`Magnipore`: Prediction of differential single nucleotide changes in the Oxford Nanopore Technologies sequencing signal of SARS-CoV-2 samples

Jannes Spangenberg[1], Christian Höner zu Siederdissen[1], Milena Žarković[1], Sandra Triebel[1], Ruben Rose[2], Christina Martínez Christophersen[3], Lea Paltzow[3], Mohsen M. Hegab[3], Anna Wansorra[3], Akash Srivastava[1], Andi Krumbholz[2,3,*], and Manja Marz[1,4,5,*]

[1]RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, Leutragraben 1, 07743 Jena, Germany
[2]Institute for Infection Medicine, Christian-Albrechts-Universität zu Kiel and University Medical Center Schleswig-Holstein, Campus Kiel, Brunswiker Straße 4, 24105 Kiel, Germany
[3]Labor Dr. Krause und Kollegen MVZ GmbH, Steenbeker Weg 23, 24106 Kiel, Germany
[4]European Virus Bioinformatics Center 2, Leutragraben 1, 07743 Jena, Germany
[5]FLI Leibniz Institute for Age Research, Beutenbergstraße 11, 07745 Jena, Germany
[*]both authors contributed equally

March 17, 2023

Abstract

Oxford Nanopore Technologies (ONT) allows direct sequencing of ribonucleic acids (RNA) and, in addition, detection of possible RNA modifications due to deviations from the expected ONT signal. The software available so far for this purpose can only detect a small number of modifications. Alternatively, two samples can be compared for different RNA modifications. We present `Magnipore`, a novel tool to search for significant signal shifts between samples of Oxford Nanopore data from similar or related species. `Magnipore` classifies them into mutations and potential modifications. We use `Magnipore` to compare SARS-CoV-2 samples. Included were representatives of the early 2020s Pango lineages (n=6), samples from Pango lineages B.1.1.7 (n=2, Alpha), B.1.617.2 (n=1, Delta), and B.1.529 (n=7, Omicron). `Magnipore` utilizes position-wise Gaussian distribution models and a comprehensible significance threshold to find differential signals. In the case of Alpha and Delta, `Magnipore` identifies 55 detected mutations and 15 sites that hint at differential modifications. We predicted potential virus-variant and variant-group-specific differential modifications. `Magnipore` contributes to advancing RNA modification analysis in the context of viruses and virus variants.

Key words: Oxford Nanopore Technologies, raw ONT sequencing signal, differential RNA modifications, SARS-CoV-2, comparative analysis

Introduction

Oxford Nanopore Technologies (ONT) provides the possibility to sequence desoxyribonucleic acid (DNA) and ribonucleic acid (RNA) directly without any amplification, which would erase nucleotide modifications. ONT uses flow cells containing nanopores with sequencing sensors. Nucleotides are pulled through nanopores. The pore is integrated into a membrane to which a voltage is applied. The measured electrical current within the nanopore is characteristic of the nucleotides within the pore. Five nucleotides are measured at a time in the most narrow part of the pore. The change in the electric current enables the basecalling procedure, which can be identified by deep learning or statistical models [1, 2, 3, 4].
Severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2) is a single positive-stranded RNA virus with a genome size of about 30kb and a 3' poly-A tail [5, 6]. Since the end of 2019, various SARS-CoV-2 variants have emerged that differ from the original Wuhan variant by defined genomic mutations and resulting amino acid substitutions, particularly in the spike protein [7]. Bioinformatics pipelines like poreCov [8] and others

[9, 10] contribute significantly to learning more about the virus and containing the COVID-19 pandemic by identifying and monitoring emerging variants. They also provide insights into potential RNA secondary structures [11, 12].

In the mammalian cell, RNA modifications influence, among other things, the formation of secondary and tertiary structures of transfer RNAs and the translation, stability, and localization of messenger RNAs (mRNAs) [13]. Previous studies have shown that positive single-stranded RNA viruses like *Picornaviruses* or *Flaviviruses* contain RNA modifications that play a significant role in viral infection and replication and even influence the host antiviral innate immunity [14, 15, 16, 17]. The replication of various coronaviruses also appears to be affected by RNA modifications [18]. The growing understanding of the importance of such modifications is accompanied by increased activities to detect them by direct RNA sequencing (DRS) [19, 20, 21, 22]. In ONT sequencing, modifications, as well as mutations, lead to shifts of the ONT signal [23, 24] and changes in the translocation speed [24].

Several machine-learning approaches have been developed to detect specific modifications. For N^6-Methyladenosine (m6A), methods employed include support-vector machines as in EpiNano [21], neural networks like m6Anet [19], as well as random forests in MINES [22]. However, due to the lack of training data, the individual ONT signals cannot be reliably assigned to one of the many modifications. So far, there are only approaches that predict a modification at a nucleotide position with a certain probability.

Another strategy is a generic detection of modification sites without assignment to a specific modification by comparing two ONT samples. Some tools, e.g., DRUMMER [25], or ELIGOS [26], focus on different patterns of sequencing errors after basecalling, introduced by RNA modifications.

Other tools, such as Nanocompore [20], Yanocomp [27], or xPore [28], analyze the raw ONT signal directly. These tools compare the raw ONT data of two samples from a control and a wild-type condition to find differentiating signals [20, 27, 28]. The control sample is thereby depleted of modifications, and the wild type should carry modifications. All three tools are limited to comparing samples of the same species and classifying differentiating signals as modifications, neglecting mutations.

Here, we present Magnipore, a Python 3 pipeline for comparative ONT data analysis on signal level, Fig. 1. Magnipore takes two reference FASTA files to compare two distantly related genomes with alignable homologous regions. The program creates signal distribution models from the raw signal in the FAST5 files. These distributions are compared between alignable regions. It predicts global differential signals between two samples. Magnipore classifies those into mutations and modifications using reference sequences detecting epitranscriptomic and transcriptomic differences. For signal segmentation and resquiggling, Magnipore uses nanopolish [29], which requires basecalled reads. The basecalls can be produced using Guppy within the pipeline, or the user can provide them. We map the basecalled reads with minimap2 [30] to segment the raw ONT signals corresponding to the bases in the reference sequence using nanopolish eventalign. For each base in the reference, we calculate a Gaussian signal distribution from all reads mapped to this base.

As a study case, we aim to identify variant-specific mutations and modifications of SARS-CoV-2 using Magnipore.

Materials and Methods

Preparation of total RNA

Nucleic acids from nasopharyngeal swabs were analyzed for the presence of SARS-CoV-2 genome equivalents using a triplex real-time RT-PCR [31]. If the sample tested positive for SARS-CoV-2, the underlying strain was determined by whole genome sequencing. For this purpose, the NEBNext® ARTIC SARS-CoV-2 Companion Kit (New England Biolabs, Ipswich, MA, USA) and a MinION system (Oxford Nanopore Technologies plc., Oxford Science Park, UK) were used according to the manufacturer's instructions. The principle of this procedure is to reverse transcribe the viral RNA into complementary desoxyribonucleic acid (cDNA) and amplify it with a mixture of primers so that the overlapping PCR fragments cover the entire viral genome. These are purified, barcoded, and sequenced. The primers are regularly adapted to newly emerging variants. Data were analyzed using the poreCov workflow and assigned to a Pango lineage [8, 32]. Aliquots of SARS-CoV-2-containing nasopharyngeal swab specimens in phosphate-buffered saline were then shaken in cell culture medium and placed in 48-well plates seeded with Vero cells, which were then incubated under standard conditions at 37°C for several days. At least one to two more passages on Vero cells followed after that. Then an aliquot of the supernatant was tested for the presence of SARS-CoV-2 RNA [31]. If SARS-CoV-2 genome equivalences were detectable, Vero cells seeded in T25 flasks were inoculated. After several days RNA was prepared using the RNeasy kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. All virus cultivation and RNA extraction steps were performed under biosafety level-3 conditions. The resulting SARS-CoV-2 isolates were also sequenced according to the ARTIC protocol and used for neutralization experiments or other purposes [32]. The sequencing of the original samples was carried out for epidemiological reasons or according to the requirements of the Coronavirus Surveillance Ordinance (CorSurV) of the Federal Republic of Germany. The provision and sequencing of viral RNA in Jena are covered by a positive vote of the Ethics Committee of the Medical Faculty of Kiel University (D467/20).

Direct RNA sequencing

The library was constructed using ONT protocol with the direct RNA sequencing and SQK-RNA002 kit. The manufacturer's instructions were followed with minor changes listed to achieve the suggested recovery aim of about 200 ng of the final library and to obtain longer reads: 1. Before library preparation, the sample was cleaned, concentrated, and size selected for longer reads with RNAClean XP beads, with a ratio of up to 4:3 (sample:beads); 2. The starting sample amount was increased from 1 µg to 4 µg RNA; 3. The amount of the RNA Control Strand (RCS) was used for samples 5 and 9 (see Tab. 1 as dilution 1:10, and for other samples not added at all; 4. As we noticed a lot of free adapters, RNA Adapter Mix (RMX), we reduced the amount from 6 µl to 3 µl; 5. For the Omicron BE.1 and BF.1 samples, incubation times were prolonged so that adapters had more time to ligate. The reaction with the RT adapter (RTA) and RMX time was prolonged to 15 min and 20 min, respectively. The library was loaded onto R9.4.1 flow cell and sequenced on a MinION Mk1B device.

`Magnipore` requirements

We recommend using a `conda` environment for `Magnipore`. Required packages and tools except `Guppy` can easily be installed with `conda` via `https://anaconda.org/JannesSP/magnipore` or with the `YAML` file in the `GitHub` repository `https://github.com/JannesSP/magnipore`. A detailed list of the packages, tools, and their versions can be found in Tab. 2.

`Magnipore` input

`Magnipore` requires two multi-`FAST5` files and two reference `FASTA` files as input for both samples (Fig. 1). In case the samples originate from the same biological sample for two different environmental conditions, the same reference file must be provided for both sample inputs. `Magnipore` requires basecalls for `nanopolish's` signal segmentation. The user provides either those basecalls with `--path_to_first_basecalls PATH` and `--path_to_sec_basecalls PATH` or `Magnipore` needs the `Guppy` binary and model path with `--guppy_bin PATH` and `--guppy_model PATH` to basecall the `FAST5` data.

Basecalling with `Guppy`

`Guppy v6.1.7` is available for ONT customers and can be obtained via their community site `https://community.nanoporetech.com`. `Magnipore` uses `Guppy` with the parameters `--disable_qscore_filtering --calib_detect -c guppy_model` if no `FASTQ` file is provided.

Mapping with `minimap2`

In the next step, `Minimap2 v2.24` [30], using the parameters `-a -x splice -k 14` maps the reads to the provided reference to run `nanopolish eventalign v0.14.0` [29]. For that we executed `nanopolish index` with the parameters to execute `nanopolish eventalign` with `--summary=summary.csv --scale-events --signal-index -t threads`.

Resquiggling with `nanopolish`

`Magnipore` needs the ONT signal segment assignment for each base to calculate a normalized signal distribution for each position. `Nanopolish eventalign` resquiggles and segments the raw ONT signal from the `FAST5` files using the mapping from `minimap2` [29]. `Nanopolish eventalign` can only process those reads that could be mapped to the reference sequence. Resquiggling the reads means that `nanopolish eventalign` corrects the bases in the basecalled reads according to the given reference sequence, integrated k-mer models from ONT, and the raw ONT signal. `Nanopolish eventalign` can not distinguish between basecalling errors or real mutations. It assumes that the reference sequence is the ground truth and will correct every base from the reads using the corresponding region in the reference sequence. Additionally to the resquiggling, `nanopolish eventalign` assigns signal segments to each reference position.

Normalization

The pico Ampere (pA) signal for each read is normalized by the subtraction of the median pA value from all measurements and scaled by its pA median absolute deviation. The normalization removes measurement biases between sequencing runs, pores, and sensors in a single sequencing run. After normalizing the raw ONT signal, we can

Table 1: SARS-CoV-2 RNAs included in this study, for details see STab. S1. Superscript [a], [b], and [c] denote different patient isolates of the same Pango lineage. All samples were sequenced using an ONT MinION Mk1B platform with an R9.4.1 flow cell. *The raw ONT data and basecalled `FASTQ`s can be found in the SRA BioProject: PRJNA907180. **All reference sequences are available via the OSF database `https://osf.io/evc6k/`. Samples were sequenced using direct RNA sequencing (DRS) on an ONT MinION platform. We show results from `Magnipore` comparisons using B.1.1.7[a] as the reference sample.

Pangolin	Nextclade (WHO)	Sample*	Reference**	Gisaid ID
B.1.513	20C	MinION DRS	MinION ARTIC	–
B.1[a]	20C	MinION DRS	MinION ARTIC	–
B.1[b]	20C	MinION DRS	MinION ARTIC	–
B.1[c]	20C	MinION DRS	MinION ARTIC	–
B.1.177.86	20E	MinION DRS	MinION ARTIC	–
B.1.1	20B	MinION DRS	MinION ARTIC	–
B.1.1.7[a]	20I (Alpha)	MinION DRS	MiSeq ARTIC	EPI_ISL_862139
B.1.1.7[b]	20I (Alpha)	MinION DRS	MiSeq ARTIC	EPI_ISL_862140
B.1.617.2	21J (Delta)	MinION DRS	MiSeq ARTIC	EPI_ISL_2500366
BA.1.17.2	21K (Omicron)	MinION DRS	MiSeq ARTIC	EPI_ISL_7929223
BA.1.18	21K (Omicron)	MinION DRS	MiSeq ARTIC	–
BA.2.9	21L (Omicron)	MinION DRS	MiSeq ARTIC	EPI_ISL_9553935
XBB.1.5	22F (Omicron)	MinION DRS	MinION ARTIC	–
BA.4.1	22A (Omicron)	MinION DRS	MinION ARTIC	–
BE.1	22B (Omicron)	MinION DRS	MinION ARTIC	–
BF.1	22B (Omicron)	MinION DRS	MinION ARTIC	–

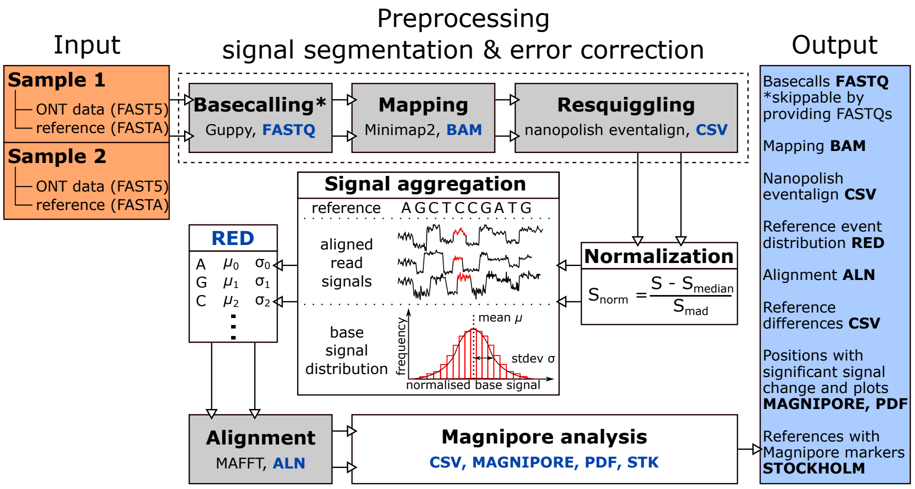

Figure 1: Workflow of `Magnipore`. A full `Magnipore` run starts with the preparation of two input samples. The incoming sample can be basecalled using `Guppy`. The `FASTQ` files are aligned to a given reference sequence using `minimap2`. The resulting `BAM`, `FASTQ`, and `FASTA` references are provided to **`nanopolish eventalign`** to resquiggle and segment the raw ONT data. The segmented read signals are then normalized. For each input sample and for each reference position, the segmented base signals from `nanopolish` are used to approximate a Gaussian signal distribution. To know which positions need to be compared, the input references will be aligned using `mafft`. If both samples share the same reference, then it must be provided in both inputs. The Gaussian signal distributions are compared for each aligned position of both samples. Unaligned positions with gaps are written to the `INDEL` file. Positions of significance surpass a significance threshold calculated according to Eqn. 1, which we will call sites. These significant sites can be ordered and filtered according to their threshold distance (TD) score, coverage, and other output values in the `Magnipore` file, which aids downstream analysis. Significant sites are also marked next to the provided reference sequences in a `STOCKHOLM` file for further analysis. Plots, as in Fig. 5, are also created.

Table 2: Tools included in this study. To run the `Magnipore` pipeline, different external tools are necessary. `Guppy` must be installed and downloaded from nanoporetech, which is only available to customers. All other tools can be installed with `conda`. `Magnipore` was tested with all listed external tools and their corresponding version. Some tools or packages are not published (U – unpublished). This is the case for h5py (`https://zenodo.org/record/6575970`), hdf5 (`https://www.hdfgroup.org/HDF5/`), hdf5plugin (`https://github.com/silx-kit/hdf5plugin`), ont vbz-hdf-plugin (`https://github.com/nanoporetech/vbz-h5py-plugin`), pandas (`https://zenodo.org/record/7658911`) and Guppy (`https://nanoporetech.com/community`).

Tool or Package	Version	Citations
biopython	1.80	[33]
h5py	3.7.0	U
hdf5	1.12.1	U
hdf5plugin	3.3.1	U
mafft	7.508	[34]
matplotlib	3.6.6	[35]
minimap2	2.24	[30]
nanopolish	0.14.0	[29]
numpy	1.23.5	[36]
ont vbz-hdf-plugin	1.0.1	U
pandas	1.5.1	U
samtools	1.16.1	[37]
scipy	1.9.3	[38]
seaborn	0.12.1	[38]
Guppy	6.1.7	U

compare the measurements across different reads and sequencing runs.

Signal aggregation per position

Then, `Magnipore` aggregates the ONT signal for each reference position from the segmented reads provided by `nanopolish eventalign` and calculates the mean and standard deviation. We employ a Gaussian distribution for each position to approximate the signal distribution from all reads in the signal aggregation step (Fig. 1). This provides a good approximation of the ONT signal distribution, as most resemble a Gaussian distribution (SFig. S4). The mean and standard deviation are stored in the reference event distribution `RED` file, containing the reference position, base, motif, signal mean, and standard deviation. The file also contains quality values such as the coverage, distribution density, how many data points are within the 99th percentile of the Gaussian distribution, and more for downstream analysis. `Magnipore` reuses this file by default for comparisons with other samples to save calculation time by default.

`Magnipore`: Position-wise signal comparison between samples

If the `RED` files for two samples are created, `mafft v7.508 [34]` auto alignment mode (`--auto`) is used to align the sample references. The mapping is used to correctly compare the signals per positions of two different individuals position-wise. The signal distributions for the alignment of position i from the reference of sample 1 and position j from the reference of sample 2 are considered to be of significance if the absolute difference of their means is larger than the average of their standard deviations, that is:

$$|\mu_{1,i} - \mu_{2,j}| > \frac{\sigma_{1,i} + \sigma_{2,j}}{2} \tag{1}$$

$$TD(i,j) = 2\frac{|\mu_{1,i} - \mu_{2,j}|}{\sigma_{1,i} + \sigma_{2,j}} \tag{2}$$

Alignment positions are called sites. Furthermore, the threshold distance (TD) score provides a convenient order for the significant sites. The significant sites are given by $TD(\cdot,\cdot) \geq 1$, and the insignificant sites are given by $TD(\cdot,\cdot) < 1$. A significant signal change between position i of one sample and j of the other sample can be caused by molecular changes between both in close proximity to the investigated site. A molecular change can be a mutation or a modification. `Magnipore` classified a significant signal change into mutation or potential modification sites using the provided reference sequences. The classification of significant signals is further described in Tab. 3. `Magnipore` is available at `https://github.com/JannesSP/magnipore`.

Table 3: Biological interpretation of Magnipore events. In all cases, base A in the middle of the 7-mers in bold letters is the base of interest with a significant signal change.

Reference alignment	Magnipore signal diff.	Designation	Symbol
mismatch	✓	base & signal mismatch mutation	▲
GGCCCUA **GGCACUA**	✗	undetected mutation or error in reference	should not occur
mismatch with gap	✓	indel mutation	▲
GGCACUACUA **GGCA---UUA**	✗	undetected mutation or error in reference	should not occur
match	✓	potential modification	🟠
GGCACUA **GGCACUA**	✗	bases & signals match	standard
match with gap	✓	potential modification near repeat deletion	🔴
GGCACUACUA **GGCA---CUA**	✗	bases & signals match	standard

Correlation with RNA secondary structures

To investigate the correlation between potential modification sites and possible RNA secondary structures, we used an alignment-based approach: We clustered complete genomes of the genus *Betacoronavirus* (taken from NCBI virus [39], submitted before June 2020) with MMSeqs2 [40] and identified for each cluster a centroid sequence as representative. We extracted a fragment of 100 *nt* around each significant Magnipore site (>40 %) derived from the genome BA.2.9 eligible for a modification. Homologous regions were used to construct a sequence-secondary-structure alignment by mLocARNA [41]. VARNA [42] and RNAalifold [43] were applied for visualization with the following parameters: `--aln --noLP --noPS --mis -r -d2 --cfactor 0.6 --nfactor 0.5`. We calculated the p-value of each prediction based on a di-nucleotide shuffling implemented in a Python script available on GitHub[1].

Results and Discussion

Magnipore classifies significant sites into mutation and potential modification

Magnipore classifies these significant sites into mutation and potential modification sites. It ignores the gaps in the reference alignment for the classifications and only uses sequence similarity. Magnipore compares the 7-mer sequence context across gaps.

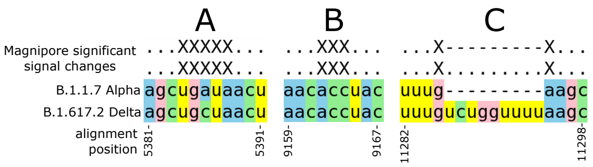

Figure 2: Magnipore provides the reference sequence with marked significant sites as a STOCKHOLM file. This file stores an alignment of marked sites with the references to visualize the context of the significant sites for further investigations. **A** shows a marked region that results from a substitution of the A in the Alpha variant to a C in Delta. **B** is an example for a site with no mutational context in the Illumina reference sequence from GISAID, that was created from the same original sample. **C** visualizes that also insertion and deletion regions result in significant sites.

Significant differential ONT signals (sites) found by Magnipore can appear in different types of events (Tab. 3): The left column in Tab. 3 depicts a comparison of the two input sequences in an alignment.

In case the two reference sequences have a mismatch with or without a gap and Magnipore identifies an expected signal change (✓), then we interpret this to be a mutation site (▲), whereas no identified signal change should not occur.

If the sequences are identical (match) but Magnipore identifies a signal change, we assume this to be a potential modification (🟠), whereas no signal change is the standard case.

In case of a gap resulting in two possible alignments of equal value, Magnipore classifies a significant signal change as a potential modification site (🔴). In both possible alignments, the sequence context is the same.

These cases can be further investigated in the output of Magnipore. Magnipore produces a STOCKHOLM file for each comparison in which all sites with a significant signal change are marked with an X in the sample alignment, see Fig. 2.

We present Magnipore using the example comparison of B.1.1.7[a] Alpha and B.1.617.2 Delta as they were the two first variants of concern in our datasets. Variants of concern may exhibit increased transmissibility and/or escape antibody-mediated neutralization [44]. These samples are therefore very interesting for research.

[1] https://github.com/klamkiew/cov_trs_structure/blob/master/diNuclShuffle.py

`Magnipore` detects 89.1% of all mutations in the SARS-CoV-2 pairwise sample comparisons

To determine the performance of `Magnipore` to detect differential ONT signals, we analyze how many mutations `Magnipore` can identify in pairwise SARS-CoV-2 comparisons. `Magnipore` with default parameters and no coverage filter can detect 89.1% of all mutations within all our SARS-CoV-2 comparisons. A mutation is a difference in the alignment of the reference sequences between two samples, including substitutions, insertions, and deletions (see Tab. 3, blue triangle). We exclude those differences where at least one of the reference sequences shows the character N instead of A, C, G, or T/U. A mutation is captured if `Magnipore` reports at least one significant site in close proximity to three bases upstream or downstream. Across 120 pairwise comparisons 13 978 non N mutations are present in the reference sequences. `Magnipore` identifies 12 460 of those using the significance threshold Eqn. 1 with Gaussian distribution models. The 12 460 identified mutations out of 13 978 form a ratio of 89.14%.

Looking at the pairwise comparisons individually, the median of detected non N mutations is 89.1% while the mean is 87.2%, Fig.4. We calculated the fraction of found non N mutations for each comparison first, then we averaged over all fractions. The lowest fraction of detected mutations is 50% in the comparison of B.1^{a} against B.1^{b} where `Magnipore` finds 2 out of 4 mutations. The highest is 100% in B.1^{a} compared to B.1^{c} with 3 mutations.

In the pairwise `Magnipore` comparisons of B.1.1.7^{a} Alpha against all other samples we filtered for a coverage of 10 and higher, Fig. 3. In the B.1.1.7^{a} Alpha comparisons the mean fraction of detected mutations is 80.07%. It ranges from 51.0%, B.1.1.7^{a} compared to XBB.1.5 Omicron, to 91.9%, B.1.1.7^{a} compared to B.1.513.

`Magnipore` is not able to detect 10.9% of the present mutations. We assume that there are three reasons why `Magnipore` misses 10.9 % of mutations: 1. Inaccurate signal segmentation and resquiggling; 2. Gaussian distributions to capture all types of signals are suboptimal; and 3. Our significance threshold to detect differential signal distributions is not optimal or the signal distribution of the mutated bases are too similar. In both cases they do not show a significant differential signal.

Inaccurate signal segmentation and resquiggling. `Magnipore` uses `nanopolish eventalign` to resquiggle and segment the raw ONT signal. The segments should correspond to a measured sequence of bases, and the segment borders resemble the signal transition of bases measured in the nanopore. Additionally, `nanopolish eventalign` uses the given reference sequence to correct errors within the basecalled reads. This way, bases are associated with the signal segments. Thereby, `nanopolish` erases true mutations and falsely corrects them using the reference bases. It is a false error correction, as haplotypes are erased, and wrong bases are associated with signal segments. That can lead to skewed or mixed signal distributions for a single reference position, SFig. S4. Another reason for mixed distributions is inaccurate signal segmentation, which can happen if the base transmission border is misplaced.

Gaussian distributions to capture all types of signals are suboptimal. We use Gaussian distribution models to approximate each sample's true signal distribution per base. In many cases, the Gaussian distribution model is a reasonable approximation. In other cases, the true signal distribution does not resemble a Gaussian distribution, SFig. S4. The signal distribution shows a mixture of multiple Gaussian distributions in these cases. Such behavior can be caused by inaccurate segmentation or false error correction from the previous case. Our current model cannot capture this behavior. These signal distributions skew the distribution model, which might be why some mutations remain undetected.

Significance threshold not optimal or mutated signals not differential. Another reason for undetected mutations could be that Eqn. 1 is a suboptimal significance threshold to detect all mutations. The signal distributions of mutated signals can sometimes be too similar and therefore do not show a differential signal. One example is the 5-mer models of AATCA and AAGCA that are provided by ONT and used by `nanopolish`. The k-mer distribution models can be found in nanoporetech's github. The 5-mer models of AATCA and AAGCA have very similar distributions, although they have different bases in the middle of the 5-mer. Their mean values are too similar and would not be detected with Eqn. 1. We also tried to use the Kullback-Leibler (KL) divergence with an empirical threshold to find differential signals, SFig. S1.

KL divergence as an alternative significance threshold. The KL divergence is a well-known statistical method to describe diverging or overlapping distributions [45]. A small KL divergence value means less diverging distributions, while

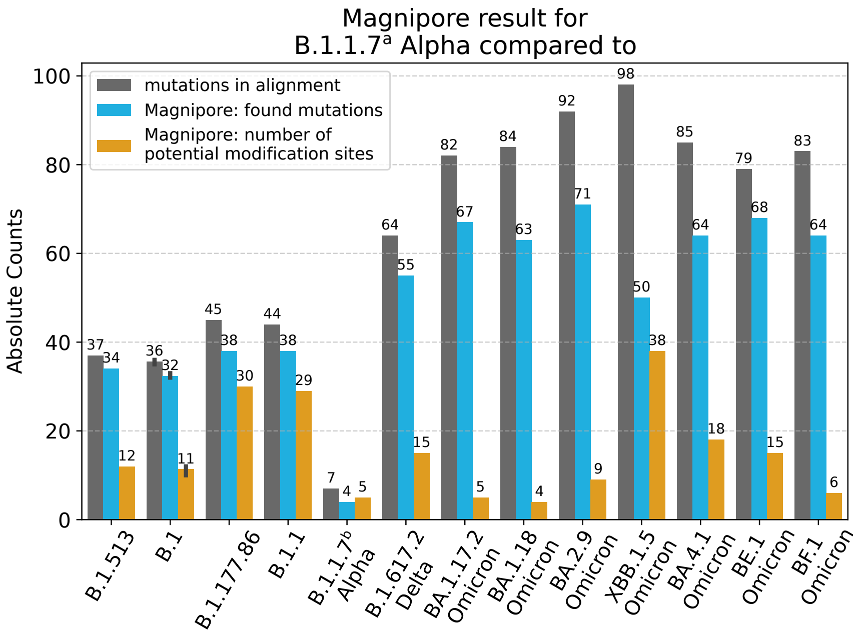

Figure 3: Results by `Magnipore` for B.1.1.7^a Alpha in comparison with all other SARS-CoV-2 strains from Tab. 1. For the comparisons we used a coverage filter of 10. Sites with a coverage below 10 were discarded. We counted the number of mutations between the reference sequences – shown in grey; blue – fraction of identified mutations with `Magnipore`; orange – potential modifications. In case of B.1, we had three samples. The standard deviation of the counts is indicated by the black dash on top of the bars.

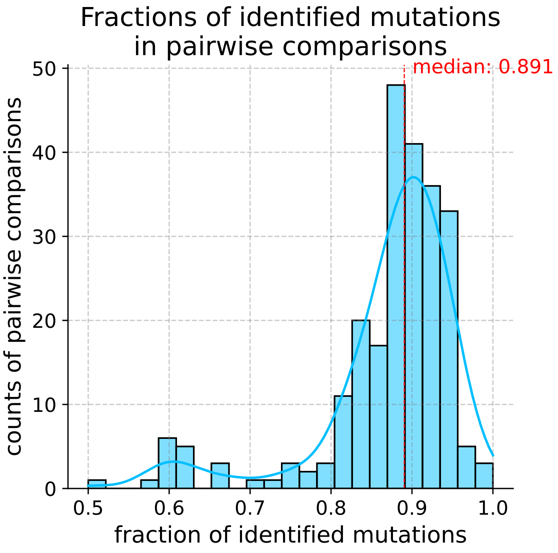

Figure 4: Distribution of fractions of identified mutations (detected non N mutations divided by non N mutations in the references) for all pairwise comparisons without filtering for the coverage. The median number of identified mutations is 0.891, the mean is 0.872.

a large value means strongly diverging distributions. We provide the KL divergence value in the `Magnipore` output and compare it to the TD score. After looking at the KL divergence distribution in the comparison of B.1.1.7^a Alpha and B.1.617.2 Delta, SFig. S1, we decided to set the cutoff for a significant signal change to 1. Each site with a KL divergence above or equal to 1 is significant. We have no gold standard or ground truth data for our samples, so we evaluated the performance of `Magnipore` according to the known mutations between the samples. There are 64 non N mutations between the B.1.1.7^a and B.1.617.2 samples, Fig. 3. Using the TD score and a coverage filter of 10, we found 146 mutation signals clustering around 55 non N mutations. Using the KL divergence threshold of 1 and a coverage filter of 10, we found 117 mutation signals clustering around 53 non N mutations. Both results share 114 mutation signals, which means with the TD score, we can detect more mutation signals and more identified mutations, SFig. S2, S3. Using the TD score we do not need to estimate a significance cutoff for every sample comparison, which we would need if we use the KL divergence. We used the TD score for the sample comparisons. The TD score is easier to calculate, more comprehensible than the KL divergence, and identifies more mutations.

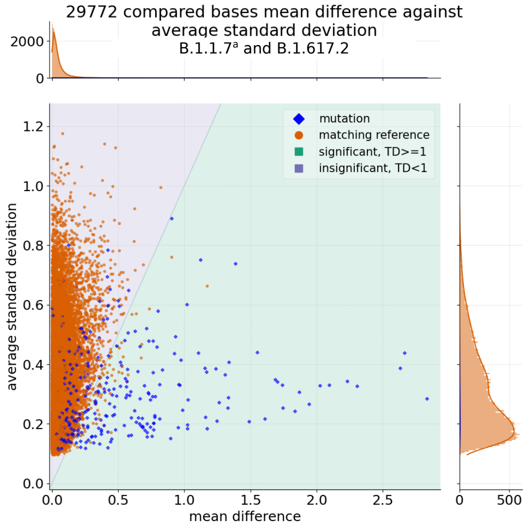

Figure 5: Position-wise signal comparison results between B.1.1.7ª Alpha and B.1.617.2 Delta. For the 29 772 compared positions the absolute mean difference is plotted against the average standard deviation. Orange – nucleotides match; blue – nucleotides mismatch. The data is not filtered for the coverage. The green and red backgrounds mark significant (signals divergent) and insignificant signal changes (signal distribution overlaps), respectively, according to Eqn. 1. The line that splits the green and red area marks the significance threshold created by Eqn. 1.

Sites with significant signal change could hint to potential modifications.

For the classification of mutations and modification, we compared for each site the genomic context in a 7-mer. Fig. 5 shows the absolute mean difference and average standard deviation from Eqn. 1 of each site as one data point for the `Magnipore` comparison of B.1.1.7ª Alpha and B.1.617.2 Delta. Most of the mismatches (blue data points) reside in the significant (green) area of `Magnipore`, as different bases result in different pA measurements. Most sites with a matching reference (orange data points) are insignificant (red background) because the same bases show similar signals and should stay consistent. However, some matching nucleotides show a significant signal difference suggesting potential differential modifications between the samples. We filtered for the coverage to ensure that inaccurate signal distributions do not cause these differential signals. A higher coverage ensures that our signal distributions have a higher data basis, so they tend to resemble the true signal distribution.

Low coverage leads to inaccurate signal distributions. The more coverage is given, the more data is used to approximate the signal distribution, and thus the Gaussian distribution has higher confidence to resemble the original signal distribution. Therefore, the coverage is one major factor for the comparison quality in `Magnipore`. Sites with low coverage have little data for the approximation of the signal distribution. SFig. S6 demonstrates the coverage drop in the region of ORF1a and ORF1b [11]. A low coverage can lead to skewed distributions that do not resemble the true signal distribution. When we execute `Magnipore` and look at all results without a coverage filter, we see a high amount of potential modification sites, especially in low coverage regions around ORF1a and ORF1b, SFig. S6 and SFig. S7. The number of potentially modified sites increases strongly in the Omicron samples when we do not use a coverage filter, SFig. S7. This results from inaccurate approximated signal distributions due to the low coverage.

Modifications can appear as mutations in reference sequences. `Magnipore` and `nanopolish eventalign` assume that the reference sequences are error-free, but this is not always true. In the ARTIC protocol RNA is converted into cDNA by reverse transcriptase. A possible modification of the RNA could introduce systematic errors in the cDNA in which only the nucleotides A, C, G, and T can be incorporated [46]. The ARTIC protocol uses the LunaScript (R) Reverse Transcriptase. According to the manufacturer, this enzyme operates at 55-65°C to melt strong secondary structures. Therefore, it is also able to transcribe areas with pronounced secondary structures. Systematic errors in the case of modified RNAs are not described. However, the introduced errors could lead to mutations between the reference sequences that do not exist in the biological sample of the ONT data [46, 47].

Potential modification sites found repeatedly could represent variant-defining sites.

We looped over all B.1.1.7ª comparisons and counted the potentially modified sites, Tab. 4. Interestingly, potential modification sites consistently reappear in multiple comparisons of the same sample, Fig. 6 and Tab. 4. Such sites could hint at variant or group-defining modifications, as these sites consistently differ between one sample and many others. A generally high density of potential differential modification sites can be observed in the genomic range from about 7 000 to 16 000 (ORFs: nsp 4-10 and RdRp, Fig. 6). These sites must be viewed cautiously, as they are

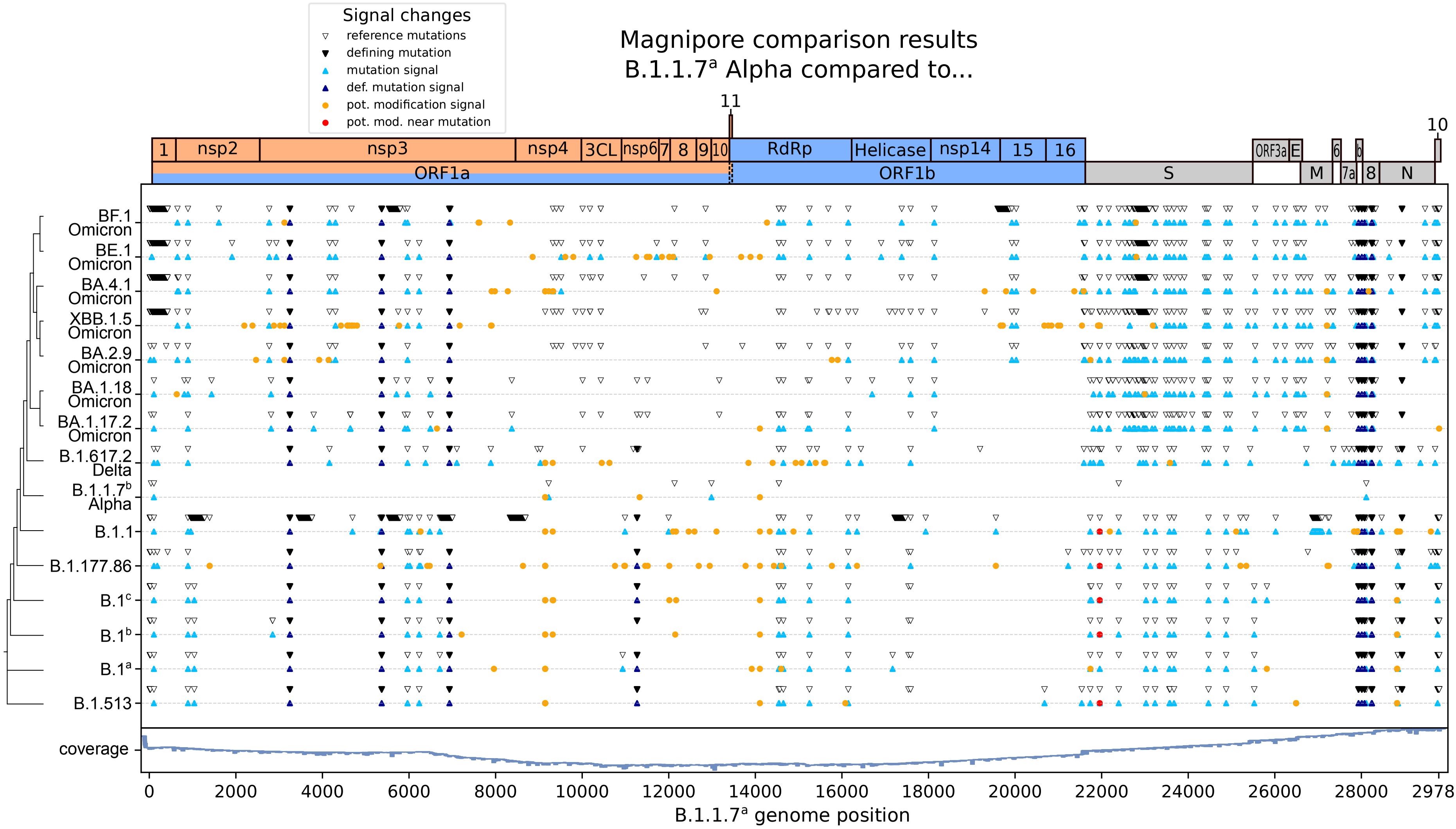

Figure 6: Magnipore sites of SARS-CoV-2 B.1.1.7^{a} Alpha comparisons to all other samples listed in Tab. 1. We marked the significant sites with a coverage of 10 or more reads on the genome of B.1.1.7^{a}. For each comparison, we add two lines: the upper one containing genome information and the lower one with Magnipore results. White – mismatches in genome including 'N' regions; black – strain defining mutations for this SARS-CoV-2 strain compared to B.1.1.7^{a} taken from the GitHub repository cov-lineages (`https://github.com/cov-lineages/`); light blue – significant site of Magnipore; dark blue – significant site of Magnipore in agreement with defining mutations; yellow – potential modification site; red – potential modification site near repeat deletion, see Tab. 3.

in the low coverage range, which could lead to false positives.

Potential differential modification sites separate pre-Omicron and Omicron. Sites 9 133, 9 134, and 9 135 (ORF: nsp 4 Fig. 6) in the B.1.1.7^{a} Alpha genome reappear consistently across multiple comparisons. These potential differential modification sites separate pre-Omicron strains from Omicron strains. In B.1.1.7^{b} comparisons, these sites do not appear as in B.1.1.7^{a}. The source for the significant signal changes is only present in the B.1.1.7^{a} comparisons.

Potential differential modification sites separate Omicron sublineages BA.1, BA.2 and BA.4. Contrarily, the significant sites 27 193 and 27 195 (high coverage region; ORF: M, Fig. 6) appear between B.1.1.7^{a} and of the Omicron samples: BA.1.17.2, BA.1.18, BA.2.9, BA.4.1, and XBB.1.5 (recombinant strain derived from BA.2). These sites can also be found in comparisons between B.1.1.7^{b} and the mentioned Omicron samples. They align with the same sites. When comparing both Alpha strains with the BA.5 sublineages BE.1 and BF.1, 27 193 and 27 195 do not appear as significant. Therefore, these sites could separate BA.1, BA.2, and BA.4 from the rest.

Further investigation is needed to clarify where the signal change is coming from. The signal distributions at 27 193 and 27 195, see SFig. S5, show the distributions of the Alpha and Omicron BA.5 samples to be very similar and tend to be bimodal (two peaks). In contrast, the other Omicron distributions are visibly different and unimodal (single peak). The source of the differential signal between these groups of variants must be investigated in further studies. Possible sources for a bimodal signal are (1) a small percentage of reads have a modification, (2) a small percentage of reads have a mutation, or (3) the segmentation tools struggle to segment the signal at this position accurately. 27 193, and 27 195 could indicate that the viral epitranscriptome differs at these sites characteristically. Sites 27 193 and 27 195 occur with a mutation in the samples' references at position 27 210 in the B.1.1.7^{a} genome. This strain shows an `'A'`, whereas the Omicron variants have a `'C'` in the alignments. Our BA.5 references do not have this mutation and do not show a potential differential modification. The `'A/C'` mutation is 24 nucleotides downstream from the potential differential modification. Hence, the nucleotides are not measured together in the pore, and the mutation should not directly influence the measured electric current. Furthermore, as the mutation is downstream in the RNA sequence, it will be sequenced before the position of the potential differential modification. The mutation has left the pore by the time the potential differential modification is sequenced. Intrinsic properties of the RNA, such as secondary or tertiary structures and RNA modifications, influence the translocation process and time (helicase stalling) [48, 49], but not the intensity of the electric current.

Potential differential modification sites separate variants of concern. Sites 28 812 and 28 815 (ORF: N, Fig. 6) seem to separate the variants of concern (VOCs: Alpha, Delta, Omicron) from the rest (except B.1.177.86). Comparisons with more pre-VOC samples are necessary to substantiate this finding. We only have samples for 4 different pre-VOCs; three B.1 samples could bias this result.

All potential differential modification sites can be found in the OSF database. To identify more defining modification candidates, we looped over all comparisons and collected potentially modified sites that appeared in at least three comparisons, Fig. 7. These sites are listed in the OSF database.

Knowledge of the localization of RNA modifications may be useful to optimize vaccines or search for antiviral targets.

COVID-19 mRNA vaccines use, among others, the modified nucleobase N1-methylpseudouridine to reduce RNA immunogenicity and increase translational efficiency [50]. Knowledge of the location and nature of RNA modifications may also be important for the development of therapeutic approaches, especially since such modifications have been described in a wide range of RNA viruses [14, 15, 16, 17, 18].

False positive potential modifications – limitation by `nanopolish eventalign`. One source for deviating signals at potential modification sites could be mutated subsets of viruses called haplotypes. Mutations get erroneously corrected as `nanopolish eventalign` tries to correct all reads according to the provided reference sequence. Signals caused by mutations in individual reads might be assigned to the wrong bases. The mutation signals then shift and skew the signal distributions at these sites, SFig. S4, which could lead to false positive potential modification sites between samples. Mistaking mutations for base-calling errors is a limitation of the resquiggle tool.

Table 4: Reappearing consistent potential modification sites in the comparisons with B.1.1.7^{a} Alpha. We count them and show a part of all sites that appeared at least twice in this table. These sites must be considered in relation to the B.1.1.7^{a} reference. Therefore, we show them on the basis of their position in this strain. All underlying data can be found as `CSV` files in the OSF database. RdRp – RNA-dependent RNA polymerase; nsp – nonstructural protein; N – nucleocapsid phosphoprotein; M – membrane glycoprotein; S – surface glycoprotein

Position	Count	5'-Motif-3'	Gene
9 133	9	AAC**A**CCT	nsp4
9 134	9	ACA**C**CTA	nsp4
9 135	9	CAC**C**TAC	nsp4
14 094	9	GTT**C**CTG	RdRp
9 308	5	TAA**A**TTT	nsp4
27 193	5	AGA**T**ATT	M
27 195	5	ATA**T**TAC	M
28 812	5	AAT**T**CAA	N
28 815	5	TCA**A**CTC	N
3 110	3	AAC**C**TTT	nsp3
7 897	2	GCA**A**AAT	nsp3
9 136	2	ACC**T**ACC	nsp4
9 320	2	CTA**A**TAT	nsp4
12 001	2	GTA**G**ACA	nsp7
12 002	2	TAG**A**CAT	nsp7
12 935	2	ACC**T**AAA	nsp9
14 586	2	TGC**T**TTT	RdRp
21 729	2	CTC**T**GGG	S
28 818	2	ACT**C**CAG	N
⋮	⋮	⋮	⋮

The exact signal distributions for each potential modification must be investigated individually to further investigate possible false positive sites, SFig. S9.

Possible influence of virus cultivation and RNA preparation techniques on the prediction of modification sites. Another reason for the observed variability could be the virus cultivation and the subsequent RNA preparations. For the propagation of SARS-CoV-2, commercially available Vero cells were used, originating from the kidney of African green monkeys [51]. It is known from experiments with arboviruses that the cell line itself influences RNA modifications. Thus, the modifications differ between Sindbis virions that were propagated in insect cells and those from mammalian cells. Virions propagated in the insect cells replicated better in the mammalian cells [52]. Although certainly not as pronounced as in the example mentioned above of a switch between insect and mammalian cells, it is conceivable that replication of SARS-CoV-2 in other mammalian cells could impact the modification pattern and, thus, the prediction of modified sites in the genome. It is also possible that the modifications of the viral RNA from the cell are different from those of the infectious virion. The RNAs we sequenced originate mainly from cell lysates but may also contain portions of extracellular virions. Another aspect is the occurrence of subpopulations *in vivo* and *in vitro*, so-called quasi-species [53], which, as explained in the previous section, can influence the results. In addition, the RNA was prepared after the occurrence of a cytopathic effect detectable by light microscopy. Standardization to the virus dose used and the cell cycle did not take place. This could be taken into account in future experiments. Likewise, different host cells and RNA preparations from purified virions of the culture supernatant should be investigated.

Modifications tend to occur in stacks of predicted RNA secondary structures.

Previous studies have shown that the RNA of SARS-CoV-2, like that of other betacoronaviruses, has conserved secondary and probably tertiary structures essential for the viral life cycle [54, 55, 56, 57, 11]. We investigated the correlation between modifications and *in silico* predicted RNA secondary structures in genomes of betacoronaviruses (including SARS-CoV-2).

`Magnipore` identified eight differentially modified sites across the entire genome which occur in at least 40% of all pairwise comparisons: 2 473, 3 932, 4 148, 15 915, 15 920, 16 143, and 27 211, 27 213 (position in BA.2.9 genome). We detected six potential differentially modified sites in a stack of base pairs, Fig. 8. The remaining two sites (15 920 and 16 143) are located in a hairpin loop and an internal loop, respectively. The alignments reveal a high sequence similarity resulting in predicted base pairs only containing one base pair type (labeled red in alignment and secondary structure in Fig. 8). Three modification sites are located in ORF1a. We predicted structural elements in the nonstructural protein (nsp) 2 gene containing a small hairpin loop with only three base pairs and a 47 nt long secondary structure with two internal loops (see Fig. 8, left bottom). This predicted formation molecule has a p-value of 2.5×10^{-3}. One modification in the nsp3 gene is located in an RNA secondary structure consisting of three stem-loops (one with an internal loop). We observed the lowest p-value (6.94×10^{-23}) compared to the other sequence-structure alignments in this

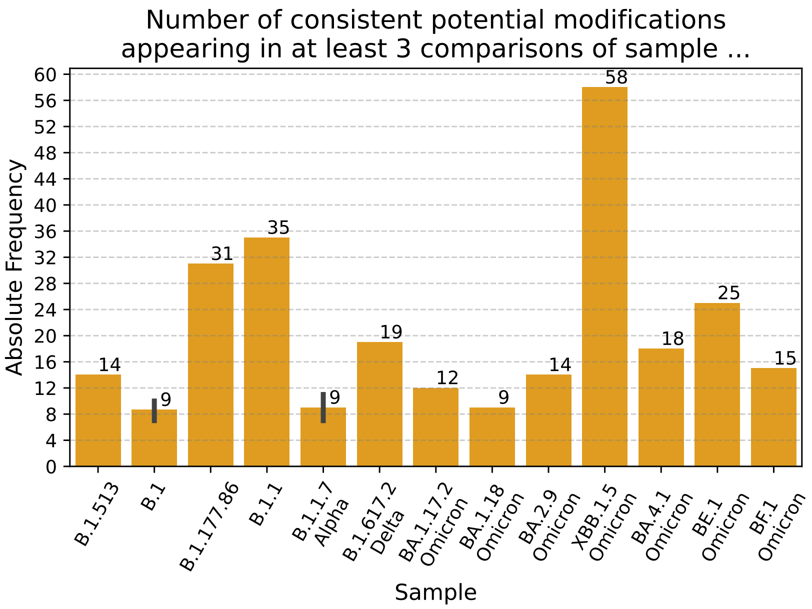

Figure 7: Consistently reappearing RNA modifications. For each site in each sample comparison against all other samples of Tab. 1, the number of significant sites classified by `Magnipore` as potentially modified is given. If a site appears in three or more sample comparisons as potentially modified, then we count this site as a consistently reappearing modification. These sites should be investigated further with higher priority as they might hint at potential defining modifications. A list of all sites can be found in the OSF database.

RNA molecule in this study. According to our computational predictions, the other modification site in the nsp3 gene may be located in a multi-loop. We observed more divergent base pairs within this structure according to the possible base pair types. Even two base pairs occur with three or four types (green and cyan colored in Fig. 8 top middle). We predicted two structures in ORF1b in the RNA-dependent RNA polymerase (RdRp) gene. The first element includes two modification sites, one in a stack and the other in the hairpin. This prediction shows the highest p-value (1.86×10^{-2}) and many predicted base pairs do not fold in two or more sequences of the alignment, assuming that this structure might not be present. The second element in the RdRp gene revealed a molecule of three stem-loops (one with several internal loops). The high level of sequence conservation in this region leads to a high number of base pairs with only one base pair type (red). However, within this RNA secondary structure, we observed three different types of basepairs (green) at one position, giving a higher confidence in the existence of the structure. The last two modification sites are located in ORF6 and present in a stacking of base pairs (see Fig. 8 bottom right). One modification is in a base pair with one possible base pair type. The other basepair does not fold in more than two sequences in the alignment. The prediction of this region revealed three stem-loops in total with a p-value of 1.36×10^{-17}. A pairing in the last element is possible with three base pair types.

The structures shown here were predicted using computational methods only and may not reflect secondary structures found in betacoronaviruses *in vivo*. Across the potential modification sites and predicted RNA secondary structures, we observed the tendency of differentially modified sites occurring in a stack of base pairs.

Comparison with similar tools

We used B.1.1.7^{a} Alpha and B.1.617.2 Delta to compare `Magnipore` with `xPore`, `Nanocompore`, and `Yanocomp`. These tools work similarly to `Magnipore` by searching for differential signals within the ONT data. We executed `Magnipore` with default parameters. We used `Guppy 6.1.7` with the `rna_r9.4.1_70bps_hac` model to basecall the SARS-CoV-2 RNA sequencing data. We use the resulting reads from `Guppy` for every tool in the comparison to establish comparable conditions. The B.1.1.7^{a} sample has 798 140 basecalled reads with around 1 308 000 000 bases, of which 1 023 000 000 could be mapped to the B.1.1.7 reference sequence. B.1.617.2 had 526 766 basecalled reads with around 67 200 000 bases, of which only 14 860 000 could be mapped to the B.1.617.2 reference sequence. We consider only sites with a coverage of at least 10 in both samples.

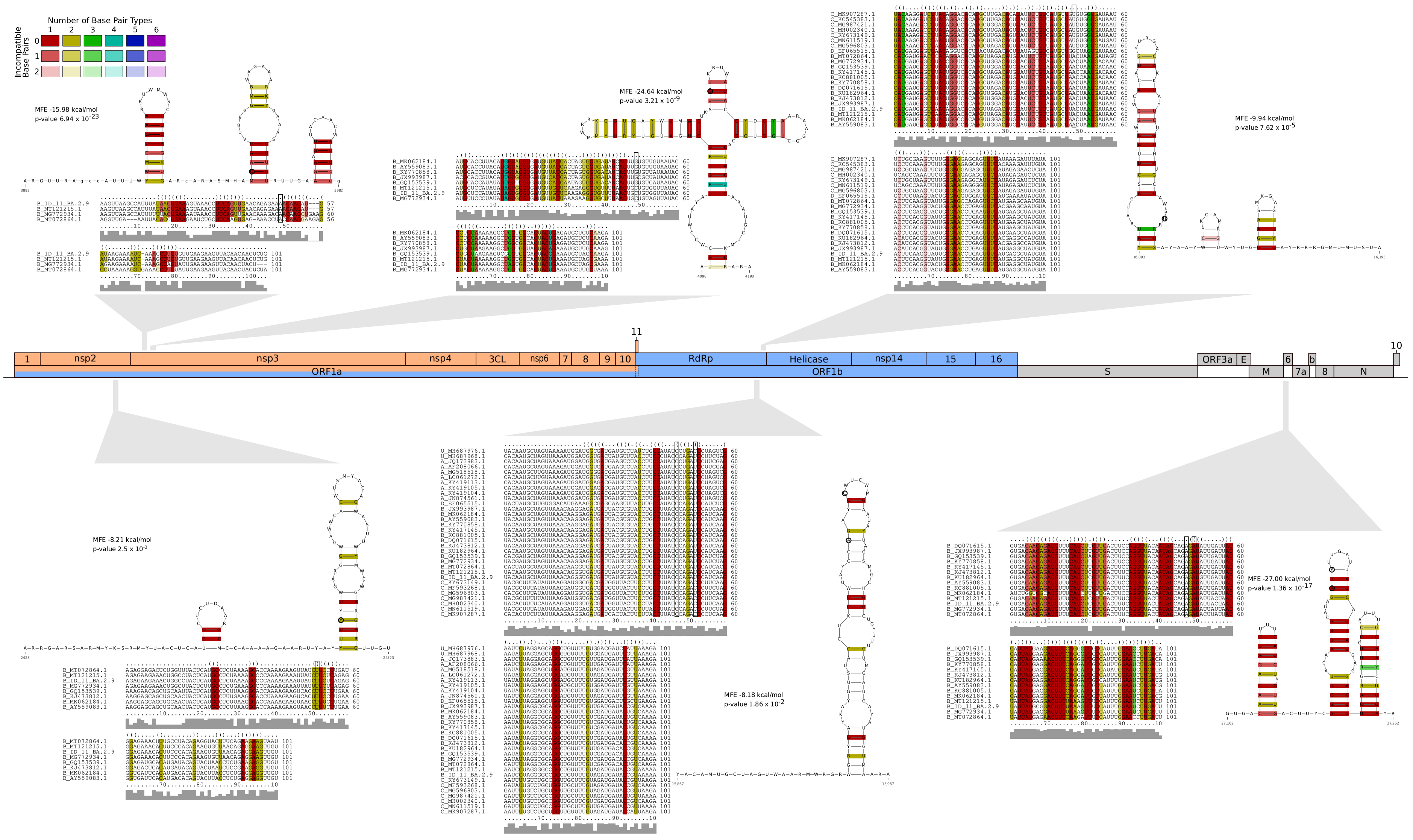

Figure 8: Predicted RNA secondary structures of betacoronaviruses in correlation with differentially modified sites in SARS-CoV-2 genome. Predicting structural elements by a computational alignment-based approach, we observed six out of eight sites in a stacking of base pairs. The remaining two are in a hairpin and internal loop. The sequence similarity in the respective regions is high. Thus, many predicted base pairs are only folded by one base pair type (colored red in alignment and secondary structure). Nevertheless, some predicted structural elements show three or four possible base pair types (green and cyan). The betacoronavirus lineages (A-D), according to ICTV [58], were added to the respective accession IDs in the alignments. Unclassified genomes are labeled with U.

For sample B.1.1.7[a] these are 29 784 of 29 788 nucleotides, and for sample B.1.617.2 25 413 of 25 413 of 29 858 bases.

Magnipore detects 161 significant sites consisting of 146 mutation and 15 potential modification sites. The 161 identified sites have a significant TD score greater or equal to 1 (see materials and methods). Magnipore classified these sites into 146 mutation and 15 potential modification sites.

Multiple adjacent sites can appear with a significant differential ONT signal caused by only one molecular modification or mutation, Fig. 2A. Likewise, the number of differing ONT signals is larger or equal to the number of naturally appearing mutations and detected mutations reported by Magnipore.

Magnipores 146 mutation sites resulted in 110 detected mutations, of which 55 are non N mutations, and a high fraction of 55 mutations include an N at the site of interest. With this result, Magnipore missed to identify 20 mutations: The ground truth (read from the alignment between B.1.1.7[a] and B.1.617.2 of the reference sequences) is 130 mutations, of which 64 are non N, and 66 are N mutations.

The 15 potential modification sites are not close to a mutation, i.e., not in the range of three nucleotides up or downstream of a mutation. At these sites, the signal significantly differs between the samples without a mutation in the reference sequence alignment. Potentially modified sites need further investigation because of software limitations like inaccurate signal segmentation or incorrect error correction by nanopolish eventalign.

For this comparison, Magnipore needed around 35 hours to process, compare, write, and plot the data of both samples on an AMD Ryzen 9 3900X 12-Core Processor with 24 threads and an NVIDIA GeForce RTX 2080 Ti. Basecalling took 1 hour and 17 minutes (B.1.1.7) and 26 minutes (B.1.617.2). Mapping and nanopolish preparation took 4 hours and 20 minutes (B.1.1.7) and 21 minutes (B.1.617.2), respectively. Creation of the distribution models for every position took 24 hours and 20 minutes (B.1.1.7) and 3 hours and 43 minutes (B.1.617.2). Comparing the position-wise signal distributions with the help of an alignment took just two minutes.

xPore misses sites and classifies mutations as modifications. xPore analyzes only 2 057 sites and mistakes mutations for modifications. As preparation for xPore, both samples are mapped to only one reference and run nanopolish eventalign. We map both samples to the B.1.1.7[a] Alpha reference. For nanopolish, the parameters signal-index and scale-events are used. We execute xpore dataprep with readcount_min 10 to filter for sites with a coverage of at least 10, compare the output to Magnipore and prepare the data for further analysis with xpore. We execute xpore diffmod with default parameters and compare the output to Magnipore. xPore provides results for 2 057 positions in the genomic range (not every position present in the ranges) of 2 to 1 710 nt and 25 347 to 29 783 nt on the B.1.1.7[a] genome, which is roughly 6.9% of all positions. For further analysis, we filtered these 2 057 sites for a p-value below 0.01 and reduced the set to 553.

Magnipore and xPore share 23 significant sites with a p-value below 0.01 by xPore and a TD score higher or equal to 1 by Magnipore (sixth bar in Fig. 9). xPore does not differentiate between modification and mutation sites, as it expects to compare the same biological sample in two different conditions. It only reports differential modification rates. xPore reports 48 significant sites out of the 553 in close proximity to 25 mutations (fifth bar in Fig. 9). Therefore, xPore mistakes 48 mutations as significantly differentially modified. These are 48 false positive sites regarding xPores task to find differential modification. Magnipore can compare different SARS-CoV-2 variants that contain mutations and reports them. Magnipore detects 161 significant positions. It detects and correctly classifies 55 mutations. Magnipore shares 23 significant sites with xPore, which are in close proximity to 20 mutations. Magnipore and xPore do not share potential modification sites.

For Nanocompore, we followed the guidelines in their documentation. As Nanocompore only works on a single reference genome, we mapped the reads from the B.1.617.2 sample to the B.1.1.7[a] reference sequence. We used the resulting mappings to execute nanopolish eventalign for both samples with the required parameters --print-read-names --scale-events --samples. After that, we ran nanocompore eventalign_collapse on both samples with default parameters and executed nanocompore sampcomp on the output files. We could not run nanocompore sampcomp without getting an error.

We ran yanocomp according to the guideline on https://github.com/bartongroup/yanocomp. The data was prepared with nanopolish eventalign using the parameters --print-read-names --scale-events and --signal-index. We executed yanocomp prep with default parameters. Finally, we used yanocomp gmmtest with default parameters and --min-read-depth 10. We could not get any

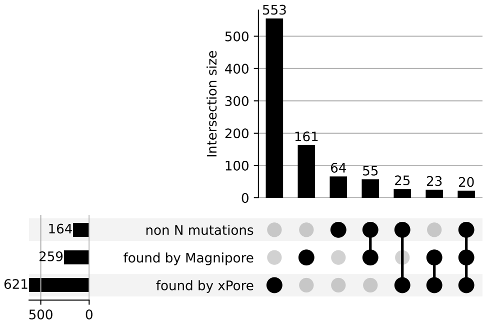

Figure 9: Upset plot showing sets and intersections between present non N mutations, found positions by `Magnipore` and found positions by `xPore`. We compared the output of `Magnipore`, `xPore`, and the set of present mutations between B.1.1.7^a Alpha and B.1.617.2 Delta. Connected dots indicate set intersections. Bars show the number of elements (positions) in this intersection. Intersections with non N mutations represent the number of identified mutations by the respective tool.

result with `yanocomp` and stopped `gmmtest` after 10 days of running.

Conclusion

Limitations Further improvement to Magnipore and its signal analysis capabilities will require a better resquiggling and segmentation algorithm of `nanopolish eventalign`. An essential task in the future will be a precise differentiation between basecalling errors, mutations, and modifications among the reads and the reference sequence of the sample. Another way to confirm the results of `Magnipore` would be to have multiple samples or replicates from the same variant. This way, we could gain confidence about consistent potential modification signals.
Multiple aspects influence the analysis of two related samples with `Magnipore`. One is the quality of the reference sequence. A high-quality reference sequence (1) should not contain ambiguous bases (N), (2) should not miss bases at the 5' and 3' endings, and (3) should not contain wrong bases (errors). `Magnipore` can also work with related reference sequences, but this will reduce the quality of the results. `Magnipore` depends heavily on the segmentation and resquiggling quality of `nanopolish eventalign`. Improvements in resquiggling and segmentation of the ONT signal would be a significant step forward in the performance of `Magnipore`, as the range of influence from mutations can be narrowed down. An evenly distributed high coverage or deep sequencing is optimal for the best analysis. `Magnipore` analyzes the raw ONT signal and only depends on the `Guppy` basecaller to map the reads to the corresponding reference sequence for the segmentation step. Other steps are independent of the basecaller. The tool can find sites with a significant signal change on a genomic level that appear consistently in different sample comparisons.

Outlook With the help of RNA sequencing and `Magnipore`, nucleotide positions that may be differentially modified can be identified. Individual sites are found across variants, while some appear specific to groups of more closely related variants. However, these are only predictions so far, and there is no experimental confirmation yet of the suspected differential RNA modifications between SARS-CoV-2 variants. From our point of view, `Magnipore` detects promising sites for potential differential RNA modifications and, in some cases, even potential defining RNA modification sites.
Findings about the localization and type of chemical modifications in the SARS-CoV-2 genome could serve as a starting point for advancing antiviral drug discovery. It is also conceivable that this knowledge will help improve the effectiveness of existing mRNA vaccines.

Funding This work was funded by (1) the German DFG Collaborative Research Centre AquaDiva (CRC 1076/3-A06 AquaDiva), NFDI 28/1 "NFDI4Microbiota", EXC 2051 "Balance of the Microverse", iDiv "AllinOneSequencing" FZT1018 (2) the German state of Thuringia via the Thüringer Aufbaubank (2021 FGI 0009), and ProDigital 2020-2024 "DigLeben" 5575/10-9 TMWBDG. (3) the Carl-Zeiss-Stiftung within the program Scientific Breakthroughs in Artificial Intelligence (project "Interactive Inference"), and Durchbrüche "Werkstatt" FKZ0563-2.8/738/2. Part of the costs for the RNA sequencing were covered by Labor Krause MVZ GmbH.

Acknowledgements The authors gratefully acknowledge the help of Sebastian Krautwurst in understanding and handling Oxford Nanopore data. The authors thank Dr. Barbara Mühlemann (Institute of Virology, Charité Berlin) for providing the Illumina reference sequences. Special thanks go to Dr. Thomas Lorentz (Labor Dr. Krause) and Prof. Helmut Fickenscher (Institute for Infection Medicine) for providing the equipment, reagents and facilities.

Author contributions Provided the SARS-CoV-2 RNAs and ARTIC-protocol generated reference sequences: AK, MMH, CMC, LP, AW, RR; RNA sequencing: MZ, AS; Tool development: JS; Algorithm design: JS, CHzS, MM; Wrote the paper: JS, CHzS, MM, MZ, AK, ST; Guided the study: CHzS, AK, MM.

Code availability The `Magnipore` pipeline is written in *python3* and available via the GitHub repository `https://github.com/JannesSP/magnipore` and via Conda `https://anaconda.org/JannesSP/magnipore`.

Data availability The raw ONT sequencing data and basecalled `FASTQ` files can be found in the SRA database under the BioProject: PRJNA907180. The output of `Magnipore`, the list of consistently reappearing potential modified positions and reference sequences can be found in the OSF database `https://osf.io/evc6k/`.

References

[1] David Deamer, Mark Akeson, and Daniel Branton. "Three decades of nanopore sequencing." In: *Nature Biotechnology* 34.5 (May 2016), pp. 518–524. DOI: `10.1038/nbt.3423`. URL: `https://doi.org/10.1038/nbt.3423`.

[2] Andrew H. Laszlo et al. "Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA." In: *Proceedings of the National Academy of Sciences* 110.47 (Oct. 2013), pp. 18904–18909. DOI: `10.1073/pnas.1310240110`. URL: `https://doi.org/10.1073/pnas.1310240110`.

[3] Jacob Schreiber et al. "Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands." In: *Proceedings of the National Academy of Sciences* 110.47 (Oct. 2013), pp. 18910–18915. DOI: `10.1073/pnas.1310615110`. URL: `https://doi.org/10.1073/pnas.1310615110`.

[4] Qian Liu et al. "NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data." In: *BMC Genomics* 20.S1 (Feb. 2019). DOI: `10.1186/s12864-018-5372-8`. URL: `https://doi.org/10.1186/s12864-018-5372-8`.

[5] Alexander E. Gorbalenya et al. "The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2." In: *Nature Microbiology* 5.4 (Mar. 2020), pp. 536–544. DOI: `10.1038/s41564-020-0695-z`. URL: `https://doi.org/10.1038/s41564-020-0695-z`.

[6] Yaara Finkel et al. "The coding capacity of SARS-CoV-2." In: *Nature* 589.7840 (Sept. 2020), pp. 125–130. DOI: `10.1038/s41586-020-2739-1`. URL: `https://doi.org/10.1038/s41586-020-2739-1`.

[7] William T. Harvey et al. "SARS-CoV-2 variants, spike mutations and immune escape." In: *Nature Reviews Microbiology* 19.7 (June 2021), pp. 409–424. DOI: `10.1038/s41579-021-00573-0`. URL: `https://doi.org/10.1038/s41579-021-00573-0`.

[8] Christian Brandt et al. "poreCov-An Easy to Use, Fast, and Robust Workflow for SARS-CoV-2 Genome Reconstruction via Nanopore Sequencing." In: *Frontiers in Genetics* 12 (July 2021). DOI: `10.3389/fgene.2021.711437`. URL: `https://doi.org/10.3389/fgene.2021.711437`.

[9] Franziska Hufsky et al. "Computational strategies to combat COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research." In: *Briefings in Bioinformatics* 22.2 (Nov. 2020), pp. 642–663. DOI: `10.1093/bib/bbaa232`. URL: `https://doi.org/10.1093/bib/bbaa232`.

[10] Sebastian Krautwurst et al. "Direct RNA Sequencing for Complete Viral Genomes." In: *Virus Bioinformatics*. Ed. by Dmitrij Frishman and Manja Marz. Chapman and Hall/CRC, 2021, pp. 35–50.

[11] Adrian Viehweger et al. "Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis." In: *Genome Research* 29.9 (Aug. 2019), pp. 1545–1554. DOI: `10.1101/gr.247064.118`. URL: `https://doi.org/10.1101/gr.247064.118`.

[12] Ramakanth Madhugiri et al. "RNA structure analysis of alphacoronavirus terminal genome regions." In: *Virus Research* 194 (Dec. 2014), pp. 76–89. DOI: `10.1016/j.virusres.2014.10.001`. URL: `https://doi.org/10.1016/j.virusres.2014.10.001`.

[13] R. Jordan Ontiveros, Julian Stoute, and Kathy Fange Liu. "The chemical diversity of RNA modifications." In: *Biochemical Journal* 476.8 (Apr. 2019), pp. 1227–1245. DOI: `10.1042/bcj20180445`. URL: `https://doi.org/10.1042/bcj20180445`.

[14] Haojie Hao et al. "N6-methyladenosine modification and METTL3 modulate enterovirus 71 replication." In: *Nucleic Acids Research* 47.1 (Oct. 2018), pp. 362–374. DOI: `10.1093/`

nar/gky1007. URL: https://doi.org/10.1093/nar/gky1007.

[15] Gianluigi Lichinchi et al. "Dynamics of Human and Viral RNA Methylation during Zika Virus Infection." In: *Cell Host Microbe* 20.5 (Nov. 2016), pp. 666–673. DOI: 10.1016/j.chom.2016.10.002. URL: https://doi.org/10.1016/j.chom.2016.10.002.

[16] Michèle Brocard, Alessia Ruggieri, and Nicolas Locker. "m6A RNA methylation, a new hallmark in virus-host interactions." In: *Journal of General Virology* 98.9 (Sept. 2017), pp. 2207–2214. DOI: 10.1099/jgv.0.000910. URL: https://doi.org/10.1099/jgv.0.000910.

[17] Graham D. Williams, Nandan S. Gokhale, and Stacy M. Horner. "Regulation of Viral Infection by the RNA Modification N6-Methyladenosine." In: *Annual Review of Virology* 6.1 (Sept. 2019), pp. 235–253. DOI: 10.1146/annurev-virology-092818-015559. URL: https://doi.org/10.1146/annurev-virology-092818-015559.

[18] Hannah M. Burgess et al. "Targeting the m6A RNA modification pathway blocks SARS-CoV-2 and HCoV-OC43 replication." In: *Genes & Development* 35.13-14 (June 2021), pp. 1005–1019. DOI: 10.1101/gad.348320.121. URL: https://doi.org/10.1101/gad.348320.121.

[19] Christopher Hendra et al. "Detection of m6A from direct RNA sequencing using a Multiple Instance Learning framework." In: (Sept. 2021). DOI: 10.1101/2021.09.20.461055. URL: https://doi.org/10.1101/2021.09.20.461055.

[20] Adrien Leger et al. "RNA modifications detection by comparative Nanopore direct RNA sequencing." In: (Nov. 2019). DOI: 10.1101/843136. URL: https://doi.org/10.1101/843136.

[21] Huanle Liu et al. "Accurate detection of m6A RNA modifications in native RNA sequences." In: *Nature Communications* 10.1 (Sept. 2019). DOI: 10.1038/s41467-019-11713-9. URL: https://doi.org/10.1038/s41467-019-11713-9.

[22] Daniel A. Lorenz et al. "Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution." In: *RNA* 26.1 (Oct. 2019), pp. 19–28. DOI: 10.1261/rna.072785.119. URL: https://doi.org/10.1261/rna.072785.119.

[23] Oguzhan Begik, John S Mattick, and Eva Maria Novoa. "Exploring the epitranscriptome by native RNA sequencing." In: *RNA* (Sept. 2022), rna.079404.122. DOI: 10.1261/rna.079404.122. URL: https://doi.org/10.1261/rna.079404.122.

[24] William Stephenson et al. "Direct detection of RNA modifications and structure using single-molecule nanopore sequencing." In: *Cell Genomics* 2.2 (Feb. 2022), p. 100097. DOI: 10.1016/j.xgen.2022.100097. URL: https://doi.org/10.1016/j.xgen.2022.100097.

[25] Jonathan S Abebe et al. "DRUMMER—rapid detection of RNA modifications through comparative nanopore sequencing." In: *Bioinformatics* 38.11 (Apr. 2022). Ed. by Valentina Boeva, pp. 3113–3115. DOI: 10.1093/bioinformatics/btac274. URL: https://doi.org/10.1093/bioinformatics/btac274.

[26] Piroon Jenjaroenpun et al. "Decoding the epitranscriptional landscape from native RNA sequences." In: *Nucleic Acids Research* 49.2 (July 2020), e7–e7. DOI: 10.1093/nar/gkaa620. URL: https://doi.org/10.1093/nar/gkaa620.

[27] Matthew T. Parker, Geoffrey J. Barton, and Gordon G. Simpson. "Yanocomp: robust prediction of m6A modifications in individual nanopore direct RNA reads." In: (June 2021). DOI: 10.1101/2021.06.15.448494. URL: https://doi.org/10.1101/2021.06.15.448494.

[28] Ploy N. Pratanwanich et al. "Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore." In: *Nature Biotechnology* 39.11 (July 2021), pp. 1394–1402. DOI: 10.1038/s41587-021-00949-w. URL: https://doi.org/10.1038/s41587-021-00949-w.

[29] Nicholas J Loman, Joshua Quick, and Jared T Simpson. "A complete bacterial genome assembled de novo using only nanopore sequencing data." In: *Nature Methods* 12.8 (June 2015), pp. 733–735. DOI: 10.1038/nmeth.3444. URL: https://doi.org/10.1038/nmeth.3444.

[30] Heng Li. "Minimap2: pairwise alignment for nucleotide sequences." In: *Bioinformatics* 34.18 (May 2018). Ed. by Inanc Birol, pp. 3094–3100. DOI: 10.1093/bioinformatics/bty191. URL: https://doi.org/10.1093/bioinformatics/bty191.

[31] Annabelle Strömer et al. “Performance of a Point-of-Care Test for the Rapid Detection of SARS-CoV-2 Antigen.” In: *Microorganisms* 9.1 (Dec. 2020), p. 58. DOI: 10.3390/microorganisms9010058. URL: https://doi.org/10.3390/microorganisms9010058.

[32] Ruben Rose et al. “Delta or Omicron BA.1/2-neutralizing antibody levels and T-cell reactivity after triple-vaccination or infection.” In: *Allergy* 77.10 (June 2022), pp. 3130–3133. DOI: 10.1111/all.15395. URL: https://doi.org/10.1111/all.15395.

[33] Peter JA Cock et al. “Biopython: freely available Python tools for computational molecular biology and bioinformatics.” In: *Bioinformatics* 25.11 (2009), pp. 1422–1423.

[34] Kazutaka Katoh and Daron M. Standley. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” In: *Molecular Biology and Evolution* 30.4 (Jan. 2013), pp. 772–780. DOI: 10.1093/molbev/mst010. URL: https://doi.org/10.1093/molbev/mst010.

[35] John D Hunter. “Matplotlib: A 2D graphics environment.” In: *Computing in science & engineering* 9.3 (2007), pp. 90–95.

[36] Charles R. Harris et al. “Array programming with NumPy.” In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[37] Petr Danecek et al. “Twelve years of SAMtools and BCFtools.” In: *GigaScience* 10.2 (Feb. 2021). giab008. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. eprint: https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf. URL: https://doi.org/10.1093/gigascience/giab008.

[38] Michael L. Waskom. “seaborn: statistical data visualization.” In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

[39] Eneida L. Hatcher et al. “Virus Variation Resource – improved response to emergent viral outbreaks.” In: *Nucleic Acids Research* 45.D1 (Nov. 2016), pp. D482–D490. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1065. eprint: https://academic.oup.com/nar/article-pdf/45/D1/D482/8846723/gkw1065.pdf. URL: https://doi.org/10.1093/nar/gkw1065.

[40] Martin Steinegger and Johannes Söding. “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.” In: *Nature Biotechnology* 35.11 (Nov. 2017), pp. 1026–1028. ISSN: 1546-1696. DOI: 10.1038/nbt.3988. URL: https://doi.org/10.1038/nbt.3988.

[41] Sebastian Will et al. “Inferring Noncoding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering.” In: *PLOS Computational Biology* 3.4 (Apr. 2007), pp. 1–12. DOI: 10.1371/journal.pcbi.0030065. URL: https://doi.org/10.1371/journal.pcbi.0030065.

[42] Kévin Darty, Alain Denise, and Yann Ponty. “VARNA: Interactive drawing and editing of the RNA secondary structure.” In: *Bioinformatics* 25.15 (Apr. 2009), pp. 1974–1975. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp250. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/15/1974/16887892/btp250.pdf. URL: https://doi.org/10.1093/bioinformatics/btp250.

[43] Ronny Lorenz et al. “ViennaRNA Package 2.0.” In: *Algorithms for Molecular Biology* 6.1 (Nov. 2011), p. 26. ISSN: 1748-7188. DOI: 10.1186/1748-7188-6-26. URL: https://doi.org/10.1186/1748-7188-6-26.

[44] Emma B Hodcroft et al. “Spread of a SARS-CoV-2 variant through Europe in the summer of 2020.” In: *Nature* 595.7869 (2021), pp. 707–712.

[45] David JC MacKay. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

[46] Vladimir Potapov et al. “Base modifications affecting RNA polymerase and reverse transcriptase fidelity.” In: *Nucleic Acids Research* 46.11 (May 2018), pp. 5753–5763. ISSN: 0305-1048. DOI: 10.1093/nar/gky341. eprint: https://academic.oup.com/nar/article-pdf/46/11/5753/25067063/gky341.pdf. URL: https://doi.org/10.1093/nar/gky341.

[47] Sven Findeiß et al. “Traces of post-transcriptional RNA modifications in deep sequencing data.” In: *Biological Chemistry* 392.4 (Apr. 2011). DOI: 10.1515/bc.2011.043. URL: https://doi.org/10.1515/bc.2011.043.

[48] Matteo Becchi et al. “RNA Pore Translocation with Static and Periodic Forces: Effect of Secondary and Tertiary Elements on Process Activation and Duration.” In: *The Journal of Physical Chemistry B* 125.4 (Jan. 2021),

pp. 1098–1106. DOI: `10.1021/acs.jpcb.0c09966`. URL: `https://doi.org/10.1021/acs.jpcb.0c09966`.

[49] Aaron M. Fleming et al. "Nanopore Dwell Time Analysis Permits Sequencing and Conformational Assignment of Pseudouridine in SARS-CoV-2." In: *ACS Central Science* 7.10 (Sept. 2021), pp. 1707–1717. DOI: `10.1021/acscentsci.1c00788`. URL: `https://doi.org/10.1021/acscentsci.1c00788`.

[50] Kellie D. Nance and Jordan L. Meier. "Modifications in an Emergency: The Role of N1-Methylpseudouridine in COVID-19 Vaccines." In: *ACS Central Science* 7.5 (Apr. 2021), pp. 748–756. DOI: `10.1021/acscentsci.1c00197`. URL: `https://doi.org/10.1021/acscentsci.1c00197`.

[51] Nicole C. Ammerman, Magda Beier-Sexton, and Abdu F. Azad. "Growth and Maintenance of Vero Cell Lines." In: *Current Protocols in Microbiology* 11.1 (Nov. 2008). DOI: `10.1002/9780471729259.mca04es11`. URL: `https://doi.org/10.1002/9780471729259.mca04es11`.

[52] John M. Crawford et al. "Host-Dependent Modifications of Packaged Alphavirus Genomic RNA Influence Virus Replication in Mammalian Cells." In: *Viruses* 14.12 (Nov. 2022), p. 2606. DOI: `10.3390/v14122606`. URL: `https://doi.org/10.3390/v14122606`.

[53] Sissy Therese Sonnleitner et al. "The mutational dynamics of the SARS-CoV-2 virus in serial passages in vitro." In: *Virologica Sinica* 37.2 (Apr. 2022), pp. 198–207. DOI: `10.1016/j.virs.2022.01.029`. URL: `https://doi.org/10.1016/j.virs.2022.01.029`.

[54] Ilaria Manfredonia et al. "Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements." In: *Nucleic Acids Research* 48.22 (Nov. 2020), pp. 12436–12452. ISSN: 0305-1048. DOI: `10.1093/nar/gkaa1053`. eprint: `https://academic.oup.com/nar/article-pdf/48/22/12436/34905944/gkaa1053.pdf`. URL: `https://doi.org/10.1093/nar/gkaa1053`.

[55] Changchang Cao et al. "The architecture of the SARS-CoV-2 RNA genome inside virion." In: *Nature Communications* 12.1 (June 2021), p. 3917. ISSN: 2041-1723. DOI: `10.1038/s41467-021-22785-x`. URL: `https://doi.org/10.1038/s41467-021-22785-x`.

[56] Tammy C. T. Lan et al. "Secondary structural ensembles of the SARS-CoV-2 RNA genome in infected cells." In: *Nature Communications* 13.1 (Mar. 2022), p. 1128. ISSN: 2041-1723. DOI: `10.1038/s41467-022-28603-2`. URL: `https://doi.org/10.1038/s41467-022-28603-2`.

[57] Ramakanth Madhugiri et al. "Structural and functional conservation of cis-acting RNA elements in coronavirus 5'-terminal genome regions." In: *Virology* 517 (2018). Nidovirus Research, pp. 44–55. ISSN: 0042-6822. DOI: `https://doi.org/10.1016/j.virol.2017.11.025`. URL: `https://www.sciencedirect.com/science/article/pii/S004268221730404X`.

[58] Peter J Walker et al. "Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022)." In: *Archives of virology* 167.11 (Nov. 2022), pp. 2429–2440. ISSN: 0304-8608. DOI: `10.1007/s00705-022-05516-5`. URL: `https://doi.org/10.1007/s00705-022-05516-5`.

[59] Candida S. Punla et al. "Are we there yet?: An analysis of the competencies of BEED graduates of BPSU-DC." In: *International Multidisciplinary Research Journal* 4.3 (Sept. 2022), pp. 50–59.

[60] Shruti Khare et al. "GISAID's role in pandemic response." en. In: *China CDC Wkly* 3.49 (Dec. 2021), pp. 1049–1051.