PNAS
www.pnas.org

Supplementary Information for

Evidence from sperm whale clans of symbolic marking in non-human cultures

Taylor A. Hersh, Shane Gero, Luke Rendell, Maurício Cantor, Lindy Weilgart, Masao Amano, Stephen M. Dawson, Elisabeth Slooten, Christopher M. Johnson, Iain Kerr, Roger Payne, Andy Rogan, Ricardo Antunes, Olive Andrews, Elizabeth L. Ferguson, Cory Ann Hom-Weaver, Thomas F. Norris, Yvonne M. Barkley, Karlina P. Merkens, Erin M. Oleson, Thomas Doniol-Valcroze, James F. Pilkington, Jonathan Gordon, Manuel Fernandes, Marta Guerra, Leigh Hickmott, Hal Whitehead

*Corresponding author: Taylor A. Hersh

Email: taylor.a.hersh@gmail.com

This PDF file includes:

Supplementary text (Method S1 and S2; Discussion S1 to S3)
Figures S1 to S13
Tables S1 to S8
SI References

Supplementary Information Text

Method S1. Coda classification using IDcall.

We fitted 2:15 (instead of 2:10) mixture components to the data during coda classification to account for the likely presence of additional coda types in this analysis compared to Hersh et al. (1), given our expanded geographic scope and dataset size. The most computationally demanding step in IDcall is call classification, in which every combination of a user-set number of mixture components and each of a family of 14 models (2) is fit to the data and compared using an information criterion. Using 2:15 mixture components and all 14 models results in 196 (i.e., 14x14) component/model combinations for each number of clicks (3–10 here), yielding 1,568 combinations total. To reduce the computational load, we performed preliminary data exploration using the 'mclust' R package (3) to determine if a subset of the 14 possible models consistently fit best, as determined by the Bayesian Information Criterion (BIC), the data for each coda click length (4). Models VVV, VVE, VEV, and VEE (see (2) for model descriptions) were typically the four best models regardless of coda click length, with model VVV almost always best (Figure S5). Subsequently, we used those four models during the call classification stage of IDcall.

Method S2. Repertoire distance calculations.

We used the 'marmap' R package (5) to calculate the distance between repertoires. Bathymetric data for the Pacific Ocean were imported into R from the ETOPO1 dataset (6) hosted on the NOAA server at one degree (60-minute) resolution. Shortest path calculations between repertoires were restricted to waters 1 km or deeper, as female sperm whales typically inhabit this water depth (7, 8). The coarseness of one degree resolution resulted in some repertoires (n=24) located in grid cells with <1 km water depth. We shifted those repertoire positions incrementally by half a degree in a cardinal direction until the water depth reached ≥1 km. Most repertoires only needed to be shifted half a degree (n=18), but a few had to be shifted one degree (n=5). A single repertoire recorded off Southern New Zealand had to be shifted one full degree southeast. One degree resolution corresponds to grid cells spanning ~111x111 km, meaning that two repertoires recorded within 111 km of each other are assigned a distance of 0 km. To achieve higher resolution in distance calculations for geographically close repertoires, we used the 'geosphere' R package (9) to calculate the great-circle distance for all repertoires within 200 km of each other. This distance matrix was merged with the other, such that the distance between repertoires separated by <200 km was calculated using the great-circle distance and the distance between repertoires separated by ≥200 km was calculated using bathymetric data with the 1 km depth minimum. These adjustments (jittering perfectly overlaid repertoires, shifting repertoires into deeper waters, etc.), coupled with sampling limitations (i.e., variable precision of localization data available for different regions; Table S5), meant that the between-repertoire distance calculations were approximate. However, the main distance distinctions of interest in this study—very close vs. very far—were well preserved.

Discussion S1. Determining the number of sperm whale clans in the Pacific Ocean.

Varying *critfact* from 3 to 20 while keeping *minrep* constant at 15 (Table S8) produces three reasonable (i.e., at least 75% of repertoires were assigned to a clan) tree scenarios: a 5-clan tree (Figures S6/S7, Table S6), a 7-clan tree (Figures 2/S8, Table S2), and an 8-clan tree (Figures S9/S10, Table S7). The tree cophenetic correlation coefficient was 0.905. Four clans are robustly delineated across scenarios: three well-known clans (the 'Four-Plus', 'Plus-One', and 'Regular' clans) (10, 11) and a clan first described as putative in Hersh et al. (1) using a much smaller and more geographically restricted dataset. As in Hersh et al. (1), variation across trees centered on repertoires dominated by shorter (mainly 3- and 4-click) codas, with these repertoires (n=85) forming one, three, or four clans, respectively, in the aforementioned trees. Codas with three or four clicks have low dimensionality (and potentially lower information content, as seen in human words; 12) and form more diffuse clusters than codas with more clicks, which may have contributed to the uncertainty of clan divisions for repertoires dominated by codas with fewer clicks. The available evidence supports the 7-clan tree (with three of these 'shorter coda' clans; see the next two paragraphs) but note that the division of repertoires dominated by shorter codas into clans is more uncertain than the division of repertoires into the four robust clans. This is clearly illustrated by the posterior probability plots in Figures 2 (7-clan tree), S6 (5-clan tree), and S9 (8-clan tree).

Photographic and acoustic work found evidence for two clans off Japan (13), with whales recorded off the Ogasawara Islands sharing dialect similarities with the Short clan (which was first documented in the eastern tropical Pacific) (10) and whales recorded off the Kumano coast belonging to

a different clan. In the 5-clan scenario of the present analysis, the Ogasawara Islands and Kumano coast repertoires are lumped into one large 'Short' clan (Figures S6/7). When we ran just the Japanese codas through IDcall, we consistently replicated the results from Amano et al. (13), with the Ogasawara Islands and Kumano coast repertoires clearly divided into two clans, even at extremely high values of *critfact* (Figure S11). This is promising, given several methodological differences between the two studies. For example, Amano et al. (13) standardized ICIs by coda length and used all recorded 3–10 click codas, while we used absolute ICIs and required 25 codas per repertoire (this latter requirement reduced our coda sample size compared to theirs). This suggests that the Ogasawara Islands and Kumano coast repertoires do indeed belong to different clans, and that any tree lumping them together (like the 5-clan tree; Figures S6/S7) is likely incorrect (but see Figure S12 for a version of the distribution map with five clans).

The 7- and 8-clan trees differ in that the 'Short' clan in the 7-clan tree (Figure 2, in red) is split into two clans in the 8-clan tree (Figure S9, in light blue and red), with repertoires primarily divided by longitude (i.e., west vs. east) (Figure S13). This longitudinal pattern fits well with within-clan acoustic drift with geographic distance, rather than two fully diverged clans. This possibility is supported by the fact that the 'Short West' clan identity coda is rarely used (Figure S9), the 'Short East' clan identity coda (Figure S9) is the same as the 7-clan tree Short clan's identity coda (Figure 2), and both the Short West and the Short East clans primarily make three click codas (Figure S10). For these reasons, the 7-clan tree is the most compelling of the three scenarios and is the focus of subsequent analyses, but we emphasize that there is clearly more uncertainty in the clan structure of whales that predominantly make shorter codas.

Discussion S2. Clan naming conventions.

We refer to the clan first described in Hersh et al. (1) as the 'Palindrome' clan, in homage to the palindromic nature of the clan's most frequently used identity codas (Figure 2, in orange). We suggest referring to the clan first described by Amano et al. (13) off the Kumano coast as the 'Rapid Increasing' clan, given that their work and ours show that very short duration codas with increasing ICIs predominate in this clan (Figure 2, in gold). The 'new' clan also makes identity codas with increasing ICIs, but the overall duration of these codas is much longer than in the Rapid Increasing clan; accordingly, we suggest referring to this clan as the 'Slow Increasing' clan (Figure 2, in purple).

In line with past work, we found that Regular clan whales frequently make codas with equally spaced (i.e., isochronous) clicks, Plus-One clan whales make codas with an extended pause before the final click, and Four-Plus clan whales make codas with a root of four isochronous clicks (10, 11). As was previously described, codas from Short clan whales do not appear to follow a rhythmic motif but are typically comprised of very few clicks (for example, the Short clan's only identity coda has three clicks) (Figures 2/S8) (1, 10, 11).

Discussion S3. Is the Short clan dialect basal?

From a production standpoint, the Short clan's isochronous, 3-click identity coda is one of the rhythmically simplest codas that can be made. Isochronous signals are prevalent in acoustic communication systems across taxa (14), which could indicate that they are evolutionarily basal throughout the animal kingdom. There is also an interesting trend in the plot of minimum spatial overlap vs. coda usage similarity in the between-clans analysis, where the plot shows a lopsided, inverted V shape (Figure S4A). Clans with low (minimum) spatial overlap have low coda usage similarity (which fits with the pattern expected from drift); clans with intermediate (minimum) spatial overlap have higher coda usage similarity; and clans with high (minimum) spatial overlap have the lowest coda usage similarity. The net effect is a decrease in coda usage similarity with increasing clan overlap, but the increase in coda usage similarity at intermediate overlap is interesting. The clan pairs with intermediate overlap may have had a more recent common ancestral dialect, as evidenced by their high coda usage similarity. Closer examination of the clan pairs with intermediate overlap shows that the Short clan is typically one of the two clans in the pairs; our most parsimonious guess is that the ancestral sperm whale dialect in the Pacific Ocean likely shared the most similarities with the modern Short clan dialect. Lastly, the Short clan is the only clan detected off southern New Zealand (Figure 1), where the codas were made by males, not females (15, 16). Male sperm whales infrequently make codas and likely mate across clans (17). We do not know if males adopt the clan dialect of the females they are with during courting or if they even make codas on these occasions, but the Short clan detections off southern New Zealand suggest that males may primarily make Short clan-style codas when they are not with females.

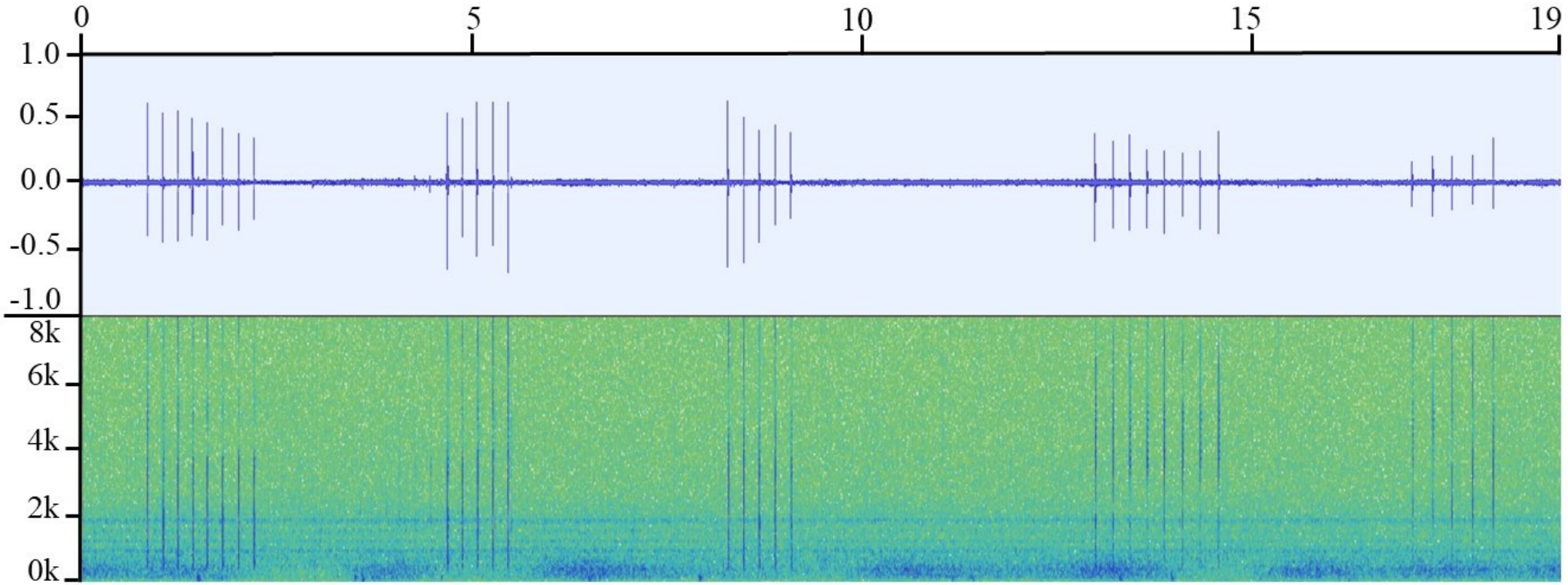


Figure S1. Waveform (top; y-axis is signal amplitude) and spectrogram (bottom; y-axis is frequency in Hz) showing five codas produced by a Regular clan sperm whale over a period of 19 s (time bar, in s, at top). From left to right, the codas contain 8 clicks, 5 clicks, 5 clicks, 8 clicks, and 5 clicks. The corresponding audio file can be accessed through the Open Science Framework (https://osf.io/ae6pd/). This segment is from a longer recording taken on May 30, 1995, off the Galápagos Islands. The spectrogram and waveform were generated in Audacity (window size: 1024, window type: Hann).

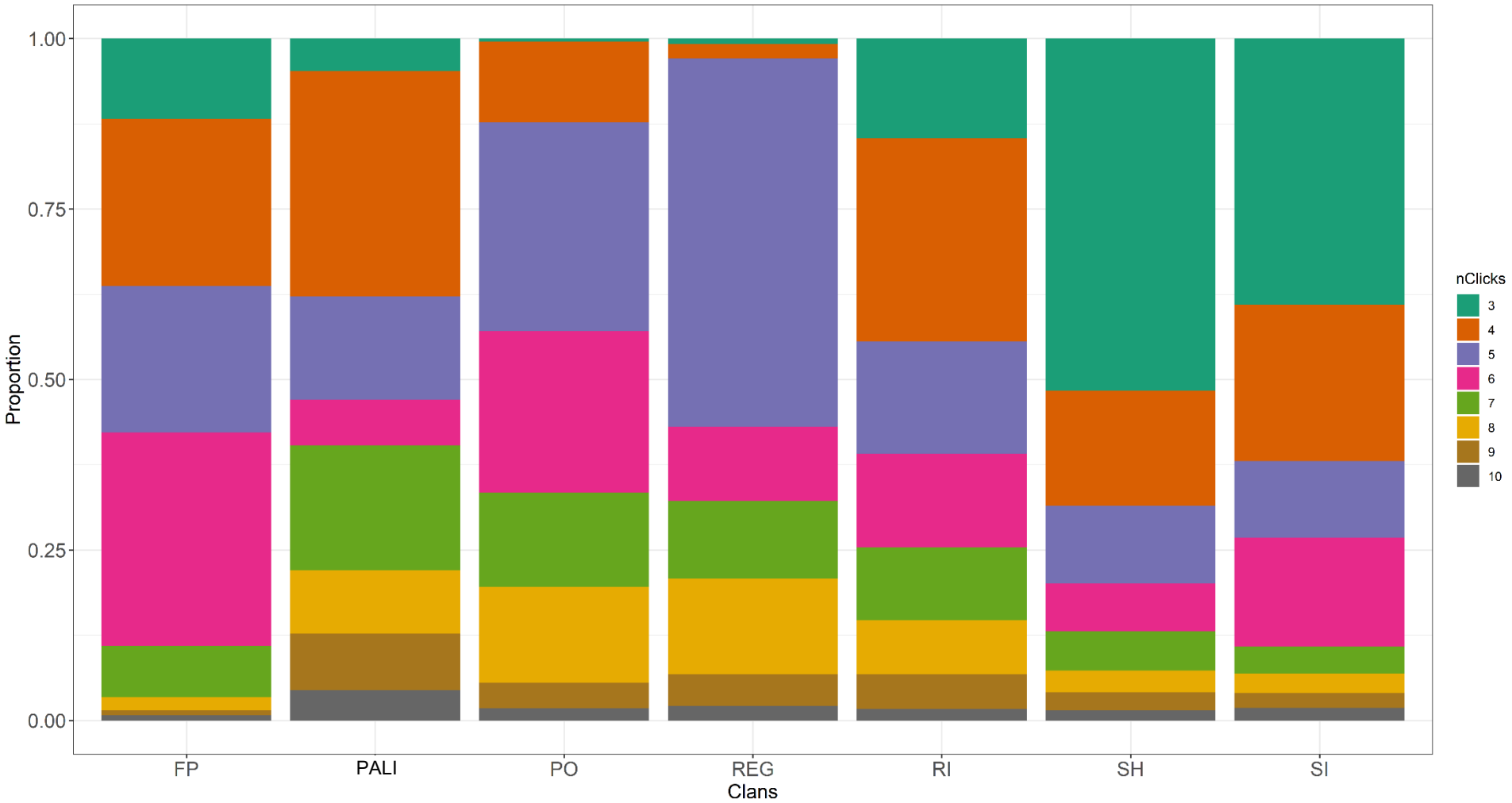
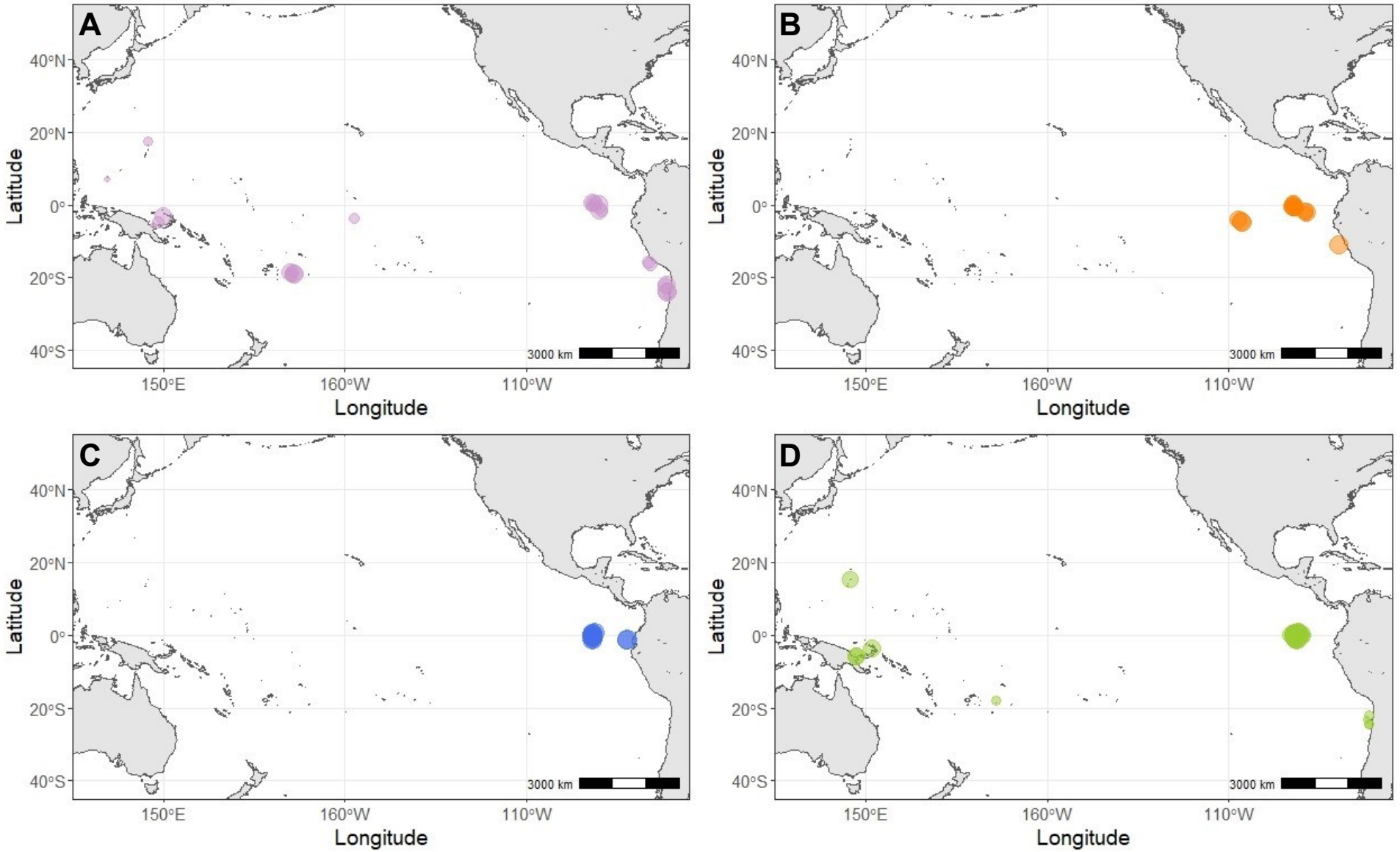


Figure S2. Proportion of clan codas made up of different click lengths in the 7-clan tree scenario. Clan abbreviations are: FP=Four-Plus, PALI=Palindrome, PO=Plus-One, REG=Regular, RI=Rapid-Increasing, SH=Short, and SI=Slow Increasing.


A
B
C
D
40°N
20°N
0°
20°S
40°S
Latitude
150°E
160°W
110°W
Longitude
3000 km

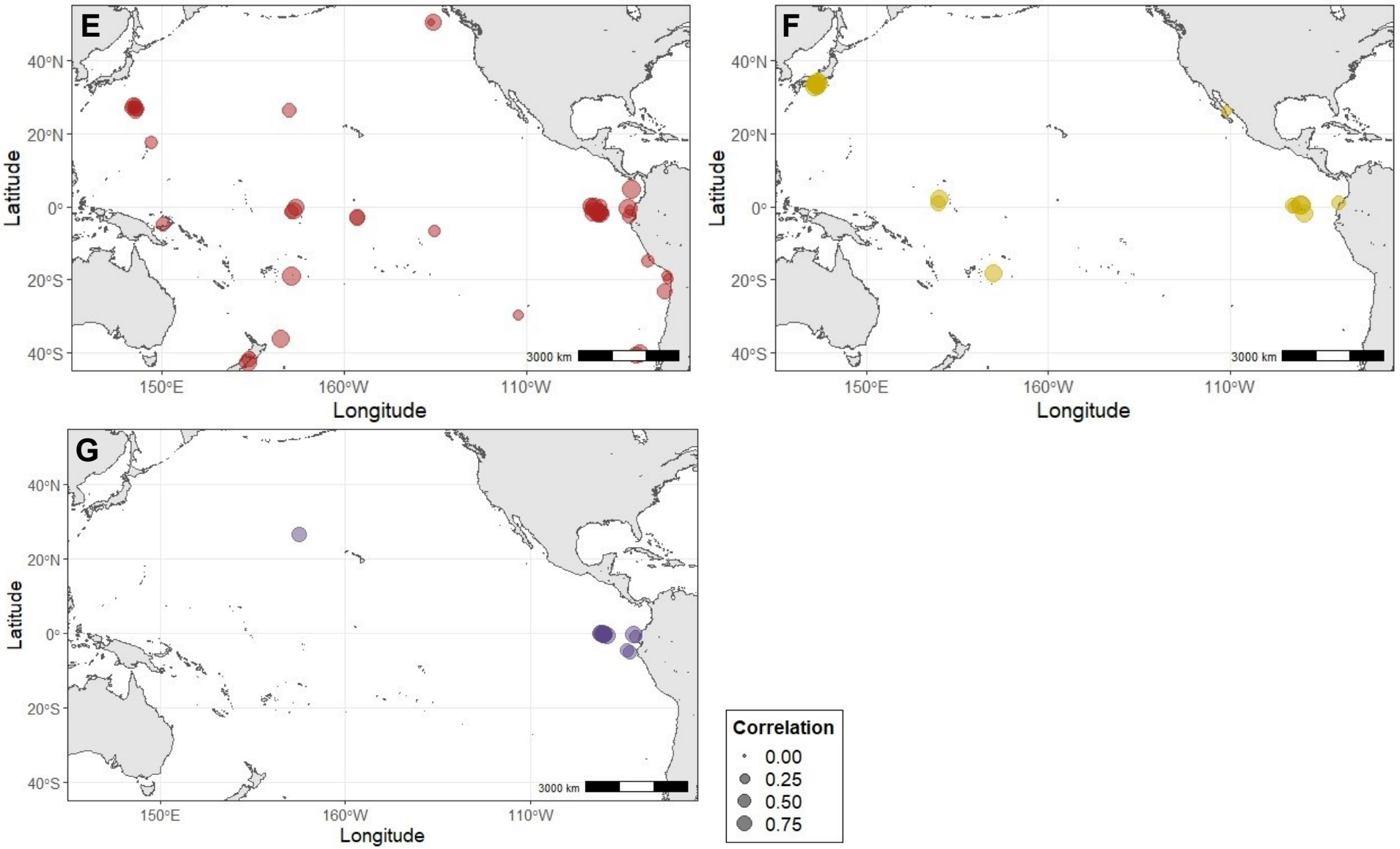


Figure S3. Maps of individual clan distributions across the Pacific Ocean using the 7-clan tree scenario clan designations. See Figure 1 for a composite map and for additional details. Panels correspond to the different clans: (A) Four-Plus, (B) Palindrome, (C) Plus-One, (D) Regular, (E) Short, (F) Rapid Increasing, and (G) Slow Increasing.

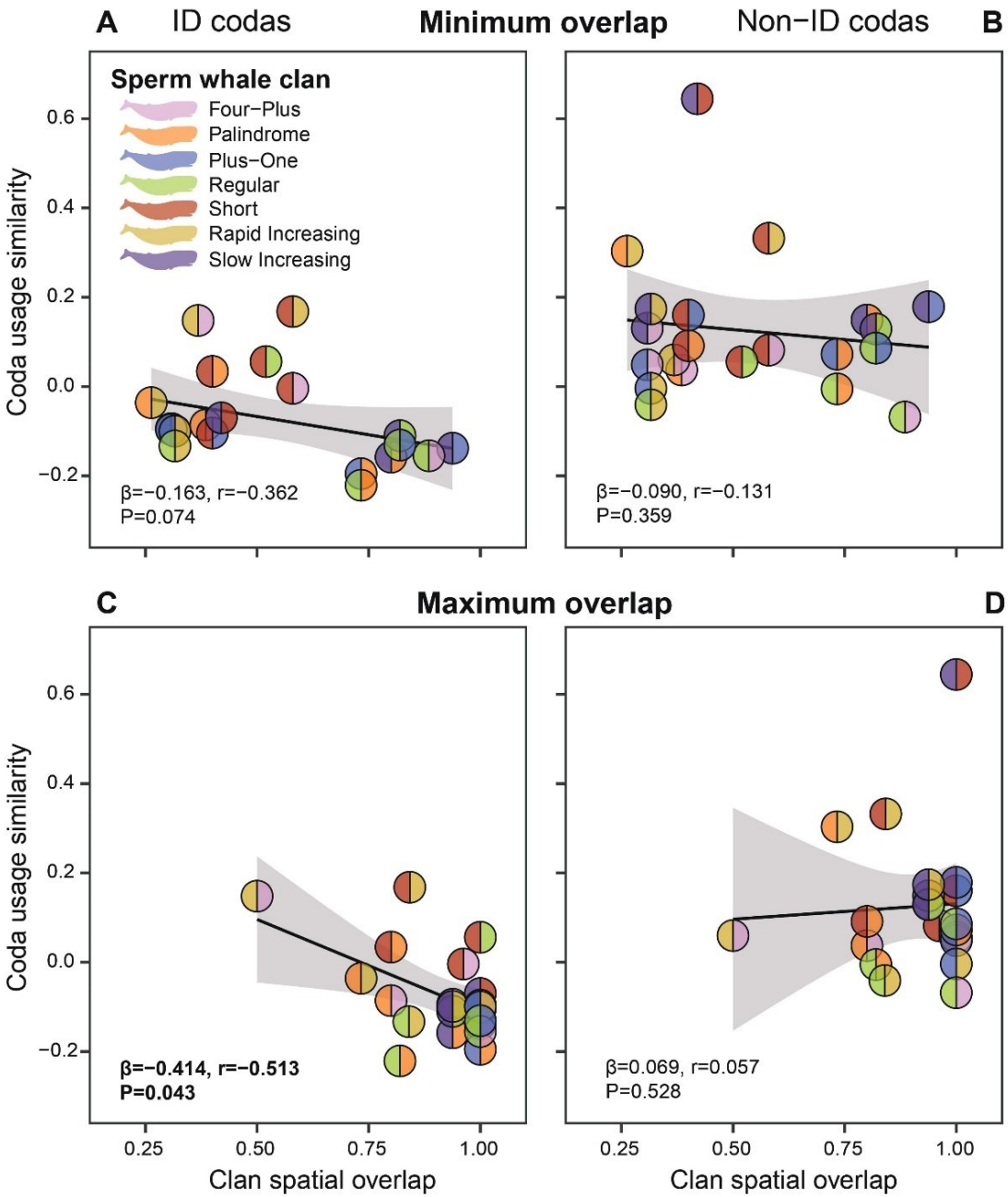
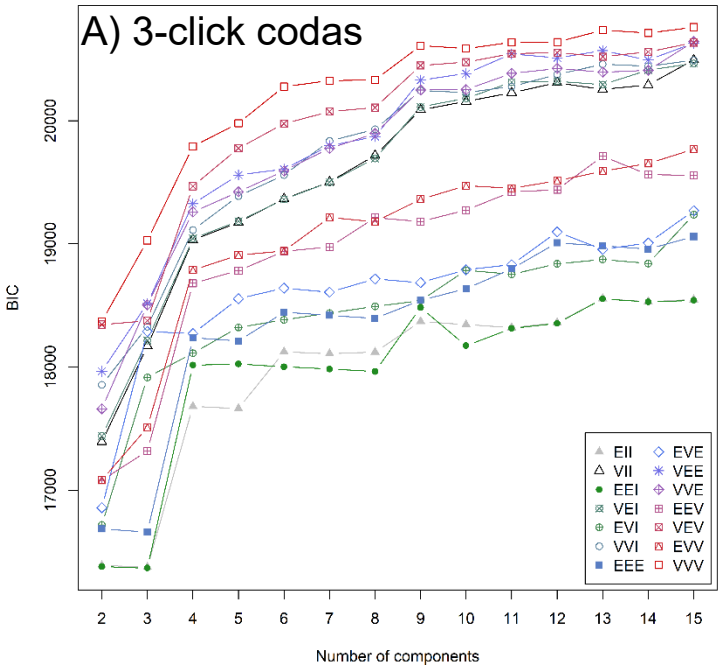
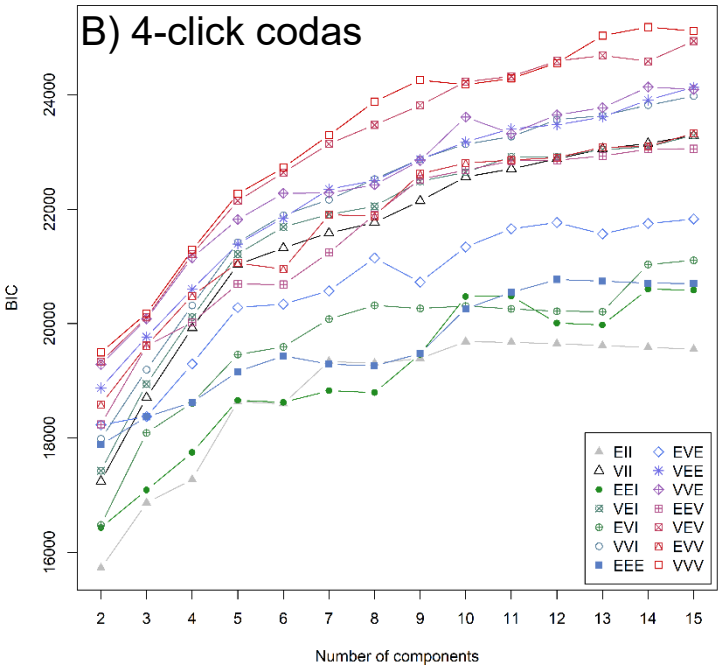
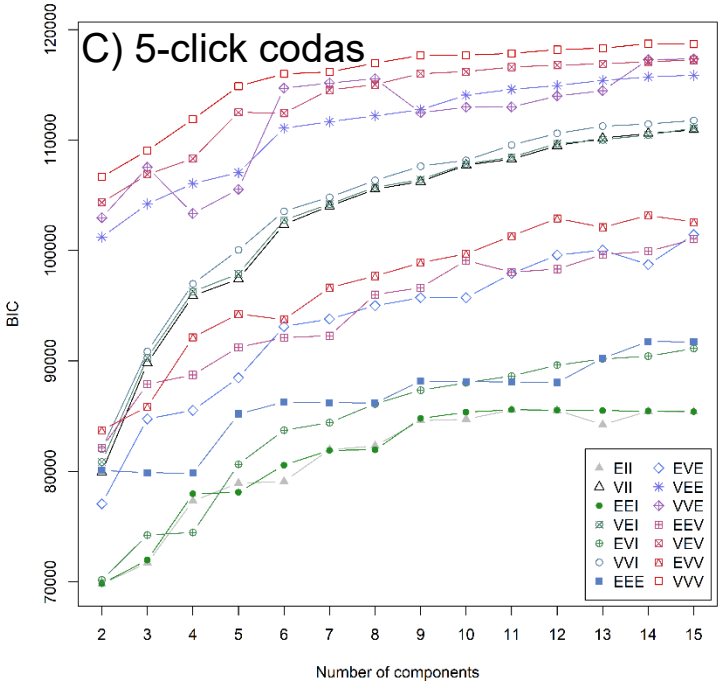
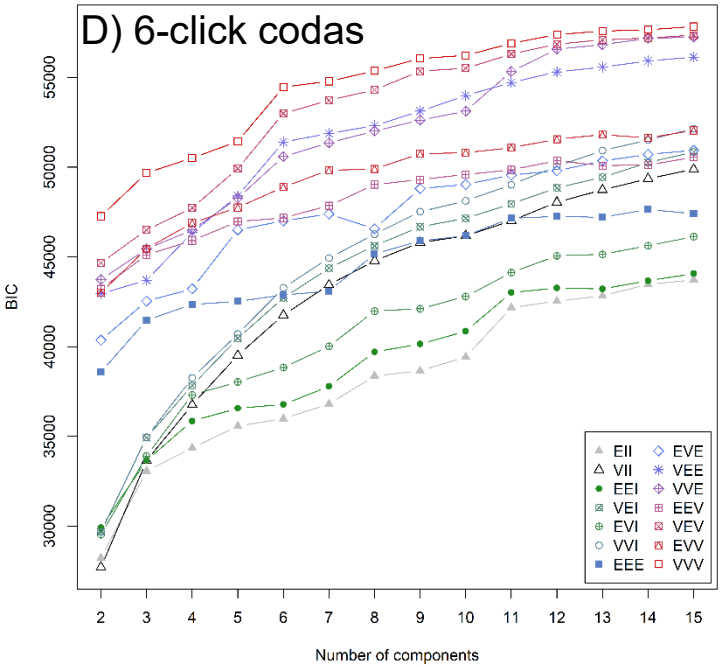


Figure S4. Minimum (A/B) and maximum (C/D) clan spatial overlap vs. coda usage similarity, calculated using identity codas (sim_{IDbt}; left panels) or non-identity codas ($sim_{nonIDbt}$; right panels). See Figure 4 for additional details. Like mean spatial overlap for identity codas (Figure 4), the maximum spatial overlap results for identity codas are significant but the minimum spatial overlap results are not. The results for non-identity codas are not significant in all cases. For all clan spatial overlap calculations, the Mantel test matric correlation coefficient is more negative for identity codas than for non-identity codas.


A) 3-click codas
BIC
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
EVE
VEE
VVE
EEV
VEV
EVV
VVV


B) 4-click codas
BIC
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
EVE
VEE
VVE
EEV
VEV
EVV
VVV


C) 5-click codas
BIC
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
EVE
VEE
VVE
EEV
VEV
EVV
VVV


D) 6-click codas
BIC
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
EVE
VEE
VVE
EEV
VEV
EVV
VVV

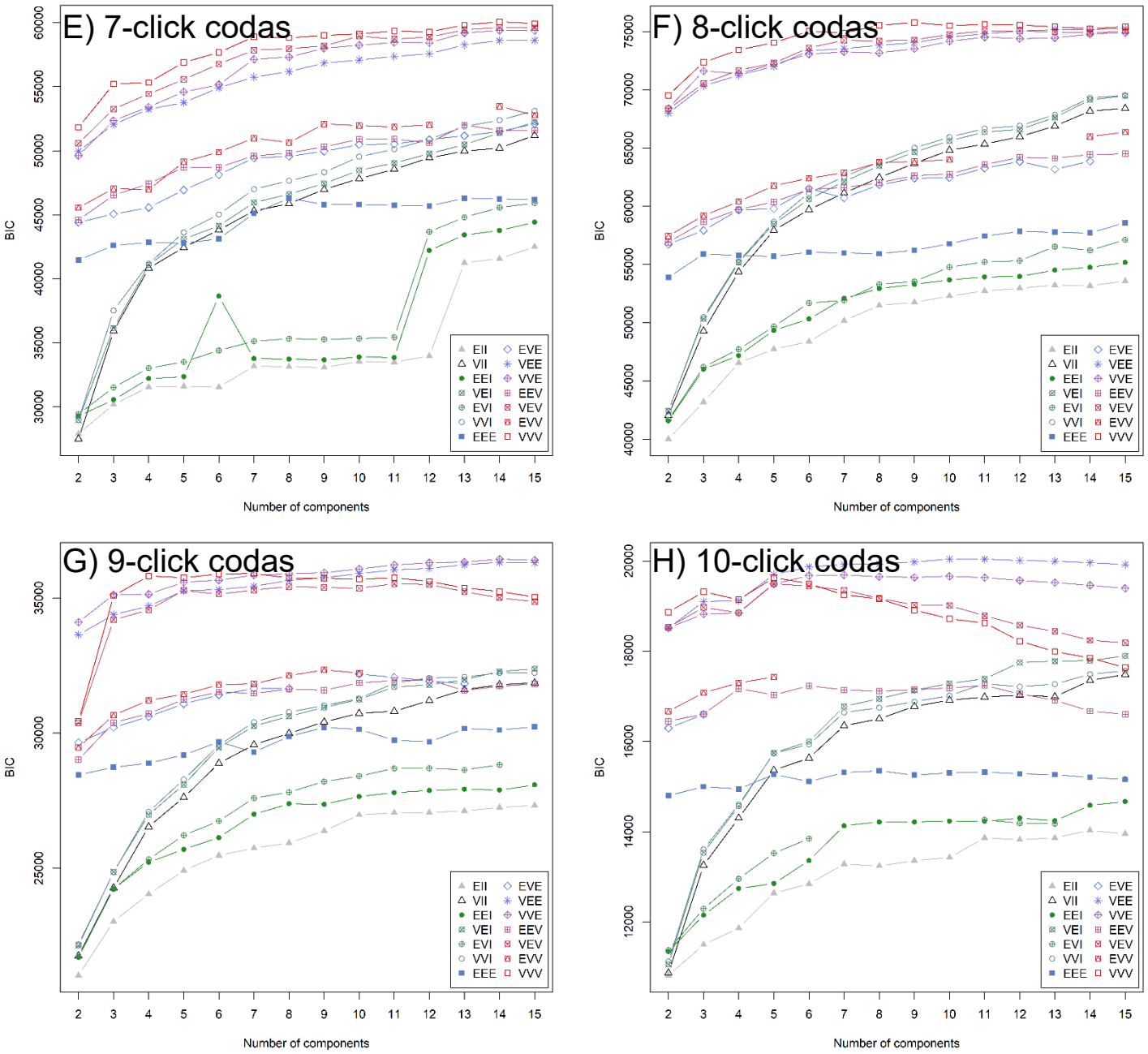


Figure S5. Results of preliminary data exploration using the 'mclust' R package (3). Plots show the BIC value (y-axis) for each of 14 models (panel insets) (2) when fitting 2 to 15 mixture components (x-axis) to 3- to 10-click codas (panels A–H, respectively). A higher BIC indicates a better fit.

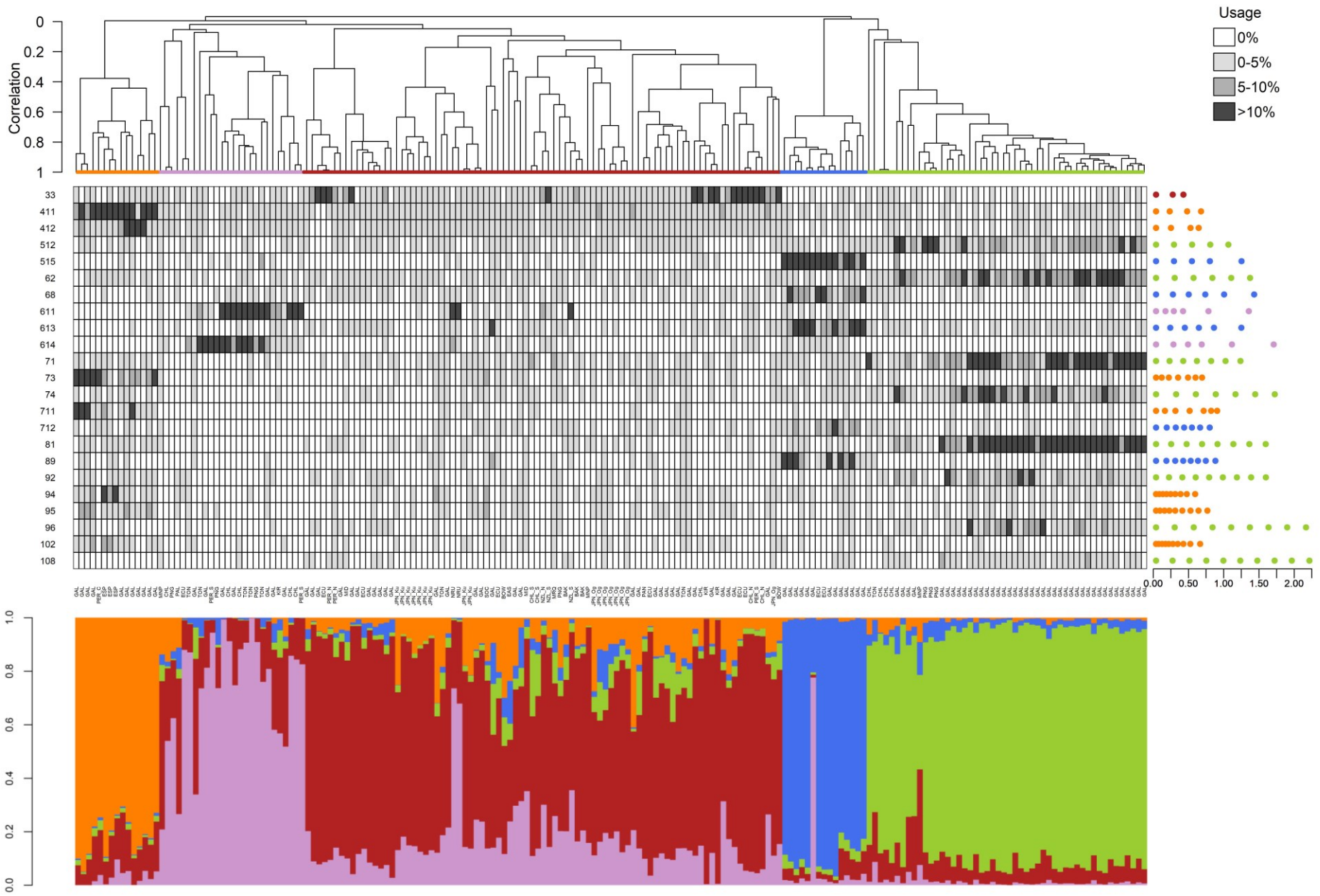


Figure S6. Example 5-clan tree with identity coda types. Colored identity clades correspond to the Palindrome (orange), Four-Plus (pink), Large Short (red), Plus-One (blue), and Regular (green) clans. See Figure 2 for additional details, Table S6 for identity coda type codes/names, and Table S1 for region abbreviations. This tree was constructed using *critfact*=12 and *minrep*=15 in IDcall (1).

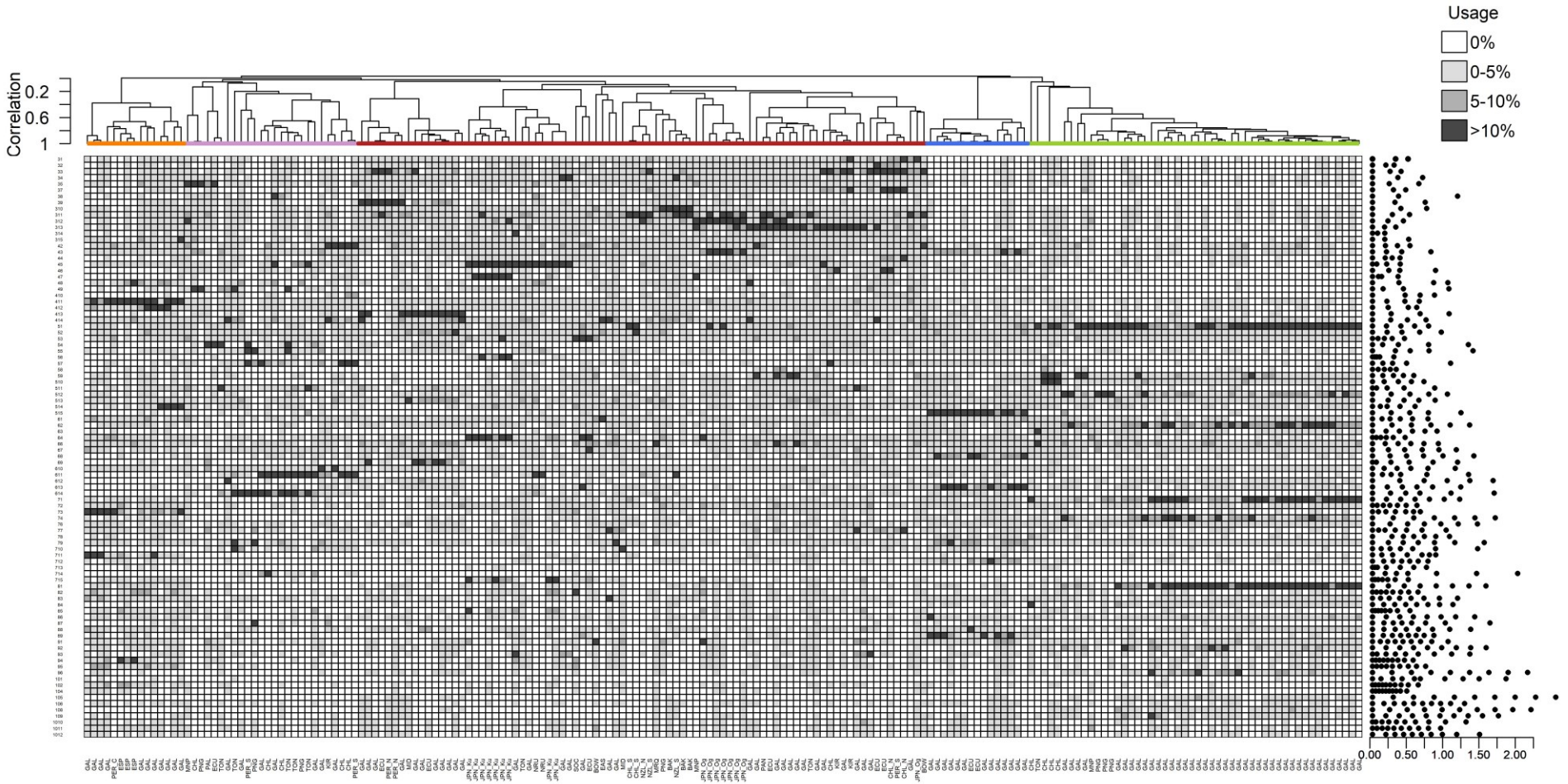


Figure S7. Example 5-clan tree with all (identity and non-identity) coda types. Colored identity clades correspond to the Palindrome (orange), Four-Plus (pink), Large Short (red), Plus-One (blue), and Regular (green) clans. See Figure 2 for additional details and Table S1 for region abbreviations. This tree was constructed using *critfact*=12 and *minrep*=15 in IDcall (1).

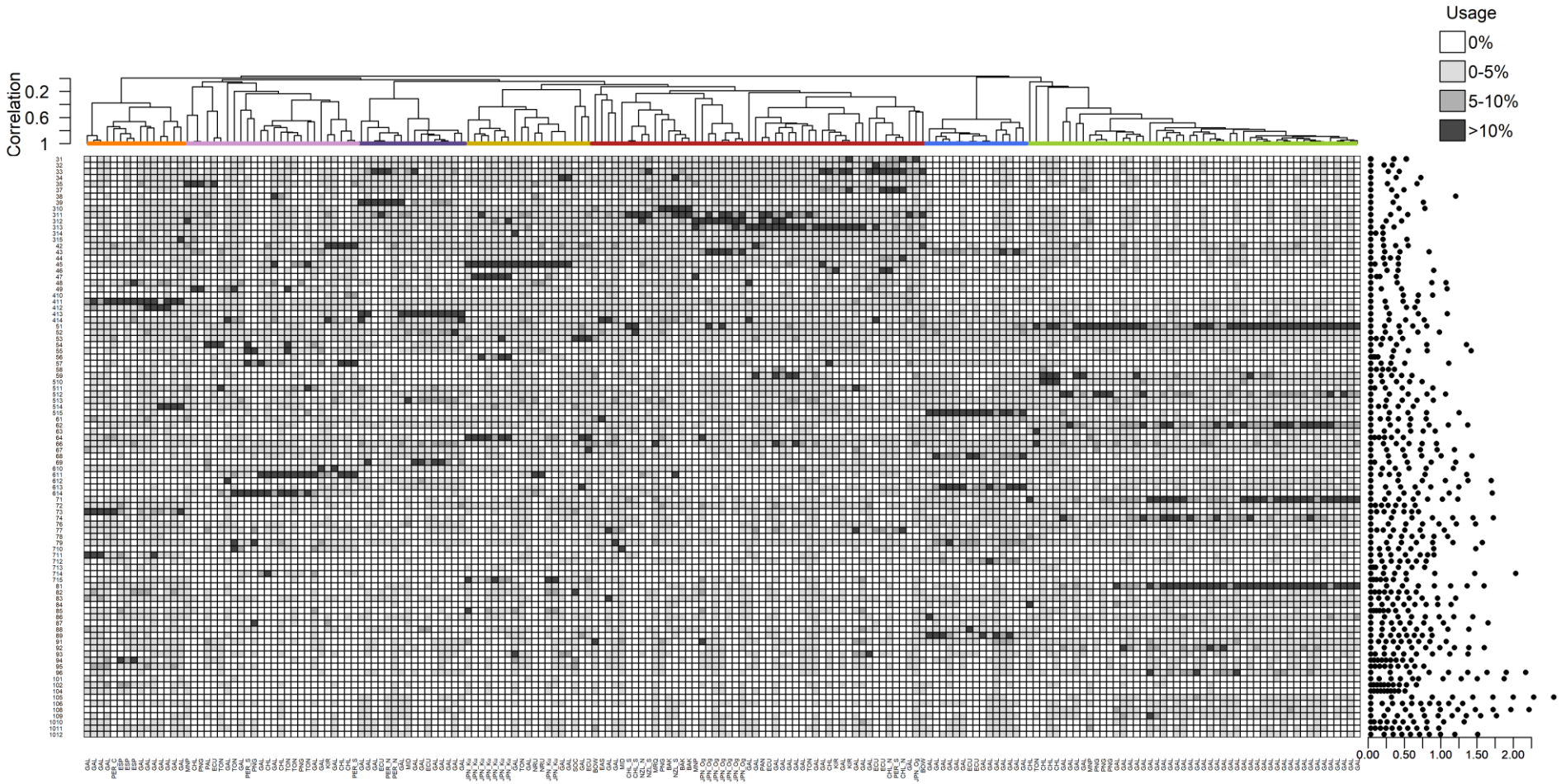


Figure S8. Example 7-clan tree with all (identity and non-identity) coda types. Colored identity clades correspond to the Palindrome (orange), Four-Plus (pink), Slow Increasing (purple), Rapid Increasing (gold), Short (red), Plus-One (blue), and Regular (green) clans. See Figure 2 for additional details and Table S1 for region abbreviations. This tree was constructed using *critfact*=5 and *minrep*=15 in IDcall (1).

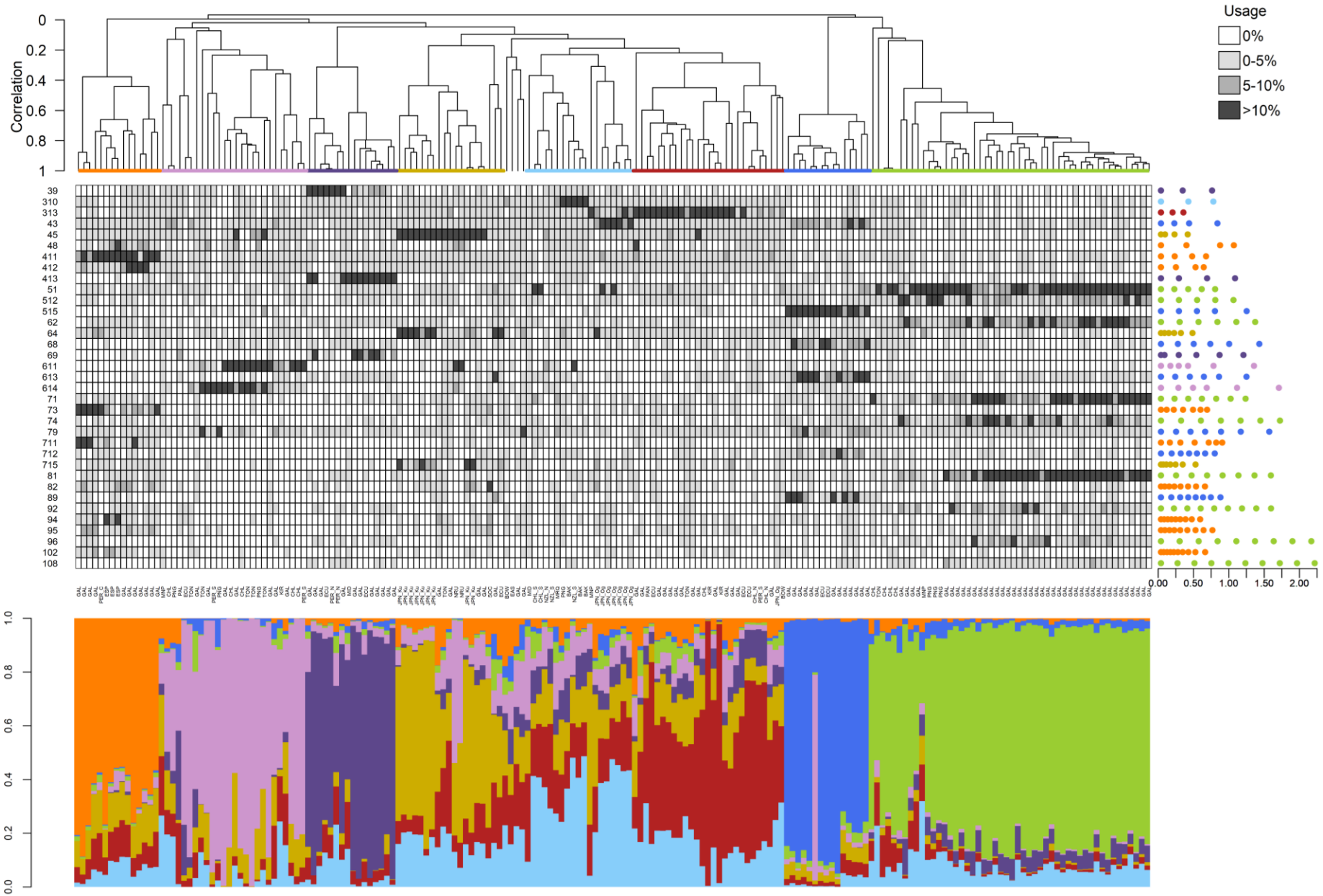


Figure S9: Example 8-clan tree with identity coda types. Colored identity clades correspond to the Palindrome (orange), Four-Plus (pink), Slow Increasing (purple), Rapid increasing (gold), Short West (light blue), Short East (red), Plus-One (dark blue), and Regular (green) clans. Four repertoires (two recorded off the Galápagos Islands, one recorded off SGaan Kinghlas-Bowie Seamount, and one recorded off Easter Island) are outliers (i.e., not assigned to a clan in this tree). See Figure 2 for additional details, Table S7 for identity coda type codes/names, and Table S1 for region abbreviations. This tree was constructed using *critfact*=6 and *minrep*=15 in IDcall (1).

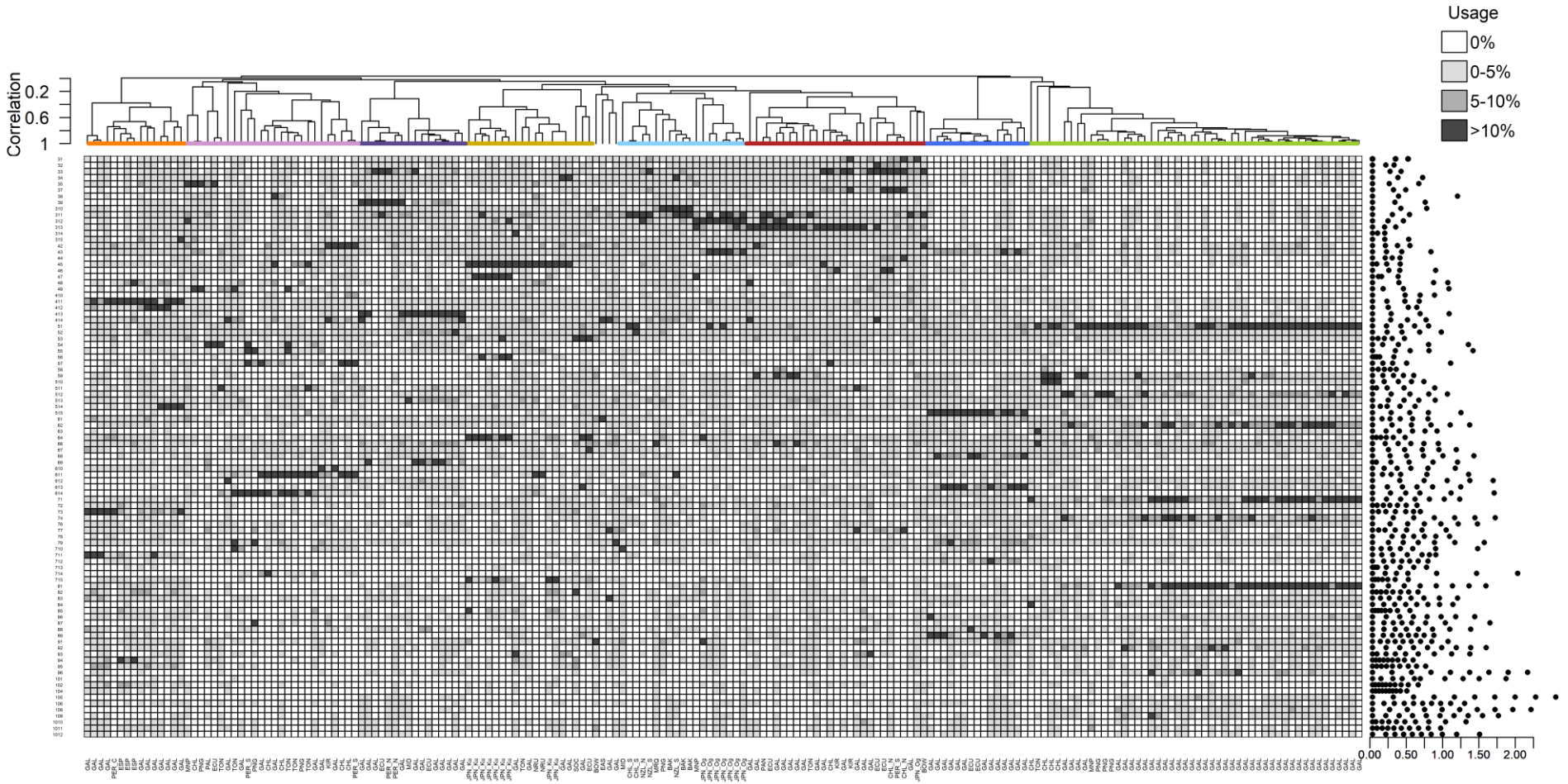


Figure S10. Example 8-clan tree with all (identity and non-identity) coda types. Colored identity clades correspond to the Palindrome (orange), Four-Plus (pink), Slow Increasing (purple), Rapid Increasing (gold), Short West (light blue), Short East (red), Plus-One (blue), and Regular (green) clans. Four repertoires (two recorded off the Galápagos Islands, one recorded off SGaan Kinghlas-Bowie Seamount, and one recorded off Easter Island) are outliers (i.e., not assigned to a clan in this tree). See Figure 2 for additional details and Table S1 for region abbreviations. This tree was constructed using *critfact*=6 and *minrep*=15 in IDcall (1).

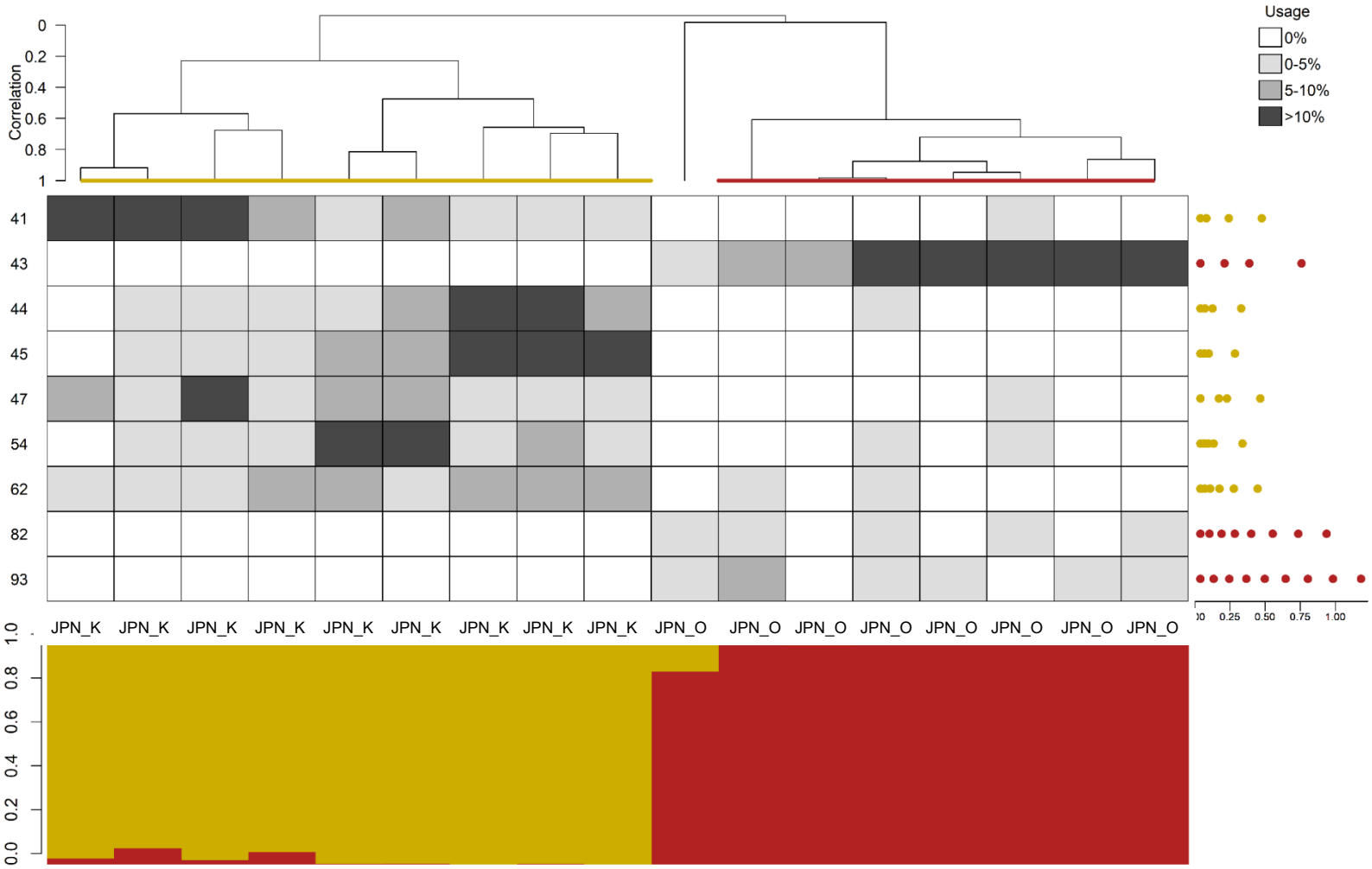


Figure S11. Example tree with identity coda types showing the division of Kumano coast (gold; JPN_K) and Ogasawara Islands (red; JPN_O) repertoires from Japan. This tree was created using *critfact*=15 and *minrep*=5 and the division of repertoires into clans was replicated at very high levels of *critfact* (e.g., *critfact*=310). The trends in coda usage seen here mirror those reported in Amano et al. (13), with Kumano coast whales making identity codas with shorter durations than Ogasawara Islands whales (see Figure 4 in Amano et al. (13)). See Figure 2 for additional details.

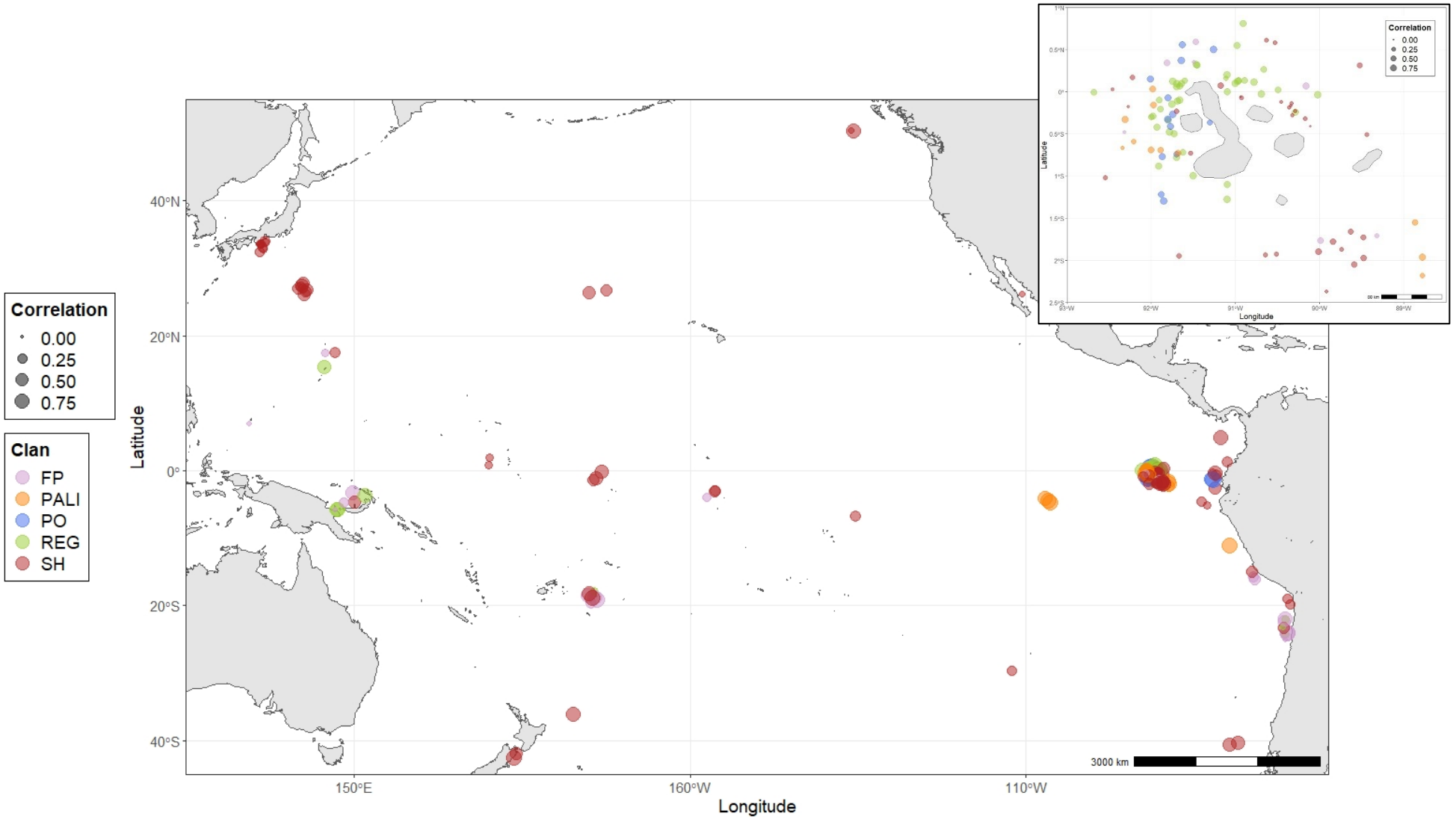


Figure S12. Composite map of sperm whale clan distributions across 23 regions in the Pacific Ocean using the 5-clan tree clan designations. Galápagos Islands inset is shown in the top right corner. See Figure 1 for additional details. Clan abbreviations are: FP=Four-Plus, PALI=Palindrome, PO=Plus-One, REG=Regular, and SH=Large Short. Note that the point size correlation scales differ for the composite map and the Galápagos Islands inset.

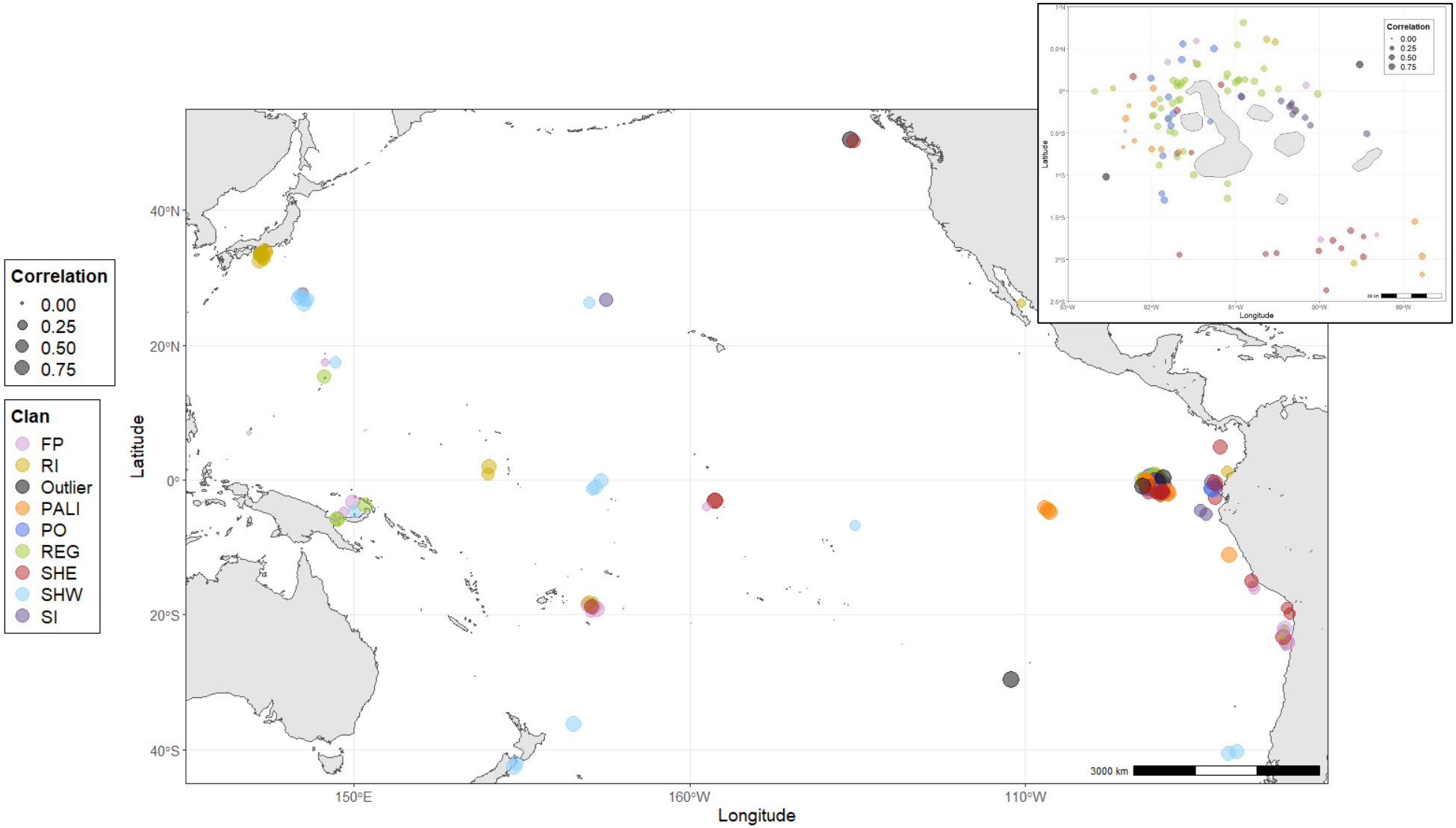


Figure S13. Composite map of sperm whale clan distributions across 23 regions in the Pacific Ocean using the 8-clan tree clan designations. Galápagos Islands inset is shown in the top right corner. See Figure 1 for additional details. Clan abbreviations are: FP=Four-Plus, RI=Rapid Increasing, Outlier (i.e., a repertoire that was not assigned to a clan), PALI=Palindrome, PO=Plus-One, REG=Regular, SHE=Short East, SHW=Short West clan, and SI=Slow Increasing. Note that the point size correlation scales differ for the composite map and the Galápagos Islands inset.

Table S1. Extracted coda and repertoire information for each region. All 3–10-click codas (third column) were included in the call classification stage of IDcall, but only repertoires with at least 25 codas (fourth column) were included in the hierarchical clustering stage. The 'Year(s)' and 'Number of repertoires' columns are based only on repertoires with at least 25 codas, which is why some years present in Table S5 are not present here.

Region (abbreviation)	Total number of codas (2–29 clicks)	Number of 3–10 click codas (% of total)	Number of 3–10 clicks from repertoires with ≥25 codas (% of total)	Year(s)	Number of repertoires
Baker Island (BAK)	278	272 (97.8%)	272 (97.8%)	1992	3
SGaan Kinghlas-Bowie Seamount (BOW)	133	126 (94.7%)	126 (94.7%)	2011 2013	1 1
Northern Chile (CHL_N)	5,538	5,523 (99.7%)	5,523 (99.7%)	1993 2000	1 11
Southern Chile (CHL_S)	183	183 (100%)	183 (100%)	1993	2
Easter Island (EAS)	90	90 (100%)	90 (100%)	1993	1
Ecuador (ECU)	779	774 (99.4%)	774 (99.4%)	1985 1991 1993	1 5 3
Equatorial South Pacific (ESP)	263	237 (90%)	215 (81.7%)	2000	3
Galápagos Islands (GAL)	10,959	10,507 (95.9%)	10,232 (93.4%)	1978 1985 1987 1989 1991 1995 1999 2000 2013 2014	1 19 24 11 3 8 10 1 13 14
Kumano coast of Japan (JPN_K)	836	829 (99.2%)	768 (91.9%)	2007 2008	6 3
Ogasawara Islands of Japan (JPN_O)	985	846 (85.9%)	697 (70.8%)	2006 2007 2008	1 2 5
Kiribati (KIR)	511	511 (100%)	511 (100%)	1992	3
Midway Atoll (MID)	250	239 (95.6%)	239 (95.6%)	2013 2017	1 1
Mariana Islands (MNP)	278	164 (59.0%)	149 (53.6%)	2007	3
Marquesas Islands (MRQ)	61	60 (98.4%)	60 (98.4%)	2000	1
Nauru (NRU)	89	88 (98.9%)	88 (98.9%)	2001	2
Northern New Zealand (NZL_N)	44	44 (100%)	44 (100%)	1993	1
Southern New Zealand (NZL_S)	157	156 (99.4%)	96 (61.1%)	1992 2010	1 1
Palau (PAL)	34	31 (91.2%)	31 (91.2%)	2012	1
Panama (PAN)	191	191 (100%)	191 (100%)	1992	1

Peru (PER)	672	665 (99.0%)	665 (99.0%)	1993	6
Papua New Guinea (PNG)	450	443 (98.4%)	443 (98.4%)	2001	7
Sea of Cortez (SOC)	207	202 (97.6%)	201 (97.1%)	1999	1
Tonga (TON)	1,249	1,248 (99.9%)	1,232 (98.6%)	1992 2003	1 7
Total	**24,237**	**23,429 (96.7%)**	**22,829 (94.2%)**	**21**	**191**

Table S2. Summary of clans and identity codas in the 7-clan tree scenario. For each repertoire, we calculated the correlation between the coda type usages of the repertoire and the median usages of the clan; the within-clan correlation was calculated by averaging these values (1). For each identity coda type, we list the number of clicks, the numeric code (used in Figure S8), and the type name (following conventions in (1, 18)). Briefly, type names refer to the overarching rhythmic pattern of clicks in each coda regardless of total duration, which is why multiple numeric codes can be linked to one type. 'R' stands for 'Regular' (i.e., all ICIs are approximately equal), '+' denotes an extended pause between clicks, 'D' stands for 'Decreasing' (i.e., ICIs become shorter throughout the coda), and 'I' stands for 'Increasing' (i.e., ICIs become longer throughout the coda).

Clan name	**Number of repertoires**	**Within-clan correlation (mean ± S.D.)**	**Number of identity codas**	**Identity coda types**		
				Number of clicks	*Numeric code*	*Type names*
Palindrome	15	0.699 ± 0.197	9	4 7 8 9 10	48, 411, 412 73, 711 82 94, 95 102	1+1++2, 2+2, 2+2 3+1+3, 3+1+3 8I 9I, 9I 10I
Four-Plus	26	0.445 ± 0.314	2	6	611, 614	4+1++1, 4+1++1
Slow Increasing	16	0.730 ± 0.175	3	3 4 6	39 413 69	2+1 4I 6I
Rapid Increasing	19	0.652 ± 0.185	2	4 7	45 715	4I 7I
Short	50	0.484 ± 0.205	1	3	313	3R
Plus-One	15	0.854 ± 0.123	6	5 6 7 8	515 68, 613 79, 712 89	4+1 5+1, 5+1 6+1, 7R 8R
Regular	50	0.763 ± 0.243	9	5 6 7 8 9 10	51, 512 62 71, 74 81 92, 96 108	5R, 5R 6R 7R, 7R 8R 9R, 9R 10R

Table S3. Number of detected clans per region and regional sampling effort under the 7-clan tree scenario. 'Number of repertoires' refers to repertoires with at least 25 codas and 'Number of codas' refers to codas with 3–10 clicks in those repertoires. Regions are ordered from most to least clans, then from most to least repertoires, and finally from most to least codas. See Table S1 for region abbreviations.

Region	Number of clans	Number of repertoires	Number of codas
GAL	7	104	10,232
ECU	5	9	774
TON	4	8	1,232
PER	4	6	665
CHL_N	3	12	5,523
PNG	3	7	443
MNP	3	3	149
KIR	2	3	511
MID	2	2	239
JPN_K	1	9	768
JPN_O	1	8	697
BAK	1	3	272
ESP	1	3	215
CHL_S	1	2	183
BOW	1	2	126
NZL_S	1	2	96
NRU	1	2	88
EAS	1	1	90
SOC	1	1	201
PAN	1	1	191
MRQ	1	1	60
NZL_N	1	1	44
PAL	1	1	31

Table S4. Clan spatial overlap values for the seven Pacific Ocean clans. The matrix gives the proportion of the row clan's repertoires that were recorded within 1,000 km of at least one of the column clan's repertoires. Note that the matrix is asymmetric, as a clan found in only one region might overlap completely with a clan that spans the ocean, while the inverse is not true (e.g., Plus-One/Short clans). For each pair of clans, minimum spatial overlap was calculated as the minimum of the two values, mean spatial overlap was calculated as the average of the two values, and maximum spatial overlap was calculated as the maximum of the two values. Without an *a priori* reason to expect minimum or maximum spatial overlap to be more/less appropriate, we have focused on mean spatial overlap in the main text. However, we include results from analyses using minimum and maximum spatial overlap in Figure S4.

Clan	Four-Plus	Rapid Increasing	Palindrome	Plus-One	Regular	Short	Slow Increasing
Four-Plus	1	0.500	0.385	0.308	0.885	0.962	0.308
Rapid Increasing	0.368	1	0.263	0.316	0.316	0.842	0.316
Palindrome	0.800	0.733	1	0.733	0.733	0.800	0.800
Plus-One	1	1	1	1	1	1	1
Regular	1	0.840	0.820	0.820	1	1	0.820
Short	0.580	0.580	0.400	0.400	0.520	1	0.420
Slow Increasing	0.938	0.938	0.938	0.938	0.938	1	1

Table S5. Regional recording metadata. See Table S1 for region abbreviations. Asterisks denote regions for which codas had previously been extracted and described, although additional codas from some of those regions were extracted in the present study. Citations with information on the recording equipment and field methods are provided in the 'Recording details' column. If citations were unavailable, coauthors provided the relevant details (see 'Coauthor(s)' column for corresponding initials). Codas from some regions, such as BOW, were recorded using stationary, bottom mounted hydrophones which enabled precise localization for each repertoire. In other regions (e.g., JPN_K, JPN_O), recording locations were given as ranges of latitude and longitude (spanning 0.1 to 0.5 decimal degrees). No GPS was available for a small number of NZL_S codas recorded in 1992 (n=36) off Kaikōura, so they were assigned the average GPS of codas recorded off Kaikōura in 2009 and 2010 by MF. Different recording systems were used across regions and years, but this does not affect our measurements of the temporal patterns of codas. By virtue of the highly collaborative nature of this project, different analysts marked codas for different regions, which could introduce some systematic variability into the results. However, most of the codas (88.2%) were marked by one of four coauthors who were trained to mark in the same lab (that of HW: LR (45.5%), TAH (19.7%), MC (12.1%), and LW (10.8%)), which should reduce the magnitude of this issue.

Region	Recording year(s)	Recording details	Funders	Coauthor(s)
BAK*	1992	(18)	National Geographic Society (NGS), Natural Sciences and Engineering Research Council of Canada (NSERC), Cetacean Society International (CSI)	LR, LW, HW
BOW	2011, 2013	(19)	Species At Risk Program of Fisheries and Oceans Canada	TD-V, JP
CHL_N*	1993, 2000	(10, 18)	NGS, NSERC, CSI	LR, LW, HW
CHL_S*	1993	(18)		LW, HW
EAS*	1993	(10, 18)		LR, LW, HW
ECU*	1985, 1991, 1993	(10, 18)		TAH, LR, LW, HW
ESP	2000	(20)	Ocean Alliances Voyage of the Odyssey (OAVO) funders (particularly The Canadian Whale Institute, The Laffey McHugh Foundation, the Pacific Life Foundation, the Marisla Foundation, and the Summit Foundation)	CJ, IK, RP, AR
GAL*	1978, 1985, 1987, 1989, 1991, 1995, 1999, 2000, 2013, 2014	(10, 11, 18, 20, 21)	NGS, NSERC, CSI, Animal Behavior Society, OAVO funders	MC, TAH, LR, CJ, IK, RP, AR, LW, HW
JPN_K*	2004, 2007, 2008	(13)	JSPS KAKENHI (grant number JP18570010)	MA
JPN_O*	1994, 1995, 1996, 2006, 2007, 2008			
KIR*	1992	(18)	NGS, NSERC, CSI	LW, HW
MID	2013, 2017	(22)	Pacific Islands Fisheries Science Center, NOAA Fisheries Office of Protected Resources, NOAA Fisheries Office of Science and Technology, Chief of Naval Operation Environmental Readiness Division and Pacific Fleet, Bureau of Ocean Energy Management	YB, KM, EO
MNP	2007	(23, 24)	U.S. Navy (NavFac Pacific/PacFleet)	OA, EF, CAH-W, TN
MRQ	2000	(20)	OAVO funders	CJ, IK, RP, AR
NRU	2001			
NZL_N*	1993	(18)	NGS, NSERC, CSI	LW, HW

NZL_S	1992, 1999, 2000, 2006, 2007, 2009, 2010, 2013, 2015	(15, 25–29)	Te Papa Atawhai, New Zealand Whale and Dolphin Trust, University of Otago	SD, MF, JG, MG, LH, ES
PAL	2012	Recordings were made from a 50 m survey vessel (*Endless Summer*) using a towed hydrophone array. The array consisted of a 70 m cable with two hydrophone elements (Reson TC4013) spaced 3 m apart. Both array channels were passed through a tunable high pass filter (Ecologic HP/27 ST Magrec Stereo Monitor Box) to reduce water and vessel noise. Corner frequencies of the high-pass filter were set between 500 Hz and 1 kHz depending on noise, and the gain was set between 10 to 20 dB. The signal was split from the high-pass filter and fed into a PC digital interface (MOTU Traveler) and multi-track digital recorder (Tascam DR-680). Files were recorded at a 192 kHz sampling rate, 24-bit resolution.	Whales Alive Australia, Australian Marine Mammal Centre	OA, EF, CAH-W, TN
PAN*	1992	(18)	NGS, NSERC, CSI	LW, HW
PER*	1993			TAH, LW, HW
PNG	2001	(20)	OAVO funders	CJ, IK, RP, AR
SOC	1999			
TON	1992, 2003	(18) Recordings in 2003 were made from a 15.3 m sailing vessel (*RV Catalyst*) using a Sony TCD-D10 Pro II DAT recorder and a 150 m custom-made towed hydrophone array (see (30) for array construction details). The array elements were EDO EC65 cylindrical elements, each amplified with a 40 dB low noise preamplifier, with on-board high pass filtering at 1592 Hz (-6 dB/octave) to reduce flow noise. Elements were spaced 5 m apart. The hydrophone array had a frequency response of 1.5 kHz–40kHz ± 4 dB. The recorder was within ± 1 dB from 20 Hz to 22 kHz. Recordings were made at a sampling rate of 48 kHz, 16 bits. Sperm whale clicks were detected in recordings using ‘Rainbow Click’ (31). Codas were then identified and measured manually.	NGS, NSERC, CSI	RA, SD, ES, LW, HW

Table S6. Summary of clans and identity codas in an exemplar 5-clan tree (created using *critfact*=12 and *minrep*=15). This table is like Table S2 except the 'Large Short' clan encompasses the Short, Rapid Increasing, and Slow Increasing clans from Table S2 and some identity codas have changed. See Table S2 for additional information.

Clan name	**Number of repertoires**	**Within-clan correlation (mean ± S.D.)**	**Number of identity codas**	**Identity coda types**		
				Number of clicks	*Numeric code*	*Type names*
Palindrome	15	0.699 ± 0.197	7	4 7 9 10	411, 412 73, 711 94, 95 102	2+2, 2+2 3+1+3, 3+1+3 9I, 9I 10I
Four-Plus	26	0.445 ± 0.314	2	6	611, 614	4+1++1, 4+1++1
Large Short	85	0.367 ± 0.210	1	3	33	1+2
Plus-One	15	0.854 ± 0.123	5	5 6 7 8	515 68, 613 712 89	4+1 5+1, 5+1 7R 8R
Regular	50	0.763 ± 0.243	8	5 6 7 8 9 10	512 62 71, 74 81 92, 96 108	5R 6R 7R, 7R 8R 9R, 9R 10R

Table S7. Summary of clans and identity codas in an exemplar 8-clan tree (created using *critfact*=6 and *minrep*=15). This table is identical to Table S2 except the Table S2 Short clan is divided into 'Short West' and 'Short East' here. Additionally, four repertoires that were assigned to the Short clan in the 7-clan tree were not assigned to a clan in the 8-clan tree. See Table S2 for additional information.

Clan name	**Number of repertoires**	**Within-clan correlation (mean ± S.D.)**	**Number of identity codas**	**Identity coda types**		
				Number of clicks	*Numeric code*	*Type names*
Palindrome	15	0.699 ± 0.197	9	4 7 8 9 10	48, 411, 412 73, 711 82 94, 95 102	1+1++2, 2+2, 2+2 3+1+3, 3+1+3 8I 9I, 9I 10I
Four-Plus	26	0.445 ± 0.314	2	6	611, 614	4+1++1, 4+1++1
Slow Increasing	16	0.730 ± 0.175	3	3 4 7	39 413 69	2+1 4I 6I
Rapid Increasing	19	0.652 ± 0.185	2	4 7	45 715	4I 7I
Short West	19	0.608 ± 0.187	1	3	310	3R
Short East	27	0.625 ± 0.166	1	3	313	3R
Plus-One	15	0.854 ± 0.123	6	5 6 7 8	515 68, 613 79, 712 89	4+1 5+1, 5+1 6+1, 7R 8R
Regular	50	0.763 ± 0.243	9	5 6 7 8 9 10	51, 512 62 71, 74 81 92, 96 108	5R, 5R 6R 7R, 7R 8R 9R, 9R 10R

Table S8. Dendrogram features as *critfact* varied from 3 to 20 and *minrep* was kept constant at 15.

critfact	Number of clans	Clan names	Total number of identity codas	Number of repertoires not assigned to a clan
3	8	Palindrome, Four-Plus, Slow Increasing, Rapid Increasing, Short West, Short East, Plus-One, Regular	34	4
4			35	
5	7	Palindrome, Four-Plus, Slow Increasing, Rapid Increasing, Short, Plus-One, Regular	32	0
6	8	Palindrome, Four-Plus, Slow Increasing, Rapid Increasing, Short West, Short East, Plus-One, Regular	33	3
7	4	Palindrome, Slow Increasing, Plus-One, Regular	25	95
8			24	
9			23	
10				
11				
12	5	Palindrome, Four-Plus, Large Short, Plus-One, Regular	23	0
13			24	
14			23	
15				
16			21	
17				
18				
19				
20			20	

SI References

1. T. A. Hersh, S. Gero, L. Rendell, H. Whitehead, Using identity calls to detect structure in acoustic datasets. *Methods Ecol. Evol.* **2021**, 1–11 (2021).
2. A. Punzo, P. D. McNicholas, Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical J.* **58**, 1506–1537 (2016).
3. L. Scrucca, M. Fop, T. B. Murphy, A. E. Raftery, mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8**, 289 (2016).
4. G. Schwarz, Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
5. E. Pante, B. Simon-Bouhet, marmap: a package for importing, plotting and analyzing bathymetric and topographic data in R. *PLoS One* **8**, e73051 (2013).
6. C. Amante, B. W. Eakins, ETOPO1 arc-minute global relief model: procedures, data sources and analysis (2009).
7. D. W. Rice, “Sperm whale. Physeter macrocephalus Linnaeus, 1758” in *Handbook of Marine Mammals*, S. H. Ridgway, R. J. Harrison, Eds. (Academic Press, 1989), pp. 177–233.
8. H. Whitehead, *Sperm whales: social evolution in the ocean* (University of Chicago Press, 2003).
9. R. J. Hijmans, E. Williams, C. Vennes, Geosphere: spherical trigonometry (2015).
10. L. E. Rendell, H. Whitehead, Vocal clans in sperm whales (Physeter macrocephalus). *Proc. R. Soc. London Ser. B Biol. Sci.* **270**, 225–31 (2003).
11. M. Cantor, H. Whitehead, S. Gero, L. Rendell, Cultural turnover among Galápagos sperm whales. *R. Soc. Open Sci.* **3**, 160615 (2016).
12. S. T. Piantadosi, H. Tily, E. Gibson, Word lengths are optimized for efficient communication. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 3526–3529 (2011).
13. M. Amano, A. Kourogi, K. Aoki, M. Yoshioka, K. Mori, Differences in sperm whale codas between two waters off Japan: possible geographic separation of vocal clans. *J. Mammal.* **95**, 169–175 (2014).
14. A. Ravignani, G. Madison, The paradox of isochrony in the evolution of human rhythm. *Front. Psychol.* **8**, 1820 (2017).
15. M. G. de M. de C. Fernandes, “Acoustic analysis of slow click function and foraging in sperm whales (Physeter macrocephalus) off Kaikoura, New Zealand,” University of Canterbury. (2016).
16. D. E. Gaskin, Sperm whales in the western South Pacific. *New Zeal. J. Mar. Freshw. Res.* **7**, 1–20 (1973).
17. L. Rendell, H. Whitehead, A. Coakes, Do breeding male sperm whales show preferences among vocal clans of females? *Mar. Mammal Sci.* **21**, 317–322 (2005).
18. L. Weilgart, H. Whitehead, Group-specific dialects and geographic variation in coda repertoire in South Pacific sperm whales. *Behav. Ecol. Sociobiol.* **40**, 277–285 (1997).
19. A. Riera, R. A. Rountree, X. Mouy, J. K. Ford, F. Juanes, Effects of anthropogenic noise on fishes at the SGaan Kinghlas-Bowie Seamount Marine Protected Area in *Proceedings of Meetings on Acoustics 4ENAL*, (Acoustical Society of America, 2016), p. 10005.
20. C. Godard, R. Clark, I. Kerr, P. Teglberg Madsen, R. Payne, Preliminary report on the sperm whale data collected during the Voyage of the Odyssey. *IWC* (2003).
21. L. Sayigh, *et al.*, The Watkins Marine Mammal Sound Database: An online, freely accessible resource in *Proceedings of Meetings on Acoustics*, (2016), p. 040013.
22. Y. Barkley, E. M. Oleson, J. N. Oswald, E. C. Franklin, Whistle classification of sympatric false killer whale populations in Hawaiian waters yields low accuracy rates. *Front. Mar. Sci.* **6**, 645 (2019).
23. K. M. Waller, *et al.*, “Marine mammal and sea turtle survey and density estimates for Guam and the Commonwealth of the Northern Mariana Islands” (2007).
24. T. F. Norris, *et al.*, “An analysis of acoustic data from the Mariana Islands Sea Turtle and Cetacean Survey (MISTCS)” (2012).
25. J. Gordon, R. Leaper, F. G. Hartley, O. Chappell, Effects of whale-watching vessels on the surface and underwater acoustic behaviour of sperm whales off Kaikoura, New Zealand. *Sci. Res. Ser.* **52**, 1–64 (1992).
26. M. Guerra, *et al.*, Changes in habitat use by a deep-diving predator in response to a coastal earthquake. *Deep. Res. Part I Oceanogr. Res. Pap.*, 103226 (2020).
27. N. Jaquet, S. Dawson, L. Douglas, Vocal behavior of male sperm whales: Why do they click? *J. Acoust. Soc. Am.* **109**, 2254–2259 (2001).

28. B. S. Miller, A. Growcott, E. Slooten, S. M. Dawson, Acoustically derived growth rates of sperm whales (Physeter macrocephalus) in Kaikoura, New Zealand. *J. Acoust. Soc. Am.* **134**, 2438–2445 (2013).
29. M. Q. Rhinelander, S. M. Dawson, Measuring sperm whales from their clicks: Stability of interpulse intervals and validation that they indicate whale length. *J. Acoust. Soc. Am.* **115**, 1826–1831 (2004).
30. J. Barlow, S. Rankin, S. Dawson, A guide to constructing hydrophones and hydrophone arrays for monitoring marine mammal vocalizations. *NOAA Tech. Memo. NMFS* **417**, 1–36 (2008).
31. D. Gillespie, An acoustic survey for sperm whales in the Southern Ocean Sanctuary conducted from the RSV Aurora Australis. *Reports Int. Whal. Comm.* **47**, 897–907 (1997).