An Investigation of Phylogeny and Recombination in SARS-CoV-2

Parker Cote, Ryan Eveloff, Weishan Li

Introduction

Emerging infectious diseases present a major challenge to the public health system. Characterizing novel pathogens enables scientists to trace their phylogeny, facilitate diagnosis, inform public health officials on policy making, and provide directions for remedy/vaccine development. Here we replicate the process of genome assembly, phylogenetic tree generation and recombination analysis on Mi-RNASeq samples obtained from a symptomatic man admitted to and hospitalized in the Central Hospital of Wuhan on 26 December 2019, one of the first victims of the COVID-19 pandemic [1].

We achieved de-novo assembly of the full genome of the SARS-Cov-2 virus from around 15,000,000 pair-ended, 150bp long Mi-RNAseq libraries and cross-confirmed results using results from multiple assemblers as well as QUAST quality assessment tool.

We also recovered the same 2 recombination events as the original paper within the spike protein of our de-novo assembly with WIV1 coronavirus and Bat SARS-like coronavirus isolate (Bat-SL-CovZC45). We subsequently generated phylogenetic trees for the 3 recombinant regions delimited by the 2 recombination events with a selection of SARS-like and betacoronavirus strains. Our trees for the first and third recombinant regions (the major parental region) were highly analogous to the ones reported in the original work, suggesting these regions of our de-novo assembly being most related to Bat-SL-CovZC45 and Bat-CovZXC21.

Interestingly, our tree for the second recombinant region (the alleged minor parental region) also suggested a close relationship with Bat-SL-CovZC45 and Bat-CovZXC21, as opposed to WIV1 coronavirus in the original paper. We do note that the second tree from the original work had multiple branches with sub 90 bootstrapping scores, while ours don’t, indicating that they may have cherry picked their second tree so that the difference is more pronounced.

Methods

De-novo Assembly

Read library QC and adaptor trimming was done using trimmomatic 0.39 [2]. The original paper did not mention the version they used, and hence we installed the newest available release 0.39 from its official github site https://github.com/usadellab/Trimmomatic.

The original paper did not mention any parameters they used, nor did they talk about the adapter file we should supply to trimmomatic. We therefore adapted the example parameters from the trimmomatic official site: keepBoth Reads LEADING:3 TRAILING:3 MINLEN:36, removing the leading and trailing 3 bases of bad quality, and ensuring that a read has a minimum length of 36.

ILLUMINACLIP:trimmomatic-master/adapters/TruSeq3-PE.fa:2:30:10:2: was specified as the adapter PE file, since the SRA accession number says the experiment was performed on MiniSeq platform (Illumina), which adapters contained in TruSeq3 file.

Command(s) used:

```
java -jar ~/bin/trimmomatic.jar PE -threads 16 -phred33 \
     -trimlog tlog.txt [read_1.fq]  [read_2.fq] \
     [out_1_paired.fq] [out_1_unpaired.fq] [out_2_paired.fq] \
     [out_2_unpaired.fq]  \
     ILLUMINACLIP:trimmomatic-master/adapters/TruSeq3-PE.fa:2:
     30:10:2:keepBoth Reads LEADING:3 TRAILING:3 MINLEN:36
```

The QC and trimming process reduces the original library of 28,282,964 pairs to 938, 207 read pairs. Note that trimmomatic also outputs the unpaired reads after QC ([out_1_unpaired.fq], [out_2_unpaired.fq]) are dropped. The assembly is carried out with these reads.

We used SPAdes 3.15.2 [3], Megahit 1.2.9 [4] and Trinity 2.8 [5] for our assembly. The original paper said they used default parameters of Megahit and Trinity. However, to make Trinity work on our side, we had to specify --no_bowtie because the software would not recognize our bowtie2 installation. For SPAdes, we additionally specified using --rna viral mode.

Command(s) used:

```
# spades
bin/spades.py --rnaviral -1 [out_1_paired.fq] -2 \
           [out_2_paired.fq] -o [assembly_dir]

# Megahit
megahit -1 [out_1_paired.fq] -2 [out_2_paired.fq] -o \
           [assembly_dir]

# Trinity
./Trinity --seqType fq --max_memory 28G  \
           --left [out_1_paired.fq] \
           --right  [out_2_paired.fq]  \
           --CPU 6  --output [assembly_dir]  --no_bowtie
```

Alignment quality accession

The assembled contigs from all assemblers are uploaded to online QUAST tool (http://cab.cc.spbu.ru/quast/) aligned against reference SARS-CoV-2 genome NC 045512.2. Contigs less than 500 base pairs of length are dropped.

Phylogenetic Analysis

Sequences of the three regions of the SARS-CoV-2 reference genome were generated by splicing the reference at the positions suggested by the authors. The first sequence spans base pairs 1-1,028 of the spike protein, the second sequence spans base pairs 1,029-1,652, and the third and final sequence spans base pairs 1,653-3,804. These sequences were separated into three different fasta files, each of which also contained the rest of the sequences used by the authors for phylogenetic analysis. We then performed multiple alignments on each of these fasta files using the L-INS-i algorithm in MAFFT v7.475. The multiple alignments were then used as the input for maximum likelihood phylogenetic tree generation for each of the three regions which was performed using a generalized time-reversible substitution model in RAxML. We then continued to use RAxML to perform a bootstrap analysis on our trees and draw bipartitions on the best tree generated for the multiple alignments associated with each region.

Command(s) used:

```
# Multiple Alignment (L-INS-i MAFFT)
linsi mafft [input.fa] > [output.aln]
```

```
# Maximum Likelihood Tree Generation (RAxML)
raxml -s [input.aln] -m GTRCAT -n [region] -p 12345

# Bootstrap Search (RAxML)
raxml -s [input.aln] -m GTRCAT -n [region_bootstrap] -b 12345

# Draw Bipartitions on Best Tree
raxml -m GTRCAT -p 12345 -f b -t RAxML_bestTree.[region] \
-z RAxML_bootstrap.[region_boostrap] -n [region_bs]
```

We then uploaded the best phylogenetic tree for each region into the Interactive Tree of Life (iTOL) to visualize our trees shown in our results.

Recombination analysis

Spike protein sequences of Bat SARS coronavirus Rf1 (DQ412042), Bat SL-CoVZC45 (MG772933), SARS-CoV Tor2 (AY274119) and Bat SARS-like coronavirus WIV1 (KF367457) were downloaded from NCBI database.

We manually extracted the 21563~25384bp region (as indicated by the NC 045512.2 SARS-CoV-2 reference annotation file to be the CDS for spike protein) from our megahit de-novo assembly and called it spike. To confirm the region is valid, we put the region through NCBI online BLAST and it aligned to the same spike protein region of multiple SARS-Cov-2 samples worldwide with no problem.

The above sequences are concatenated to a single fasta file and passed to mafft for multiple alignment, mafft strategy is set to --auto, output file format is set to fasta. The multiple alignment file was passed to RDP5 [6] via the open menu. The original paper uses RDP4, though we used RDP5 due to compatibility issues with Windows 10. Lastly, we ran automated RDP with RDP5.

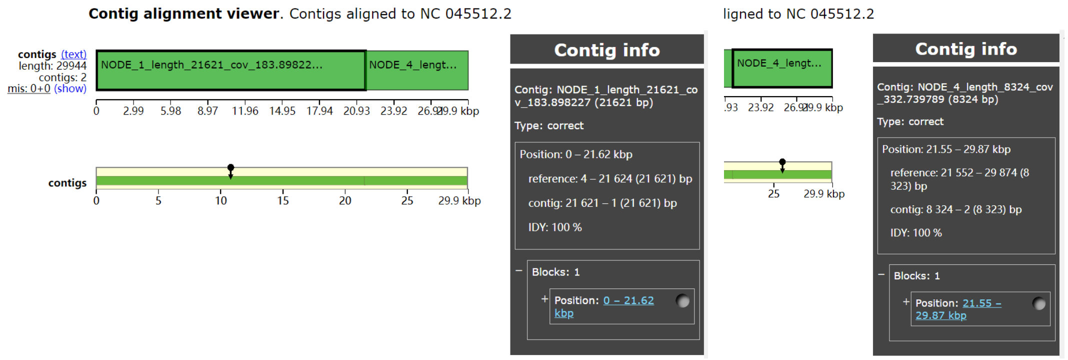
We selected reference_spike (our de-novo assembly) as the potential recombinant sequence and all others as potential parental sequences, then generated a plot using Window size = 200 and Step size = 20. Right click on the RDP plot and select save as .csv file to extract the actual plot data so we may plot ourselves. The original paper used Simplot for similarity plot. However, we were unable to download any usable version of Simplot (the official download server is down, and we did not find any mirrors). So we generated an initial similarity plot with

RDP5 and later used an in-house python script for analysis and visualization (code available here).

Results

Assembler Benchmarking

SPAdes



Contig alignment viewer. Contigs aligned to NC 045512.2

ligned to NC 045512.2

contigs (text)
length: 29944
contigs: 2
mis: 0+0 (show)

Contig info

Contig: NODE_1_length_21621_cov_183.898227 (21621 bp)

Type: correct

Position: 0 – 21.62 kbp

reference: 4 – 21 624 (21 621) bp

contig: 21 621 – 1 (21 621) bp

IDY: 100 %

− Blocks: 1

+ Position: 0 – 21.62 kbp

Contig info

Contig: NODE_4_length_8324_cov_332.739789 (8324 bp)

Type: correct

Position: 21.55 – 29.87 kbp

reference: 21 552 – 29 874 (8 323) bp

contig: 8 324 – 2 (8 323) bp

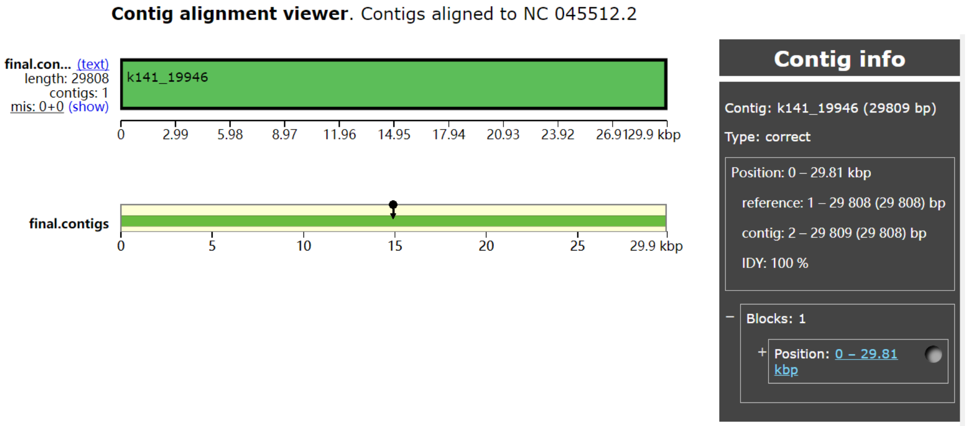
IDY: 100 %

− Blocks: 1

+ Position: 21.55 – 29.87 kbp

Fig. 1a: Icarus genome browser view of SPAdes assembly - The longest contig from SPAdes assembly was 21621 base pairs. The longest contig plus a 8324 base pair long contig spans the full SARS-CoV-2 genome. However, SPAdes failed to provide a single full contig that spans the viral genome.

Megahit



Contig alignment viewer. Contigs aligned to NC 045512.2

final.con... (text)
length: 29808
contigs: 1
mis: 0+0 (show)

Contig info

Contig: k141_19946 (29809 bp)

Type: correct

Position: 0 – 29.81 kbp

reference: 1 – 29 808 (29 808) bp

contig: 2 – 29 809 (29 808) bp

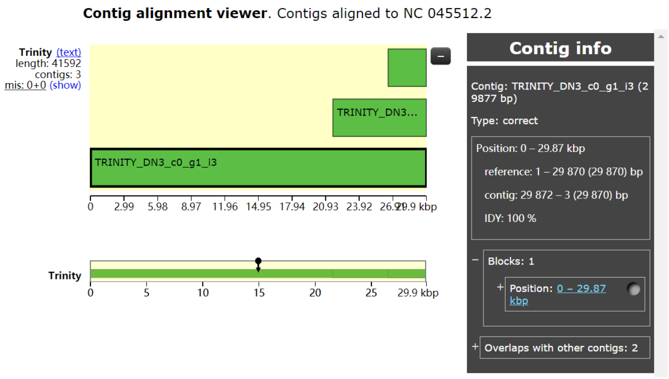
IDY: 100 %

− Blocks: 1

+ Position: 0 – 29.81 kbp

Fig. 1b Icarus genome browser view of Megahit assembly - The longest contig from Megahit assembly was 29808 base pairs, spanning the full SARS-CoV-2 genome with 1 extra nucleotide.

Trinity



Contig alignment viewer. Contigs aligned to NC 045512.2

Trinity (text)
length: 41592
contigs: 3
mis: 0+0 (show)

Contig info

Contig: TRINITY_DN3_c0_g1_i3 (29877 bp)

Type: correct

Position: 0 – 29.87 kbp

reference: 1 – 29 870 (29 870) bp

contig: 29 872 – 3 (29 870) bp

IDY: 100 %

Blocks: 1

Position: 0 – 29.87 kbp

Overlaps with other contigs: 2

Fig. 1c Icarus genome browser view of Trinity assembly - The longest contig from Trinity assembly was 29877 base pairs, spanning the full SARS-CoV-2 genome with 7 extra nucleotides. Interestingly, Trinity is the only assembler that produced overlapping contigs (the 2 other ones from the figure).

The performance statistics of each of the assemblers are presented below. The overall performance in terms of output contig quantity and length are generally similar. It seems like regardless of how long the assemblers run, the number of contigs they are able to produce in each length range given the same input reads are within the same order of magnitude. And since all 3 produced enough contigs to span the entire viral genome, the most efficient one, SPAdes, also producing least total unaligned length, might be most suitable for assembly for pathogen characterization purposes.

Spec\Assembler	SPAdes	Megahit	Trinity
Successful genome assembly	Yes	Yes	Yes
Single contig spanning entire genome	No	Yes	Yes
# contigs >= 0bp	115698	29318	152940
# contigs >= 500bp	4010	7458	7305
# contigs >= 1000bp	757	1432	1548
# contigs >= 5000bp	7	15	29
N50	803	817	869
N75	605	605	616

Total unaligned length	3305344	6276410	6370783
# unaligned contigs	4008	4008	7302
Time	2 hrs	6 hrs	24 hrs

.Phylogenetic Trees

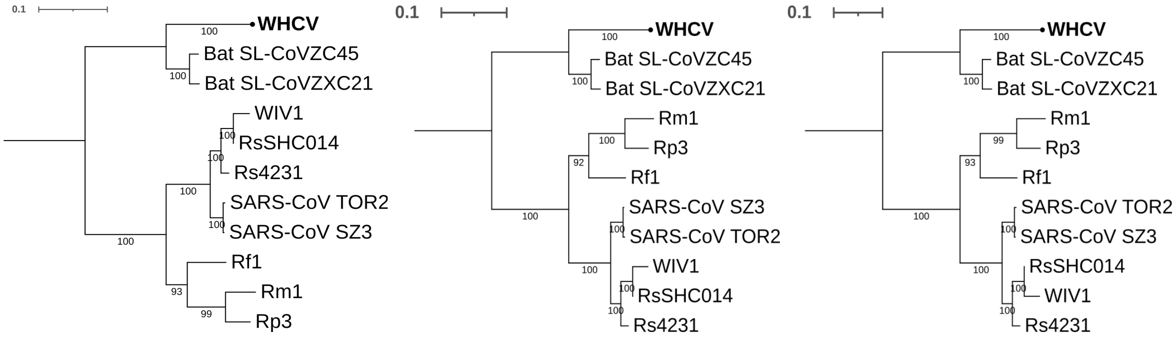


Fig. 2a Visualized phylogenetic trees for each of the three regions of the spike protein identified by the authors. Left is the region spanning base pairs 1-1,028 of the spike protein. Center is the region spanning base pairs 1,029-1,652 of the spike protein. Right is the region spanning base pairs 1,653-3,804 of the spike protein. All three regions seem to show similar branching patterns with high bootstrap values.

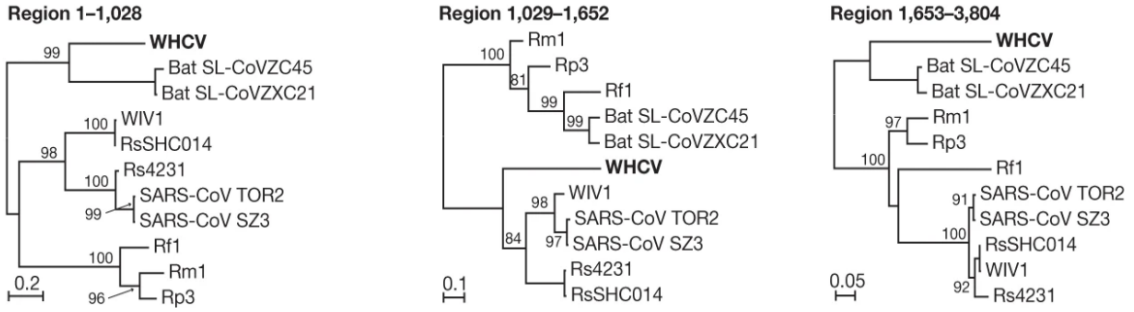


Fig. 2b Original phylogenetic trees from the authors. Left is the region spanning base pairs 1-1,028 of the spike protein. Center is the region spanning base pairs 1,029-1,652 of the spike protein. Right is the region spanning base pairs 1,653-3,804 of the spike protein. The first and third regions are very similar to one another and the trees we were able to generate, while the second region suggests a significantly different phylogeny. However, the bootstrap values for many important branches in the second tree generated by the authors show bootstrap values significantly lower than ours.

Recombination

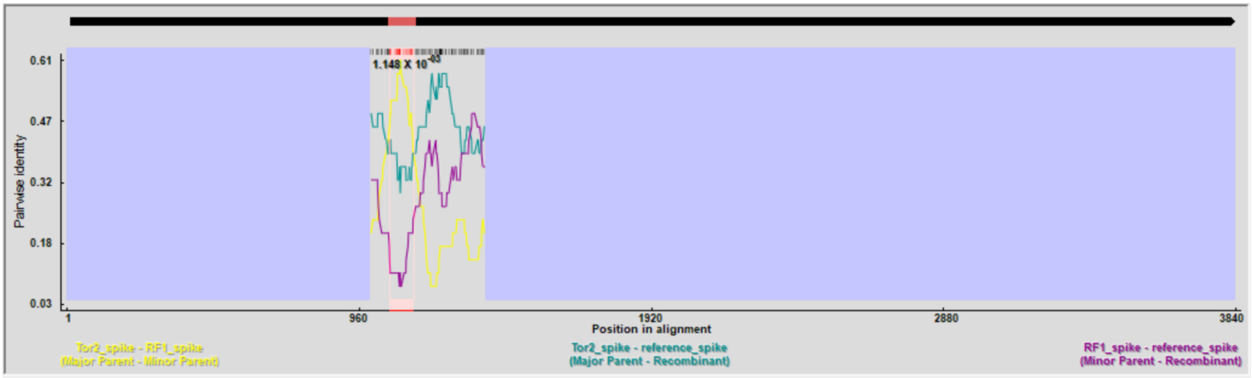


Fig. 3a RDP5 recombination analysis - de-novo with Tor2 strain (pine green), region bounded by red box spanning 1,061 - 1,136.

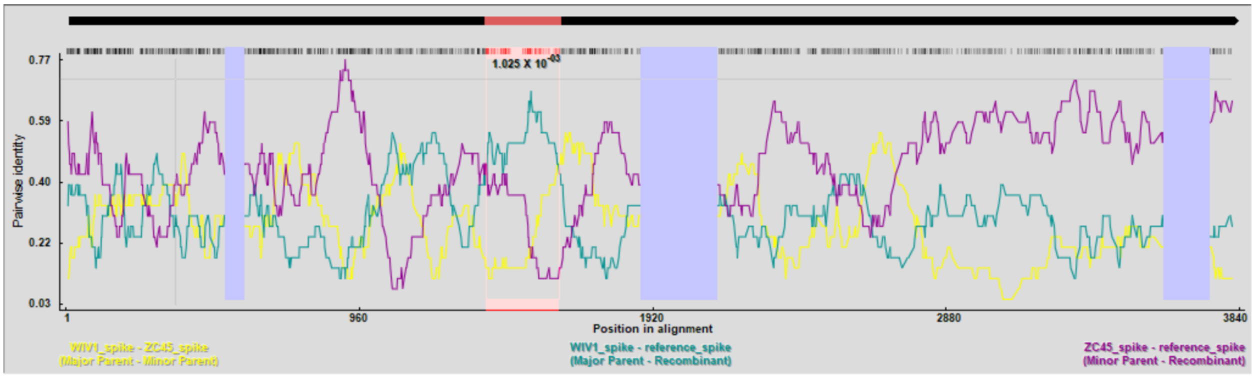


Fig. 3b RDP5 recombination analysis - de-novo with WIV1 strain (pine green), region bounded by red box spanning 1,372 - 1,615. In regions other than the 1,061 - 1,136 (Tor2) and 1,372 - 1,615 (WIV1) windows, we can see that our de-novo assembly is most identical to ZC45-spike (purple curve).

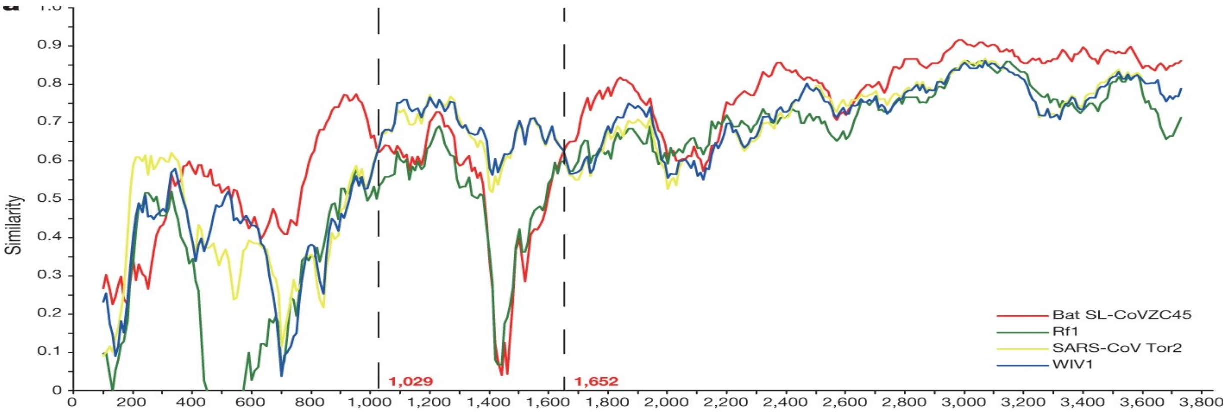


Fig. 3c Original analysis - Marks 1,029 - 1,652 region marked as recombinant. (The authors only provided a similarity plot but not the analytical details from RDP).

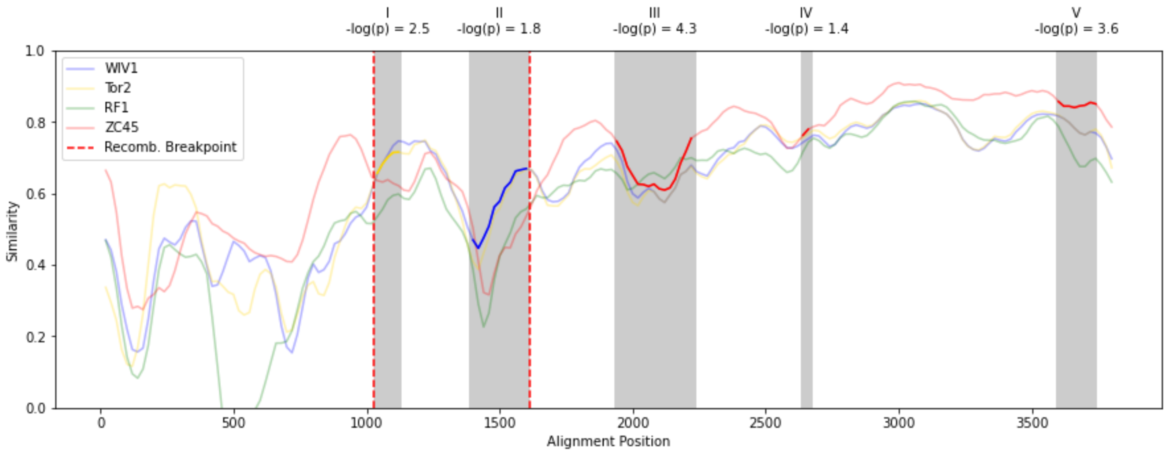


Fig. 3d: Final Recombination Analysis. Our improved rendition of the original figure. Each recombination event is highlighted with a dark grey, and each strain (by color) is plotted with its percent similarity to SARS-CoV-2 plotted as a function of its position in the spike protein. Additionally, each recombination event is annotated with a numeral and associated negative log-scaled p-value.

Our RDP analysis discovered a recombination event starting at ~1,061 bp of the de-novo spike genome with Tor2 strain as well as a recombination event with WIV1 strain ending at ~1,615 bp, both agreeing with the discoveries of the original paper. In addition, our analysis identified a more detailed version of the event, since we were able to attribute the recombinant region to 2 distinct parents. The authors found that “in phylogenies of the nucleotide fragments from 1 to 1,029 and from 1,652 to the end of the sequence, WHCV was most closely related to bat SL-CoVZC45 and bat SL-CoVZXC21, whereas in the region of nucleotides 1,030 to 1,651 (the RBD region) WHCV grouped with SARS-CoV and bat SARS-like CoVs (WIV1 and RsSHC014) that are capable of direct human transmission” [1].

Figure 3d is our final reconstruction of the published plot, but with multiple additions that we feel enhance its message. First, we highlighted all recombination regions found from our RDP analysis and attached their respective p-values. Interestingly, the two recombinations captured in the “breakpoint region” arbitrarily created by the original authors contained neither of the two most significant recombination events. Events III and V both occurred between SARS-CoV-2 and Bat SL-CoVZC45, with -log10(p)-values of 4.3 and 3.6, respectively. With that in mind, we argue that the original authors’ analysis yielded incomplete results and the arbitrary “breakpoint region” should either be recalculated using empirical means or removed altogether. In our figure, we used the region from the beginning of recombination event I to the end of recombination event II as the “breakpoint region” as it most closely resembles the one in the original figure.

Discussion

Limitations

Most notably, our analyses are limited to only one of the original cases of SARS-CoV-2 in China and do not account for variants/mutations over the last 18 months. Additionally, p-values obtained from RDP5 are statistically opaque as RDP5 does not make the algorithm responsible available to the end-user during runtime.

Challenges

We initially attempted to use Simplot (as used in the original publication), but there were no mirrors available and the original download link had been inoperable for a number of years. We then attempted to use RDP4, but it was incompatible with our version of Windows 10. Next, we began developing an in-house recombination tool; however, window size calculation and tuning other parameters was a difficult process. When we found RDP5 to work satisfactorily, we opted to use that instead. Similarly, PhyML would not run on our data, but given that we were already using new tools for assembly, we simply switched to RAxML.

Future Improvements

The single most beneficial improvement we can foresee is an open-source and readily available recombination analysis tool. RDP and Simplot are both antiquated, slow, and generally unusable. Speaking more towards the science, using more analogous pathogens (MERS-CoV, pangolin SARS-CoV, etc.) or even variants of SARS-CoV-2 could expand the scope of our conclusions. Alternatively, increasing bootstrapping or ML tree iterations could yield more statistically relevant results.

Appendix

Contributor	Role
Parker Cote	Phylogenetics
Ryan Eveloff	Recombination
Weishan Li	Assembly
Equal Contribution	Report, Presentation

References

1. Wu, F., Zhao, S., Yu, B. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269 (2020). https://doi.org/10.1038/s41586-020-2008-3
2. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170
3. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-477. doi:10.1089/cmb.2012.0021
4. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015 May 15;31(10):1674-6. doi: 10.1093/bioinformatics/btv033. Epub 2015 Jan 20. PMID: 25609793.
5. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644-652. Published 2011 May 15. doi:10.1038/nbt.1883
6. Darren P Martin, Arvind Varsani, Philippe Roumagnac, Gerrit Botha, Suresh Maslamoney, Tiana Schwab, Zena Kelz, Venkatesh Kumar, Ben Murrell, RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets, *Virus Evolution*, Volume 7, Issue 1, January 2021, veaa087, https://doi.org/10.1093/ve/veaa087