LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-JRNL-494293

Bioinformatics for Microbial Genotyping of Equine Encephalitis Viruses, Orthopox Viruses, and Hantaviruses

S. N. Gardner, C. J. Jaing

August 19, 2011

Journal of Virological Methods

Disclaimer

Bioinformatics for Microbial Genotyping of Equine Encephalitis Viruses, Orthopox Viruses, and Hantaviruses

Shea N. Gardner[1]* and Crystal J. Jaing[1]
[1]Global Security, Lawrence Livermore National Laboratory, Livermore, CA 94550
*Author for correspondence

Gardner26@LLNL.gov, phone 925-422-4317, fax 925-422-6736

Abstract

Microbial genotyping is essential for investigating pathogen forensics, tracking epidemics, and understanding evolutionary processes. We performed phylogenetic analyses and designed genotyping assays for all available genomes from five viral species complexes or genera: Western, Eastern, and Venezuelan equine encephalitis virus complexes, Hanta virus genus segments L, M, and S, and Orthopox virus genus. For each virus group, whole genome Multiple Sequence Alignments (MSAs) and phylogenetic trees were built. PCR signatures composed of primer pairs or TaqMan triplets were designed and mapped to the nodes of the trees for sub-type or strain specific PCR-based identification. In addition, SNPs were identified and mapped to the nodes of phlyogenetic trees, and SNP microarray probes were designed to enable highly multiplexed genotyping of an unsequenced sample by SNP array hybridization. Near-perfect isolate resolution was possible for all viruses analyzed computationally using either SNPs or PCR signatures. More tree nodes were represented by SNP loci than by PCR signatures, as PCR signatures more often represented subsets of strains that did not correspond to a branch of the tree. However, while PCR genotyping is possible, the number of PCR signatures needed to characterize an unknown relative to the tree can be very large. SNP microarrays are a suitable alternative, as arrays enable highly multiplexed, high resolution genotyping of an unknown in a single hybridization assay. All TaqMan signatures, SNPs and microarray probes are available as supplementary information.

Introduction

Forensic characterization of select agent viruses requires the detection of reliably measured molecular variations between related viral strains. Critical characteristics of viral typing include universality, sensitivity, specificity, efficiency, reproducibility and resolution. Although genomic sequencing provides the highest resolution of measuring molecular variations and the cost of sequencing is rapidly decreasing, it is not yet feasible to type every strain of every viral pathogen of interest by sequencing. Additionally, modern sequencing platforms are geared for efficiently sequencing human genomes and are quite expensive if used to sequence a single viral sample. Currently some PCR assays have been developed on selected viruses for detection at the family level, but there is a lack of sensitive and reliable assays for forensic discrimination of select agent viruses at the strain and isolate level.

One approach pioneered at Lawrence Livermore National Laboratory (LLNL) is the use of high-density microarrays to detect bacterial and viral pathogens and to do forensic analysis of biothreat agents. Using this approach, we built an array to determine the presence/absence of genes known or suspected to be tied to mechanisms of virulence and antibiotic-resistance, or vectors that could indicate bacterial genetic engineering (1,2). In prior reports to the Department of Homeland Security (DHS), we showed microarray results demonstrating that this “functional forensics fingerprinting” can readily discriminate between species and strains of organisms for which we have sufficient knowledge of those mechanisms. We also developed a Lawrence Livermore Microbial Detection Array to detect any sequenced virus or bacteria within 24 hours (3). We used this array to identify a contaminating pig virus from a rotavirus vaccine, a vaccine used worldwide to prevent rotavirus infection in infants (4). We also used this array to diagnose viral infections from human clinical samples (5). This work is helping to rapidly transition microbial forensics from a marker-based science into a sequence-based one, whether the sequence data is determined via whole-genome sequencing or via microarray probes.

We have demonstrated how arrays with all known SNPs can provide accurate forensic determination at a low cost (unpublished reports to DHS). We developed a high-throughput approach called kSNP for polymorphism detection and assay validation, which combines automated analysis of draft and finished genome sequence data with rapid microarray development and testing (6). Candidate

SNP loci are first identified via sequence analysis, and then used to design probes for a high-density oligo microarray, which can be quickly fabricated commercially by Maskless Array Synthesis (Roche NimbleGen) (7) or ink-jet (Agilent) technologies (8). The array is then tested with a panel of strains of the species of interest. This dual-platform use of sequencing and arrays easily distinguishes true polymorphisms from loci that appear polymorphic due to sequencing errors encountered with modern platforms. PCR primer sets can then be designed for the validated loci; in addition, the array itself can be used for rapid genotyping of uncharacterized isolates. We applied this approach to several hundred bacterial and viral genomes, and empirically tested it on six species and dozens of isolates as part of our work for the National BioForensic Analysis and Countermeasures Center (NBFAC). Our estimated call error rates for microarray probes averaged 0.5% for SNP loci when the array was hybridized with various strains of *B. anthracis* (unpublished data).

The overall goal of this project was to design high resolution and cost-effective genotyping assays for strain level forensic discrimination of select agent viruses, addressing a significant capability gap for the viral bioforensics and law enforcement community. We used a multipronged approach combining phylogenetic analysis, TaqMan signatures, single nucleotide polymorphism (SNP) markers and microarray probes to comprehensively characterize the viral sequences of Western, Eastern, and Venezuelan equine encephalitis viruses, hantavirus L, M, and S segments, and orthopoxviruses. When combined with lab data, the bioinformatics presented here should facilitate future studies to correlate genotype with phenotypes such as virulence, transmission and antigenicity.

Methods

MSAs

MSAs were built with MUSCLE (http://www.drive5.com/muscle/ ; (9)) (WEE, VEE, EEE and Hantavirus) or MAUVE (http://gel.ahabs.wisc.edu/mauve/ ; (10)) (Orthopoxvirus) when MUSCLE exceeded memory before completion. Table 1 summarizes the fraction of the genome conserved across all available sequences for each target set based on the MSAs. Maximum likelihood phylogenetic trees were constructed from the alignments using RAxML v7.2.7 (http://wwwkramer.in.tum.de/exelixis/software.html ; (11)) with the GTR substitution matrix and the GAMMA-based likelihood as described in the RAXmL manual (–m GTRGAMMA option, other parameters were RAxML defaults, and the best scoring maximum likelihood tree is reported).

Data files containing the complete fasta header information for all of the genome identifiers shown on the trees are provided as supplementary information. The genome identifiers on the trees have an appended "_#" to ensure that each sequence identifier fed into the kSNP software is unique, since not all genomes had strain information.

PCR signatures

Signature candidates were designed with Multiple Set Clustering (12) and Primux (13), an improved version of the MPP algorithm described in (14) to include degenerate bases. These signatures included primers and probes with and without degenerate bases. Amplicons were 80-250 bp, primer T_m's approximately 60-65°C, primer lengths 18-22 bases, probes 18-30 bases, and probe T_m's approximately 68-73°C.

TaqSim software (http://staff.vbi.vt.edu/dyermd/publications/files/TaqSim_Help.pdf) was used to predict which signatures should detect which targets, assuming no mismatches of primers and probes to targets. Signatures were either primers only or TaqMan triplets (primers with a probe). Many of the primer pairs were conserved among many sequences within a target set, and adding probes provided additional discrimination. Mix and match alternative combinations of forward and reverse primers and

probes were considered different signatures, even if some shared a primer or probe. Hundreds tothousands of signatures were designed for the sequences in each target set (Table 1).

Signatures were mapped to the nodes of the MSA-based tree. Signatures predicted to detect any cluster of targets that did not correspond to a node, that is, the exact set of sequences down a branch on the tree, were considered homoplastic. Whether a signature maps to a node or is homoplastic depends on the tree topology and the root ancestor/descendant relationship. Alternative roots for a tree do not change the tree branching patterns; they do change the ancestor/descendant relationships. Since no outgroups were used for building the trees, the MSA-based trees were unrooted. Therefore, we drew each tree using the root that minimized the number of homoplastic signatures: Signatures were mapped to each node for all possible roots of the tree. The root for which the tree had the fewest homoplastic signatures was selected as the best root.

SNPs

We applied the kSNP software to find single nucleotide polymorphisms (SNPs) in whole genome data (6).This is an alignment free method based on k-mer (oligos of length k) analyses. A SNP locus is defined by the sequence context of length k surrounding the SNP (k-1)/2 bases either side of the SNP with a variant SNP allele at the central base. This representation of a SNP locus is based on surrounding sequence information rather than positional information in a genome. It differs from traditional alignment-based concepts of a SNP locus, and it allows us to consider draft genomes which are available only as contig fragments in which positional information relative to the complete genome is not known. kSNP is also useful for viruses in which there may be highly divergent and poorly alignable regions among a large group of sequences, and conserved regions only exist among small subgroups of sequences. There is no bias that otherwise results from the choice of a reference sequence or from considering only a subset of regions of the genome that can be easily or quickly aligned. kSNP scales to hundreds of bacterial or viral genomes, and can be used for finished and/or draft genomes available as unassembled contigs. The method is fast to compute, finding SNPs and building a SNP-based phylogeny in seconds to hours. Here, SNP-based trees were calculated from a simple hamming distance metric of the number of SNP allele differences between each target sequence, and SNP alleles were mapped to the nodes. For moderately diverse target sets like viral species complexes, SNP-based trees are consistent with MSA-based trees. However, in cases where target sets contain viruses of different species, SNP trees cannot always distinguish higher level relationships between different viral species since variations are larger than single nucleotide changes. In such cases when the correspondence between the SNP-based tree and the MSA tree was poor, we mapped SNP alleles to the MSA-based tree instead.

To improve accuracy, memory efficiency, and speed, we made some modifications from the kSNP software described in (6): 1) k-mer enumeration with a suffix array was replaced with a new hash table implementation using the Jellyfish software (15); 2) Calls to blastn (16) to find all candidate loci in all genomes were replaced with calls to MUMmer (17); 3) trees were rooted so as to maximize the number of SNP alleles that map to nodes of the tree (to minimize homoplastic SNPs), as described above for PCR signatures.

SNP analysis was performed with k=13 and k=25. k=13 identified more SNP alleles than k=25. Fewer SNPs were found with the larger k because a longer length of conserved sequence surrounding the SNP is required. With these divergent viruses, shorter k means that SNPs in closer proximity to one another can be found, thus reducing the stringency for conservation surrounding a SNP. A value of k=13 for viruses should provide better resolution of unsequenced novel isolates than k=25, so all results reported below are for k=13.

Microarray Probe Design

Microarray probes were designed for every SNP. Probe design strategy maximized sensitivity and specificity based on extensive prior lab testing on a Roche NimbleGen microarray platform, where we demonstrated 100% SNP allele call rates and 99.5% accuracy (in prep, and unpublished reports for DHS). We determined that maximum sensitivity and SNP discrimination accuracy result if the SNP base is at the 13th position from the 5' end of the probe (the end farthest from the array), probes are between 32 and 40 bases long, and length varies so as to equalize hybridization free energy (ΔG) to the extent possible within the allowable length range. Probes shorter than 32 bases have high false negative rates, and longer probes are inefficient at discriminating single base mismatches. We found that ΔG is a better predictor of hybridization than T_m. Probe candidates with hybridization free energy below ΔG=-43 kcal/mol were shortened until either their ΔG exceeded -43 kcal/mol or they reached the minimum 32 bases. Probes were designed around the SNP on both the plus and minus strands, for all four possible SNP alleles, and all surrounding sequence variants. We design probes for both the plus and minus strands; these are not the reverse complements of one another because the SNP does not lie at the center of the probe. There are probes for each of the four variants on each strand, so at least eight probes per SNP locus. In addition, any sequence variation outside of the k-mer SNP context of conserved bases is captured in multiple alternative probes for that allele, so there may be more than 8 probes per SNP locus, although for a given hybridization, only the probe variant with the best signal is used for assessing the SNP allele at the 13th position. Finally, probes are trimmed from the 3' end to remove any N's or other degenerate bases, and omitted altogether if doing so results in a probe less than 32 bases. If a probe is a subsequence of any other, only the shorter of the two is kept. If necessary to fit on the desired array format, probes can be omitted for alleles not represented in the target sequences, e.g. for biallelic SNPs and some of the possible probe variations outside the conserved k-mer context for some alleles. Pruning to the subset to detect only observed alleles in the available genomes dropped the probe counts by over 70%. Both unpruned and full sets of probes are provided as supplementary data. These probe counts could be further reduced, for example, by including probes for only a subset of the SNPs for each node or homoplastic group.

All data presented/discussed here is available as supplementary information or by contacting the authors.

Results

Trees based on MSA's or SNPs are shown in Figures 1-7, showing either branch lengths that scale with genetic distance (MSA trees) or the number of PCR signatures or SNP alleles that map to each node (in blue at the node) and leaf (in brackets after the strain name). Counts of homoplastic PCR signatures or SNPs cannot be shown on the tree. Table 1 summarizes the numbers of genomes, PCR signatures, SNPs, and homoplastic markers for each viral target set.

The vast majority of PCR signatures are strain-specific (mapping to the leaves) and or homoplastic, detecting combinations of sequences that do not map exactly to a node of the MSA trees. The homoplastic signatures nonetheless may be useful; some signatures may detect most of the members of a clade. For the Orthopoxviruses, the majority of signatures are homoplastic, predicted to detect various subsets of the target sequences.

Where branch lengths are very short, usually PCR signatures can be found. Few signatures map to the more distantly related targets separated by long branch lengths like those seen between different species. Many of the antigenic subgroups and species within the equine encephalitis virus complexes and Hantavirus genus do have PCR signatures at forensic level resolution. In cases where zero signatures map to a node of interest, one would need to use a combination of homoplastic signatures or leaf-level signatures to detect and discriminate all the desired targets.

For the Orthopoxviruses, there are three variola-specific PCR signatures, three variola minor specific signatures, two vaccinia, seven camelpox, one taterapox, and six ectromelia specific signatures.

Cowpox falls into two clades. There are some subtype and strain signatures, but due to the relatively high level of conservation across this double-stranded DNA virus, combinations of homoplastic signatures would likely be required to classify an unknown down to a forensic level using PCR signatures.

Almost all targets had substantially higher fraction of nodes without PCR signatures than without SNPs (Figure 8). Likewise, most targets had a larger fraction of PCR signatures that were homoplastic compared to the fraction of SNPs that were homoplastic. Homoplastic PCR signatures often were predicted to detect only a subset of the strains down a branch. The one exception was Hanta segment L, for which the vast majority of PCR signatures were genome-specific, that is, they mapped to the leaves, and so were not considered homoplastic.

Comparisons between MSA and SNP trees for each viral group

The MSA and SNP trees for WEE viruses are nearly the same, except that the MSA tree has the Fort Morgan and Buggy Creek as nearest neighbors to the Highlands J viruses, but the SNP tree has the branch to Fort Morgan and Buggy Creek viruses from the root, since there were no SNP loci present in both clusters to indicate that they are more closely related to one another than to other species in the WEE virus complex. There are no PCR signatures at this node either, nor is there a single PCR signature that classifies all Sindbis viruses together.

MSA and SNP trees for VEE are the same for the upper part of the trees, but SNPs cannot correctly place the divergent Pixuna, Cabassou relative to the others, since each is very divergent from other members of the species complex, and branches off near the root of the tree. These isolates do share some SNPs with other isolates, but only homoplastic SNPs that do not correspond to any nodes of the SNP-based tree. Nor are there PCR signatures for many of the nodes at the bottom of the tree that lead to the Cassabou and other divergent isolates.

MSA and SNP trees for EEE are similar. The BeAr strain and the PE-3 and PE-0 are relatively different from the other larger group of EEE viruses that cluster tightly at the top of the tree, and it is difficult to tell whether BeAr or the PE-3/PE-0 cluster is closer to the large cluster. For EEE, only the Patent WO2005000881 and Florida91-4697 sequences cannot be discriminated based on SNPs. The patent sequence is a live attenuated vaccine derived from the Florida91-4697 parent isolate. It has five deleted codons at the furin cleavage site. All other branches and strains can be discriminated based on SNPs, most by dozens of loci.

The SNP-based tree for OPXV does not correspond to the MSA tree or other published work (REF), and most of the internal nodes have zero SNPs that map to them. The SNP tree captures strain-level relationships, but SNPs are not adequate to determine relationships among different species. Mapping SNPs instead to the MSA-based tree shows that SNPs do correspond to most of the nodes of the MSA tree, showing that there are species- and clade-specific SNPs. The Orthopox tree shows that there are 3067 species-specific SNP alleles for variola and 442 subtype-specific SNP alleles for variola minor. There are hundreds of clade-specific SNPs at most levels of the tree, which should enable high confidence classification of an unknown isolate.

OPXV illustrates a case where SNPs can provide excellent forensic genotyping capabilities but only if they are combined with the tree derived from a full genome sequence alignment, since SNPs alone do not build an adequate tree. Of the nearly 800 PCR signatures we generated, very few of them mapped to internal nodes of the MSA tree, while the vast majority mapped to leaves or to strain groupings that did not correspond exactly to a branch of the tree. Therefore PCR signatures may be a poor choice for phylogenetic subtyping of OPXV compare to SNPs; To categorize an unknown to any given branch would require testing against combinations of many PCR signatures, since most nodes have no PCR signature corresponding exactly to the strains down that branch.

MSA and SNP trees for Hanta L segment have nearly identical branching topology for species within the genus. The SNP tree cannot distinguish the branch structure among some species (nodes at the far left of the tree) where there are zero SNP loci in common between branches, but there are

nevertheless sufficient SNPs for most branches and all leaves to place an unknown on a tree based on SNP allele calls, whether it be the MSA- or the SNP-based tree. Few of the internal nodes have PCR signatures, so again one would need to test a large panel of PCR signatures capturing the variation near the leaves against an unknown to place it on the tree.

As for OPXV, MSA-based trees for Hantavirus segments M and S showed inter-species relationships more clearly than did SNP-based trees, since although SNPs could differentiate species, they could not determine higher level clustering between species. Therefore, the SNPs were mapped to the MSA tree instead of the SNP tree. Segments M and S have over a hundred sequences each, making visualization on a tree difficult. There are dozens of species, as well as many "unclassified Hantavirus" sequences. Most SNPs map close to the leaves of the trees (species, subspecies, and strain) and do not map to higher level branches shared across species, but SNPs do allow clear clustering by species. Considering the challenge of finding genus-level conserved primers or employing the hundreds of assays it would take to classify the three segments of an unknown hantavirus by any single-plex approach such as PCR, the value of a microarray is clear for species and strain level characterization, as well as identification of new segment recombinant strains.

Discussion and Conclusions

We performed whole genome alignment, SNP discovery and microarray design, and PCR signature design for genotyping all available finished and draft genomes of WEE, EEE, VEE, Hantavirus L, M and S segments, and Orthopox. Almost every available genome can be discriminated on the basis of SNPs or PCR signatures. Trees based on SNPs alone were compared with those based on the full MSA, showing good correspondence at the strain level within a species, and that species clustered separately. For very large and diverse groups of sequences spanning an entire genus, however, SNPs were insufficient to build the topology at higher levels like the clustering between different species. For these target sets, mapping SNPs to a tree built from the full MSA proved optimal.

We include plots showing trees with branch lengths scaled by genetic distance, as well as cladograms with PCR signature and allele counts, since comparing the trees illustrate the difficulty of obtaining PCR signatures for long branches. In many cases, only the shortest branches have corresponding PCR signatures. Therefore, to place an unknown sample on the tree using PCR signatures, one usually cannot traverse the tree from the top down (root to leaf) because there are no or very few PCR signatures at the more conserved levels. Instead, the sample must be tested against many leaf level and homoplastic signatures for classification as to the most similar leaf. This issue is less problematic for SNP genotyping, since SNP alleles exist to discriminate more of the conserved nodes, and, more importantly, highly multiplexed analysis like arrays makes it possible to determine all SNP alleles in an unknown in a single assay.

Every node may not have SNP alleles that correspond exactly to that node, but those nodes are nonetheless differentiable by combining information from a number of homoplastic SNP loci. As such, the microarray approach that queries all known SNPs for an unknown isolate in a single assay is much more powerful than a set of single-plex assays for a limited number of canonical SNPs or PCR signatures that map to a small handful of the nodes of a tree. As our knowledge of viral diversity expands, the need for larger multiplexes for genotyping unknowns grows apace. Due to the large numbers of signatures and thus reactions that would be required to classify an unknown down to the subtype and strain level, TaqMan PCR reactions for forensic classification will likely be labor intensive and expensive. Using a PCR-based technology that allows higher levels of multiplexing is an option, such as a bead based assay (e.g. Luminex), although the multiplexing levels are still in the tens of assays rather than the hundreds of thousands available from a microarray.

One possible disadvatage of a SNP array compared to PCR genotyping signatures is lower array sensitivity compared to PCR. However, considering that a small sample would need to be subdivided

into aliquots for many PCR reactions, the sensitivity difference between arrays and PCR might disappear. Whole genome preamplification with random hexa- or nona-mers or virus-specific whole genome amplification with conserved degenerate primers that tile across the viral genome are possible methods to increase sensitivity prior to array hybridization (1,5). In other work, we have designed conserved, degenerate primer sets that tile across whole genomes and which should amplify all sequence variants of the viruses discussed here (manuscript in prep).

Acknowledgement

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

References:

1. Jaing, C., Gardner, S.N., McLoughlin, K., Mulakken, N., Alegria-Hartman, M., Banda, P., Williams, P., Gu, P., Wagner, M., Manohar, C. et al. (2008) A functional gene array for detection of bacterial virulence elements. PLoS ONE, 3(5), e2163. doi:2110.1371/journal.pone.0002163.
2. Allen, J.E., Gardner, S.N. and Slezak, T.R. (2008) DNA signatures for detecting genetic engineering in bacteria. Genome Biology, 9, 56.
3. Gardner, S., Jaing, C., McLoughlin, K. and Slezak, T. (2010) A Microbial Detection Array (MDA) for viral and bacterial detection. BMC Genomics, 11, 668doi:610.1186/1471-2164-1111-1668.
4. Victoria, J.G., Wang, C., Jones, M.S., Jaing, C., McLoughlin, K., Gardner, S. and Delwart, E.L. (2010) Viral Nucleic Acids in Live-Attenuated Vaccines: Detection of Minority Variants and an Adventitious Virus. Journal of Virology, 84, 6033-6040.
5. Erlandsson, L., Rosenstierne, M.W., McLoughlin, K., Jaing, C. and Fomsgaard, A. (2011) The Microbial Detection Array combined with random Phi29-amplification used as a diagnostic tool for virus detection in clinical samples. PLoS ONE, http://dx.plos.org/10.1371/journal.pone.0022631.
6. Gardner, S. and Slezak, T. (2010) Scalable SNP Analyses of 100+ Bacterial or Viral Genomes. J Forensic Res, 1, 107, doi:110.4172/2157-7145.1000107.
7. Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R. and Cerrina, F. (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat. Biotechnol., 17, 974-978.
8. Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R. et al. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol., 19, 342-347.
9. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, 32, 1792-1797.
10. Darling, A.C.E., Mau, B., Blatter, F.R. and Perna, N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Research, 14, 1394-1403.
11. Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 22, 2688-2690.
12. Gardner, S.N., Kuczmarski, T.A., Vitalis, E.A. and Slezak, T.R. (2003) Limitations of TaqMan PCR for detecting divergent viral pathogens illustrated by hepatitis A, B, C, and E viruses and human immunodeficiency virus. Journal of Clinical Microbiology, 41, 2417-2427.
13. Hysom, D., Naraghi-Arani , P., Williams, P. and Gardner, S.N. Skip the Alignment: Degenerate, Multiplex Primer and Probe Design Using K-mer Matching Instead of Alignments. In prep.
14. Gardner, S.N., Hiddessen, A.L., Williams, P.L., Hara, C., Wagner, M.C. and Colston Jr., B.W. (2009) Multiplex primer prediction software for divergent targets. Nucl. Acids Res, gkp659.
15. Marcais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers Bioinformatics 27(6), 764-770.
16. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L. (2009) BLAST+: architecture and applications. BMC Bioinformatics, 10:421.
17. Kurtz S., Phillippy A., Delcher A.L., Smoot M., Shumway M., Antonescu C., Salzberg S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.* **5(2)**:R12.

Table 1: Target count, conservation across available genomes, numbers of signatures designed, number of homoplastic signatures that do not correspond to a node or leaf of the MSA tree, number of SNP loci, number of homoplastic SNP loci based on SNP hamming distance tree (EEE, VEE, WEE, Hanta L) or MSA tree (Hanta M and S and OPXV), number of SNP probes, number of SNP probes for the alleles observed among available genomes, and the resolution of the SNP

Organism	# Target sequences	Fraction of genome conserved	# of PCR signatures	# of homoplastic PCR signatures	# SNP loci	# homo-plastic SNP loci	# SNP probes	# probes for observed alleles only	Strain resolution based on SNPs
EEE	11	0.63	9,065	4,134	936	15	15565	4588	EEE_Florida91-4697 and EEE_PatentWO2005000881 cannot be resolved by SNPs
VEE	38	0.47	13,577	4,510	4421	285	96,945	26,430	All resolved
WEE	36	0.37	5,926	1,595	3382	177	68,257	19,745	All resolved
Hanta segment L	32	0.36	2,011	64	4992	243	107,753	27,591	All resolved
Hanta segment M	129	0.05	4,712	809	9914	1084**	271,017	70,145	All resolved
Hanta segment S	265	0.10	945	365	9066	1806**	343,621	89,768	Hantavirus_Hantaan_AP1168 and Hantavirus_Hantaan_AP1371 cannot be resolved by SNPs
Orthopox	121	0.49	791	471	82493	10781**	2,354,505	625,296	Taterapox_NC_008291.1 and Taterapox_Dahomey cannot be resolved by SNPs

* the 2 SNP unresolvable taterapox genomes are most likely duplicate genomes that resulted because we used the pre-publication sequences from collaborators as well as the sequences that were later entered in Genbank. Without confirmation from our collaborators that the genomes are different versions of the same isolate, we keep them both in our database.

**Homoplastic SNP counts based on mapping SNPs to the MSA tree rather than the SNP-based tree.

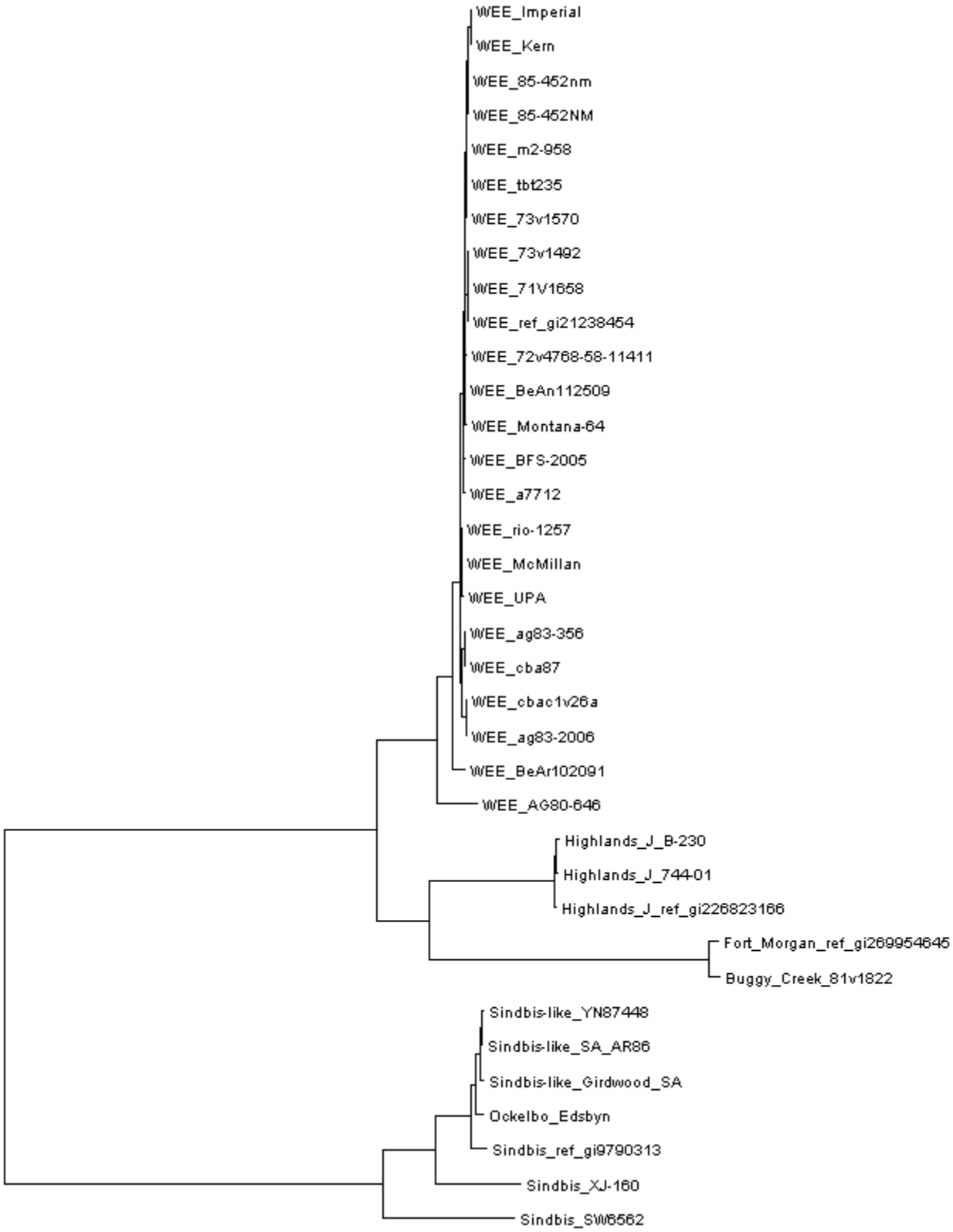

Figure 1A: Maximum likelihood phylogenetic tree for WEE based on a whole genome multiple sequence alignment, showing branch lengths scaled by genetic distance.

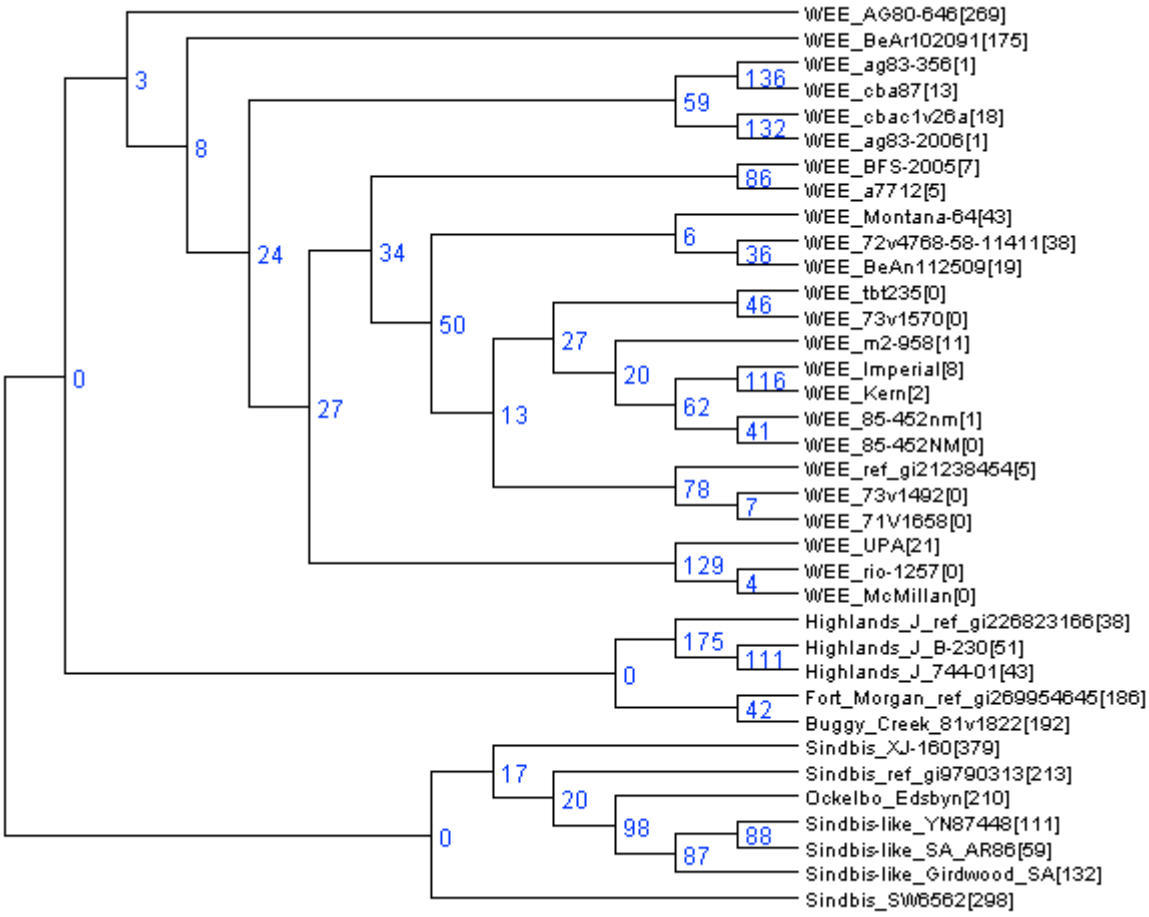

Figure 1B: The MSA tree for WEE (from Figure 1A) with counts of the number of signatures (TaqMan or primer pairs) that map to the sequences on that branch. Strain-specific signature counts are in brackets after strain names. Branch lengths are not to scale (cladogram format). The longer branches in Figure 1A have zero signatures, and only the short branches have many PCR signatures, so it is difficult to view the PCR signature counts on the phylogram in Figure 1A with accurate branch lengths.

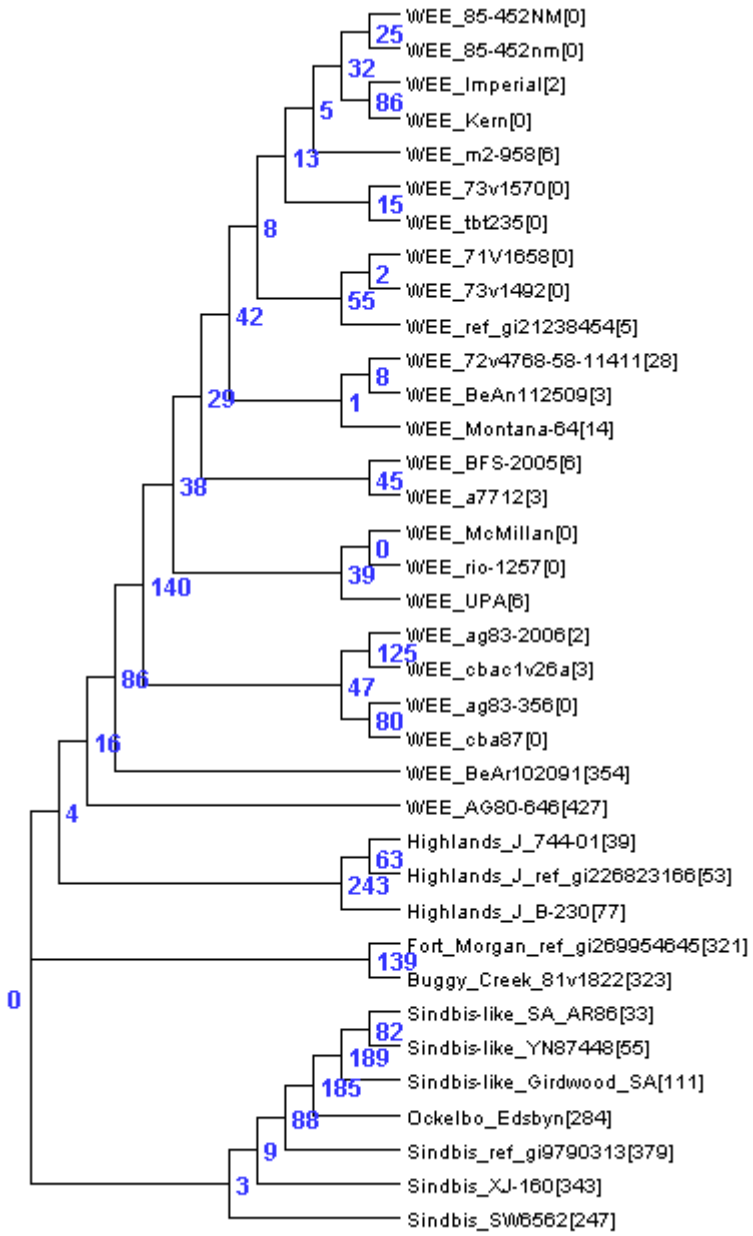

Figure 1C: SNP-based tree for WEE with counts at each node of the number of SNP alleles shared by the genomes down that branch, and genome-specific allele counts shown in brackets after the strain name.

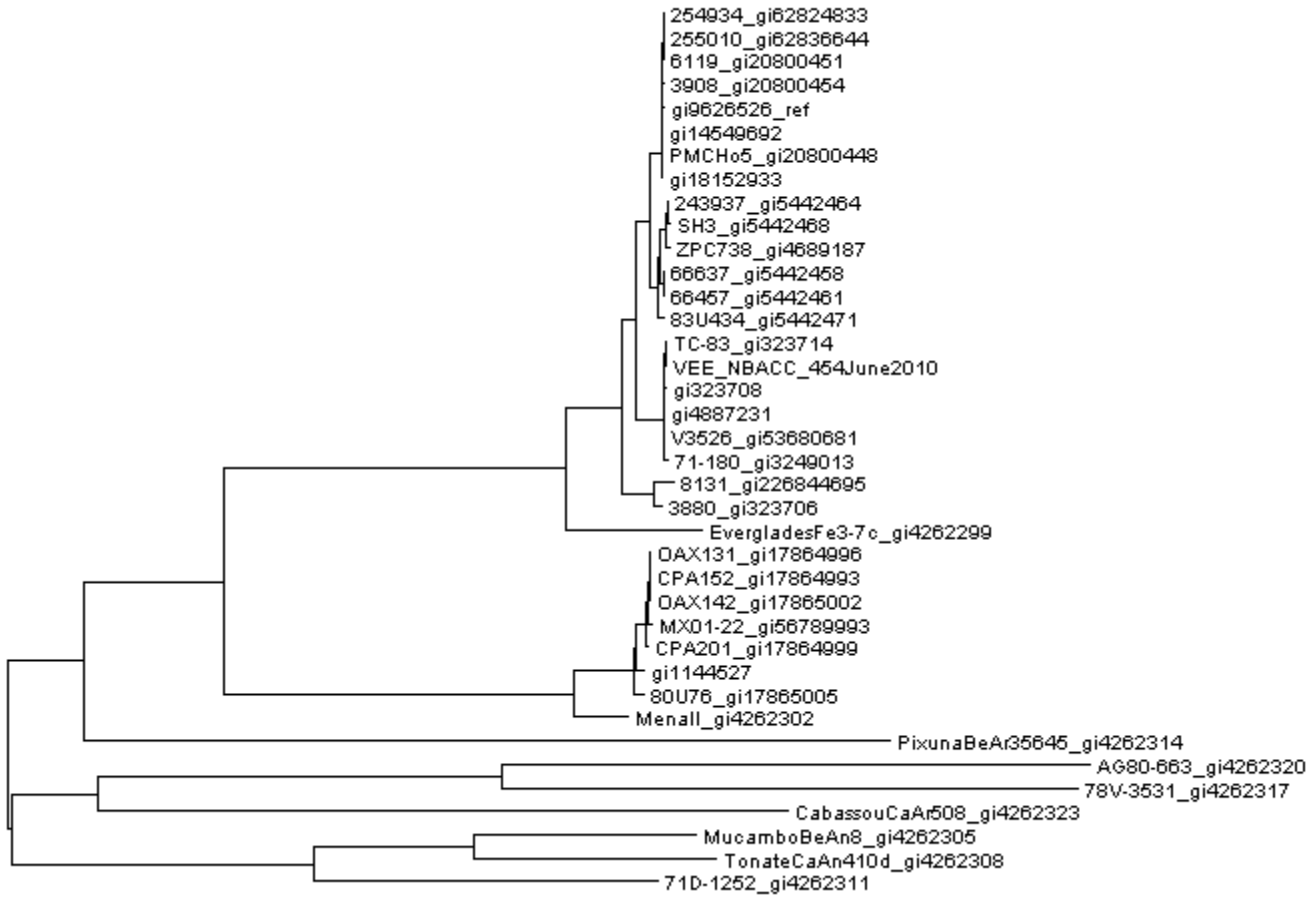

Figure 2A: MSA-based tree for VEE, as in Figure 1A.

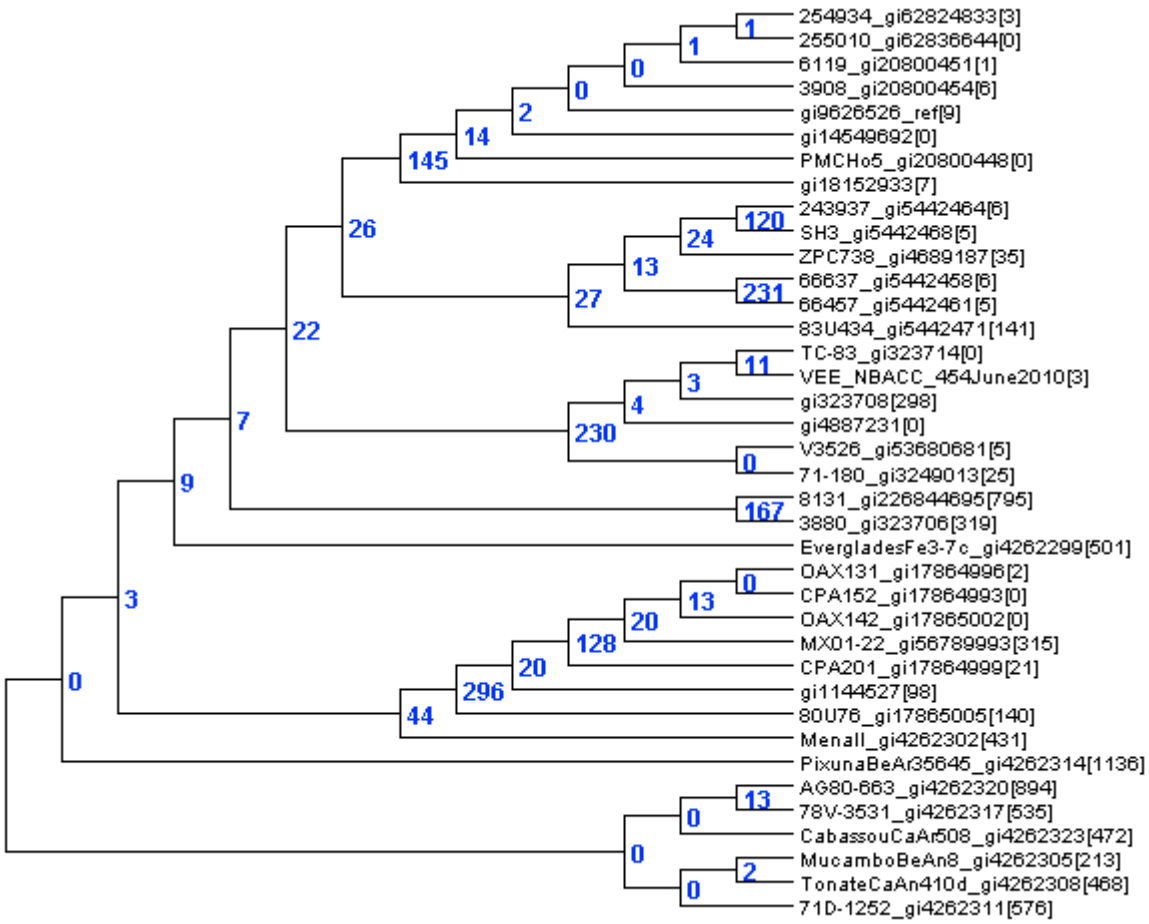

Figure 2B: PCR signature counts mapped to nodes, as in Figure 1B, for VEE.

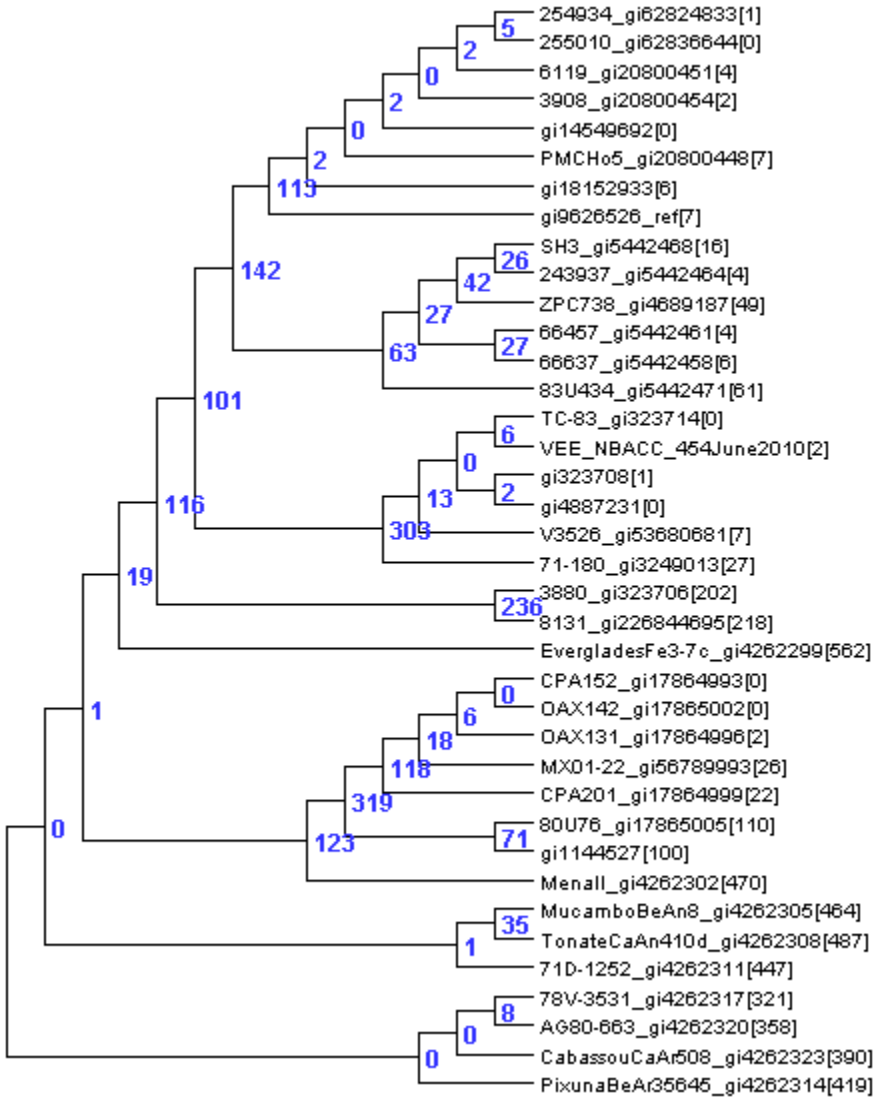

Figure 2C: SNP-based tree for VEE, as in Figure 1C.

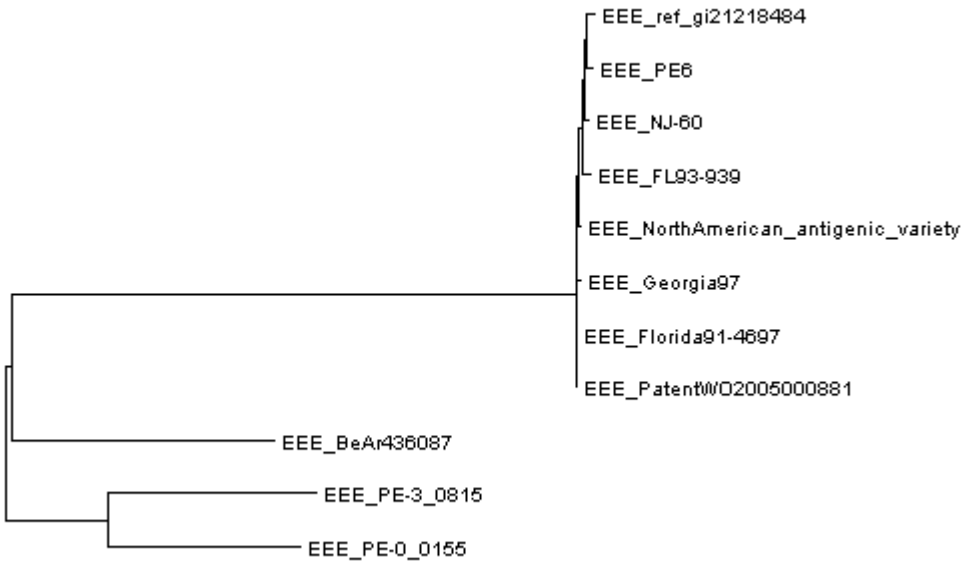

Figure 3A: MSA tree as in Figure 1A, for EEE.

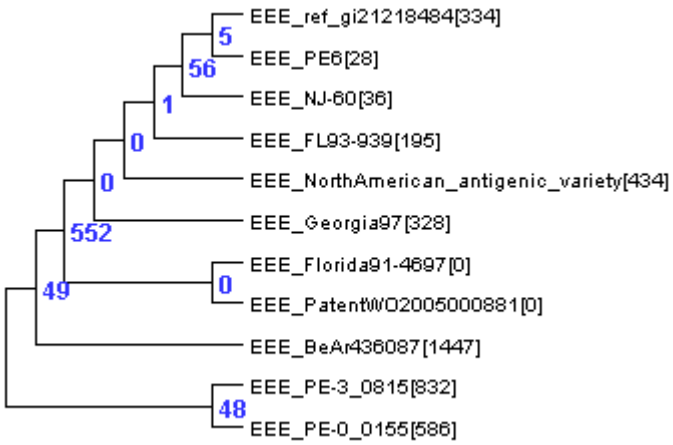

Figure 3B: PCR signature counts mapped to nodes, as in Figure 1B, for EEE.

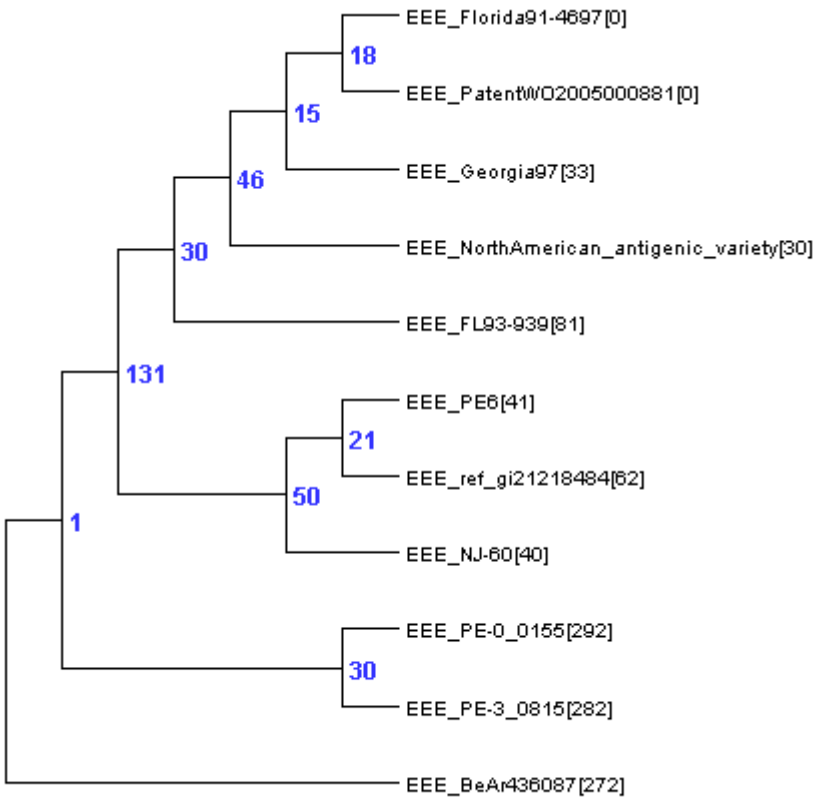

Figure 3C: SNP tree for EEE, as in Figure 1C.

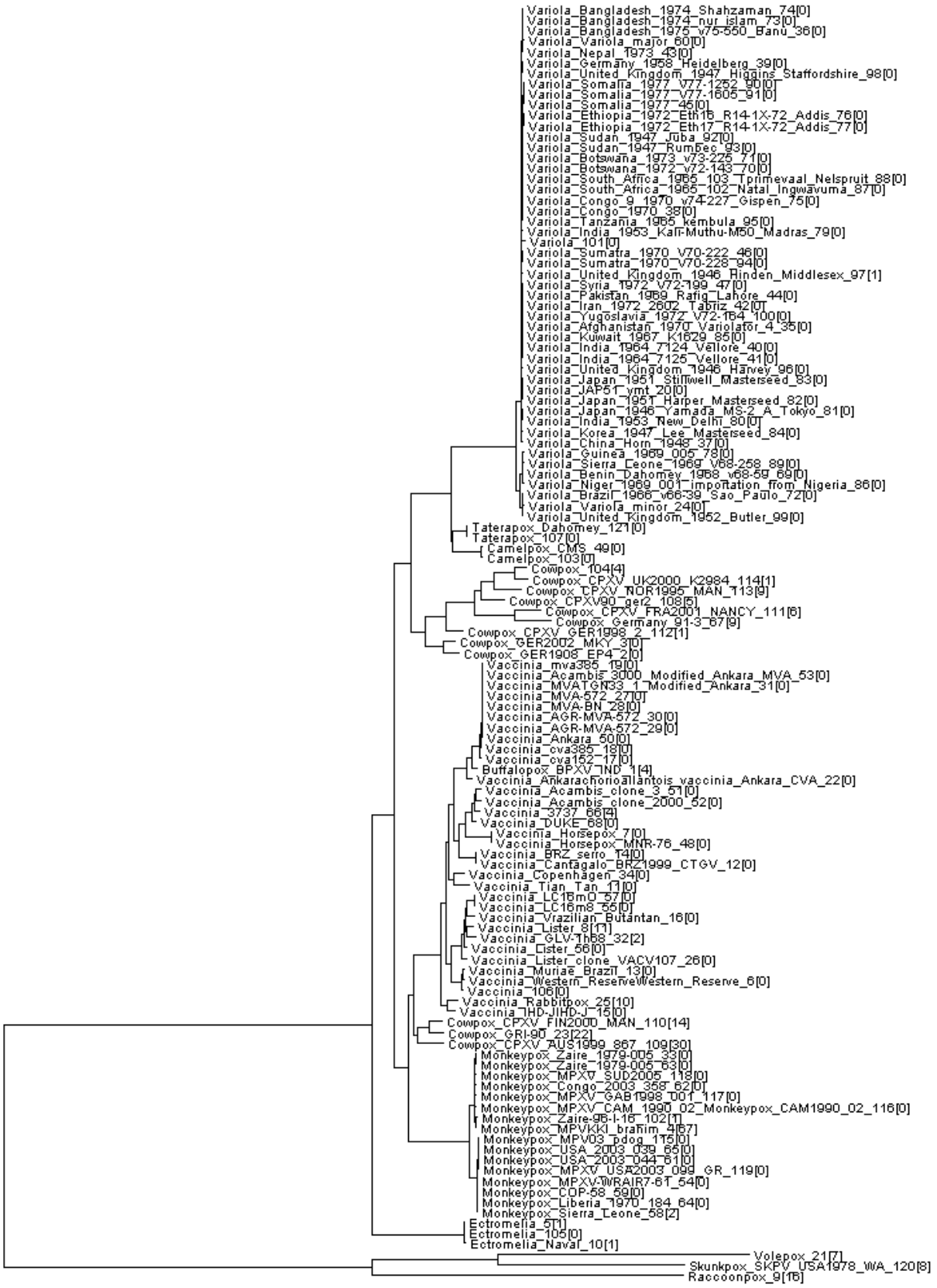

Figure 4A: MSA based tree for OPXV, as in Figure 1A.

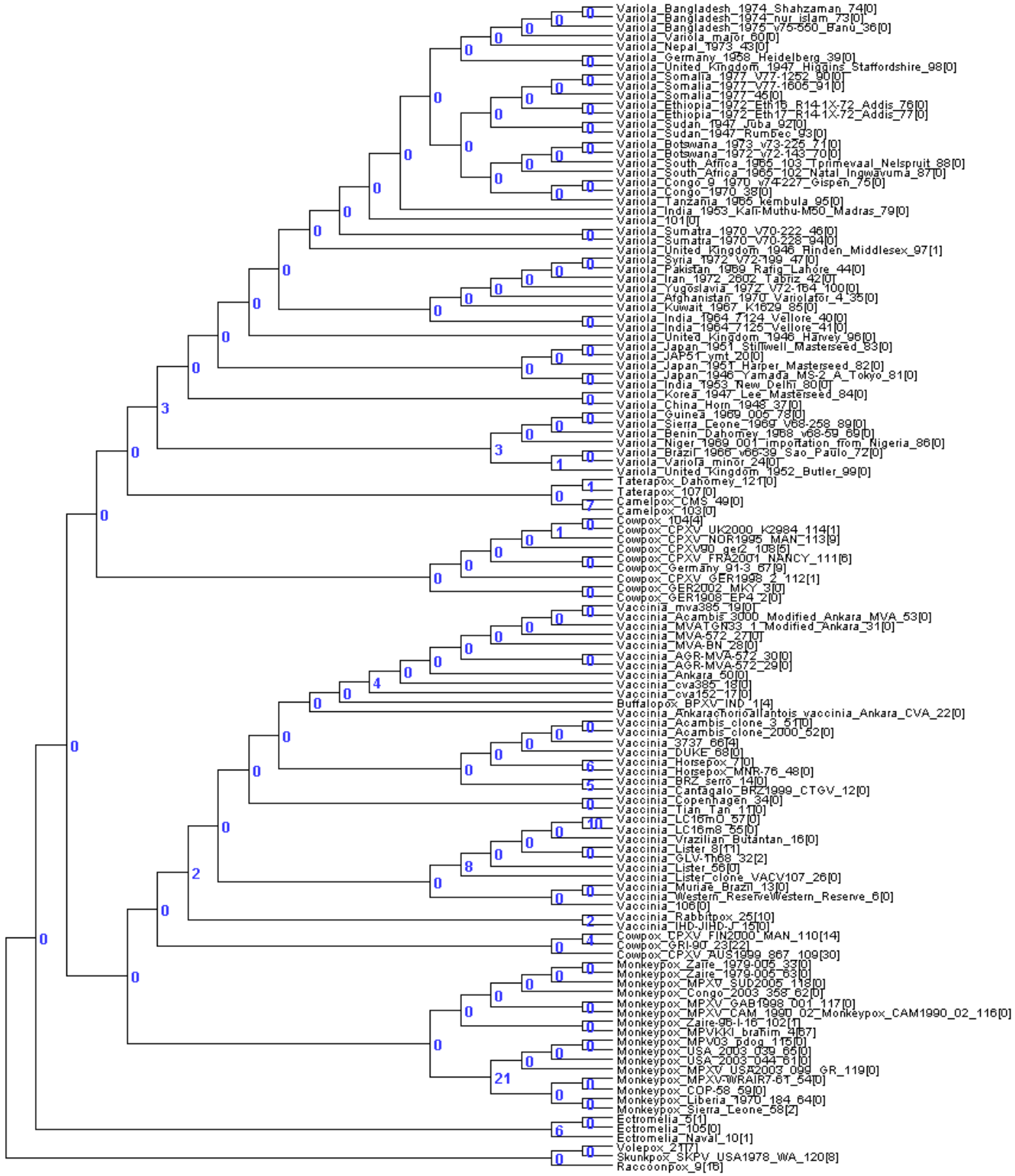

Figure 4B: PCR signature counts mapped to nodes, as in Figure 1B, for OXPV.

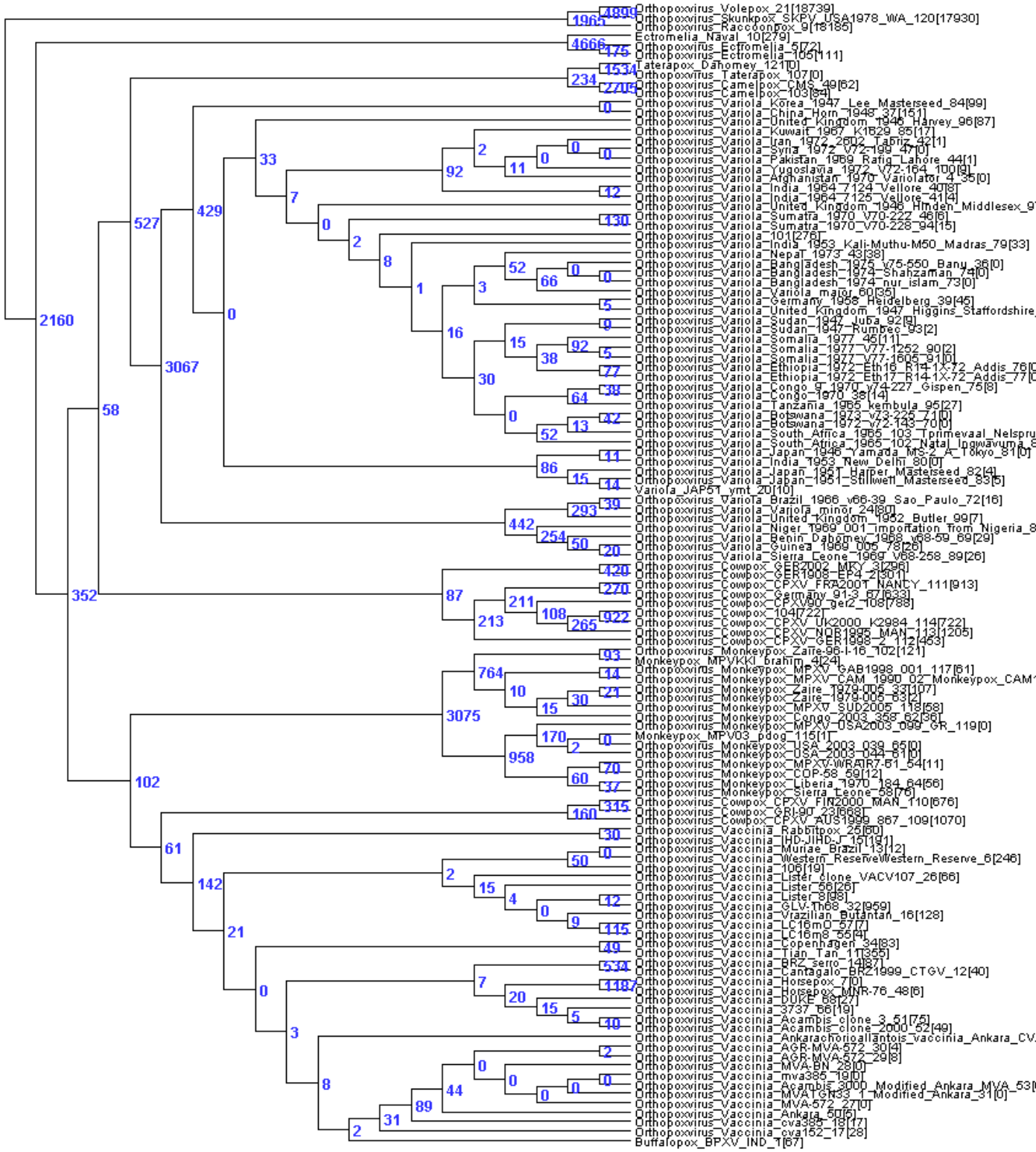

Figure 4C: MSA-based tree for Orthopox with SNP allele counts mapped onto tree.

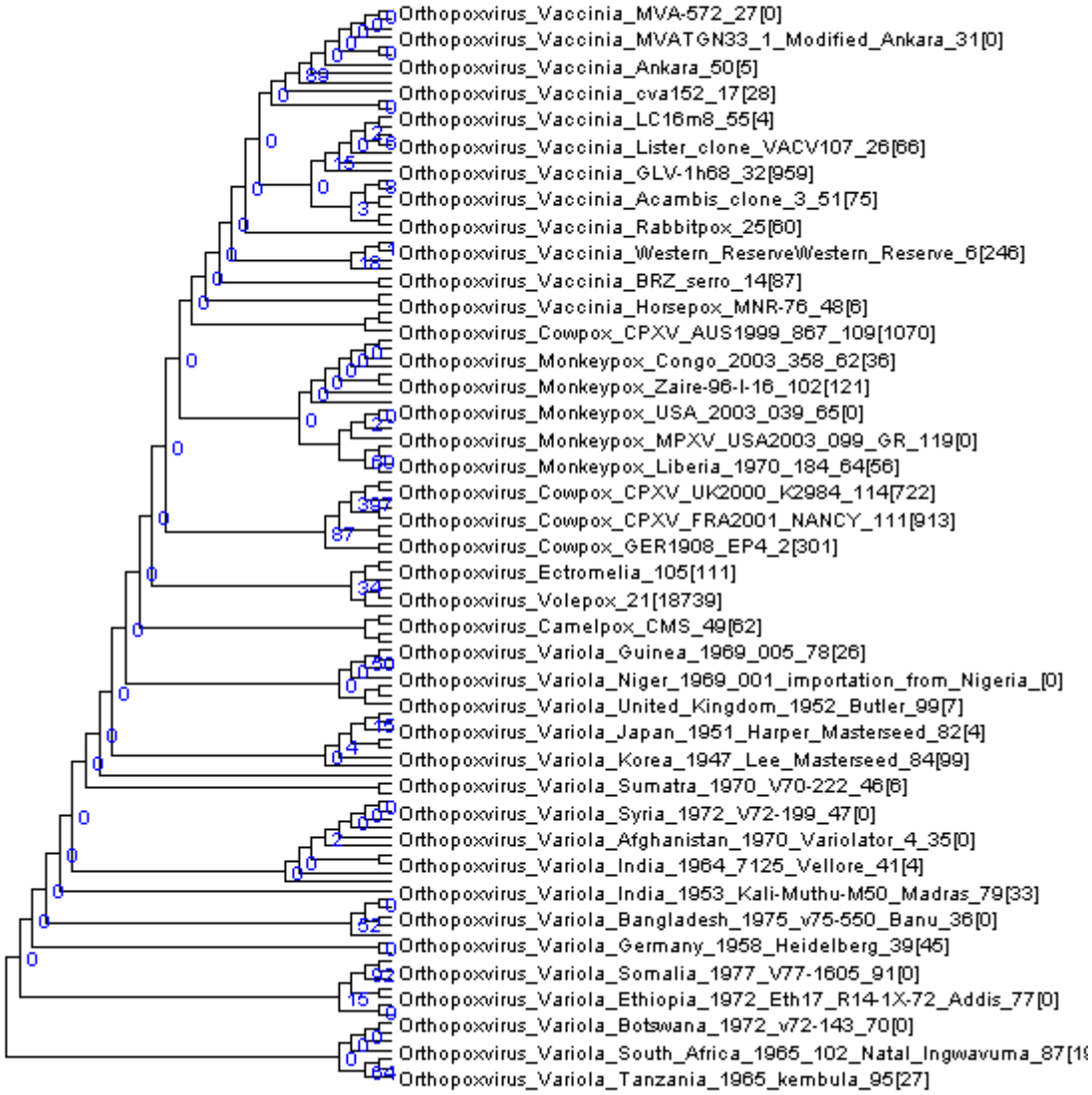

Figure 4D: SNP-based tree for OPXV, with SNPs mapped onto tree, as in Figure 1C. Not all strains or node SNP counts are shown, but most are zero, since the SNP-based tree is a poor estimate of phylogeny compared with the MSA-based tree in Figure 4C.

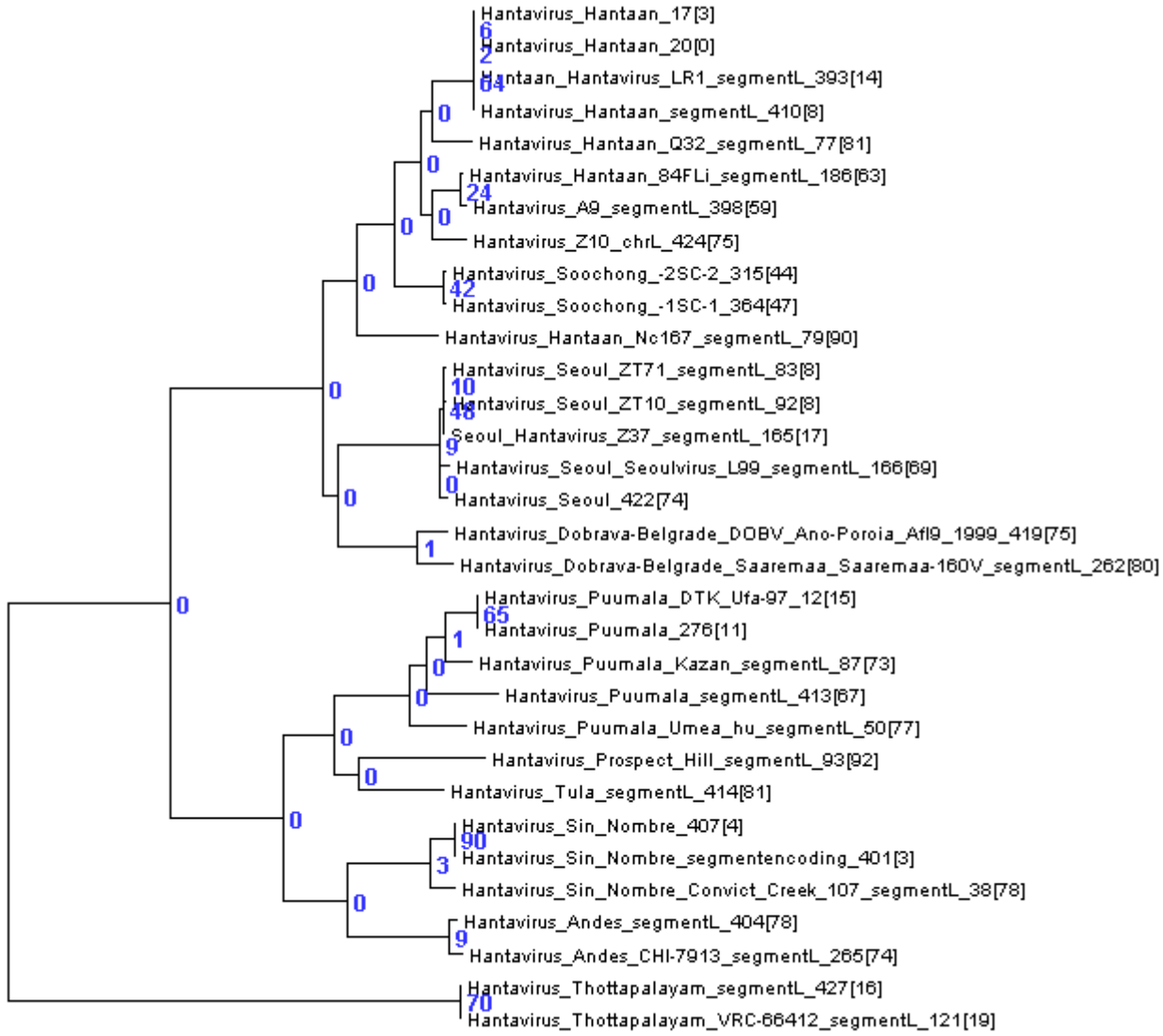

Figure 5A: MSA based tree for Hanta segment L with accurate branch lengths, as in Figure 1A, and also showing PCR signature counts at the nodes and in brackets after strain names, as in Figure 1B.

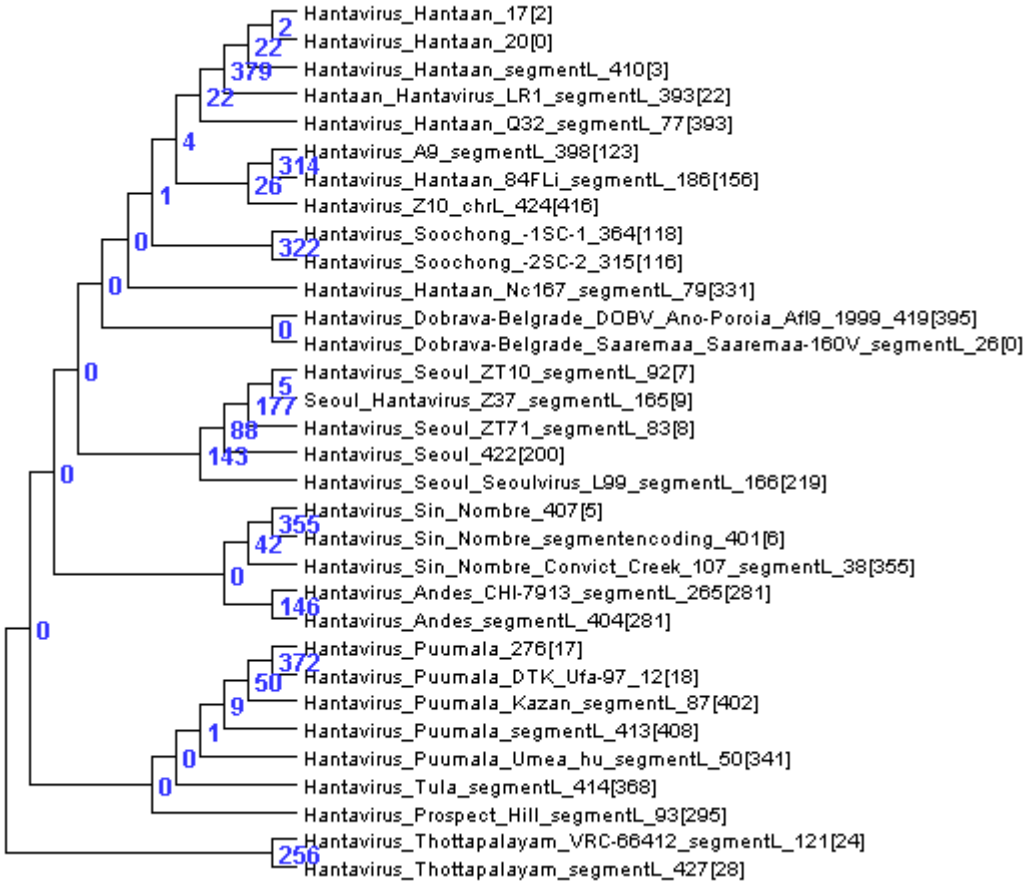

Figure 5B: SNP-based tree for Hantavirus segment L, as in Figure 1C.

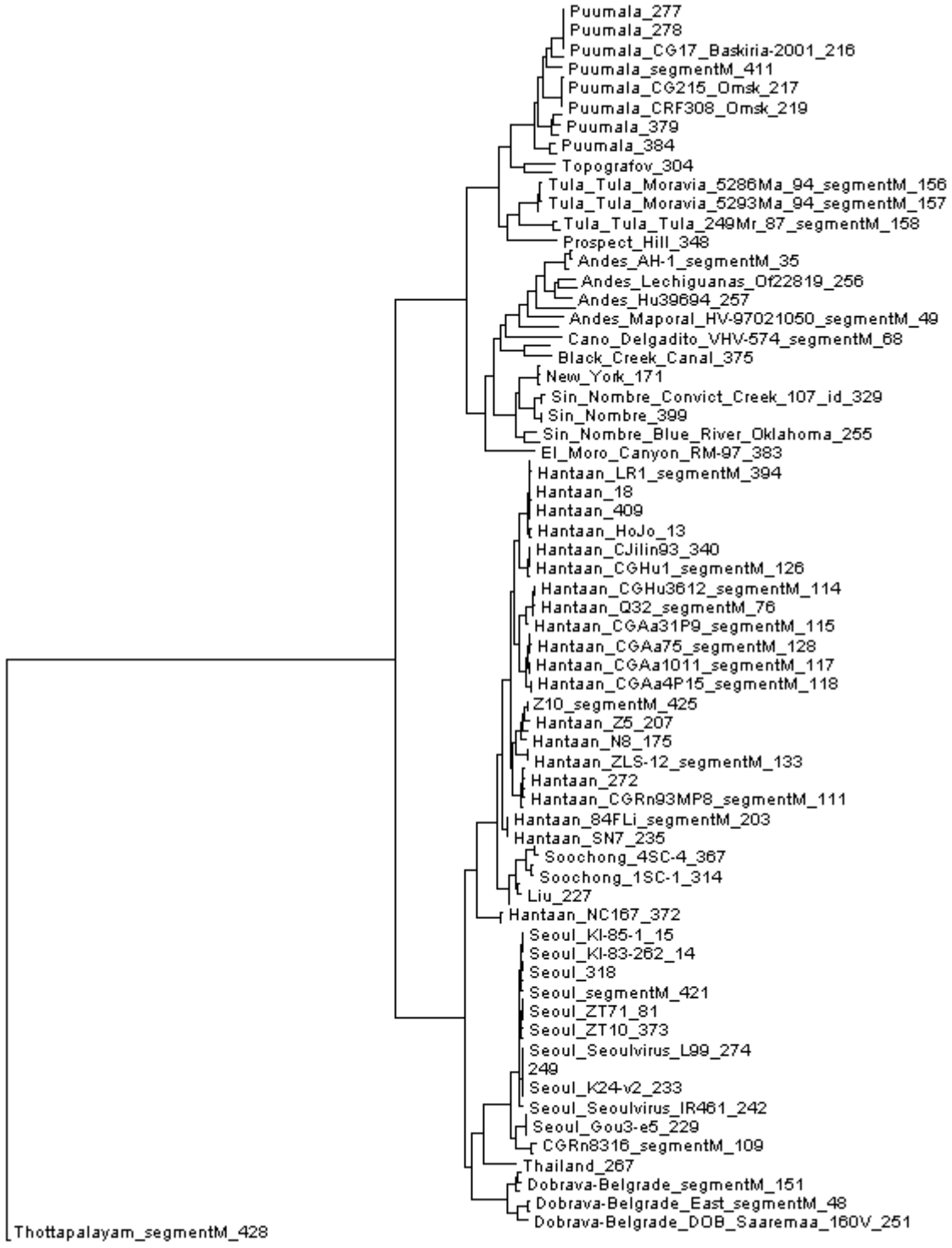

Figure 6A: MSA tree for Hanta segment M. Some strain names are not shown, for better legibility.

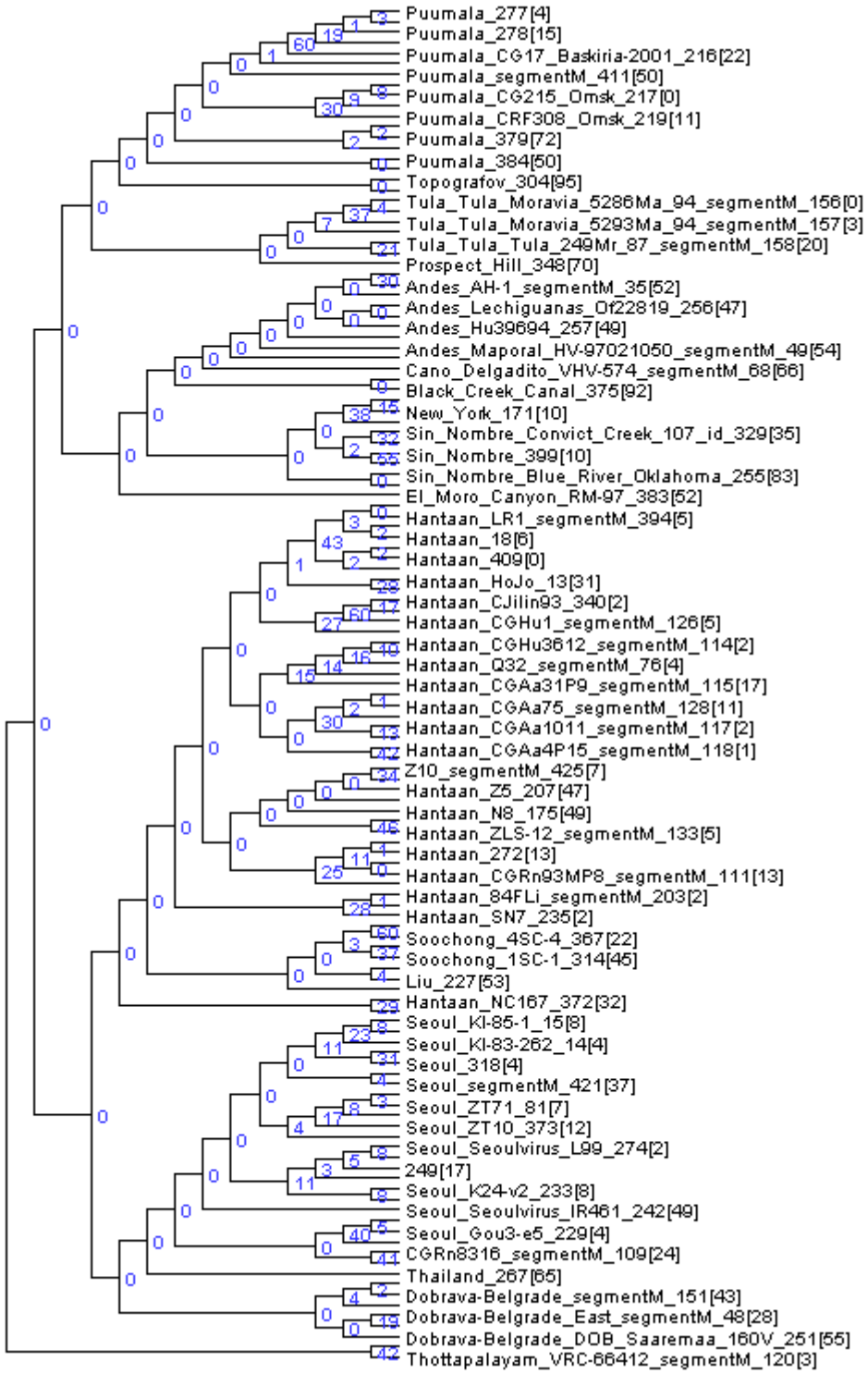

Figure 6B: MSA-based tree for Hanta segment M with PCR signature counts mapped onto tree. Not all strains are shown.

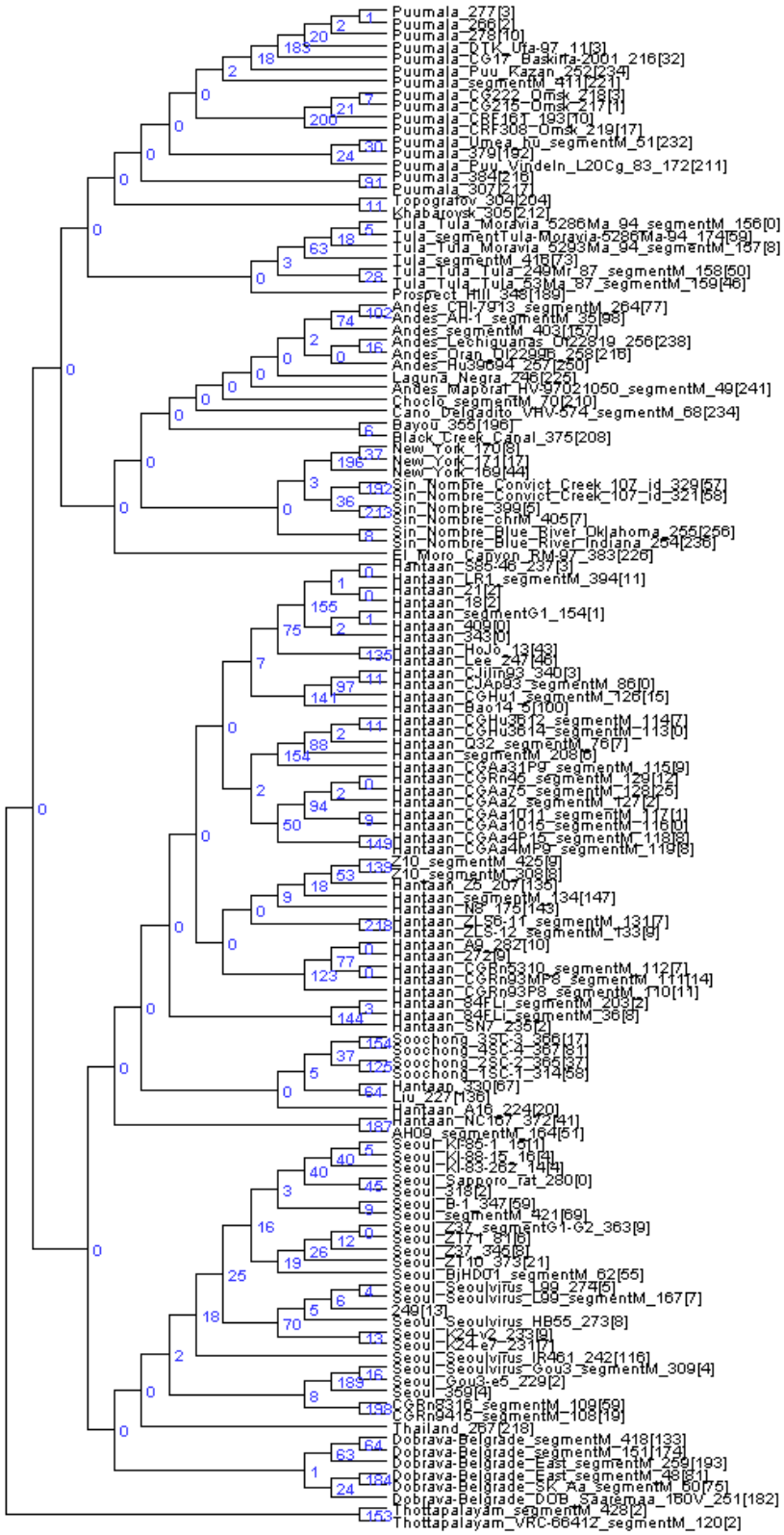

Figure 6B: MSA-based tree for Hanta segment M with SNP allele counts mapped onto tree. SNPs were insufficient for determining branching relationships at the higher (inter-species) levels, as can be seen by all the zeros on nodes at the left of the tree, so we show the SNP counts mapped onto the MSA-based tree rather than the SNP-based tree.

Figure 7A: MSA tree for Hanta segment S. Some strain names are not shown, for better legibility.

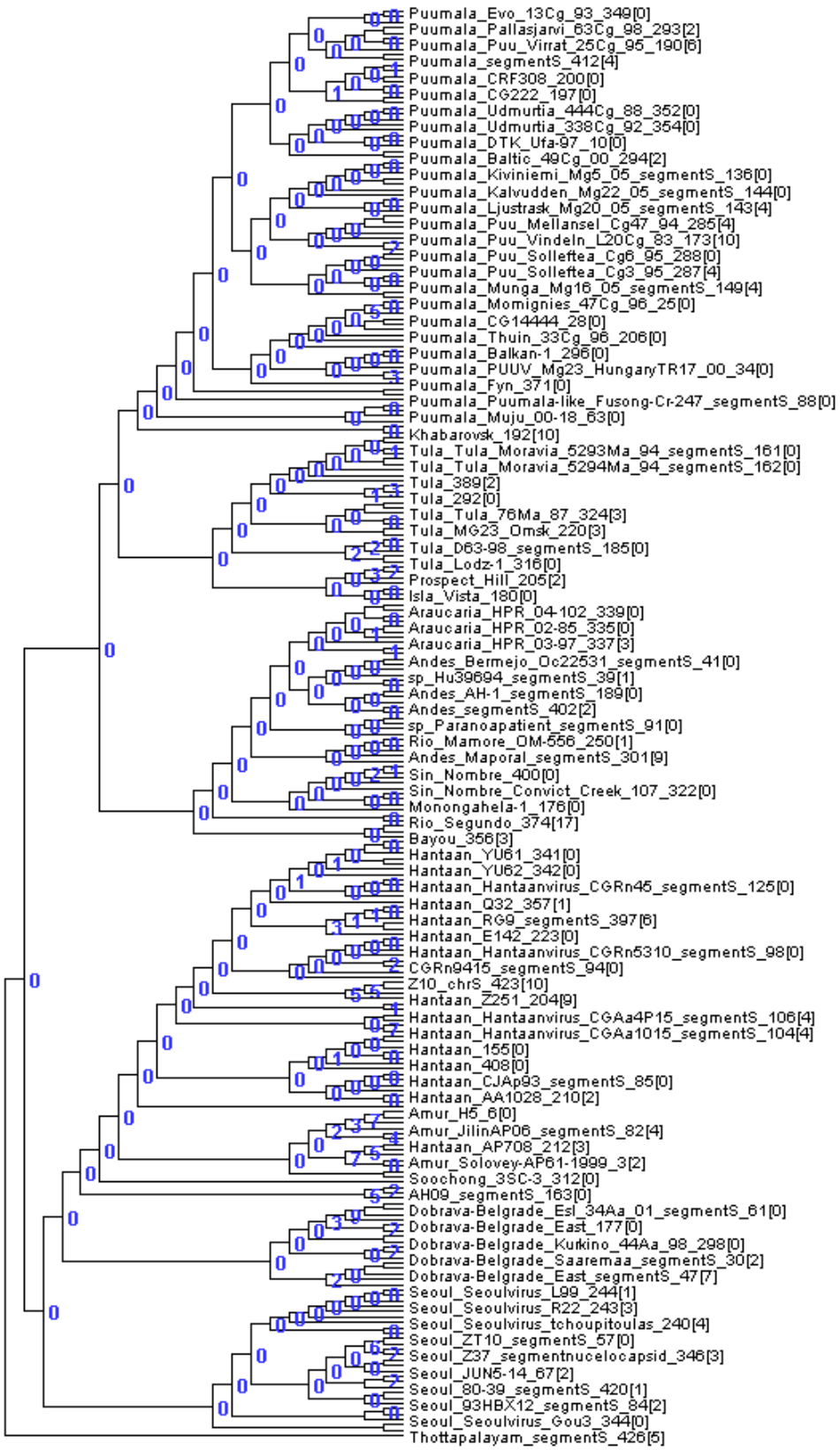

Figure 7B: PCR signatures mapped onto MSA tree for Hanta segment S.

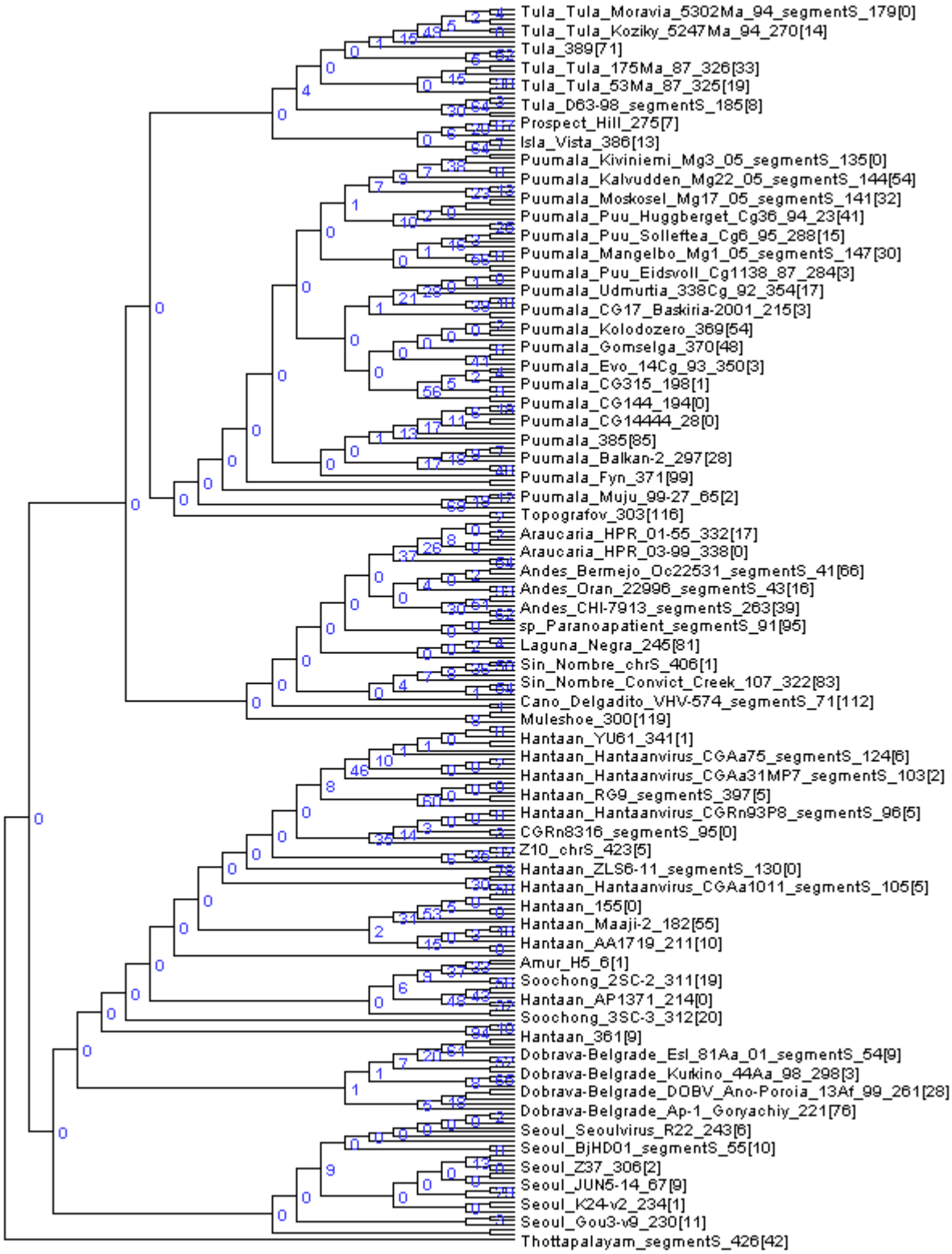

Figure 7C: SNP tree for Hantavirus segment S. Some sequences and nodes are not shown, for better legibility.

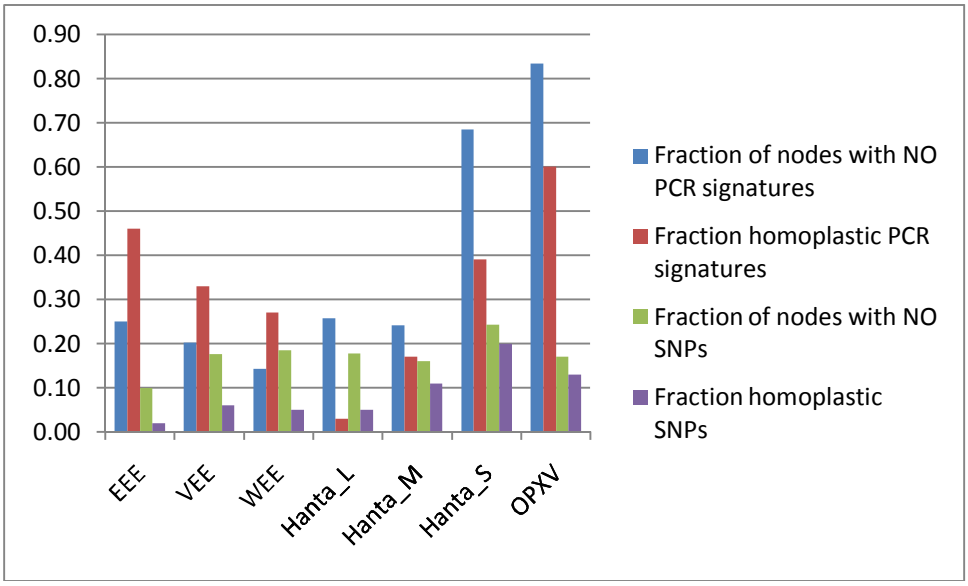

Figure 8: Chart showing that SNPs have better representation on tree nodes than do PCR signatures, since PCR signatures have a higher fraction of nodes with zero PCR signatures and more homplastic PCR signatures than do SNPs.