# Tiled-ClickSeq for targeted sequencing of complete coronavirus genomes with simultaneous capture of RNA recombination and minority variants.

Elizabeth Jaworski[1,2], Rose M. Langsjoen[1], Barbara Judy[3], Patrick Newman[3], Jessica A. Plante[4,5,6], Kenneth S. Plante[4,5,6], Aaron L. Miller[3], Yiyang Zhou[1], Daniele Swetnam[1], Jianli Dong[6,7], Ping Ren[7], Rick B. Pyles[3], Thomas Ksiazek[4,5], Vineet D. Menachery[5,6,7], Scott C. Weaver[5,6,7], Andrew Routh*[1,6,8]

1) Department of Biochemistry and Molecular Biology, The University of Texas Medical Branch, Galveston, TX, USA
2) ClickSeq Technologies LLC, Galveston, TX, USA
3) Department of Pediatrics, University of Texas Medical Branch, Galveston, TX, USA
4) Department of Pathology, University of Texas Medical Branch, Galveston TX, USA
5) World Reference Center for Emerging Viruses and Arboviruses, University of Texas Medical Branch, Galveston, TX, USA
6) Institute for Human Infections and Immunity, University of Texas Medical Branch, Galveston, TX, USA
7) Department of Microbiology and Immunology, The University of Texas Medical Branch, Galveston, TX, USA
8) Sealy Centre for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, Texas, USA.

*Correspondence: alrouth@utmb.edu

## ***Abstract***

High-throughput genomics of SARS-CoV-2 is essential to characterize virus evolution and to identify adaptations that affect pathogenicity or transmission. While single-nucleotide variations (SNVs) are commonly considered as driving virus adaption, RNA recombination events that delete or insert nucleic acid sequences are also critical. Whole genome targeting sequencing of SARS-CoV-2 is typically achieved using pairs of primers to generate cDNA amplicons suitable for Next-Generation Sequencing (NGS). However, paired-primer approaches impose constraints on where primers can be designed, how many amplicons are synthesized and requires multiple PCR reactions with non-overlapping primer pools. This imparts sensitivity to underlying SNVs and fails to resolve RNA recombination junctions that are not flanked by primer pairs. To address these limitations, we have designed an approach called '*Tiled-ClickSeq*'. Tiled-ClickSeq uses hundreds of tiled-primers spaced evenly along the virus genome in a single reverse-transcription reaction. The other end of the cDNA amplicon is generated by azido-nucleotides that stochastically terminate cDNA synthesis, obviating the need for a paired-primer. A sequencing adaptor containing a Unique Molecular Identifier (UMI) is appended using click-chemistry and a PCR reaction using Illumina adaptors generates a final NGS library. Tiled-ClickSeq provides complete genome coverage, including the 5'UTR, at high depth and specificity to virus on both Illumina and Nanopore NGS platforms. Here, we analyze multiple SARS-CoV-2 isolates and simultaneously characterize minority variants, sub-genomic mRNAs (sgmRNAs), structural variants (SVs) and D-RNAs. Tiled-ClickSeq therefore provides a convenient and robust platform for SARS-CoV-2 genomics that captures the full range of RNA species in a single, simple assay.

# ***Introduction***

Virus genomics and Next-Generation Sequencing (NGS) are an essential component of viral outbreak responses (1). Reconstruction of consensus genetic sequences is essential to identify adaptations correlated with changes in pathogenicity or transmission (2). In addition to single nucleotide variations, studies of SARS-CoV-2 have identified numerous genomic structural variants (SVs) (3) that arise due to non-homologous RNA recombination. SVs typically comprise small insertions/deletions that nonetheless allow the variant genome to independently replicate and transmit. Numerous SVs have been described for CoVs including deletions of the accessory open reading frames (aORFs) (4, 5) and changes in spike protein observed in the B.1.1.7 and other variants of concern (6). Adaptation of SARS-CoV-2 also occurs during passaging in cell-culture, such as small deletions that arise near the furin cleavage site of spike protein during amplification on Vero cells (7). These deletions can alter the fitness and virulence of SARS-CoV-2 isolates and thus must be genetically characterized prior to passaged stock use in subsequent studies.

Similar to SVs, non-homologous RNA recombination also gives rise to Defective-RNAs (D-RNAs), also known as Defective Viral Genomes (DVGs). D-RNAs have been observed in multiple studies of coronaviruses (CoVs), including mouse hepatitis virus (MHV) (8-11), bovine CoV (12), avian infectious bronchitis virus (IBV) (13), human CoV 299E (14-17). We recently demonstrated that SARS-CoV-2 is >10-fold more recombinogenic in cell culture than other CoVs such as MERS (18) and generates abundant D-RNAs containing RNA recombination junctions that most commonly flank U-rich RNA sequences. D-RNAs may change the fitness, disease outcomes and vaccine effectiveness for SARS-CoV-2 similar to other respiratory pathogens such as influenza and RSV (19). Together, these findings highlight the need to identify these RNA changes and their impact on SARS-CoV-2 infection and pathogenesis.

Whole genome sequencing can be achieved through a range of approaches including non-targeted (random) NGS of virus isolates amplified in cell culture or directly from patient samples. However, when input material is limited, low viral genome copy numbers necessitate a template-targeted approached followed by molecular amplification by PCR or iso-thermal amplification to generate sufficient nucleic acid for sequencing. Generally, these require knowledge of the virus genome and the design pairs of primers that anneal to the target genome. Perhaps the most popular method for SARS-CoV-2 sequencing is the ‘ARTIC’ approach (20), which can reliably identify SNVs and minority variants present in as little as 3% of genomes (21). However, the requirement for pairs of primers constrains where amplicons can be designed and imparts sensitivity to single nucleotide variants (SNVs). Multiple PCR reactions containing different pools of paired-primers must also be performed in order to obtain cDNA amplicons of the correct size and to prevent the interaction or mis-priming of PCR primers. Importantly, pairs of primers that do not flank RNA recombination junctions will be unable to detect unexpected or unpredicted RNA recombinant species. Finally, paired-primer approaches also necessitate the re-design and validation of alternative sets of primer-pairs for each specific NGS platform used (e.g. Illumina amplicons are 200-500 nts, Nanopore amplicons are ~2000-5000nts).

To address these limitations and optimize the ability of NGS to quantify all types of viral genetic variants, we have combined ‘*ClickSeq’* with tiled-amplicon approaches. ClickSeq (22, 23) is a click-chemistry based platform for NGS that prevents artifactual sequence chimeras in the output data (24). Using ClickSeq, the 3’end of an amplified cDNA segment is generated by the stochastic incorporation of terminating 3’ azido-nucleotides (AzNTPs) during reverse transcription. A downstream adaptor is *‘click-ligated’* onto the cDNA using copper-catalyzed azide-alkyne cycloaddition (CuAAC). Therefore, ‘*Tiled-ClickSeq*’ only requires one template-specific primer per cDNA amplicon. To achieve whole genome sequencing of a virus isolate or sample, multiple tiled primers are designed evenly along the virus genome. Only one pool of RT-primers is required, even when >300 template specific primers and their corresponding cDNA amplicons are generated in the same reaction. This simplifies the assay design, and importantly removes constraints imposed in paired-primer strategies (25). Furthermore, the same primer set can be used for both Illumina and Nanopore platforms even when requiring different cDNA amplicon sizes. The library construction allows for additional quality control features including the use of unique molecular identifiers (UMIs) in the ‘click-adaptor’ as well as the ability to identify each RT-primer that gives rise to specific cDNA amplicon when using paired-read NGS.

Here, we utilize the Tiled-ClickSeq method to analyze multiple isolates of SARS-CoV-2 and demonstrate that ‘*Tiled-ClickSeq*’ accurately reconstructs full-length viral genomes. The method also captures recombinant RNA species including sgmRNAs, SVs and D-RNAs. Overall, Tiled-ClickSeq therefore provides a convenient and robust platform for full genetic characterization of viral isolates.

# Methods

## Viruses and RNA extraction.

For WRCEVA isolates, viral RNA was obtained from supernatant materials of viral isolates amplified on Vero cells originally obtained from nasopharyngeal swab samples that tested positive in clinical laboratory assays for SARS-CoV-2 RNA, as described previously (26). The use of deidentified human samples was approved by the UTMB IRB under protocol 20-0088. The recombinant wild-type and 'PRRA-deletion' mutant SARS-CoV-2 are based on the sequence of USA-WA1/2020 isolate provided by the WRCEVA as previously described (27, 28). Wild-type and mutant SARS-CoV-2 were titrated and propagated on Vero E6 cells. RNAs were extracted from either total cellular materials or supernatants as indicated in the main text.

## SARS-CoV-2 reverse transcription primer design

A 'first' tiled-primer set (v1) containing 71 primers was designed cognate to the WA-1 SARS-CoV-2 genome (accession number: NC_045512.2) using the *primalseq* webserver (21) (http://primal.zibraproject.org/) with an amplicon distance of approximately 500nt in between each primer pair. We used only the 'right' primer sequences generated by *primalseq* and appended the Illumina p7 adaptor to these (e.g. GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT + NNNN + TGTCTCACCACTACGACCGTAC). We also included an additional primer designed to target the 3'-most 25 nts of the SARS-CoV-2 genome. A 'second' tiled-primer set was synthesized in a similar fashion using 326 loci described previously (29). A 'third' tiled-primer pool (v3) was generated by combining the v1 and v2 pools. Primers used in this study are provided as BED files with their loci and corresponding sequence in **SData 1**. Each primer was pooled in equimolar ratios to yield a SARS-CoV-2 specific primer pool used for the RT step of Tiled-ClickSeq.

## ***ClickSeq Library preps***

Random-primer ClickSeq NGS libraries were synthesized as described in previously published protocols from our lab (23, 30). For Tiled-ClickSeq, we make two important adjustments: 1) firstly, the primers used to initiate reverse transcription comprise pools of 10s-100s of virus-specific primer oligos; and 2) we anneal the RT-primers to the RNA template by incubating the RNA + primer mixture at 65ºC for 5mins, followed by a slow-cool of 1 degree per minute to a final temperature or 12ºC. RT-enzyme mixes are added at 12ºC and primer extension is performed for 10 mins at 55ºC. All subsequent steps of the Tiled-ClickSeq reaction, comprising RT cleanup, click-ligation, PCR amplification and cDNA library size-selection are identical to those used in the random-primed ClickSeq method and described previously (23, 30). The i5 'Click-Adaptor' was a reverse complement of the full Illumina Universal Adaptor sequence, plus an additional twelve 'N's at 5'-end to provide a Unique Molecular Identifier and functionalized with a 5'-hexynyl group (IDT). Final NGS libraries containing fragment sizes ranging 300-700 nts were pooled and sequenced on Illumina MiSeq, MiniSeq or NextSeq platforms using paired-end sequencings.

## ***Nanopore Sequencing***

Final cDNA libraries generated by the Tiled-ClickSeq protocol, although containing Illumina adaptors, are compatible with the Direct Sequencing by Ligation Kit (LSK-109) provided by Oxford Nanopore Technologies. cDNA library fragments >600nts in length are gel extracted and processed for nanopore sequencing using the manufacturer's protocols. The addition of demultiplexing barcodes can be achieved using the Native Barcoding by Ligation module (NBD104), again following the manufacturer's protocols. Single-plex or pooled cDNA libraries with ONT adaptors were loaded onto MIN-FLO109 flowcells on a MinION Mk1C and sequenced using the MinKNOW controller software for >24 hours. Raw FAST5 reads were base-called and demultiplexed using *Guppy*.

## Bioinformatics

All batch scripts and custom python scripts used in this manuscript are available in **SData 1**. Specific command-line entries and parameters can be found therein.

For Illumina reads, raw data were filtered and trimmed using *fastp (31)* to remove Illumina adaptors, quality filter reads and extract Unique Molecular Identifiers (UMIs). A custom python3 script was written to split the raw 'forward'/R1 reads into multiple individual FASTQ files depending upon the tiled-sequencing primer that is present in the first 30 nts of the 'reverse'/R2 paired-read. These split FASTQ files were then trimmed using *cutadapt* (32) to remove primer-derived sequences from the R1 reads. After trimming, all the split R1 files were re-combined to yield a final processed dataset. These reads were mapped to the WA-1 strain (NC_045512.2) of SARS-CoV-2 using *bowtie2* (33) and a new reference consensus genome was rebuilt for each dataset using *pilon* (34). Next, we mapped the processed read data to the reconstructed reference genome using *ViReMa* (35) to map to both the virus and the host (*chlSab2*) genome. SAM files were manipulated using *samtools (36)* and de-duplicated using *umi-tools (37)*. Minority variants were extracted using the *mpileup* command and a custom python3 script to count nucleotide frequency at each coordinate to find minority variants. Mapped data were visualized using the Tablet Sequence Viewer (38).

For Nanopore reads, *porechop* (https://github.com/rrwick/Porechop) was used to remove Illumina adaptor sequences and reads greater than 100nts in length were retained. These were mapped to the WA-1 SARS-CoV-2 genome (NC_045512.2) using *minimap2* (39) with the -*splice* option selected. Output SAM files were processed using *samtools (36)* and *bedtools (40)* to generate coverage maps.

## ***Data Availability Statement:***

All raw sequencing data (Illumina and Nanopore in FASTQ format) are available in the NCBI Small Read Archive with BioProject PRJNA707211. Consensus genomes for WRCEVA SARS-CoV-2 isolates reported in this manuscript are deposited at GenBank (MW047307-MW047318 and MW703487-MW703490).

# Results

## Overview of sequencing strategy

Most tiled approaches for complete viral genomes sequencing from viral isolates require the design of pairs of primers that generate pre-defined overlapping amplicons in multiple pools (**Fig 1A,B**). However, this can prevent the detection of recombinant viral genomic materials such as sub-genomic mRNAs (sgmRNAs) or Defective-RNAs (D-RNAs). To overcome these issues, we designed a template directed tiled-primer approach to reverse transcribe segments of the SARS-CoV-2 genome based upon the ‘ClickSeq’ method for NGS library synthesis (30). Instead of random-hexamer or oligo-dT primers as used in ClickSeq and Poly(A)-ClickSeq, respectively (41), we use multiple ‘tiled’ RT-primers designed at regular interval along the viral genome (**Fig 1C**). In ‘*Tiled-ClickSeq*’, pooled primers initiate a reverse transcription in a reaction that has been supplemented with 3’-azido-nucleotides (AzNTPs). This yields stochastically terminated 3’-azido-cDNA fragments, which can be click-ligated onto a hexynyl-functionalized Illumina i5 sequencing adaptor (**Fig 1D**). After click-ligation, the single-stranded triazole-linked cDNA is PCR-amplified using indexing p7 adaptors to fill in the ends of the NGS library, yielding the final library schema shown in **Figure 1E**. We designed the click-adaptor with an additional 12 random nucleotides at its 5’ end. As each adaptor can only be ligated once onto each unique cDNA molecule, this provides a unique molecular identifier (UMI) (42). Due to the stochastic termination of cDNA synthesis in the RT step, a random distribution of cDNA fragments is generated from each primer, giving rise to the hypothetical read coverage depicted in **Figure 1F**. The lengths of these fragments, and thus the obtained read coverage can be optimized to ensure overlapping read data from each amplicon by adjusting the ratio of AzNTPs to dNTPs in the RT reaction (30). With this approach, we found that we could robustly make NGS libraries from as little as 8ng of total cellular RNA with only 18 PCR cycles (**Fig 1G**). Final libraries are excised from agarose gels (300-600nt cDNA size), pooled, and are compatible with Illumina sequencing platforms. A computational pipeline was compiled into a batch script (**SData 1**) depicted by the flow-chart in **Figure 1H**.

### ***Validation with WA-1 Strain***

To test this approach, we obtained 200ng RNA from an SARS-CoV-2 isolate deposited at the World Reference Center for Emerging Viruses and Arboviruses (WRECEVA) at UTMB (26) and performed Tiled-ClickSeq using a 1:35 AzNTP:dNTP mix. NGS libraries were sequenced on an Illumina MiSeq (2x150 reads). Reads were quality processed using *fastp (31)* and mapped to the virus genome using *bowtie2 (33)*. A 'saw-tooth' pattern of read coverage over the genome was generated (**Fig 2A**, orange plot) with 'teeth' appearing as expected upstream of each tiled primer. Peaks of coverage for each 'tooth' ranged from ~13000x to ~100x. Overall, we obtained genome coverage >25X from nucleotide 3 to 29823 (50nts from the 3' end of the genome). This depth is sufficient to reconstruct a consensus genome sequence which was found to be identical to that already deposited (MT020881) for this isolate (43).

When using paired-end sequencing, the 'forward'/'R1' read is derived from the click-adaptor and contains the UMI. The 'reverse'/'R2' read is derived directly from the tiled primer (see schematic in **Figure 1E**). We wrote a custom python3 script to split all the forward 'R1' reads into multiple individual FASTQ files based upon which primer generated each fragment. The mapping coverage obtained from five individual tiled-primers is shown in **Figure 2B**. The coverage for each primer (denoted by individual colours in **Fig 2B**) spans approximately 500-

600 nts and extends 5'-wards from the tiled RT-primer. Read coverage from each primer overlaps the read coverage of the upstream primer. This allows for continuous gap-free read coverage over the viral genome which, importantly, allows a downstream cDNA amplicon to provide sequence information over and beyond an upstream primer. Additionally, we can determine the frequency with which each primer either successfully maps to the viral genome, mis-primes from the host RNA, or gives rise to adaptor-dimers or other sequencing artifacts. This information can be used to identify primers that yield poor viral priming efficiency and therefore a more specific primer can be designed and substituted as needed.

For nanopore sequencing, we also synthesized Tiled-ClickSeq libraries but using a 1:100 AzNTP:dNTP ratio to generate cDNA amplicons of increased lengths. We retained cDNA fragments >600nts, yielding a few nanograms of dsDNA. This library, though containing the Illumina adaptors, can nonetheless be used as input in the default Oxford Nanopore Technologies (ONT) Ligation-Sequencing protocol (LSK-109) that appends ONT adaptors directly onto the ends of A-tailed dsDNA fragments. We sequenced this library using an ONT MinION device and obtained 279,192 reads greater than 1kbp in length. These were mapped to the WA-1 viral genome using *minimap2* yielding continuous genome coverage (**Fig 2A**, blue). A similar profile of read coverage to the Illumina data was observed, with peaks of coverage upstream of tiled-primer sites. The deeper dips in coverage were avoided however, due to the longer reads lengths that give greater overlap between cDNA amplicons.

## ***Genome reconstruction of 12 Isolates: ClickSeq, Tiled-ClickSeq and Nanopore-Tiled-ClickSeq***

To validate the suitability of Tiled-ClickSeq for whole virus genome reconstruction, we obtained RNA extracted from 12 outgrowth samples of SARS-CoV-2 deposited at WRCEVA from nasopharyngeal swabs collected between March and April 2020. We synthesized 12 Tiled-ClickSeq libraries and 12 random-primer ClickSeq libraries in parallel. These were submitted for sequencing on a NextSeq (2x150) yielding ~2-5M reads per sample (**Table 1**). Random-primed ClickSeq data were quality-filtered and adaptor trimmed using *fastp (31)* retaining only the forward R1 reads. Tiled-ClickSeq read data were processed and mapped following the scheme in **Figure 1H**.

**Table 1: Read counts and mapping rates for random-primed versus Tiled-ClickSeq approaches**

| ***Sample*** | ***Outgrowth CT*** | ***ClickSeq Reads*** | ***Virus Mapped*** | **% Viral Reads** | ***Tiled v1 Reads*** | ***Virus Mapped*** | **% Viral Reads** |
|---|---|---|---|---|---|---|---|
| **WRCEVA_00501** | 12.9 | 4,665,869 | 116,036 | 2.5% | 2,359,795 | 2,204,750 | 93.4% |
| **WRCEVA_00502** | 12.9 | 4,989,513 | 118,260 | 2.4% | 1,962,581 | 1,820,925 | 92.8% |
| **WRCEVA_00505** | 12.7 | 3,894,325 | 71,809 | 1.8% | 2,779,672 | 2,482,854 | 89.3% |
| **WRCEVA_00506** | 12.5 | 4,979,989 | 108,532 | 2.2% | 2,395,750 | 2,148,256 | 89.7% |
| **WRCEVA_00507** | 12.9 | 5,659,073 | 161,059 | 2.8% | 2,056,670 | 1,867,012 | 90.8% |
| **WRCEVA_00508** | 16.8 | 3,987,009 | 91,452 | 2.3% | 1,787,418 | 1,433,005 | 80.2% |
| **WRCEVA_00509** | 17.1 | 4,057,928 | 57,424 | 1.4% | 2,202,661 | 1,856,633 | 84.3% |
| **WRCEVA_00510** | 16.2 | 5,328,829 | 65,281 | 1.2% | 2,040,332 | 1,601,544 | 78.5% |
| **WRCEVA_00513** | 16.0 | 4,391,175 | 69,169 | 1.6% | 1,641,213 | 1,455,991 | 88.7% |
| **WRCEVA_00514** | 12.9 | 4,340,084 | 84,211 | 1.9% | 2,089,241 | 1,902,748 | 91.1% |
| **WRCEVA_00515** | 15.7 | 5,416853 | 102,179 | 1.9% | 2,205,166 | 1,915,129 | 86.8% |
| **WRCEVA_00516** | 17.4 | 4,290,929 | 61,017 | 1.4% | 1,988,939 | 1,715,448 | 86.2% |

In the Tiled-ClickSeq data, after UMI deduplication, each isolate had an average coverage between 4,500-7,500 reads and a coverage of 25 reads in greater than 99.5% (29753/29903 nts) of the SARS-CoV-2 genome. Read coverage was also obtained covering the

5'UTR of each strain (>25 reads for all isolates from nucleotide 3 onwards (**Fig. 3A** and **B**)). When using paired-primer approaches, the 5'UTR is ordinarily obscured by the 5'-most primer used in each pool (nts 30-54 for the ARTIC primer set depicted in **Fig. 3A**). As the 5' end is resolved here due to stochastic incorporation of a single AzNTP in a template-specific manner, the entirety of the viral genome can be resolved. We reconstructed reference genomes from mapped reads using *pilon (34)* requiring 25x coverage for variant calling. In all cases, the reconstructed reference genomes were identical with or without controlingl for PCR duplicates using the UMIs. We found 5-12 SNVs per viral genome (**SData 2**), including the prevalent D614G (A23403G) spike adaptation, which enhances SARS-CoV-2 transmission (44), in 11 out of the 12 isolates (**Fig. 3C**).

Genome reconstructions was similarly performed using the random-primed ClickSeq data reads. Identical genomes to the Tiled data were obtained for 11 out of 12 isolates, with only one SNV difference in one sample (WRCEVA _000510: T168C). In this case, the read coverage was too low in the random-primed data for *pilon* to report an SNV. Nevertheless, visual inspection of the mapped data revealed that all nucleotides at this locus were indeed C's, as reported for the Tiled-ClickSeq data. Phylogenetic tree reconstruction using NextStrain (45) placed 10 of the isolates in the A2a clade (**Fig 3D**). Three of these isolates (WRCEVA_00506, WRCEVA_00510, WRCEVA_00515) were most closely related to European ancestors. Two isolates (WRCEVA_00508, WRCEVA_00513) were Clade B/B1 most closely related to Asian ancestors. Together, these data thus supported a model for multiple independent introductions of SARS-CoV-2 into the USA and subsequently into Galveston, Texas.

We also retained cDNA fragments >600bps from the Tiled-ClickSeq libraries and sequenced these using an ONT MinION device. We used the ONT native barcoding kit to multiplex the 12 samples and the Ligation-Sequencing protocol (LSK-109) to generate final libraries. Reads were mapped with *minimap2 (39)* yielding at least 100x coverage over >99.6% of the genome for each isolate (**SFig 1A,B)**. Again, reference genomes were reconstructed from the mapped data using *pilon* (**SData 3**). With the exception of WRCEVA_000514 which contained a single additional SNV (C14220T), the reference genomes reconstructed from the nanopore data were identical to those generated from the Tiled-ClickSeq Illumina data. These data illustrate that Tiled-ClickSeq performs as well as random-primed methods either on Illumina or Nanopore platforms for whole genome reconstruction.

## ***Minority Variants***

Our initial primer design (v1) (**Fig 4A**, blue plots) successfully yielded coverage suitable for complete genome reconstruction. However, some regions still received low coverage with fewer than a 100 deduplicated reads, preventing identification of minority variants in these regions. Therefore, we redesigned our primer scheme by adding an additional 326 primers (v2) previously reported (29) for tiled coronavirus sequencing (**SData 1**) to make a pool comprising a total of 396 unique primers (v3). We re-sequenced the 12 WRCEVA isolates analyzed as described above plus an additional four that subsequently became available. An example of mapping coverage for isolate WCREVA_000508 is illustrated in **Fig 4A**, where the coverage over the viral genome is more even with less extreme ranges of read depth.

Using the R2 read, we can determine which primer gives rise to each R1 read and trim primer-derived nucleotides from the R1 read. This is an important quality control as it prevents the assignment (or failure thereof) of SNVs and/or the mapping of recombination events due to primer mis-priming. If reads are mapped without trimming away the primer-derived nucleotides found in the R1 read (as depicted in **Fig 4B**), we see numerous high frequency (2-50%) minority variants. The majority of these apparent minority variants overlap primer-target sites and are likely artefactual. Furthermore, the same high-frequency events are often seen across multiple independent samples. To control for this, we map reads after trimming away primer-derived nucleotides from the R1 reads as per our pipeline described above (schematic in **Fig 1H**). Finally, to control for PCR duplication events, we make use of the UMIs embedded in the click-adaptor. The final de-duplicated mapped, primer-trimmed reads (**Fig 4C**) provide a robust readout of minority variants in these isolates (**Table 2** and **SData 4**). Across 10 WRCEVA isolates we found only 26 minority variants present at >2% all of which were unique within this dataset. Six isolates reported no minority variants at all.

***Table 2: Minority variants and rates (>2%) found across 16 WRCEVA isolates***

| | | | ***Read*** | | | | | ***Variant*** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Sample*** | ***Nt*** | ***Nuc*** | ***Depth*** | ***A*** | ***U*** | ***G*** | ***C*** | ***Rate*** | ***Location*** | ***Result*** |
| WRCEVA_000501 | 12049 | C | 2116 | 0 | **95** | 1 | *2020* | 4.5% | ORF1ab | N3928K |
| WRCEVA_000502 | 10207 | C | 2240 | 0 | **118** | 0 | *2122* | 5.3% | - | - |
| WRCEVA_000502 | 16050 | U | 3853 | 0 | *3322* | 0 | **531** | 13.8% | - | - |
| WRCEVA_000502 | 17489 | A | 4597 | *4433* | **162** | 1 | 1 | 3.6% | ORF1ab | E5742V |
| WRCEVA_000502 | 21526 | A | 8749 | *6508* | 0 | **2240** | 1 | 25.6% | ORF1ab | I7088V |
| WRCEVA_000503 | 14220 | C | 1638 | 1 | **463** | 0 | *1174* | 28.3% | - | - |
| WRCEVA_000504 | 1556 | A | 2828 | *2499* | 0 | **328** | 1 | 11.6% | ORF1ab | I431V |
| WRCEVA_000504 | 27925 | C | 2857 | 0 | **134** | 0 | *2723* | 4.7% | ORF8 | T11I |
| WRCEVA_000507 | 19515 | A | 2393 | *2295* | 1 | **97** | 0 | 4.1% | - | - |
| WRCEVA_000508 | 9756 | G | 1376 | **28** | 0 | *1348* | 0 | 2.1% | ORF1ab | R3164H |
| WRCEVA_000508 | 26056 | G | 2092 | 0 | **86** | *2006* | 0 | 4.1% | ORF3a | D222Y |
| WRCEVA_000508 | 27556 | G | 2066 | **128** | 0 | *1938* | 0 | 6.2% | ORF7a | A55T |
| WRCEVA_000509 | 11956 | C | 1962 | 0 | **199** | 0 | *1763* | 10.1% | - | - |
| WRCEVA_000509 | 17245 | C | 4062 | 2 | **470** | 0 | *3590* | 11.6% | ORF1ab | R5661C |
| WRCEVA_000509 | 18005 | U | 5408 | 1 | *4949* | **458** | 0 | 8.5% | ORF1ab | L5915R |
| WRCEVA_000509 | 25569 | U | 3448 | 4 | *3326* | **113** | 5 | 3.5% | - | - |
| WRCEVA_000509 | 27919 | U | 839 | 0 | *809* | 0 | **30** | 3.6% | ORF8 | I9T |
| WRCEVA_000509 | 28767 | C | 2011 | 0 | **109** | 0 | *1902* | 5.4% | N | T165I |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| WRCEVA_000511 | 3003 | U | 2880 | **79** | *2787* | 1 | 13 | 2.7% | ORF1ab | V913E |
| WRCEVA_000511 | 10738 | U | 4580 | 0 | *4440* | 0 | **140** | 3.1% | - | - |
| WRCEVA_000511 | 25892 | U | 133 | 0 | *130* | 0 | **3** | 2.3% | ORF3a | I167T |
| WRCEVA_000511 | 28001 | G | 1414 | 1 | **29** | *1384* | 0 | 2.1% | - | - |
| WRCEVA_000513 | 27046 | C | 5539 | 0 | **138** | 0 | *5401* | 2.5% | M | T175M |
| WRCEVA_000514 | 11603 | A | 5405 | *5075* | 0 | **330** | 0 | 6.1% | ORF1ab | M3780V |
| WRCEVA_000514 | 26526 | G | 525 | 0 | **20** | *505* | 0 | 3.8% | M | A2S |

## ***RNA Recombination: sgmRNAs, Structural variants and Defective RNAs***

To characterize RNA recombination, we used our bespoke *ViReMa* pipeline (35) to map RNA recombination events in NGS reads that correspond to either sgmRNAs, SVs or D-RNAs. *ViReMa* can detect agnostically a range of expected and unusual RNA recombination events including deletions, insertions, duplications, inversions as well as virus-to-host chimeric events and provides BED files containing the junction sites and frequencies of RNA recombination events. We mapped the Tiled-ClickSeq data to the corrected reference genome for each WRCEVA isolate using *ViReMa*. We also took total cellular RNA and RNA extracted from the supernatants of *Vero* cells transfected with RNA derived from an *in vitro* infectious clone of SARS-CoV-2 (icSARS-CoV-2) (27). These clone-derived RNAs contained either the WT SARS-CoV-2, or were engineered with a deletion near the furin cleavage site of the spike protein, which we recently demonstrated is a common adaption to Vero cells and which alters SARS-CoV-2 pathogenesis in mammalian models of infection (28).

The identities and frequencies of the 13 most abundant RNA recombination events are illustrated in **Figure 5A.** We found all the expected sgmRNAs previously annotated for SARS-CoV-2 (46) as well as non-canonical sgmRNAs. An overview of mapped data over the SARS-CoV-2 illustrating large recombination events (depicted by the blue horizontal lines) is provided in **SFig. 2**. We found that sgmRNAs were highly enriched in the cellular fractions from expressed icSARS-CoV-2 isolates (comprising >95% of the total viral genetic materials) but were relatively depleted in the supernatant fraction. This reflects a strong restriction of the packaging of these RNA species into virions. In the icSARS-CoV-2 samples, Tiled-ClickSeq and *ViReMa* accurately reported the expected deletion (Δ23603^23616). Interestingly, we also identified small structural variants (Δ23583^23599) in seven of the WRCEVA isolates with a frequency of 2-50%, similar to reports of the selection of variants containing deletions at this site after *in vitro* passaging on Vero cells (47). We also found a novel SV in one isolate (WRCEVA_000504: Δ27619^27642) present in 3.5% of the reads resulting in an 8 amino acid deletion in ORF7a. We additionally identified a small number of micro-indels (**Table 3**) in some isolates.

***Table 3: Micro-indels and rates (>2%) found across 16 WRCEVA isolates***

| ***Sample*** | ***MicroInDel*** | ***Nucs*** | ***Variant Rate*** | ***Location*** | ***Result*** |
|---|---|---|---|---|---|
| WRCEVA_000502 | Δ519^523 | UGGUU | 2.2% | ORF1AB | Frameshift |
| WRCEVA_000504 | Δ29686^29693 | CAGUGUGU | 3.5% | 3'UTR | - |
| WRCEVA_000505 | Δ519^523 | UGGUU | 2.9% | ORF1AB | Frameshift |
| WRCEVA_000506 | Δ519^523 | UGGUU | 3.8% | ORF1AB | Frameshift |
| WRCEVA_000509 | Δ1237^1239 | UCA | 2.9% | ORF1AB | ΔH325 |
| WRCEVA_000510 | Δ686^694 | AAGUCAUUU | 5.1% | ORF1ab | ΔLSF141-143 |
| WRCEVA_000511 | Δ519^523 | UGGUU | 3.7% | ORF1AB | Frameshift |
| WRCEVA_000511 | Δ10811^10813 | CUU | 3.1% | ORF1AB | ΔL3516 |
| WRCEVA_000512 | Δ29750^29759 | GAUCGAGUG | 10.0% | 3'UTR | - |

Finally, we observed thousands of RNA recombination events corresponding to D-RNAs (BED files for each sample are provided in **SData 5**). Despite their individual low frequencies, these events (displayed as a Recombination Heatmap in **Figure 5B**) reveals interesting features of D-RNAs of SARS-CoV-2. Apparent duplication events or insertions were most commonly observed with recombination events enriched around the 3'UTR of the genome, consistent with our previous characterization of RNA recombination in distinct coronavirus isolates including MHV, MERS and SARS-CoV-2 (48). Finally, large deletions comprising RNA recombination events stretching from nucleotides ~6000-7000 to the 3'UTR were also observed, again, consistent with our previous observations. Altogether, these results demonstrate RNA recombination is a common and conserved feature of SARS-CoV-2 and that the emergence of D-RNAs is prevalent source of genetic diversity amongst these isolates and is captured using Tiled-ClickSeq.

## ***Discussion***

Tiled-ClickSeq provides a simple method for whole genome sequencing of virus isolates such as SARS-CoV-2 that can simultaneously map SNVs, minority variants as well as recombination events. Importantly, having only a single template-targeted primer per amplicon provides the opportunity to sequence any RNA template regardless of what expected or unknown sequence is found upstream, including recombinant RNA molecules such as sgmRNAs and D-RNAs. The targeted approach requires a relatively small number of reads to be collected, allowing 10s of samples to be processed on a MiSeq platform or potentially 100s on a single flowcells of a NextSeq. Furthermore, the same library preps can be used as input in Oxford Nanopore Sequencing pipelines to yield longer reads, providing the convenience and portably inherent to the platform. We demonstrated that this method can reconstruct full length SARS-CoV-2 genomes in a manner equivalent to random-primed methods. Full length-genome sequencing is achieved, including the 5'UTR, missed in the bulk of current high-throughput sequencing efforts, removing the need for 5'RACE.

The design of Tiled-ClickSeq imparts built-in quality control tools, including UMIs in the click-adaptor and the opportunity to use paired-end sequencing to identify the primer that gives rise to each amplicon. In addition to controlling for aberrant SNVs, minority and structural variants, this information can be used to determine the relative sensitivity and specificity of each primer in the primer mix allowing the scheme to be pruned and optimized. Our final primer scheme contained over 390 unique SARS-CoV-2 primers. This purposely thorough design demonstrates how the Tiled-ClickSeq pipeline can accommodate complex mixtures of overlapping primers within the same RT reaction. This built-in redundancy reduces the chance of primer dropout due to the presence of SNVs, SVs or recombination events found in primerannealing sites. This feature is especially important considering the emergence of SARS-CoV-2 variants with deletions and mutations that disrupt sequencing efforts (49). Interestingly, we detected very few minority variants in in our samples present above 2%. This is consistent with other reports of minority variant detection in SARS-CoV-2 isolates and likely reflects the well-characterized activity of the coronavirus ExoN-nsp14 as a ‘proof-reader’ enzyme (50). As a result, the greatest source of genetic diversity in coronavirus isolates may well be due to RNA recombination.

On the nanopore sequencing platform, we could obtain sequence reads within the same day as RNA extraction. While the baseline accuracy rate of the nanopore platform prevents the reliable annotation of minority variants present at <5%, this platform can reconstruct novel SARS-CoV-2 variants as well as identify abundant sgmRNAs longer reads. Nanopore sequencing also allows for identification of long-range epistatically linked variants. Epistatic linkage can also be computationally leveraged to identify minority variants present at levels below the baseline error-rate of the sequencing platform, for example, using *CliqueSNV* (51) or *CoVaMa (52)*. Therefore, the nanopore platform in combination with Tiled-ClickSeq provides a robust pipeline for high-throughput SARS-CoV-2 variant detection with minimal infrastructure.

The ARTIC protocol contains a primer cognate to the 5’UTR of SARS-CoV-2 (nts 30-54) to capture and quantitate sub-genomic mRNAs (53). However, recombination events including non-canonical sgmRNAs will be missed by primer-pools that do not happen to flank RNA recombination junctions. In contrast, Tiled-ClickSeq is capable of ‘agnostically’ detecting any unanticipated RNA recombination including D-RNAs that can be characterized by RNA recombination events in or between any region of the viral genome, often in an unpredictable manner. As ClickSeq was originally designed to avoid artefactual recombination with fewer than

3 artefactual chimeric reads found per million reads, Tiled-ClickSeq provides a useful tool to identify D-RNAs and to robustly characterize rates of RNA recombination. Together, using the Tiled ClickSeq approach, we have the opportunity to identify rare and unexpected recombination events and are not biased by the limitation of primer-pair approaches. Coupled with its cross-sequencing platform capabilities, the work highlights the utility of Tiled-ClickSeq for analysis of SARS-CoV-2.

## ***Acknowledgements***

This work was funded by: a Technology Commercialization Program grant (UTMB) to EJ and ALR; a 'Data collection grant' and 'COVID-19 funding' from the Institute for Human Infections & Immunity (UTMB) to ALR; NIH grant R21AI151725 from to ALR; NIH Grant R01AI153602 and R00AG049042 to VDM; NIH grant R24 AI120942 "World Reference Center for Emerging Viruses and Arboviruses" to SW; and The Sealy and Smith Foundation for financial support to SW and UTMB. RML was supported by an NRSA Clinical and Translational Science (NCATS) TL-1 Training Core Award (TL1TR001440). We thank Steve Widen and Jill Thompson (UTMB) for NGS support.

## ***Conflict of Interest Statement***

E.J. and A.R. are co-founders and owners of 'ClickSeq Technologies LLC', a Texas-based Next-Generation Sequencing provider offering ClickSeq kits and services including the methods described in this manuscript. E.J. and A.R have a patent-pending on the method and use of single-primer tiled sequencing.

# Figure Legends

**Figure 1: Schematic of Tiled-ClickSeq and Computational Pipeline: A)** Schematic of SARS-CoV-2 genome with two examples of sub-genomic mRNAs. **B)** Paired-primer approaches typically generate short amplicons flanked by upstream and downstream primers that are PCR amplified in non-overlapping pools. **C)** Tiled-ClickSeq uses a single pool of primers at the reverse-transcription step with the upstream site generated by stochastic termination by azido-nucleotides. **D)** 3'-azido-blocked single-stranded cDNA fragments are 'click-ligated' using copper-catalyzed azide alkyne cycloaddition (CuAAC) to hexynyl functionalized Illumina i5 sequencing adaptors. Triazole-linked ssDNA is PCR amplified to generate a final cDNA library. **E)** The structure of the final cDNA is illustrated indicating the presence of the i5 and i7 adaptors, the 12N unique molecular identifier (UMI), the expected location of the triazole linkage, and the origins of the cDNA in the reads including the tiled primer-derived DNA, which is captured using paired-end sequencing. **F)** The hypothetical read coverage over a viral genome is indicated in red, yielding overlapping 'saw-tooth' patterns of sequencing coverage. Longer fragment lengths with more extensive overlapping can be obtained using decreased AzNTP:dNTP ratios. **G)** Final cDNA libraries are analyzed and size-selected by gel electrophoresis (2% agarose gel). Duplicates of libraries synthesized from 8, 80 and 800 ng of input SARS-CoV-2 RNA input are shown. **H)** Flowchart of the data processing and bioinformatic pipeline. Input data is in Blue, output data are in Green, scripts/processes are Purple.

**Figure 2: Read coverage over the SARS-CoV-2 genome using Tiled-ClickSeq: A)** Read coverage obtained from Tiled-ClickSeq over the whole viral genome is depicted when

sequencing using an Illumina MiSeq (orange) or on an Oxford Nanopore Technologies MinION device (blue). A ‘saw-tooth’ pattern of coverage is observed with ‘teeth’ upstream of tiled-primers, indicated at the bottom of the plot by short black lines. **B)** Zoomed in read coverage of nts 1-2400 of the SARS-CoV-2 genome with read coverage from five individual primers coloured to illustrate coverage from downstream amplicons overlapping the primer-binding sites of upstream tiled-primers.

**Figure 3: Genome Reconstruction of 12 SARS-CoV-2 isolates deposited at the World Reference Center for Emerging Viruses and Arboviruses (WRCEVA): A)** Read coverage is depicted over the 5’ UTR of the SARS-CoV-2 genome for each isolate revealing capture of this region. The 5’-most primer from the ARTICv3 protocol at nts-30-54 is illustrated. **B)** Snapshot of read data from Tiled-ClickSeq is depicted using the Tablet Sequencing Viewer from WRCEVA_000508 over the same region of the 5’UTR as **A)**. **C)** The most common single-nucleotide variants (SNVs) found in complete genome reconstructions from all 12 isolates are illustrated and colour-coded to depict the underlying viral protein. **D)** Phylogenetic tree of 12 WRCEVA isolates with their corresponding clade indicated.

**Figure 4: Additional tiled-primers improves read coverage and allows identification of minority variants: A)** Read coverage obtained from Tiled-ClickSeq over the whole viral genome is depicted using an Illumina MiSeq when using the original primers as in **Fig 2** (v1 - blue) or with an additional 326 tiled-primers (v3 - pink). Tiled-primers are indicated at the bottom of the plot by short blue (v1) or pink (v3) lines. **B)** The rates of mismatching nucleotides

found in mapped NGS reads is depicted across the SARS-CoV-2 genome for isolate WRECVA_000508 prior to trimming the tiled primers from forward/'R1' reads and without PCR deduplication. **C)** The rates of mismatching is also depicted after data quality processing to remove PCR duplicates and primer-derived nucleotides in the reads, revealing 3 minority variants in this sample with frequencies >2%.

**Figure 5: Tiled-ClickSeq identifies sub-genomic mRNAs, structural variants and Defective-RNAs: A)** A table of the most common RNA recombination events found using Tiled-ClickSeq in this study. The recombination junctions are indicated on the left of the table, with their relative frequencies indicated in the table and colour-matched for each sample analyzed. All canonical sgmRNAs are found with their open-reading frame (ORF) indicated, in addition to one non-canonical sgmRNAs (*). Three common structural variants including two deletions in spike protein and a deletion in ORF7a were also detected. **B)** Unique RNA recombination events are plotted for 16 WRCEVA isolates as a scatter plots whereby the upstream 'donor' site is plotted on the y-axis and a downstream 'acceptor' site is plotted on x-axis. The read count for each unique RNA recombination event is indicated by the size of the point, while the number of samples in which this each RNA recombination event is found is indicated by the color. Insertions/duplication/back-splicing events are found above the x=y axis, while deletions and RNA recombination events yielding sgmRNAs are found below.

# Supplementary Figure Legends

**Supplementary Figure 1: Read coverage of tiled nanopore data over 12 SARS-CoV-2 isolates: A)** Read coverage obtained from Tiled-ClickSeq over the whole viral genome for 12 WRCEVA isolates is depicted when using an Oxford Nanopore Technologies MinION device. Tiled-primers (v1) are indicated at the bottom of the plot by short blue lines. **B)** Read count mapping statistics for each isolate are shown in the table.

**Supplementary Figure 2: IGV snapshot of Tiled-ClickSeq data over icSARS-CoV-2 delta PRRA:** A full-view of the SARS-CoV-2 genome with mapped Tiled-ClickSeq reads is depicted using Integrative Genomics Viewer. Individual reads are illustrated with short grey lines. Recombination events mapped by ViReMa are illustrated by light-blue lines. Common variants engineered into the icSARS clone are indicated by vertical coloured striations.

**Supplementary Data 1**: Annotations and sequences of tiled-primers used in this manuscript are provided in BED format.

**Supplementary Data 2**: Batch scripts provided all computational tools and parameters used and python3 scripts used in this study are provided.

**Supplementary Data 3**: A summary of all Single-Nucleotide Variants (SNVs) detected for all samples sequenced in this study are provided. Each unique sample/isolate is listed, together with the SNVs relative to the WA-1 (NC_045512.2) strain in different NGS library preparation methods and sequencing platforms. The accession number for each reconstructed genome deposited in GenBank is also indicated.

**Supplementary Data 4**: The frequency of all mapped nucleotides at each genome coordinate for each WRCEVA isolate is provided. The reference genome, nucleotide coordinate and expected reference Nucleotide is provided. Total read coverage and the numbers of each non-reference nucleotide are also shown. Finally, the mismatch/error rate at each site is provided which reveals minority variants in each isolate.

**Supplementary Data 5**: BED files of RNA recombination events detected by ViReMa in the Tiled-ClickSeq data from each WRCEVA isolate.

# References

1. Grubaugh ND, Saraf S, Gangavarapu K, Watts A, Tan AL, Oidtman RJ, et al. Travel Surveillance and Genomics Uncover a Hidden Zika Outbreak during the Waning Epidemic. Cell. 2019 Aug 22;178(5):1057-71 e11.
2. Gussow AB, Auslander N, Faure G, Wolf YI, Zhang F, Koonin EV. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. Proc Natl Acad Sci U S A. 2020 Jun 30;117(26):15193-9.
3. Yi H. 2019 novel coronavirus is undergoing active recombination. Clin Infect Dis. 2020 Mar 4.
4. Su YC, Anderson DE, Young BE, Zhu F, Linster M, Kalimuddin S, et al. Discovery of a 382-nt deletion during the early evolution of SARS-CoV-2. bioRxiv. 2020:2020.03.11.987222.
5. Muth D, Corman VM, Roth H, Binger T, Dijkman R, Gottula LT, et al. Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. Scientific reports. 2018 Oct 11;8(1):15177.
6. Kemp S, Harvey W, Datir R, Collier D, Ferreira I, Carabelli A, et al. Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/V70. bioRxiv. 2020:2020.12.14.422555.
7. Ogando NS, Dalebout TJ, Zevenhoven-Dobbe JC, Limpens RW, van der Meer Y, Caly L, et al. SARS-coronavirus-2 replication in Vero E6 cells: replication kinetics, rapid adaptation and cytopathology. bioRxiv. 2020:2020.04.20.049924.
8. Makino S, Taguchi F, Fujiwara K. Defective interfering particles of mouse hepatitis virus. Virology. 1984 Feb;133(1):9-17.
9. Makino S, Fujioka N, Fujiwara K. Structure of the intracellular defective viral RNAs of defective interfering particles of mouse hepatitis virus. J Virol. 1985 May;54(2):329-36.
10. Makino S, Shieh CK, Keck JG, Lai MM. Defective-interfering particles of murine coronavirus: mechanism of synthesis of defective viral RNAs. Virology. 1988 Mar;163(1):104-11.
11. Makino S, Shieh CK, Soe LH, Baker SC, Lai MM. Primary structure and translation of a defective interfering RNA of murine coronavirus. Virology. 1988 Oct;166(2):550-60.
12. Chang RY, Hofmann MA, Sethna PB, Brian DA. A cis-acting function for the coronavirus leader in defective interfering RNA replication. J Virol. 1994 Dec;68(12):8223-31.
13. Penzes Z, Tibbles KW, Shaw K, Britton P, Brown TD, Cavanagh D. Generation of a defective RNA of avian coronavirus infectious bronchitis virus (IBV). Defective RNA of coronavirus IBV. Adv Exp Med Biol. 1995;380:563-9.
14. Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Hölzer M, et al. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome Res. 2019 09;29(9):1545-54.
15. Banerjee S, Repass JF, Makino S. Enhanced accumulation of coronavirus defective interfering RNA from expressed negative-strand transcripts by coexpressed positive-strand RNA transcripts. Virology. 2001 Sep 1;287(2):286-300.
16. Joo M, Banerjee S, Makino S. Replication of murine coronavirus defective interfering RNA from negative-strand transcripts. J Virol. 1996 Sep;70(9):5769-76.
17. Kim YN, Lai MM, Makino S. Generation and selection of coronavirus defective interfering RNA with large open reading frame by RNA recombination and possible editing. Virology. 1993 May;194(1):244-53.
18. Gribble J, Pruijssers AJ, Agostini ML, Anderson-Daniels J, Chappell JD, Lu X, et al. The coronavirus proofreading exoribonuclease mediates extensive viral recombination. bioRxiv. 2020:2020.04.23.057786.

19. Vignuzzi M, Lopez CB. Defective viral genomes are key drivers of the virus-host interaction. Nat Microbiol. 2019 Jun 3.

20. Tyson JR, James P, Stoddart D, Sparks N, Wickenhagen A, Hall G, et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. bioRxiv. 2020 Sep 4.

21. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. 2019 Jan 8;20(1):8.

22. Routh A, Head SR, Ordoukhanian P, Johnson JE. ClickSeq: Fragmentation-Free Next-Generation Sequencing via Click Ligation of Adaptors to Stochastically Terminated 3'-Azido cDNAs. J Mol Biol. 2015 Jun 24.

23. Jaworski E, Routh A. ClickSeq: Replacing Fragmentation and Enzymatic Ligation with Click-Chemistry to Prevent Sequence Chimeras. Methods Mol Biol. 2018;1712:71-85.

24. Gorzer I, Guelly C, Trajanoski S, Puchhammer-Stockl E. The impact of PCR-generated recombination on diversity estimation of mixed viral populations by deep sequencing. J Virol Methods. 2010 Oct;169(1):248-52.

25. Itokawa K, Sekizuka T, Hashino M, Tanaka R, Kuroda M. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. PloS one. 2020;15(9):e0239403.

26. Harcourt J, Tamin A, Lu X, Kamili S, Sakthivel SK, Murray J, et al. Severe Acute Respiratory Syndrome Coronavirus 2 from Patient with Coronavirus Disease, United States. Emerg Infect Dis. 2020 Jun;26(6):1266-73.

27. Xie X, Muruato A, Lokugamage KG, Narayanan K, Zhang X, Zou J, et al. An Infectious cDNA Clone of SARS-CoV-2. Cell Host Microbe. 2020 May 13;27(5):841-8 e3.

28. Johnson BA, Xie X, Bailey AL, Kalveram B, Lokugamage KG, Muruato A, et al. Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. Nature. 2021 Jan 25.

29. Guo L, Boocock J, Tome JM, Chandrasekaran S, Hilt EE, Zhang Y, et al. Rapid cost-effective viral genome sequencing by V-seq. bioRxiv. 2020:2020.08.15.252510.

30. Routh A, Head SR, Ordoukhanian P, Johnson JE. ClickSeq: Fragmentation-Free Next-Generation Sequencing via Click Ligation of Adaptors to Stochastically Terminated 3'-Azido cDNAs. J Mol Biol. 2015 Aug 14;427(16):2610-6.

31. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i90.

32. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17:10-2.

33. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012 Mar 04;9(4):357-9.

34. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one. 2014;9(11):e112963.

35. Routh A, Johnson JE. Discovery of functional genomic motifs in viruses with ViReMa-a Virus Recombination Mapper-for analysis of next-generation sequencing data. Nucleic Acids Res. 2014 Jan;42(2):e11.

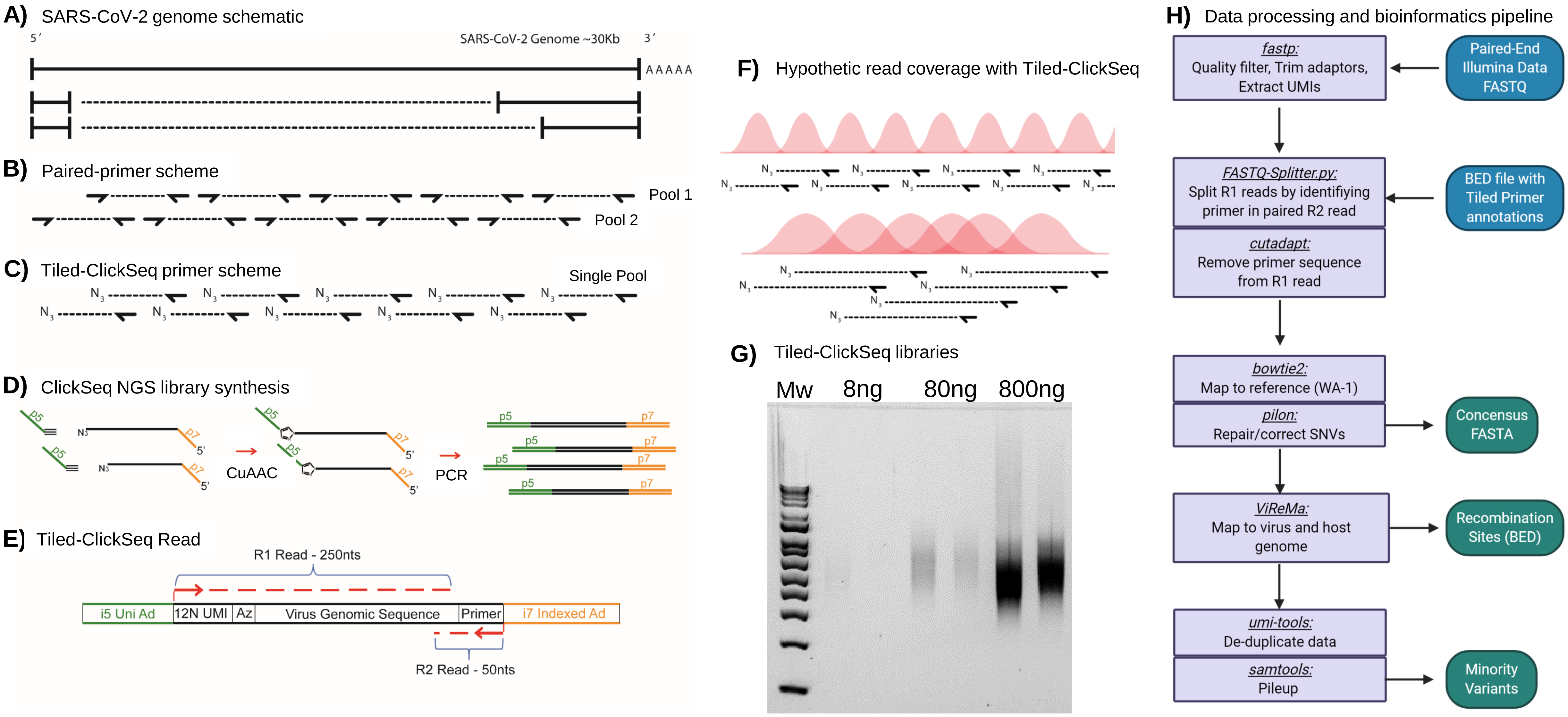
36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

37. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017 Mar;27(3):491-9.

38. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, et al. Tablet--next generation sequence assembly visualization. Bioinformatics. 2010 Feb 1;26(3):401-2.

39. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016 Jul 15;32(14):2103-10.
40. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014 Sep 8;47:11 2 1-34.
41. Routh A, Ji P, Jaworski E, Xia Z, Li W, Wagner EJ. Poly(A)-ClickSeq: click-chemistry for next-generation 3-end sequencing without RNA enrichment or fragmentation. Nucleic Acids Res. 2017 Jul 7;45(12):e112.
42. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci U S A. 2011 Dec 13;108(50):20166-71.
43. Harcourt J, Tamin A, Lu X, Kamili S, Sakthivel SK, Murray J, et al. Isolation and characterization of SARS-CoV-2 from the first US COVID-19 patient. bioRxiv. 2020 Mar 7.
44. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness and neutralization susceptibility. bioRxiv. 2020 Sep 2.
45. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018 Dec 1;34(23):4121-3.
46. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell. 2020 Apr 18.
47. Klimstra WB, Tilston-Lunel NL, Nambulli S, Boslett J, McMillen CM, Gilliland T, et al. SARS-CoV-2 growth, furin-cleavage-site adaptation and neutralization using serum from acutely infected, hospitalized COVID-19 patients. bioRxiv. 2020:2020.06.19.154930.
48. Gribble J, Stevens LJ, Agostini ML, Anderson-Daniels J, Chappell JD, Lu X, et al. The coronavirus proofreading exoribonuclease mediates extensive viral recombination. PLoS pathogens. 2021 Jan;17(1):e1009226.
49. Plante JA, Mitchell BM, Plante KS, Debbink K, Weaver SC, Menachery VD. The Variant Gambit: COVID's Next Move. Cell Host & Microbe. 2021 2021/03/01/.
50. Smith EC, Denison MR. Coronaviruses as DNA wannabes: a new model for the regulation of RNA virus replication fidelity. PLoS pathogens. 2013;9(12):e1003760.
51. Knyazev S, Tsyvina V, Shankar A, Melnyk A, Artyomenko A, Malygina T, et al. CliqueSNV: An Efficient Noise Reduction Technique for Accurate Assembly of Viral Variants from NGS Data. bioRxiv. 2020:264242.
52. Routh A, Chang MW, Okulicz JF, Johnson JE, Torbett BE. CoVaMa: Co-Variation Mapper for disequilibrium analysis of mutant loci in viral populations using next-generation sequence data. Methods. 2015 Sep 25.
53. Parker MD, Lindsey BB, Leary S, Gaudieri S, Chopra A, Wyles M, et al. periscope: sub-genomic RNA identification in SARS-CoV-2 Genomic Sequencing Data. bioRxiv. 2020:2020.07.01.181867.

# Figure 1 – Schematic of Tiled-ClickSeq and Computational Pipeline



A) SARS-CoV-2 genome schematic

B) Paired-primer scheme

C) Tiled-ClickSeq primer scheme

D) ClickSeq NGS library synthesis

E) Tiled-ClickSeq Read

F) Hypothetic read coverage with Tiled-ClickSeq

G) Tiled-ClickSeq libraries

H) Data processing and bioinformatics pipeline

Paired-End Illumina Data FASTQ

*fastp:* Quality filter, Trim adaptors, Extract UMIs

BED file with Tiled Primer annotations

*FASTQ-Splitter.py:* Split R1 reads by identifiying primer in paired R2 read

*cutadapt:* Remove primer sequence from R1 read

*bowtie2:* Map to reference (WA-1)

*pilon:* Repair/correct SNVs

Concensus FASTA

*ViReMa:* Map to virus and host genome

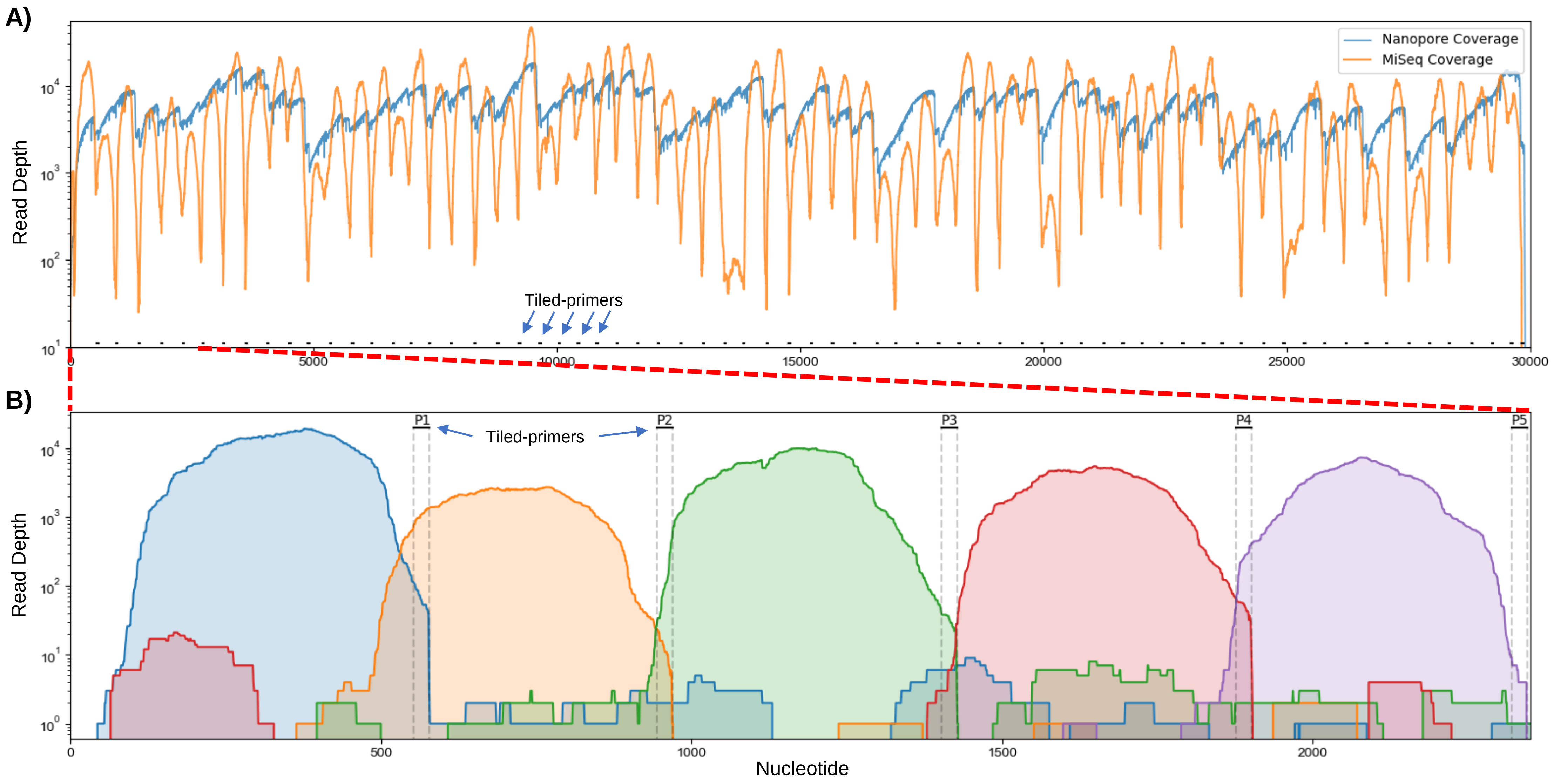
Recombination Sites (BED)

*umi-tools:* De-duplicate data

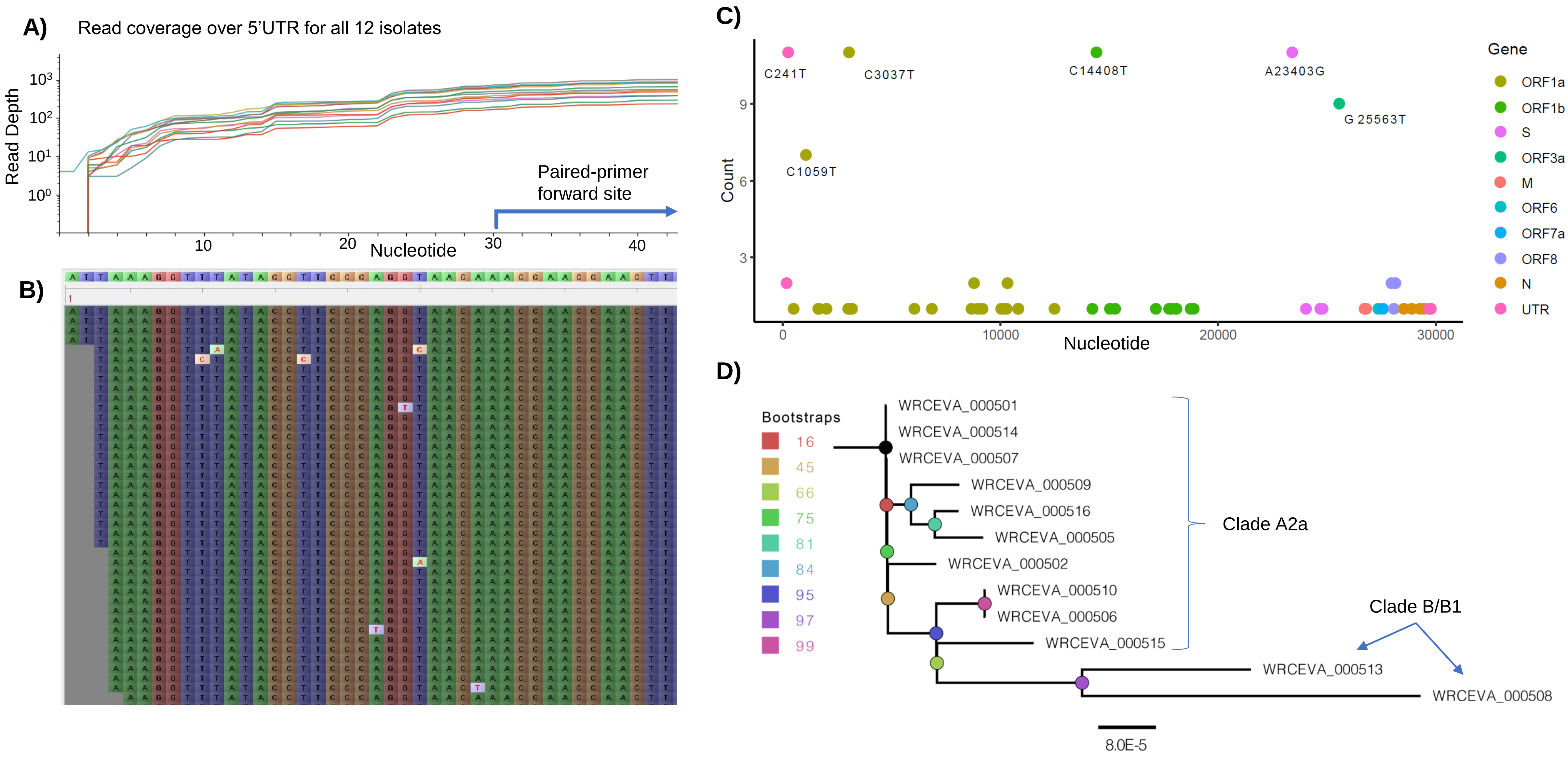
*samtools:* Pileup

Minority Variants

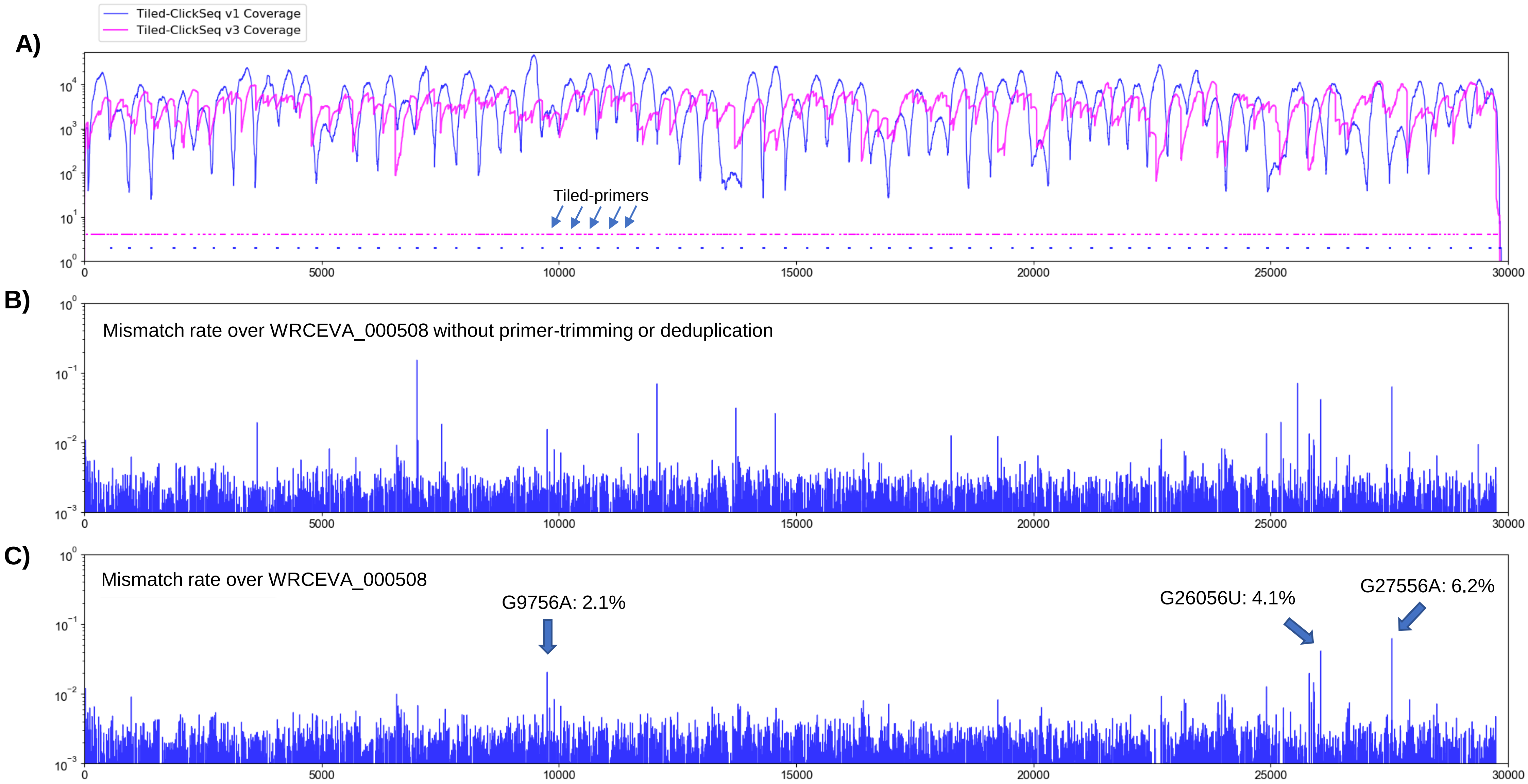
# Figure 2 – Read Coverage over genome with Illumina and Nanopore Sequencing



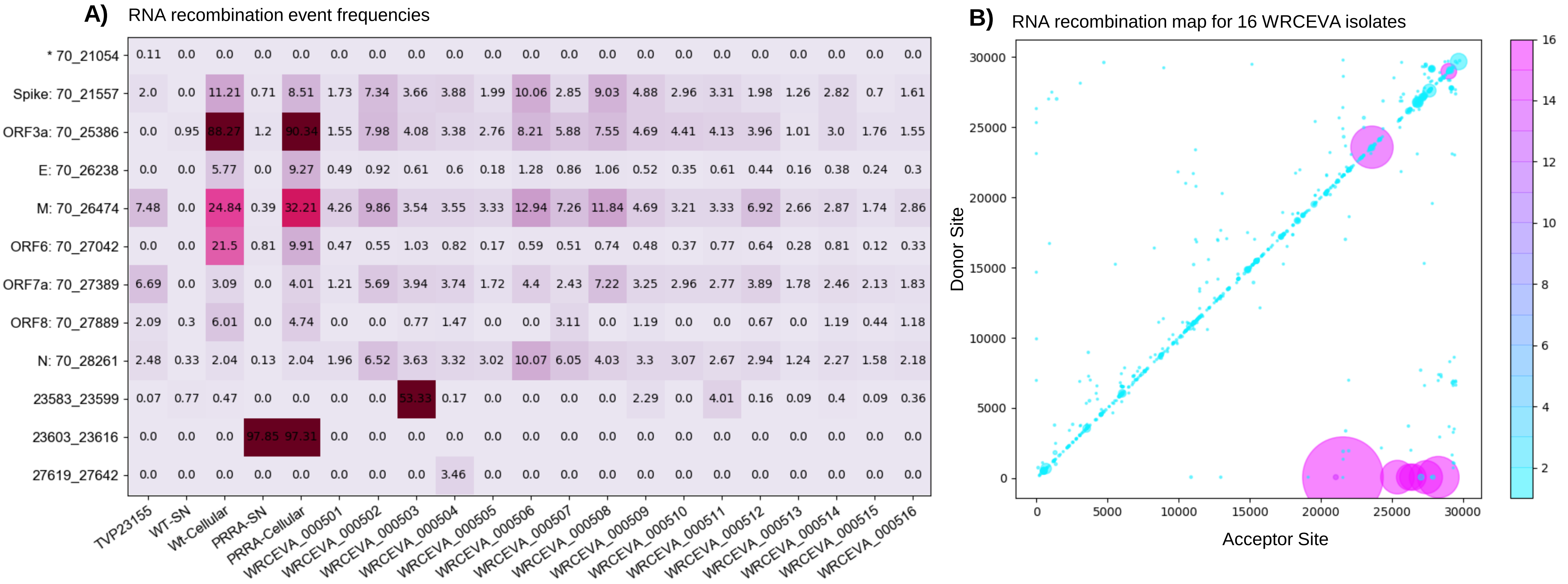
# Figure 3 – Genome Reconstruction with 12 isolates



# Figure 4 – Improved coverage with extra tiled primers



# Figure 5 – Identification of sgmRNAs and structural variants

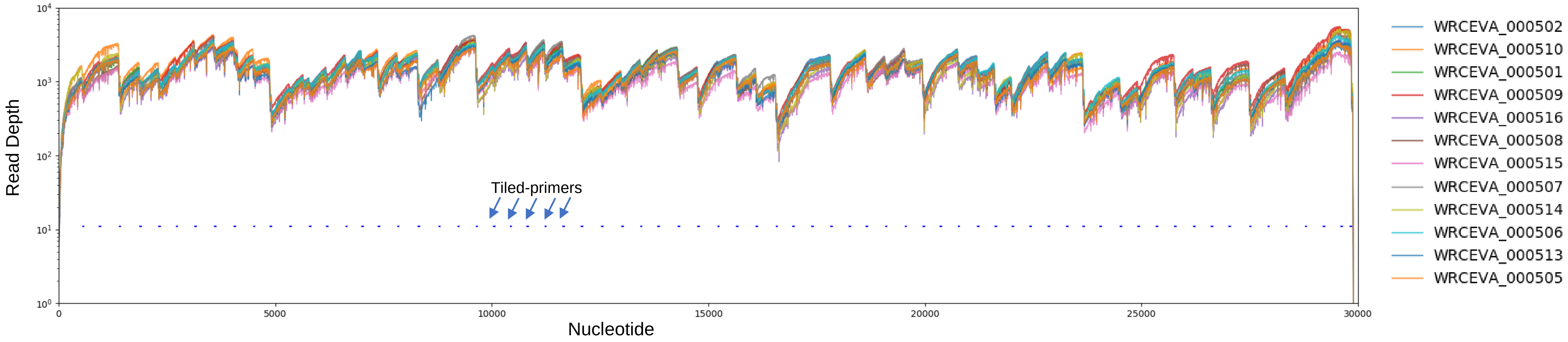


**A)** RNA recombination event frequencies

| | TVP23155 | WT-SN | Wt-Cellular | PRRA-SN | PRRA-Cellular | WRCEVA_000501 | WRCEVA_000502 | WRCEVA_000503 | WRCEVA_000504 | WRCEVA_000505 | WRCEVA_000506 | WRCEVA_000507 | WRCEVA_000508 | WRCEVA_000509 | WRCEVA_000510 | WRCEVA_000511 | WRCEVA_000512 | WRCEVA_000513 | WRCEVA_000514 | WRCEVA_000515 | WRCEVA_000516 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| * 70_21054 | 0.11 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Spike: 70_21557 | 2.0 | 0.0 | 11.21 | 0.71 | 8.51 | 1.73 | 7.34 | 3.66 | 3.88 | 1.99 | 10.06 | 2.85 | 9.03 | 4.88 | 2.96 | 3.31 | 1.98 | 1.26 | 2.82 | 0.7 | 1.61 |
| ORF3a: 70_25386 | 0.0 | 0.95 | 88.27 | 1.2 | 90.34 | 1.55 | 7.98 | 4.08 | 3.38 | 2.76 | 8.21 | 5.88 | 7.55 | 4.69 | 4.41 | 4.13 | 3.96 | 1.01 | 3.0 | 1.76 | 1.55 |
| E: 70_26238 | 0.0 | 0.0 | 5.77 | 0.0 | 9.27 | 0.49 | 0.92 | 0.61 | 0.6 | 0.18 | 1.28 | 0.86 | 1.06 | 0.52 | 0.35 | 0.61 | 0.44 | 0.16 | 0.38 | 0.24 | 0.3 |
| M: 70_26474 | 7.48 | 0.0 | 24.84 | 0.39 | 32.21 | 4.26 | 9.86 | 3.54 | 3.55 | 3.33 | 12.94 | 7.26 | 11.84 | 4.69 | 3.21 | 3.33 | 6.92 | 2.66 | 2.87 | 1.74 | 2.86 |
| ORF6: 70_27042 | 0.0 | 0.0 | 21.5 | 0.81 | 9.91 | 0.47 | 0.55 | 1.03 | 0.82 | 0.17 | 0.59 | 0.51 | 0.74 | 0.48 | 0.37 | 0.77 | 0.64 | 0.28 | 0.81 | 0.12 | 0.33 |
| ORF7a: 70_27389 | 6.69 | 0.0 | 3.09 | 0.0 | 4.01 | 1.21 | 5.69 | 3.94 | 3.74 | 1.72 | 4.4 | 2.43 | 7.22 | 3.25 | 2.96 | 2.77 | 3.89 | 1.78 | 2.46 | 2.13 | 1.83 |
| ORF8: 70_27889 | 2.09 | 0.3 | 6.01 | 0.0 | 4.74 | 0.0 | 0.0 | 0.77 | 1.47 | 0.0 | 0.0 | 3.11 | 0.0 | 1.19 | 0.0 | 0.0 | 0.67 | 0.0 | 1.19 | 0.44 | 1.18 |
| N: 70_28261 | 2.48 | 0.33 | 2.04 | 0.13 | 2.04 | 1.96 | 6.52 | 3.63 | 3.32 | 3.02 | 10.07 | 6.05 | 4.03 | 3.3 | 3.07 | 2.67 | 2.94 | 1.24 | 2.27 | 1.58 | 2.18 |
| 23583_23599 | 0.07 | 0.77 | 0.47 | 0.0 | 0.0 | 0.0 | 0.0 | 53.33 | 0.17 | 0.0 | 0.0 | 0.0 | 0.0 | 2.29 | 0.0 | 4.01 | 0.16 | 0.09 | 0.4 | 0.09 | 0.36 |
| 23603_23616 | 0.0 | 0.0 | 0.0 | 97.85 | 97.31 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 27619_27642 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.46 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

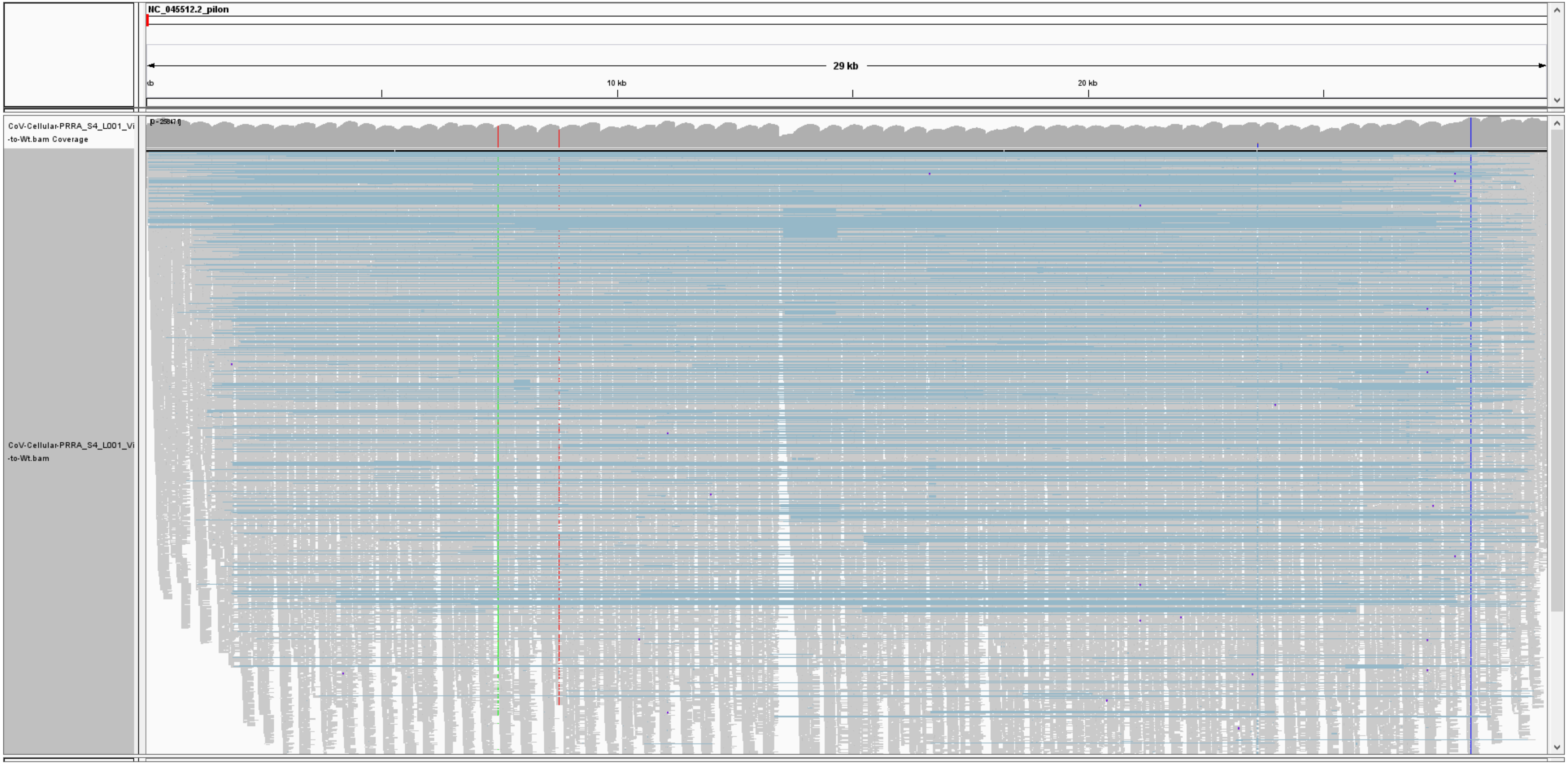
**B)** RNA recombination map for 16 WRCEVA isolates

# Sfig 1 – Read coverage of tiled nanopore data over 12 SARS-CoV-2 isolates



| | | Virus Mapped | Nts > 100 Read | % Nts > 100 | Unmapped |
|---|---|---|---|---|---|
| Isolate Name | Total Reads | Reads | Coverage | Read Coverage | Reads |
| WRCEVA_000502 | 68623 | 65954 | 29806 | 99.68% | 2669 |
| WRCEVA_000510 | 95468 | 79421 | 29814 | 99.70% | 16047 |
| WRCEVA_000501 | 70508 | 64145 | 29802 | 99.66% | 6363 |
| WRCEVA_000509 | 77441 | 71092 | 29779 | 99.59% | 6349 |
| WRCEVA_000516 | 68623 | 60795 | 29798 | 99.65% | 7828 |
| WRCEVA_000508 | 76040 | 63468 | 29815 | 99.71% | 12572 |
| WRCEVA_000515 | 64917 | 48641 | 29803 | 99.67% | 16276 |
| WRCEVA_000507 | 76159 | 72440 | 29809 | 99.69% | 3719 |
| WRCEVA_000514 | 84382 | 75143 | 29814 | 99.70% | 9239 |
| WRCEVA_000506 | 72199 | 67867 | 29811 | 99.69% | 4332 |
| WRCEVA_000513 | 67028 | 57866 | 29799 | 99.65% | 9162 |
| WRCEVA_000505 | 61718 | 58723 | 29805 | 99.67% | 2995 |

# Sfig 2 – IGV snapshot of Tiled-ClickSeq data over icSARS-CoV-2 delta PRRA


NC_045512.2_pilon
29 kb
kb
10 kb
20 kb
CoV-Cellular-PRRA_S4_L001_Vi
-to-Wt.bam Coverage
[0 - 25847]
CoV-Cellular-PRRA_S4_L001_Vi
-to-Wt.bam