | | | |
| 40. HPIV3_Chile_3_2004 - HN gene | hemagglutinin-neuraminidase CDS | | | | |
| 41. HPIV3_Chile_43_2012 - HN gene | hemagglutinin-neuraminidase CDS | | | | |

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

Batch Rename

Rename Method

Aspect to Rename: Sequences in Sequence List
Property to rename: Name of Sequence

☒ Replace with
Name
None
None
None
None

☐ Add
to end

☐ Remove
1 character(s) from end

☒ Advanced

☐ Replace entire field
☒ Replace parts matching: - F gene
Ignore Case
Regex

With:
Add Property...

Help
Fill From Basic
Fewer Options
OK
Cancel

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 are 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 Masatoshi Nei and Koichiro Tamura. Common heuristics implemented in programs run 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 plethora 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 dataset, we will selectively pull documents by length.

The screenshot shows the Geneious Prime interface. On the left is a file explorer with a tree view showing various folders like 'Local', '229E_corona', 'bartonella', etc. The main area displays a list of sequences. A search filter box at the top right is highlighted with a red rectangle, showing 'Match all of the following:' with two criteria: 'All Fields contains bartonella rpoB' and 'Sequence Length is greater or equal 700'. Below this, a table lists sequences with columns 'Name' and 'Description'. The 'Sequence View' tab is active, showing a detailed view of a sequence with annotations for 'rpoB CDS', 'rpoB gene', and 'rpoB mRNA'. The sequence is visualized with a color-coded bar at the top and a detailed view of the sequence below.

*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.

The screenshot shows the Geneious Prime software interface. On the left, a project tree under 'Local' shows a folder named 'bartonella' containing a 'ref' subfolder. The main window displays a 'Sequence View' of a selected sequence, 'Bartonella quintana - AF165994.1'. The sequence is shown as a green bar with a scale from 1 to 4,152 bp. The 'Annotations' tab is active, showing a single annotation for 'rpoB_extraction' spanning the entire sequence. The 'Tools' menu is open, and 'Annotate > Extract' is selected.

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

The screenshot shows the 'Extract Annotations' dialog box in Geneious Prime. The dialog is open over a table of sequences. The 'Annotation name' field is set to 'rpoB gene'. The 'What to extract' dropdown is set to 'Annotated region'. The 'Intersecting Annotations' dropdown is set to 'Include fully contained'. The 'Extraction Context' section has 'Include upstream (5') flanking bases' and 'Include downstream (3') flanking bases' both set to 0. The 'OK' button is highlighted.

| Name | Description | Modified | Sequence... | % GC | Topology | Molecule Ty... | AT... |
|---|--|-------------|-------------|-------|----------|----------------|-------|
| Bartonella sp. - MN529433.1 | Bartonella sp. strain B41042 DNA-directed RNA polymerase subunit beta (rpoB) gene, partial cds | 20 Apr 2020 | 720 | 44.0% | linear | DNA | N |
| Bartonella sp. - MT876358.1 | Bartonella sp. clone Sdra1Br15 RNA polymerase beta-subunit (rpoB) gene, partial cds | 29 Sep 2020 | 718 | 42.8% | linear | DNA | N |
| Bartonella sp. SK163 - KY232287.1 | Bartonella sp. SK163 DNA-directed RNA polymerase subunit beta (rpoB) gene, partial cds | 30 Jul 2017 | 716 | 43.2% | linear | DNA | K |
| uncultured Bartonella sp. - MF498842.1 | Uncultured Bartonella sp. clone DMS131002-RN3 RNA polymerase beta-subunit (rpoB) gene, part... | 26 Dec 2017 | 713 | 43.9% | linear | DNA | N |
| Candidatus Bartonella odocoilei - MW076526.1 | Candidatus Bartonella odocoilei isolate BLMCMT009 beta subunit of RNA polymerase (rpoB) gen... | 07 Jul 2021 | 712 | 43.1% | linear | DNA | N |
| Candidatus Bartonella odocoilei - MW076525.1 | Candidatus Bartonella odocoilei isolate BLMCMT008 beta subunit of RNA polymerase (rpoB) gen... | 07 Jul 2021 | 712 | 43.0% | linear | DNA | N |
| Candidatus Bartonella odocoilei - MW076524.1 | Candidatus Bartonella odocoilei isolate BLMCMT007 beta subunit of RNA polymerase (rpoB) gen... | 07 Jul 2021 | 712 | 43.1% | linear | DNA | N |
| Candidatus Bartonella odocoilei - MW076523.1 | Candidatus Bartonella odocoilei isolate BLMCMT006 beta subunit of RNA polymerase (rpoB) gen... | 07 Jul 2021 | 712 | 43.4% | linear | DNA | N |
| Bartonella sp. SK165 - KY232288.1 | Bartonella sp. SK165 DNA-directed RNA polymerase subunit beta (rpoB) gene, partial cds | 30 Jul 2017 | 711 | 43.0% | linear | DNA | K |
| Bartonella sp. clone BP42By3 RNA polymerase beta-subunit (rpoB) gene, partial cds | | 29 Sep 2020 | 706 | 42.1% | linear | DNA | N |
| Bartonella henselae - MN107417.1 | Bartonella henselae isolate 06 RNA polymerase beta subunit (rpoB) gene, partial cds | 22 Oct 2019 | 703 | 43.5% | linear | DNA | N |
| Bartonella sp. - MT876369.1 | | 29 Sep 2020 | 702 | 41.6% | linear | DNA | N |
| uncultured Bartonella sp. - MT876369.1 | | 05 Oct 2021 | 701 | 43.1% | linear | DNA | N |
| Bartonella sp. - MT876369.1 | | 20 Apr 2020 | 701 | 43.7% | linear | DNA | N |

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.

| | Modified | Sequence L... | % GC ▲ | Topology |
|--|----------------------|---------------|--------|----------|
| ✓ a strain IBS 382T/CIP 105477, whole genome shotgun sequence | 12 Oct 2021 10:01 am | 4,152 | 0.0% | linear |
| ✓ strain PEB0150 NODE_3_len_493521_cov_111_441_ID_5.3, whole genome sho... | 12 Oct 2021 10:01 am | 4,140 | 0.0% | linear |
| ✓ strain PEB0149 NODE_3_len_493701_cov_79_7004_ID_5.6, whole genome sho... | 12 Oct 2021 10:01 am | 4,140 | 0.0% | linear |
| ✓ strain PEB0122 NODE_4_len_500662_cov_84_3588_ID_7.4, whole genome sho... | 12 Oct 2021 10:01 am | 4,140 | 0.0% | linear |
| ✓ iformis CAR600-02 acOtt-supercont1.2, whole genome shotgun sequence | 12 Oct 2021 10:01 am | 4,152 | 0.0% | linear |
| ✓ iformis Cond044 acOtG-supercont1.2, whole genome shotgun sequence | 12 Oct 2021 10:01 am | 4,152 | 0.0% | linear |
| ✓ iformis CUSCO5 acOtx-supercont1.2, whole genome shotgun sequence | 12 Oct 2021 10:01 am | 4,152 | 0.0% | linear |
| ✓ iformis Hosp800-02 acPFO-supercont1.2, whole genome shotgun sequence | 12 Oct 2021 10:01 am | 4,152 | 0.0% | linear |
| ✓ iformis INS contig007, whole genome shotgun sequence | 12 Oct 2021 10:01 am | 4,152 | 0.0% | linear |
| ✓ iformis strain USM-LMMB 06 scaffold6.1, whole genome shotgun sequence | 12 Oct 2021 10:01 am | 4,152 | 0.0% | linear |
| ✓ iformis strain USM-LMMB 07 scaffold6.1, whole genome shotgun sequence | 12 Oct 2021 10:01 am | 4,152 | 0.0% | linear |
| ✓ iformis Peru38 acOtB-supercont1.4, whole genome shotgun sequence | 12 Oct 2021 10:01 am | 4,152 | 0.0% | linear |
| ✓ iformis San Pedro600-02 adjKC-supercont1.1, whole genome shotgun sequ... | 12 Oct 2021 10:01 am | 4,152 | 0.0% | linear |
| ✓ iformis str. Heidi Mejia adjKv-supercont1.1, whole genome shotgun seque... | 12 Oct 2021 10:01 am | 4,152 | 0.0% | linear |

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.

| Local 3 | description | Modified | Sequence... | % GC |
|---|---|-------------|-------------|-------|
| 229E_corona 119 | icutured Bartonella sp. clone 106MM18RPOB RNA polymerase beta subunit (rpoB) gene, parti... | 09 Feb 2021 | 800 | 44.3% |
| bartonella 1500 | rtionella sp. isolate Apf9 RNA polymerase B-subunit (rpoB) gene, partial cds | 28 Aug 2020 | 801 | 45.2% |
| ref 1 | rtionella sp. isolate Apf34 RNA polymerase B-subunit (rpoB) gene, partial cds | 28 Aug 2020 | 801 | 45.2% |
| trimmed 875 (862 unread) | rtionella cooperplainsensis clone 88.1 RNA polymerase beta-subunit (rpoB) gene, partial cds | 29 Sep 2020 | 803 | 43.8% |
| dolyce 2 | rtionella queenslandensis strain THCR-041 RNA polymerase beta subunit (rpoB) gene, p... | 07 Jan 2014 | 803 | 43.9% |
| eonycteris 200 | rtionella sp. clone Ca.c16.13Br7 RNA polymerase beta-subunit (rpoB) gene, partial cds | 29 Sep 2020 | 803 | 41.8% |
| geneious_tutorial 14 | rtionella cooperplainsensis clone 88.2 RNA polymerase beta-subunit (rpoB) gene, partial cds | 29 Sep 2020 | 804 | 43.9% |
| IAV 7 | VERIFIED: Bartonella sp. enrichment culture isolate Mi-BA38 RNA polymerase beta sub... | 25 Nov 2014 | 804 | 43.0% |
| iav_epitope 23 | rtionella sp. clone SKaur1Br6 RNA polymerase beta-subunit (rpoB) gene, partial cds | 29 Sep 2020 | 804 | 42.0% |
| ngs 1006 (998 unread) | rtionella queenslandensis strain SKR-002 RNA polymerase beta subunit (rpoB) gene, par... | 07 Jan 2014 | 805 | 43.6% |
| ngs2 445 (407 unread) | rtionella queenslandensis strain THSKR-010 RNA polymerase beta subunit (rpoB) gene, ... | 07 Jan 2014 | 805 | 43.8% |
| ngs_spades 1 | VERIFIED: Bartonella sp. enrichment culture isolate Cq-BAF37 RNA pol... | 25 Nov 2014 | 805 | 46.0% |
| NL63 7 | rtionella sp. clone Ce.e17.1Br3 RNA polymerase beta-subunit (rpoB) gene, partial cds | 29 Sep 2020 | 805 | 42.1% |
| OC34 9 | rtionella sp. clone Al.a16.4Br21 RNA polymerase beta-subunit (rpoB) gene, partial cds | 29 Sep 2020 | 805 | 42.0% |
| pcv2 6 | rtionella sp. clone Al.a16.4Br27 RNA polymerase beta-subunit (rpoB) gene, partial cds | 29 Sep 2020 | 805 | 42.1% |
| picornavirus 15 | rtionella sp. clone Ca.c16.13Br6 RNA polymerase beta-subunit (rpoB) gene, partial cds | 29 Sep 2020 | 806 | 41.8% |
| PIV 295 | rtionella taylorii clone 138 RNA polymerase beta-subunit (rpoB) gene, partial cds | 29 Sep 2020 | 806 | 44.5% |
| PRRSV 2 | rtionella washoensis subsp. cynomysii rpoB gene for RNA polymerase beta subunit, parti... | 27 Dec 2012 | 806 | 42.9% |
| sars2_cambodia 1 | icutured Bartonella sp. clone 081MM18RPOB RNA polymerase beta subunit (rpoB) gene... | 09 Feb 2021 | 806 | 44.0% |
| archive 2 | rtionella grahamii strain CL09QHWL RNA polymerase beta subunit (rpoB) gene, partial c... | 01 Feb 2021 | 807 | 43.2% |
| exp2_trimmed_5prime3prime 1 | rtionella henselae isolate 611(3) RNA polymerase beta subunit (rpoB) gene, partial cds | 25 Jun 2018 | 807 | 43.7% |
| gisaid_hcov-19_2021_10_08_05 assembled to Sev | | | | |
| gisaid_hcov-19_2021_10_08_05 assembled to Sev | | | | |
| ref 1 | | | | |
| swine_yvonne 396 | | | | |
| test 530 | | | | |
| Sample Documents 623 | | | | |
| Reference Features 841 | | | | |
| Deleted Items 1758 (523 unread) | | | | |
| Shared Databases | | | | |
| Operations | | | | |
| NCBI | | | | |
| Gene | | | | |

No document selected

Select documents in the table above to view

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 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.