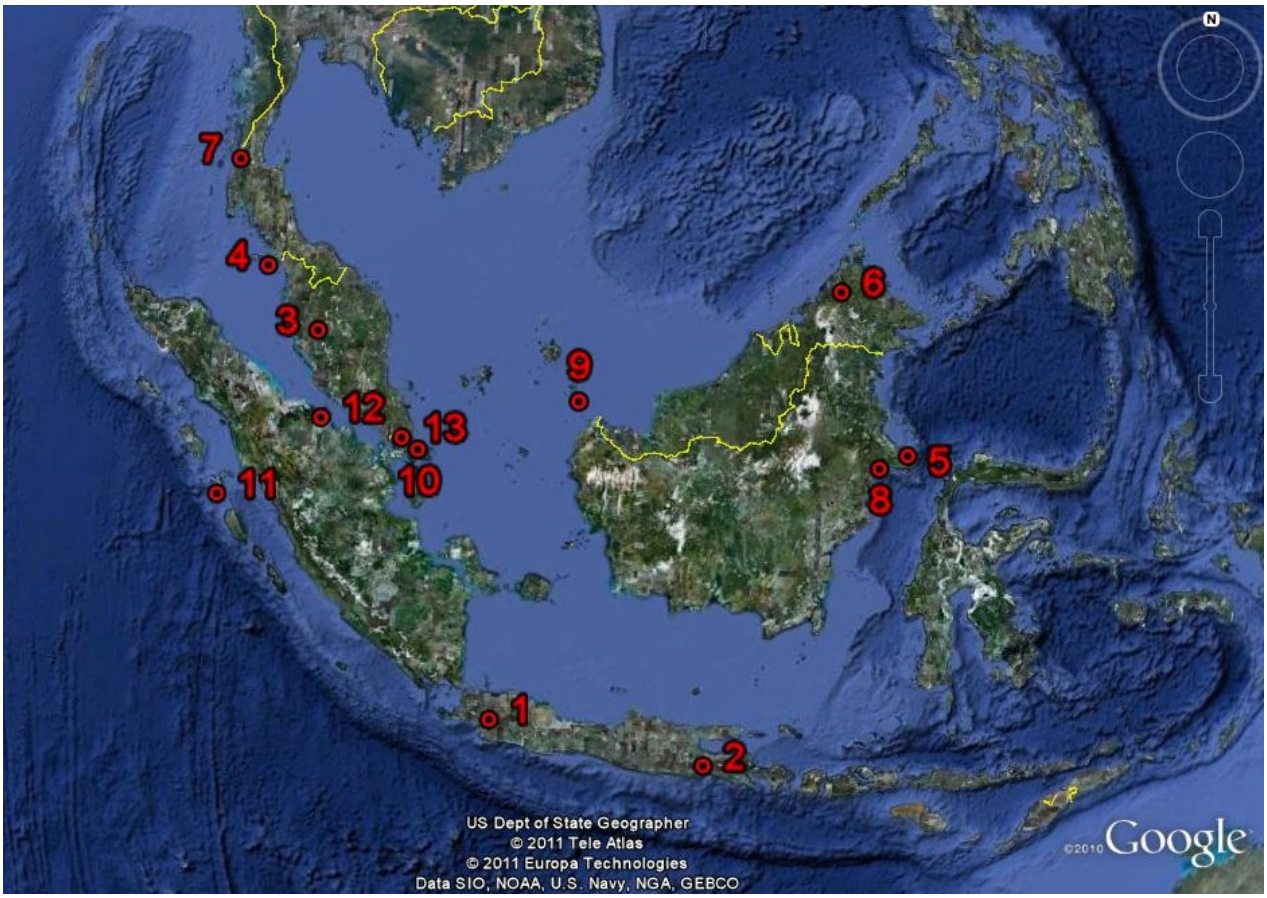
Supplementary Material for

Efficient Cross-Species Capture Hybridization and Next-Generation DNA Sequencing of Mitochondrial Genomes from Non-Invasively Sampled Museum Specimens.

Victor C. Mason, Gang Li, Kristofer M. Helgen, and William J. Murphy

Contents:

Figure S1. Geographic origin of USNM colugo specimens sampled.



Preliminary sequence library analysis

The 2˚ selected amplified libraries from specimens 12 (USNM 143327), and 6 (USNM 317119) were cloned and sequenced using Sanger sequencing to evaluate mtDNA content and enrichment, prior to processing the remaining samples for NextGen library production. The coverage and distributions of selected products are shown in Figure S2 (a&b), relative to a linear depiction of the colugo mitochondrial reference genome. 92.3% of the high-quality sequences from specimen 12 mapped to the reference mtDNA sequence, with a relatively even distribution of DNA fragments (Figure S2a). By contrast, 71.9% of the sequences from specimen 6 aligned to the reference genome. Sequence coverage for specimen 6 was more biased, with more reads aligning to the highly conserved 12S rRNA region of the genome (the left end of the linear genome) (Figure S2b).

Sequencing Results Summary:

192 sequences (99% pass-rate)

183 (95%) high-quality sequences after trimming

155 (85%) mtDNA

86 from spec. 12 (92% selection rate)

68 from spec. 6 (71% selection rate)

28 (15%) non-mtDNA

9 CynSINEs (colugo-specific SINE element).

19 genomic DNA or no specific match in GenBank

3 numts

Figure S2. Mitochondrial genome coverage based on preliminary Sanger sequencing of 96 clones from specimens a) 12, and b). 6 relative to the reference sequence of a Bornean colugo (GenBank accession number AJ428849) represented by the long green arrow at the top of the figure. The smaller sequences (short red and green arrows) are random fragments recovered after 2 rounds of mtDNA selection that were cloned into a plasmid vector. Genome coverage is depicted by the green, blue, and light blue bar at the bottom of the figure. Green regions represent areas of the genome that are covered with more than 1 sequence, the dark blue checkered areas represent that the reference sequence is covered by 1 sequence, and the light blue means no sequence coverage with respect to the reference.

Figure S2a : Distribution of reads from specimen 12 mapped onto the colugo reference genome.

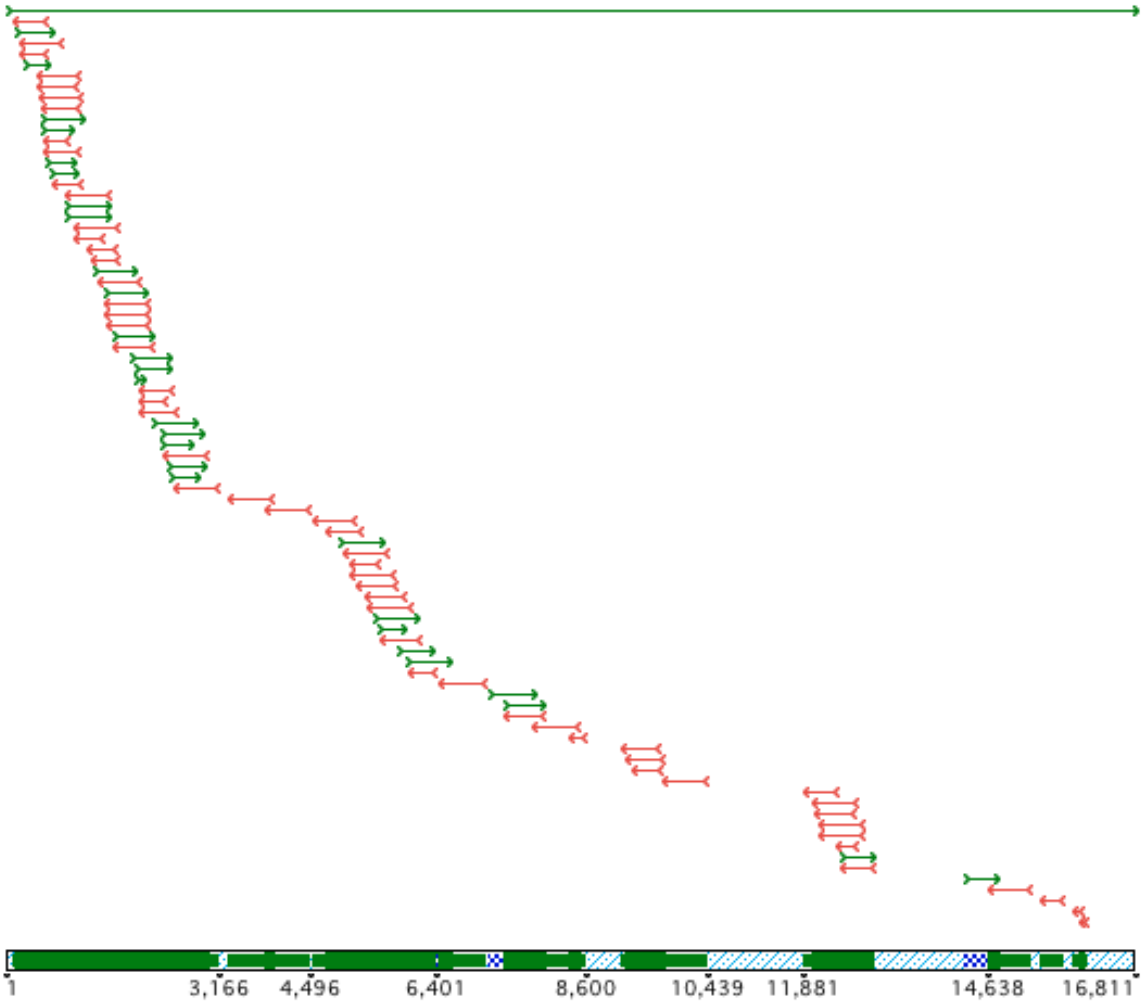
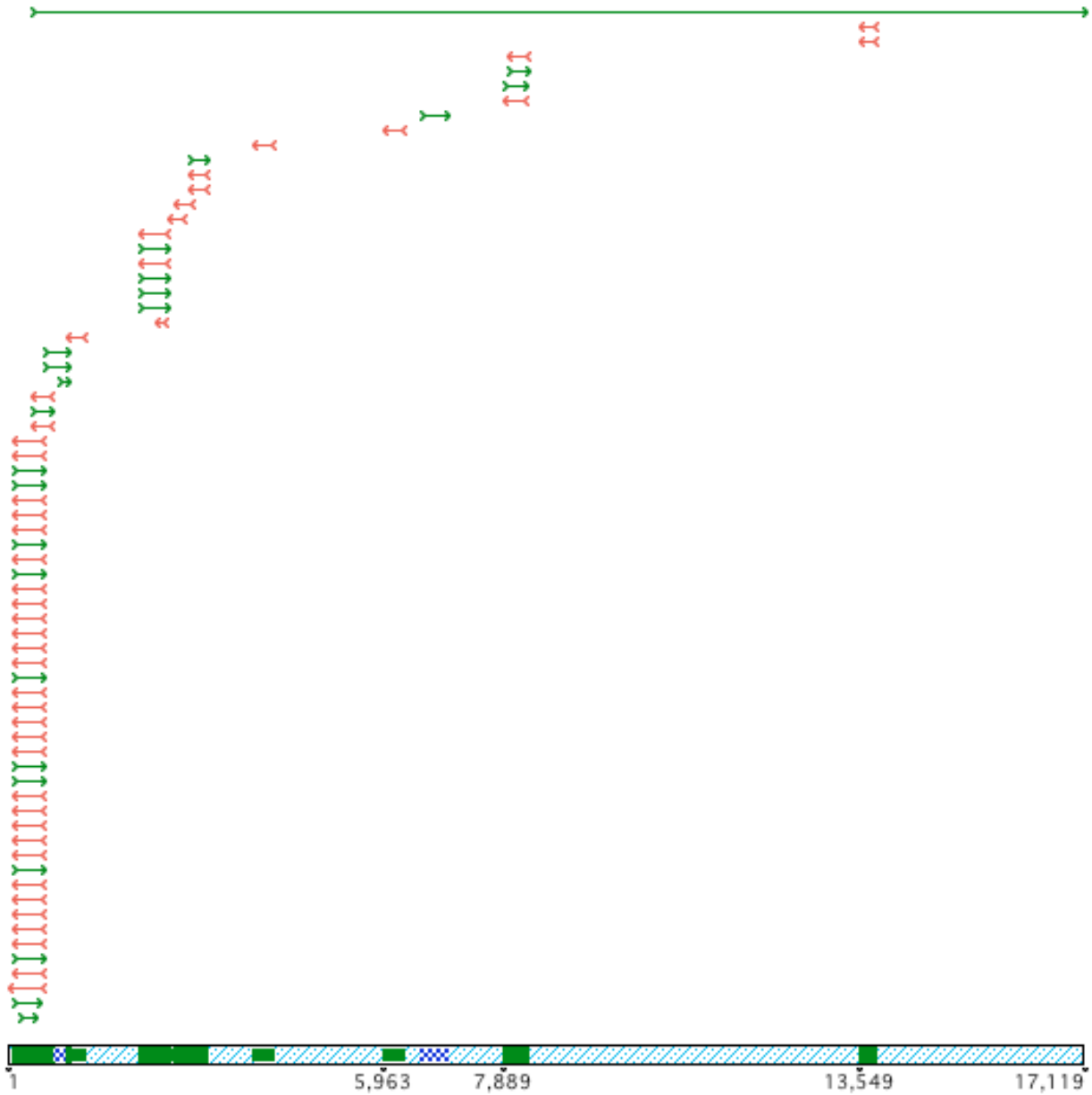


Figure S2b. Distribution of reads from specimen 6 mapped onto the colugo reference genome



Phylogenetic Robustness

To evaluate the potential effect of sequence coverage on the dataset, and our resulting phylogenetic conclusions, consensus sequences for each individual were subjected to variable sequence depth enforcements (d5, d10, d15, d25) to remove any bases that were not supported under the sequence depth criterion. Sequences from each category (d5, d10, d15, d25) were used to make an alignment for each depth. Each alignment at each depth was used to generate a maximum likelihood (ML) tree in RaxML 7.0.3 (Figure S3, next page). Little to no variation is observed in overall tree topology between the different depth enforcements, which supports robustness of the dataset. Support metrics on each node are based on 1,000 bootstrap replicates.

Figure S3. Maximum likelihood phylogenetic trees at variable sequence depth (d) enforcements. A) ML tree constructed from alignment d10. B) ML tree constructed from alignment d15. C) ML tree constructed from alignment d25.

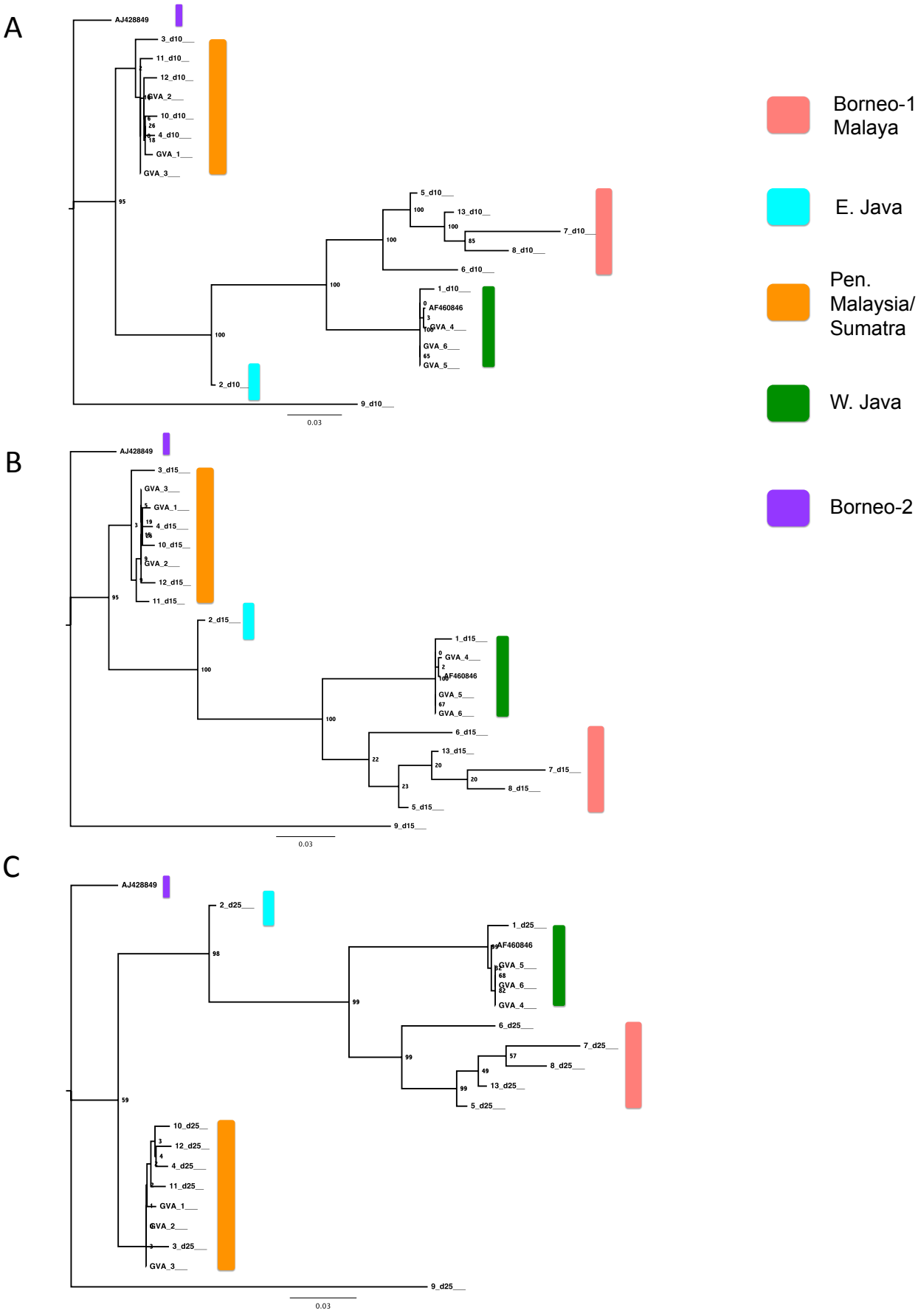


Figure S4. Maximum likelihood phylogenetic trees constructed from alignments where each site is present 30%, 50%, and 70% of sites across individuals. Analyses (B, D, E) are shown where the Natuna Islands specimen (9) was removed to examine effects of long-branch attraction on the phylogenetic support levels. The 50% tree is shown in Fig. 5, while the Natuna (9)-excluded tree for 50% is shown in E.

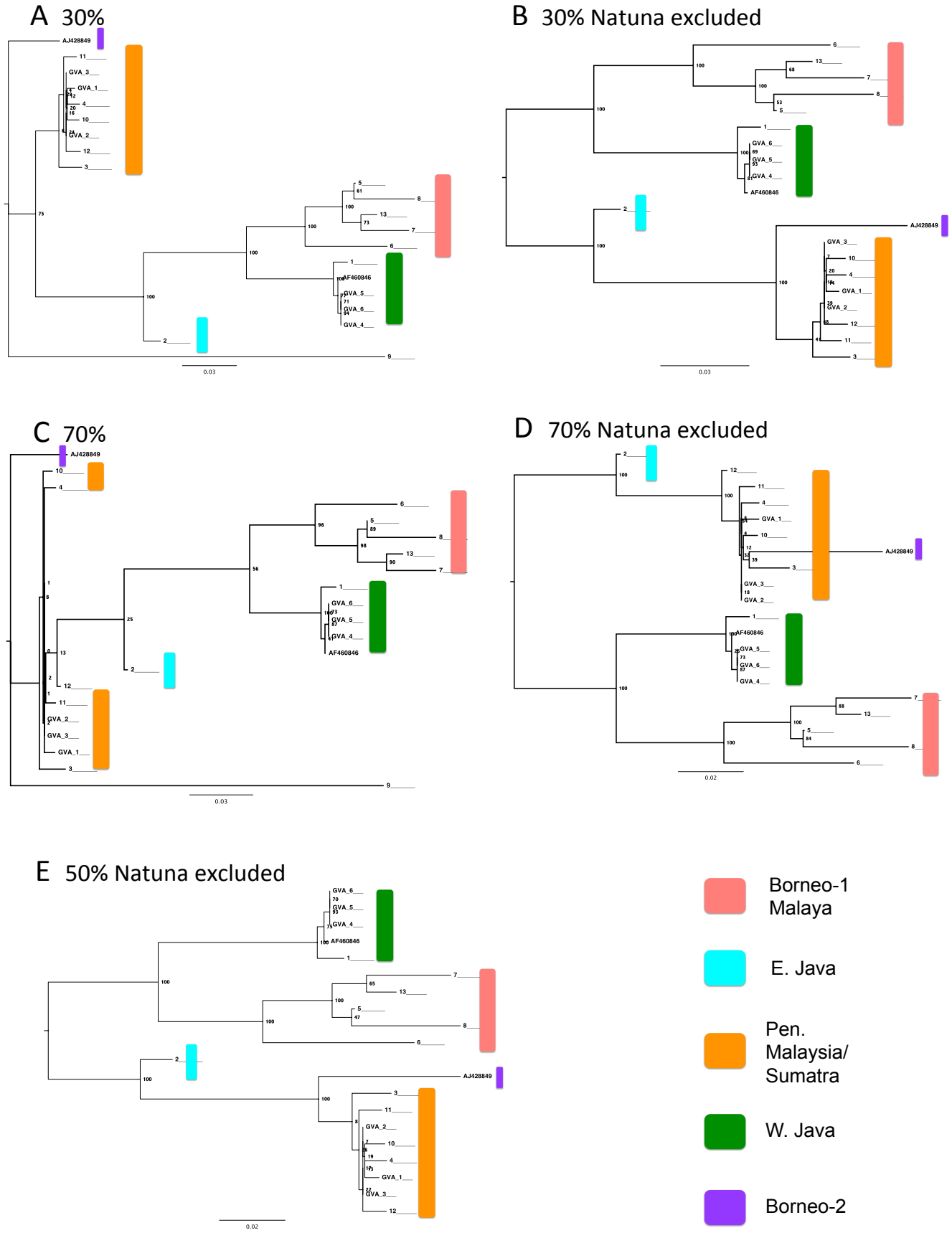


Figure S5. Biallelic SNP locations for each individual and SNP statistics. Only SNPs with minor allele frequency ≥20% were included. A) Major allele frequencies are plotted in their relative position across the mitochondrial genome for each individual. B) SNP Statistics.

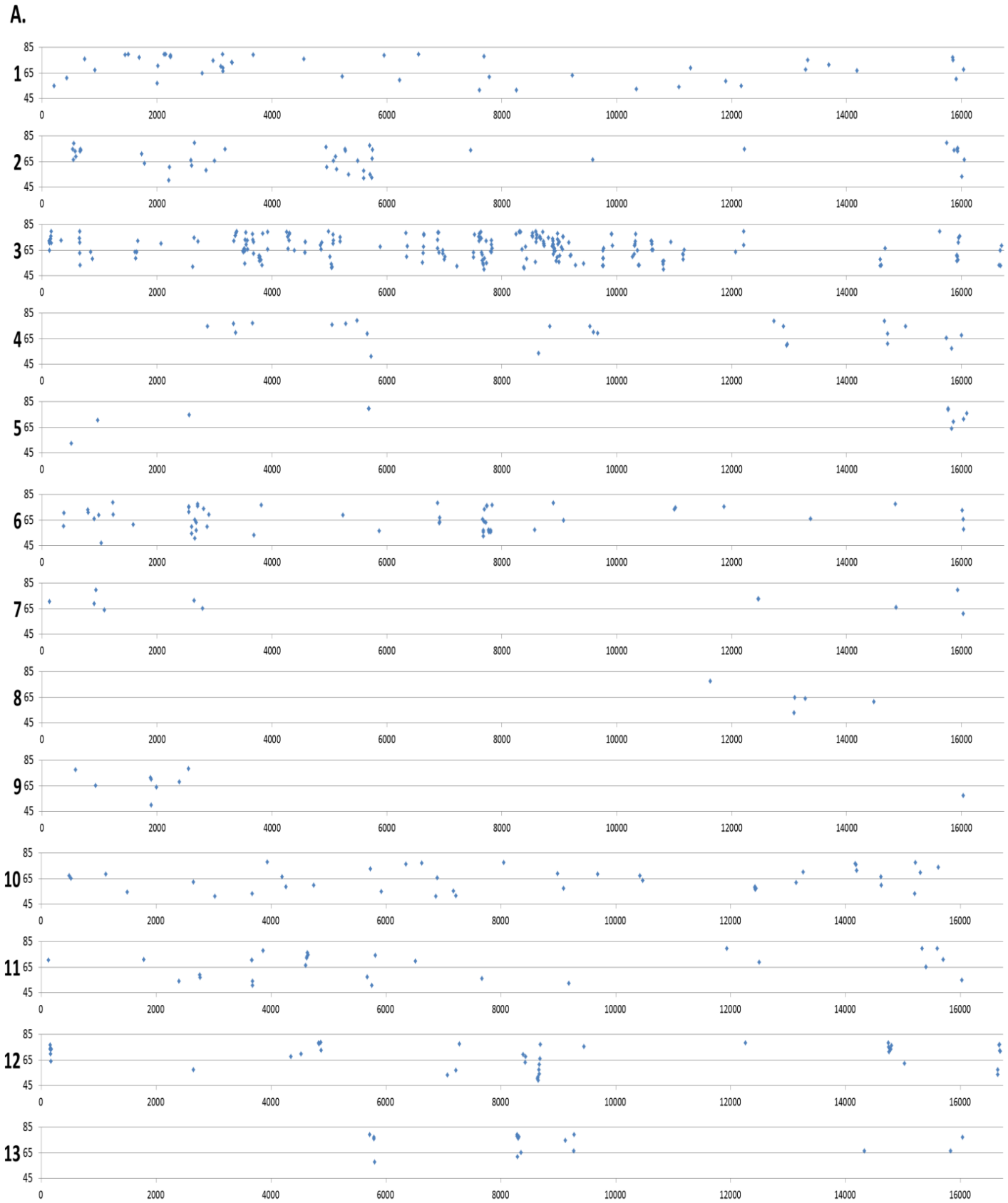


Fig. S5B. SNP statistics for each individual.

Specimen	# of SNPs	# of AA altering SNPS	# of sites analyzed	% of sites w/minor allele ≥20%	% of SNP sites that are AA altering	% of SNPs that are AA altering
1	46	8	15099	0.305	0.053	17.39
2	44	7	7956	0.553	0.088	15.91
3	225	37	15141	1.486	0.244	16.44
4	25	12	13044	0.192	0.092	48.00
5	11	0	4807	0.229	0.000	0.00
6	58	12	10098	0.574	0.119	20.69
7	12	3	3809	0.315	0.079	25.00
8	5	0	4687	0.107	0.000	0.00
9	9	0	2292	0.393	0.000	0.00
10	40	20	15807	0.253	0.127	50.00
11	28	9	11819	0.237	0.076	32.14
12	43	11	14918	0.288	0.074	25.58
13	16	4	3167	0.505	0.126	25.00
Average:	43.23	9.46	9434.15	0.42	0.08	21.24

Figure S6. Unabbreviated sequence depth maps of each individual.

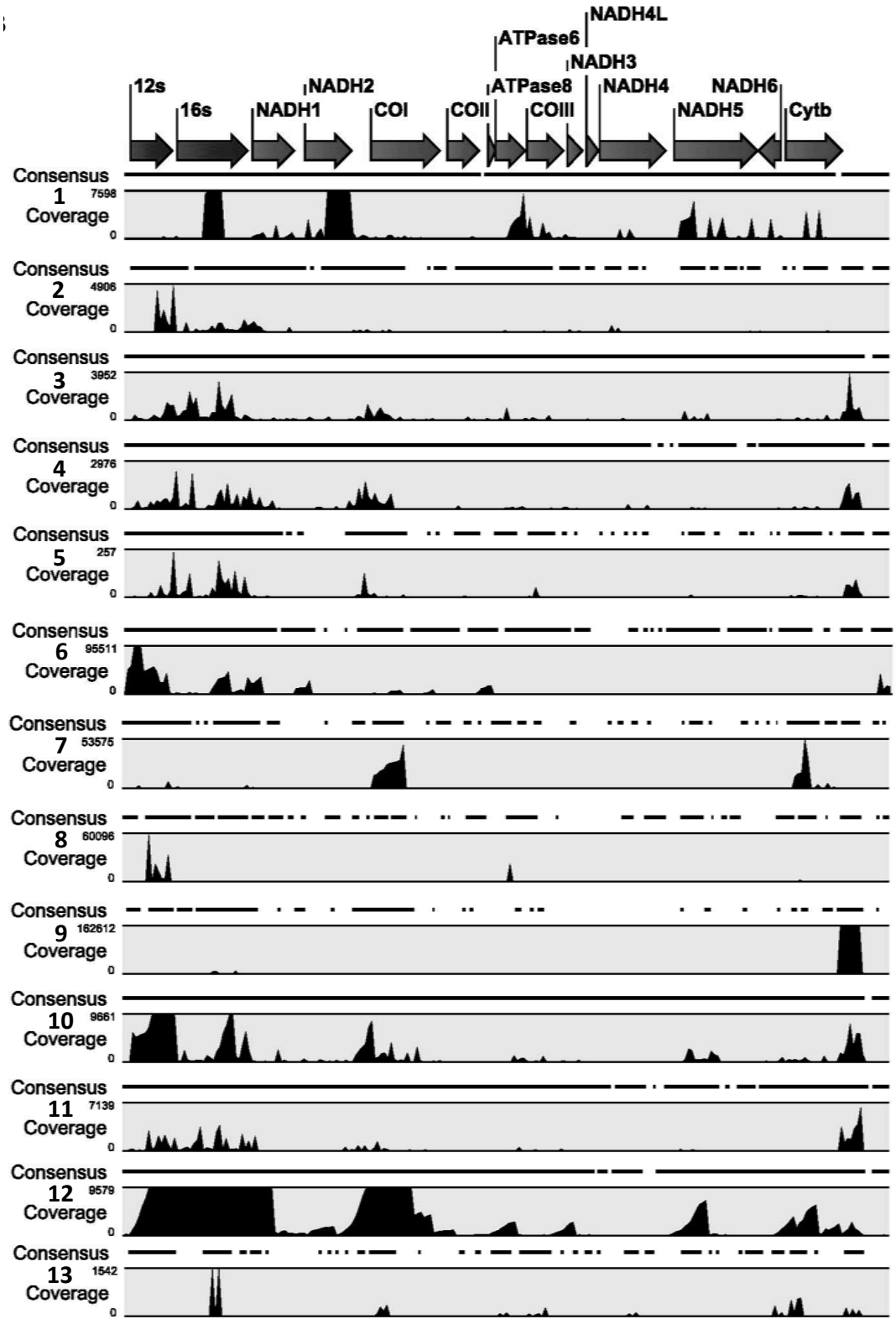


Table S1. Quantification of initial colugo DNA extracts

ID	ng/µL	260/280	Midpoint of DNA Smear
1	65.21	1.76	400
2	16.17	1.86	175
3	205.41	1.80	650
4	9.46	1.92	200
5	64.14	1.84	250
6	9.85	1.67	150
7	126.61	1.58	125
8	225.35	1.84	350
9	15.35	1.93	200
10	24.26	1.67	300
11	236.65	1.64	350
12	234.29	1.88	250
13	16.71	1.41	75

Table S2. Categorical classification of reads that did not align to the reference sequence.

I.D.	GVA_mtDNA	Potential_Numts	other_mtDNA	GVA_Nuclear	Human Nuclear	Potential_GVA_Nuclear	Other_Nuclear	Bacterial
1	0.06	2.83	2.75	0.17	0.00	0.05	0.00	0.00
2	6.46	0.00	3.47	0.00	0.59	0.66	1.20	0.54
3	3.85	0.96	3.10	0.00	0.00	0.00	0.03	0.00
4	2.04	0.20	1.14	0.00	0.00	0.00	0.00	0.00
5	4.95	0.93	1.00	0.00	0.00	0.00	0.00	0.00
6	20.11	0.00	2.80	0.01	0.00	3.68	0.16	0.00
7	2.78	0.00	1.95	0.00	0.26	2.49	1.05	2.24
8	5.19	0.82	0.46	0.00	5.80	0.00	0.09	0.30
9	0.00	0.00	16.21	0.00	0.00	0.02	10.40	0.41
10	4.50	0.70	2.77	0.01	2.69	0.00	0.15	0.00
11	5.34	0.01	0.96	0.00	0.00	0.00	0.04	0.02
12	3.05	0.78	1.73	0.02	0.00	2.37	0.00	0.28
13	0.09	0.00	0.00	0.00	4.23	0.58	0.00	0.34
Average:	4.49	0.56	2.95	0.27	0.78	0.76	1.01	0.32

Table S3. Average sequence depth per site. The first column lists the number of sites that meet the requirement of depth 5, the second lists the total number of nucleotides that meet the depth requirement, the third lists the depth/site.

Specimen	Number of Sites	Total Nucleotide Depth	Depth/site
1	15069	5826393	386.65
2	8040	2083649	259.16
3	15182	3736938	246.14
4	13058	2905341	222.50
5	4825	165591	34.32
6	10230	31392326	3068.65
7	3851	9103187	2363.85
8	3929	9493576	2416.28
9	2256	22605458	10020.15
10	15864	10165381	640.78
11	12035	4938501	410.34
12	14967	16563249	1106.65
13	2739	517622	188.98
ALL Individuals	122,045	119,497,212	979.12

Table S4. Assessment of chemical damage. *A*) The total number of transitions and transversions are shown for each individual, where "TC" indicates a transition from thymine in the reference sequence to cytosine in the specimen consensus sequence. *B*) The number of complementary base transitions CT + GA and TC + AG. A chi-square test was used to test for significance. *C*) The percentage of sites with CT and GA transitions in each consensus sequence at depth 5.

A.

Number of Directional Substitutions													
I.D.	Reference:	TC	TA	TG	CT	CA	CG	AT	AC	AG	GT	GC	GA
1	AF460846	51	2	0	57	0	0	1	2	22	3	1	16
2	AJ428849	157	6	3	189	10	0	11	10	85	3	3	75
3	AJ428849	221	5	4	244	9	2	9	7	132	2	2	131
4	AJ428849	165	3	1	209	10	1	9	6	101	2	1	99
5	AF460846	106	13	1	71	9	0	0	7	54	1	3	38
6	AF460846	280	13	3	190	14	1	11	9	135	1	4	108
7	AJ428849	74	5	5	91	6	2	6	4	42	0	1	37
8	AF460846	126	5	5	127	16	4	12	10	61	4	3	47
9	AJ428849	32	16	2	51	17	2	19	6	26	1	3	26
10	AJ428849	216	5	1	273	10	1	8	7	137	2	1	134
11	AJ428849	161	3	1	188	8	1	7	5	93	2	1	101
12	AJ428849	202	6	1	248	13	0	10	7	135	2	1	124
13	AF460846	68	2	1	47	4	1	1	6	39	0	2	24
	Total:	1859	84	28	1985	126	15	104	86	1062	23	26	960

B.

ID	CT + GA	TC + AG	P-value
1	73	73	1.000
2	264	242	0.328
3	375	353	0.415
4	308	266	0.080
5	109	160	0.002
6	298	415	<0.001
7	128	116	0.442
8	174	187	0.494
9	77	58	0.102
10	407	353	0.050
11	289	254	0.133
12	372	337	0.189
13	71	107	0.007
Total:	2945	2921	0.754

C.

ID	Total # of Sites	Total CT and GA Transitions	Percentage of Sites Affected by CT and GA
1	15099	73	0.48
2	7956	264	3.32
3	15141	375	2.48
4	13044	308	2.36
5	4807	109	2.27
6	10098	298	2.95
7	3809	128	3.36
8	4687	174	3.71
9	2292	77	3.36
10	15807	407	2.57
11	11819	289	2.45
12	14918	372	2.49
13	3167	71	2.24
Total	122644	2945	2.40
		Average:	2.62

Table S5 Open reading frame analysis.

ID	NADH1	NADH2	COI	COII	ATPase8	ATPase6	COIII	NADH3	NADH4L	NADH4	NADH5	NADH6	CYTB
1	2	1	1	3	2	2	1	1	2	4	2	1	4
2	4	3	4	3	4	5	5	3	4	5	5	5	3
3	1	2	1	2	1*	1	1	1	1	4	5	4	2
4	1	3	3	3	4	3	2	2	4	5	5	5	2
5	5	4	4	5	n/a	5	5	5	n/a	n/a	5	n/a	4
6	2*	2	2	3	4	3	2	3	4	5	5	5	5
7	4	n/a	3	5	n/a	5	n/a	3	n/a	5	5	n/a	5
8	5	5	n/a	3	4	3	4	n/a	n/a	3*	5	n/a	5
9	n/a	4	5	n/a	n/a	n/a	5	n/a	n/a	n/a	5	5	5
10	1	1	1	4	2	1	1	1	1	2	3	2	2
11	2	2	2	3	4	3	1	2	3	4	5	5	4
12	1	1	1	1	1	1	1	1	1	4	5	4	1
13	n/a	5	5*	n/a	n/a	5	5	n/a	n/a	5	5	5	5

LEGEND:
1 = Complete CDS with intact ORF (i.e. premature stop codon)
2 = Incomplete (i.e. gaps in) CDS, start and stop codons present, and intact partial ORF
3 = Incomplete CDS with the correct stop codon, no start codon present, and intact partial ORF
4 = Incomplete CDS with the correct start Codon, no stop codon present, and intact partial ORF
5 = Incomplete CDS with no start or stop codon present, and intact partial ORF
* = Premature stop codon in CDS
n/a = No sequence present

Table S6. Mitochondrial sequence divergence between Sunda colugos

	AJ428849	AF460846	1	2	3	4	5	6	7	8	9	10	11	12	13	GVA_4
AJ428849																
AF460846	0.124															
1	0.127	0.011														
2	0.074	0.067	0.065													
3	0.054	0.119	0.121	0.057												
4	0.050	0.124	0.126	0.052	0.018											
5	0.096	0.069	0.071	0.083	0.092	0.095										
6	0.113	0.086	0.088	0.093	0.106	0.107	0.043									
7	0.080	0.095	0.097	0.067	0.067	0.070	0.041	0.078								
8	0.137	0.100	0.099	0.121	0.130	0.142	0.034	0.063	0.059							
9	0.101	0.132	0.128	0.109	0.107	0.103	0.124	0.131	0.066	0.206						
10	0.051	0.125	0.128	0.053	0.019	0.012	0.094	0.114	0.069	0.139	0.105					
11	0.051	0.121	0.122	0.046	0.018	0.014	0.092	0.107	0.070	0.136	0.100	0.014				
12	0.052	0.127	0.129	0.046	0.021	0.013	0.098	0.112	0.069	0.141	0.103	0.013	0.014			
13	0.112	0.074	0.082	0.069	0.106	0.111	0.033	0.076	0.043	0.030	0.106	0.113	0.108	0.119		
GVA_4	0.125	0.003	0.011	0.067	0.121	0.125	0.070	0.087	0.097	0.101	0.130	0.127	0.122	0.129	0.076	

The number of base substitutions per site between individual sequences are shown. Analyses were conducted using the Maximum Composite Likelihood model [1]. The analysis involved 16 nucleotide sequences. All positions with less than 50% site coverage were eliminated. That is, fewer than 50% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 14008 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 [2]

1. Tamura K., Nei M., and Kumar S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences (USA) 101:11030-11035.

2. Tamura K., Peterson D., Peterson N., Stecher G., Nei M., and Kumar S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution (In Press)

Table S7 Pairwise sequence divergence from the probe sequence

Approximate Genetic Divergence (ML distance) from Probe (GVA4)						
(5,459bp)70%*			(14,008bp)50%*			(16,051bp)10%*
0.2%	AF		0.3%	AF		0.3%	AF
1.0%	1		1.1%	1		1.1%	1
6.4%	2		6.7%	2		6.6%	2
6.6%	5		7.0%	5		7.0%	5
7.7%	6		7.6%	13		7.4%	13
7.9%	13		8.7%	6		8.8%	6
9.1%	7		9.7%	7		9.8%	7
9.5%	3		10.1%	8		10.4%	8
9.8%	AJ		12.1%	3		12.3%	11
9.9%	11		12.2%	11		12.4%	3
10.0%	4		12.5%	4		12.6%	4
10.0%	10		12.5%	AJ		12.9%	AJ
10.0%	12		12.7%	10		13.0%	10
10.2%	8		12.9%	12		13.0%	9
11.8%	9		13.0%	9		13.1%	12
	Legend: Phylogenetic groupings						
	W. Java						
	E. Java						
	Borneo-1/Malaya						
	Pen.Malaysia/Sumatra						
	Natuna Islands						
	Borneo-2						
*The length of the alignment matrix based on sites for which ≥X% of the individuals possess a base at that site.							

Table S8. Primer pairs used to amplify Sunda colugo mtDNA genome probe fragments.

			Reference Seq: AJ428849		
	Start	End	Forward Primer Sequence (start)	Reverse Primer Sequence (end)	Size (bp)
1	25	1084	GCAAGGTACTGAAAATACCAAGATG	TGAAATCTTCCGGGTGTAGG	1060
2	847	1958	CAAAGGAGGATTTAGCAGTAAATTAAG	TGCTAGAGGTGATGTTTTTGG	1112
3	1621	2681	GCCACCAATTAAGATAGCGTTC	CTAACAAGCCCTGCTCTTGG	1061
4	2426	3809	CTCGATGTTGGATCAGGACA	TTCTCAGGAGTGGGTTCGAT	1382
5	3573	4328	CGAGCTTCATACCCACGATT	GGCTAGTTTTTGTCATGTCAGG	756
6	4062	5340	AACCCACGATCAACAGAAGC	AGGGTGAGGTGGCTGAGTAA	1279
7	5171	6473	CTACTTCTCCCGCCTCCAAG	TGTGCTACTACGTAATATGTGTCGTG	1303
8	6319	7221	GCTACACTGCACGGAGGAA	TGGTTTCTACTATTTGGGCATTT	903
9	6975	8348	AAAAACATTACATGACTTCGTCAGA	GGTGTGCCTTGGGGTAGAAG	1374
10	7777	8951	CCACAATGAAATGCCACAAC	TGGAGCTAGGCTTGAGTGGT	1175
11	8559	9664	CACCGTAGCCCTAATCCAAG	ACGTGATGGCCACTAGGAAA	1106
12	8864	9867	ACGATACGGAATAATTCTCTTCA	AATGGGTCGAAACCAGTTGT	1004
13	9628	10837	CCCTTCTCCATAAAATTTTTCC	TTTTGGTAGTCAGAGGTGAAGTC	1210
14	10550	11697	GAAGCAACACTAATCCCAACC	TTGAAAGTAAGAAAGCCATATTTTT	1148
15	11334	12602	CAGCATTCTCCTGATCAAACA	AGTGTGGTGAGGGCACCTA	1269
16	12431	13869	TACACCCGTGACTTCCCTCT	TACTGCCATGGCTATTGAGG	1439
17	13660	14806	GTAGAATCCCCATGAAAATAACC	GGGATTTTGTCTGAGTTTGATG	1147
18	14660	15919	AGACAAAGCCACCCTCACAC	GCATGGCCCTGAAGTAAGAA	1260
19	15349	16734	CTCCCCAGGACAATCAAGG	GCTTCAGGCCAAAATTCAAA	1386