nature communications

Article

https://doi.org/10.1038/s41467-025-56126-z

Short-term evolutionary implications of an introgressed size-determining supergene in a vulnerable population

Received: 29 March 2024

Accepted: 2 January 2025

Published online: 27 January 2025

Check for updates

Pierre Lesturgie[1,2,3] ✉, John S. S. Denton[4], Lei Yang[1], Shannon Corrigan[1], Jeff Kneebone[5], Romuald Laso-Jadart[2,6], Arve Lynghammar[7], Olivier Fedrigo[8], Stefano Mona[2,6,9] ✉ & Gavin J. P. Naylor[1,9] ✉

The Thorny Skate (*Amblyraja radiata)* is a vulnerable species displaying a discrete size-polymorphism in the northwest Atlantic Ocean (NWA). We conducted whole genome sequencing of samples collected across its range. Genetic diversity was similar at all sampled sites, but we discovered a ~ 31 megabase bi-allelic supergene associated with the size polymorphism, with the larger size allele having introgressed in the last ~160,000 years B.P. While both Gulf of Maine (GoM) and Canadian (CAN) populations exhibit the size polymorphism, we detected a significant deficit of heterozygotes at the supergene and longer stretches of homozygosity in GoM population. This suggests inbreeding driven by assortative mating for size in GoM but not in CAN. Coalescent-based demographic modelling reveals strong migration between regions maintaining genetic variability in the recombining genome, preventing speciation between morphs. This study highlights short-term context-dependent evolutionary consequences of a size-determining supergene providing new insights for the management of vulnerable species.

Chromosomal inversions prevent recombination, thereby maintaining the specific allelic arrangement within the genes they encompass, leading to enhanced fitness[1–5]. Suites of genes in inversions are often referred to as supergenes, as they can lead to the Mendelian inheritance of complex phenotypes[6,7]. While the existence of supergene systems has long been acknowledged[8], the accessibility of whole genome sequencing (WGS) data has significantly amplified their detection. The presence of supergene-associated traits has been shown in several systems including sociality in ants[9–13], migratory behavior and adaptation to salinity and temperature in Atlantic cod[14–16], and wing morphology and pattern coloration in butterflies[17,18]. Supergenes are maintained in populations by an interplay between demographic and selective processes[6], the relative contributions of which can be difficult to disentangle without careful reconstruction of the demographic history of populations based on neutral markers. Varying selection in space and time is often invoked as one of the mechanisms to explain the long-term persistence of supergenes[1]. Yet, the limited variability resulting from recombination suppression alongside complex phenotypes may impede swift adaptative responses to rapid environmental shifts. This could critically impact survival

[1]Florida Museum of Natural History, Dickinson Hall, 1659 Museum Road, Gainesville, FL 32611, USA. [2]Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France. [3]cE3c – Centre for Ecology, Evolution and Environmental Changes, CHANGE–Global Change and Sustainability Institute, Department of Animal Biology, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisbon, Portugal. [4]Department of Ichthyology, American Museum of Natural History, Central Park West @ 79th Street, New York, NY 10024, USA. [5]Anderson Cabot Center for Ocean Life at the New England Aquarium, 1 Central Wharf, Boston, MA 02110, USA. [6]EPHE, PSL Research University, Paris, France. [7]UiT The Arctic University of Norway, Faculty of Biosciences, Fisheries and Economics, The Norwegian College of Fishery Science, NO-9037 Breivika, Tromsø, Norway. [8]Colossal Biosciences, Dallas, TX, USA. [9]These authors contributed equally: Stefano Mona, Gavin J. P. Naylor. ✉e-mail: pierrelesturgie@outlook.fr; stefano.mona@mnhn.fr; gjpnaylor@gmail.com

potential in threatened species, underscoring the potential role that supergenes might play in driving short-term evolutionary dynamics of species.

The thorny skate (*Amblyraja radiata*) is a demersal species found on the continental shelf from South Carolina in the Northwest Atlantic (NWA) via Greenland and Iceland to the British Isles and the Barents Sea in the Northeast Atlantic (NEA, Fig. 1)[19]. In the NWA, the thorny skate has long been taken as target catch or bycatch in commercial fisheries. Their abundance has declined steeply in waters off Canada and the US Gulf of Maine over the last 50 years[19,20]. This decline prompted stringent conservation measures that eliminated directed harvest in large geographic areas. However, the Gulf of Maine population has shown no signs of recovery despite two decades of reduced fishing mortality[19,20]. The factors impeding population recovery in the Gulf of Maine remain unknown[21–24]. Complicating this conservation issue are two regionally sympatric size morphs present in the NWA, each displaying characteristic growth curves[25–29]. The large morph is restricted to the NWA and reaches a maximum size of 105 cm total length (TL). The small morph occurs over the species' entire range and reaches a maximum size of 72 cm TL[28]. To date, the genetic underpinnings of this morphological polymorphism have eluded detection. Mitochondrial data suggest weak population genetic structure and isolation by distance across the entire range[30,31] and microsatellite data show regional genetic differentiation between the NWA and NEA[32]. However, neither data type indicates genetic differentiation between large and small morphotypes[31–33].

Here we sought to understand the genomic and/or environmental origins of the size polymorphism in the NWA and any implications for

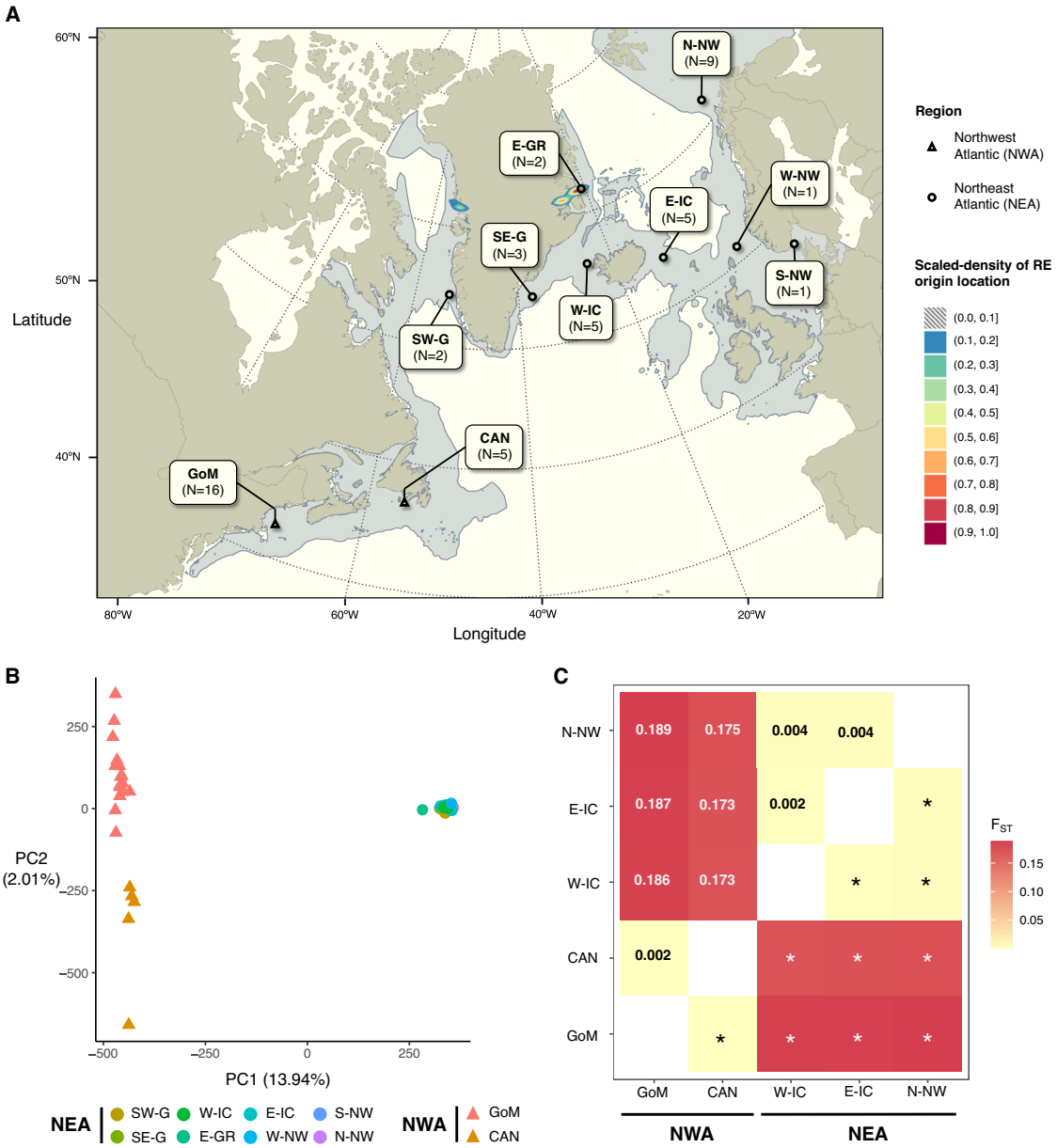

Fig. 1 | Whole genome sample scheme and population structure of thorny skates. Panel (**A**): sampling locations, from left to right: Gulf of Maine (GoM); Newfoundland, Canada (CAN); Southwest Greenland (SW-G); Southeast Greenland (SE-G); East Greenland (E-GR); Western Iceland (W-IC); Eastern Iceland (E-IC); Northern Norway (N-NW); Western Norway (W-NW) and Southern Norway (S-NW). Range distribution of the thorny skate is filled in shaded blue (obtained from[31]). The sequential coloured areas represent the scaled density of the range expansion origin inferred using the TDOA algorithm computed 100 times. The map was made using R v4.4.0[87]. Panel (**B**): Principal Component Analysis (PCA) using all individuals. Panel (**C**): Heatmap of pairwise F_{ST} values between sampling locations with $N \geq 5$ (upper left) and associated significance evaluated using 1000 permutations for each pairwise comparison (lower right).

https://doi.org/10.1038/s41467-025-56126-z

the conservation of the species. We first established a high-quality reference genome for the thorny skate based on a combination of long read (PacBio), Hi-C, Bionano, and Illumina short read sequencing in collaboration with the Vertebrate Genomes Project (https://vertebrategenomesproject.org). Using this high-quality reference assembly, we carried out whole-genome population resequencing (Illumina ~24x coverage) of 49 individuals from representative locations spanning the species' distribution range (Fig. 1A). This approach allowed us to discover a ~ 31 megabase (Mb) supergene on chromosome 2, likely contained in a chromosome inversion, that was polymorphic in the NWA. We expanded the scope of our study and screened for the genotype of the supergene in 465 individuals across the range of the species to characterize the distribution of the supergene's alleles and investigate its association with size. To better understand the origin, maintenance, and allelic distribution of the supergene, we reconstructed the demographic history of the species based on analysis of millions of putatively neutral genome-wide Single Nucleotide Polymorphisms (SNPs). PacBio long read sequencing of an individual of the closely related species *A. hyperborea* revealed the supergene to be present in at least one of the congeners. When this information was combined with the extant geographic distribution of both alleles and the historical reconstruction of demography, we were able to infer that the supergene was most likely transmitted to *A. radiata* through cross-species introgression. Our findings showed significantly higher level of inbreeding, driven by assortative mating for size polymorphism, in the highly vulnerable Gulf of Maine population. We further discussed how assortative mating could hamper its recovery, presenting a particular challenge for thorny skate conservation and management in the NWA.

Results

Summary statistics

Mapping rate, total number of reads, average coverage, average depth and average mapping quality computed for each individual before filtering is presented in Table S1. 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). Additional description of summary statistics can be found in the supplementary material.

Population structure

We used Principal Component Analysis (PCA) of SNP variation to explore population structure. The first axis (~14% of total variance) revealed two clusters, corresponding to the NEA and NWA regions, respectively (Fig. 1B). The second axis (~2% of total variance) separated NWA individuals sampled from the Gulf of Maine (GoM) with those from Newfoundland (CAN). The sparse non-Negative Matrix Factorization (sNMF) algorithm[34] further identified K = 2 as the most likely number of ancestral populations with individual ancestry coefficients perfectly matching the clusters detected by the PCA (Fig. 1B, S1). We further ran both the PCA and sNMF within each cluster separately. The first two PC axes explained only ~5% and ~4% of the total variance in NWA and NEA respectively (Fig. S2), and in both datasets K = 1 was the most likely number of ancestral populations (Fig. S1). However, both algorithms harbored signatures of fine scale population structure driven by the clinal distribution of genetic variation within both regions (Fig. S1 and S2). All pairwise F_{ST} comparisons were statistically significant ($p \leq 0.001$) and generally consistent with the results provided by the clustering algorithms. Values ranged from 0.002 to 0.004 in intra-cluster comparisons (i.e., within NEA and within NWA) and from 0.173 to 0.189 when comparing NEA vs NWA sampling sites (Fig. 1C).

Table 1 | Summary statistics for each sampling location

		N_{WG}	N_{SC}	N_{SNPs}	N_{sites}	θ_π	θ_w	TD	$N_{HB/HB}$	$N_{HB/HS}$	$N_{HS/HS}$	f_{HB}
NWA	GoM	16	282	13,926,040	439,435,032	0.0063	0.0079	-0.7903	**76 (f = 0.27)**	**10 (f = 0.04)**	**196 (f = 0.69)**	0.29
	CAN	5	38	11,872,562	604,725,402	0.0063	0.0069	-0.4877	3 (*f* = 0.08)	23 (*f* = 0.60)	12 (*f* = 0.32)	0.38
NEA	SW-G	2	7	–	–	–	–	–	0 (*f* = 0)	0 (*f* = 0)	7 (*f* = 1)	0
	SE-G	3	0	–	–	–	–	–	–	–	–	–
	E-GR	2	2	–	-	-	-	-	0 (*f* = 0)	0 (*f* = 0)	2 (*f* = 1)	0
	W-IC	5	50	10,275,797	558,262,795	0.0058	0.0065	-0.5318	0 (*f* = 0)	0 (*f* = 0)	50 (*f* = 1)	0
	E-IC	5	34	9,982,473	535,940,022	0.0059	0.0066	-0.5212	0 (*f* = 0)	0 (*f* = 0)	34 (*f* = 1)	0
	W-NW	1	0	–	–	–	–	–	–	–	–	–
	S-NW	1	5	–	–	–	–	–	0 (*f* = 0)	0 (*f* = 0)	5 (*f* = 1)	0
	N-NW	9	47	12,708,852	538,100,423	0.0057	0.0069	−0.6983	0 (*f* = 0)	0 (*f* = 0)	47 (*f* = 1)	0

Number of individuals sampled for the whole genome study (N_{WG}) and the screening study (N_{SC}). For each sampling location with $N_{WG} \geq 5$: total number of SNPs (N_{SNPs}), total number of sites (N_{sites}), mean pairwise difference (θ_π) and Watterson's estimator of genetic diversity (θ_w) both scaled by N_{sites} and Tajima's *D* (TD). For sampling sites included in the screening study ($N_{SC} > 0$), total number of individuals carrying each genotype ($N_{HB/HB}$, $N_{HB/HS}$, $N_{HS/HS}$) and HB allele frequency (f_{HB}) are reported. Number of individuals carrying each genotype in bold are not in Hardy-Weinberg equilibrium (HW two-sided exact-test: $p < 0.001$).

https://doi.org/10.1038/s41467-025-56126-z

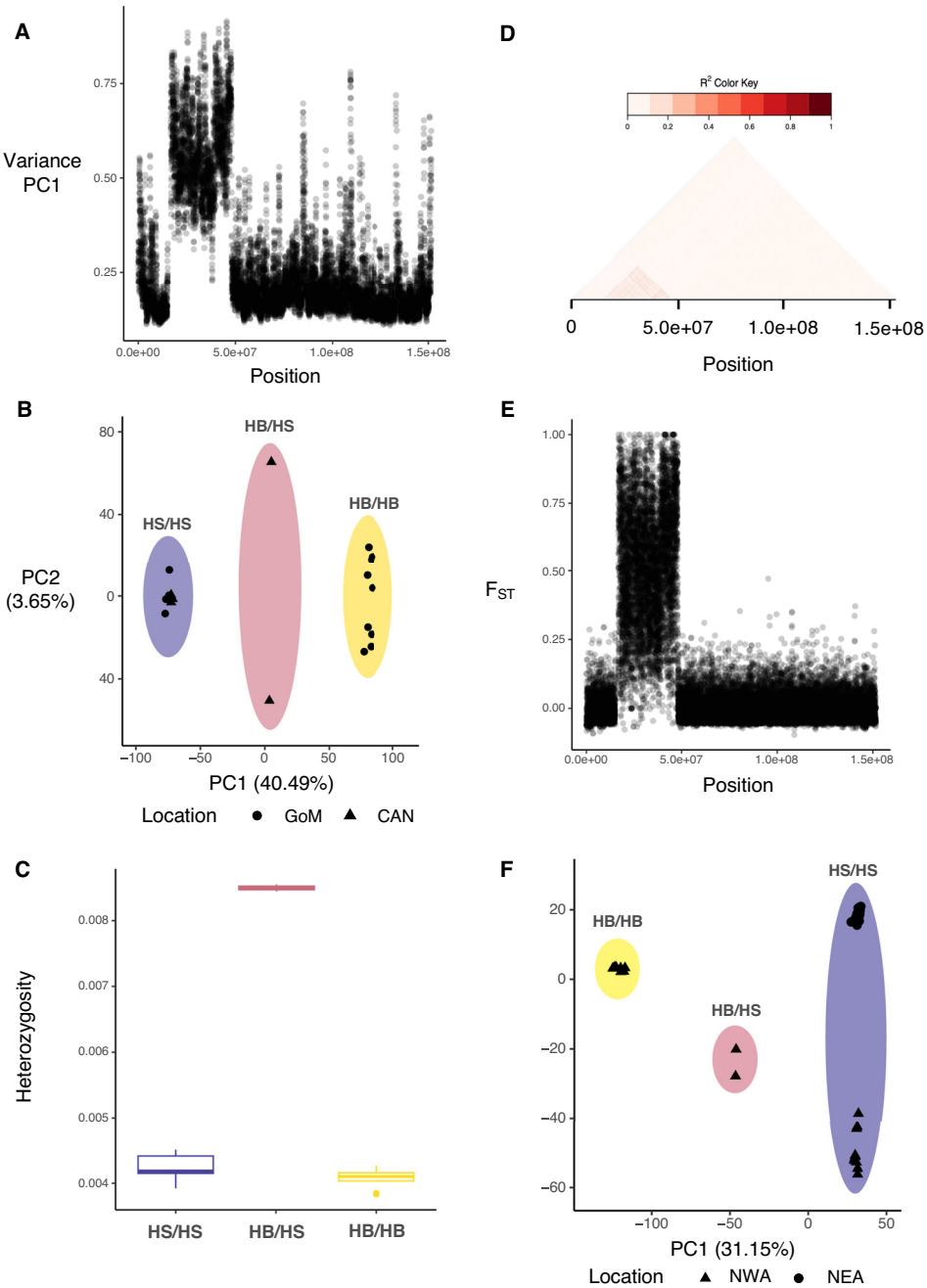

Fig. 2 | Detection of a supergene in chromosome 2. Panel (**A**): sliding windows analysis of the proportion of variance explained by the first axis of a PCA computed on chromosome 2 in NWA samples. Each dot corresponds to a window. Panel (**B**): local PCA within the supergene region (17,000,000-48,000,000) of chromosome 2 including only NWA individuals. Panel (**C**): proportion of heterozygotes within the supergene region for HB/HB ($N = 9$), HB/HS ($N = 2$) and HS/HS ($N = 10$) genotypes. The line within the boxplot represents the median value with the lower and upper hinges being the first (Q1) and third (Q3) quartiles. The upper whisker goes from the hinge to the largest value no further than 1.5 times the distance between Q1 and Q3 and the lower whisker from the hinge to the lowest value no further than 1.5 times the distance between Q1 and Q3. Data beyond the whisker are represented as points. Panel (**D**): heatmap of the pairwise linkage disequilibrium between SNPs. Color gradient represents the value of the r^2 correlation between SNPs. Panel (**E**): sliding window F_{ST} between HB/HB and HS/HS individuals from NWA. Panel (**F**): local PCA within the supergene region including all individuals. Dot shape in Panels **B** and **F** represents the sampling location and colour represents the genotype at the supergene: HS/HS (purple), HB/HS (red) and HB/HB (yellow).

Detection of a supergene

Genome wide analyses highlighted geography as the main driver of population differentiation and did not detect any differences between the large and small morphs in the NWA (Fig. S2). However, when we used a genomic sliding windows analysis of PCA over the pooled NWA sample (to compute the percentage of total variance explained by the first axis within each window), we identified a ~ 31 Mb region in chromosome 2 (coordinates: ~17 Mb - ~48 Mb) that was strikingly different from the genome-wide average (Fig. 2A). Local PCA computed within this region displayed three clusters segregated by the first axis (Fig. 2B). The two most distant clusters on the first axis were characterized by an excess of the two alternative homozygous genotypes, while individuals in the middle of the first axis displayed an excess of heterozygous genotypes (Fig. 2C). This result was corroborated by the sNMF, which found K = 2 as the most likely number of ancestral populations with individuals showing an excess of heterozygotes

https://doi.org/10.1038/s41467-025-56126-z

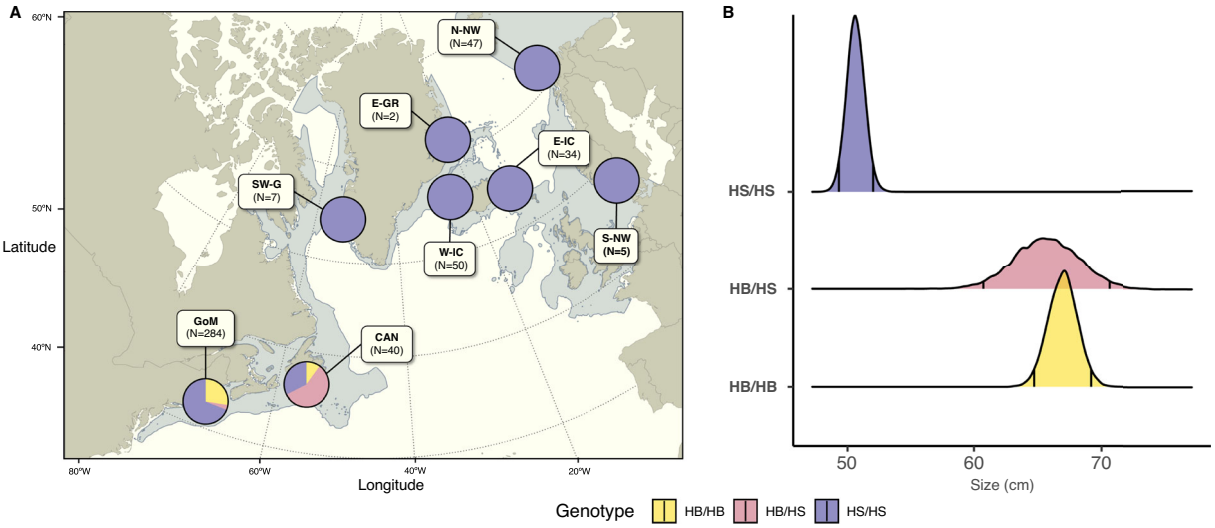

Fig. 3 | Distribution of supergene's genotypes and their association with size. Panel (**A**): distribution of the genotypes across sampling locations obtained using the screening approach, from left to right: Gulf of Maine (GoM); Newfoundland, Canada (CAN); Southwest Greenland (SW-G); East Greenland (E-GR); Western Iceland (W-IC); Eastern Iceland (E-IC); Northern Norway (N-NW) and Southern Norway (S-NW). Range distribution of the thorny skate is filled in shaded blue (obtained from[31]). The map was made using R v4.4.0[87]. Panel (**B**): Posterior distribution of the size as estimated by model GenoMat for each genotype. 2.5% and 97.5% quantiles of the credible interval are represented in each distribution by vertical bars. Colours represent the genotype of the supergene: HS/HS (purple), HB/HS (Red) and HB/HB (yellow).

being almost exactly half admixed between them (Fig. S3). Finally, we investigated Linkage Disequilibrium (LD) in both the pooled sample and in the two clusters separately (Fig. 2D and S3): the region is characterized by strong LD in the pooled sample when compared to the rest of chromosome 2. Conversely, LD values were similar to the rest of the genome (or lower) when computed within each previously identified cluster. Additionally, F_{ST} values between the two clusters characterized by an excess of homozygosity reached up to ~1 in the region (suggestive of fixation) while remaining distributed around ~0 outside (Fig. 2E). Collectively, these results suggest that this region is an inversion, where recombination has been suppressed.

Given the occurrence of 226 annotated genes in the 17–48 Mb region of chromosome 2 of the reference genome, we hereafter refer to this region as a supergene, characterized by two haplotypes (HB and HS) inherited in a Mendelian fashion[6]. Individuals with reference homozygous genotypes are HS/HS, individuals with alternative homozygous genotypes are HB/HB, and heterozygote individuals are HB/HS. Preliminary results suggested that the size of individuals was different between the two homozygous genotypes: HB/HB had an average size of ~71.7 cm and HS/HS of ~53.9 cm. However, sample sizes were too low (HB/HB: $N = 9$, HS/HS: $N = 10$) to model confounding factors such as sex and maturity. When a local PCA was run including the NEA samples, the first axis segregated NEA and HS/HS individuals from HB/HB and HB/HS individuals before the NEA-NWA geographical divergence detected by the genome-wide structure analyses (Fig. 2F). The same pattern was observed when computing the individual ancestry with the sNMF (Fig. S4), which implies that NEA individuals are all HS/HS and the divergence between HB and HS alleles predates the split between the NEA and NWA regions.

Genotype screening and size association

To further investigate the association between the supergene genotypes and size, we first selected two regions each with ≥ 4 SNPs discriminating HB and HS alleles within the supergene and further screened by PCR and Sanger sequencing of 501 individuals (465 after filtering, see supplementary results) throughout the species' distribution range (Table 1). HB was absent in the NEA, consistent with the lack of body size polymorphism in this part of the range. Conversely, HB reached a frequency of 0.29 and 0.38 in the GoM and CAN sampling sites, respectively (Table 1 and Fig. 3A). However, the distribution of genotypes in the two sampling sites was strikingly different: the GoM displayed a strong deficit in heterozygotes (only 10 out of 282 individuals, Hardy-Weinberg exact-test: $p < 0.001$), while CAN was in Hardy-Weinberg equilibrium (Table 1). We then investigated the relationship between size and genotype controlling for maturity and/or sex using linear models in a Bayesian framework (Table S2) using the 241 GoM individuals with no missing information on any trait. Leave-One-Out cross validation indicated that the model including size and maturity provided the best fit (see supplementary results). Posterior distribution of size for HB/HB and HB/HS individuals largely overlapped, with median values and 95% credibility intervals (averaged over the levels of maturity) of 66.95 cm (95% CI [64.73; 69.17] cm) for the former and 65.61 cm (95% CI [60.81; 70.50] cm) for the latter. Conversely, size for HS/HS individuals was strikingly lower (median value of 50.68 cm, 95% CI [49.28; 52.07] cm) and associated with disjunct posterior distribution from HB carriers (Fig. 3B).

Range expansion and genomic inbreeding

Population structure (Fig. 1B) suggested strong genetic differences between NEA and NWA. To test for signatures of Range Expansion (RE) and understand the colonization history of the species, we examined genomic signatures of RE by fitting the directionality index between each pair of individuals to the time difference of arrival (TDOA) location algorithm[35]. The density distribution of the center of origin of the RE computed over 100 independent runs was always found in NEA region, off the coast of Greenland, with more than 80% of runs indicating an origin off the eastern coast of Greenland (Fig. 1A). We then investigated the distribution of Runs of Homozygosity (ROH). Length and number of ROH depends on both species-wide demographic processes and on intra-population levels of genomic inbreeding. Consistently with the RE evidence, we found larger ROH in areas away from the putative center of origin (Fig. S5). For instance, the number (N_{ROH}) and genomic coverage (SUM_{ROH}) were always the lowest in the two Iceland sampling sites (W-IC and E-IC) and strikingly higher in the GoM followed by CAN and N-NW (Fig. S5). The significantly larger N_{ROH} and SUM_{ROH} in GoM than in CAN, despite the similar amount of genetic

https://doi.org/10.1038/s41467-025-56126-z

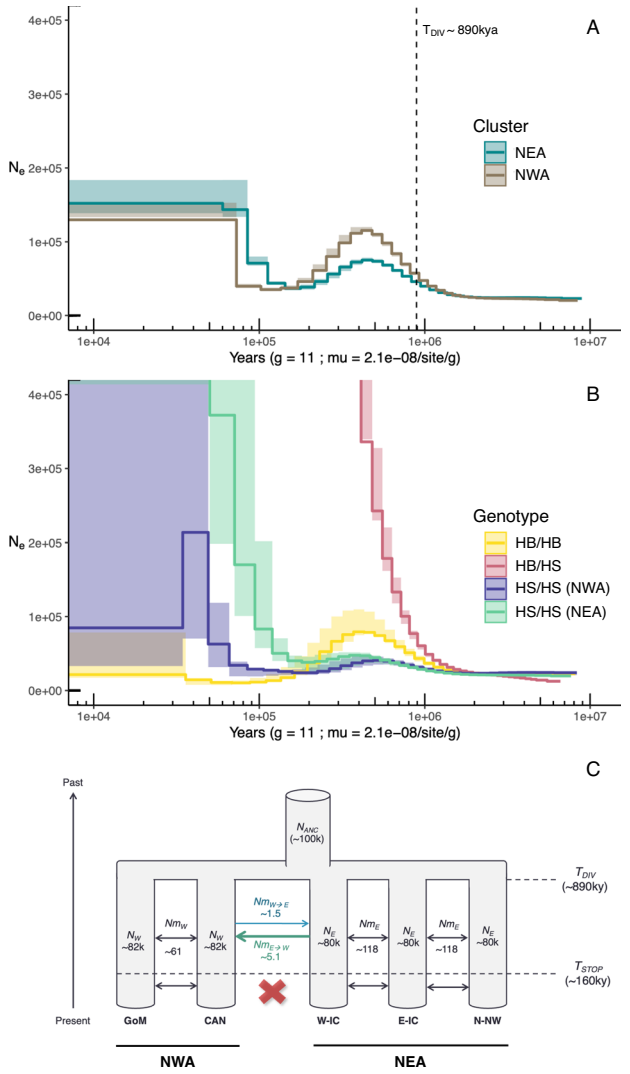

Fig. 4 | Global and within supergene historical demography. PSMC computed using the whole genome data from two individuals representing the NEA (turquoise) and NWA (brown) regions (**A**) and within the chromosome 2 supergene in four individuals: HB/HB (Yellow), HB/HS (Red), HS/HS from the NWA (Blue), HS/HS from the NEA (Green) (**B**). Shaded areas represent the 95% confidence interval computed on 100 bootstraps replicates. The vertical dotted line in panel (**A**) represents T_{DIV} (divergence between the NEA and the NWA) as estimated by fastsimcoal under IMM-5-NM-STOP model. Panel (**C**): Demographic model IMM-5-NM-STOP with maximum likelihood estimates for each parameter.

diversity (Table 1) and low genetic differentiation (Fig. 1), strongly suggest higher inbreeding in the former as consequence of different mating strategy.

Historical demography

The restricted distribution of HB might be the consequence of neutral and/or selective forces. To better understand the origin, maintenance, and historical demography of the size polymorphism, we first ran the Pairwise-Sequential Markovian Coalescent (PSMC) algorithm[36] on each of the 49 whole-genome sequenced individuals. PSMC curves were almost identical for every individual at the regional scale, but the dynamics differed between the NEA and NWA regions, whose trajectory started to diverge ~1 million years B.P. (Fig. 4A, S6). While the exact date of divergence between the two trajectories may be inaccurate[37], there is a clear separation of the evolutionary trajectories between the NEA and NWA, consistent with population structure (Fig. 1B).

To further investigate patterns of colonization, migration and divergence between and within the two meta-populations (NEA and NWA; Fig. 4C and S7), we investigated five demographic scenarios using fastsimcoal2[38]. The AIC criterion computed after choosing the best out of 10 replicates of each model indicated IMM-5-NM-STOP as the most likely (Fig. S7). IMM-5-NM-STOP depicts two meta-populations composed (in this order) of GoM and CAN sampling locations (NWA meta-population) and W-IC, E-IC and N-NW sampling locations (NEA meta-population). Deme effective sizes were highly similar between the NEA and NWA demes (N_E ~ 80,000, 95% CI [74,595; 81,106] and N_W ~ 82,000, 95% CI [78,482; 90,962], Table S3). However, demes were twice as connected in the NEA than in the NWA despite largely overlapping confidence intervals (Nm_E~118 95% CI [76.74; 144.91] and Nm_W ~ 61 95% CI [49.67; 124.05], respectively, Fig. 4C) suggesting high local connectivity within both meta-populations. Going backward in time, the two meta-populations were isolated until T_{CH} ~ 160,000 (95% CI [47,000; 163,000]) years B.P. when began an asymmetrical exchange of migrants three times greater from the NEA to the NWA than the reverse ($Nm_{E→W}$ ~ 5.1, 95% CI [4.88; 13.73] and $Nm_{W→E}$ ~ 1.5, 95% CI [1.69; 5.56] per generation). All lineages finally merged into an ancestral population of N_{ANC} ~ 101,000 (95% CI [99,116; 106,634]) at T_{DIV} ~ 891,000 (95% CI [800,000; 920,000]) years B.P., corresponding to the NEA-NWA divergence time (Table S3).

Origin of the supergene

Population structure results within the supergene revealed that the divergence between alleles HB and HS pre-dates the divergence between the NEA and NWA regions (Figs. 1 and 2). This could be due to an introgression of allele HB[39,40]. To test this hypothesis, we first computed a PCA within the supergene using *A. radiata* and a single representative of the congeneric species *A. hyperborea*. The analysis showed that the *A. hyperborea* individual was homozygous for allele HB as it clustered with HB/HB individuals from GoM, consistent with an introgression of HB (Fig. S8). We then computed rooted phylogenetic trees in 500 kb windows across chromosome 2 including *A. hyperborea* to test whether each tree followed a "species-tree" (Fig. 5A) or an "introgression tree" (Fig. 5B) using topology weighting[41]. The analysis strongly supported an introgression of HB: topological support was always of 100% for the introgression tree in the supergene region and almost always of ~100% for the species tree outside the region (Fig. 5C). This clearly suggests that HB allele is more closely related to *A. hyperborea* than to the HS allele. In addition, the Time to the Most Recent Common Ancestor (T_{MRCA}) extracted from each window's consensus tree was similar along the chromosome (Fig. 5D), which would be consistent with *A. hyperborea* being the donor species of HB. The PSMC computed in both HS/HS and HB/HB individuals within the supergene was strikingly different from the trajectory estimated over the whole genome (Fig. 4A, B). Similarly, the one hundred PSMC curves obtained by randomly sampling each time a 31 Mb region in the genome from both HB/HB and HS/HS individuals were incongruent with the supergene's PSMC curves but consistent with the genome-wide estimates (Fig. S9). Following[42], we further estimated the divergence time between alleles HB and HS at ~1.5 million years B.P. by computing the PSMC in heterozygote individuals in the supergene region (Fig. 4B).

Discussion

The striking size polymorphism in the thorny skate[26–29,43] offers a rare opportunity to improve our understanding of how the presence of two (or more) morphs may affect the population dynamics through time. We first identified a ~ 31 Mb size-determining supergene characterized by two alleles, HB and HS (the reference genome is built from an individual with the HS/HS genotype), between which recombination is prevented. HB was only found in the NWA where genotype distribution was strikingly spatially differentiated, with a significant deficit of

https://doi.org/10.1038/s41467-025-56126-z

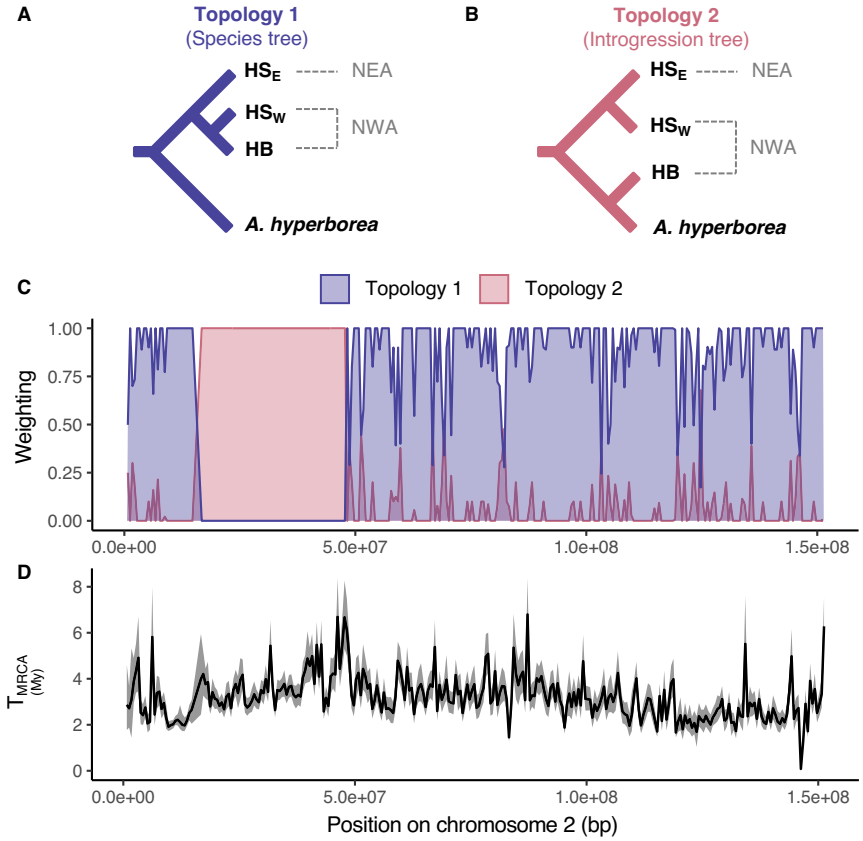

Fig. 5 | Characterization of the supergene's introgression. Species tree (**A**) and introgressed tree (**B**) topologies tested using TWISST algorithm. Panel (**C**): topology weighting in 500 kb non-overlapping windows along chromosome 2. Panel (**D**): distribution of the age in 500 kb non-overlapping windows along chromosome 2. The curve represents the median T_{MRCA} age across all individuals and the shaded area the associated 95% highest probability density interval for each window.

heterozygotes detected in the GoM but not in CAN. Heterozygotes and homozygotes for HB had an average size of ~67 cm, while homozygotes for HS had an average size of ~51 cm (Fig. 3B), suggesting the dominance of HB. Supergenes are known to determine a wide variety of traits[9–17], but this is a compelling example of the clear association between a supergene and a continuous quantitative phenotype in a group of individuals living in sympatry in a vertebrate species. Notably, size is a continuous trait with polygenic determinism across various species[44–46]. It is likely that the near-discrete size-determinism revealed in our study involves several genes spanning the supergene region, but we cannot exclude the interplay between them and other genes across the genome[47,48]. Given the substantial length of the supergene (~31 Mb) and the presence of numerous genes within it (226), it is also possible that this region controls multiple phenotypes, as exemplified in ants, cods, and butterflies[9–17]. More phenotypic data will be required to better characterize the consequences of this supergene. In addition, the mechanisms inherent to the variation in size should be investigated in the future as it can arise from different non-exclusive molecular processes, e.g., gene silencing, promoter and enhancer placement, differences in transcriptional efficiencies and their downstream impacts on dosage dependent epistasis and pleiotropy, as well as to a different number of genes present in the two alleles. Characterizing differences in gene expression linked to the observed genotypes as well as obtaining high quality de novo assembly of the HB haplotype to uncover fine scale structural variation (such as indels and multiple inversions), will provide valuable knowledge in the future.

Supergenes are maintained in space and time through a combination of demographic (neutral) and selective processes[6]. To shed light on the distribution of the two supergene alleles and of the maintenance of genotypes in the NWA, we investigated the historical demography of *A. radiata* through its whole range. Defining a comprehensive demographic scenario explaining the whole history of a species is a challenging task, but WGS data provides resourceful basis upon which to test competing models of demographic history. First, clustering algorithms, F_{ST} and PSMC analyses supported the unambiguous signature of long-term divergence between the NEA and NWA, and weak but spatially continuous genetic differentiation within each region (Figs. 1, 4, S1, S2), consistent with recent findings based on mitogenomic variation[31]. We further detected signatures of range expansion, with shared derived allele frequencies supporting an origin close to Iceland/Greenland and a two-wave colonization: first eastward to the European coasts, and then westward into the NWA (Fig. 1). This is consistent with the distribution of ROH, which were never larger than ~1.41 Mb and were fewer and shorter in the peripheral populations of Greenland (Fig. S5). Range expansions are indeed characterized by a series of founding events that result in higher genetic drift in populations far away from the origin of the expansion[49]. This pattern is consistent with the GoM, CAN and N-NW as the more derived and E-IC and W-IC as the more ancestral populations. The most likely scenario as inferred under the framework of fastsimcoal2[50] highlights a divergence between the NWA and NEA occurring ~900,000 years B.P., corresponding to the colonization time of the NWA. After colonization, the NWA and NEA metapopulations remained connected by asymmetrical migration, the rate of exchange being ~3 times higher from NEA to NWA than the reverse, until becoming isolated ~160,000 years B.P. The modelling of the genome wide diversity suggests that HB either originated or introgressed in the NWA regions more recently than ~160,000 years B.P., as this allele is absent in the NEA.

Sequencing of the congeneric species *A. hyperborea* provided strong evidence that HB did not originate in *A. radiata*, but rather

introgressed from a donor species. Indeed, *A. hyperborea* variants were more frequent in HB than in HS, which was confirmed by a topology weighting approach strikingly grouping HB/HB individuals and the one *A. hyperborea* individual in the supergene region (Fig. 5). Furthermore, clustering algorithms indicate that HB/HB individuals are separated from all HS/HS independently of their geographic origin (NWA or NEA) (Fig. 2F and S4), in stark contrast with the genome wide results (Fig. 1), and closer to *A. hyperborea* (Fig S8). This clearly suggests that the HB-HS divergence predates the NEA-NWA divergence. Indeed, the PSMC estimated the divergence between HB and HS at ~1.5 million years B.P. (Fig. 4B), likely corresponding to the separation between *A. radiata* and the donor species. Given the demographic scenario, the time separation between HB and HS as well as their present-day spatial distribution, we believe that the time when migration stopped between the NEA and NWA provides a reasonable upper limit to the HB introgression into NWA individuals. The donor species could be *A. hyperborea*, in line with the similar distribution of the T_{MRCA} inside and outside the supergene (Fig. 5). However, this would need confirmation by studying other *Amblyraja* species. For instance, we notice that no other *Amblyraja* species exhibit size polymorphism, and their size distribution generally resembles that of HB carriers except for *A. doellojuradoi*, which reaches a maximum size closer to that of HS/HS[51]. This suggests that multi-species investigations will be required to understand the dynamics of the evolution of this supergene, similar to studies on the origin of the social supergene in fire ants and the timing of associated introgression[39,52,53].

Demographic modelling of genome wide data highlighted high connectivity between the GoM and CAN sampling sites (Nm~61, Fig. 4C and Table S3), which explains why the two sites have similar level of genetic variability, as shown by θ values, Tajima's D (Table 1), as well as the genome-wide distribution of coalescence times estimated by the PSMC. However, demographic modelling cannot account for two differences observed between GoM and CAN: the genotype distribution at the supergene and the genome-wide excess of ROH in the former. Moreover, neither HB/HB nor HS/HS individuals in the NWA show a coalescence rate dynamic over time within the supergene consistent with the genome-wide pattern (Fig. 4, S6, S9). Supergenes can promote local adaptations even when gene flow is high[54], and more generally, local adaptations are usually maintained by various selective pressures[1,6,55]. Here, we argue that positive assortative mating for size polymorphism could explain both the observed deficit in heterozygotes in the GoM and the significantly greater inbreeding observed through ROH in GoM than in CAN (Fig. S5). We note that it is possible that the supergene controls other traits than size, which may also be under negative selection against heterozygotes in the GoM and thus explain their deficit. However, negative selection could not explain the inbreeding excess, suggesting that positive assortative mating remains the most parsimonious hypothesis to reconcile all the observed genetic characteristics. This is consistent with the previously discussed physical incompatibility in mating between larger and smaller skates in the GoM[31,32] due to evident differences in maximum size and size-at-maturity[25,26,28]. These differences tend to disappear northwards as skates sampled off the coast of Newfoundland (i.e., CAN sampling site) do not show a bi-modal distribution of size at first maturity as in the GoM, but rather a unimodal distribution associated with larger variance[28]. Maturity can covary with the environment[56] and could be a key factor in explaining possible mating between the size morphs in CAN but not in the GoM. This would have considerable implications for the trajectory and conservation of the NWA population in the context of climate induced environmental change as increasing sea temperature can alter age and size-at-maturity[57].

Positive assortative mating can have strong short-term consequences. First, it can lead to sympatric speciation[58]; however, high connectivity between the GoM and CAN could maintain recombination between HB and HS carriers (Figs. 2, 4 and Table S3) and thus explain the absence of neutral divergence between small and large individuals in the GoM (Fig. 1, S2). Second, it decreases the probability to find a mate in comparison to panmictic scenarios. We hypothesize that this could progressively lead to an Allee effect[59,60] in the GoM. For instance, Allee effects in marine organisms can occur due to exploitation[61], and we argue here that past overfishing coupled with positive assortative mating could have led to an extinction vortex in the GoM, explaining the recovery trends observed in CAN (Newfoundland) but not in the GoM. All in all, this highlights the significant short-term and context-dependent effects related to phenotypes determined by a supergene locus in a Vulnerable population. We stress that our hypothesis will require direct validation in the future, as our data cannot provide explicit evidence of an Allee effect linked to supergene-determined phenotypes.

We discovered a size-determining supergene and demonstrated its introgression in *A. radiata* in the last ~160,000 years from a donor species. Our work brings light to new findings of general interest in evolutionary biology: (i) we provide a unique direct example of a continuous quantitative trait whose distribution is largely explained by a simple Mendelian inheritance in a vertebrate species. It is likely that other genomic regions (in conjunction with the environment) contribute to the determination of size, but the cluster of genes controlling such a complex trait will provide an opportunity to dissect its genetic architecture. (ii) Supergenes are known to alter mating patterns as a function of spatial heterogeneity[7], but for the first time we are able to hypothesize a link between a supergene system and population survival. This is crucial as one relevant question in evolutionary biology is how supergene alleles are maintained through time. In addition, here, we speculate that changing environmental conditions (warming temperatures) may ultimately lead to the loss of HB (as, in the worst scenario, HS would still be present in NEA). (iii) Finally, our study demonstrates (once more) the importance of reconstructing the neutral evolutionary history of a species, an essential background needed to uncover complex non-neutral processes. The inferred demographic scenario was of paramount importance not only to interpret the spatial distribution of the two alleles of the supergene but also to issue a hypothesis for genotype distribution in the NWA, which in turn carry profound implications for the conservation of the thorny skate[1].

Methods

Whole genome sequencing

Forty-nine *Amblyraja radiata* individuals were sampled from ten regions throughout the North Atlantic including the US Gulf of Maine (GoM, $N = 16$), Newfoundland, Canada (CAN, $N = 5$), South West Greenland (SW-G, $N = 2$), South East Greenland (SE-G, $N = 3$), East Greenland (E-GR, $N = 2$), Western Iceland (W-IC, $N = 5$), Eastern Iceland (E-IC, N = 5), Western Norway (W-NW, $N = 1$), Southern Norway (S-NW, $N = 1$), and Northern Norway (N-NW, $N = 9$). Genomic DNA was extracted using the E.Z.N.A. Tissue DNA Kit (Omega Bio-Tek, Inc., Norcross, GA, USA) following the manufacturer's instructions. The extracted DNAs were then sent to the Next-Generation Sequencing (NGS) Core of the University of Florida's Interdisciplinary Center for Biotechnology Research (UF ICBR) for quality control. After that, libraries were prepared, pooled, and loaded on the Illumina NovaSeq 6000 platform for whole genome sequencing with S4 flow cell and 2×151 setup.

Bioinformatics

We used the reference genome of the thorny skate available from the NCBI website (sAmbRad1.1.pri; accession GCF_010909765.2). The genome was first masked using the Chondrichthyes database in a first run of *RepeatMasker* v.4.1.0[62]. We then created a de novo database for

A. radiata by using *RepeatModeler* v.2.0.3[63] on the genome masked at the first step. Then, we masked the repeated elements annotated in the de novo database by running *RepeatMasker* a second time on the initially masked genome. We finally extracted a bed-file of the masked regions further used in downstream bioinformatic analyses.

Illumina reads for the 49 samples were trimmed for adapter and quality using *bbduk* from *bbmap* v.38.44 suite (sourceforge.net/projects/bbmap/). After checking for quality using *FastQC* v0.11.7[64], reads were mapped against the reference genome using bwa mem algorithm v.0.7.17[65] with -M option. Mapped reads were sorted and indexed using samtools v.1.10[66] and then marked for duplicates using Picard v.2.21.2 MarkDuplicates[67]. Except for the PSMC analysis (see below), indexed reads were fed to the haplotypecaller algorithm in GATK v.4.1.9.0[68] for variant discovery using the -gvcf option to obtain individual variant calling files (VCF) with annotations for all sites. Individual VCFs were then combined using CombineGVCF to build different datasets according to the downstream analysis (see below). Joint calling was then performed for each dataset using GenotypeGVCF by including both monomorphic and polymorphic sites (all-sites argument) which are necessary for scaling genetic diversity correctly. We then selected the 49 identified autosomes and removed the regions annotated as repeats using the bed-file produced by the repeat masking step. By combining VariantFiltration and SelectVariant GATK's scripts, we then filtered out sites with Mapping Quality <40 and marked genotypes as missing if genotypic depth (i.e., depth per individual and per site) was below 6 or over 50. We further removed chromosome 2 and 8 for all genome-wide historical demographic analyses after genomic scans identified two potential large chromosomal inversions (Table S4). Additional filters were applied on the resulting VCF depending on the analysis (described below).

Population structure

Population structure datasets were filtered using a combination of vcftools v.0.1.16[69], bcftools v.1.15 and custom python v.3.8 scripts, keeping only bi-allelic SNPs with a missing data rate of less than 20% and discarding SNPs heterozygous in more than 80% individuals (removing potential paralogous loci) and with a minor allele frequency <0.05. VCFs were binned by only selecting SNPs that were at least 1 kb apart to account for linkage disequilibrium. Depending on the analysis and on the scale of investigation, we built different datasets. We first built a dataset including all individuals: ALL ($N = 49$). Based on global population structure (see results), two additional datasets were created to investigate fine scale structure: the NWA dataset ($N = 21$), which only included individuals from CAN and GoM sampling locations, and the NEA dataset ($N = 28$) including all the remaining individuals (Fig. 1 and Table 1). We performed a PCA and ran the sNMF algorithm, both implemented in the R package LEA[34] on each dataset separately. The sNMF algorithm is a clustering algorithm that determines the most likely number of K ancestral populations best describing the genomic variability and infers individual admixture proportions under the selected model. The algorithm was run 10 times with values of K ranging from 1 to 6, and we chose the most likely model as the one associated with the lowest cross-entropy value. We built a final dataset to quantify genetic differentiation between sampling locations (dataset FST, $N = 40$) using F_{ST} metrics, including only sampling locations with $N \geq 5$ (Fig. 1, Table 1). We computed Hudson's estimator of pairwise-F_{ST}[70] between each location using a custom R script and evaluated significance by randomly permuting individuals 1000 times for each comparison.

Genetic diversity

Genetic diversity datasets were filtered using custom bash and python scripts, keeping only biallelic sites with no missing data and removing indels and SNPs heterozygous in more than 80% individuals. We built one dataset per sampling location of $N \geq 5$ from which we computed the folded site frequency spectrum (SFS) using a custom python script. Using custom R scripts, we then computed Tajima's D (TD)[71], and two estimators of θ, namely the mean pairwise difference θ_π[72] and Watterson's θ_w[73], both standardized by the total number of called sites (i.e., monomorphic sites included). We investigated the influence of binning the dataset on the reconstructed SFS and genetic diversity estimates by sampling regions of 100 bp (to account for monomorphic sites) apart from 1 kb, 10 kb, 50 kb or 100 kb in the GoM (the sampling location with the more individuals). All statistics remained similar (see supplementary results, Fig. S10) and since accuracy in demographic inferences is improved by having more data[74], we decided not to bin datasets for historical demographic reconstructions (see below).

Detection of a supergene

Considering the high degree of genetic differentiation between the NEA and NWA but low within NWA region (see results), we performed genomic scans at the NWA scale (including the GoM and CAN sampling locations) to find regions putatively related to the size polymorphism. We first scanned the genome using an approach based on PCAs which does not require any prior grouping information (here, phenotypes). PCA can detect genomic regions of more than average global population structure and is particularly useful when looking for association with a complex trait which requires a large sample size for a robust characterization. We ran the algorithm implemented by the Hierfstat package[75] on each chromosome in sliding windows of 100 kb with a jump of 10 kb on the NWA dataset. The proportion of variance explained by PC1 was extracted for windows with more than 50 SNPs and plotted against the location on the chromosomes. This analysis revealed a region of high divergence on chromosome 2 (see Results) in which we further computed linkage disequilibrium using the r^2 statistic between SNPs 50 kb apart (to avoid computational burden) using the LDheatmap R package[76]. Local PCA, sNMF, and the analysis of genotype distribution highlighted the occurrence of a bi-allelic supergene (HB and HS) in which all the three possible genotypes (HB/HB, HB/HS, HS/HS) were present (Fig. 2). Additionally, we computed sliding windows of pairwise-F_{ST} between the two clusters composed of individuals homozygote at the supergene in the NWA dataset (see Fig. 2) using windows of 10 kb with a jump of 5 kb. Finally, we ran the local PCA and sNMF in the supergene region using the ALL dataset.

Genotype screening and size association

Preliminary assessment of the relationship between size and haplotypes suggested an association between the supergene genotypes and size (see Results). To directly test the relationship between size and supergene polymorphism, we identified two regions: from 25,075,452 to 25,075,619 (167 bp) and from 41,404,405 to 41,404,539 (134 bp) with respectively 5 and 4 SNPs discriminating the two supergene alleles. Primers were designed in 500 bp flanking our target regions and used to amplify 501 individuals sampled from the whole range (Table 1). The regions with five and four discriminating SNPs were hereafter referred to as "Region 051" and "Region 034", respectively. The primers designed for "Region 051" (051 F: 5′- CGG CAG TTS ACC ATC TTA GA -3′; 051 R: 5′- GCT TGT AAC CAC ACT GCT -3′) are targeting a fragment of ~280 bp in length. The primers designed for "Region 034" (034 F: 5′- GTA TGG AGT ACC ACC TTG AAT G -3′; 034 R: 5′- GGT TGA TGT ATC TGC TGT AAG -3′) are targeting a fragment of ~760 bp in length. PCR reactions were carried out in 25 μL tubes by adding 14.775 μL of PCR grade water, 2.5 μL of PCR buffer, 2.0 μL of MgCl2 (25 mM), 2.0 μL of dNTP mix (2.5 mM each), 0.8 μL of each primer (10 μM), 0.125 μL of GoTaq® Hot Start Polymerase (Promega, Madison, WI, USA; 5 U/μL) and 2 μL of DNA template. The reaction mix was denatured at 94 °C for

2 min, followed by 35 cycles of denaturation at 94 °C for 30 s, annealing at 52 °C (Region 051) or 52 °C (Region 034) for 30 s and extension at 72 °C for 60 s. PCR products were sent off to Retrogen Inc. (San Diego, CA, USA) for purification and sequencing. Genotypes for the 9 discriminating SNPs were attributed by visual assessment of base sequencing peaks. Genotypes were attributed a NA value when the peak was ambiguous. Individuals with missing genotypes or for which the supergene genotype could not be determined were discarded (Supplementary Data 1 reports the filtered genotype data). We then tested whether the genotypes at the supergene were at the Hardy Weinberg equilibrium in sampling locations where the supergene was polymorphic using the HardyWeinberg R package[77] exact-test.

We tested for the association between size and haplotype by accounting for maturity stage and sex. We filtered out individuals with missing information for any of these traits, which yielded only individuals from the GoM ($N = 241$). We designed three linear models using the Bayesian framework implemented in the R package brms[78]. The richest model in parameters, model GenoMatSex, included Genotype, Maturity and Sex traits as determining variables ("Size ~ Genotype + Maturity + Sex"). The two other models were nested within GenoMatSex, removing the variable "Sex" for model GenoMat and both "Sex" and "Maturity" variables for model Geno. Four MCMC runs, each of 10,000 iterations with 1,000 warmup samples and a thinning of 4 iterations were performed for each model, using flat priors. We assessed which model was the most accurate by performing the approximate leave-one-out (LOO) cross validation implemented in brms. Median values and the 95% credible interval of the posterior distributions of size under the best model were averaged on the levels of the other variables by using emmeans R package[79].

Range expansion and genomic inbreeding

Range expansions (RE) occur by serial founding events leading to the fixation of derived alleles along the colonization process[35]. Areas located further away from the origin of the RE are therefore expected to display stronger linkage disequilibrium and higher frequency of fixed derived alleles. These patterns can subsequently be used to infer the colonization dynamics of a species. To investigate this, we first investigated the signatures of shared derived alleles across the whole range. However, such analysis requires polarizing the allelic state. To that end, we performed long-read sequencing (PacBio HiFi) of one individual of the congeneric species *Amblyraja hyperborea* using the PacBio Sequel IIe System at the NGS Core of UF ICBR. A total of two SMRT cell runs (each generates 3–5 million reads) have been performed. HiFi long reads were then mapped using the pbmm2 v.1.3.0 align subcommand (https://github.com/PacificBiosciences/pbmm2) with the HIFI preset. Mapped reads were then sorted and indexed using samtools. Variants were called using deepvariant v.1.4.0[80] by applying the PacBio model and specifying polymorphic and monomorphic sites as output. Using bcftools, we then merged the long-read VCF to a dataset including all the short-read *A. radiata* individuals previously processed for genetic diversity analyses. We then filtered out non-bi-allelic sites and polarized the remaining variable sites based on the *A. hyperborea* individuals state (i.e., the outgroup individual, hereafter referred to as OG) by: (1) discarding sites heterozygous in OG; (2) recoding all allele(s) as ANC (ancestral, coded as "0") when corresponding to the allele for which OG is homozygote and as DER (derived, coded as "1") otherwise. The derived allele frequency was calculated per individual and the average value for each sampling location was reported. Based on the number of derived alleles per individual and per site, we calculated the directionality index (ψ)[35,81], which is the pairwise difference between shared derived alleles and is expected to be different from 0 if there is a signature of range expansion. The TDOA location algorithm of Peter and Slatkin[35,81] was run on the pairwise ψ matrix to identify the RE origin. We ran the algorithm 100 times using one random individual from each location and displayed results as a density of the location of RE.

We then assessed inbreeding by investigating Runs of Homozygosity (ROH) signatures using the HMM model implemented in bcftools-ROH[82]. The analysis was run for each sampling location with $N \geq 5$ separately, by specifying bcftools-ROH to estimate allele frequencies from the observed genotypes. ROH were classified into three arbitrary length categories to investigate changes in signals related to the length of ROH: ROH shorter than 10 kb, ROH of length between 10 kb and 20 kb, and ROH larger than 20 kb. We then plotted both the number (N_{ROH}) and the sum of ROH (SUM_{ROH}) for each class and for each sampling location.

Historical demography

We first investigated the variation of the coalescence rate through time by using the Pairwise Sequential Markovian Coalescent (PSMC) model on each individual. To that end, we first called SNPs from each bam file using bcftools to obtain one VCF per individual. Each VCF was masked for repeats using bedtools[83]. Using scripts provided with the PSMC[36], we filtered for the depth of coverage using the parameters -d 6 and -D 50 and extracted the consensus sequence that was fed to the PSMC algorithm using the following parameters: -t15 -N25 -r5 -p "6 + 30*2 + 4 + 6". Because PSMC curves were highly similar in each cluster (see results), we computed 100 bootstraps for one individual in each cluster (i.e., one for the NEA and one for the NWA).

We investigated historical demography scenarios (Fig. 4 and S7) by using the composite likelihood approach of fastsimcoal 2.7[38] to further investigate migration and divergence patterns at the scale of the range distribution. Model IMM-5 depicts an Isolation-Migration Metapopulation scenario with 5 demes connected in a one-dimensional stepping-stone fashion (i.e., migrants are only exchanged with direct neighbors). The five demes refer to the sampling locations with $N \geq 5$ (Table 1) and are spread across two metapopulations corresponding to the NWA (GoM and CAN) and NEA (W-ICE, E-ICE and NOR) genetic clusters (see results). In the NWA, the two demes have a size of N_W and exchange Nm_W migrants per generation with each other. Similarly, demes have a size of N_E in the NEA and exchange Nm_E migrants per generation with the closest neighbor. The two regions are connected by an asymmetrical exchange of migrant of $Nm_{W\to E}$ from CAN to W-ICE and $Nm_{E\to W}$ from W-ICE to CAN (Fig. S7). Going backwards in time, all demes merge into an ancestral population of size N_{ANC} at T_{DIV} generations ago. Two scenarios based on IMM-5 topology were additionally tested. Going back in time, the IMM-5-NM-CH model describes a change in connectivity between the NWA and NEA happening T_{CH} generations ago, going from $Nm_{W\to E\text{-}MOD}$ and $Nm_{E\to W\text{-}MOD}$ from the present to T_{CH} to $Nm_{W\to E\text{-}ANC}$ to $Nm_{E\to W\text{-}ANC}$ from T_{CH} to T_{DIV}. IMM-5-NM-STOP is similar to IMM-5-NM-CH but NWA and NEA are isolated from the present to T_{CH}, with the two regions being then connected by $Nm_{W\to E}$ and $Nm_{E\to W}$ from T_{CH} to T_{DIV}. Additional scenarios were tested to investigate whether adding unsampled demes better depicted the genetic variability due to meta-population structure because the five sampled demes do not cover the whole range distribution of the species. IMM-20 is similar to IMM-5 but includes unsampled demes so that each of the two regions are composed of $D = 10$ demes, resulting in a 20-demes 1D-stepping-stone matrix. From west to east, the GoM and CAN are respectively sampled at demes 2 and 8 in the NWA and W-IC, E-IC and N-NW at demes 4, 5 and 8 in the NEA, and the asymmetrical gene flow from the NWA to NEA occurs from demes 10 and 11 and vice-versa. IMM-30 is similar to IMM-20, but the NEA is composed of $D = 20$ demes, with W-IC, E-IC and N-NW, respectively sampled at deme 8, 10 and 17. This model was investigated to account for the likely different number of demes in each metapopulation given the larger geographical area covered by the NEA cluster (i.e., from Greenland to Norway, see results). The fastsimcoal algorithm is based on the modelling of a set of two-dimensional SFS

(2D-SFS) between sampled locations. To that end, we built a dataset with all individuals from CAN, E-IC, and W-IC and a random subset of 5 individuals from N-NW and the GoM to get a balanced sampling scheme. Using a custom R script, we processed the dataset using the same filters as for the genetic diversity datasets and computed a 2D-SFS between each sampling location. The set of observed SFS were maximized using 100,000 coalescent simulations (-n 100,000), 40 expectation-maximization cycles (-L 40) and by considering at least 10 entry counts in the SFS to perform parameter estimation (-C 10). We performed 10 independent runs for each scenario and used the best run (i.e., with the highest likelihood) to compute the AIC to perform model selection. Moreover, we computed the likelihood distribution for each scenario by simulating 100 replicates under the previously estimated best set of parameters: this procedure is necessary to determine whether the different scenarios are distinguishable or not. We then computed a confidence interval for parameter values for the model with the lowest AIC. To that end, we calculated 100 non-parametric bootstrapped 2D-SFS by randomly sampling blocks of 10,000 bp with replacement using a custom R script. Each set of bootstrapped 2D-SFS were maximized following the same procedure applied to the observed set of 2D-SFS. The 95% confidence intervals were calculated from the distribution of the best ML estimates for each bootstrap set. All historical demography inferences were performed using a mutation rate $\mu = 2.01e{-}8$ per site and per generation following a generation time of 11 years (average time at maturity, COSEPAC, 2012) and the genomic mutation rate estimated for the chondrichthyan species *Carcharhinus melanopterus*[37].

Origin of the supergene

We ran the PSMC algorithm within the supergene region for a randomly selected representative HB/HB, HS/HS and HS/HB individual to detect when the divergence between the two haplotypes occurred. The PSMC estimates the distribution of coalescence times along the genome between two chromosomes: in a non-recombining block such as our supergene, this amounts to compute the divergence between the two alleles, which graphically corresponds to the time when the Ne suddenly increased to infinite in heterozygous individuals[42]. To investigate whether the PSMC run in the supergene region reproduced the signal in the whole genome datasets, we randomly sampled 100 regions of 31 Mb (the size of the supergene) spread in the genome on which we ran the PSMC. We used the dataset polarized with *A. hyperborea* to perform sliding window analyses of the ancestral allele frequency for HB/HB and HS/HS NWA individuals in chromosome 2. We then performed a PCA using the ALL dataset that was merged with the *A. hyperborea* individual data (MERGED dataset) and subset to take only SNPs within the supergene. The dataset was filtered similarly to population structure datasets. We then computed rooted phylogenetic trees per 500 kb non-overlapping window along the chromosome 2 to investigate discrepancies between the supergene and the genome-wide population trees using the MERGED dataset. All missing data were filtered out as well as data from individuals heterozygous at the supergene (GN19146 and GN19147), resulting in windows of ~367 SNPs on average (Total $N_{SNP} = 109{,}501$). For each window, trees were inferred using BEAST v2.7.6[84] using a MCMC chain length of 50,000,000 iterations and a thinning of 1000 iterations. All windows individually had a GTR substitution model with gamma distributed rates across sites ($\alpha = 0.05$ and $\beta = 10.0$). The tree topology followed a birth-death prior with unscaled tree type, and with both birth and death rates uniformly distributed (respectively lower=0.0 and upper=1000.0 and lower=0.0 and upper=1.0). We then computed a consensus tree per window with TreeAnnotator v2.7.5, using default parameters and a burnin of 40%. To investigate how the topology of the tree varied across the chromosome, we used the topology weighting algorithm implemented in TWISST[41]. The fifteen possible rooted topologies (for 4 taxa) were investigated, and weights are reported for the two most common topologies found across the chromosome (Fig. 5A-B): (1) a scenario where HB is introgressed ((HS_{NEA}, HS_{NWA}), (HB, *A. hyperborea*)) and (2) a species-tree scenario with *A. hyperborea* as outgroup ((HS_{NEA}, (HB, HS_{NWA})), *A. hyperborea*).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw sequence data generated in this study have been deposited on GeneBank under the BioProject ID: PRJNA1189519 (https://www.ncbi.nlm.nih.gov/bioproject/1189519).

Code availability

Custom R and Python scripts used in this study are publicly available at https://github.com/PierreLesturgie/Py-PopGen[85] and https://github.com/PierreLesturgie/R-PopGen[86].

References

1. Wellenreuther, M. & Bernatchez, L. Eco-evolutionary genomics of chromosomal inversions. *Trends Ecology Evolut.* **33**, 427–440 (2018).
2. Faria, R., Johannesson, K., Butlin, R. K. & Westram, A. M. Evolving Inversions. *Trends Ecology Evolut.* **34**, 239–248 (2019).
3. Hoffmann, A. A. & Rieseberg, L. H. Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Ann. Rev. Ecol. Evolut. Syst.* **39**, 21–42 (2008).
4. Kirkpatrick, M. How and why chromosome inversions evolve. *PLoS Biol.* **8**, e1000501 (2010).
5. Dobzhansky, T. & Sturtevant, A. H. Inversions in the chromosomes of drosophila pseudoobscura. *Genetics* **23**, 28–64 (1938).
6. Thompson, M. J. & Jiggins, C. D. Supergenes and their role in evolution. *Heredity (Edinb.)* **113**, 1–8 (2014).
7. Schwander, T., Libbrecht, R. & Keller, L. Supergenes and complex phenotypes. *Curr. Biol.* **24**, R288–R294 (2014).
8. Clarke, C. A. & Sheppard, P. M. Super-genes and mimicry. *Heredity (Edinb.)* **14**, 175–185 (1960).
9. Chapuisat, M. Supergenes as drivers of ant evolution. *Myrmecol N.* **33**, 1–18 (2023).
10. Brelsford, A. et al. An ancient and eroded social supergene is widespread across Formica Ants. *Curr. Biol.* **30**, 304–311.e4 (2020).
11. Lagunas-Robles, G., Purcell, J. & Brelsford, A. Linked supergenes underlie split sex ratio and social organization in an ant. *Proc. Natl. Acad. Sci.* **118**, e210142711 (2021).
12. Avril, A., Purcell, J., Brelsford, A. & Chapuisat, M. Asymmetric assortative mating and queen polyandry are linked to a supergene controlling ant social organization. *Mol. Ecol.* **28**, 1428–1438 (2019).
13. Kay, T., Helleu, Q. & Keller, L. Iterative evolution of supergene-based social polymorphism in ants. *Philos. Trans. R. Soc. B: Biol. Sci.* **377**, 20210196 (2022).
14. Berg, P. R. et al. Adaptation to low salinity promotes genomic divergence in Atlantic Cod (Gadus morhua L.). *Genome Biol. Evol.* **7**, 1644–1663 (2015).
15. Barney, B. T., Munkholm, C., Walt, D. R. & Palumbi, S. R. Highly localized divergence within supergenes in Atlantic cod (Gadus morhua) within the Gulf of Maine. *BMC Genomics* **18**, 271 (2017).
16. Berg, P. R. et al. Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Sci. Rep.* **6**, 23246 (2016).

17. Joron, M. et al. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**, 203–206 (2011).
18. Joron, M., Wynne, I. R., Lamas, G. & Mallet, J. Variable selection and the coexistence of multiple mimetic forms of the Butterfly Heliconius numata. *Evol. Ecol.* **13**, 721–754 (1999).
19. Kulka, D. W. et al. Amblyraja radiata. *The IUCN Red List of Threatened Species* (2020).
20. Sosebee, K. A., Miller, A., O'brien, L., Mcelroy, D. & Sherman, S. Update of thorny skate (*Amblyraja radiata*) commercial and survey data. *Northeast Fisheries Science Center reference document; 16-08* https://doi.org/10.7289/V5/RD-NEFSC-16-08 (2016).
21. Grieve, B. D., Hare, J. A. & McElroy, W. D. Modeling the impacts of climate change on thorny skate (Amblyraja radiata) on the Northeast US shelf using trawl and longline surveys. *Fish. Oceanogr.* **30**, 300–314 (2021).
22. Pennino, M. G., Guijarro-García, E., Vilela, R., Del Río, J. L. & Bellido, J. M. Modeling the distribution of thorny skate (*Amblyraja radiata*) in the southern grand banks (Newfoundland, Canada). *Can. J. Fish. Aquat. Sci.* **76**, 2121–2130 (2019).
23. Swain, D. P. & Benoît, H. P. Change in habitat associations and geographic distribution of thorny skate (*Amblyraja radiata*) in the southern Gulf of St Lawrence: Density-dependent habitat selection or response to environmental change? *Fish. Oceanogr.* **15**, 166–182 (2006).
24. Swain, D. P., Jonsen, I. D., Simon, J. E. & Davies, T. D. Contrasting decadal trends in mortality between large and small individuals in skate populations in Atlantic Canada. *Can. J. Fish. Aquat. Sci.* **70**, 74–89 (2013).
25. Sosebee, K. A. Maturity of skates in Northeast United States waters. *J. Northwest Atl. Fish. Sci.* **35**, 141–153 (2005).
26. Sulikowski, J. A. et al. Age and growth estimates of the thorny skate (*Amblyraja radiata*) in the western Gulf of Maine. *Fishery Bull.* **103**, 161–168 (2005).
27. Mcphie, R. P. & Campana, S. E. Reproductive characteristics and population decline of four species of skate (Rajidae) off the eastern coast of Canada. *J. Fish. Biol.* **75**, 223–246 (2009).
28. Templeman, W. *Differences in Sexual Maturity and Related Characteristics Between Populations of Thorny Skate (Raja Radiata) in the Northwest Atlantic*. *J. Northw. Atl. Fish. SeL* **7** http://journal.nafo.int (1987).
29. Templeman, E. G. Variations in Numbers of Median Dorsal Thorns and Rows of Teeth in Thorny Skate (*Raja radiata*) of the Northwest Atlantic. *J. Northwest Atl. Fish. Sci.* **5**, 171–179 (1984).
30. Chevolot, M. et al. Population structure and historical demography of the thorny skate (*Amblyraja radiata*, Rajidae) in the North Atlantic. *Mar. Biol.* **151**, 1275–1286 (2007).
31. Denton, J. S. S. et al. Mitogenomic evidence of population differentiation of thorny skate, *Amblyraja radiata*, in the North Atlantic. *J. Fish Biol.* https://doi.org/10.1111/jfb.15689 (2024).
32. Lynghammar, A., Præbel, K., Bhat, S., Fevolden, S. & Christiansen, J. Widespread physical mixing of starry ray from differentiated populations and life histories in the North Atlantic. *Mar. Ecol. Prog. Ser.* **562**, 123–134 (2016).
33. Lynghammar, A. et al. DNA barcoding of the northern Northeast Atlantic skates (Chondrichthyes, Rajiformes), with remarks on the widely distributed starry ray. *Zool. Scr.* **43**, 485–495 (2014).
34. Frichot, E. & François, O. LEA: An R package for landscape and ecological association studies. *Methods Ecol. Evol.* **6**, 925–929 (2015).
35. Peter, B. M. & Slatkin, M. Detecting range expansions from genetic data. *Evolution (N. Y)* **67**, 3274–3289 (2013).
36. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
37. Lesturgie, P., Planes, S. & Mona, S. Coalescence times, life history traits and conservation concerns: An example from four coastal shark species from the Indo-Pacific. *Mol. Ecol. Resour.* **22**, 554–566 (2022).
38. Excoffier, L. et al. Fastsimcoal2: Demographic inference under complex evolutionary scenarios. *Bioinformatics* **37**, 4882–4885 (2021).
39. Stolle, E. et al. Recurring adaptive introgression of a supergene variant that determines social organization. *Nat. Commun.* **13**, 1180 (2022).
40. Jay, P. et al. Supergene Evolution Triggered by the Introgression of a Chromosomal Inversion. *Curr. Biol.* **28**, 1839–1845.e3 (2018).
41. Martin, S. H. & Van Belleghem, S. M. Exploring evolutionary relationships across the genome using topology weighting. *Genetics* **206**, 429–438 (2017).
42. Cahill, J. A., Soares, A. E. R., Green, R. E. & Shapiro, B. Inferring species divergence times using pairwise sequential markovian coalescent modelling and low-coverage genomic data. *Philos. Trans. R. Soc. B: Biol. Sci.* **371**, 20150138 (2016).
43. Templeman, W. Migrations of Thorny Skate, *Raja radiata*, Tagged in Newfoundland. *J. Northwest Atl. Fish. Sci.* **5**, 55–63 (1984).
44. Bouwman, A. C. et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat. Genet* **50**, 362–367 (2018).
45. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet* **46**, 1173–1186 (2014).
46. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
47. Jones, R. T., Salazar, P. A., Ffrench-Constant, R. H., Jiggins, C. D. & Joron, M. Evolution of a mimicry supergene from a multilocus architecture. *Proc. R. Soc. B: Biol. Sci.* **279**, 316–325 (2012).
48. Errbii, M. et al. Evolutionary genomics of socially polymorphic populations of Pogonomyrmex californicus. *BMC Biol.* **22**, 109 (2024).
49. Slatkin, M. & Excoffier, L. Serial founder effects during range expansion: A spatial analog of genetic drift. *Genetics* **191**, 171–181 (2012).
50. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP Data. *PLoS Genet* **9**, e1003905 (2013).
51. Last, P. R. et al. *Rays of the World*. (CSIRO PUBLISHING, 2016).
52. Yan, Z. et al. Evolution of a supergene that regulates a trans-species social polymorphism. *Nat. Ecol. Evol.* **4**, 240–249 (2020).
53. Helleu, Q., Roux, C., Ross, K. G. & Keller, L. Radiation and hybridization underpin the spread of the fire ant social supergene. https://doi.org/10.1073/pnas (2022).
54. Schaal, S. M., Haller, B. C. & Lotterhos, K. E. Inversion invasions: When the genetic basis of local adaptation is concentrated within inversions in the face of gene flow. *Philos. Trans. R. Soc. B: Biol. Sci.* **377**, 20210200 (2022).
55. Berdan, E. L. et al. Genomic architecture of supergenes: Connecting form and function. *Philos. Trans. R. Soc. B: Biol. Sci.* **377**, 20210192 (2022).
56. Martin, S. B. & Leberg, P. L. Influence of environmental stress on age- and size-at-maturity: Genetic and plastic responses of coastal marsh fishes to changing salinities. *Can. J. Fish. Aquat. Sci.* **68**, 2121–2131 (2011).
57. Niu, J., Huss, M., Vasemägi, A. & Gårdmark, A. Decades of warming alters maturation and reproductive investment in fish. *Ecosphere* **14** (2023).
58. Straw, R. M. Hybridization, homogamy, and sympatric speciation. *Evolution (N. Y)* **9**, 441–444 (1955).
59. Allee, W. C. The social life of animals. *W.W. Norton & Company, inc*. https://doi.org/10.5962/bhl.title.7226 (1938).

60. Stephens, P. A., Sutherland, W. J. & Freckleton, R. P. *What Is the Allee Effect?* vol. 87 https://www.jstor.org/stable/3547011 (1999).
61. Gascoigne, J. & Lipcius, R. N. Allee effects in marine systems. *Mar. Ecol. Prog. Ser.* **269**, 49–59 (2004).
62. Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0*. 2013-2015 http://www.repeatmasker.org.
63. Smit, AFA, Hubley, R. *RepeatModeler Open-1.0*. 2008-2015 http://www.repeatmasker.org.
64. Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Babraham Bioinformatics* (2010).
65. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **00**, 1–3 (2013).
66. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
67. Broad Institute. Picard Toolkit. *GitHub Repository* https://broadinstitute.github.io/picard/ (2019).
68. McKenna, A. et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
69. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
70. Hudson, R. R. *Properties of a Neutral Allele Model with Lntragenic Recombination*. *POPULATION BIOLOGY* vol. 23 (1983).
71. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
72. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
73. Watterson, G. A. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul Biol.* **7**, 256–276 (1975).
74. Felsenstein, J. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? *Mol. Biol. Evol.* **23**, 691–700 (2006).
75. Goudet, J. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **5**, 184–186 (2005).
76. Shin, J.-H., Blay, S., Mcneney, B. & Graham, J. *LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms*. *JSS J. Stat. Softw.* **16** http://www.jstatsoft.org/ (2006).
77. Graffelman, J. *Journal of Statistical Software Exploring Diallelic Genetic Markers: The HardyWeinberg Package*. 64 http://www.jstatsoft.org/ (2015).
78. Bürkner, P. C. Bayesian Item Response Modeling in R with brms and Stan. *J. Stat. Softw.* **100**, 1–54 (2021).
79. Lenth, R. emmeans: Estimated Marginal Means, aka Least-Squares Means. *American Statistician* vol. 34 (2024).
80. Poplin, R. et al. A universal snp and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983 (2018).
81. Peter, B. M. & Slatkin, M. The effective founder effect in a spatially expanding population. *Evolution (N. Y)* **69**, 721–734 (2015).
82. Narasimhan, V. et al. BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749–1751 (2016).
83. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
84. Bouckaert, R. et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).
85. Lesturgie, P. Python programs to analyse and filter population genomics data in "*Short-term Evolutionary Implications of an Introgressed Size-Determining Supergene in a Vulnerable Population*". PierreLesturgie/Py-PopGen v1.0.0, https://doi.org/10.5281/zenodo.14170049 (2024).
86. Lesturgie, P. R scripts to analyse and filter population genomics data. PierreLesturgie/R-PopGen v1.0.0, https://doi.org/10.5281/zenodo.14170055 (2024).
87. https://www.R-project.org/ R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Acknowledgements

We are grateful to the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12), to the HPC resources of the SACADO MeSU platform at Sorbonne Université (https://sacado.sorbonne-universite.fr/mesu/), the University of Florida Research Computing (http://www.rc.ufl.edu) and the Plateforme de Calcul Intensif et Algorithmique (PCIA), Muséum National d'Histoire Naturelle, Centre National de la Recherche Scientifique (http://uar2700.mnhn.fr/fr/pcia-9024) for providing computational and storage resources that have contributed to the research results reported in this publication. We also acknowledge the Vertebrate Genome Project and The College of Charleston who contributed to the research results reported in this study. This work was funded by Lenfest Foundation Award GR-000002881 (Pew Charitable Trusts Contract ID #30884) attributed to GN; NSF grant DEB-1541556 attributed to GN and NSF grant IOS-2232269 attributed to GN. This research was conducted without financial support from Colossal Biosciences. We are especially grateful to Jason Landrum, Charlotte Hudson and the team at the Lenfest foundation for all their support over the course of this project. We thank Klara Jakobsdóttir, Carolyn Miri, W. David McElroy, the Norwegian Institute of Marine Research, Jan Yde Poulsen, Heino Fock and the TUNU-Programme at UiT The Arctic University of Norway for sample collection. This study complies with all relevant ethical regulations.

Author contributions

Conceptualization: P.L., S.M., G.N. Data Curation: P.L. Formal Analysis: P.L. Funding Acquisition: G.N. Investigation: P.L., J.D., L.Y., S.C., J.K., O.F. Methodology: P.L. Resources: J.K., A.L., G.N. Software: P.L. Validation: P.L., G.N., S.M. Visualization: P.L. Supervision: S.M., G.N. Writing—original draft: P.L., S.M., G.N. Writing—review and editing: P.L., J.D., L.Y., S.C., J.K., R.L.J., A.L., O.F., S.M., G.N. All authors gave final approval for publication and agree to be held accountable for the work performed therein.

Competing interests

O.F. is employed by Colossal Biosciences. The company had no role in the design, execution, or funding of this research. The remaining Authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-56126-z.

Correspondence and requests for materials should be addressed to Pierre Lesturgie, Stefano Mona or Gavin J. P. Naylor.

Peer review information *Nature Communications* thanks Chris Jiggins, Xin Liu, and Sankar Subramanian for their contribution to the peer review of this work. A peer review file is available.