Genomic insights into the historical and contemporary demographics of the grey reef shark

Supplementary information

Materials and methods

SNP calling, filtering, SFS calculation & length recalibrations

De novo assembly and SNP calling were performed using the iPyRAD pipeline v. 0.7.28 (Eaton and Overcast, 2016). First, raw reads were quality checked: those with more than 5 low quality bases (Q < 20) were removed, adaptors were filtered, and only reads which after adaptor removal were >50 bp long were retained. A clustering threshold of 0.9 was used, the maximum number of SNPs per locus was set at 5, the minimum read depth for base calling was set to 6 and maximum read depth to 1000 (this was filtered more strictly by read depth in the following step). The pipeline produced 15,769 loci (averaging ~63.5-bp in length) containing a total of 42,939 SNPs (0-5 SNPs per locus). Then, read depth per locus and per individual was calculated and only loci in the core 90% of the distribution of mean read depth (9.05 < read depth < 37.8) were retained, eliminating potential paralogs and very low coverage data. All individuals had an average read depth above 10. Furthermore, SNPs with more than 10% missing data, more than two alleles, or heterozygosity of more than 0.5 were removed. The introduction of technical replicates was used to test for SNP reproducibility, and only SNPs which had identical genotypes in >99.5% of the 215 technical replicate pairs were retained. The technical

replicates were also used to estimate heterozygous discovery rate using the perl script repMatchStats.pl (available from https://github.com/z0on/2bRAD_denovo), which ranged from 0.8 to 0.98 (with most technical replicates pairs above 0.94). Finally, some individuals were removed due to excess heterozygosity (probably due to laboratory contamination or cryptic hybridization). The final filtered dataset used as the starting point for further analyses contained 7,952 loci with a total of 14,935 SNPs.

To standardize genetic diversity indices per site, as well as accurately estimate the timing of inferred demographic processes, the total number (i.e. length) of all bases sequenced for this study needed to be determined. Because the SNP calling procedures used with restriction-site associated sequencing are not typically applied to monomorphic loci, assumptions were made using ratios of various types of loci in the initial (unfiltered) dataset to back-calculate the total number of sequenced bases. The overarching assumption, applied at multiple levels of our analyses when justified (and described in each case below) was that the proportion of monomorphic to polymorphic loci, or polymorphic to total loci (polymorphic plus monomorphic), should remain constant throughout filtering. Considering that 1,192 of the initial 15,769 sequenced loci were monomorphic and thus removed at the beginning of SNP calling, it was assumed that a similar proportion of monomorphic loci to the 7,952 (polymorphic) loci that passed quality filtering (see Materials and Methods section 2.1) could have passed as well. Similarly, 791 singleton loci (loci with only one SNP) were filtered out during the technical replicate step, and we assume a similar proportion of those singleton loci would

have actually been acceptable monomorphic loci. The final number of loci could then be calculated as:

filtered loci + (initial monomorphic loci ÷ initial loci) • filtered loci +

(initial singleton loci ÷ initial loci) • singletons removed at technical replicate step

while the total number of base pairs sequenced (given the average of ~63.48 per locus) was calculated as:

(7,952+ (1,192 ÷ 15,769) • 7,952) + ((2,520 ÷ 15,769) • 791) • ~63.48 = 551,040

This overall total of 551,040 base pairs was then re-calibrated at every sampling location to account for varying levels of missing data at the population scale, as well as for each analysis that used different sets of assumptions about the independence of variable sites.

One way to account for varying levels of missing data across populations is to downwardly project their sample size (i.e. number of haplotypes surveyed) in SFS calculations (which otherwise cannot incorporate sites with missing data). Following projection, information from fewer samples in the population is then needed to determine which frequency bin a SNP falls into. This increases the number of segregating sites that can be analysed (i.e. decreases the amount of unusable missing data), which is recommended practice in SFS-based demographic history methods (Gutenkunst et al. 2009). For populations with sufficiently large sample sizes, the increase in statistical power from including more SNPs in downstream analyses is likely to outweigh any power lost by decreasing the number of samples considered (if the level of missing data is sufficiently high as well). Therefore, where possible, we downwardly projected the

sample size of each population considered in our genetic diversity and population history analyses to maximize the number of segregating sites, as well as increase the accuracy and informativeness of our frequency distributions. This increased the number of analysable segregating sites by hundreds to thousands in populations that were sufficiently sampled, as can be seen in Table S1 by comparing the S_{pop} and S_{thin} values with projection and without (in parentheses). Cocos (Keeling) had too few samples for thinning to be effective (i.e. including all 5 samples still yielded the most segregating sites), and the number analysable segregating sites in the Herald Cays population only increased marginally following the slight downwards projection (8 to 7 samples).

Following downwards projection, there may still be sites which do not have enough information (i.e. too much missing data) in a population to be included in the final SFS. The original number of segregating sites in the filtered dataset from which all of the population site frequency spectra were calculated (14,935) is therefore greater than the number of sites that can actually be included in the SFS of a given population. Since we assume that the initial ratio of polymorphic to total (polymorphic and monomorphic) sites remains constant, the loss of information about certain sites in a specific population means that its length of sequenced base pairs decreases proportionally as well. So, for the SFS of each population (or pair of populations for joint population history modelling), the recalibration of total sequenced base pairs was calculated as:

$$L \bullet (\Sigma SFS_{projected} \div S) = L_{pop}$$

where L is the total number of sequenced base pairs calculated for the entire dataset (551,040), S is the total number of SNPs in the entire (filtered) dataset (i.e. the number of sites in L that were polymorphic: 14,935), and $\Sigma SFS_{projected}$ is the number of sites in S that were observed in a sufficient number of individuals in the population (the downwardly-projected sample size) to be included in its SFS (including those from S that are monomorphic in this population). The recalibrated lengths for each population (L_{pop}) were between 507,982 and 547,203 (Table S1), and therefore only marginally different from the original L of 551,040. To report per site values, diversity indices for each population were divided by L_{pop}.

The other three analyses performed in this study operate under the simplifying assumption that all SNPs are independent (i.e. not in linkage disequilibrium). The primary way to conform to these assumptions when using restriction-site associated sequencing data (without a reference genome) is to only consider one SNP per locus. This kind of filtering, called thinning, minimizes the biasing effect of linkage disequilibrium between SNPs on the same locus. Following filtering to include only sites that were polymorphic within populations (or pair of populations for joint population history modelling) using VCFTOOLS, (Danecek et al. 2011) site frequency spectra with only one random SNP per locus were created using 'easySFS' (Overcast, 2018). However, because we assume that the initial ratio of polymorphic to total sites remains constant, this thinning alters total sequence length (which is important calibrating the timing of events in coalescent SFS-based population modelling). This loss of polymorphism means our total sequence length must decrease proportionally. Similar to

the equation used to calculate L_{pop} above, the total sequence length of each population was therefore recalibrated after thinning as:

$$L_{pop} \bullet (S_{thin}/S_{pop}) = L_{pop(thinned)}$$

where S_{pop} is the number of SNPs observed in the population before thinning (i.e. the number of polymorphic sites in $\Sigma SFS_{projected}$ from the previous recalibration) and S_{thin} is the number of SNPs remaining after filtering to include only one per locus (i.e. the number of loci with variable sites) in the population (Table S1). The site frequency spectra and final lengths calculated for each population (or pair of populations) were then used in our population history modelling as described in the main text.

The contemporary N_e analysis also requires the use of independent (unlinked) loci. However, this is not a coalescent- or SFS-based analysis, so neither site frequency spectra nor information about total sequenced base pairs is needed. In this analysis, data from each individual population was converted into GENEPOP (Raymond and Rousset, 1995) format using the 'radiator' package in R (Gosselin, 2017) and then analyzed using the option to only consider sites that on different chromosomes (or in this case, loci).

Plotting

All figures (including maps) were produced in R using the packages 'ggplot2' (Wickham, 2016), 'ggthemes' (Arnold, 2019), 'ggrepel' (Slowikowski, 2019), 'ggnewscale' (Campitelli, 2020), 'scales' (Wickham and Seidel, 2020), 'sp' (Bivand et al. 2013), 'raster' (Hijmans, 2019), 'marmap' (Pante and Simon-Bouhet, 2013), 'rnaturalearth' (South, 2017), 'sf' (Pebesma, 2018), 'mapview' (Appelhans et al. 2020), and 'patchwork' (Pedersen, 2020).

References

Appelhans T, Detsch F, Reudenbach C, Woellauer S (2020). *mapview: Interactive Viewing of Spatial Data in R*.

Arnold JB (2019). *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*.

Bivand RS, Pebesma E, Gomez-Rubio V (2013). *Applied spatial data analysis with R*, 2nd edn. Springer, NY.

Campitelli E (2020). *ggnewscale: Multiple Fill and Colour Scales in 'ggplot2'*.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* (2011). The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Eaton DAR, Overcast IA (2016). *ipyrad: interactive assembly and analysis of RADseq data sets — ipyrad documentation*.

Gosselin T (2017). *radiator: RADseq Data Exploration, Manipulation and Visualization using R*.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics* **5**: e1000695.

Hijmans RJ (2019). *raster: Geographic Data Analysis and Modeling*.

Overcast I (2018). *isaacovercast/easySFS*.

Pante E, Simon-Bouhet B (2013). marmap: A Package for Importing, Plotting and Analyzing Bathymetric and Topographic Data in R. *PLOS ONE* **8**: e73051.

Pebesma EJ (2018). Simple features for R: standardized support for spatial vector data. *R J* **10**: 439.

Pedersen TL (2020). *patchwork: The Composer of Plots*.

Raymond M, Rousset F (1995). GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism. *Journal of Heredity* **86**: 248–249.

Slowikowski K (2019). *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*.

South A (2017). *rnaturalearth: World Map Data from Natural Earth*.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag: New York.

Wickham H, Seidel D (2020). *scales: Scale Functions for Visualization*.

TABLE S1 List of model names, parameters, and parameter estimates from the best replicates of *Carcharhinus amblyrhynchos* two-population models (listed in order of increasing AIC). LogLik = log likelihood, AIC = Akaike Information Criterion, $\theta = 4N_{REF}\mu$ where N_{REF} is the ancestral population size and μ the mutation rate over the entire sequenced length. N_{AE} = ancestral population after expansion (only for "AE" models), N_1 = size of population 1 (Misool), N_2 = size of population 2 (North GBR), s = size of population 2 at time of split (for models allowing growth in population 2), T_{AE} = time of ancestral population size change, T_1 = time of split, T_2 = time at which the gene flow scenario changes (for SC, AM, and 2EP models), m_{12anc} and m_{21anc} = ancestral migration rate (from T_1 to T_2), m_{12} and m_{21} = contemporary migration rates. Population sizes are relative to N_{REF} and times are given in units of $2 \bullet N_{REF}$ generations. Migration rates are given as $M_{ij} = 2 \bullet N_{REF} \bullet m_{ij}$, where m_{ij} is the proportion of individuals in population i that is made up of migrants from population j at a given generation.

Model	LogLik	AIC	Δ_{AIC}	θ	N_{AE}	N_1	N_2	s	T_{AE}	T_1	T_2	m_{12anc}	m_{21anc}	m_{12}	m_{21}
SC_B	-467.16	948.32	0	82.58	–	17.54	31.97	0.14	–	9.29	1.16	–	–	2.8	0.75
IM_{AEB}	-466.74	949.48	1.16	87.58	14.92	23.19	26.99	0.01	5.36	0.89	0	–	–	1.02	3
SC_{AE}	-466.74	949.49	1.17	58.54	16.91	20.7	27.66	–	8.85	1.09	0.14	–	–	5.77	1.24
SC_{AEB}	-466.28	950.56	2.24	71.22	13.88	20.02	71.97	0.76	6.99	0.93	0.2	–	–	3.86	1.58
$2EP_B$	-466.77	951.53	3.22	120.76	–	16.77	20.21	0.01	–	3.29	0.61	8.44	96.72	1.37	4.24
IM_{AE}	-468.97	951.93	3.62	54.52	18.22	25.05	29.34	–	9.48	1.41	0	–	–	2.02	0.33
$2EP_{AE}$	-466.78	953.56	5.24	66.42	14.99	18.18	24.41	–	7.67	0.97	0.11	0.05	0.01	6.95	1.5
$2EP_{AEB}$	-465.97	953.95	5.63	527.24	0.73	3.04	4.61	0.04	1.87	1.15	0.18	0.8	0.08	13.24	9.05
AM_{AE}	-469.02	954.05	5.73	72.76	13.78	18.99	22.06	–	6.85	1.03	0	2.71	0.45	–	–
2EP	-469.1	954.2	5.89	72	–	23.22	13.74	–	–	7.79	0.19	1.9	0	0.83	3.68
IM_B	-471.62	955.25	6.93	66.48	–	18.27	26.25	0.26	–	9.47	0	–	–	2.91	0.06
AM_{AEB}	-469.19	956.38	8.07	73.81	13.62	20.1	52.37	0.99	6.73	0.99	0.01	2.7	0.74	–	–
AM_B	-473.06	960.12	11.81	220.42	–	12.15	98.24	0.4	–	1.59	0.07	22.41	2.71	–	–
SI_{AE}	-475.56	961.12	12.8	219.52	6.13	19.74	10.52	–	1.6	0.1	0	0	0	–	–
SI_{AEB}	-474.93	961.86	13.55	223.43	6.1	20.08	99.33	0.12	1.56	0.1	0	0	0	–	–
IM	-479.24	968.49	20.17	251.42	–	4.86	5.79	–	–	1.31	0	–	–	11.43	0
AM	-478.88	969.75	21.43	248.62	1	6.46	5.86	–	–	1.32	0.02	15	24.8	–	–
SC	-479.24	970.49	22.17	251.42	–	4.86	5.79	–	–	0.18	1.13	0	0	11.44	0
SI	-742.31	1490.61	542.3	612.9	–	7.94	7.49	–	–	0.07	–	–	–	–	–
SI_B	-741.96	1491.92	543.6	615.34	–	8.15	6	0.98	–	0.07	–	–	–	–	–

TABLE S2 Summary table of actual and projected sample sizes, total number of segregating sites, and total number of sequenced base pairs in thinned and un-thinned SFSs for each *C. amblyrhynchos* population (listed in increasing order of longitude). Un-thinned values were used for diversity index calculations while thinned values were used for population history modelling. *S* values in parentheses indicate the lower number of segregating sites that would have been kept without projection (for comparison) and come from the 'easySFS' preview outputs (SFS_preview_thin.txt and SFS_preview_all.txt for each population's thinned and un-thinned datasets in our analyses). The thinning procedure randomly selects one SNP per locus, so the total (thinned) *S* per population may vary by a couple of SNPs each time this is run with the same vcf file (as is done in our demographic modelling analyses). Cocos (Keeling) had too few samples for thinning to be effective to gain segregating sites for demographic modelling (i.e. including all 5 samples still yielded the most segregating sites). L_{pop} and $L_{pop(thinned)}$ were calculated using the formulae described above.

*Petit Astrolabe was projected to have 5971 SNPs at a sample size of both 92 and 90 in its un-thinned SFS, so the larger sample size of 92 was selected to retain more information. For its thinned SFS, however, a sample size projection to 90 yielded more segregating sites.

Population	Actual sample size (inds., *n*)	Un-thinned			Thinned		
		Projected size (*n*)	S_{pop}	L_{pop}	Projected size (*n*)	S_{thin}	$L_{pop(thinned)}$
Chagos	22 *(44)*	34	1824 (*1237*)	534511	34	1587 (*1044*)	465148
Cocos (Keeling)	5 *(10)*	10	1947 (*1947*)	517465	10	1647 (*1647*)	437732
Ningaloo	23 *(46)*	38	4976 (*3503*)	536466	38	3704 (*2601*)	399275
Rowley	24 *(48)*	42	5189 (*4086*)	539049	42	3870 (*3049*)	401958
Scott	24 *(48)*	42	5186 (*3951*)	532371	42	3831 (*2898*)	393210
Misool	23 *(46)*	40	5190 (*4017*)	537536	40	3837 (*2956*)	397425
North GBR	19 *(38)*	26	3937 (*1411*)	521929	26	2999 (*1074*)	397560
Herald Cays	8 *(16)*	14	2995 (*2948*)	545727	14	2404 (*2364*)	438086
South GBR	21 *(42)*	26	3566 (*1759*)	507982	26	2764 (*1369*)	393684
Chesterfield	40 *(80)*	72	5590 (*4223*)	544251	72	4098 (*3079*)	399027
Entrecasteaux	54 *(108)*	96	6018 (*3964*)	547203	96	4341 (*2843*)	394730
Northern Lagoon	51 *(102)*	88	5889 (*3657*)	547055	88	4267 (*2662*)	396381
Grand Astrolabe	46 *(92)*	80	5523 (*3328*)	544915	80	4080 (*2450*)	402546
Petit Astrolabe	51 *(102)*	92*	5971 (*3860*)	543292	90*	4284 (*2765*)	389816
Southern Lagoon	39 *(78)*	68	5291 (*3453*)	542923	68	3936 (*2537*)	403910
Walpole	27 *(54)*	48	4677 (*3593*)	540414	48	3561 (*2719*)	411445
Matthew	36 *(72)*	62	4457 (*3179*)	544103	62	3415 (*2430*)	416972
North GBR + Misool	19+23 (*38+46*)	26,40	6029	509643	26,40	4226	357727

TABLE S3 Summary of genetic diversity and effective population size (N_e) results for each *Carcharhinus amblyrhynchos* population (listed in increasing order of longitude). The four largest focal values from each analysis are in bold. Final N_e estimates from STAIRWAY PLOT are from the most recent coalescent time inferred, which depends on the number of sequences analysed in a given population (with information from a greater number of sequences, more coalescent events must be inferred, typically at more recent time points). The final value from the location with the fewest individuals analysed in STAIRWAY PLOT (8; Herald Cays) was inferred to be from 2,940 years ago, while the final values from several sampling locations with the most sequences analysed was inferred to be from just over 100 years ago. The estimates from LDNE are of a different kind. While STAIRWAY PLOT estimates "long-term" N_e (from coalescent models over long periods of time), LDNE estimates "short-term" N_e, essentially assessing the N_e of the previous generation (the gene pool that produced the individuals sampled). All upper 95% confidence interval limits estimated by LDNE were infinite except for at Matthew. See main text for further discussion of the timescales covered by our effective population size analyses.

Population	Diversity indices		STAIRWAY PLOT			LDNE (NeEstimator)		
	π	D	**Final N_e**	Lower 95% CI	Upper 95% CI	**N_e**	Lower 95% CI	Upper 95% CI
Chagos	0.000723	−0.514	28891.5	5121.7	41047.6	466.3	157.4	∞
Cocos (Keeling)	0.001282	–0.182	–	–	–	–	–	–
Ningaloo	**0.001499**	–1.221	101470.5	24949.7	138356.2	**3884.7**	334.6	∞
Rowley	**0.001476**	–1.277	**135240.3**	25634.9	153438.2	750.3	288.8	∞
Scott	**0.001505**	–1.285	**132911.6**	26008.6	147212.8	1589.5	591.3	∞
Misool	**0.001486**	–1.304	**194916.1**	56426.7	494678.8	1748.4	431.5	∞
North GBR	0.001420	–1.132	**169239.3**	98373.3	522511.8	482.5	33.7	∞
Herald Cays	0.001419	–0.812	111353.1	32215.4	207913.2	–	–	–
South GBR	0.001385	–0.993	99221.2	21461.0	152602.1	70.2	16.9	∞
Chesterfield	0.001415	–1.172	113394.3	26102.5	139106.3	**5528.6**	2056.2	∞
Entrecasteaux	0.001400	–1.188	102809.2	19814.1	113421.8	**4400.9**	2095.6	∞
Northern Lagoon	0.001373	–1.232	104105.8	20163.6	115034.0	**4973.8**	2115.8	∞
Grand Astrolabe	0.001376	–1.143	46222.8	18749.3	104094.3	507.5	229.6	∞
Petit Astrolabe	0.001403	–1.206	109191.4	20991.8	118066.9	408.6	179.7	∞
Southern Lagoon	0.001391	–1.122	95646.5	19313.9	105695.9	691.1	249.9	∞
Walpole	0.001387	–1.066	71518.2	18588.6	102554.2	502.0	228.8	∞
Matthew	0.001370	–0.769	30855.7	4320.1	68399.5	160.8	90.5	546.9

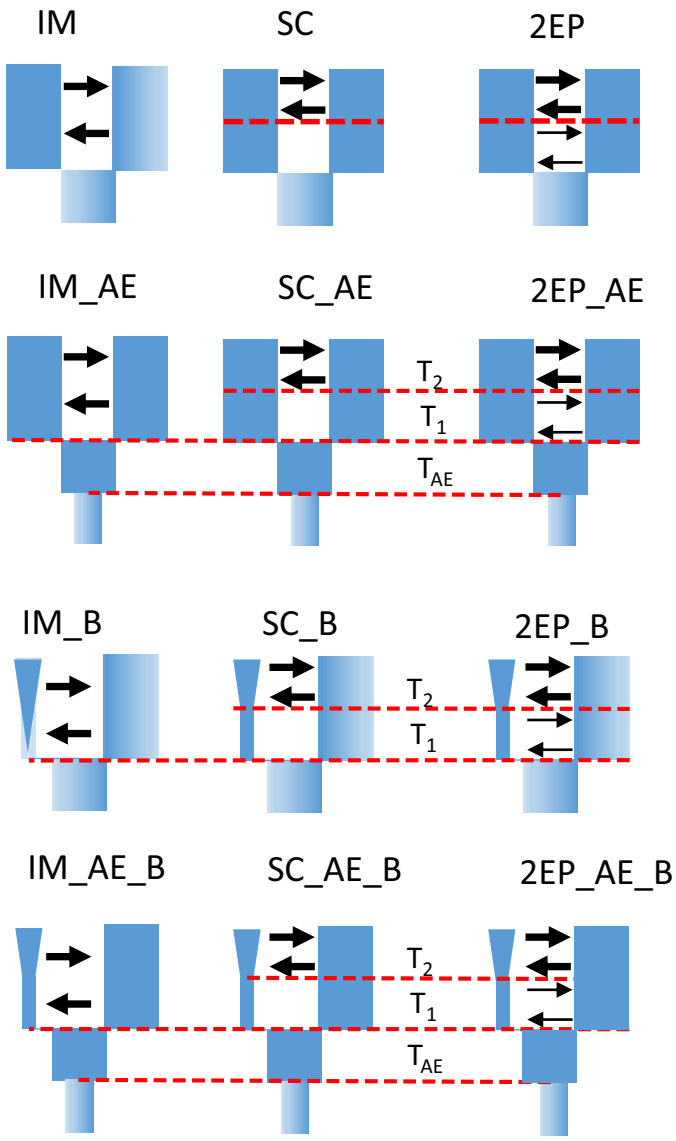

FIGURE S1 Graphical representation of all *Carcharhinus amblyrhynchos* two-population models tested in MOMENTS (except the SI and AM models which performed much worse than all others).

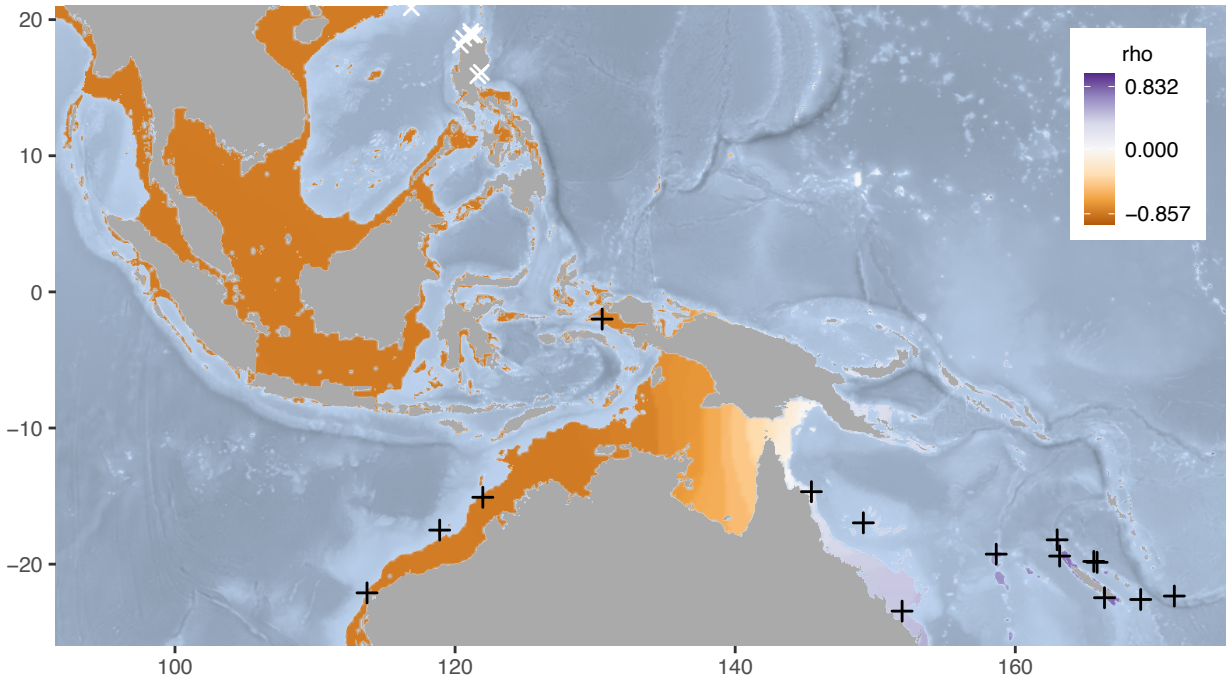

FIGURE S2 Spatial distribution of the Spearman's ρ (rho) coefficient between genetic diversity and distance by sea from the 15 *Carcharhinus amblyrhynchos* sampling locations (black crosses) other than Chagos and Cocos (Keeling) for 7,521 points shallower than 200 meters below sea level. The lowest values of rho (indicating the most likely centre of origin) were -0.857 in the northern Coral Triangle (white *X*'s).