# Interim Report on Multiple Sequence Alignments and TaqMan Signature Mapping to Phylogenetic Trees

**Project Title: Forensic TaqMan and Microarray Assays for Viral Genotyping**


**Contributors**

Shea Gardner and Crystal Jaing

Lawrence Livermore National Laboratory (LLNL), Livermore, CA

**Principal Investigator and Correspondent**

Crystal Jaing

925-424-6574, jaing2@llnl.gov



LLNL-TR-543651

**March 28, 2012**





This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

## Introduction

The goal of this project is to develop forensic genotyping assays for select agent viruses, addressing a significant capability gap for the viral bioforensics and law enforcement community. We used a multipronged approach combining bioinformatics analysis, PCR-enriched samples, microarrays and TaqMan assays to develop high resolution and cost effective genotyping methods for strain level forensic discrimination of viruses. We have leveraged substantial experience and efficiency gained through year 1 on software development, SNP discovery, TaqMan signature design and phylogenetic signature mapping to scale up the development of forensics signatures in year 2. In this report, we have summarized the Taqman signature development for South American hemorrhagic fever viruses, tick-borne encephalitis viruses and henipaviruses, Old World Arenaviruses, filoviruses, Crimean-Congo hemorrhagic fever virus, Rift Valley fever virus and Japanese encephalitis virus.

## Methods

Multiple sequence alignments (MSAs) were built for all available full length sequences of South American hemorrhagic fever viruses, tick-borne encephalitis viruses and henipaviruses filoviruses, Japanese Encephalitis viruses, Rift Valley fever viruses, Crimean-Congo hemorrhagic fever viruses. Phylogenetic trees were built from the MSAs. Primer pairs and TaqMan signatures were designed and mapped to the nodes of the trees.

In aligning the genomes, we encountered three issues that required us to build specialized software for preprocessing the target sequence sets: 1) sequences provided in both plus and minus orientations; 2) taxonomic groups that had multi-species, multi-segment sequences with little or no homology at the nucleotide level; and 3) large target sets with up to hundreds of sequences. Often, a few of the sequences for a virus species or complex are given in opposite (reverse complemented) directions compared to the other sequences of the species. In experiments with various alignment software, we were unable to find a tool which automatically detected and corrected sequences provided in the reverse complement direction. For example, Mauve (http://gel.ahabs.wisc.edu/mauve/; Darling et al. 2004) can handle rearrangements so theoretically should handle plus and minus sequence orientations. However, Mauve also relies on MUMs (maximum unique matches), which sets of divergent viruses (e.g. New or Old World Arenaviruses) do not have, with the result that Mauve fails except for closely related strains. MUSCLE (http://www.drive5.com/muscle/; Edgar et al. 2004) will align divergent sequences in mixed directions, but the results are bogus, which makes for undetectable errors: a genome will look like an outlier, but really it is just reported in the reverse complement direction. With hundreds of sequences in some target sets, curation by hand was not feasible.

Therefore, we used the USEARCH software (http://www.drive5.com/usearch/; Edgar, 2010) with the –uclust –rev options to cluster sequences under the desired taxonomy node by major groups (e.g. by segment). Then we used tblastx to determine the correct orientation of each sequence relative to a reference, and correct the target sets into groups where all sequences represent the same segment in the same orientation. Multiple sequence alignments were built with MUSCLE and the best maximum likelihood tree was generated from each alignment using RAxML (http://icwww.epfl.ch/~stamatak/index-Dateien/software/RAxML-Manual.7.0.4.pdf ; Stamatakis 2006) with 1000 bootstrap inferences.

Primer pairs and TaqMan sequences were generated with Primux using the run_Primux_triplet script (manuscript submitted), searching for minimal sets of conserved primers such that all targets have an amplicon, and then finding 1) conserved minimal set probes on those amplicons; and 2) the least conserved probes on those amplicons to maximize genotype discrimination. Up to 20 iterations of this process were attempted to generate signatures in which no primers overlapped (to avoid near-duplicate solutions that differ by only a few bases), although signature amplicons could overlap. In some cases fewer than 20 iterations were possible because too few primer candidates passed design criteria for $T_m$, sequence complexity, hairpin and primer dimer avoidance, maximum homopolymer length, etc. listed below.

- Maximum number of degenerate bases per primer or conserved probe=2
- Maximum number of degenerate bases per genotyping probe=0
- Primer length=18-25 bases
- Probe length=20=30 bases
- Amplicon length=80-300 bp
- GC% of primers and probes=20-80%
- Maximum length of homopolymer run=4
- Minimum primer or probe sequence entropy based on trimer frequencies=3.5
- Unafold NN Tm calculations
- Primer Tm=60-65C
- Probe Tm=68-75C
- Minimum hairpin free energy=-5.0 kcal/mol
- Minimum primer dimer free energy=-6.0 kcal/mol
- Minimum allowable distance between a degenerate base and the 3' end of a primer=3

Viruses analyzed are listed below, with the number of sequences available and the average sequence length.

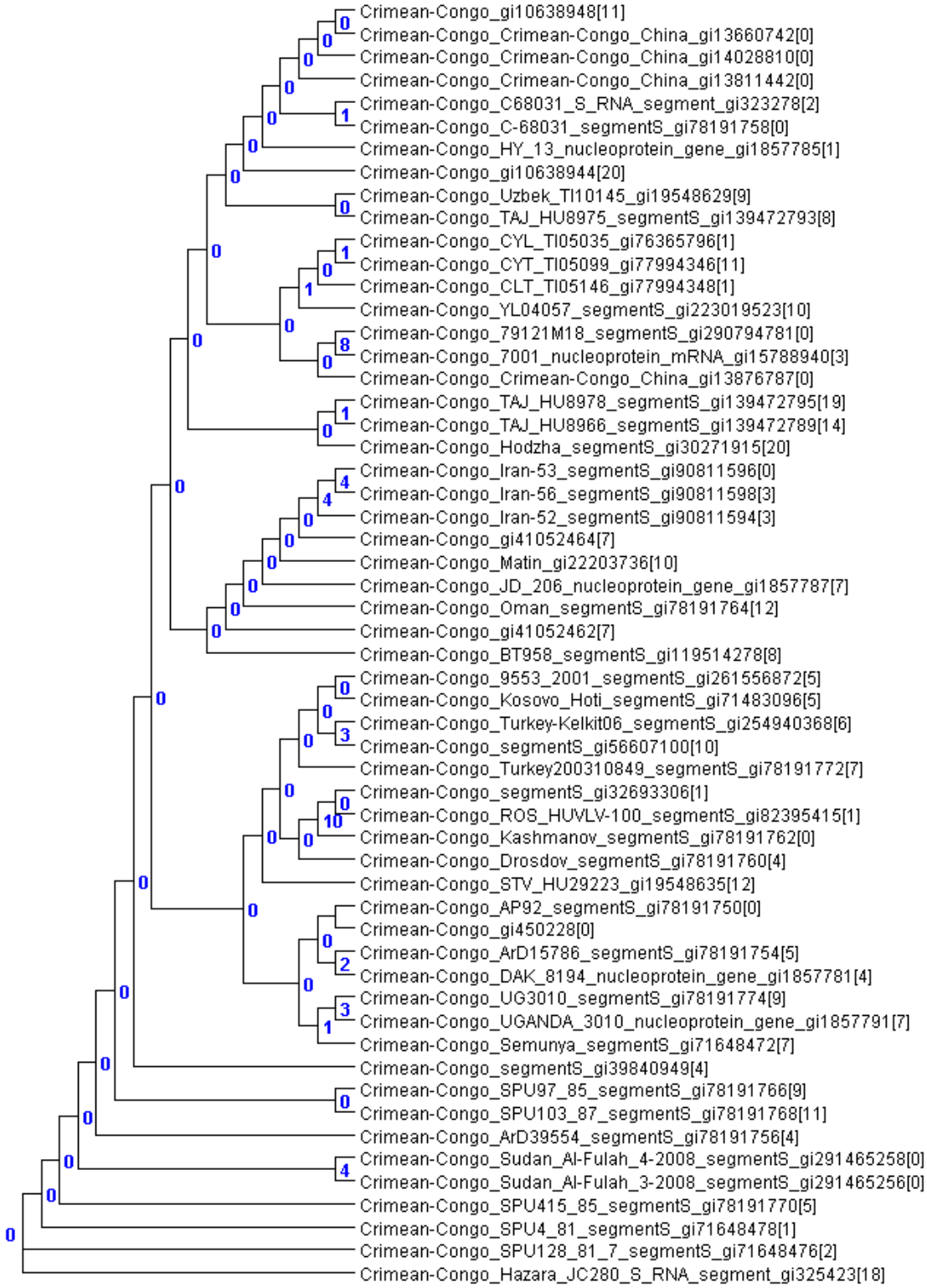
| Organism | # sequences | Avg Length | # of signatures | # of homoplastic signatures |
|---|---|---|---|---|
| CCHF_S | 56 | 1668 | 646 | 267 |
| CCHF_M | 49 | 5314 | 551 | 132 |
| CCHF_L | 31 | 12113 | 1582 | 879 |
| RVF_S | 89 | 1684 | 216 | 77 |
| RVF_M | 69 | 3885 | 219 | 69 |
| RVF_L | 62 | 6404 | 206 | 70 |
| Ebola | 22 | 18659 | 261 | 42 |
| Marburg | 31 | 19115 | 169 | 91 |
| Hendra | 10 | 18234 | 100 | 26 |
| Nipah | 9 | 18247 | 69 | 21 |
| Junin_L | 12 | 7114 | Included in NW Arena analyses | |
| Machupo_L | 5 | 7141 | | |
| Junin_S | 26 | 3410 | | |
| Machupo_S | 13 | 3432 | | |

| | | | | |
|---|---|---|---|---|
| JEV | 144 | 10968 | 746 | 286 |
| NW_Arena_S | 100 | 3396 | 898 | 183 |
| NW_Arena_L | 42 | 7107 | 478 | 31 |
| OW_Arena_S | 54 | 3547 | 2348 | 1338 |
| OW_Arena_L | 45 | 7199 | 2640 | 1415 |
| TBEV | 67 | 10840 | 765 | 182 |

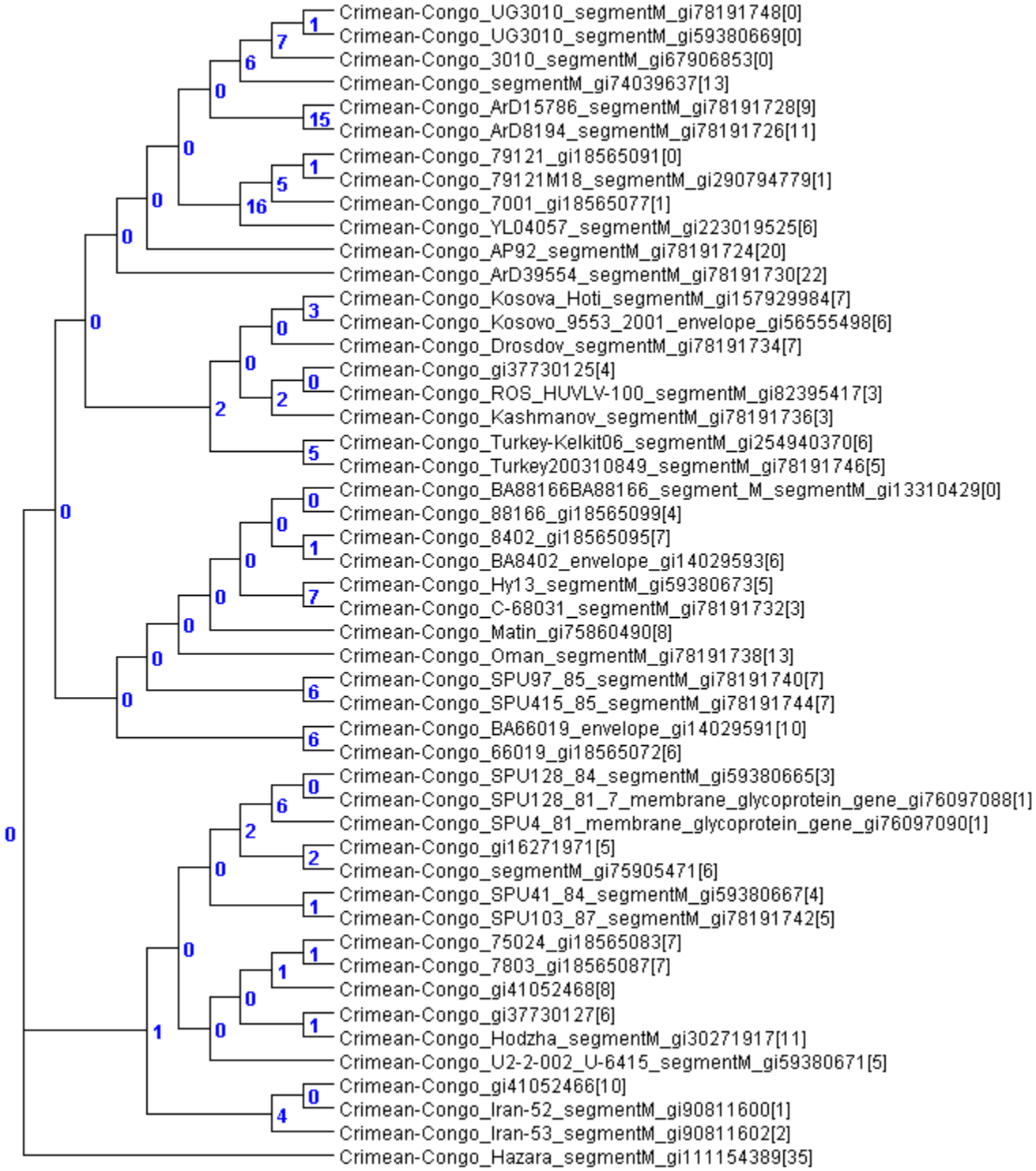
Abbreviations: CCHF=Crimean-Congo hemorrhagic fever, RVF=Rift Valley fever, JEV=Japanese encephalitis virus, NW_Arena=New World Arenavirus, OW_Arena=Old World Arenavirus, TBEV=tick-borne encephalitis virus.

Shown below are the phylogenetic trees for each virus group. In blue at the nodes and at the end of the genome identifiers in brackets [#] are indicated the numbers of TaqMan signatures or primer pairs that specifically amplify the genome or genomes down the branch. While many nodes do not have signatures mapping to them from those generated, there are additional "homoplastic" signatures which are shared by groups of genomes that do not exactly match a node of the tree. All signatures generated and the groups of genomes which they are predicted to detect will be provided separately. Due to the complexity of these large trees, genotyping an unknown could require many TaqMan signature reactions: low-plex PCR based methods are not recommended for large, diverse virus groups. One TaqMan signature only represents a single "locus". Doing viral forensics using a microarray format is more efficient and provides higher confidence genotyping, since a single array can test thousands of loci simultaneously.

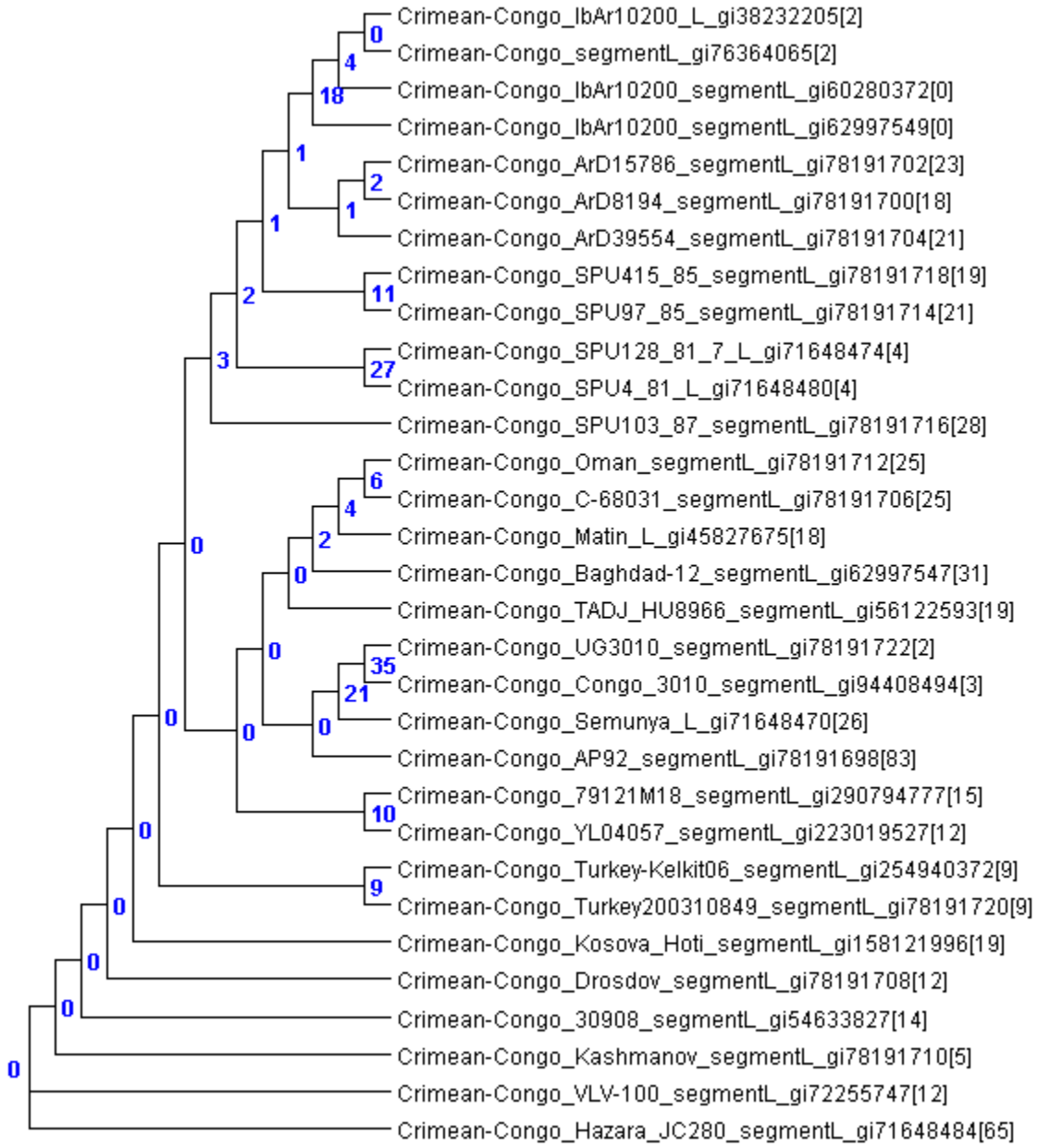
CCHF S
Number_signatures: 646
Number_Homoplastic_signatures: 267



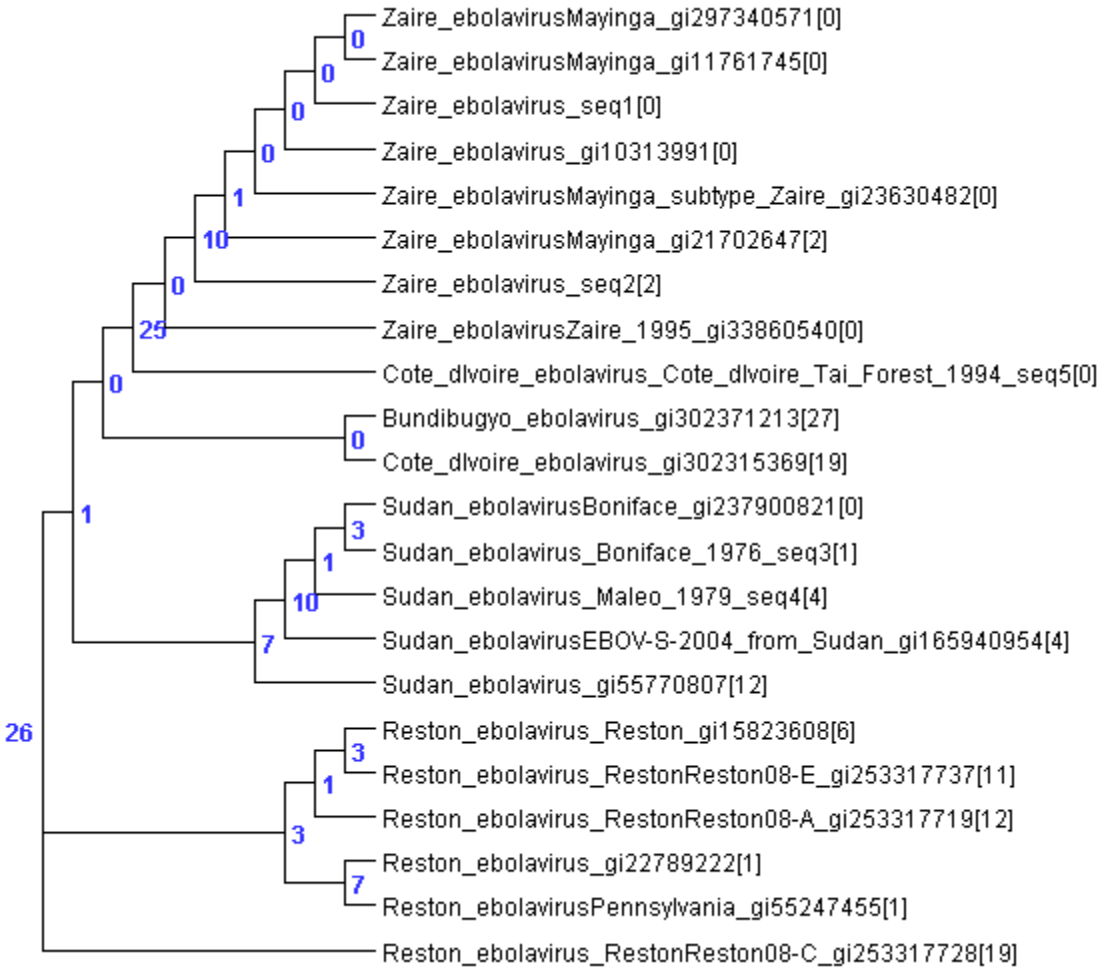
CCHF M
Number_signatures: 551
Number_Homoplastic_signatures: 132



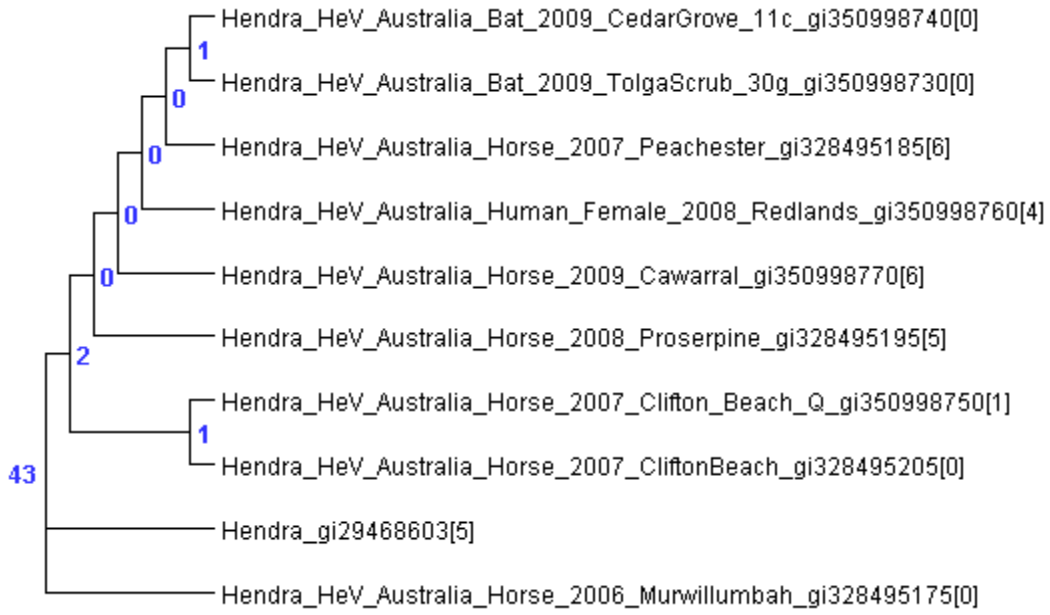
CCHF L
Number_signatures: 1582
Number_Homoplastic_signatures: 879



Ebola
Number_signatures: 261
Number_Homoplastic_signatures: 42

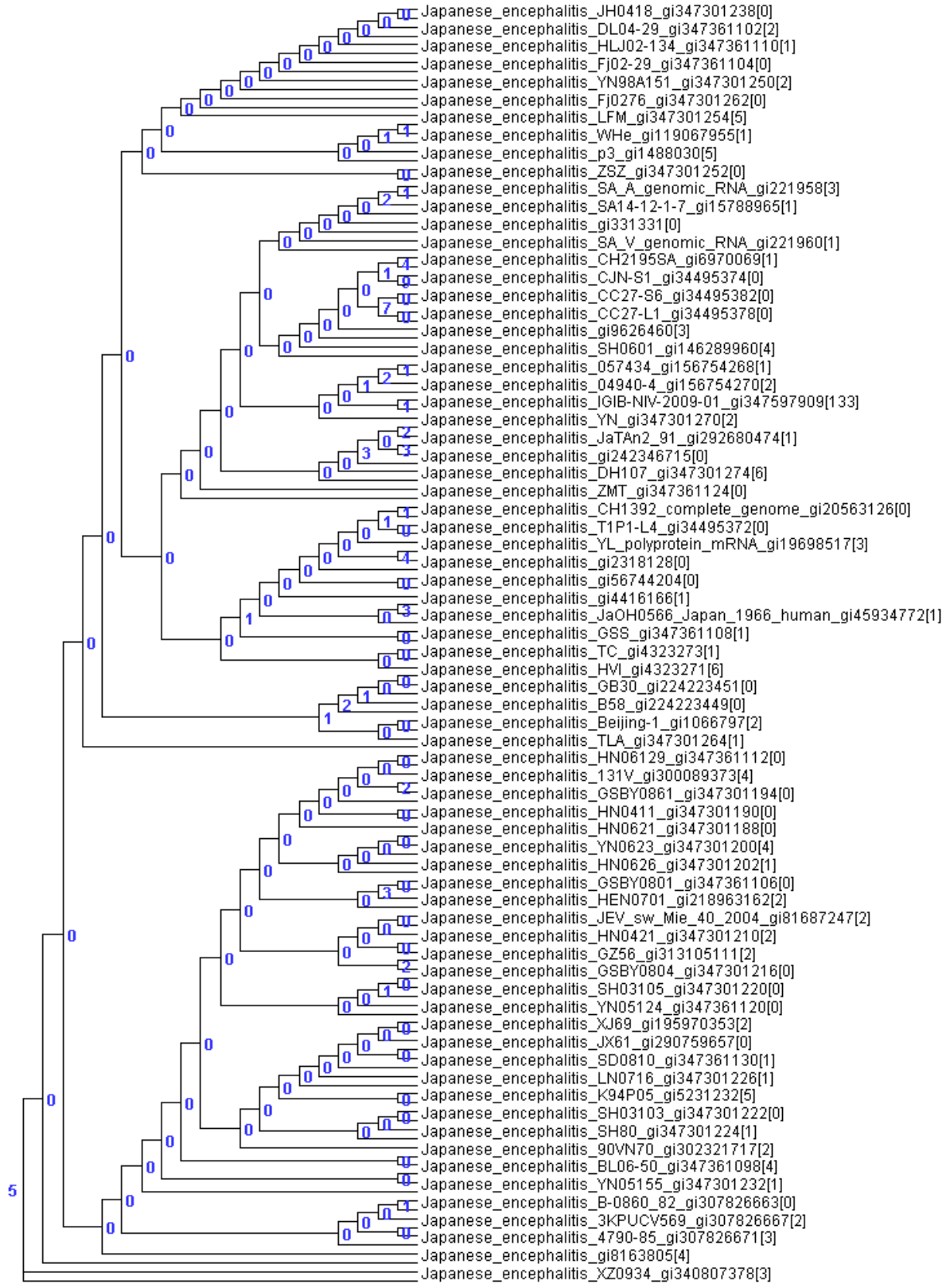


Hendra
Number_signatures: 100
Number_Homoplastic_signatures: 26

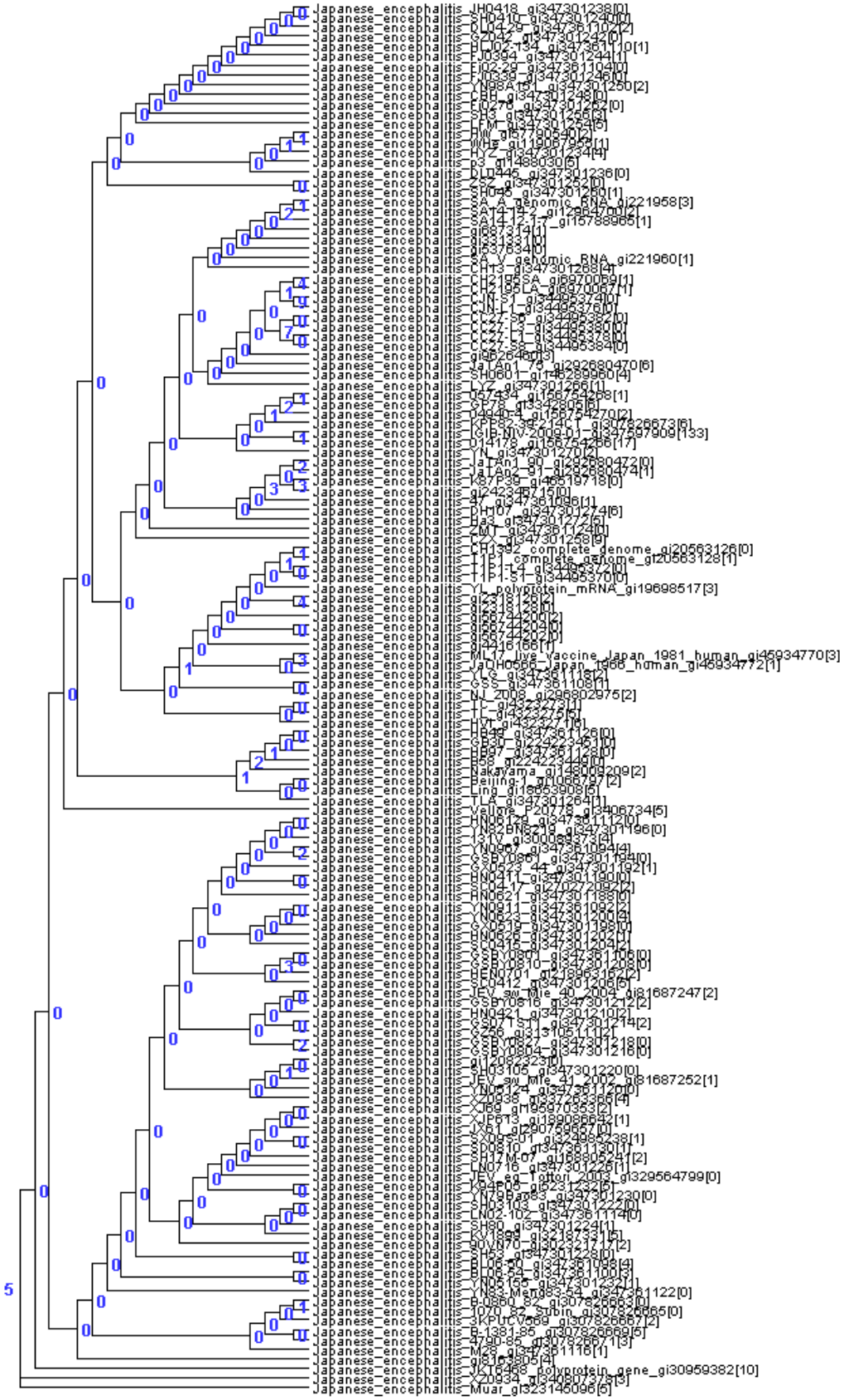


Nipah
Number_signatures: 69
Number_Homoplastic_signatures: 21

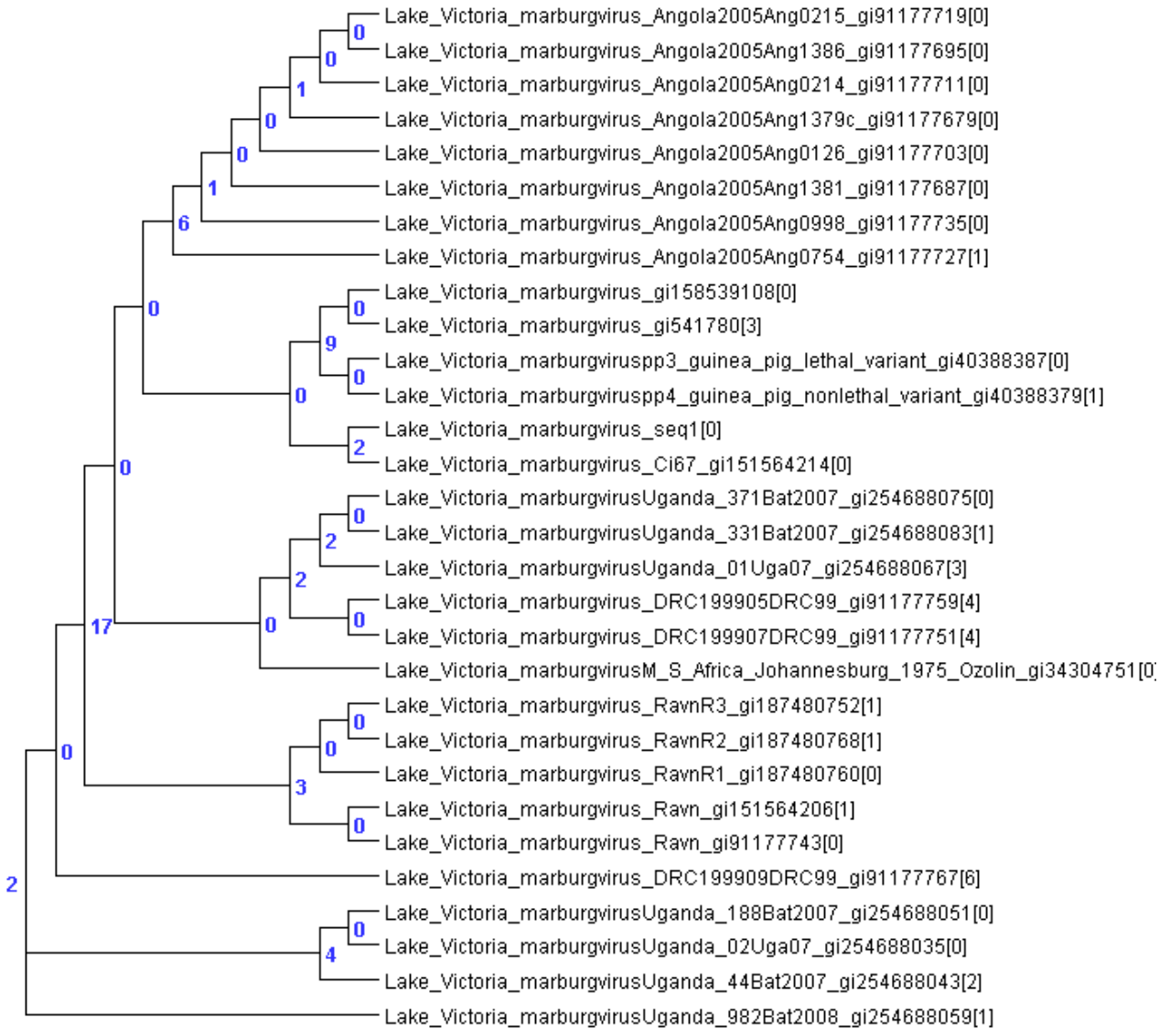
Japanese encephalitis (sparse labels)
Number_signatures: 746
Number_Homoplastic_signatures: 286



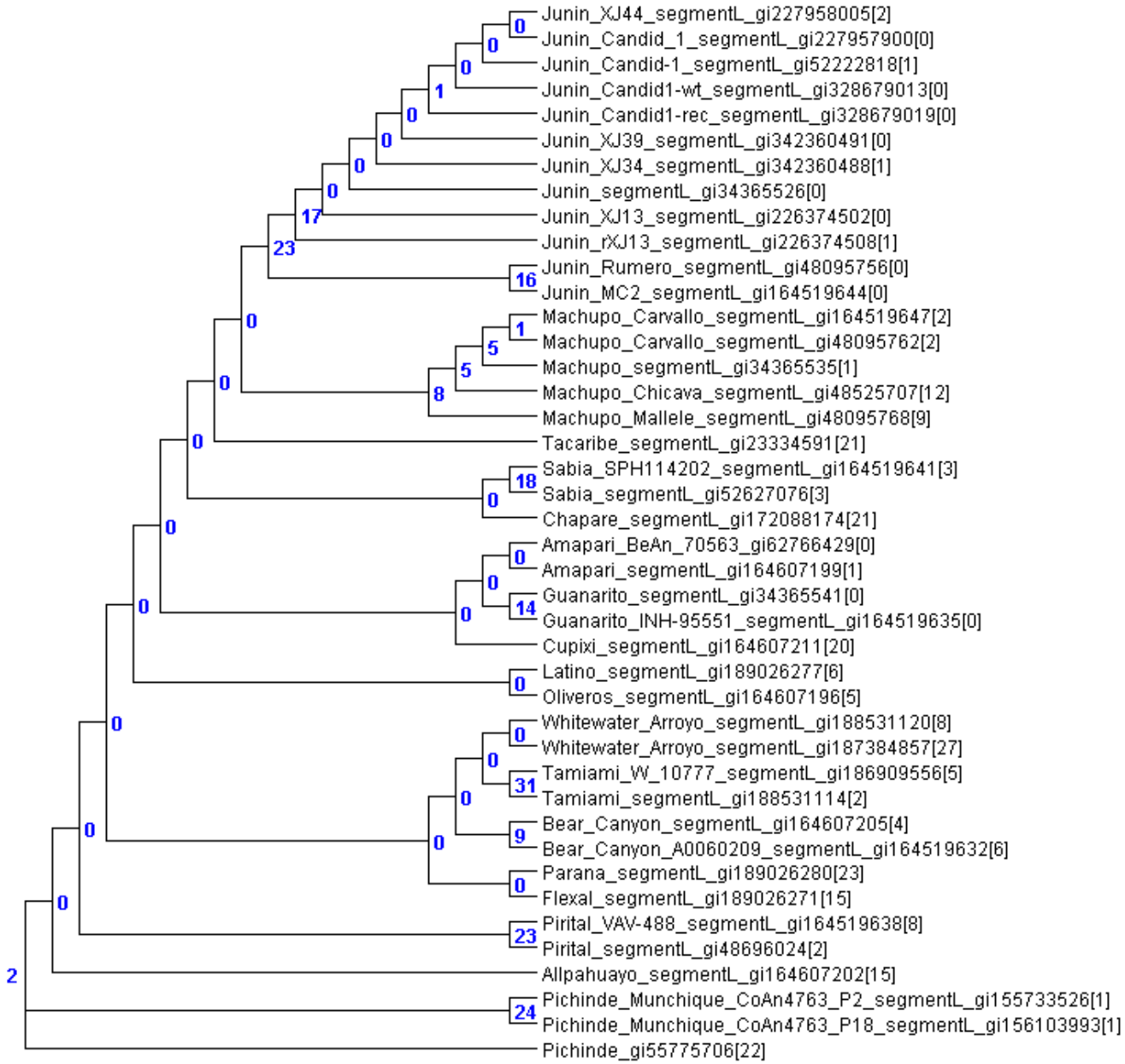
Japanese encephalitis (all labels)



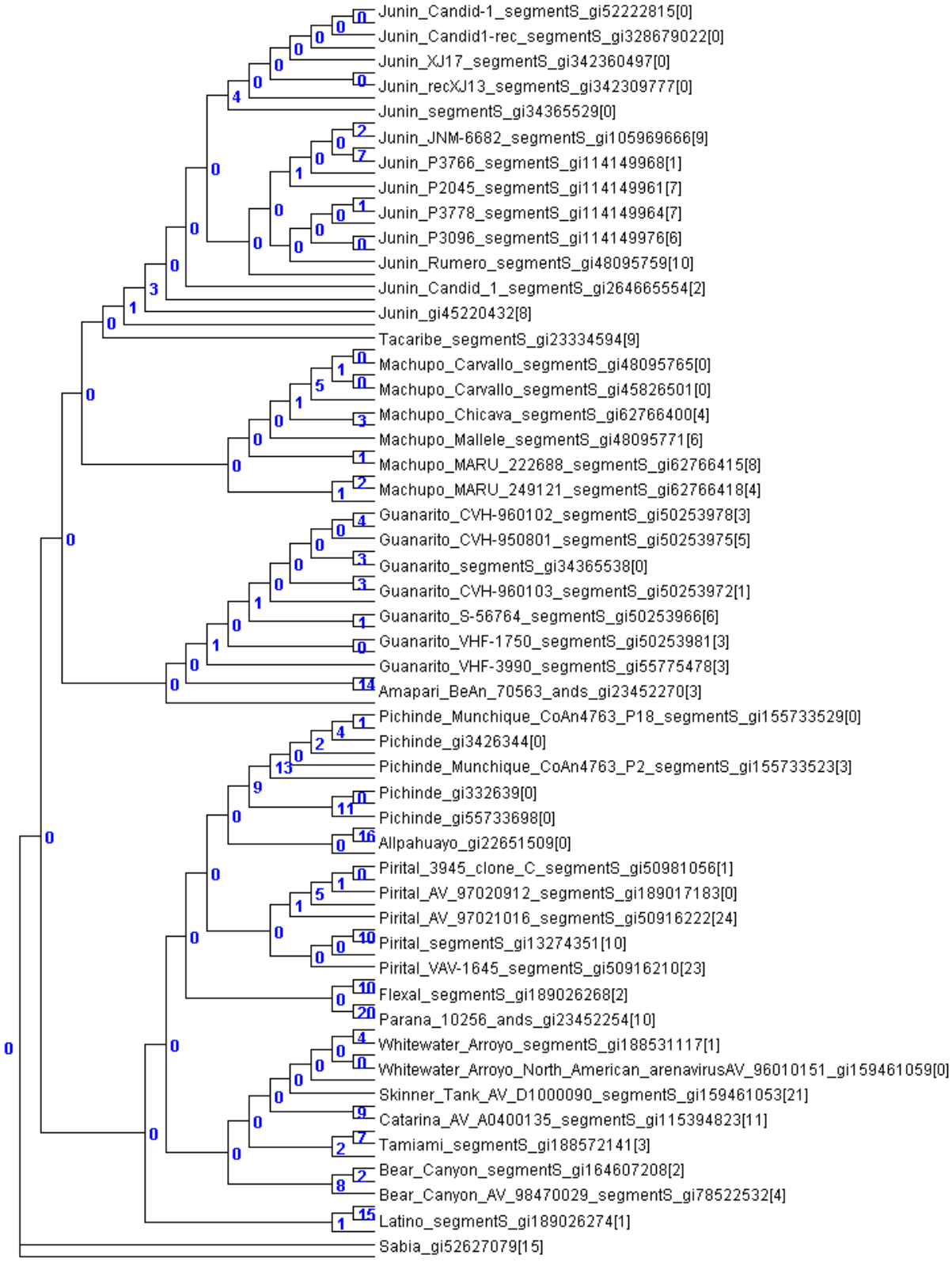
Marburg
Number_signatures: 169
Number_Homoplastic_signatures: 91



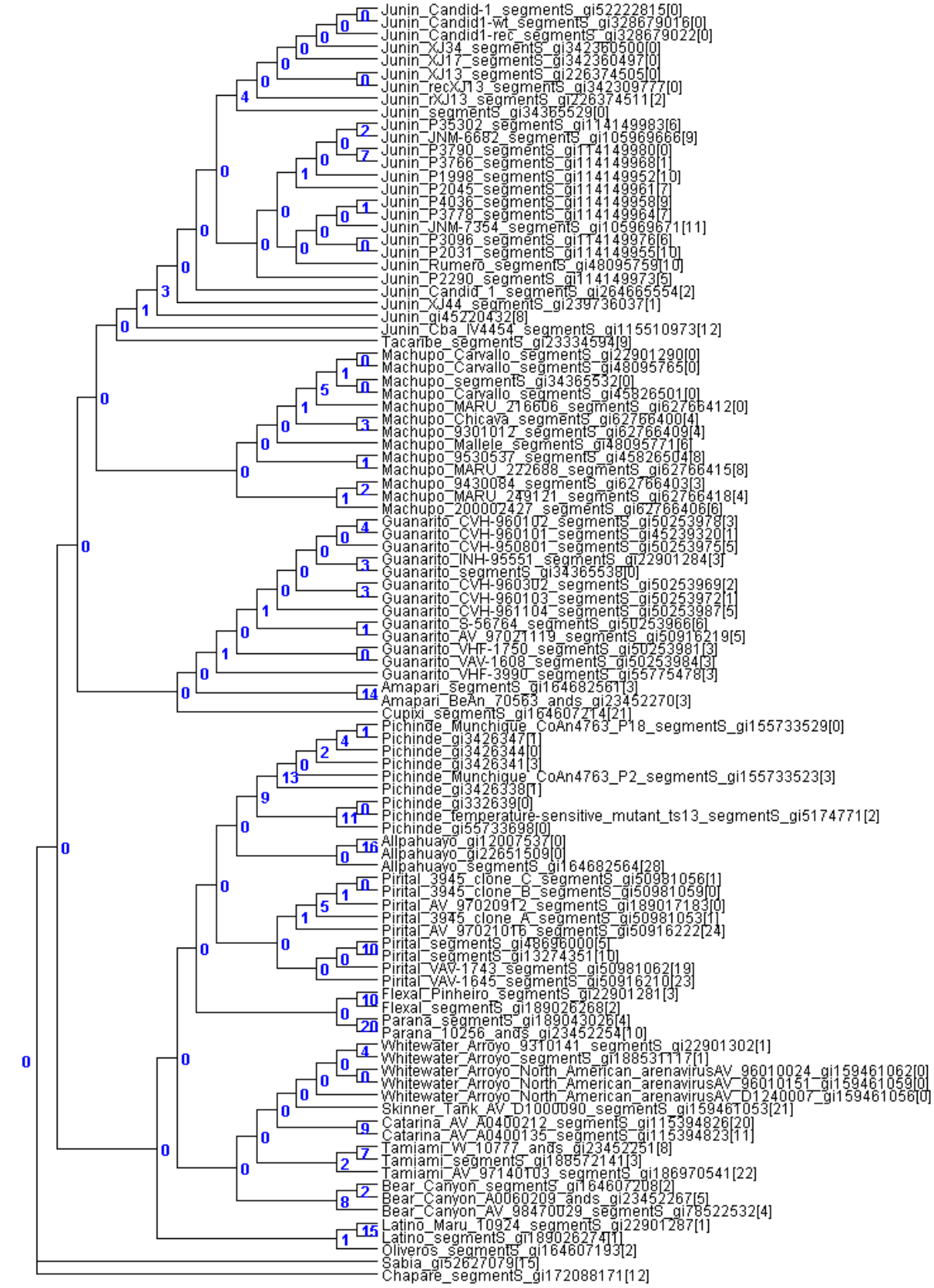
New World Arenaviruses Segment L
Number_signatures: 478
Number_Homoplastic_signatures: 31



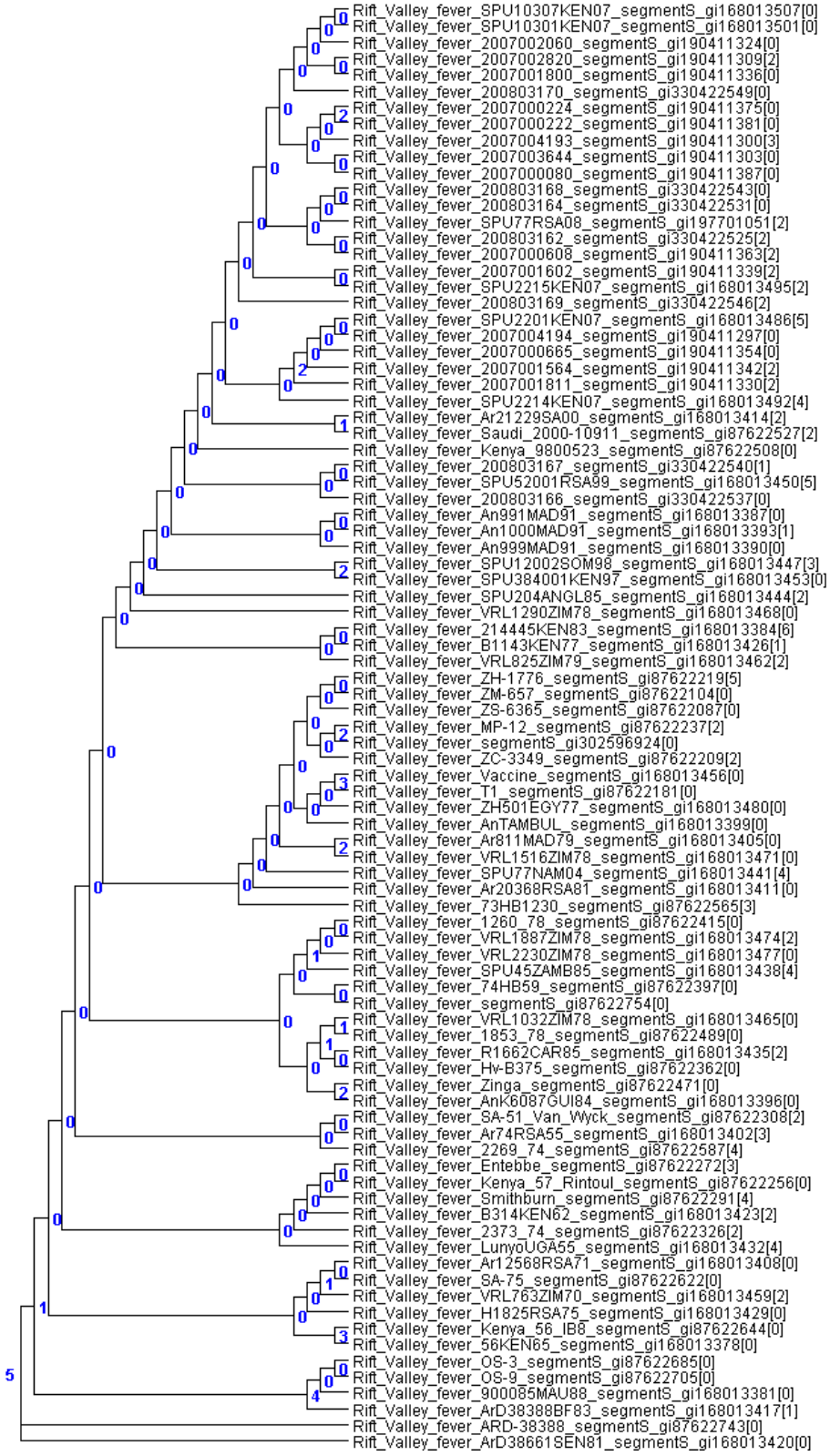
New World Arenavirus Segment S (sparse labels)
Number_signatures: 898
Number_Homoplastic_signatures: 183



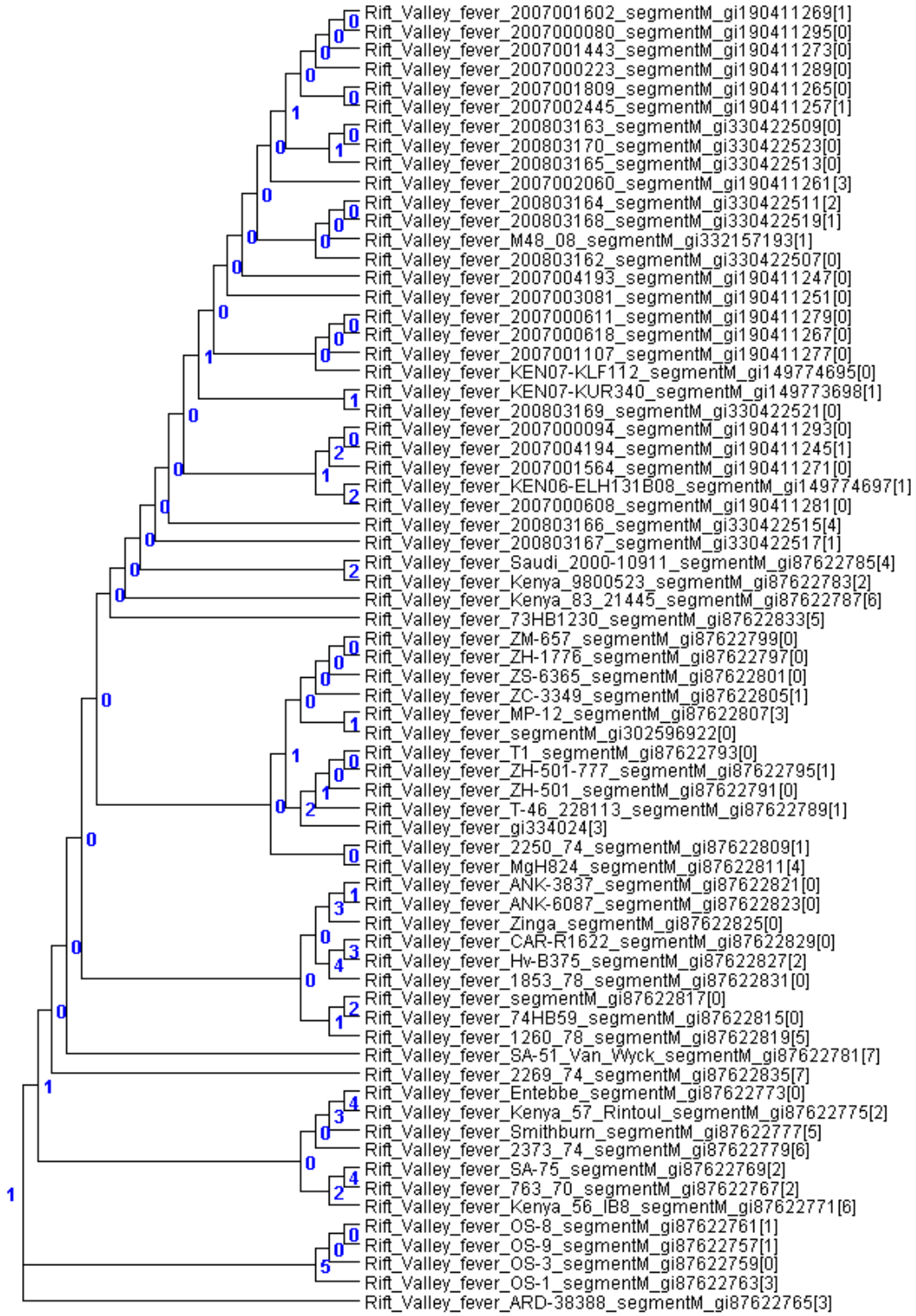
New World Arenaviruses Segment S (all labels)



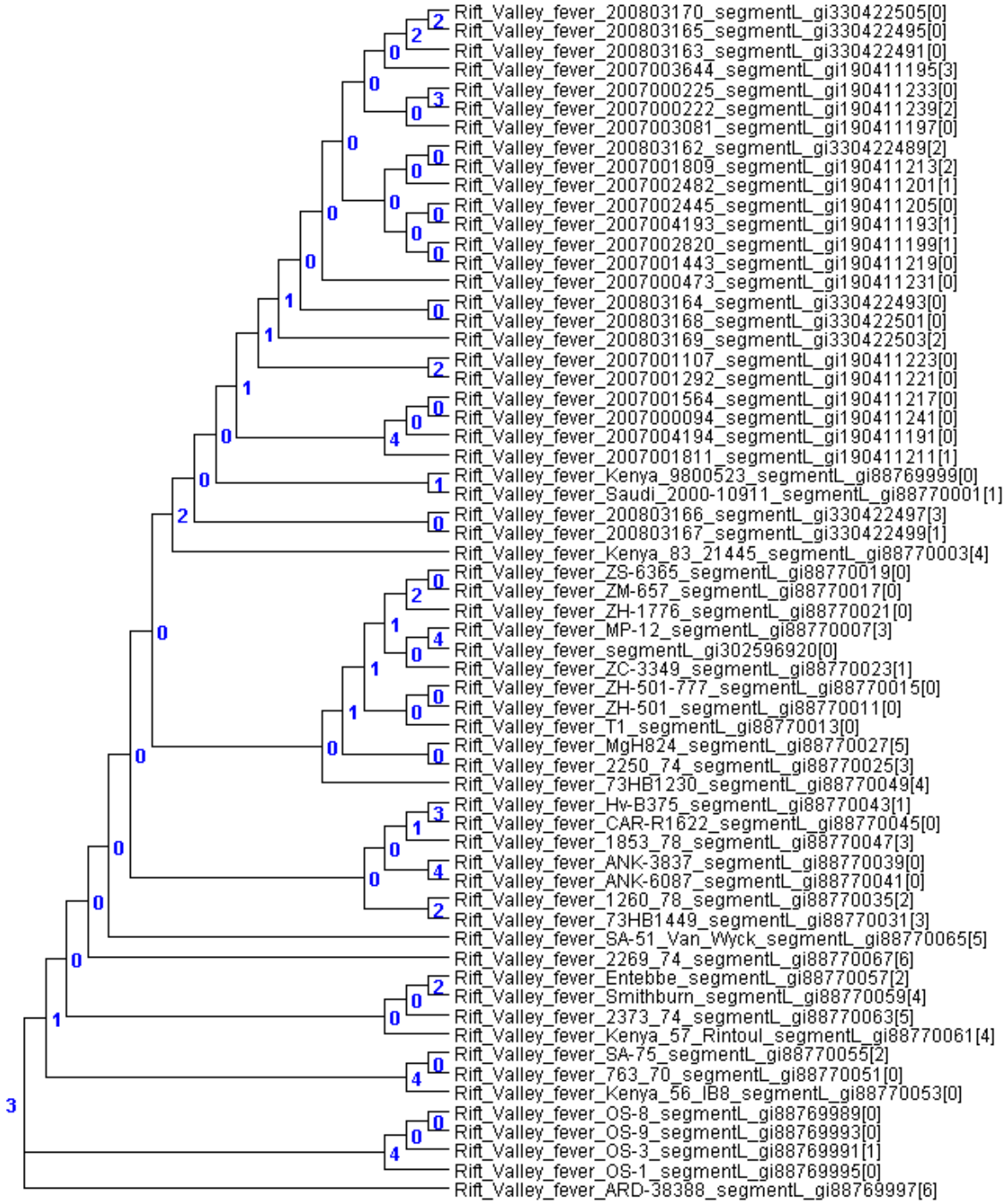
RVF segment S
Number_signatures: 216
Number_Homoplastic_signatures: 77



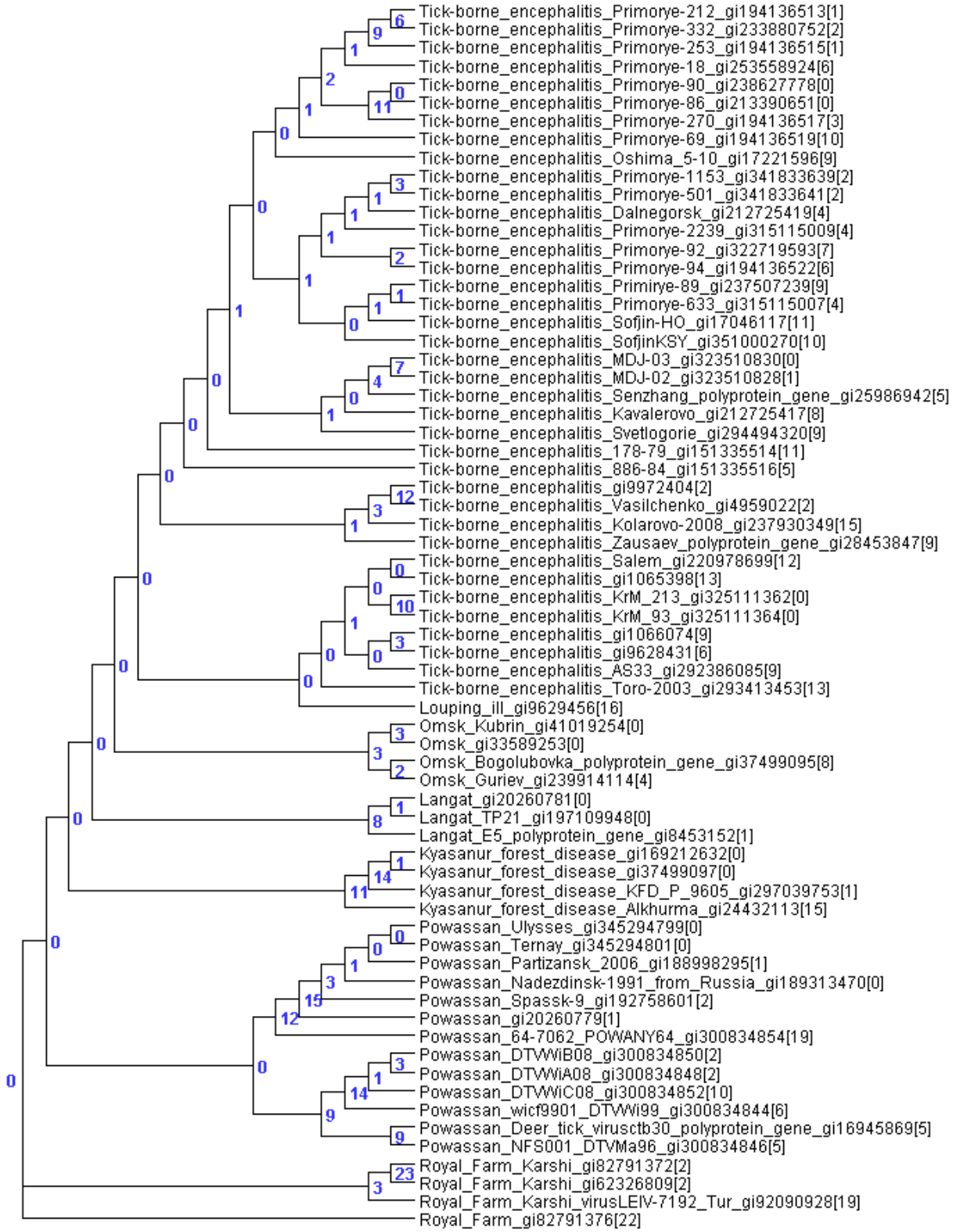
RVF Segment M
Number_signatures: 219
Number_Homoplastic_signatures: 69



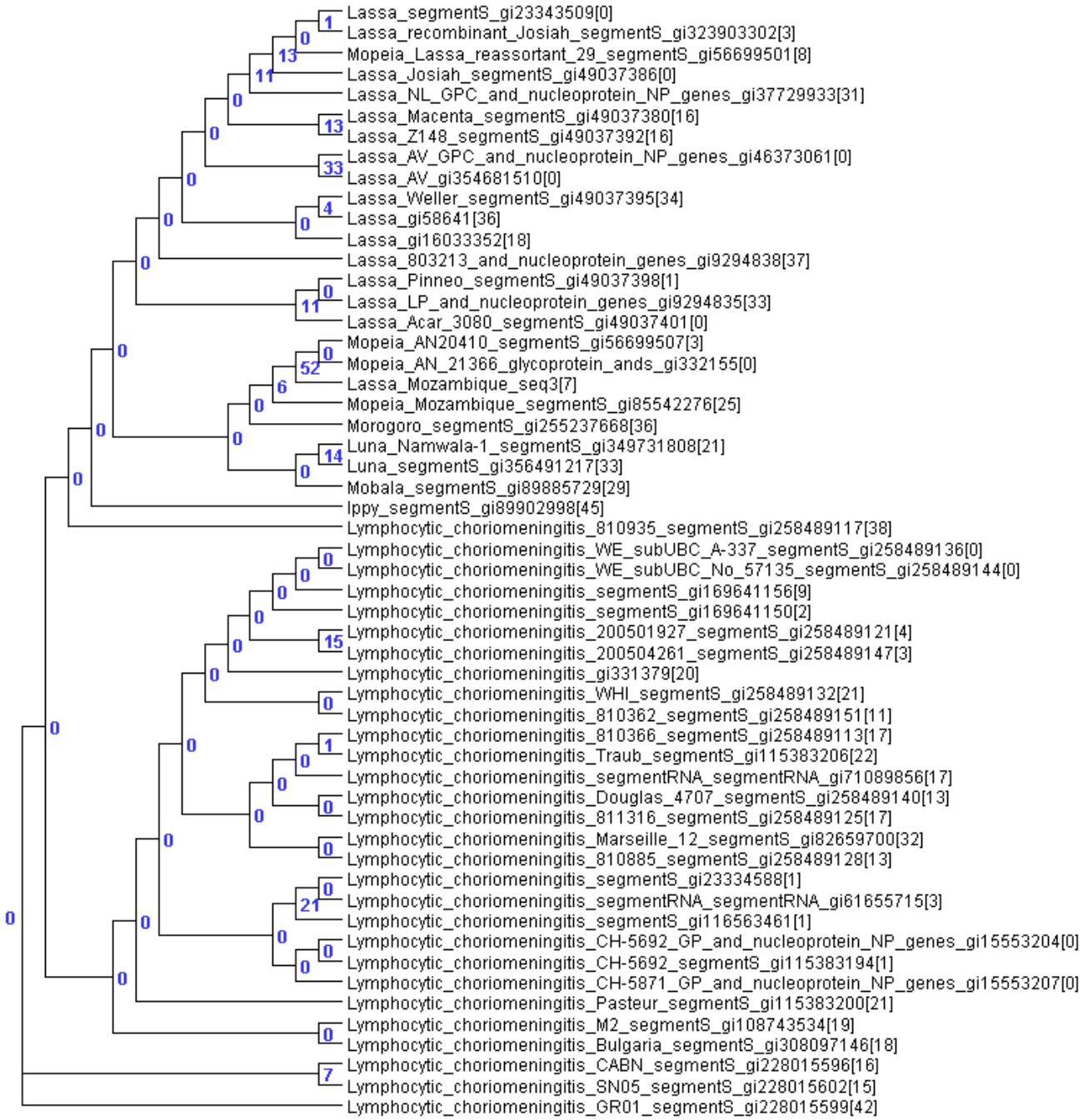
RVF Segment L
Number_signatures: 206
Number_Homoplastic_signatures: 70



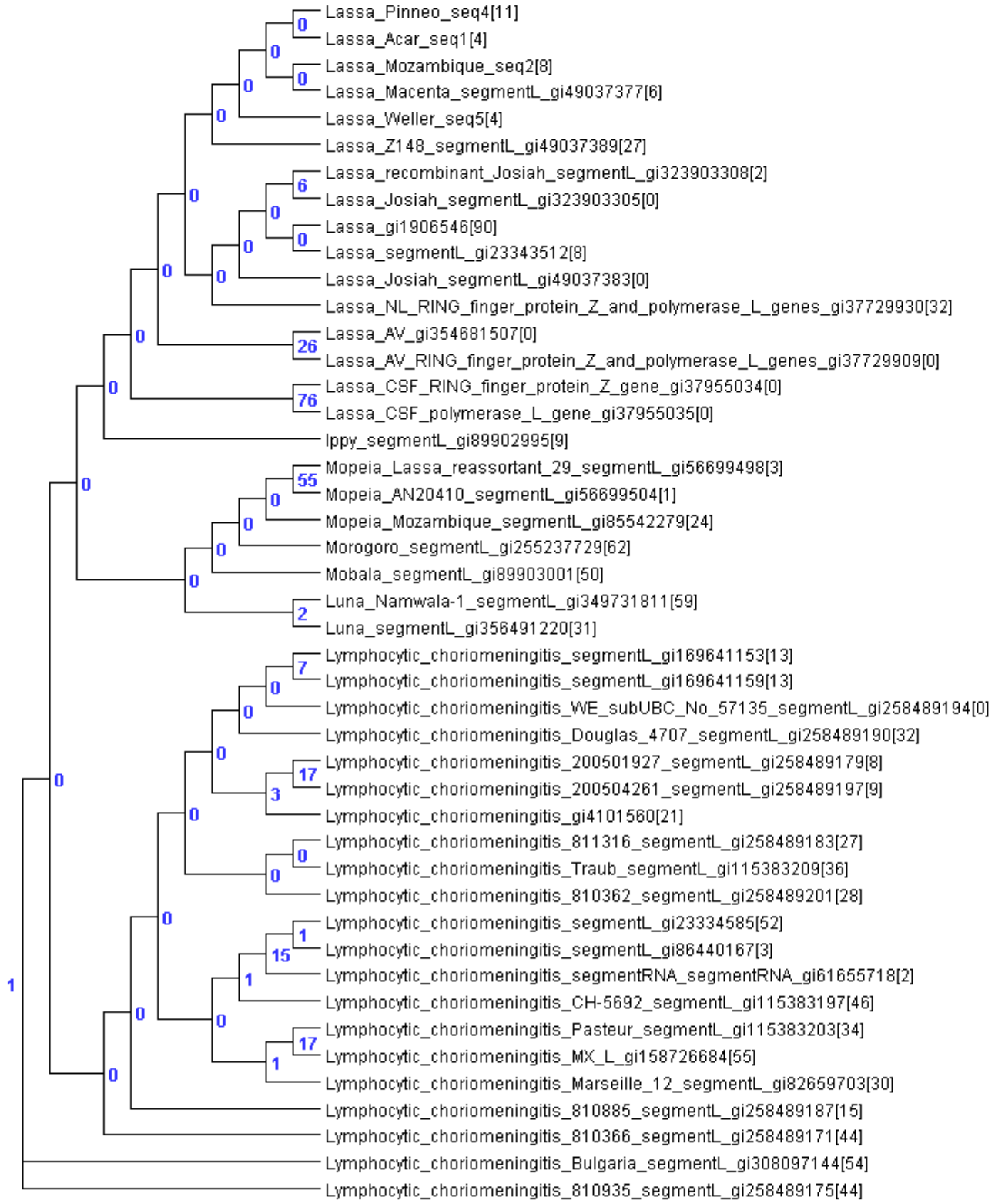
Tick-borne encephalitis complex
Number_signatures: 765
Number_Homoplastic_signatures: 182



Old World Arenavirus Segment S
Number_signatures: 2348
Number_Homoplastic_signatures: 1338



Old World Arenavirus Segment L
Number_signatures: 2640
Number_Homoplastic_signatures: 1415



# References

Aaron C.E. Darling, Bob Mau, Frederick R. Blatter, and Nicole T. Perna. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*. 14(7):1394-1403.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput.*Nucleic Acids Res*. 32(5):1792-1797.

Edgar,R.C. (2010), Search and clustering orders of magnitude faster than BLAST, *Bioinformatics*. doi: 10.1093/bioinformatics/btq461

Stamatakis, A. 2006. RAxML-VI-HPC:Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics.* 22(21):2688–2690.