Metadata check

Diagnostic plot to estimate min−length filter for phylogenetic analysis (all data = 4173)

Rubella clade (n=3783)

Table 1: Date check

accession	strain	date	date_released	date_updated
EU016200	RVs/Nijmegen.NET/05.29/TT	2029-01-XX	2007-07-30	2016-07-26

Nextclade compared with GenBank annotations:

Nextclade Genotype results

Original Genotype from GenBank	1A	1B	1C	1D	1E	1F	1G	1H	1I	1J	2A	2B	2C	unassigned	NA
1A	29	1												32	
1B		24					1							2	
1C			24												
1D				4						3				3	
1E					1643										16
1F						11									
1G		5					107	17						29	
1H								19						1	
1I									7						
1J										35				1	
2A											5				
2B												1495		1	4
2C													1		
NA	13	13	26	11	117	2	37	9	14	57	4	91	18	111	130
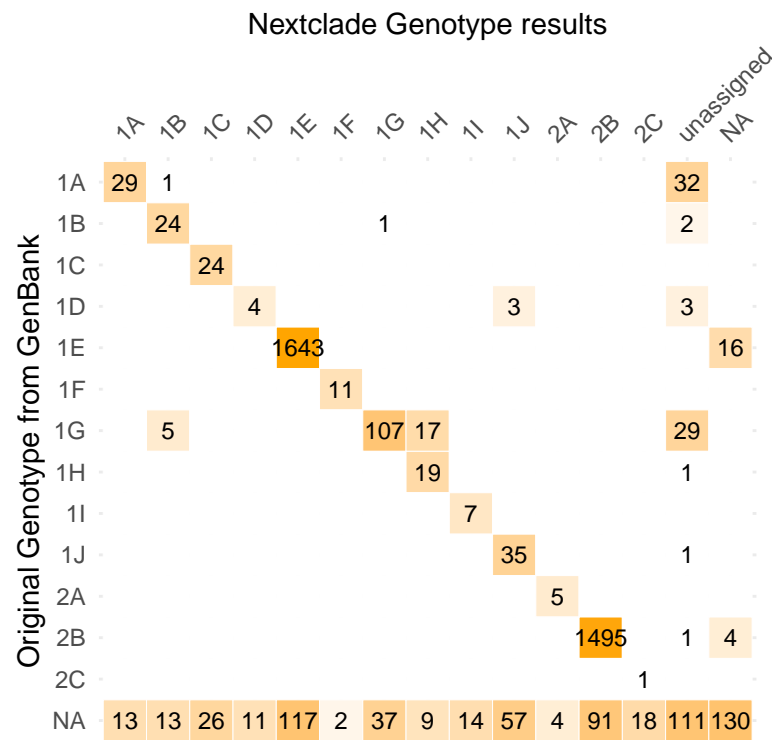
Nextclade alignment check

```
"alignmentParams": {
    "minSeedCover": 0.01,
    "terminalBandwidth": 100,
    "excessBandwith": 18,
    "minMatchLength": 20
},
```

Frequency of deletions along rubella
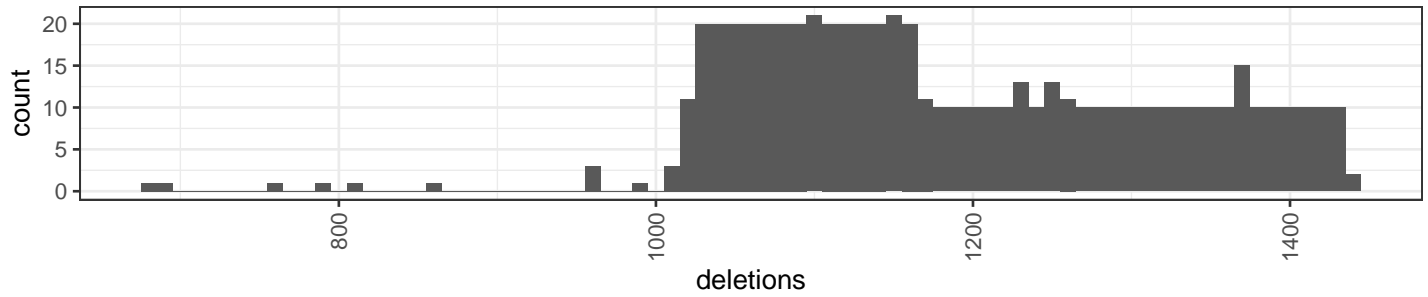
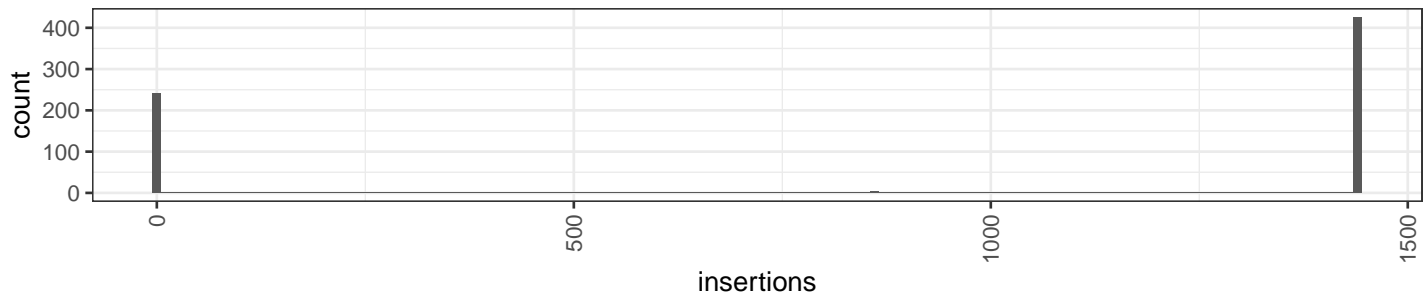## Frequency of insertions along rubella

Table 2: Top Frameshifts

frameShifts	n
E1:290-481	4
E1:228-266	1
E1:232-481	1
E1:266-267	1
E1:272-331	1
E1:312-333	1
E1:342-359	1
E1:367-481	1
E1:385-416	1
E1:390-481	1

Missing Data

```
"missingData": {
  "enabled": true,
  "missingDataThreshold": 20, # Top threshold, scale the next set of missing data from 0 to 100
  "scoreBias": 4  # First 4 not penalized
},
```
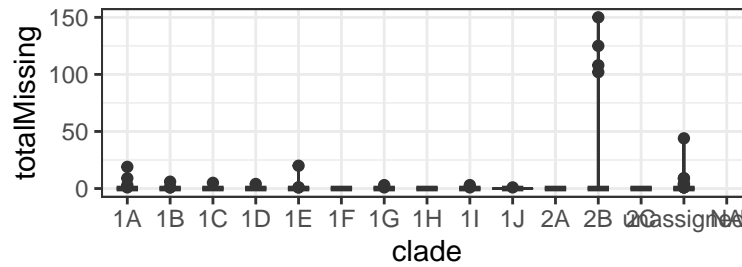

Table 3: Records above missing threshold

seqName	totalMissing
MK754430	150
MK754427	125
MK754428	108
MK754429	102
L16231	44

Mixed Sites (M)

```
"mixedSites": {
  "enabled": true,
  "mixedSitesThreshold": 4  # More than this number of mixed sites results in a "bad" score
},
```
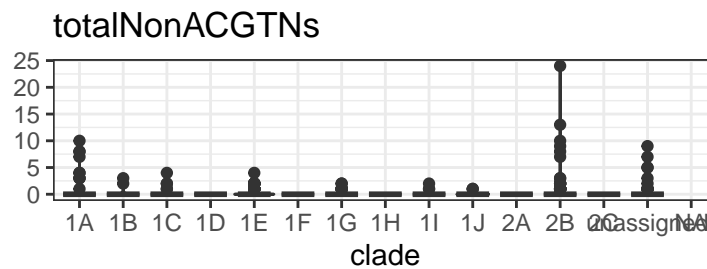
Table 4: Records above totalNonACGTNs threshold

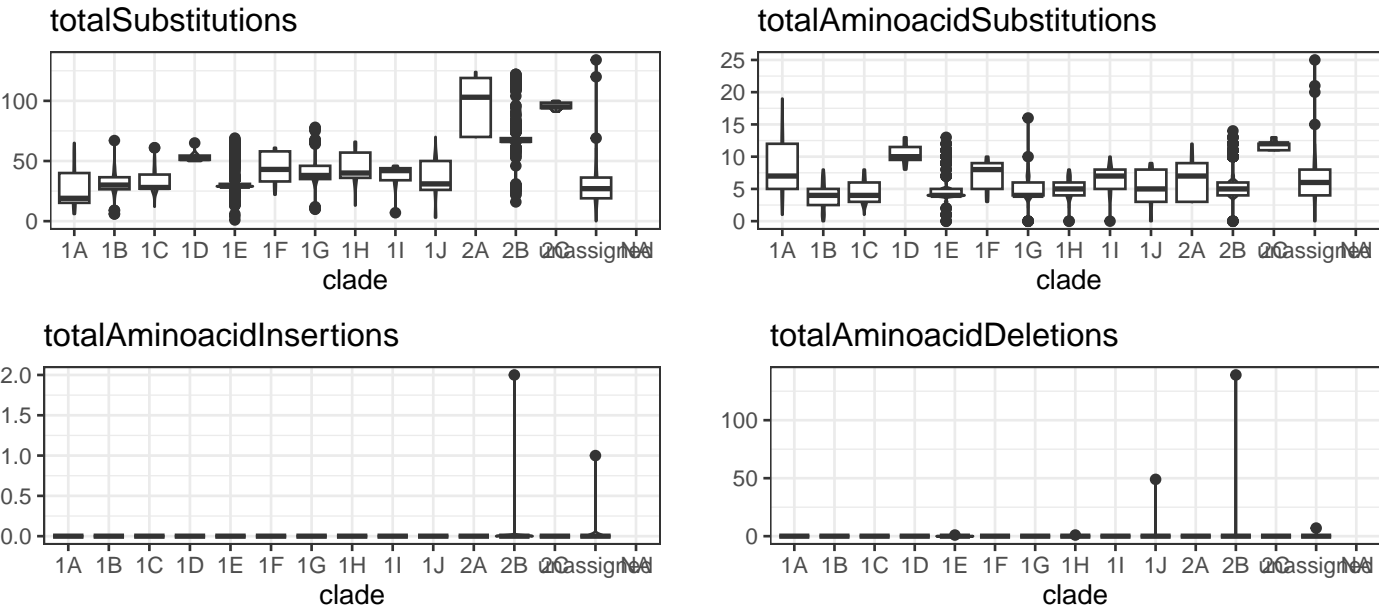| seqName	totalNonACGTNs
PV035063	24
PV035062	13
KT369573	10
OQ296620	10
L19420	9
PV035058	9
OM022831	8
PP536671	8
PV035060	8
MK780810	7
PV035061	7
U58299	7
AF039106	5
L16230	5

Private Mutations

```
"privateMutations": {
  "enabled": true,
  "cutoff": 45,
  "typical": 4,
  "weightLabeledSubstitutions": 1,
  "weightReversionSubstitutions": 1,
  "weightUnlabeledSubstitutions": 1
},
```
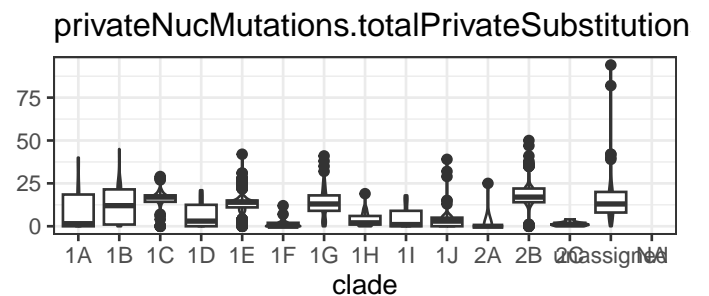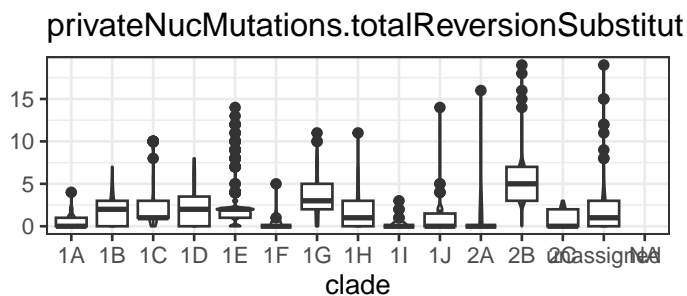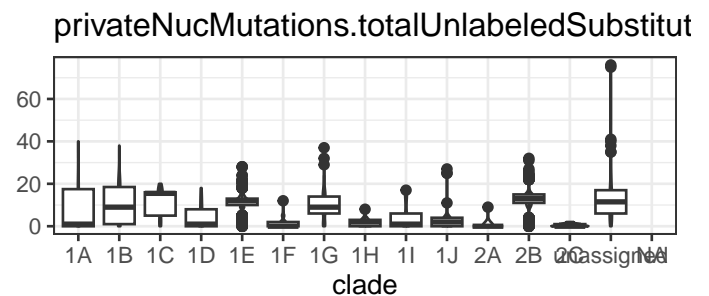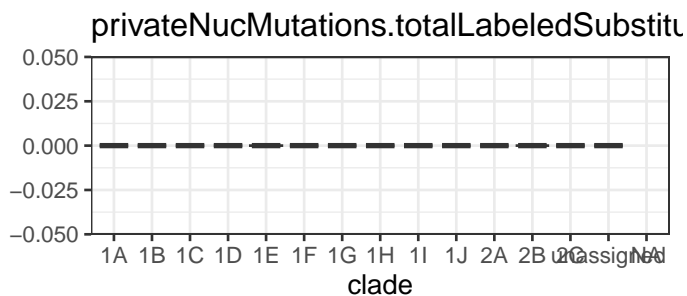

Mutation clusters

```
"snpClusters": {
  "enabled": true,
  "clusterCutOff": 10,
  "scoreWeight": 50,
  "windowSize": 50
}
```

totalSubstitutions

totalAminoacidSubstitutions

totalAminoacidInsertions

totalAminoacidDeletions

6

Stop Codons

```
# Example, not set in rubella
"stopCodons": {
      "enabled": true,
      "ignoredStopCodons": [
         {
            "geneName": "ORF8",
            "codon": 26
         },
         {
            "geneName": "ORF8",
            "codon": 67
         }
      ]
}
```

qc.stopCodons.totalStopCodons
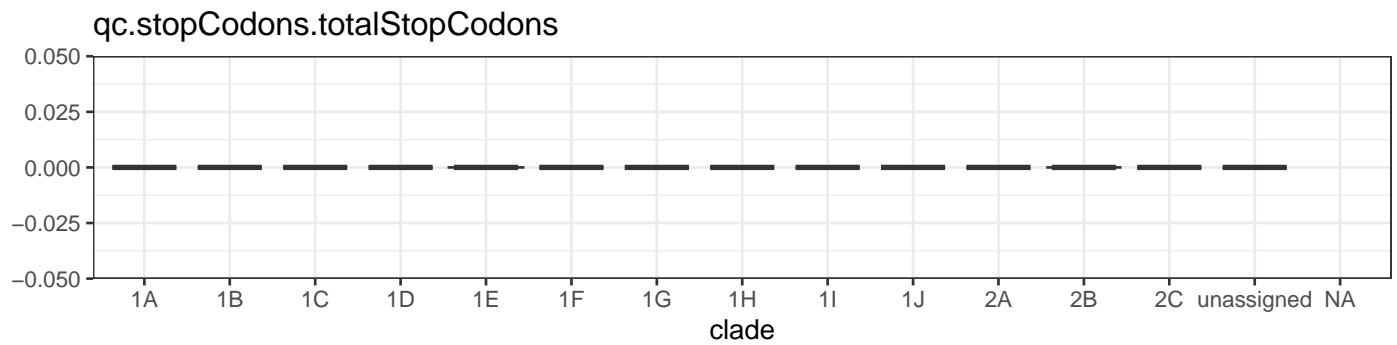
Frameshifts

```
"frameShifts": {
   "enabled": true,
   # Example known frame shifts, enabled in rubella but no ignored ones set
   "ignoredFrameShifts": [
      {
         "geneName": "ORF3a",
         "codonRange": {
            "begin": 256,
            "end": 276
         }
      },
      {
         "geneName": "ORF3a",
         "codonRange": {
            "begin": 258,
            "end": 276
         }
      }
   ]
},
```
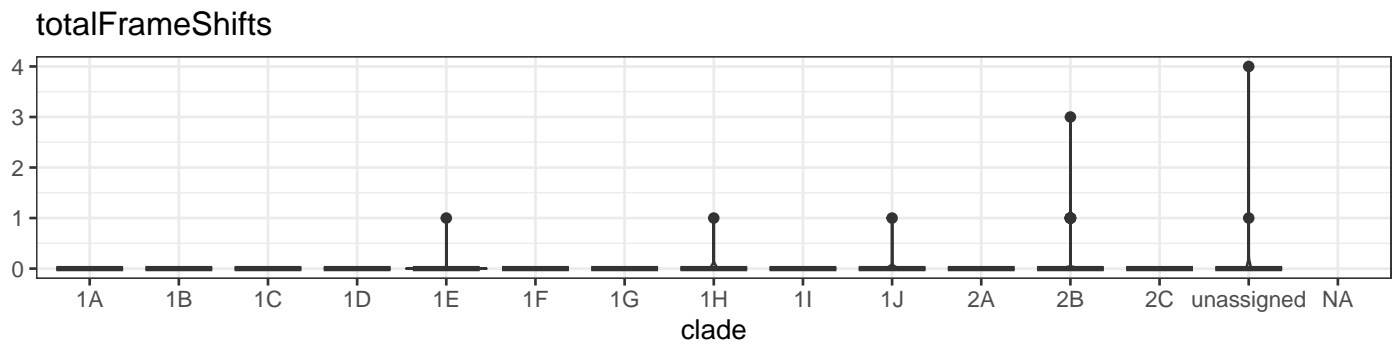
totalFrameShifts

Table 5: Top Frameshifts

frameShifts	n
E1:290-481	4
E1:228-266	1
E1:232-481	1
E1:266-267	1
E1:272-331	1
E1:312-333	1
E1:342-359	1
E1:367-481	1
E1:385-416	1
E1:390-481	1

Table 6: Records with top frameShifts

seqName	frameShifts
FR693375	E1:290-481
FR693376	E1:290-481
FR693377	E1:290-481
FR693378	E1:290-481