Supplementary Material

Pneumovirus genome misassemblies associated with duplications in glycoprotein genes

Stephanie Goya, Alex Greninger

Methods

FASTQ datasets

FASTQ files of pneumovirus containing nucleotide duplications in the attachment glycoprotein G gene were downloaded from NCBI Sequence Read Archive (SRA):

- SRR22580744: RSV-A with 72-nt duplication, sequenced by metagenomic paired end NextSeq 2000 Illumina sequencer. GenBank consensus sequence OP965711.1
- SRR22580750: RSV-B with 60-nt duplication, sequenced by metagenomic paired end NextSeq 2000 Illumina sequencer. GenBank consensus sequence OP965698.1
- SRR17439934: MPV with 111-nt duplication, sequenced by metagenomic single end MiSeq Illumina sequencer. GenBank consensus sequence MK167040.2
- SRR17439935: MPV with 180-nt duplication, sequenced by metagenomic single end MiSeq Illumina sequencer. GenBank consensus sequence MK167039.2

Reference genomes

Reference genomes of each viral species, both with and without the nucleotide duplication in the attachment glycoprotein G gene, were downloaded from NCBI GenBank:

- RSV-A
  - NC_038235: NCBI Reference Sequence RSV-A strain A2 (1961), no 72-nt duplication.
  - OR795328.1: Clinical isolate (2009), no 72-nt duplication.
  - PP109421.1: Clinical isolate (2017), contains 72-nt duplication.
- RSV-B
  - NC_001781: NCBI Reference Sequence RSV-A strain B1 (1985), no 60-nt duplication.
  - JQ582843.1: Clinical isolate (2002), no 60-nt duplication.
  - OP975389.1: Clinical isolate (2019), contains 60-nt duplication.
- MPV
  - OL794474.1: Clinical isolate (2016), no nucleotide duplications.
  - OL794481.1: Clinical isolate (2015), contains 111-nt duplication.
  - OL794465.1: Clinical isolate (2016), contains 180-nt duplication.

Bioinformatic Pipeline

1. Read preprocessing: Each FASTQ file underwent sequencing adapter trimming using cutadapt v3.4, followed by quality filtering with BBDuk v39.01 (parameters:

forcetrimleft=10, minavgquality=20). In RSV-A dataset, an additional assay was performed removing reads shorter than 72nt (minlength=72).

2. Reference-based Mapping: Quality-filtered FASTQ files were mapped against their respective reference genomes using BWA MEM v0.7.18 with default parameters.
3. Reference-guided de novo assembly: For additional analysis, a hybrid approach was used where quality-filtered FASTQ files were assembled de novo with MetaSPAdes v3.15.5. the resulting contigs (in FASTA format) were then mapped to the reference genomes using BWA MEM.
4. Visualization: Reference-based assemblies were visualized using IGV v2.17.4 and Geneious Prime v2023.01.

Consensus genomes from databases and phylogenetic tree construction

1. RSV Consensus Genomes

All complete RSV consensus genomes submitted from January 01, 2022 to November 16, 2024 were downloaded from GISAID EpiRSV (https://gisaid.org/) with the filters “complete (>14,900nt length)” and “low coverage excl (exclude sequences with >5%N)”. Genomes were aligned using mafft v7.511, and the alignments were visualized with Aliview. Subset of genomes lacking the 72 or 60 nucleotide duplication in the G gene in RSV-A or RSV-B, respectively, were selected for further analysis.

2. Phylogenetic Analysis

Maximum likelihood trees for RSV genomes missing the nucleotide duplication were built using NextClade (https://clades.nextstrain.org, accessed on November 20, 2024). Trees also included a set of reference sequences containing representative genomes of all RSV lineages. Trees were visualized using auspice v 0.12.0 (https://auspice.us/).

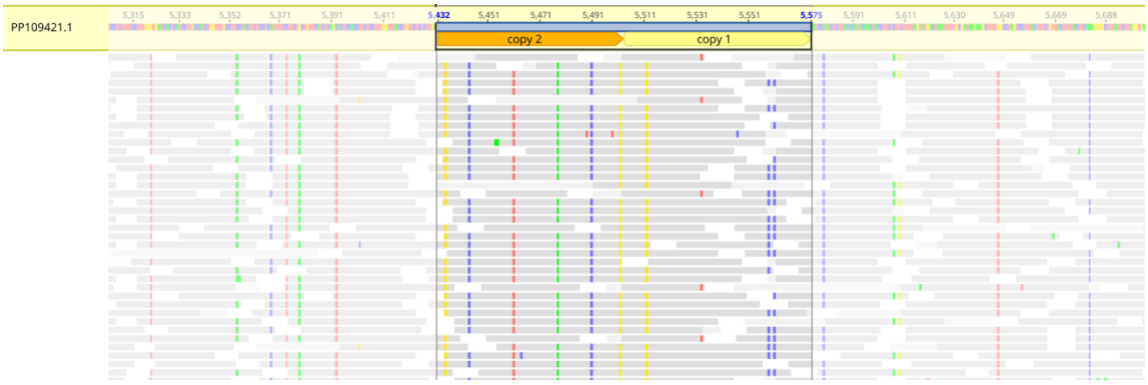
Results

RSV-A

The results presented below describe the mapping of sequencing reads from a given RSV-A sample containing the 72-nt duplication against three different reference genomes.

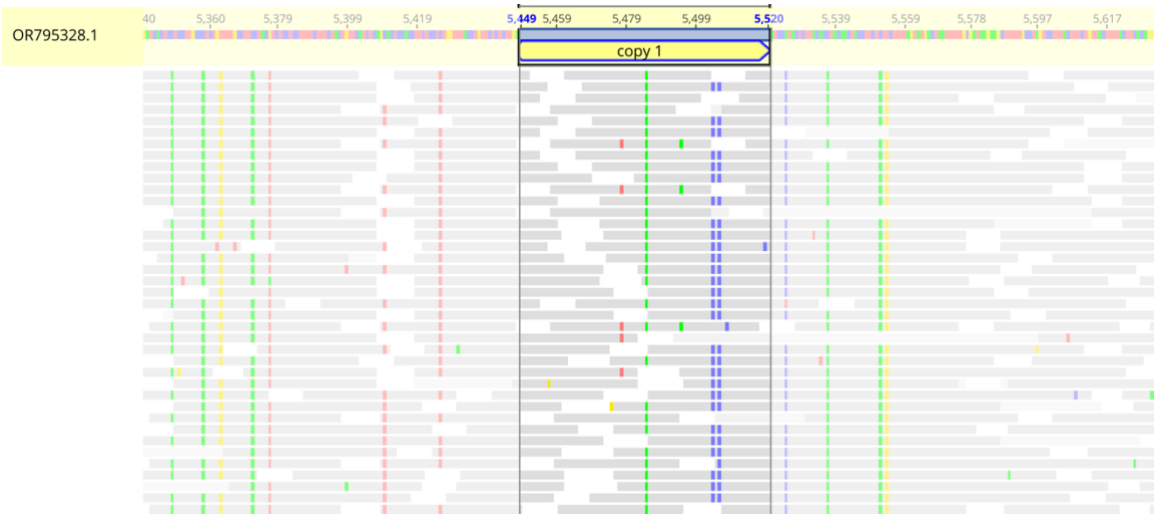
Mapping of RSV-A with the 72-nt Duplication Against PP109421.1 (containing the 72-nt duplication).

The RSV-A sequence containing the 72-nt duplication was mapped against **PP109421.1** (clinical isolate from 2017). The alignment shows uniform coverage across the reference genome, including the duplicated region. Reads span both copies of the duplication, with some reads mapping between the two copies. This indicates that the duplication is well represented in the sequencing data, as shown in the figure below.

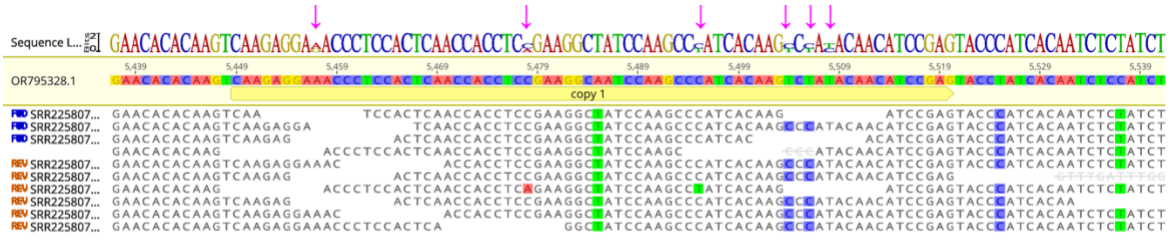


Mapping of RSV-A with the 72-nt Duplication Against OR795328.1 (no 72-nt duplication).

Next, the RSV-A sequence was mapped against OR795328.1, a strain that lacks the 72-nt duplication. While the overall coverage appears uniform, the alignment reveals an interesting pattern: only short reads map to the first copy region, and no reads cover the entire region that spans both copies. Specifically, the reads that map to copy 1 are partial and do not extend across the full length of the duplication.



Two nucleotide combinations were detected at six specific coordinates, where the sequence appears to be a combination of the nucleotides from both copies. The zoomed-in figure below highlights, with pink arrows, the positions where these combined nucleotides occur. The sequence logo illustrates the nucleotide frequency per position, showing that the first copy (copy 1) contains multiple nucleotide variations at these specific sites. The consensus sequences for copy 1 and copy 2 are displayed, with differences between the copies highlighted in pink.



>RSV-A consensus copy 1
CAAGAGGAGACCCTCCACTCAACCACCTCCGAAGGCTATCCAAGCCCATCACAAGCCCATACAACATCCGAG
>RSV-A consensus copy 2
CAAGAGGAAACCCTCCACTCAACCACCTCAGAAGGCTATCCAAGCCTATCACAAGTCTACACAACATCCGGT

<u>Mapping of RSV-A with the 72-nt Duplication Against NC_038235 (Strain A2)</u>

Finally, the RSV-A sequence was mapped against NC_038235, strain A2, which lacks the 72-nt duplication. In this case, significant soft clipping was observed in the region corresponding to the original copy (copy 1). This is clearly shown in the figure below, where the clipped regions are indicated by a light gray strikethrough. The coverage in the copy 1 region was sparse, and no reads fully span the copy 1 region.

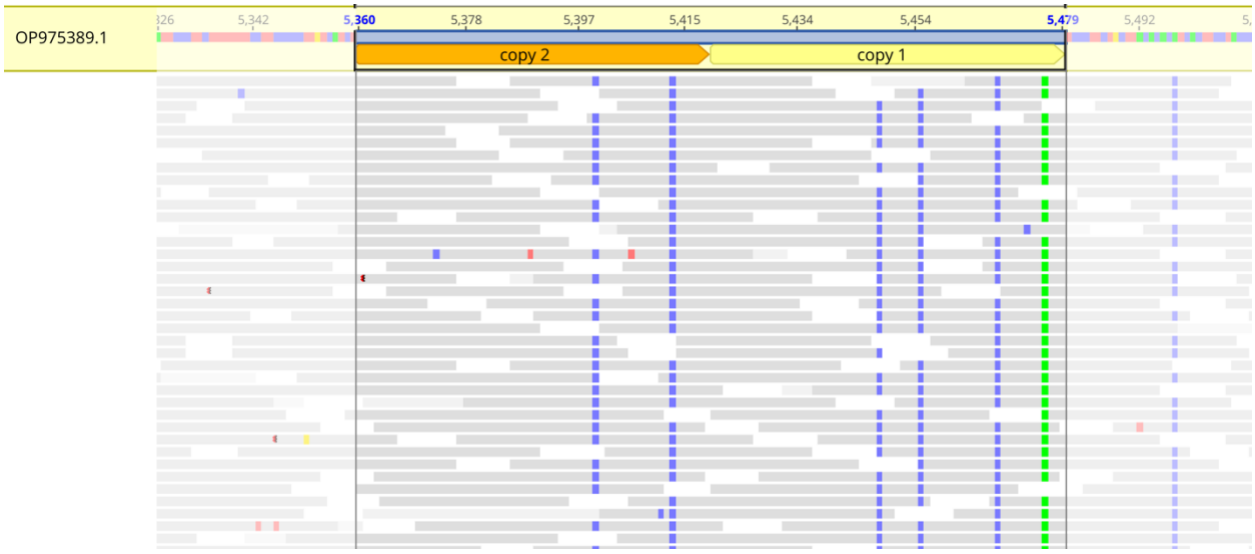
Furthermore, the zoomed-out figure below highlights the absence of read depth in the region corresponding to the 72-nt duplication in this reference-based assembly. After the duplication region, no additional coverage is detected, further supporting the absence of the duplication in this reference.

RSV-B

The following describes the results of mapping RSV-B FASTQ sequences against three different reference genomes.

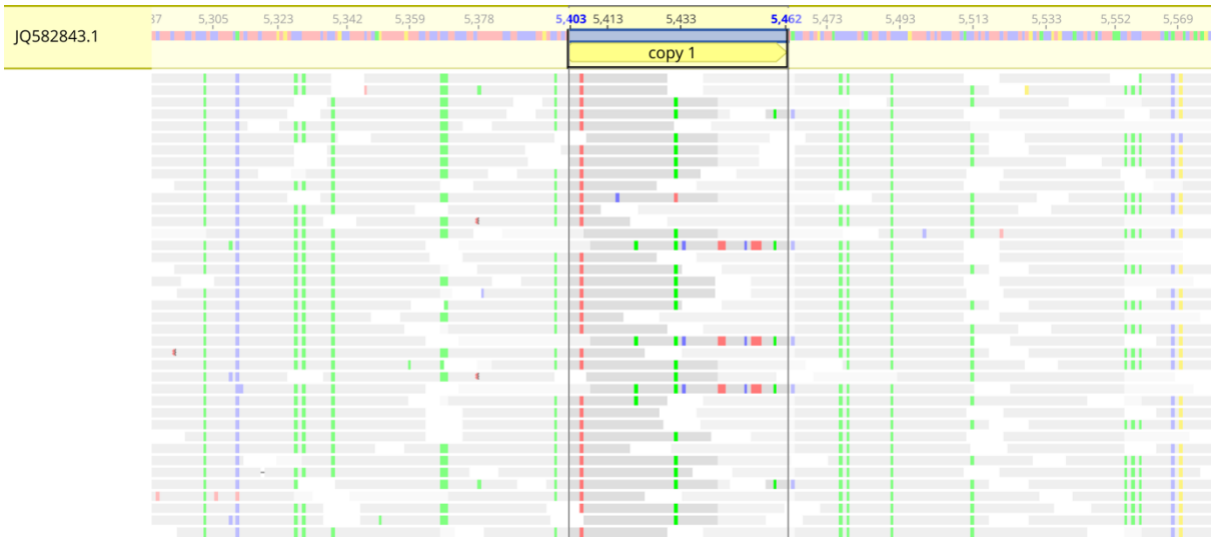
Mapping of RSV-B with the 60-nt Duplication Against OP975389.1

When mapping RSV-B containing the 60-nt duplication against **OP975389.1**, the alignment shows no evidence of bias. Sequencing reads map uniformly across the entire reference, with some reads spanning the intersection of the two duplicated copies. This indicates that both copies of the duplication are well represented, as illustrated in the figure below.



Mapping of RSV-B with the 60-nt Duplication Against JQ582843.1

In contrast, when mapping RSV-B against JQ582843.1, which lacks the 60-nt duplication, the alignment shows a distinctive pattern. Reads in the region of the duplication are primarily short and do not extend before or after the duplicated region, as shown in the next figure by the light gray region in some reads mainly at the 3' end of the copy 1 region. This suggests that the duplication region is poorly covered in this mapping. In addtiion, many of the reads in this region are soft-clipped at the location of the duplication.



Mapping of RSV-B with the 60-nt Duplication Against NC_001781 (Strain B1)

Finally, when RSV-B was mapped against NC_001781 (RSV-B strain B1), which also lacks the 60-nt duplication, a similar pattern to the JQ582843.1 alignment was observed. The region corresponding to the duplication contains primarily short reads, which do not extend before or after the duplication, as shown in the figure below. Moreover, two nucleotides at specific positions appear to be a mixture of the nucleotides from copy 1 and copy 2.

This indicates that the reads are capturing sequence information from both copies of the duplicated region, even though the reference genome does not contain the duplication.

Reference-guided de novo assemblies

The use of hybrid methodologies—first performing de novo assembly and then mapping the resulting contigs to a reference—allows for more accurate detection of nucleotide duplications in RSV genomes.

In the case of RSV-A, the de novo assembly of the FASTQ reads generated contigs of up to 3700-nt length, which, when mapped against OR795328.1 (a reference genome without the 72-nt duplication), correctly identified the presence of a 72-nt insertion. This allowed for the accurate annotation of the duplication in the consensus sequence.

It is important to note that the thresholds used for reference-guided de novo assemblies should not be the same as those used for read mapping against a reference genome. In a hybrid assembly approach, the concept of coverage depth differs, as the data now consists of contigs rather than individual reads. Each contig represents the consensus of hundreds or even thousands of reads. Therefore, a position mapped by a single contig is still considered sufficiently accurate for

annotation. This highlights the need for careful consideration of coverage metrics when interpreting results from hybrid assemblies.