MacVector 17.5

for Mac OS X

Technical Note

Data mining with MacVector confirms pangolins as potential intermediary hosts during the evolution of SARS-CoV-2, the coronavirus infectious agent of COVID-19

Abstract

We used MacVector to scan NGS datasets of pangolin lung RNA deposited in the NCBI SRA for similarity to the recently described human SARS-CoV-2 coronavirus, the causative agent of the worldwide COVID-19 pandemic. Many of the samples (6/20) contained at least some reads with significant similarity to the human virus. Earlier published reports suggested that the Pangolin CoV-2 virus may have been involved in a recombination event during the transmission between bats and humans. Our results do support this hypothesis, but clearly show that the bat CoV-RaTG13 genome is overall much more closely related to human SARS CoV-2 than the genome of CoV-2 from pangolins.

Introduction

The current COVID-19 coronavirus outbreak has caused considerable disruption to world financial markets and travelling due to quarantine requirements. While the primary medical focus has to be on responding to the current outbreak, it is crucial to understand how the outbreak evolved so that we can take steps to reduce future incidences.

An early report on the analysis of the first published human SARS-CoV-2 sequences demonstrated that they had significant sequence similarity to a known bat coronavirus, suggesting that bats were the potential source of the disease. However, a more recent publication has indicated that a coronavirus found in pangolins may have been involved in the transmission from bat to humans. Here we confirm some of those observations and show how MacVector can be used as a datamining tool to retrieve and analyze matching low abundance reads from large input datasets.

Materials and Methods

All analyses were performed using MacVector 17.5.3 on either a Late 2014 27” iMac with a 4 GHz Intel i7 processor and 32 GB RAM running macOS Mojave (10.14) or on a 2018 15” MacBook Pro with a 2.9 GHz Intel i9 processor and 32 GB RAM running macOS Catalina (10.15). These are both modest machines, illustrating that you can perform quite sophisticated data mining experiments on cheap hardware.

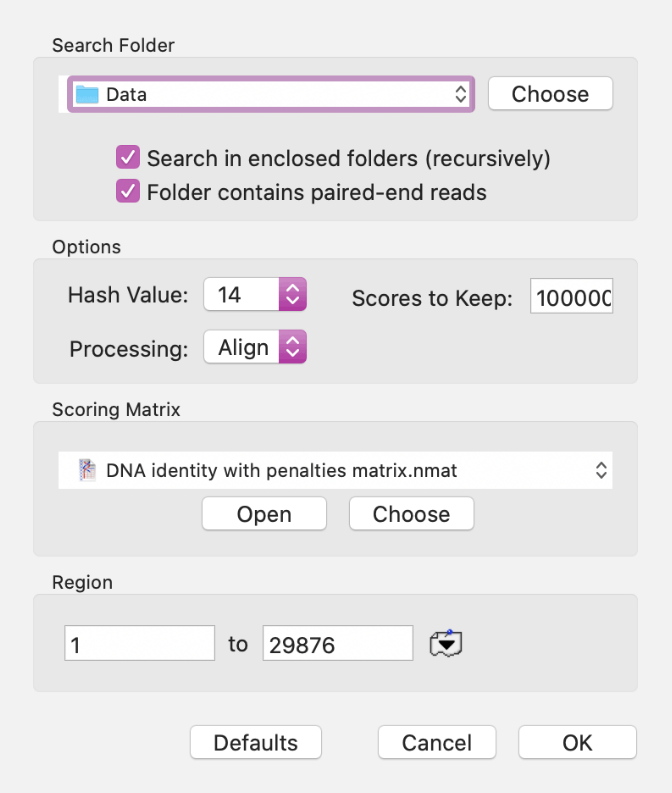
Results

Identifying Pangolin CoV-2 Analogue RNASeq Reads Using MacVector

The NCBI SRA "Sequence Read Archive" is an excellent source of Next Generation Sequencing (NGS) data for hundreds of thousands of sequencing projects. For this analysis we focused on a project where RNA was recovered and sequenced from the lungs of 20 pangolins from Asia. These were initially downloaded in paired interleaved fasta format and each renamed appropriately and placed in a single folder on a Mac computer.

We separately searched for and downloaded a few SARS-CoV-2 sequences using the **Database | Online Keyword Search...** function in MacVector using "SARS-CoV-2" as the search term. We focused on a few isolates from the USA, in particular `2019-nCov/USA-AZ1`, along with one of the original isolates from Wuhan China (`BetaCoV/Wuhan/IPBCAMS-WH-04/2019`) though other isolates were considered too, with similar results.

`2019-nCov/USA-AZ1` was used to run a **Database | Align to Folder** search against a folder containing the 20 fasta-formatted interleaved Pangolin NGS RNA-Seq data files. To speed up the search and ensure all of the hits were retained, the following settings were used;



Search Folder

Data

☑ Search in enclosed folders (recursively)

☑ Folder contains paired-end reads

Options

Hash Value: 14

Scores to Keep: 10000C

Processing: Align

Scoring Matrix

DNA identity with penalties matrix.nmat

Region

1 to 29876

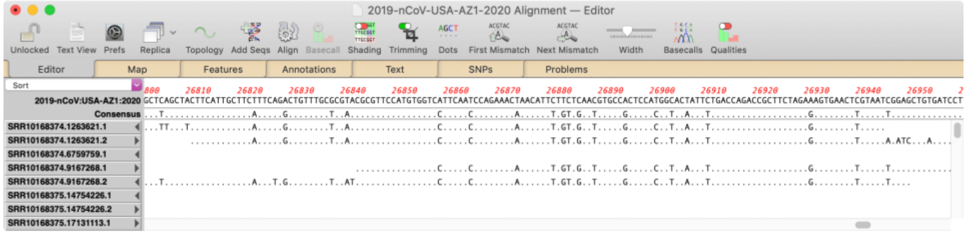
Note the **Hash Value** of 14 to speed up the search and the increase of **Scores to Keep** to 100,000 to ensure we retain all relevant hits.

After the search completed, (two days, but this can be speeded up by duplicating the search sequence and setting multiple searches going at one time to subsets of the fasta data) a total of 1,583 hits were found. These were extracted into a pair of fasta-format files by selecting rows in the **Folder Description List** tab and choosing **Database | Retrieve to File**.

The extracted hits were then aligned to `2019-nCov/USA-AZ1` using **Analyze | Align to Reference**. 125 of the hits failed to align to the sequence using the default settings. Additional analysis of a random sample of these indicated that they had extremely limited sequence similarity with

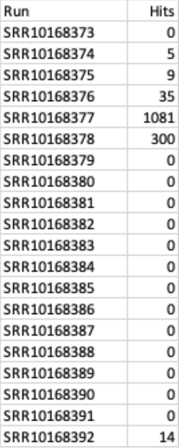
`2019-nCov/USA-AZ1` and BLAST searches demonstrated that they were not related to coronavirus sequences.

The remaining RNA-Seq reads aligned well with `2019-nCov/USA-AZ1`, such as this selection of reads from the SRR10168374 data set (shown with the **Dots** toolbar button turned on to help accentuate differences between the reads and the reference);



```
2019-nCoV:USA-AZ1:2020 GCTCAGCTACTTCATTGCTTCTTTCAGACTGTTTGCGCGTACGCGTTCCATGTGGTCATTCAATCCAGAAACTAACATTCTTCTCAACGTGCCACTCCATGGCACTATTCTGACCAGACCGCTTCTAGAAAGTGAACTCGTAATCGGAGCTGTGATCCT
Consensus
SRR10168374.1263621.1
SRR10168374.1263621.2
SRR10168374.6759759.1
SRR10168374.9167268.1
SRR10168374.9167268.2
SRR10168375.14754226.1
SRR10168375.14754226.2
SRR10168375.17131113.1
```

The number of hits found in each dataset is presented below;



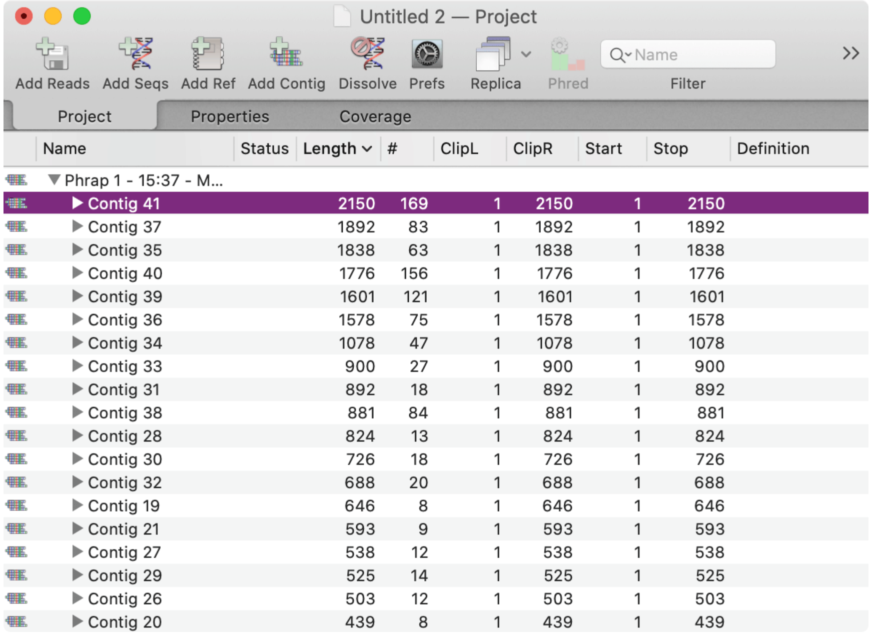
Run	Hits
SRR10168373	0
SRR10168374	5
SRR10168375	9
SRR10168376	35
SRR10168377	1081
SRR10168378	300
SRR10168379	0
SRR10168380	0
SRR10168381	0
SRR10168382	0
SRR10168383	0
SRR10168384	0
SRR10168385	0
SRR10168386	0
SRR10168387	0
SRR10168388	0
SRR10168389	0
SRR10168390	0
SRR10168391	0
SRR10168392	14

Assembly of Pangolin SARS-Cov-2 Genome

For assembling the CoV-2 genome we focused on the SRR10168376, SRR10168377 and SRR10168378 datasets. We re-downloaded those datasets in fastq format and repeated the **Align to Folder** searches using `2019-nCov/USA-AZ1` and `Bat CoV-RaTG13`, a coronavirus from bats that has been previously shown to be related to human SARS-CoV-2. Again, pairs of hits were aligned to `2019-nCov/USA-AZ1` and `Bat CoV-RaTG13` to confirm they contained coronavirus sequence and any non-matching reads were removed from the dataset.

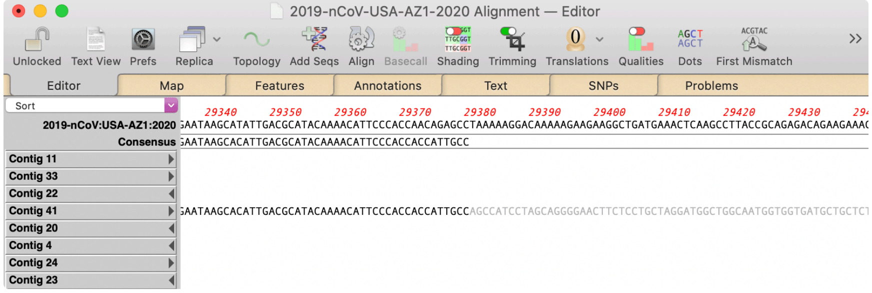
Assembly with phrap

Coronaviruses have a single positive strand RNA genome of around 30,000 nt in length which is transcribed to generate a variety of spliced RNA products in infected cells. This leads to a significant technical problem when trying to assemble a genome from RNA-Seq data as the reads will be a combination of genomic single stranded RNA and spliced or otherwise processed mRNA. To illustrate this, the hits from SRR10168377 were added to a new **Assembly Project** and assembled with ***phrap*** using the **Short Read Defaults** settings.



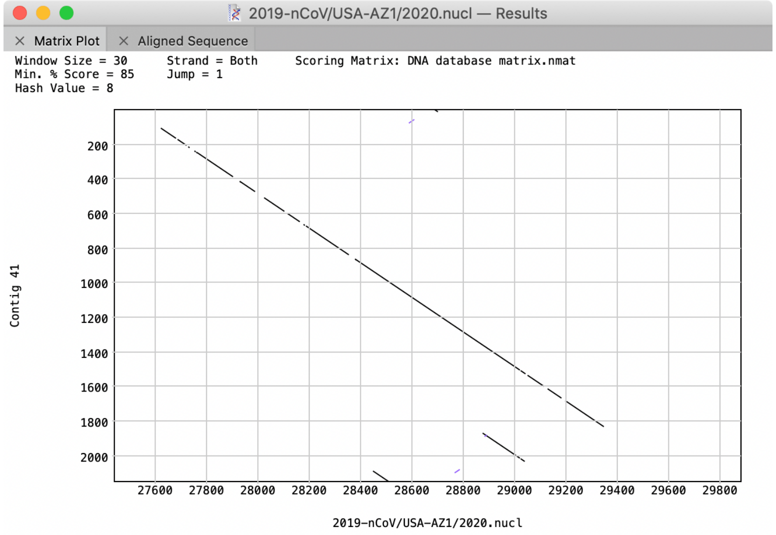
Name	Status	Length	#	ClipL	ClipR	Start	Stop	Definition
Phrap 1 - 15:37 - M...								
Contig 41		2150	169	1	2150	1	2150	
Contig 37		1892	83	1	1892	1	1892	
Contig 35		1838	63	1	1838	1	1838	
Contig 40		1776	156	1	1776	1	1776	
Contig 39		1601	121	1	1601	1	1601	
Contig 36		1578	75	1	1578	1	1578	
Contig 34		1078	47	1	1078	1	1078	
Contig 33		900	27	1	900	1	900	
Contig 31		892	18	1	892	1	892	
Contig 38		881	84	1	881	1	881	
Contig 28		824	13	1	824	1	824	
Contig 30		726	18	1	726	1	726	
Contig 32		688	20	1	688	1	688	
Contig 19		646	8	1	646	1	646	
Contig 21		593	9	1	593	1	593	
Contig 27		538	12	1	538	1	538	
Contig 29		525	14	1	525	1	525	
Contig 26		503	12	1	503	1	503	
Contig 20		439	8	1	439	1	439	

The contigs were exported as individual MacVector `.nucl` files by selecting them in the **Project** tab and choosing **File | Export Selected Contigs to...**. They were then aligned to `2019-nCov/USA-AZ1` using **Analyze | Align to Reference**. Many of the contigs showed extensive regions of misalignment, illustrated for `Contig 41` below by the greyed-out sequence;



```
                      29340     29350     29360     29370     29380     29390     29400     29410     29420     29430     294
2019-nCoV:USA-AZ1:2020 GAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCCTAAAAAGGACAAAAAGAAGAAGGCTGATGAAACTCAAGCCTTACCGCAGAGACAGAAGAAAC
             Consensus GAATAAGCACATTGACGCATACAAAACATTCCCACCACCATTGCC
Contig 11
Contig 33
Contig 22
Contig 41              GAATAAGCACATTGACGCATACAAAACATTCCCACCACCATTGCCAGCCATCCTAGCAGGGGAACTTCTCCTGCTAGGATGGCTGGCAATGGTGGTGATGCTGCTCT
Contig 20
Contig 4
Contig 24
Contig 23
```

The problems were visualized more graphically by performing an **Analyze | Create Dot Plot | Pustell DNA Matrix** analysis between `2019-nCov/USA-AZ1` and `Contig 41`;



```
Window Size = 30     Strand = Both     Scoring Matrix: DNA database matrix.nmat
Min. % Score = 85    Jump = 1
Hash Value = 8
```

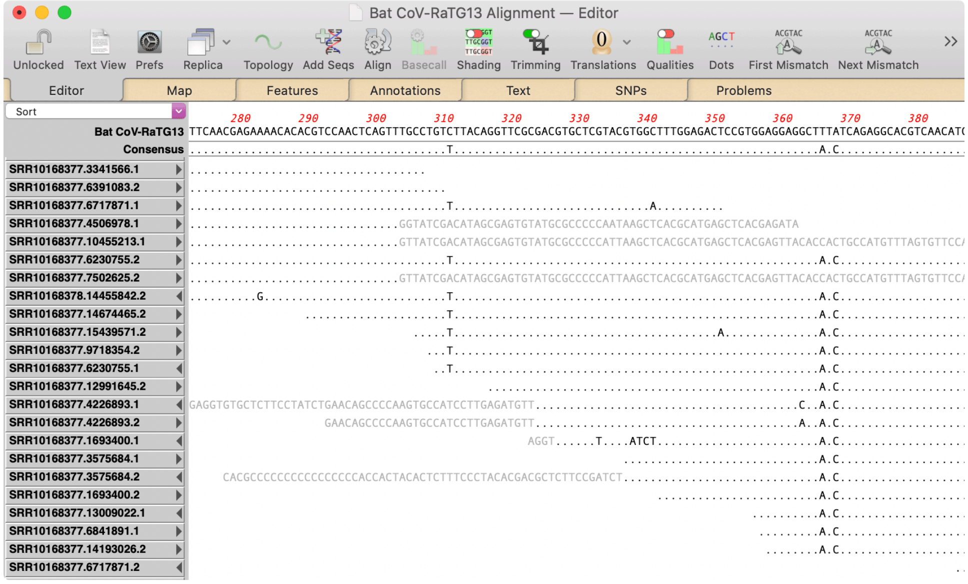
The zoomed-in view clearly shows how, instead of a single diagonal running from top left to bottom right, there are at least three diagonals with the end of `Contig 41` having similarity to dis-contiguous sections of `2019-nCov/USA-AZ1`. This is undoubtedly the result of spliced RNA variants interfering with the assembly algorithm.

Accordingly, we used an alternative strategy to determine the genomic sequence by aligning the reads directly against human and bat coronavirus genomes.

Assembly via Align to Reference

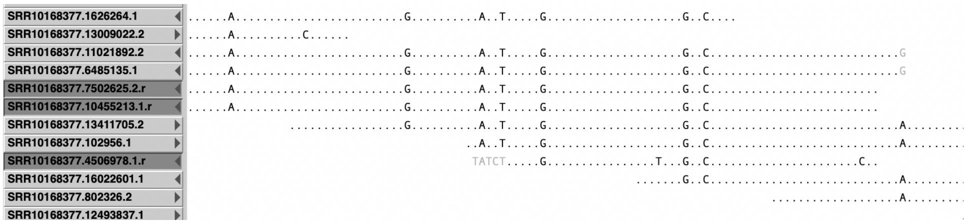
Preliminary experiments revealed that (a) none of the original sets of Pangolin RNASeq reads contained enough coronavirus samples to fully cover the SARS-CoV-2 genome and (b) there were minimal differences between CoV-2 reads from the different samples. Accordingly, to help obtain maximum genome coverage, the CoV-2-matching reads from SRR10168376, SRR10168377 and SRR10168378 were pooled for the initial assembly procedure.

The reads were aligned to `2019-nCov/USA-AZ1` and also separately to `Bat CoV-RaTG13` using the default **Align to Reference** settings. The presence of multiple spliced RNA species in the samples was immediately apparent from viewing the alignments;



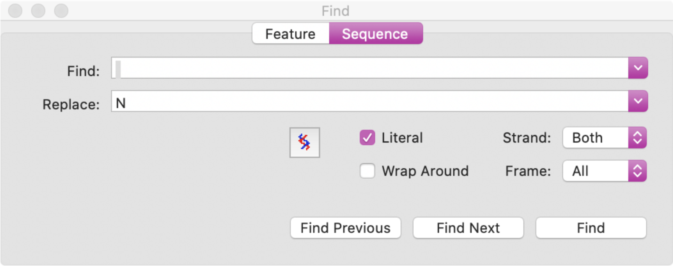
```
                             280       290       300       310       320       330       340       350       360       370       380
        Bat CoV-RaTG13 TTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTCTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGCTTTATCAGAGGCACGTCAACAT
             Consensus ......................................T......................................................A.C..................
SRR10168377.3341566.1  ...................................
SRR10168377.6391083.2  ......................................
SRR10168377.6717871.1  ......................................T.............................A..........
SRR10168377.4506978.1  ...............................GGTATCGACATAGCGAGTGTATGCGCCCCCAATAAGCTCACGCATGAGCTCACGAGATA
SRR10168377.10455213.1 ...............................GTTATCGACATAGCGAGTGTATGCGCCCCCATTAAGCTCACGCATGAGCTCACGAGTTACACCACTGCCATGTTTAGTGTTCCA
SRR10168377.6230755.2  ......................................T......................................................A.C..................
SRR10168377.7502625.2  ...............................GTTATCGACATAGCGAGTGTATGCGCCCCCATTAAGCTCACGCATGAGCTCACGAGTTACACCACTGCCATGTTTAGTGTTCCA
SRR10168378.14455842.2 ..........G...........................T......................................................A.C..................
SRR10168377.14674465.2                  .....................T......................................................A.C..................
SRR10168377.15439571.2                                  .....T.......................................A..............A.C..................
SRR10168377.9718354.2                                     ...T......................................................A.C..................
SRR10168377.6230755.1                                      ..T......................................................A.C..................
SRR10168377.12991645.2                                             .................................................A.C..................
SRR10168377.4226893.1  GAGGTGTGCTCTTCCTATCTGAACAGCCCCAAGTGCCATCCTTGAGATGTT.......................................C..A.C..................
SRR10168377.4226893.2                      GAACAGCCCCAAGTGCCATCCTTGAGATGTT.......................................A..A.C..................
SRR10168377.1693400.1                                                    AGGT......T....ATCT........................A.C..................
SRR10168377.3575684.1                                                                   .............................A.C..................
SRR10168377.3575684.2       CACGCCCCCCCCCCCCCCCCACCACTACACTCTTTCCCTACACGACGCTCTTCCGATCT.............................A.C..................
SRR10168377.1693400.2                                                                        ........................A.C..................
SRR10168377.13009022.1                                                                                     ..........A.C..................
SRR10168377.6841891.1                                                                                       .........A.C..................
SRR10168377.14193026.2                                                                                       ........A.C..................
SRR10168377.6717871.2                                                                                                                       ..
```

While MacVector does have an alignment mode tuned to standard splicing events, many of these chimeric reads were found to not follow typical splicing rules. However, MacVector 17.5.3 has an option accessible through a right-click (or <ctrl>-click) menu option **Cut Clipped Residues and Re-align**. This cuts all significant "trimmed" (greyed out in the above figure) residues, renames them with the addition of a ".r" (Right) or ".l" (Left) and re-aligns them against the reference with the last used parameters. As the new reads are kept selected after assembly, it's easy to see how they often have extensive identity to other regions of the reference sequence;



```
SRR10168377.1626264.1     ......A.........................G..........A..T....G....................G..C....
SRR10168377.13009022.2    ......A..........C......
SRR10168377.11021892.2    ......A.........................G..........A..T....G....................G..C...........................G
SRR10168377.6485135.1     ......A.........................G..........A..T....G....................G..C...........................G
SRR10168377.7502625.2.r   ......A.........................G..........A..T....G....................G..C........................
SRR10168377.10455213.1.r  ......A.........................G..........A..T....G....................G..C........................
SRR10168377.13411705.2                   .................G..........A..T....G....................G..C...........................A.........
SRR10168377.102956.1                                                     ..A..T....G....................G..C...........................A.........
SRR10168377.4506978.1.r                                                  TATCT.....G.................T..G..C......................C..
SRR10168377.16022601.1                                                                           .......G..C...........................A.........
SRR10168377.802326.2                                                                                                 ..................A.........
SRR10168377.12493837.1
```

After two rounds of **Cut Clipped Residues and Re-align** we obtained an alignment with a consensus sequence that had spaces where no reads overlapped the reference and gaps to maintain the alignment. To save the consensus without gaps, but with the spaces replaced by N's we double-clicked on the reference to select the entire sequence then used **Edit | Copy** followed by **File | New From Clipboard** to create a new sequence. To replace the spaces with Ns, we invoked **File | Find | Replace...**, and clicked on the **Literal** button so that we could replace "<space>" with "N";



Find:
Replace: N
☑ Literal
☐ Wrap Around
Strand: Both
Frame: All

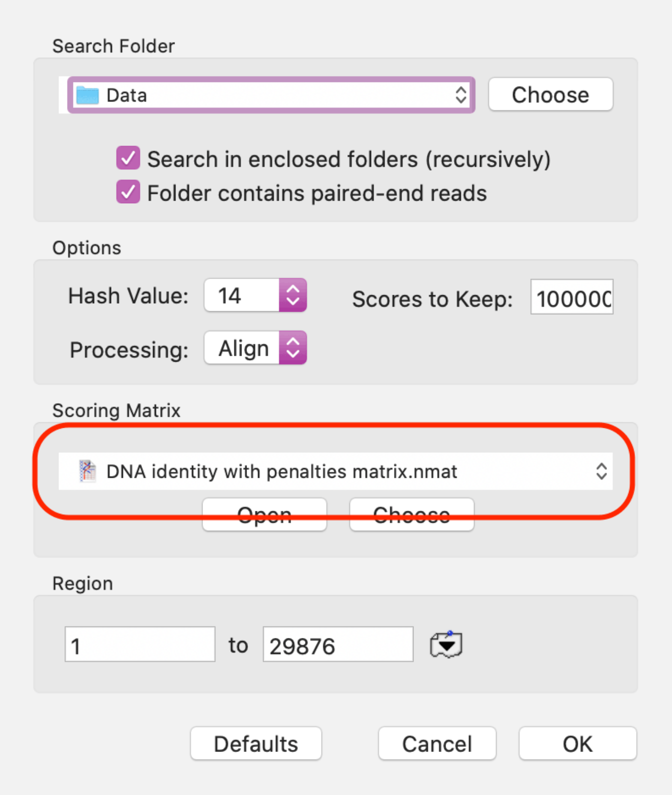
Similarly, we deleted any gap characters ("-") to generate a final consensus for each of the human and bat reference sequences.

Generating a Combined Consensus

The two consensus sequences from the bat and human **Align to Reference** alignments were then aligned using *ClustalW* to generate a combined consensus. The Ns in the two consensus sequences do interfere with getting a clean alignment so some minor editing was required after the alignment completed to optimize the assembly. Finally, we opened the **Prefs** dialog and checked the **No spaces in consensus** checkbox in the **Consensus** tab to replace space characters with Ns. Finally, we exported the combined consensus by clicking on the **Consensus** toolbar button and checking the **Create Consensus Sequence** checkbox.

Confirming the Consensus

The consensus from the *ClustalW* alignment was then used to scan the CoV-2-containing Pangolin datasets once again with **Align to Folder** to ensure that every appropriate read had been recovered. One important change from the initial search was to use the *DNA Identities with Penalties* scoring matrix. This dramatically reduced the number of false positives while ensuring that reads with short overlaps to the ends of consensus sequences were identified.



Search Folder

Data

☑ Search in enclosed folders (recursively)

☑ Folder contains paired-end reads

Options

Hash Value: 14

Scores to Keep: 10000C

Processing: Align

Scoring Matrix

DNA identity with penalties matrix.nmat

Region

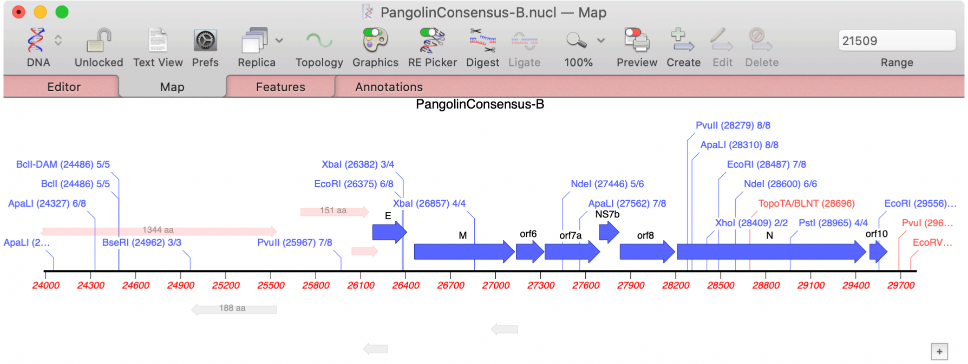
1 to 29876

All matching reads were then realigned to the combined consensus using the *Align to Reference* function as described above.

We then carefully walked through the alignment in the **Editor** to be sure that all residues in the consensus were accounted for and attempted to resolve any ambiguities where possible by editing and/or local realignment. Where there were differences in the reads from the 3 different pangolin datasets, we prioritized the SRR10168377 reads to resolve ambiguities. The final consensus was 29,802 nt in length with 3,351 N's added in 36 gapped segments to maintain alignment with the human and bat CoV-2 sequences with 7 other ambiguities due to discrepancies between the reads. Average read coverage was 7.0x.

Automatic Feature Annotation

Despite the large number of gaps, we annotated the final consensus sequence using MacVector's **Database | Auto-annotate Sequence** function using a collection of annotated human and bat CoV-2 genomes downloaded from *Entrez* as source sequences. As expected, there were too many gaps for the large non-structural ~21kb *orf1ab* coding region to be automatically identified, but many of the smaller structural proteins including the Envelope (*E*), Membrane (*M*) and Nucleocapsid (*N*) proteins were directly annotated;



The Spike (*S*) protein is partially present – some of the corresponding ORF can be seen on the left side of the **Map** above, but several large gaps in the consensus meant that a definitive amino acid sequence could not be automatically obtained.

Comparison with Human and Bat CoV-2 Genomes

The RNA genome sequences from bat (`Bat CoV-RaTG13`), a representative human CoV-2 sequence (`BetaCoV/Wuhan/IPBCAMS-WH-04/2019`) and pangolin CoV-2 were compared within MacVector using the multiple sequence alignment functionality with automatic alignment using ClustalW. Because the pangolin CoV-2 genome was incomplete due to a lack of reads, the longest 15 segments were compared with the following results;

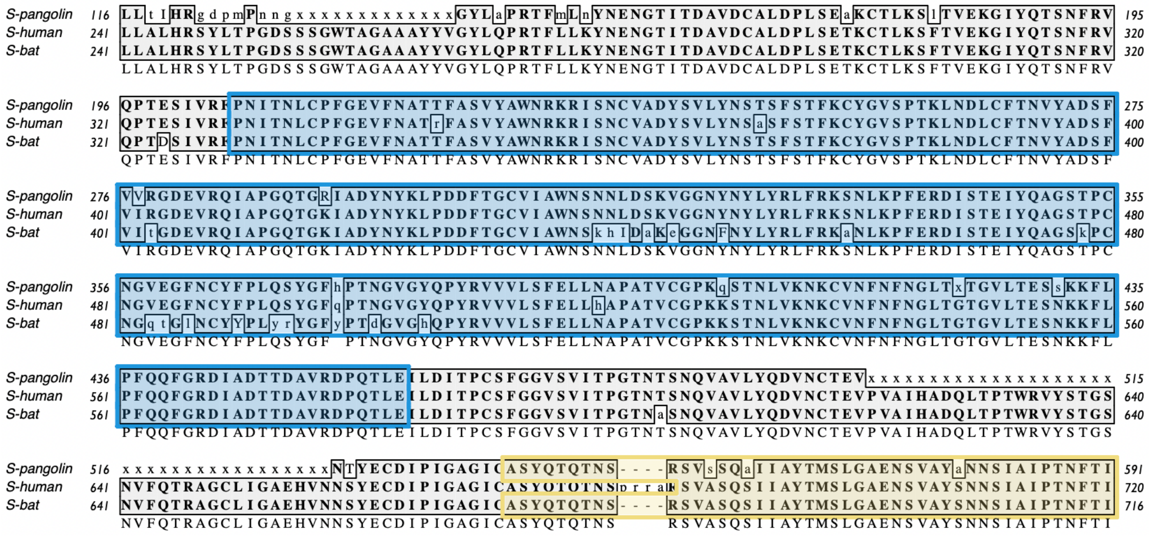
	Pangolin vs Human			***Pangolin vs Bat***			***Bat vs Human***		
Location	*Length*	*Mismatches*	*Identity %*	*Length*	*Mismatches*	*Identity %*	*Length*	*Mismatches*	*Identity %*
121	1414	125	91.2	1414	118	91.7	1414	49	96.5
1554	617	75	87.7	617	71	88.3	617	38	93.8
2863	402	71	82.8	402	72	82.5	400	28	93.0
4210	538	69	87.2	538	67	87.5	538	23	95.7
4753	1324	151	88.4	1324	147	88.7	1324	59	95.5
7267	519	49	90.6	519	41	92.1	519	14	97.3
8227	2015	192	90.3	2015	184	90.7	2015	85	95.8
11198	3388	232	93.2	3388	246	92.7	3388	81	97.6
14686	3419	307	91.0	3419	313	90.8	3419	70	98.0
18221	1549	167	89.2	1540	166	89.3	1539	43	97.2
19849	2356	350	86.0	2356	351	85.9	2333	110	95.3
22392	1063	136	87.1	1067	177	83.6	1065	112	89.8
23569	1559	157	89.9	1547	130	91.6	1559	95	93.9
25667	3730	198	94.7	3730	200	94.6	3730	110	97.1
29504	410	9	97.3	410	8	97.6	419	6	98.5
Total	**24303**	**2288**	**90.6**	**24286**	**2291**	**90.6**	**24279**	**923**	**96.2**

The overall identity between the pangolin and human CoV-2 genomes was just under 91% whereas the bat and human genomes shared over 96% identity. Every individual pangolin vs human segment exhibited less identity than the corresponding bat vs human segment.

The amino acid sequences of the E, M and N proteins from bat, human and pangolin CoV-2 were aligned using ClustalW. The short E protein was identical across the three species. The pangolin M and N proteins had 3 and 10 changes respectively compared to the human proteins, whereas the bat M and N proteins had just 1 and 4 changes compared to the human proteins.

A partial amino acid sequence of the Spike (S) protein from the pangolin CoV-2 was determined by translating from an AUG start codon near the predicted beginning of the protein and terminating at a run of N's about 300 nt short of the expected termination codon, which was not covered by reads. The 1047 amino acid protein (compared to 1270 for the bat and human proteins) contained 51 "X" residues due to ambiguities in the RNA sequence. The S proteins from all three genomes were then annotated using the MacVector implementation of the online

InterProScan function. Specifically, each protein was annotated with the *Spike Receptor Binding* domain and the *Coronavirus-2* domain. The sequences were then aligned, and the domains visualized in the MSA Picture tab;



```
S-pangolin 116 LLtIHRgdpmPnngxxxxxxxxxxxxxGYLaPRTFmLnYNENGTITDAVDCALDPLSEaKCTLKSlTVEKGIYQTSNFRV 195
S-human    241 LLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRV 320
S-bat      241 LLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRV 320
               LLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRV

S-pangolin 196 QPTESIVRFPNITNLCPFGEVFNATTFASVYAWNRKRISNCVADYSVLYNSTSFSTFKCYGVSPTKLNDLCFTNVYADSF 275
S-human    321 QPTESIVRFPNITNLCPFGEVFNATrFASVYAWNRKRISNCVADYSVLYNSaSFSTFKCYGVSPTKLNDLCFTNVYADSF 400
S-bat      321 QPTDSIVRFPNITNLCPFGEVFNATTFASVYAWNRKRISNCVADYSVLYNSTSFSTFKCYGVSPTKLNDLCFTNVYADSF 400
               QPTESIVRFPNITNLCPFGEVFNATTFASVYAWNRKRISNCVADYSVLYNSTSFSTFKCYGVSPTKLNDLCFTNVYADSF

S-pangolin 276 VVRGDEVRQIAPGQTGRIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPC 355
S-human    401 VIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPC 480
S-bat      401 VItGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSkhIDaKeGGNFNYLYRLFRKaNLKPFERDISTEIYQAGSkPC 480
               VIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPC

S-pangolin 356 NGVEGFNCYFPLQSYGFhPTNGVGYQPYRVVVLSFELLNAPATVCGPKqSTNLVKNKCVNFNFNGLTxTGVLTESsKKFL 435
S-human    481 NGVEGFNCYFPLQSYGFqPTNGVGYQPYRVVVLSFELLhAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFL 560
S-bat      481 NGqtGlNCYYPLyrYGFyPTdGVGhQPYRVVVLSFELLNAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFL 560
               NGVEGFNCYFPLQSYGF PTNGVGYQPYRVVVLSFELLNAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFL

S-pangolin 436 PFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVxxxxxxxxxxxxxxxxxxxx 515
S-human    561 PFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGS 640
S-bat      561 PFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNaSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGS 640
               PFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGS

S-pangolin 516 xxxxxxxxxxxxxxxxxNTYECDIPIGAGICASYQTQTNS----RSVsSQaIIAYTMSLGAENSVAYaNNSIAIPTNFTI 591
S-human    641 NVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSprraRSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI 720
S-bat      641 NVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS----RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI 716
               NVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS    RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
```

Here, the 253aa *Spike Receptor Binding* domain is outlined and filled in blue. The pangolin sequence has just a single X within the domain (at position 421). Other than that, the pangolin and human sequences have 8 differences (96.8% identity) whereas the bat and human sequences have 20 differences (92.1% identity), with most discrepancies (10) clustered between 477 and 505 where there is just a single discrepancy between pangolin and human.

A detailed examination of the potential role of the *Spike Receptor Binding* domain in CoV-2 host specificity is described in this publication and is not considered in more detail here.

Comparison with Published Pangolin CoV-2 Genome

The final pangolin sequence was aligned with the published sequence (MT084071) using ClustalW. A certain amount of editing was required to generate an optimal alignment due to differences in the number of N's inserted in the non-covered regions of MT084071. The MT084071 genome contains 25,491 unambiguous residues compared to the 26,444 unambiguous residues in the sequence we determined. There were 43 non-N mismatches between the two sequences.

We then used MacVector's **Align to Reference** function to align our entire collection of CoV-2 containing pangolin reads against the MT084071 genome to examine the discrepancies between the two sequences. In almost every case, we determined that the sequence we had determined was more likely to be correct. While some of the discrepancies were due to our use of additional non-SRR10168377 reads to help improve coverage, others appeared to be simply reads in the SRR10168377 dataset that the original authors had missed.

Conclusions

We have shown here that MacVector can be used on modest Macintosh personal computers to data mine existing NGS datasets and retrieve low abundance reads representing viruses and potentially other infectious agents. The tools within MacVector allowed the construction of a high-quality pangolin SARS-CoV-2 genome and comprehensive RNA and analysis confirmed that the viral genome shares around 90% identity with a typical human SARS-CoV-2 strain, compared to over 96% identity between the bat SARS-CoV-2 strain `Bat CoV-RaTG13` and human strains. However, protein analysis of the predicted pangolin CoV-2 Spike (S) protein confirmed that the *Spike Receptor Binding* domain shares more similarity with the human *Spike Receptor Binding* domain than does the equivalent bat domain, supporting the hypothesis that recombination between bat and pangolin viruses may have been an intermediate step during the jump of SARS-CoV-2 from animals to humans.