Supplementary information for:

A Size-determining Supergene Hampers a Vulnerable Population Recovery

Pierre Lesturgie[1,2], John Denton[1], Lei Yang[1], Shannon Corrigan[1], Jeff Kneebone[3], Romuald Laso-Jadart[2], Arve Lynghammar[4], Olivier Fedrigo[5], Stefano Mona[2,6,*], Gavin JP Naylor[1,*]

[1]Florida Museum of Natural History, Dickinson Hall, 1659 Museum Road, Gainesville, FL 32611

[2]Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, EPHE-PSL, Université PSL, CNRS, SU, UA, Paris, France

[3]Anderson Cabot Center for Ocean Life at the New England Aquarium, 1 Central Wharf, Boston, MA 02110, jkneebone@neaq.org, 617-226-2424

[4]UiT The Arctic University of Norway, Faculty of Biosciences, Fisheries and Economics, The Norwegian College of Fishery Science, NO-9037 Breivika, Norway

[5]Colossal Biosciences, Dallas TX, United States

[6]EPHE, PSL Research University, Paris, France

[*]These authors contributed equally to this work.

Supplementary results

Summary statistics

Depth of coverage was on average ~17x. After filtering and binning, we performed population structure analyses using ~1.15 to ~1.19 million SNPs. Genetic diversity estimates were computed from the Site Frequency Spectrum (SFS) in sampling locations with N ≥ 5 specimens, based on ~9.98 to ~13.93 million SNPs after filtering (Table 1). We investigated the effect of binning on genetic diversity estimates and on the shape of the SFS by computing the normalized SFS as in [1]. To that end, we sampled genomic regions of 100 bp (to account for monomorphic sites) apart from 1kb, 10kb, 50kb or 100kb in GoM (the site with the largest sample size). Genetic diversity estimates as well as the shape of the SFS were similar at different levels of binning (Fig. S9), thus we decided to keep all SNPs to increase accuracy in demographic inferences. Genetic diversity estimates (θ_π, θ_w) were highly similar in the whole range of the thorny skate, with θ_π ranging from 0.0057 to 0. 0063 and θ_w from 0.0065 to 0.0079. Tajima's D ranged from -0.49 to -0.79 (Table 1).

Haplotype screening and linear modelling

501 individuals were screened by PCR and Sanger sequencing in two regions with ≥ 4 SNPs discriminating the two alleles (HB and HS) of the supergene. Thirty-one individuals with ambiguous genotypes (i.e., where it was not possible to determine genotypes at all discriminating SNPs) were discarded. Linear modelling was performed using a Bayesian framework implemented in the R library *brms*[2]. We tested three models: 1) *GenoMatSex*, the richest model in terms of variables, included Maturity and Sex along with Genotype as dependent variables: "Size ~ Genotype + Maturity + Sex"; 2) *GenoMat* ("Size ~ Genotype + Maturity") and 3) Geno ("Size ~ Genotype"). The two latter models were nested within the more complex one. Modelling was

performed on a subset of individuals for which Maturity and Sex were available. To avoid bias due to population structure and to maximize the sample size, we only retained N=243 individuals from GoM. We run each model four independent times for 10,000 MCMC iterations with 10% burn-in and a thinning of 4 iterations. All models converged for all parameters in all four runs ($\hat{R}$=1) with effective sample sizes (ESS) ≥ 8,000 for any parameters (when pooling all four runs within each model). The Leave-One-Out cross validation displayed model *GenoMat* as the most accurate, even though the expected log pointwise predictive density (ELPD) for *GenoMatSex*, the richest model, was highly similar (difference of -0.3). The posterior predictive check was assessed by using 100 posterior draws from *HaploMat* model and suggested high adequacy between observed and predicted data.

Historical demographic modelling

Five demographic scenarios were investigated (Fig. S7). A set of three models were first tested to specifically investigate patterns of migration and divergence between the two meta-populations (NEA and NWA): IMM-5, IMM-5-NM-STOP and IMM-5-NM-CH, the latter the most parameter rich (Fig. S7). We note that while IMM-5-NM-STOP was the model with the lowest AIC, its likelihood distribution computed for the set of ML parameters' value slightly overlaps with that of IMM-5-NM-CH (Fig. S7). This would suggest that the two models cannot be statistically distinguished [3]. However, the ML values estimated under the two models support the same biological scenario. Going backward in time, the migration rate between the two meta-populations estimated under IMM-5-NM-CH is very close to 0 until T_{CH} (~141ky), suggesting non-significant exchange of migrants in this time frame which overlaps that of IMM-5-NM-STOP (where T_{CH} ~160ky). Similarly, migration rates sharply increase between T_{CH} and T_{DV} to values similar to those

estimated under IMM-5-NM-STOP (Table S2). Finally, we note that the AIC values for IMM-5 was larger, suggesting that modelling a change in connectivity significantly improves our understanding of the demographic dynamics of the thorny skate. Remarkably, all three scenarios were highly consistent in the estimates of the divergence time between the two metapopulations NEA and NWA and in the intra-region estimates of connectivity (Table S2). We further ran a second set of scenarios including *ghost* demes in order to account for the unsampled demes in both metapopulations. Two scenarios were investigated based on IMM-5 topology: IMM-20, with two one-dimensional-matrices of 10 demes exchanging migrants in a stepping-stone fashion (one matrix per region) and IMM-30 in which NEA region was represented by D=20 demes and NWA by D=10 demes. The two scenarios were strikingly less likely than the IMM-5-*like* models (Fig. S5) suggesting that introducing *ghost* demes does not improve our understanding of *A. radiata* historical demography.

References

1. Lapierre, M., Lambert, A. & Achaz, G. Accuracy of Demographic Inferences from the Site Frequency Spectrum: The Case of the Yoruba Population. *Genetics* **206**, 439–449 (2017).
2. Bürkner, P. C. Bayesian Item Response Modeling in R with brms and Stan. *J Stat Softw* **100**, (2021).
3. Meier, J. I. *et al.* Demographic modelling with whole-genome data reveals parallel origin of similar Pundamilia cichlid species after hybridization. *Mol Ecol* **26**, 123–141 (2017).

Supplementary figures legends

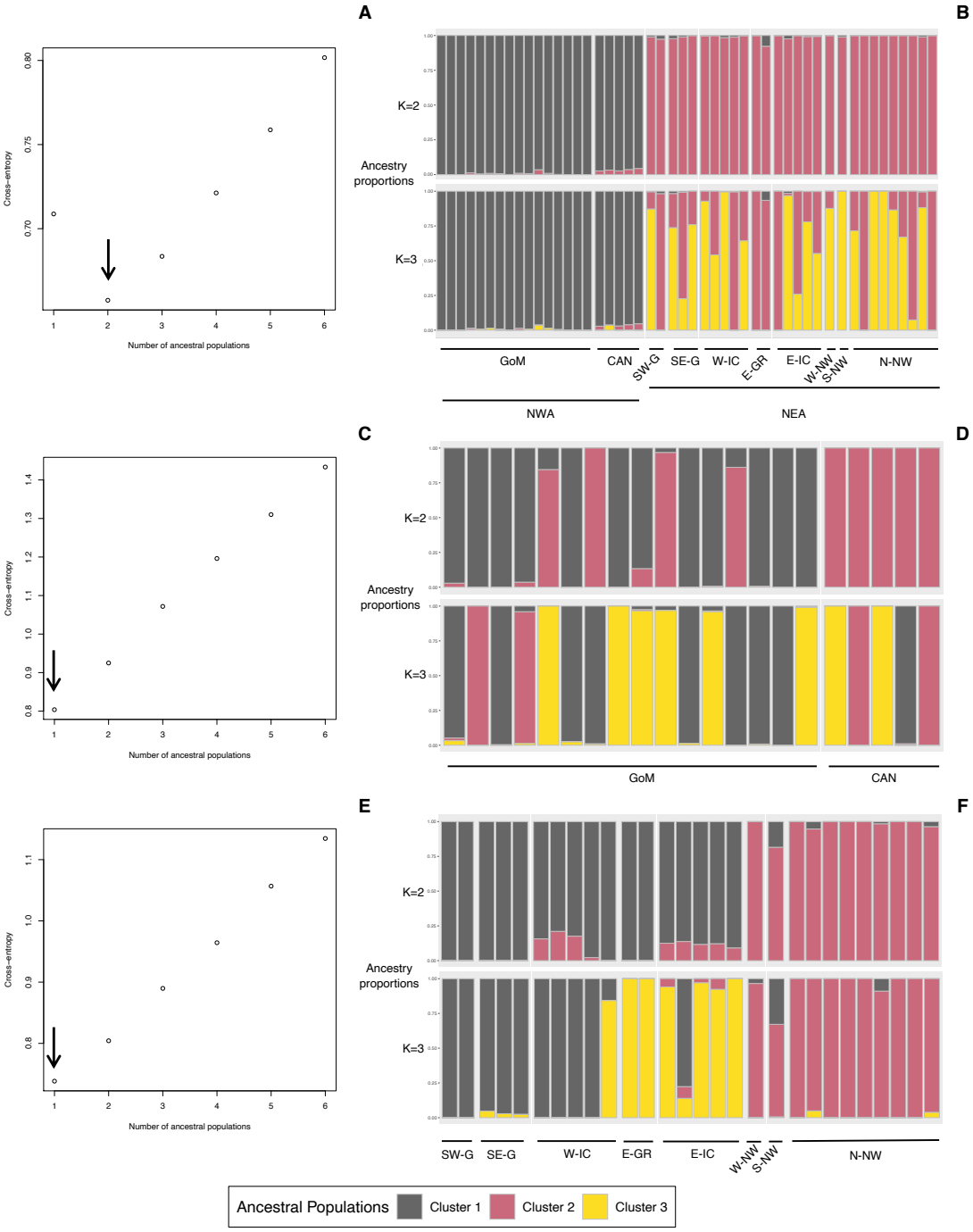

Figure S1. Panels A-F: Cross entropy criterion with an arrow indicating the most likely number of ancestral populations K when using all individuals (A), only GoM and CAN individuals (Cluster WEST, C) or only SWG, SEG, W-IC, E-GR, E-IC, W-NW, S-NW and N-NW individuals (Cluster EAST, E) and corresponding admixture proportions for each individual estimated for K=2 and K=3 ancestral populations when using all individuals (B), WEST Cluster individuals (D) or EAST Cluster individuals (F).

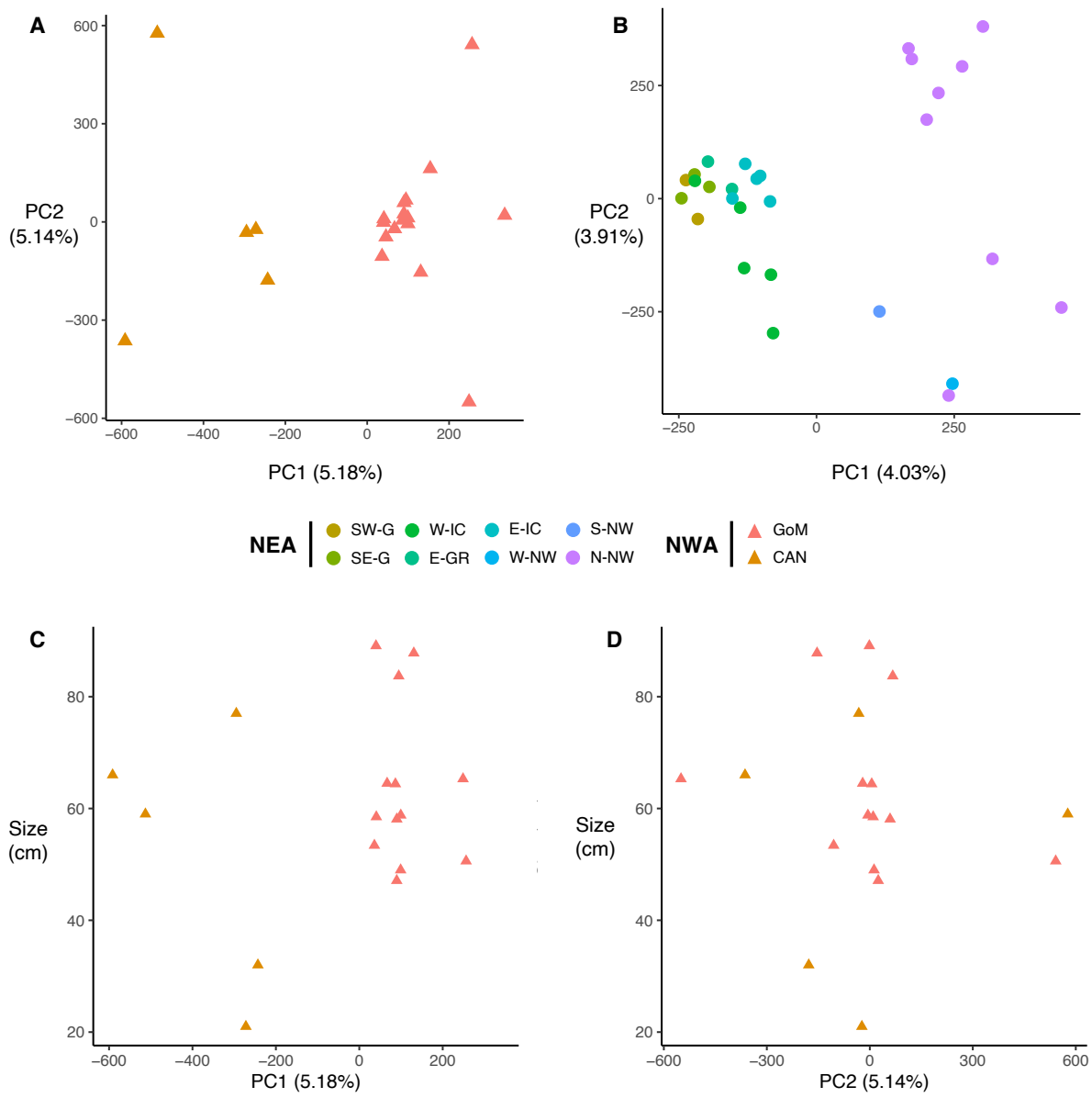

Figure S2. Population structure within each cluster. Panels A-B: PCA using only GoM and CAN individuals (cluster NWA, panel A), and only individuals from SW-G, SE-G, W-IC, E-GR, E-IC, W-NW, S-NW, and N-NW (cluster NEA, panel B). Panels C-D: distribution of Size (in cm) along the PC1 axis (C) and PC2 axis (D) within NWA.

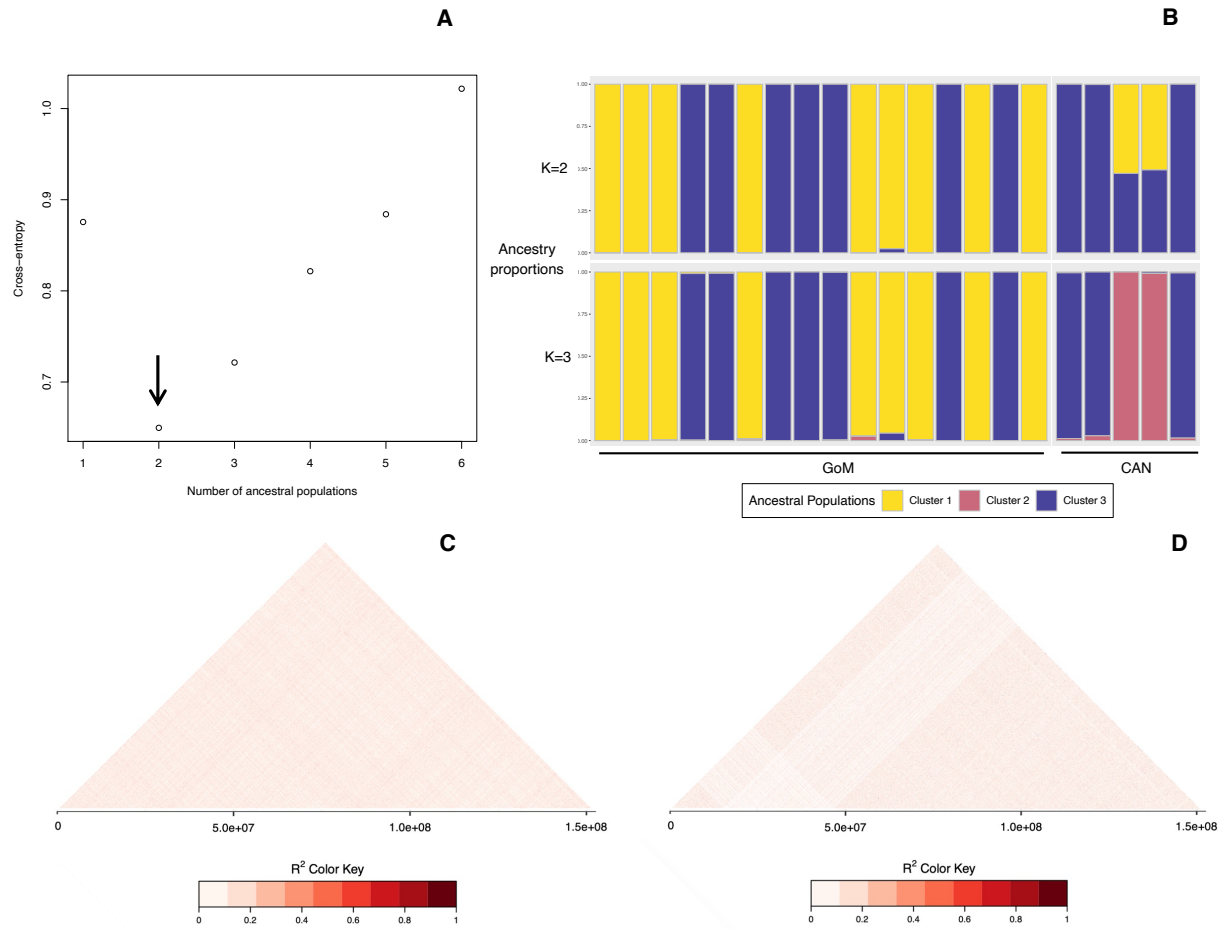

Figure S3. Genetic structure in the NWA within the chromosome 2 inversion (17-48Mb region). Panel A: Cross entropy criterion with an arrow indicating the most likely number of ancestral populations K. Panel B: admixture proportions for each individual estimated for K=2 and K=3 ancestral populations. Panels C-D: Heatmaps of the pairwise linkage disequilibrium between SNPs for HS/HS individuals (C) or HB/HB individuals (D). Color gradients represent the value of the r^2 correlation between SNPs.

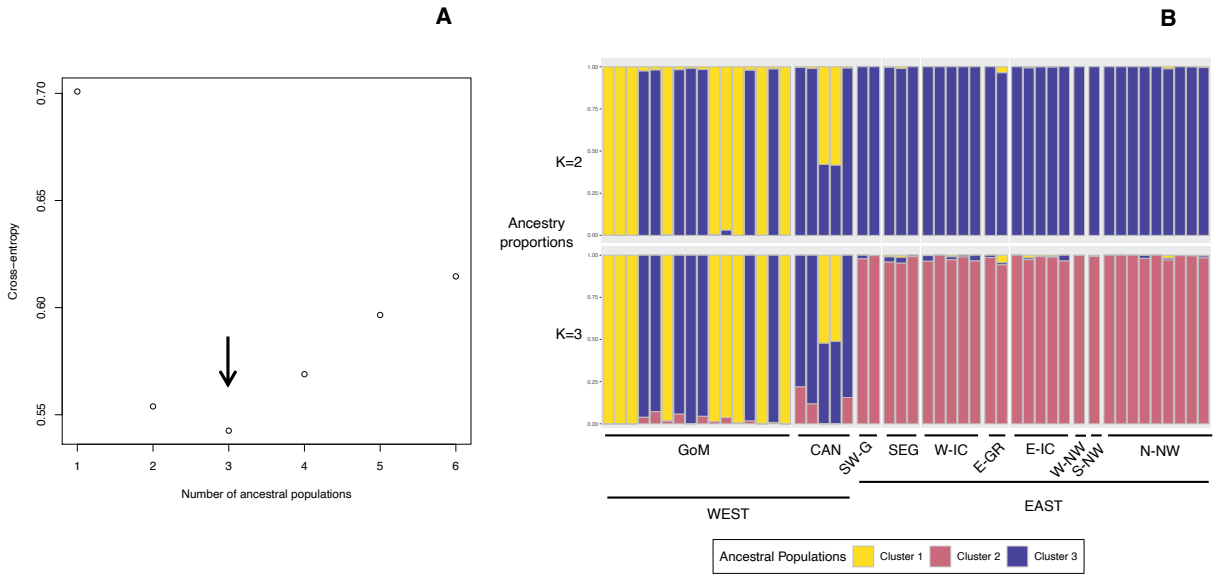

Figure S4. Genetic structure in the whole range within the chromosome 2 inversion (17-48Mb region). Panel A: Cross entropy criterion with an arrow indicating the most likely number of ancestral populations K. Panel B: admixture proportions for each individual estimated for K=2 and K=3 ancestral populations.

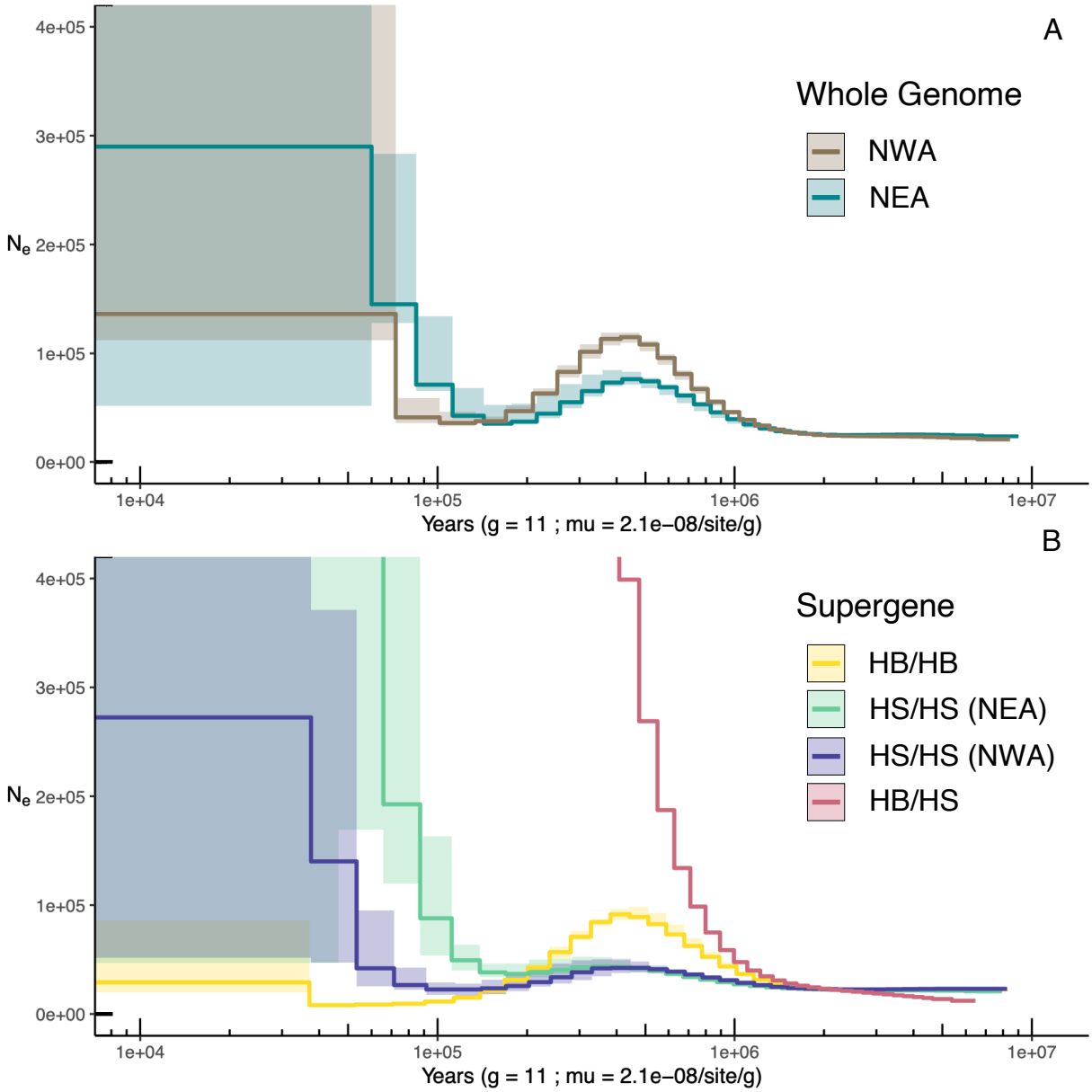

Figure S5. Variation of the coalescence rate through time as estimated by the PSMC algorithm. Panels A and B: inference on the whole genome of NWA (Brown) and NEA (Turquoise) individuals (A) and within the chromosome 2 supergene region (B) for HB/HB (Yellow), HS/HS for NEA (Green), HS/HS for NWA (Blue) and HB/HS (Yellow). The shaded areas and the continuous line represent respectively the distribution and the median of the coalescence rate computed over the 49 individuals at each time interval.

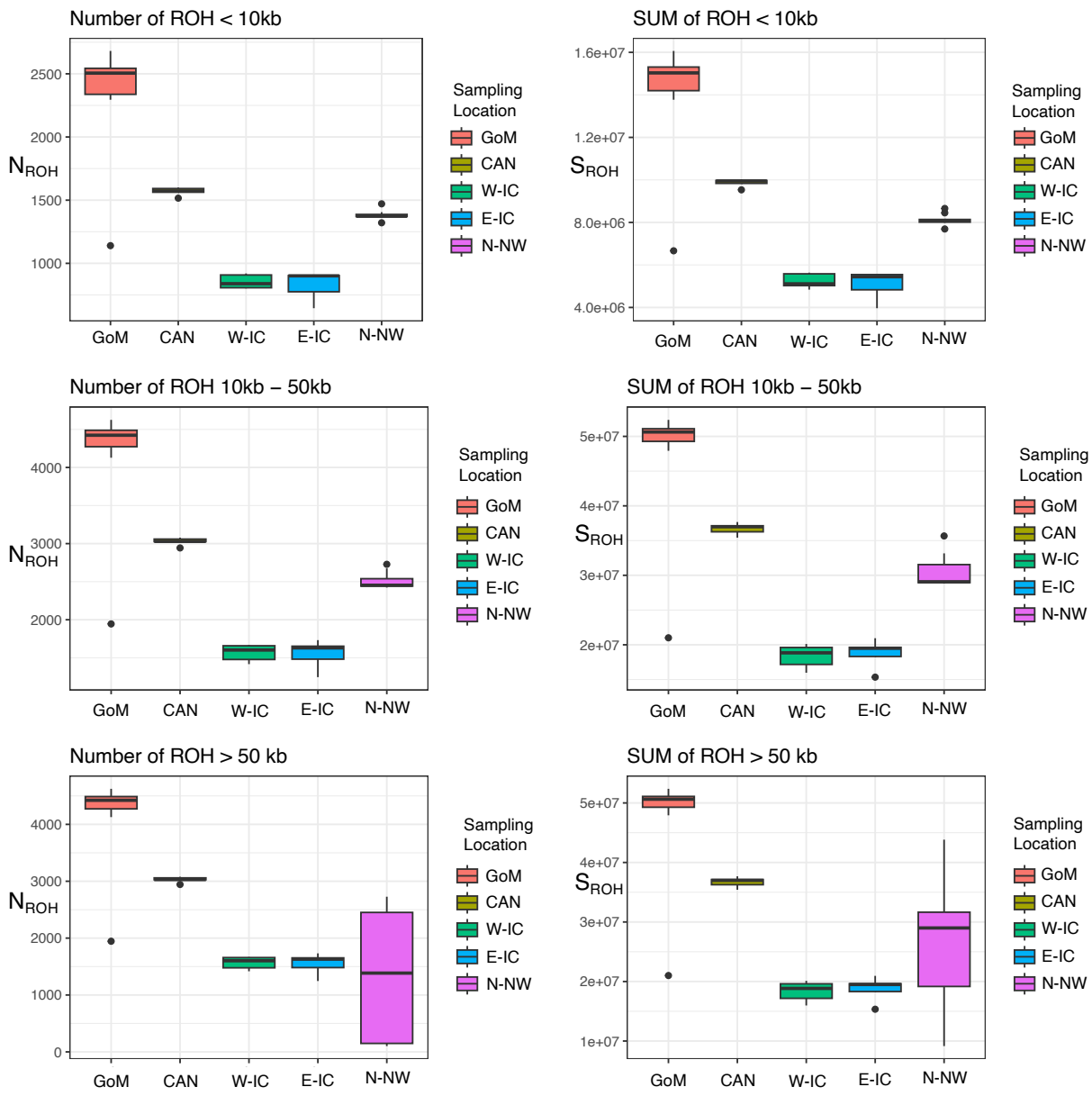

Figure S6. Distribution of Runs of Homozygosity (ROH) in sampling locations with N≥5. Number of ROH (panels A1-A3) and sum of the ROH (panels B1-B3) for different ROH size classes: below 10kb (A1 & B1), between 10kb and 20kb (A2 & B2) and over 20kb (A3 & B3).

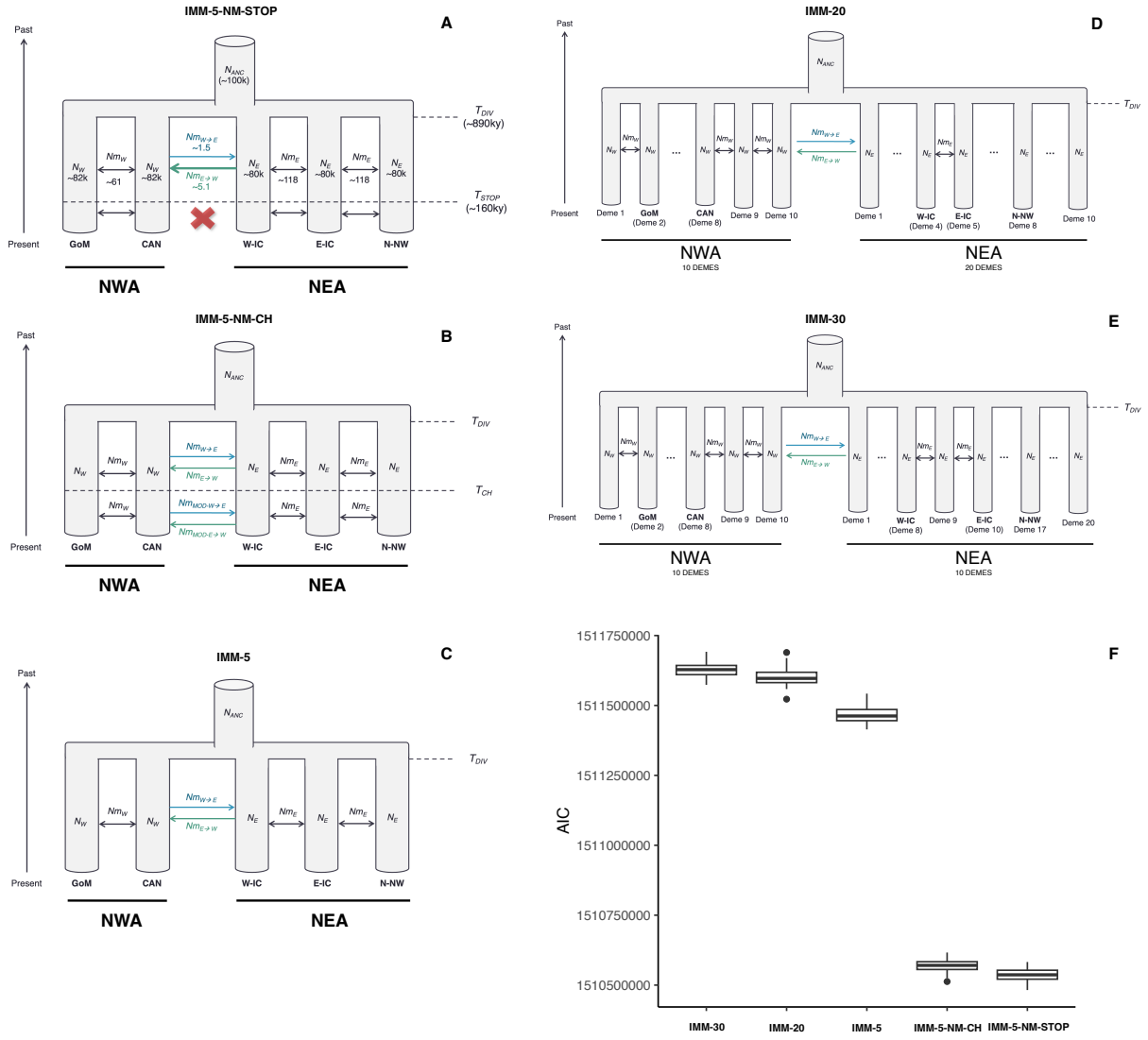

Figure S7. Demographic scenarios tested (panels A-E) and associated AIC values (Panel F).

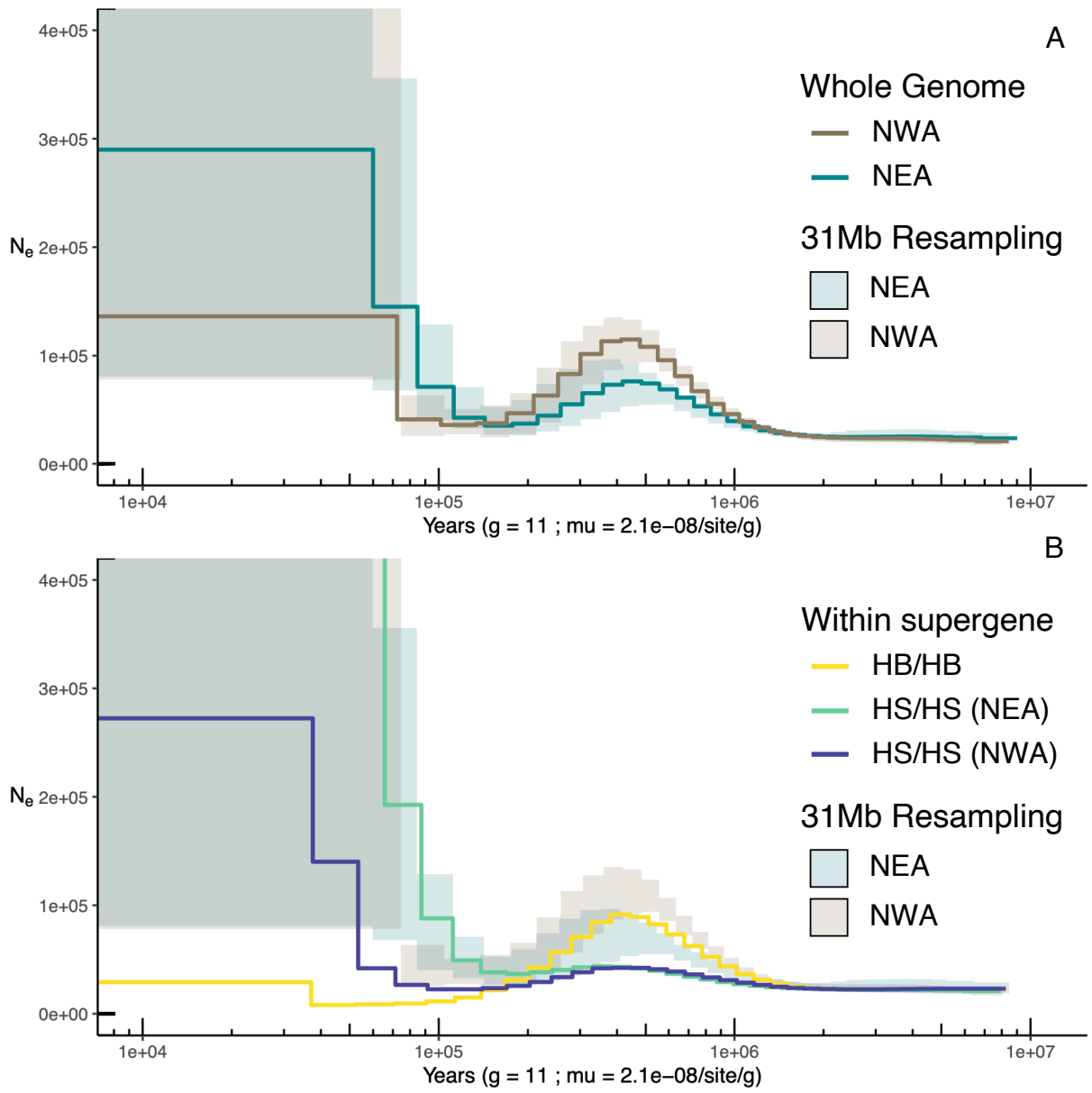

Figure S8. Variation of the coalescence rate through time in random resampled regions of 31Mb. Panels A and B: inference on the whole genome of NWA (Brown) and NEA (Turquoise) individuals (A) and within the chromosome 2 supergene region (B) for HB/HB (Yellow), HS/HS for NEA (Green), HS/HS for NWA (Blue) and HB/HS (Yellow). The shaded areas represent the 95% quantiles of the distribution of random resampling of 31Mb regions across the genome in an NEA (Turquoise) and NWA (Brown) individual.

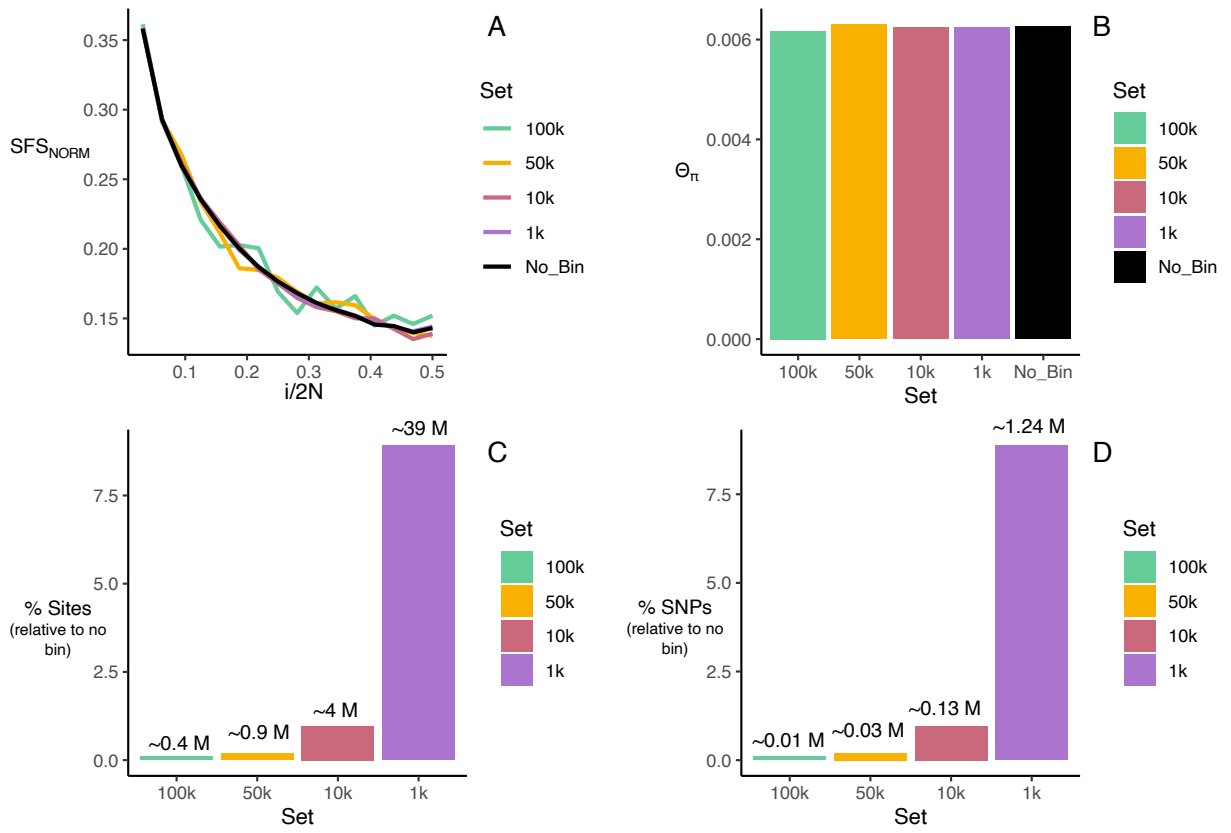

Figure S9. Influence of binning on summary statistics computed in GoM (N=16). Panel A: normalized SFS. Panel B: mean pairwise difference. Panels C and D: barplots of percentage of Sites (C) and SNPs (D) relative to the reference dataset (no binning) with the observed number of Sites and SNPs indicated above each bar. Each color represents a different level of binning: regions separated by 100kb (green), 50kb (orange), 10kb (red), 1kb (purple). Reference dataset (no binning) is presented in black.