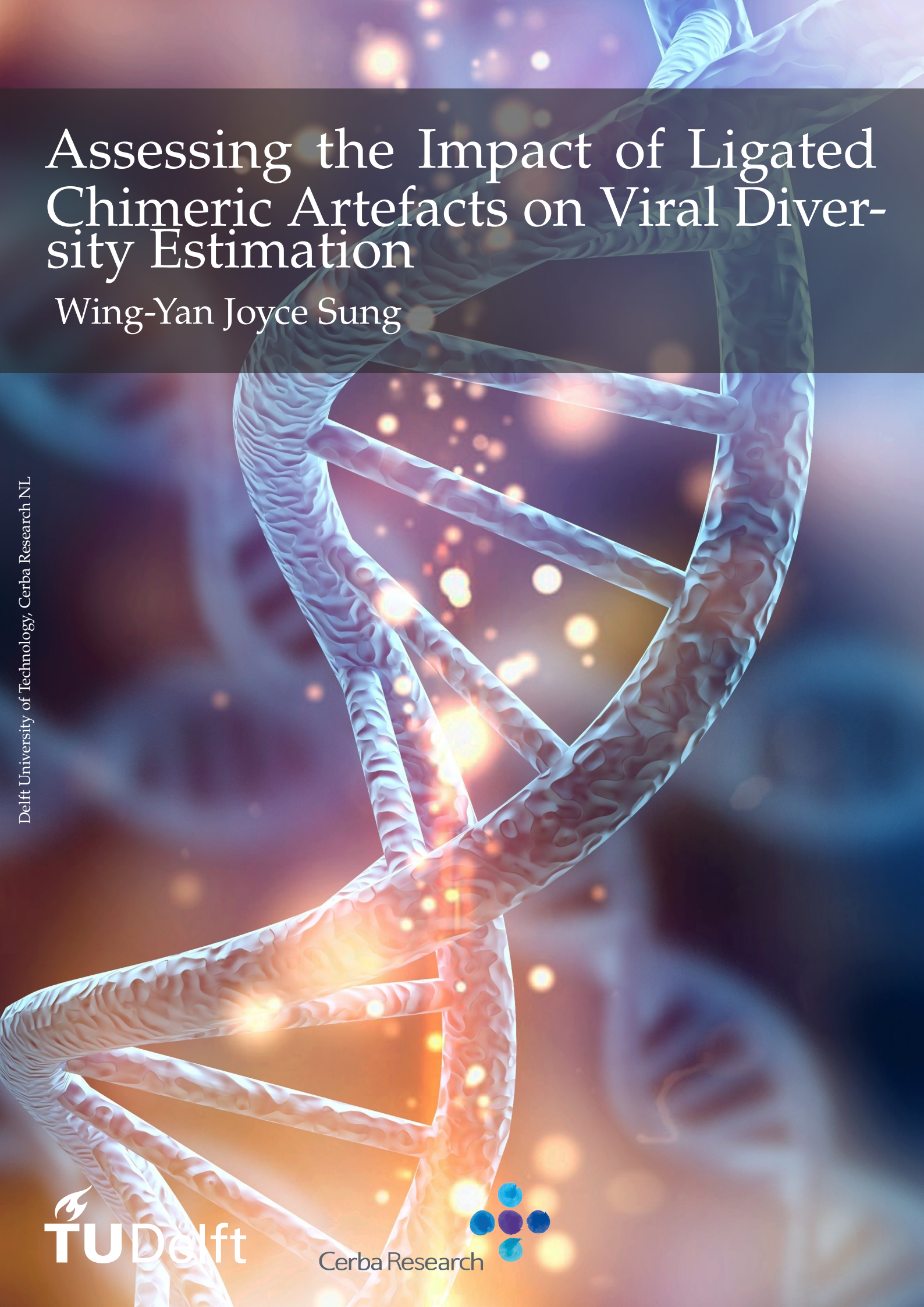
Assessing the Impact of Ligated Chimeric Artefacts on Viral Diversity Estimation

Wing-Yan Joyce Sung

Delft University of Technology, Cerba Research NL



TU Delft

Cerba Research

Assessing the Impact of Ligated Chimeric Artefacts on Viral Diversity Estimation

by

Wing-Yan Joyce Sung

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday July 11, 2025 at 10:00 AM.

Student number:	5011825	
Project duration:	November 11, 2024 – July 11, 2025	
Thesis committee:	Dr. T. E. P. M. F. Abeel,	TU Delft, Thesis Advisor
	Dr. J. A. Baaijens,	TU Delft, Supervisor
	Dr. M. Khosla,	TU Delft
	MSc. M. Weber,	Cerba Research NL

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Cerba Research

Abstract

Reliable estimation of intra-host viral diversity is essential for understanding viral evolution, treatment resistance, and outbreak dynamics. However, technical artefacts introduced during sample preparation and sequencing can distort variant frequencies and lead to incorrect conclusions. One such group of artefacts is ligated chimeric reads, also referred to as ligation chimeras, formed when full-length DNA molecules are erroneously joined during library preparation. Ligation chimeras are currently poorly characterized and their impact on downstream analyses is largely unknown. In this thesis, we developed a modular and reproducible computational pipeline to detect, quantify, and analyze ligated chimeras in amplicon-based viral sequencing datasets. We applied this pipeline to both public and internal datasets, evaluating the prevalence and structural patterns of chimeras and their impact on viral diversity estimates. Our results show that ligated chimeras are widespread, disproportionately affect specific amplicons, and can introduce substantial allele frequency shifts and spurious variants. This means that common filtering strategies in current pipelines risk discarding true low-frequency variants or failing to remove artefactual ones. These findings highlight the importance of chimera-aware preprocessing to ensure accurate viral diversity estimation from long-read sequencing data.

1. Introduction

Fast-mutating RNA viruses pose significant challenges in both clinical and public health settings. Their ability to evolve rapidly allows them to evade host immune responses [1], influence the severity of disease [2], and develop resistance to antiviral treatments [3]. These characteristics complicate efforts to control viral infections and highlight the need to monitor how viruses change over time. Understanding the diversity of viruses within a host can provide crucial insights for designing effective treatments, predicting disease progression, and preventing transmission [4].

The adaptability of RNA viruses stems from their high mutation rates, fast replication, and large population sizes [5]. These factors drive rapid viral evolution and lead to the accumulation of genetic variation within an infected individual. This means that even within a host the virus cannot be described with a single genome, but instead the virus exists as a population of closely related genetic variants [6]. Each unique version of the viral genome in this population is known as a viral haplotype.

Genetic differences between haplotypes can take several forms. The most common are single nucleotide variants (SNVs), which are single-letter changes in the genome sequence [7]. Insertions and deletions (indels) involve the addition or loss of small genome segments (1-50bp). In some cases, larger structural variants (SVs) occur, involving rearrangements or duplications of extended regions of the genome (>50bp). Viruses can also undergo recombination, where segments of genetic material are exchanged between different viral genomes within the same host [8]. These different types of variation all contribute to viral diversity and can influence the biological behavior of the virus.

Viral haplotypes can be studied using sequencing, in which reads are produced. These are sequences of nucleotides derived from fragments of viral genomes. Several sequencing technologies are used to study viral diversity, each with distinct strengths and limitations. Short-read sequencing platforms, such as those developed by Illumina, are widely used due to their low error rates and high throughput [9]. However, their short read length (typically 100–300 bp) limits the ability to detect long-range variation [10]. In contrast, third-generation sequencing platforms such as Oxford Nanopore Technologies (ONT) produce much longer reads, often spanning several kilobases. Although these technologies have higher raw error rates, their extended read lengths make them more suitable for detecting structural variants and recombination events, which are often difficult to resolve with short-read data [11].

Besides sequencing errors, there are many other sources of bias which complicate accurately measuring intra-host viral diversity. Sources of bias include variation in initial virus concentration in the sample [12], differences in sample preparation [13], sequencing errors [14] and computational inferences [15]. One problematic group of technical errors that complicate viral diversity estimation is formed by chimeric artefacts. These artefacts are concatenated sequences which are not native to the sample and can be found in both long and short reads [16, 17].

A distinct subtype of chimeras, which we will refer to as ligation chimeras, was first characterized by White *et al.* [16]. These ligation chimeras consist of entire DNA templates that are concatenated after enrichment but before sequencing. The authors hypothesize that these chimeras are found in both short and long read sequencing techniques, however can be observed more easily in long reads due to their length. Unlike other subtypes of chimeras, which have been more thoroughly studied [8, 17], ligation chimeras remain poorly understood and under-addressed. Most current viral sequencing workflows, either take no specific measures [18–24], or simple length-based filters are applied to remove unexpectedly long reads that may represent chimeric artifacts [25–28]. However, such filtering can reduce sequencing depth and thus sensitivity for detecting low-frequency variants. Moreover, undetected ligation chimeras may introduce false variants or distort variant frequencies, complicating accurate viral diversity assessment. Only a few tools have been proposed to detect and remove ligation chimeras [29, 30] and the impact of ligation chimeras on downstream analyses has not been systematically studied, leaving a gap in current best practices and possibly leading to skewed viral diversity estimates (Figure 1).

Therefore, in this thesis, we investigate the effect of ligated chimeric artefacts on viral diversity estimation in Nanopore sequencing data (Fig. 1). Specifically, we address three research questions:

- How prevalent are ligated chimeric artefacts in viral samples sequenced using Nanopore R10 flow cells?
- What patterns or properties characterize the composition of ligated chimeric reads?
- How do different chimeric read pre-processing strategies mitigate the effect of ligation

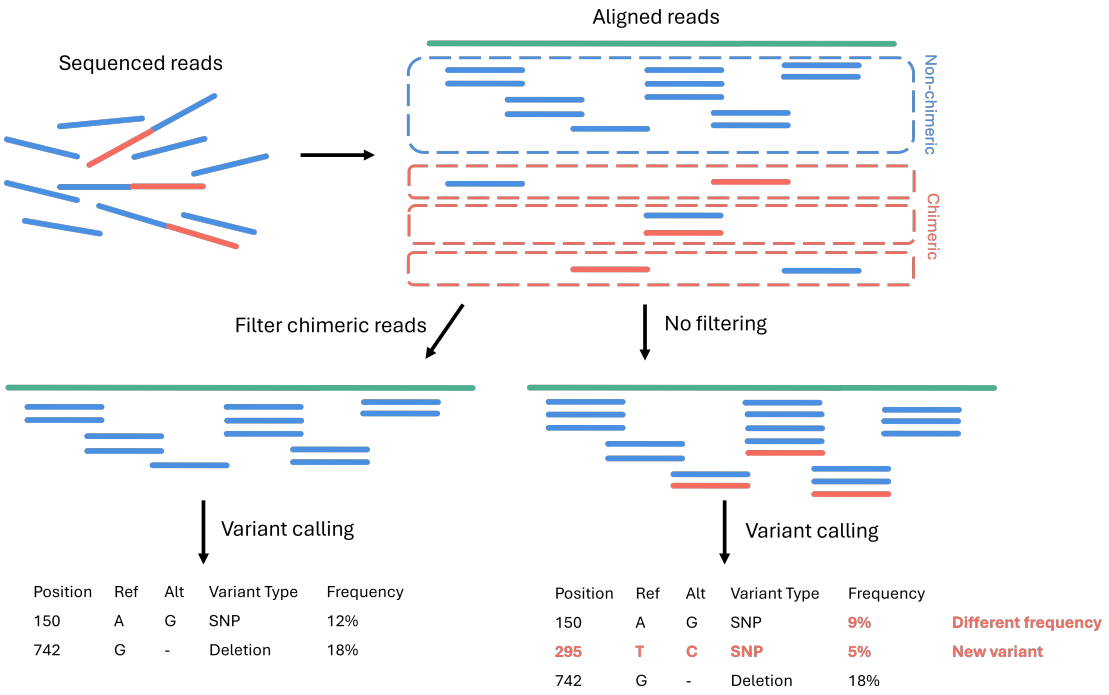


Figure 1: Conceptual overview of how ligated chimeric artefacts may impact variant calling. Sequencing reads are generated from viral genomes; ligated chimeric reads consist of concatenated fragments. Reads can be grouped into non-chimeric (blue) or chimeric (orange) categories, based on alignment. We can attempt variant calling including or excluding chimeric reads. Chimeric reads may introduce spurious variants or distort variant frequencies, potentially biasing diversity estimates. The extent and nature of this effect are investigated in this thesis.

chimeras and influence downstream diversity estimates?

To address these questions, we aim to develop a modular and reproducible computational pipeline. The goal of this framework is to standardize the preprocessing and analysis of long-read viral sequencing data. By automating each step and encapsulating dependencies, the pipeline should support consistent application across datasets and sequencing runs, reducing the risk of manual error and improving reproducibility. Utilizing a general pipeline structure also enables its use for future datasets to assess the number of and impact of ligation chimeras as sequencing platforms, basecallers, and protocols evolve. By implementing a modular structure, steps can be replaced according to users needs or as computational tools improve. As ligation chimeras remain poorly characterized and may persist as a challenge in emerging technologies, having a flexible and reproducible toolset is critical for ensuring reliable viral diversity estimates over time.

We systematically characterize and quantify ligated chimeric reads and their effects in datasets generated using state-of-the-art laboratory workflows. By doing so, we aim to improve the reliability of viral diversity inference. Our findings help inform current and future best practices for data pre-processing in viral genomics, supporting more accurate downstream analyses.

2. Background

Before diving into the specifics of ligation chimeras, we need to understand how we retrieve sequencing data from a viral sample and where in that process different types of artefacts can emerge. From there, we consider the category of chimeric reads in nanopore sequencing, and how ligation chimeras fit into this picture. Then we will take a look at the current landscape of viral diversity estimation, how it's typically approached, and why chimeric reads might interfere with those estimates. Finally, we focus on a specific analysis, called variant calling.

2.1. Sample preparation and genome sequencing

Before a viral sample can be sequenced, the sample needs to undergo several preparation steps. Firstly, the viral genetic material must be isolated, since the original collected sample contains mostly host, often human or animal, genomic material. After isolation and before sequencing, the

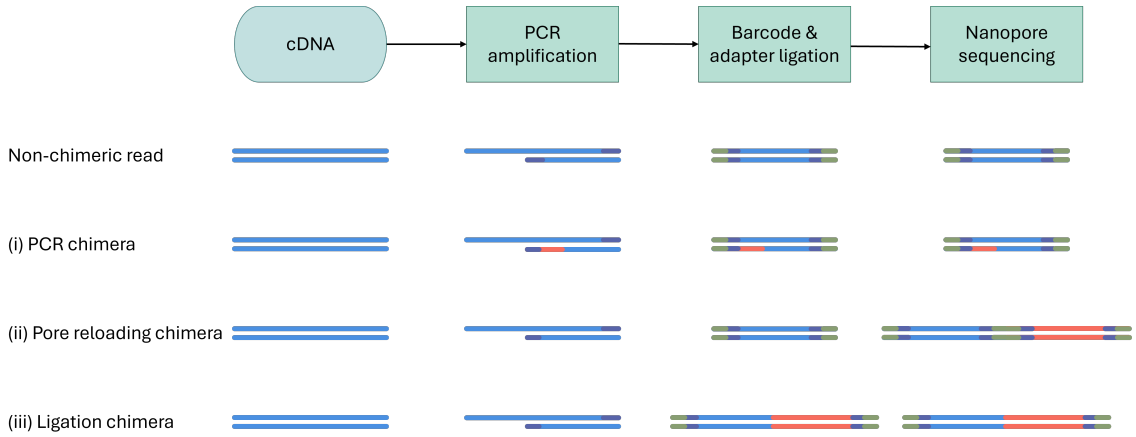


Figure 2: Illustration of the three types of chimeric artefacts in long-read viral sequencing: (i) PCR chimeras formed during amplification, (ii) pore reloading chimeras caused by missed open pore signals between consecutive reads, and (iii) ligation chimeras resulting from the end-to-end joining of distinct DNA molecules prior to sequencing.

sample undergoes library preparation. This is a process in which short DNA sequences called adapters and barcodes are attached to the DNA fragments of interest. These adapters are required for the sequencing device to recognize, capture, and process each fragment. The barcodes can optionally be attached to allow multiple samples to be sequenced together in a single run and later distinguished computationally.

Between isolation and library preparation, amplification can optionally be performed. For viral samples amplification is often necessary to get enough sequencing data, as the concentration of viral material in the sample is often very low [7, 31]. Polymerase chain reaction (PCR) amplification is commonly used; in this process, many copies of a specific DNA region are created. PCR uses short primers, to mark the target sequence, and an enzyme, to build new strands from the original template. Primers are short DNA sequences that match the target DNA. They bind to the DNA, after which the enzyme extends the primer, copying the target DNA sequence. By performing cycles, in which copies of the target sequence, become targets themselves, the number of molecules with the sequence of interest grows exponentially. When sequencing the entire viral genome, primers are usually designed using a tiling amplicon scheme [32–34]. In this scheme, the target regions which are amplified, also called amplicons, are designed such that adjacent regions overlap and the whole genome is covered.

In each of the preparation steps between collection and sequencing of the sample, errors might occur, which can affect downstream analyses of the sequencing data. For instance, during PCR amplification, the enzyme might not copy the sequence correctly at one position during the extension process, imitating a mutation. If this occurs early in the cycling process and the erroneous molecule is amplified in subsequent cycles, the error is copied and passed on to many daughter molecules [35]. This leads to systematic amplification of artefactual variants, which can then appear in the sequencing data at low frequencies. In viral datasets, where estimating intra-host diversity or identifying minor variants is a key goal, these PCR-induced errors can inflate diversity estimates or lead to false detection of low-frequency mutations [36]. Nanopore sequencing itself is prone to basecalling errors, especially in areas with high GC-content or homopolymers, which can particularly affect variant calling [37]. In addition to these errors, chimeric reads, which can be formed during amplification, ligation, or even sequencing, represent another important source of technical noise.

2.2. Chimeric artefacts

The first type of chimeras (Figure 2), PCR chimeras can form during amplification when an incomplete extension product re-anneals to a different template and continues elongating, resulting in hybrid molecules (Fig. 2) [38, 39]. Such chimeras have been shown to falsely link mutations that occur distinct strains [40], leading to incorrect haplotype inference [8, 41]. Several approaches have been developed to detect and mitigate PCR chimeras. Laboratory strategies such as limiting PCR cycles and using high-fidelity enzymes can reduce their formation [35, 42], additionally computational tools can be used to identify and remove them from sequencing data [17]. Although no method guarantees complete removal, the mechanisms by which PCR chimeras form are well

understood, and their effects on downstream analyses, such as haplotype inference or variant calling, have been extensively studied [8]. This allows researchers to interpret results with appropriate caution and apply filtering strategies where necessary.

A different type of chimera can form through rapid pore reloading in nanopore sequencing. These occur when two DNA molecules are sequenced in quick succession. The resulting continuous signal may be interpreted by the base caller as a single read. These artefacts generally contain barcode and adapter sequences in the middle (Fig. 2), making them easier to detect and split. Modern base callers such as Guppy and Dorado include modules to handle these, and additional tools like Porechop [43] or DeepChopper [44] can further refine detection. Although not entirely eliminated, pore reloading chimeras are relatively well-characterised and can be effectively managed with current tools.

A last type of chimeras is formed by ligation chimeras, which are not as well understood. These arise during the library preparation stage, when two or more full-length DNA templates are accidentally ligated together (Fig. 2) [16]. These artefacts do not contain any barcodes or adapters in the middle and can resemble true long reads or biological recombinants, making them harder to detect than pore reloading chimeras. White *et al.* [16] were among the first to report on and characterize these ligation chimeras. Their analysis, based on synthetic constructs and PCR amplification, contained ligation chimeras in 1.7% of reads. Additionally they showed that these artefacts generally form through the ligation of similar amplicons. This process is possibly facilitated by secondary structures or sequence similarity bringing molecule ends into close proximity. They noted that repeated amplicons were overrepresented in chimeric reads, suggesting non-random formation patterns. However, no analysis was done on downstream effects of these artefacts. Similarly Wick *et al.* [45] finds ligation chimeras in 1.4% of reads, and shows that their presence is likely caused by ligation steps of the protocol of the nanopore native barcoding kit, when using the rapid barcoding kit ligation chimeras were found in only 0.1% of reads.

As we focus on ligation chimeras in this thesis, we sometimes refer to ligation chimeras as simply chimeras.

2.3. Current methods for ligation chimeras

Several tools have been proposed to detect and handle ligation chimeras. *yacrd* [29] identifies potential chimeras without relying on a reference genome, based on the assumption that chimeric regions are poorly supported by other reads. However, in amplified viral data, the molecules available for ligation come from a limited pool of amplicons. As a result, junctions between specific amplicons may occur repeatedly, causing yacrd to misinterpret these artefacts as well-supported and fail to flag them as chimeric. The reference-based tool *Liger2Liger* (`https://github.com/rlorigro/Liger2LiGer`) detects disjoint alignments within a read and marks such reads and alignments as chimeric. However, it requires the distance between alignments to exceed 50 kbp—longer than many entire viral genomes, rendering it unsuitable for detecting ligation chimeras in viral samples. The simulator Meta-NanoSim [30] also identifies disjoint alignments within reads and uses them to infer and simulate chimeric structures. Despite these approaches, current methods remain unable to reliably detect ligation chimeras in viral samples.

2.4. Impact of chimeras on viral diversity estimation

The effect of different types of errors on viral diversity analyses depends on the type of analyses. Viral diversity can be estimated at different spatial resolutions: single nucleotide variants (SNVs), local haplotypes and global haplotypes. SNVs indicate a variation in a single position, local haplotypes can be used to describe a set of variations which appear together in a contiguous area, whereas global haplotypes describe the variations which appear together over the whole genome. We will focus on the detection of SNVs, since this is a prerequisite to the reconstruction of haplotypes.

SNVs are detected in viral populations in an analysis called minor variant calling. In this analysis we attempt to find all SNVs, compared to a reference sequence, and the frequency at which it occurs. Most pipelines use general-purpose variant callers such as LoFreq [46], FreeBayes [47], or bcftools, often applying thresholds for allele frequency and read depth. However, these tools are typically designed for diploid or human genomes and may not perform optimally on viral data due to differences in coverage depth, error profiles, and genome organization. Accurate alignment to a high-quality reference is also critical, as reference errors can lead to false positives or missed variants. To address these challenges, virus-specific tools like iVar [48] and VirVarSeq have been

developed to better account for the unique properties of viral datasets, particularly those generated from amplicon-based protocols. Despite these advances, minor variant calling remains sensitive to various technical artefacts, including PCR errors, index hopping, and basecalling inaccuracies, which can introduce spurious low-frequency variants or obscure true signals, underscoring the need for careful filtering and validation strategies.

The potential impact of ligation chimeras on minor variant calling has not been assessed but is likely to be substantial in datasets where they are prevalent. Currently, ligation chimeras are either left unfiltered [18–24] or removed using simple length thresholds [25–28], as they tend to be longer due to the concatenation of two or more amplicons. Filtering based on read length might indiscriminately remove ligated but genuine reads. This risks discarding true biological signals, particularly if ligated chimeras are more likely to form for specific sequences. However, not accounting for chimeras at all could lead to introduction of spurious variants, especially if chimeric reads contain more errors or simply due to misalignments. As a result, both approaches, leaving chimeras unfiltered or filtering by length, can compromise the accuracy of variant detection and interpretation in viral sequencing studies. As of yet, no unified and validated method exists to handle ligation chimeras to ensure accurate variant calling.

Variant calls of a viral sample give an indication of the genetic diversity of the virus within a host. This diversity can reflect evolutionary processes, such as mutation, selection, and genetic drift, and it has implications for disease progression, treatment outcome, and transmission potential [49]. Several metrics exist to quantify viral genetic diversity and can be broadly divided into three categories [49]: (i) richness indices, (ii) abundance-based indices, and (iii) pairwise distance-based indices. These collectively measure the number of observed variants, the distribution of their frequencies, and their genetic distances, respectively. Together these metrics provide a means of comparing datasets processed differently but also help in assessing which handling of chimeric reads yields results that are more consistent with expected biological variability.

In summary, PCR-derived, pore-reloading, and ligation-induced chimeras each introduce different types of bias into Nanopore-based viral diversity estimation. While PCR and pore reloading chimeras have been studied and can be mitigated with existing tools, no methods exist to detect ligation chimeras. Additionally no characterization of the reads and their impact on downstream analysis has been done. Their potential to distort diversity estimates motivates this thesis, which investigates the prevalence, characteristics and the impact of different pre-processing strategies of ligation chimeras on downstream viral diversity analysis.

3. Methods

In order to assess the impact of ligated chimeric artefacts on viral diversity estimation, we create a modular pipeline framework for quantification, characterization and impact assessment (Fig. 3). First we focus on quantification of ligated chimeric artefacts based on alignment. Then we characterize the found chimeric artefacts, by identifying their composite sequences. Lastly, we assess the impact by creating datasets including and excluding the chimeric reads and compare variant calls and resulting viral diversity estimates. We compare these to a proxy ground truth constructed by leveraging known metadata and similarity of replicates in some experiments.

3.1. Reproducibility & Flexibility

To ensure reproducibility and flexibility, we implement all steps of the computational pipeline within a modular and containerized workflow. The pipeline uses Snakemake [50], a workflow management system that enables reproducible and scalable execution of bioinformatics analyses. Each step is defined as an independent rule, allowing users to easily modify, add, or replace components based on experimental needs or tool preferences. We manage software dependencies through a Docker container image, which ensure consistent execution across computing environments and prevent version conflicts. The modular design supports testing of alternative methods, and application to a wide range of datasets.

3.2. Pre-processing of sequencing reads

Before quantification, characterization or impact assessment, raw sequencing reads need to be preprocessed to remove adapters and low-quality sequences. Basecalling and demultiplexing was performed before reads enter the pipeline using Dorado or Guppy (Table 1). In the first step of the

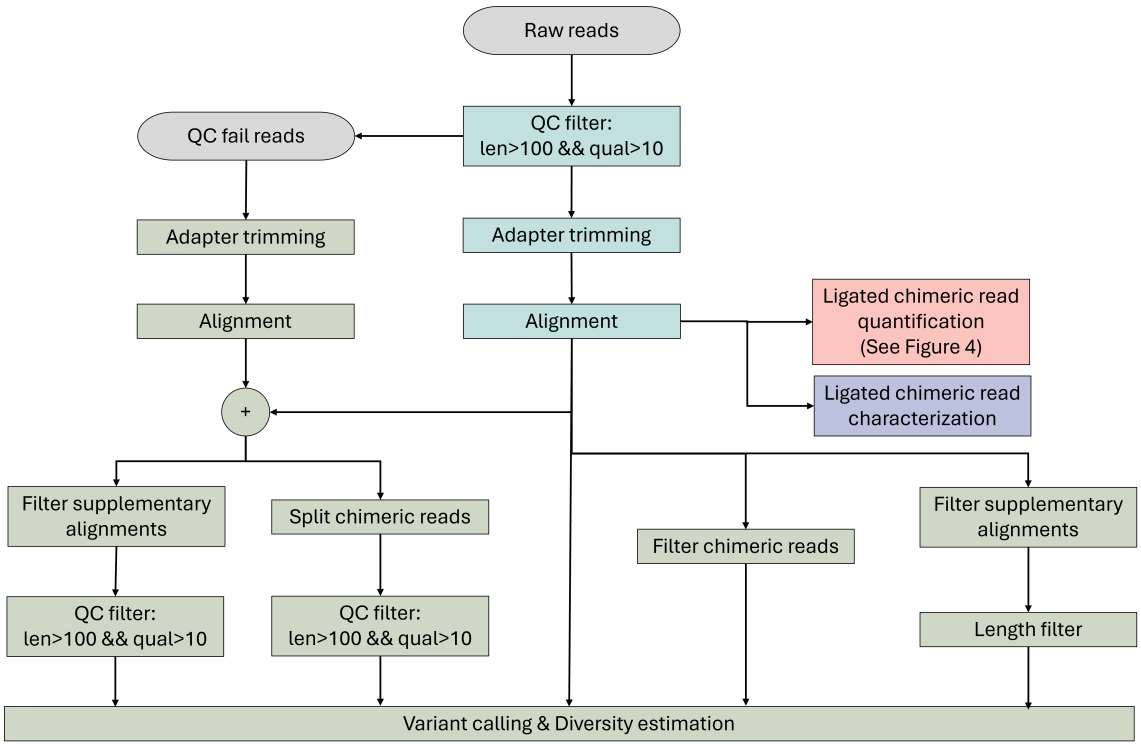


Figure 3: An overview of the computational pipeline. All samples first undergo pre-processing, which includes quality and length filtering, adapter trimming and alignment and pictured in teal. The alignment can be for different downstream analyses, including ligated chimeric reads quantification, characterization and impact assessment of ligated chimeric reads on viral diversity estimates. For impact assessment, some of the reads not passing the quality filter might be reconsidered. Impact of ligated chimeric reads is assessed by applying different pre-procesing methods for ligation chimeras to the alignment and estimating viral diversity with each dataset, these steps are coloured in green.

pipeline, we remove reads shorter than a 100 base pairs or with a mean quality score lower than 10. Then, we perform adapter trimming with Porechop [43], which also removes any remaining barcodes, which might have been missed during demultiplexing. By additionally splitting on middle adapters and barcodes, any pore reloading chimeras remaining after base calling are split into multiple reads in this step. We then align reads with minimap2 [51] using the recommended nanopore preset. The reads are aligned to the reference sequence which was used to identify the primer positions for the samples.

3.3. Quantification of ligated chimeric artefacts

We identify ligated chimeric reads based on alignment structure. Reads containing supplementary alignments, indicated by the "SA" tag in the BAM file, are categorized as ligated chimeric reads. Each chimeric read has a primary alignment and one or more supplementary alignments. Each supplementary alignment represents a non-contiguous mapping to the reference genome. We consider all these reads with supplementary alignments ligation chimeras and not any other type of chimera. Since we assume all pore reloading chimeras have already been split by the basecaller or Porechop and we expect PCR chimeras to align contiguously [52].

Next we assess whether chimeric reads occur at the same rate for every amplicon. For this we compared for each amplicon, the number of chimeric reads which contained the amplicon, to the number of reads non-chimeric reads containing the amplicon. We need to normalize with the non-chimeric reads per amplicon, as not all amplicons are amplified equally [52]. To determine the number of successfully amplified reads per amplicon, we first filter the non-chimeric reads, to exclude reads which have soft clips larger than a 100 bp. These soft clips are not expected in successfully amplified amplicons (Figure 4). We then categorize the remaining reads based on their alignment positions.

An alignment is considered to correspond to an amplicon if its start lies within 20 base pairs of the forward primer's start, and its end lies within 20 base pairs of the reverse primer's end. We use a position-based approach contrary to a sequence-based approach, due to the error rate of

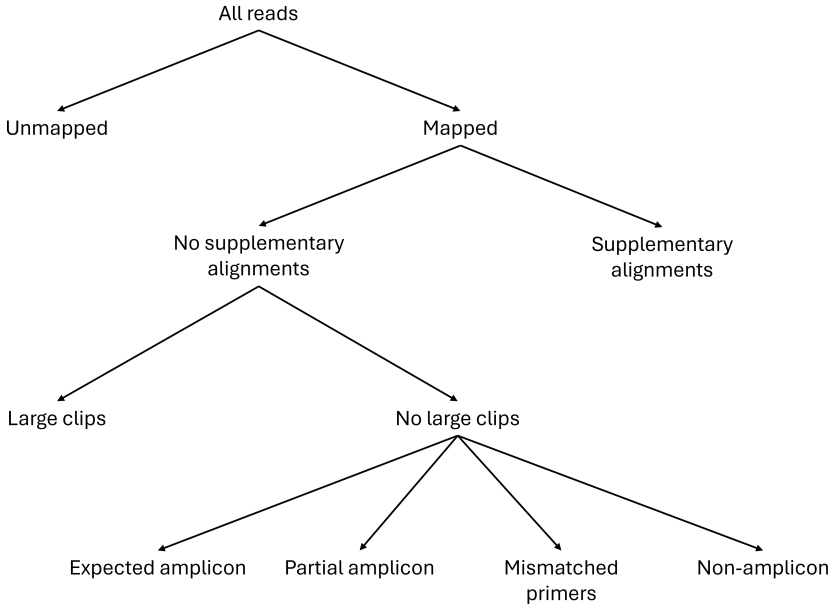


Figure 4: Read categorization based on alignment structure and position. In alignment mapped and unmapped reads are separated. We then categorize all mapped reads which contain supplementary alignments as chimeric. The non-chimeric reads can then be further categorized to get a detailed picture of the sample composition.

the nanopore sequencer and the decline in quality at read ends. The 20 base pair threshold was selected to accommodate typical alignment variability and soft-clipping observed near primer regions in long-read sequencing data, allowing for robust identification of reads corresponding to expected amplicons.

A segment is considered a partial amplicon when its' alignment start and end position are within the positions of a primer pair, but do not pass the proximity threshold. When a read has mismatched primers, the start of the alignment is in range of a forward primer and the end is in range of a reverse primer, however, the forward and reverse primers are not for the same amplicon. As this should only be possible within a PCR pool, we only categorize these type of reads when the mismatched primers are from the same pool. Lastly, reads for which the alignment start and end do not lie within the positions of a primer pair and does not correspond to a pair of mismatched primers, are categorized as non-amplicon reads.

3.4. Characterization of ligated chimeric artefacts

In the characterization section of the pipeline, we focus on the composition of the chimeric reads. Specifically, the number of segments the read is composed of and what patterns can be found in the combination of segments. To characterize the reads, we first count the number of alignments in each read. Then for each alignment in the read we determine from which amplicon it originates using the same method as described in Section 3.3.

To test whether certain combinations of amplicons occur significantly more often than expected, we analyzed pairwise combinations between subsequent segments within each chimeric read. This analysis assumes that ligation occurs randomly and uniformly across all available amplicons, meaning all pairwise combinations are equally likely. We first create an observed ligation matrix O, where each entry $O_{i,j}$ represents the number of times amplicons i and j were observed ligated in adjacent positions within chimeric reads, including partial amplicons.

Then to evaluate statistical significance, we compute the expected number of ligations between amplicons i and j under the null model of random ligation:

$$E_{i,j} = \begin{cases} \frac{a_i \cdot a_j}{\sum_{k=1}^{N} \sum_{l=1}^{k-1} a_k \cdot a_l + \sum_k^N \binom{a_k}{2}} \cdot T & \text{if } i \neq j \\ \frac{\binom{a_i}{2}}{\sum_{k=1}^{N} \sum_{l=1}^{k-1} a_k \cdot a_l + \sum_k^N \binom{a_k}{2}} \cdot T & \text{if } i = j \end{cases} \tag{1}$$

where a_i and a_j denote the number of observed occurrences of amplicons i and j in chimeric and non-chimeric reads, and $T = \frac{1}{2} \cdot \sum_i^N \sum_j^N O_{i,j}$ is the total number of chimeric events in the dataset.

We then determine significant under- or over-representation of amplicon pairs using the Z-statistic:

$$Z_{i,j} = \frac{O_{i,j} - E_{i,j}}{\sqrt{E_{i,j} + \varepsilon}} \tag{2}$$

where ε is a small constant added for numerical stability.

Two-sided p-values are derived from the standard normal distribution:

$$p_{i,j} = 2 \cdot (1 - \Phi(|Z_{i,j}|)) \tag{3}$$

To account for multiple hypothesis testing across all pairwise comparisons, p-values are adjusted using the Benjamini–Hochberg correction to control the false discovery rate (FDR). Amplicon pairs with adjusted p-values below 0.05 are considered statistically significant. Pairs with $Z_{i,j} > 0$ are comparatively over-represented, and those with $Z_{i,j} < 0$are comparatively under-represented relative to the null model.

To ensure statistical reliability, comparisons for which the expected number of ligations $E_{i,j}$ is below 5 are masked and excluded from significance testing and multiple testing correction. This threshold is a standard heuristic used to ensure that the normal approximation underlying the Z-statistic remains valid, as low expected counts can lead to unstable variance estimates and inflated test statistics.

3.5. Datasets

In this thesis, we make use of seven datasets containing viral nanopore reads from public and non-public sources to perform quantification and characterization. Public datasets enable us to investigate the size and generality of chimeric reads, while internal datasets enable us to perform more fine-grained analysis, as more sample information is available.

The internal datasets include the hMPV-A, hMPV-B, RSV-A and RSV-B datasets and were provided by Cerba Research NL in BAM format as generated by the basecaller. Samples include labstrains in dilution series, as well as clinical samples. The read files were converted to fastq format using bedtools bamtofastq.

Public datasets were retrieved from the NCBI database in fastq format. The ZIKV [53] and two public RSV [54] datasets are available under the accession numbers PRJNA1035959 and PRJNA1257940, respectively. The RSV-A and -B datasets were combined in the BioProject. To categorize a sample as RSV-A or RSV-B, we subsampled a 1000 reads and aligned them to both a RSV-A and RSV-B reference genome. The reference to which more reads aligned was the subtype the sample was categorized as.

Virus	# Samples	Avg amplicon length (bp)	Basecaller	Nanopore sequencing kit
hMPV-A	22	866	Dorado	SQK-NBD114-24
hMPV-B	24	868	Dorado	SQK-NBD114-24
RSV-A	32	972	Dorado	SQK-NBD114-24
RSV-B	44	905	Dorado	SQK-NBD114-24
ZIKV	36	454	Guppy	SQK-NBD112-96
RSV-A public	52	840	Guppy	SQK-NBD114-96
RSV-B public	37	507	Guppy	SQK-NBD114-96

Table 1: Summary of nanopore sequencing parameters for different viral samples, including the number of samples, average amplicon length, basecaller used and sequencing kit used.

For the non-public data, first RNA was isolated from the collected sample, after which reverse transcription was performed. The material was then amplified in two multiplex PCR reaction pools amplifying non-overlapping fragments in each pool. After amplification, library preparation was performed according to the standard workflow for Native Barcoding Kit 24 V14. Sequencing was performed on a R10.4.1 flowcell and basecalled and demultiplexed using Dorado.

3.5.1 Simulated datasets

In order to validate the behaviour of our pipeline we simulate a dataset with all identified possible chimeras. First we sample chimera-free reads from an existing dataset, where we sample 4 types of reads for each amplicon:

1. The whole amplicon in forward direction
2. The whole amplicon in reverse direction
3. Part of the amplicon in forward direction
4. Part of the amplicon in the reverse direction

These reads are categorized as described in Section 3.3, with here additionally utilizing the alignment orientation. From the chimera-free reads we create all possible chimeras in the dataset by concatenating any two reads in any order. We then perform quantification as described in 3.3, in which all reads should be categorized as chimeras.

3.6. Viral diversity estimation

To assess the impact of chimeric reads on viral diversity estimation, we first derive five versions of the dataset, each differing in how chimeric reads were handled (Figure 3). We then perform variant calling on each dataset. From the variant calls we can quantify the diversity of the sample using several viral diversity estimate metrics. To assess which dataset provides the most accurate estimate we use replicates to approximate a ground truth, which we can compare the diversity estimates of each dataset against.

3.6.1 Dataset creation

The datasets are created after read alignment, to decrease computational costs of repeated pre-processing (Figure 3). The first dataset is the original dataset, with chimeric reads included as-is. In the second dataset, chimeric reads are split into multiple reads, each corresponding to one of their alignments. The third dataset retains only the primary alignment from each chimeric read. The fourth dataset excludes all chimeric reads entirely. The last dataset contains solely primary alignments and is filtered on length, based on the designed amplicon sizes. To create the datasets (ii), split chimeric reads, and (iii), with only primary alignments of chimeric reads, we first align any reads which have been filtered out in the QC step in pre-processing. Any chimeric reads from these are added to the preliminary dataset (Figure 3), as some of the alignments in the reads might have an higher average read quality compared to the read as a whole. For dataset (ii), for each supplementary read we create an unique read name by adding a suffix to its original read name and change its flag to the primary alignment flag "0x0". For the primary alignment, we cut down the read to the aligned part. This is to avoid supplementary alignments possibly counting twice, as they are present in the primary alignment although soft clipped. Then to create the dataset (iii), we filter all alignments with the supplementary read flag "0x2048" out. Lastly we again filter on minimum length of 100 bp and a minimal quality of 10 for preliminary datasets (ii) and (iii) to get the final datasets.

To create dataset (iv) we filter on chimeric reads, by filtering out any reads which have the "SA" tag. Lastly for dataset (v) we first filter out all supplementary alignments using the alignment flags. We then apply a length filter to exclude sequencing reads that are either too long or too short compared to the expected range of amplicon sizes. Specifically, we remove any reads that are longer than the maximum amplicon size or shorter than the minimum amplicon size. To allow for minor variations, we incorporate a margin equal to 10% of the average amplicon size when determining the acceptable length range.

Before performing any viral diversity analyses, we trim primers for all datasets using iVar [48]. We then perform variant calling on each dataset using iVar [48]. Using a minimum frequency threshold of 3% and at least 30 supporting reads, in order to exclude sequencing errors.

3.6.2 Viral diversity estimate metrics

To quantify how different chimeric read handling strategies influence viral diversity estimation, we compute several established diversity indices on each derived dataset. These metrics aim to capture different aspects of the intra-host viral population structure and are crucial for evaluating the biological fidelity of variant calls. We adopt one metric of each of three categories of diversity indices: (i) richness indices, (ii) abundance-based indices, and (iii) pairwise distance-based indices. These collectively measure the number of observed variants, the distribution of their frequencies, and their genetic distances, respectively.

For the richness index, we compute the number of single nucleotide variants (SNVs) passing our quality thresholds across the viral genome. This basic richness index serves as a measure of the

mutational landscape complexity but is sensitive to sequencing depth and minor variant detection thresholds.

Additionally for more detailed insight, abundance-based indices, such as the Shannon entropy, can be calculated. The Shannon entropy $\bar{H}$ provides a measure of the uncertainty or unpredictability in nucleotide composition at each site and is independent of the number of SNVs called. Thus representing diversity on the average position, unlike the richness index which is an indication of the diversity of the entire genome. Higher entropy values indicate greater diversity. The Shannon entropy of a sample is calculated as:

$$\bar{H} = \frac{1}{L} \sum_{i=1}^{L} \sum_{a \in \{A,C,G,T\}} -p_{i,a} \log_2 p_{i,a} \tag{4}$$

Where $p_{i,a}$ is the proportion of nucleotide a at site i and L is the total number of sites in the genome or region analyzed.

Lastly, we can compute pairwise distance-based indices, which account for genetic similarity between strains in the sample. Diversity is estimated by comparing all strains to the most abundant strain in the population. This gives us the population nucleotide diversity π, which is defined as the average number of nucleotide differences per site between all possible pairs of reads. Since, unlike entropy-based measures, π accounts for the genetic distance between sequences, it offers a more robust estimate less affected by sequencing depth variability. The population nucleotide diversity can be calculated as:

$$\pi = \frac{1}{L} \frac{2}{n(n-1)} \sum_{i=1}^{L} \sum_{\alpha \neq \beta} n_{i,\alpha} n_{i,\beta} \tag{5}$$

with L as the length of the sequence and $n_{i,\alpha}$ the number of reads supporting nucleotide $\alpha \in \{A, C, G, T\}$ at site i.

When comparing diversity metrics across datasets, higher richness or entropy may indicate that more minor variants are picked up using certain datasets. However, this could also indicate the presence of more noise due to sequencing artifacts, including unaccounted chimeric reads, in the dataset. Conversely, lower diversity might suggest under-detection or over-stringent filtering. Together, these metrics provide a complementary set of metrics to evaluate the impact of chimeric read handling on within-host viral diversity estimation. Employing multiple metrics allows for a more nuanced interpretation of diversity patterns than any single measure alone [49].

3.6.3 Ground truth

To assess which of the datasets leads to the most accurate variant calls and viral diversity estimates, we need to compare the datasets to the ground truth. However, this is not available for any of the datasets used, therefore, we use replicate samples to approximate the ground truth.

In our internal datasets replicates of labstrain samples are available for all datasets. The replicates are taken from the same original sample, the cultured lab strain, and then worked up in a separate PCR and individually barcoded. For the adapter ligation samples on the same sequencing runs are pooled, after which they are sequenced in multiplex. In total the labstrains are sequenced 18 times using 3 different dilutions. This means we have 6 samples per dilution, divided equally over 2 different sequencing runs. As we expect the most accurate results from the high viral load, we use the six replicates from the high viral load to approximate the ground truth.

Using the six high viral load replicates, we perform variant calling on each of the samples separately using the unfiltered dataset. Any SNV which is present in at least half of the replicates is included in our proxy ground truth. The allele frequency in the ground truth set is calculated as the average of the allele frequencies over all sets in which the particular SNV was called.

To evaluate the accuracy of variant calling across different datasets, we employ three commonly used metrics: precision, recall, and F1-score. These metrics provide a quantitative framework for comparing the called SNVs against the proxy ground truth. Precision reflects the proportion of SNVs called by a dataset that are also present in the ground truth, indicating the rate of false positives. Recall measures the proportion of SNVs in the ground truth that are successfully identified by a dataset, highlighting the rate of false negatives. The F1-score represents the harmonic mean of precision and recall, offering a single summary statistic that balances both error types. High precision suggests reliable calls with few false positives, while high recall indicates

comprehensive detection of true variants. A high F1-score requires both a high precision and high recall. These metrics are described by below equations

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Where TP = true positives, FP = false positives, and FN = false negatives.

Additionally we compare the viral diversity estimate metrics described in Section 3.6.2 between the ground truth and all readsets to determine which preprocessing method most faithfully preserves the true viral population structure.

4. Results

4.1. Supplementary alignments are present in 80% of ligation chimeras

In order to determine whether our pipeline can successfully quantify the number of chimeric reads in a dataset, we validate its performance using a simulated set created as described in Section 3.5.1. For this we use two samples from the internal hMPV-A assay to sample chimera-free reads from, of which one sample clinical and one from a cultured labstrain. Additionally we use two samples from the public ZIKV assay, as this uses much shorter amplicons than the other assays tested.

Our pipeline finds 80% of possible chimeras in the dataset for all tested samples (Table 2). Of the simulated chimeras classified as non-chimeric, approximately half of the reads contain large soft clips. This means that in alignment one segment is effectively clipped off. The great majority of the remaining miscategorized reads contain large indels, less than 0.1% did not contain a large indel.

Sample	% Large indels	% Large clips	% Supplementary aligned
hMPV-A (labstrain)	10.8%	10.0%	79.2%
hMPV-A (clinical)	8.4%	9.8%	81.8%
ZIKV (sample 1)	7.4%	13.0%	80.0%
ZIKV (sample 2)	6.4%	13.6%	80.1%

Table 2: For each sample, the table shows the percentage of the simulated chimeras in each alignment structure. We consider three types (i) reads which are primary aligned and contain a large indel, (ii) reads which are primary aligned and contain a large clip and (iii) reads with a supplementary alignment.

We find equal amounts of chimeras for which the segments have the same orientation as opposite orientation (Figure 5). In Figure 5 we can clearly see that when segments are very far apart, the read is aligned with a supplementary alignment. However, when segments are closer together on the genome, they may be aligned supplementary, one of the segments might be clipped of or they might be primary aligned with a large indel. The alignment structure is likely dependent on the size and quality of matches in the alignment and the precise mapping parameters used in minimap2.

Our pipeline processes a sample of 2 million reads (500–800 bp) in approximately 3–4 hours using 4 CPU cores. This includes chimeric read quantification, characterization, and viral diversity estimation across five pre-processing strategies. Runtime can likely be reduced to under an hour with increased parallelization, making the approach feasible for clinical applications—especially given that this reflects a relatively large dataset. In future work, overall runtime may be further improved by replacing the current adapter trimming tool, which accounts for the majority of processing time.

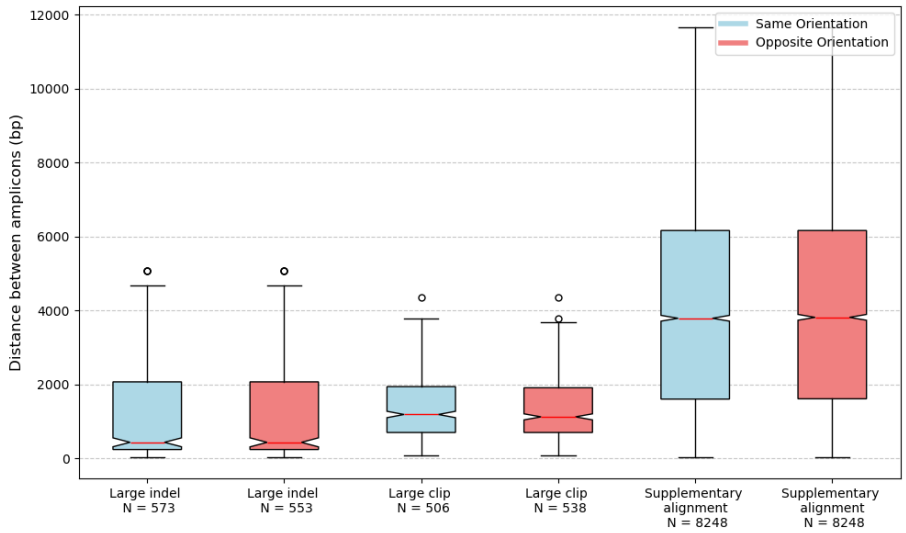


Figure 5: Theoretic genomic distance for the simulated chimeric reads, grouped by alignment structure. Each box shows the variability in the genomic distance between the two segments the simulated chimera consists of. Boxes represent three different alignment structures, (i) reads which are primary aligned and contain a large indel, (ii) reads which are primary aligned and contain a large clip and (iii) reads with a supplementary alignment. Of these groups (i) and (ii) would be considered non-chimeric and group (iii) would be considered chimeric. The groups are further divided into two categories based on the orientation of the segments in the chimera, these can align in the same orientation or opposite orientations. For each box the number of reads in the category is annotated in the label.

4.2. Ligated chimeric artefacts are prevalent in all datasets

To assess the prevalence of ligated chimeric artefacts in the most recent ONT chemistry, we quantify the fraction of chimeras in both internal and public datasets, which were sequenced using the R10 flowcell, using the methodology described in Section 3.3. Chimeric reads were detected in all datasets, indicating that ligation artefacts are a general feature of amplicon-based nanopore sequencing (Figure 6). In internal datasets, total chimeric fractions ranged from approximately 25% to over 40%. Public datasets showed substantially lower average total chimera fractions, between 1.3% and 6%, but still consistently contained measurable levels of chimeric reads. The differences in the number of chimeric reads found between different datasets may reflect variation in the virus sequenced, the assay protocol, assay execution or basecalling settings.

While there are significant differences in the number of chimeric reads found for different viruses, these differences are also present for datasets which study the same virus but use different assays. With the public RSV-A and RSV-B containing 1.8% and 1.4% respectively, while internal datasets contain 25.6% and 28.2% chimeric reads. This difference is much larger than the difference in chimeric fractions between the internal hMPV-A and RSV-A assay, 28.5% and 25.6% respectively, which use the same assay design, apart from amplicon scheme which is virus dependent. This suggests assay execution and design has a larger impact than the studied virus on the number of chimeric reads formed.

Assay design can vary in multiple ways, however, as we focus on ligation chimeras, two important design choices lie in the chosen amplicon scheme and the library preparation protocol. The amplicon scheme dictates the diversity and length of the DNA molecules present in the sample after amplification, which may influence the rate at which chimeric molecules form. To study the effect of the chosen amplicon scheme on the ligation chimera formation rate, different amplicon schemes should be compared over the same samples. White *et al.* [16] and Wick *et al.* [45] show that difference in the library preparation protocol impact the number of chimeric reads. Wick *et al.* [45] found a smaller number of chimeras when using the rapid barcoding kit (0.03% and 0.14%) from nanopore compared to the ligation barcoding kit (1.41% and 0.88%). The authors hypothesize that the choice of a barcoding kit lacking a ligation step, such as the rapid barcoding kit, might reduce chimera formation. The native barcoding kit, which was used for all studied samples, contains a DNA ligation step. Notably, however, the public RSV assays replace a DNA ligation step from the standard nanopore protocol in the library preparation. While this might explain the differences in the fraction of chimeric reads found, it is confounded by differences in chosen amplicon scheme and assay execution.

The relatively large fraction of chimeric reads found in most data contrasts with earlier

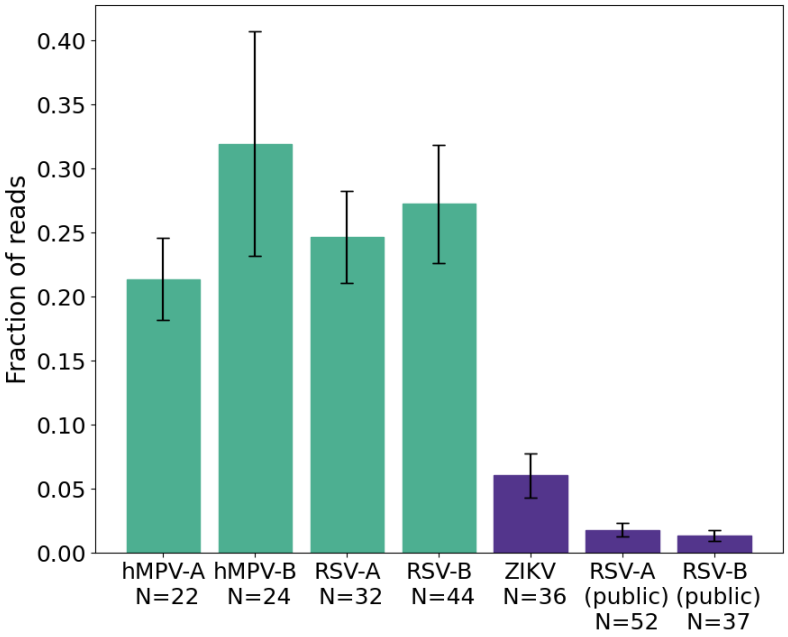


Figure 6: Average fraction of chimeric reads per sample for each assay. Sample counts per assay are indicated below each bar. Internal datasets are pictured in teal and public datasets in purple.

observations, where much lower prevalence of 1.7%-3.2% of chimeric reads in their datasets was reported [16, 30, 45, 55, 56]. We find 10x chimeric reads in internal datasets and 2x more chimeric reads in the public ZIKV dataset. Only the public RSV assays contain chimeric fractions corresponding to fractions expected from literature. This suggest that ligation chimeras may be substantially underreported in existing literature, and their potential to distort variant frequencies or mimic recombination events has likely been underestimated.

4.3. Some amplicons are more prone to chimera formation

Ligated chimeric reads were found in varying numbers across all datasets, we next assessed whether this burden was uniformly distributed across amplicons to investigate whether chimera formation is influenced by amplicon-specific properties. Additionally if some amplicons are disproportionally affected, accuracy of downstream analysis might be especially compromised in these regions.

To assess the uniformity of the chimeric burden across amplicons, we quantified the proportion of chimeric reads per amplicon, including whole and partial amplicon segments. In Figure 7 can be seen that across the internal RSV-B dataset, some amplicons consistently exhibited low chimeric fractions, while others were heavily affected. In some cases, the sequencing depth of the amplicon attributed to chimeric reads can be twice as high compared to the average amount of sequencing depth attributed to chimeric reads across all amplicons. Such patterns can be seen across all analyzed datasets (Supplementary Material A). This analysis makes clear that certain amplicons are disproportionately impacted, suggesting that chimera formation is influenced by amplicon-specific properties.

4.4. Partial amplicons are overrepresented in chimeric reads

Having quantified the prevalence of chimeric reads, we next examine their internal composition by assessing the number of alignments per chimeric read across datasets. By studying the internal composition more closely, we might gain insights into the formation of these chimeras. Additionally, a more detailed overview of the composition of the chimeric reads can enhance future detection.

The majority of chimeric reads contain exactly two alignments, while the frequency of reads with more than two segments decreases exponentially (Figure 8). This pattern is consistently observed across all viral datasets, regardless of sample origin or sequencing depth. These findings suggest that most ligation-derived chimeras result from the fusion of two DNA molecules, while more complex chimeras consisting of more than two segments are comparatively rare.

We then investigate whether the alignment segments within a chimeric read correspond to known amplicons. While many alignments within chimeric reads could be matched to expected

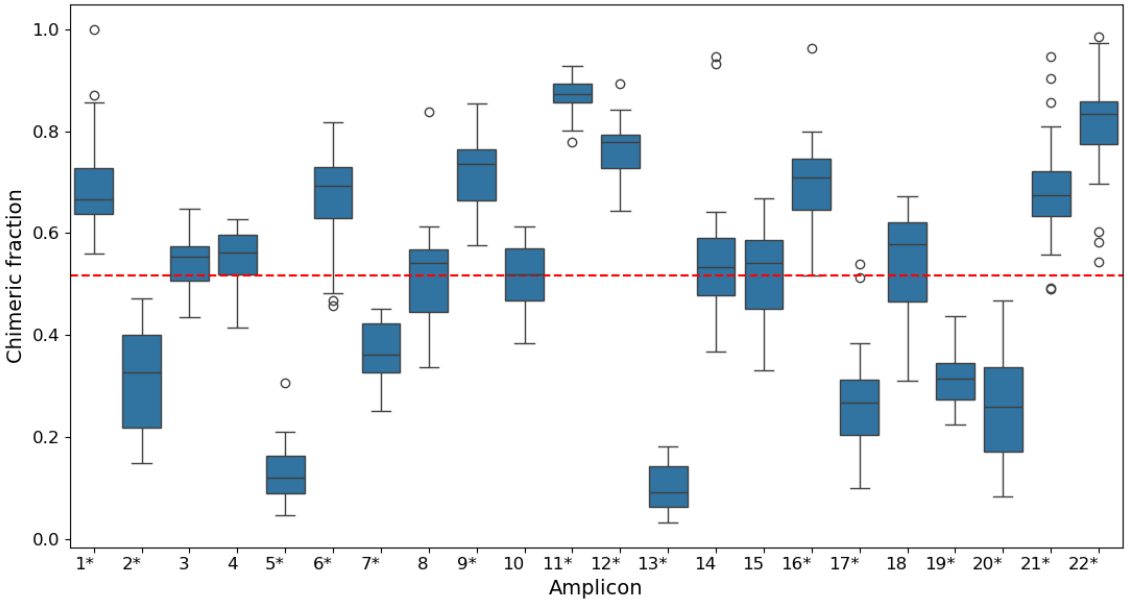


Figure 7: Chimeric fraction per amplicon in the internal RSV-B dataset. The red dotted line represents the average chimeric fraction across all amplicons. Amplicons that are significantly more or less affected are marked with an asterisk (*) in the label. Significance was determined using the Mann–Whitney U test, comparing each amplicon to the overall distribution.

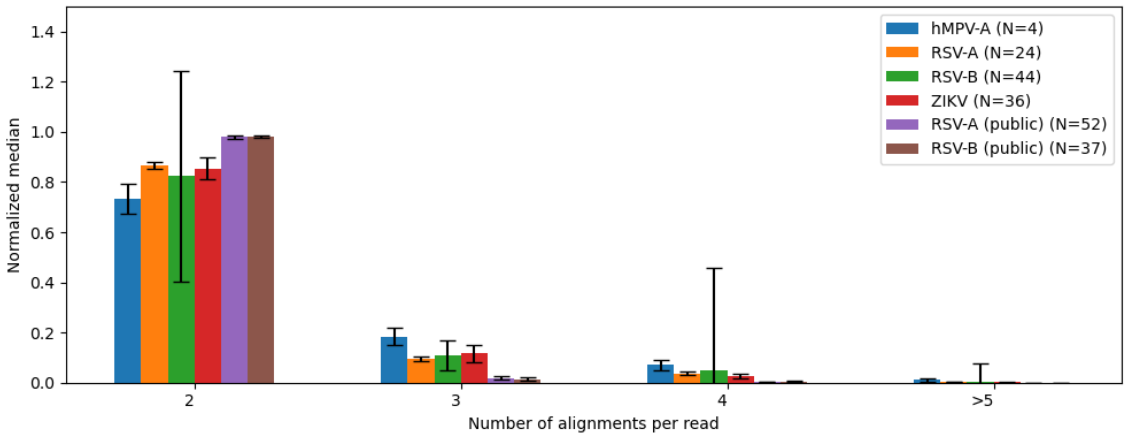


Figure 8: Average number of segments in supplementary alignments. Bar heights represent the normalized median fraction of chimeric reads with different numbers of alignments within the read, aggregated across all samples within each dataset.

Virus	Total segments	Fraction whole amplicon	Fraction partial amplicon	Fraction non-amplicon
hMPV-A	153,702	0.475	0.493	0.010
hMPV-B	226,060	0.402	0.366	0.230
RSV-A	359,505	0.481	0.473	0.044
RSV-B	472,163	0.413	0.535	0.040
ZIKV	11,659	0.496	0.397	0.096
RSV-A (public)	9,593	0.205	0.732	0.035
RSV-B (public)	4,138	0.427	0.508	0.057

Table 3: For each virus, the table shows the total number of alignment segments found in chimeric reads, the fraction of these segments that could be assigned to a whole or amplicon based on reference primer positions.

Virus	Chimeric			Non-chimeric		
	%	Average length	Proximity primer	%	Average length	Proximity primer
hMPV-A	43.6	336	2	11.8	551	2
hMPV-B	40.1	307	3	12.3	490	2
RSV-A	51.2	386	2	19.9	620	2
RSV-B	59.4	380	2	29.1	763	2
ZIKV	44.4	245	2	24.3	339	5
RSV-A (public)	76.8	380	2	8.5	602	1
RSV-B (public)	52.2	286	2	6.5	394	1

Table 4: For each virus, the table shows the average total number of partial amplicon segments found in an sample, from chimeric as well as non-chimeric reads. As well as how much of the chimeric and non-chimeric reads are made up out of the partial amplicons and a characterization based on length and alignment position. For the alignment position we evaluate the median distance to the closest primer across all partial amplicons.

amplicons, many segments correspond to partial amplicons (Table 3). A large portion of these partial amplicons start or end at a primer region, but are shorter than the expected amplicons. Compared to the proportion of partial amplicons in non-chimeric reads, these partial amplicons are overrepresented in chimeric reads (Table 4).

Some of the instances of partial amplicons might be attributed to misalignment. However, in the validation sets only 1.3% of whole amplicons were miscategorized as partial amplicons. Thus misalignment is not expected to be the cause for the majority of the partial amplicons as this is comparatively rare.

An alternative potential explanation for this overrepresentation of partial amplicons is that these identified ligation chimeras which contain partial amplicons are actually not ligated chimeric artefacts at all. These could be biological structural variants or recombinants, or possibly PCR chimeras which do not align contiguously [57]. However, across all labstrains a similarly large fraction of these partial amplicons was found as in clinical samples. While biological structural variants or recombinants cannot be ruled out entirely as labstrains are cultured before sequencing, this type of biological variation is not expected in labstrains [58]. Additionally, biological structural events are comparatively rare and do not fully explain the large fraction of partial amplicons in chimeric reads. For non-contiguously aligning PCR chimeras, we expect a pattern in where the partial amplicons start or end, which do not correspond to primer position, while a few of these are visible, the majority of these do not fall within such a pattern (Supplement B).

It is unclear what is the cause of the overrepresentation of partial amplicons in ligation chimeras. One remaining possibility is that partial amplicons are more prone to chimera formation. Especially shorter fragments of partial amplicons, as the average size of partial amplicons found in chimeras is much smaller (Table 4) These partial amplicon fragments have been known to arise in PCR and nanopore sequencing. These fragments might be caused by early stopping of PCR, occurring more often in certain spots due to secondary structures or problematic regions [59]. Fragments might also be caused by early termination of sequencing by the pore. Since chimeras are longer molecules, Brejová *et al.* [26] found that longer templates are more often fragmented or fail to produce reads covering the entire target. Alternatively, ligation chimera formation might create partial amplicons, if ligation chimeras are created in secondary structures. White *et al.* [16] hypothesizes ligation chimeras might be formed by DNA molecules with similar ends forming secondary structures, eventually resulting in chimeric reads.

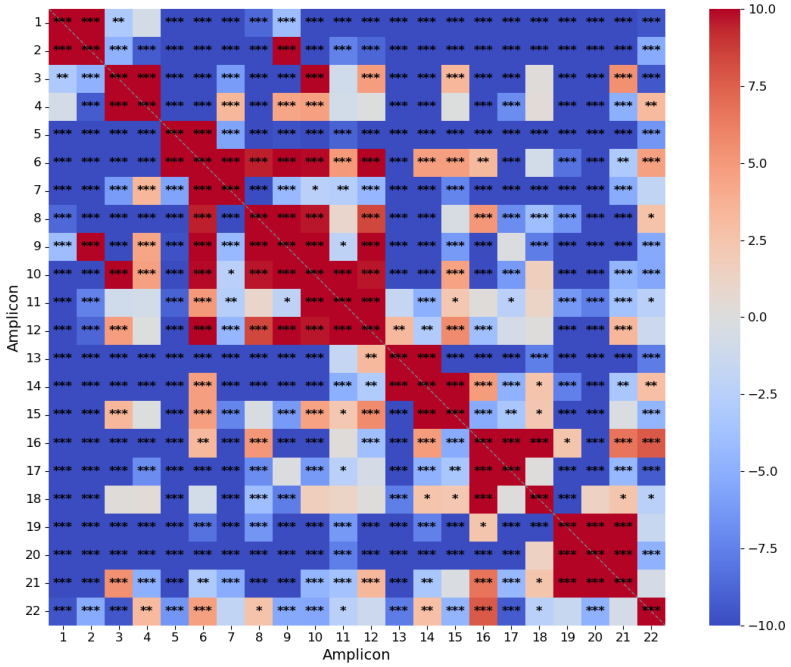


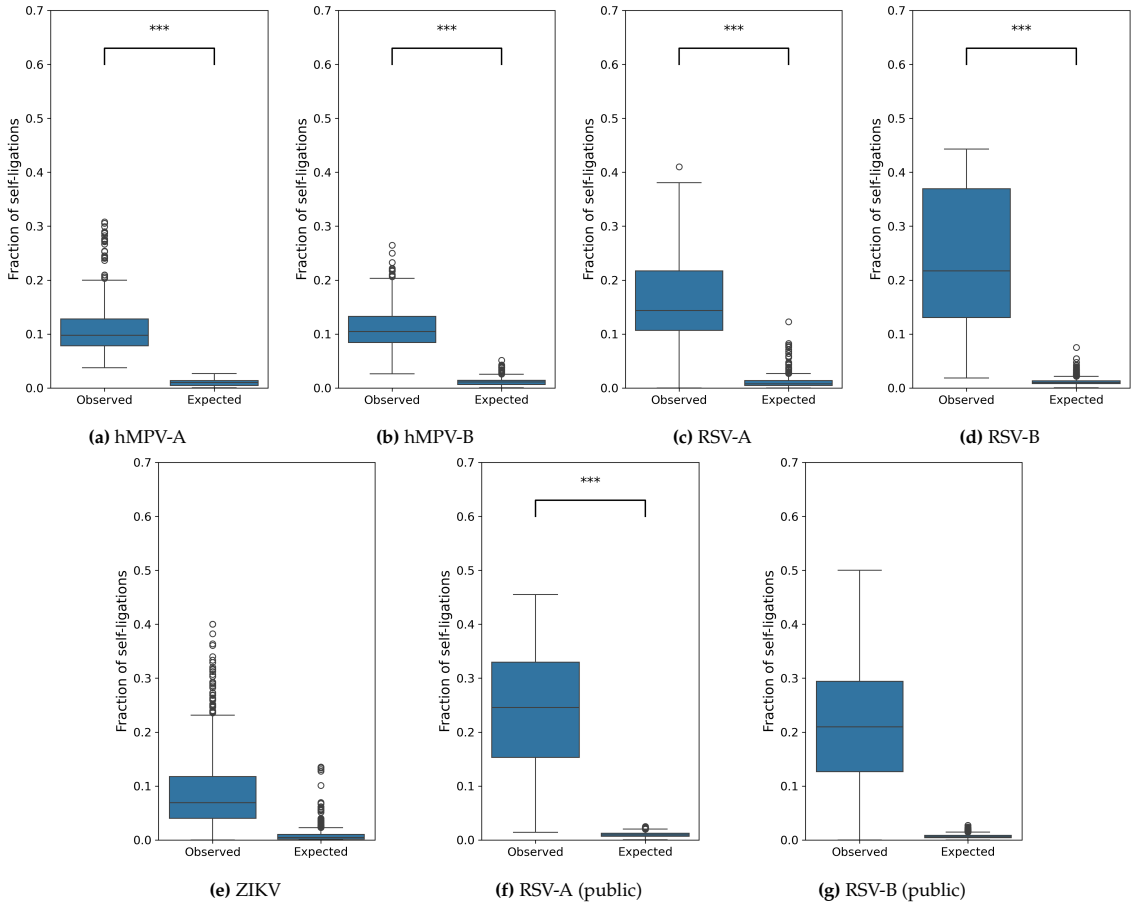
Figure 9: Chimeric events for an internal RSV-B sample. Each cell represent a chimeric event between a segment from the amplicon on the row and a segment from the amplicon on the column. Cells are colored by their over- or underrepresentation compared to the null hypothesis that ligation is completely random. Asterisks indicate statistical significance: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

4.5. Ligated chimeric artefacts formation is not random

To characterize the composition of chimeric reads in even more detail, we assessed how often combinations of certain amplicons appeared within chimeric reads, taking into account whole and partial amplicons. We compare the frequency of observed ligations to expected ligations under a model of random ligation, based on amplicon availability in the sample as described in Section 3.4. In many samples we find a high proportion of significantly over- or under-represented ligation events relative to the null model, which suggests that ligation does not occur randomly based solely on amplicon abundance (Figure 9). This indicates that biological or technical factors systematically influence ligation frequencies.

First we investigate self-ligations, these are chimeric events formed using segments from the same amplicon. The expected number of self-ligations per amplicon were calculated as described in Section 3.4, summing over all amplicons yields the total number of expected self-ligations per sample. Significant differences between observed and expected self-ligation fractions are evaluated using the Wilcoxon signed-rank test, a non-parametric test appropriate for paired data without assuming normality. To ensure statistical reliability, only amplicons with an expected self-ligation count above a defined threshold are included in the analysis. This approach minimizes the influence of low-count events, which may otherwise lead to inflated or unstable fraction estimates. We find significantly more self-ligations than expected under a model of random ligation across all assays, but the public RSV-B and ZIKV datasets (Figure 10). The lack of significance in the public RSV-B and ZIKV datasets likely reflects higher variability and sample heterogeneity, rather than an absence of increased self-ligation.

This corresponds with the findings from White *et al.* [16], the authors similarly found that self-ligation was more common than other ligations. They found a majority of 75% of repeated amplicons and hypothesizes this could be caused by DNA molecules with similar ends forming secondary structures which might result in chimeric reads. While repeated amplicons are overrepresented compared to other combinations of amplicons in our assays, self-ligations do not represent the majority of all chimeric reads. This difference in the number of self-ligations is likely attributed to experimental design between the assays studied in this thesis compared to White *et al.* [16]. In the experiment of White *et al.* [16] each sample only contained one amplicon, meaning each amplicon was barcoded separately. In contrast, for all viral samples studied in this thesis all amplicons of one viral sample are barcoded at once. Our results suggest that even in the presence of other ligation candidates, repeated segments are still more likely to form.



(a) hMPV-A **(b)** hMPV-B **(c)** RSV-A **(d)** RSV-B

(e) ZIKV **(f)** RSV-A (public) **(g)** RSV-B (public)

Figure 10: Observed versus expected fraction of self-ligations across different assays. Each subplot shows the distribution of self-ligation fractions in experimental data ("Observed") compared to a random ligation model ("Expected"). Asterisks indicate statistical significance: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

Aside from repeated segments, many samples contained other combinations of amplicons which were overrepresented in the model of random ligation. To assess reproducibility of these ligation patterns, we computed pairwise correlations between Z-score matrices after excluding diagonal entries, which represent self-ligation events. These are consistently overrepresented across all samples and could artificially inflate similarity metrics. By focusing on off-diagonal interactions, we specifically test for reproducibility in ligation patterns beyond self-ligation. We find high pairwise correlation between Z-score matrices across all samples, with a median pairwise Pearson correlation of at least 0.88 for all assays, except for the ZIKV assay which had a median pairwise Pearson correlation of 0.60. The strong similarity in ligation patterns across samples suggests that chimera formation is not random, but likely driven by systematic biases, including amplicon abundance, sequence features, or consistent technical factors in sample preparation. These findings indicate that chimeras arise in a reproducible and protocol-dependent manner.

4.6. Inclusion of ligation chimeras improves SNV recall

To assess the impact of ligated chimeric artefacts on variant calling, we call SNVs using the different pre-processed datasets described in Section 3.6.1 and compare to a proxy ground truth for all internal datasets.

The unfiltered and split chimera sets achieved significantly higher recall compared to the length filtered and no chimeras sets (Figure 11, Supplement C). This indicates more true positives SNVs were called in the readsets containing chimeras. This increase in sensitivity likely arises because chimeric reads contribute additional support to low-frequency variants that may otherwise fall below allele frequency or read depth thresholds. There are no significant differences, nor a clear trend in precision between datasets (Figure 11, Supplement C). In some assays precision is increased when including chimeras, while for other assays precision decreases when including chimeras. This indicates that inclusion of chimeras can sometimes also lead to increased false positive calls. The significantly improved recall and small differences in precision result in significantly higher

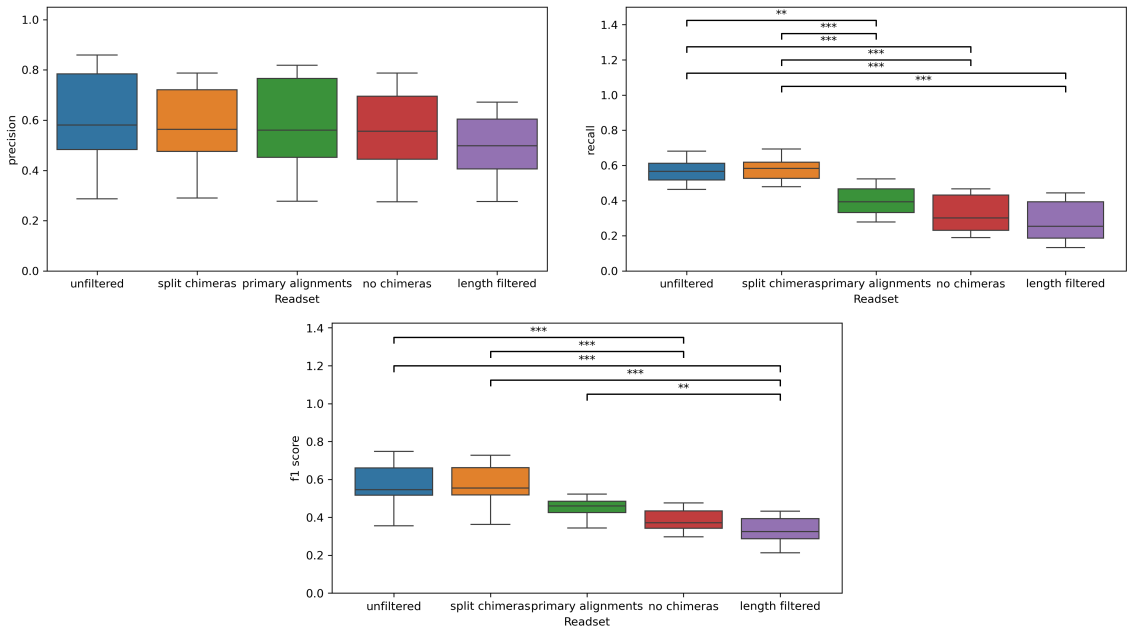


Figure 11: Precision, recall and F1-score of variant calls of replicate samples of the internal RSV-B labstrain. Datasets are compared which have undergone different filtering strategies for chimeric reads. Statistically significant differences between datasets are indicated with brackets; asterisks indicate significance levels: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***). Results for additional assays are provided in Supplement C

F1-scores for readsets that included chimeras (Figure 11).

Across all assays we find no significant differences in the performance of the split chimeras and unfiltered dataset. Similarly, no significant differences are found in the performance of the length filtered and no chimeras datasets. The dataset containing only the primary alignments of the chimeric reads generally performs worse than the datasets containing chimeras, and better than the datasets containing no chimeras (Figure 11, Supplement C).

A key factor contributing to these differences appears to be sequencing depth. Length filtering, which removes long reads that are likely to be chimeric, leads to a substantial reduction in sequencing depth. For the internal datasets, sequencing depth often decreased to less than half of that observed in unfiltered datasets, an example can be seen in Figure 12. This reduced sequencing depth limits the statistical power to detect true low-frequency variants and likely contributes to the decreased recall in these datasets.

While inclusion of chimeric reads increases performance, the pre-processing method which should be chosen is still dependent on the goal of the analysis. Chimeric reads should be included if the goal is to uncover new variants as this increases recall greatly. However, if it is essential no false positives are called, exclusion of chimeric reads might still be more suitable, despite the decrease in overall performance. The results of the datasets excluding chimeric reads also vary less among each other and thus might also be chosen if reproducibility is very important.

4.7. Removal of chimeric reads leads to lower diversity estimates

Finally, we assess the impact of ligation chimeras on viral diversity estimates. We use the metrics described in Section 3.6.2 to compare the diversity estimates between datasets.

For the richness estimate, we see that generally in datasets with chimeric reads more SNVs are called (Figure 13, Supplement D). These calls include true positives, as seen by the increase in recall and can improve the richness estimate (Figure 13a,22,23). However, in some cases these additional variant calls also appear include false positives, which can lead to overestimating the richness of the sample as seen for the internal hMPV-A dataset (Figure 21). Removal of all chimeras, as done in the no chimeras and length filtered datasets, leads to underestimation of the richness of the sample. However, using the datasets excluding chimeras does lead to a more consistent viral diversity estimate across labstrains, suggesting these datasets might be more robust to noise.

When additionally including clinical samples, we observe fewer significant differences in richness index (Figure 13, 21, 22, 23) and for public datasets no significant differences are found. However, in the ZIKV dataset the general trend of higher richness in datasets containing the

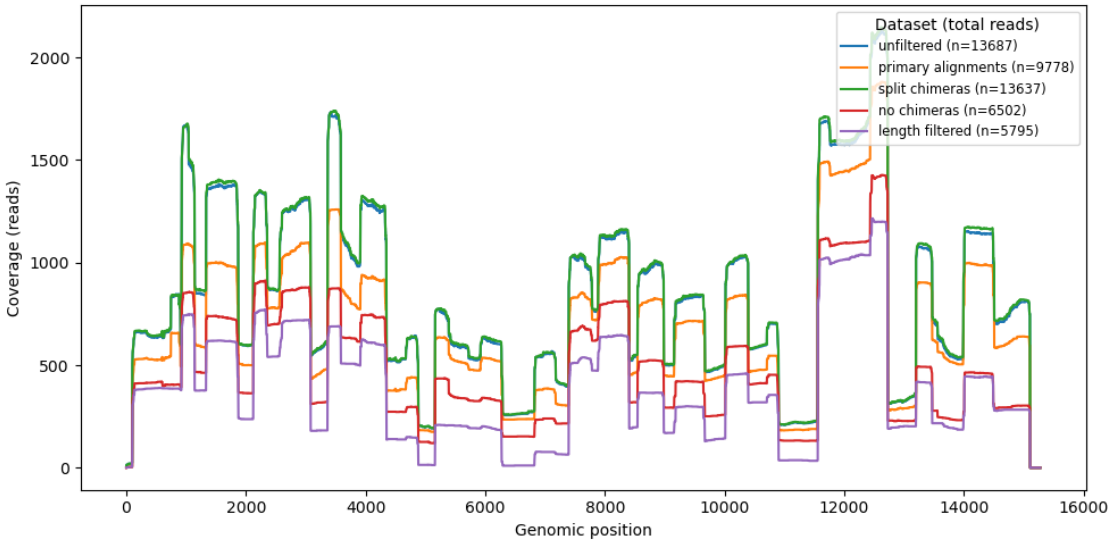


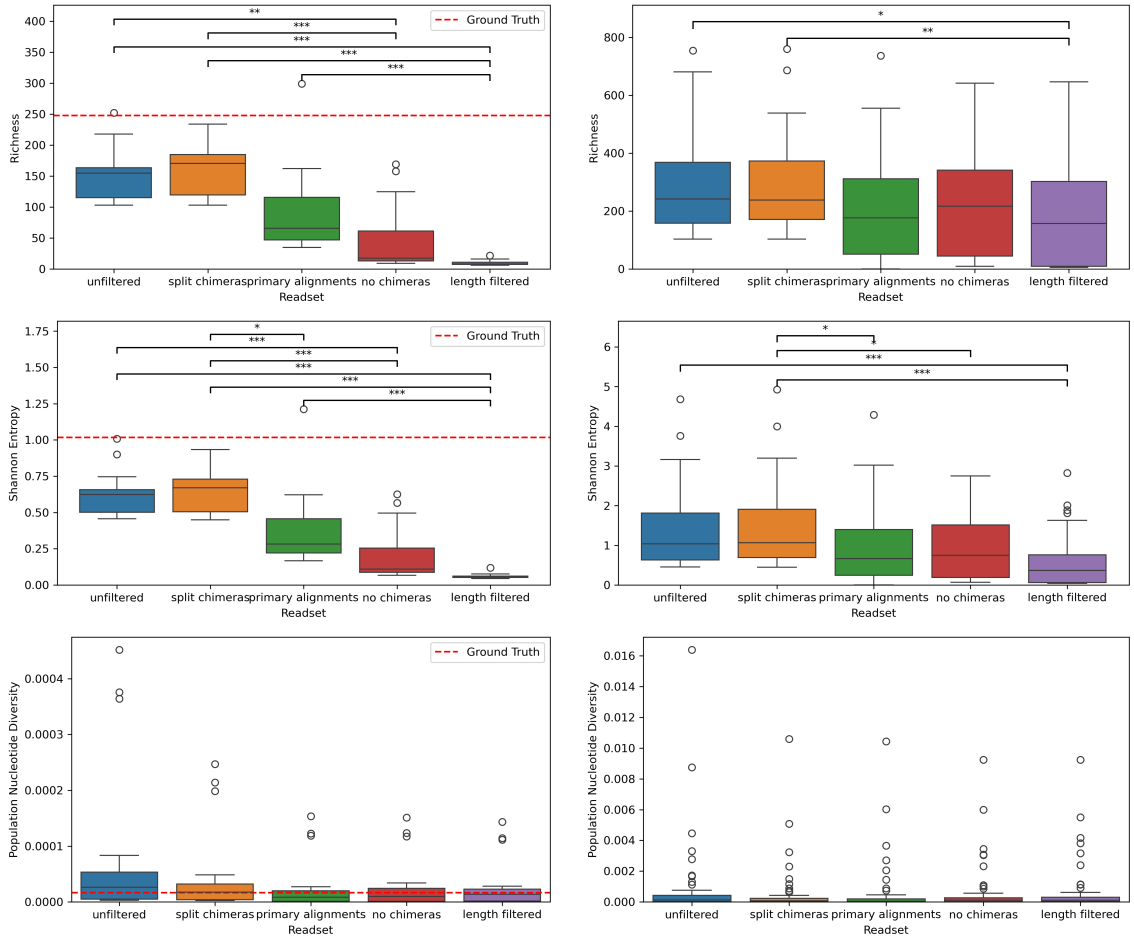
Figure 12: Sequencing depth across the genome for a RSV-B sample. A comparison across filtering approaches. Total alignment counts per dataset are shown in the legend.

chimeric reads seems to be present (Figure 14). The differences in impact on the richness index when additionally including clinical samples is likely due in part to greater variation in richness between the samples. In the public datasets, no significant differences are observed, which may be more strongly attributed to the overall lower fraction of chimeric reads. Since the datasets differ primarily in how chimeric reads are handled. As a result, the differences between datasets are inherently smaller, and filtering strategies have less impact on the observed richness. Thus, for the ZIKV assay, which contains 6% chimeric reads, the same general trend can be observed when comparing dataset. For the public RSV-A and RSV-B assay, which contain less than 2% chimeric reads, the richness index appears almost identical across the datasets.

The abundance-based index, the Shannon entropy, is also increased for readsets including chimeras. This indicates that not only more variants are called, but the frequencies of minor variants might also be increased when calling with the support of chimeric reads. For all internal datasets inclusion of chimeric reads results in an Shannon entropy estimate closer to the proxy ground truth estimate (Figure 13, 21, 22, 23). In contrast to the richness index, significant differences in Shannon entropy remain for most assays when also comparing with clinical samples.

The population nucleotide diversity is similar across most datasets, this implies that the found pairwise distance between strains is not significantly affected by the inclusion or exclusion of chimeric reads. This suggests that while chimeras lead to the detection of more SNVs and may increase the frequency of some minor alleles, they do not drastically alter the overall average number of nucleotide differences per site. In other words, the additional variants introduced or boosted by chimeric reads are either too low in frequency or too evenly distributed to significantly shift the population-level pairwise diversity. It is also possible that many of these additional SNVs occur at sites already contributing to diversity, thus having a limited net effect on the genome-wide estimate of the population-level pairwise diversity.

However, in two of the assays a lower population nucleotide diversity in datasets only including primary alignments of chimeric reads is found, compared to both datasets including and excluding chimeric reads fully (Figure 21, 22). This might be an artefact of reference bias. In minimap2 the alignment with the highest alignment score, i.e. the most similar to the reference genome, will be marked as the primary alignment and all other alignments will be marked supplementary [51]. As the original labstrain was very similar to the reference used, it can be assumed that the majority strain is the most similar to the reference genome. Thus by only including primary alignments, additional support is provided for the majority strain, but not for the minority strains. This results in a lower population nucleotide diversity estimate compared to datasets including entire chimeric reads, which provide additional support for the minority strains as well as the majority strain. When comparing to datasets which do not include chimeric reads at all, if the primary alignments largely support the majority strain, the majority strain is comparatively overrepresented, resulting in a lower population nucleotide diversity estimate. This effect remains when the clinical samples are additionally considered (Figure 21, 22), however, the hMPV assays contain relatively few clinical



(a) Replicate labstrain samples, the viral diversity estimate of the ground truth is indicated with the red dotted line.

(b) Replicate labstrain and clinical samples

Figure 13: Viral diversity estimate metrics of the internal RSV-B assay using different filtering strategies for chimeric reads. Statistically significant differences between datasets are indicated with brackets; asterisks indicate significance levels: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***). Results for additional assays are provided in Supplement C

samples, 4 and 6, compared to the 18 labstrains.

Overall our results show that the inclusion of ligation chimeras impacts diversity estimates, largely through increased sequencing depth. However, we also observe significantly higher Shannon entropy in datasets that include chimeric reads, indicating a broader effect on the variant frequency distribution beyond simple changes in read depth. These effects are more pronounced in datasets with higher chimeric read fractions.

Both datasets which include chimeric reads, unfiltered and split chimera datasets, estimate similar viral diversity. Although both the length filtered and no chimeras dataset exclude chimeras, use of the length filtered dataset generally results in lower diversity estimates.

5. Discussion

We have developed a pipeline to quantify and characterize ligated chimeric reads in a dataset based on alignment structure and position. As well as study the impact of these ligation chimeras on viral diversity estimates. Our results provide compelling evidence that ligated chimeric artefacts are a consistent and non-trivial feature of Nanopore amplicon sequencing across diverse viral samples. By characterizing their prevalence and structure, we find that chimera formation is likely dependent on amplicon-specific features. By studying the influence of using chimeric reads for variant calling, we show that these artefacts can significantly alter viral diversity estimates, when included compared to excluded. However, the current work also presents several limitations that open up important avenues for future research.

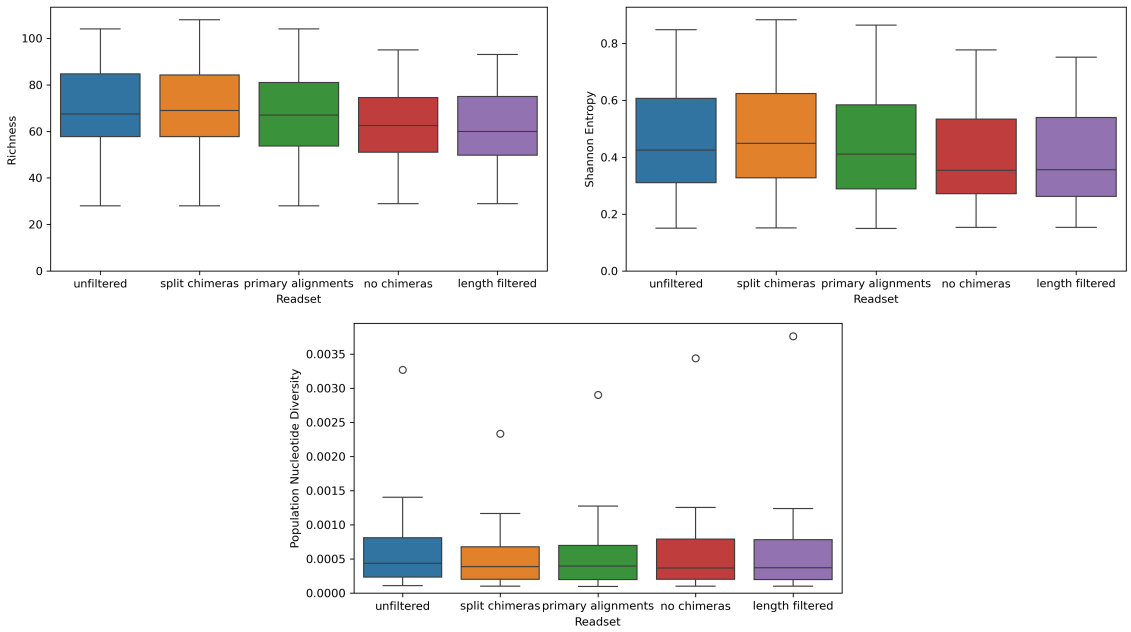


Figure 14: Viral diversity estimate metrics of the public ZIKV assay using different filtering strategies for chimeric reads. Statistically significant differences between datasets are indicated with brackets; asterisks indicate significance levels: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

5.1. Current detection of ligated chimeric reads is limited

Approximately 80% of our simulated chimeric reads could be found with our strategy of using supplementary alignments. Performance might be improved by optimizing the parameters used for mapping. For instance, more stringent gap penalties or increased sensitivity could shift reads currently aligned with large indels or soft clips into the supplementary alignment category. This could reduce misclassification and help better capture the true number of chimeras.

In our current framework, performance could be improved by post-processsing large indels, by checking whether breakpoints correspond to known amplicon junctions. While post-processing large indels is feasible, we chose not to implement this step in the current framework. Correct implementation of this approach requires suitable thresholds for what would be considered a large indel and what an appropriate margin to a primer position would be. These considerations are out of scope for the current project, but could be considered in future research to improve detection of ligation chimeras. An alternative strategy could be to map to a multifasta, where each reference is an individual amplicon, this might reveal more simplified read alignment behaviors that can be tackled more straightforward downstream.

By focusing on supplementary alignments instead, we prioritize a generalizable method that captures the majority of chimeric reads without requiring additional logic or risking overfitting to specific artefact patterns. However, it is important to note that this strategy inherently underestimates the total number of chimeras. Future refinement of alignment and classification criteria may help address these limitations and provide a more comprehensive quantification of chimeric reads.

5.2. Further experiments are needed to draw overall conclusions on the presence of ligation chimeras

We find that ligated chimeric reads are prevalent across all studied viral datasets. To draw overall conclusions on the presence of ligated chimeric artefacts on a larger scale, further experiments are needed. We have made the pipeline publicly available, enabling other researchers to apply it to broader datasets and assess the prevalence and characteristics of ligated chimeric reads across diverse viral sequencing efforts.

The number of ligation chimeras found, varied greatly between assays. Specifically, the internal datasets contained far more chimeras than the public datasets. To investigate the cause of this large difference, confounding factors need to be separated. Factors which might influence the number of ligated chimeric reads include the initial sample composition and the assay design and execution. Comparing the number of ligation chimeras found in samples which originate from the

same original sample but are processed by different laboratories performing the same protocol, can give more insight into the effect of assay execution. The effect of assay design can be studied in more detail by testing various assays on the same samples. Specifically, different amplicon tiling strategies can be tested, as well as different adapter ligation conditions or even kits [54, 60].

Furthermore this analysis can be extended to a broader range of samples and assays considered in this thesis. It is unclear to what extent ligation chimeras are specific to the amplicon-based library preparation protocols used in this study. Our data are derived from multiplex PCR schemes using many PCR cycles. It would be informative to assess the presence of similar artefacts in datasets generated using random hexamer priming, capture-based enrichment, or direct RNA sequencing. These alternative strategies may introduce less bias during ligation or may reveal different artefact profiles altogether. A systematic comparison of these methods could inform best practices for viral sequencing, particularly in clinical or metagenomic settings.

Additionally, although this study focuses on Nanopore sequencing, ligation chimeras may not be exclusive to this platform. In Illumina amplicon datasets, PCR chimeras are well documented [35, 52], but ligation-induced artefacts have not been reported. However, White *et al.* [16] theorizes that ligated chimeric artefacts do form during library preparation for Illumina sequencing, but are difficult to detect due to the short read length. Ligation chimeras might explain index switching in Illumina reads. If present, such artefacts could also influence viral diversity estimates in short-read viral pipelines. Cross-platform validation studies, ideally using matched samples, could determine whether this issue is generalizable and evaluate how current software tools handle such artefacts. To apply the current developed pipeline to Illumina data, large adaptations are needed, as this pipeline assumes a read to span the whole amplicon.

5.3. Recombination and other types of chimeras confound the detection of ligation chimeras

In the current pipeline chimeras containing partial amplicons cannot yet be confidently considered as ligation chimeras. These categories might instead represent biological structural events or PCR chimeras which align non-contiguously. This limitation is particularly significant in the context of recombinogenic RNA viruses such as coronaviruses and enteroviruses, where recombination has known biological relevance [61]. However, the large amounts ligation chimeras containing partial amplicons found in labstrains indicate, that certainly not all reads containing partial amplicons are biological or PCR chimeras. By investigating the composition and formation of ligation chimeras, we might gain more insight into how to differentiate ligation chimeras. However, it is possible that some of these artefacts are effectively indistinguishable from biological events.

Patterns in combinations of segments might shed more light on the process of chimera formation. Our results suggest that chimera formation is dependent on amplicon specific features. The chemical properties of an amplicon are determined by its' sequence. Investigating which sequences or combinations of sequences are more likely to form chimeras, can provide hints as to the origin of these chimeras. Additionally if strong correlations between amplicon sequence and its' chimera formation rate are found, this could enable predictive modeling of chimera-prone regions and inform primer designs or ligation conditions.

Predictions might also be used to aid in distinguishing ligated chimeric artefacts from biological events. For example by providing confidence estimates on whether a structural event is biological or artifical, based on the expected number of chimeras in that region. However, these kinds of approaches risk removing biological recombinants or structural variants.

5.4. Toward broader analysis and more complex samples

While the current work demonstrates that ligated chimeric reads can significantly impact viral diversity estimates, generally leading to more accurate results, several limitations should be considered when interpreting these results. Because the ground truth is based on unfiltered data, it may more closely resemble readsets that include chimeric reads. This could introduce bias in our evaluation of accuracy and agreement across filtering strategies, and should be kept in mind when interpreting the apparent improvements observed in chimeric-inclusive datasets and could mean that the higher diversity estimates are actually an overestimation compared to real diversity. Additionally, for the ground truth we assume that all observed diversity reflects either technical artefacts or biological variation within a known, clonal background. In real-world clinical or environmental datasets, however, samples may contain multiple co-circulating viral lineages or

recombinants. In such cases, distinguishing between biological variation and artefacts becomes even more challenging. Future research should aim to assess the impact of ligation chimeras on samples which have more realistic and unbiased intra-sample diversity. This can be done using well-characterized synthetic mixtures or standardized benchmarking datasets [62].

Additionally, viral diversity in this study was assessed using only single-nucleotide variant calls, without reconstruction of full haplotypes. As a result, the metrics used, richness, Shannon entropy, and nucleotide diversity, reflect unlinked site-level variation and do not capture linkage information that may be critical for understanding functional or evolutionary interactions among mutations. Since chimeric reads may falsely link alleles from different genomic regions, their impact may be even more pronounced in haplotype-based diversity estimates [40]. Future work could extend this analysis to local or global haplotypes, using tools such as VILOCA [63] or Strainline [64], to assess how chimera handling strategies affect inferred viral population structures.

Our findings show that viral diversity metrics are sensitive to pre-processing choices with respect to ligation artefacts. While this sensitivity can be leveraged to maximize recall and detect emerging variation, it might also lead to overestimation of diversity in the presence of chimeric reads. A better understanding of the mechanisms driving chimera formation, coupled with improved methods to distinguish artefacts from true recombination events, will be essential for accurately estimating intra-host diversity in viral genomics and accurately informing clinical practice.

6. Conclusion

In this thesis, we developed a modular pipeline framework for the quantification, characterization, and impact assessment of ligated chimeric reads in viral sequencing data. Applying this framework to a range of Nanopore-based viral amplicon datasets, we demonstrate that ligated chimeric artefacts are a prevalent and non-trivial feature of current sequencing protocols. These artefacts often exhibit non-random, structured patterns, suggesting that chimera formation is influenced by amplicon-specific properties.

Through controlled experiments, we show that the inclusion of chimeric reads in variant calling can significantly increase recall. These changes propagate to higher estimated diversity metrics such as richness and Shannon entropy, which are commonly used to quantify intra-host viral variation. Our results highlight that different strategies for handling chimeric reads can yield markedly different variant profiles and diversity estimates.

Our findings underscore the importance of transparent and context-aware pre-processing in viral genomics workflows. While no single approach is optimal for all applications, our pipeline enables systematic evaluation of chimera-related biases and offers a starting point for method selection tailored to specific research goals. Future work may further refine chimera classification, integrate haplotype-level analysis, and extend validation to additional sequencing platforms and library preparation protocols.

7. Code and data availability

The code developed for this thesis is publicly available on GitHub at: `https://github.com/JoyceS13/qc_pipeline`.
The study is based on a combination of internal and public datasets. The internal data was provided by Cerba Research NL and is not publicly available. Public datasets used in this research are available from previously published studies and can be accessed under the respective accession numbers provided in the text.

8. AI Disclosure Statement

In the course of this thesis, I used AI-based tools to support both writing and coding tasks. Specifically, ChatGPT was used to assist with drafting and refining written content, while ChatGPT and GitHub Copilot were employed to support code development. All outputs generated with these tools were critically reviewed and edited by me to ensure their accuracy, originality, and compliance with academic standards.

Glossary

Adapter (Sequencing Adapter) A short artificial DNA sequence ligated to DNA fragments during library preparation to enable sequencing and indexing.

Adapter Trimming The removal of artificial adapter sequences from raw sequencing reads.

Alignment The process of matching sequencing reads to a reference genome to determine their origin or similarity.

Amplicon A DNA fragment that has been amplified through PCR; typically corresponds to a specific genomic region of interest.

Barcode (Index Sequence) A short, unique DNA sequence added to each sample during library preparation. Barcodes allow multiple samples to be sequenced together in the same run, and then later computationally separated (demultiplexed) based on their barcode.

Base Pair (bp) A unit of length in DNA or RNA sequences, consisting of two nucleotides bonded together (A–T or G–C).

Chimeric Read A sequencing read composed of segments from different DNA fragments, usually introduced during sample preparation.

Coverage (Depth) The average number of times each nucleotide is sequenced.

Demultiplexing The process of separating sequencing reads from different samples after a pooled sequencing run, based on unique barcode sequences added during library preparation.

Haplotype A combination of alleles or variants at multiple loci that are inherited together; in viral sequencing, it refers to the sequence of an individual viral genome within a population.

Library Preparation The process of preparing DNA or RNA samples for sequencing, involving adapter and possibly barcode ligation.

Ligation Chimera A chimeric read formed during library preparation when two DNA fragments are accidentally joined, alternatively referred to as ligated chimeric artefacts, or sometimes simply chimera.

Minor Variant A genetic variant present at low frequency within a sample, often relevant for detecting mixed infections or evolution.

Nucleotide The basic structural unit of DNA and RNA, consisting of a base (A, T/U, C, or G), a sugar, and a phosphate group.

PCR (Polymerase Chain Reaction) A laboratory technique used to amplify DNA sequences, enabling the production of millions of copies from a small DNA sample.

Population Nucleotide Diversity (π) A measure of the average number of nucleotide differences per site between any two sequences in a population.

Primary Alignment The highest-scoring alignment for a read, considered the most accurate placement.

Primer A short single-stranded DNA sequence used to initiate DNA synthesis during PCR by binding to the target DNA.

Recombination A biological process where genetic material is rearranged, e.g. viral replication.

Reference Bias A tendency to favor sequences more similar to the reference genome, potentially masking true biological variation.

Richness The number of unique variants observed in a sample, regardless of their abundance.

Secondary Alignment An alternative alignment for a read, typically lower scoring than the primary but still plausible, which uses the same region of the read as the primary alignment.

Sequencing depth The average number of times each nucleotide is sequenced.

Shannon Entropy A diversity index considering both the number and relative abundance of variants, reflecting evenness in the population.

SNV (Single Nucleotide Variant) A change in a single nucleotide position in the genome.

Variant Calling The process of identifying genetic variants from sequencing data.

Viral Diversity The amount of genetic variation within a viral population, often quantified by richness, entropy, or nucleotide diversity.

References

1. Kuroda, M. *et al.* Characterization of Quasispecies of Pandemic 2009 Influenza A Virus (A/H1N1/2009) by De Novo Sequencing Using a Next-Generation DNA Sequencer. *PLoS ONE* **5** (ed Jacobson, S.) e10256. ISSN: 1932-6203. https://dx.plos.org/10.1371/journal.pone.0010256 (2025) (Apr. 23, 2010).
2. Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439,** 344–348. ISSN: 0028-0836, 1476-4687. https://www.nature.com/articles/nature04388 (2025) (Jan. 2006).
3. Mason, S., Devincenzo, J. P., Toovey, S., Wu, J. Z. & Whitley, R. J. Comparison of antiviral resistance across acute and chronic viral infections. *Antiviral Research* **158,** 103–112. ISSN: 01663542. https://linkinghub.elsevier.com/retrieve/pii/S0166354218301219 (2025) (Oct. 2018).
4. Simen, B. B. *et al.* Low-Abundance Drug-Resistant Viral Variants in Chronically HIV-Infected, Antiretroviral Treatment–Naive Patients Significantly Impact Treatment Outcomes. *Journal of Infectious Diseases* **199,** 693–701. ISSN: 0022-1899, 1537-6613. https://academic.oup.com/jid/article-lookup/doi/10.1086/596736 (2025) (Mar. 1, 2009).
5. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* **9,** 267–276. ISSN: 1471-0056, 1471-0064. https://www.nature.com/articles/nrg2323 (2025) (Apr. 2008).
6. Domingo, E., Sheldon, J. & Perales, C. Viral Quasispecies Evolution. *Microbiology and Molecular Biology Reviews* **76,** 159–216. ISSN: 1092-2172, 1098-5557. https://journals.asm.org/doi/10.1128/MMBR.05023-11 (2025) (June 2012).
7. Posada-Cespedes, S., Seifert, D. & Beerenwinkel, N. Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus Research* **239,** 17–32. ISSN: 01681702. https://linkinghub.elsevier.com/retrieve/pii/S0168170216304130 (2025) (July 2017).
8. Di Giallonardo, F. *et al.* Next-Generation Sequencing of HIV-1 RNA Genomes: Determination of Error Rates and Minimizing Artificial Recombination. *PLoS ONE* **8** (ed Wainberg, M.) e74249. ISSN: 1932-6203. https://dx.plos.org/10.1371/journal.pone.0074249 (2025) (Sept. 18, 2013).
9. Quail, M. *et al.* A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* **13,** 341. ISSN: 1471-2164. http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-341 (2025) (2012).
10. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13,** 36–46. ISSN: 1471-0056, 1471-0064. https://www.nature.com/articles/nrg3117 (2025) (Jan. 2012).
11. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* **21,** 30. ISSN: 1474-760X. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1935-5 (2025) (Dec. 2020).
12. Illingworth, C. J. R. *et al.* On the effective depth of viral sequence data. *Virus Evolution* **3.** ISSN: 2057-1577. https://academic.oup.com/ve/article/doi/10.1093/ve/vex030/4629376 (2025) (July 1, 2017).
13. Nasir, J. A. *et al.* A Comparison of Whole Genome Sequencing of SARS-CoV-2 Using Amplicon-Based Sequencing, Random Hexamers, and Bait Capture. *Viruses* **12,** 895. ISSN: 1999-4915. https://www.mdpi.com/1999-4915/12/8/895 (2025) (Aug. 15, 2020).
14. Zanini, F., Brodin, J., Albert, J. & Neher, R. A. Error rates, PCR recombination, and sampling depth in HIV-1 whole genome deep sequencing. *Virus Research* **239,** 106–114. ISSN: 01681702. https://linkinghub.elsevier.com/retrieve/pii/S0168170216304221 (2025) (July 2017).
15. McCrone, J. T. & Lauring, A. S. Measurements of Intrahost Viral Diversity Are Extremely Sensitive to Systematic Errors in Variant Calling. *Journal of Virology* **90** (ed Dermody, T. S.) 6884–6895. ISSN: 0022-538X, 1098-5514. https://journals.asm.org/doi/10.1128/JVI.00667-16 (2025) (Aug. 2016).
16. White, R., Pellefigues, C., Ronchese, F., Lamiable, O. & Eccles, D. Investigation of chimeric reads using the MinION. *F1000Research* **6,** 631. ISSN: 2046-1402 (2017).

17. Edgar, R. C. *UCHIME2: improved chimera prediction for amplicon sequencing* Sept. 12, 2016. http://biorxiv.org/lookup/doi/10.1101/074252 (2024).

18. Croville, G. *et al.* An amplicon-based nanopore sequencing workflow for rapid tracking of avian influenza outbreaks, France, 2020-2022. *Frontiers in Cellular and Infection Microbiology* **14,** 1257586. ISSN: 2235-2988. https://www.frontiersin.org/articles/10.3389/fcimb.2024.1257586/full (2025) (Jan. 22, 2024).

19. Brinkmann, A. *et al.* AmpliCoV: Rapid Whole-Genome Sequencing Using Multiplex PCR Amplification and Real-Time Oxford Nanopore MinION Sequencing Enables Rapid Variant Identification of SARS-CoV-2. *Frontiers in Microbiology* **12,** 651151. ISSN: 1664-302X. https://www.frontiersin.org/articles/10.3389/fmicb.2021.651151/full (2025) (July 1, 2021).

20. Dong, X. *et al. An improved rapid and sensitive long amplicon method for nanopore-based RSV whole genome sequencing* Jan. 6, 2025. http://biorxiv.org/lookup/doi/10.1101/2025.01.06.631406 (2025).

21. Lumley, S. F. *et al.* Whole genome sequencing of hepatitis B virus using tiled amplicon (HEPTILE) and probe based enrichment on Illumina and Nanopore platforms. *Scientific Reports* **15,** 5795. ISSN: 2045-2322. https://www.nature.com/articles/s41598-025-87721-1 (2025) (Feb. 17, 2025).

22. Lee, G.-Y. *et al.* Molecular diagnosis of patients with hepatitis A virus infection using amplicon-based nanopore sequencing. *PLOS ONE* **18** (ed Li, Y.) e0288361. ISSN: 1932-6203. https://dx.plos.org/10.1371/journal.pone.0288361 (2025) (July 12, 2023).

23. Bull, R. A. *et al.* Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nature Communications* **11,** 6272. ISSN: 2041-1723. https://www.nature.com/articles/s41467-020-20075-6 (2025) (Dec. 9, 2020).

24. Kalendar, R. *et al.* Universal whole-genome Oxford nanopore sequencing of SARS-CoV-2 using tiled amplicons. *Scientific Reports* **13,** 10334. ISSN: 2045-2322. https://www.nature.com/articles/s41598-023-37588-x (2025) (June 26, 2023).

25. Xu, X. *et al.* High-resolution and real-time wastewater viral surveillance by Nanopore sequencing. *Water Research* **256,** 121623. ISSN: 00431354. https://linkinghub.elsevier.com/retrieve/pii/S0043135424005244 (2025) (June 2024).

26. Brejová, B. *et al.* Nanopore sequencing of SARS-CoV-2: Comparison of short and long PCR-tiling amplicon protocols. *PLOS ONE* **16** (ed Abd El-Aty, A. M.) e0259277. ISSN: 1932-6203. https://dx.plos.org/10.1371/journal.pone.0259277 (2025) (Oct. 29, 2021).

27. A. Pater, A. *et al.* High throughput nanopore sequencing of SARS-CoV-2 viral genomes from patient samples. *Journal of Biological Methods* **8,** 1. ISSN: 2326-9901. https://polscientific.com/journal/JBM/8/4/10.14440/jbm.2021.360 (2025) (Sept. 27, 2021).

28. Liu, H. *et al.* Assessment of two-pool multiplex long-amplicon nanopore sequencing of SARS-CoV-2. *Journal of Medical Virology* **94,** 327–334. ISSN: 0146-6615, 1096-9071. https://onlinelibrary.wiley.com/doi/10.1002/jmv.27336 (2025) (Jan. 2022).

29. Marijon, P., Chikhi, R. & Varré, J.-S. yacrd and fpa: upstream tools for long-read genome assembly. *Bioinformatics* **36** (ed Birol, I.) 3894–3896. ISSN: 1367-4803, 1367-4811. https://academic.oup.com/bioinformatics/article/36/12/3894/5823296 (2024) (June 1, 2020).

30. Yang, C. *et al.* Characterization and simulation of metagenomic nanopore sequencing data with Meta-NanoSim. *GigaScience* **12,** giad013. ISSN: 2047-217X. https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad013/7080817 (2025) (Mar. 20, 2023).

31. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature Protocols* **12,** 1261–1276. ISSN: 1754-2189, 1750-2799. https://www.nature.com/articles/nprot.2017.066 (2025) (June 2017).

32. Salzberg, S. L. *et al.* Genome Analysis Linking Recent European and African Influenza (H5N1) Viruses. *Emerging Infectious Diseases* **13,** 713–718. ISSN: 1080-6040, 1080-6059. http://wwwnc.cdc.gov/eid/article/13/5/07-0013_article.htm (2025) (May 2007).

33. Yu, Q. *et al.* PriSM: a primer selection and matching tool for amplification and sequencing of viral genomes. *Bioinformatics* **27,** 266–267. ISSN: 1367-4811, 1367-4803. https://academic.oup.com/bioinformatics/article/27/2/266/284697 (2025) (Jan. 15, 2011).

34. Gohl, D. M. *et al.* A rapid, cost-effective tailed amplicon method for sequencing SARS-CoV-2. *BMC Genomics* **21,** 863. ISSN: 1471-2164. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-07283-6 (2025) (Dec. 2020).

35. Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V. & Polz, M. F. PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. *Applied and Environmental Microbiology* **71,** 8966–8969. ISSN: 0099-2240, 1098-5336. https://journals.asm.org/doi/10.1128/AEM.71.12.8966-8969.2005 (2025) (Dec. 2005).

36. Orton, R. J. *et al.* Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. *BMC Genomics* **16,** 229. ISSN: 1471-2164. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1456-x (2025) (Dec. 2015).

37. Delahaye, C. & Nicolas, J. Sequencing DNA with nanopores: Troubles and biases. *PLOS ONE* **16** (ed Andrés-León, E.) e0257521. ISSN: 1932-6203. https://dx.plos.org/10.1371/journal.pone.0257521 (2025) (Oct. 1, 2021).

38. Saiki, R. K. *et al.* Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. *Science* **239,** 487–491. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.2448875 (2025) (Jan. 29, 1988).

39. Meyerhans, A., Vartanian, J.-P. & Wain-Hobson, S. DNA recombination during PCR. *Nucleic Acids Research* **18,** 1687–1691. ISSN: 0305-1048, 1362-4962. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/18.7.1687 (2025) (1990).

40. Fang, G., Zhu, G., Burger, H., Keithly, J. S. & Weiser, B. Minimizing DNA recombination during long RT-PCR. *Journal of Virological Methods* **76,** 139–148. ISSN: 01660934. https://linkinghub.elsevier.com/retrieve/pii/S0166093498001335 (2025) (Dec. 1998).

41. Scott, G. *et al.* Long Amplicon Nanopore Sequencing for Dual-Typing RdRp and VP1 Genes of Norovirus Genogroups I and II in Wastewater. *Food and Environmental Virology* **16,** 479–491. ISSN: 1867-0334, 1867-0342. https://link.springer.com/10.1007/s12560-024-09611-5 (2024) (Dec. 2024).

42. Omelina, E. S., Ivankin, A. V., Letiagina, A. E. & Pindyurin, A. V. Optimized PCR conditions minimizing the formation of chimeric DNA molecules from MPRA plasmid libraries. *BMC Genomics* **20,** 536. ISSN: 1471-2164. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5847-2 (2025) (S7 July 2019).

43. *PoreChop* https://github.com/rrwick/Porechop.

44. Li, Y. *et al. A Genomic Language Model for Chimera Artifact Detection in Nanopore Direct RNA Sequencing* Oct. 25, 2024. http://biorxiv.org/lookup/doi/10.1101/2024.10.23.619929 (2025).

45. Wick, R. R., Judd, L. M., Wyres, K. L. & Holt, K. E. Recovery of small plasmid sequences via Oxford Nanopore sequencing. *Microbial Genomics* **7.** ISSN: 2057-5858. https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000631 (2025) (Aug. 31, 2021).

46. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research* **40,** 11189–11201. ISSN: 1362-4962, 0305-1048. https://academic.oup.com/nar/article/40/22/11189/1152727 (2025) (Dec. 1, 2012).

47. Garrison, E. & Marth, G. *Haplotype-based variant detection from short-read sequencing* July 20, 2012. arXiv: 1207.3907[q-bio]. http://arxiv.org/abs/1207.3907 (2025).

48. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biology* **20,** 8. ISSN: 1474-760X (Jan. 8, 2019).

49. Fuhrmann, L., Jablonski, K. P. & Beerenwinkel, N. Quantitative measures of within-host viral genetic diversity. Artwork Size: 7 p. Medium: application/pdf Publisher: ETH Zurich, 7 p. http://hdl.handle.net/20.500.11850/493904 (2025) (Aug. 2021).

50. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10.** Publisher: F1000 Research Ltd, 33. ISSN: 2046-1402. https://f1000research.com/articles/10-33/v2 (2025) (Apr. 19, 2021).

51. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34** (ed Birol, I.) 3094–3100. ISSN: 1367-4803, 1367-4811. https://academic.oup.com/bioinformatics/article/34/18/3094/4994778 (2025) (Sept. 15, 2018).

52. Kebschull, J. M. & Zador, A. M. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*, gkv717. ISSN: 0305-1048, 1362-4962. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv717 (2025) (July 17, 2015).

53. Rivera-Franco, N. *et al.* Genomic variability in Zika virus in GBS cases in Colombia. *PLOS ONE* **19** (ed Ruiz-Saenz, J.) Publisher: Public Library of Science (PLoS), e0313545. ISSN: 1932-6203. https://dx.plos.org/10.1371/journal.pone.0313545 (2025) (Nov. 19, 2024).

54. Gao, R. *et al. Tiled PCR amplification-based Whole Genome Sequencing and Phylogenetic Classification Accelerate the Implementation of Respiratory Syncytial Virus Genomic surveillance in Canada as a Pilot Study* Nov. 25, 2024. http://biorxiv.org/lookup/doi/10.1101/2024.11.22.624816 (2025).

55. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6,** 100. ISSN: 2046-1402 (2017).

56. Xu, Y. *et al.* Detection of Viral Pathogens With Multiplex Nanopore MinION Sequencing: Be Careful With Cross-Talk. *Frontiers in Microbiology* **9,** 2225. ISSN: 1664-302X. https://www.frontiersin.org/article/10.3389/fmicb.2018.02225/full (2025) (Sept. 19, 2018).

57. Odelberg, S. J., Weiss, R. B., Hata, A. & White, R. Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. *Nucleic Acids Research* **23,** 2049–2057. ISSN: 0305-1048, 1362-4962. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/23.11.2049 (2025) (1995).

58. Pérez-Losada, M., Arenas, M., Galán, J. C., Palero, F. & González-Candelas, F. Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution* **30,** 296–307. ISSN: 15671348. https://linkinghub.elsevier.com/retrieve/pii/S156713481400478X (2025) (Mar. 2015).

59. Liu, Z. *et al.* Transient stem-loop structure of nucleic acid template may interfere with polymerase chain reaction through endonuclease activity of Taq DNA polymerase. *Gene* **764,** 145095. ISSN: 03781119. https://linkinghub.elsevier.com/retrieve/pii/S0378111920307642 (2025) (Jan. 2021).

60. Wick, R. Badread: simulation of error-prone long reads. *Journal of Open Source Software* **4,** 1316. ISSN: 2475-9066. http://joss.theoj.org/papers/10.21105/joss.01316 (2025) (Apr. 4, 2019).

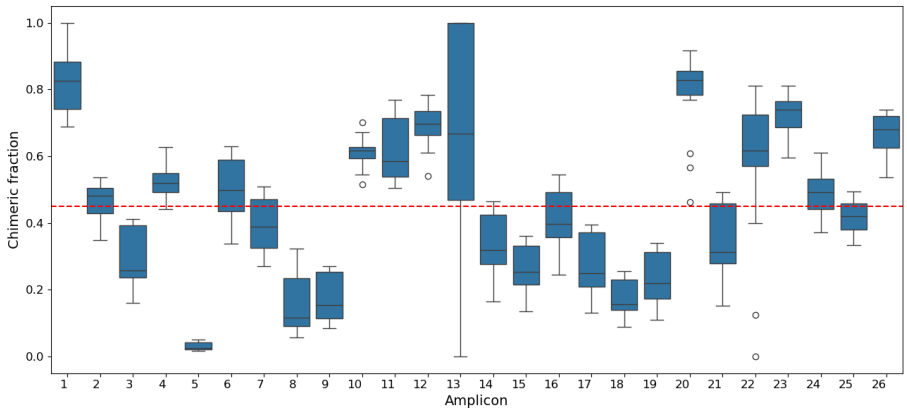
61. Simon-Loriere, E. & Holmes, E. C. Why do RNA viruses recombine? *Nature Reviews Microbiology* **9,** 617–626. ISSN: 1740-1526, 1740-1534. https://www.nature.com/articles/nrmicro2614 (2025) (Aug. 2011).

62. Knyazev, S. *et al.* Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction. *Nucleic Acids Research* **49,** e102–e102. ISSN: 0305-1048, 1362-4962. https://academic.oup.com/nar/article/49/17/e102/6313236 (2025) (Sept. 27, 2021).

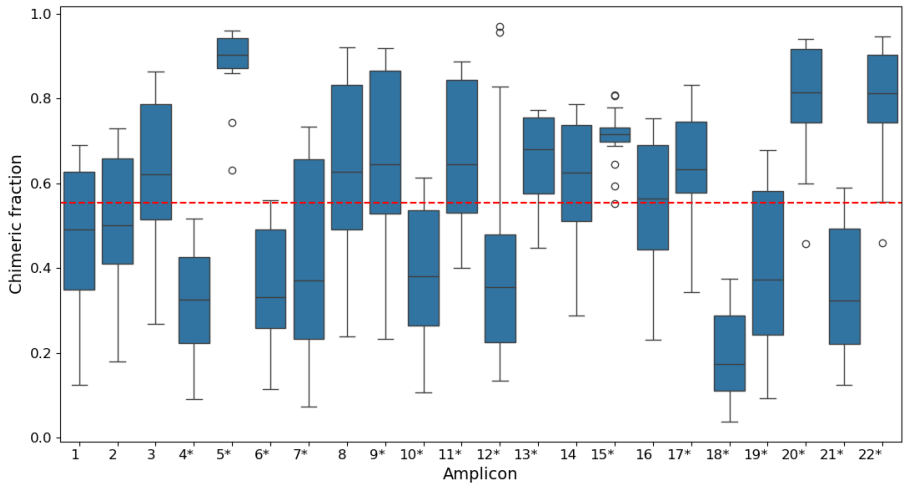
63. Fuhrmann, L., Langer, B., Topolsky, I. & Beerenwinkel, N. *VILOCA: Sequencing quality-aware haplotype reconstruction and mutation calling for short- and long-read data* June 9, 2024. http://biorxiv.org/lookup/doi/10.1101/2024.06.06.597712 (2025).

64. Luo, X., Kang, X. & Schönhuth, A. Strainline: full-length de novo viral haplotype reconstruction from noisy long reads. *Genome Biology* **23,** 29. ISSN: 1474-760X. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02587-6 (2025) (Jan. 20, 2022).

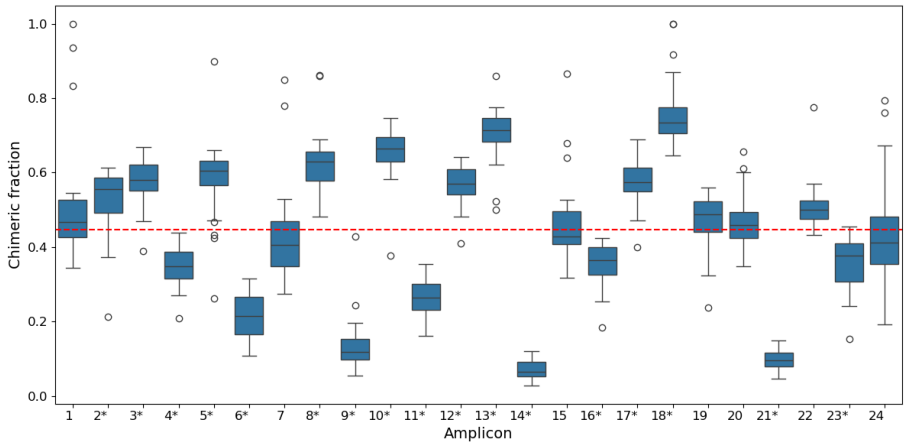
A. Chimeric burden across amplicons



(a) hMPV-A

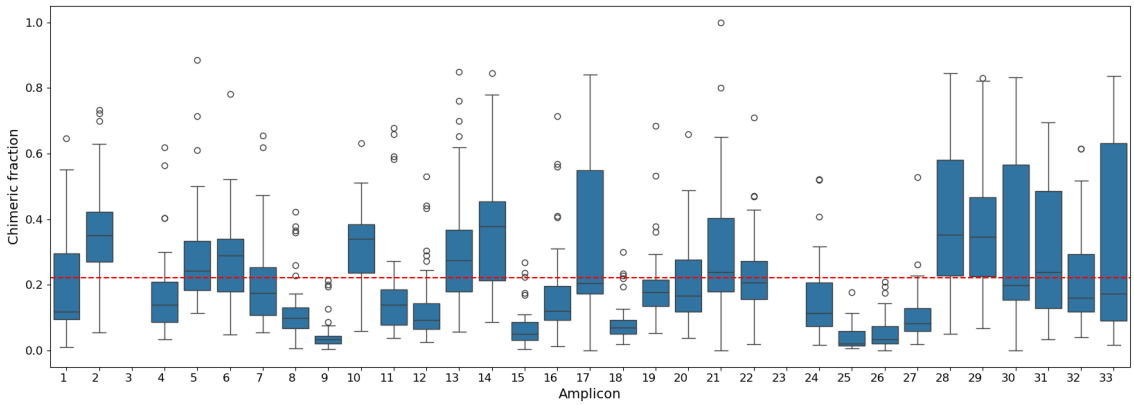


(b) hMPV-B

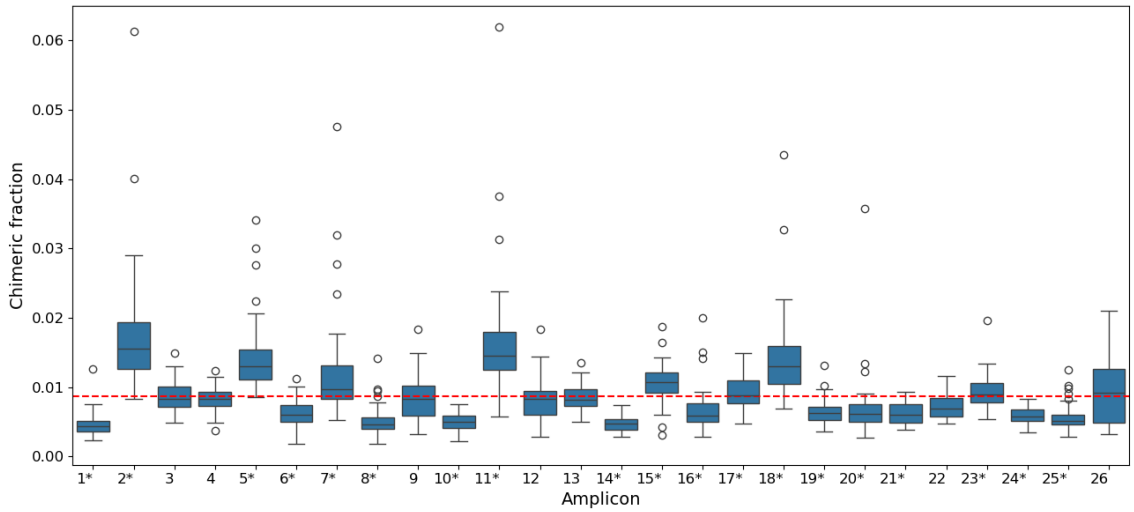


(c) RSV-A

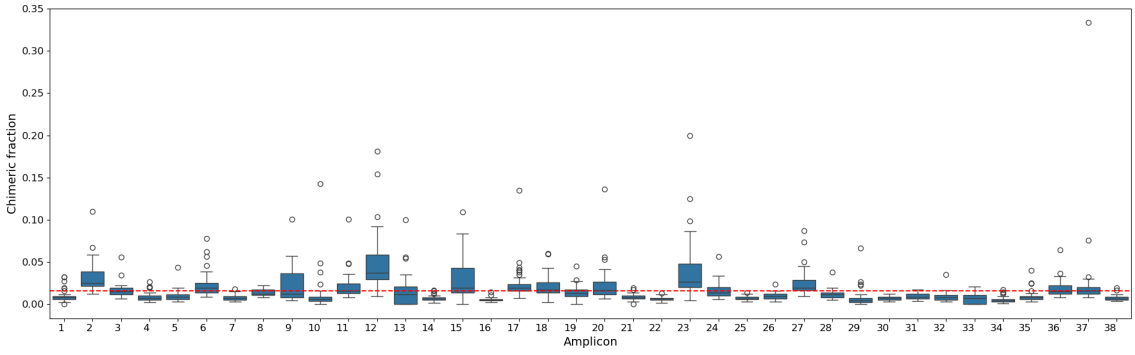
Figure 15: Chimeric fraction per amplicon. The red dotted line represents the average chimeric fraction across all amplicons. Amplicons that are significantly more or less affected are marked with an asterisk (*) in the label. Significance was determined using the Mann–Whitney U test, comparing each amplicon to the overall distribution.



(a) ZIKV



(b) RSV-A (public)



(c) RSV-B (public)

Figure 16: Chimeric fraction per amplicon. The red dotted line represents the average chimeric fraction across all amplicons. Amplicons that are significantly more or less affected are marked with an asterisk (*) in the label. Significance was determined using the Mann–Whitney U test, comparing each amplicon to the overall distribution.

B. Alignment patterns of partial amplicon segments in ligated chimeric reads

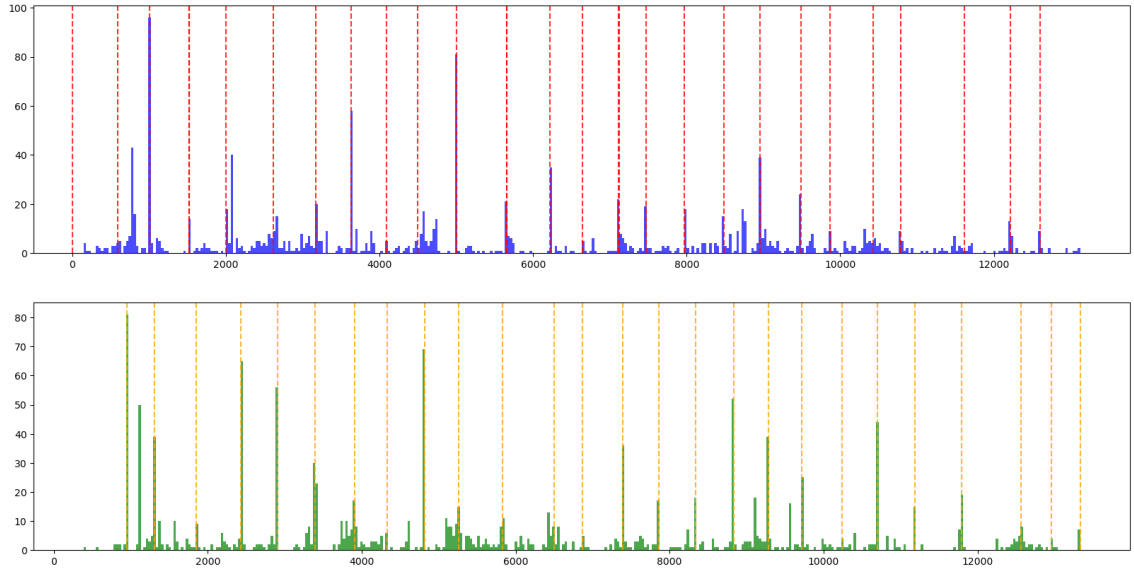


Figure 17: In the top panel the start positions of alignments of partial amplicons segments in chimeric reads are binned. In the bottom panel the ending positions are pictured. In both panels dotted lines indicate primer positions, for the top panel forward primers are shown and for the bottom panel reverse primers are shown. Many segments start or end on primer positions, there are some positions which do not correspond to primer positions, where a larger proportion of segments start or end. These positions might indicate formation of PCR chimeras which align non-contiguously. To confirm the presence of PCR chimeras, sequence similarity of the reference around the junction should be found.

C. Variant calling on additional assays

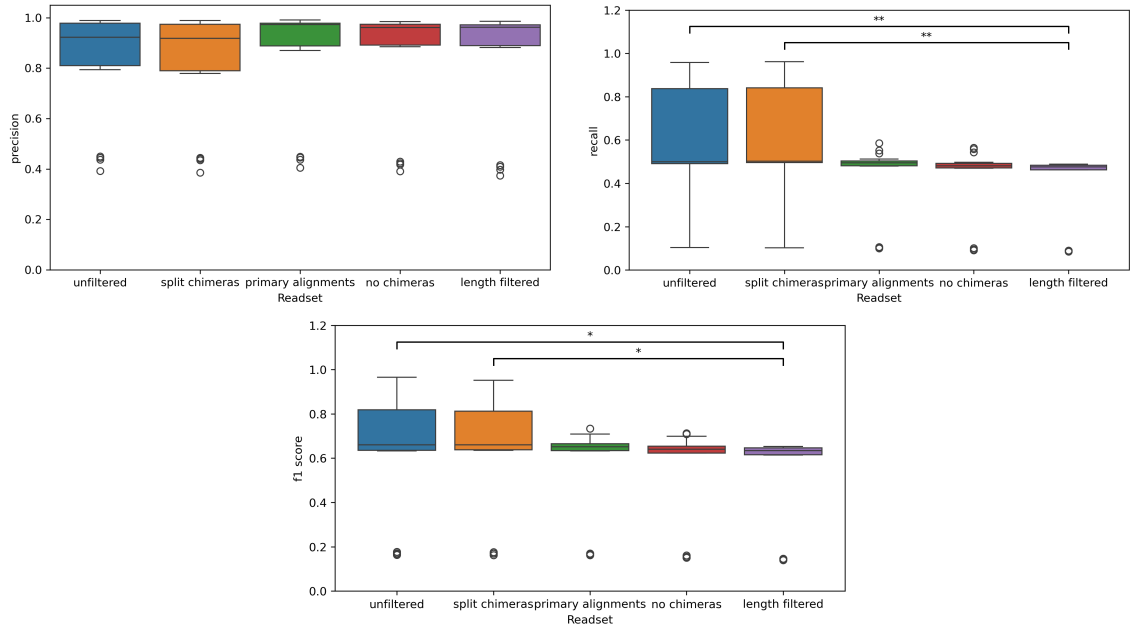


Figure 18: Precision, recall and F1-score of variant calls of replicate samples of the internal hMPV-A labstrain. Datasets are compared which have undergone different filtering strategies for chimeric reads. Statistically significant differences between datasets are indicated with brackets; asterisks indicate significance levels: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

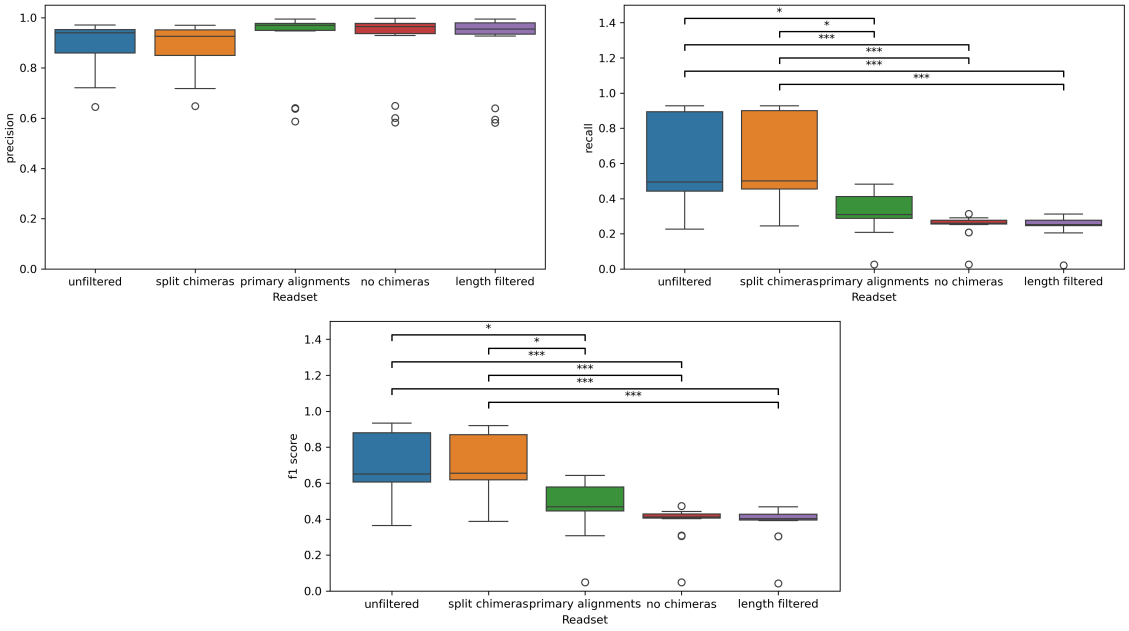


Figure 19: Precision, recall and F1-score of variant calls of replicate samples of the internal hMPV-B labstrain. Datasets are compared which have undergone different filtering strategies for chimeric reads. Statistically significant differences between datasets are indicated with brackets; asterisks indicate significance levels: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

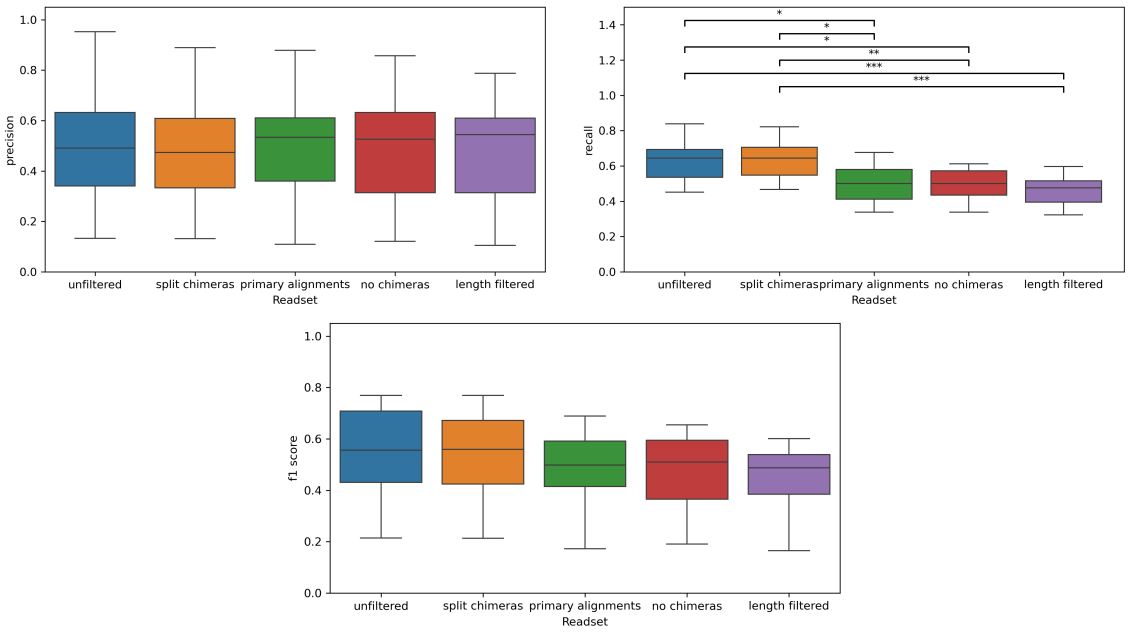
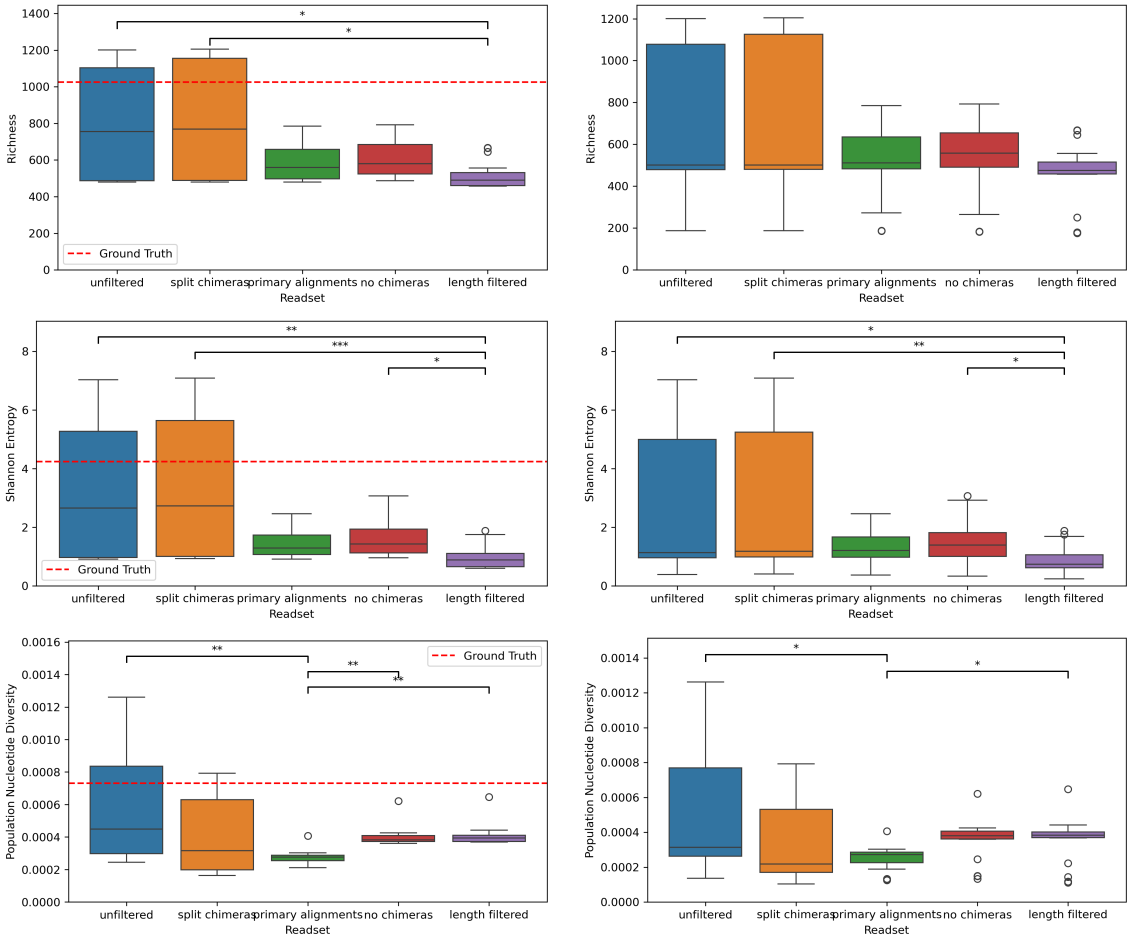


Figure 20: Precision, recall and F1-score of variant calls of replicate samples of the internal RSV-A labstrain. Datasets are compared which have undergone different filtering strategies for chimeric reads. Statistically significant differences between datasets are indicated with brackets; asterisks indicate significance levels: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

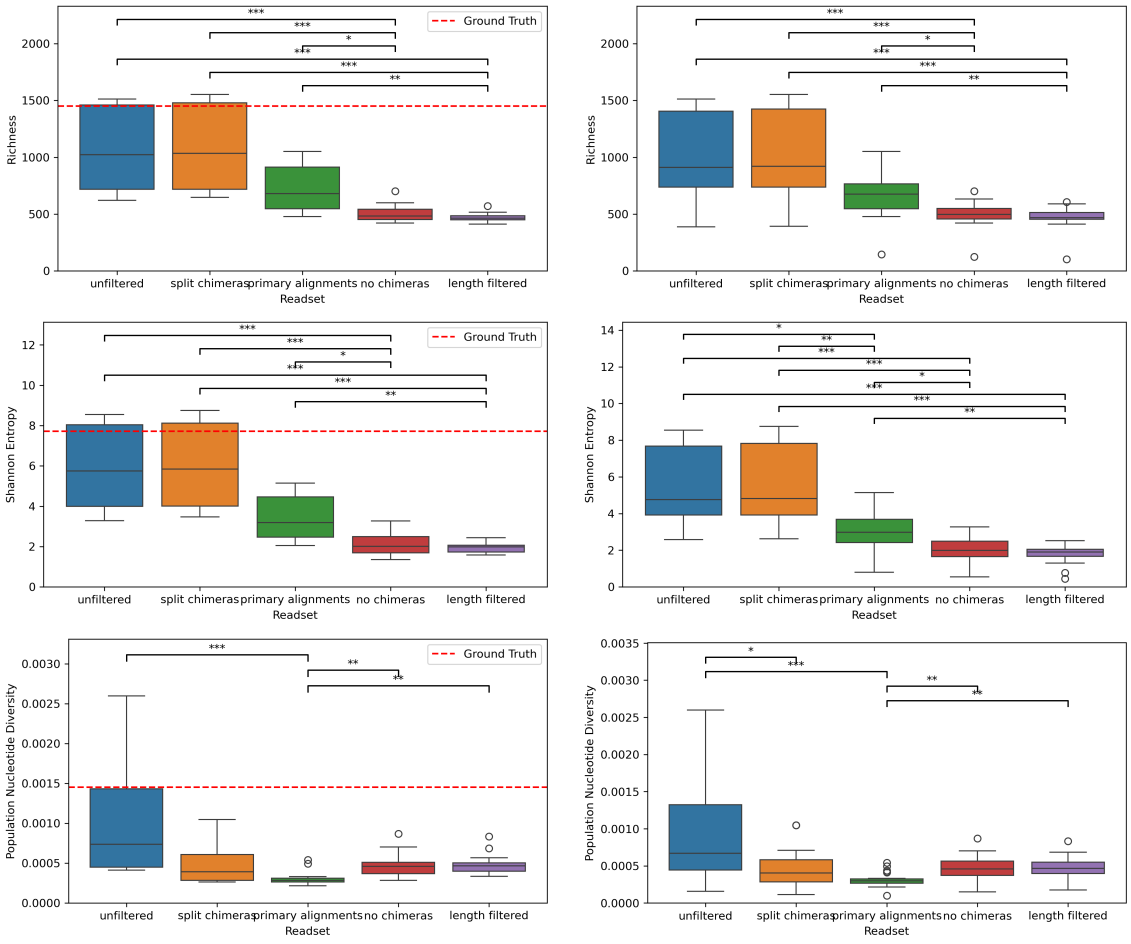
D. Viral diversity estimation on additional assays



(a) Replicate labstrain samples, the viral diversity estimate of the ground truth is indicated with the red dotted line.

(b) Replicate labstrain and clinical samples

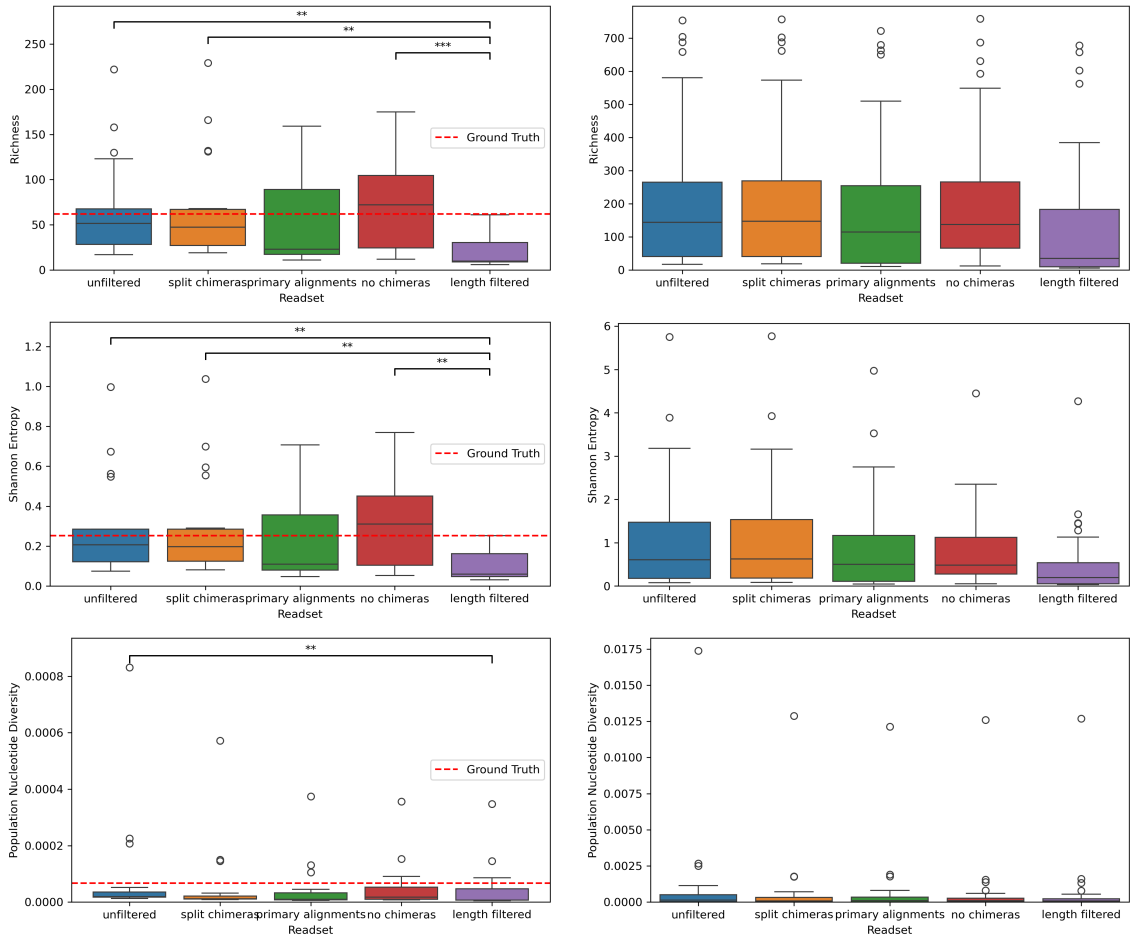
Figure 21: Viral diversity estimate metrics of the internal hMPV-A assay using different filtering strategies for chimeric reads. Statistically significant differences between datasets are indicated with brackets; asterisks indicate significance levels: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).



(a) Replicate labstrain samples, the viral diversity estimate of the ground truth is indicated with the red dotted line.

(b) Replicate labstrain and clinical samples

Figure 22: Viral diversity estimate metrics of the internal hMPV-B assay using different filtering strategies for chimeric reads. Statistically significant differences between datasets are indicated with brackets; asterisks indicate significance levels: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).



(a) Replicate labstrain samples, the viral diversity estimate of the ground truth is indicated with the red dotted line.

(b) Replicate labstrain and clinical samples

Figure 23: Viral diversity estimate metrics of the internal RSV-A assay using different filtering strategies for chimeric reads. Statistically significant differences between datasets are indicated with brackets; asterisks indicate significance levels: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

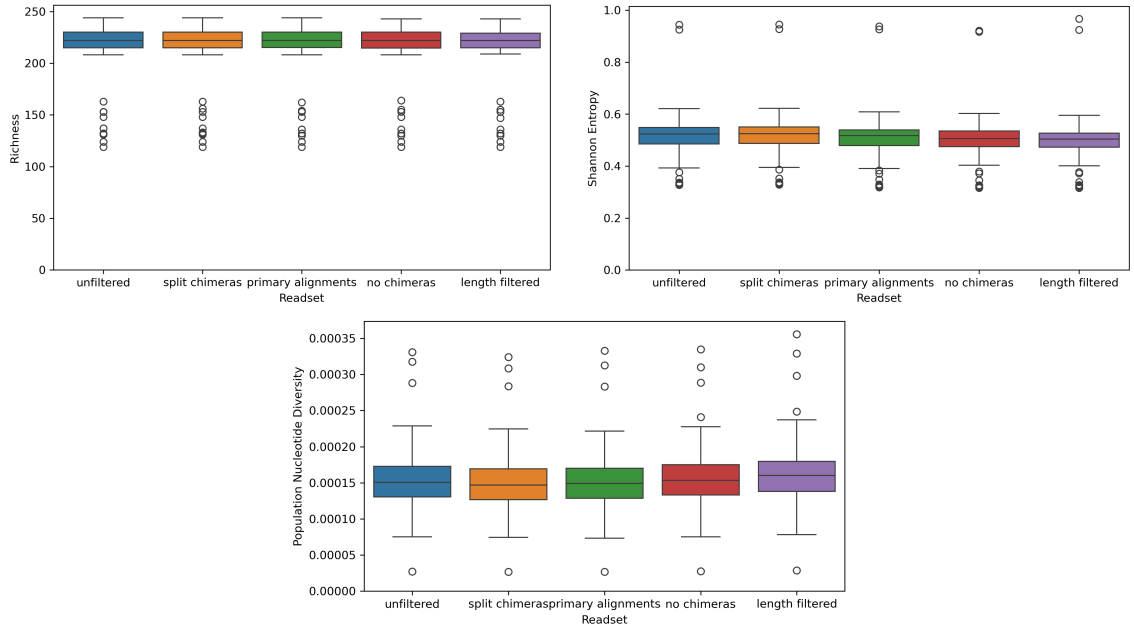


Figure 24: Viral diversity estimate metrics of the public RSV-A assay using different filtering strategies for chimeric reads. Statistically significant differences between datasets are indicated with brackets; asterisks indicate significance levels: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

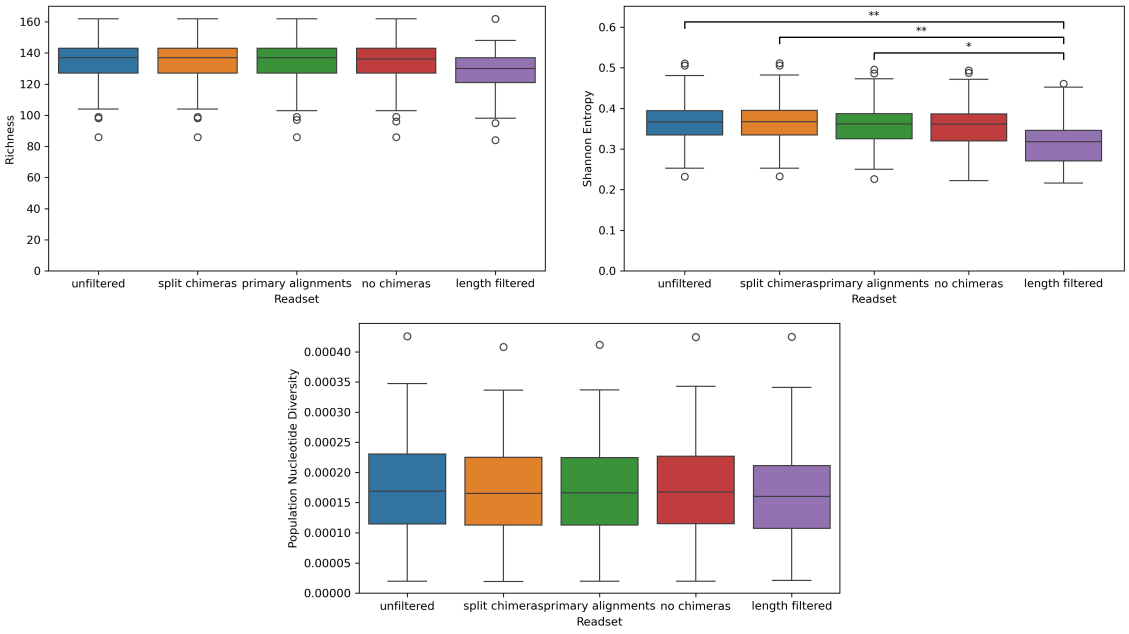


Figure 25: Viral diversity estimate metrics of the public RSV-B assay using different filtering strategies for chimeric reads. Statistically significant differences between datasets are indicated with brackets; asterisks indicate significance levels: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).