Molecular Ecology
December 2024, Volume 33, Issue 24, Pages e17277 (21p.)
https://doi.org/10.1111/mec.17277
https://archimer.ifremer.fr/doc/00875/98678/

Divergence and gene flow history at two large chromosomal inversions underlying ecotype differentiation in the long-snouted seahorse

Meyer Laura [1, *], Barry Pierre [1, 2], Riquet Florentine [3], Foote Andrew [4], Der sarkissian Clio [5], Cunha Regina L. [6], Arbiol Christine [1], Cerqueira Frédérique [1], Desmarais Erick [1], Bordes Anaïs [1], Bierne Nicolas [1], Guinand Bruno [1], Gagnaire Pierre-alexandre [1, *]

[1] ISEM, Univ Montpellier, CNRS, EPHE, IRD Montpellier, France
[2] CIBIO-InBIO Centro de Investigação em Biodiversidade e Recursos Genéticos Universidade do Porto Vairão ,Portugal
[3] Ifremer, RBE-ASIM, Station de La Tremblade La Tremblade, France
[4] Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences University of Oslo Oslo ,Norway
[5] Centre for Anthropobiology and Genomics of Toulouse, CNRS University of Toulouse Paul Sabatier Toulouse ,France
[6] Centre of Marine Sciences-CCMAR University of Algarve Faro ,Portugal

* Corresponding authors : Laura Meyer, email address : laura.meyer@umontpellier.fr ; Pierre-Alexandre Gagnaire, email address : pierre-alexandre.gagnaire@umontpellier.fr

Abstract :

Chromosomal inversions can play an important role in divergence and reproductive isolation by building and maintaining distinct allelic combinations between evolutionary lineages. Alternatively, they can take the form of balanced polymorphisms that segregate within populations until one arrangement becomes fixed. Many questions remain about how inversion polymorphisms arise, how they are maintained over the long term, and ultimately, whether and how they contribute to speciation. The long-snouted seahorse (Hippocampus guttulatus) is genetically subdivided into geographic lineages and marine-lagoon ecotypes, with shared structural variation underlying lineage and ecotype divergence. Here, we aim to characterize structural variants and to reconstruct their history and suspected role in ecotype formation. We generated a near chromosome-level genome assembly and described genome-wide patterns of diversity and divergence through the analysis of 112 whole-genome sequences from Atlantic, Mediterranean, and Black Sea populations. By also analysing linked-read sequencing data, we found evidence for two chromosomal inversions that were several megabases in length and showed contrasting allele frequency patterns between lineages and ecotypes across the species range. We reveal that these inversions represent ancient intraspecific polymorphisms, one likely being maintained by divergent selection and the other by pseudo-overdominance. A possible selective coupling between the two inversions was further supported by the absence of specific haplotype combinations and a putative functional interaction between the two inversions in reproduction. Lastly, we detected gene flux eroding divergence between inverted alleles at varying levels for the two inversions, with a likely impact on their dynamics and contribution to divergence and speciation.

Keywords : ancestral recombination graph, ecotypes, evolutionary history, gene flux, inversions, whole-genome resequencing

Introduction

How genetic differences accumulate between nascent species and contribute to the build-up of reproductive isolation remains a fundamental question in evolutionary biology (Westram et al., 2022b). Mechanisms that bring unlinked or distant sites into linkage disequilibrium and that maintain strong allelic associations in the presence of gene flow are key drivers of reproductive isolation (Butlin, 2005; Ortiz-Barrientos et al., 2016; Tigano & Friesen, 2016). Large structural variants (SVs) such as chromosomal inversions combine these two properties, since they both establish and maintain divergent haplotypes by reducing inter-haplotype recombination (Faria & Navarro, 2010; Hoffmann & Rieseberg, 2008; Mérot et al., 2020a; Wellenreuther et al., 2019; Zhang et al., 2021). For these reasons, large SVs have often received special attention in speciation genomics studies, particularly those investigating ecotype formation (e.g. Atlantic cod, Berg et al., 2016; rough periwinkles, Faria et al., 2019b; deer mice, Harringmeyer & Hoekstra, 2022; seaweed flies, Mérot et al., 2020b; sunflowers, Todesco et al., 2020; monkeyflowers, Lowry & Willis, 2010). Despite the growing number of studies reporting the importance of inversions for divergence between ecotypes in a variety of organisms, many questions remain about the conditions in which these inversions arise, how they are maintained over the long term, and how they contribute to reproductive isolation and the progress towards speciation (Faria et al., 2019a; Kirkpatrick, 2010; Wellenreuther & Bernatchez, 2018; Westram et al., 2022a).

The most immediate role of inversions is generally ascribed to recombination suppression, which protects combinations of locally advantageous or co-adapted alleles locked within an inverted region (Butlin, 2005; Kirkpatrick, 2010; Noor et al., 2001; Rieseberg, 2001). Several studies have directly linked inversions to traits that differ between ecotypes, with functional implications in local adaptation or reproductive isolation (Campbell et al., 2021; Funk et al., 2021; Gould et al., 2017; Hager et al., 2022; Jay et al., 2021; Lundberg et al., 2023). These tend to be complex phenotypes that likely involve several genes inherited as a single block (Lamichhaney et al., 2016; Matschiner et al., 2022; Schwander et al., 2014; Thompson & Jiggins, 2014). Large inversions spanning megabase-sized chromosomal segments can facilitate the emergence of such complex phenotypes by building linkage disequilibrium between multiple distant functional sites (Thompson & Jiggins, 2014). However, inversions can also be expected to carry other types of selected sites, including co-adapted gene complexes and deleterious mutations that are captured by chance (Berdan et al., 2021; Jay et al., 2021; Nei et al., 1967). These different types of selected loci, as well as the potential interactions between them, imply that the dynamics of an inversion can be driven by multiple simultaneous processes (Faria et al., 2019a; Guerrero et al., 2012). Furthermore, the allelic content of an inversion is not fixed and can change over time, for example through the accumulation of new mutations or the exchange of genetic material by gene flux between inverted haplotypes (Cheng et al., 2012; Faria et al., 2019a; Matschiner et al., 2022; Navarro et al., 1997; Schaeffer & Anderson, 2005). This can occur through double crossover events with the formation of an inversion loop, or through non-crossover gene conversion during the repair of double-stranded breaks (Korunes & Noor, 2019; Matschiner et al., 2022). The resulting changes in allelic content could affect the processes that control the frequency of an inversion (Berdan et al., 2021), and ultimately, its evolutionary trajectory and contribution to reproductive isolation.

Considering the evolutionary history of inversions provides a particularly useful framework for understanding their role in speciation (Faria et al., 2019a). This framework aims to disentangle the roles of different evolutionary processes acting at different stages – from the appearance of inversions to their maintenance and long-term fate – in order to go beyond a simple inventory of the inversions present in a given system. In their review on the evolution of inversions, Faria et al. (2019a) distinguished two main classes of inversion polymorphisms, which they refer to as Type I and Type II. Type I inversions are expected to show frequency differences and divergence between populations, with the possible presence

of polymorphism within populations due to gene flow. This pattern may be caused by divergent selection acting on alternate haplotypes conferring local adaptation and/or a form of bistable selection involving frequency dependence, incompatibility selection or assortative mating. A Type I polymorphism may show signs of underdominance either due to direct effects (loss of unbalanced recombinant gametes during meiosis) or due to the allelic contents of the inversion (Faria et al., 2019a; Kirkpatrick & Barton, 2006). Over time, Type I inversions likely contribute to the accumulation of Dobzhansky-Müller (DM) incompatibilities and reinforcement mechanisms via coupling with other reproductive isolation polymorphisms (Kulmuni et al., 2020; Navarro & Barton, 2003). Alternatively, inversions can be maintained by a form of balancing selection (Jay et al., 2021; Mérot et al., 2020a; Wellenreuther & Bernatchez, 2018; Yeaman, 2013; Yeaman & Whitlock, 2011). Faria et al. (2019a) recognise such Type II inversions as polymorphisms that are maintained by frequency-dependence, disassortative mating, or antagonistic pleiotropy. Type II inversions might also contain overdominant loci and deleterious mutations in repulsion phase linkage (i.e. pseudo-overdominance), leading to heterokaryotype advantage (Marion & Noor, 2023). Indeed, some inversions can be expected to suffer from high mutation load, due to deleterious mutations hitchhiking with selected genes, or simply accumulating in low recombination regions (Hill & Robertson, 1966; Nei et al., 1967). These various processes associated with Type I and Type II polymorphisms are not necessarily mutually exclusive and might act in concert to shape the evolutionary dynamics of a given system. In particular, the role of potential interactions between different types of inversions remains a key question that needs to be addressed with empirical data.

Here, we aim to disentangle the different processes that affect the evolution of chromosomal inversions involved in ecotype differentiation in the long-snouted seahorse (*Hippocampus guttulatus*). Seahorses inhabit shallow coastal waters and littoral lagoons of the Northeast Atlantic, Mediterranean and Black Seas. Throughout their distribution, seahorse populations occupy a wide variety of habitats, spanning a gradient of environmental conditions from lagoon to marine habitats. These habitats notably differ in the composition of their species communities associated with the seagrass beds inhabited by seahorses (*Zostera*, *Cymodocea*, *Posidonia*), reflecting the ecological impact of variations in physical and chemical parameters across the marine-lagoon gradient (Perez-Ruzafa et al. 2017). The long-snouted seahorse is genetically subdivided into semi-isolated lineages separated by sharp eco-geographical boundaries (Riquet et al., 2019). In the Northeast Atlantic, seahorse populations are split into a northern and southern genetic lineage, while Mediterranean populations display ecotype divergence between marine and lagoon habitats. Although lagoon and marine ecotypes are morphologically similar, they show strong genetic differences concentrated in a large genomic island of divergence which also differentiates the two Atlantic lineages (Riquet et al., 2019). The genomic architecture of partial reproductive isolation between Mediterranean ecotypes on the one hand, and Atlantic lineages on the other hand, therefore probably relies on shared structural variation.

We introduce the first whole-genome study in *H. guttulatus*, aiming to characterise and understand the origin and evolution of the genomic landscape of divergence between lineages and ecotypes. We generate a near chromosome-level assembly of the *H. guttulatus* genome, and describe genetic variation across habitats throughout the range distribution of the species. We found the presence of not a single, but two inversions of megabase size, which may have played an important role in ecotype formation in the Mediterranean Sea. We then attempt to reconstruct the evolutionary history of these two inversions, especially focusing on their origin, divergence and subsequent evolution, including gene exchange between the inverted alleles. Our study highlights interactions between two inversions in the same system and the impact of gene flux between arrangements.

Materials and methods

Sampling and DNA extraction

A total of 112 samples from different habitats across the species range were used for whole-genome resequencing (Supplementary Table S1). Different DNA extraction protocols were applied to three types of samples that either originated from published studies, our own laboratory collection, or museum collections. (*i*) Samples from published studies (Riquet et al. 2019; Barry et al 2022) consisted of non-lethal fin or tail clips preserved in 95% ethanol at -20 °C. Their genomic DNA was extracted using the Nucleospin Tissue kit ® (Macherey–Nagel, Germany) following the manufacturer's protocol. (*ii*) Dried seahorse samples were obtained from private collections by means of a call to the public in a local newspaper. The approximate date (1935-2010) and collection site were recorded based on information provided by the donors. The dorsal fin of each dried seahorse was scratched to collect c.a. 20 µg of tissue powder. Genomic DNA was extracted using a standard CetylTrimethyl Ammonium Bromide (CTAB) Chloroform:Isoamyl alcohol (24:1) protocol (Doyle & Doyle, 1987). (*iii*) Lastly, four alcohol-preserved museum samples (1856-1898) were provided by the National Museum of Natural History (Paris, France). Tissue fragments (1-2 mm) were obtained by internal needlepoint sampling (Haÿ et al., 2020) and subjected to overnight enzymatic digestion (40 µL proteinase K at 25 mg/mL for 500 µL volume, at 56 °C). DNA was extracted using a phenol-chloroform method (Campos & Gilbert, 2012) in a dedicated clean lab facility located at the Institute of Evolutionary Science of Montpellier (ISEM, France).

Assembly and annotation of the *H. guttulatus* reference genome

We performed high-coverage linked-read sequencing of an Atlantic *H. guttulatus* individual from Hossegor lagoon (Bay of Biscay, Hgutt_GA_13) to generate a high-quality reference genome assembly (hereafter referred to as Hgutt_V1). Fresh gill and muscle tissue were solubilized in a 25 ml solution of TNES-Urea (10 mM Tris-HCl, 120 mM NaCl, 10 mM EDTA, 0.5% SDS, 4 M urea, PH 8) during 4 weeks at 20°C. High molecular weight genomic DNA (HMW gDNA) was isolated using three phenol-chloroform followed by two chloroform extractions after digestion with proteinase K (150 µg/ml). Final precipitation was performed using two volumes of 100% Ethanol. The resulting pellet was washed several times in 80% ethanol and resuspended in ultrapure water by heating to 65°C, before being kept at 40°C during 4 days. The length distribution of extracted DNA molecules was assessed by electrophoresis on a TapeStation Genomic DNA ScreenTape assay (Agilent Technologies). Single-stranded DNA damage was treated with the NEBNext FFPE DNA Repair mix and repaired DNA was then subjected to size selection to remove fragments shorter than 40 kb using a PippinHT instrument (Sage Science) with a 0.75% Agarose Gel Cassette. HMW gDNA was submitted to the 10x Genomics linked-read library preparation following the Chromium Genome Reagent Kit v2 protocol at the MGX sequencing facility (CNRS, Montpellier, France). The genome library was sequenced to ~100X on a S1 lane of an Illumina NovaSeq6000 in 150 bp paired-end mode by Genewiz Inc (USA), generating ~0.3 billion reads.

Raw demultiplexed reads were deduplicated using *nubeam-dedup* (Dai and Guan 2020) and processed with *process_10xReads.py* (https://github.com/ucdavis-bioinformatics/proc10xG) to extract the barcode sequence of each read pair and the number of read pairs associated to each barcode. The distribution of the number of read pairs per barcode was then analysed to identify rare barcodes potentially generated by sequencing errors and over-represented barcodes (Supplementary Fig. S1). A total of 143.7 million paired-end reads carrying 1.455 million retained barcodes were finally extracted using *proc10xG* scripts and used for linked-read-based *de novo* genome assembly using the *Supernova-2.1.1* software package (Weisenfeld et al., 2017). Assembled scaffolds were outputted in pseudohap style, with a minimum size set to 1 kb. Assembly statistics of the Hgutt_V1 reference genome were computed and visualised with the

BlobToolKit v3.5.2 software suite (Challis et al., 2020), combined with an assessment of genome assembly completeness with *BUSCO 5.4.4* (Manni et al., 2021) using the actinopterygii_odb10 fish dataset containing 3640 conserved genes. Whole-genome alignment was performed with *Minimap2* (Li, 2018) and visualised using *D-GENIES* (Cabanettes & Klopp, 2018) to compare and anchor *H. guttulatus* scaffolds to the chromosome-scale assembly of the closely related *H. erectus* (Li et al., 2021), provided by the authors. We also used the genome sequence of a north Atlantic long-snouted seahorse (UK), which was assembled by Iridian Genomes by ordering and orienting pre-assembled contigs based on other fish reference genomes (Accession PRJNA481552, hereafter called HguttRefA).

We used *RepeatModeler2* (Flynn et al., 2020) for *de novo* repeat finding and identification of the unique transposable element families present in the seahorse genome. In addition, we searched for tandem repeats (TRs) following the strategy developed in Melters et al. (2013), using the same parameter values to run *Tandem Repeats Finder* v4.09.1 (Benson, 1999) within the *pyTanFinder* pipeline (Kirov et al., 2018). We then used *RepeatMasker* v4.0.5 (http://repeatmasker.org) to perform repeat annotation and masking of the identified repeat elements. Structural gene annotation was performed using the RNA-Seq pipeline in *Braker2* v2.1.6 (Brůna et al., 2021), which manages the training of the gene prediction tools *GeneMark-ET* (Lomsadze et al., 2014) and *AUGUSTUS* (Stanke et al., 2008). In brief, we used seven RNA-Seq libraries from NCBI's SRA (Accessions SRX565152-57, SRX20881937), totalling 27.8 Gb of Illumina short reads, and mapped them to the soft-masked reference genome with *HISAT2* v2.0.4 (Kim et al., 2019). RNA-Seq alignment information was used to iteratively train *GeneMark-ET* to generate initial gene structures, which were passed to *AUGUSTUS* along with RNA-Seq mapping information to generate final gene predictions. Functional annotation of the predicted gene coding sequences was finally conducted using *eggNOG-mapper* v2 (Cantalapiedra et al., 2021), which relies on precomputed Orthologous Groups (OGs) to transfer functional information using phylogenetically refined orthology assignments.

Library preparation and whole-genome resequencing

Whole-genome sequencing (WGS) libraries were prepared for 86 samples using the Ovation Ultralow System V2 library preparation kit (NuGEN/Tecan) from 100 ng DNA input (when possible) following the manufacturer's instructions. We used unique dual indexing to minimise the effect of index-hopping and PCR cycles were adapted to the amount of input DNA (9-20 cycles). Libraries were pooled in equimolar ratio and sequenced to different coverage depths on a single S4 flow cell on a NovaSeq6000 instrument (Illumina) to generate 150 bp paired-end reads.

Sequence processing and alignment

To complement our dataset, WGS data for an additional 26 seahorse samples were obtained from Barry et al. (2022) (TruSeq DNA PCR-free libraries sequenced on Illumina NovaSeq6000, GenBank Sequence Read Archive, accession BioProject ID PRJNA777424). Therefore, our final combined dataset consisted of 112 samples (hereafter referred to as the "full dataset", Supplementary Table S1). Raw demultiplexed reads were processed using fastp (v0.23.1) (Chen et al., 2018) with the "*--merge*" option, in order to stitch together paired-end reads with overlapping sections. Both merged and unmerged reads were aligned to our reference genome using BWA-MEM (BWA v0.7.17; (H. Li, 2013). Picard (v2.26.8) («Picard toolkit», 2019, Broad Institute) was used for sorting read alignments, marking duplicates and adding read groups. DNA damage patterns in older dried and museum samples were visualised using PMDtools (v0.60) (Skoglund et al., 2014) (Supplementary Figure S2).

Variant calling of medium- to high-coverage samples

Forty-eight samples with sufficient coverage (~5-50X) were selected for variant calling in such a way that this subset contained five individuals in certain key locations (from now on referred to as the "GATK dataset", Supplementary Table S1, indicated by variant calling = Yes). Variants were called using the GATK best practices workflow (McKenna et al., 2010; Van der Auwera et al., 2013) without performing Variant and Base Quality Score Recalibration (VQSR and BQSR). Firstly, individual GVCF files were created from bam files with HaplotypeCaller (GATK v.4.1.8.0). This information was then stored in a GVCF database using GenomicsDBImport, and VCF files (one file per scaffold) were generated with GenotypeGVCFs. After concatenation, the resulting VCF was filtered for indels, multiallelic SNPs and missing data ("*--max-missing 0.9*"). Our detailed workflow and commands used are provided in Supplementary File S2. An Rmarkdown report produced with flex dashboard (Sievert et al., 2022), showing detailed results across various bioinformatic steps and analyses is provided in the Supplementary Appendix (Meyer_et_al_Results_Appendix.html).

Population structure

A Principal component analysis (PCA) was carried out on samples with sufficient coverage (>3X) from the full dataset without calling genotypes. To accomplish this, genotype likelihoods were calculated using ANGSD (v0.933) (Korneliussen et al., 2014) with the GATK model ("*-GL 2*") and the following parameters: "*-doMajorMinor 1 -minMapQ 30 -minQ 20 -doMaf 1 -doCounts 1 -minMaf 0.05 -uniqueonly 1 -remove_bads 1 -C 50 -baq 1 -doCov 1 -doIBS 2 -makeMatrix 1 -ref referencegenome_Hgutt_V1.fa*". In addition, the first 5 bp were trimmed off of reads (*-trim 5*) to account for DNA damage patterns due to cytosine deamination in historical samples. The analysis was restricted to sites with a strong probability of being SNPs ("*-SNP_pval 1e-6*"). Finally, sites were not considered if the total depth across samples was greater than twice the sum of mean coverage per sample ("*-setMaxDepth*"), in order to avoid artefacts in problematic regions that received unexpectedly high coverage (low-complexity and duplicated regions), or if less than half of the individuals had data ("*-minInd*"). The same filters were used in subsequent analyses conducted with ANGSD, unless specified otherwise. To perform a PCA while minimising the effect of linkage disequilibrium, we used ANGSD with the "*-sites*" option and provided a file containing one randomly selected SNP per 10 kb window in every 500 kb interval. To complement this genotype likelihood-based approach, a PCA was also performed on called genotypes from the GATK dataset using the R package SNPRelate (v1.28.0) (Zheng et al., 2012). Lastly, we conducted local PCA on the GATK dataset to capture variation in population structure along the genome in order to identify outlying patterns indicating the presence of putative chromosomal inversions. To this end, local PCA was performed in non-overlapping windows of 5 kb using the R package lostruct (v0.0.0.9, Li & Ralph, 2019).

Analysis of large structural variants

Genomic regions that were identified in local PCA as being candidates for the presence of large SVs (i.e. three groups that persist over many consecutive windows) were more specifically tested for evidence of chromosomal inversions in linked-read sequencing data. Firstly, we used abrupt signal shifts flanking outlying regions in the local PCA to determine the location of putative inversion breakpoints. We performed manual curation of the reference genome assembly in those regions following (Rhie et al., 2021), using the alignment to the genome assembly of *H. erectus* to confirm the suspected breakpoints. We then used *MTG-Link* v2.4.1 (Guichard et al., 2023) to perform local re-assembly of the linked-reads mapping in the regions surrounding breakpoints. The 10X linked-reads data were preprocessed with *EMA* v0.6.2 (Shajii et al., 2018) and *LRez* v2.2.4 (Morisse et al., 2021b) before running *MTG-Link* with default settings. The locally reassembled contigs were finally aligned against the Hgutt_V1 reference genome

with *Minimap2* (H. Li, 2018) to check for local consistency between assemblies and identify possible mis-joins. We characterised repeat content around the inversion breakpoints, looking specifically for the presence of recombinogenic sequences such as long inverted repeats (LIRs). Finally, we used *Leviathan* v1.0.2 (Morisse et al., 2021a) for calling SVs using linked-read data information connecting distant regions that share a higher number of barcodes than expected based on their distance. We ran *Leviathan* on the 10X linked-read data from the Hgutt_V1 assembly mapped to the HguttRefA genome.

In addition, we followed the haplotagging library construction protocol described in Meier et al. (2021) to generate complementary linked-read sequencing data for alternate genotypes at the two suspected inversions. Namely, we constructed haplotagging libraries for a DD/AA Mediterranean and a CC/AB Atlantic individual, and sequenced them with 2*50 bp paired-end reads to an average coverage depth of 12X.

Finally, we performed a functional enrichment analysis of the gene sets contained in each inversion using ShinyGO v0.77 (Ge et al., 2020) with a false discovery rate (FDR) threshold of 0.05.

Genomic landscape of divergence and introgression

We characterised the genomic landscape of divergence between localities and habitats using the GATK dataset. Genetic differentiation (*F*ST), nucleotide diversity (π) and absolute genetic divergence (*d*XY) were calculated in 25 kb windows using the popgenWindows.py script (Martin, 2018; https://github.com/simonhmartin/genomics_general). We then sought to characterise two highly divergent regions on Chr2 and Chr12, which represented suspected inversions. We used BAMscorer v1.4 (Ferrari et al., 2022) to assign inversion genotypes in all samples (including low-coverage samples). This firstly consisted of classifying all samples in the GATK dataset with regards to their haplotypes, as based on PCA groupings and individual heterozygosity values calculated with VCFtools (v0.1.16) (Danecek et al., 2011). This reference database was then used to score alignments and to ascertain allelic state and haplotype for all samples.

In order to test for introgression with other seahorse species and to determine evolutionary relationships across the genome, we obtained whole-genome resequencing data for *H. erectus*, *H. hippocampus*, *H. zosterae*, *H. capensis* and *H. comes* from Li et al. (2021) (NCBI BioProjects accession code PRJNA612146), including data for one individual from each species. Reads were processed and aligned to the Hgutt_V1 reference genome as described above. ANGSD was used to call genotypes and to produce a VCF ("*-doBcf*") containing seven individuals (two *H. guttulatus* that were homozygous for the four alternate inversion haplotypes, i.e. AA-DD and BB-CC, and five individuals from other species). This VCF was filtered to include only fixed sites (no heterozygous genotypes) and was used as input for performing topology weighting using the Twisst pipeline (Martin & Van Belleghem, 2017). This method performs iterative sampling of subtrees and quantifies the proportion of trees that match a certain taxon topology, allowing to study complex evolutionary relationships along the genome (e.g. patterns of introgression). We generated unrooted phylogenies using PhyML (Guindon et al., 2010) as implemented in the script *phyml_sliding_windows.py* (https://github.com/simonhmartin/twisst) for windows of 50 SNPs, setting a minimum of 25 non-missing genotypes per individual per window ("*--minPerInd 25"*) and performed topology weighting using the *twisst.py* script. Additionally, we produced a genome-wide consensus tree, as well as consensus trees for the two inversions using the averageTree function from phytools (Revell, 2012).

Inferring inversion history from the Ancestral Recombination Graph

We used tsinfer (v0.3.0) (Kelleher et al., 2019) to describe the diversity of genetic relationships between sequenced individuals along the genome. This method takes ancestral recombination events into account to infer the genome-wide sequence of correlated local genealogies, called the Ancestral Recombination Graph (ARG). Given accurate inference, the ARG provides a complete description of the available information on the evolutionary history of a set of related sequences, represented in terms of recombination and coalescence events. Inferring the ARG with tsinfer requires an input VCF file for a set of phased, diploid genomes, and the ancestral state of each mutation present in the INFO field. We carried out statistical phasing and missing data imputation of per-chromosome VCF files using SHAPEIT (v4.2.2) (Delaneau et al., 2019), assuming constant recombination rate of 1 cM per Mb (Stapley et al., 2017) and an effective population size (*Ne*) estimated based on genome-wide diversity values (Charlesworth, 2009). To infer ancestral states, we used BLAST searches to determine the allelic state of each SNP position in each of three outgroup species (*H. erectus, H. kuda* [Accession GCA_901007745.1]*, H. comes* [Accession GCA_001891065.2]) using the BLAST+ package (v2.2.28) (Camacho et al., 2009). Flanking regions of 100 bp on either side of each SNP were blasted to the reference genome of each outgroup using *blastn*. The top hit was inspected to determine the state of the homologous position in each outgroup using a program which was developed for this purpose (https://gitlab.mbb.cnrs.fr/ibonnici/snom). Thereafter, est-sfs (v2.04) (Keightley & Jackson, 2018) was run using the Kimura 2-parameter model to infer ancestral state probabilities and phased VCFs were annotated with the most likely ancestral variant using BCFtools (v1.9) (Danecek et al., 2021).

We reconstructed the ARG by running tsinfer on our phased, oriented VCF containing all chromosomes. Sample objects were created using the CYVCF2 library (v0.30.18; Pedersen & Quinlan, 2017) as in the tskit tutorial (https://tskit.dev/tsinfer/docs/stable/tutorial.html). The tree sequence was then inferred using a constant recombination rate of 1e-8 per bp and a mismatch ratio of 1. The ages of the ancestral nodes in the inferred trees were estimated using tsdate (v0.1.5) (Wohns et al., 2022) using parameters "*Ne=200000*", "*timepoints=100*" and "*mutation_rate=1e-8*". The *Ne* value used was calculated from the nucleotide diversity of the most divergent region of the genome (i.e. Chr12 inversion). This choice did not affect the distribution of inferred coalescent times in the genome background, while enabling tsdate to infer ancient node ages within inversions.

The inferred ARG was used for two purposes. First, we described variation in coalescence times across the genome using time to the most recent common ancestor (TMRCA) for randomly selected combinations of sample haplotypes (i.e. tree leaves) extracted from genealogies along the tree sequence. Secondly, we characterised local ancestry for each individual haplotype within inversion regions. The signal of an inversion was shown to appear in the ARG as clades of samples that persist over a more extended region of the genome than would otherwise be expected (Ignatieva et al., 2023). Here, we assumed that, within an inversion, the two branches immediately below the oldest node of each local tree represent the two clades of non-inverted and inverted extended haplotypes. To assign local ancestry to each branch, a clade was considered to belong to non-inverted or inverted ancestry if it satisfied one of two empirically established conditions: (*i*) the clade contained more than 75% of the haplotype copies from all of the homokaryotypes of a given haplotype, and reciprocally for the alternate haplotype in the other clade; (*ii*) the clade contained 100% of the copies from all of the homokaryotypes of that haplotype, and the same clade contained less than 75% of the copies from the alternate haplotype. Local ancestry switches along individual haplotypes were used to identify inter-haplotype recombination events within inversions, indicative of gene conversion or introgression.

Results

Reference genome

We obtained a near chromosome-level genome assembly of the *Hippocampus guttulatus* Hgutt_V1 reference genome. The assembly contained 3,878 scaffolds spanning 451 Mb (424 Mb in scaffolds of at least 10 kb, longest scaffold 28.8 Mb, scaffold N50 = 18.1 Mb, N90 = 5.65 Mb, Supplementary Fig. S3), which is close to the genome size of 424 Mb predicted by GenomeScope (Barry et al., 2022). Genome completeness was very high (Busco C: 95.6% [S: 94.4%, D: 1.2%], F: 1.6%, M: 2.8%, n: 3640) and comparable to the HguttRefA and the *H. erectus* assemblies (Supplementary Fig. S4a and S4b). All scaffolds showed strong homology and conserved synteny relationships with the chromosome-level genome assembly of *H. erectus* (Supplementary Fig. S5), suggesting that Hgutt_V1 is a high-quality assembly. Scaffolds anchoring to the *H. erectus* chromosome assembly produced 22 pseudomolecules that were used to characterise the genomic landscape of divergence (Fig. 2).

The repeat landscape showed that interspersed repeats add up to 33% of the genome, with 13% being occupied by unknown repeats (Supplementary Fig. S6). Among these unknown repeats, the clustering analysis of similar tandem repeats revealed a highly abundant TR of approximately 500 bp in length, present in more than 5,500 copies across the genome, with an accumulated abundance of 2.7 Mb. The consensus monomer of this TR, hereafter called Hgutt_Tan9, is a short sequence of 37 bp present in all chromosomes (Supplementary Fig. S7). Structural gene annotation based on RNA-Seq data predicted a total of 25,770 coding gene sequences, of which 13,346 unique gene names were identified by functional annotation.

Overall genetic structure

We produced genome resequencing data of heterogeneous quality (ranging from 1X to ~50X, Supplementary Table S1) and therefore characterised overall genetic structure using a genotype likelihood-based approach after controlling for low-coverage individuals (<3X) (Fig. 1B and 1C). We found evidence for pronounced population structure, even between certain sampling locations that were in close geographic proximity (inset map, Fig. 1A). This genetic structure could largely be ascribed to markers that are in tight linkage disequilibrium (LD), since PCA performed on unlinked markers revealed a different pattern (Fig. 1C). Here, all Mediterranean marine samples grouped together with Black Sea samples, separately from the different Mediterranean lagoon populations (Bz, Th & Mu). In the Atlantic, we also discern a northern (Br & Ga) and a southern (Fa) cluster of samples. By contrast, when all markers are considered (Fig. 1B), samples do not group together based on geographical origin. Instead, we observe that the individuals from a given location are grouped into either 1, 2 or 3 clusters that are organised along two orthogonal axes. The clusters from different locations were sometimes shifted away from each other (e.g. the three groups of Bs and Ma mirror those of Br and Ga), although they likely captured the same underlying variation. Observing these replicated three-cluster patterns in the PCA is consistent with the presence of large chromosomal inversions, with the three groups representing different inversion genotypes - homokaryotypes with two inverted alleles, heterokaryotypes carrying both haplotypes, and non-inverted homokaryotypes. The segregation of samples along two orthogonal axes, each showing replicated three-cluster patterns, is thus indicative of two polymorphic inversions containing large numbers of sites in strong LD.

In line with this finding, we observed heterogeneous landscapes of differentiation along the genome (Fig. 2). Background genomic differentiation between northern and southern Atlantic lineages ranged between 0.06 and 0.15 (first and third quartiles of F_{ST} distribution) (Fig. 2A). Differentiation was weaker between

Mediterranean marine and lagoon locations, with *F*ST values ranging between 0 and 0.03 (Fig. 2B). For other similar contrasts (i.e. northern vs. southern Atlantic lineage, or Mediterranean marine vs. lagoon comparison), the values were not substantially affected by the choice of the specific locations being compared (Supplementary Fig. S8, S9 & S10). These patterns of relatively weak genome-wide differentiation contrasted with high *F*ST and *d*XY values found on two chromosomes - Chr2 (Fig. 2C) and Chr12 (Fig. 2D). Chr12 presented an 8.2 Mb long plateau of high *F*ST and *d*XY values in both Atlantic and Mediterranean comparisons, a pattern that has previously been associated with inversions segregating at different frequencies between populations. For Chr2, a differentiation plateau was only observed in the Mediterranean marine vs. lagoon comparison. This pattern was also less clear, as the highest values did not occur in one contiguous block, but were split across scaffolds and even showed discontinuities within scaffolds.

Inversions differentiate ecotypes in the Mediterranean Sea

The presence of an inversion for the block of high differentiation on Chr12 was first supported by the alignment of Hgutt_V1 against HguttRefA (generated from the opposite homokaryotype, (Supplementary Fig. S11), as well as with the *H. erectus* assembly. This analysis showed the presence of an 8.2 Mb long inverted segment between the two *H. guttulatus* assemblies, with a putative inversion breakpoint located near position 1.699 Mb on scaffold 14 of Hgutt_V1, which corresponds to the abrupt signal shift in divergence at the beginning of the block. Linked-read based local reassembly of a 3 kb region centred on the putative breakpoint confirmed the contiguity of the Hgutt_V1 reference. In the middle of this 3 kb region, we found a long inverted repeat (LIR) consisting of two inverted arrays of the Tan9 monomer, with the internal spacer overlapping the inversion breakpoint (Supplementary Fig. S12). The other end of the inversion at the end of scaffold 14 (near 9.877 Mb) also contained a tandem repeat of Tan9 monomers (Supplementary Fig. S6). Linked-read sequencing data obtained from three individuals showing the three possible genotypes (AA, AB and BB) showed mapping patterns that were consistent with the presence of a large chromosomal inversion (Supplementary Fig. S12). Finally, the analysis of linked-reads from a BB genotype mapped to the HguttRefA assembly allowed the direct detection of a 8.2 Mb inversion at the breakpoints expected from all previous analyses.

Although we were not able to perform the same detailed analysis for Chr2, since the two *H. guttulatus* assemblies both carry the same haplotype, other patterns suggested that a second inversion is highly likely to be present on this chromosome. Performing PCA separately on Chr2 and Chr12, we observed three clusters along PC1 with the middle group presenting higher heterozygosity than the outer groups (Fig. 3A & 3B). Using BAMscorer (Ferrari et al., 2022), we characterised inversion genotypes in all samples except for one extremely low-quality individual (Hgutt_Se_1898_94, see Supplementary Appendix). These scores were consistent with the groupings observed from the PCA (Fig. 3A & 3B) - that is, samples that were classified as heterokaryotypes were in the middle PCA group, and homokaryotypic samples were in the outer groups. These patterns confirm that the multiple blocks of high *F*ST on Chr2 are in perfect LD, and that they collectively segregate as a single contiguous variant, despite the discontinuous signature observed in the divergence landscape (Fig. 2C). Consequently, our results indicate that there are two polymorphic inversions segregating in seahorse populations, and that these play an important role in the differentiation between lineages and ecotypes. As for their gene content, the 8.2 Mb inverted segment on Chr12 contained a total of 409 annotated genes showing a significant 3.7-fold enrichment for the estrogen signalling pathway (Enrichment FDR=0.042, 11 of 134 annotated genes in the pathway). The high-LD region on Chr2 had a total of 419 annotated genes and showed a significant 2.5-fold enrichment for the neuroactive ligand-receptor interaction pathway (Enrichment FDR=0.047, 20 of 372 annotated genes in the pathway).

For each inversion, we characterised the relative frequencies of the three alternative genotypes across

sampling locations and habitats (Fig. 3C & 3F). We found that the inversion on Chr12 (B12) was present in one of the two homozygous states (either AA or BB) in almost all locations (Fig. 3F). Atlantic samples from the northern part of Biscay (Br), the Gulf of Cadiz (Fa) and samples from Mediterranean lagoons (Th, Mu & Bz) exclusively presented homokaryotypes for the A allele, while Mediterranean marine sites only presented homokaryotypes for the B allele. Fine-scale variation in inversion genotypes was especially pronounced in the Mediterranean Sea, since samples that were collected only a few kilometres apart showed differential fixation for B12 between marine and lagoon habitats (inset map, Fig. 3F). Samples for which precise location information was not available (grey circles) were either AA or BB homokaryotypes, with A alleles probably hailing from lagoon habitats in these locations and B alleles from the sea. Samples of unknown origin ("Xx") mostly carried the lagoon haplotype, probably reflecting frequent incidental catches in lagoon habitats during artisanal and non-professional fishing activities. There were only two sampling locations which were polymorphic for B12: the Hossegor marine lake in the Bay of Biscay (Ga) and Varna in the Black Sea (Bs), which presented samples from each of the three inversion genotypes (AA, AB and BB). The inversion located on Chr2 (B2) showed different distribution patterns compared to that of B12, since this polymorphism was only found in Mediterranean lagoons (Fig. 3C). Both in the Atlantic and in the Black Sea, all samples were fixed for haplotype C. The alternative D haplotype was found at varying frequencies (28% to 70%) in Thau, Bizerte and Murcia lagoons. All samples caught in the Bizerte lagoon were CD heterokaryotypes (n=5), while 3 out of 5 samples from Murcia were DD homokaryotypes. All samples carrying one or two D alleles had been classified as an AA homokaryotype for B12 (including those from unknown habitats), indicating that the Chr2 polymorphism is private to Mediterranean lagoon populations.

Our results thus indicate that two large chromosomal inversions covering roughly 9% of the seahorse genome largely drive ecotype differentiation in the Mediterranean Sea. Lagoon populations are fixed at B12 for the A haplotype, whereas populations in the sea only carry the alternate B haplotype. In addition, there is a second inversion on Chr2 (with haplotypes C or D) which differentiates marine and lagoon ecotypes based on a polymorphism which is private to the lagoon ecotype. Even though the B12 inversion is also polymorphic in the Atlantic, it does not show differential fixation between habitat types in this zone. Instead, Atlantic populations to the north of Hossegor carry the A allele, while both A and B alleles are present from Hossegor to the south of Portugal (see also Riquet et al., 2019). Interestingly, the B haplotype on B12 and the D haplotype on B2 were never found together in the same individual (Fig. 3G), suggesting a possible negative interaction between the two alleles. In what follows, we sought to further characterise the respective histories of the two inversions, including their origins and ages.

Evolutionary history of the two inversions

An important question regarding the origin of the inversions is whether the heterogeneous divergence landscape has resulted from the differential erosion of past genome-wide divergence, or alternatively, from the emergence of newly divergent regions (i.e. secondary contact *versus* primary divergence). To address this question, we studied the distribution of coalescence times along the genome as inferred by tsdate (v0.1.5). Outside the inversion regions, we did not find a strong signature of ancient coalescence in the form of deep time to the most recent common ancestor (TMRCA) (Fig. 4E). Only the two chromosomes carrying inversions showed the presence of TMRCA values that were up to 5 times older than in the genome background, which also corresponded to their high *d*XY values (Fig. 2). What is more, except for the inversions, regions of high *F*ST were not shared between Atlantic lineages and Mediterranean ecotype contrasts, suggesting a lack of parallelism in the genome background (Supplementary Fig. S13).

Topology weighting conducted using Twisst (Martin & Van Belleghem, 2017) revealed that the most predominant topologies – both in the genome background as well as in the inversions – were consistent with the currently accepted genome-wide phylogeny of the genus (Li et al., 2021; Stiller et al., 2022). However, they simultaneously showed high levels of incomplete lineage sorting (ILS) near the ancestral nodes of the genus. If one of the two haplotypes on either Chr2 or Chr12 had been introduced through introgression from a closely related species, we would have expected the tree topologies in this region to differ from genome-wide topologies. In this case, an introgressed haplotype carried by a *H. guttulatus* individual would have grouped closer to the donor species than with the alternate haplotype of *H. guttulatus*. The topologies that did not group alternate haplotypes together within *H. guttulatus* amounted to only 4.57% and 15.7% of all topologies that were observed in the B2 and B12 regions, respectively (see Fig. 4C & 4D for the five most frequent of these topologies). Even when the two *H. guttulatus* haplotypes were not in the same grouping, they were generally not placed in distant positions (i.e. they still occurred in the same sub-branch). These topologies did not show any particular haplotype tending to group with a potential donor species, and both haplotypes showed shifted positions, most probably due to ILS. Given these findings, we conclude that the two inversion polymorphisms in *H. guttulatus* were not likely introduced through introgression with a closely related species.

Our results indicate that the B2 and B12 polymorphisms are most likely explained by chromosomal inversion events that took place within the *H. guttulatus* lineage. These inversions have been maintained as intraspecific polymorphisms for a long enough period of time for divergence to have accumulated between haplotypes. The divergence between A and B haplotypes on Chr12 was particularly high, since *dXY* values for comparisons between opposite homokaryotypes ranged between 0.69 and 0.76% (first and third quartiles). Absolute divergence between opposite B2 homokaryotypes was slightly lower and showed more variance (first to third quartile, 0.36 to 0.57%) (Supplementary Fig. S9). TMRCA inferred in the B12 region were also higher than for the B2 region, and consensus phylogenies constructed with PhyML showed a deeper split for haplotypes A and B (Fig. 4C) than for C and D (Fig. 4B). We conclude that the inversions probably do not have the same age, and that the B12 polymorphism emerged before B2. Furthermore, it should be noted that divergence might have been underestimated in our study due to reference bias, since reads from AA individuals were mapped to our B reference (Hgutt_V1).

In addition to these inversions, other intra- and inter-chromosomal rearrangements were evidenced in alignments between our reference assembly (Hgutt_V1), HguttRefA and the *H. erectus* assembly. Due to the abundance of these structural rearrangements, we were not able to determine the ancestral arrangement for inversions B2 and B12 through comparison with the *H. erectus* assembly. These analyses showed that *H. erectus* chromosomes Chr11 and Chr12 were fused in the HguttRefA assembly to form chromosome JAOYMQ010000004.1 (Supplementary Fig. S11). This could potentially indicate that the A haplotype (carried by HguttRefA) on Chr12 is involved in a chromosomal fusion with Chr11 (contrary to the B haplotype), or alternatively, that this might represent a technical artefact of the HguttRefA assembly, which is based on short-read data. As for B2, we speculate that the discontinuous signal of high *F*ST associated with this inversion (Fig. 2C) might be due to additional rearrangements that have altered the collinearity between the C and D haplotypes. The D haplotype might present a different intra-chromosomal structure to the C haplotype, resulting in a heterogeneous divergence landscape when mapped to the C reference. However, our analyses did not allow us to confirm this hypothesis, since both *H. guttulatus* assemblies were CC homokaryotypes. Alternatively, regions of low *F*ST and *d*XY could be due to local erosion of divergence through recombination between inverted haplotypes.

We tested for gene flux between inverted and non-inverted haplotypes, which can be expected in the form of double crossover events or gene conversion. Plotting PC1 coordinates from local PCA along inversion regions allowed us to locate genome windows which were candidates for gene flux. This was based on local deviations from the three-cluster pattern typical of inversions (three groups of horizontal lines representing the three genotypes, Fig. 5A and 5D). B12 showed the classical three-cluster pattern at

a large scale, whereas B2 showed a more “choppy” landscape, with long stretches of “irregular” patterns separating three-cluster windows. For both inversions, these deviated windows were further examined for evidence of local gene flux. We used inferred ARG trees to determine local ancestry along both inversions (Fig. 5B, 5C & 5E). We determined which trees showed the expected topology of reciprocal monophyly between non-inverted and inverted haplotypes grouping in two different clades (Fig. 5B), and which trees showed a discordant topology (Fig. 5C). Local trees grouping opposite haplotypes in the same clade indicated regions which were locally introgressed with a tract from the alternate haplotype. This allowed us to perform chromosome painting for each non-recombinant block associated with a particular tree. This approach revealed large-scale haplotypes corresponding to inverted and non-inverted ancestry, interspersed with local ancestry variation at a finer scale. Chromosome painting strongly reflected the patterns observed in the local PCA and *F*ST landscapes (see Supplementary Fig. S14, S15a and S15b for painted chromosomes of the entire GATK dataset). For lower quality samples, we found evidence for phasing errors in the form of switches between maternal and paternal haplotypes, which were visible in heterokaryotic samples. These errors did not prevent us from locating introgressed tracts in homokaryotypes, which spanned up to ~100 kb for B12 and up to ~500 kb for B2.

Discussion

We present the first genome-wide study of intraspecific diversity in *H. guttulatus*, investigating multiple aspects of population structure related to geography, habitat type and genome architecture. We further characterise divergence between two previously described geographical lineages in the Atlantic and between marine-lagoon ecotypes in the Mediterranean Sea (Riquet et al., 2019). Our results reveal a sharp contrast between low levels of genome background differentiation and the substantial haplotype frequency differences observed at two megabase-scale chromosomal inversions differentiating Mediterranean ecotypes. We characterise the origin of these inversions and discuss the possible mechanisms responsible for their long-term maintenance. Lastly, we reveal patterns suggestive of gene flux between inverted alleles, and address how this process could shape the dynamics and evolutionary fate of seahorse inversions.

Evolutionary origin of seahorse inversions

Our results indicate that inversions B2 and B12 have been maintained as polymorphisms for hundreds of thousands of generations, that is, well beyond the mean coalescence time inferred within populations (<100k generations, Fig. 4E). Consistently, these regions also showed remarkably high levels of raw divergence compared to the rest of the genome (*d*XY up to 1%, *vs.* maximum of 0.25% elsewhere). Moreover, the levels of nucleotide diversity associated with each haplotype were, given their respective frequencies, consistent with genome-wide average values. It is therefore likely that sufficient time has elapsed since the appearance of the inverted haplotypes for them to have progressed towards mutation-drift equilibrium. It should however be noted that estimating coalescence times with *tsdate* is known to be hampered by technical limitations (e.g., underestimation of deeper coalescence times; Brandt et al., 2022) that prevented our ability to precisely determine and compare the ages of B2 and B12.

Our finding that the two seahorse inversions represent ancient polymorphisms thus raises the question of their origin. One way that a divergent inverted haplotype may be introduced is through hybridisation with a closely related species (e.g., Hsieh et al., 2019; Jay et al., 2018). We did not find evidence for introgressed ancestry at either of the inversions, since topology analysis did not support similarity between any inversion haplotypes and a potential donor species (Fig. 4 B-E). Instead, all major sources of genealogical conflict could be explained by ILS occurring at deep nodes in the *Hippocampus* genus phylogeny. Alternatively, the inversions could have resisted re-homogenisation by gene flow upon

secondary contact between two divergent lineages, due to recombination suppression (Lundberg et al., 2023; Rafajlović et al., 2021; Yeaman, 2013). However, we did not find any regions of deep coalescence (i.e. expected remnants of recombined divergent ancestries) comparable to that found in the inverted regions. It is thus most likely that the inversions have emerged within the *H. guttulatus* lineage and that they have remained polymorphic at the species level for a long period of time.

Molecular mechanisms facilitating the emergence of inversions have been studied in detail using long-read sequencing in deer mice, showing that inversion breakpoints tend to occur in centromeric and telomeric regions and that they are sometimes flanked by LIRs (Harringmeyer & Hoekstra, 2022). Here, we were able to directly demonstrate, using linked-read data, that the 8.2 Mb inversion B12 occurs near a chromosome extremity and that it has its breakpoints in an approximately 1 kb LIR bearing a Tan9 monomer. Tandem repeats containing Tan9 are widespread throughout the *H. guttulatus* genome and may therefore facilitate ectopic recombination, leading to an increased rate of formation of new structural variants.

Maintenance of polymorphisms B12 and B2

A chromosomal inversion might undergo a series of evolutionary changes from the time of its appearance to its loss or fixation (Faria et al. 2019a). We thus sought to distinguish between the mechanisms that might maintain the two inversion polymorphisms in the seahorse, and to provide explanations for their respective patterns. Our results support that the B12 polymorphism is most comparable in its characteristics to Type I inversion polymorphisms (i.e. divergent between populations), while B2 probably calls on the interaction of forces linked to both Type I and Type II (i.e. within-population variation) polymorphisms.

The B12 polymorphism corresponds to the genomic island identified by Riquet et al. (2019) (Supplementary Fig. S16) as the main region differentiating geographical lineages in the Atlantic and ecotypes in the Mediterranean Sea. B12 shows local fixation for a given haplotype within almost all populations (Fig. 3F), in accordance with the pattern expected for Type I polymorphisms (Faria et al. 2019a). In this scenario, within-population polymorphism is not maintained unless there is sufficient gene flow, but among-population variation is conferred by divergent selection on alternate haplotypes. Given the differential fixation of B12 haplotypes between adjacent lagoon (A haplotype) and marine (B haplotype) sites in the Mediterranean Sea, it is possible that the inversion is either underdominant, under negative epistasis, or affected by selection against locally disadvantageous haplotypes. The general lack of recombination between inverted haplotypes ensures that the genes in the A and B haplotypes are inherited together, collectively forming a barrier to gene flow. Over a long period of time, the accumulation of new mutations between inverted haplotypes leads to high levels of divergence, as is reflected by high *d*XY values observed between AA and BB homokaryotypes (Fig. 2B). Although there is little opportunity for gene flow between alternate haplotypes, recombination can proceed normally within every population that is fixed for a given haplotype, since crossing over formation is unimpeded in homokaryotypes. In this way, the accumulation of mutation load in inverted regions is not expected to be substantially higher than in the collinear genome (Berdan et al. 2021). Empirical examples are provided by inversions in deer mice that show local fixation and lack of mutation load (Harringmeyer & Hoekstra, 2022), as well as inversions in sunflower populations fixed for one arrangement having lower mutation load compared to polymorphic populations (Huang et al., 2022).

If inversion B12 is a Type I polymorphism under divergent selection, a strong association of each haplotype with either the lagoon or marine habitats would be expected across the entire range. However, a few seahorse samples from Atlantic (Hossegor, Bay of Biscay) and Black Sea (Varna) populations showed a breakdown of the strong association between B12 genotype and habitat type. In these

populations, both haplotypes A and B were present and heterokaryotypes were locally common (4 out of 11 and 3 out of 5 samples, respectively), as already observed by Riquet et al. (2019). A possible explanation for this surprising pattern involves a predominantly intrinsic barrier effect of B12, and possible epistatic interactions with selected loci present on other chromosomes, such as the B2 inversion (see below).

In contrast to B12, the B2 inversion is maintained as a polymorphism which is private to Mediterranean lagoons. Classical work that has addressed the long-term maintenance of inversion polymorphisms reported an important role of balancing selection in maintaining within-population variation (e.g. reviewed in Llaurens et al., 2017; Wellenreuther & Bernatchez, 2018). This results in Type II polymorphisms *sensu* Faria et al. (2019a), which could explain the distribution of the B2 polymorphism in the Mediterranean Sea. We observed intermediate haplotype frequencies and relatively high frequencies of CD genotypes in Mediterranean lagoons (Th: 12 out of 36; Mu: 3 out of 5; Bz: 5 out of 5), possibly pointing to a form of heterozygote advantage. The B2 inversion is nearly three times longer than B12, making it more likely to have captured multiple recessive deleterious mutations (especially due to reduced effective population size) that are inherited as a single block (Connallon & Olito, 2022). For inversions that suffer from high mutation load, theory predicts that polymorphism could be maintained by pseudo-overdominance (Berdan et al., 2021), which is also supported by empirical studies in insects (Jay et al., 2021; Yang et al., 2002). For these reasons, it is likely that compensation of the recessive mutation load in heterokaryotypes could contribute to the persistence of the B2 polymorphism in Mediterranean lagoons.

The simultaneous action of forces associated with both Type I and Type II polymorphisms has been shown to potentially give rise to a range of different inversion equilibrium frequencies (Faria et al., 2019a). For instance, selection for heterokaryotypes combined with local adaptation can result in polymorphism in one environment and fixation in the other. A similar commixture of forces could be at work to explain the fixation of B2 haplotype C in marine *H. guttulatus* populations, while polymorphism is maintained in Mediterranean lagoons. If the B2 polymorphism is only maintained by pseudo-overdominance, we could ask why the D haplotype is not present outside of Mediterranean lagoons. The answer may be connected to local adaptation alleles that are carried in the haplotypes of B2, or in epistatic association with B12. If D provides local advantage in lagoons, or if D/A combinations are favoured over C/A, it could help maintain polymorphism in these environments. On the other hand, D would be outcompeted elsewhere (e.g. in Mediterranean marine habitats) due to lack of selective advantage. Although current knowledge does not allow us to answer all questions regarding the maintenance of the B2 polymorphism, we suspect that it involves an interaction of forces which could include local selection and a form of balancing selection.

Long-term fate of inversion polymorphisms in the long-snouted seahorse

Theoretical work (reviewed in, e.g. Hoffmann & Rieseberg, 2008) and empirical studies (e.g. Huang et al., 2020; Lohse et al., 2015; Noor et al., 2001) have found that, under certain conditions, the presence of inversion polymorphisms may facilitate speciation. In fish, multiple empirical studies have highlighted the role of chromosomal inversions in local adaptation, ecotype formation and speciation (Berg et al., 2016; Cayuela et al., 2020; Jones et al., 2012; Le Moan et al., 2021; Matschiner et al., 2022; Pettersson et al., 2019; Tigano et al., 2021). Mediterranean ecotypes of long-snouted seahorse show differential fixation at inversion B12, which represents a Type I polymorphism that could facilitate speciation between marine and lagoon ecotypes. We found evidence for very low levels of gene flux taking place in this inversion, suggesting that gene exchange in heterokaryotypes is either rare or selected against. Since exchanged segments were generally longer than a few kilobases and occurred in the middle of the inversion, they most likely resulted from double crossovers rather than gene conversion. This contrasts to what was

found for cod inversions where gene conversion predominated (Matschiner et al., 2022), but is in line with theoretical expectations for long inversions (Navarro et al., 1997). The introgressed tracts found within B12 across the species range were generally small in size (up to 50 kb), suggesting that introgression took place a long time ago and that the remaining tracts represent recombined segments that have passed the filter of selection. This is consistent with the view that multiple selected mutations have accumulated within the inversion over the long term (e.g., DM incompatibilities; Navarro & Barton, 2003). A key question remains whether the barrier effect of B12 also reduces effective gene flow between lagoon and marine ecotypes at unlinked neutral loci, which may be facilitated by coupling with premating isolation mechanisms. Some level of genome-wide differentiation has been observed between Mediterranean ecotypes, thus indicating a significant reduction in effective gene flow at a small spatial scale relative to dispersal. However, the contribution of B12 to gene flow reduction is more uncertain in the Atlantic and the Black Sea. Although a weak association has been observed between B12 and habitat type along the Portuguese coasts, no associated genetic structure was detected in the genomic background (Riquet et al., 2019).

The B12 polymorphism could facilitate speciation between seahorse ecotypes over the long term, but the fate of the B2 polymorphism is less straightforward to predict. Since the frequencies of this inversion may result from a balance between local and balancing selection, it is unclear whether B2 will eventually undergo differential fixation between habitats, or universal fixation of one arrangement. In contrast to B12, we found ample evidence for the erosion of divergence in B2 through gene flux, as illustrated by the detection of many introgressed segments and “suspension bridge” patterns in the *F*ST landscape (Fig. 2C). The intensity of gene flux was heterogeneous along this chromosome, since we found evidence for relatively recent double crossover events (i.e. larger introgressed segments than in B12) as well as one region where low *F*ST values (comparable to the genome background) are bordered by high differentiation segments, indicating that divergence has been eroded over time. Gene flux is therefore likely to impact the fate of this inversion by purging some of its mutation load (Berdan et al., 2021) and weakening the pseudo-overdominance that could be maintaining polymorphism. Because the large-scale assembly of Chr2 is still only partly resolved, we cannot rule out the possibility that additional chromosomal rearrangements have affected its divergence landscape and evolutionary trajectory. For example, multiple inversions occurring in the same region might have extended or restructured the block of high LD, as is the case for adjacent (Jay et al., 2021) or nested inversions (Maggiolini et al., 2020).

The framework laid out by Faria et al. (2019a) does not specifically consider the dynamics of multiple inversions co-existing in the same system, but leaves such interactions as an outstanding question. The presence of other polymorphic inversions could indeed impact the establishment, maintenance, and long-term fate of a given inversion, for example through epistatic interactions or coupling. The two inversion polymorphisms segregating in *H. guttulatus* might present a case study for such potential interactions. Epistatic interactions may explain why there is a strong association between B12 and habitat type in the Mediterranean, while the absence of the B2 polymorphism in the Atlantic would only produce weak associations of B12 with habitat. Interestingly, the two molecular pathways that were enriched within the inversions have interconnected functions in reproduction. The estrogen signalling pathway enriched in B12 has been shown to play a role in seahorse sexual dimorphism, gonad development, and stimulation of parturition in pregnant seahorses (Qin et al., 2019; Whittington et al., 2015; H. Zhang et al., 2022). This involves the regulation of expression of the two estrogen receptor genes, *esr1* and *esr2*, which are both present on B12. In mice, estrogenic regulation by *esr1* induces expression changes in a population of neurons that mediate estrogen feedback mechanisms affecting the neuroactive ligand-receptor interaction pathway (Göcz et al., 2022), corresponding to the pathway that was enriched in B2. Functional incompatibilities among genes that interact in the regulation of these pathways may potentially disrupt the reproductive function in incompatible genotypes at the two inversions involved in seahorse ecotype differentiation. Future directions for research could address these aspects by focusing on polymorphic populations, attempting to characterise the fitnesses of different karyotypes in different environments, and

by quantifying gene expression and mutation load. Studying gene flux may also reveal more about the loci contained in each arrangement, as we should expect erosion of divergence in regions carrying selectively neutral or disadvantageous mutations, and maintenance in regions that are under divergent selection.

Acknowledgements

The data were partly produced and analysed with the support of the *GenSeq* genotyping and sequencing platform and the *MBB* Montpellier Bioinformatics Biodiversity platform, both being supported by ANR program ”Investissements d’avenir” (ANR-10-LABX-04-01). We thank Rémy Dernat and Khalid Belkhir for their assistance in data storage, management and processing, and Iago Bonnici for support on variant orientation. DNA extraction of historical samples was carried out in a clean lab at ISEM, Montpellier (Plateforme d’ADN dégradé, LabEx CeMEB). We thank the Montpellier GenomiX platform for constructing the reference genome's 10X Chromium library. We are grateful to the National Museum of Natural History (MNHN Paris, Agnès Dettaï) and colleagues who provided us with samples, as well as those who facilitated or participated in sampling: Jorge Palma and Rita Castilho (CCMAR, Portugal), Cristina Mena (Hippocampus association, Spain), Patrick Louisy (peau-bleue association, CPIE Bassin de Thau, France), Lucy Woodall (University of Oxford, UK), and citizens from the Sète region who provided dried seahorses samples. We thank Huixian Zhang of the Key Laboratory of Tropical Marine Bio-Resources and Ecology, South China Sea Institute of Oceanology, for providing the *Hippocampus erectus* reference genome. Finally, we thank Thomas Broquet for his inputs on the manuscript. This work was supported by the ANR grant CoGeDiv ANR-17-CE02-0006-01 to PAG, and by a Languedoc-Roussillon Region “Chercheur(se)s d'avenir” grant to NB (Connect7 project), with the support of the Occitanie Regional Council’s program «Key challenge BiodivOc». Mobility during the project was partly funded by an ESEB Godfrew Hewitt Mobility Award as well as a Laura Corrigan conservation grant.

References

Barry, P., Broquet, T., & Gagnaire, P.-A. (2022). Age-specific survivorship and fecundity shape genetic diversity in marine fishes. *Evolution letters*, *6*(1), 46-62.

Benson, G. (1999). Tandem repeats finder : A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573. https://doi.org/10.1093/nar/27.2.573

Berdan, E. L., Blanckaert, A., Butlin, R. K., & Bank, C. (2021). Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLOS Genetics*, *17*(3), e1009411. https://doi.org/10.1371/journal.pgen.1009411

Berg, P. R., Star, B., Pampoulie, C., Sodeland, M., Barth, J. M. I., Knutsen, H., Jakobsen, K. S., & Jentoft, S. (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep23246

Brandt, Y. C., D., Wei, X., Deng, Y., Vaughn, A. H., & Nielsen, R. (2022). Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics, 221*(1), iyac044. https://doi.org/10.1093/genetics/iyac044

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

Butlin, R. K. (2005). Recombination and speciation. *Molecular Ecology*, *14*(9), 2621-2635. https://doi.org/10.1111/j.1365-294X.2005.02617.x

Cabanettes, F., & Klopp, C. (2018). D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *6*, e4958. https://doi.org/10.7717/peerj.4958

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+ :

Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Campbell, M. A., Anderson, E. C., Garza, J. C., & Pearse, D. E. (2021). Polygenic Basis and the Role of Genome Duplication in Adaptation to Similar Selective Environments. *Journal of Heredity, 112*(7), 614-625. https://doi.org/10.1093/jhered/esab049

Campos, P. F., & Gilbert, T. M. P. (2012). DNA extraction from formalin-fixed material. *Methods in Molecular Biology (Clifton, N.J.)*, *840*, 81-85. https://doi.org/10.1007/978-1-61779-516-9_11

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, *38*(12), 5825-5829. https://doi.org/10.1093/molbev/msab293

Cayuela, H., Rougemont, Q., Laporte, M., Mérot, C., Normandeau, E., Dorant, Y., Tørresen, O. K., Hoff, S. N. K., Jentoft, S., Sirois, P., Castonguay, M., Jansen, T., Praebel, K., Clément, M., & Bernatchez, L. (2020). Shared ancestral polymorphisms and chromosomal rearrangements as potential drivers of local adaptation in a marine fish. *Molecular Ecology*, *29*(13), 2379-2398. https://doi.org/10.1111/mec.15499

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit – Interactive quality assessment of genome assemblies. *G3 Genes|Genomes|Genetics*, *10*(4), 1361-1374. https://doi.org/10.1534/g3.119.400908

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195-205. https://doi.org/10.1038/nrg2526

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Cheng, C., White, B. J., Kamdem, C., Mockaitis, K., Costantini, C., Hahn, M. W., & Besansky, N. J. (2012). Ecological genomics of *Anopheles gambiae* along a latitudinal cline: a population-resequencing approach. *Genetics, 190*(4), 1417-1432. https://doi.org/10.1534/genetics.111.137794

Connallon, T., & Olito, C. (2022). Natural selection and the distribution of chromosomal inversion lengths. *Molecular Ecology*, *31*(13), 3627-3641. https://doi.org/10.1111/mec.16091

Dai, H., & Guan, Y. (2020). The Nubeam reference-free approach to analyze metagenomic sequencing reads. *Genome Research, 30*(9), 1364-1375. https://doi.org/10.1101/gr.261750.120

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., Zagury, J.-F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-13225-y

Doyle, J. J., & Doyle, J. L. (Eds.). (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, *19*, 11-15.

Faria, R., Johannesson, K., Butlin, R. K., & Westram, A. M. (2019a). Evolving Inversions. *Trends in Ecology & Evolution*, *34*(3), 239-248. https://doi.org/10.1016/j.tree.2018.12.005

Faria, R., Chaube, P., Morales, H. E., Larsson, T., Lemmon, A. R., Lemmon, E. M., Rafajlović, M., Panova, M., Ravinet, M., Johannesson, K., Westram, A. M., & Butlin, R. K. (2019b). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. *Molecular Ecology*, *28*(6), 1375-1393. https://doi.org/10.1111/mec.14972

Faria, R., & Navarro, A. (2010). Chromosomal speciation revisited : Rearranging theory with pieces of evidence. *Trends in Ecology & Evolution*, *25*(11), 660-669. https://doi.org/10.1016/j.tree.2010.07.008

Ferrari, G., Atmore, L. M., Jentoft, S., Jakobsen, K. S., Makowiecki, D., Barrett, J. H., & Star, B. (2022). An accurate assignment test for extremely low-coverage whole-genome sequence data. *Molecular Ecology Resources*, *22*(4), 1330-1344. https://doi.org/10.1111/1755-0998.13551

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*, *117*(17), 9451-9457. https://doi.org/10.1073/pnas.1921046117

Funk, E. R., Mason, N. A., Pálsson, S., Albrecht, T., Johnson, J. A., & Taylor, S. A. (2021). A supergene underlies linked variation in color and morphology in a Holarctic songbird. *Nature Communications, 12*(1), Article 1. https://doi.org/10.1038/s41467-021-27173-z

Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics (Oxford, England)*, *36*(8), 2628-2629. https://doi.org/10.1093/bioinformatics/btz931

Göcz, B., Takács, S., Skrapits, K., Rumpler, É., Solymosi, N., Póliska, S., ... & Sárvári, M. (2022). Estrogen differentially regulates transcriptional landscapes of preoptic and arcuate kisspeptin neuron populations. *Frontiers in Endocrinology, 13,* 960769.

Gould, B. A., Chen, Y., & Lowry, D. B. (2017). Pooled ecotype sequencing reveals candidate genetic mechanisms for adaptive differentiation and reproductive isolation. *Molecular Ecology, 26*(1), 163-177. https://doi.org/10.1111/mec.13881

Guerrero, R. F., Rousset, F., & Kirkpatrick, M. (2012). Coalescent patterns for chromosomal inversions in divergent populations. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1587), 430-438. https://doi.org/10.1098/rstb.2011.0246

Guichard, A., Legeai, F., Tagu, D., & Lemaitre, C. (2023). MTG-Link: Leveraging barcode information from linked-reads to assemble specific loci. *BMC Bioinformatics*, *24*(284). https://doi.org/10.1186/s12859-023-05395-w

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307-321. https://doi.org/10.1093/sysbio/syq010

Hager, E. R., Harringmeyer, O. S., Wooldridge, T. B., Theingi, S., Gable, J. T., McFadden, S., Neugeboren, B., Turner, K. M., Jensen, J. D., & Hoekstra, H. E. (2022). A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. *Science, 377*(6604), 399-405. https://doi.org/10.1126/science.abg0718

Harringmeyer, O. S., & Hoekstra, H. E. (2022). Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. *Nature Ecology & Evolution*, 1-15. https://doi.org/10.1038/s41559-022-01890-0

Haÿ, V., Mennesson, M. I., Dettaï, A., Bonillo, C., Keith, P., & Lord, C. (2020). Needlepoint non-destructive internal tissue sampling for precious fish specimens. *Cybium: Revue Internationale d'Ichtyologie*, *44*(1), 73-79. https://doi.org/10.26028/cybium/2020-441-010

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetics Research*, *8*(3), 269-294. https://doi.org/10.1017/S0016672300010156

Hoffmann, A. A., & Rieseberg, L. H. (2008). Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annual review of ecology, evolution, and systematics*, *39*, 21-42. https://doi.org/10.1146/annurev.ecolsys.39.110707.173532

Hsieh, P., Vollger, M. R., Dang, V., Porubsky, D., Baker, C., Cantsilieris, S., Hoekzema, K., Lewis, A. P., Munson, K. M., Sorensen, M., Kronenberg, Z. N., Murali, S., Nelson, B. J., Chiatante, G., Maggiolini, F. A. M., Blanché, H., Underwood, J. G., Antonacci, F., Deleuze, J.-F., & Eichler, E. E. (2019). Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science*, *366*(6463), eaax2083. https://doi.org/10.1126/science.aax2083

Huang, K., Andrew, R. L., Owens, G. L., Ostevik, K. L., & Rieseberg, L. H. (2020). Multiple chromosomal inversions contribute to adaptive divergence of a dune sunflower ecotype. *Molecular Ecology*, *29*(14), 2535-2549. https://doi.org/10.1111/mec.15428

Huang, K., Ostevik, K. L., Elphinstone, C., Todesco, M., Bercovich, N., Owens, G. L., & Rieseberg, L. H. (2022). Mutation load in sunflower inversions is negatively correlated with inversion heterozygosity. *Molecular Biology and Evolution*, *39*(5), msac101. https://doi.org/10.1093/molbev/msac101

Ignatieva, A., Favero, M., Koskela, J., Sant, J., & Myers, S. R. (2023). The distribution of branch duration and detection of inversions in ancestral recombination graphs. *bioRxiv* doi: 10.1101/2023.07.11.548567

Jay, P., Whibley, A., Frézal, L., Rodríguez de Cara, M. Á., Nowell, R. W., Mallet, J., Dasmahapatra, K. K., & Joron, M. (2018). Supergene evolution triggered by the introgression of a chromosomal inversion. *Current Biology*, *28*(11), 1839-1845.e3. https://doi.org/10.1016/j.cub.2018.04.072

Jay, P., Chouteau, M., Whibley, A., Bastide, H., Parrinello, H., Llaurens, V., & Joron, M. (2021). Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. *Nature Genetics*, *53*(3), 288-293. https://doi.org/10.1038/s41588-020-00771-1

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), Article 7392. https://doi.org/10.1038/nature10944

Keightley, P. D., & Jackson, B. C. (2018). Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. *Genetics*, *209*(3), 897-906. https://doi.org/10.1534/genetics.118.301120

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., & McVean, G. (2019). Inferring whole-genome histories

in large population datasets. *Nature Genetics*, *51*(9), Article 9. https://doi.org/10.1038/s41588-019-0483-y

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907-915. https://doi.org/10.1038/s41587-019-0201-4

Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, *173*(1), 419-434. https://doi.org/10.1534/genetics.105.047985

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLOS Biology*, *8*(9), e1000501. https://doi.org/10.1371/journal.pbio.1000501

Kirov, I., Gilyok, M., Knyazev, A., & Fesenko, I. (2018). Pilot satellitome analysis of the model plant, *Physcomitrella patens*, revealed a transcribed and high-copy IGS related tandem repeat. *Comparative Cytogenetics*, *12*(4), 493-513. https://doi.org/10.3897/CompCytogen.v12i4.31015

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Korunes, K. L., & Noor, M. A. F. (2019). Pervasive gene conversion in chromosomal inversion heterozygotes. *Molecular Ecology, 28*(6), 1302-1315. https://doi.org/10.1111/mec.14921

Kulmuni, J., Butlin, R. K., Lucek, K., Savolainen, V., & Westram, A. M. (2020). Towards the completion of speciation: The evolution of reproductive isolation beyond the first barriers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1806), 20190528. https://doi.org/10.1098/rstb.2019.0528

Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoeppner, M. P., Kerje, S., Gustafson, U., Shi, C., Zhang, H., Chen, W., Liang, X., Huang, L., Wang, J., Liang, E., Wu, Q., Lee, S. M.-Y., Xu, X., Höglund, J., … Andersson, L. (2016). Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nature Genetics, 48*(1), Article 1. https://doi.org/10.1038/ng.3430

Le Moan, A., Bekkevold, D., & Hemmer-Hansen, J. (2021). Evolution at two time frames: Ancient structural variants involved in post-glacial divergence of the European plaice (*Pleuronectes platessa*). *Heredity*, *126*(4), Article 4. https://doi.org/10.1038/s41437-020-00389-3

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (arXiv:1303.3997). arXiv. http://arxiv.org/abs/1303.3997

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094-3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Ralph, P. (2019). Local PCA shows how the effect of population structure differs along the genome. *Genetics*, *211*(1), 289-304. https://doi.org/10.1534/genetics.118.301747

Li, C., Olave, M., Hou, Y., Qin, G., Schneider, R. F., Gao, Z., Tu, X., Wang, X., Qi, F., Nater, A., Kautt, A. F., Wan, S., Zhang, Y., Liu, Y., Zhang, H., Zhang, B., Zhang, H., Qu, M., Liu, S., … Lin, Q. (2021). Genome sequences reveal global dispersal routes and suggest convergent genetic adaptations in seahorse evolution. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-21379-x

Llaurens, V., Whibley, A., & Joron, M. (2017). Genetic architecture and balancing selection: The life and death of differentiated variants. *Molecular Ecology*, *26*(9), 2430-2448. https://doi.org/10.1111/mec.14051

Lohse, K., Clarke, M., Ritchie, M. G., & Etges, W. J. (2015). Genome-wide tests for introgression between cactophilic *Drosophila* implicate a role of inversions during speciation. *Evolution*, *69*(5), 1178-1190. https://doi.org/10.1111/evo.12650

Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, *42*(15), e119. https://doi.org/10.1093/nar/gku557

Lowry, D. B., & Willis, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLOS Biology*, *8*(9), e1000500. https://doi.org/10.1371/journal.pbio.1000500

Lundberg, M., Mackintosh, A., Petri, A., & Bensch, S. (2023). Inversions maintain differences between migratory phenotypes of a songbird. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-36167-y

Maggiolini, F. A. M., Sanders, A. D., Shew, C. J., Sulovari, A., Mao, Y., Puig, M., Catacchio, C. R., Dellino, M., Palmisano, D., Mercuri, L., Bitonto, M., Porubský, D., Cáceres, M., Eichler, E. E., Ventura, M., Dennis, M. Y., Korbel, J. O., & Antonacci, F. (2020). Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome Research*, *30*(11), 1680-1693. https://doi.org/10.1101/gr.265322.120

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO:Assessing genomic data quality and beyond. *Current Protocols*, *1*(12), e323. https://doi.org/10.1002/cpz1.323

Marion, S. B., & Noor, M. A. F. (2023). Interrogating the roles of mutation–selection balance, heterozygote advantage, and linked selection in maintaining recessive lethal variation in natural populations. *Annual Review of Animal Biosciences*, *11*(1), 77-91. https://doi.org/10.1146/annurev-animal-050422-092520

Martin, S. H., & Van Belleghem, S. M. (2017). Exploring evolutionary relationships across the genome using topology weighting. *Genetics*, *206*(1), 429-438. https://doi.org/10.1534/genetics.116.194720

Matschiner, M., Barth, J. M. I., Tørresen, O. K., Star, B., Baalsrud, H. T., Brieuc, M. S. O., Pampoulie, C., Bradbury, I., Jakobsen, K. S., & Jentoft, S. (2022). Supergene origin and maintenance in Atlantic cod. *Nature Ecology & Evolution, 6*(4), Article 4. https://doi.org/10.1038/s41559-022-01661-x

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, 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

Meier, J. I., Salazar, P. A., Kučka, M., Davies, R. W., Dréau, A., Aldás, I., Box Power, O., Nadeau, N. J., Bridle, J. R., Rolian, C., Barton, N. H., McMillan, W. O., Jiggins, C. D., & Chan, Y. F. (2021). Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(25), e2015005118. https://doi.org/10.1073/pnas.2015005118

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., Sebra, R., Peluso, P., Eid, J., Rank, D., Garcia, J. F., DeRisi, J. L., Smith, T., Tobias, C., Ross-Ibarra, J., Korf, I., & Chan, S. W. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, *14*(1), R10. https://doi.org/10.1186/gb-2013-14-1-r10

Mérot, C., Llaurens, V., Normandeau, E., Bernatchez, L., & Wellenreuther, M. (2020a). Balancing selection via life-history trade-offs maintains an inversion polymorphism in a seaweed fly. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-14479-7

Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020b). A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*, *35*(7), 561-572. https://doi.org/10.1016/j.tree.2020.03.002

Morisse, P., Legeai, F., & Lemaitre, C. (2021a). LEVIATHAN: Efficient discovery of large structural variants by leveraging long-range information from Linked-Reads data. *bioRxiv* doi: 10.1101/2021.03.25.437002

Morisse, P., Lemaitre, C., & Legeai, F. (2021b). LRez: A C++ API and toolkit for analyzing and managing Linked-Reads data. *Bioinformatics Advances*, *1*(1), vbab022. https://doi.org/10.1093/bioadv/vbab022

Navarro, A., Betrán, E., Barbadilla, A., & Ruiz, A. (1997). Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics*, *146*(2), 695-709. https://doi.org/10.1093/genetics/146.2.695

Navarro, A., & Barton, N. H. (2003). Accumulating postzygotic isolation genes in parapatry: A new twist on chromosomal speciation. *Evolution, 57*(3), 447-459. https://doi.org/10.1111/j.0014-3820.2003.tb01537.x

Nei, M., Kojima, K.-I., & Schaffer, H. E. (1967). Frequency changes of new inversions in populations under mutation-selection equilibria. *Genetics*, *57*(4), 741-750. https://doi.org/10.1093/genetics/57.4.741

Noor, M. A. F., Grams, K. L., Bertucci, L. A., & Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences*, *98*(21), 12084-12088. https://doi.org/10.1073/pnas.221274498

Ortiz-Barrientos, D., Engelstädter, J., & Rieseberg, L. H. (2016). Recombination rate evolution and the origin of species. *Trends in Ecology & Evolution*, *31*(3), 226-236. https://doi.org/10.1016/j.tree.2015.12.016

Pedersen, B. S., & Quinlan, A. R. (2017). cyvcf2 : Fast, flexible variant analysis with Python. *Bioinformatics*, *33*(12), 1867-1869. https://doi.org/10.1093/bioinformatics/btx057

Perez-Ruzafa, A., Marcos, C., Pérez-Ruzafa, I., & Pérez-Marcos, M. (2011). Coastal lagoons: "Transitional ecosystems" between transitional and coastal waters. *Journal of Coastal Conservation, 15,* 369-392. https://doi.org/10.1007/s11852-010-0095-2

Pettersson, M. E., Rochus, C. M., Han, F., Chen, J., Hill, J., Wallerman, O., Fan, G., Hong, X., Xu, Q., Zhang, H., Liu, S., Liu, X., Haggerty, L., Hunt, T., Martin, F. J., Flicek, P., Bunikis, I., Folkvord, A., & Andersson, L. (2019). A chromosome-level assembly of the Atlantic herring genome-detection of a supergene and other signals of selection. *Genome Research*, *29*(11), 1919-1928. https://doi.org/10.1101/gr.253435.119

Picard toolkit. (2019). In *Broad Institute, GitHub repository*. Broad Institute. https://broadinstitute.github.io/picard/

Qin, G., Luo, W., Tan, S., Zhang, B., Ma, S., & Lin, Q. (2019). Dimorphism of sex and gonad-development-related genes in male and female lined seahorse, Hippocampus erectus, based on transcriptome analyses. *Genomics*, *111*(3), 260-266. https://doi.org/10.1016/j.ygeno.2018.11.008

Rafajlović, M., Rambla, J., Feder, J. L., Navarro, A., & Faria, R. (2021). Inversions and genomic differentiation after secondary contact: When drift contributes to maintenance, not loss, of differentiation. *Evolution*, *75*(6), 1288-1303. https://doi.org/10.1111/evo.14223

Revell, L. J. (2012). Phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*(2), 217-223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), Article 7856. https://doi.org/10.1038/s41586-021-03451-0

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, *16*(7), 351-358. https://doi.org/10.1016/s0169-5347(01)02187-5

Riquet, F., Liautard- Haag, C., Woodall, L., Bouza, C., Louisy, P., Hamer, B., Otero- Ferrer, F., Aublanc, P., Béduneau, V., Briard, O., Ayari, T. E., Hochscheid, S., Belkhir, K., Arnaud- Haond, S., Gagnaire, P.-A., & Bierne, N. (2019). Parallel pattern of differentiation at a genomic island shared between clinal and mosaic hybrid zones in a complex of cryptic seahorse lineages. *Evolution*, *73*(4), 817-835. https://doi.org/10.1111/evo.13696

Schaeffer, S. W., & Anderson, W. W. (2005). Mechanisms of Genetic Exchange Within the Chromosomal Inversions of Drosophila pseudoobscura. *Genetics, 171*(4), 1729-1739. https://doi.org/10.1534/genetics.105.041947

Schwander, T., Libbrecht, R., & Keller, L. (2014). Supergenes and Complex Phenotypes. *Current Biology, 24*(7), R288-R294. https://doi.org/10.1016/j.cub.2014.01.056

Shajii, A., Numanagić, I., & Berger, B. (2018). Latent variable model for aligning barcoded short-reads improves downstream analyses. *Research in computational molecular biology : Annual International Conference, proceedings. 10812*, 280-282.

Sievert, C., Iannone, R., Allaire, J. J., & Borges, B. (2022). flexdashboard: R Markdown Format for Flexible Dashboards. https://pkgs.rstudio.com/flexdashboard/, https://github.com/rstudio/flexdashboard/.

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., & Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes : Patterns and processes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 372*(1736), 20160455. https://doi.org/10.1098/rstb.2016.0455

Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S., Krause, J., & Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, *111*(6), 2229-2234. https://doi.org/10.1073/pnas.1318934111

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics (Oxford, England)*, *24*(5), 637-644. https://doi.org/10.1093/bioinformatics/btn013

Stiller, J., Short, G., Hamilton, H., Saarman, N., Longo, S., Wainwright, P., Rouse, G. W., & Simison, W. B. (2022). Phylogenomic analysis of Syngnathidae reveals novel relationships, origins of endemic diversity and variable diversification rates. *BMC Biology*, *20*(1), 75. https://doi.org/10.1186/s12915-022-01271-w

Thompson, M. J., & Jiggins, C. D. (2014). Supergenes and their role in evolution. *Heredity*, *113*(1), Article 1. https://doi.org/10.1038/hdy.2014.20

Tigano, A., & Friesen, V. L. (2016). Genomics of local adaptation with gene flow. *Molecular Ecology*, *25*(10), 2144-2164. https://doi.org/10.1111/mec.13606

Tigano, A., Jacobs, A., Wilder, A. P., Nand, A., Zhan, Y., Dekker, J., & Therkildsen, N. O. (2021). Chromosome-level assembly of the Atlantic silverside genome reveals extreme levels of sequence diversity and structural genetic variation. *Genome Biology and Evolution*, *13*(6), evab098. https://doi.org/10.1093/gbe/evab098

Todesco, M., Owens, G. L., Bercovich, N., Légaré, J.-S., Soudi, S., Burge, D. O., Huang, K., Ostevik, K. L., Drummond, E. B. M., Imerovski, I., Lande, K., Pascual-Robles, M. A., Nanavati, M., Jahani, M., Cheung, W., Staton, S. E., Muños, S., Nielsen, R., Donovan, L. A., … Rieseberg, L. H. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, *584*(7822), Article 7822. https://doi.org/10.1038/s41586-020-2467-6

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From fastQ data to high-confidence variant calls : the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1), 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Wang, Y., & Leung, F. C. C. (2006). Long inverted repeats in eukaryotic genomes: Recombinogenic motifs determine genomic plasticity. *FEBS Letters*, *580*(5), 1277-1284. https://doi.org/10.1016/j.febslet.2006.01.045

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, *27*(5), 757-767. https://doi.org/10.1101/gr.214874.116

Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, *33*(6), 427-440. https://doi.org/10.1016/j.tree.2018.04.002

Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, *28*(6), 1203-1209. https://doi.org/10.1111/mec.15066

Westram, A. M., Faria, R., Johannesson, K., Butlin, R., & Barton, N. (2022a). Inversions and parallel evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1856). https://doi.org/10.1098/rstb.2021.0203

Westram, A. M., Stankowski, S., Surendranadh, P., & Barton, N. (2022b). What is reproductive isolation? *Journal of Evolutionary Biology*, *35*(9), 1143-1164. https://doi.org/10.1111/jeb.14005

Whittington, C. M., Griffith, O. W., Qi, W., Thompson, M. B., & Wilson, A. B. (2015). Seahorse Brood Pouch Transcriptome Reveals Common Genes Associated with Vertebrate Pregnancy. *Molecular Biology and Evolution*, *32*(12), 3114-3131. https://doi.org/10.1093/molbev/msv177

Wohns, A., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., & McVean, G. (2022). A unified genealogy of modern and ancient genomes. *Science*, 375. https://doi.org/10.1126/science.abi8264

Yang, Y.-Y., Lin, F.-J., & Chang, H. (2002). Comparison of recessive lethal accumulation in inversion-bearing and inversion-free chromosomes in *Drosophila*. *Zoological Studies*, *41*(3): 271-282.

Yeaman, S. (2013). Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(19), E1743-1751. https://doi.org/10.1073/pnas.1219381110

Yeaman, S., & Whitlock, M. C. (2011). The genetic architecture of adaptation under migration-selection balance. *Evolution*, *65*(7), 1897-1911. https://doi.org/10.1111/j.1558-5646.2011.01269.x

Zhang, H., Liu, Y., Qin, G., & Lin, Q. (2022). Identification of neurohypophysial hormones and the role of VT in the parturition of pregnant seahorses (*Hippocampus erectus*). *Frontiers in Endocrinology*, *13*. https://doi.org/10.3389/fendo.2022.923234

Zhang, L., Reifová, R., Halenková, Z., & Gompert, Z. (2021). How important are structural variants for speciation? *Genes*, *12*(7), 1084. https://doi.org/10.3390/genes12071084

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics (Oxford, England)*, *28*(24), 3326-3328. https://doi.org/10.1093/bioinformatics/bts606

Data accessibility and benefit-sharing statement

Sequence reads for the reference genome assembly and whole-genome re-sequencing data have been deposited in the GenBank Sequence Read Archive under the accession code BioProject ID PRJNA777424. Supplementary File S2 contains scripts and commands used for all bioinformatic analyses. Full detailed metadata and results may be consulted in the supplied Rmarkdown reports in the Supplementary Appendix (Meyer_et_al_Results_Appendix.html). Regarding the benefits generated, we reanalysed genetic samples from countries (other than France) that were published in a previous study (Riquet et al., 2019) where collaborators who provided samples were already included as co-authors. The results of this research have been shared with the sample providers and the broader scientific community, in particular through the sharing of genomic sequences on NCBI public databases. Other genetic resources will be made available on request.

Author contributions

LM, BG, NB and P-AG conceived the study. LM wrote the manuscript with inputs from all co-authors.

Sampling was conducted by NB, FR, PB, RC and P-AG. CA performed HMW DNA extraction and preparation for the reference genome 10X library. DNA extraction for WGS libraries was performed by LM, PB, and FR. Aspects of degraded/historical DNA extraction and analysis were handled by AF, CDS, BG and LM. WGS library preparation was done by LM and FC. FC, AB and ED were responsible for haplotagging library construction and sequencing. P-AG produced the reference genome and analysed linked-read sequencing data. PB provided scripts for genotyping high-coverage samples. LM performed all other bioinformatics and population genomic analyses. P-AG managed financial support.

Tables and Figures

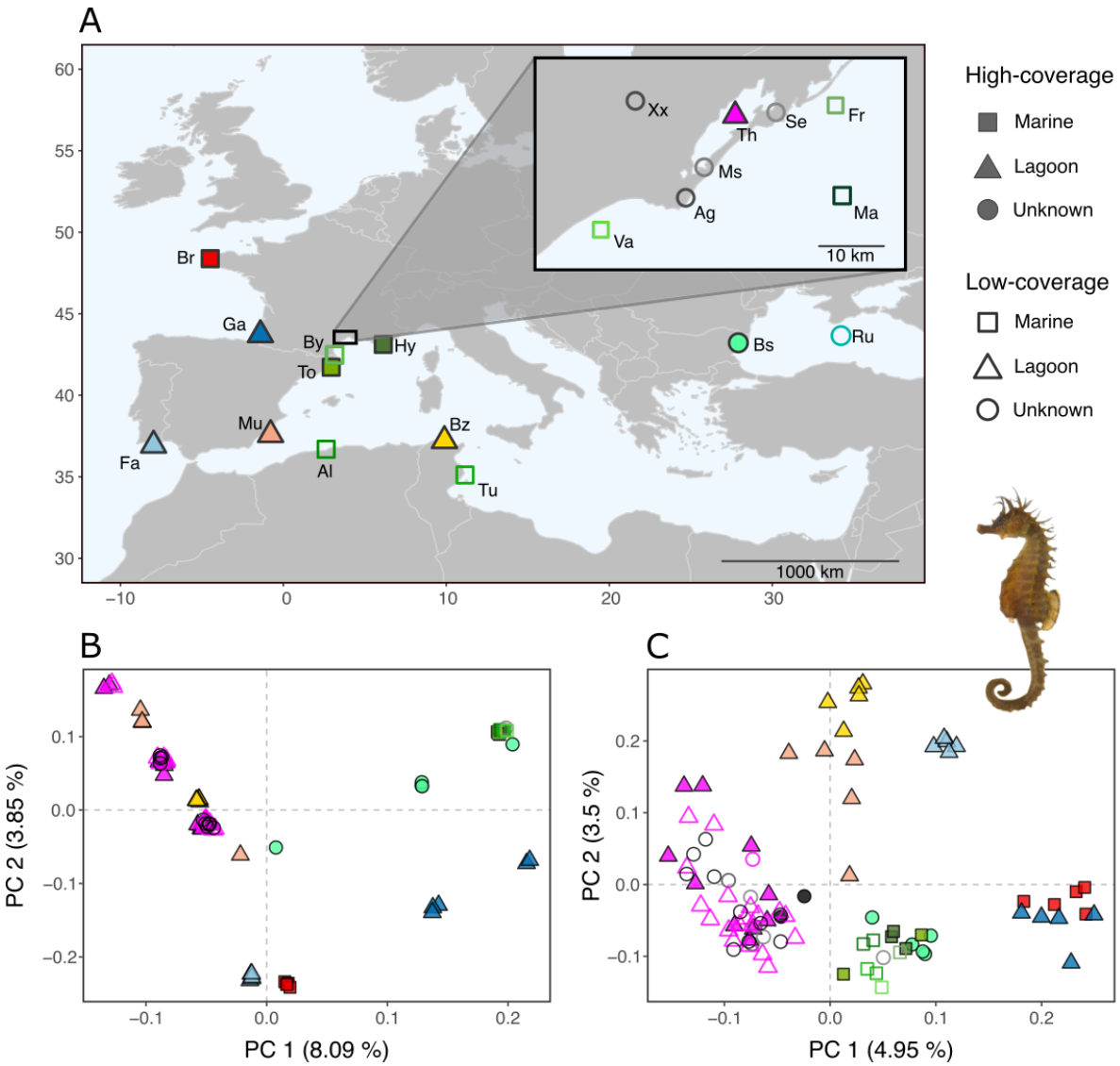


Fig. 1. A) Sampling map for the full dataset of *H. guttulatus* samples (n=112). Triangles: lagoon sites. Squares: marine sites. Circles: unknown habitat type. Filled shapes: medium to high-coverage samples for which individual genotype calling was performed (the GATK dataset). Unfilled shapes: low-coverage samples for which a genotype likelihood-based approach was used. Br: Brest. Ga: Hossegor. Fa: Faro. Mu: Murcia. To: Tossa de mar. By: Banyuls. Va: Valras. Ag: Agde. Ms: Marseillan. Th: Thau lagoon. Se: Sète. Fr: Frontignan. Hy: Hyères. Ma: unknown Mediterranean marine site. Al: unknown Algerian site. Bz: Bizerte lagoon. Tu: unknown Tunisian site. Bs: Varna. Ru: unknown Russian site. Xx: unknown northwestern Mediterranean location. B & C) PCA of 89 individuals with >3X coverage based on IBS distances calculated in ANGSD. PCA was either performed using all genome-wide markers (1,478,955 SNPs) (B), or using a subset of markers at linkage equilibrium (898 SNPs) (C). © Seahorse picture Iglésias 2013.

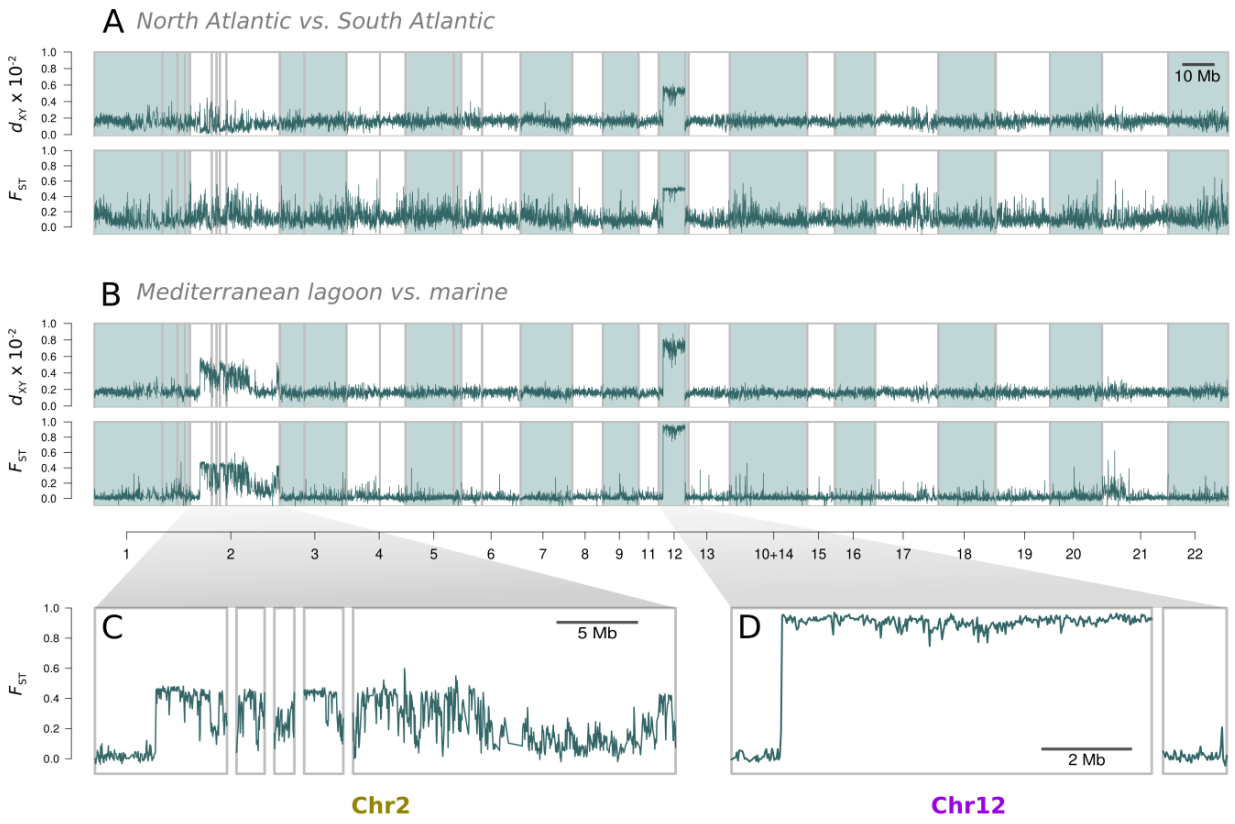


Fig. 2. Genomic landscape of absolute divergence (d_{XY} x 10^{-2}) and differentiation (F_{ST}) calculated in 25 kb non-overlapping sliding windows between a north Atlantic (Ga) and a south Atlantic (Fa) population (A) and between a Mediterranean lagoon (Mu) and a Mediterranean marine (To and Hy) population (B), using 5 high-coverage samples per population. *Hippocampus guttulatus* scaffolds (grey rectangles) were aligned to the chromosome-level assembly of *H. erectus* and are displayed according to their homology with the 22 *H. erectus* chromosomes (marked by alternating white and blue rectangles). Regions of high divergence clustered on two chromosomes: Chr2 (~32 Mb) (C) and Chr12 (~11 Mb) (D), which each carry a large chromosomal inversion. Exact ordering and orientation of *H. guttulatus* scaffolds within chromosomes 2 and 12 of *H. erectus* could not be determined due to multiple rearrangements between species (Supplementary Figure S5).

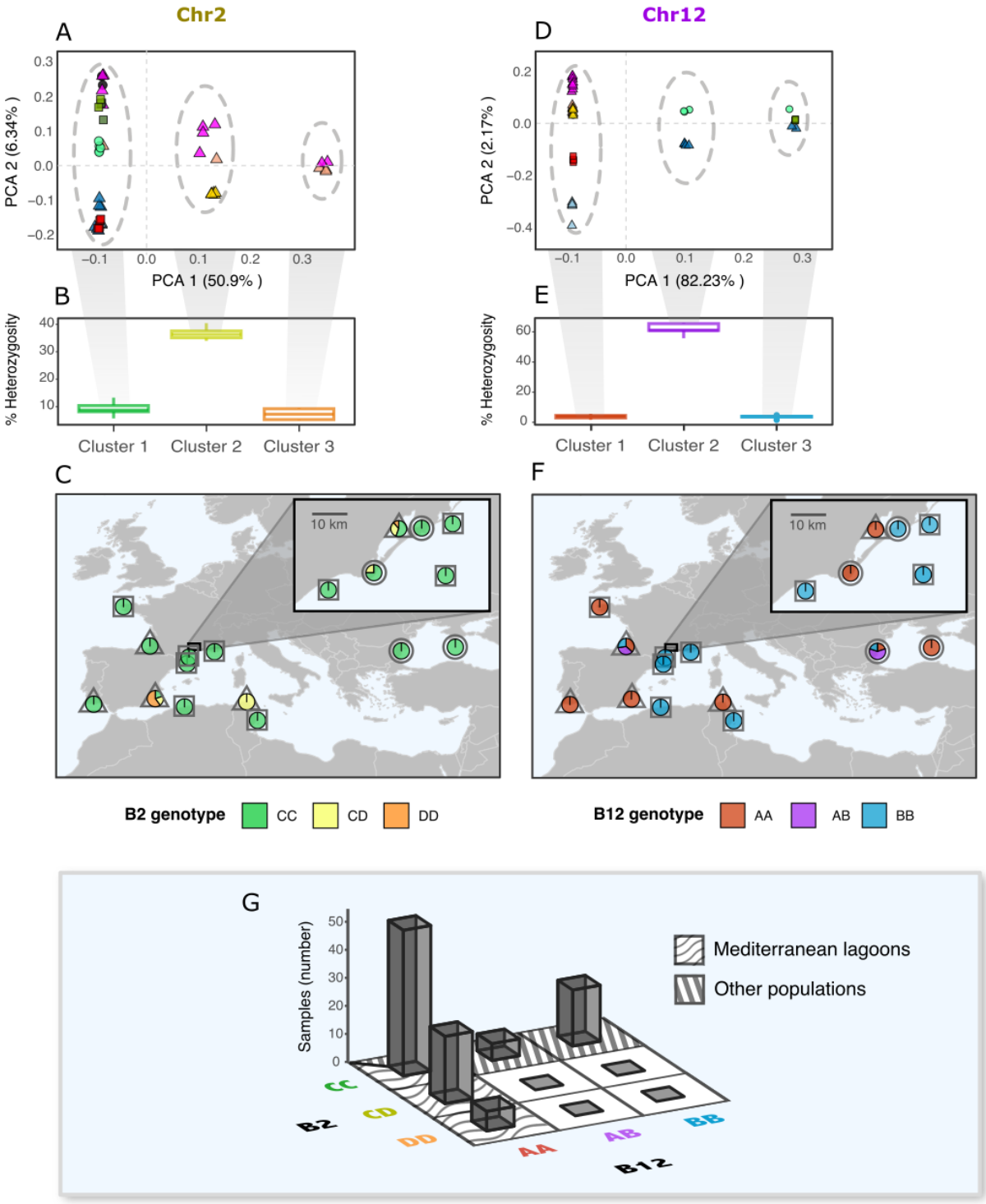


Fig. 3. Molecular diversity patterns observed for Chr2 (A-C) and Chr12 (D-F), which carry large inversions (B2 and B12). A & D) Chromosome-wide Principal Component Analysis (PCA) of 48 medium- to high-coverage samples (the GATK dataset) using the same symbols as in Fig. 1. The three clusters for Chr2 (A) and Chr12 (D) are illustrated by grey ellipses. B & E) Boxplots of observed heterozygosity (percentage of heterozygous sites) for each cluster. Heterozygosity was calculated at a chromosome-wide scale and averaged per sample. C & F) Maps of inversion genotype frequencies with respect to habitat type. Pie charts represent inversion genotypes and grey symbols indicate marine (square), lagoon (triangle) or uncharacterised (circle) habitats. G) Observed genotypic combinations at inversions B2 and B12,

showing distinct associations between Mediterranean lagoons (B12 fixed for the A haplotype and B2 polymorphic) and other populations (B2 fixed for the C haplotype, B12 locally polymorphic), and the complete absence of 4 genotype combinations.

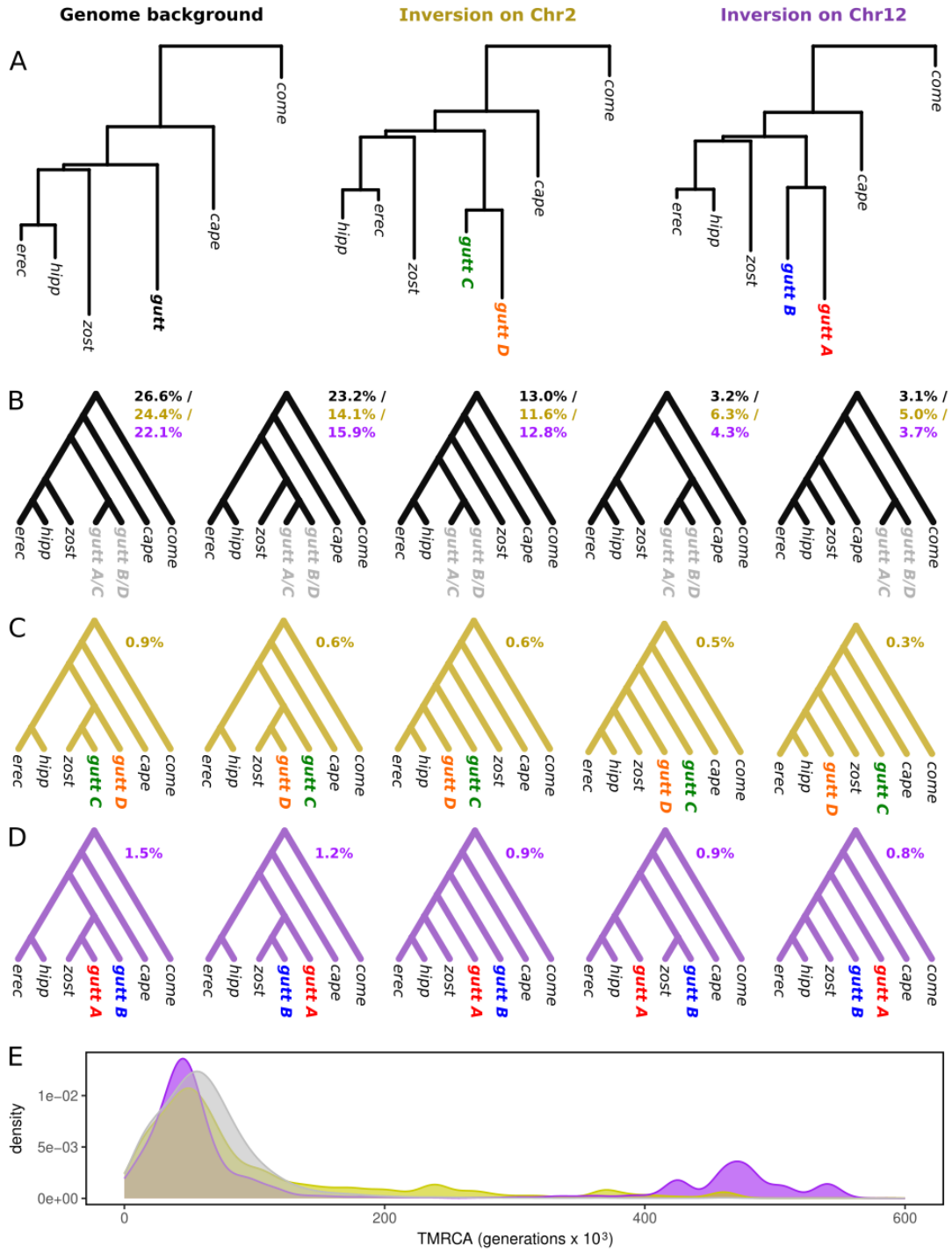


Fig. 4. A) Maximum likelihood phylogenies for different regions of the seahorse genome. B) The five most predominant 7-taxon topologies identified with Twisst in the genome background as well as in the inversions. Relative abundance among all topologies for each specific region is written as a percentage for the genome background (top), for inversion B2 (middle, leaves *gutt C* and *gutt D*) and for inversion B12 (bottom, leaves *gutt A* and *gutt B*). C-D) Amongst topologies not placing alternate haplotypes as sister lineages, the five most common are shown for inversions B2 (C) and B12 (D). E) Distribution of coalescent times estimated with *tsdate* (v0.1.5) in the genome background (grey), on Chr2 (yellow) and on Chr12 (purple). The TMRCA (in units of 1k generations) was extracted for any random pair of

haplotypes in every local genealogy along the genome. Abbreviations: *Come*: *H. comes. Cape*: *H. capensis. Zost*: *H. zosterae. Hipp*: *H. hippocampus. Erec*: *H. erectus*.

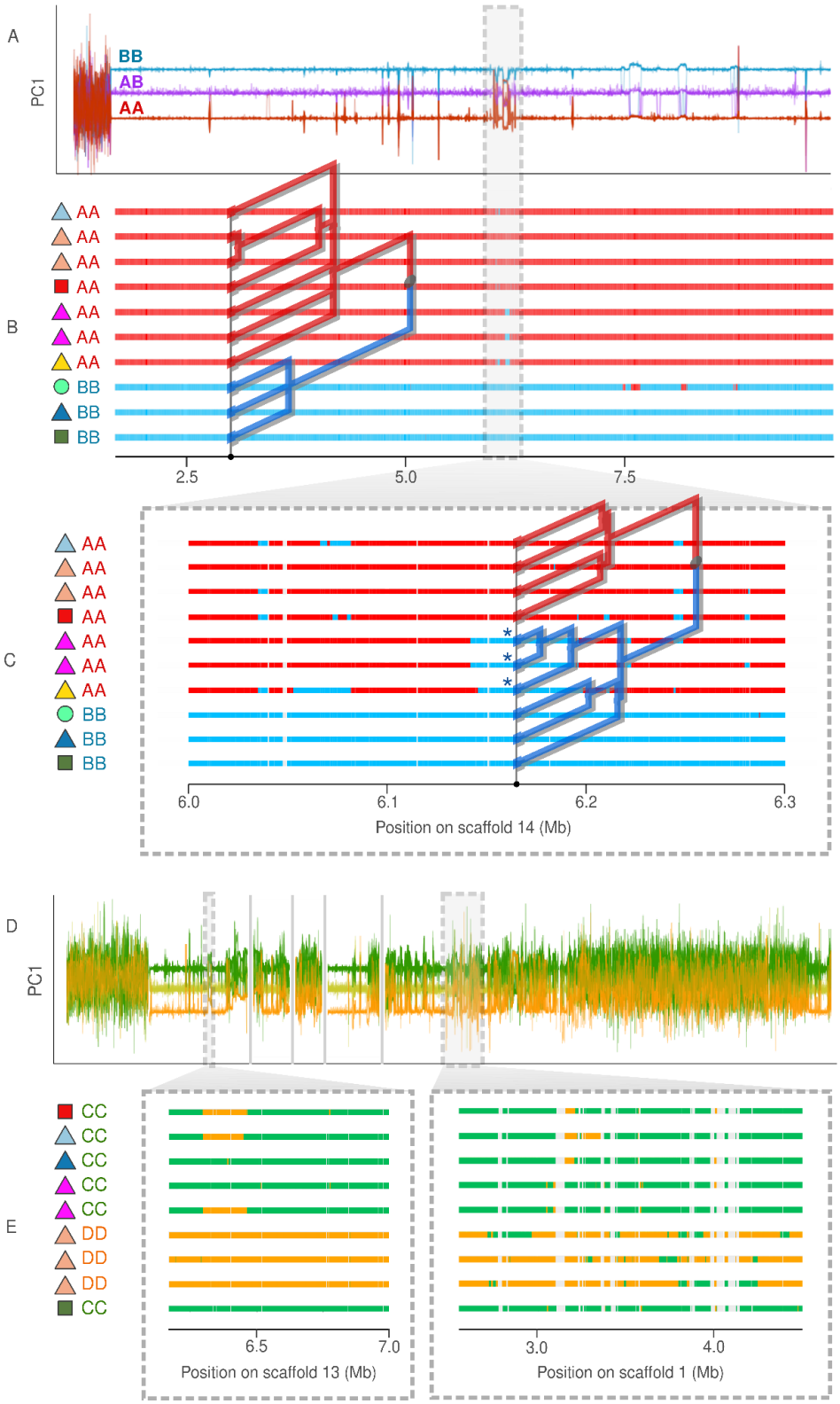


Fig. 5. Gene flux between the inverted and non-inverted haplotypes on Chr12 (A-C) and on Chr2 (D-E). A & D) PC1 coordinates from local PCA plotted for non-overlapping 5 kb window along the inversions. Individual lines represent samples and are coloured according to inversion genotype. Vertical grey lines indicate the edges of scaffolds composing Chr2. B & C & E) Chromosome painting showing local

ancestry within inversion regions as determined by our ARG-based approach. The superposed trees were constructed with *tsinfer* (v0.3.0) and illustrate how local ancestry at a given position was determined for each haplotype leaf. Horizontal coloured bars represent a subsample of all haplotypes and colours indicate inverted or non-inverted ancestry. Symbols on the left of each chromosome indicate sampling location and habitat type (see Fig. 1A), as well as the large-scale inversion genotype for each individual. C) Enlarged view of an introgressed region of B12, where certain A haplotypes (red) locally show B ancestry (blue). The tree at position 6.175 Mb shows that three recombinant A haplotypes (marked by asterisks) are grouped in the same clade as the B haplotypes. E) Enlarged views of two regions on Chr2 showing different levels of introgression.

SUPPLEMENTARY MATERIALS

Supplementary File S1

Supplementary tables and figures

- Supplementary Table S1. Samples used in the study (full dataset).
- Supplementary Fig S1. Number of reads per barcode (10X Chromium linked-read sequencing).
- Supplementary Fig S2. PMDtools
- Supplementary Fig S3. Snailplot of our reference genome.
- Supplementary Fig S4a and S4b. Snailplots of other genomes used.
- Supplementary Fig S5: D-genies alignment of our reference genome vs. H. erectus.
- Supplementary Fig S6: repeats
- Supplementary Fig S7: Tan9
- Supplementary Fig S8: *F*ST distributions.
- Supplementary Fig S9: *d*XY distributions.
- Supplementary Fig S10: Pi distributions.
- Supplementary Fig S11: D-genies alignment of Hgutt_V1 vs. HguttRefA
- Supplementary Fig S12. Haplotagging
- Supplementary Fig S13. *F*ST co-plot
- Supplementary Fig S14: all painted chromosomes (Chr12)
- Supplementary Fig S15a and S15b: all painted chromosomes (Chr2)
- Supplementary Fig S16: SNPs Riquet mapped scaffold 14

Supplementary Table S1: Full dataset of samples used in the study. For our sequenced samples, codes are composed of the species name abbreviation, the approximate sampling locality, approximate age and the sample number. The GATK dataset consisted of samples for which variant calling was performed (Variant Calling = “yes”). The last column shows mean coverage depth per individual. Br: Brest. Ga: Hossegor. Fa: Faro. Mu: Murcia. To: Tossa de mar. By: Banyuls. Va: Valras. Ag: Agde. Ms: Marseillan. Th: Thau lagoon. Se: Sète. Fr: Frontignan. Hy: Hyères. Ma: unknown Mediterranean marine site. Al: unknown Algerian site. Bz: Bizerte lagoon. Tu: unknown Tunisian site. Bs: Varna. Ru: unknown Russian site. Xx: unknown location.

Sample	Location	Habitat	Origin	Type	Library	Variant Calling	Mean cov
Hgutt_Ru_1856_77	Ru	unknown	Museum	Ethanol preserved	OVATION	No	5.15
Hgutt_Al_1864_80	Al	marine	Museum	Ethanol preserved	OVATION	No	2.16
Hgutt_Se_1898_81	Se	unknown	Museum	Ethanol preserved	OVATION	No	3.42
Hgutt_Se_1898_94	Se	unknown	Museum	Ethanol preserved	OVATION	No	8.16
Hgutt_Ag_1935_17	Ag	unknown	Citizen collect	Dried	OVATION	No	4.09
Hgutt_Ag_1935_19	Ag	unknown	Citizen collect	Dried	OVATION	No	1.54
Hgutt_Ag_1935_59	Ag	unknown	Citizen collect	Dried	OVATION	No	4.74
Hgutt_Ag_1935_70	Ag	unknown	Citizen collect	Dried	OVATION	No	4.88
Hgutt_Th_1935_66	Th	lagoon	Citizen collect	Dried	OVATION	No	4.14
Hgutt_Va_1935_22	Va	marine	Citizen collect	Dried	OVATION	No	4.00
Hgutt_Ba_1950_41	Th	lagoon	Citizen collect	Dried	OVATION	No	4.16
Hgutt_Ma_1960_60	Ma	marine	Citizen collect	Dried	OVATION	No	1.37
Hgutt_Ms_1960_21	Ms	unknown	Citizen collect	Dried	OVATION	No	4.20
Hgutt_Ro_1960_23	Th	lagoon	Citizen collect	Dried	OVATION	No	1.27
Hgutt_Ro_1960_43	Th	lagoon	Citizen collect	Dried	OVATION	No	4.39
Hgutt_Th_1960_11	Th	lagoon	Citizen collect	Dried	OVATION	No	3.91
Hgutt_Th_1960_24	Th	lagoon	Citizen collect	Dried	OVATION	No	4.30
Hgutt_Th_1960_42	Th	lagoon	Citizen collect	Dried	OVATION	No	3.11
Hgutt_Th_1960_44	Th	lagoon	Citizen collect	Dried	OVATION	No	4.76
Hgutt_Th_1960_95	Th	lagoon	Citizen collect	Dried	OVATION	No	1.61
Hgutt_Th_1964_45	Th	lagoon	Citizen collect	Dried	OVATION	No	4.46
Hgutt_Th_1964_65	Th	lagoon	Citizen collect	Dried	OVATION	No	3.93

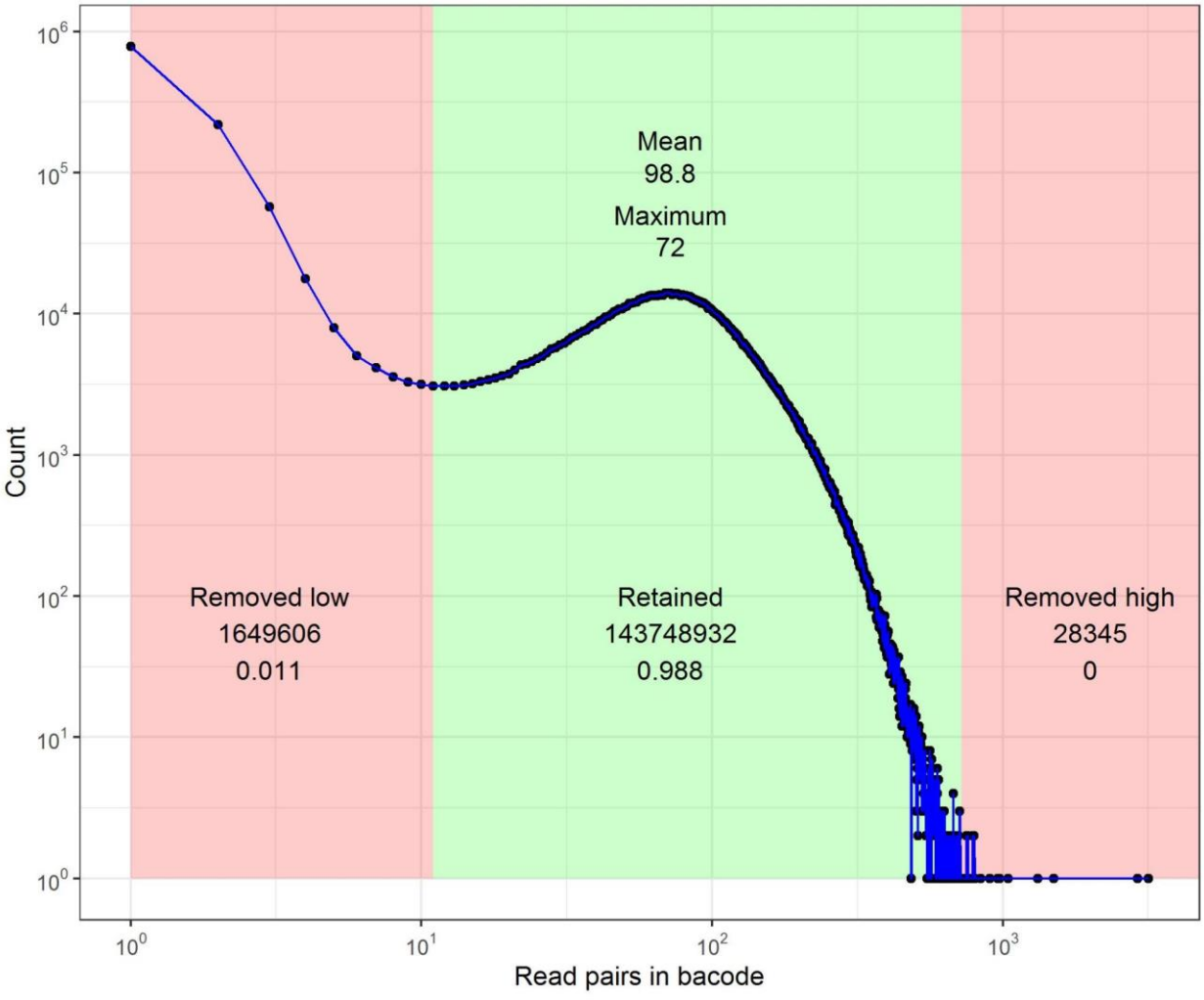
Hgutt_Th_1970_30	Th	lagoon	Citizen collect	Dried	OVATION	No	4.03
Hgutt_Th_1970_46	Th	lagoon	Citizen collect	Dried	OVATION	No	3.82
Hgutt_Il_1975_47	Th	lagoon	Citizen collect	Dried	OVATION	No	4.40
Hgutt_Il_1975_48	Th	lagoon	Citizen collect	Dried	OVATION	Yes	6.66
Hgutt_Il_1975_68	Th	lagoon	Citizen collect	Dried	OVATION	Yes	6.95
Hgutt_Bo_1980_50	Th	lagoon	Citizen collect	Dried	OVATION	No	4.30
Hgutt_By_1980_67	By	marine	Citizen collect	Dried	OVATION	No	1.33
Hgutt_Th_1980_12	Th	lagoon	Citizen collect	Dried	OVATION	No	1.10
Hgutt_Th_1980_49	Th	lagoon	Citizen collect	Dried	OVATION	No	1.58
Hgutt_Th_1980_69	Th	lagoon	Citizen collect	Dried	OVATION	No	4.59
Hgutt_Tu_1980_25	Tu	lagoon	Citizen collect	Dried	OVATION	No	4.78
Hgutt_Tu_1980_31	Tu	lagoon	Citizen collect	Dried	OVATION	No	4.46
Hgutt_Me_1982_51	Th	lagoon	Citizen collect	Dried	OVATION	No	4.63
Hgutt_Fr_1985_52	Fr	marine	Citizen collect	Dried	OVATION	No	6.07
Hgutt_Ma_1985_61	Ma	marine	Citizen collect	Dried	OVATION	No	4.89
Hgutt_Ms_1985_18	Ms	unknown	Citizen collect	Dried	OVATION	No	3.91
Hgutt_Xx_1985_71	Xx	unknown	Citizen collect	Dried	OVATION	No	6.15
Hgutt_Ma_1990_63	Ma	marine	Citizen collect	Dried	OVATION	No	4.62
Hgutt_Th_1990_72	Th	lagoon	Citizen collect	Dried	OVATION	Yes	6.46
Hgutt_Xx_1990_13	Xx	unknown	Citizen collect	Dried	OVATION	No	5.3
Hgutt_Xx_1990_14	Xx	unknown	Citizen collect	Dried	OVATION	No	2.57
Hgutt_Xx_1990_15	Xx	unknown	Citizen collect	Dried	OVATION	No	1.46
Hgutt_Xx_1990_16	Xx	unknown	Citizen collect	Dried	OVATION	No	3.45
Hgutt_Xx_1990_26	Xx	unknown	Citizen collect	Dried	OVATION	No	4.5
Hgutt_Xx_1990_27	Xx	unknown	Citizen collect	Dried	OVATION	No	4.33
Hgutt_Xx_1990_32	Xx	unknown	Citizen collect	Dried	OVATION	No	4.67
Hgutt_Xx_1990_33	Xx	unknown	Citizen collect	Dried	OVATION	No	4.82
Hgutt_Xx_1990_34	Xx	unknown	Citizen collect	Dried	OVATION	No	4.72
Hgutt_Xx_1990_53	Xx	unknown	Citizen	Dried	OVATION	No	4.98

			collect				
Hgutt_Xx_1990_54	Xx	unknown	Citizen collect	Dried	OVATION	No	1.77
Hgutt_Th_1991_55	Th	lagoon	Citizen collect	Dried	OVATION	No	5.2
Hgutt_Ms_1995_64	Ms	unknown	Citizen collect	Dried	OVATION	No	5.28
Hgutt_Th_1995_56	Th	lagoon	Citizen collect	Dried	OVATION	No	4.80
Hgutt_Th_1995_76	Th	lagoon	Citizen collect	Dried	OVATION	No	1.61
Hgutt_Xx_1995_20	Xx	unknown	Citizen collect	Dried	OVATION	Yes	7.03
Hgutt_Xx_1995_38	Xx	unknown	Citizen collect	Dried	OVATION	No	1.33
Hgutt_Xx_1995_39	Xx	unknown	Citizen collect	Dried	OVATION	No	3.66
Hgutt_Xx_1995_73	Xx	unknown	Citizen collect	Dried	OVATION	Yes	8.35
Hgutt_Th_1996_28	Th	lagoon	Citizen collect	Dried	OVATION	No	4.12
Hgutt_Th_1996_74	Th	lagoon	Citizen collect	Dried	OVATION	No	1.54
Hgutt_Th_1996_75	Th	lagoon	Citizen collect	Dried	OVATION	No	4.97
Hgutt_Bs_2006_83	Bs	unknown	Riquet et al. 2019	Fresh	OVATION	Yes	8.17
Hgutt_Bs_2006_85	Bs	unknown	Riquet et al. 2019	Fresh	OVATION	Yes	11.06
Hgutt_Bs_2006_87	Bs	unknown	Riquet et al. 2019	Fresh	OVATION	Yes	11.2
Hgutt_Bs_2006_89	Bs	unknown	Riquet et al. 2019	Fresh	OVATION	Yes	9.97
Hgutt_Bs_2006_91	Bs	unknown	Riquet et al. 2019	Fresh	OVATION	Yes	11.71
Hgutt_Th_2007_36	Th	lagoon	Citizen collect	Dried	OVATION	Yes	7.39
Hgutt_Th_2007_57	Th	lagoon	Citizen collect	Dried	OVATION	No	5.94
Hgutt_Me_2010_58	Th	lagoon	Citizen collect	Dried	OVATION	Yes	7.85
Hgutt_Br_2014_06	Br	marine	Riquet et al. 2019	Fresh	OVATION	Yes	7.75
Hgutt_Br_2014_07	Br	marine	Riquet et al. 2019	Fresh	OVATION	Yes	8.18
Hgutt_Br_2014_08	Br	marine	Riquet et al. 2019	Fresh	OVATION	Yes	6.68
Hgutt_Br_2014_09	Br	marine	Riquet et al. 2019	Fresh	OVATION	Yes	6.30
Hgutt_Br_2014_10	Br	marine	Riquet et al. 2019	Fresh	OVATION	Yes	7.04
Hgutt_Bz_2014_82	Bz	lagoon	Riquet et al. 2019	Fresh	OVATION	Yes	9.81
Hgutt_Bz_2014_84	Bz	lagoon	Riquet et al. 2019	Fresh	OVATION	Yes	10.23
Hgutt_Bz_2014_86	Bz	lagoon	Riquet et al. 2019	Fresh	OVATION	Yes	9.40

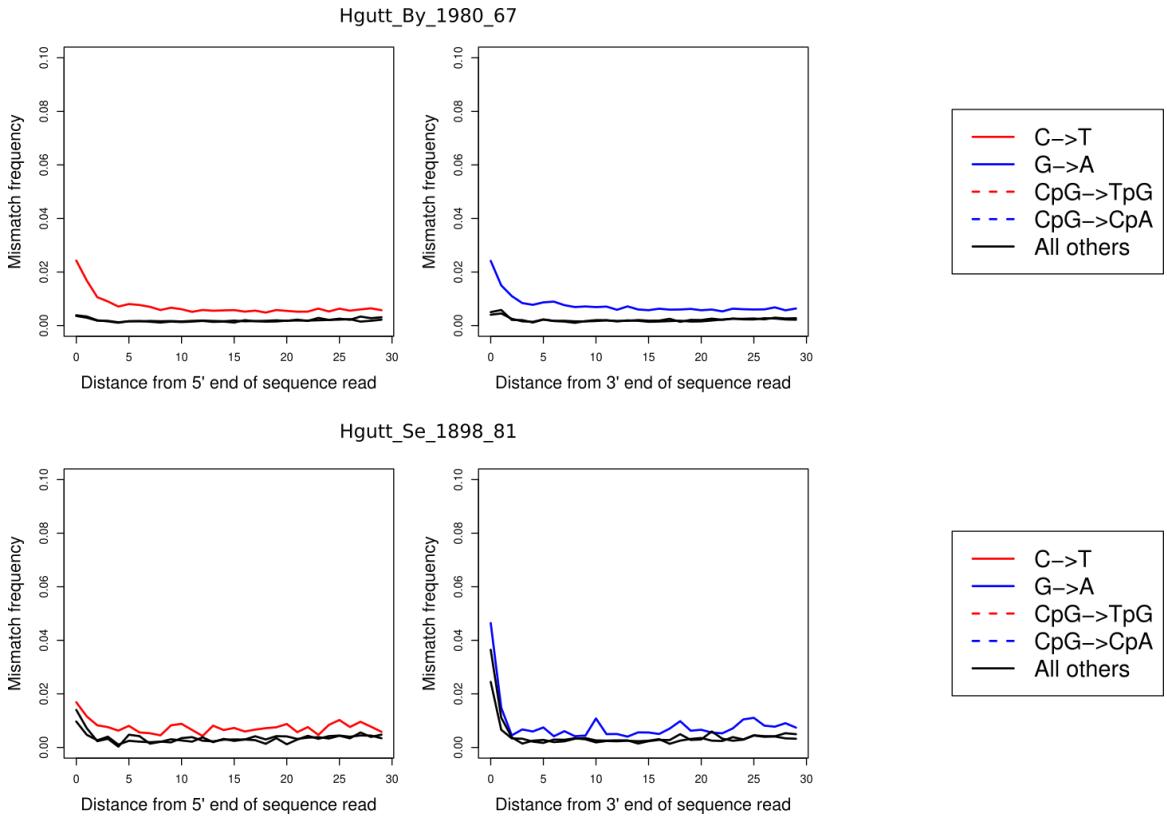
Hgutt_Bz_2014_88	Bz	lagoon	Riquet et al. 2019	Fresh	OVATION	Yes	11.03
Hgutt_Bz_2014_90	Bz	lagoon	Riquet et al. 2019	Fresh	OVATION	Yes	8.83
Hgutt_Li_2015_1	Th	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	39.82
Hgutt_Li_2015_3	Th	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	54.64
Hgutt_Li_2015_4	Th	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	51.15
Hgutt_Li_2015_5	Th	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	47.56
Hgutt_Li_2015_6	Th	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	47.57
Hgutt_To_2016_01	To	marine	Riquet et al. 2019	Fresh	OVATION	Yes	7.83
Hgutt_To_2016_02	To	marine	Riquet et al. 2019	Fresh	OVATION	Yes	6.98
Hgutt_Hy_2017_03	Hy	marine	Riquet et al. 2019	Fresh	OVATION	Yes	5.75
Hgutt_Hy_2017_04	Hy	marine	Riquet et al. 2019	Fresh	OVATION	Yes	7.41
Hgutt_Hy_2017_05	Hy	marine	Riquet et al. 2019	Fresh	OVATION	Yes	8.18
Hgutt_Ga_2019_10	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	No	2.89
Hgutt_Ga_2019_11	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	50.33
Hgutt_Ga_2019_12	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	No	1.77
Hgutt_Ga_2019_1	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	No	1.35
Hgutt_Ga_2019_2	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	No	1.34
Hgutt_Ga_2019_3	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	31.01
Hgutt_Ga_2019_4	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	No	1.28
Hgutt_Ga_2019_6	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	17.7
Hgutt_Ga_2019_7	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	46.32
Hgutt_Ga_2019_8	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	5.74
Hgutt_Ga_2019_9	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	13.36
Hgutt_Mu_2019_1	Mu	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	48.45
Hgutt_Mu_2019_2	Mu	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	43.72
Hgutt_Mu_2019_3	Mu	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	20.79
Hgutt_Mu_2019_4	Mu	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	33.14
Hgutt_Mu_2019_5	Mu	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	43.94
Hgutt_Fa_2020_2	Fa	lagoon	Barry et	Fresh	TruSeq	Yes	43.61

			al. (2022)				
Hgutt_Fa_2020_3	Fa	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	48.43
Hgutt_Fa_2020_4	Fa	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	45.89
Hgutt_Fa_2020_5	Fa	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	21.86
Hgutt_Fa_2020_6	Fa	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	47.96

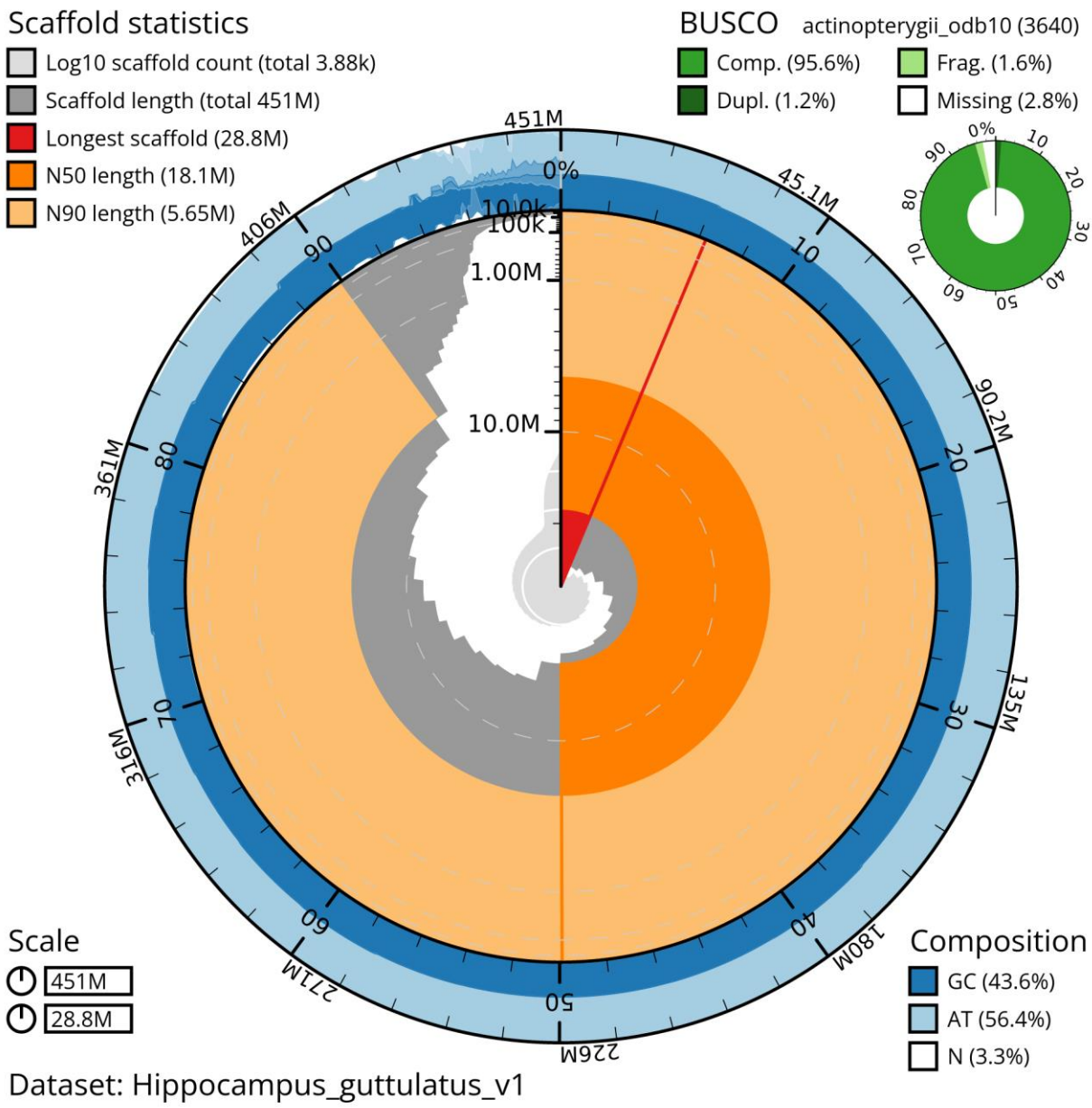
Supplementary Figure S1: Count distribution of the number of read pairs per barcode in the deduplicated 10X Chromium linked-read sequencing data generated for the reference genome assembly (Hgutt_V1). Molecular barcodes associated with low numbers of read pairs (left part of the distribution shaded in red) reflect noise in the 10X data (e.g. sequencing errors within barcodes), and were therefore excluded before performing *de novo* assembly (1,649,606 read pairs, 1.1% of the total sequencing data). Over-represented molecular barcodes on the right end of the distribution (28,345 read pairs) were also excluded. The resulting average number of read pairs per retained barcode was approximately 100, with a maximum occurrence at 72.



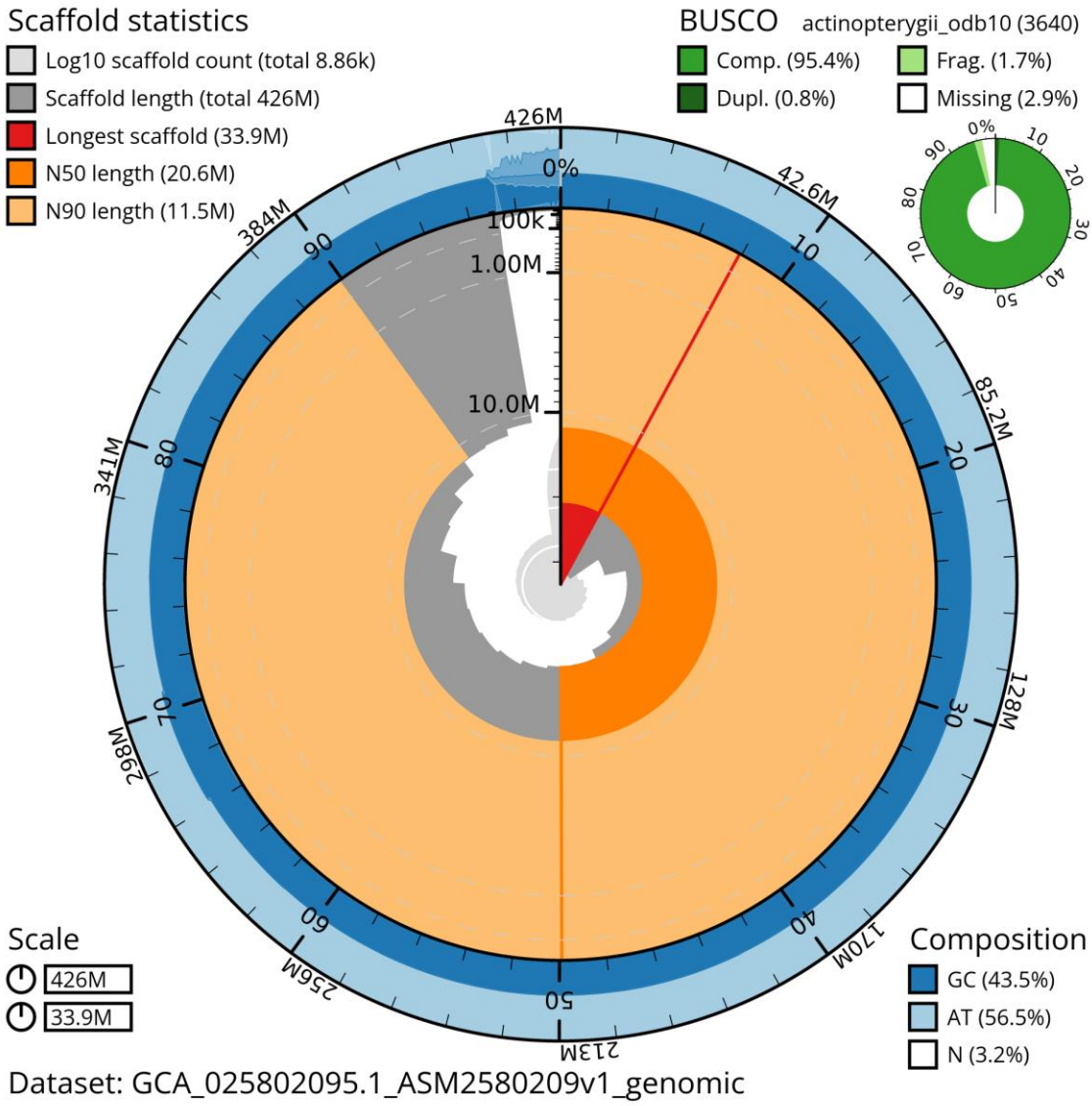
Supplementary Figure S2: DNA damage patterns (cytosine deamination) visualised using PMDtools (v0.60) (Skoglund et al., 2014). Results are shown for two historical samples.



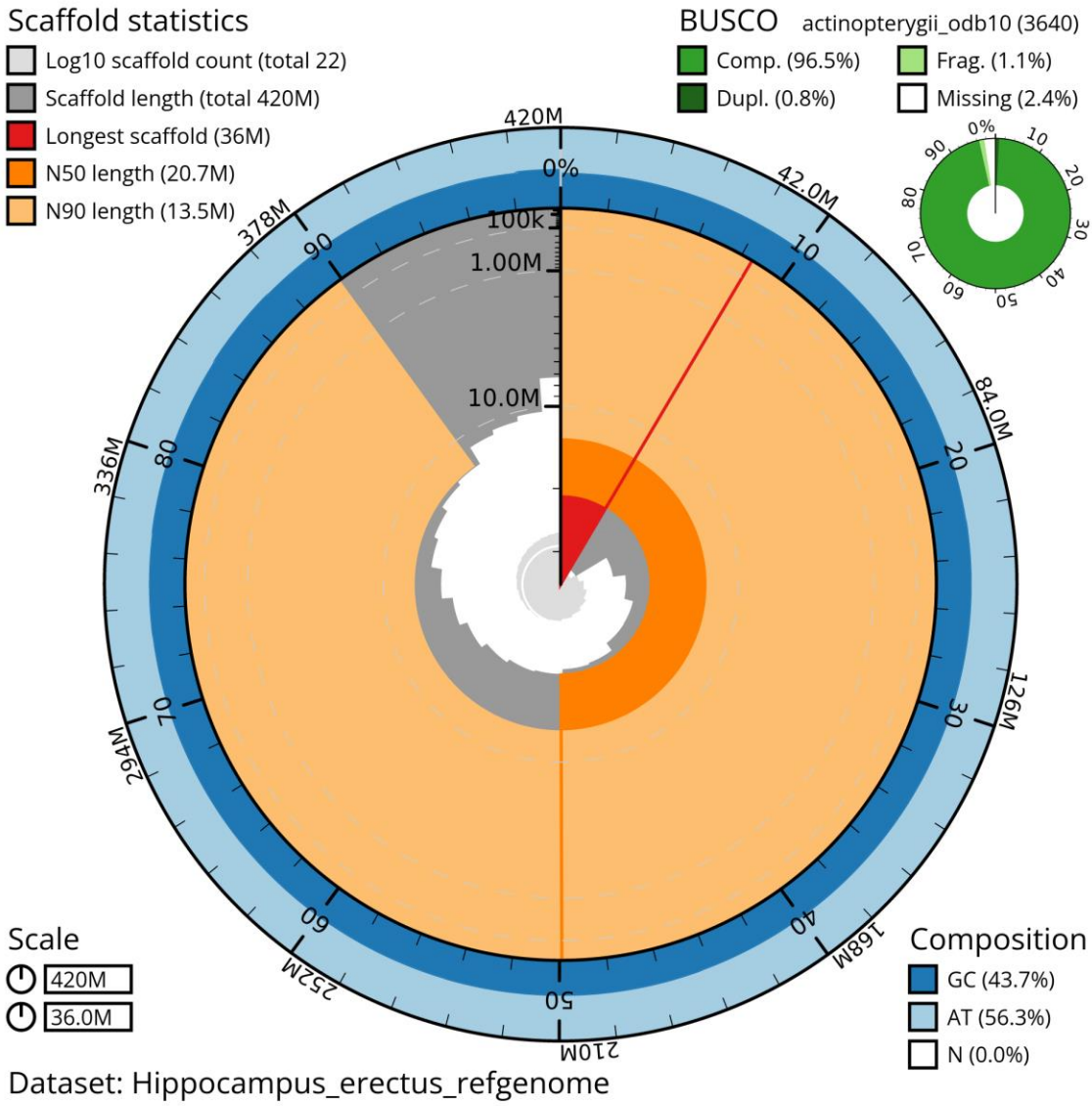
Supplementary Figure S3: Snail plot describing the assembly statistics of the *Hippocampus guttulatus* Hgutt_V1 reference genome. The genome contiguity is shown in a circle representing the full assembly length of 451 Mb, with the N50 length in dark orange and the N90 length in light orange. The longest scaffold was 28.8 Mb. The completeness BUSCO scores (in shades of green) and base composition (percentage of GC in dark blue, AT in light blue, and N in light grey) appear in top right and bottom left panels, respectively. The plot was generated using *Blobtoolkit* (https://blobtoolkit.genomehubs.org/).



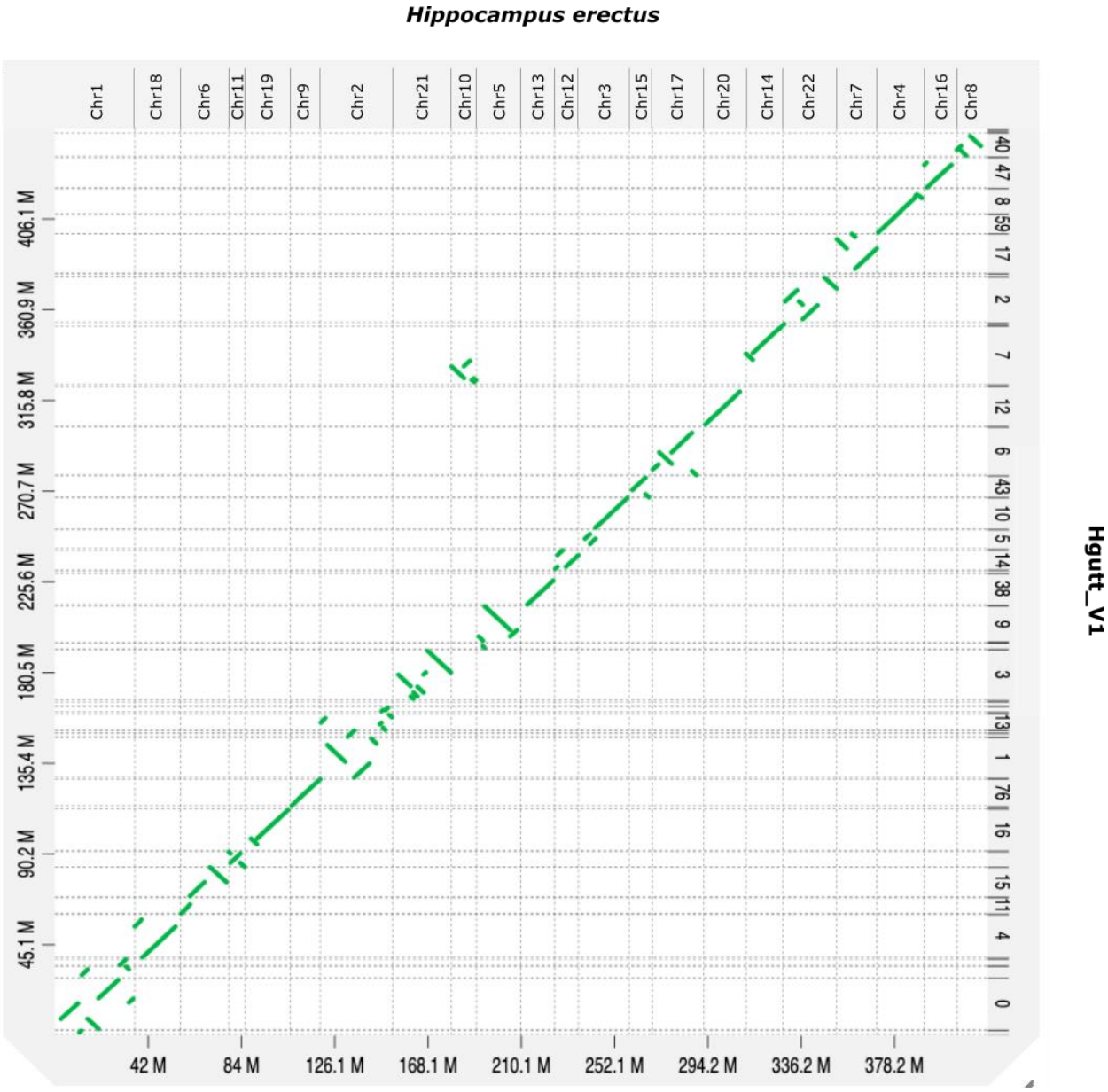
Supplementary Figure S4a: Snail plot describing the assembly statistics of the *Hippocampus guttulatus* GCA_025802095.1 reference genome (HguttRefA, Jones et al in prep). (Enlarged and split into S4a and S4b)



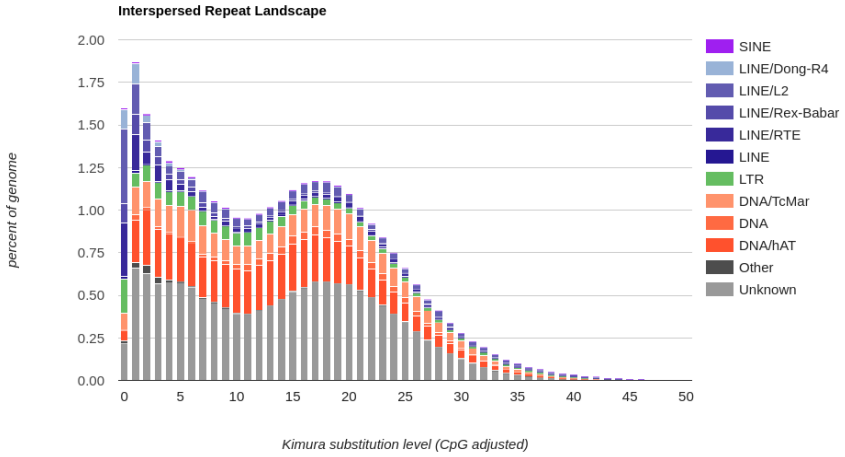
Supplementary Figure S4b: Snail plot describing the assembly statistics of the *Hippocampus erectus* reference genome (Li et al 2021). (Enlarged and split into S4a and S4b)

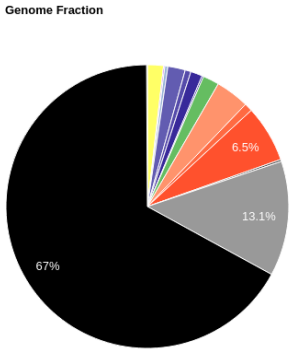


Supplementary Figure S5: Genomic alignment dot plot showing the comparison between the chromosome-level assembly of *Hippocampus erectus* (top) and our *H. guttulatus* Hgutt_V1 genome assembly (right). Only shown here are alignment matches passing a minimum size filter and a similarity threshold of 50%. All *H. guttulatus* scaffolds match a single *H. erectus* chromosome (with varied number of intrachromosomal rearrangements), except for scaffold 7 that matches both Chr10 and Chr14 of the *H. erectus* genome. The plot was generated using *D-GENIES* (https://github.com/genotoul-bioinfo/dgenies). (Modified labels)

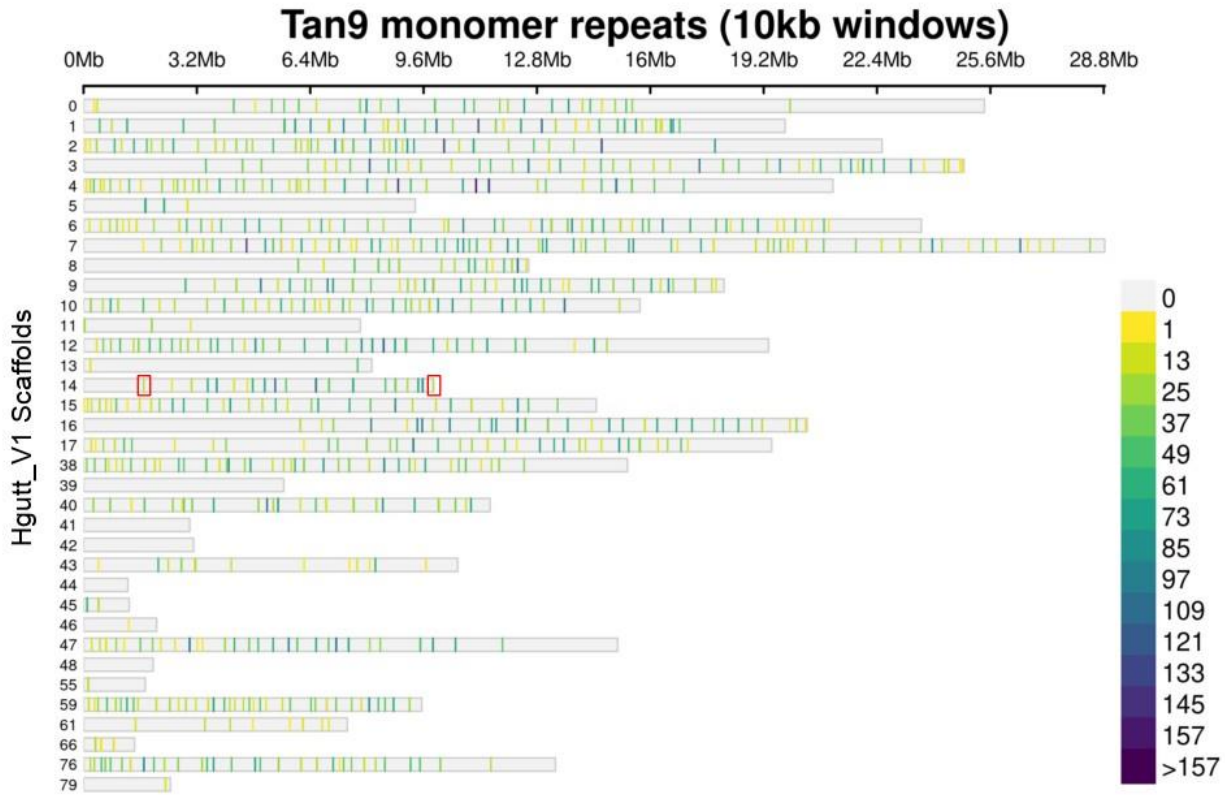


Supplementary Figure S6: Repeat landscape showing the percentage of the Hgutt_V1 genome occupied by different categories of interspersed repeat elements (right), and detailed as a function of weighted average Kimura divergence in alignments for each repeat family (left).

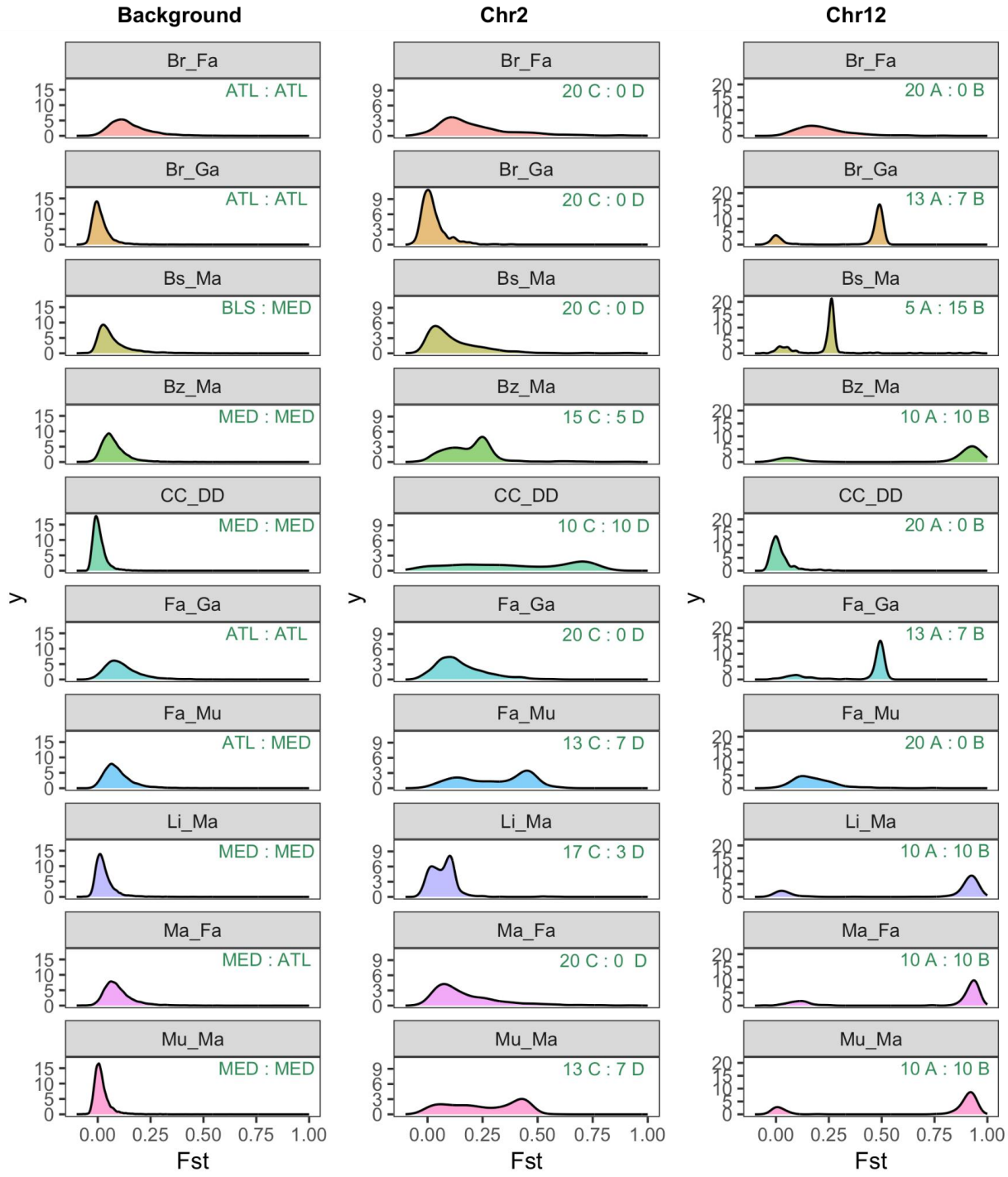

Interspersed Repeat Landscape
percent of genome
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
30
35
40
45
50
Kimura substitution level (CpG adjusted)
SINE
LINE/Dong-R4
LINE/L2
LINE/Rex-Babar
LINE/RTE
LINE
LTR
DNA/TcMar
DNA
DNA/hAT
Other
Unknown


Genome Fraction
6.5%
13.1%
67%

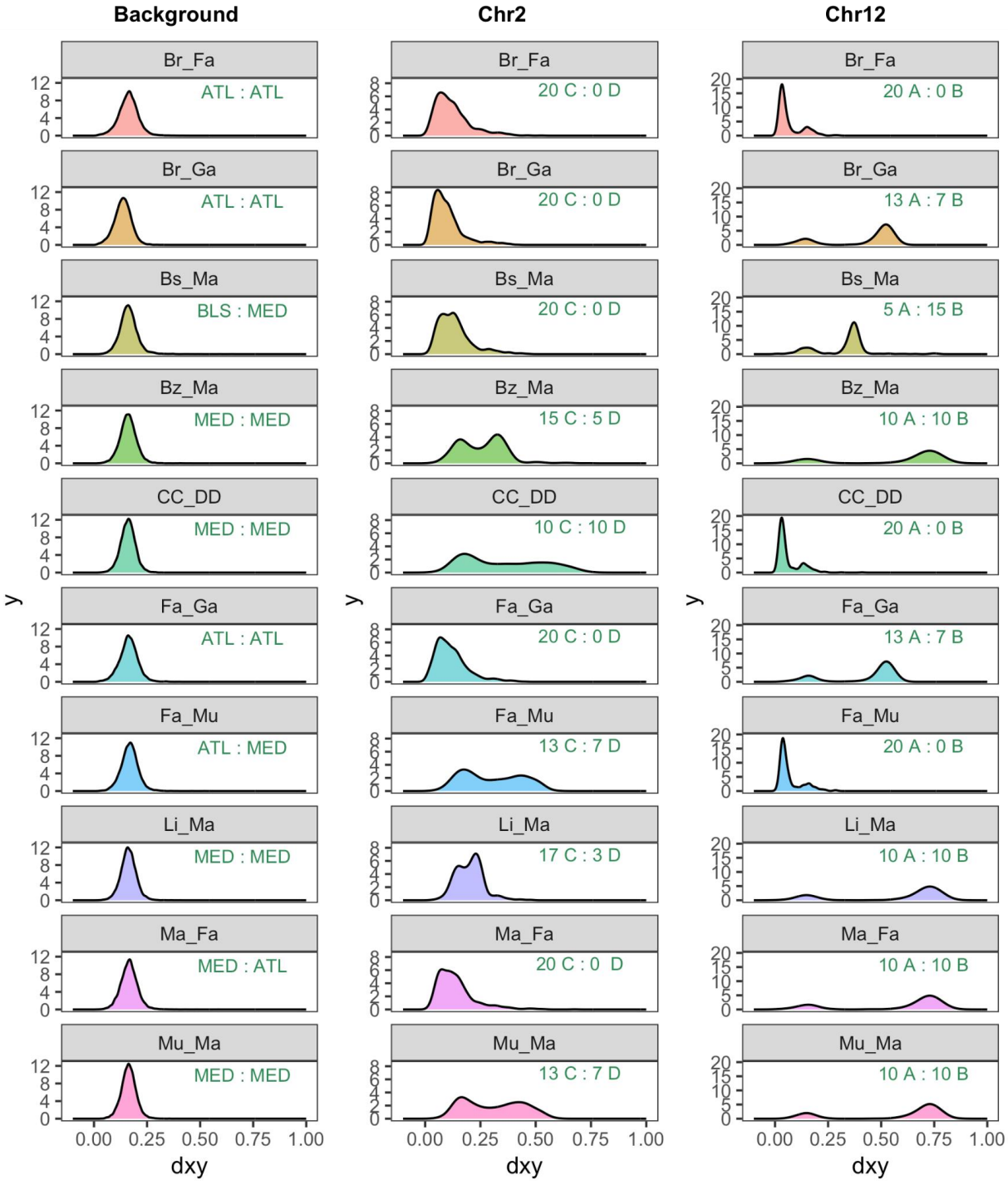
Supplementary Figure S7: Repeat landscape showing the number of Tan9 monomer repeats within 10kb windows across Hgutt_V1 scaffolds longer than 1Mb. The consensus sequence of the Tan9 monomer (37bp, GTCGTTTTTTTCGGCTAAAAACGCCTTACTATACATG) was searched with *blastn* with parameters set to M=1, N=-1, Q=2, R=2, W=8, following (Melters et al 2013). Red boxes indicate the Tan9 repeats where the two inversion breakpoints map on Scaffold 14 (Chr12), delimiting an 8.2 Mb-long inversion.



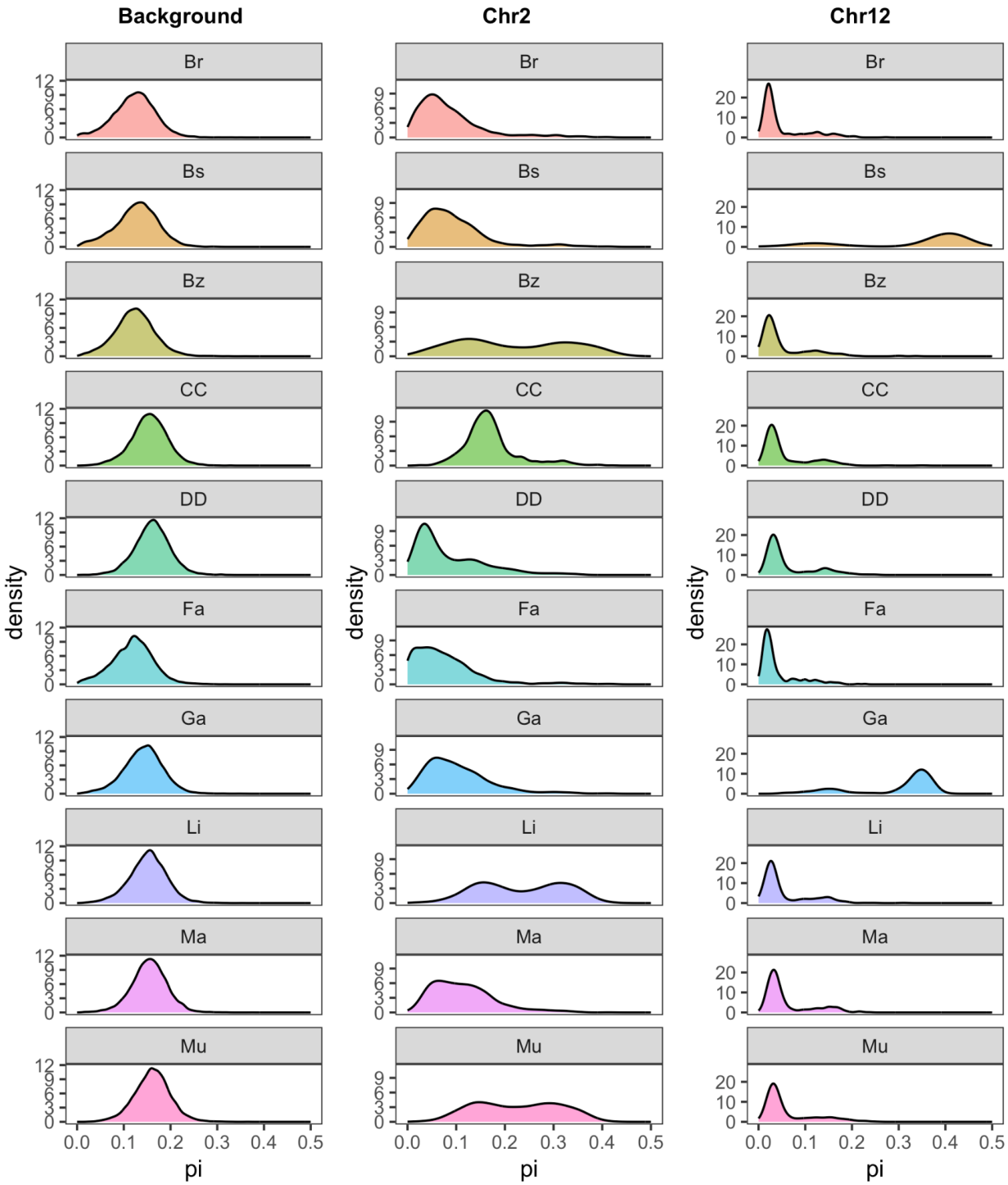
Supplementary Figure S8: Distributions of *F*ST values displayed separately for the genome background, Chr2 and Chr12. Black text gives information about the populations being compared: oceanic basins ("Background" column), total number of C/D alleles in the comparison ("Chr2" column), or total number of A/B alleles in the comparison ("Chr12" column). *F*ST was calculated for each population pair using the popgenWindows.py script in 25 kb windows (Martin, 2018; https://github.com/simonhmartin/genomics_general). The "Ma" population is here represented by individuals from the Mediterranean marine sites Tossa de mar and Hyères. (Labels enlarged. Background removed)



Supplementary Figure S9: Distributions of *d*XY values displayed separately for the genome background, Chr2 and Chr12 (for details see previous figure). (Labels enlarged. Background removed)

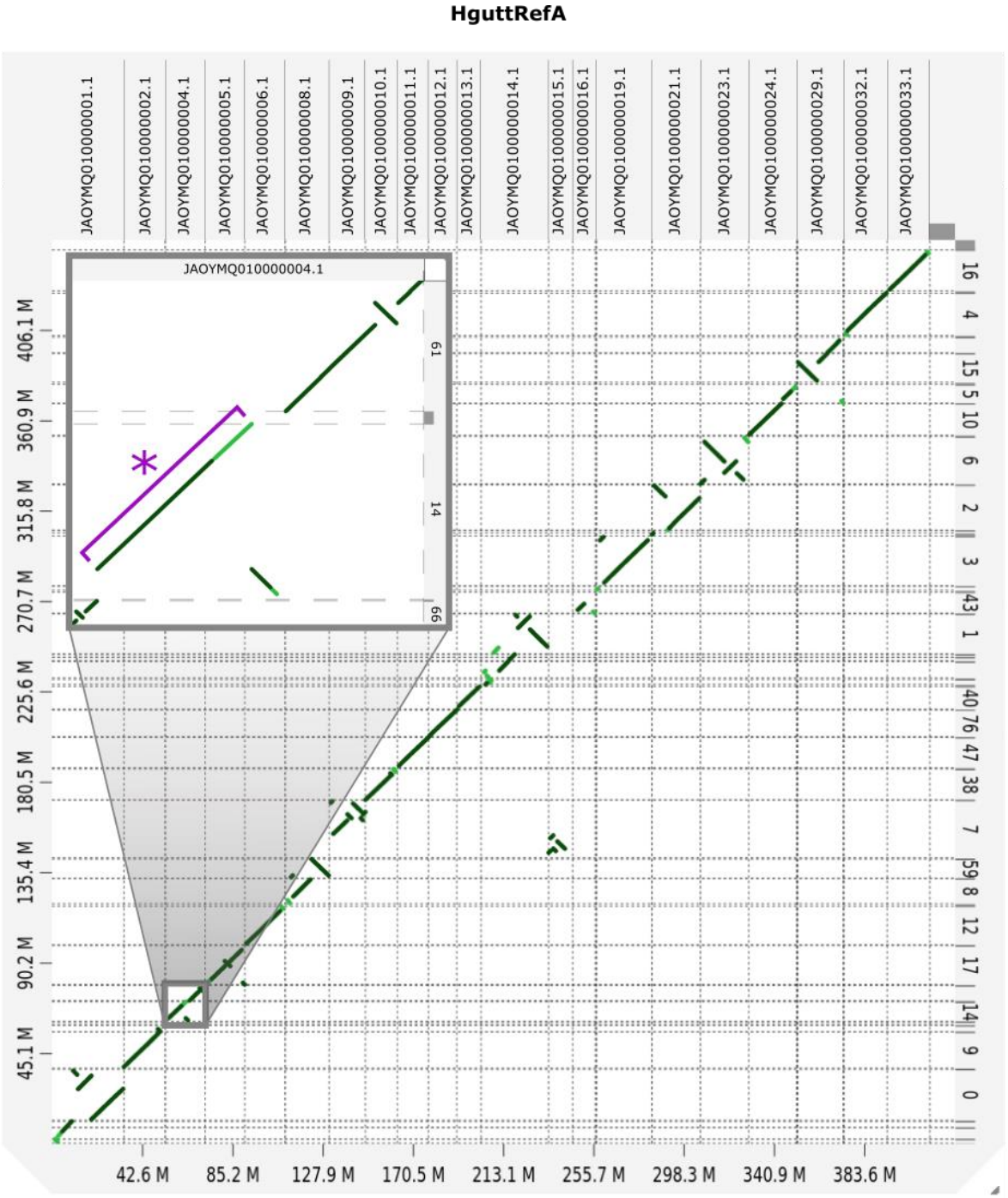


Supplementary Figure S10: Distributions of nucleotide diversity (π) values displayed separately for the genome background, Chr2 and Chr12. Nucleotide diversity was calculated for each population or haplotype group (for details see Supplementary Fig. S8). (Labels enlarged. Background removed)



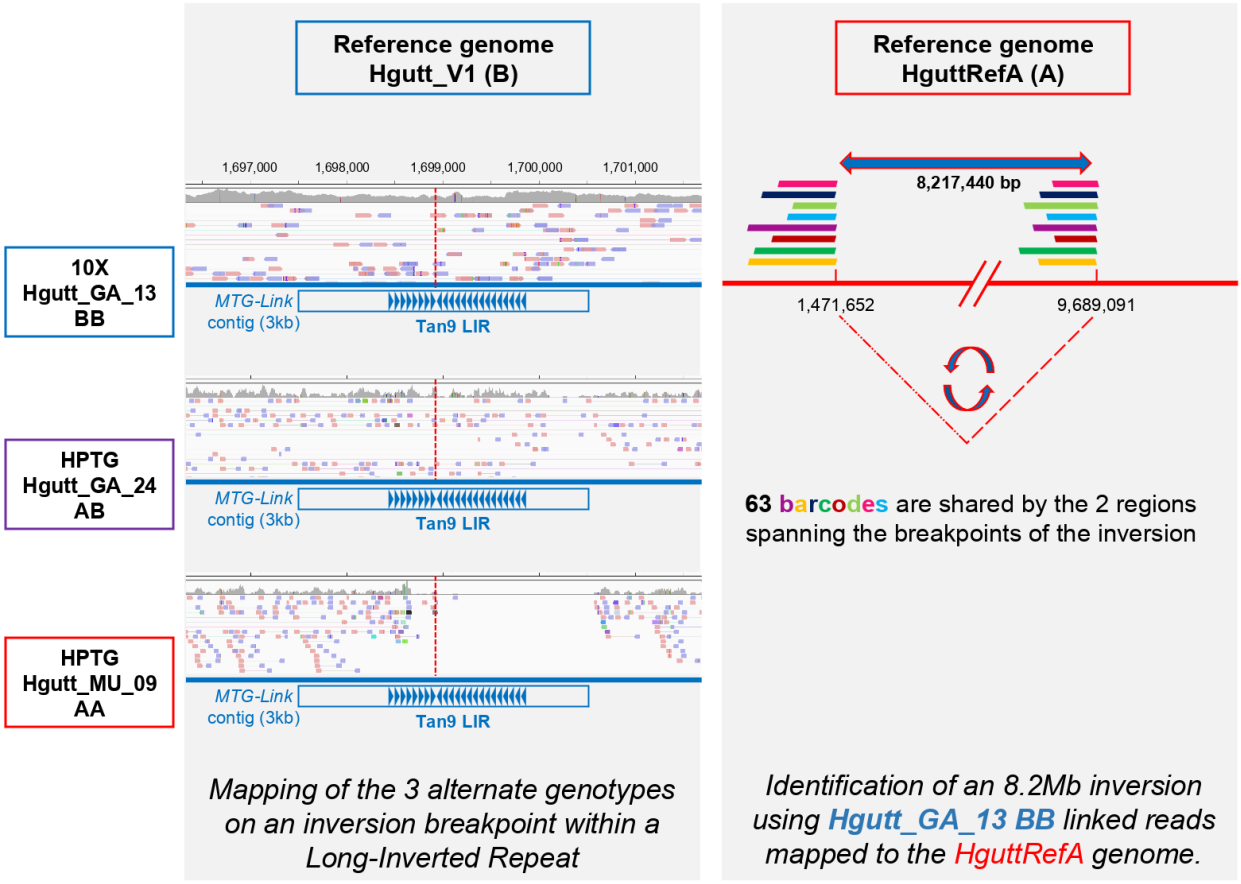
Supplementary Figure S11: Genomic alignment dot plot showing the comparison between the chromosome-level assembly of *Hippocampus guttulatus* from the English Channel (HguttRefA) (top) and

our Hgutt_V1 genome assembly (right). Only shown here are alignment matches passing a minimum size filter and a similarity threshold of 50%. The zoomed section shows chromosome JAOYMQ010000004.1 which corresponds to *H. erectus* Chr11 (scaffold 61) and Chr12. The junction between Chr11 and Chr12 within chromosome JAOYMQ010000004.1 is possibly an artefact of the assembly pipeline (i.e. the junction is not supported by reads from *H. guttulatus*). The purple asterisk indicates the 8.2 Mb-long region that is inverted between the two *H. guttulatus* reference assemblies. The plot was generated using *D-GENIES* (https://github.com/genotoul-bioinfo/dgenies). (Modified labels)

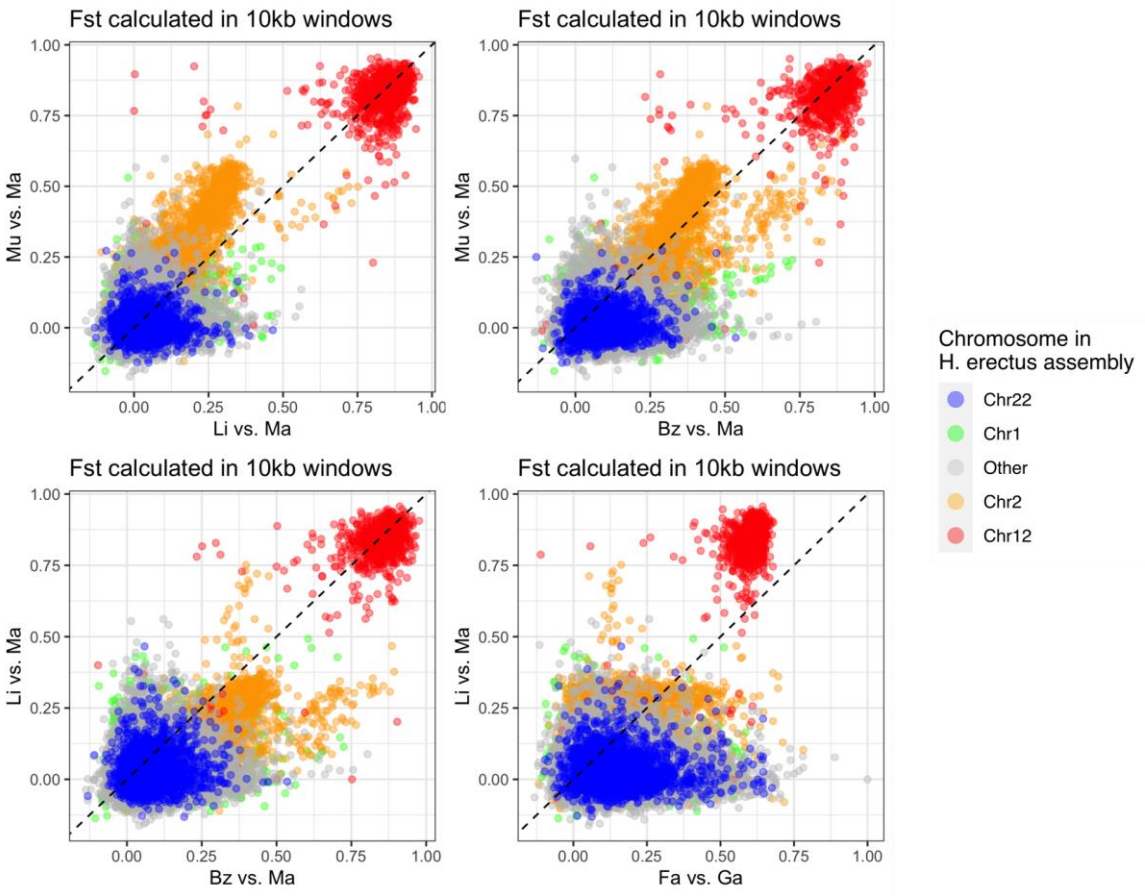


Supplementary Figure S12: Direct identification of an 8.2 Mb inversion on Chr12 and characterisation of inversion breakpoints. *Left panel:* The region around the inversion breakpoint localised near position 1,699 Mb (read dotted line) on scaffold 14 of Hgutt_V1 (Blue horizontal line, B arrangement) was

reassembled using *MTG-Link*, resulting in a 3kb contig matching the reference (blue rectangle). A long inverted repeat of Tan9 monomers (blue triangles) was found centred at the breakpoint position. Linked-read sequencing data obtained for three individuals with alternate B12 genotypes (i.e. the BB individual used for reference genome assembly sequenced at >100X coverage, plus an AA and an AB obtained by haplotagging and sequenced at ~12X coverage) were mapped against the Hgutt_V1 genome assembly. Horizontal links in each local alignment show linked paired-end reads sharing the same molecular barcode. Long synthetic molecules spanning the 3kb breakpoint region were found for both the BB and the AB genotypes, but not for the AA genotype that showed a ~1.5kb mapping gap directly after the breakpoint. *Right panel:* Linked-reads from the BB genotype were mapped against the HguttRefA assembly (Red horizontal line, A arrangement), and analysed with *Leviathan* to search for structural variants. A total of 63 long synthetic molecules (coloured segments) were found shared by two regions spanning the inversion breakpoints and separated by 8.2 Mb. Read orientation supported a structural variant of type inversion.

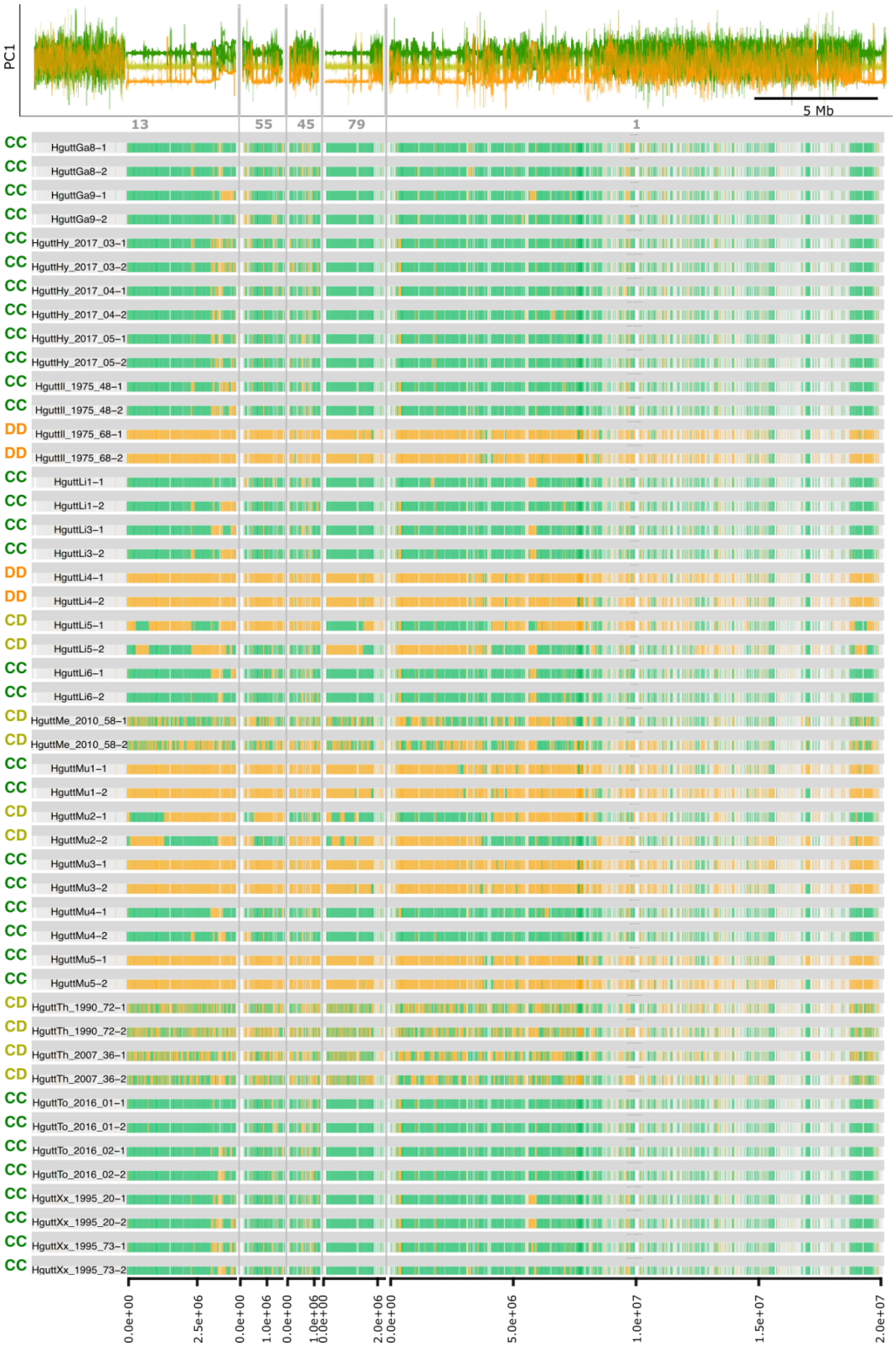


Supplementary Figure S13: *F*ST co-plots between different Mediterranean and Atlantic lineages. (Axis labels and text enlarged)

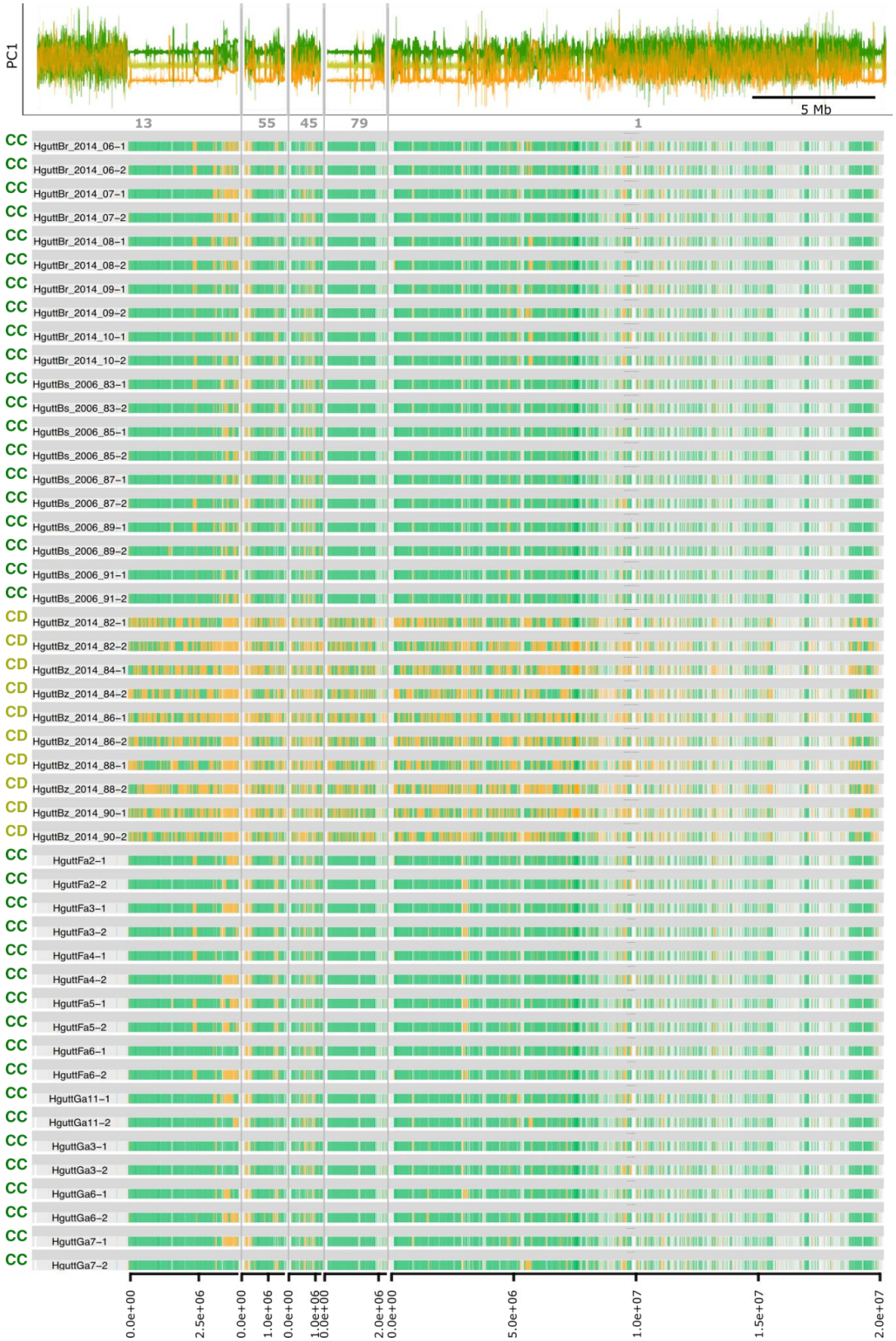


Supplementary Figure S14: Chromosome painting in the inversion region on Chr12 (scaffold 14). Codes indicate inversion genotypes and individual sample names of which the last characters ("-1" or "-2") refer to phased parental haplotypes. Inversion heterokaryotypes show phasing errors in the form of haplotype switches. The top panels show the first principal component of local PCA along the inversion. (Enlarged x-axis)

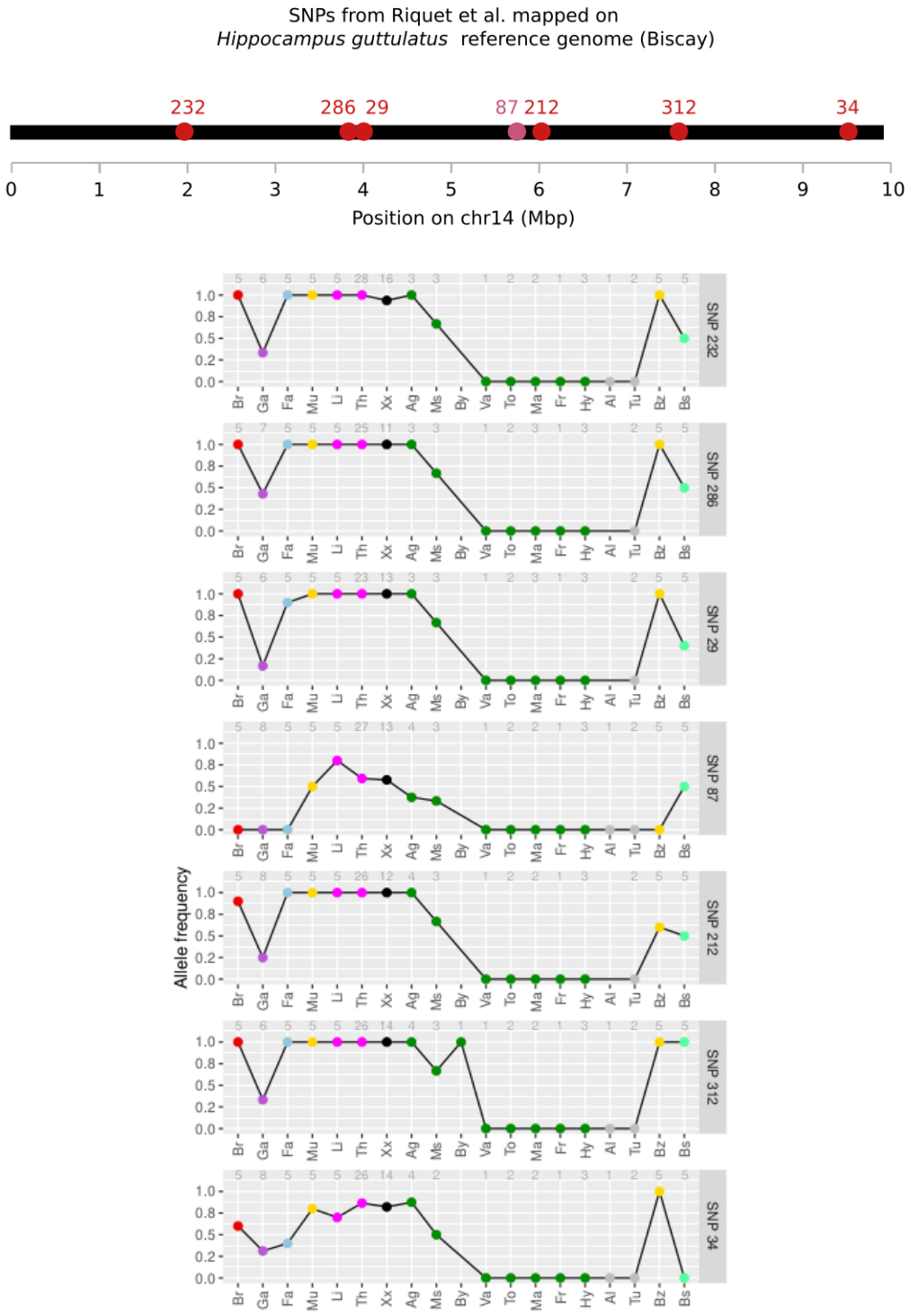
Supplementary Figure S15a: Chromosome painting in the inversion region on Chr2. Grey font indicates scaffold number. Grey areas represent blocks where it was not possible to determine ancestry. (Added x-axis)



Supplementary Figure S15b: Chromosome painting in the inversion region on Chr2 (continued).



Supplementary Figure S16: Outlier SNPs identified in Riquet et al. (2019) mapped onto scaffold 14 (*H. erectus* Chr12) of our reference genome (Hgutt_V1). Allele frequencies calculated from our whole-genome data (n=112) are shown per sampling location. Grey numbers indicate sample sizes.



Supplementary File S2

Scripts and commands:

Reference genome assembly and repeat annotation

Step	Command
nubeam-dedup	#Deduplication of 10x genomics reads nubeam-dedup -i1 Hgutt_S1_L001_R1_001.fastq.gz -i2 Hgutt_S1_L001_R2_001.fastq.gz -o1 Hgutt_DEDUP_S1_L001_R1_001.fastq.gz -o2 Hgutt_DEDUP_S1_L001_R2_001.fastq.gz -z 6 -r 0
process_10xReads	#Process raw 10x genomics reads process_10xReads.py -a -o Hgutt_DEDUP_PROC_S1 -1 Hgutt_DEDUP_S1_L001_R1_001.fastq.gz -2 Hgutt_DEDUP_S1_L001_R2_001.fastq.gz
custom R script	#Identification of rare and over-represented barcodes Rscript --vanilla Autoset_Filters.R Hgutt_DEDUP_PROC_S1_barcodes.txt barocode_list.txt
filter_10XReads	#Filters reads for status and barcodes filter_10xReads.py -m 11 -n 720 -o Hgutt_DEDUP_FILTERED_S1 -B Hgutt_DEDUP_PROC_S1_barcodes.txt -1 Hgutt_DEDUP_PROC_S1_R1_001.fastq.gz -2 Hgutt_DEDUP_PROC_S1_R2_001.fastq.gz
regen_10XReads	#Return reads to origin format regen_10xReads.py -o Hgutt_DEDUP_REGEN_S1 -1 Hgutt_DEDUP_FILTERED_S1_R1_001.fastq.gz -2 Hgutt_DEDUP_FILTERED_S1_R2_001.fastq.gz
Supernova-2.1.1	#Generate whole genome *de novo* assembly supernova run --id Hgutt_DEDUP_FILTERED_S1_all --description DEDUP_FILTERED_S1_all_reads --fastqs /DEDUP_FILTERED -- maxreads=all --accept-extreme-coverage #Generate FASTA output for the assembly supernova mkoutput --style=pseudohap2 -- asmdir=Hgutt_DEDUP_FILTERED_S1_all/outs/assembly -- outprefix=Hippocampus_guttulatus_v1 --minsize=1000 --index --headers=short
BlobToolKit (v3.5.2) BUSCO (v5.4.4)	#Create a new BlobDir dataset blobtools create --fasta Hippocampus_guttulatus_v1.fasta --meta Hippocampus_guttulatus_v1.yaml /DATASETS/Hippocampus_guttulatus_v1 #Run BUSCO analysis busco -i ./Hippocampus_guttulatus_v1.fasta -l actinopterygii -o Hgutt_v1_BUSCO -m geno #Add results of BUSCO analysis to the BlobToolKit dataset blobtools add --busco /Hgutt_v1_BUSCO/run_actinopterygii_odb10/full_table.tsv /DATASETS/Hippocampus_guttulatus_v1

	#View results blobtools view --local /DATASETS/Hippocampus_guttulatus_v1
RepeatModeler2	#Create database for RepeatModeler BuildDatabase -name Hippocampus_guttulatus_v1 Hippocampus_guttulatus_v1.fasta #Run RepeatModeler nohup RepeatModeler -database Hippocampus_guttulatus_v1 -threads 20 - LTRStruct >& run.out &
Tandem Repeats Finder (v4.09.1)	#Run pyTanFinder pyTanFinder.py Hippocampus_guttulatus_v1.fasta -minM 50 -maxM 2000 - minMN 2 -minA 100000 -px HguttV1_Tan -tp ./trf409.linux64
RepeatMasker (v4.0.5)	#Run RepeatMasker RepeatMasker -lib Hippocampus_guttulatus_v1-families.fa -pa 12 -a Hippocampus_guttulatus_v1.fasta #Calculate divergence for each repeat family calcDivergenceFromAlign.pl -s Hippocampus_guttulatus_v1.divsum Hippocampus_guttulatus_v1.fasta.align #Generate Repeat Landscape createRepeatLandscape.pl -div ./Hippocampus_guttulatus_v1.divsum -g 424000000 > ./Hgutt_Repeat_Landscape.html #Softmasking for Braker2 structural annotation RepeatMasker -lib Hippocampus_guttulatus_v1-families.fa -pa 12 -gff -xsmall - a Hippocampus_guttulatus_v1.fasta
trim-galore (v0.6.7-1)	#Trim reads trim_galore --paired RNAseqfile_R1.fq RNAseqfile_R2.fq
HISAT2 (v2.2.1)	#Index the softmasked reference genome hisat2-build Hippocampus_guttulatus_v1.fasta.masked /Hisat2_db/Hippocampus_guttulatus_v1_masked #Align RNA-Seq reads hisat2 -p 20 -x /Hisat2_db/Hippocampus_guttulatus_v1_masked -1 Trimmed_RNAseqfile_1.fq -2 Trimmed_RNAseqfile_2.fq -S aligned.sam
Samtools (v1.12)	#Convert sam to sorted bam samtools sort aligned.sam -o aligned_sorted.bam
Braker2 (v2.1.6)	#Run Braker2 braker.pl --species=Hippocampus_guttulatus -- genome=Hippocampus_guttulatus_v1.fasta.masked -- bam=aligned_1_sorted.bam,aligned_2_sorted.bam,aligned_3_sorted.bam,align ed_4_sorted.bam,aligned_5_sorted.bam,aligned_6_sorted.bam,aligned_7_sort

	ed.bam

Whole-genome sequencing data

Some of the following commands were ran using snakemake (v7.1.1), for which snakefiles can be found at https://github.com/pierrebarry/life_tables_genetic_diversity_marine_fishes.

Step	Command
fastp (v0.23.1)	fastp -i {input.raw_R1} -I {input.raw_R2} -o {output.fastp_R1} -O {output.fastp_R2} --trim_poly_g --correction --low_complexity_filter –html {output.report_html} --json {output.report_json} –report_title {wildcards.sample} --thread $ --dont_overwrite --merge –merged_out {output.merged}
bwa mem (v0.7.17)	#paired reads bwa mem -M -t $ referencegenome_{wildcards.species} {input.fastq_R1} {input.fastq_R2} > {output.align_sam_paired}) #merged reads bwa mem -M -t $ referencegenome_{wildcards.species} {input.fastq_merged} > {output.align_sam_merged}) #combine sam files picard MergeSamFiles I={input.align_sam_paired} I={input.align_sam_merged} O={output.sam_final}
Sam to bam file (picard v2.26.8)	picard SortSam -I {input.align_sam} -O {output.align_bam_picard} -SO coordinate -CREATE_INDEX true -VALIDATION_STRINGENCY LENIENT -TMP_DIR tmp
Mark duplicates (picard v2.26.8)	picard -Xmx$g MarkDuplicates -I {input.align_bam_picard} -O {output.markdup_picard} -ASSUME_SORTED TRUE -REMOVE_DUPLICATES FALSE -CREATE_INDEX TRUE -METRICS_FILE {wildcards.sample}_duplicate_metrics.txt -VALIDATION_STRINGENCY LENIENT -TMP_DIR tmp
Add read group (picard v2.26.8)	picard AddOrReplaceReadGroups -I {input.markdup_picard} -O {output.markdup_rg_picard} -RGPL ILLUMINA -RGLB lib -RGPU genewiz -RGSM {wildcards.sample}
Get stats (samtools v1.10, htslib v1.10.2)	samtools flagstat {input.bamfile} > {output.samtools_flagstat} samtools stats -d {input.bamfile} > {output.samtools_stats}

pmdtools (v0.50)	for i in *.bam do echo pmdtools running on: $i samtools view $i | python /path/pmdtools --deamination > $i_pmdtools.txt done #Plotted with their Rscript plotPMD.R
GATK HaplotypeCaller (GATK 4.1.8.0)	gatk HaplotypeCaller -R {input.reference_genome} -I {input.markdup_rg_picard} -O {output.gvcf_first} -bamout {output.realign_bam} -ERC GVCF -G StandardAnnotation -G AS_StandardAnnotation -G StandardHCAnnotation --tmp-dir tmp
GATK GenomicsDBImport	gatk --java-options '-Xms$g -Xmx$G' GenomicsDBImport --sample-name-map name_Hgutt.txt –genomicsdb-workspace-path {output.joint_genotyping_files} --tmp-dir /home/lmeyer/tmp #name_Hgutt.txt contained sample names and the list of *_gvcf_first.g.vcf files
GATK GenotypeGVCFs	gatk --java-options '-Xms$g -Xmx$G' GenotypeGVCFs -R {input.reference_genome} -V gendb://Joint_Genotyping_{wildcards.interval} -G StandardAnnotation -O {output.joint_gvcf_first}
Merge VCFs (vcftools v0.1.16)	vcf-concat -f /path/list_vcf_files.txt | bgzip -c > {output.vcf_first}
Vcf filtering (bcftools v1.9, vcftools v0.1.16)	#Filter VCF around indels bcftools filter -g 5 --output {output.vcf_indel5bp} {input.vcf} #Keep biallelic snps vcftools --gzvcf {input.vcf_indel5bp} --remove-indels --max-alleles 2 --recode --stdout | bgzip > {output.vcf_indel5bp_snponly} #Filter missing data vcftools --gzvcf {input.vcf_indel5bp_snponly} –max-missing {params.per} --recode --stdout | bgzip > {output.vcf_indel5bp_snponly_missing} #Remove sites with extremely high depth vcftools --exclude-positions {list} –vcfgz {output.vcf_indel5bp_snponly_missing} --recode --recode-INFO-all --out {output.vcf_indel5bp_snponly_missing_depth}
ANGSD (v0.933)	#ANGSD all SNPs angsd -bam {bam_file_list} -GL 2 -trim 5 -setMaxDepth {value} -doMajorMinor 1 -minMapQ 30 -minQ 20 -doMaf 1 -SNP_pval 1e-6 -minInd 45 -doCounts 1 -minMaf 0.05 -doGlf 2 -uniqueonly 1 -remove_bads 1 -C 50 -baq 1 -doCov 1 -doIBS 2 -makeMatrix 1 -ref referencegenome_Hgutt_V1.fa -out {prefix} -P $ 2>log.txt

ANGSD (v0.933)	#ANGSD markers in linkage equilibrium angsd -bam {bam_file_list} -sites {list_sites} -rf chrs.txt -GL 2 -trim 5 -setMaxDepth {value} -doMajorMinor 1 -minMapQ 30 -minQ 20 -doMaf 1 -minInd 40 -doCounts 1 -minMaf 0.05 -doGlf 2 -uniqueonly 1 -remove_bads 1 -C 50 -baq 1 -doCov 1 -doIBS 2 -makeMatrix 1 -ref referencegenome_Hgutt_V1.fa -out {prefix} -P $ 2>log.txt

SNPRelate v1.28.0, SeqVartools v1.38.0, R v4.3.0)	showfile.gds(closeall=TRUE)
	VCF_PATH="/path/file.vcf.gz"
	
	if (file.exists("file.gds")==F){
	vcf.fn <- VCF_PATH
	seqVCF2GDS(vcf.fn, "file.gds")
	}
	
	#OPEN GDS (can start from here)
	genofile <- seqOpen("/path/file.gds")
	
	#Perform PCA
	pca <- snpgdsPCA(genofile, num.thread=5,autosome.only = T, maf=0.05)

lostruct (v0.0.0.9)	snps<-vcf_windower("/path/file.bcf",size=5000,type='bp') pcs <- eigen_windows(snps,k=2) write.table(pcs,"filename.txt")
genomics_general (python v3.9.13)	#Parse VCF and generate .geno format python /path/parseVCF.py -i file.vcf --skipIndels --minQual 30 --gtf flag=DP min=2 max=100 -o file.geno.gz # popgenWindows.py #diversity and divergence along the genome #Example for contrast between pops Mu and Ma python /path/popgenWindows.py --windType coordinate -w 25000 -m 100 -g file.geno.gz -o file.csv.gz -f phased -T $ -p Mu HguttMu1,HguttMu2,HguttMu3,HguttMu4,HguttMu5 -p Ma Hgutt_Hy_2017_03,Hgutt_Hy_2017_04,Hgutt_Hy_2017_05,Hgutt_To_2016_01,Hgutt_To_2016_02 #Scripts available at https://github.com/simonhmartin/genomics_general
Heterozygosity (vcftools v0.1.16)	vcftools --gzvcf file.vcf.gz --het --out prefix
BAMscorer (v1.4)	#select SNPs BAMscorer select_snps file.vcf.gz output_prefix --numchrom $ #Manual step to get 3 files with the individuals of each karyotype

	#{OUT}_AA_individuals.txt #{OUT}_BB_individuals.txt #{OUT}_db_individuals.txt #score bam files BAMscorer score_bams file.vcf.gz output_prefix path_to_bams
Twisst	#Produce a VCF with ANGSD including other spp. angsd -bam bamlist -r $ -ref referencegenome_Hgutt_V1.fa -GL 2 -doPost 1 -doGeno 1 -trim 5 -setMaxDepth {value} -doMajorMinor 1 -minMapQ 30 -minQ 20 -doMaf 1 -minInd 8 -doCounts 1 -doGlf 2 -uniqueonly 1 -remove_bads 1 -C 50 -baq 1 -doIBS 2 -doCov 1 -makeMatrix 1 -doBcf 1 --ignore-RG 0 -out {prefix} -P $ 2>log.txt #Remove lines with heterozygous genotypes bcftools view file.bcf| grep -v "0/1" > file_noHet.vcf #Remove lines where all species are "0/0" bcftools view file_noHet.vcf| grep "1/1" > tmp.vcf bcftools view -h file_noHet.vcf > VCF_header cat VCF_header tmp.vcf > file_noHet_variable.vcf #also added the header rm tmp.vcf #Make it a "phased" VCF bcftools view -H file_noHet_variable.vcf | sed 's/\/|/g' > tmp.vcf cat VCF_header tmp.vcf > file_noHet_variable_phased.vcf #also added the header rm tmp.vcf VCF_header bgzip file_noHet_variable_phased.vcf #Convert to .geno format python /path/parseVCF.py -i file_noHet_variable_phased.vcf.gz --skipIndels --gtf flag=DP min=2 max=100 | gzip > file_noHet_variable_phased.geno.gz #Infer the trees python path/phyml_sliding_windows.py --minPerInd 25 -T $ -g file_noHet_variable_phased.geno.gz --prefix prefix.phyml_bionj.w50 -w 50 --windType sites --model GTR --optimise n #Get topology weights python /path/twisst.py -t prefix.phyml_bionj.w50.trees.gz prefix.phyml_bionj.w50.weights.csv.gz -g A -g B -g C -g D -g E -g F -g G -g H --groupsFile groups.tsv #Scripts can be found at https://github.com/simonhmartin/twisst
SHAPEIT (v4.2.2)	#without gmap for i in $(cat scaffold_list); do shapeit4 \ --input file.bcf \ --region ${i} \ --effective-size X \ --output prefix_phased_${i}.vcf.gz \

	--log phasing.log; done

tsinfer & tsdate	#See the tutorial: https://tskit.dev/tsinfer/docs/stable/tutorial.html
	
	#create samples using tsinfer_create_input_files.py
	#the function add_diploid_sites iterates over the variants in a CYVCF2.VCF object and adds them to a tsinfer sample data file.
	
	#run tsinfer using tsinfer_infer.py
	#The tree object is inferred with the tsinfer.infer function
	
	#run tsdate using tsdate_infer_part1.py
	#Use tsdate.build_prior_grid to specify a prior
	#Run tsdate.date to date the nodes
	
	#extract TMRCA using tsinfer_extract_info_v2.py
	#Randomly sample two leaves in each topology and extract information such as TMRCA
	
	a=1
	block_start=0
	for tree_index in range(1,len(breaks)-2):
	a=a+1
	block_start=block_start+1
	for times in range(1,3):
	inds_list = random.sample(range(96), 2)
	ind1 = inds_list[0]
	ind2 = inds_list[1]
	IND1+=[ind1]
	IND2+=[ind2]
	BLOCK_LENGTH+=[breaks[a]-breaks[block_start]]
	POSITION+=[breaks[a]]
	time=dated_ts.at_index(tree_index).tmrca(ind1,ind2)
	DIVERGENCE+=[time]
	POP1+=[corr[ind1]]
	POP2+=[corr[ind2]]

Analysis of SVs

Step	Command
EMA (v0.6.2)	#Interleave read files parallel -j20 --bar 'paste <(pigz -c -d {} | paste - - - -) <(pigz -c -d {= s:_R1_:_R2_: =} | paste - - - -) | tr "\t" "\n" | ema count -w ./barcode_list.txt -o {/.} 2>{/.}.log' ::: *_R1_*.gz

	#Preprocess 10x data and insert BX:Z tags paste <(pigz -c -d *_R1_*.gz | paste - - - -) <(pigz -c -d *_R2_*.gz | paste - - - -) | tr "\t" "\n" | ema preproc -w ./barcode_list.txt -b -n 500 -t 20 -o output_dir *.ema-ncnt 2>&1 | tee preproc.log #Concatenate files cat ema-bin-* > Hgutt_DEDUP_REGEN_S1_Interleaved_BXed.fastq
LRez (v2.2.4)	#Build barcode index LRez index fastq -f Hgutt_DEDUP_REGEN_S1_Interleaved_BXed.fastq -o Barcode_Index.bci -t 20
BWA (v0.7.17)	#Map linked-reads to the reference genomes Hgutt_V1 and HguttRefA bwa mem -C -t 20 Hippocampus_guttulatus_v1.fasta -p Hgutt_DEDUP_REGEN_S1_Interleaved_BXed.fastq.gz > Hgutt_10XLR_Aligned_V1.sam bwa mem -C -t 20 GCA_025802095.1_ASM2580209v1_genomic.fna -p Hgutt_DEDUP_REGEN_S1_Interleaved_BXed.fastq.gz > Hgutt_10XLR_Aligned_RefA.sam #Sort and convert to bam and make index samtools sort Hgutt_10XLR_Aligned_V1.sam -@ 4 -O bam -l 0 -m 2G -o Hgutt_10XLR_Aligned_V1.bam samtools index -b Hgutt_10XLR_Aligned_V1.bam samtools sort Hgutt_10XLR_Aligned_RefA.sam -@ 4 -O bam -l 0 -m 2G -o Hgutt_10XLR_Aligned_RefA.bam samtools index -b Hgutt_10XLR_Aligned_RefA.bam
MTG-Link	#Generate the input GFA file bed2gfa.py -bed Hgutt_14_1698000_1700000.bed -fa Hippocampus_guttulatus_v1.fasta -out Hgutt_14_1698000_1700000.gfa #Run MTG-Link mtglink.py DBG -gfa Hgutt_14_1698000_1700000.gfa -bam Hgutt_10XLR_Aligned_V1.bam -fastq Hgutt_DEDUP_REGEN_S1_Interleaved_BXed.fastq -index Barcode_Index.bci -k 61 51 41 31 21 -t 20
Minimap2 (v2.26)	# Align assembled contig on the reference genome minimap2 -a Hippocampus_guttulatus_v1.fasta Hgutt_14_1698000_1700000.gfa.14_0-1698000-L+_14_1700000-9878394-R+.g2000.flank10000.occ2.k61.a3.bxu.insertions_filtered_quality.fasta > 3kb_contig.sam
Leviathan (v1.0.2)	#Extract data mapping to Chr4.1 of HguttRefA samtools view -b Hgutt_10XLR_Aligned_RefA.bam JAOYMQ010000004.1 > Hgutt_10XLR_Aligned_RefA_4.1.bam samtools index Hgutt_10XLR_Aligned_RefA_4.1.bam #Build LRez barcode index LRez index bam -p -b Hgutt_10XLR_Aligned_RefA_4.1.bam -o

	Hgutt_10XLR_Aligned_RefA_4.1.bci #Run Leviathan on Chromosome 4.1 LEVIATHAN -b Hgutt_10XLR_Aligned_RefA_4.1.bam -i Hgutt_10XLR_Aligned_RefA_4.1.bci -g GCA_025802095.1_ASM2580209v1_genomic.fna -o Hgutt_10XLR_Aligned_RefA_4.1.vcf

Supplementary Appendix

See the file Meyer_et_al_Results_Appendix.html