Genomic landscape of introgression from the ghost lineage in a gobiid fish uncovers the generality of forces shaping hybrid genomes

Kato, Shuya
Fisheries Laboratory, Graduate School of Agricultural and Life Sciences, The University of Tokyo

Arakaki, Seiji
Amakusa Marine Biological Laboratory, Kyushu University

Nagano, Atsushi J.
Department of Life Sciences, Faculty of Agriculture Ryukoku University

Kikuchi, Kiyoshi
Fisheries Laboratory, Graduate School of Agricultural and Life Sciences, The University of Tokyo

他

出版情報：Molecular Ecology. 33 (20), pp.e17216-, 2024-10. Wiley
バージョン：

Received: 14 July 2023 | Revised: 23 September 2023 | Accepted: 26 October 2023

DOI: 10.1111/mec.17216

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Genomic landscape of introgression from the ghost lineage in a gobiid fish uncovers the generality of forces shaping hybrid genomes

Shuya Kato[1] | Seiji Arakaki[2] | Atsushi J. Nagano[3,4] | Kiyoshi Kikuchi[1] | Shotaro Hirase[1]

[1]Fisheries Laboratory, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Hamamatsu, Shizuoka, Japan

[2]Amakusa Marine Biological Laboratory, Kyushu University, Amakusa, Kumamoto, Japan

[3]Department of Life Sciences, Faculty of Agriculture, Ryukoku University, Ōtsu, Shiga, Japan

[4]Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, Japan

Correspondence
Shuya Kato and Shotaro Hirase, Fisheries Laboratory, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Hamamatsu 431-0214, Shizuoka, Japan.
Email: tumagroikekatuo142@gmail.com and cashirase@g.ecc.u-tokyo.ac.jp

Funding information
Japan Society for the Promotion of Science (KAKENHI), Grant/Award Number: 18H02493, 22H00377 and 22J12643

Handling Editor: Philine G. D. Feulner

Abstract

Extinct lineages can leave legacies in the genomes of extant lineages through ancient introgressive hybridization. The patterns of genomic survival of these extinct lineages provide insight into the role of extinct lineages in current biodiversity. However, our understanding on the genomic landscape of introgression from extinct lineages remains limited due to challenges associated with locating the traces of unsampled 'ghost' extinct lineages without ancient genomes. Herein, we conducted population genomic analyses on the East China Sea (ECS) lineage of *Chaenogobius annularis*, which was suspected to have originated from ghost introgression, with the aim of elucidating its genomic origins and characterizing its landscape of introgression. By combining phylogeographic analysis and demographic modelling, we demonstrated that the ECS lineage originated from ancient hybridization with an extinct ghost lineage. Forward simulations based on the estimated demography indicated that the statistic γ of the HyDe analysis can be used to distinguish the differences in local introgression rates in our data. Consistent with introgression between extant organisms, we found reduced introgression from extinct lineage in regions with low recombination rates and with functional importance, thereby suggesting a role of linked selection that has eliminated the extinct lineage in shaping the hybrid genome. Moreover, we identified enrichment of repetitive elements in regions associated with ghost introgression, which was hitherto little known but was also observed in the re-analysis of published data on introgression between extant organisms. Overall, our findings underscore the unexpected similarities in the characteristics of introgression landscapes across different taxa, even in cases of ghost introgression.

KEYWORDS
demographic modelling, genomic landscape, ghost introgression, hybridization, phylogeography

1 | INTRODUCTION

Dramatic environmental changes during the quaternary have shaped the current patterns of biodiversity by influencing the divergence, distribution and survival of biological lineages (Avise, 2000; Nogués-Bravo et al., 2010; Sandel et al., 2011). While the biodiversity we see today represents lineages that managed to survive these environmental changes, numerous extinct lineages remain inaccessible without rare records like fossils. However, it has become clear in recent years that some of these extinct lineages have left legacies in the genomes of extant lineages through past introgressive hybridization, thereby contributing some parts of the current diversity. Since the discovery that introgression from ancient hominids has contributed some part of our own genomes (Green et al., 2010; Reich et al., 2010), there has been an increasing number of studies suggesting introgression from extinct lineages across various taxa in recent years (e.g. Ai et al., 2015; Barlow et al., 2018; Frei et al., 2022; Gopalakrishnan et al., 2018; Kuhlwilm et al., 2019; Palkopoulou et al., 2018; Ru et al., 2018; Zhang et al., 2019). This is not surprising given that introgression between lineages is ubiquitous among organisms (Mallet, 2005; Taylor & Larson, 2019). Nevertheless, the contribution of extinct lineages is suggested to be underestimated for many species due to the challenges associated with accounting for the presence of unsampled 'ghost' lineages in the absence of fossils or literature (Ottenburghs, 2020; Zhang et al., 2019), even though they may be an important piece of the diversity of extant organisms.

Exploring the genomic landscape of introgression from extinct lineages, which encompasses the genomic regions in which extinct lineages have or have not survived, holds the potential to illuminate how extinct lineages have contributed to the genomic evolution of extant lineages (Martin & Jiggins, 2017). Genomic regions derived from the extinct lineages can not only construct some parts of the extant genomes but also harbour variants that have accumulated in the evolutionary history unique to extinct lineages. These variants may help the recipients adapt to new environments (Frei et al.,2023; Gittelman et al., 2016; Wang et al., 2020; Zeberg & Pääbo, 2021) or, conversely, introduce deleterious mutation loads (Harris & Nielsen, 2016; Zeberg & Pääbo, 2020). Furthermore, the characteristics of regions where extinct lineages failed to be introgressed may shed light on the presence of barriers or selective process that excluded the genomes of extinct lineages (Schumer et al., 2018). Uncovering the trends in genomic survival of extinct lineages may provide insight into genomic regions where genealogical stability or instability may exist across the tree of life, given the underestimated action of extinct lineages in rewriting the genomic genealogy through introgression.

However, there are several difficulties in detecting regions introgressed from extinct lineages. Firstly, as previously mentioned, accounting for the presence of extinct lineages is often difficult because most extinct lineages are unsampled ghost lineage lacking available ancient DNA. Even when we suspect the contribution of extinct ghost lineages, inferring their landscapes of introgression remains challenging since prevalent methods for detecting and localizing introgression signals rely on genomic information from the parental lineages (Hibbins & Hahn, 2022; Martin et al., 2015; Patterson et al., 2012). Although several methods have been developed to detect introgressed loci without using one of parental lineages (Guan, 2014; Hammer et al., 2011), their application to non-human organisms is limited, perhaps due to their reliance on detailed information such as admixture timings or complex implementation procedures (but see Kuhlwilm et al., 2019. While a few excellent studies have overcome these difficulties and successfully identified some regions derived from extinct lineages (Ai et al., 2015; Frei et al., 2022; Kuhlwilm et al., 2019; Zhang et al., 2019), they often focused solely on specific functions or adaptive effects of such regions, paying little attention to the landscape that shapes the entire hybrid mosaic genome. Therefore, our understanding of the genomic landscape of extinct lineage ancestry is largely limited to humans, where extensive work has been conducted leveraging abundant ancient genomes, and little is known about the role of ghost introgression from extinct lineages in the genome evolution of extant species.

Recent advances in analytical methods for demographic inference have provided new avenues for detecting traces of extinct ghost lineages in non-model organisms. As highlighted by Ottenburghs (2020) and Hibbins and Hahn (2022), model-based demographic inference using backward in-time coalescent simulations enable the comparison of different scenarios regarding the presence and forms of ghost introgression. Indeed, model-based approaches have played a major role in unravelling the demography with ghost introgression in recent studies (Kuhlwilm et al., 2019; Ru et al., 2018; Zhang et al., 2019). Furthermore, recent developments in forward in-time simulations enable us to create genetic data that follow a particular demography in a scalable manner (Haller et al., 2019). Thus, once a plausible demography for ghost introgression can be estimated via demographic modelling, forward in-time simulations can produce datasets that mimic the observed data, enabling the evaluation of whether well-known existing methods are applicable to characterize introgression from the extinct ghost lineage.

Chaenogobius annularis Gill, 1859, a common intertidal goby inhabiting the temperate rocky coast around the Japanese Archipelago (Akihito et al., 2013), provides a valuable opportunity for exploring the genomic landscape of introgression from an extinct ghost lineage. This species exhibits a distinct phylogeographic structure, reflecting recurring cycles of isolation and connection between marginal seas in the North-Western Pacific Ocean (Hirase & Ikeda, 2015; Hirase, Ikeda, et al., 2012; Kato et al., 2021). Within this species, three allopatric lineages have been identified: the Sea of Japan (SJ) lineage and the Pacific Ocean (PO) lineage, which are ascribed to the past isolation of the Sea of Japan (Hirase, Ikeda, et al., 2012), and the East China Sea (ECS) lineage, endemic to the western coast of Kyushu Island (Figure 1a; Kato et al., 2021). The ECS lineage exhibits remarkable mitonuclear discordance, with mitochondrial DNA (mtDNA) of the ECS lineage diverging deeply from the PO lineage, whereas microsatellite polymorphism patterns suggested that nuclear DNA is highly similar to that of the SJ lineage (Kato et al., 2021). One of

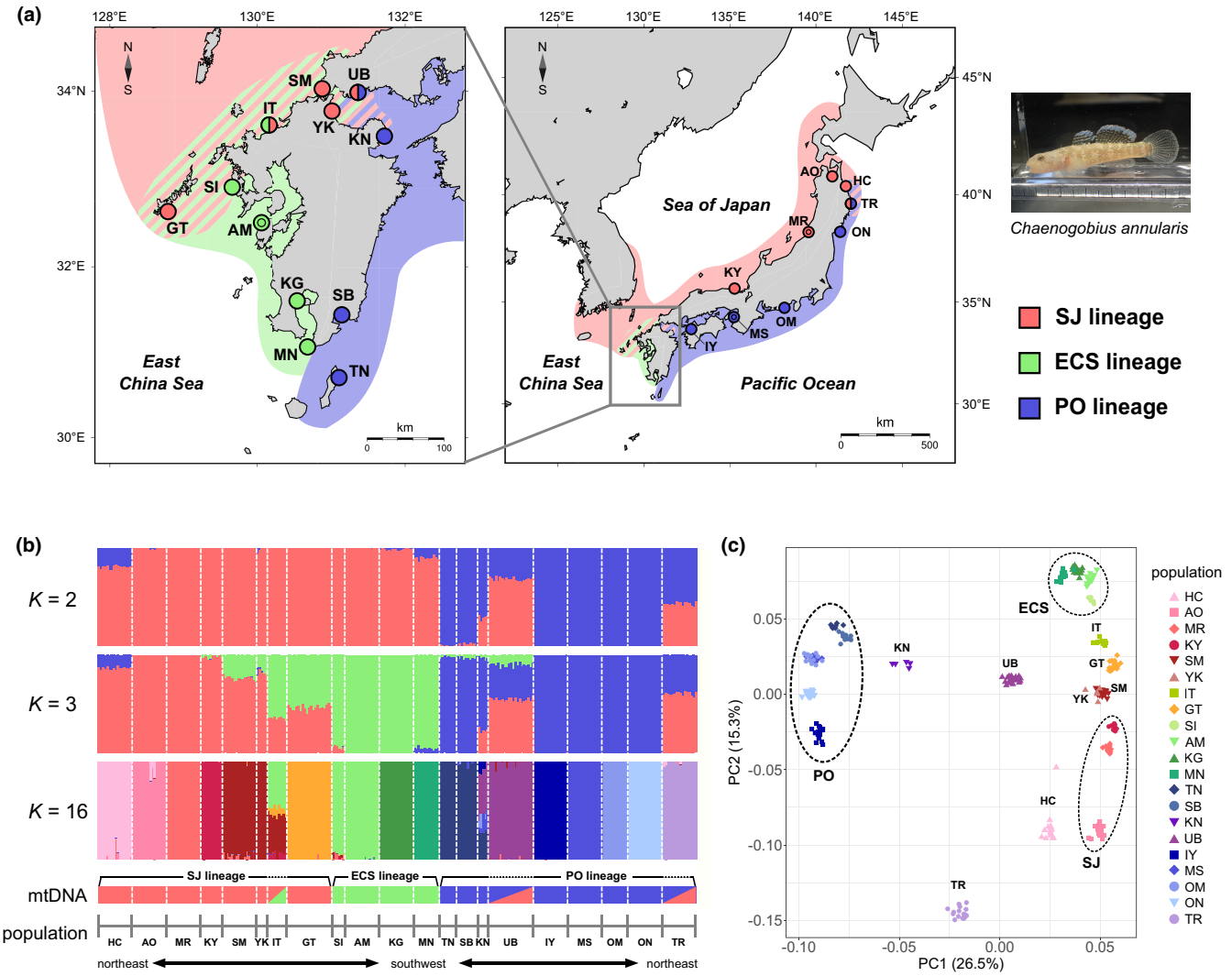

FIGURE 1 Population structure of *Chaenogobius annularis*. (a) Sampling locations of *C. annularis* and the geographic distribution of the three allopatric lineages. Circles indicate the sampling locations used in this study, with single circles representing sites used for ddRAD-seq only and double circles representing populations used for both ddRAD-seq and whole genome re-sequencing. The colours in the circles show the mitochondrial lineage affiliation based on Kato et al. (2021), while the colours in the background outline the estimated distributional range of the three lineages (single colour) and the hybrid zones between lineages (stripe) as suggested by the population structure analysis in previous studies (Hirase et al., 2021; Kato et al., 2021) and this study. Photograph of *C. annularis* by the first author. (b) Assignment plots inferred from clustering analysis using ADMIXTURE based on 3667 SNPs from the ddRAD-seq data. Each vertical bar represents an individual assigned to *K* clusters (results shown for $K=2$, 3 and 16). The colours in horizontal bars below the plots indicate the mitochondrial lineage affiliation [pink: Sea of Japan (SJ) lineage; green: East China Sea (ECS) lineage; blue: Pacific Ocean (PO) lineage]. (c) Principal component analysis based on 3667 SNPs from the ddRAD-seq data. The percentage in brackets on each axis represents genomic variance explained by each principal component (PC). The dotted circles indicate the lineage affiliation according to the results of the clustering analysis at $K=3$ (see Figure 1b). The population codes (HC, AO, ..., TR) used in panels a–c are described in Table S1.

the most common explanations for such a remarkable mitonuclear discordance is a hybrid origin (Toews & Brelsford, 2012). However, a scenario in which the ECS lineage simply arose from direct hybridization between SJ and PO is unlikely given that the level of divergence in mtDNA is deeper than that in microsatellites. Specifically, if this scenario is correct, the level of divergence in mtDNA should be equal to or lower than that of microsatellites as microsatellites generally have a faster mutation rate than mtDNA (Allio et al., 2017; Ellegren, 2000). Therefore, we previously suggested that this discordance in the ECS lineage may have originated from hybridization involving an extinct ghost lineage that diverged from the PO lineage (Kato et al., 2021). Although these low-resolution markers do not provide detailed information regarding its demography or the involvement of the extinct ghost lineage, the ECS lineage presents an intriguing case of a ghost extinct lineage contributing to diversity by establishing a new intra-specific lineage.

In this study, we aimed to test the hypothesis of ghost introgression origin in the ECS lineage of *C. annularis* and to characterize the genomic landscape of introgression from the extinct ghost lineage using genome-wide single-nucleotide polymorphisms (SNPs) and

whole genome re-sequencing data. First, we re-assessed the overall phylogeographic structure of the species and the demography of the three main lineages. Second, we employed demographic modelling to examine multiple different scenarios concerning the formation history of the ECS lineage to ascertain the occurrence of ghost introgression. Finally, we assessed whether local introgression signals can be detected in our dataset by forward simulations and then investigated the characteristics of the landscape of introgression from the ghost lineage in the ECS genome. Our results revealed that the ECS lineage originated through introgression with the extinct ghost lineage and further suggested that its hybrid mosaic genome was characterized by a selection that excluded minor parent ancestry, consistent with findings in hybrid genomes among extant organisms.

2 | MATERIALS AND METHODS

2.1 | Sample collection and DNA/RNA extraction

A total of 292 individuals of *C. annularis* were collected from 21 localities along the coast of Japan for DNA analysis (Figure 1a, Table S1). All of these samples were collected in our previous studies (Hirase et al., 2021; Hirase & Ikeda, 2014, 2015; Hirase, Ikeda, et al., 2012; Kato et al., 2021). One individual of *Chaenogobius gulosus* from Kagoshima Prefecture was also newly collected in this study and used as an outgroup (Table S1). Additionally, two individuals of *C. annularis* (one female and one male) from Kanagawa Prefecture were newly collected and used for RNA sequencing (Table S1). All samples, including those collected in the past, were collected by hand net and euthanized using ice water immersion. These two species are common species in the region and do not require any permits for collection.

Genomic DNA was extracted from a piece of fin or muscle preserved in 99% ethanol using the phenol/chloroform method (Asahida et al., 1996) or Gentra Puregene Tissue Kits (Qiagen). For RNA extraction, 11 tissues (brain, eyes, gills, heart, intestine, liver, muscle, ovary and spleen from one female, and brain and testis from one male) were dissected and immediately stored in Gene Keeper RNA&DNA stabilization solution (Nippon Gene) at −30°C. RNA extraction was performed at Macrogen Japan Corporation using QIAzol® Lysis Reagents (Qiagen) and RNeasy® Mini Kits (Qiagen).

2.2 | ddRAD-sequencing

We combined newly generated with previously published double-digest restriction-site-associated DNA sequencing (ddRAD-seq) data to create a dataset that roughly covered the entire distributional range of the species (Figure 1a, Tables S1 and S2). For this study, we sequenced 13 populations of *C. annularis* (151 individuals) and one individual of *C. gulosus*. Previously published data from eight populations (128 individuals) were obtained from Hirase et al. (2021) (accession numbers: DRR174909, DRR175781–DRR175796, DRR175830–DRR175860, DRR175876–DRR175955). Sequence data generated in this study have been deposited in the DDBJ databases under accession numbers DRR489922–DRR490073. Libraries for both datasets were prepared following Peterson et al. (2012) with slight modifications as described in Sakaguchi et al. (2015), using *EcoRI* and *BglII* as restriction enzymes. The newly generated data consisted of 150 bp × 2 paired end reads sequenced on an Illumina Hiseq X platform, whereas the previously published data comprised 51 bp single end reads sequenced on an Illumina Hiseq 2500 platform.

All raw sequence reads were filtered using Trimmomatic v0.33 (Bolger et al., 2014) to remove low-quality regions (*phred33*, *LEADING:19*, *TRAILING:19*, *SLIDINGWINDOW:30:20*, *AVGQUAL:20*, *MINLEN:51*) and Illumina adapters (*ILLUMINACLIP TruSeq3-PE.fa:2:30:10* for paired end reads and *ILLUMINACLIP TruSeq3-SE.fa:2:30:10* for single end reads). The filtered reads were then mapped to the reference genome of the SJ lineage of *C. annularis* (GCA_015082035.1; Hirase et al., 2021) using BWA version 0.7.15 (Li & Durbin, 2009) with the mem and -R options. The resulting alignments were exported as BAM files and sorted and indexed by samtools v.1.7 (Li et al., 2009). To mitigate the effects of mismapping, only uniquely mapped reads were retained from the BAM files, employing sambamba v.0.6 (Tarasov et al., 2015). Genotypes for variant and invariant sites were called using samtools v.1.7 mpileup and bcftools v.1.9 (Danecek et al., 2021) call. VCF records were jointly genotyped per population and then filtered after merging the populations as required for each analysis. Five distinct filtered SNP datasets were created. Details regarding the merging and filtering settings used in each analysis are provided in Supplementary Notes 1 and summarized in Figure S1 and Table S3.

2.3 | Whole genome re-sequencing

We performed whole genome re-sequencing (WGS) on 18 individuals of *C. annularis* (six individuals from one population each for the SJ, ECS and PO lineages) and one individual of outgroup *C. gulosus*. Library construction and sequencing were conducted at Beijing Genomics Institute (BGI), and sequencing was performed using an Illumina Novaseq 6000 to generate 150 bp × 2 paired end reads (Table S2). Sequenced reads were filtered at BGI using SOAPnuke (Chen et al., 2018) to remove adaptor sequences, low-quality reads with 40% or more bases having a quality value below 10 and reads containing 1% or more *N* bases. These filtered sequence data have been deposited in the DDBJ databases under accession numbers DRR489903–DRR489921.

Filtered reads were mapped to the reference genome using BWA version 0.7.15 with the mem and -R options. Alignments were exported as BAM files and sorted and indexed using samtools v.1.7. To mitigate the effects of mismapping, only uniquely mapped reads were retained from the BAM files, employing sambamba v.0.6. PCR duplicates were identified and removed using GATK v. 4.1.9 (McKenna et al., 2010) MarkDuplicatesSpark with the option

remove-all-duplicates true. The mean depth of coverage scored × 21.6 (SD = 0.853) after these processing.

Genotypes for variant and invariant sites were called using GATK v. 4.1.9 HaplotypeCaller, GenomicsDBImport and GenotypeGVCF with the *-all-sites* option. VCF records were jointly genotyped per population, and subsequently merged to create two datasets: '*C. annularis*-only' dataset and 'with-outgroup' dataset. After merging, indels were removed using VCFtools v.0.1.16 (Danecek et al., 2011) and hard filtering was applied using GATK VariantFiltration with the following parameters: *MQ* < 40.0, *QD* < 2.0, *SOR* > 4.0, *FS* > *60.0*, *MQRankSum* < −12.5 or *ReadPosRankSum* < −8.0. We retained the loci that did not meet any of the above conditions. Following hard filtering, we obtained biallelic SNPs by filtering using VCFtools v0.1.16 with the following parameters: *max-missing 1.0*, *min-alleles 2*, *max-alleles 2*, *minDP 8*, *max-meanDP XX* (*XX*: twice the mean depth of called sites, with *XX* = *40.33* for the *C. annularis*-only dataset and *XX* = 33.5492 for the with-outgroup dataset). These merging and filtering steps yielded in 12,241,098 SNPs in the *C. annularis*-only dataset and 17,943,129 SNPs in the with-outgroup dataset (Table S3).

We also obtained mitochondrial sequences by mapping the filtered WGS reads to the mitochondrial genome (mitogenome) reference sequence of *C. annularis* (accession number: OM830225). The mapping, processing and genotyping were performed following the same methods as whole genome genotyping. Briefly, we retained only uniquely mapped reads after BWA mapping and then removed PCR duplicates using GATK v. 4.1.9 MarkDuplicatesSpark. Variant and invariant sites were jointly called per population using GATK v. 4.1.9 HaplotypeCaller, GenomicsDBImport and GenotypeGVCF. We then applied hard filtering using GATK VariantFiltration employing the same parameter as those in whole genome genotyping. We retained the loci that did not violate any of the conditions in the hard filtering step, and further filtering was conducted using VCFtools v0.1.16 with the following parameters: *--remove-indels --minDP 30 --minGQ 40*. Finally, we merged all VCF records and converted VCF to fasta format by vcf2phylip.py (https://github.com/edgardomortiz/vcf2phylip). Note that we excluded our own *C. gulosus* data from this genotyping because of poor sequence coverage across the mitogenome.

2.4 | Repeats and gene annotation

To identify repetitive elements, we initially generated a de novo transposable elements (TE) library of the reference genome using RepeatModeler v 2.0.3 (Flynn et al., 2020). We then used RepeatMasker v 4.1.2 (Smit et al., 2010) to detect repeats using both the de novo library and the Dfam database.

Gene annotation was performed using BRAKER v 2.1.5 (Brůna et al., 2021), a fully automated pipeline of ab initio gene prediction with training by RNA sequencing (RNA-seq) data and protein sequences. We performed RNA-seq on 11 tissues of *C. annularis* to train the gene prediction model. Library construction and sequencing were performed at Macrogen Japan Corporation. Sequencing was performed using an Illumina Novaseq 6000 to generate 101 bp × 2 paired end reads. These sequence data have been deposited in the DDBJ databases under accession numbers DRR490074–DRR490084. The sequenced reads were mapped to the reference genome (soft-masked for repeats) using a spliced aligner STAR (Dobin et al., 2013), and alignments were exported as sorted BAM files. Subsequently, gene prediction was performed on the soft-masked reference genome using BRAKER2 employing the RNA-seq bam files. Additionally, we separately performed gene prediction using BRAKER2 with the downloaded protein sequences from six teleost species: *Periophthalmus magnuspinnatus*, *Sphaeramia orbicularis*, *Oryzias latipes*, *Gasterosteus aculeatus* and *Danio rerio*.

The two gene prediction sets were integrated using the TSEBRA module (Gabriel et al., 2021) with the default settings, resulting in 39,542 predicted genes. To minimize inaccurate gene predictions, we created a conservative gene annotation with 16,095 predicted genes by selectively restoring only the genes that exactly matched the RNA-seq hintfile using the *selectSupportedSubsets.py* script provided in BRAKER2. For functional enrichment analysis, we annotated the conserved predicted genes using blastp v.2.11.0+ against peptide sequences of *O. latipes*, *G. aculeatus* and *D. rerio*, obtained from Ensemble BioMart. The blast search employed an *E*-value cutoff of 10^{-10} and we considered the match with the lowest *E* value for each predicted gene. We also detected oxidative phosphorylation (OXPHOS) genes (except for complex2) and mitochondrial ribosomal (mitoribosomal) genes, which form a complex together with the products of genes in mtDNA. We searched these genes by blastp against the peptide sequences of *P. magnuspinnatus*, *O. latipes*, *G. aculeatus* and *D. rerio* in the same way. We selected genes that received support from blastp results from two or more species. Through this process, we identified 77 OXPHOS and 60 mitoribosomal genes and were considered representative of nuclear genes encoding mitochondrial-targeted proteins (N-mt genes).

2.5 | Population recombination rate estimation

The population recombination rate across genome for each lineage was estimated using LDhelmet v. 1.10 (Chan et al., 2012) with the *C. annularis*-only dataset of WGS data. The population recombination rates for each lineage were estimated only for scaffolds larger than 100 kb (1116/1897 scaffolds, which represent more than 95.7% of the total genome). Briefly, we first performed read aware phasing using SHAPEIT v2.r904 (Delaneau et al., 2012) and generated FASTA sequences of each haplotype from the phased VCF using *vcf-2fasta* tool in vcflib 1.0.0 (Garrison et al., 2022). Subsequently, haplotype configuration files were created using the *find_confs* module with the option *-w 50*. Likelihood lookup tables and padé coefficient files were generated using the average Watterson's θ values in the 10 kb sliding window calculated by R package *PopGenome* (Pfeifer et al., 2014; 0.00079 for SJ, 0.0015 for ECS and 0.0020 for PO) for

parameters of -*t*, and recommended values for others (–*r* 0.0 0.1 10.0 1.0100.0, –x 11). We performed 1,000,000 iterations with 100,000 burn-in iterations for the rjMCMC procedure with the option *-w 50* and *-b 10*. The obtained mean population recombination rates per bp were then smoothed for each non-overlapping sliding window (10, 30, 50 and 100 kb). We also calculated the average population recombination rate for three lineages.

2.6 | Identification of potentially deleterious SNPs

We searched for potentially deleterious SNPs on the WGS *C. annularis*-only dataset. Firstly, we used SnpEff v5.1d (Cingolani et al., 2012) to annotate the 12,241,098 SNPs based on our conservative gene annotation and extract SNPs with non-synonymous mutations. We then determined potential deleterious mutations using PROVEAN analysis (Choi et al., 2012), which evaluates amino acid mutations based on the principle that phylogenetically conserved amino acids can be important. We used PROVEAN v.1.1.5 and scored all non-synonymous mutations detected by SnpEff. To account for the possibility of deleterious mutations in the individual used as the reference genome, we considered alternative alleles as potential deleterious mutations for SNPs with PROVEAN score ≤−4.1 and reference alleles as potential deleterious mutations for SNPs with PROVEAN score ≥4.1.

2.7 | Population genetic analyses

Population structure was investigated using ddRAD-seq dataset 1 (Figure S1). We performed principal component analysis (PCA) on all 21 populations of *C. annularis* using PLINK v1.90 (Chang et al., 2015) to visualize genetic relationships among individuals. We also performed clustering analysis of all 21 populations using ADMIXTURE v 1.3.0 (Alexander et al., 2009), while varying the assumption of the number of clusters *K* from 1 to 21. One run was made for each *K* value. Appropriate *K* value was inferred based on the minimized cross-validation error rate. In subsequent analyses, populations with at least 10% assignment to minor clusters in the clustering analysis at $K=3$ (which corresponds to the three lineages of *C. annularis*) were considered presumptive hybrid populations between clusters. Conversely, populations not meeting this criterion were considered as presumptive 'core' populations representing each lineage.

Diversity indices were calculated using ddRAD-seq dataset 2 (Figure S1). Average nucleotide diversity (π) within each population was calculated with pixy v 1.2.7 (Korunes & Samuk, 2021). Nucleotide diversity within each lineage was also calculated using the presumptive 'core' population for each lineage. Genetic differentiations between populations or between lineages were measured by calculating the pairwise mean wcF_{ST} (F_{ST} in Weir & Cockerham, 1984) using VCFtools.

2.8 | Phylogenetic analysis

The phylogenetic relationships among the populations were examined using ddRAD-seq dataset 3 (Figure S1), which excluded presumptive hybrid populations. We constructed maximum likelihood (ML) tree on concatenated SNP alignment data using RAxML v 8.2.12 (Stamatakis, 2014) with the GTRGAMMA model. Node support was assessed through 1000 bootstrap replicates. To demonstrate the mitonuclear discordance among the three lineages, we obtained partial *cytochrome b* sequences of mtDNA for the same populations from Kato et al. (2021) and constructed a neighbour-joining tree (Saitou & Nei, 1987) using MEGA X (Kumar et al., 2018) using the method described in Kato et al. (2021). The phylogenetic discordance between nuclear DNA and mtDNA was also examined by ML trees using the WGS data. We constructed ML trees for both nuclear SNPs in the with-outgroup dataset and whole mitogenome sequences using RAxML with the same setting as described earlier. For the mitogenome sequence analysis, we downloaded the complete mitogenome sequences of *C. gulosus* (NC_027193.1) to serve as outgroups. These mitogenome sequences were aligned using MAFFT version 7 (Katoh & Standley, 2013).

2.9 | Detecting introgression

Signals of introgression between lineages or populations were tested by analysing the excess of allele sharing patterns among four taxa (one hybrid population, two parental populations and the outgroup) for both the ddRAD-seq (dataset 4; Figure S1) and the WGS data (with-outgroup dataset). Firstly, we calculated *D* statistics (Durand et al., 2011; Green et al., 2010) using *Dtrios* command in Dsuite v 0.5 r47 (Malinsky et al., 2021). The significance of *D* statistics was assessed using the block jackknife method and adjusted with the Bonferroni correction. To summarize the pattern of significant signals of introgression, we applied the *f-branch* analysis (Malinsky et al., 2018) implemented in Dsuite based on the significant results of *Dtrios* ($p<.05$ after Bonferroni correction) for ddRAD-seq dataset. Phylogenetic relationships between populations for *Dtrios* and *f-branch* were inferred from the results of phylogenetic analysis and clustering analysis. We also conducted HyDe analysis (Blischak et al., 2018) on the same datasets. HyDe can test the signal of hybridization and calculate estimated γ, the proportion of contribution from one parent to the hybrid population, by utilizing allelic patterns of BBAA as well as ABBA and BABA. We ran the analysis using the run_hyde.py script for all possible triplets of populations, and the significance of the signal of hybridization was adjusted with the Bonferroni correction. Since our focus was on the introgression between the three lineages, we focused only on signals occurring among different lineages. As these analyses supported the hybrid origin of the ECS lineage for the WGS data, we also observed the allele frequency of the ECS lineage at 1,580,393 fixed sites between the SJ and PO lineages (*C. annularis*-only dataset) to visualize the hybrid mosaic genome of the ECS lineage.

2.10 | Demographic estimation

The effective population size of each lineage across time was estimated using pairwise sequentially Markovian coalescent (PSMC; Li & Durbin, 2011). Consensus genome sequences were called separately for each individual. After performing local realignment on the bam file using GATK v. 3.8.1 RealignerTargetCreator and IndelRealigner, consensus sequences were called by samtools mpileup and bcftools call with skipping indels. The consensus sequences were then converted to fastq format using the *vcfutil.pl* script, excluding regions with depths less than 1/3 of the average read depth or greater than two times the average read depth. Fastq files were converted to psmcfa format. We then ran PSMC with the following parameters: *-N30, -t 10, -r 4, -p "6+23*2+6"*. These parameters ensured at least 10 recombination events occur in every atomic interval at the 20th iteration of the PSMC. We applied 1 year for the generation time of the species (Sasaki & Hattori, 1969) and a mutation rate of 3.5×10^{-9} per site per generation. This mutation rate was estimates for cichlids (Malinsky et al., 2018) and was almost identical to the values adopted in the study of mudskippers (You et al., 2014), which belong to the same family as *C. annularis*. We performed 100 bootstrap replicates on 500 kb segments for each analysis. Global average surface temperature (Snyder, 2016) and Red Sea relative sea level (Grant et al., 2012) were used as environmental data for the Late Pleistocene to interpret the estimated population size changes.

We conducted demographic modelling based on the site frequency spectrum (SFS) using fastsimcoal2 v 2.7 (Excoffier et al., 2013, 2021) to estimate the demographic history of the three lineages and the origin of the ECS lineage. We performed this analysis only for ddRAD-seq data (dataset 5) due to computational constraints, but our preliminary analysis on a small number of models showed that the use of WGS did not change the main results (data not shown). The folded multidimensional observed SFS was calculated using *easySFS.py* (https://github.com/isaacovercast/easySFS). Given our interest in determining if any ghost lineages were involved in the hybridization events that formed the ECS lineage, we examined 16 models with 10 demographies that focused on the order of the timing of divergence or introgression between the ECS lineage and the two parental lineages (Figure S11). It is important to note that several models represented virtually identical demography, just with different assumptions for the ghost lineages. This approach allowed us to confirm the robustness of our results and demonstrate that they were not influenced by the presence or absence of ghost lineages or the assumptions made about them. Although we did not include population size changes other than those occurring at the timing of divergence and introgression in the aforementioned analysis, it has been reported that not accounting for recent population size changes in demographic modelling can introduce biases in model selection and parameter estimation (Momigliano et al., 2021). Therefore, we also performed separate analyses for the 16 models incorporating a certain model for recent population size changes for each lineage inferred from the results of PSMC. This included continuous population size shrinking and subsequent sudden expansion for the SJ lineage, as well as sudden population size change for the ECS lineage and the PO lineage. To ensure that the results obtained were not affected by recent gene flow, we further tested the models incorporating recent or ongoing migration between the SJ and ECS lineages (Figure S12). Due to high computational demands for computing gene flow in every generation, we restricted this additional modelling to model 6a and model 7a without recent population size change, which is important for testing the hypothesis of ghost introgression. We used a mutation rate of 3.5×10^{-9} per site per generation in all models and applied a generation time of 1 year. Further details about each model are available from Dryad (https://doi.org/10.5061/dryad.7wm37pw09). Run configuration setting on fastsimcoal2 are provided in Supplementary Notes 2.

2.11 | Detecting genome-wide signals of introgression

We tried to estimate the genomic landscape of introgression in the ECS genome using a sliding window approach. As our objective was to evaluate the entire genome in terms of introgression, we calculated the statistic γ in HyDe analysis, estimators for admixture rate after considering the effect of incomplete lineage sorting (ILS), in non-overlapping windows of 10, 30, 50 and 100 kb. Initially, we calculated the pseudo counts of ABBA, BABA and BBAA allelic patterns for each window that contained at least 100 biallelic SNPs using the python script slightly modified from the *ABBABABAwindows.py* script (Martin et al., 2015). We then calculated the statistic γ for each window using the equation: $\gamma=(\mathrm{ABBA}-\mathrm{BABA})/(\mathrm{ABBA}+\mathrm{BBAB}-2\times\mathrm{BABA})$ (see equation 3 in Blischak et al., 2018). For windows where the denominator was less than or equal to zero, we considered the statistic γ to be incalculable due to an excess of ILS, and we denoted γ as NA. We note that the windows with γ that largely deviated from the range of $0\le\gamma\le1$ were also affected by excess of ILS. We also calculated two well-known statistics to locate introgression, f_d (Martin et al., 2015) and f_{dM} (Malinsky et al., 2015), using the original function of the *ABBABABAwindows.py* script.

Although all three statistics have been well tested, their performance when applied to ghost introgression has not been verified. Furthermore, the statistic γ in HyDe analysis was not originally designed for application to small sliding windows. To address these concerns, we roughly estimated the extent to which each statistic could reflect the strength of introgression from the ghost lineage across windows in the simulated data which imitated the diversity in the observed data (Figure 5a). Herein, we simulated five taxa, namely SJ, ECS1 (SJ-derived ECS), ECS2 (PO-derived ECS; equivalent to the ghost lineage in the best models of demographic modelling analysis), PO and outgroup using forward-in-time population simulation software SLiM v 3.7.1 (Haller & Messer, 2019). Firstly, we simulated 7500 kb of chromosomes 100 times to determine the demographic parameters used in the simulation. The simulations were performed using the tree-sequence recording mode in SLiM, and VCFs were generated for the same number of

randomly selected individuals as in the actual WGS data (six for SJ, ECS1, ECS2 and PO, and one for outgroup) by putting mutations according to a mutation rate of 8.0×10^{-8} per site per generation. The demographic pattern to be simulated was determined based on the results of demographic analyses. The variants in the simulated VCF were restricted to biallelic SNPs using VCFtools v0.1.16. To mimic the introgression from the ghost lineage, we created a hypothetical ECS lineage for each simulated VCF by replacing 30% of the regions from ECS1 with corresponding regions from ECS2. We adjusted the demographic parameters to ensure that the simulated VCFs exhibited diversity indices (mean wcF_{ST} between populations and total SNP number) similar to those in the observed data. Further details about the SLiM setting are available from Dryad. Secondly, we simulated small chromosomes of various sizes (10, 30, 50 and 100 kb) 100 times each to evaluate how well the statistic γ can infer the rate of introgression. We generated VCF records for the simulated chromosomes using the same settings and demographic parameters as above. We then created a hypothetical ECS lineage for each simulated VCF by replacing 0%, 10%, ..., 90% and 100% regions from ECS1 with corresponding regions from ECS2. We calculated γ, f_d and f_{dM} in the simulated VCFs, which consisted of SJ, hypothetical ECS, PO and outgroup, using the same method employed for the observed data. The performance was evaluated by comparing the statistics with the percentage of the regions that were actually replaced. We also determined the threshold for the window in which introgression is predominant by searching for the range of γ that minimized the sum of false positive and missing rates for windows where ≥50% and ≥90% regions were derived from ECS2.

2.12 | Characterizing the genomic regions introgressed from the ghost lineage in the ECS lineage

To characterize the estimated landscape of introgression in the ECS genome, we categorized each window according to the value of γ ($-3.0 \le \gamma < -1/7$, $-1/7 \le \gamma < 0.1$, $0.1 \le \gamma < 0.3$, $0.3 \le \gamma < 0.5$, $0.5 \le \gamma < 0.7$, $0.7 \le \gamma < 0.9$, $0.9 \le \gamma < 8/7$ and $8/7 \le \gamma \le 4.0$; these categorical divisions were determined based on the relative frequencies of the ABBA, BBAA and BABA allelic patterns). The category with γ close to zero represented regions that were likely to be resistant to introgression, whereas the categories with γ close to one comprised regions where the ECS and the PO lineages are very similar because of introgression. We also examined the categories with $\gamma < -3.0$ and $\gamma =$ NA, where the BABA allelic pattern was predominant. Category with $\gamma > 4.0$ were not analysed as they either had negligible representation or were absent in our dataset (<0.028%). We focused on several properties to characterize each γ category, including the proportion of coding sequences, the proportion of coding sequences of representative N-mt genes (OXPHOS and mitoribosomal genes), the population recombination rate (ρ), the proportion of repetitive sequences and the number of potential deleterious alleles per coding sequence length. The population recombination rate and the number of deleterious alleles were observed for the SJ, ECS and PO lineages. The population recombination rate was also observed for the mean of the three lineages.

We performed three different types of permutation tests to compare the observed values of each γ category to the genomic background. First, we used 'random permutation' created by randomly obtaining the same number of windows as each γ category. Second, we applied 'circular permutation' approach (Nouhaud et al., 2022) to reduce the effect of adjacent windows sharing ancestry and genomic features. In this approach, we randomly shifted the window coordinates of each γ category by 10 or more windows to create subsets that preserved the relative positions of the windows. Third, we invented 'controlled permutation' approach in which one variable was roughly controlled in order to reduce the effect of correlations between characteristics. In this approach, we ranked the control variables in each window and created permutations with similar control variables with each γ category by obtaining windows with rankings that randomly varied the ranking of each gamma category by ±1% of all windows. For all permutation tests, the significance of deviations from genomic background was assessed based on the number of permutations that exhibited difference exceeding the difference between the observed value and the mean of the permutations (equivalent to a two-sided test). Random permutations and circular permutations were used for all analyses, while controlled permutations were used only to compare proportion of repetitive sequences when proportion of coding sequences or average population recombination rate were controlled.

We also tried to characterize the regions likely to be introgressed based on the γ threshold determined through simulations. We conducted gene ontology (GO) enrichment analysis for the predicted genes in the presumptive introgressed regions using the *enrichGO* function in the R package *clusterProfiler*. Analyses were performed for each window size using Ensembl Gene IDs obtained with blastp against *O. latipes*, *G. aculeatus* and *D. rerio*. Ensembl Gene IDs of all predicted genes with confirmed homology to transcripts of each species were used as background gene lists. The statistical significance was adjusted by Bonferroni correction.

To explore the generality of the relationship between introgressed regions and the proportion of repetitive sequence (see Section 3 Results), we re-analysed published data from two papers for which results of sliding window analysis on introgression were available. The first dataset involved introgression between two Japanese sticklebacks (*G. nipponicus* and *G. aculeatus*; Ravinet et al., 2018), which included f_d and G_{MIN} (Geneva et al., 2015) for non-overlapped 10 kb sliding windows, as well as the 'valley' and 'non-valley' regions for G_{MIN} estimated by Hidden Markov Model. The second dataset focused on introgression between sympatric *Heliconius* butterflies (Martin et al., 2019), which included f_d for overlapped 100 kb sliding windows. Although Martin et al. (2019) focused on introgression both between *H. timareta* and eastern populations of *H. melpomene* and between *H. cydno* and western populations of *H. melpomene*, we only re-analysed the dataset focusing on the introgression between the former pair (sets 4–6

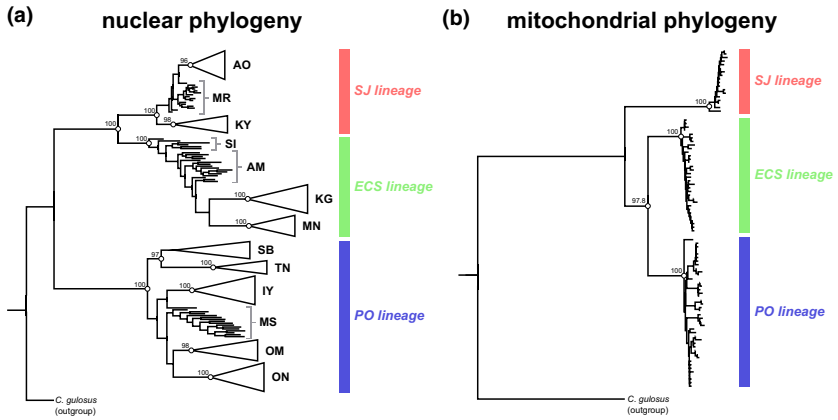

FIGURE 2 Mitonuclear discordance in phylogenetic relationships among the three allopatric lineages of *Chaenogobius annularis*. (a) Maximum likelihood tree generated by RAxML for 13 populations based on 2347 nuclear SNPs from the ddRAD-seq data. Monophyletic clades formed by a single population are shown collapsed. The population codes are shown in Table S1. (b) Neighbour-joining tree using MEGAX for the same 13 populations based on mitochondrial cytochrome *b* sequences (1025 bp) obtained by Kato et al. (2021). Bootstrap probabilities over 90% of 1000 re-samplings are shown for major nodes both in a and b. SJ lineage: Sea of Japan lineage; ECS lineage: East China Sea lineage; PO lineage: Pacific Ocean lineage.

in Martin et al., 2019), where the minor and major parentages in hybridization are clear (Martin et al., 2013). For both datasets, we annotated repetitive sequences within the reference genome used in these two studies (BROAD S1 for sticklebacks and Hmel2.5 for butterflies) using the method described in Section 2.4. We then calculated the proportion of repetitive sequence in each f_d decile category and G_{MIN} valley and non-valley category using the same method as in this study.

3 | RESULTS

3.1 | Re-assessment of the population structure of *Chaenogobius annularis*

Understanding the population structure and demography of the target species is essential for subsequent evolutionary analyses. To recapture an accurate depiction of the population structure of *C. annularis*, we first performed clustering analysis and PCA on SNPs obtained from ddRAD-seq. The clustering analysis showed three distinct clusters corresponding to three mtDNA lineages when assuming $K=3$ (Figure 1b). If we referred to these clusters as the SJ cluster, the ECS cluster and the PO cluster based on their mtDNA lineage affiliation, the ECS cluster was assigned to the same cluster as the SJ cluster at $K=2$. The PCA showed congruent result with the clustering analysis, with PC1 distinguishing the PO cluster from the other populations, and PC2 and PC3 distinguishing the SJ cluster from the ECS cluster (Figure 1c, Figure S2). In the clustering analysis at $K=3$, the previously identified hybrid populations between three lineages (HC, IT, GT, KN, UB and TR; Hirase & Ikeda, 2015; Kato et al., 2021) were detected as admixed populations between two or three clusters, while some nearby populations (SM, YK) also showed some signatures of admixture between clusters. The UB population showed admixture between three clusters. These presumptive hybrid populations were also positioned between the respective clusters in the PCA plot.

The cross-validation error rate, a measure of the optimal number of clusters in a clustering analysis, continued to decrease progressively until $K=16$ (Figure S3). Each sample tended to be assigned to a separate cluster for each location as the number of K increased (Figure 1b, Figure S4). This implied that *C. annularis* has a well-subdivided population structure, suggesting that the restricted gene flow observed at a small spatial scale between generations shown by Hirase, Kanno, et al. (2012) has persisted throughout demographic history. Among the presumptive hybrid populations at $K=3$, the TR population tended to be assigned to its own cluster at $K\geq4$, and the populations in the southern part of the Sea of Japan, excluding IT, showed same trend at $K\geq6$ (Figure S4). The TR population was also distinguishable from the other populations by PC2 in the PCA (Figure 1c), suggesting its unique genome composition. Conversely, the IT and UB hybrid populations were often assigned as a mixture of clusters from the southern Sea of Japan and neighbouring populations (e.g. SM and AM, or SM and IY), suggesting that these populations have likely arisen from more recent hybridization events with their neighbouring counterparts.

We constructed a ML tree using ddRAD-seq data, excluding the presumptive hybridization populations. The ML tree revealed that the populations in the SJ, ECS and PO clusters formed three well-supported clades (with 100% bootstrap values). This confirms the existence of the three lineages reported by Kato et al. (2021) in the nuclear genome (Figure 2a). The ECS clade formed a monophyletic lineage with 100% bootstrap value with the SJ clade. Contrastingly, the mtDNA phylogenetic trees showed monophyly of the ECS and PO lineages with a 100% bootstrap value (Figure 2b). These results show the discordance in phylogenetic relationships between mtDNA and nuclear DNA in the ECS lineage. Within the SJ and PO clades, southern populations tended to be placed in more basal positions (Figure 2a), suggesting a south-to-north expansion

of this species. Genetic differentiations between populations were relatively high (Table S4; wcF_{ST}=0.064–0.532). Differentiation between the ECS and the PO lineages (wcF_{ST}=0.212) was higher than that between the ECS and SJ lineages (wcF_{ST}=0.196; Table S5), which is consistent with the monophyly of the SJ and ECS lineages in the nuclear phylogeny. Nucleotide diversity within each lineage was lowest in the SJ lineage and highest in the PO lineage (Figure S5a), consistent with previous findings from mtDNA and microsatellite DNA analyses in our previous studies (Kato et al., 2021). The presumptive hybrid populations tended to exhibit higher diversity than that of their parental lineages (Figure S5b).

Based on these results, we performed whole genome sequencing on the three lineages, using MR, AM and MS as representative populations. The ML trees constructed using nuclear SNPs and mitogenome obtained from the WGS data exhibited the same pattern of mitonuclear discordance as observed in the ddRAD-seq data (for the SJ, ECS and PO populations respectively; Figure S6).

3.2 | Hybrid origin of the ECS lineage

The remarkable mitonuclear discordance observed in the ECS lineage may have originated from an ancient hybridization event rather than ILS. We therefore tested the hybrid origin of the ECS lineage by examining the allele sharing pattern of these three lineages. For the WGS data, *D* statistics showed a significant signal of introgression from the PO lineage to the ECS lineage (*D* statistics=0.234, adjusted $p=2.3\times10^{-16}$; Table 1). The HyDe analysis also detected a significant signal of hybridization in the ECS lineage (adjusted p=~.0; Table 1). The estimated γ value was 0.9034, indicating that the majority of the ECS lineage genome derived from the SJ lineage. However, the ddRAD-seq data failed to detect the hybrid origin of the ECS lineage in the *D* statistics or HyDe analysis (Figure S7, Tables S6 and S7). This may be due to the limited number of sites available in the ddRAD-seq data (5477 SNPs), which may lack the power necessary for robust detection in the HyDe analysis (Blischak et al., 2018). Additionally, the challenge of detecting ancient hybrid events compared to more recent ones may contribute to this discrepancy as the analysis of the ddRAD-seq data detected significant signals of hybridization between lineages in the presumptive hybrid populations (e.g. IT, GT, KN, UB, TR; Figure S7).

Sites fixed between the SJ and PO lineages showed a high level of homozygosity (88.2%–88.6%) in each individual in the ECS lineage, with 68.0% of these sites fixed to either SJ or PO allele within the ECS lineage (Table S8, Figure S8). These results suggest that the ECS lineage originated from ancient introgression and that its hybrid mosaic genome has reached genetic stabilization (Gompert et al., 2014).

3.3 | Demographic history of the three lineages and the formation of the ECS lineage

To unravel the demographic history of the three lineages, we initially inferred the historical population size change prior to demographic modelling. The PSMC analysis showed distinct differences in population size history among the three lineages (Figure 3, Figure S9). Notably, the SJ lineage was estimated to have experienced a severe bottleneck approximately 10,000–100,000 years ago, coinciding with the last glacial period (Hughes et al., 2013), and characterized by low temperature and reduced sea levels (Figure S10). The effective population size of the SJ lineage reached the lowest point during the last glacial maximum (mean $N_e \pm$ SD = 23,840 ± 1392 at the nadir). Subsequently, the SJ lineage rapidly expanded its population size after the last glacial period (mean $N_e \pm$ SD = 1,344,664 ± 591,803). While the ECS lineage showed some degree of population of expansion, and the PO lineage exhibited modest population decline during the same period, these trends were not as pronounced as the SJ lineage. A similar demographic pattern, with only the lineage in the Sea of Japan experiencing remarkable population size changes, has been observed in several coastal species in this region (Kojima et al., 2004; Kokita & Nohara, 2011; Ravinet et al., 2018). This is consistent with the severe environmental changes in the Sea of Japan during the glacial periods (Gorbarenko & Southon, 2000; Oba et al., 1991). The previous study on *C. annularis* based on mtDNA (Hirase et al., 2016) inferred a bottleneck at an earlier time than in this study. Given that the estimates from the demographic modelling analysis are also younger than Hirase et al. (2016), this may possibly be due to the differences in mutation rate assumptions.

We then conducted coalescent demographic modelling of the three lineages to test whether the ECS lineage originated from hybridization with an extinct ghost lineage. Demographic modelling for 16 models with 10 demographies (Figure S11) showed that the most supported models were the three models (models 7a, 7b and 7c) representing the demography wherein the ECS lineage emerged through hybridization between a lineage that diverged from the PO lineage and a lineage recently diverged from the SJ lineage (Figure 4a,b, Table S9). Although the best model with the lowest AIC values (Akaike, 1973) was model 7a, the distributions of AIC values among these three models overlapped to

TABLE 1 Allele sharing pattern analyses to test the hybrid origin of the East China Sea (ECS) lineage based on WGS data. BBAA, ABBA and BABA indicate counts of alleles showing each pattern when assuming four taxa as (((SJ, ECS), PO), outgroup).

			Dsuites				*HyDe*		
BBAA	ABBA	BABA	*D* statistics	*Z* score	Adjusted *p*	f_4 ratio	γ	*Z* score	Adjusted *p*
853,632	204,943	127,294	0.233717	66.9879	2.30E-16	0.085886	0.90341955	138.96807	~.0

Note: The adjusted *p* value indicates the significance after Bonferroni correction.

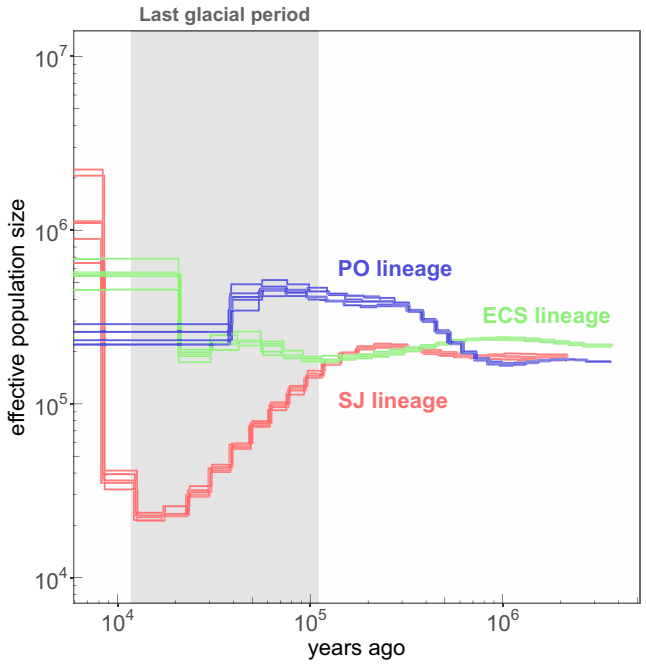

FIGURE 3 Past effective population sizes of the three allopatric lineages estimated using PSMC based on 18 re-sequenced genomes. Each line indicates an individual, and the colour of the line represents the lineage [pink: Sea of Japan (SJ) lineage; green: East China Sea (ECS) lineage; blue: Pacific Ocean (PO) lineage]. The grey background denotes the last glacial period.

some extent (Table S9). In all three models, the majority of the ECS lineage was derived from the SJ lineage (Figure S13, Tables S10–S12), thereby supported the notion that the ECS lineage originated from hybridization with a ghost lineage that branched from the PO lineage. These results were robust despite different ghost lineage assumptions, as the three models that varied solely in their treatment of ghost lineages garnered nearly equal support, and because other demographies were not supported even if ghost lineages were assumed. The same results were supported in demographic modelling incorporating recent population size changes estimated by PSMC (Figure 4d,e, Figure S14, Tables S13–S16). The best model accounting for recent changes in population size exhibited lower AIC values than those of the best model without such consideration. We suspected that these results could be due to the effect of the young tract brought about by the recent gene flow between the SJ and ECS lineages, but models with ghost introgression consistently showed lower AICs than the alternative models even when incorporating gene flow (Figure S15, Table S17). For both modelling with and without accounting for recent population changes, the best estimates demonstrated that the time parameters T2 and T3 are very far apart, with no overlap in their 95% confidence intervals (Figure 4c,f). Since T2 represents the divergence time between the PO and the ghost lineages and T3 represents the divergence time of the SJ and ECS lineages, this result indicated that the ghost lineage in these models represented an ancient ghost lineage diverged much earlier than the ECS lineage. The estimated parameter for the timing of the hybridization event (T4) is around 100 kya. Nevertheless, care must be taken in interpreting the absolute timing, as the estimated parameter for divergence times between the SJ and PO lineages (T1) were three times younger than the estimates derived from mtDNA (Hirase et al., 2016). Estimates and 95% confidence intervals of other parameters in the best models are shown in Figures S16 and S17.

3.4 | Characteristics of genomic landscape of introgression from the ghost lineage in the ECS lineage

To characterize the genomic regions introgressed from the ghost lineage, we tried to estimate the genomic landscape of introgression in the ECS genome using a sliding window approach. We first performed forward-in-time simulations to evaluate the performance of different statistics in capturing the landscape of introgression (Figure 5a, Figure S18). Although our neutral simulations did not perfectly replicate the genetic diversity of our WGS dataset, they served as benchmarks for comparing the performance of each statistic. Sliding window analysis on the simulation dataset revealed that the statistics γ in HyDe analysis showed a higher correlation with the actual proportion of different ancestry than those exhibited by other statistics f_d or f_{dM} (Table S18), suggesting that γ can better represent the difference in the proportion of introgressed regions within windows (Figure 5b, Figure S19). This advantage of γ may be ascribed to its utilization of the allelic pattern BBAA, which is not employed by f_d and f_{dM}. We note that, however, this result is based solely on our simplified simulations and does not negate the utility of f_d and f_{dM}. The performance of γ improved with increasing window size, with a particularly notable improvement from 10 to 30 kb (Figure S19). As the smaller the window size, the higher the expected resolution per genome; and therefore, window sizes ranging from 30 to 50 kb seems optimal for our dataset. The sum of the false positive and missing rates for windows with an introgression rate ≥50% was minimized in the range of $\gamma \geq 0.15$ for all window sizes, while for windows with introgression rate ≥90% was minimized in the range of $\gamma \geq 0.65$ for 10–30 kb and $\gamma \geq 0.55$ for 50–100 kb.

Therefore, we calculated the sliding window γ and examined characteristics of the regions likely to be introgressed and not introgressed using γ as an indicator of the introgressed proportion in the ECS genome (Figure 5c). Based on the performance of γ observed in the simulations, we specifically focused on window sizes of 30 kb or larger. We found several consistent patterns associated with γ values for each window size (Figure 5d–f). Firstly, the regions with γ close to zero ($-1/7 \leq \gamma < 0.1$), indicating resistance to introgression, exhibited higher proportions of coding sequences, while the presumptive introgressed regions with $0.1 \leq \gamma$ tended to show lower proportions of coding sequences (Figure 5d, Figures S20 and S21). For window sizes of 30 kb or larger, the percentage of coding sequences decreased as the γ category approached one. These patterns significantly deviated from the genomic background in several categories. Secondly, the average population recombination rate was significantly lower

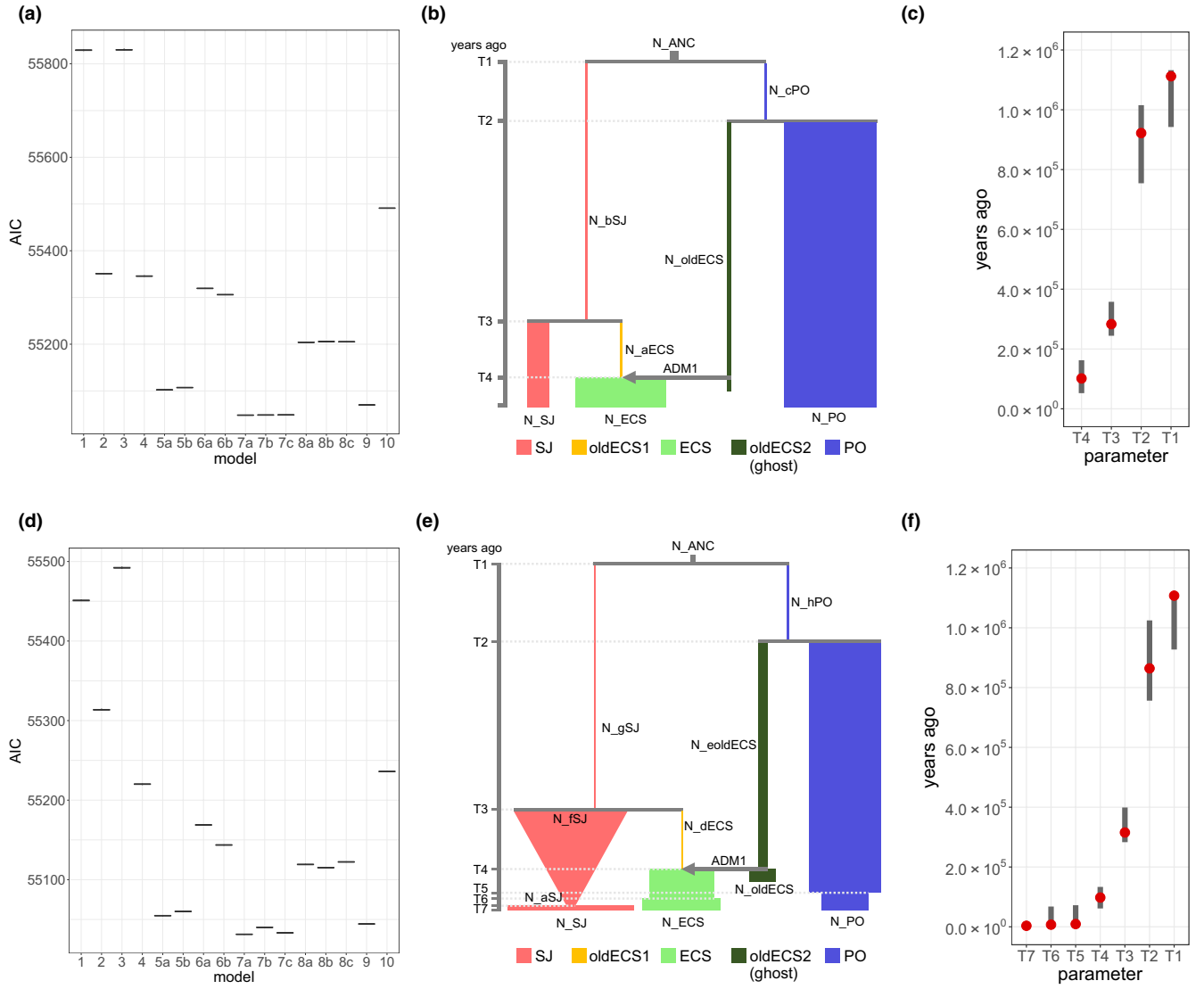

FIGURE 4 Demographic modelling using fastsimcoal2. Panels (a–c) represent the results of modelling without recent population size changes, while panels (d–f) represent the results of modelling incorporating recent population size changes. (a,d) show Box plots of AIC for 16 demographic models. Further details about each model can be found in Figure S11 and Dryad. Panels b, e present the best demographic model for Figure 4a,d respectively (model 7a for both). The thickness and height of the rectangles reflect the magnitude of the maximum likelihood parameter values for the effective population size (N_e) and the divergence time respectively. The arrows indicate introgression events from one lineage to another. N_* refers to the parameters for N_e at each time interval, T^* denotes the parameters for the divergence time and ADM1 represents the parameters for admixture rate from the oldECS2 lineage to the ECS lineage. Panels c, f display the estimated parameter values for the divergence time along with their corresponding 95% confidence intervals in the best models. The red points indicate the values in the maximum likelihood parameter sets and the grey bars represent the 95% confidence intervals of the parameter based on 100 rounds of block bootstrapping.

in regions with γ close to zero ($-1/7 \le \gamma < 0.1$) and higher in regions with moderately high γ ($0.1 \le \gamma < 0.5$) than that of the genomic background (Figure 5e, Figures S22 and S23). Categories with γ close to 1 showed low recombination rates for window sizes below 30 kb, but not for sizes above 50 kb. These trends were also observed in the population recombination rate within each of the three lineages (Figures S24–S29). Thirdly, the proportion of repetitive sequences within each category effectively characterized the estimated landscape of introgression, with the proportion increasing as γ approached closer to one from zero (Figure 5f, Figures S30 and S31). Long interspersed nuclear element (LINE) was particularly enriched in the introgressed regions (Figure S32). The enrichment of repetitive sequences in the high γ category was still pronounced even when controlling for the density of coding sequences and population recombination rate (Figures S33 and S34). These three features (coding sequence density, local population recombination rate and repetitive sequence density) showed significant correlation with the γ value for each window size (Table S19). To explore whether similar trends in the introgression landscape of repetitive sequences are observed in other organisms, we analysed published sliding window

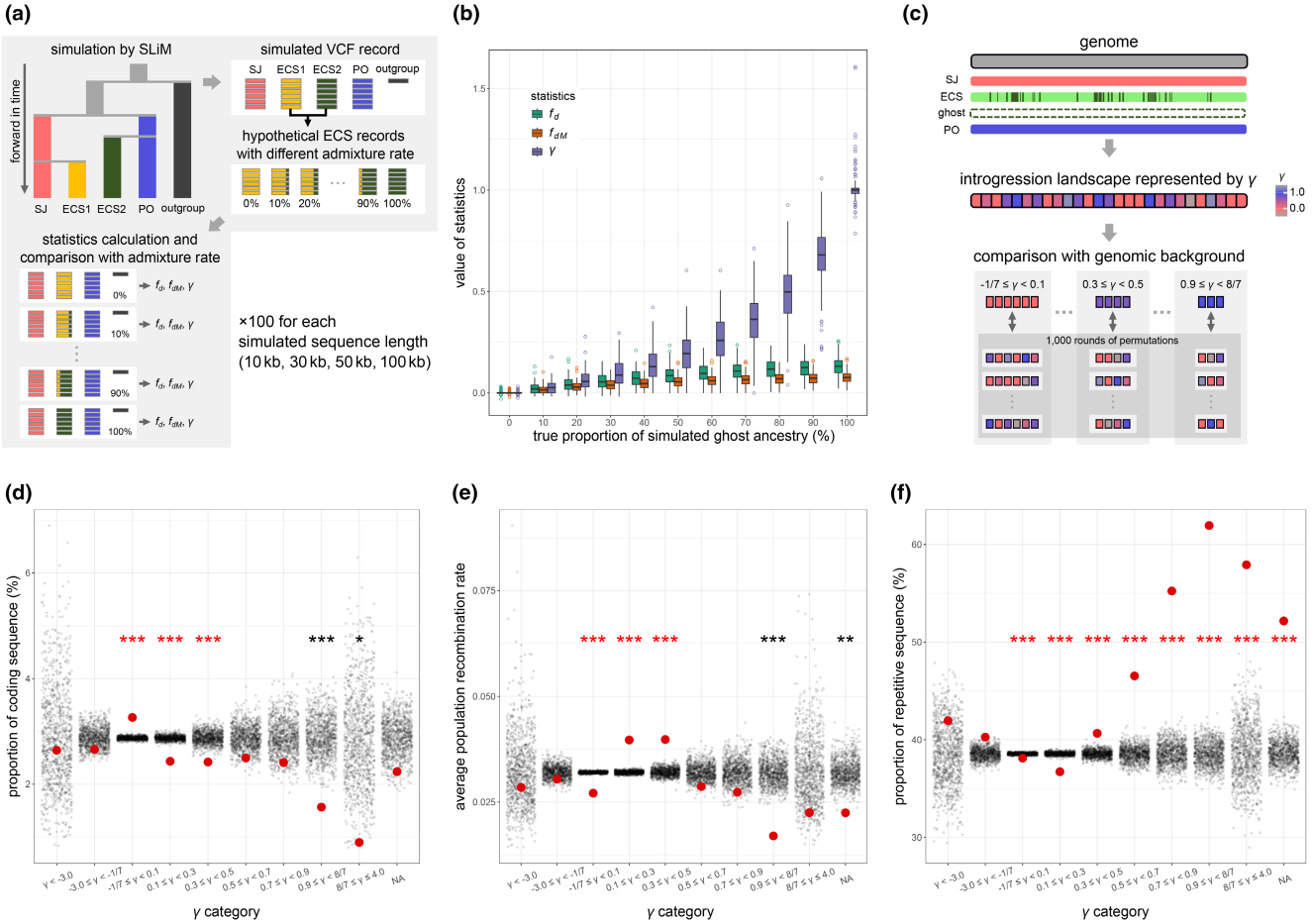

FIGURE 5 Genomic landscape of introgression in the East China Sea (ECS) lineage. (a) Schematic diagram of the workflow illustrating the process of performance evaluation of statistics in inferring the level of ghost introgression using forward simulations that imitate our dataset. (b) Box plots show the performance of three statistics, f_d, f_{dM} and γ, in inferring the level of ghost introgression in our 30 kb simulations. Each statistic was calculated for 100 simulated sequences. An open circle represents an outlier that is located more than 1.5 times the interquartile range away from the box. Refer to Figure S18 for the results obtained with other sequence lengths. (c) A workflow schematic diagram depicting the characterization of the genomic landscape of introgression in the ECS lineage. (d–f) Characteristics of the landscape of introgression compared to the genomic background, focusing on the proportion of coding sequences (d), population recombination rate averaged for the three lineages (e) and proportion of repetitive sequences (f). The results shown here are based on 30 kb windows. See Figures S20, S22 and S30 for the complete results obtained with other window sizes. The red points indicate the observed values for the windows in specific γ categories, while grey points represent values obtained through random permutations. The significance was assessed by comparing the number of permutations whose difference from the mean of 1000 permutations exceeded the difference between the observed value and the mean of the permutations. ***0/1000, **<0/1000, *<50/1000. The red asterisks are shown for categories that consistently showed significant signals in the same direction across all window sizes (see Figures S20, S22 and S30).

data from Japanese sticklebacks and *Heliconius* butterflies. Our analyses revealed a similar correlation between the proportion of repetitive sequences and the estimated local ancestry in these species complexes (Figures S35–S37). Differences in the decile in which repetitive sequences were enriched probably reflect differences in the level of introgression in each system and the resulting differences in the distribution of f_d (Table S20). In all three organisms, an increase or decrease in TEs seemed to drive this correlation (Figures S36 and S37).

We also focused on the functional properties of the introgressed regions using the threshold for >50% introgression determined by the above simulations. GO analysis showed no significant enrichment of GO terms, except for analysis in the 10 and 100 kb windows using zebrafish GeneID (Table S21). Similar results were obtained when using the >90% introgression threshold (Table S22). Since no significant enrichment was consistently detected across multiple analyses, there is no strong evidence that genes with specific functions were susceptible to introgression. We expected that N-mt genes may be co-introgressed with mtDNA as a mechanism to alleviate incompatibilities arising from mitonuclear discordance (Beck et al., 2015; Pritchard & Edmands, 2013; Sloan et al., 2017), but no mitochondria-related GO terms were detected in any of the analyses. Furthermore, the proportion of coding sequence of N-mt genes did not significantly differ from the genomic background across each γ category (Figures S38 and S39). We also investigated the presence of potentially deleterious

alleles in the ECS lineage within each γ category. We found that regions with γ close to zero consistently exhibited a lower number of deleterious alleles relative to the length of the coding sequences than those exhibited by the background (Figures S40 and S41). Conversely, several categories that were likely affected by introgression showed significantly higher numbers of deleterious alleles in 30, 50 and 100 kb windows. Similar trends were observed for the number of deleterious alleles in the SJ and PO lineages (Figures S42–S45).

4 | DISCUSSION

4.1 | The formation history of the ECS lineage with ghost introgression

The primary objective of this study was to test the hypothesis of the ghost introgression origin of the ECS lineage of *C. annularis*. Our analyses based on extensive sampling and high-throughput data confirmed the existence of three allopatric lineages, the SJ, PO and ECS lineages, and the presence of the mitonuclear discordance in the ECS lineage. Allelic pattern analysis showed the hybrid origin of the ECS lineage, supporting our previous speculations that the mitonuclear discordance has its origin in hybridization (Kato et al., 2021). By explicitly incorporating the ghost lineage into the demographic modelling, we showed that the ECS lineage was most likely formed by hybridization with the ghost lineage that diverged from the PO lineage at early time. Notably, despite our comprehensive sampling efforts (Kato et al., 2021), including this study, we did not identify any extant populations corresponding to the ghost lineage, indicating that it is indeed an extinct lineage rather than only an unsampled population. Although the contribution of the extinct lineage to the genome of the ECS lineage is limited to mtDNA and a few parts of nuclear genome, it certainly plays a role in genome construction and contributes to the current diversity and phylogeographic patterns of this species. The demographic modelling analysis, along with the stabilized hybrid mosaic genome of the ECS lineage, suggests that the ECS lineage was formed by an ancient hybridization event during the Pleistocene. This formation history of the ECS lineage, together with the occurrence of hybrid populations between lineages, suggests the multiple rounds of isolation and connection events. Therefore, our results emphasize that the complex geological history of the North-Western Pacific has influenced the current patterns of coastal biodiversity.

4.2 | Characterizing the genomic landscape of introgression from the extinct ghost lineage

It is generally difficult to detect regions derived from unsampled ghost lineages, as many of the existing methods for introgression detection assume the availability of samples from the parental lineages (Hibbins & Hahn, 2022; Patterson et al., 2012). However, in case where other analyses support the existence of a ghost lineage and its relationship with extant lineages, as in this study, these methods can prove useful for identifying or localizing introgression from ghost lineages. Using forward simulations based on the estimated demography, we showed that the statistic γ in HyDe analysis employed in a sliding window approach, which is easy to apply to non-model species, offers a practical means of estimating local ancestry for ghost introgression in this species. As a similar approach of attempting to detect ghosts by leveraging knowledge of demography has been effective in hominin introgression (Durvasula & Sankararaman, 2019), combining backward simulation for demographic estimation and forward simulation for evaluation will expand the possibilities for studying ghost lineages in non-model species (Ottenburghs, 2020).

We investigated the genomic landscape of introgression in the ECS lineage estimated using the sliding window γ and identified noteworthy patterns in regions that resisted introgression from the extinct ghost lineage. The regions with γ values close to zero, indicating minimal contribution from the extinct lineage, exhibited a lower population recombination rate and a higher proportion of coding sequences than that of the genomic background. The pattern of reduced introgression in low-recombination regions is the best-known feature of the genomic landscapes of introgression commonly observed in various taxa (human: Sankararaman et al., 2014; Schumer et al., 2018, baboon: Vilgalys et al., 2022, mouse: Janoušek et al., 2015, swordtail, Schumer et al., 2018; Langdon et al., 2022, stickleback: Ravinet et al., 2018; Yamasaki et al., 2020, *Heliconius* butterfly: Martin et al., 2019, monkeyflower: Aeschbacher et al., 2017, Oak: Fu et al., 2022). This pattern, which is unexpected in neutral conditions, is considered to be the result from linked selection that purges multiple regions where the donor is harmful in the recipient genome (Schumer et al., 2018). Furthermore, given that the accumulation of genetic incompatibilities reinforces reproductive isolation between divergent lineages, this pattern is often interpreted as the effect of isolation barriers at multiple loci to eliminate minor parent ancestry (Martin et al., 2019; Ravinet et al., 2018; Schumer et al., 2018). Indeed, the high gene density in non-introgressed regions observed in this study as well as several other studies (Fu et al., 2022; Sankararaman et al., 2014; Schumer et al., 2016; Vilgalys et al., 2022) coincides with the notion that incompatibilities should be linked with functional importance. Therefore, it is plausible that the landscape of introgression in the ECS lineage has been shaped by selection against the extinct ghost lineages, indicating the presence of pre-existing incompatibilities at the time of hybridization.

These aforementioned features consistently indicate the potential deleterious effects of introgression, such as incompatibility resulting from the extinct ghost lineage. However, it is also possible that the mutation loads accumulated in the extinct lineage was the target of selection (Kim et al., 2018). If the extinct ghost lineage carried a high level of mutation load, we might see differences in the number of deleterious alleles between introgressed regions and non-introgressed regions, only in the ECS lineage. Although the number of potentially deleterious alleles in the ECS lineage certainly tended to be low in the non-introgressed region

and somewhat high in the introgressed region, similar trends were observed for deleterious alleles in the SJ and the PO lineages. Since the deleterious mutation search conducted by PROVEAN is based on phylogenetic conservation and the target variants were limited to non-synonymous polymorphisms within the species, in our case, the consistent results among the three lineages seemed to reflect the strength of evolutionary constraints rather than actual mutation loads. This perspective aligns with findings that conserved elements have shown resistance to introgression in hybrid populations of swordtail fish (Schumer et al., 2016). Hence, the landscape of introgression from the ghost lineage in the ECS lineage may have been formed by selection, which eliminated minor parent contributions in regions under strong evolutionary constraints, along with the local recombination rates that either amplified or attenuated the effect of selection. The consistent findings across diverse taxa suggest that the evolutionary forces shaping the genome-wide ancestry may be widely shared throughout the tree of life, including cases of ghost introgression.

While the landscape of introgression may be characterized primarily by the selective forces eliminating minor parent ancestry, it is also possible that the introgressed regions are favoured through adaptive introgression or the avoidance of incompatibilities. Adaptive introgression has been well documented in a few extensively studied system, such as the wing colour patterns of the *Heliconius* butterflies (Pardo-Diaz et al., 2012) and the adaptation of the human population through ancient hominid alleles (Racimo et al., 2015). There are also limited examples of potential co-introgression of N-mt genes and mtDNA to resolve mitonuclear incompatibilities (Beck et al., 2015; Morales et al., 2018). However, we were not able to provide strong evidence supporting the accumulation of genes with specific functions in highly introgressed regions (regions with γ close to one). Although we speculated that N-mt genes and mtDNA may be co-introgressed because highly differentiated mtDNA of the ghost lineage was placed in the different nuclear genomic background, but N-mt genes was not significantly enriched in introgressed regions. We note that our observations only reflect overall trends across genomes and do not rule out the possibility that a small number of genes were adaptively introgressed or co-introgressed. In fact, signals of adaptive introgression can be restricted to a few small genomic regions in well-studied systems (Moest et al., 2020). Moreover, Moran et al. (2021), preprint available on bioRxiv: https://doi.org/10.1101/2021.07.13.452279 recently identified only two N-mt genes in OXPHOS complex I as responsible for a lethal mitonuclear hybrid incompatibility in natural hybrid populations of swordtail fish. Nevertheless, our analyses indicate that these factors are not at least the primary determinants shaping the genomic landscape of introgression from the ghost lineage in the ECS lineage. To gain a comprehensive understanding of the functional roles of the region inherited from the ghost lineage, further investigations focused on specific functions and genes, coupled with phenotypic observations, may be warranted.

We also found that the proportion of repetitive sequences, especially TEs, characterized the landscapes of introgression well. We consistently found that as the estimated proportion of introgression in windows increased, so did the density of repetitive sequences. To explore the generality of this relationship, we re-analysed the data of introgression between extant lineages (Japanese sticklebacks and *Heliconius* butterflies; Martin et al., 2019; Ravinet et al., 2018) and observed similar trends in both cases. Although few studies have focused on the relationship between introgression and repetitive sequences, it was recently reported that the introgressed loci were in the vicinity of TEs in the nine-spined sticklebacks in the White Sea (Nedoluzhko et al., 2022). It is unclear why the estimated introgressed regions are repeat rich. One possible reason is that the density of repetitive sequences may simply reflect some correlation with gene density (Wright et al., 2003) or recombination rate (Kent et al., 2017). However, this seems unlikely because the density of repetitive sequences prominently characterized the landscape of introgression even after controlling for the coding sequence density or recombination rate. It is also possible that genotyping errors associated with the difficulty of genotyping in and around repeat-rich regions caused biases by creating allele patterns that do not follow the species tree, but the similar trends observed across studies using different data and variant calling methods suggest some other effects that cannot be explained by this factor alone. For instance, chromosomal structures linked to repetitive sequences or certain functions of TEs may facilitate introgression of surrounding DNA (Serrato-Capuchina & Matute, 2018). It is now clear that some repetitive sequences are important for gene regulation (Bonchev & Parisod, 2013; Chuong et al., 2017), and the inheritance of repeat-rich regions from local lineages may have been partially adaptive, as suggested by the study of introgression between oak trees (Fu et al., 2022). However, this study provides genome-wide trends for only a few species, and further extensive and detailed investigations are required to test the generality, underlying mechanisms, functions and potential biases associated with the relationship between introgression and repetitive sequences.

In conclusion, our study has shed light on the origin of the ECS lineage in *C. annularis*, originated through ancient hybridization with an extinct ghost lineage, as well as characterized the genomic landscape of introgression from this extinct ghost lineage. Our study represents one of the most detailed phylogeographic studies to date in the coastal region of the North-Western Pacific, and it stands as the first example to elucidate the patterns of genomic survival of extinct ghost lineages. The trends observed in regions where the extinct lineage did not introgress coincide with those of the genomic landscapes of introgression between extant organisms, suggesting the role of selection in eliminating the extinct lineage ancestry as a minor parent to shape the hybrid genome. Additionally, we found an intriguing and less-known feature, positive correlation between the density of repetitive elements and the estimated rate of introgression. This correlation appears to be common in certain instances of introgression between extant lineages. Our findings underscore the unexpected similarities in the characteristics of introgression landscapes across different taxa, even in cases involving unsampled 'ghost' lineages.

AUTHOR CONTRIBUTIONS

Shuya Kato and Shotaro Hirase designed the study. Shuya Kato, Shotaro Hirase and Seiji Arakaki collected samples. Shuya Kato and Atsushi J. Nagano performed wet laboratory experiments and collected data. Shuya Kato wrote the manuscript with the help of Shotaro Hirase and Kiyoshi Kikuchi. All authors approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We express great gratitude to the members of the laboratory for their helpful comments on this research. We are grateful to Satoko Kondo for conducting the ddRAD-seq analyses and Dr. Mark Ravinet for helpful discussion prior to the analyses. We also offer our thanks to the anonymous reviewers for their valuable suggestions. This work was supported by the Japan Society for the Promotion of Science (KAKENHI 18H02493, 22H00377, 22J12643). Computations were partially performed on the NIG supercomputer at the ROIS National Institute of Genetics.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Raw sequence reads newly obtained in this study are deposited in the DNA Data Bank of Japan (DDBJ). The DDBJ accession numbers are DRR489922–DRR490073 for ddRAD-seq, DRR489903–DRR489921 for whole genome re-sequencing and DRR490074–DRR490084 for RNA-seq. Accession numbers for each sequence read used in this study are listed in Table S2. Metadata of all sequence reads are also stored in DDBJ (BioProject: PRJDB16162). Genotyping data of each dataset, gene annotation file, results of demographic modelling and sliding window analysis and the scripts used in this study are available on Dryad https://doi.org/10.5061/dryad.7wm37pw09.

BENEFIT-SHARING STATEMENT

Benefits Generated: Although our research deals only with samples collected in Japan, benefits from this research accrue from the sharing of our data and results on public databases as described earlier.

ORCID

Shuya Kato https://orcid.org/0000-0002-6661-7586
Kiyoshi Kikuchi https://orcid.org/0000-0001-5435-198X

REFERENCES

Aeschbacher, S., Selby, J. P., Willis, J. H., & Coop, G. (2017). Population-genomic inference of the strength and timing of selection against gene flow. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(27), 7061–7066. https://doi.org/10.1073/pnas.1616755114

Ai, H., Fang, X., Yang, B., Huang, Z., Chen, H., Mao, L., Zhang, F., Zhang, L., Cui, L., He, W., Yang, J., Yao, X., Zhou, L., Han, L., Li, J., Sun, S., Xie, X., Lai, B., Su, Y., … Huang, L. (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature Genetics*, *47*(3), 217–225. https://doi.org/10.1038/ng.3199

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings of the second international symposium on information theory* (pp. 267–281). Akadémiai Kiado.

Akihito, S. K., Ikeda, Y., & Aizawa, M. (2013). Gobioidei. In T. Nakabo (Ed.), *Fishes of Japan with pictorial keys to the species* (3rd ed., pp. 1347–1608, 2109–2211). Tokai University Press (in Japanese).

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Allio, R., Donega, S., Galtier, N., & Nabholz, B. (2017). Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: Implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. *Molecular Biology and Evolution*, *34*(11), 2762–2772. https://doi.org/10.1093/molbev/msx197

Asahida, T., Kobayashi, T., Saitoh, K., & Nakayama, I. (1996). Tissue preservation and total DNA extraction form fish stored at ambient temperature using buffers containing high concentration of urea. *Fisheries Science*, *62*(5), 727–730. https://doi.org/10.2331/fishsci.62.727

Avise, J. C. (2000). *Phylogeography: The history and formation of species*. Harvard University Press.

Barlow, A., Cahill, J. A., Hartmann, S., Theunert, C., Xenikoudakis, G., Fortes, G. G., Pajimans, L. A. J., Rabeder, G., Frischauf, C., Grandal-d'Anglade, A., García-Vázquez, A., Murtskhvaladze, M., Saarma, A. P., Skrbinšek, T., Bertorelle, G., Gasparian, B., Bar-Oz, G., Pinhasi, R., … Hofreiter, M. (2018). Partial genomic survival of cave bears in living brown bears. *Nature Ecology & Evolution*, *2*(10), 1563–1570. https://doi.org/10.1038/s41559-018-0654-8

Beck, E. A., Thompson, A. C., Sharbrough, J., Brud, E., & Llopart, A. (2015). Gene flow between *Drosophila yakuba* and *Drosophila santomea* in subunit V of cytochrome *c* oxidase: A potential case of cytonuclear cointrogression. *Evolution*, *69*(8), 1973–1986. https://doi.org/10.1111/evo.12718

Blischak, P. D., Chifman, J., Wolfe, A. D., & Kubatko, L. S. (2018). HyDe: A python package for genome-scale hybridization detection. *Systematic Biology*, *67*(5), 821–829. https://doi.org/10.1093/sysbio/syy023

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bonchev, G., & Parisod, C. (2013). Transposable elements and microevolutionary changes in natural populations. *Molecular Ecology Resources*, *13*(5), 765–775. https://doi.org/10.1111/1755-0998.12133

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, *3*(1), lqaa108. https://doi.org/10.1093/nargab/lqaa108

Chan, A. H., Jenkins, P. A., & Song, Y. S. (2012). Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genetics*, *8*(12), e1003090. https://doi.org/10.1371/journal.pgen.1003090

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., Zhang, X., Wang, J., Yang, H., Fang, L., & Chen, Q. (2018). SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience*, *7*(1), 1–6. https://doi.org/10.1093/gigascience/gix120

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, *7*(10), e46688. https://doi.org/10.1371/journal.pone.0046688

Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics*, *18*(2), 71–86. https://doi.org/10.1038/nrg.2016.139

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, J. S., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3. *Fly (Austin)*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Delaneau, O., Marchini, J., & Zagury, J. F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, *9*(2), 179–181. https://doi.org/10.1038/nmeth.1785

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252. https://doi.org/10.1093/molbev/msr048

Durvasula, A., & Sankararaman, S. (2019). A statistical model for reference-free inference of archaic local ancestry. *PLoS Genetics*, *15*(5), e1008175. https://doi.org/10.1371/journal.pgen.1008175

Ellegren, H. (2000). Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics*, *16*(12), 551–558. https://doi.org/10.1016/S0168-9525(00)02139-9

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*(10), e1003905. https://doi.org/10.1371/journal.pgen.1003905

Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). *fastsimcoal2*: Demographic inference under complex evolutionary scenarios. *Bioinformatics*, *37*(24), 4882–4885. https://doi.org/10.1093/bioinformatics/btab468

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Frei, D., De-Kayne, R., Selz, O. M., Seehausen, O., & Feulner, P. G. (2022). Genomic variation from an extinct species is retained in the extant radiation following speciation reversal. *Nature Ecology & Evolution*, *6*(4), 461–468. https://doi.org/10.1038/s41559-022-01665-7

Frei, D., Reichlin, P., Seehausen, O., & Feulner, P. G. (2023). Introgression from extinct species facilitates adaptation to its vacated niche. *Molecular Ecology*, *32*(4), 841–853. https://doi.org/10.1111/mec.16791

Fu, R., Zhu, Y., Liu, Y., Feng, Y., Lu, R. S., Li, Y., Li, P., Kremer, A., Lascoux, M., & Chen, J. (2022). Genome-wide analyses of introgression between two sympatric Asian oak species. *Nature Ecology & Evolution*, *6*(7), 924–935. https://doi.org/10.1038/s41559-022-01754-7

Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M., & Stanke, M. (2021). TSEBRA: Transcript selector for BRAKER. *BMC Bioinformatics*, *22*(1), 1–12. https://doi.org/10.1186/s12859-021-04482-0

Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S., & Prins, P. (2022). A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Computational Biology*, *18*(5), e1009123. https://doi.org/10.1371/journal.pcbi.1009123

Geneva, A. J., Muirhead, C. A., Kingan, S. B., & Garrigan, D. (2015). A new method to scan genomes for introgression in a secondary contact model. *PLoS One*, *10*(4), e0118621. https://doi.org/10.1371/journal.pone.0118621

Gittelman, R. M., Schraiber, J. G., Vernot, B., Mikacenic, C., Wurfel, M. M., & Akey, J. M. (2016). Archaic hominin admixture facilitated adaptation to out-of-Africa environments. *Current Biology*, *26*(24), 3375–3382. https://doi.org/10.1016/j.cub.2016.10.041

Gompert, Z., Lucas, L. K., Buerkle, C. A., Forister, M. L., Fordyce, J. A., & Nice, C. C. (2014). Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology*, *23*(18), 4555–4573. https://doi.org/10.1111/mec.12811

Gopalakrishnan, S., Sinding, M. H. S., Ramos-Madrigal, J., Niemann, J., Castruita, J. A. S., Vieira, F. G., Carøe, C., de Manuel Montero, M., Kuderna, L., Serres, A., González-Basallote, V. M., Liu, Y. H., Wang, G. D., Marques-Bonet, T., Mirarab, S., Fernandes, C., Gaubert, P., Koepfli, K. P., Budd, J., … Gilbert, M. T. P. (2018). Interspecific gene flow shaped the evolution of the genus *Canis*. *Current Biology*, *28*(21), 3441–3449. https://doi.org/10.1016/j.cub.2018.08.041

Gorbarenko, S. A., & Southon, J. R. (2000). Detailed Japan Sea paleoceanography during the last 25 kyr: Constraints from AMS dating and $\delta^{18}O$ of planktonic foraminifera. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *156*(3–4), 177–193. https://doi.org/10.1016/S0031-0182(99)00137-6

Grant, K. M., Rohling, E. J., Bar-Matthews, M., Ayalon, A., Medina-Elizalde, M., Bronk Ramsey, C., Satow, C., & Roberts, A. P. (2012). Rapid coupling between ice volume and polar temperature over the past 150,000 years. *Nature*, *491*(7426), 744–747. https://doi.org/10.1038/nature11593

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H. Y., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., … Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Guan, Y. (2014). Detecting structure of haplotypes and local ancestry. *Genetics*, *196*(3), 625–642. https://doi.org/10.1534/genetics.113.160697

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., & Ralph, P. L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, *19*(2), 552–566. https://doi.org/10.1111/1755-0998.12968

Haller, B. C., & Messer, P. W. (2019). SLiM 3: forward genetic simulations beyond the Wright–Fisher model. *Molecular Biology and Evolution*, *36*(3), 632–637. https://doi.org/10.1093/molbev/msy228

Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., & Wall, J. D. (2011). Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(37), 15123–15128. https://doi.org/10.1073/pnas.1109300108

Harris, K., & Nielsen, R. (2016). The genetic cost of Neanderthal introgression. *Genetics*, *203*(2), 881–891. https://doi.org/10.1534/genetics.116.186890

Hibbins, M. S., & Hahn, M. W. (2022). Phylogenomic approaches to detecting and characterizing introgression. *Genetics*, *220*(2), iyab173. https://doi.org/10.1093/genetics/iyab173

Hirase, S., & Ikeda, M. (2014). Long-term vicariance and post-glacial expansion in the Japanese rocky intertidal goby *Chaenogobius annularis*. *Marine Ecology Progress Series*, *499*, 217–231. https://doi.org/10.3354/meps10641

Hirase, S., & Ikeda, M. (2015). Hybrid population of highly divergent groups of the intertidal goby *Chaenogobius annularis*. *Journal of Experimental Marine Biology and Ecology*, *473*, 121–128. https://doi.org/10.1016/j.jembe.2015.08.010

Hirase, S., Ikeda, M., Kanno, M., & Kijima, A. (2012). Phylogeography of the intertidal goby *Chaenogobius annularis* associated with paleo-environmental changes around the Japanese archipelago. *Marine Ecology Progress Series*, *450*, 167–179. https://doi.org/10.3354/meps09584

Hirase, S., Kanno, M., Ikeda, M., & Kijima, A. (2012). Evidence of the restricted gene flow within a small spatial scale in the Japanese common intertidal goby *Chaenogobius annularis*. *Marine Ecology*, *33*(4), 481–489. https://doi.org/10.1111/j.1439-0485.2012.00512.x

Hirase, S., Takeshima, H., Nishida, M., & Iwasaki, W. (2016). Parallel mitogenome sequencing alleviates random rooting effect in phylogeography. *Genome Biology and Evolution*, *8*(4), 1267–1278. https://doi.org/10.1093/gbe/evw063

Hirase, S., Tezuka, A., Nagano, A. J., Sato, M., Hosoya, S., Kikuchi, K., & Iwasaki, W. (2021). Integrative genomic phylogeography reveals signs of mitonuclear incompatibility in a natural hybrid goby population. *Evolution*, *75*(1), 176–194. https://doi.org/10.1111/evo.14120

Hughes, P. D., Gibbard, P. L., & Ehlers, J. (2013). Timing of glaciation during the last glacial cycle: Evaluating the concept of a global 'last glacial maximum'(LGM). *Earth-Science Reviews*, *125*, 171–198. https://doi.org/10.1016/j.earscirev.2013.07.003

Janoušek, V., Munclinger, P., Wang, L., Teeter, K. C., & Tucker, P. K. (2015). Functional organization of the genome may shape the species boundary in the house mouse. *Molecular Biology and Evolution*, *32*(5), 1208–1220. https://doi.org/10.1093/molbev/msv011

Kato, S., Arakaki, S., Kikuchi, K., & Hirase, S. (2021). Complex phylogeographic patterns in the intertidal goby *Chaenogobius annularis* around Kyushu Island as a boundary zone of three different seas. *Ichthyological Research*, *68*, 86–100. https://doi.org/10.1007/s10228-020-00772-4

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kent, T. V., Uzunović, J., & Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1736), 20160458. https://doi.org/10.1098/rstb.2016.0458

Kim, B. Y., Huber, C. D., & Lohmueller, K. E. (2018). Deleterious variation shapes the genomic landscape of introgression. *PLoS Genetics*, *14*(10), e1007741. https://doi.org/10.1371/journal.pgen.1007741

Kojima, S., Hayashi, I., Kim, D., Iijima, A., & Furota, T. (2004). Phylogeography of an intertidal direct-developing gastropod *Batillaria cumingi* around the Japanese Islands. *Marine Ecology Progress Series*, *276*, 161–172. https://doi.org/10.3354/meps276161

Kokita, T., & Nohara, K. (2011). Phylogeography and historical demography of the anadromous fish *Leucopsarion petersii* in relation to geological history and oceanography around the Japanese archipelago. *Molecular Ecology*, *20*(1), 143–164. https://doi.org/10.1111/j.1365-294X.2010.04920.x

Korunes, K. L., & Samuk, K. (2021). PIXY: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*(4), 1359–1368. https://doi.org/10.1111/1755-0998.13326

Kuhlwilm, M., Han, S., Sousa, V. C., Excoffier, L., & Marques-Bonet, T. (2019). Ancient admixture from an extinct ape lineage into bonobos. *Nature Ecology & Evolution*, *3*(6), 957–965. https://doi.org/10.1038/s41559-019-0881-7

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Langdon, Q. K., Powell, D. L., Kim, B., Banerjee, S. M., Payne, C., Dodge, T. O., Moran, B., Fascinetto-Zago, P., & Schumer, M. (2022). Predictability and parallelism in the contemporary evolution of hybrid genomes. *PLoS Genetics*, *18*(1), e1009914. https://doi.org/10.1371/journal.pgen.1009914

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Malinsky, M., Challis, R. J., Tyers, A. M., Schiffels, S., Terai, Y., Ngatunga, B. P., Miska, E. A., Durbin, R., Genner, M. J., & Turner, G. F. (2015). Genomic islands of speciation separate cichlid ecomorphs in an east African crater lake. *Science*, *350*(6267), 1493–1498. https://doi.org/10.1126/science.aac9927

Malinsky, M., Matschiner, M., & Svardal, H. (2021). Dsuite-fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, *21*(2), 584–595. https://doi.org/10.1111/1755-0998.13265

Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F., & Durbin, R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution*, *2*(12), 1940–1955. https://doi.org/10.1038/s41559-018-0717-x

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, *20*(5), 229–237. https://doi.org/10.1016/j.tree.2005.02.010

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., & Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*, *23*(11), 1817–1828. https://doi.org/10.1101/gr.159426.113

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2015). Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, *32*(1), 244–257. https://doi.org/10.1093/molbev/msu269

Martin, S. H., Davey, J. W., Salazar, C., & Jiggins, C. D. (2019). Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biology*, *17*(2), e2006288. https://doi.org/10.1371/journal.pbio.2006288

Martin, S. H., & Jiggins, C. D. (2017). Interpreting the genomic landscape of introgression. *Current Opinion in Genetics & Development*, *47*, 69–74. https://doi.org/10.1016/j.gde.2017.08.007

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella1, K., Altshuler, D., Gabriel1, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Moest, M., Van Belleghem, S. M., James, J. E., Salazar, C., Martin, S. H., Barker, S. L., Moreira, G. R. P., Mérot, C., Joron, M., Nadeau, N. J., Steiner, F. M., & Jiggins, C. D. (2020). Selective sweeps on novel and introgressed variation shape mimicry loci in a butterfly adaptive radiation. *PLoS Biology*, *18*(2), e3000597. https://doi.org/10.1371/journal.pbio.3000597

Momigliano, P., Florin, A. B., & Merilä, J. (2021). Biases in demographic modeling affect our understanding of recent divergence. *Molecular*

Biology and Evolution, *38*(7), 2967–2985. https://doi.org/10.1093/molbev/msab047

Morales, H. E., Pavlova, A., Amos, N., Major, R., Kilian, A., Greening, C., & Sunnucks, P. (2018). Concordant divergence of mitogenomes and a mitonuclear gene cluster in bird lineages inhabiting different climates. *Nature Ecology & Evolution*, *2*(8), 1258–1267. https://doi.org/10.1038/s41559-018-0606-3

Moran, B. M., Payne, C. Y., Powell, D. L., Iverson, E. N., Banerjee, S. M., Langdon, Q. K., Liu, F., Matney, R., Singhal, K., Leib, R. D., Hernandez-Perez, O., Corbett-Detig, R., Frydman, J., Schartl, M., Hvird, J. C., & Schumer, M. (2021). A lethal genetic incompatibility between naturally hybridizing species in mitochondrial complex I. *BioRxiv*, 1–49. https://doi.org/10.1101/2021.07.13.452279

Nedoluzhko, A., Sharko, F., Tsyganova, S., Boulygina, E. S., Slobodova, N., Teslyuk, A., Galindo-Villegas, J., & Rastorguev, S. (2022). Intergeneric hybridization of two stickleback species leads to introgression of membrane-associated genes and invasive TE expansion. *Frontiers in Genetics*, *13*, 863547. https://doi.org/10.3389/fgene.2022.863547

Nogués-Bravo, D., Ohlemüller, R., Batra, P., & Araújo, M. B. (2010). Climate predictors of late quaternary extinctions. *Evolution*, *64*(8), 2442–2449. https://doi.org/10.1111/j.1558-5646.2010.01009.x

Nouhaud, P., Martin, S. H., Portinha, B., Sousa, V. C., & Kulmuni, J. (2022). Rapid and predictable genome evolution across three hybrid ant populations. *PLoS Biology*, *20*(12), e3001914. https://doi.org/10.1371/journal.pbio.3001914

Oba, T., Kato, M., Kitazato, H., Koizumi, I., Omura, A., Sakai, T., & Takayama, T. (1991). Paleoenvironmental changes in the Japan Sea during the last 85,000 years. *Paleoceanography*, *6*(4), 499–518. https://doi.org/10.1029/91PA00560

Ottenburghs, J. (2020). Ghost introgression: Spooky gene flow in the distant past. *BioEssays*, *42*(6), 2000012. https://doi.org/10.1002/bies.202000012

Palkopoulou, E., Lipson, M., Mallick, S., Nielsen, S., Rohland, N., Baleka, S., Karpinski, E., Ivancevic, A. M., To, T. H, Kortschak, R. D., Raison, J. R., Qu, Z., Chin, T. J., Alt, K. W., Claesson, S., Dalén, L., MacPhee, R. D. E., Meller, H., Roca, A. L., … Reich, D. (2018). A comprehensive genomic history of extinct and living elephants. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), E2566–E2574. https://doi.org/10.1073/pnas.1720554115

Pardo-Diaz, C., Salazar, C., Baxter, S. W., Merot, C., Figueiredo-Ready, W., Joron, M., McMillan, W. O., & Jiggins, C. D. (2012). Adaptive introgression across species boundaries in *Heliconius* butterflies. *PLoS Genetics*, *8*(6), e1002752. https://doi.org/10.1371/journal.pgen.1002752

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient Swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, *31*(7), 1929–1936. https://doi.org/10.1093/molbev/msu136

Pritchard, V. L., & Edmands, S. (2013). The genomic trajectory of hybrid swarms: Outcomes of repeated crosses between populations of *Tigriopus californicus*. *Evolution*, *67*(3), 774–791. https://doi.org/10.1111/j.1558-5646.2012.01814.x

Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, *16*(6), 359–371. https://doi.org/10.1038/nrg3936

Ravinet, M., Yoshida, K., Shigenobu, S., Toyoda, A., Fujiyama, A., & Kitano, J. (2018). The genomic landscape at a late stage of stickleback speciation: High genomic divergence interspersed by small localized regions of introgression. *PLoS Genetics*, *14*(5), e1007358. https://doi.org/10.1371/journal.pgen.1007358

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., … Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature*, *468*(7327), 1053–1060. https://doi.org/10.1038/nature09710

Ru, D., Sun, Y., Wang, D., Chen, Y., Wang, T., Hu, Q., Abbott, R. J., & Liu, J. (2018). Population genomic analysis reveals that homoploid hybrid speciation can be a lengthy process. *Molecular Ecology*, *27*(23), 4875–4887. https://doi.org/10.1111/mec.14909

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Sakaguchi, S., Sugino, T., Tsumura, Y., Ito, M., Crisp, M. D., Bowman, D. M., Nagano, A. J., Honjo, M. N., Yasugi, M., Kudoh, H., Matsuki, Y., Suyama, Y., & Isagi, Y. (2015). High-throughput linkage mapping of Australian white cypress pine (*Callitris glaucophylla*) and map transferability to related species. *Tree Genetics & Genomes*, *11*(6), 1–12. https://doi.org/10.1007/s11295-015-0944-0

Sandel, B., Arge, L., Dalsgaard, B., Davies, R. G., Gaston, K. J., Sutherland, W. J., & Svenning, J. C. (2011). The influence of late quaternary climate-change velocity on species endemism. *Science*, *334*(6056), 660–664. https://doi.org/10.1126/science.1210173

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., & Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, *507*(7492), 354–357. https://doi.org/10.1038/nature12961

Sasaki, T., & Hattori, J. (1969). Comparative ecology of two closely related sympatric gobiid fishes living in tide pools. *Japanese Journal of Ichthyology*, *15*(4), 143–155. (in Japanese with English abstract). https://doi.org/10.11369/jji1950.15.143

Schumer, M., Cui, R., Powell, D. L., Rosenthal, G. G., & Andolfatto, P. (2016). Ancient hybridization and genomic stabilization in a swordtail fish. *Molecular Ecology*, *25*(11), 2661–2679. https://doi.org/10.1111/mec.13602

Schumer, M., Xu, C., Powell, D. L., Durvasula, A., Skov, L., Holland, C., Blazier, J. C., Sankaraman, S., Andolfatto, P., Rosenthal, G. G., & Przeworski, M. (2018). Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science*, *360*(6389), 656–660. https://doi.org/10.1126/science.aar3684

Serrato-Capuchina, A., & Matute, D. R. (2018). The role of transposable elements in speciation. *Genes*, *9*(5), 254. https://doi.org/10.3390/genes9050254

Sloan, D. B., Havird, J. C., & Sharbrough, J. (2017). The on-again, off-again relationship between mitochondrial genomes and species boundaries. *Molecular Ecology*, *26*(8), 2212–2236. https://doi.org/10.1111/mec.13959

Smit, A. F. A., Hubley, R., & Green, P. (2010). *RepeatMasker Open-4.0. 2013–2015*. http://www.repeatmasker.org

Snyder, C. W. (2016). Evolution of global temperature over the past two million years. *Nature*, *538*(7624), 226–228. https://doi.org/10.1038/nature19798

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*, *31*(12), 2032–2034. https://doi.org/10.1093/bioinformatics/btv098

Taylor, S. A., & Larson, E. L. (2019). Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology & Evolution*, *3*(2), 170–177. https://doi.org/10.1038/s41559-018-0777-y

Toews, D. P., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, *21*(16), 3907–3930. https://doi.org/10.1111/j.1365-294X.2012.05664

Vilgalys, T. P., Fogel, A. S., Anderson, J. A., Mututua, R. S., Warutere, J. K., Siodi, I. L. I., Kim, S. Y., Voyles, T. N., Robinson, J. A., Wall, J. D., Archie, E. A., Alberts, S. C., & Tung, J. (2022). Selection against admixture and gene regulatory divergence in a long-term primate field study. *Science*, *377*(6606), 635–641. https://doi.org/10.1126/science.abm4917

Wang, M. S., Wang, S., Li, Y., Jhala, Y., Thakur, M., Otecko, N. O., Si, J. F., Chen, H. M., Shapiro, B., Nielsen, R., Zhang, Y. P., & Wu, D. D. (2020). Ancient hybridization with an unknown population facilitated high-altitude adaptation of canids. *Molecular Biology and Evolution*, *37*(9), 2616–2629. https://doi.org/10.1093/molbev/msaa113

Weir, B. S., & Cockerham, C. C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.2307/2408641

Wright, S. I., Agrawal, N., & Bureau, T. E. (2003). Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. *Genome Research*, *13*(8), 1897–1903. https://doi.org/10.1101/gr.1281503

Yamasaki, Y. Y., Kakioka, R., Takahashi, H., Toyoda, A., Nagano, A. J., Machida, Y., Møller, P. R., & Kitano, J. (2020). Genome-wide patterns of divergence and introgression after secondary contact between *Pungitius* sticklebacks. *Philosophical Transactions of The Royal Society B Biological Sciences*, *375*(1806), 20190548. https://doi.org/10.1098/rstb.2019.0548

You, X., Bian, C., Zan, Q., Xu, X., Liu, X., Chen, J., Wang, J., Qiu, Y., Li, W., Zhang, X., Sun, Y., Chen, S., Hong, W., Li, Y., Cheng, S., Fan, G., Shi, C., Liang, J., Tang, Y. T., … Shi, Q. (2014). Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. *Nature Communications*, *5*(1), 1–8. https://doi.org/10.1038/ncomms6594

Zeberg, H., & Pääbo, S. (2020). The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature*, *587*(7835), 610–612. https://doi.org/10.1038/s41586-020-2818-3

Zeberg, H., & Pääbo, S. (2021). A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(9), e2026309118. https://doi.org/10.1073/pnas.2026309118

Zhang, D., Tang, L., Cheng, Y., Hao, Y., Xiong, Y., Song, G., Qu, Y., Rheindt, F. E., Alström, P., Jia, C., & Lei, F. (2019). "Ghost introgression" as a cause of deep mitochondrial divergence in a bird species complex. *Molecular Biology and Evolution*, *36*(11), 2375–2386. https://doi.org/10.1093/molbev/msz170

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kato, S., Arakaki, S., Nagano, A. J., Kikuchi, K., & Hirase, S. (2023). Genomic landscape of introgression from the ghost lineage in a gobiid fish uncovers the generality of forces shaping hybrid genomes. *Molecular Ecology*, *00*, 1–20. https://doi.org/10.1111/mec.17216