

Supplemental information

Scalable regional genomic surveillance reveals previously unrecognized SARS-CoV-2 recombinant diversity and limited onward transmission

Ana Tzvetkova, Anne Pohlmann, Kathrin Lehmann, Tilo Sasse, Veronika Balau, Martin Beer, Lars Kaderali, Katja Verena Goller, Nils-Olaf Hübner, Karsten Becker and Christian Kohler

Contents

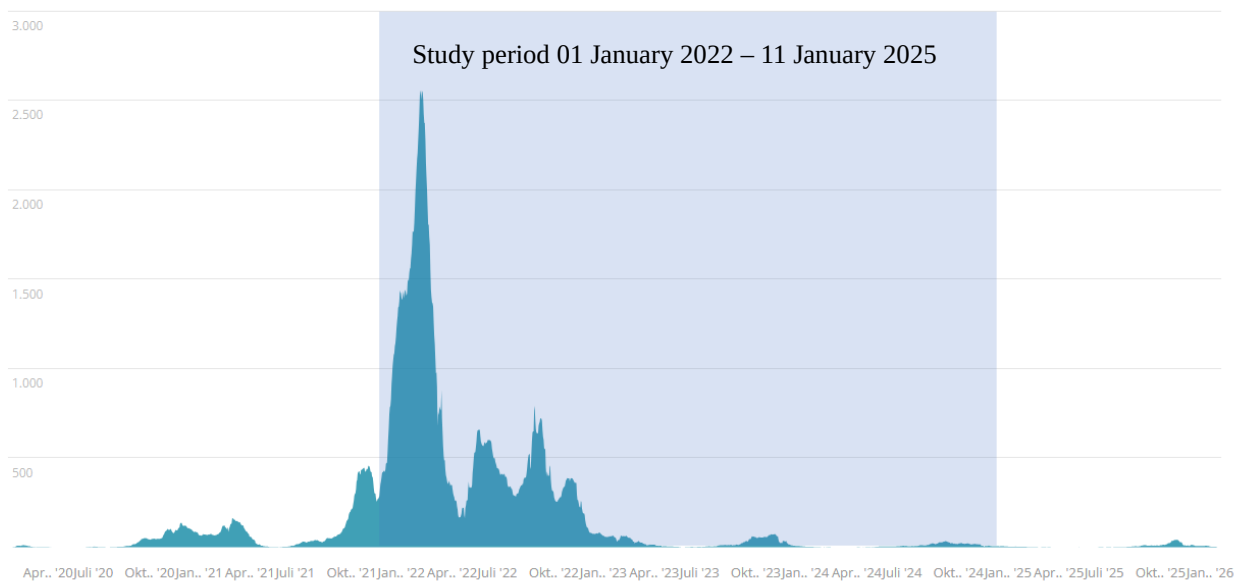
Supplemental figures S1-S4

Supplemental Methods

Supplemental Figure S1

COVID-19 7-day incidence in Mecklenburg-Western Pomerania (M-V)

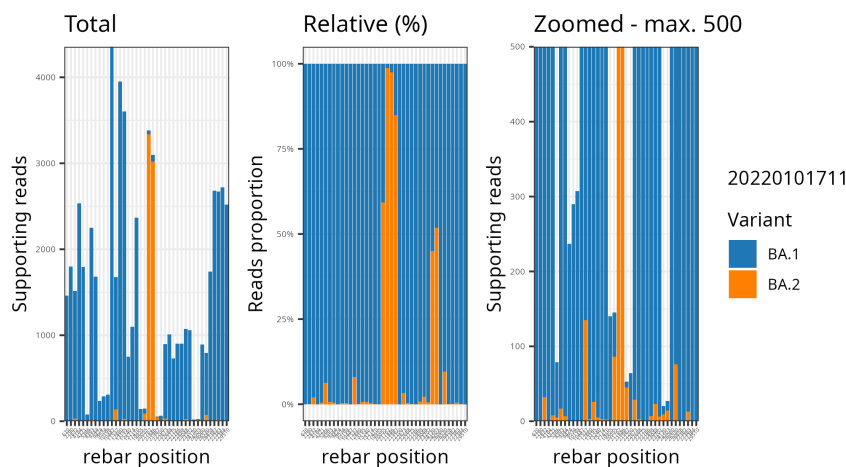
Seven-day COVID-19 incidence in Mecklenburg–Western Pomerania (M-V) from April 2020 to January 2026. The study period (1 January 2022 to 11 January 2025) is highlighted in light blue. The figure was adapted from data obtained from <https://corona-zahlen-heute.de/deutschland/mecklenburg-vorpommern/> (date 25 February 2026). Use of the figure was permitted and complies with applicable copyright regulations.



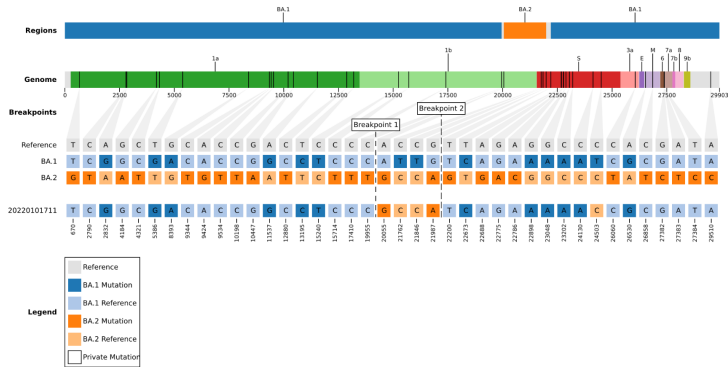
Detection of putative Novel Recombinant Lineages (NRLs).

Shown are results from REBAR and raw read-level analyses for all putative novel recombinant lineages (NRLs) classified as true recombinants, as well as twelve circulating recombinant circulating lineages (CRLs) as controls. REBAR (bottom panel) identifies lineage-specific nucleotide positions characteristic of the putative parental lineages, assigns them to their respective genomic backgrounds (blue and orange), and infers recombination breakpoints based on transitions between lineage-defining patterns. These signals were validated using raw read-level analyses (top panel), in which variant frequencies at each genomic position were quantified from BAM file data to distinguish co-infections from true recombination events. For each position, the total number of reads (Total; zoomed view up to 500 reads) and corresponding relative variant frequencies (%) are shown. Variants occurring at frequencies >90% were considered the consensus base at the respective genomic position.

Sample: 20220101711



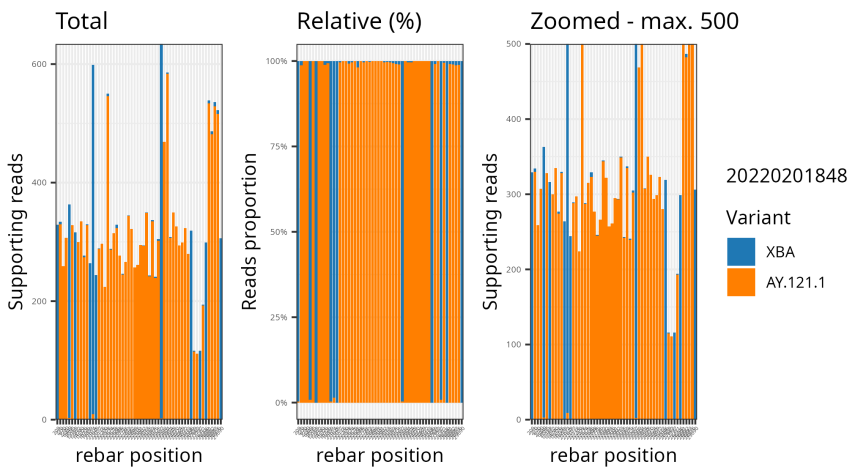
Raw-Read-analysis



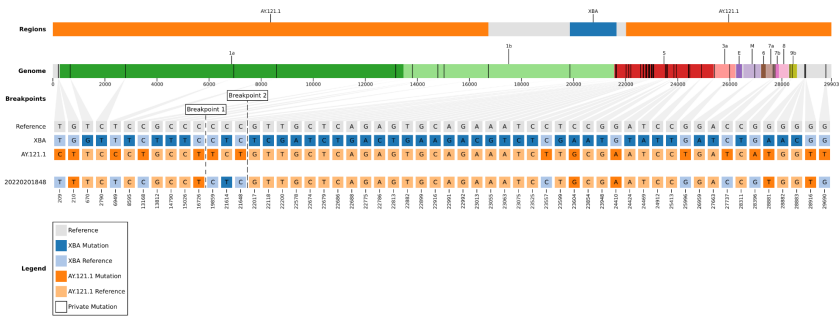
REBAR analysis

Supplemental Figure S2

Sample: 20220201848

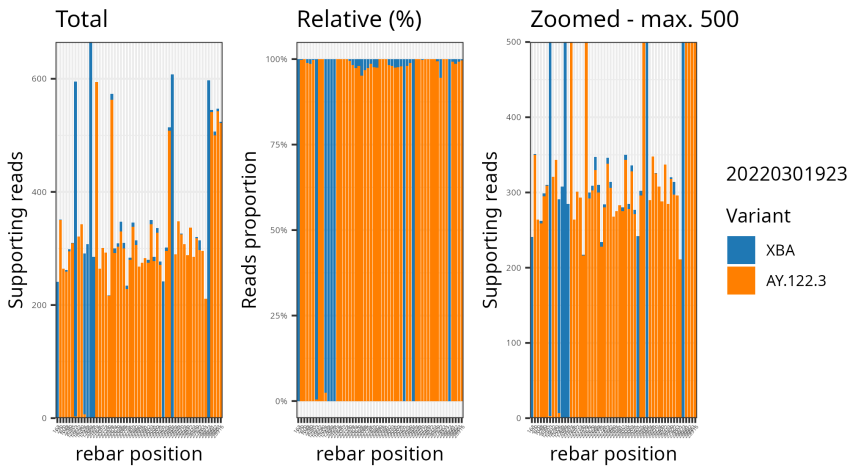


Raw-Read-analysis

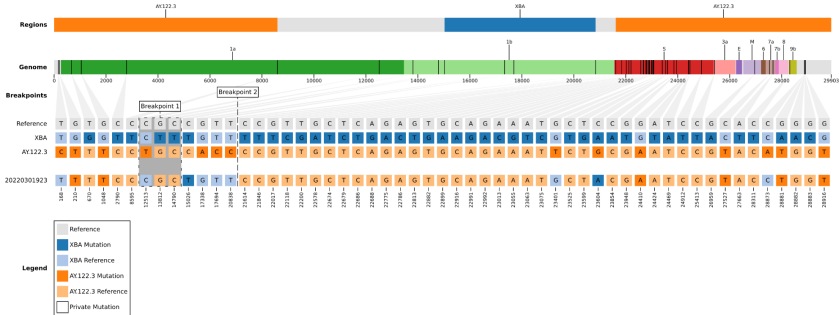


REBAR analysis

Sample: 20220301923



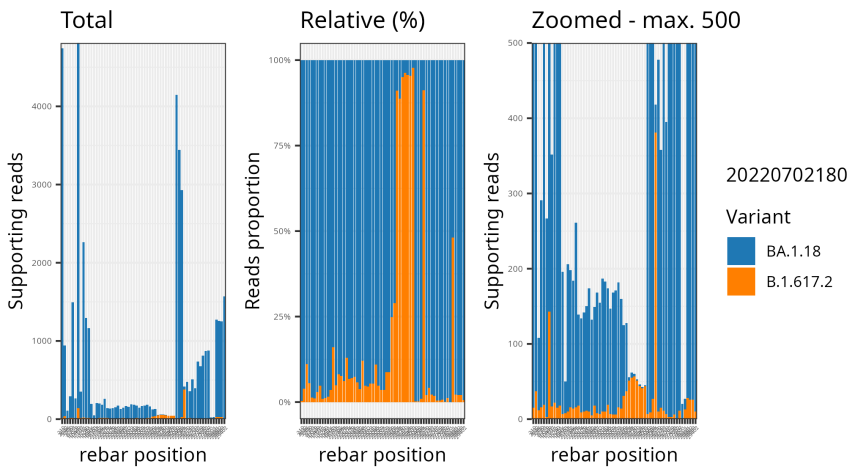
Raw-Read-analysis



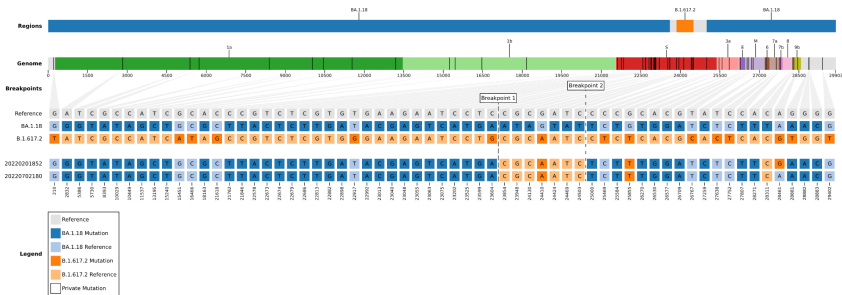
REBAR analysis

Supplemental Figure S2

Sample: 20220702180

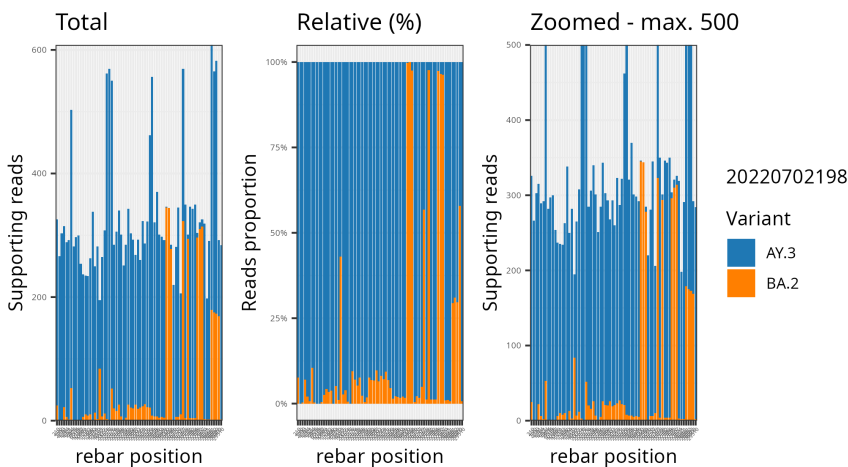


Raw-Read-analysis

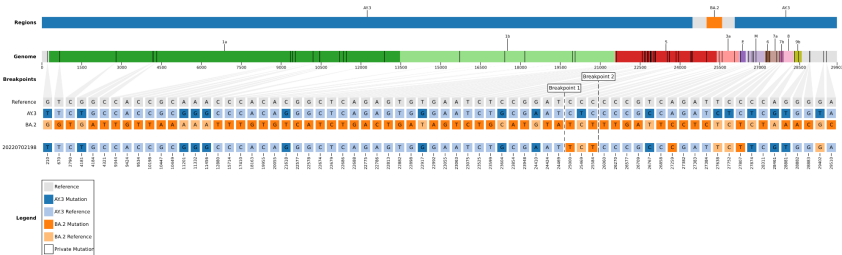


REBAR analysis

Sample: 20220702198



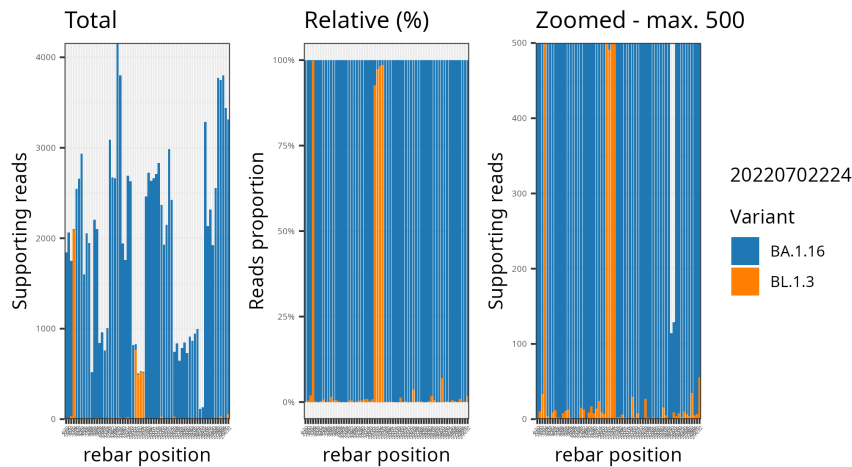
Raw-Read-analysis



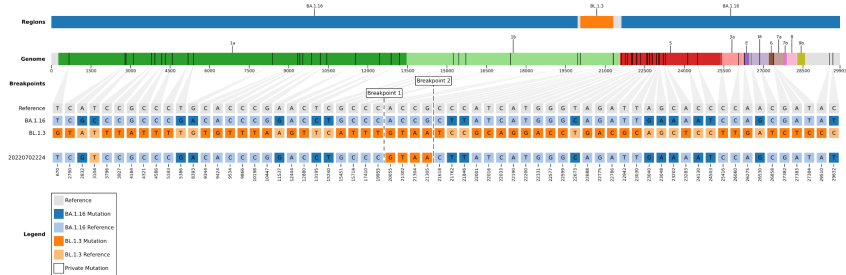
REBAR analysis

Supplemental Figure S2

Sample: 20220702224

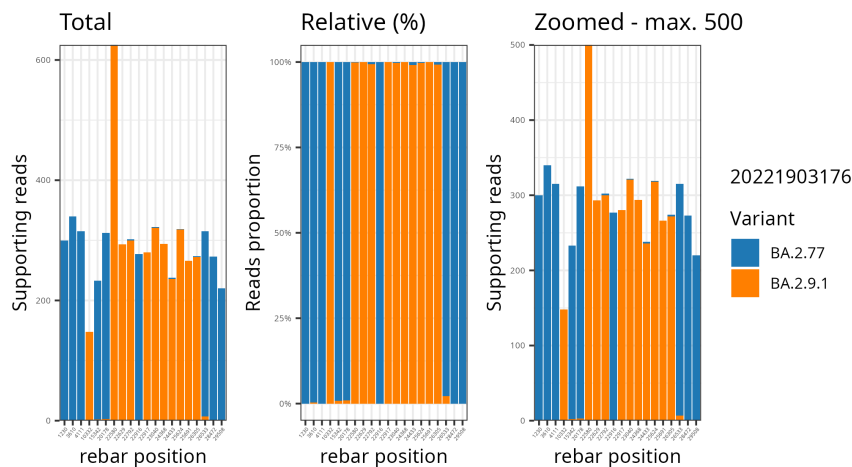


Raw-Read-analysis

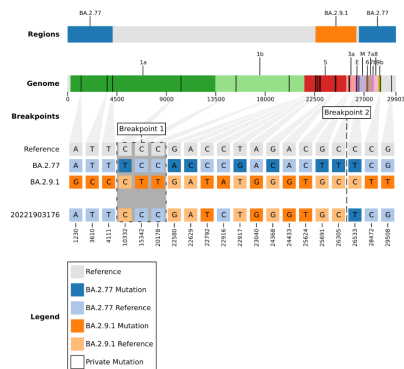


REBAR analysis

Sample: 20221903176



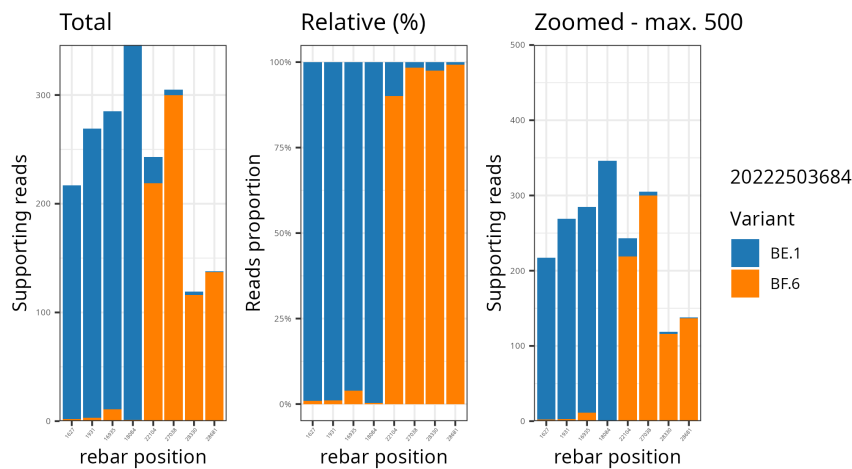
Raw-Read-analysis



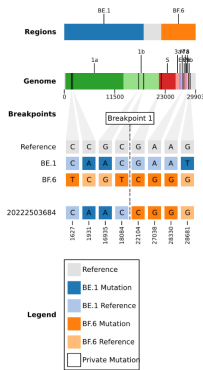
REBAR analysis

Supplemental Figure S2

Sample: 20222503684

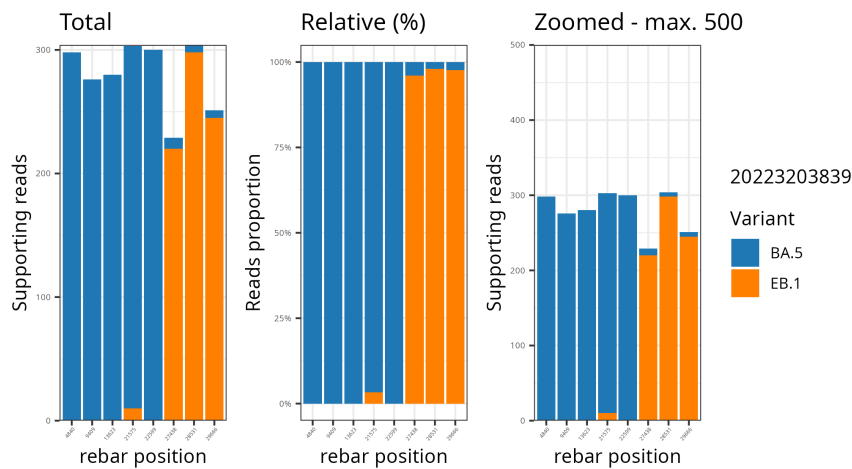


Raw-Read-analysis

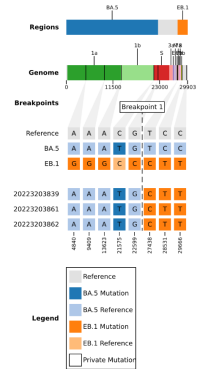


REBAR analysis

Sample: 20223203839



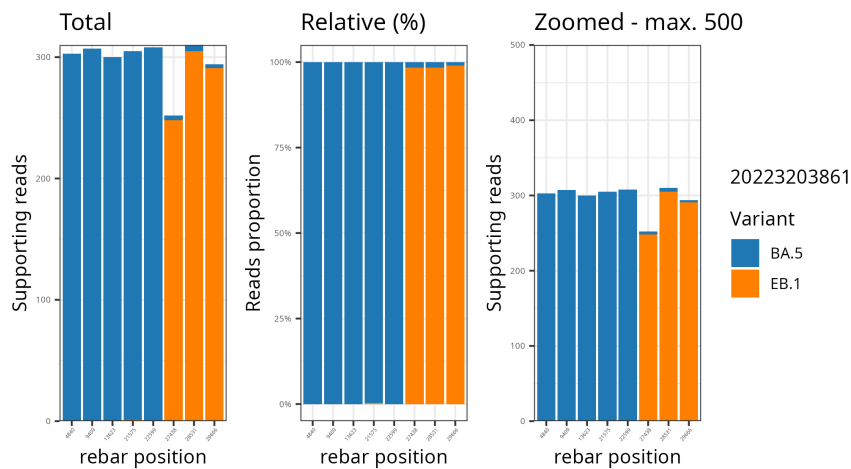
Raw-Read-analysis



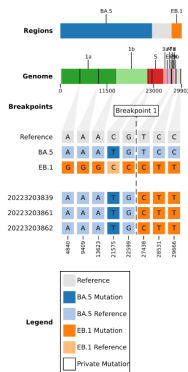
REBAR analysis

Supplemental Figure S2

Sample: 20223203861

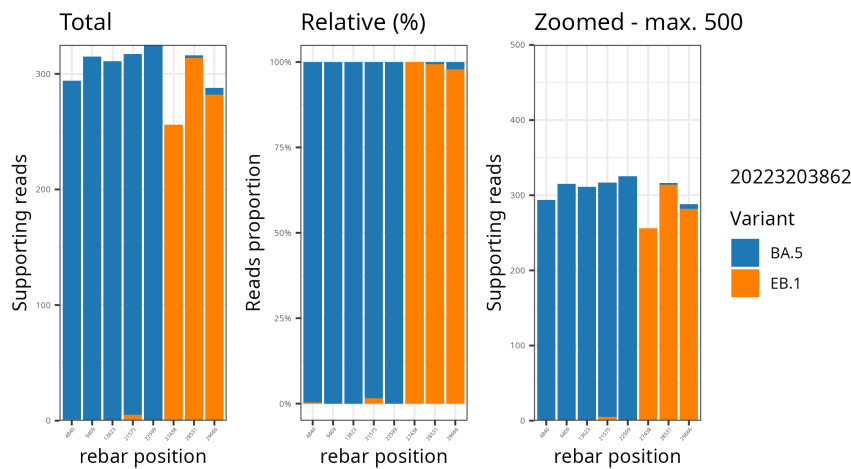


Raw-Read-analysis

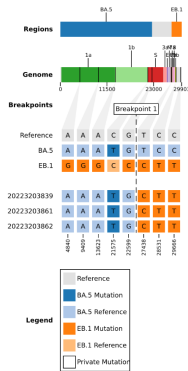


REBAR analysis

Sample: 20223203862



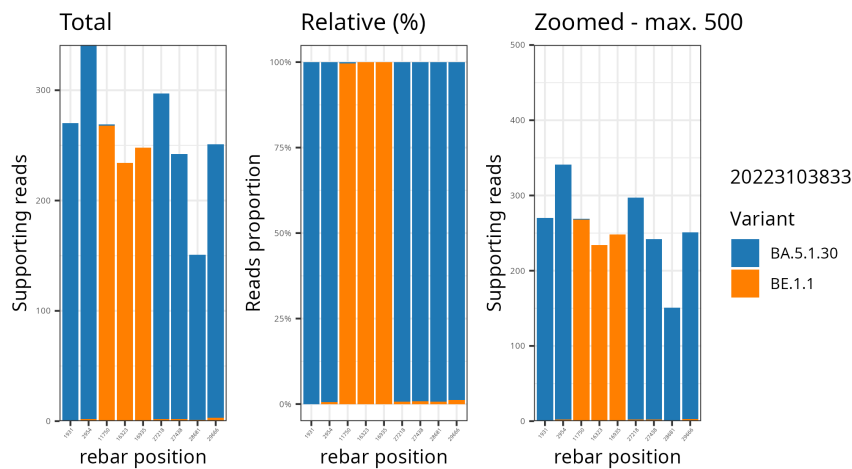
Raw-Read-analysis



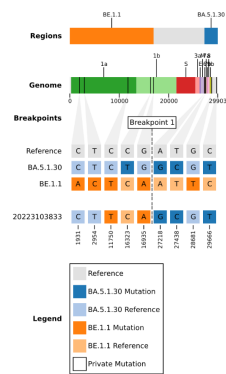
REBAR analysis

Supplemental Figure S2

Sample: 20223103833

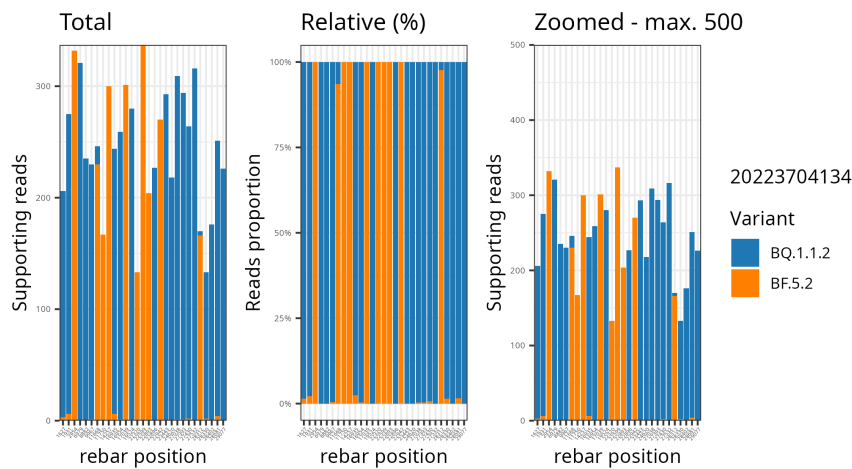


Raw-Read-analysis

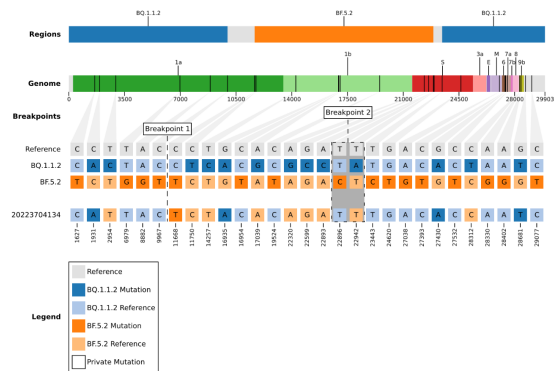


REBAR analysis

Sample: 20223704134



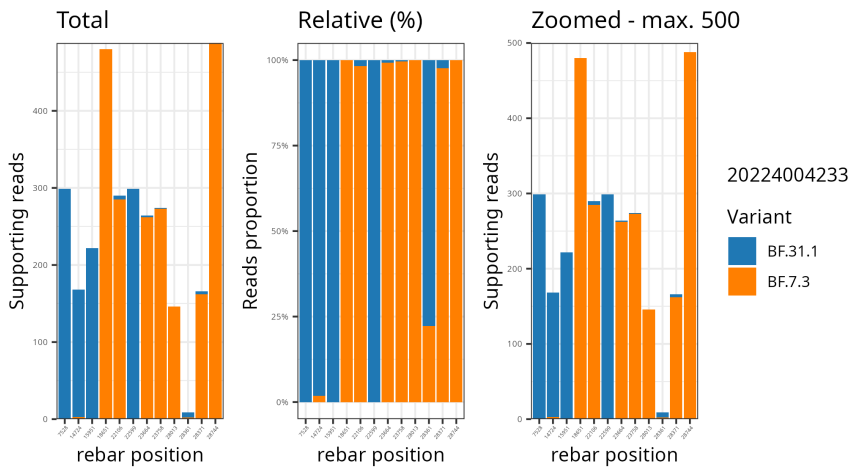
Raw-Read-analysis



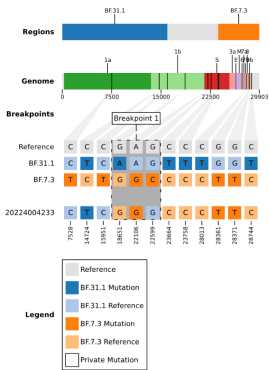
REBAR analysis

Supplemental Figure S2

Sample: 20224004233

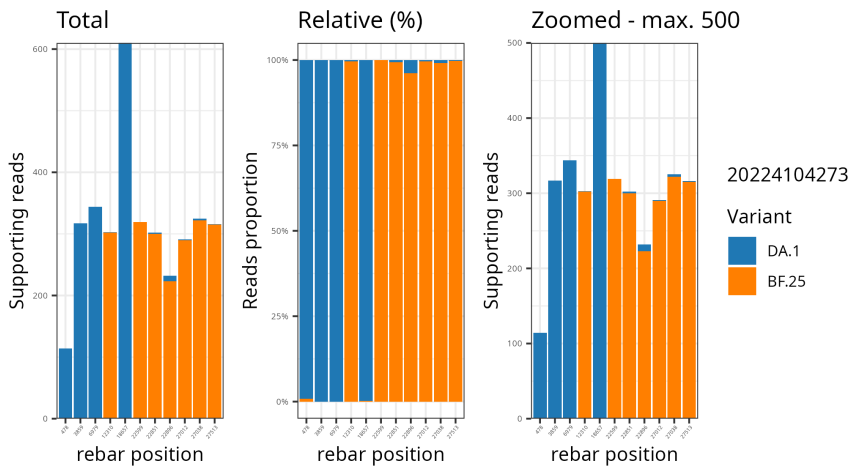


Raw-Read-
analysis

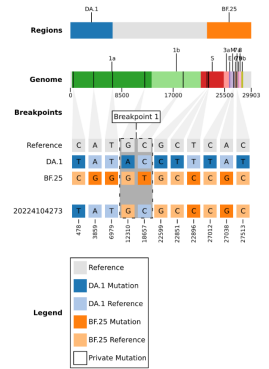


REBAR
analysis

Sample: 20224104273



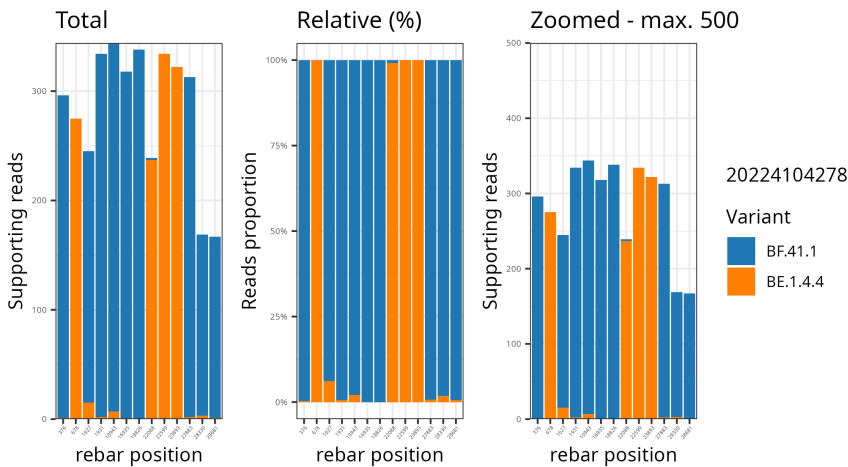
Raw-Read-
analysis



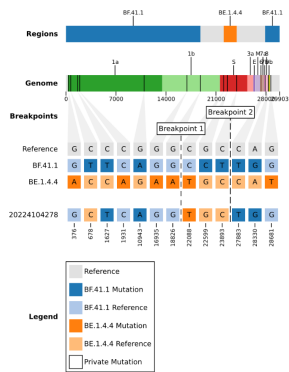
REBAR
analysis

Supplemental Figure S2

Sample: 20224104278

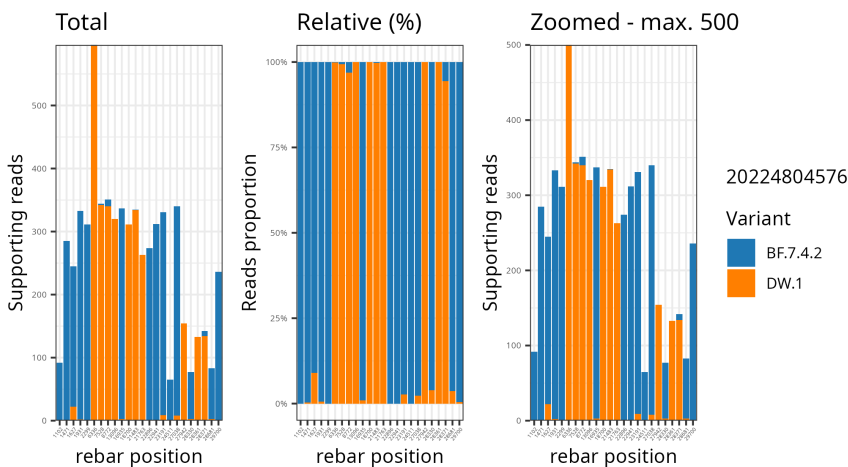


Raw-Read-analysis

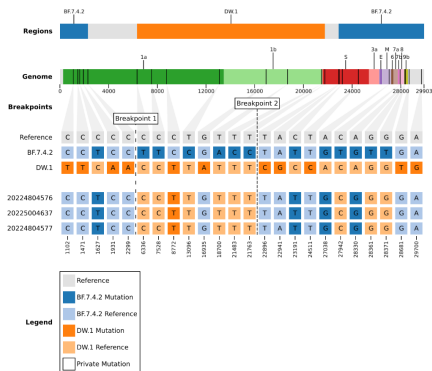


REBAR analysis

Sample: 20224804576



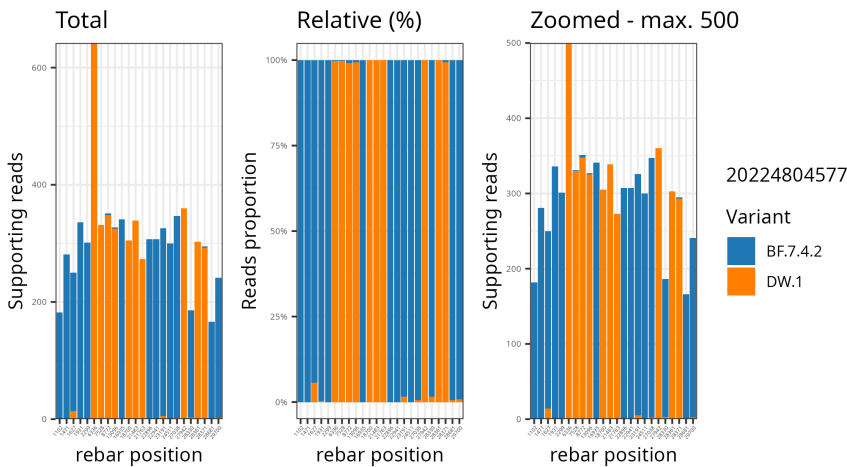
Raw-Read-analysis



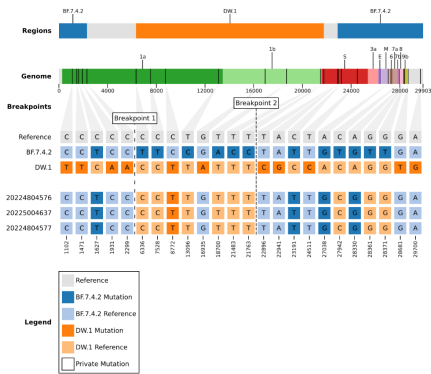
REBAR analysis

Supplemental Figure S2

Sample: 20224804577

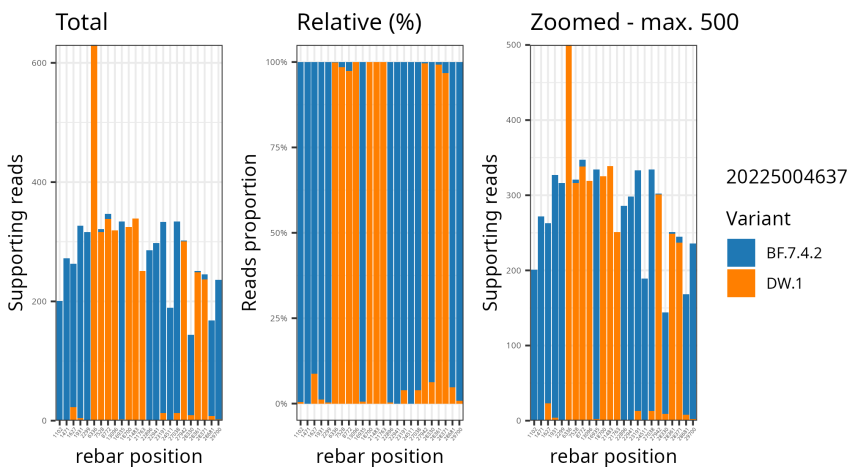


Raw-Read-analysis

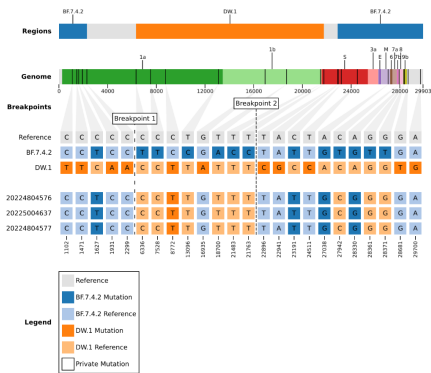


REBAR analysis

Sample: 20225004637



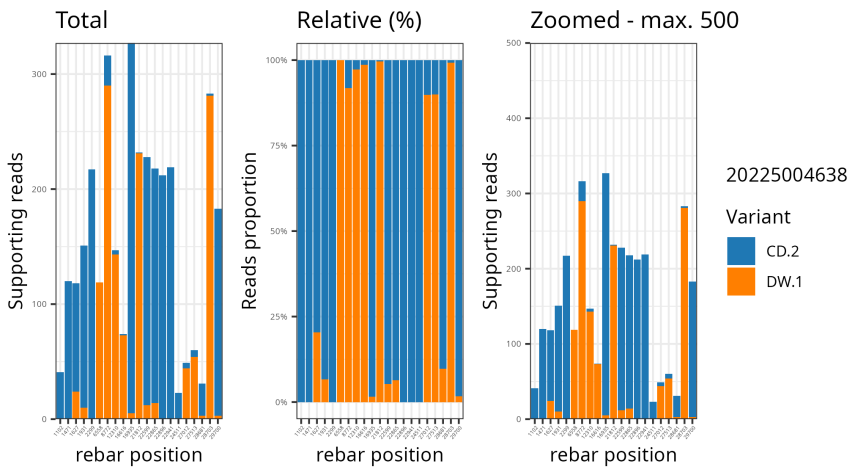
Raw-Read-analysis



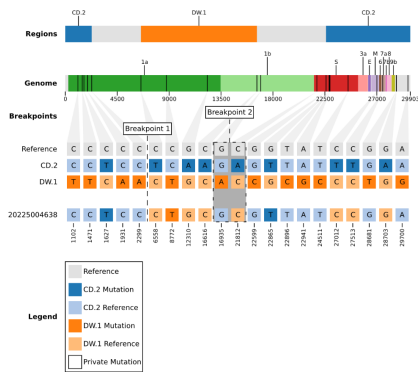
REBAR analysis

Supplemental Figure S2

Sample: 20225004638

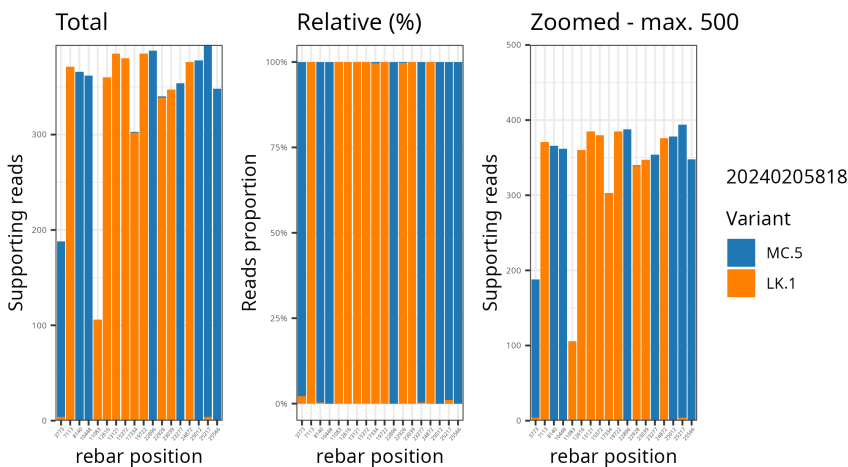


Raw-Read-analysis

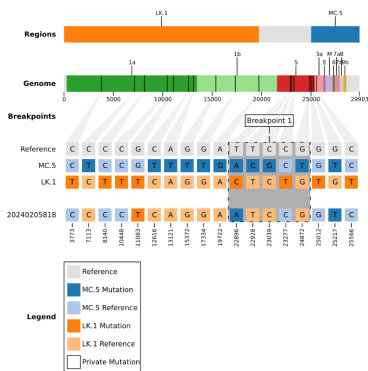


REBAR analysis

Sample: 20240205818



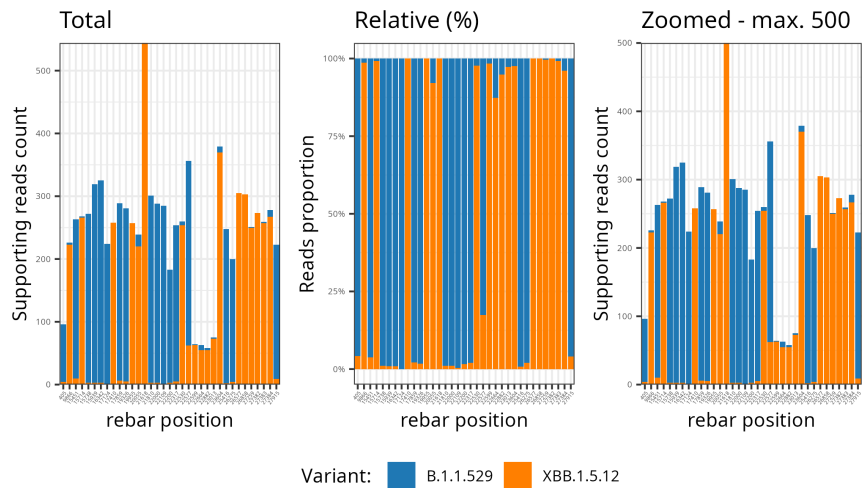
Raw-Read-analysis



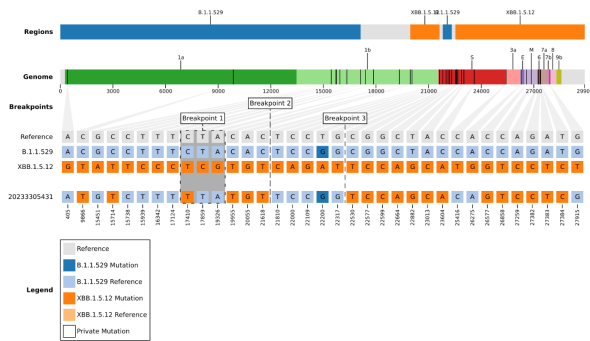
REBAR analysis

Supplemental Figure S2

Sample: 20233305431

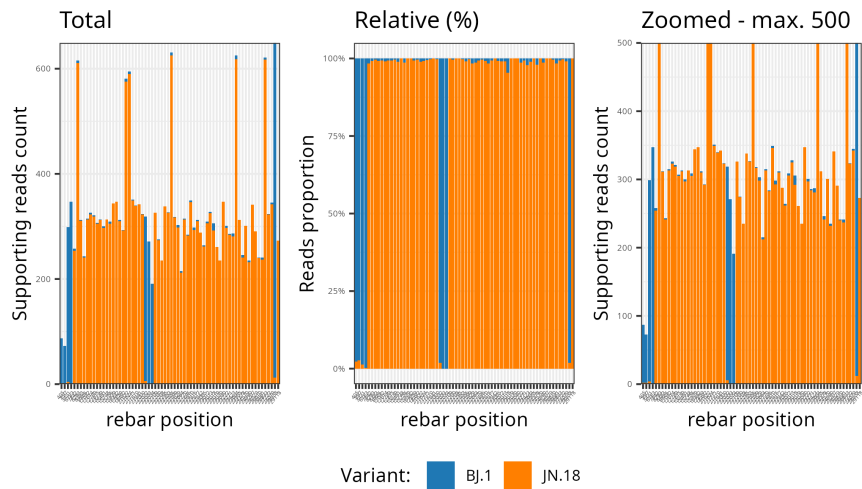


Raw-Read-
analysis

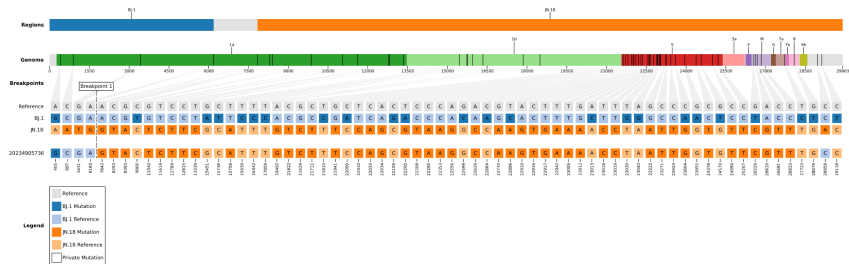


REBAR
analysis

Sample: 20240205736



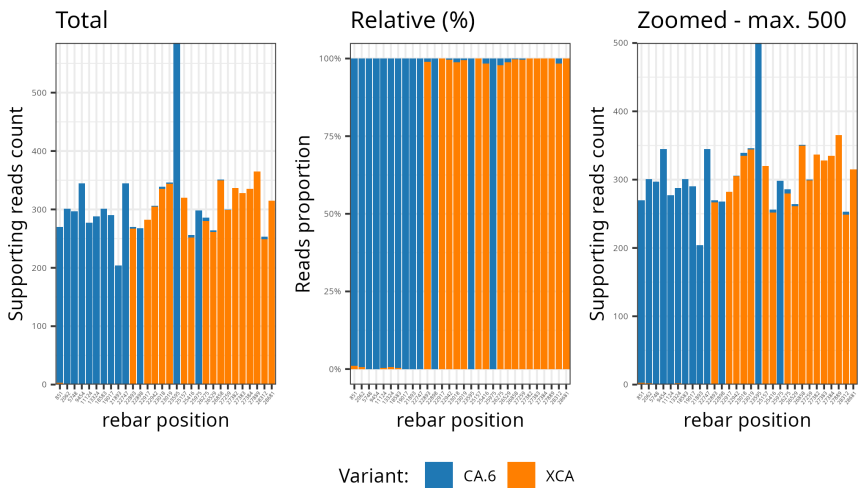
Raw-Read-
analysis



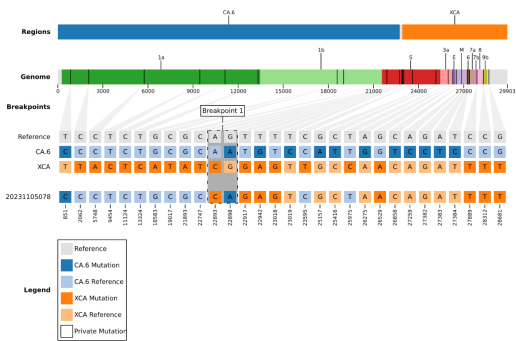
REBAR
analysis

Supplemental Figure S2

Sample: 20231105078



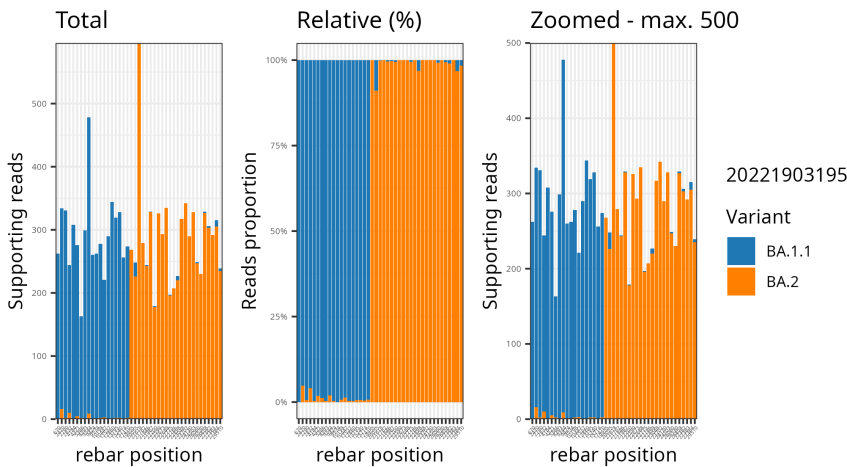
Raw-Read-
analysis



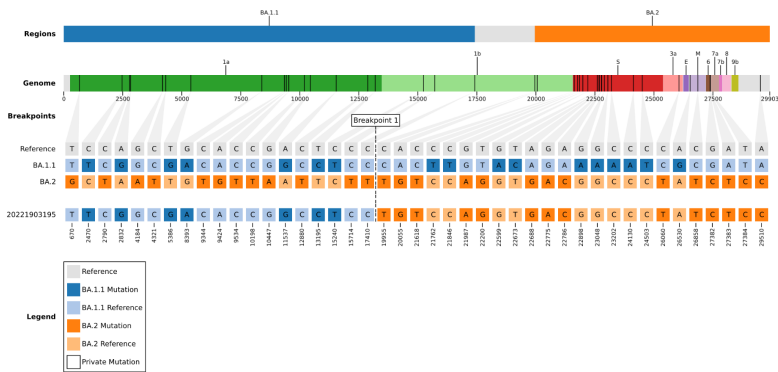
REBAR
analysis

Supplemental Figure S2

Control XAL: 20221903195

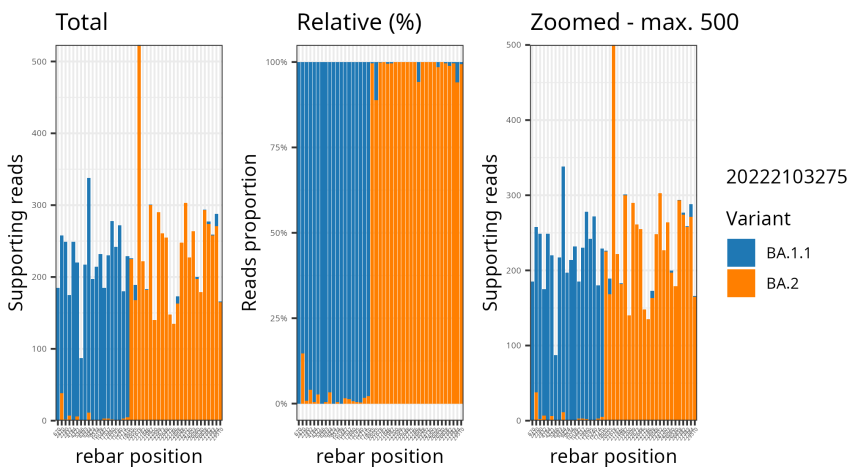


Raw-Read-analysis

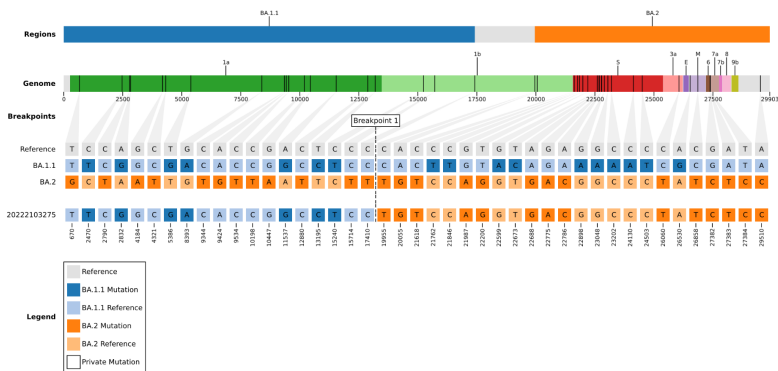


REBAR analysis

Control XM: 20222103275



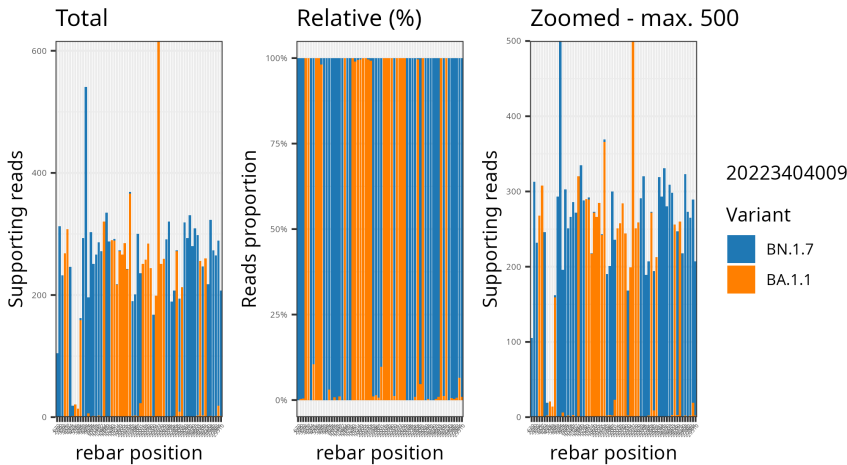
Raw-Read-analysis



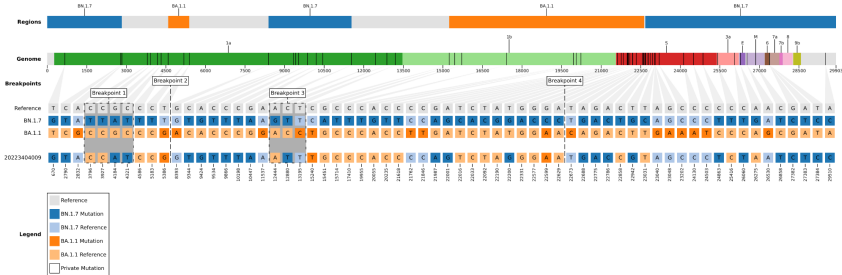
REBAR analysis

Supplemental Figure S2

Control XAK: 20223404009

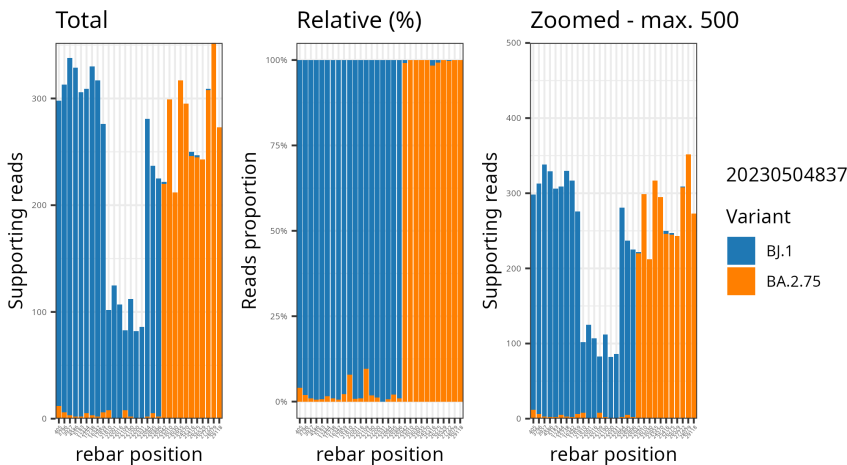


Raw-Read-analysis

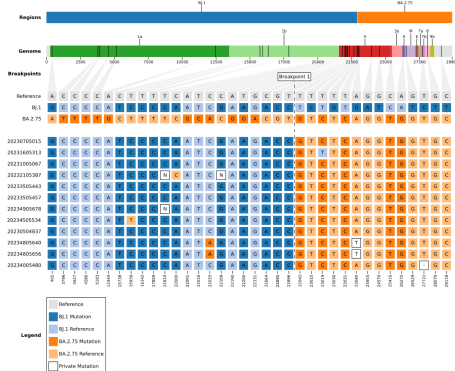


REBAR analysis

Control XBB: 20230504837



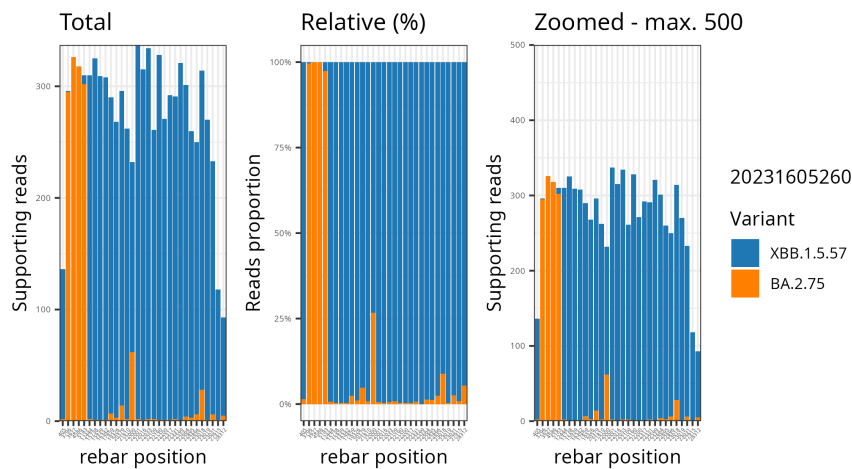
Raw-Read-analysis



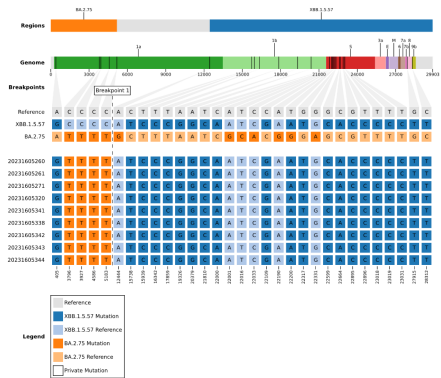
REBAR analysis

Supplemental Figure S2

Control XBL: 20231605260

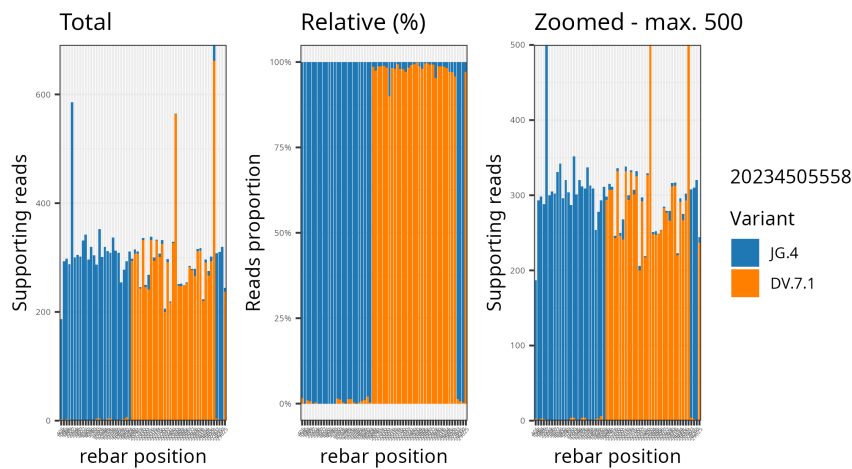


Raw-Read-analysis

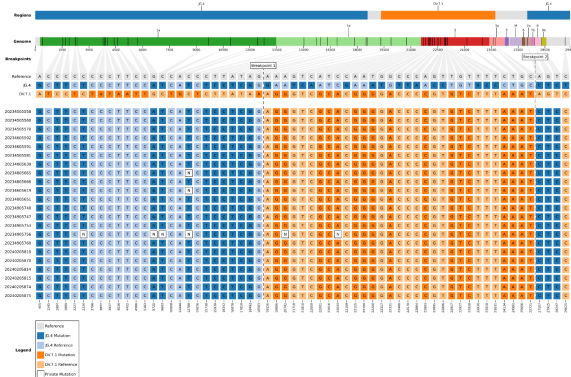


REBAR analysis

Control XCT.1:
20234505558



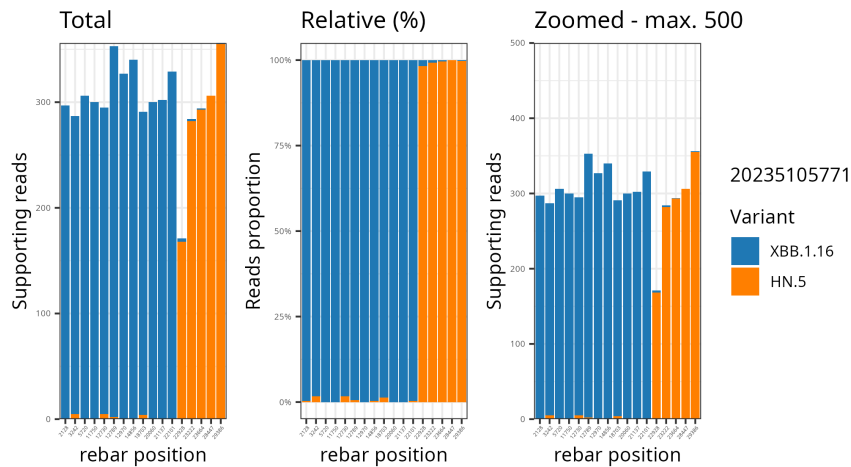
Raw-Read-analysis



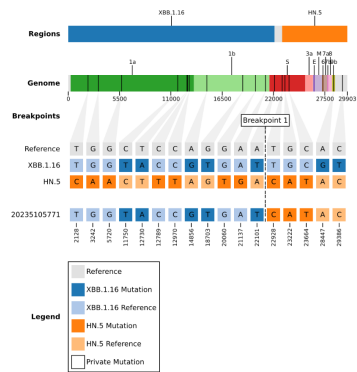
REBAR analysis

Supplemental Figure S2

Control XDA.1: 20235105771

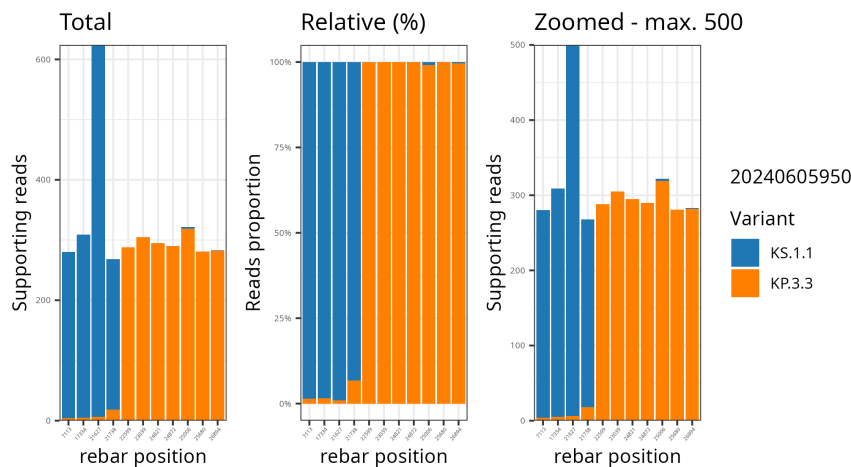


Raw-Read-analysis



REBAR analysis

Control XEC: 20240605950

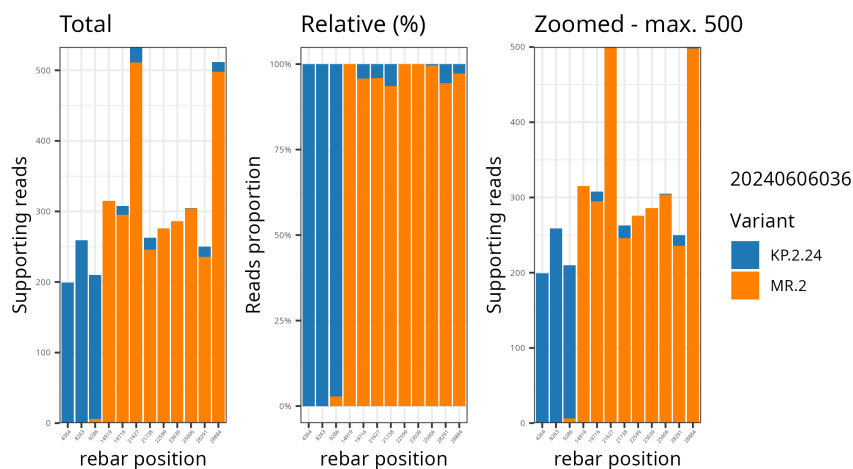


Raw-Read-analysis

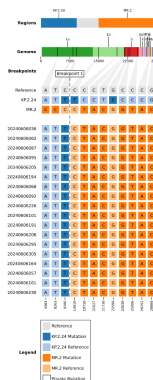


REBAR analysis

Control XEK: 20240606036

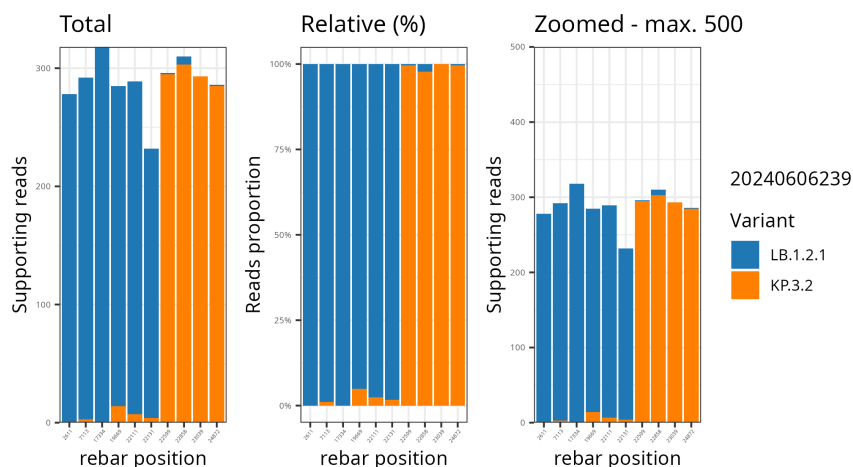


Raw-Read-analysis

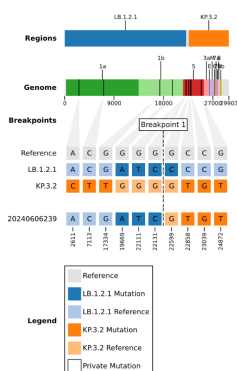


REBAR analysis

Control XDY: 20240606239



Raw-Read-analysis



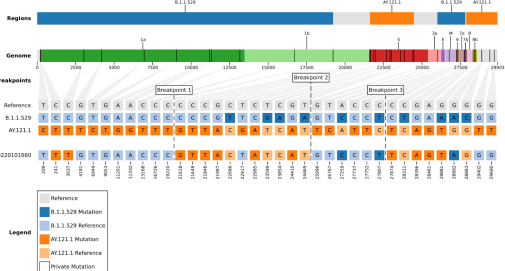
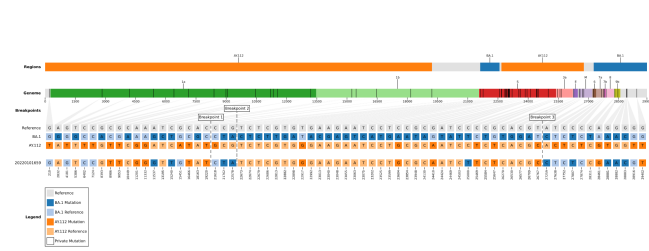
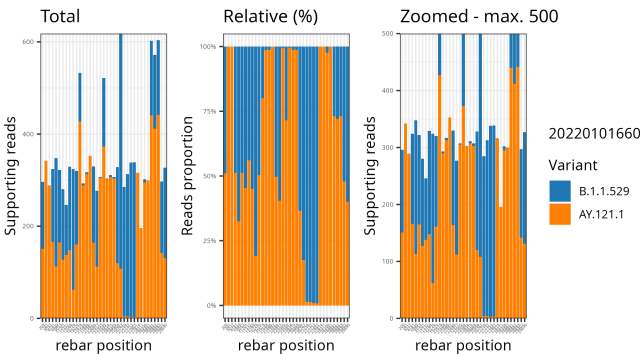
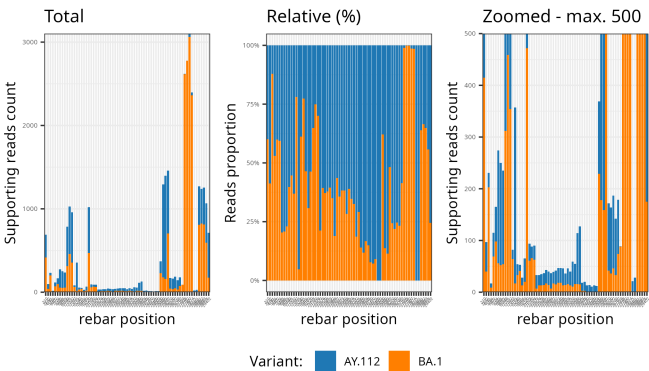
REBAR analysis

Detection of co-infections, contamination, and ambiguous sequences.

Shown are results from REBAR and raw read-level analyses for all putative co-infection or contamination cases. REBAR (bottom panel) identifies lineage-specific nucleotide positions characteristic of the putative parental lineages, assigns them to their respective genomic backgrounds (blue and orange), and infers putative recombination breakpoints based on transitions between lineage-defining patterns. These signals were validated using raw read-level analysis (top panel), in which variant frequencies at each genomic position were quantified from BAM file data to distinguish co-infections from true recombination events. For each position, the total number of reads (Total; zoomed view up to 500 reads) and corresponding relative variant frequencies (%) are shown. Variants with frequencies >90% were considered the consensus base. Samples were classified as potential co-infections and/or contamination events (designated “coco”) when more than 25% of the combined lineage-defining positions of the two putative parental lineages identified by REBAR exhibited intermediate allele frequencies (>10% and <90%). Genomic regions were classified as “unclear” when a consistent variant signal was present but sequencing depth was low (<40 reads). These regions are indicated accordingly. Overall, several samples showed patterns consistent with either co-infection or contamination, with key genomic positions exhibiting insufficient read coverage and therefore classified as “coco” or “unclear”.

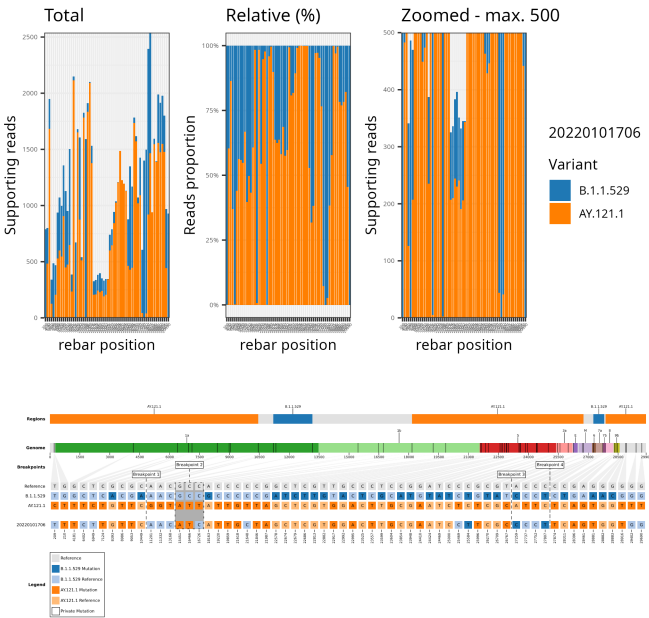
sample: 20220101659-coco/unclear

sample: 20220101660-coco

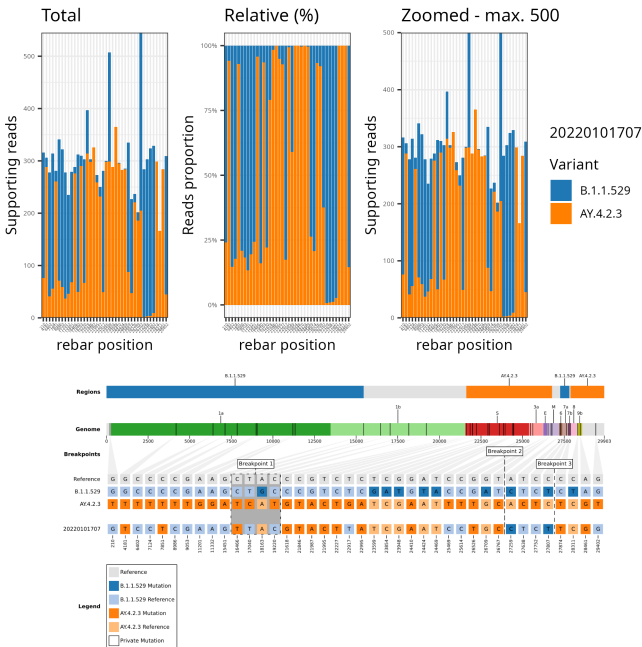


Supplemental Figure S3

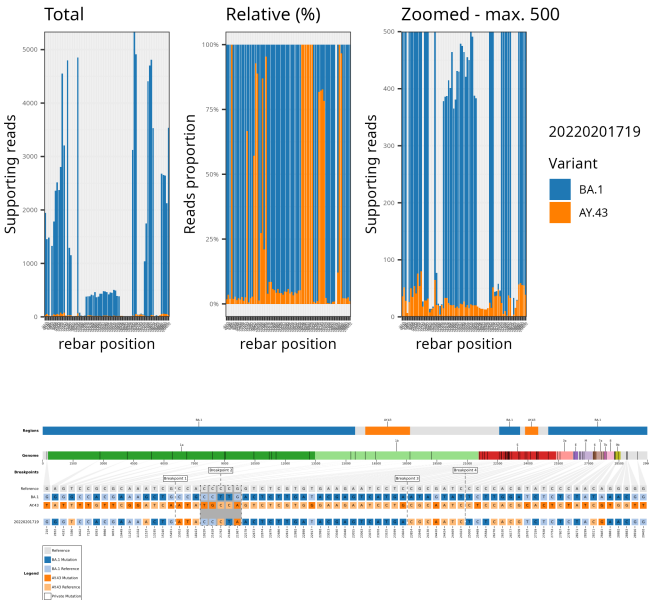
sample: 20220101706-coco



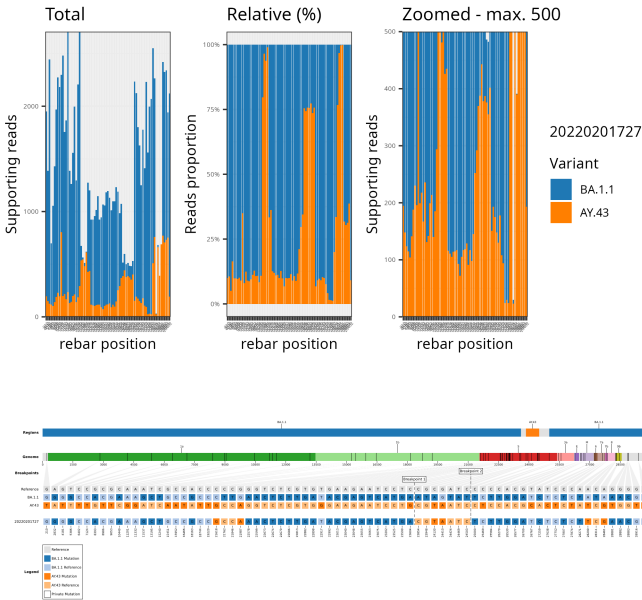
sample: 20220101707-coco



sample: 20220201719-unclear



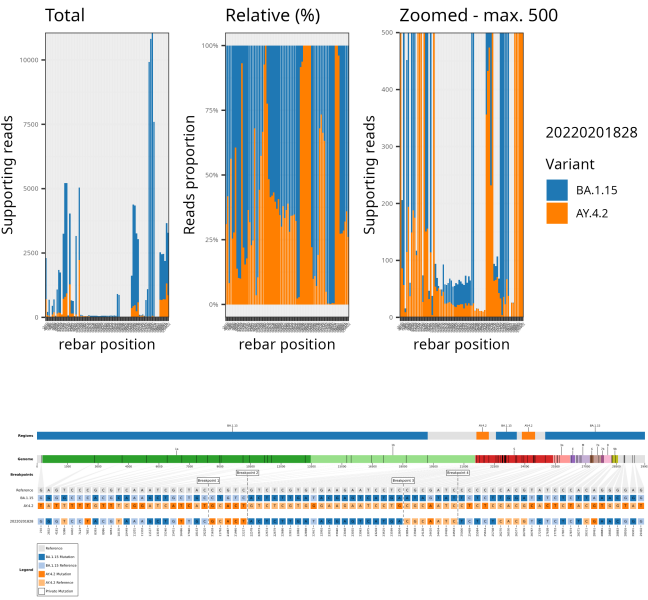
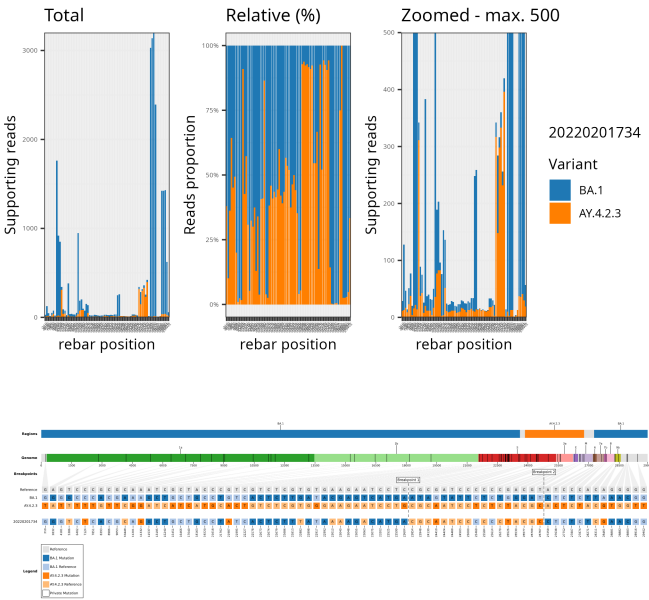
sample: 20220201727-coco



Supplemental Figure S3

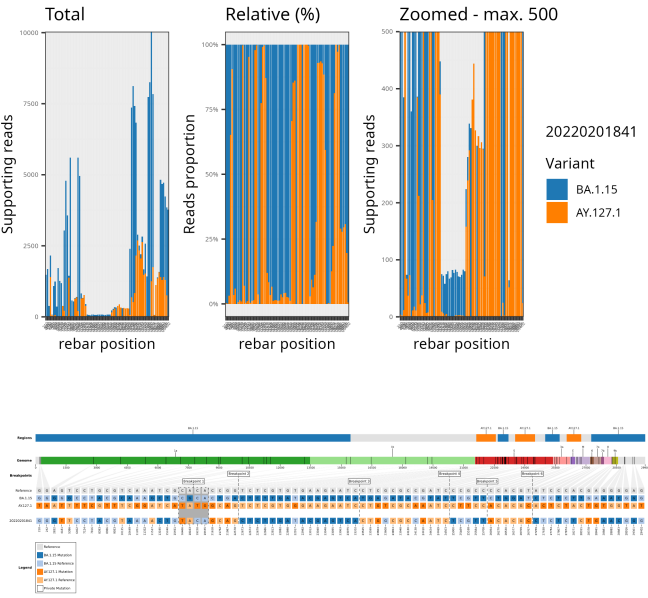
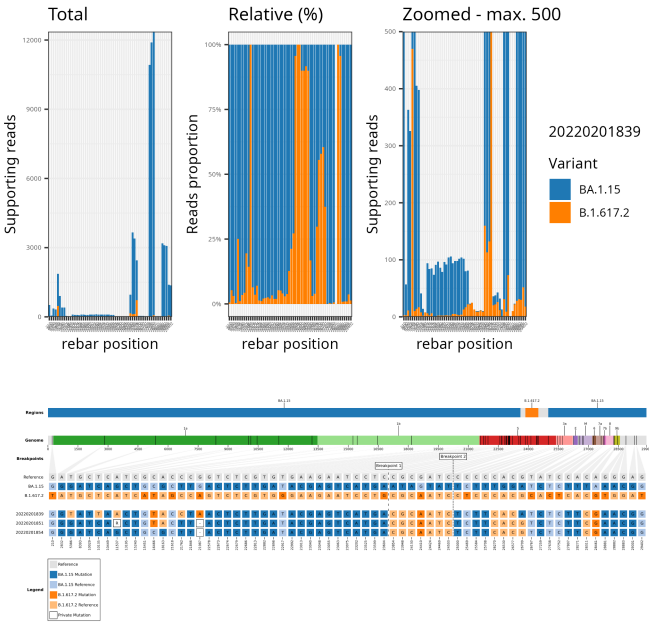
sample: 20220201734-coco/unclear

sample: 20220201828-coco/unclear



sample: 20220201839-coco/unclear

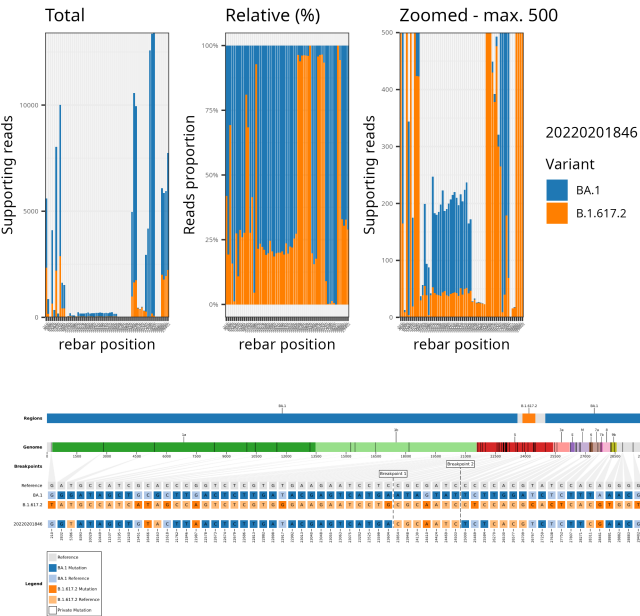
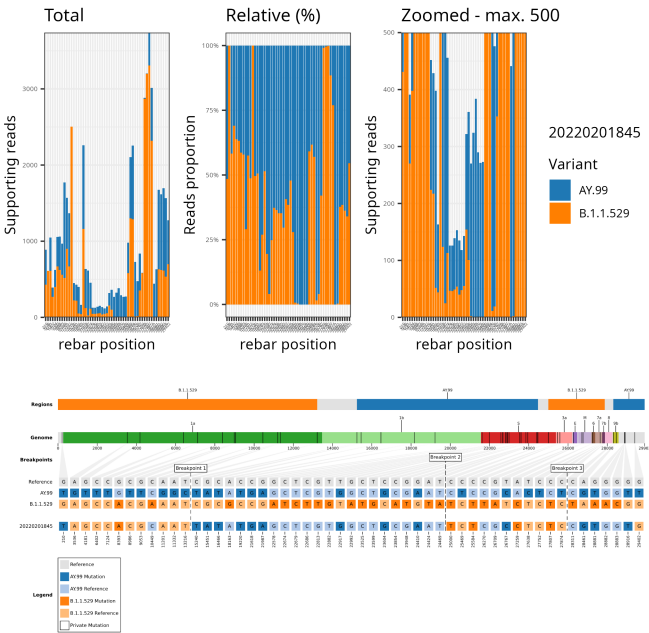
sample: 20220201841-coco



Supplemental Figure S3

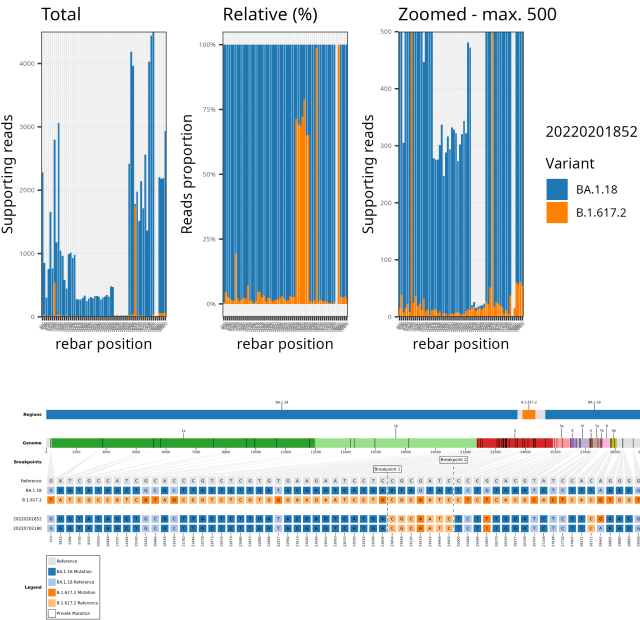
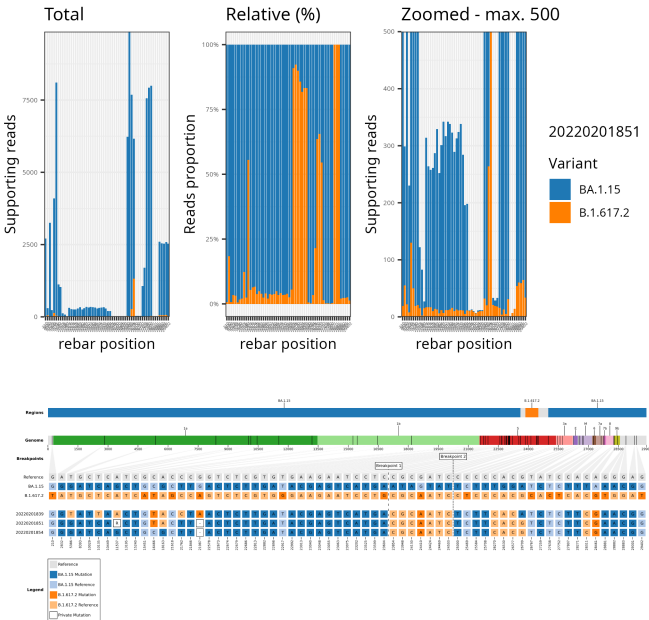
sample: 20220201845-coco

sample: 20220201846-coco/unclear



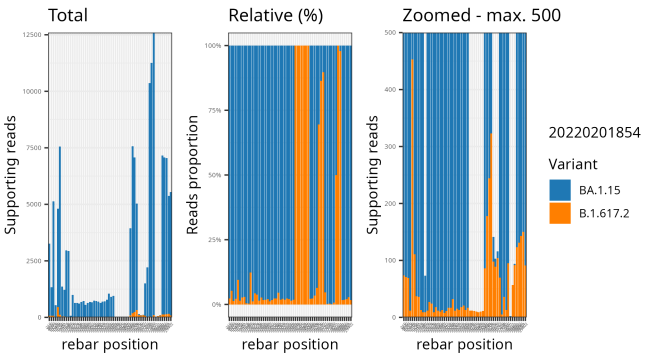
sample: 20220201851-unclear

sample: 20220201852-unclear

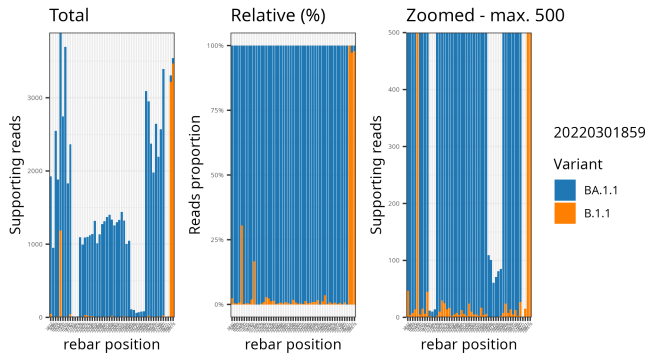


Supplemental Figure S3

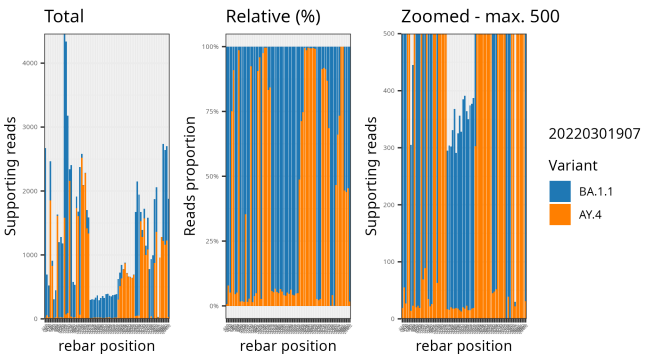
sample: 20220201854-unclear



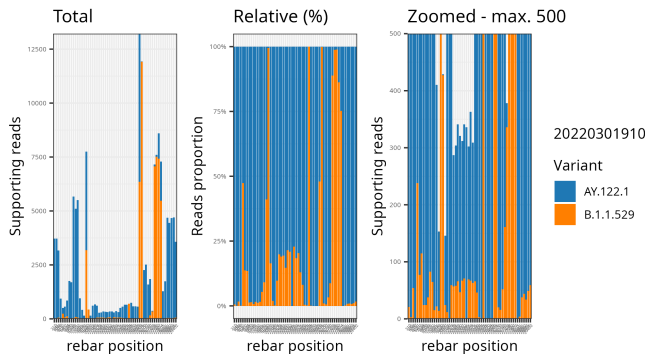
sample: 20220301859-unclear



sample: 20220301907-coco

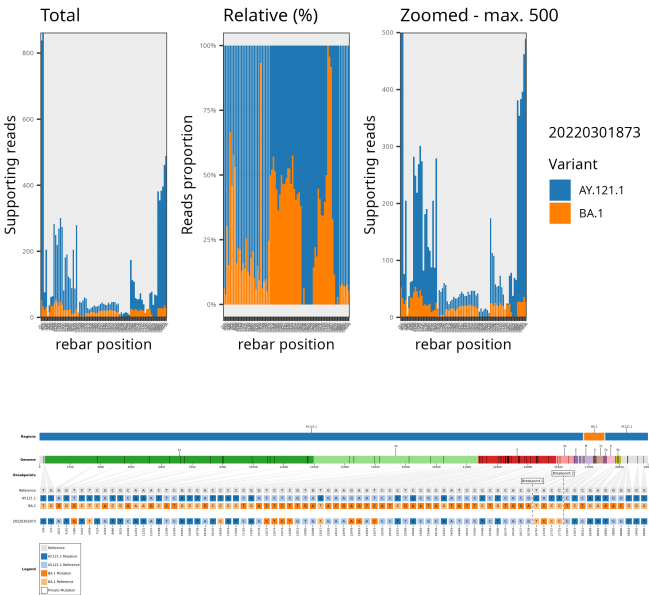


sample: 20220301910-coco

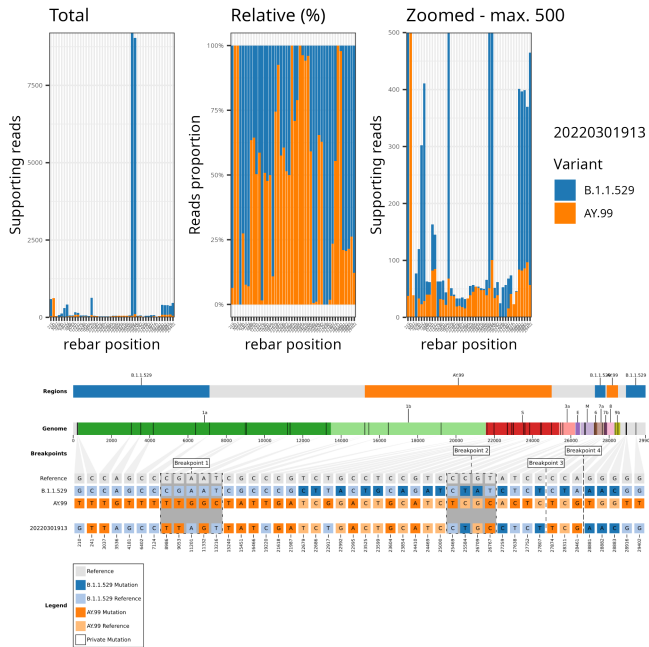


Supplemental Figure S3

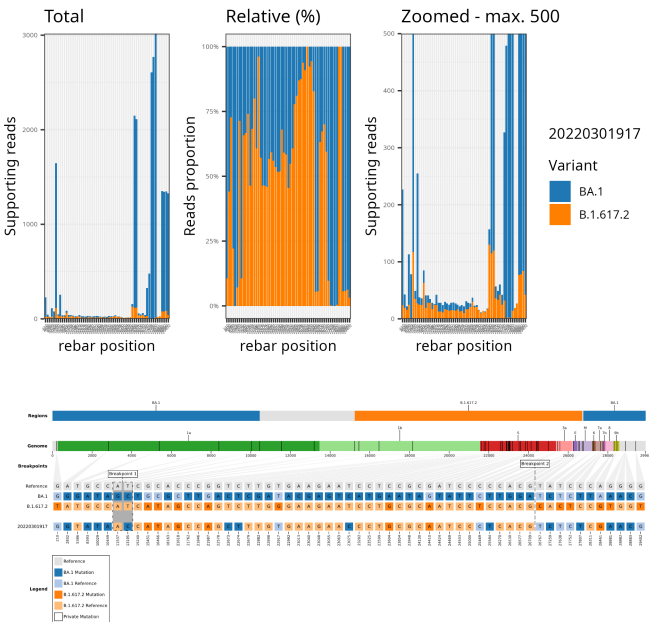
sample: 20220301873-coco



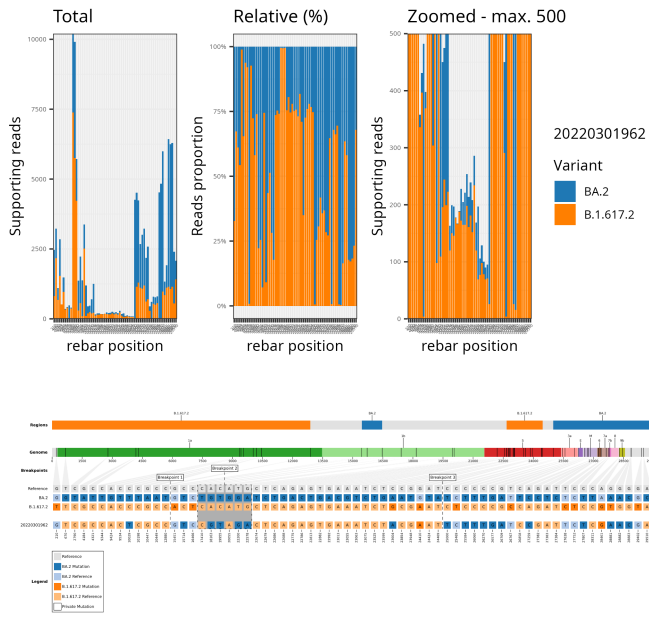
sample: 20220301913-coco



sample: 20220301917-coco/unclear

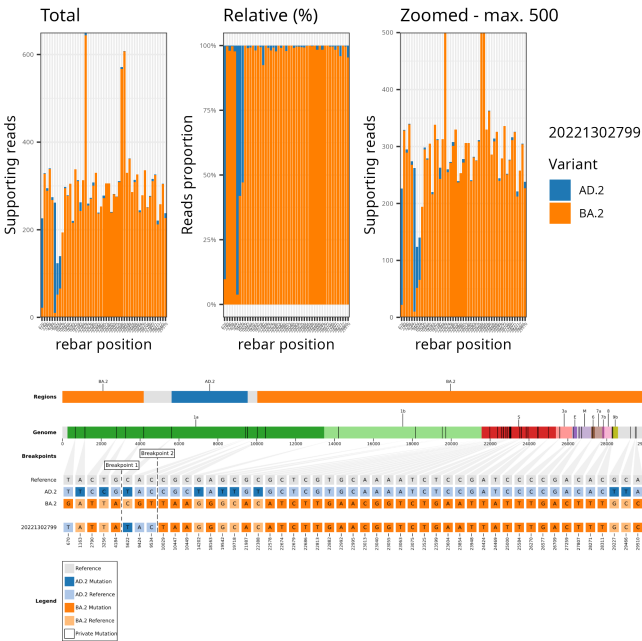


sample: 20220301962-coco

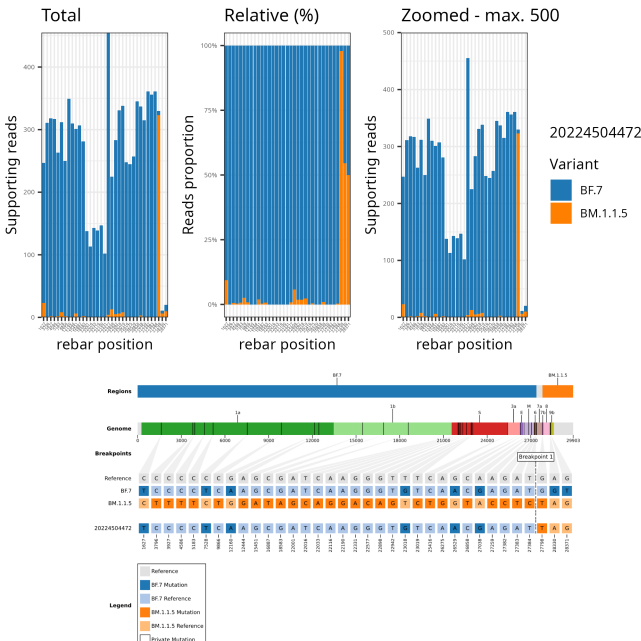


Supplemental Figure S3

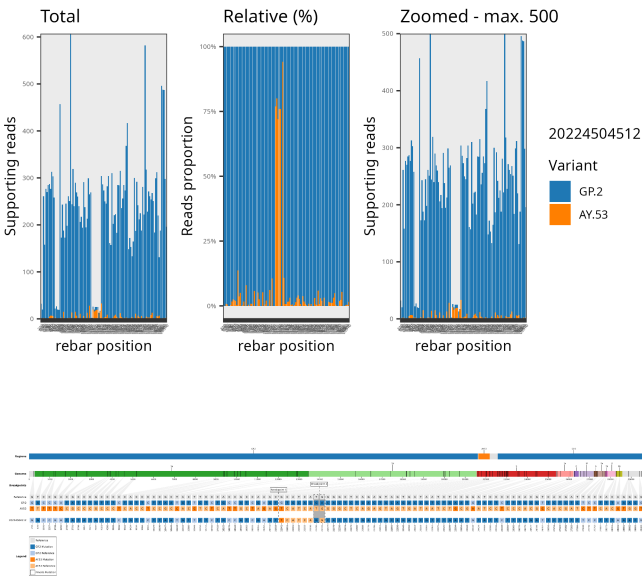
sample: 20221302799-coco



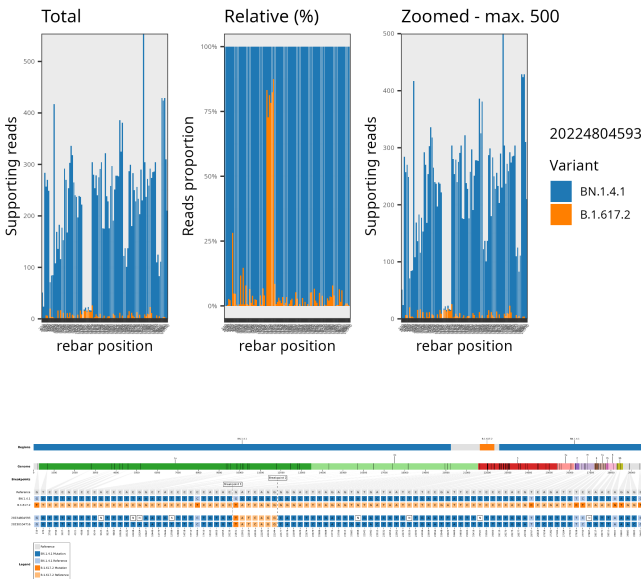
sample: 20224504472-coco/unclear



sample: 20224504512-unclear

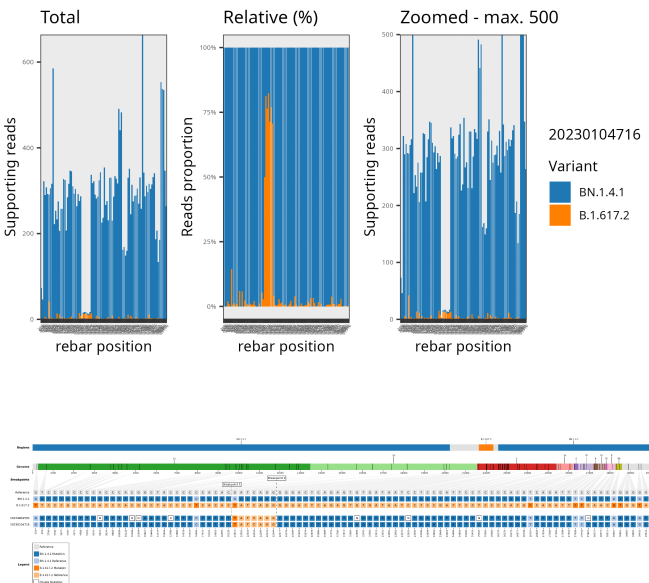


sample: 20224804593-unclear

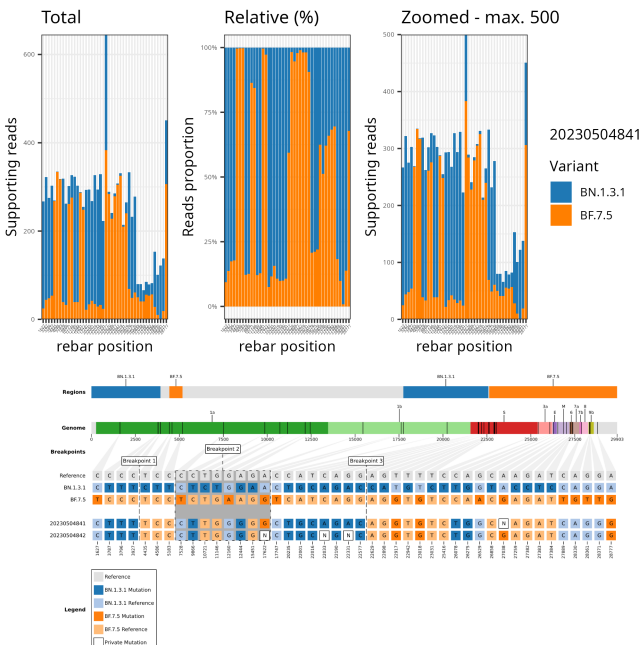


Supplemental Figure S3

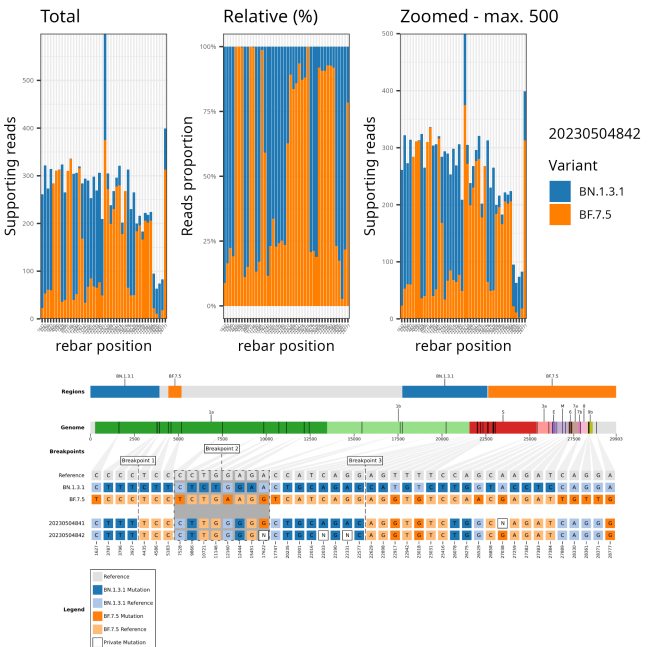
sample: 20230104716-unclear



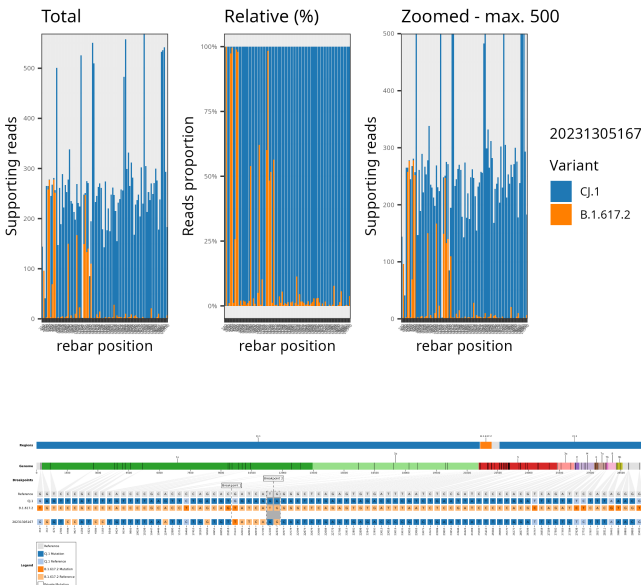
sample: 20230504841-coco



sample: 20230504842-coco

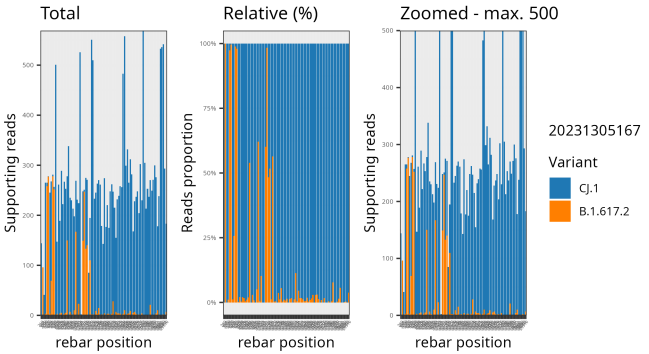


sample: 20230604918-coco

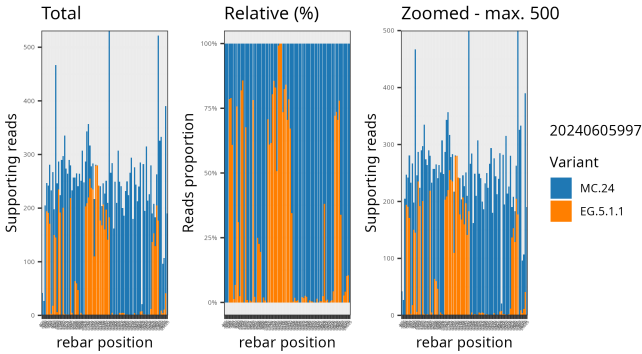


Supplemental Figure S3

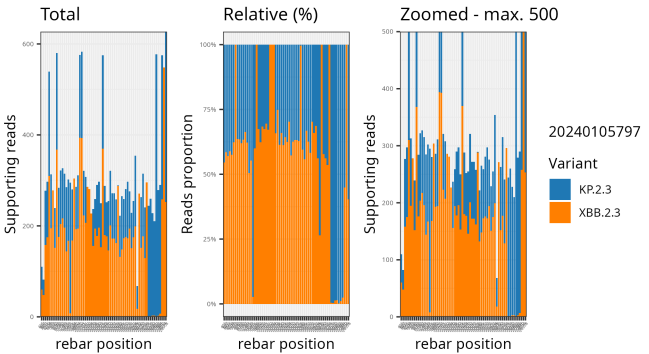
sample: 20231305167-coco



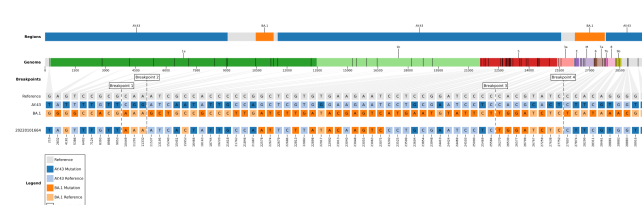
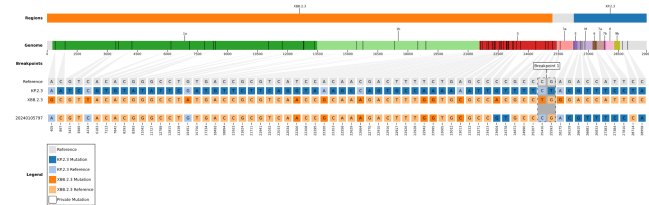
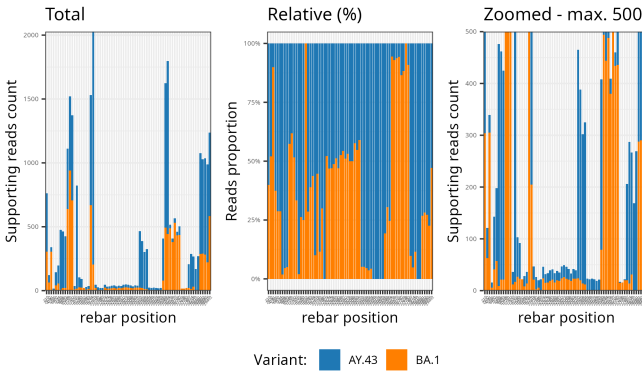
sample: 20240605997-coco

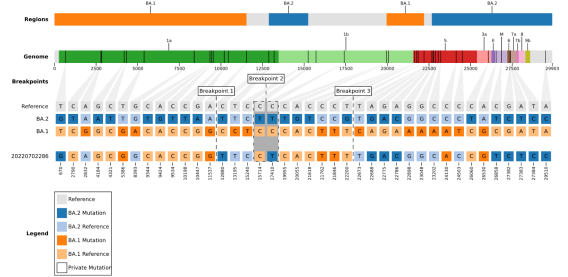


sample: 20240105797-coco



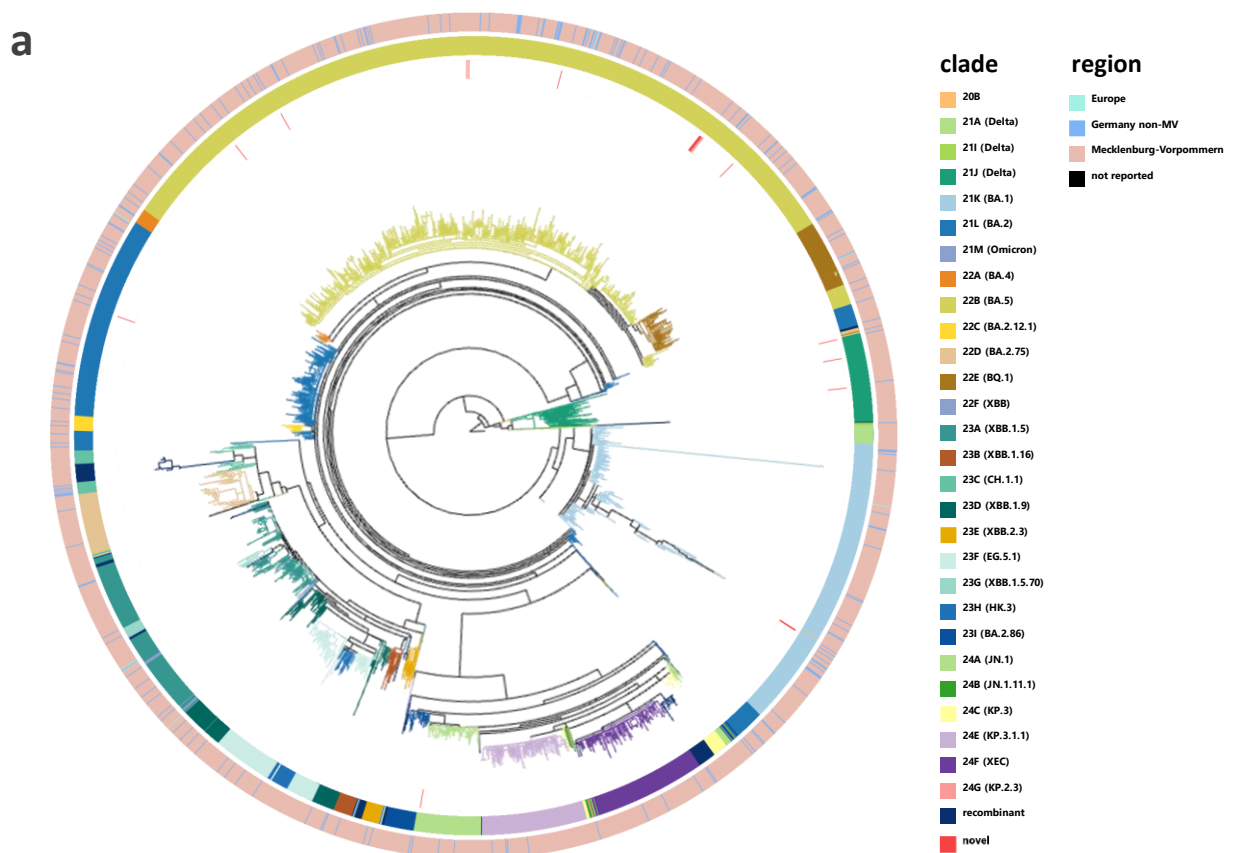
sample: 20220101664-coco/unclear





Spatially and temporally resolved maximum-likelihood phylogeny of 3,360 SARS-CoV-2 genomes sampled in M-V between January 2022 and January 2025.

a Spatially resolved maximum-likelihood phylogeny of SARS-CoV-2 genomes sampled in Mecklenburg-Western Pomerania, including sequences from individuals residing outside the state (other German federal states and European countries). The tree was inferred using RAxML v8.2.12 with the Wuhan reference genome (MN908947.3; clade 19A) as outgroup. Outer, middle, and inner rings denote Nextstrain clade assignments, geographic origin, and novel recombinant lineages (NRLs), respectively. A total of 20 recombinant Pango lineages (n = 75 sequences; XAK, XAL, XBC.1.3.1, XBF, XBL, XBW, XCA, XCH.1, XCT, XCT.1, XDA.1, XDD, XDQ, XDY, XEK, XEK.1, XEK.2, XEL, XEU, XM) were grouped into a single recombinant clade.



Spatially and temporally resolved maximum-likelihood phylogeny of 3,360 SARS-CoV-2 genomes sampled in M-V between January 2022 and January 2025.

b Time-resolved maximum-likelihood phylogeny of the same dataset. The tree is stratified by SARS-CoV-2 clades and highlights novel recombinant lineages (NRLs; red) and additional rare recombinant lineages (dark blue, $n = 75$ sequences; XAK, XAL, XBC.1.3.1, XBF, XBL, XBW, XCA, XCH.1, XCT, XCT.1, XDA.1, XDD, XDQ, XDY, XEK, XEK.1, XEK.2, XEL, XEU, XM). Multiple sequence alignment was performed using MAFFT v7.490, and phylogenetic inference was conducted with RAxML v8.2.12 using the Wuhan reference genome (MN908947.3) as outgroup. Time-scaled phylogenies were reconstructed and visualized using Nextstrain (Augur and Auspice). Clade assignments are indicated in the figure legend.



Supplemental Methods

Permutation-based framework

Breakpoint enrichment and statistical analysis

Genomic interval enrichment was assessed using a permutation-based sliding window approach implemented in NumPy (v1.26.4) within the Python environment. Analyses were performed on the SARS-CoV-2 reference genome Wuhan-Hu-1 (MN908947) with a total length of $G = 29,903$ bp, using a predefined set of genomic intervals $I_i = [s_i, e_i]$, where s_i and e_i denote the start and end positions of interval i , respectively, and N is the total number of intervals.

Sliding window analysis

The genome was partitioned into overlapping windows of fixed size $L = 100$ bp with a step size of $s = 10$ bp. Windows were defined as:

$$w_j = [1 + (j - 1)s, \min((j - 1)s + L, G)]$$

where w_j denotes window j , $j = 1, \dots, m$ indexes all windows, s is the step size (10 bp), L is the window length (100 bp), and G is the genome length (29,903 bp). The total number of windows is denoted by m :

$$m = \lceil \frac{G - L}{s} \rceil + 1$$

For each window w , the number of overlapping intervals was calculated as:

$$C_w = \sum_{i=1}^N 1(I_i \cap w \neq \emptyset)$$

20 where C_w is the observed number of intervals overlapping window w , $I_i=[s_i, e_i]$ is interval i , N
21 is the total number of intervals, and $1(\cdot)$ is the indicator function, which equals 1 if the condition
22 is true and 0 otherwise.

23 **Permutation testing**

24 Statistical significance was assessed using a non-parametric permutation framework with $K =$
25 10,000 permutations. In each iteration $k \in \{1, \dots, K\}$, all intervals were randomly repositioned
26 along the genome while preserving their original lengths $l_i = e_i - s_i + 1$, generating permuted
27 interval sets $I_i^{(k)}$. For each permutation, window-level overlap counts $C_w^{(k)}$ were recomputed.
28 Empirical p-values were calculated as:

$$29 \quad p_w = \frac{1 + \sum_{k=1}^K 1(C_w^{(k)} \geq C_w)}{K + 1}$$

30 where p_w is the empirical p-value for window w , $C_w^{(k)}$ is the overlap count in permutation k ,
31 and
32 K is the total number of permutations (10,000).

33 **Validity of the permutation framework**

34 The permutation strategy assumes that genomic interval positions are exchangeable along the
35 genome under the null hypothesis, conditional on the observed interval length distribution. By
36 preserving interval lengths while randomizing positions, the method maintains intrinsic
37 properties of the dataset while removing positional structure.

38 This approach tests whether the observed clustering or depletion of intervals within genomic
39 windows deviates from a uniform spatial distribution. As no parametric assumptions are made,
40 the method constitutes a distribution-free (non-parametric) test. The number of permutations
41 ($K = 10,000$) ensures stable estimation of empirical p-values, particularly in the tail of the
42 distribution.

43 **Multiple testing correction**

To account for multiple testing across all windows (m total windows), p-values were adjusted using the Benjamini–Hochberg procedure. Let the ordered p-values be:

$$p(1) \leq p(2) \leq \dots \leq p(m)$$

where $p_{(i)}$ denotes the i -th smallest p-value.

The corresponding false discovery rates (FDR) were calculated as:

$$q_{(i)} = \min_{j \geq i} \left(\frac{m}{j} \cdot p_{(j)} \right)$$

where $q_{(i)}$ is the adjusted p-value (FDR) corresponding to rank i , m and is the total number of tested windows. Monotonicity was enforced to ensure that q-values are non-increasing with increasing rank.

Base-level signal assignment

Base-level interval coverage was computed for each genomic position $x \in \{1, \dots, G\}$ as the number of overlapping intervals. To assign statistical significance at single-base resolution, each position x was assigned the most significant (i.e., minimal) p-value and FDR among all overlapping windows:

$$p_{\text{base}}(x) = \min\{p_w | x \in w\}$$

$$\text{FDR}_{\text{base}}(x) = \min\{q_w | x \in w\}$$

where $p_{\text{base}}(x)$ and $\text{FDR}_{\text{base}}(x)$ denote the assigned p-value and FDR at genomic position x , respectively. This results in a conservative estimate of local enrichment. These base-level values were used for visualization in the corresponding figure (Figure 3b).

Output generation and reproducibility

Two output files were generated. The first file contains window-level results, including the genomic start and end positions of each window, the observed counts C_w , the empirical p-values p_w , and the corresponding FDR-adjusted values. The second file contains base-level results,

68 reporting for each genomic position $x = 1, \dots, 29,903$ the interval coverage as well as the
69 assigned p-values and FDR values. Both datasets are provided in Supplementary Table S10.

70