Supplementary material

RefSeq accession	Strain	Lineage	Comments
NZ_CP0009427.1	96121	1	
NC_021194.1	EAI5/NITR206	1	Reference-guided assembly: AL123456 .1
NC_021740.1	EAI5	1	Reference-guided assembly: NC_000962.2
NC_002755.2	CDC1551	4	
NC_020089.1	7199-99	4	
NC_020559.1	Erdman = ATCC 35801	4	
NC_022350.1	Haarlem	4	
NZ_CP010337.1	22115	4	
NZ_CP010338.1	37004	4	Removed from RefSeq, Reference-guided assembly: NC_017524
NZ_CP010339.1	22103	4	
NZ_AP014573.1	Kurono	4	
NC_018143.2	H37Rv	4	
NZ_CP007027.1	H37RvSiena	4	Reference-guided assembly: NC_000962.3
NZ_CP009480.1	H37Rv	4	Reference-guided assembly: NC_018143.2
NC_009525.1	H37Ra	4	
NC_000962.3	H37Rv	4	Reference genome
NZ_CP010330.1	F28	4	
NC_009565.1	F11	4	
NC_017524.1	CTRI-2	4	
NC_016768.1	KZN 4207	4	
NC_012943.1	KZN 1435	4	
NC_018078.1	KZN 605	4	
NZ_CP010340.1	26105	3	Removed from RefSeq
NZ_CP009100.1	ZMC13-264	2	Reference-guided assembly: NC_000962.3
NZ_CP009101.1	ZMC13-88	2	Reference-guided assembly: NC_000962.3
NZ_CP007803.1	K	2	
NZ_CP007809.1	KIT87190	2	
NZ_CP011510.1	Beijing	2	
NC_021054.1	Beijing/NITR203	2	Reference-guided assembly: AL123456.1
NZ_CP012090.1	W-148	2	
NZ_CP016888.1	SCAID 252.0	2	Reference-guided assembly: CP012506.1, NC_000962.1
NZ_CP016794.1	SCAID 320.0	2	Reference-guided assembly: CP012506.1, NC_000962.1
NZ_CP012506.2	SCAID 187.0	2	
NC_021251.1	CCDC5079	2	
NZ_HG813240.1	49-02	2	
NC_017522.1	CCDC5180	2	
NZ_CP002885.1	CCDC5180	2	
NZ_CP002882.1	BT2	2	
NZ_CP009426.1	96075	2	
NZ_CP002883.1	BT1	2	
NZ_CP002871.1	HKBS1	2	

Table S1: Description of dataset used in this study. Genomes span the four lineages as follows: Lineage 1 (East-Asian, 3 strains), Lineage 2 (Beijing, 18 strains), Lineage 3 (Central Asian, 1 strain) and Lineage 4 (Euro-American, 19 strains). The genomes are on average 4.407Mb (4.379Mb - 4.439Mb) long and contain on average 3,997 (3,869 - 4,069) proteins each. RefSeq accession number are shown with strain description and lineage classification. The comments show observations extracted from the GenBank description file.

Region ID	Alignment length (nt)	Variable sites	Parsimony informative sites	Genomic positions	Gene content	Recombined segments terminal	Recombined segments internal	HoT
31	32,195	197	87	672,740 - 704,243	30	6	0	98,60%
32	348,664	2,302	1,385	932,202 - 1,262,959	292	69	12	97.99%
33	301,251	1,373	641	1,262,960 - 1,546,124	239	32	0	97.06%
34	6,197	13	5	1,546,125 - 1,552,321	4	0	0	99.87%
35	1,164,225	13,573	10,675	1,992,287 - 2,796,180	722	259	107	50.41%
40	611,340	4,187	2,788	2,796,181 - 3,378,377	534	125	38	98.28%
44	255,921	2,060	1,684	3,480,373 - 3,711,736	214	80	55	99.14%
45	49,583	1,900	1,867	3,711,737 - 3,731,431	14	26	24	59.25%
46	108,774	4,354	746	3,750,543 - 3,842,343	76	54	4	98.19%
53	97,243	661	435	3,846,810 - 3,931,936	75	30	7	98.86%
83	339,288	1,480	685	1 - 334,774	283	34	8	99.61%
90	341,537	1,423	797	336,862 - 672,739	301	39	2	96.23%
91	172,754	1,336	968	704,244 - 873,312	171	31	1	98.07%
92	22,429	563	552	873,313 - 895,563	19	13	0	99.41%
93	39,322	901	665	895,564 - 932,201	32	14	0	98.08%
94	489,247	4,501	2,642	1,552,322 - 1,990,641	380	136	36	93.95%
95	102,535	283	97	3,380,709 - 3,480,372	92	6	0	99.21%
109	469,196	1,927	792	3,950,754 - 4,411,531	408	39	5	99.13%

Table S2: Description of the 18 universal genomic regions identified by progressiveMauve and realigned with MAFFT. Region numbering is determined by the first strain in the alignment (here: NC_000962.3). Variable sites represent the number of positions in the alignment with at least two nucleotides. Parsimony informative sites represent the number of positions in the alignment where at least 2 different nucleotides are each found in at least two different strains. Genomic positions are delimitations of the specific region in NC_000962.3. Gene content is the number of genes found for the specific region in NC_000962.3. Recombined segments are number of segments found by ClonalFrameML, in the specific region, on internal or terminal branches. HoT is the score computed by the HoT procedure and is the percentage of reliable positions in the alignment.

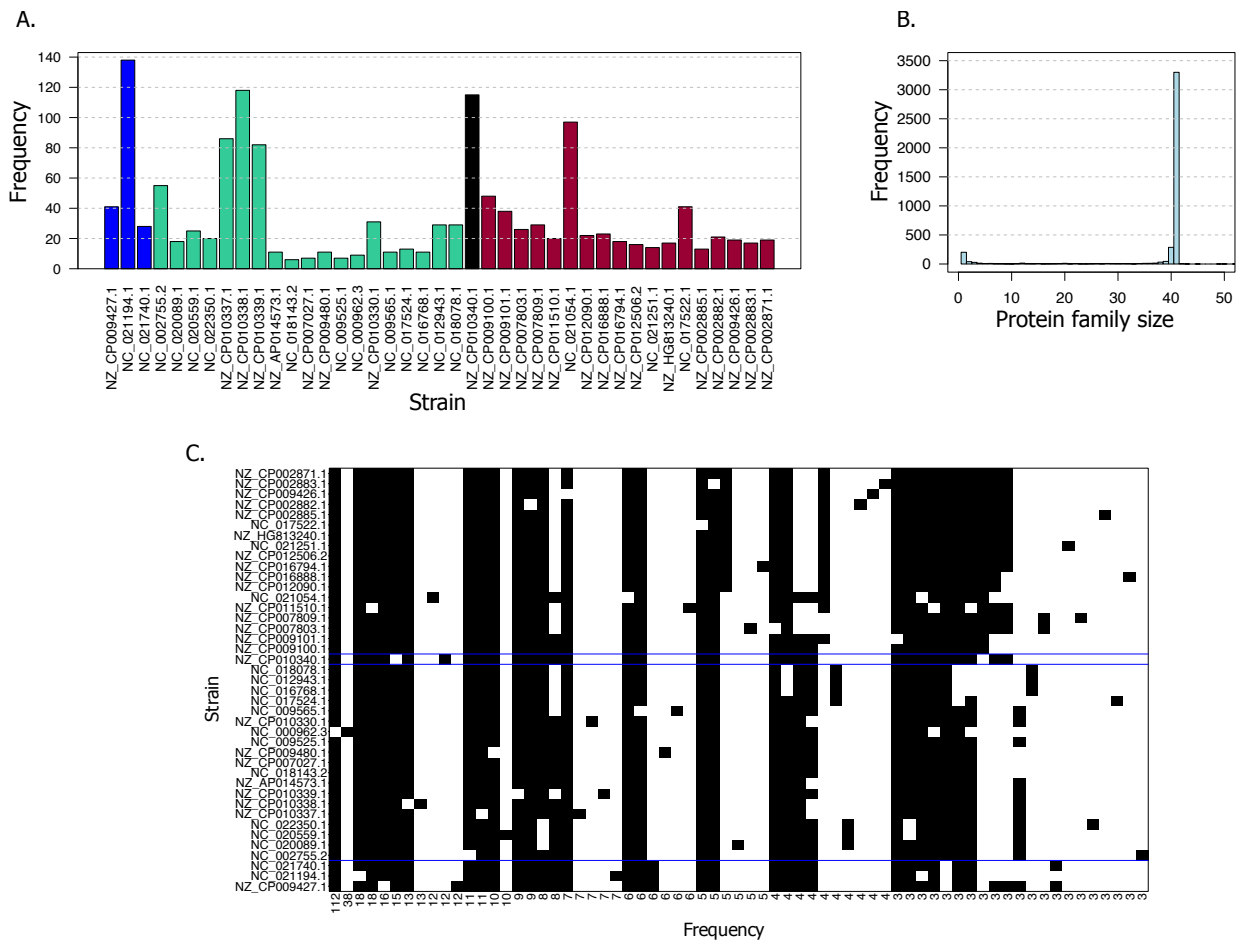

Figure S1: (A) Number of pseudogenes that extend protein families per strain. Colors denote lineages as follows: blue for lineage 1, red for lineage 2, black for lineage 3 and green for lineage 4.. (B) Distribution of extended protein family sizes. The total number of homologous families is 4,272 of which 3,382 (79%) are complete families (i.e., found at least once in each genome), 3,293 (77%) are complete single-copy families (i.e. having exactly one representation in each genomes), and the remaining 890 are partial (i.e., absent in at least one genome). Not shown are 96 families of size more than 50. (C) Presence-absence matrix of 559 protein families grouped in 67 different patterns, ordered by decreasing frequencies. Black indicates presence in the genome, while white indicates absence. The blue lines differentiate the strains according to their lineages, from top to bottom : lineage 2, 3, 4 and 1. Not shown are 331 protein families grouped in 293 different patterns represented less than 3 times.

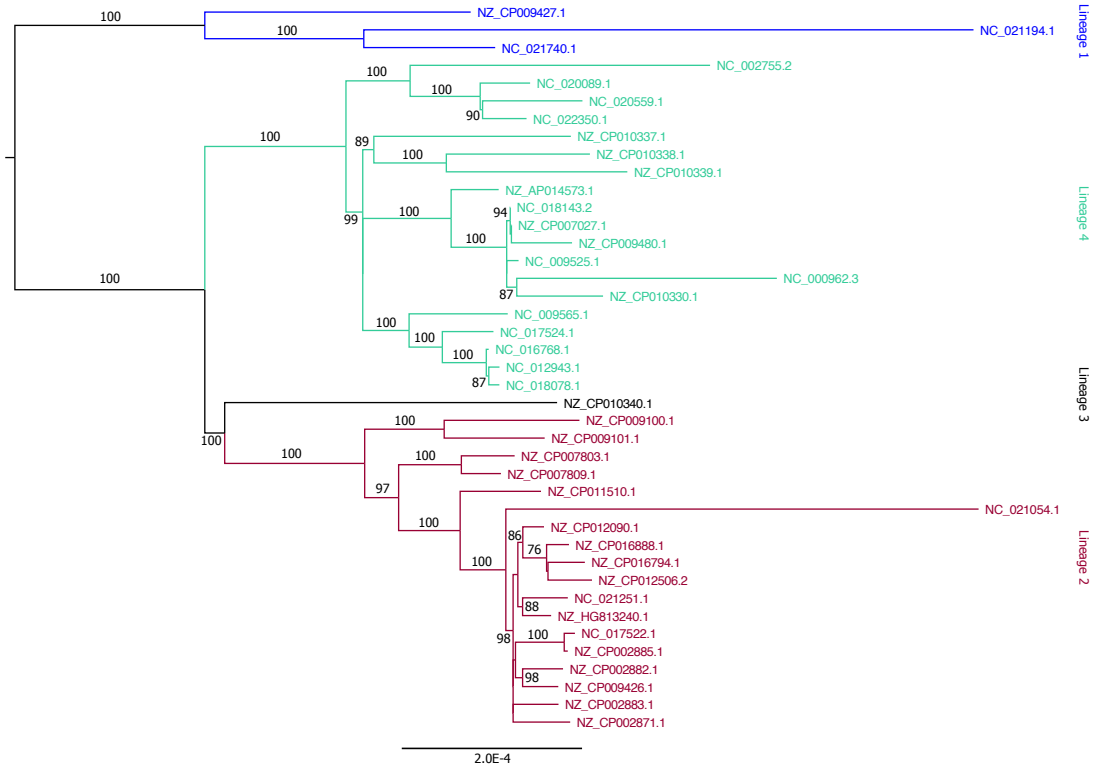

Figure S2: Reference pylogeny. Phylogenetic tree inferred from the amino acid alignment of the 2,650 complete single-copy protein families. Bootstrap values based on 100 replicates are shown for values of at least 75.

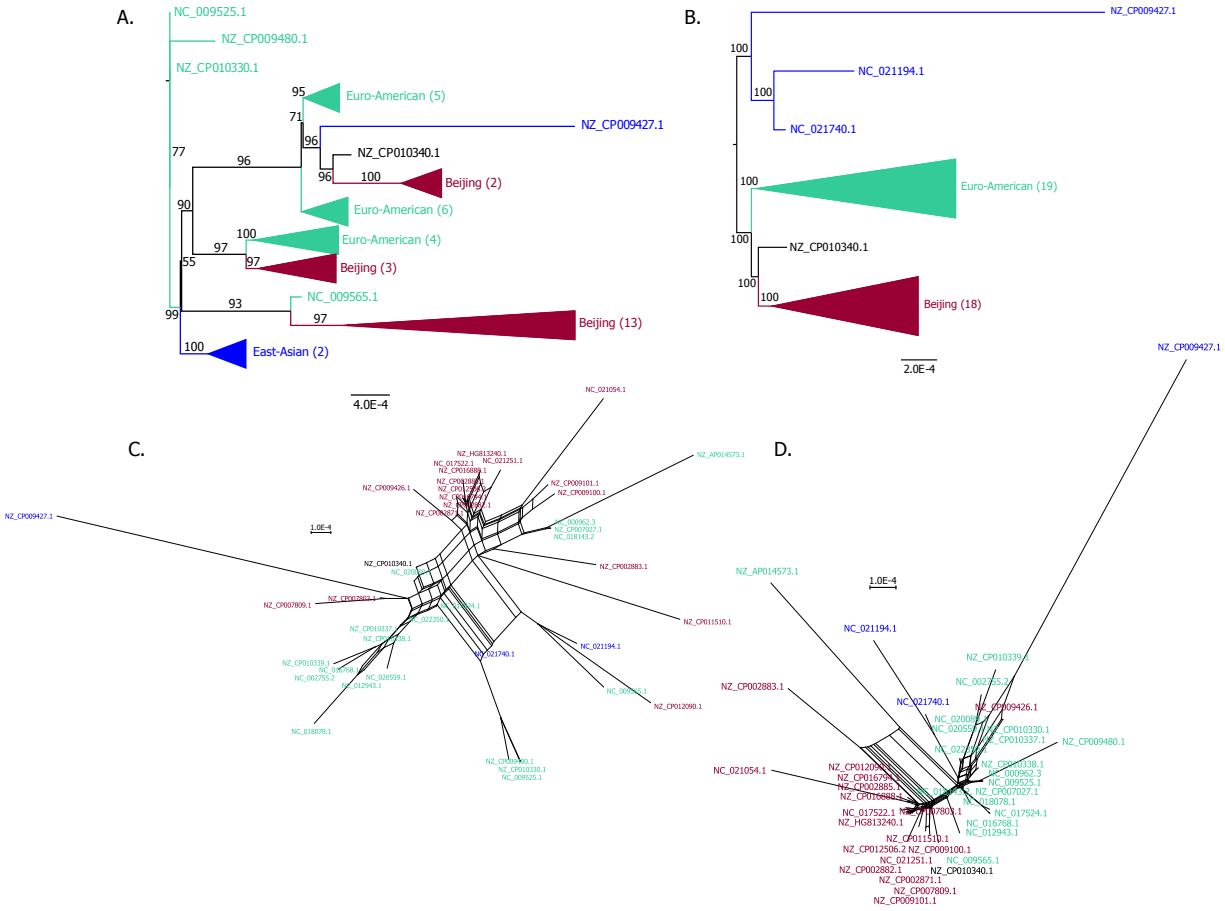

Figure S3: Phylogeny of universal genomic region 35. (A) Phylogeny estimated from complete alignment (1,164,225nt), variable sites: 13,573 (1.2%), parsimony informative sites: 10,675 (0.92%), HoT score: 50.41%. (B) Phylogeny estimated from alignment without unreliable columns identified by the HoT procedure (586,857nt). variable sites: 4,144 (0.71%), parsimony informative sites: 2,506 (0.43%).(C) Splits network of complete alignment. (D) Splits network of alignment after removal of unreliable columns identified by the HoT procedure. The HoT procedure efficiently identified and eliminated the incongruent signal.

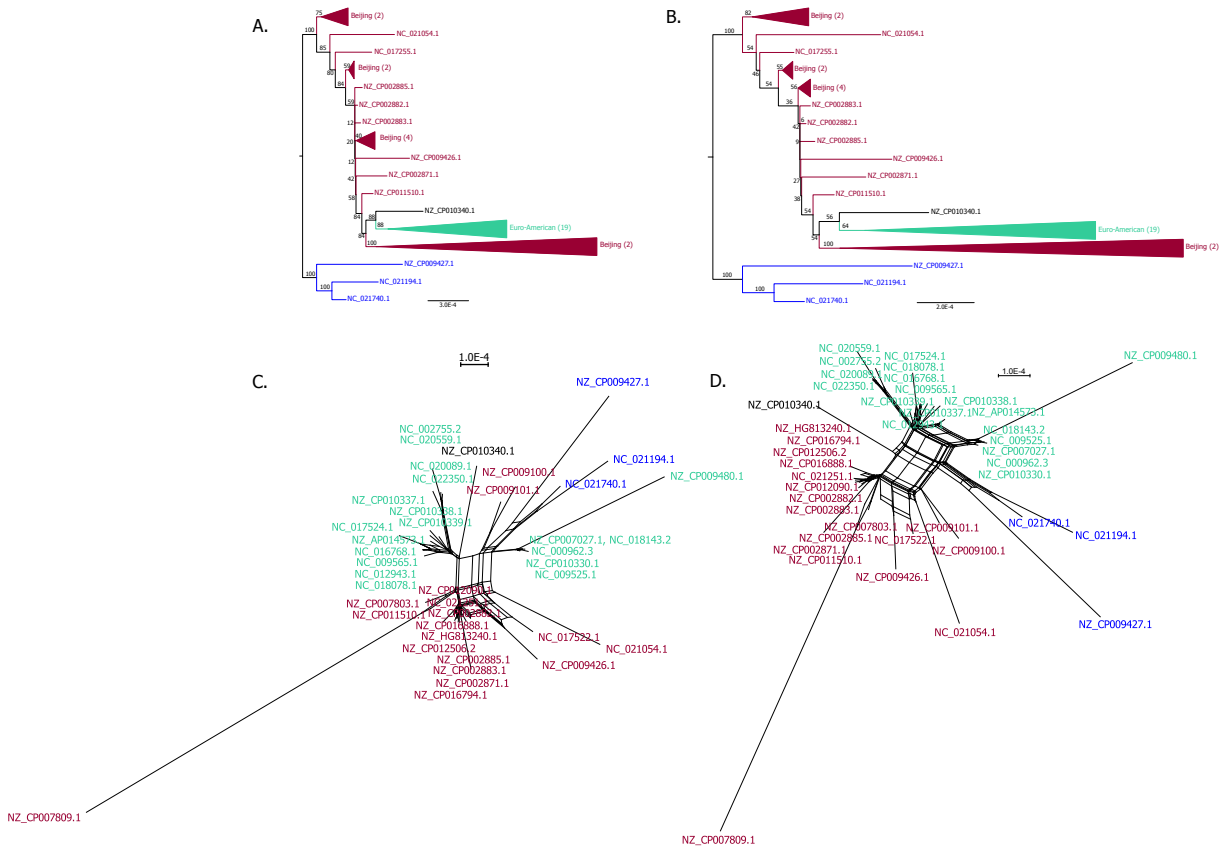

Figure S4: Phylogeny of universal genomic region 32. (A) Phylogeny estimated from complete alignment (348,664nt), variable sites: 2,302 (0.66%), parsimony informative sites: 1,385 (0.4%), HoT score: 97.99%. (B) Phylogeny estimated from alignment without unreliable columns identified by the HoT procedure (341,675nt), variable sites: 1,975 (0.58%), parsimony informative sites: 1,106 (0.32%). There are no branches with bootstrap support of at least 75 that conflict the monophyly of each established lineage. (C) Splits network of complete alignment. (D) Splits network of alignment after removal of unreliable columns identified by the HoT procedure.

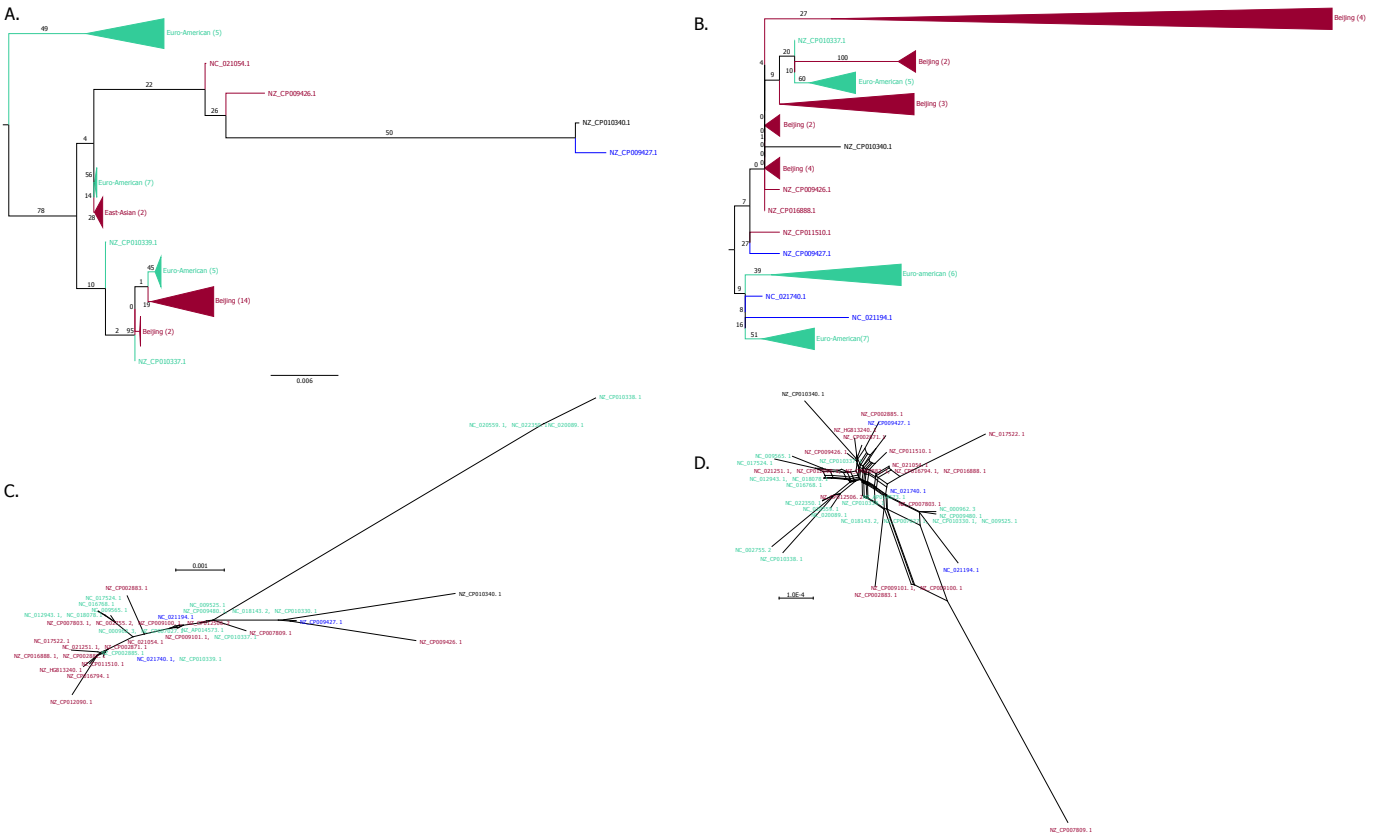

Figure S5: Phylogeny of universal genomic region 45. (A) Phylogeny estimated from complete alignment (49,583nt), variable sites: 1900 (3.83%), parsimony informative sites: 1867 (3.76%), HoT score: 59.25%. (B) Phylogeny estimated from alignment without unreliable columns identified by the HoT procedure (29,377nt), variable sites: 107 (0.36%), parsimony informative sites: 78 (0.26%).(C) Splits network of complete alignment. (D) Splits network of alignment after removal of unreliable columns identified by the HoT procedure. The HoT procedure removed the incongruent signal as no branches with bootstrap support of at least 75 conflict the monophyly of each established lineage.

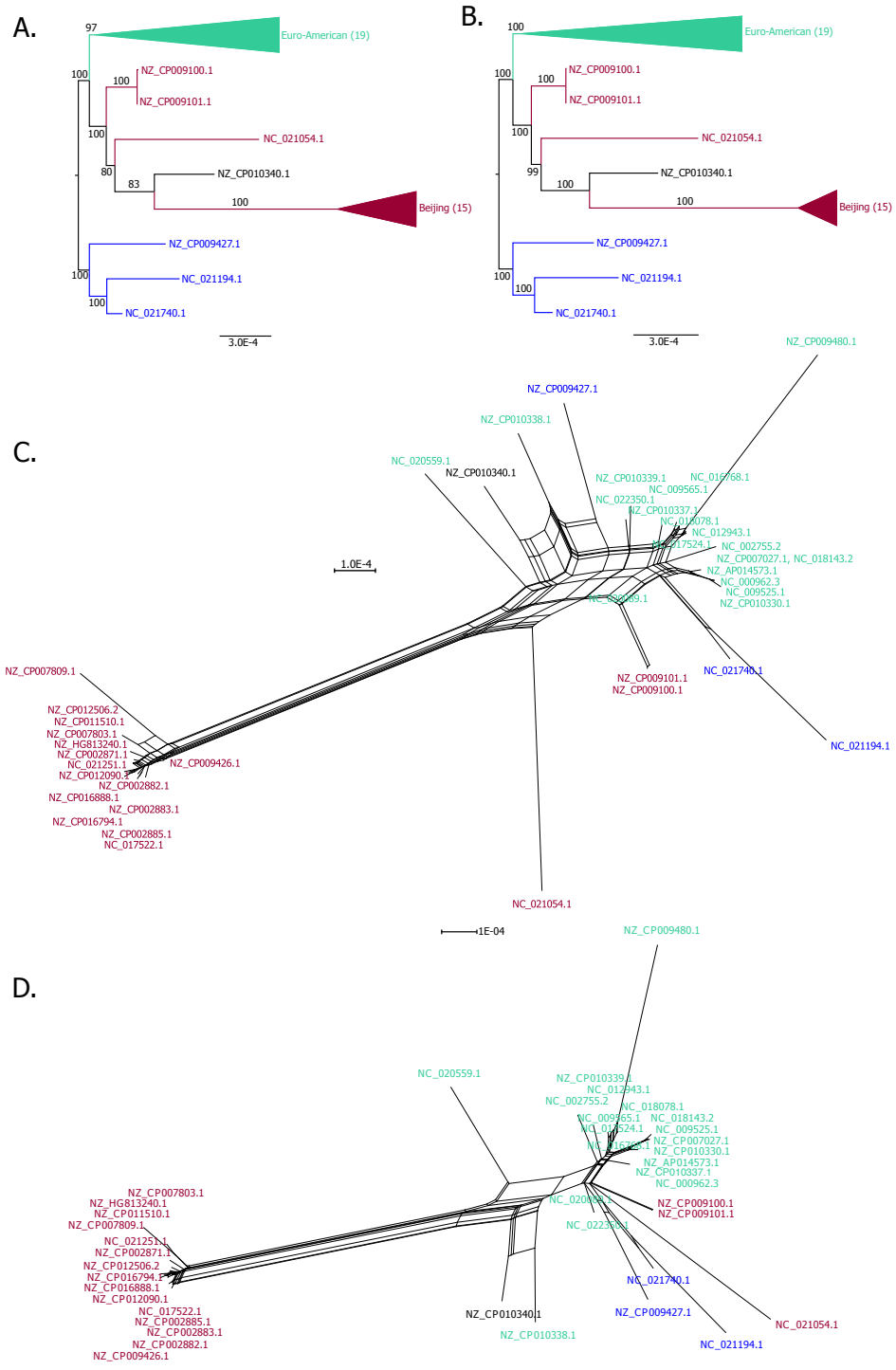

Figure S6: Phylogeny of universal genomic region 40. (A) Phylogeny estimated from complete alignment (611,340nt), variable sites: 4,187 (0.68%), parsimony informative sites: 2,788 (0.46%), HoT score: 98.28%. (B) Phylogeny estimated from alignment without unreliable columns identified by the HoT procedure (600,805nt), variable sites: 3,720 (0.62%), parsimony informative sites: 2,439 (0.41%). (C) Splits network of complete alignment. (D) Splits network of alignment after removal of unreliable columns identified by the HoT procedure. Three lineage 2 strains (NZ_CP009100.1, NZ_CP009101.1 and NC_021054.1), that are reference-guided assemblies, are misplaced and show a strong bias towards the reference (NC_000962.3). This conflict remains after the HoT procedure.

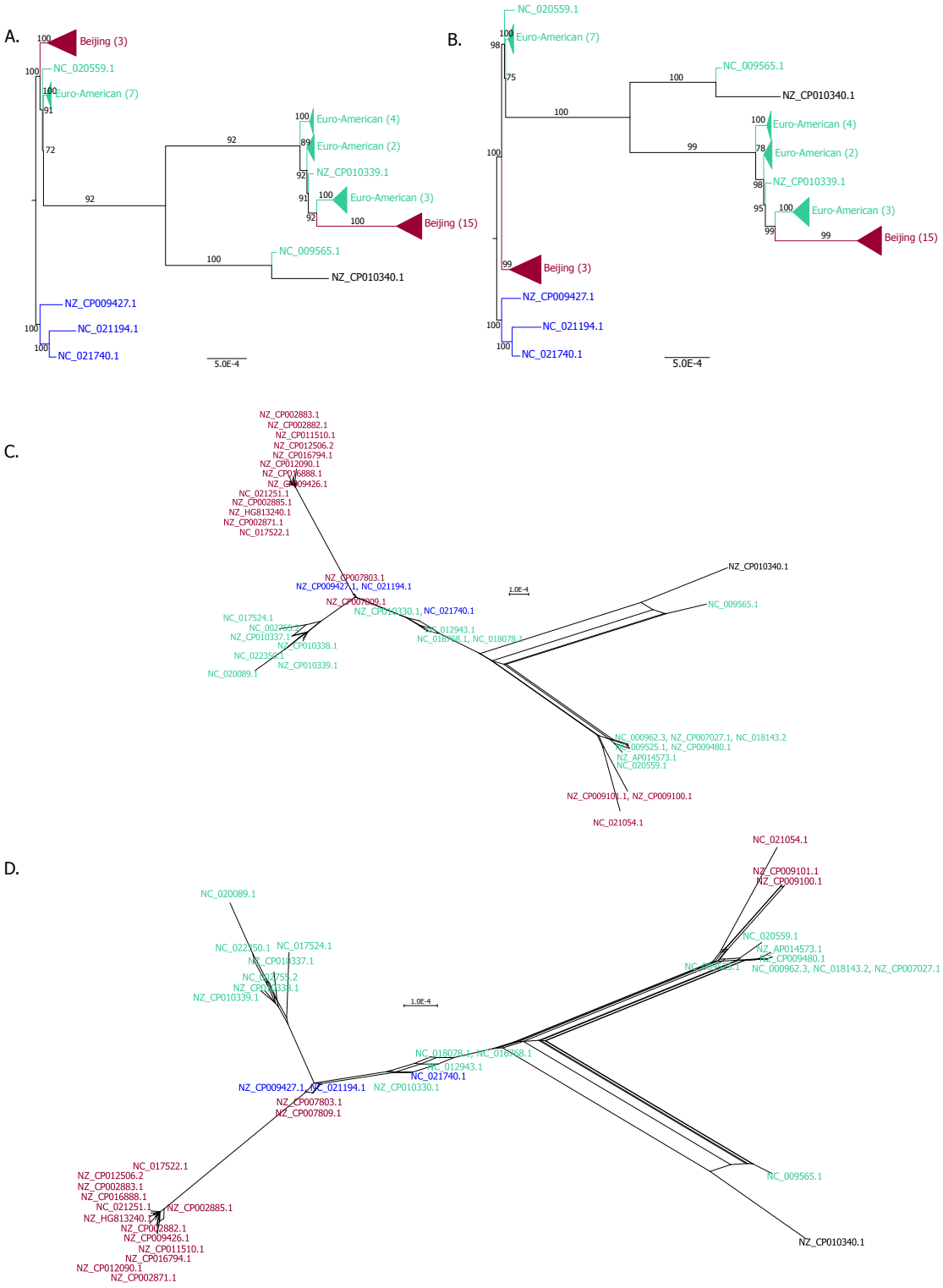

Figure S7: Phylogeny of universal genomic region 44. (A) Phylogeny estimated from complete alignment (255,921nt), variable sites: 2,060 (0.80%), parsimony informative sites: 1,684 (0.66%), HoT score: 99.14%. (B) Phylogeny estimated from alignment without unreliable columns identified by the HoT procedure (253,734nt), variable sites: 1,803 (0.71%), parsimony informative sites: 1,450 (0.57%).(C) Splits network of complete alignment. (D) Splits network of alignment without unreliable columns identified by the HoT procedure. Three misplaced lineage 2 strains (NZ_CP009100.1, NZ_CP009101.1 and NC_021054.1) are reference-guided assemblies and show a strong bias towards the reference used (NC_000962.3), strain NZ_CP010340.1 was removed from RefSeq.

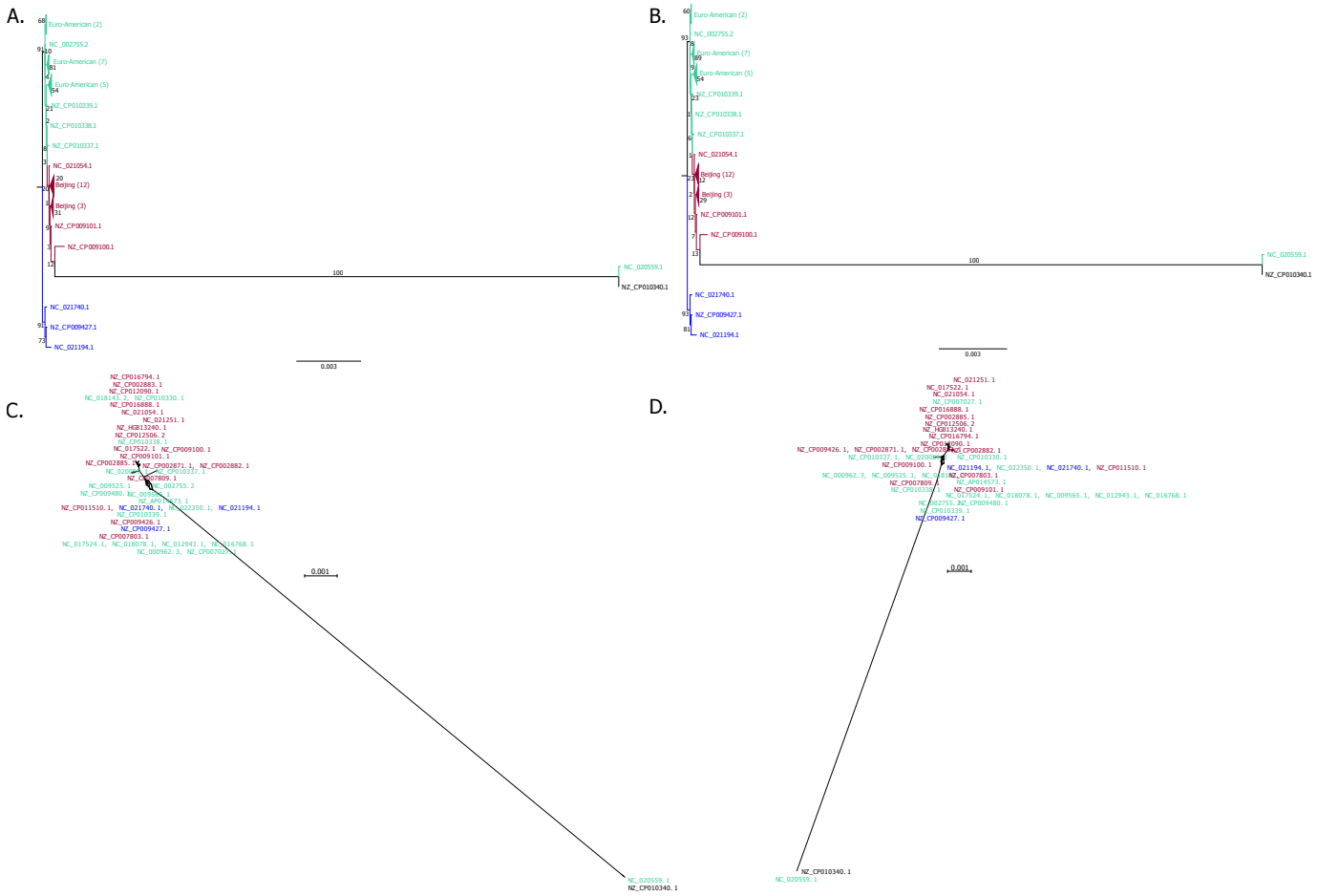

Figure S8: Phylogeny of universal genomic region 92. (A) Phylogeny estimated from complete alignment (22,429nt), variable sites: 563 (2.51%), parsimony informative sites: 552 (2.46%), HoT score: 99.41%. (B) Phylogeny estimated from alignment without unreliable columns identified by the HoT procedure (22,296nt), variable sites: 531 (2.38%), parsimony informative sites: 520 (2.33%).(C) Splits network of complete alignment. (D) Splits network of alignment after removal of unreliable columns identified by the HoT procedure. The misplaced strain NZ_CP010340.1, removed from RefSeq, remain after HoT procedure.

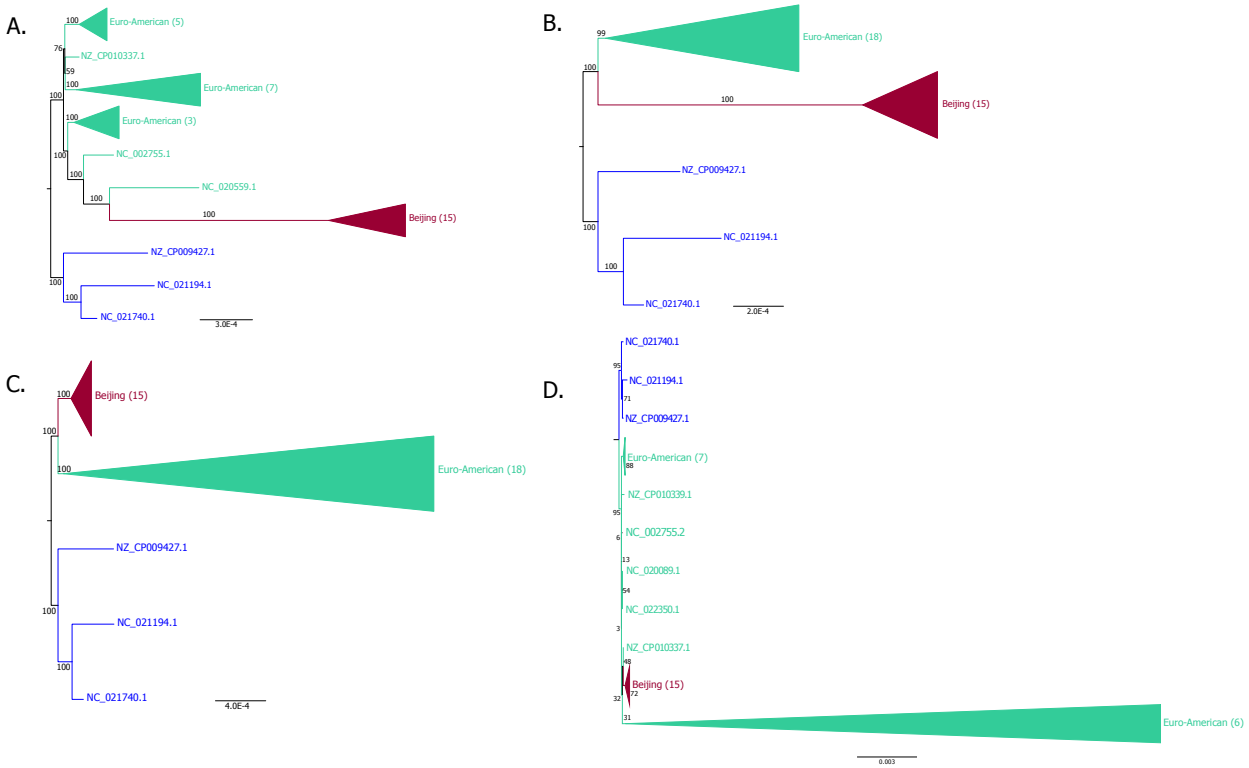

Figure S9: Phylogenies of regions 40, 44 and 92 without five strains that were misplaced in the splits network and presumbly show alignments of low quality (NZ_CP010338.1, NZ_CP010340.1, NZ_CP009100.1, NZ_CP009101.1 & NC_021054.1). (A) Phylogeny of universal genomic region 40 estimated from complete alignment (608,690nt), variable sites: 3,063 (0.5%), parsimony informative sites: 1,942 (0.32%), HoT score: 98.51%. (B) Phylogeny of universal genomic region 40 estimated from alignment without unreliable columns identified by the HoT procedure (599,593nt), variable sites: 2,603 (0.43%), parsimony informative sites: 1,583 (0.26%). The incongruent phylogenetic signal is efficiently eliminated by the HoT procedure. (C) Phylogeny of universal genomic region 44 estimated from complete alignment (255,921nt), variable sites: 2,060 (0.80%), parsimony informative sites: 1,684 (0.66%). The incongruent phylogenetic signal is efficiently eliminated when removing problematic assemblies. (D) Phylogeny of universal genomic region 92 estimated from complete alignment (22,411nt), variable sites: 562 (2.51%), parsimony informative sites: 553 (2.47%). The incongruent phylogenetic signal is efficiently eliminated when removing problematic assemblies as no branches with bootstrap support of at least 75 conflict the monophyly of each established lineage.