Patterns of *hAT* transposable element insertions linked to chromosomal inversions in Atlantic cod (*Gadus morhua*)

Robin Aasegg Araya

Master Thesis

Ecology and Evolution

30 credits

Centre for Ecological and Evolutionary Synthesis

Department of Biosciences

Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

December 2022

2022

Patterns of *hAT* transposable element insertions linked to chromosomal inversions in Atlantic cod (*Gadus morhua*)

Robin Aasegg Araya

http://www.duo.uio.no/

Print: Reprosentralen, Universitetet i Oslo

Acknowledgements

To my supervisors, Kjetill Sigurd Jakobsen, Ole Kristian Tørresen and William Brynildsen Reinar. Thank you for sharing your knowledge, your support, and your invaluable guidance. Kjetill, thank you for letting me become a part of the genomics group at CEES. Your ideas, enthusiasm and valuable insight have been inspiring throughout this project. Thank you, Ole, for always being helpful with analyses and addressing essential questions along the way that have helped me keep on track. William, thank you for your patience and guidance. You have always been open to discussions and brainstorming. I appreciate all your time helping me with bioinformatics and figuring out each step in the process.

Thank you, Marius Maurstad and José Cerca, for being available to chat about the project, sharing your input and bringing creative ideas to the analyses. Thank you to everyone I have gotten to know at the Centre for Ecological and Evolutionary Synthesis (CEES) for all the fun coffee breaks and for making me feel included. I highly value the time I have spent here.

To Ingrid, your patience, unconditional support, and encouragement mean everything to me.

Thank you, Bændik Ola, for day-to-day talk about nonsense, without exception (I think), during the last two months. And, of course, Sigurd for providing care packages during the most challenging times. To my family, thank you for always supporting me, and to my friends, thank you for (mostly) always supporting me.

Lastly, I would like to thank Marine S. O. Brieuc, Helle T. Baalsrud, and Tara J. Daughton for providing information on inversion breakpoints in the NEAC and NCC assemblies and Siv Hoff for providing the NCC assembly.

Table of Contents

Abstract

Chromosomal inversions control complex phenotypes by capturing co-adapted alleles and maintaining stable polymorphisms in interbreeding populations. In the Atlantic cod (*Gadus morhua*, L. 1758), four large-scale inversions on linkage groups (LGs) 1, 2, 7 and 12 have been linked to local adaptation and behavioural ecotypes. LG1 and LG7 are most prominent in separating the Northeast Arctic cod (NEAC) and Norwegian coastal cod (NCC) ecotypes. Here, using the current reference assemblies of the NEAC, NCC and Celtic cod genomes, we find closely related TEs in the corresponding breakpoints for each LG, respectively, and only for the inverted (derived) haplotypes on LG1 and LG7 in NEAC and LG12 in NCC. This is consistent with a scenario where the inversions arose due to the activity of transposable elements (TEs). In particular, multiple copies of a DNA TE named *hAT*-1784 have inserted into breakpoints of the LG1 and LG7 inversions in NEAC. Our data support independent insertions of different members of the *hAT*-1784 in each of the LG1 and LG7 inversions during two major TE expansions within the cod genome. We find the *hAT*-1784 TEs in breakpoints to be long (mean length above 2,000 bp) and reversely oriented, further suggesting that the LG1 and LG7 inversions originated by ectopic recombination between *hAT* elements.

Introduction

Two ecotypes of Atlantic cod (*Gadus morhua*, L. 1758), the Northeast Arctic cod (NEAC) and the Norwegian coastal cod (NCC), exhibit several fundamental differences in behavioural and life-history traits (Rollefsen, 1953). NEAC migrates from the feeding grounds in the Barents Sea to the Lofoten islands and other parts of the Northern Norway coast for spawning, while NCC spawns and feeds along most of the Norwegian coast, including Lofoten, without performing any long-distance migrations (Brander, 2005) (**Figure 1a**; Materials and Methods). Although both ecotypes share spawning grounds, early genetic analyses have suggested that NEAC and NCC show some, but not large, overall population structuring (Reiss et al., 2009). However, the structuring is most profoundly evident in four large-scale chromosomal inversions located on linkage groups (LGs) 1, 2, 7 and 12 (Berg et al., 2016, 2017; Matschiner et al., 2022). These inversions are also present in the Baltic, North Sea, Icelandic, coastal Norwegian and Canadian populations (Berg et al., 2015, 2016; Kirubakaran et al., 2016; Sodeland et al., 2016; Barth et al., 2017; Berg et al., 2017; Matschiner et al., 2022), and are linked to adaptation to environmental conditions (temperature, light conditions, oxygen, salinity (Berg et al., 2015)) and behavioural and reproductive traits (Kirubakaran et al., 2016).

In NEAC and NCC, the inversions on LG1 and LG7 have been suggested to be associated with their distinct behavioural and reproductive adaptations (Hemmer-Hansen et al., 2013; Berg et al., 2016, 2017; Matschiner et al., 2022). The inversion on LG1, a double inversion, is believed to significantly contribute to generating the different migratory behaviours between NEAC and NCC (Berg et al., 2016, 2017; Matschiner et al. 2022). However, it is currently unknown to what extent the other inversions, including LG7, also play a role here (Matschiner et al. 2022).

A chromosomal inversion occurs when a region on a chromosome is flipped and placed in a reverse orientation compared to the rest of the chromosome (**Figure 1b**; Materials and Methods). Inversions create regions with suppressed recombination, both by reducing the number of homologous sites for recombination in the inverted region and by affecting fitness of individuals carrying the inverted haplotype. In heterozygote carriers, crossover between an inverted and non-inverted haplotype may result in non-viable gametes due to the production of duplication- and deletion products, reducing recombination (reviewed in Wellenreuther & Bernatchez, 2018; Villoutreix et al., 2021). Recombination is also reduced for individuals carrying two copies (homozygotes) of the derived, inverted haplotype. This is mainly due to the selection against breaking up the linkage of beneficial allele combinations captured by the inversion (reviewed in Schwander et al., 2014), but also because the origin of an inversion is expected to be equal to a severe bottleneck, since the initial frequency of an inversion in the population is, by definition, low. Thus, inversions provide a mechanism for maintaining large, stable polymorphisms within panmictic populations (Kapun et al., 2016; Lamichhaney et al., 2016; Matschiner et al., 2022). Double crossover and gene conversion can still break up associations within an inversion (Stump et al., 2007; Matschiner et al., 2022). However, recombination rates tend to decrease near breakpoints (Stump et al., 2007).

In Atlantic cod, studies have identified large regions in LGs 1, 2, 7 and 12 (4-17 Mb) of strongly linked SNPs (Berg et al., 2015, 2016; Matschiner et al., 2022), and these regions have been confirmed to represent four distinct chromosomal inversions (Berg et al., 2016; Kirubakaran et al., 2016; Sodeland et al., 2016; Berg et al., 2017; Kirubakaran et al., 2020; Matschiner et al., 2022). Matschiner et al. (2022) mapped these four inversions onto prior assemblies of the NEAC and NCC genomes, suggesting that NEAC carries the derived and inverted haplotypes on LGs 1 and 7 while NCC carries the ancestral and non-inverted haplotypes (**Figure 1c;** Materials and Methods**)**. In contrast, the NCC reference genome carries the derived (inverted) arrangements on LGs 2 and 12, while NEAC has the ancestral

arrangement (Matschiner et al., 2022). Comparison with the reference genome of the North Sea (Celtic Sea) cod reference shows that Celtic cod carries the same arrangements as NCC in all four inversion haplotypes (Kirubakaran et al., 2020).

The differentiation of inversion haplotypes in the Northeast Atlantic cod is also displayed in ecotype divergence in the Northwest Atlantic cod, suggesting all four inversions occurred before their split >100,000 years ago (Berg et al., 2017). This has been confirmed by recent phylogenomic analyses, with age estimations ranging from ~0.40 million years ago (Ma) for the youngest inversion in LG12, to ~1.66 Ma for the oldest inversion on LG7 (Matschiner et al., 2022). A remaining key question is how the inversions in Atlantic cod arose.

There are a few molecular mechanisms that can cause an inversion, including (i) non-homologous end joining (NHEJ) (Ranz et al., 2007) and (ii) ectopic recombination, i.e., recombination between homologous sequences that are not at the same position on homologous chromosomes (Ling & Cordaux, 2010; Harringmeyer & Hoekstra, 2022). Additionally, inverted segments may enter a population by introgression (Jay et al., 2018). However, in Atlantic cod, none of the four inversions appear to have introgressed (Matschiner et al., 2022), and non-homologous end joining (i) seems to have a low frequency. Thus, ectopic recombination may be a plausible mechanism.

If ectopic recombination is the inversion mechanism in Atlantic cod, one would expect to find homologous sequences in the breakpoints of the derived, inverted haplotypes. Indeed, inversions tend to display high frequencies of repetitive DNA around breakpoints, and several studies have linked genomic spreading of transposable elements (TEs) with inversions (Caceres et al., 1999; reviewed in Böhne et al., 2008; Lamichhaney et al., 2016; Sharma & Peterson, 2022). TEs are mobile genetic elements, and closely related TEs share sequence similarities and structures (e.g., terminal inverted repeats [TIRs] and long terminal repeats [LTRs]). Thus, crossover events between remote TEs or TE fragments can create genomic rearrangements. For instance, TE activity has induced complex chromosomal rearrangements in strains of maize appearing after only five generations (Sharma & Peterson, 2022). Furthermore, TEs contribute to genome size variations in teleost fishes (Auvinet et al., 2020; Symonová & Suh, 2019) and have been associated with inversions in maize (Zhang et al., 2014), Drosophila (Cáceres et al., 1999), yeast (Sarilar et al., 2014), and indirectly in deer mice through detection of inverted repeats (Harringmeyer & Hoekstra, 2022).

TEs often occur in high copy numbers throughout the genome. Thus, multiple identical or near identical TEs exist in different genomic locations at any time and may facilitate a recombination event. However, depending on the TE type, genomic regions with high TE densities tend to correlate negatively with recombination rates (Kent et al., 2017; Peona, Palacios-Gimenez, et al., 2021). TE accumulation can further reduce recombination rates by heterochromatin formation and negative GC bias caused by methylation-spreading (reviewed in Kent et al., 2017; Symonová & Suh, 2019). Hence, to infer a causal relationship of TEs and Atlantic cod inversions, it is critical to examine not only the TE content of breakpoints, but also the general sequence characteristics of breakpoints.

Here, we investigate the TE content in the four chromosomal inversions of LGs 1, 2, 7, and 12 in Atlantic cod to address the plausibility of TEs being causative agents of the inversions. We have undertaken a strategy of investigating the content of related TEs and general sequence characteristics (including GC content and nucleotide statistics) of inversion breakpoints to explore the origins of chromosomal inversions. Our analyses reveal that multiple elements of a DNA *hAT* family, a subgroup of TIR TEs, have transposed into the breakpoints of the derived inversion haplotypes in LGs 1 and 7 in NEAC during two major TE expansions. We find closely related TEs within breakpoints of all derived inversion haplotypes, in support of the hypothesis that TEs have contributed to the inversion origins in Atlantic cod.

Materials and Methods

Data

In this study, we used genome assemblies from three ecotypes of Atlantic cod, namely the Northeast Arctic cod (NEAC) (NCBI accession ID: GCA_902167405.1), Norwegian coastal cod (NCC, Brieuc et al., in prep) and Celtic cod (NCBI accession ID: GCA_010882105.1). The three Atlantic cod assemblies were produced by long-read sequencing to reference-quality (Kirubakaran et al., 2020; Brieuc et al., in prep). Reference-quality assemblies are well-suited to study TEs and other repetitive DNA (Peona et al., 2021).

Repeat annotation

To best capture the TE makeup in Atlantic cod, we produced a *de novo* repeat library with RepeatModeler2 to obtain a species-specific library of TEs for Atlantic cod (Flynn et al., 2020).

RepeatModeler identifies TEs by counting high-frequency short regions of homology (i.e., interspersed repeats) (Price et al., 2005) and by a pairwise alignment approach of putatively related TEs (Bao & Eddy, 2002). We ran RepeatModeler v 2.0.1 on the gadMor3.0 assembly and included the built-in LTR structural discovery pipeline (Ellinghaus et al., 2008; Ou et al., 2018) to increase LTR detection. The final library was classified using the Dfam database v 3.1 (Storer et al., 2021) and was used to mask the NEAC, NCC and Celtic cod assemblies with RepeatMasker v 4.1.2-p1.

Filtering and grouping of data

We used the annotation table from the RepeatMasker output in TE analyses of the Atlantic cod assemblies. Annotation files were cleaned and grouped according to the hierarchical classification scheme proposed by Wicker et al. (2007). Repeats on unplaced scaffolds were filtered out. We only included interspersed repeats in further analyses, and TE families were named by combining the superfamily name and unique discovery number from RepeatModeler.

Determining inversion breakpoints in NEAC and NCC

Approximate inversion breakpoint coordinates were previously decided for LGs 1, 2 and 7 of the NEAC and NCC assemblies (Brieuc et al., in prep), using a similar approach as Matschiner et al. (2022) by aligning contigs and investigating where contigs split in two. The breakpoints on LG2 have been confirmed with HiFi-sequencing reads mapped onto the reference genomes (Daughton et al., in prep.). The breakpoints in LG12 were not known for either assembly. **Figure 1c** illustrates the NEAC and NCC inversion haplotypes with breakpoints and breakpoint labelling conventions that are used throughout this paper.

Alignments of PacBio reads and NCC scaffolds show that an assembly error had flipped the ~12.5-16.3Mb region on LG7 in NCC, overlapping a breakpoint (Brieuc et al., in prep). We confirmed the LG7 assembly error with the Celtic cod assembly using dot-plots (Dgenies v.1.4 (Cabanettes & Klopp, 2018)) since the haplotype of NCC and Celtic cod is expected to be similarly oriented. To determine the breakpoints of LG12 and the missing LG7 breakpoint (A_{NCC}), we first used SAMtools v.1.9 with faidx to obtain the flipped sequence on LG7 in NCC (12.501.067-16.288.916 bp). The sequence was reverse-complemented and merged back with

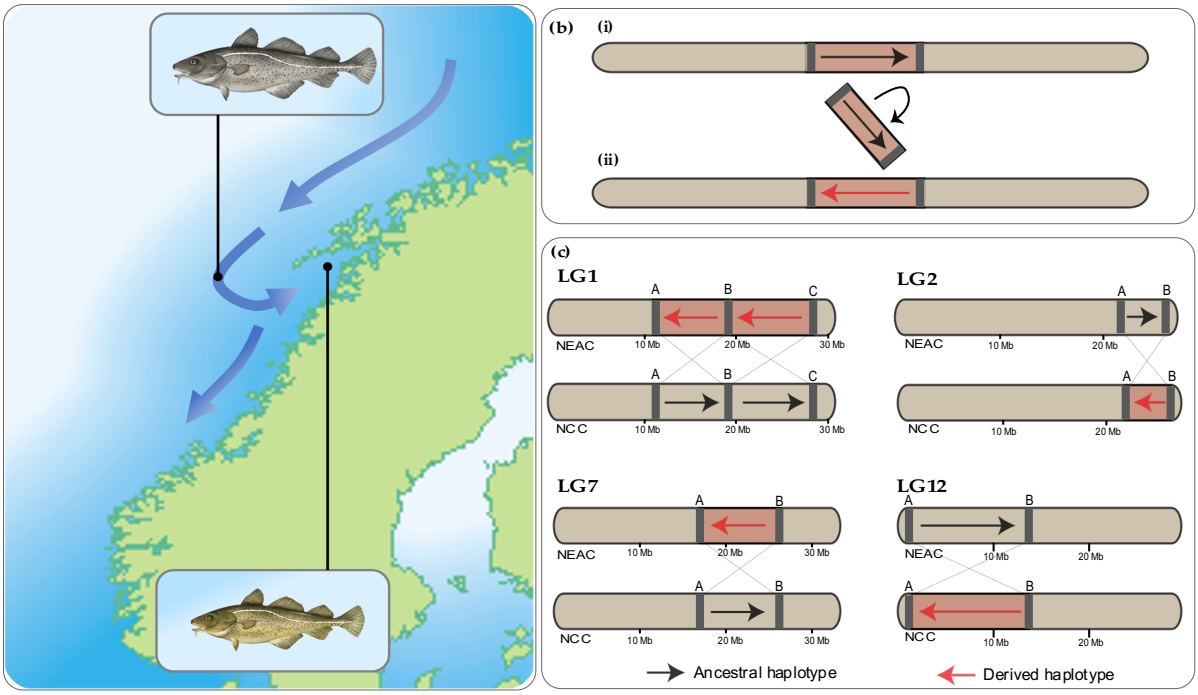


Figure 1: a, Map of the shared spawning site of NEAC and NCC off the Lofoten islands in Norway. The migratory pattern of NEAC is shown with arrows, adapted from Matschiner et al. (2022). Map source: Aotearoa (CC BY-SA 3.0). Fish images was retrieved from public domain. **b,** Illustration of a chromosomal inversion. (i) Black arrows indicate ancestral arrangement, while (ii) the flipped, reverse segment with a red arrow indicates the derived arrangement. **c,** Inversion haplotypes in Atlantic cod on LGs 1, 2, 7 and 12 (Kirubakaran et al., 2020; Matschiner et al., 2022; Brieuc et al., in prep). Breakpoints of inversions are labelled A/B from left to right for NEAC and NCC The inversion on LG1 is a double inversion; hence breakpoints are labelled A/B/C. All breakpoints will be applied the ecotype name in subscript to distinguish between NEAC and NCC. Grey diagonal lines illustrate synteny between breakpoints. NEAC carry the derived, inverted haplotypes on LGs 1 and 7, while NCC carries the derived, inverted haplotypes on LGs 2 and 12 (Matschiner et al., 2022).

LG7. Second, we dot-plotted LGs 7 and 12 in NEAC against NCC with the D-GENIES v.1.4 web application (Cabanettes & Klopp, 2018), similar to the procedure in Kirubakaran et al. (2020). D-GENIES was run with default settings. We cross-referenced the NCC breakpoints with LGs 7 and 12 in the Celtic cod assembly to evaluate the breakpoint coordinates with dot-plots. This revealed another possible assembly error in NCC on LG12 in the segment 13.947.727-15.505.346 bp. Hence, this region was reverse-complemented using the same procedure as in LG7. From the final dot-plots, we characterised each breakpoint using the innermost coordinate of the inversion breakpoint. Finally, we aligned and double-checked breakpoint estimates in Brieuc et al. (in prep) on LGs 1 and 2 in NEAC and NCC with D-GENIES.

Aligning inversion breakpoints to investigate TE content

Breakpoints of each inversion were extended 50 kb up- and downstream and aligned using D-GENIES v.1.4 with default settings to evaluate sequence similarity. If TEs were to associate with an inversion, we expected the breakpoints to align and to find related TE copies within the aligned regions. To limit the investigation to a reasonable number of candidate sequences, we applied a cut-off value of 0.25 sequence identity in keeping alignments for further investigation. Alignments were investigated for TEs of shared family origin, using corresponding coordinates of annotated repeats from the filtered annotation table (see *Filtering and grouping of data*), and ignored TE fragments of <100 bp that we regarded less likely to facilitate ectopic recombination.

Classification of unclassified repeats in breakpoints

We attempted to characterise unclassified TEs residing within both breakpoints of an inversion using TE-Aid (Goubert et al., 2022). TE-Aid is a pipeline that assists in manually curating putative TEs. TE-Aid was run on the unclassified TE consensus sequences, providing information on (i) BLAST hits against the gadMor3.0 reference genome, (ii) coverage depth of BLAST hits, (iii) structural characteristics (e.g., repeated sequences such as TIRs and LTRs), and (iv) putative ORFs and coding TE proteins using the protein database from RepeatMasker, containing 18,011 predicted TE proteins. Additionally, we scanned the unclassified consensus sequences for ORFs using 'transeq' from EMBOSS (Rice et al., 2000) and screened hits for known TE proteins using the Pfam database (Mistry et al., 2021).

Summary statistics on TEs in breakpoints

To perform an in-depth description of breakpoint characteristics and its TE content, we performed three distinct statistical analyses on all inversion breakpoints. First, we plotted mean TE density and frequency on LGs 1, 2, 7 and 12 using non-overlapping sliding windows of 50 kb and comparing each measure to the chromosomal means. Second, we measured the GC content in all inversion breakpoints (+/- 50 kb) using non-overlapping sliding windows of sizes 500 bp and 1 kb, comparing the breakpoint GC content to a genome-wide percentage of GC. Lastly, we statistically evaluated the chance of finding the same TE families within both breakpoints of an inversion. We estimated the mean discovery count of related TEs in randomly selected regions on LGs 1, 2, 7 and 12, iteratively increasing the threshold for TE length and

window size for TE discovery. Each simulation was run 1000 times on each parameter. For each of the four LGs, the TE discovery count of increasing TE lengths (0 bp, 100 bp, 500 bp, 1000 bp, 2000 bp) within the different window sizes (10 – 100 kb) in randomly selected LG tracts was compared to the true count in the respective inversion breakpoints.

Phylogenetic analyses of the hAT-1784 family in multiple breakpoints

We investigated the evolutionary relationship of TEs of the DNA *hAT* family named *hAT*-1784 in Atlantic cod. This TE family appeared in multiple inversion breakpoints, and we performed phylogenetic analyses of *hAT*-1784 using two different phylogenetic approaches. First, we performed phylogenetic tree inference of *hAT*-1784 elements from NEAC, NCC and Celtic cod. To carry out the analysis, we aligned all *hAT* elements (473 single-copies in total) using MAFFT v.7.508. The resulting multiple sequence alignment (MSA) was trimmed with Gblocks v.0.91b to retain semi-conserved regions and remove the shortest *hAT* copies with no alignment. Phylogenetic tree inference was performed with IQ-TREE 2 (Minh et al., 2020), producing an unrooted tree for the *hAT*-1784 family. IQ-TREE uses maximum likelihood (ML) for phylogenetic inference and was run with default settings, and we obtained branch supports using the implemented ultrafast bootstrap (UFBoot2) with 1000 replicates (Hoang et al., 2018). Second, we analysed all 473 single-copy *hAT*-1784 elements with OrthoFinder to infer homology relationships between family members. OrthoFinder infers orthogroups, i.e., clusters of sequences descended from a common ancestor, applying an all-vs-all pairwise alignment and clustering approach based on sequence similarity (Emms & Kelly, 2015). It identifies orthogroups independent of sequence length and phylogenetic distance and attempts to generate gene trees within orthogroups (Emms & Kelly, 2019). OrthoFinder was run with default settings for nucleotide sequences.

Results

Broad-scale similarity in TEs and tandem repeat content of Atlantic cod ecotypes

To allow a comparison of repetitive DNA in NEAC, NCC and Celtic cod ecotypes, we first created a species-specific repeat library of the *G. morhua* reference (NEAC) genome (gadMor3.0) with RepeatModeler2 (Flynn et al., 2020). The final library masked 37.2% of gadMor3.0, of which 29.4% was interspersed repeats (**Supplementary Table 1**). This indicates a more complete repeat library and higher resolution in repetitive regions in gadMor3.0, compared to the previous Atlantic cod assembly, gadMor2, which had 22.9% interspersed

repeats (Tørresen et al., 2017). To compare the composition and age distribution of TEs in NEAC, NCC and Celtic cod genomes, the repeat library was used to mask the NCC and Celtic cod assembly (**Supplementary Tables 2-3**). A broad comparison of the annotation efforts shows similar TE make-up in the NEAC (gadMor3.0), NCC and Celtic cod assemblies (**Supplementary Figure 1**).

Deciding on breakpoint coordinates in NEAC and NCC

Breakpoint coordinates have previously been estimated in NEAC and NCC by alignment of NEAC PacBio reads and NCC scaffolds (Brieuc et al., in prep.), but the breakpoints in LG12 and LG7 (A_{NCC}) were either missing or misplaced. To allow for breakpoint comparison between the cod ecotypes, we dot-plotted LGs 7 and 12 in NEAC against the homologous LGs in NCC with D-GENIES and cross-referenced with homologous LGs in the Celtic cod assembly that share inversion haplotypes with NCC. By dot-plotting LG12 of NEAC against LG12 in NCC, we decided the inversion to be between ~0.5-13.7Mb, corresponding to ~0.7-13.6Mb in NCC. By dot-plotting LG7, the correct coordinate for A_{NCC} was found to be ~15.5Mb (**Figure 2** and **Supplementary Table 4**). The same approach was used to validate estimates in Brieuc et al. (in prep.), which confirmed the breakpoint coordinates on LGs 1 and 2, and LG 7 (A_{NEAC}, B_{NEAC} and B_{NCC}) (**Figure 2** and **Supplementary Table 4**).

TE-density, frequency, and GC-content of inversion breakpoints of NEAC and NCC

Our TE-density scans revealed TE-density and frequency levels above LG-specific averages in breakpoints of the derived haplotypes on LG1, LG7 and LG12 (B_{NCC}), in addition to breakpoints of the ancestral haplotypes on LG1 C_{NCC}, LG7 B_{NCC} and LG12 A_{NEAC} and B_{NEAC} (**Figure 3**, frequency plot in **Supplementary Figure 2**). However, overall GC content in breakpoints did not show any significant deviations compared to the genome-wide background for NEAC nor NCC, except for regions with low %GC around the LG1 B_{NEAC} and LG7 A_{NEAC} breakpoints (**Supplementary Figures 3-6**). High TE densities and low %GC implies lower recombination rates around these two breakpoints. Furthermore, reduced %GC can indicate older TE age distributions around the breakpoints. Young TEs are often silenced upon arrival by host silencing mechanisms, regularly leading to local methylation spreading and subsequent cytosine conversion (Fryxell & Zuckerkandl, 2000). Thus, the low %GC in LG1 B_{NEAC} and LG7 A_{NEAC} may indicate an accumulation of older TEs in these breakpoints.

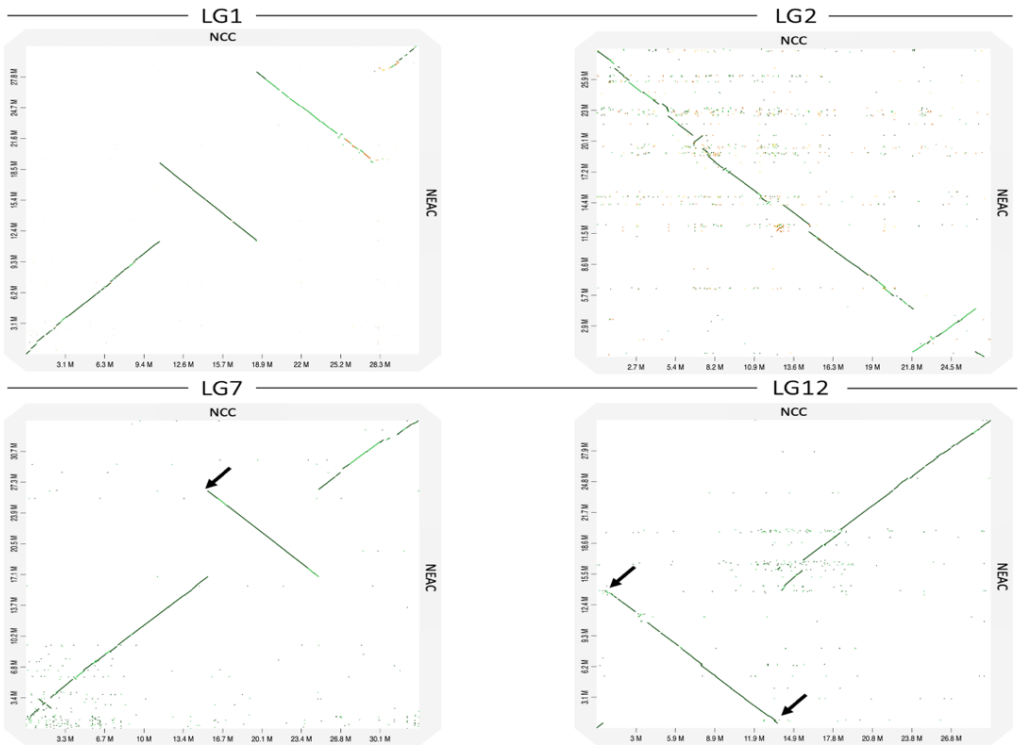


Figure 2: Alignments of LG1, 2, 7 and 12 of NCC on the x-axes and NEAC on the y-axes. Inversion breakpoints on LG7 (A_{NCC}) and LG12 (A_{NEAC}, B_{NEAC}, A_{NCC} and B_{NCC}) (black arrows), were estimated by dot-plotting LGs 7 and 12 of NEAC against NCC homologs, where the innermost coordinate of each breakpoint was chosen.

Multiple copies of a hAT-family reside in inversion breakpoints on LG1 and LG7

To investigate the hypothesis that the Atlantic cod inversions originated by ectopic recombination of TEs, we evaluated sequence similarity by aligning the breakpoints of respective inversion haplotypes (+/- 50 kb) in NEAC and NCC with D-GENIES v.1.4 (Cabanettes & Klopp, 2018) (**Supplementary Figures 7-8**). Alignments with sequence identity >0.25 were screened for related TEs. Strikingly, we found multiple copies of a family of DNA *hAT* elements residing in the breakpoints of both the derived LG1 (B_{NEAC}/C_{NEAC}) and LG7 inversions, named *hAT*-1784. Single-copies of *hAT*-1784 display highly repetitive patterns when aligned (**Figure 4a**), and they appear in both reverse and direct orientation within the different breakpoints of the derived LG1 and LG7 haplotypes (**Figure 4b**). Moreover, we find that the LG1 and LG7 breakpoints contain relatively long *hAT*-1784 elements (mean length of 2042 bp and 3514 bp, respectively), often alongside or overlapping one or more fragmented elements. The *hAT*-1784 family also appear within the non-inverted LG1 A_{NCC} breakpoint. However, it was not found in LG1 B_{NCC}, C_{NCC}, nor any of the LG7 breakpoints of NCC, all of which are the syntenic breakpoint regions of where we find *hAT* elements in the NEAC inversions.

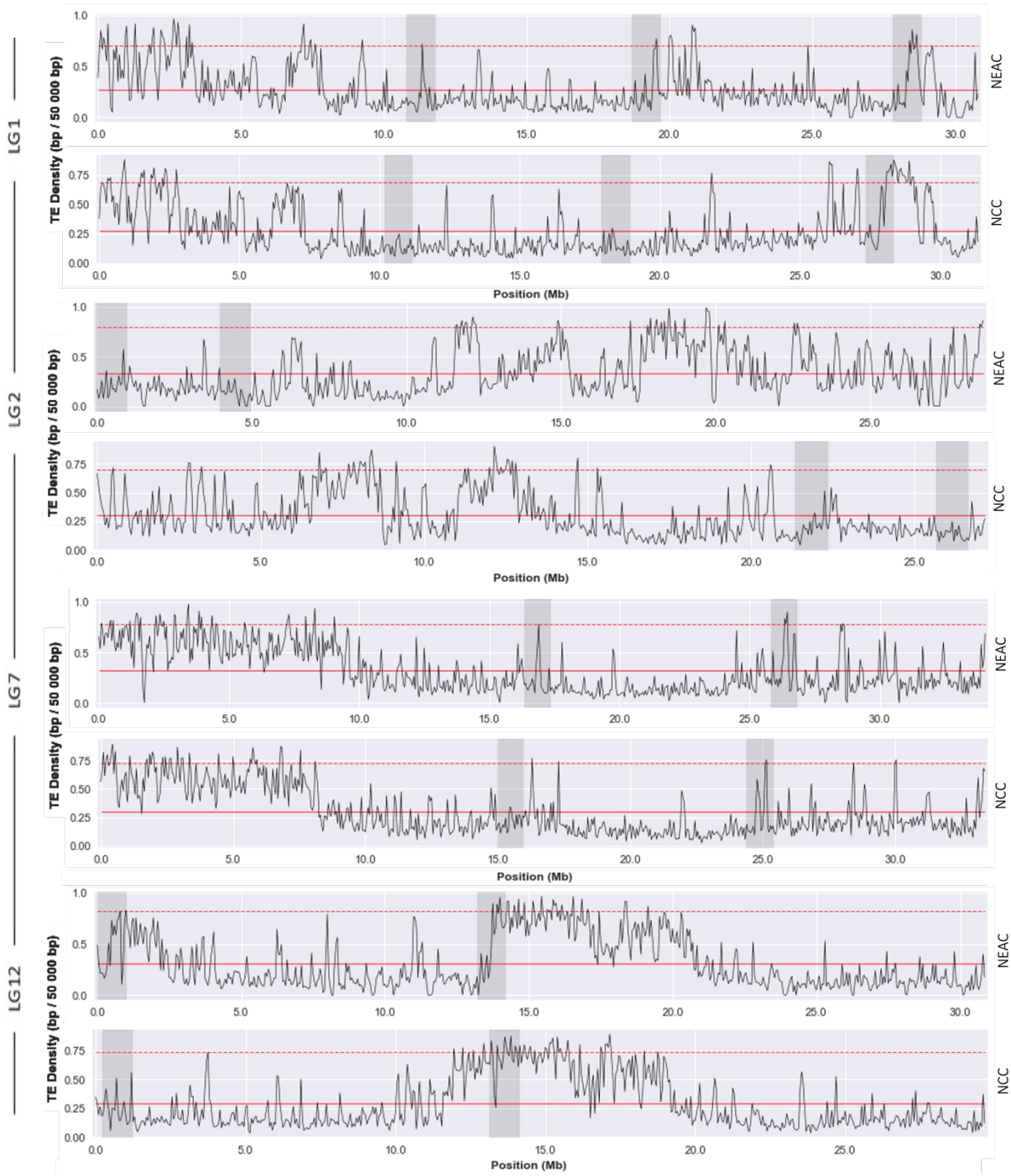


Figure 3: Mean TE densities in LGs 1, 2, 7 and 12 in NEAC (top) and NCC (bottom). Here, we plot density (TEs per bp) per non-overlapping 50 kb sliding window in all four LGs. The red solid line shows mean TE densities across the specific LG, the red dashed line indicates 2 standard deviations from the mean, and the grey shaded areas denote the breakpoints +/- 500 kb. LGs 1 and 7 are derived haplotypes in NEAC. LGs 2 and 12 are derived haplotypes in NCC. Note that the difference in the LG2 position in NEAC and NCC is due to the LG2 being placed in the opposite direction in the gadMor3.0 assembly, relative to NCC.

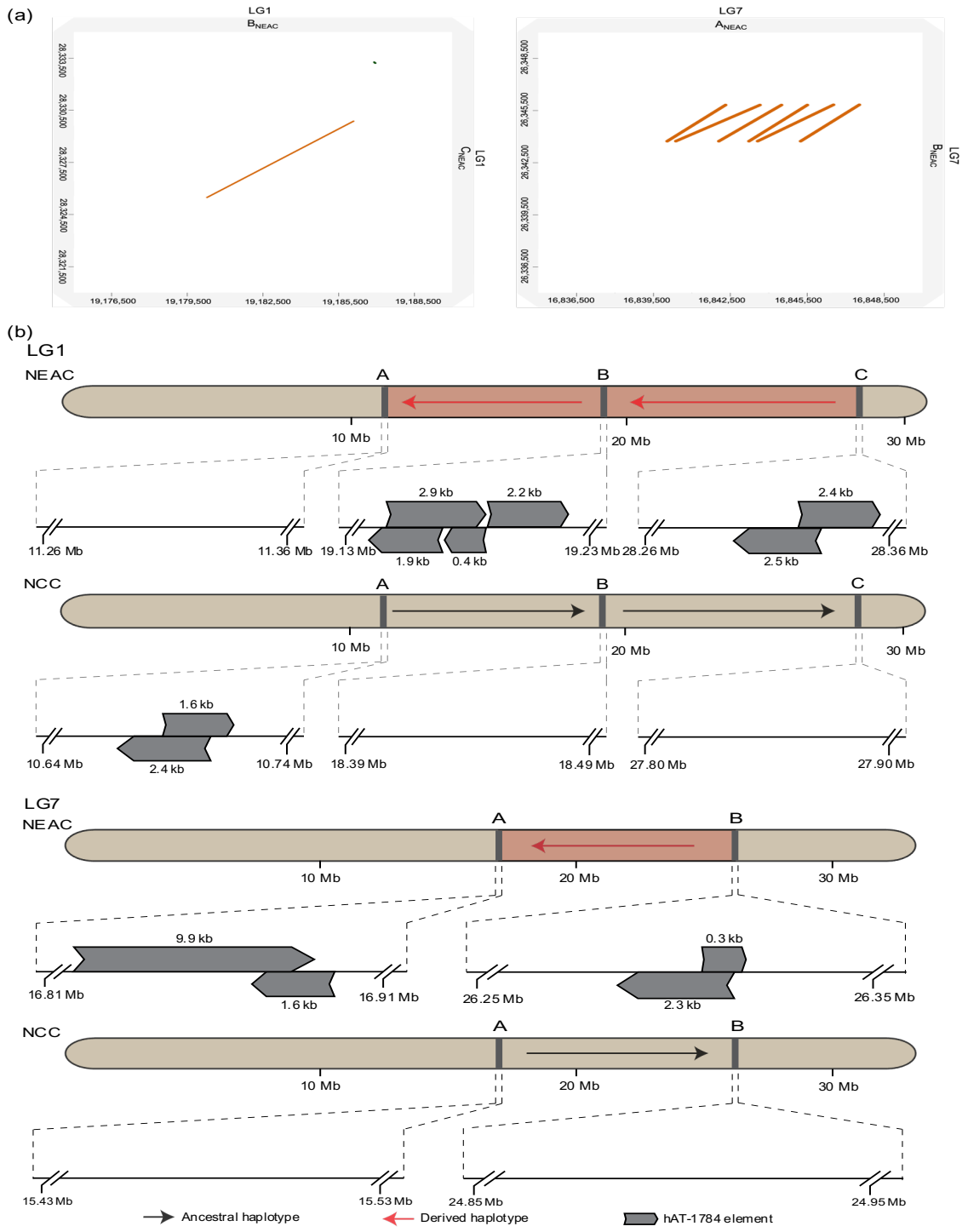


Figure 4: a, Dot-plot alignments of breakpoints in the derived haplotypes on LG1 (B_{NEAC} vs C_{NEAC}) and LG7 (A_{NEAC} vs B_{NEAC}) to visualise sequence similarity between the inversion breakpoints of each of the two inversions. The lines represent sequence similarity between the LG1 breakpoint sequences and LG7 breakpoint sequences (taken from the full alignment in **Supplementary Figure 7 b, d**). Alignments display ~0.5 sequence identity and consist of *hAT*-1784 elements. **b,** Genomic locations of *hAT*-1784 elements in breakpoints of LG1 and LG7. LGs are illustrated in brown, inverted or non-inverted orientations are shown by the black and red arrows, respectively. *hAT* elements are shown as grey bars with its copy length (in kb), where the direction indicates strand orientation (direct or reverse). Breakpoints are zoomed in within a 100 kb window, showing only regions that contain *hAT*-elements in either haplotype.

Closely related TEs reside in breakpoints of all four, derived haplotypes

We applied the same approach as described for identifying the *hAT*-1784 elements, to all inversion breakpoints. Our analyses revealed the presence of related TEs of shared family origin within both breakpoints in all four derived, inverted haplotypes (**Table 1**). Most of these TE families consist of several hundred genome copies (see the genomic distribution in **Supplementary Figure 9**), and they show diverse consensus lengths ranging from ~400 to 4400 bp (for an overview of single-copies, see **Supplementary Table 5**). We found TE families exclusively residing in the breakpoints of the derived haplotypes of LGs 1, 7 and 12, and not in the ancestral haplotypes. In contrast, the TE families in the derived LG2 breakpoints were also found in the breakpoints of the ancestral arrangement. Furthermore, the syntenic breakpoints of the ancestral LG2 arrangement hold the highest abundance of different TE-families, of which the majority of families are LTRs.

Unclassified breakpoint TEs were scrutinized with TE-Aid (Goubert et al., 2022). This revealed the presence of long terminal repeats in the Unc-41 family residing in the ancestral, non-inverted LG2 haplotype, suggesting that it might be a LTR retrotransposon (**Supplementary Figure 10**). No further classification was possible for the remaining breakpoint TEs (see **Supplementary Figures 11-14**).

Table 1 – TE families present in both breakpoints of the same inversion haplotype. Classification, whole genome copy number, breakpoint locations and length of consensus are provided in discrete columns.

LG	**Transposable Element**			**Genome copies**		**Breakpoint locations**		**Length of consensus**	**Relative orientation[a]**
	Order	**Superfamily**	**Family**	**NEAC**	**NCC**	**NEAC**	**NCC**		
1	TIR	*hAT*	*hAT*-1784	166	159	B,C*	A	1723 bp	Dir / Inv
	Unc	Unc	Unc-753	273	282	B,C*	-	467 bp	Inv
	Unc	Unc	Unc-915	161	159	A,C*	-	434 bp	Inv
2	LTR	Gypsy	Gypsy-428	56	52	A,B	A,B*	1817 bp	Dir
	LINE	L2	L2-123	754	755	A,B	A,B*	709 bp	Inv
	LTR	Gypsy	Gypsy-342	146	127	A,B	-	4408 bp	Inv
	Unc	Unc	Unc-41	1667	1742	A,B	-	970 bp	Inv
	Unc	Unc	Unc-3148	302	304	A,B	-	879 bp	Inv
7	TIR	*hAT*	*hAT*-1784	166	159	A,B*	-	1723 bp	Dir / Inv
12	Unc	Unc	Unc-307	1176	1208	-	A,B*	555 bp	Dir

*Derived haplotype. [a]Describes the relative positions of complementary TE copies in breakpoints, either in the direct, same orientation (Dir), inverted orientation (Inv), or several individual copies existing in both direct and inverted orientation (Dir/Inv). If the TE family exists in the breakpoints of both haplotypes, relative orientation refers to the TE copies in the derived haplotype. Unc = Unclassified.

To statistically evaluate the probability of finding copies of the same TE family within both breakpoints, we performed statistical simulations comparing the count of TEs found in breakpoints to obtaining the same results in randomly selected regions on the chromosome. Our results show a higher count of long (1-2 kb) related TEs in breakpoints of the derived haplotypes in LGs 1 and 7 in NEAC, and LG2 in NCC, compared to the mean count of related TEs on a chromosome-wide background (**Supplementary Figures 15-16**). Hence, our results suggest that the probability of finding long TEs of the same family (1-2 kb in length) by chance is low (e.g., the occurrence of at least one TE family longer than 1000 bp within 100 kb windows in LG7 in NEAC was found in only 0.20 incidences). This further supports the association between inversion breakpoints and longer TEs.

Evolutionary relationship of the hAT-1784 family in breakpoints of LG1 and LG7 inversions

Following the discovery of multiple copies of the *hAT*-1784 family in breakpoints of the derived LG1 and LG7 inversions in NEAC, we analysed the evolutionary relationship of all 473 *hAT*-1784 family members in NEAC, NCC and Celtic cod with phylogenetic approaches. Prior phylogenomic analyses on SNP data have suggested separate origins for the derived inversions on LG1 and LG7 – ~0.61 Ma and ~1.66 Ma – respectively (Matschiner et al., 2022). If the insertions of *hAT*-1784 single-copies temporally associated with the inversion origins, we would expect to observe separate expansions for the TEs in breakpoints of the different derived inversions.

Our phylogenetic analyses revealed that *hAT*-1784 have undergone two major expansions in the Atlantic cod genome (**Figure 5**). In the phylogenetic tree of the *hAT*-1784 elements (**Figure 5**), branch lengths indicate divergence time between sequences (i.e., phylogenetic distance), and we observe two events of rapid bursts of short branch lengths, separated by longer branch lengths with less diversification, indicative of two TE expansions. Interestingly, we observe that the *hAT*-1784 copies residing in breakpoints of LGs 1 and 7 have transposed during separate expansions, colour-coded in **Figure 5**. This finding suggests that if *hAT*-1784 caused the inversions, it happened due to two separate genomic radiations of *hAT*-1784.

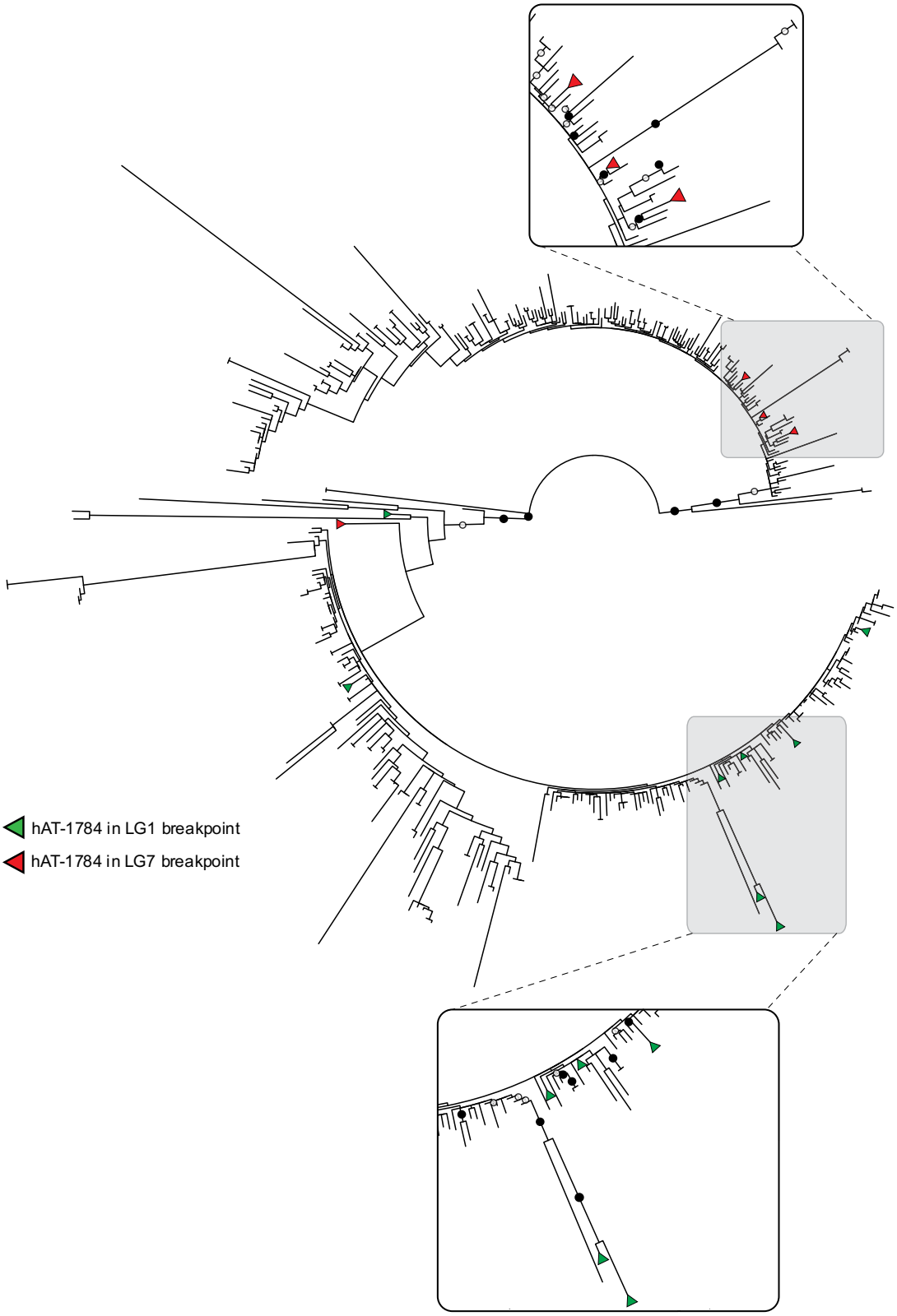


Figure 5 – Phylogeny of *hAT*-1784 in NEAC, NCC and Celtic cod, using ML tree estimation (see Materials and Methods). Each branch tip is a single *hAT*-1784 element, and branch lengths represent phylogenetic distance. Green triangles: *hAT* elements in LG1 breakpoints, red triangles: *hAT* elements in LG7 breakpoints. Clusters of breakpoint *hATs* are highlighted. For simplicity, only bootstrap support in highlighted regions and for the long branches separating the two *hAT*-1784 expansions is shown. Black dots show bootstrap support >90, white dots show bootstrap support >75.

TE-families often consist of several hundreds of single-copy elements, both full-length and fragmented TEs. This implies that in the 473 *hAT*-1784 copies, we would expect more closely related copies to form clusters. We therefore inferred orthogroups with OrthoFinder (Emms & Kelly, 2019) for all *hAT*-1784 copies in the NEAC, NCC and Celtic cod ecotypes. Out of 473 input sequences, ~90% were assigned in orthogroups. Of these orthogroups, one noticeable feature was that several of the breakpoint *hAT* elements cluster in the same orthogroups, indicating a shared ancestry of several breakpoint TEs (**Supplementary Table 6**). Inferring the gene trees of the orthogroups containing breakpoint *hAT* elements, we find that breakpoint *hATs* in the LG1 and LG7 inversions of NEAC share orthologs with both NCC and Celtic cod (**Figure 6**). However, there seem to have been private expansions of *hAT* copies in NEAC, all involving transposition into different breakpoints of the derived inversion haplotypes.

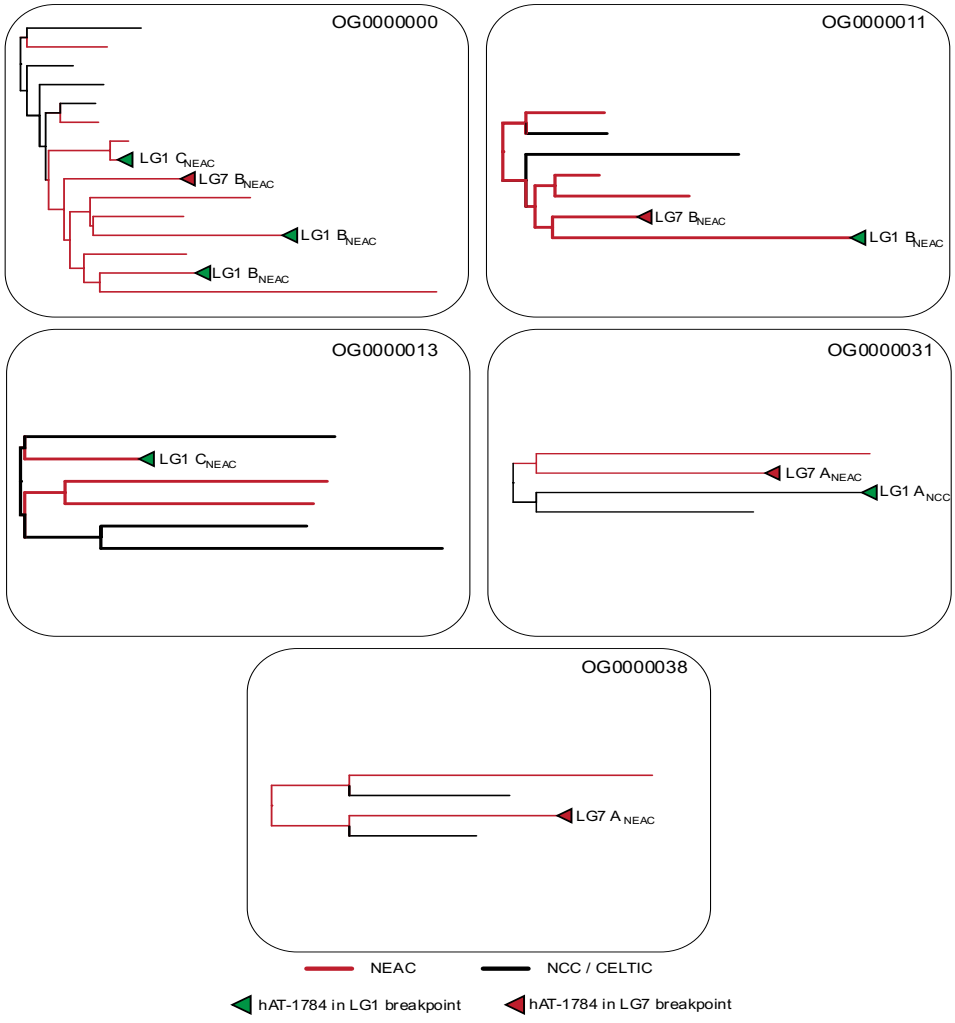


Figure 6 – Gene trees for the orthogroups involving *hAT*-1784 elements in breakpoints of the derived inversion in NEAC (see Materials and Methods). Each branch tip is a single *hAT*-1784 element. Green triangles: *hAT* elements in LG1 breakpoints. Red triangles: *hAT* elements in LG7 breakpoints. Red branches show *hAT* elements in NEAC. Black branches are *hAT* elements in NCC or Celtic cod.

Discussion

Here, we have investigated the potential influence of TEs in facilitating the origin of the four chromosomal inversions in Atlantic cod on LGs 1, 2, 7 and 12. To do this, we created a *de novo* repeat library for Atlantic cod and performed comparative genomic and phylogenetic analyses on the TE content in inversion breakpoints of the different ecotypes NEAC and NCC. We found related TEs in the breakpoints of all the derived, inverted haplotypes, of which one DNA *hAT*-family was present both in the breakpoints of the LG1 and LG7 inversions in NEAC.

Proliferation of DNA hAT elements promotes chromosomal rearrangements

By comparing the TE content in inversion breakpoints, we discovered closely related TEs in the LG1 and LG7 inversion breakpoints, all belonging to a single DNA *hAT* family that we named *hAT*-1784. We found breakpoint *hAT* elements in both direct and reverse orientation of one another, in line with other findings in distantly related species of inverted repeats in inversion breakpoints such as deer mice and maize (Harringmeyer & Hoekstra, 2022; Sharma & Peterson, 2022)**.** The *hAT* superfamily is highly abundant in teleost genomes (Gao et al., 2016) and has been linked to chromosomal rearrangements in a wide range of species (Lyttle & Haymer, 1992; Sarilar et al., 2014; Sharma & Peterson, 2022; Zhang & Peterson, 2004). Previous studies in fungi suggest that *hAT* elements induce chromosomal inversions by recombination, as opposed to NHEJ (Sarilar et al., 2014). *hAT* elements contain characteristic TIRs (Atkinson, 2015), and other DNA classes with TIRs are known to facilitate inversions by TIR-TIR recombination events (Ling & Cordaux, 2010). The *hAT*-1784-elements we located in the breakpoints are fairly long (mean length of 2042 bp and 3514 bp, **Figure 4b, Supplementary Table 5**), and recombination probability is known to increase with TE length (Petrov et al., 2011). Indeed, studies on inversion breakpoints in the human genome have reported that ~70% of the detected inversions were driven by ectopic recombination of larger homologous segments (≥200 bp) (Kidd et al., 2010). In light of their structural features, their length, and data from other species, it is likely that the *hAT*-1784 elements may facilitate ectopic recombination (Sharma & Peterson, 2022), and thus may have been involved in the LG1 and LG7 inversions in Atlantic cod. This is supported by finding *hAT*-1784 in the breakpoints of only the derived arrangements in LG1 and LG7, knowing that TE expansions can cause broad-scale genomic rearrangements within few generations (Sharma & Peterson, 2022). Moreover, inversions can establish in a population because they generate beneficial breakpoint mutations when they emerge (reviewed in Villoutreix et al., 2021). In such a

scenario, one would expect the breakpoint sequences to remain conserved over time due to a selective advantage, which is in line with our observation of fairly long *hAT*-1784 elements exclusively residing in the derived LG1 and 7 breakpoints. To further evaluate the association of *hAT*-1784, investigating population data on NEAC breakpoints will help to evaluate the population frequency of *hAT*-1784 around inversion breakpoints and determine if these TEs are fixed or not.

However, the *hAT*-1784-elements in LG 1 and 7 resided within breakpoints with above-average TE density- and frequency levels (**Figure 3** and **Supplementary Figure 2**). We also found TE density- and frequency peaks within other inversion breakpoints. TE accumulation has been suggested to arise in regions of suppressed recombination (Peona, Palacios-Gimenez, et al., 2021). For instance, the non-recombining avian W chromosome show TE densities more than five times greater than the mean genome-wide TE density (Peona, Palacios-Gimenez, et al., 2021). Thus, the observed TE density peaks in Atlantic cod breakpoints could represent low-recombination refugium for TEs, rather than TEs having facilitated the origin of the inversions, or even both. However, density- and frequency peaks were not solely confined to breakpoints, and we also observed peaks in some syntenic breakpoints of the ancestral haplotypes (e.g., LG1 C_{NCC} and LG7 B_{NCC}). Moreover, the TE content of the different density peaks was largely heterogeneous. Hence, time estimates are needed to differentiate between the TE insertions around breakpoints and further elucidate the association of *hAT*-1784 (and other TEs) with inversion origins, for instance by comparing breakpoint TE composition with a suitable outgroup (e.g., on homologous chromosomes in the haddock genome).

hAT-1784 has undergone two major expansions in Atlantic cod

Our phylogenetic analyses indicated that the Atlantic cod genome has experienced two major expansions of the *hAT*-1784 family. Interestingly, we found that the *hAT* elements that reside in the breakpoints of LG1 and LG7 formed separate clusters within the different expansions, meaning they potentially inserted in the two inversions during separate expansions. Furthermore, we found that the breakpoint *hATs* often clustered together in orthogroups (most similar elements) that had privately expanded within the NEAC genome and that *hAT*-1784 was absent from syntenic breakpoints in NCC. The inversions in LGs 1 and 7 likely originated on separate occasions (~0.61 Ma and ~1.66 Ma, respectively) (Matschiner et al., 2022), and the two largely separated expansions of *hAT*-1784 are therefore consistent with a scenario

where the inversions originated following the *hAT*-insertions. However, although our *hAT* phylogeny suggests they transposed on two distinct occasions, it does not provide age estimates to distinguish between different TE ages. Our GC analyses suggests older TE content in LG1 B_{NEAC} and LG7 A_{NEAC} (GC content of ~37% and ~20%, respectively, **Supplementary Figures 3-4**), compared to all the other breakpoints in NEAC and NCC. This fits with age estimates of the NEAC inversions, where the LG7 inversion is estimated to be one million years older than the LG1 inversions (Matschiner et al., 2022). However, we did not observe lowered GC content in the other LG1 and 7 breakpoints. Instead, the negative GC bias in LG1 B_{NEAC} and LG7 A_{NEAC} could reflect the innate low %GC of *hAT* elements (Symonová & Suh, 2019), or it may indicate low recombination rates due to silencing of larger TE accumulations (Kent et al., 2017), consistent with the low-recombination refugium hypothesis. However, given the TE refugium hypothesis, we would expect to observe lower GC content also in the other LG1 and 7 breakpoints (i.e., LG1 A_{NEAC}, C_{NEAC} and LG7 B_{NEAC}) with high TE density, suggesting that the clustering of *hAT* elements within the different expansions (**Figure 5**) was due to successful integration in low-recombining breakpoints. Our %GC analysis did not, however, show a negative GC bias following the TE density and frequency patterns.

To further investigate the *hAT* elements as candidates and to illuminate the different scenarios discussed here, an in-depth investigation of each *hAT*-1784 element and adjacent sequence directionalities are needed (e.g., protocols in Guillén & Ruiz, 2012). This may allow us to determine the directionality of flanking sequences of the TEs and evaluate if they are inverted or not. Age estimations of the *hAT* elements are also needed to evaluate if they were inserted prior to or following the inversion. For instance, by comparing the *hAT* elements in NEAC with an outgroup whose split with Atlantic cod predates the inversions (e.g., polar cod or Arctic cod), infer orthogroups and rooted orthogroup gene trees. Suppose the TEs in breakpoints cluster with *hATs* from the outgroup species; in that case, it is possible to estimate when these elements coalesce by applying mutation rates (e.g., the mutation rate of the cod inversions, estimated in Matschiner et al., 2022) and divergence between sequences (similar to the approach in Jedlicka et al., 2020).

Related TEs within breakpoints of ancestral and derived haplotypes

Our comparative analyses of Atlantic cod inversion breakpoints revealed that closely related TEs appeared in the breakpoints of all four derived inversion haplotypes. Our results agree with

previous findings of TEs in inversion breakpoint sequences (Cáceres et al., 1999; Mathiopoulos et al., 1998; Sharma & Peterson, 2022), suggesting a predominant role of TEs in facilitating chromosomal inversions. However, as opposed to the described *hAT* elements exclusively residing in the derived haplotypes of LG 1 and 7, the LTR/Gypsy- and LINE/L2-family in the LG2 inversion breakpoints appeared in both the ancestral and derived haplotypes. This could imply that the two TE families were fixed prior to the inversion event or that they have inserted in those regions following the inversion event.

We also found accumulations of another LTR/Gypsy family, a putative LTR family, as well as an unclassified TE family exclusively residing in the ancestral LG2 haplotype (**Table 1**). It is plausible that these families transposed after the inversion originated. The inversion is located near the end of LG2, and it is tempting to speculate that the proliferation of diverse TE families in the LG2 breakpoints might result from successful TE invasion of telomeric and subtelomeric regions with usually low recombination rates, where TE insertions have more neutral effects on the host genome (reviewed in Kent et al., 2017).

In conclusion, we reveal that related TEs of the DNA *hAT* family, *hAT*-1784, reside within the breakpoints of the derived inversions on LG1 and LG7 in NEAC. We find that *hAT*-1784 most likely inserted into the breakpoints of the LG1 and LG7 inversions during separate TE expansions. Our results suggest that *hAT* elements are candidates to have facilitated inversion origins in NEAC by ectopic recombination. Furthermore, we show that related TEs reside within the breakpoints of all derived inversion haplotypes in NEAC and NCC. Thus, our study presents a potential new, exciting example of the role of TEs in facilitating chromosomal rearrangements, underlining TEs as important drivers of genomic evolution and in the genetic differentiation between the NEAC and NCC ecotypes.

REFERENCES

Atkinson, P. W. (2015). HAT transposable elements. *Microbiology Spectrum*, *3*(4). https://doi.org/10.1128/microbiolspec.MDNA3-0054-2014

Auvinet, J., Graça, P., Dettai, A., Amores, A., Postlethwait, J. H., Detrich, H. W., 3rd, Ozouf-Costaz, C., Coriton, O., & Higuet, D. (2020). Multiple independent chromosomal fusions accompanied the radiation of the Antarctic teleost genus Trematomus (Notothenioidei:Nototheniidae). *BMC Evolutionary Biology*, *20*(1), 39.

Bao, Z., & Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research*, *12*(8), 1269–1276.

Barth, J. M. I., Berg, P. R., Jonsson, P. R., Bonanomi, S., Corell, H., Hemmer-Hansen, J., Jakobsen, K. S., Johannesson, K., Jorde, P. E., Knutsen, H., Moksnes, P.-O., Star, B., Stenseth, N. C., Svedäng, H., Jentoft, S., & André, C. (2017). Genome architecture enables local adaptation of Atlantic cod despite high connectivity. *Molecular Ecology*, *26*(17), 4452–4466.

Berg, P. R., Star, B., Pampoulie, C., Bradbury, I. R., Bentzen, P., Hutchings, J. A., Jentoft, S., & Jakobsen, K. S. (2017). Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity*, *119*(6), 418–428.

Berg, Paul R., Jentoft, S., Star, B., Ring, K. H., Knutsen, H., Lien, S., Jakobsen, K. S., & André, C. (2015). Adaptation to low salinity promotes genomic divergence in Atlantic cod (Gadus morhua L.). *Genome Biology and Evolution*, *7*(6), 1644–1663.

Berg, Paul 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), 23246.

Brander, K. (2005). Spawning and life history information for North Atlantic cod stocks. *ICES Coo*, *274*, 1–152.

Böhne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C., & Volff, J.-N. (2008). Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *16*(1), 203–215.

Cabanettes, F., & Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *6*, e4958.

Cáceres, M., Ranz, J. M., Barbadilla, A., Long, M., & Ruiz, A. (1999). Generation of a widespread Drosophila inversion by a transposable element. *Science (New York, N.Y.)*, *285*(5426), 415–418.

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, *9*(1), 18.

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 157.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238.

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

Fryxell, K. J., & Zuckerkandl, E. (2000). Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Molecular Biology and Evolution*, *17*(9), 1371–1383.

Gao, B., Shen, D., Xue, S., Chen, C., Cui, H., & Song, C. (2016). The contribution of transposable elements to size variations between four teleost genomes. *Mobile DNA*, *7*(1), 4.

Goubert, C., Craig, R. J., Bilat, A. F., Peona, V., Vogan, A. A., & Protasio, A. V. (2022). A beginner's guide to manual curation of transposable elements. *Mobile DNA*, *13*(1), 7.

Guillén, Y., & Ruiz, A. (2012). Gene alterations at Drosophila inversion breakpoints provide prima facie evidence for natural selection as an explanation for rapid chromosomal evolution. *BMC Genomics*, *13*, 53.

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

Hemmer-Hansen, J., Nielsen, E. E., Therkildsen, N. O., Taylor, M. I., Ogden, R., Geffen, A. J., Bekkevold, D., Helyar, S., Pampoulie, C., Johansen, T., FishPopTrace Consortium, & Carvalho, G. R. (2013). A genomic island linked to ecotype divergence in Atlantic cod. *Molecular Ecology*, *22*(10), 2653–2667.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522.

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: CB*, *28*(11), 1839-1845.e3.

Jedlicka, P., Lexa, M., & Kejnovsky, E. (2020). What can long terminal repeats tell us about the age of LTR retrotransposons, gene conversion and ectopic recombination? *Frontiers in Plant Science*, *11*, 644.

Kapun, M., Fabian, D. K., Goudet, J., & Flatt, T. (2016). Genomic Evidence for Adaptive Inversion Clines in Drosophila melanogaster. *Molecular Biology and Evolution*, *33*(5), 1317–1336.

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

Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., Kallicki, J., Kaul, R., Wilson, R. K., & Eichler, E. E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, *143*(5), 837–847.

Kirubakaran, T. G., Andersen, Ø., Moser, M., Árnyasi, M., McGinnity, P., Lien, S., & Kent, M. (2020). A nanopore based chromosome-level assembly representing Atlantic cod from the Celtic sea. *G3 (Bethesda, Md.)*, *10*(9), 2903–2910.

Kirubakaran, T. G., Grove, H., Kent, M. P., Sandve, S. R., Baranski, M., Nome, T., De Rosa, M. C., Righino, B., Johansen, T., Otterå, H., Sonesson, A., Lien, S., & Andersen, Ø. (2016). Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*, *25*(10), 2130–2143.

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), 84–88.

Ling, A., & Cordaux, R. (2010). Insertion sequence inversions mediated by ectopic recombination between terminal inverted repeats. *PloS One*, *5*(12), e15654.

Lyttle, T. W., & Haymer, D. S. (1992). The role of the transposable element hobo in the origin of endemic inversions in wild populations of Drosophila melanogaster. *Genetica*, *86*(1–3), 113–126.

Mathiopoulos, K. D., della Torre, A., Predazzi, V., Petrarca, V., & Coluzzi, M. (1998). Cloning of inversion breakpoints in the Anopheles gambiae complex traces a transposable element at the inversion junction. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(21), 12444–12449.

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), 469–481.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419.

Ou, S., Chen, J., & Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research*, *46*(21), e126.

Peona, V., Blom, M. P. K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., Liachko, I., Haryoko, T., Jønsson, K. A., Zhou, Q., Irestedt, M., & Suh, A. (2021). Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular Ecology Resources*, *21*(1), 263–286.

Peona, V., Palacios-Gimenez, O. M., Blommaert, J., Liu, J., Haryoko, T., Jønsson, K. A., Irestedt, M., Zhou, Q., Jern, P., & Suh, A. (2021). The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *376*(1833), 20200186.

Petrov, D. A., Fiston-Lavier, A.-S., Lipatov, M., Lenkov, K., & González, J. (2011). Population genomics of transposable elements in Drosophila melanogaster. *Molecular Biology and Evolution*, *28*(5), 1633–1644.

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics (Oxford, England)*, *21 Suppl 1*(Suppl 1), i351-8.

Ranz, J. M., Maurin, D., Chan, Y. S., von Grotthuss, M., Hillier, L. W., Roote, J., Ashburner, M., & Bergman, C. M. (2007). Principles of genome evolution in the Drosophila melanogaster species group. *PLoS Biology*, *5*(6), e152.

Reiss, H., Hoarau, G., Dickey-Collas, M., & Wolff, W. J. (2009). Genetic population structure of marine fish: mismatch between biological and fisheries management units. *Fish and Fisheries (Oxford, England)*, *10*(4), 361–395.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European molecular biology open software suite. *Trends in Genetics: TIG*, *16*(6), 276–277.

Rollefsen, G. (1953). Observations on the Cod and Cod Fisheries of Lofoten. *ICES*. http://hdl.handle.net/11250/101129

Sarilar, V., Bleykasten-Grosshans, C., & Neuvéglise, C. (2014). Evolutionary dynamics of hAT DNA transposon families in Saccharomycetaceae. *Genome Biology and Evolution*, *7*(1), 172–190.

Schwander, T., Libbrecht, R., & Keller, L. (2014). Supergenes and complex phenotypes. *Current Biology: CB*, *24*(7), R288-94.

Sharma, S. P., & Peterson, T. (2022). Complex chromosomal rearrangements induced by transposons in maize. *Genetics*. https://doi.org/10.1093/genetics/iyac124

Sodeland, M., Jorde, P. E., Lien, S., Jentoft, S., Berg, P. R., Grove, H., Kent, M. P., Arnyasi, M., Olsen, E. M., & Knutsen, H. (2016). “islands of divergence” in the Atlantic cod genome represent polymorphic chromosomal rearrangements. *Genome Biology and Evolution*, *8*(4), 1012–1022.

Storer, J., Hubley, R., Rosen, J., Wheeler, T. J., & Smit, A. F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, *12*(1), 2.

Stump, A. D., Pombi, M., Goeddel, L., Ribeiro, J. M. C., Wilder, J. A., della Torre, A., & Besansky, N. J. (2007). Genetic exchange in 2La inversion heterokaryotypes of Anopheles gambiae. *Insect Molecular Biology*, *16*(6), 703–709.

Symonová, R., & Suh, A. (2019). Nucleotide composition of transposable elements likely contributes to AT/GC compositional homogeneity of teleost fish genomes. *Mobile DNA*, *10*(1), 49.

Tørresen, O. K., Star, B., Jentoft, S., Reinar, W. B., Grove, H., Miller, J. R., Walenz, B. P., Knight, J., Ekholm, J. M., Peluso, P., Edvardsen, R. B., Tooming-Klunderud, A., Skage, M., Lien, S., Jakobsen, K. S., & Nederbragt, A. J. (2017). An improved

genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics*, *18*(1), 95.

Villoutreix, R., Ayala, D., Joron, M., Gompert, Z., Feder, J. L., & Nosil, P. (2021). Inversion breakpoints and the evolution of supergenes. *Molecular Ecology*, *30*(12), 2738–2755.

Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, *33*(6), 427–440.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews. Genetics*, *8*(12), 973–982.

Zhang, J., & Peterson, T. (2004). Transposition of reversed Ac element ends generates chromosome rearrangements in maize. *Genetics*, *167*(4), 1929–1937.

Zhang, J., Zuo, T., Wang, D., & Peterson, T. (2014). Transposition-mediated DNA re-replication in maize. *ELife*, *3*, e03724.

Supplementary information

Table of Contents

Supplementary Tables

Repeat content in Atlantic cod ecotypes

Supplementary Table 1 – The repeat content of the Northeast Atlantic cod (NEAC) genome assembly. All numbers are from output estimates from running RepeatMasker on the NEAC assembly with the *de novo* NEAC repeat library and adopted to the Wicker classification system (Wicker et al., 2007).

Group	Num. elements[a]	Coverage (Mb)	Coverage[b] (%)
Unknown	316,828	67.5	10.05
DNA	287,972	63.4	9.46
LTR	161,993	42.6	6.35
LINE	57,155	22.0	3.28
Total interspersed repeats[c]	835,118	197.1	29.41
Tandem repeats	605,957	50.7	7.57
Total	**1,447,606[d]**	**249.3[e]**	**37.22[e]**

[a]Most repeats fragmented by insertions or deletions are counted as one element by RepeatMasker. [b]Groups of elements with less than 1% coverage in the assembly are not shown, but they are included in the total count.. [c]Total of all annotated interspersed repeats, including Unknown, DNA, LTR, LINE and SINE. [d]This is the sum of all annotated repeats from this table, including small RNAs. [e]Refers to bases masked by RepeatMasker.

Supplementary Table 2 – The repeat content of the Norwegian Coastal Cod (NCC) genome assembly (Brieuc et al., in prep.). All numbers are from output estimates from running RepeatMasker on the NCC assembly with the *de novo* NEAC repeat library and adopted to the Wicker classification system (Wicker et al., 2007).

Group	Num. elements[a]	Coverage (Mb)	Coverage[b] (%)
Unknown	326,619	68.3	10.05
DNA	307,885	64.8	9.53
LTR	171,369	41.3	6.07
LINE	59,285	21.1	3.10
Total interspersed repeats[c]	876,962	197.3	29.03
Tandem repeats	656,471	55.2	8.12
Total	**1,537,507[d]**	**253.0[e]**	**37.23[e]**

[a]Most repeats fragmented by insertions or deletions are counted as one element by RepeatMasker. [b]Groups of elements with less than 1% coverage in the assembly are not shown, but they are included in the total count.. [c]Total of all annotated interspersed repeats, including Unknown, DNA, LTR, LINE and SINE. [d]This is the sum of all annotated repeats from this table, including small RNAs. [e]Refers to bases masked by RepeatMasker.

Supplementary Table 3 – The repeat content of the Celtic cod genome assembly (Kirubakaran et al., 2020). All numbers are from output estimates from running RepeatMasker on the NCC assembly with the *de novo* NEAC repeat library and adopted to the Wicker classification system (Wicker et al., 2007).

Group	**Num. elements**[a]	**Coverage (Mb)**	**Coverage**[b] **(%)**
Unknown	320,021	67.7	9.90
DNA	298,393	67.9	9.96
LTR	170,264	42.8	6.26
LINE	58,990	22.3	3.26
Total interspersed repeats[c]	859,484	193.0	28.20
Tandem repeats	637,297	58.6	8.56
Total	**1,502,844**[d]	**253.4**[e]	**38.42**[e]

[a]Most repeats fragmented by insertions or deletions are counted as one element by RepeatMasker. [b]Groups of elements with less than 1% coverage in the assembly are not shown, but they are included in the total count. [c]Total of all annotated interspersed repeats, including Unknown, DNA, LTR, LINE and SINE. [d]This is the sum of all annotated repeats from this table, including small RNAs. [e]Refers to bases masked by RepeatMasker.

Breakpoint coordinates in NEAC and NCC

Supplementary Table 4 – Estimated coordinates for breakpoint regions in NEAC and NCC. New coordinate estimates were made for LG7 (A_{NCC}), and LG12 (A_{NCC}, B_{NCC}, and A_{NEAC}, B_{NEAC}).

LG	Cod ecotype	Arrangement	Inversion breakpoint		
			A	B	C[a]
1	NEAC	Derived	11,304.580	19,183,000	28,309,406
	NCC	Ancestral	10,687.767	18,435,406	27,848,889
2[b]	NEAC	Ancestral	476,583	4,467,322	-
	NCC	Derived	21,854,529	26,155,172	-
7	NEAC	Derived	16,855,144	26,338,536	-
	NCC	Ancestral	***15,476,000***	24,895,985	-
12	NEAC	Ancestral	***500,973***	***13,705,000***	-
	NCC	Derived	***740,576***	***13,632,000***	-

Bold italics: Breakpoint coordinates that were estimated in this study using D-GENIES (Cabanettes & Klopp, 2018).
[a]Breakpoint C is due to a double inversion on LG1, creating three breakpoints consisting of two outermost breakpoints (A and C) and one innermost breakpoint (B). [b]The large discrepancy between the inversion coordinates on NEAC and NCC is due to the LG2 in the gadMor3 assembly being placed in the opposite direction.

Overview of TEs in breakpoints of derived inversion haplotypes

Supplementary Table 5 – Coordinates and mean length of related TEs residing within breakpoints of the derived inversion haplotypes of NEAC and NCC on LGs 1, 2, 7 and 12.

TE family	Inversion (LG)	Num. elements in inversion breakpoints	Mean length (bp)
hAT-1784[a]	1	6	2042
	7	4	3514
Unc-753	1	2	240
Unc-915	1	2	333
Gypsy-428[a]	2	2	413
L2-123[a]	2	5	179
Unc-307	12	2	521

[a]Also present in at least one ancestral syntenic breakpoint.

Orthogroups with breakpoint hAT-1784 elements

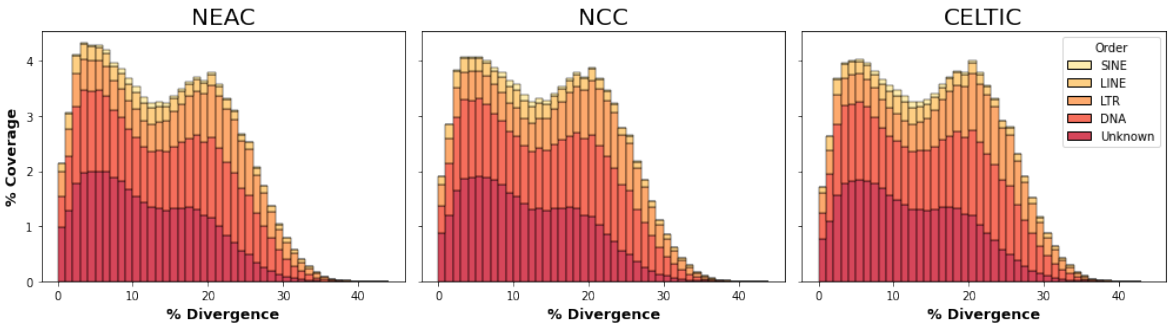
Supplementary Table 6 – Orthogroups of hAT-1784 that contain elements in inversion breakpoints. Orthogroupc were inferred with Orthofinder (Emms & Kelly, 2015, 2019) TE copies from each orthogroup are shown in coordinates (LG: Start – End) and ecotype affiliation.

Orthogroup	**TE copies in orthogroup (LG: Start-End)**		
	NEAC	**NCC**	**Celtic**
OG0000000	1: 19,183,901-19,186,096*	4: 4,115,463-4,118,682	12: 5,813,800-5,816,281
	1: 19,180,987-19,183,838*	7: 11,195,670-11,198,218	
	1: 28,327,476-28,329,870*	10: 8,233,730-8,236,204	
	4: 3,747,283-3,749,855		
	4: 38,087,461-38,089,390		
	5: 19,210,597-19,212,623		
	7: 2,126,370-2,128,676		
	7: 26,343,511-26,345,790*		
	7: 29,353,630-29,355,856		
	17: 10,773,874-10,776,348		
	22: 13,344,637-13,349,203		
OG0000011	1: 19,183,462-19,183,883*	15: 16,058,119-16,058,417	23: 16,804,413-16,804,731
	7: 26,345,504-26,345,838*		
	11: 20,845,522-20,845,813		
	15: 15,761,047-15,761,374		
	19: 9,791,341-9,791,665		
OG0000013	1: 28,325,484-28,328,002*	4: 36,384,673-36,386,798	9: 1,353,933-1,356,437
	2: 1,135,855-1,137,891		20: 12,360,391-12,363,071
	7: 29,355,303-29,358,520		
OG0000031	7: 16847707-16849277*	1: 10687734-10689294[b]	7: 17631435-17632974
	20: 12508287-12509847		
OG0000038	7: 16,838,883-16,848,755*	22: 13,236,655-13,248,560	19: 9,826,175-9,835,850
	11: 11,626,938-11,641,238		
OG0000141[a]	1: 19,180,283-19,182,158*		

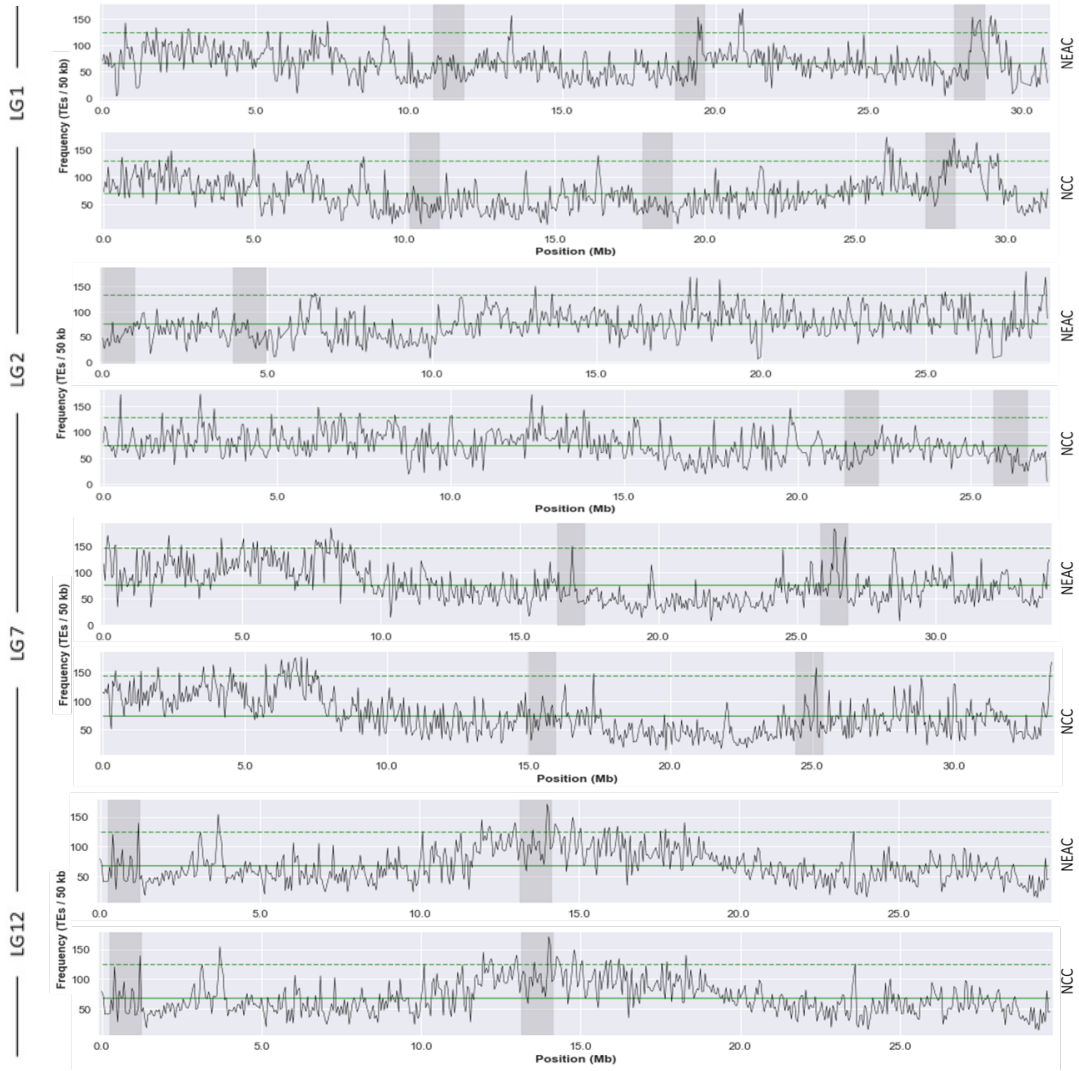
*TEs in breakpoints. [a]Sequence was not assigned to any orthogroup. [b]TE located in syntenic breakpoint of ancestral non-inverted haplotype.

Supplementary Figures

TE statistics in Atlantic cod

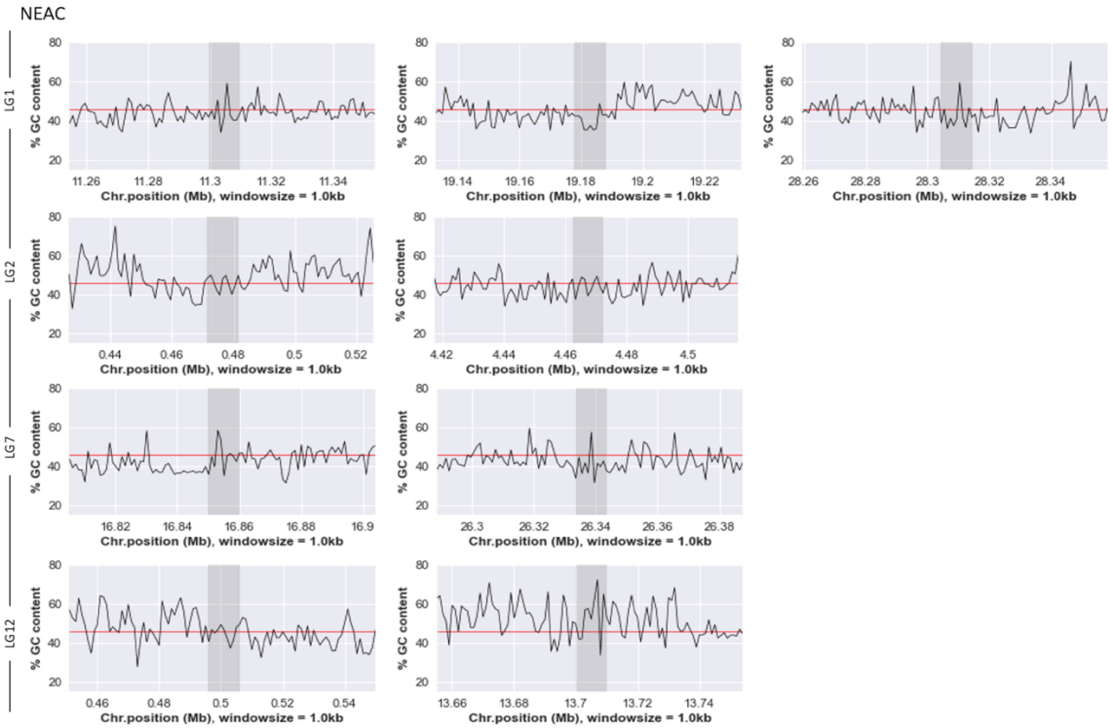


Supplementary Figure 1 – Repeat landscape of NEAC, NCC and Celtic cod. The plots show percentage divergence of SINEs, LINEs, LTRs, DNA-elements and Unknowns from its consensus sequence on the x-axis, and genome coverage (%) on the y-axis.

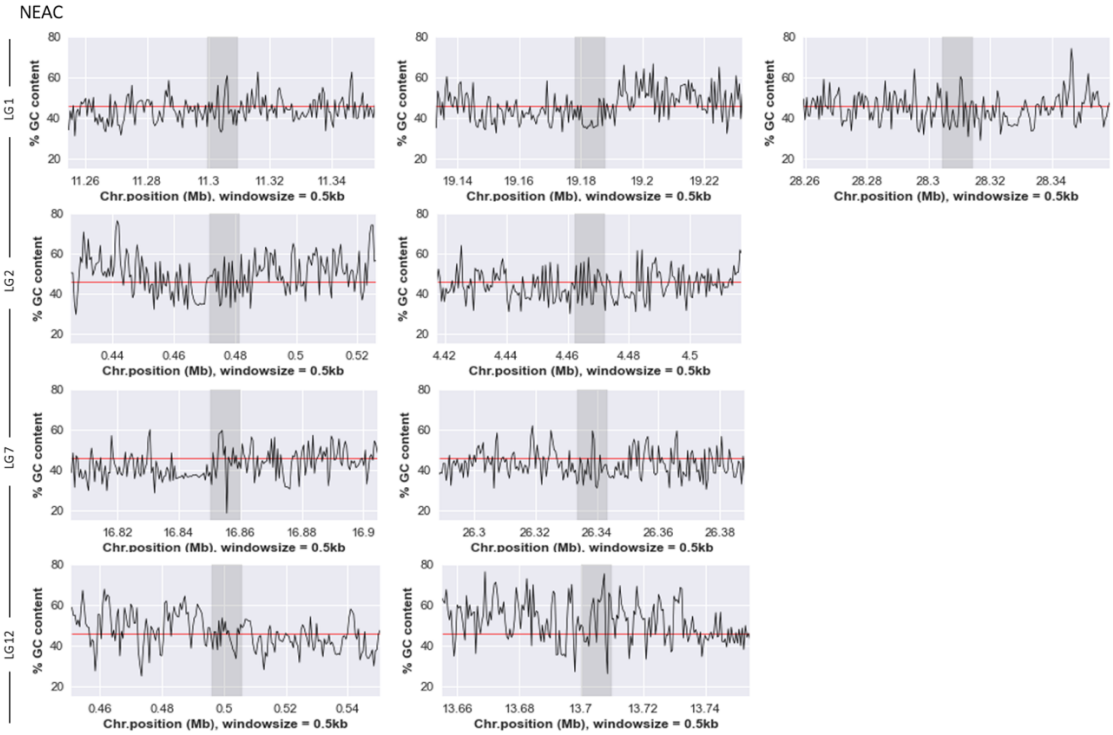


Supplementary Figure 2 – Frequency of TEs in LGs 1, 2, 7 and 12 in NEAC. TE frequency is plotted as TE single-copies per non-overlapping 50 kb sliding window in all four LGs. The green solid line is mean TE frequency across the LG, the green dashed line indicates 2 standard deviations from the mean, and the grey shaded areas illustrate breakpoints +/- 500 kb. LGs 1 and 7 are derived haplotypes in NEAC. LGs 2 and 12 are derived haplotypes in NCC.

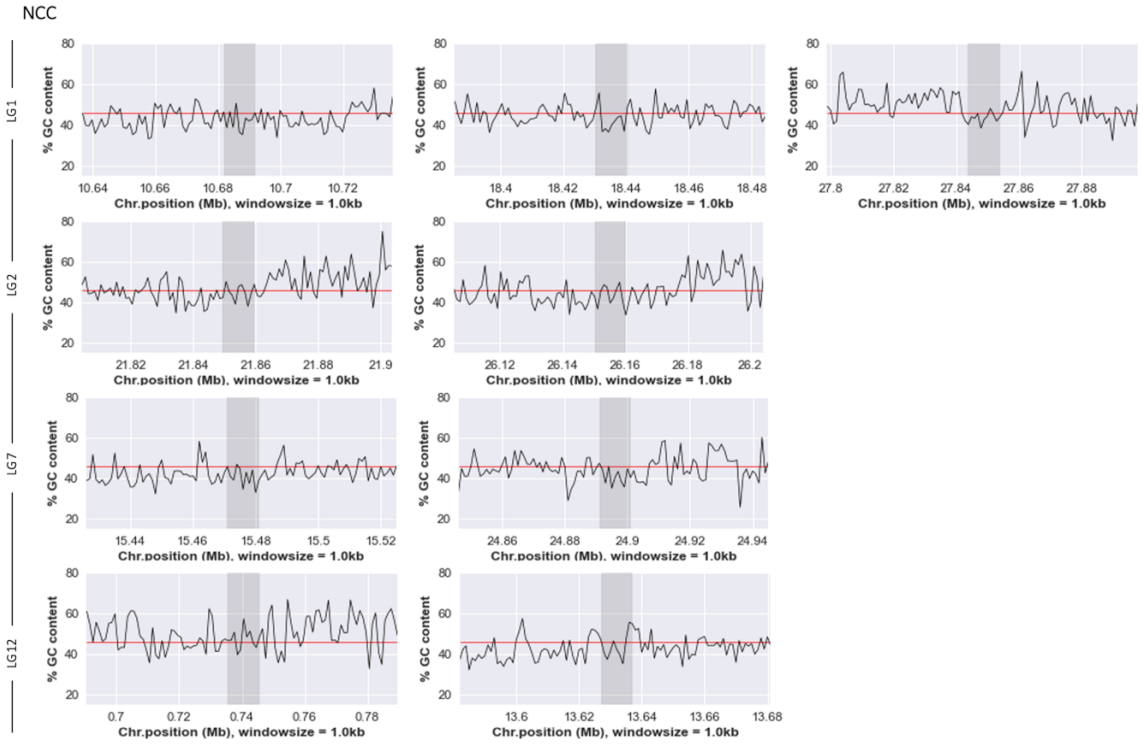
GC content in NEAC and NCC inversion breakpoints



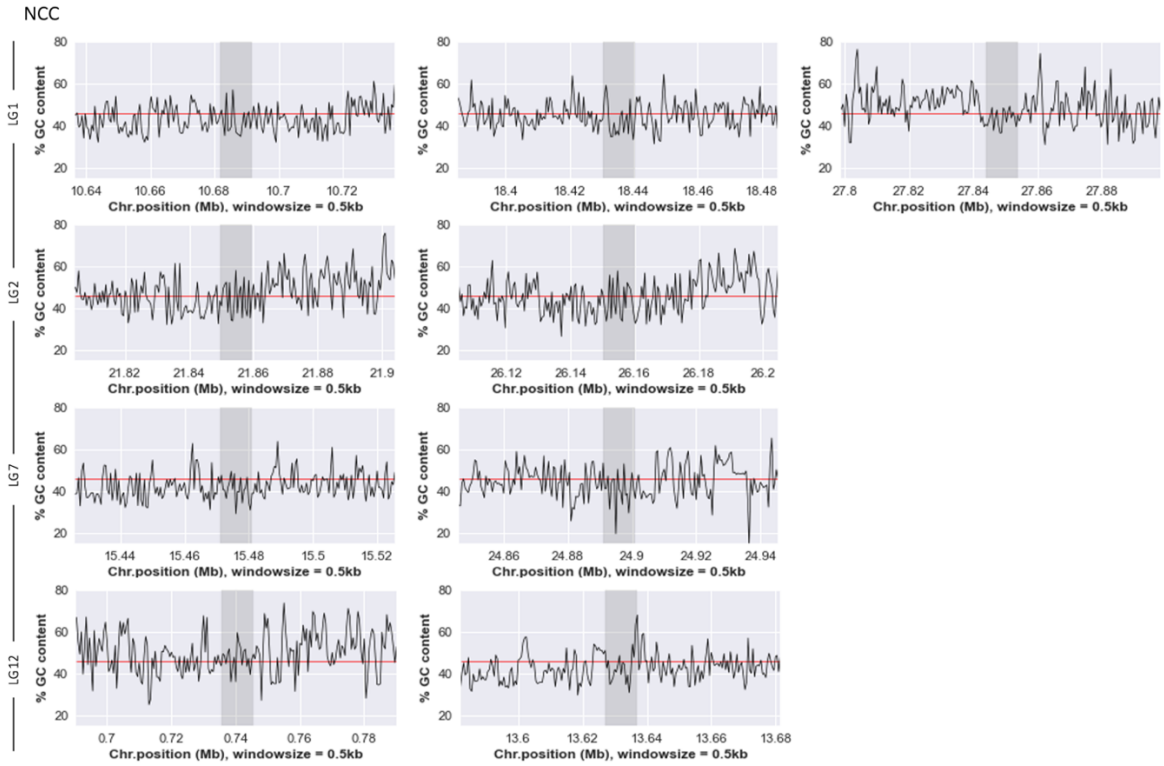
Supplementary Figure 3 – GC content in breakpoint regions (+/- 50 kb) in inversion haplotypes on LGs 1, 2, 7 and 12 in NEAC. % GC is plotted in non-overlapping 1.0 kb sliding windows. The red line is mean %GC for the whole genome. The shaded areas indicate the breakpoint coordinates from **Supplementary Table 4**, +/- 5 kb. LG1 is a double inversion and therefore the three breakpoints are shown.



Supplementary Figure 4 – GC content in breakpoint regions (+/- 50 kb) in inversion haplotypes on LGs 1, 2, 7 and 12 in NEAC. % GC is plotted in non-overlapping 1.0 kb sliding windows. The red line is mean %GC for the whole genome. The shaded areas indicate the breakpoint coordinates from **Supplementary Table 4**, +/- 5 kb. LG1 is a double inversion and therefore the three breakpoints are shown.

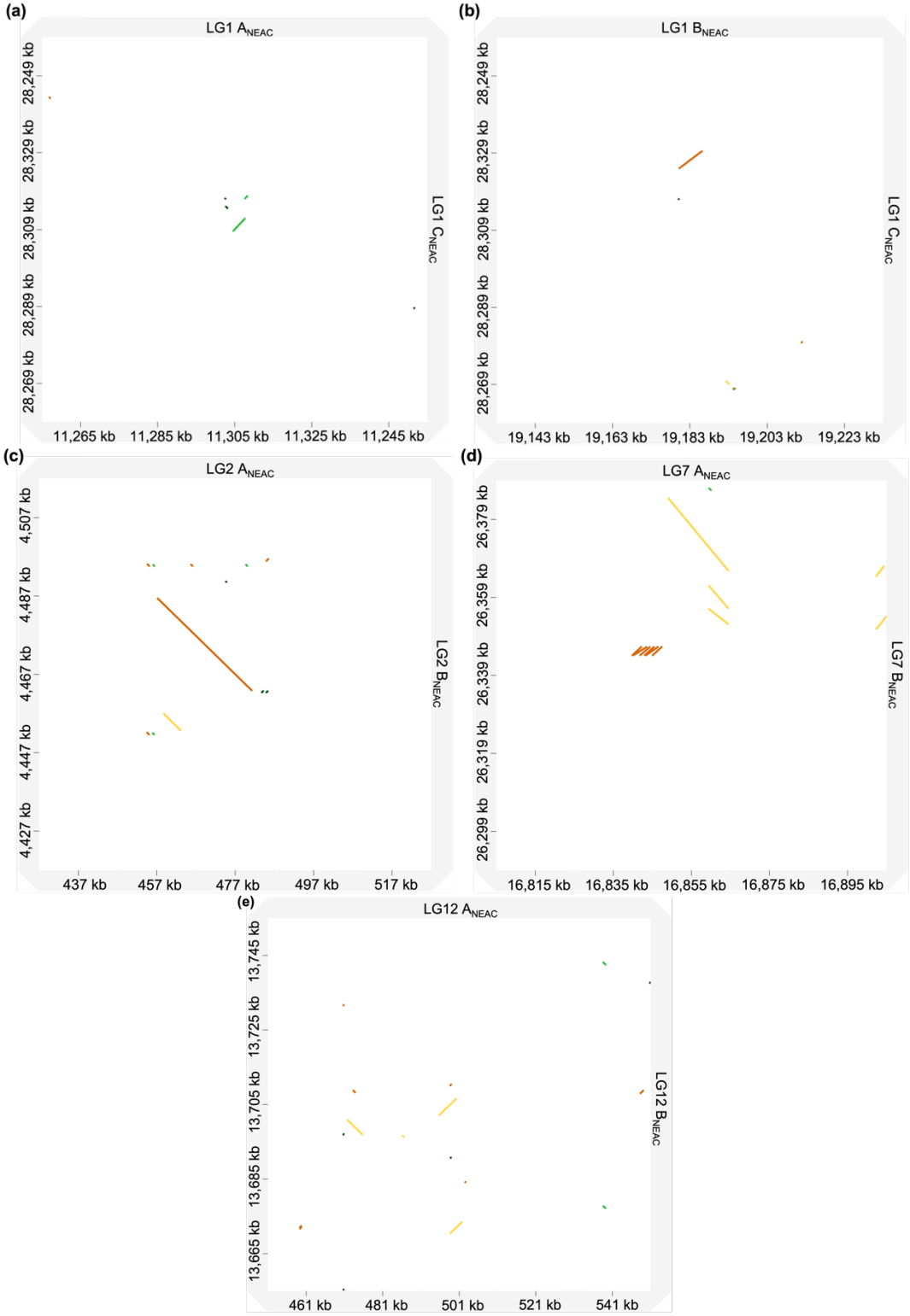


Supplementary Figure 5 – GC content in breakpoint regions (+/- 50 kb) in inversion haplotypes on LGs 1, 2, 7 and 12 in NCC. % GC is plotted in non-overlapping 1.0 kb sliding windows. The red line is mean %GC for the whole genome. The shaded areas indicate the breakpoint coordinates from **Supplementary Table 4**, +/- 5 kb. LG1 is a double inversion and therefore the three breakpoints are shown.

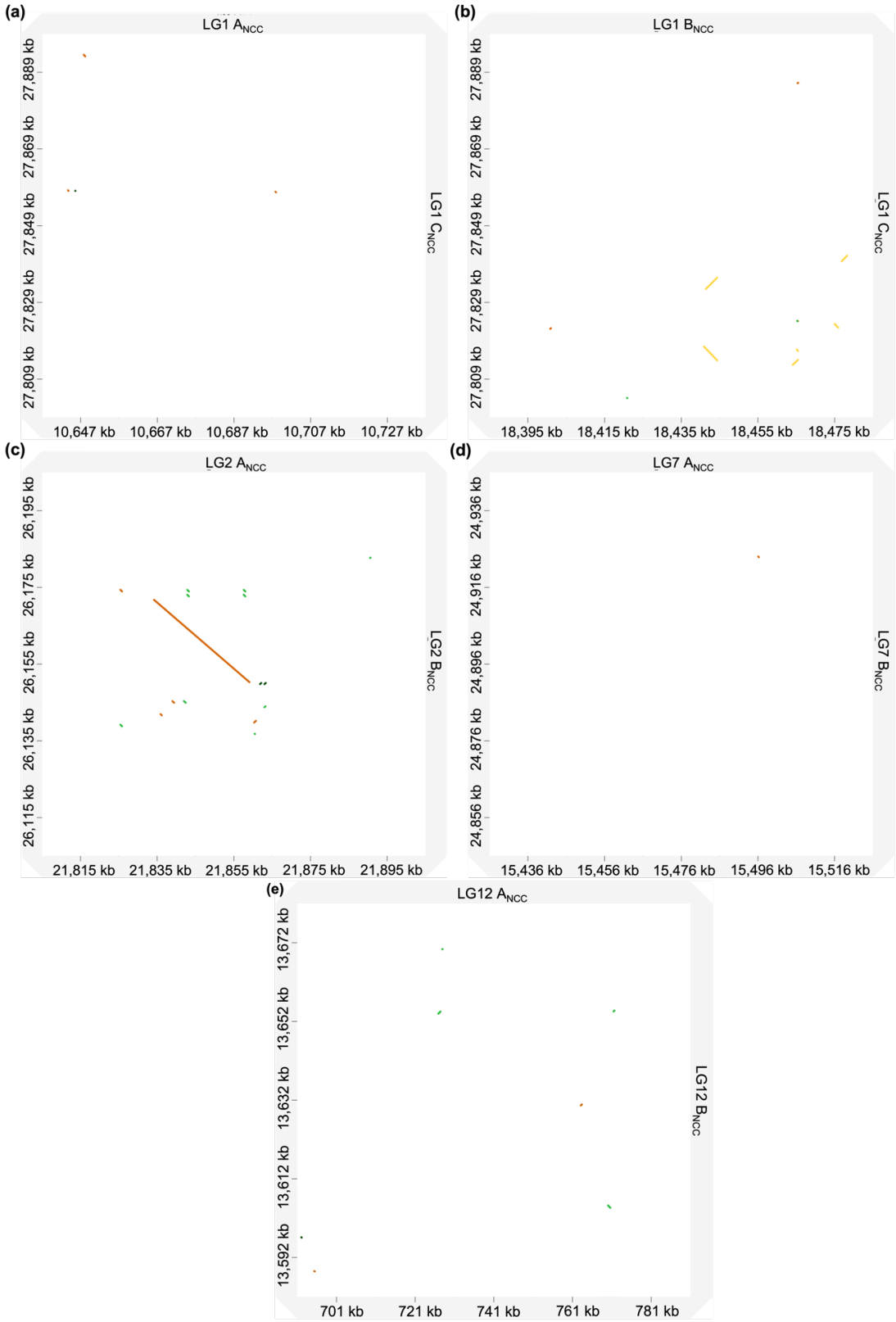


Supplementary Figure 6 – GC content in breakpoint regions (+/- 50 kb) in inversion haplotypes on LGs 1, 2, 7 and 12 in NEAC. % GC is plotted in non-overlapping 1.0 kb sliding windows. The red line is mean %GC for the whole genome. The shaded areas indicate the breakpoint coordinates from **Supplementary Table 4**, +/- 5 kb. LG1 is a double inversion and therefore the three breakpoints are shown.

Dot-plot alignments of breakpoints in NEAC and NCC

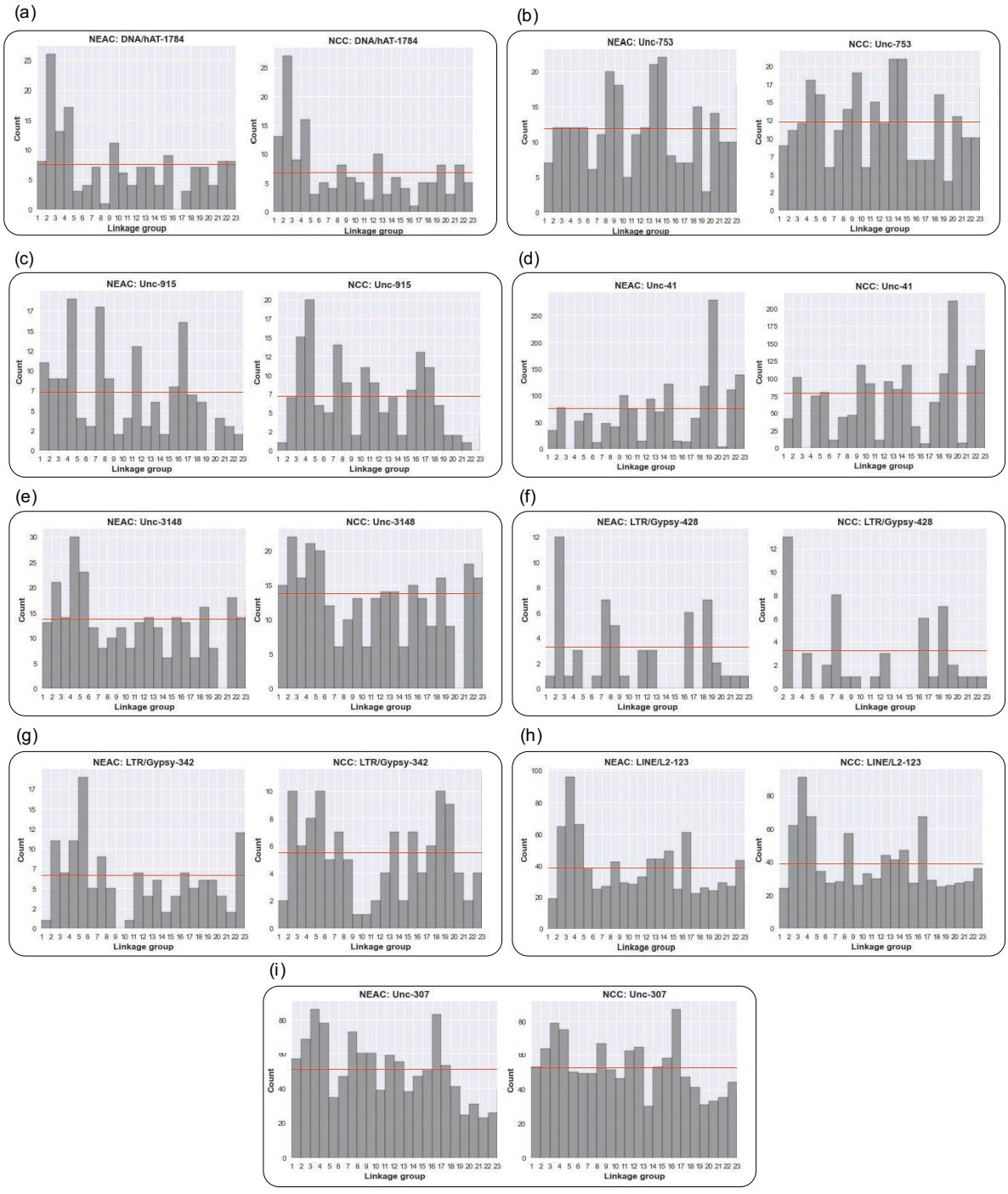


Supplementary Figure 7 – Alignment of breakpoints in inversion haplotypes of NEAC (coordinates from **Suppl. Table 4**, +/- 50 kb) using D-GENIES v.1.4 (Cabanettes & Klopp, 2018). Colour codes represent sequence similarity, yellow = 0-0.25, red = 0.26-0.5, light green = 0.51-0.75, dark green = 0.76-1. (a) LG1 A_{NEAC} vs C_{NEAC} (derived). (b) LG1 B_{NEAC} vs C_{NEAC} (derived). (c) LG2 B_{NEAC} vs C_{NEAC} (ancestral). (d) LG7 B_{NEAC} vs C_{NEAC} (derived). (e) LG12 B_{NEAC} vs C_{NEAC} (ancestral). LG1 A_{NEAC} vs B_{NEAC} did not provide any significant similarity and is not included.



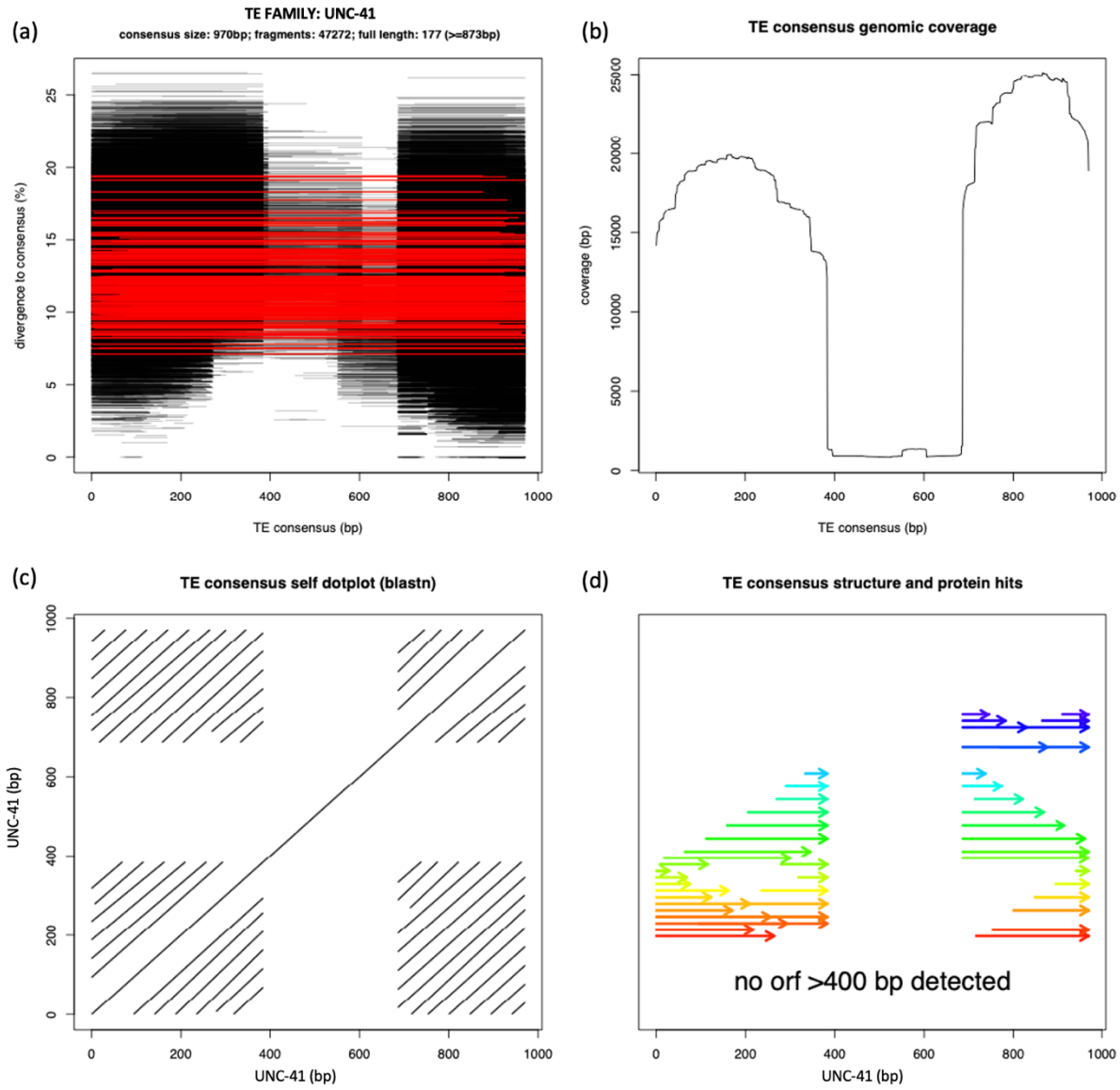
Supplementary Figure 8 – Alignment of breakpoints in inversion haplotypes of NCC (coordinates from **Suppl. Table 4**, +/- 50 kb) using D-GENIES v.1.4 (Cabanettes & Klopp, 2018). Colour codes represent sequence similarity, yellow = 0-0.25, red = 0.26-0.5, light green = 0.51-0.75, dark green = 0.76-1. (a) LG1 A_{NCC} vs C_{NCC} (ancestral). (b) LG1 B_{NCC} vs C_{NCC} (ancestral). (c) LG2 B_{NCC} vs C_{NCC} (derived). (d) LG7 B_{NCC} vs C_{NCC} (ancestral). (e) LG12 B_{NCC} vs C_{NCC} (derived). LG1 A_{NCC} vs B_{NCC} did not provide any significant similarity and is not included.

Genomic distribution of breakpoint TEs

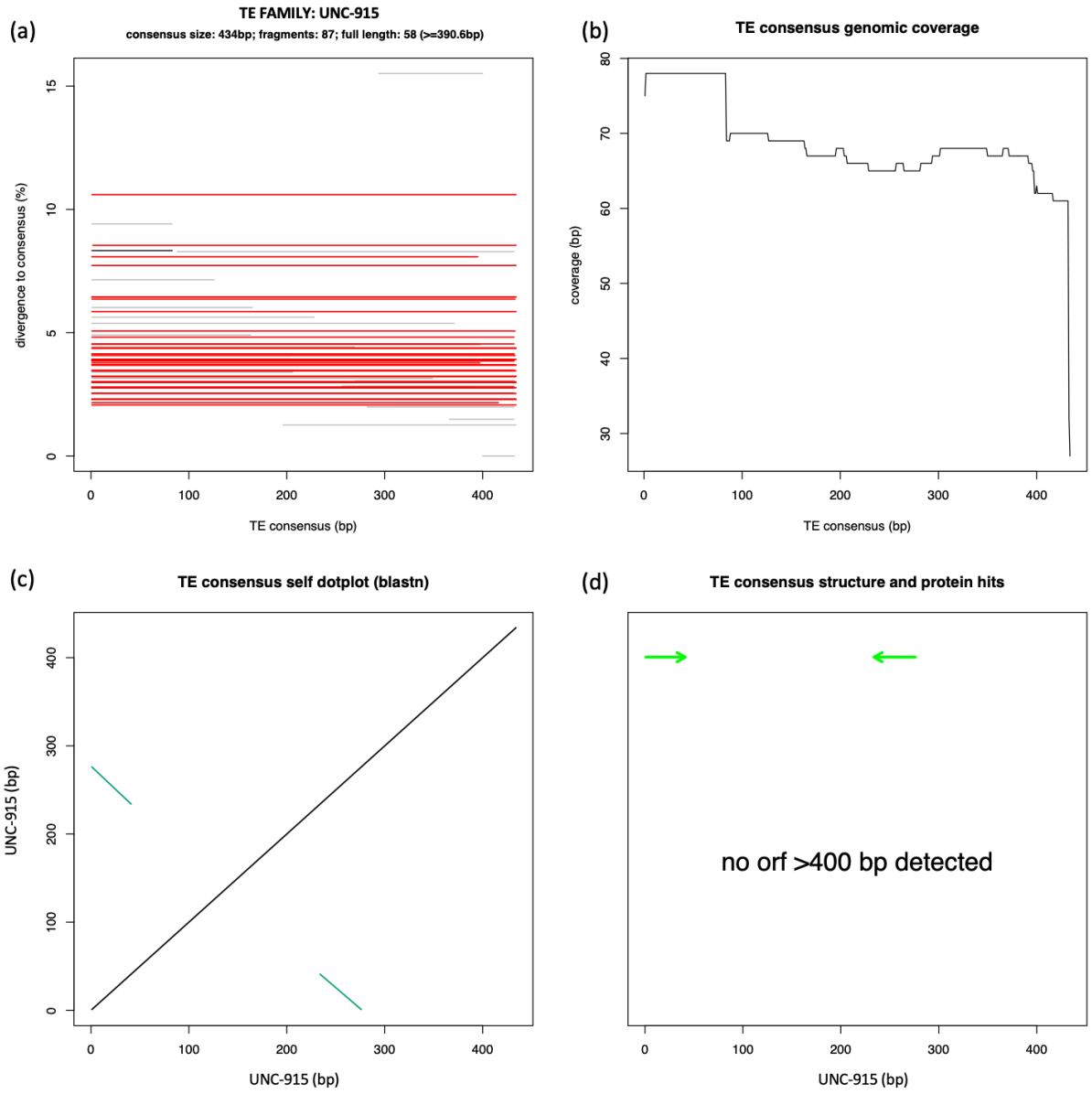


Supplementary Figure 9 – Genome distribution of the TE families residing in breakpoints of inversion haplotypes for NEAC and NCC. The red line illustrates mean count of the TE family. **(a)** hAT-1784, **(b)** Unc-753, **(c)** Unc-915, **(d)** Unc-41, **I** Unc-3148, **(f)** Gypsy-428, **(g)** Gypsy-342, **(h)** L2-123, **(i)** Unc-307.

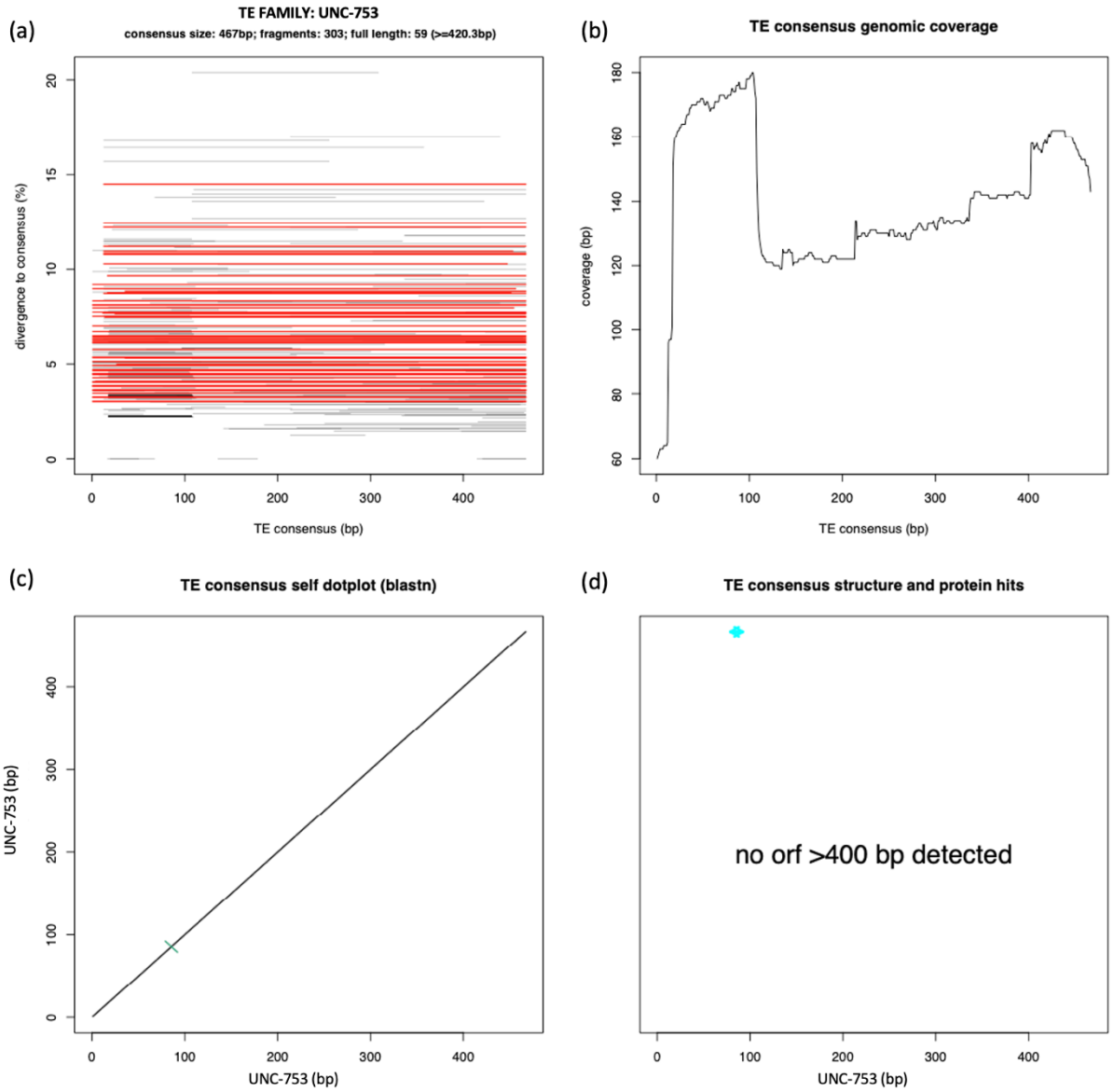
Manual curation of unclassified TEs in breakpoints



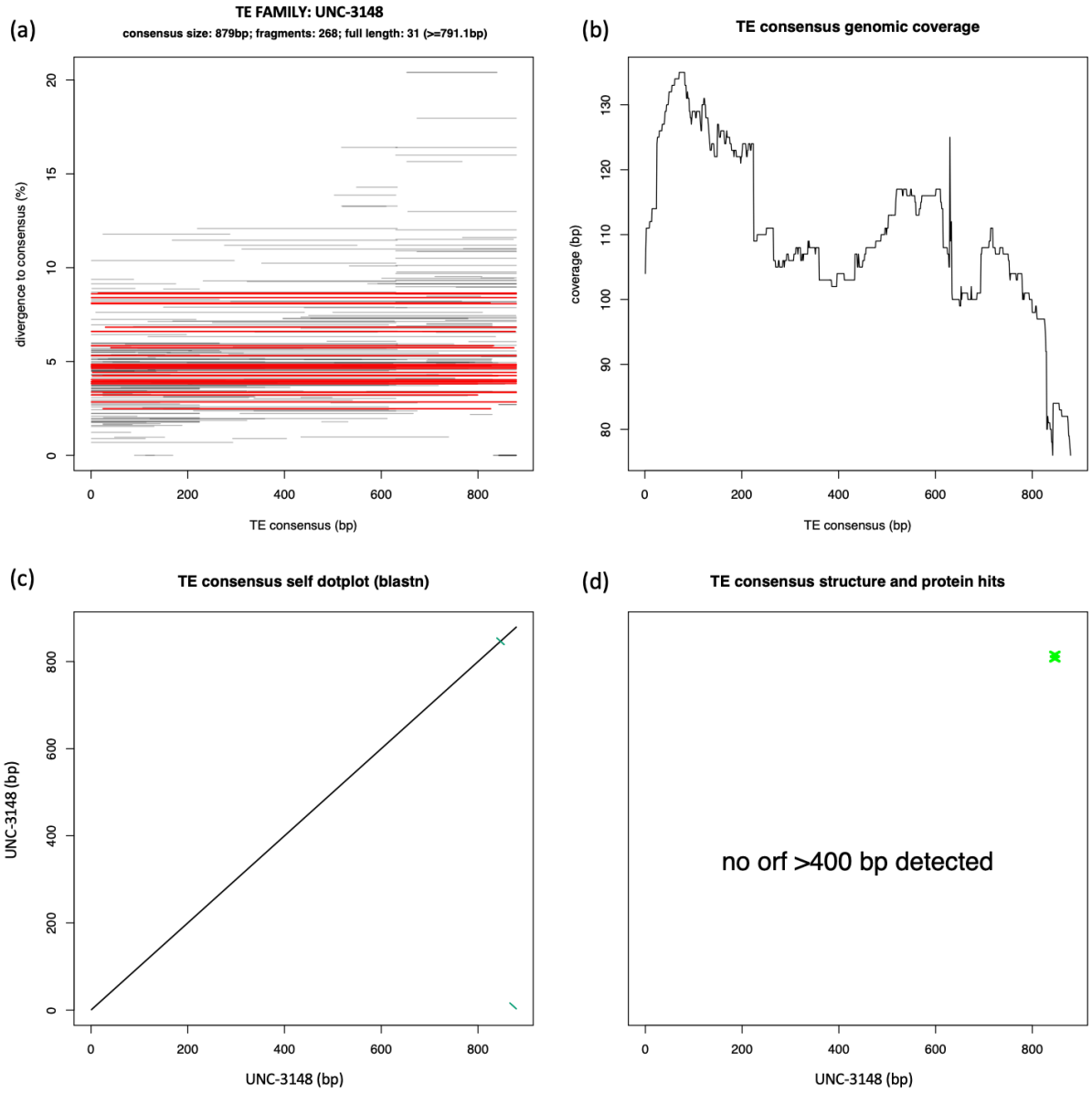
Supplementary Figure 10 – Output from TE-Aid (Goubert et al., 2022), (visualising the Unc-41 family residing in the ancestral LG2 haplotype in NEAC. **(a)** Fragment and divergence plot after blasting consensus sequence against gadMor3. Horizontal lines are genomic hits relative to the consensus, and the position on the y-axis represents divergence from consensus. Red lines are considered 'full-length' hits (> 90% of the consensus). **(b)** Sequence coverage of genomic hits from blast relative to the position along the TE consensus. **(c)** Self dot-plot of the TE consensus for revealing repetitive sequence patterns and structures (e.g., TIRs and LTRs) **(d)** Putative ORFs and corresponding peptides found in TE sequence. Arrows represent micro-homologies and the repetitive structures shown in the self-alignment in (c).



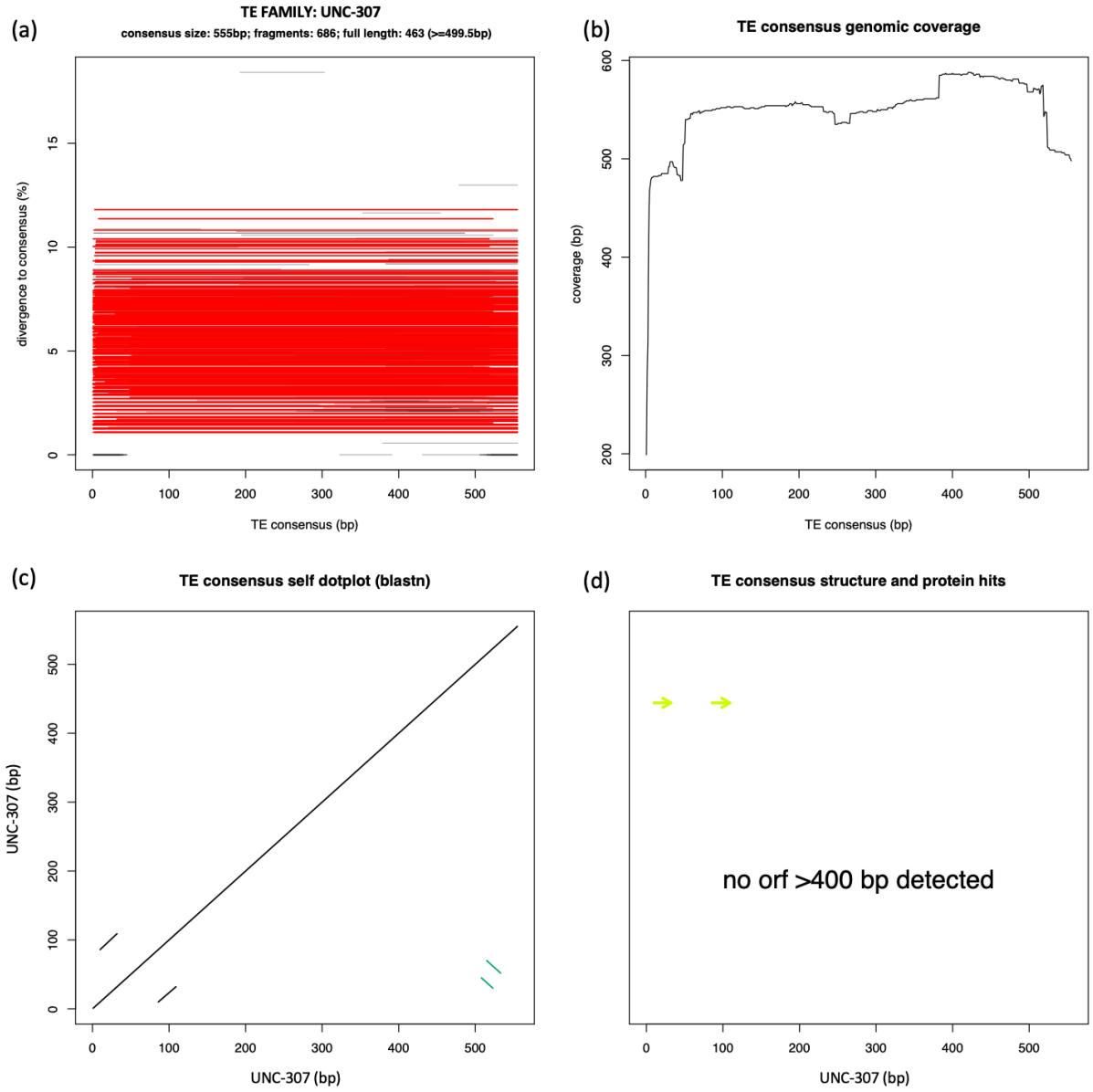
Supplementary Figure 11 – Output from TE-Aid (Goubert et al., 2022), (visualising the Unc-915 family residing in the ancestral LG2 haplotype in NEAC. **(a)** Fragment and divergence plot after blasting consensus sequence against gadMor3. Horizontal lines are genomic hits relative to the consensus, and the position on the y-axis represents divergence from consensus. Red lines are considered 'full-length' hits (> 90% of the consensus). **(b)** Sequence coverage of genomic hits from blast relative to the position along the TE consensus. **(c)** Self dot-plot of the TE consensus for revealing repetitive sequence patterns and structures (e.g., TIRs and LTRs) **(d)** Putative ORFs and corresponding peptides found in TE sequence. Arrows represent micro-homologies and the repetitive structures shown in the self-alignment in (c).



Supplementary Figure 12 – Output from TE-Aid (Goubert et al., 2022), (visualising the Unc-753 family residing in the ancestral LG2 haplotype in NEAC. **(a)** Fragment and divergence plot after blasting consensus sequence against gadMor3. Horizontal lines are genomic hits relative to the consensus, and the position on the y-axis represents divergence from consensus. Red lines are considered 'full-length' hits (> 90% of the consensus). **(b)** Sequence coverage of genomic hits from blast relative to the position along the TE consensus. **(c)** Self dot-plot of the TE consensus for revealing repetitive sequence patterns and structures (e.g., TIRs and LTRs) **(d)** Putative ORFs and corresponding peptides found in TE sequence. Arrows represent micro-homologies and the repetitive structures shown in the self-alignment in (c).

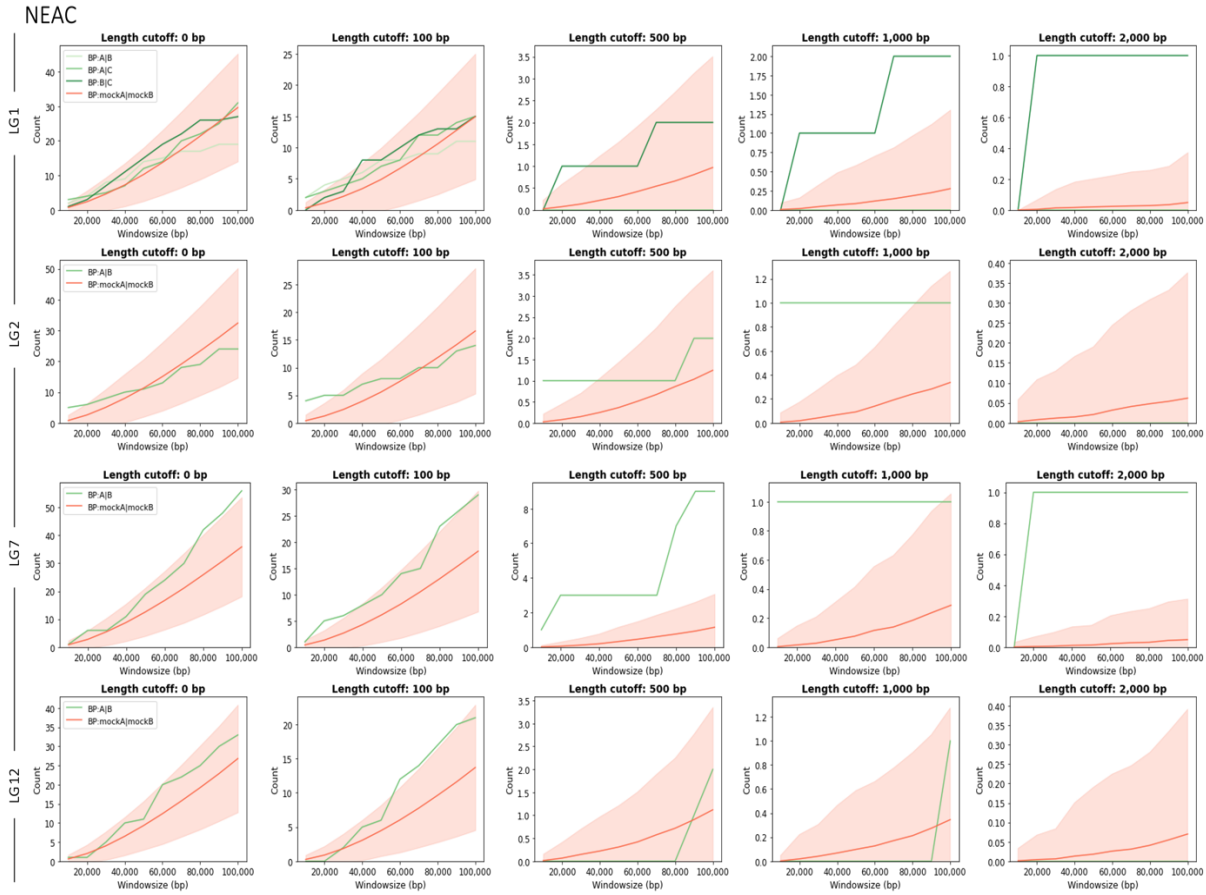


Supplementary Figure 13 – Output from TE-Aid (Goubert et al., 2022), (visualising the Unc-3148 family residing in the ancestral LG2 haplotype in NEAC. **(a)** Fragment and divergence plot after blasting consensus sequence against gadMor3. Horizontal lines are genomic hits relative to the consensus, and the position on the y-axis represents divergence from consensus. Red lines are considered 'full-length' hits (> 90% of the consensus). **(b)** Sequence coverage of genomic hits from blast relative to the position along the TE consensus. **(c)** Self dot-plot of the TE consensus for revealing repetitive sequence patterns and structures (e.g., TIRs and LTRs) **(d)** Putative ORFs and corresponding peptides found in TE sequence. Arrows represent micro-homologies and the repetitive structures shown in the self-alignment in (c).

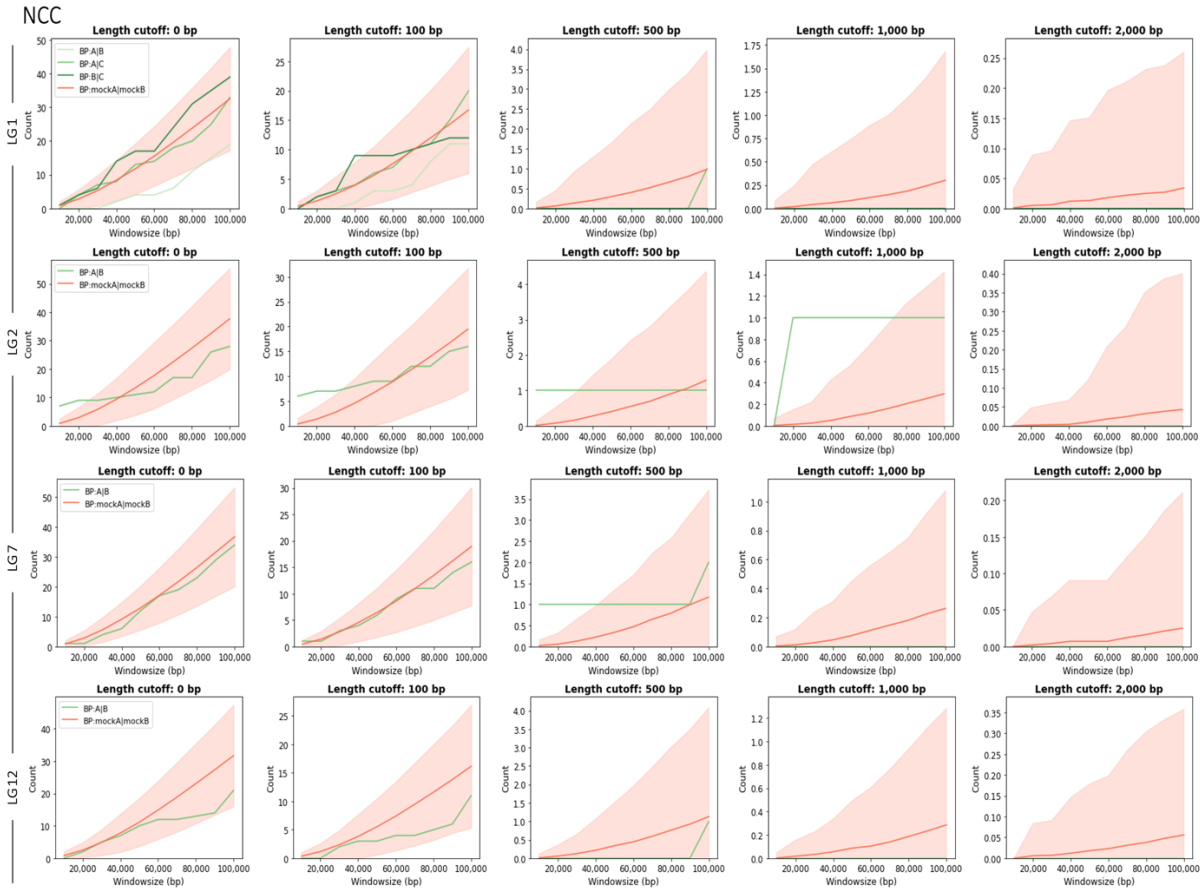


Supplementary Figure 14 – Output from TE-Aid (Goubert et al., 2022), (visualising the Unc-307 family residing in the ancestral LG2 haplotype in NEAC. **(a)** Fragment and divergence plot after blasting consensus sequence against gadMor3. Horizontal lines are genomic hits relative to the consensus, and the position on the y-axis represents divergence from consensus. Red lines are considered 'full-length' hits (> 90% of the consensus). **(b)** Sequence coverage of genomic hits from blast relative to the position along the TE consensus. **(c)** Self dot-plot of the TE consensus for revealing repetitive sequence patterns and structures (e.g., TIRs and LTRs) **(d)** Putative ORFs and corresponding peptides found in TE sequence. Arrows represent micro-homologies and the repetitive structures shown in the self-alignment in (c).

Statistical evaluations of finding breakpoint-related TEs



Supplementary Figure 15 – Plots for simulating the chance of finding related TEs (i.e., copies of the same family) in breakpoints compared to a chromosome-wide background on LGs 1, 2, 7 and 12 in NEAC. The count of related TEs of length 0 bp, 100 bp, 500 bp, 1000 bp and 2000 bp are plotted within window sizes of 10-100 kb. Green lines show the true count for related TEs within the increasingly larger window sizes around breakpoints. Red lines show the mean count of related TEs within 2000 randomly distributed 'mock' breakpoints on each of LGs 1, 2, 7 and 12. Standard deviations are shown within the pink shaded areas.



Supplementary Figure 16 – Plots for simulating the chance of finding related TEs (i.e., copies of the same family) in breakpoints compared to a chromosome-wide background on LGs 1, 2, 7 and 12 in NCC. The count of related TEs of length 0 bp, 100 bp, 500 bp, 1000 bp and 2000 bp are plotted within window sizes of 10-100 kb. Green lines show the true count for related TEs within the increasingly larger window sizes around breakpoints. Red lines show the mean count of related TEs within 2000 randomly distributed 'mock' breakpoints on each of LGs 1, 2, 7 and 12. Standard deviations are shown within the pink shaded areas.

SUPPLEMENTARY REFERENCES

Cabanettes, F., & Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *6*, e4958.

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 157.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238.

Goubert, C., Craig, R. J., Bilat, A. F., Peona, V., Vogan, A. A., & Protasio, A. V. (2022). Correction: A beginner's guide to manual curation of transposable elements. *Mobile DNA*, *13*(1), 15.

Kirubakaran, T. G., Andersen, Ø., Moser, M., Árnyasi, M., McGinnity, P., Lien, S., & Kent, M. (2020). A nanopore based chromosome-level assembly representing Atlantic cod from the Celtic sea. *G3 (Bethesda, Md.)*, *10*(9), 2903–2910.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews. Genetics*, *8*(12), 973–982.