 doi: 10.1111/mec.14111

Chironomus riparius (Diptera) genome sequencing reveals the impact of minisatellite transposable elements on population divergence

ANN-MARIE OPPOLD,*† HANNO SCHMIDT,† MARCEL ROSE,* SÖREN LUKAS HELLMANN,‡ FLORIAN DOLZE,‡ FABIAN RIPP,‡ BETTINA WEICH,‡ URS SCHMIDT-OTT,§ ERWIN SCHMIDT,‡ ROBERT KOFLER,¶ THOMAS HANKELN‡ and MARKUS PFENNINGER*†

Molecular Ecology Group, Institute for Ecology, Evolution & Diversity, Goethe-University Frankfurt am Main, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Hessen, Germany, †Senckenberg Biodiversity and Climate Research Centre, Georg-Voigt-Str. 14-16, 60325 Frankfurt am Main, Hessen, Germany, ‡Institute of Molecular Genetics, Johannes Gutenberg-University, Johann-Joachim-Becherweg 30a, 55128 Mainz, Rheinland-Pfalz, Germany, §Department of Organismal Biology and Anatomy, University of Chicago, 920 E. 58th Street, 1061C, Chicago, IL 60637, USA, ¶Institut für Populationsgenetik, Vetmeduni Vienna, 1210 Vienna, Austria

Abstract

Active transposable elements (TEs) may result in divergent genomic insertion and abundance patterns among conspecific populations. Upon secondary contact, such divergent genetic backgrounds can theoretically give rise to classical Dobzhansky–Muller incompatibilities (DMI), thus contributing to the evolution of endogenous genetic barriers and eventually causing population divergence. We investigated differential TE abundance among conspecific populations of the nonbiting midge *Chironomus riparius* and evaluated their potential role in causing endogenous genetic incompatibilities between these populations. We focussed on a *Chironomus*-specific TE, the minisatellite-like *Cla-element*, whose activity is associated with speciation in the genus. Using a newly generated and annotated draft genome for a genomic study with five natural *C. riparius* populations, we found highly population-specific TE insertion patterns with many private insertions. A significant correlation of the pairwise F_{ST} estimated from genomewide single-nucleotide polymorphisms (SNPs) and the F_{ST} estimated from TEs is consistent with drift as the major force driving TE population differentiation. However, the significantly higher *Cla-element* F_{ST} level due to a high proportion of differentially fixed *Cla-element* insertions also indicates selection against segregating (i.e. heterozygous) insertions. With reciprocal crossing experiments and fluorescent in situ hybridization of *Cla-elements* to polytene chromosomes, we documented phenotypic effects on female fertility and chromosomal mispairings. We propose that the inferred negative selection on heterozygous *Cla-element* insertions may cause endogenous genetic barriers and therefore acts as DMI among *C. riparius* populations. The intrinsic genomic turnover exerted by TEs may thus have a direct impact on population divergence that is operationally different from drift and local adaptation.

Keywords: endogenous selection, genome draft, insect genome, Pool-Seq, speciation, transposon

Received 14 September 2016; revision received 23 February 2017; accepted 6 March 2017

Correspondence: Ann-Marie Oppold,
E-mail: ann-marie.oppold@senckenberg.de

Introduction

Population divergence involves the accumulation of genetic differences across the genome, including protein-coding gene sequence, noncoding DNA, gene synteny, and large-scale genomic architecture (Maheshwari & Barbash 2011). Such divergence is often driven by ecological adaptation. However, endogenous genetic barriers as a result of intrinsic genetic incompatibilities can also impede neutral gene flow and thus substantially contribute to population differentiation (Bierne *et al.* 2011). Dobzhansky–Muller incompatibilities (DMI) result from negative genetic interactions between two or more loci (Dobzhansky 1937; Muller 1942), although the model does neither specify the molecular nature of the genetic elements that can lead to hybrid incompatibility (Castillo & Moyle 2012), nor whether genetic drift, local adaptation or endogenous sources of selection, such as genomic conflict or transposition processes, caused divergence. Transposable elements (TEs), which make up a large fraction of most eukaryotic genomes, have been proposed as major driver of genome evolution and attributed an important role in speciation (Kazazian 2004; Werren 2011). The idea that TEs can directly influence hybrid fertility is well supported in *Drosophila* (Kidwell 1985; Yannopoulos *et al.* 1987; Blumenstiel & Hartl 2005), where three different TEs (P-, I-, and hobo element; Kavi *et al.* 2005; Kofler *et al.* 2015; Simmons *et al.* 2015) are responsible for hybrid dysgenesis. Hybrids of *D. melanogaster* strains show elevated rates of mutation, chromosomal rearrangements, male recombination or sterility of females due to gonadal dysfunction (Bingham *et al.* 1982; Bucheton *et al.* 1984; Galindo *et al.* 1995). Similar TE-mediated hybrid breakdown phenomena have been observed in plants (Martienssen 2010) and vertebrates (Dion-Cote *et al.* 2014).

As TEs evolve rapidly, transpose throughout the genome and may lead to chromosomal rearrangements by ectopic recombination, it is reasonable to consider them as a common driver of genomic diversification among conspecific populations, eventually leading to speciation (Montgomery *et al.* 1991; Jurka *et al.* 2011). In general, TE insertions in the genome are deleterious (Pasyukova *et al.* 2004). Therefore, positive selection for mechanisms that suppress TE activity is to be expected (Simkin *et al.* 2013). When different TE suppression mechanisms evolve in two populations, hybrids may suffer from unstable genome constellations or uncontrolled TE spread, which could result in reduced hybrid fitness (Crespi & Nosil 2013). Besides the sensitive balance between TE proliferation and its suppression by the host genome, divergent insertion and abundance patterns of TEs might also create DMI in hybrids of different populations (Bierne *et al.* 2011). Hybridization among divergent populations may thus provoke substantial genomic stress (Fontdevila 1992).

Previous investigations of incompatibilities caused by TEs mostly focussed on classical TE types, such as transposase-encoding DNA transposons (e.g. P- and hobo element in *Drosophila*, Boussy & Periquet 1993) or LINE (long interspersed nuclear element) retrotransposons (e.g. I-element in *Drosophila*, Moschetti *et al.* 2010). Whether DMI, and hence reproductive isolation (RI), can also be caused by nonclassical TEs, which can be noncoding and organized as tandemly repeated, interspersed minisatellite clusters, remains unexplored. Such mobile repeat clusters might reduce fitness or even cause abortive development because they are often associated with chromosomal heterochromatin, thereby possibly affecting genomic stability (Ilkova *et al.* 2013), chromosome pairing (Hernandez-Hernandez *et al.* 2008), recombination (Talbert & Henikoff 2010) and expression of neighbouring genes by epigenetic position effects (Pezer & Ugarkovic 2012). During cell division, improper chromatid pairing may eventually lead to developmental problems and reduced hybrid fitness (Hailu *et al.* 1999; Oleszczuk & Lukaszewski 2014).

TEs have been invoked in the speciation of the morphologically cryptic nonbiting midge sister species *Chironomus riparius* and *C. piger* (formerly called *C. thummi thummi* and *C. t. piger*, Hankeln *et al.* 1994; Schmidt 1984). Depending on the crossing direction, two incompatibility syndromes involving hybrid dysgenesis have been described in these closely related midges, the Rud and the HLE syndrome (Hägele 1984, 1985, 1987, 1995; Hägele & Oschmann 1987; Hägele & Lachmann 1992; Hägele *et al.* 1995; Hägele & Kasper-Sonnenberg 2000). The latter also occurs on an intraspecific level in different *C. riparius* populations as aberrant traits such as reduced egg hatch and chromosome aberrations occurring with population-specific intensity (Hägele 1995). However, the molecular basis of the syndromes has so far not been investigated. On the genome level, a markedly different genome size and organization of repetitive DNA are suspected to contribute to the above syndromes (Schmidt 1984). The *C. riparius* genome is about 27% larger compared to that of *C. piger*, the latter presumably representing the ancestral chromosomal karyotype (Keyl 1965). One-fifth of the accumulated DNA in *C. riparius* is due to the transposable *Cla-element* sequence (Schmidt 1981). *Cla-elements* occur either as tandem-repetitive minisatellite-like clusters, often containing more than 20 repeat units (Hankeln *et al.* 1994), or as monomeric repeats that share characteristics with SINE (short interspersed element) retrotransposons (Hankeln & Schmidt 1987). The exact mechanism of transposition of these sequences is unclear, and it is unknown whether *Cla-elements* are still active. In *C. piger* and the more distantly related *C. luridus*, *Cla-element* clusters are mostly

restricted to centromeric regions, whereas in *C. riparius*, they additionally occur across most chromosomal arms (Schmidt 1984; Hankeln & Schmidt 1987; Hankeln *et al.* 1994). The AT-rich 120-bp sequence of the *Cla-element* has the potential to form hairpin structures and induce a strong sequence-directed DNA curvature, and these effects are expected to increase when *Cla-elements* build multimeric clusters (Schmidt 1981, 1984; Israelewski 1983; Hankeln 1990). Such structures may be particularly relevant with regard to potential effects on chromosome integrity, centromere formation and synapsis of homologous chromosomes. Comparable sequence-inherent curvature of the mouse major centromeric satellite DNA also seems important for heterochromatin formation (Radic *et al.* 1987). DNA hairpins are involved in several biological processes and may affect transcription, recombination and replication in pro- and eukaryotes (Mariappan *et al.* 1996; Bikard *et al.* 2010), as well as kinetochore formation in human centromeres (Catasti *et al.* 1994). In *C. riparius*, common chromosomal breakpoints were reported to colocalize with heterochromatic regions that contain blocks of tandem-repetitive DNA such as the *Cla-element* (Bovero *et al.* 2002). As the *Cla-element* potentially affects chromosomal architecture, function and genome evolution in *C. riparius*, its genomic spread might be involved in previously reported incompatibilities between the sister species *C. riparius* and *C. piger* (Schmidt 1984) and, therefore, a cause of speciation in the genus.

In this study, we attempted to elucidate whether diverging patterns of *Cla-element* distribution contribute disproportionately to population divergence and may result in DMI between populations by creating endogenous genetic barriers among *C. riparius* populations. To this end, we sequenced, assembled and annotated a *C. riparius* draft genome. In a genomewide population comparison, we investigated the population differentiation at single-nucleotide polymorphism (SNP) loci and at *Cla-element* insertion sites to see if the *Cla-element* distribution and frequency are more divergent than expected from SNPs between populations. We then hybridized geographically distant *C. riparius* field populations and investigated by in situ hybridization of *Cla-element* clusters to polytene chromosomes whether interpopulation hybrid individuals reveal chromosomal abnormalities that are consistent with DMI. In laboratory experiments, we tested whether we find evidence for reduced hybrid fitness.

Material and methods

Sequencing, assembly, scaffolding and annotation of a Chironomus riparius draft genome

Approximately 50 larvae of a *C. riparius* laboratory culture, established several decades ago, were used for extraction of genomic DNA. After quality checks of DNA, three paired-end and two mate-pair sequencing libraries of 3 and 5.5 kb insert size were prepared (Illumina, CA, USA) and run on Illumina HiSeq2500 and MiSeq instruments (Institute of Molecular Genetics, University of Mainz, Germany; StarSEQ, Mainz, Germany). Sequence reads were quality-checked, quality-processed and then assembled using the PLATANUS v1.2.1 pipeline that can effectively handle high-throughput data from heterozygous samples (Kajitani *et al.* 2014). We chose kmer sizes between $k = 32$ and $k = 84$. Iterative scaffolding of contigs was conducted by PLATANUS with default parameters and SSPACE 3.0 (Boetzer *et al.* 2011), the contig-extension option enabled. Completeness and integrity of the draft genome were examined by back-mapping of reads with BWA v0.7.10-r789 (Li & Durbin 2009) using the *bwa mem* algorithm, and by analyses using the tools BUSCO v1.1b1 (Simao *et al.* 2015) and REAPR 1.0.18 (Hunt *et al.* 2013).

The genome sequence was then annotated using three iterations of the MAKER2 v2.31.8 pipeline (Cantarel *et al.* 2008; Holt & Yandell 2011). For annotation of protein-coding genes, we supplied the pipeline with a reference transcriptome assembled from cDNA sequence data of embryonic, larval and adult *C. riparius* specimen and additionally with the Swiss-Prot database (obtained on 2016-01-13 from uniprot.org). Repeat detection was based on a custom repeat library. See Methods S1 (Supporting Information) for more information on the whole process of generating the 'CRIP_Laufer' draft genome sequence and annotation. The draft genome, including raw reads and assembly, is available at the European Nucleotide Archive (ENA) under the study Accession no. PRJEB15223 (raw data: ERS1460012-ERS1460016; assembly: ERS1474014). The annotation file as well as access to a *C. riparius* genome browser is available upon request.

Sampling of natural populations

We sampled a total of five natural *C. riparius* populations from across Europe (Table S1, Supporting Information). Larvae were collected alive to establish laboratory cultures, as well as preserved in ethanol for pooled sequencing. Species identification was done with COI barcoding (Pfenninger *et al.* 2007) of F_1 clutches from the laboratory culture and all preserved larvae. To rule out the misidentification of possible hybrids, a fragment of the nuclear-encoded mitochondrial 39S ribosomal protein L44 was added as nuclear marker. Barcoding of the two markers was performed as described in Oppold *et al.* (2016).

Pooled sequencing of natural populations

The Pool-Seq approach allowed us to obtain unbiased estimates of allele frequencies in the entire genome by sequencing a population pool of many (>100) diploid individuals to relatively low coverage (30×), thus making probable that each locus per chromosome in the pool is sequenced only once (Futschik & Schlötterer 2010). Population genomic data were obtained of each of the five natural *C. riparius* populations (Table S1, Supporting Information). We dissected head capsules from all barcoded larvae of the respective populations, pooled them and extracted the genomic DNA using the DNeasy Blood & Tissue Kit (QIAGEN, Hilden, Germany). DNA concentration was measured with the Qubit® dsDNA BR Assay Kit in a Qubit® fluorometer, and quality was assessed by gel electrophoresis. Sequencing libraries of the individual pools were constructed using the TruSeq DNA Nano Library Prep Kit (Illumina, CA, USA) tagged with different multiplex barcodes (StarSEQ, Mainz, Germany). For each library, the same insert size was selected, ranging from 180 to 680 (average 380). All pools were sequenced as 100-bp paired-end sequences on an Illumina HiSeq 2500 platform (Institute of Molecular Genetics, University of Mainz, Germany). Adapter clipping and end-quality trimming on the raw sequences were performed with TRIMMOMATIC (ILLUMINACLIP:adapters.fa:2:30:10:8, Bolger *et al.* 2014) using a sliding-window approach (window size 4, Phred quality 20). Trimmed reads were inspected with FASTQC (v0.11.2; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Inference of population differentiation and heterozygosity from genomewide SNPs

To estimate the genomewide population differentiation and heterozygosity based on SNPs, we followed the POPOOLATION2 pipeline (Kofler *et al.* 2011b). To compare the results, mapping of Pool-Seq reads of the five natural populations was performed with two alternative BWA mapping algorithms. First, we used *bwa aln* with a maximum number of 1 gap per alignment, and a mismatch probability of 0.01 without seeding as recommended by Kofler *et al.* (2011a). Paired-end information was then linked with *bwa sampe*. Better mapping success, however, was retrieved with the *bwa mem* algorithm and an adapted seed size (*-k 30*). With a seed size of 30, we still obtained a high stringency level whilst increasing the amount of properly paired reads and drastically speeding up the analysis. Mapping success was inspected with QUALIMAP v2.0 (Okonechnikov *et al.* 2016). Downstream processing of the bam files strictly followed the pipeline of POPOOLATION2 (Kofler *et al.* 2011b). Mean coverage was calculated with R v3.2.3 (R-Core T 2015) and *samtools depth* (SAMTOOLS utilities; version 1.1, Li *et al.* 2009). To ensure comparable mean coverage between the data, the bam file of the Lorraine population (NMF) was downsampled to the average coverage of the population means with PICARD (DownSampleSam.jar; v1.119, available at http://picard.sourceforge.net). SNP calling was performed on the sync file that combined all five Pool-Seq data sets. To calculate population-specific heterozygosity, we called SNPs with the *snp-frequency-diff.pl* script of the POPOOLATION2 package (Kofler *et al.* 2011b) with a minimum count of 4, and a coverage ranging from a minimum of 20 to the upper 2% of the respective data set (i.e. MF: 50, MG: 37, NMF: 49, SI: 50, SS: 45). From the output (rc-file), we calculated allele frequencies and heterozygosity for all population-specific SNPs, as well as for population-private SNPs. To estimate the pairwise F_{ST} for each SNP between all possible population pairs, we used the *fst-sliding.pl* script of the POPOOLATION2 package (parameters: –min-count 4 –min-coverage 20 –max-coverage 50, 37, 49, 50, 45 –min-covered-fraction 1.0 –window-size 1 –step-size 1 –pool-size 168:112:105:155:118 –suppress-noninformative). Mean heterozygosity and mean F_{ST}s for each pairwise comparison were then calculated in R.

BLAST analysis with Cla-element consensus sequence

As required for the POPOOLATIONTE package (v51, Kofler *et al.* 2012), we prepared a custom library of 108 *Chironomus*-specific transposable elements and repeats, including the 117-bp *Cla-element* consensus sequence (Schmidt 1984; Appendix S1, Supporting Information). To test whether all *Cla-element* variants present in our data can be mapped to this *Cla-element* consensus sequence, we first performed a BLAST search (Altschul *et al.* 1990) with the five Pool-Seq data sets blasted against all available *Cla-element* sequences from NCBI (gi|6456818, gi|7099, gi|89994172, gi|89994177, gi|89994178, gi|89994179; BLAST options: -outfmt 6 -max_target_seqs 100000000 -evalue 1e-30). The output of each population to each *Cla-element* sequence was filtered for a minimum alignment length of 80 bp. We then merged the filtered output that contained a single BLAST hit per population to at least one *Cla-element* sequence (filtering duplicates with *find-uniq-BLAST-hits.pl*, Dryad doi:10.5061/dryad.39s6k/1). This resulted in a file containing all the *Cla-element* variation present in our five data sets; 98.4% of these data could successfully be back-mapped (*bwa bwasw*, see below) to the *Cla-element* consensus sequence in our TE library, which demonstrated that this sequence is a suitable reference for our study.

Estimating TE abundance and frequency with POPOOLATIONTE

The custom TE library (Appendix S1, Supporting Information) was first used for masking the genome with REPEATMASKER v4.0.5 (Smit *et al.* 2013-2015). The masked genome was then combined with the TE library and further used as combined reference sequence to map Pool-Seq data sets. We generated a TE hierarchy (Appendix S2, Supporting Information) by obtaining the insert name, order, family and subfamily for every entry in the TE library, as required input for POPOOLATIONTE. Paired-end sequence reads of the five Pool-Seq data sets were mapped as singletons to the combined reference sequence with BWA's Smith–Waterman alignment algorithm *bwasw* (Li & Durbin 2010). Paired-end information was post hoc recovered with the *samro.pl* script from POPOOLATIONTE. The resulting paired-end bam files were subsampled to an equal number of 50 million reads using *samtools view* in order to avoid biased estimations of TE abundance when comparing the different populations. Next, we sorted the five sam files (*samtools sort*) and identified and localized TE insertions and their frequency in each sample separately with POPOOLATIONTE (Kofler *et al.* 2012). The exact genomic localization of TE insertions estimated by POPOOLATIONTE may be slightly inaccurate and thus result in a slight variation of the genomic position of identical TE insertions in different Pool-Seq data sets. In order to compare TE abundance between samples, we clustered TE insertion sites from the five bam files within a specific genomic window to one single TE insertion using a custom script (Appendix S3, Supporting Information). For the identification of the optimal window size, we calculated the exponential one-phase decay model of increasing window size with the number of TE insertions in GraphPad Prism® (version 5, GraphPad Software, San Diego, USA; Fig. S1, Supporting Information). A plateau was reached at a window size of about 450 bp for a model using all TEs excluding the *Cla-element*, as well as for the model using solely the *Cla-element*. All downstream analyses were therefore based on homogenization of TE insertions within a 450-bp window across populations. Please note that all population genetic analyses of TE insertions were solely based on the frequency of their presence/absence in a population, thus making them perfectly comparable to biallelic SNP markers.

Based on the genome annotation file, we investigated whether TE insertions were localized close to coding regions, in introns or exons (*PosinGenes.py*: Dryad doi:10.5061/dryad.39s6k/3, *PosinExons.py*: Dryad doi:10.5061/dryad.39s6k/2).

Cla-element cluster size estimation

The *Cla-element* is a tandem-repetitive element, building large clusters in the genome of *C. riparius* (Hankeln *et al.* 1994). With POPOOLATIONTE, it is, however, not possible to distinguish monomer insertions from clusters (Kofler *et al.* 2012). To quantify the mean *Cla-element* cluster size, we first predicted an expected number of reads per monomer insertion by combining the number of identified insertion sites with the fraction of reads that cover one monomer and the genomewide mean coverage adjusted to the mean *Cla-element* frequency (Table 1, calculation steps outlined in column 1). We extracted the reads mapping to the *Cla-element* directly from the mappings to the TE combined reference (using *samtools view*). Comparison of observed to expected values finally gave an average cluster size estimate. We randomly selected low-frequency *Cla-element* insertions from the POPOOLATIONTE output and validated them visually with the INTEGRATIVE GENOMICS VIEWER v2.3.68 (Robinson *et al.* 2011; Thorvaldsdottir *et al.* 2013).

Scan for Cla-element transcripts

As an additional line of evidence for potential ongoing activity of the *Cla-element*, we scanned the publicly available transcriptome data sets SRR834592 and SRR834593 (Schmidt *et al.* 2013) for *Cla-element*-containing transcripts by mapping those reads to the consensus sequence of the *Cla-element* dimer with the *bwa bwasw* algorithm.

Analysis of population-private Cla-element insertions

To visualize the overlap of *Cla-element* insertions between the five populations, we plotted Venn diagrams on the basis of all *Cla-element* insertion sites (online tool: http://bioinformatics.psb.ugent.be/webtools/Venn/ accessed 29.01.2016). The numbers of population-private *Cla-element* insertions suggested a pattern that might be linked to geographical location or environmental conditions. Previous studies in *C. riparius* have revealed that genetic diversity and the population mutation parameter theta are correlated with ambient temperatures (Oppold *et al.* 2016). Considering new TE insertions as mutations in a classical sense, the rate of TE transposition might also correlate with temperature. Therefore, we tested population-private *Cla-element* insertions as pairwise comparisons (chi-square tests with Benjamini–Hochberg correction for multiple comparisons in R) and their relation to the annual mean temperature of the respective natural population site (climate data were derived from the WorldClim database; Hijmans *et al.* 2005) as a nonlinear exponential

Table 1 Estimation of *Cla-element* cluster size and genomewide copy number in the different *Chironomus riparius* populations (MF: Rhone-Alpes, MG: Hessen, NMF: Lorraine, SI: Piemont, SS: Andalusia)

		MF	MG	NMF	SI	SS
A	Insertion sites	849	538	807	803	772
B	Mapped reads to Cla (observed value)	1 918 365	1 494 769	2 117 883	1 761 580	1 889 805
C	Mean read length	79.0	75.5	75.5	79.0	75.5
D	Fraction of monomer coverage by one read	1.48	1.55	1.55	1.48	1.55
E	Genomewide mean coverage	24	23	22	24	23
F	Mean Cla frequency	0.84	0.87	0.84	0.84	0.82
G = A*D*E*F	Predicted number of reads per monomer (expected value)	25 390	16 710	23 163	24 078	22 476
H = B/G	Predicted mean cluster size (observed/expected)	75.6	89.5	91.4	73.2	84.1
I = B/(D*E*F)	Estimated genomewide copy number	64147.2	48126.3	73786.1	58748.5	64909.4

one-phase association (best fit model) in GraphPad Prism®. A nonlinear one-phase association model accounts for an expected upper threshold of new TE insertions due to the limited genome size and the evolution of a transposition–selection balance (Barron *et al.* 2014).

Analysing population differentiation and heterozygosity from genomewide SNPs, TEs and Cla-elements

Based on the frequency estimates of all TE insertions among populations, we calculated expected heterozygosity and pairwise F_{ST} for TEs (excluding *Cla-elements*) and solely for *Cla-element* insertions. F_{ST} was calculated as the reduction in heterozygosity in subpopulations (H_S) compared to heterozygosity of the total population (H_T): $F_{ST} = (H_T - H_S)/H_T$. As mentioned above, the biallelic nature of TE insertions (present/absent) makes F_{ST} estimated from them directly comparable to the same values of biallelic SNP markers.

To test the effect of drift on different genomic elements, we correlated mean pairwise F_{ST} calculated from TEs and *Cla-elements* with mean pairwise F_{ST} from genomewide SNPs. Isolation by distance (IBD) was tested for by correlating each of the three groups of mean pairwise F_{ST} with the geographical distance between the sampling sites of populations. Statistical significance was assessed with Mantel's tests with 9999 random permutations (ade4 package, R).

Mean heterozygosity was calculated for TEs, and *Cla-elements*, respectively, in order to compare them to genomewide heterozygosity calculated for SNPs. Heterozygosity reveals the proportion of segregating sites for SNPs, TEs or *Cla-elements*. We also compared the proportion of fixed population-private TE or *Cla-element* insertions (insertions that are invariant in one population, i.e. all individuals of a population carry this insertion, and absent in all other populations) to fixed population-private SNPs. If only neutral drift governed the evolution of TEs, different types of genetic elements (*Cla-elements*, TEs or SNPs) should show equal heterozygosity and proportion of population-private loci. Statistical significance of differences in mean heterozygosity and proportion of fixed population-private insertions between the three groups (*Cla-elements*, TEs, SNPs) was tested using repeated-measures ANOVA with Tukey post hoc test.

To infer whether potential endogenous selection on TEs affected levels of population divergence at SNPs in their vicinity, we compared the mean F_{ST} of SNPs within 50-, 100- and 500-bp upstream and downstream of *Cla-elements* and other TE insertion sites with the F_{ST}s from 10 000 randomly chosen SNPs from other locations in the genome, respectively. These intervals were chosen because, in *Drosophila melanogaster*, decay of local linkage disequilibrium is strongest within the first 200 bp (Franssen *et al.* 2015), and due to similar genetic constitutions (genome size, chromosome number, population size), this might also apply to *C. riparius*.

Hybridization experiment

To experimentally investigate potential intraspecific RI, we hybridized individuals from the three geographically most distant populations (MG – Hessen, SI – Piemonte, SS – Andalusia). Egg clutches of the natural populations were barcoded (see above) to start pure *C. riparius* laboratory cultures that were acclimatized to standard conditions for at least three generations. The hybridization experiment comprised two subsequent experimental generations (F_0-F_1), whereas the hybridization happened during the adult stage of F_0 (Fig. S2, Supporting Information). During the F_0 of the

experiment, we hybridized the populations in both crossing directions and investigated their fitness (clutches per female, eggs per clutch, fertility of clutches). In F_1 generation of the experiment, that is first hybrid generation, we screened for indications of hybrid dysgenesis by following life-trait parameters: mortality, mean emergence time (EmT_{50}), clutches per female, eggs per clutch, fertility of clutches and population growth rate (PGR). The resulting offspring of the experimental F_1 therefore is the second hybrid generation. Life-trait parameters of the different experimental populations were statistically compared via one-way ANOVA with Tukey post-test (mortality, EmT_{50}, PGR) in GraphPad Prism®. Population-specific life-trait parameters (clutches per female, fraction of fertile clutches) were compared as contingency tables with Fisher's exact test in R because there was only a single value per population: ratio between clutches per female or fraction of fertile clutches and the total sum of clutches per population, with the null hypothesis that ratios between populations are the same.

To start F_0, we raised 360 larvae of each of the three parental pure populations, distributed over six replicates and under constant conditions at 20 °C (Oppold *et al.* 2016). Larvae were fed with a developmental stage-adjusted amount of ground TetraMin (Tetra GmbH, Germany). After emergence, six hybrid populations and the parental pure populations (summing up to nine populations in total) were started by, respectively, combining 50 male and 50 female midges in reproduction cages. Egg clutches were removed daily from the reproduction cages and stored separately in six-well plates with medium under the same conditions to document the early development of each clutch and define its fertility with the successful hatch of larvae as criteria. F_1 was started when all clutches of the populations were hatched. L1 larvae of a respective population were pooled, and 200 larvae were distributed over five replicates during larval development. After emergence, again 50 individuals of each sex per population were transferred to respective reproduction cages.

Fluorescent in situ hybridization of Cla-elements

Polytene chromosomes of *C. riparius* were dissected from salivary glands of L4 larvae from the F_1 generation of the hybridization experiment. We sampled ten larvae from each of the nine experimental populations, that is noncrossed pure individuals as well as hybrids from all crossing directions. The preparation of chromosomes and in situ hybridization of *Cla-elements* followed Hankeln & Schmidt (1987) and Schmidt *et al.* (1988). To guarantee the equal treatment of samples, all were prepared in one batch. As a *Cla-element* probe, we used the plasmid clone pAD 2903 containing a *Cla-element* dimer (Hankeln 1990), which was DIG-labelled (DNA Labeling Kit©, Roche Life Science, USA). FITC-conjugated anti-DIG (Roche Life Science, USA) was used as an antibody to detect hybridization sites. Chromosomes were DAPI-stained (Roche Life Science, USA). Preparations were inspected with a fluorescence microscope system (Olympus BX61 + Olympus BX UCB, Olympus, Japan). Pictures of DAPI-stained chromosomes were overlaid by pictures of the *Cla-element* FISH in Photoshop CS6 Extended (Adobe, USA).

Results

Chironomus riparius draft genome

We generated about 350 million Illumina sequence reads from five sequencing libraries. On average, 71.5 % of the reads passed quality filtering steps, resulting in a total 250 million reads and approximately 130× genome coverage (Table S2, Supporting Information). The assembly resulted in 41 974 contiguous sequences ≥500 bp with a GC content of 31%. These contigs could be linked to yield 5292 scaffolds ≥1000 bp with an N50 value of 272 065 bp (NG50 = 227 750). The total length of the draft genome is 180 652 019 bp (16 % N's, Table S3, Supporting Information), which covers about 90% of the estimated genome size of 200 Mb (Schmidt-Ott *et al.* 2009). On average, 91.9% of the reads used for assembly could successfully be mapped back to the draft genome (Table S4, Supporting Information). The draft genome contains 93% of a set of 2675 core arthropod genes (Table S5, Supporting Information). Using read pair information, 550 (0.1%) scaffolding positions were identified that cannot be resolved unambiguously.

The genome was annotated using a reference transcriptome assembled from 14 cDNA data sets (Table S6, Supporting Information). We found predictions for 13 093 protein-coding genes with an average of 4.7 exons per gene. The exons have an average length of 355 bp and all exons together sum up to 11% of the whole genome (Table S7, Supporting Information). Additionally, 71 990 regions were marked as repetitive element sequences, including 582 potential *Cla-element* positions.

Cla-element sequence variability

When analysing the *Cla-element* sequences obtained from the different populations, we did not find a population-specific nucleotide sequence pattern. We randomly selected 100k *Cla-element* sequences that produced a BLAST hit against one of the *Cla-element*

sequences from the database. These were aligned to the *Cla-element* reference sequence of the custom TE library and the 95% consensus alignment comprised a sequence length of 664 bases containing 430 gaps (data not shown). Thus, variability between *Cla-element* monomers was very high, both on the level of nucleotide substitutions and for insertions and deletions. We did not find any sequence clustering, which would be indicative of *Cla-element* subfamily structure. Due to large variation of homopolymer stretches across the *Cla* sequence, no position of the sequence was invariant throughout the alignment.

Transposable element families in Chironomus riparius

We identified 19 different TE families summing up to an overall number of 5645 insertion sites within all five natural *C. riparius* populations (Fig. 1). The abundance of TE families differed among populations. Alu, Hind and *Cla* minisatellite-like families, as well as the CTRT1 SINE element, the NLRCth1 retrotransposon and the TECth1 DNA transposon, each occurred at more than 100 insertion sites across the genome in all populations (Fig. 1). The majority of TE insertions were fixed within populations (disregarding the number of empty sites that arise from comparisons between populations, Fig. S3, Supporting Information). However, in all populations we also found low to intermediate insertion frequencies for all TE families, except for the coding tandem repeats originating from Balbiani Ring genes. A subtelomeric repeat known from *C. dilutus* (gi|6525125 - gi|6525127) could not be localized in the draft genome but half of the sequence was found to be associated with the *C. thummi* DNA for TsB and TsC telomeric tandem repeat (gi|9557882 and gi|9557883), as one read mate mapped to either of the two sequences in the TE library.

Among all populations, there were only eight TEs inserted close to or into gene regions (Table S8, Supporting Information), which is significantly less than expected by chance (Table S9, Supporting Information). Besides insertions into pseudogenes or weak support for the insertion due to low frequencies, only two insertions were localized in predicted genes: the Balbiani Ring repeat was found in the Balbiani Ring protein 2 gene as expected and an Alu element was detected in a retrovirus-related polymerase polyprotein gene (Table S8, Supporting Information).

TE frequency distribution

In line with the result for all TEs in general, the *Cla-element* was fixed at most insertion sites (Fig. S3, Supporting Information). However, intermediate- and low-frequency insertions were also identified (Fig. S4, Supporting Information) and the mean frequency of segregating sites ranged from 0.57 to 0.64 in the different *C. riparius* populations. We visually validated most of the low-frequency insertions to approve a sufficient coverage and appropriate mapping results in IGV. In all populations, we found low-frequency insertions below 0.25: 16 sites in MG, 16 sites in NMF, 21 sites in MF, 18 sites in SI and 50 sites in SS. For the two southern populations (SI – Piemont, SS – Andalusia), *Cla-element* insertions at a frequency of 0.004 were detected in regions with particularly high genome coverage (>1000×) likely resulting from copy number variations.

Estimated Cla-element cluster size

Absolute numbers of *Cla-element* insertions varied between populations, ranging from 539 insertions in the Hessian population (MG) to 803 insertions in the Piemonte population (SI, Fig. 1). Based on these numbers of *Cla-element* insertions per population, and the number of reads that were mapped to the *Cla-element* sequence in the TE library, we estimated the mean *Cla-element* cluster size to be on average 82.8 monomers, ranging from 73.2 (SI) to 89.5 (MF) monomers (Table 1, factor H). The genomewide number of *Cla* monomer copies was estimated upon the number of mapped reads, ranging between 48 126 in MG and 73 786 in NMF (Table 1, factor I).

Cla-element transcripts

Low-frequency insertions (here <0.25) might be due to recent transposition activity of the element or purifying selection acting on the element. To test for the possibility of ongoing transposition activity, we searched RNA transcriptome data for evidence of *Cla-elements* transcription, as these might be expected for active SINE-like retrotransposons. The two 454 transcriptome data sets from the SRA contained 29 and 27 *Cla-element* sequence reads (representing 0.01% of the data), most of them present as dimers, none showing traces of polyadenylation or other adjacent mRNA sequences (Table S10, Supporting Information).

Population-private Cla-element insertions

By comparing the *Cla-element* insertion sites between the different populations, we obtained presence–absence patterns. From a total of 1341 *Cla-element* loci, 329 (96 of them fixed) were shared among all populations and are therefore likely ancestral *Cla-element* insertion sites for the species (Fig. S5, Supporting Information).

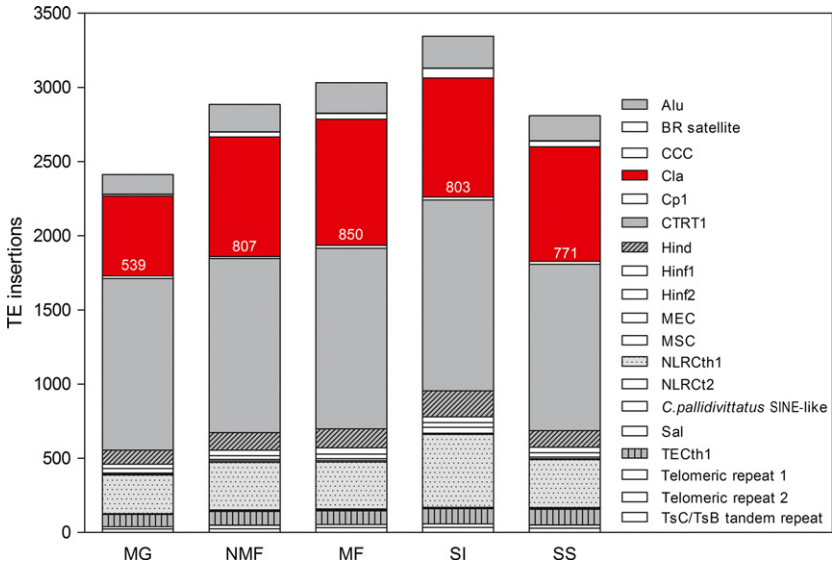

Fig. 1 Numbers of detected loci of all transposable element families as represented in the custom TE library (Appendix S1, Supporting Information) across the whole genome of *Chironomus riparius*. Orientation of the legend refers to the order of the vertical stacked bars. *Cla-element* loci are highlighted in red and numbers of insertion sites are given. All TEs with >100 loci are highlighted in grey and/or pattern.

The relation between population-private and total *Cla-element* loci differed significantly between populations (Fig. S6, Supporting Information). Only 23 private *Cla-element* insertions were identified in Hessen (MG), whilst the highest number of 118 private insertions was found in the Andalusian population (SS), suggesting a clinal association. The two northern populations had significantly less private *Cla-element* insertions than the three remaining populations from sites further south. Relating the number of population-private *Cla-element* insertions to the annual mean temperature of the respective location showed a strong exponential one-phase association (regression $r^2 = 0.84$, Fig. S6, Supporting Information).

Population differentiation and heterozygosity from SNPs, TEs or Cla-elements

To investigate the structure of the five geographically widespread populations, we estimated the population differentiation based on different genomic elements (Fig. S7, Supporting Information). On the basis of genomewide SNPs, the mean population differentiation turned out to be very low with the F_{ST} ranging between 0.025 and 0.07, whereas the level of differentiation was at least fourfold higher when considering mean pairwise F_{ST} calculated from TEs (0.22 to 0.28). This level of differentiation was even higher when considering the *Cla-element* alone (maximum $F_{ST} = 0.31$ in population pair SI: SS, i.e. Piemont and Andalusia). There was no significant difference or visible trend regarding the levels of differentiation in close vicinity to *Cla-elements* or TEs compared to randomly drawn SNPs from other genome positions.

We found a significant correlation between mean pairwise F_{ST} calculated from SNPs and calculated from TEs or *Cla-elements* (TEs: $P = 0.0322$, *Cla-elements*: $P = 0.0183$, Fig. 2A). Correlating geographical distance with the mean pairwise F_{ST} did not reveal an IBD pattern for any genomic element (Fig. 2B).

Mean population heterozygosity of private *Cla-element* insertions was significantly lower compared to mean population heterozygosity of private TE insertions excluding *Cla-element*s, and private genomewide SNPs (Fig. 3A). On average, within all populations more than 50% of private *Cla-element* insertions were fixed (Figs 3B and S8, Supporting Information). This proportion is significantly higher than the ratio of fixed (relative to all) population-private TE insertions and genomewide SNPs (Fig. 3B). On average, less than 30% of population-private TE insertions, excluding *Cla-element*s, were fixed. This is, however, still significantly higher than the fixation of population-private genomewide SNPs (~5%).

Chromosomal rearrangements in hybrid individuals

FISH of the *Cla-element* probe was successfully performed for multiple chromosomal sets of each pure and hybrid population (see Table S11, Supporting Information for detailed numbers). As expected for *C. riparius*, *Cla-element* bands were distributed across all arms of all chromosomes (Fig. S9, Supporting Information). To identify differences in banding patterns between populations, all chromosomes were carefully inspected by multiple investigators of this study. To rule out artefacts of sample preparation or differential staining, we required each of our findings to be supported by at least two independent chromosomes. Copy number variations (CNV) inside *Cla-element* clusters were defined as an obvious difference in bandwidth between the two homologous chromosomes. If the *Cla-element*

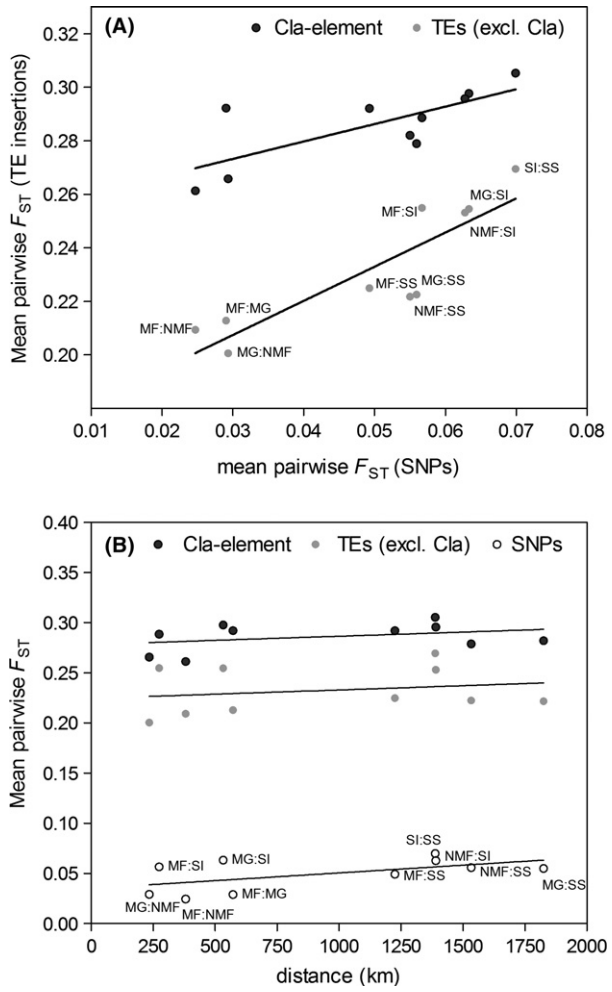

Fig. 2 (A) Mean pairwise F_{ST}s measured with genomewide SNPs correlated to mean pairwise F_{ST}s on the basis of TE insertions (Mantel's test $P = 0.0322$) and *Cla-element* insertions (Mantel's test $P = 0.0183$). (B) Isolation-by-distance pattern as correlation of the geographical distance between the populations with mean pairwise F_{ST}s on the basis of genomewide SNPs (ns), TE insertions (ns), *Cla-element* insertions (ns).

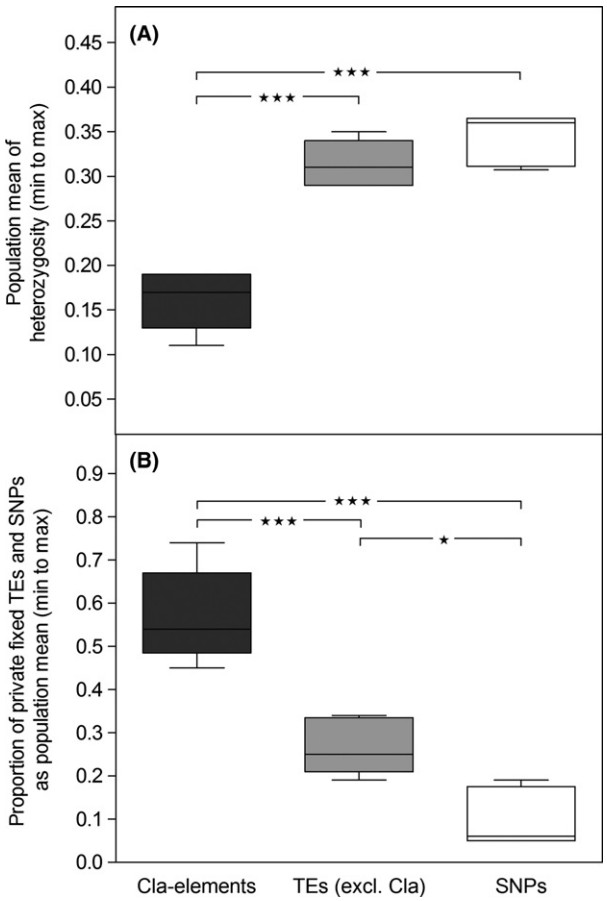

Fig. 3 (A) Population mean of heterozygosity on private *Cla-element* insertions, TE insertions excluding *Cla-elements*, and genomewide SNPs, respectively. Heterozygosity is significantly lower for *Cla-element*s compared to TEs and SNPs (repeated-measures ANOVA with Tukey post hoc test, significance levels shown as asterisks). (B) Population mean of the proportion of private fixed to all private *Cla-element* and TE insertions, and SNPs (repeated-measures ANOVA with Tukey post hoc test).

locus was clearly restricted to only one of the homologous chromosomes, we assumed a heterozygous site. In all hybrid populations, we found CNVs or heterozygous *Cla-element* bands that were either homozygous or absent in the pure parental populations, respectively (Fig. 4, marked by asterisks and arrows). Most of these structural differences occurred close to pericentromeric regions of chromosome I or II. Furthermore, partial asynapsis of homologous chromosomes was documented twice on chromosome II in different individuals of the hybrid population MG×SI (Fig 4, circled). Both chromosomal arms were affected by this aberration, which occurred in conjunction with multiple CNVs and heterozygous *Cla-element* bands. A similar effect could be documented on chromosome I of the hybrid population MG×SS, albeit only in a single individual (Fig. 4, bottom).

Fertility of hybrid populations

The three natural *C. riparius* populations were successfully crossed in both directions resulting in six hybrid populations. Mortality, EmT_{50}, PGR and number of eggs per clutch did not generally differ between pure and hybrid populations (Fig. S10, Supporting Information). Mating between individuals from different populations did not affect the fertility of resulting egg clutches (Fig. 5, top left). However, fertility of clutches in the following hybrid offspring generation (F_1 of the experiment) of SI♂×SS♀ and SS×MG (both crossing directions) decreased significantly compared to at least

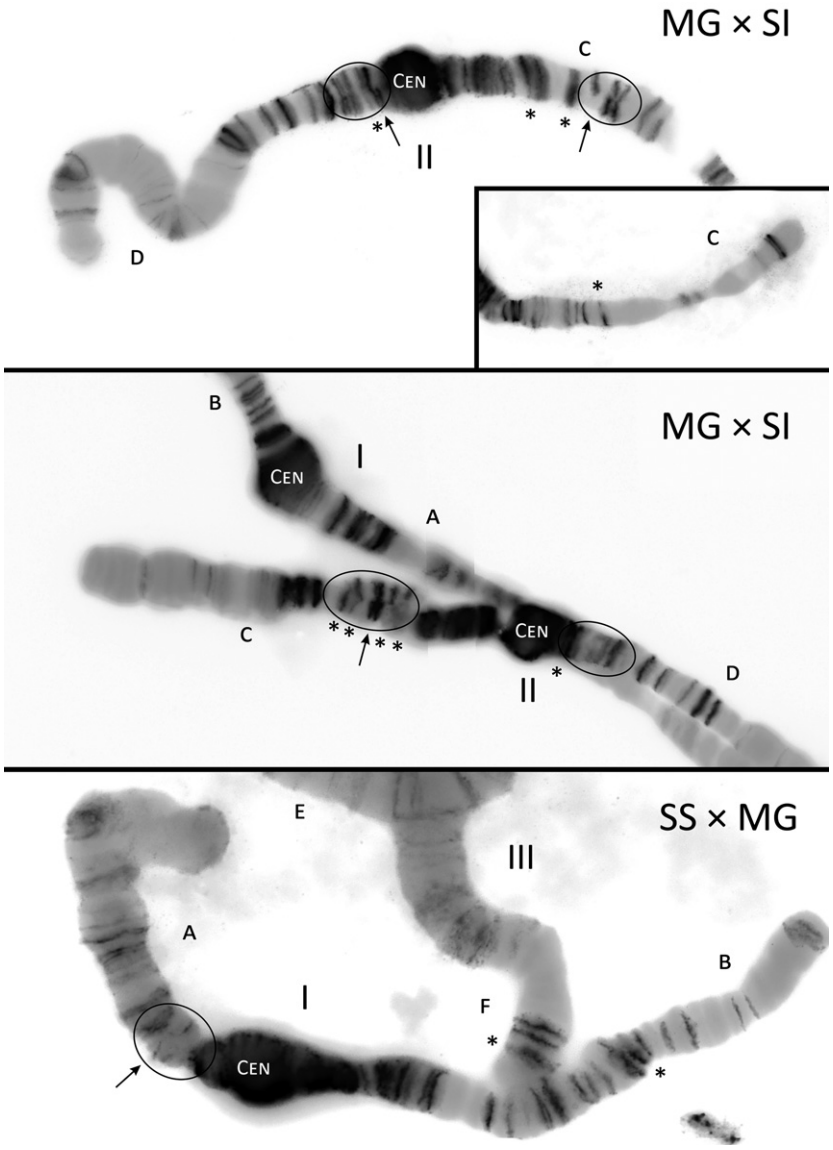

Fig. 4 Exemplary fluorescent in situ hybridization (FISH) micrograph of the *Cla-element* probe hybridized to polytene chromosomes of different *Chironomus riparius* hybrid populations. Pictures of DAPI-stained chromosomes were overlaid by the FISH signal and converted to grey scale. Copy number variations (asterisks), heterozygous *Cla-element* bands (arrows) and regions of chromosome asynapsis (circles) are highlighted. Chromosomes are designated by Roman numbers (I, II, III), chromosome arms are denoted by letters (A, B, C, D, F), centromeres are labelled (CEN). Graphical insert of chromosomal arm II C (top) results from an additional picture of the identical chromosome.

one of the respective pure populations (Fig. 5, top right). Hybridization during the first experimental generation F_0 resulted in a significant reduction of clutches per female in at least two hybrid populations, although some females of parental pure populations also failed to oviposit (Fig. 5, bottom left). Hybrid females of all but one hybrid population in the second experimental generation produced more than one clutch, significant for both SI×SS crosses (Fig. 5, bottom right).

Discussion

Draft genome as prerequisite for TE analyses

Computational analysis of genomewide TE activity requires a relatively well-assembled and scaffolded draft genome (Ewing 2015). In our draft genome version, we increased the maximum N50 value for genomes of the genus *Chironomus* by the factor of 35 and for all chironomids by the factor of 20 (Table S12, Supporting Information). In direct comparison with a recently published draft genome of *C. riparius* (Vicoso & Bachtrog 2015), our version covers a higher fraction of the genome (90% instead of 77.5%) is less fragmented (5292 instead of 29 677 scaffolds) and has a substantially increased N50 value (272 065 instead of 7097). From a BUSCO analysis, we assume that the genome is almost complete in terms of its gene set (93 % of core arthropod genes found). The annotation produced a reasonable number of protein-coding genes. This is comparable to *C. tentans* from the same genus as well as *Drosophila melanogaster* (15 120 and 13 907 protein-coding genes, respectively;

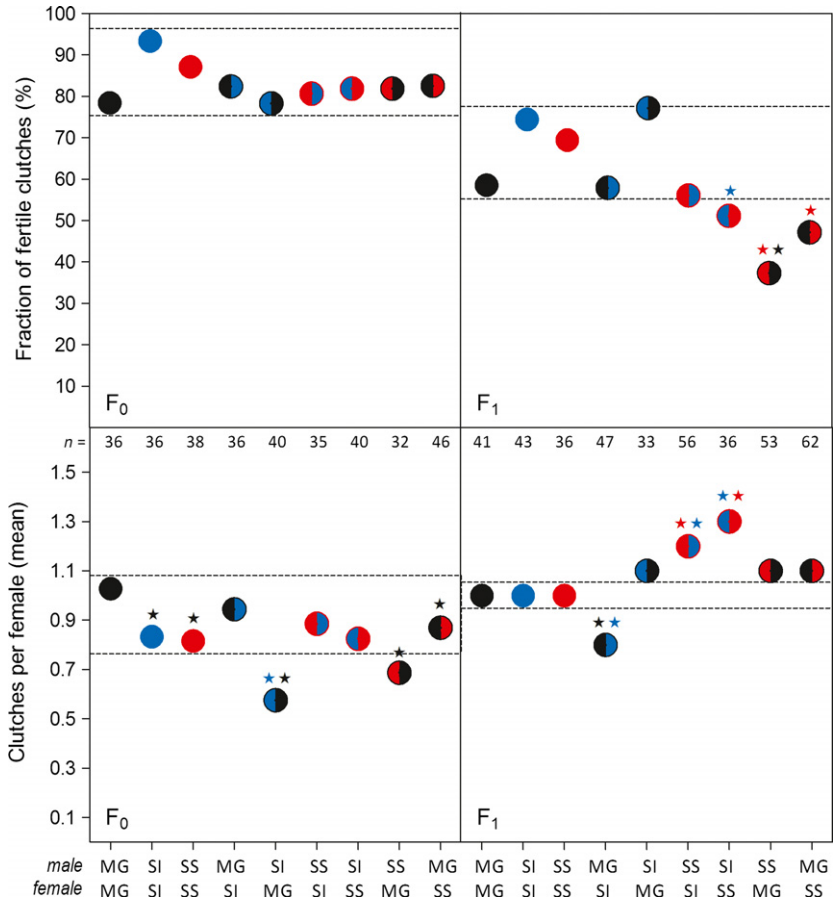

Fig. 5 Female fitness parameters in two subsequent generations (F_0 left, and F_1 right) of the hybridization experiment with three *Chironomus riparius* populations MG – Hessen, SI – Piemonte, SS – Andalusia: fraction of fertile clutches (top) and clutches per female (bottom) are indicated. Hybrid populations were compared to both of their parental pure populations and asterisks with the according colour of each parental pure population mark significant differences ($P < 0.05$, one asterisk for each comparison, e.g. red asterisk if significantly different to SI). Parental pure populations were compared among each other (respective colour code). Horizontal dashed lines outline the range of variation set by the pure populations. Number of females to calculate clutches per female is given for each population (*n*).

dos Santos *et al.* 2015; Table S7, Supporting Information; Kutsenko *et al.* 2014). The difference of approximately 30 MB of sequence length between the presented and the estimated genome size of 200 MB (Schmidt-Ott *et al.* 2009) is in the most part likely to be due to centromers, telomers and other large repetitive clusters that cannot be assembled with the existing data.

Selection against heterozygous Cla-elements as potential endogenous genetic barrier among Chironomus riparius populations

Using the draft genome and pooled genomic sequences, we have shown that population divergence can be stronger at the level of TEs than expected from genome-wide divergence averages. This observation, together with the special properties of the *Cla-element*, indicates a potential role of this TE in building endogenous genetic barriers among *C. riparius* populations. Despite the large geographical distance between populations, the population differentiation based on genomewide SNPs is very low. This is probably due to a large effective population size and high gene flow in this species rather than short divergence time (Oppold *et al.* 2016). The population differentiation on the level of *Cla-elements*, and TEs in general, was significantly stronger than anticipated. This pattern cannot be explained by differences in rates at which new SNPs or TE insertions occur because both represent biallelic markers, and hence, F_{ST} estimates are directly comparable (Edelaar & Björklund 2011). If the mutation rate is at least two orders of magnitude smaller than the migration rate, a relation that we assume given the high gene flow, F_{ST}, is hardly affected by mutation (Edelaar & Björklund 2011). Many of the highly differentiated TE loci would have been flagged as under divergent selection in an F_{ST}-outlier analysis, indicating that evolutionary forces other than neutral drift are acting on them. However, the significant correlation of mean pairwise F_{ST} measured from genomewide SNPs with those estimated from all TEs in general and the *Cla-element* in particular (Fig. 2A) suggests that drift nevertheless shaped the overall population differentiation pattern of these genomic elements. Many private insertions indicate that either ongoing TE activity after population split or purifying selection by differential purging (Charlesworth & Charlesworth 1983) played a major role in TE population dynamics (reviewed in Barron *et al.* 2014). We

therefore focussed on population-private TE and *Cla-element* insertions compared to population-private SNPs. For the majority of the latter, we can reasonably assume that they occurred after the respective population splits. If neutral, the vast majority of private markers should follow the same population genetic trajectories, governed by the respective demographic population parameters. Given the low overall differentiation (mean $F_{ST(SNPs)} = 0.05$), indicative of high interpopulation gene flow (Slatkin 1987), we would expect strictly neutral private TE insertions to segregate within populations at about the level of private neutral SNPs. This is not the case, as shown by the significant difference in heterozygosity and the higher proportion of fixed population-private TEs, in particular for the *Cla-elements* (Figs 3 and S8, Supporting Information). Rather, non-neutral, that is endogenous selective processes acting on TEs in general, and on the *Cla-elements* in particular, seems necessary to explain the observed patterns. We therefore suggest that the heterozygous state of tandem-repetitive *Cla-element* clusters as such in an individual may be deleterious. To explain the presence of fixed population-private insertions, there has to be a mechanism through which new insertions can initially escape negative selection. It can reasonably be assumed that short *Cla-element* monomer insertions in nongenic regions are nearly neutral and may, thus, by chance drift to frequencies at which homozygous insertions occur. The latter is necessary to start expansion in a tandem-like fashion by unequal crossing-over. This process finally enhances the formation of large tandem-repetitive clusters because generally the frequency of unequal crossing-over events or slipped strand mispairing increases with the length of an allele (Levinson *et al.* 1985; Levinson & Gutman 1987). Depending on their population frequency, such clusters might either become fixed or eliminated by selection. We are, however, not able to test this hypothesis as our data do not allow for length comparisons of *Cla-element* clusters. Moreover, there is neither information about rates of transposition and expansion to clusters, nor about the mode of transposition and suppression of *Cla-elements* in order to parameterize an appropriate model.

As a potentially deleterious mechanism, secondary structures formed by tandem-repetitive *Cla-element* clusters due to hairpin loop formation and the marked inherent DNA curvature (Israelewski 1983; Schmidt 1984; Hankeln 1990) might lead to complex epistatic interactions triggering DMI (Brown & O'Neill 2010) and ectopic recombination events, as observed for classical TEs in *Drosophila* (Petrov *et al*. 2011; Barron *et al*. 2014). Moreover, chromosomal regions containing *Cla-element* clusters are associated with known chromosome breakpoints in *C. riparius* (Bovero *et al.* 2002), and also in our study chromosomal pairing aberrations were observed in conjunction with heterozygous *Cla* clusters (Fig. 4). Ectopic recombination events occur preferentially between heterozygous TEs (Montgomery *et al.* 1991) and negative fitness effects are hence expected to occur preferentially in individuals from interpopulation crossings (Rizzon *et al.* 2002). This inevitably leads to differential fixation of TE insertions among populations. Particularly interpopulation hybrids during the first few generations are expected to harbour an increased amount of heterozygous insertion sites and, thus, experience the consequences of multiple endogenous incompatibilities, as observed here. The adjacent genetic variation of *Cla-elements* did not reveal a hitchhiking signal. This could have two alternative reasons, depending on the expansion process of *Cla-element* monomers during cluster formation. If the monomer of a single haplotype expands and is subsequently selected, a hitchhiking signal would be expected. If, however, and this appears more probable, monomer expansion happens in multiple haplotypes, these would be selected for their size only, producing a signal comparable to soft selective sweep (Messer & Petrov 2013) with hitchhiking effects being more difficult to detect. The pattern found in our data is therefore consistent with both the possibility that *Cla-element* expansion happens in multiple haplotypes and the possibility that most instances are relatively old events where recombination and mutation have already erased any initially present signs of hitchhiking during the fixation process, as predicted by Bierne *et al.* (2011).

In general, the abundance of TEs in the *C. riparius* genome reflects the expected distribution pattern of significantly less TEs inserted into protein-coding regions than expected by chance, supporting their deleterious influence on conserved coding regions (Pasyukova *et al.* 2004).

Here we propose a TE-based mechanism for the induction of DMI as basis for endogenous selection. This phenomenon does not seem to only arise from classical TEs and antagonistic elements, but also from transposable tandem-repetitive minisatellite-like sequences with a presumably passive mode of transposition. The mechanism is based on chromosome structure and potential steric consequences and thus is different from the concept of genomic conflict (Crespi & Nosil 2013). To our best knowledge, such a process has so far not been empirically investigated.

Postpopulation split transpositional activity of the Cla-element in Chironomus riparius

To contribute to population divergence, TEs must have either been active after the population split or TEs segregating at the time of the population split must have differentially been fixed thereafter. Several aspects of

our data point towards a postsplit transposition activity of the *Cla-element* in the analysed *C. riparius* populations. First, there are many population-private *Cla-element* insertion sites resulting in a divergent genomic TE pattern among populations (Fig. 1). A scenario of differential fixation is only conceivable if the presplit *Cla-element* was subject to different selective forces than today. If the same negative selection pressure inferred to act on heterozygous *Cla-element* insertions in the current populations applied in the past, most of the *Cla-element* insertions would have been fixed as they are today and could not get lost differentially. Sequence comparisons of the *Cla* consensus sequence revealed high similarity to the sister species *C. piger* (3.7% Cla-element divergence, Hankeln & Schmidt 1987) and also to the more distantly related *C. luridus* (Ross *et al.* 1997). Hence, it is unlikely that structural features of the element on which selection may act were any different in the ancestral *C. riparius* population. We therefore suggest that the *Cla-element* has been active after the establishment of the different populations at their respective geographical site. Regarding the high variability of the *Cla* sequence identified via the BLAST analysis, it might be argued that the *Cla-element* has generally accumulated too many mutations to maintain activity until today. However, classical TEs such as the ETn elements in the mouse genome are transposed from a few master copies of intact feature architecture, which will ensure continued active transposition (Ribet *et al.* 2004). With the method applied here, we cannot distinguish sequence variability in regard of the respective insertion site, and thus, no statement about the *Cla* sequence at population-private or nonfixed sites can be made. Furthermore, the transposition mechanism of the *Cla-element* is still unknown, as are the necessary sequence features for activity maintenance.

The second indication of ongoing activity comes from the existence of insertion sites still segregating in the populations. These sites might have been introduced, either by *de novo* insertion or migration, recently enough to avoid fixation by drift. An active TE will constantly create new insertions that will start at low frequencies in the population and that are exclusive to this population. Only about 25% of all *Cla-element* loci were shared among all populations and are therefore likely to belong to the most ancient species-specific insertions (Fig. S5, Supporting Information). The detected population-specific low-frequency (<0.25) *Cla-element* insertions can be best explained with active transposition of the element after the population split. However, negative selection against segregating (i.e. heterozygous) *Cla-element* insertions could also be the reason for their relatively low amount.

In the two southernmost populations (SI – Piemonte, SS – Andalusia), we detected insertions at very low frequencies (0.003–0.01) in different genomic regions with high coverage, respectively. On the one hand, this might be an artefact due to *Cla-element* CNV. On the other hand, these regions might be heterochromatic with a high content of so-far-uncharacterized repetitive elements that are not well assembled and not masked in the draft genome, thus explaining the high coverage. Such regions, presumably evolving under neutral conditions, will allow new insertions to easily accumulate (Cridland *et al.* 2013). Therefore, these low-frequency *Cla-element* insertions might be traces of recent transposition events in these two populations. The higher proportion of private *Cla-element* insertions and low-frequency insertions in the southern populations (SI – Piemonte, SS – Andalusia) suggest that the *Cla-element* is more active there. Consistent with this, significantly less population-private *Cla-element* insertions were found in northern populations (MG – Hessen, NMF – Lorraine). Additionally, this pattern is positively correlated (exponential one-phase association) to the annual mean temperature at the respective location (Fig. S6, Supporting Information). These results fit very well to a recent study showing that the evolutionary 'speed' of the very same *C. riparius* populations depends on the temperature-dependent number of generations per year and the possibly also increased mutation rate (Oppold *et al.* 2016). It is conceivable that the observed quantitative differences in population TE content are also temperature-related. However, more populations have to be investigated to support this finding and rule out potential impact of other processes, such as population expansion effects or bottlenecks.

The third indication of ongoing activity is the presence of monomeric or dimeric *Cla-element* transcripts in *C. riparius*, as evidenced by transcriptome data (Table S10, Supporting Information), northern hybridization (Hankeln & Schmidt 1987) and RT–PCR (Martinez-Guitarte *et al.* 2012) further hints at an ongoing *Cla-element* transposition, presumably *via* RNA intermediates that are then reverse transcribed and reintegrated into the genome. Monomeric *Cla-elements* are structurally similar to classical SINE retroelements (Hankeln & Schmidt 1987). However, SINEs depend on active LINEs (long interspersed nuclear elements) for their retrotransposition machinery (Wicker *et al.* 2007). In fact, there is indirect evidence for ongoing transpositional activity of at least one LINE in *C. riparius* (NLRCth1; Zampicinini *et al.* 2011) that could provide the necessary enzymatic components for passive *Cla-element* transposition. It remains unclear, however, if *Cla-elements* predominantly retrotranspose as mono- or oligomers. Most *Cla* sequences are organized as minisatellite-like, interspersed clusters, which might also use DNA-based transposition, for example by excision and

re-integration of extrachromosomal copies (reviewed in Kazazian 2004).

Taken together, population-private *Cla-element* abundance, *Cla-element* insertions at low and intermediate frequencies in the respective populations, and the presence of *Cla* transcripts strongly suggest recent transpositional activity of the element in the investigated populations. The effect of drift or purifying selection acting on *Cla-element* insertions could alternatively explain the divergent abundance and segregating insertion sites, but not the *Cla* transcripts.

Incompatibilities on the phenotypic level of chromosomes among Chironomus riparius populations

In situ hybridization of *Cla-elements* to polytene chromosomes of the different pure and hybrid populations showed an additive *Cla-element* pattern in F_1 hybrid individuals as expected. A drastic burst of *Cla-element* activity creating new loci during early embryo development was not observed, in contrast to the activation of the P-element in hybrid dysgenesis of *D. melanogaster*, when only one strain is carrying the P-element (Rio 1990). The additive effect of the *Cla-element* patterns in hybrids led to several heterozygous sites, sometimes showing partial asynapsis of the homologous chromosomes in this region (Fig. 4). The same effect is already known from hybrids when *C. riparius* is crossed with *C. piger*, the sister species that is free of *Cla-element* insertions across chromosomal arms (Hägele 1984; Schmidt 1984). Asynapsis of homologous chromosomes is very pronounced in crosses of the sister species, which might be due to the stronger difference in *Cla-element* distribution patterns. This could explain the weaker effect in hybrids of conspecific *C. riparius* populations where not all *Cla-element* insertions automatically produce heterozygous sites in hybrids. However, because the in situ hybridization required viable larvae relatively late in their development, our sample is inevitably biased towards individuals with non- or only mildly deleterious chromosome rearrangements. It is possible that the observed infertile clutches were infertile because of deleterious heterozygous insertion sites.

Based on our findings and the correlation of *Cla-element* distribution with sites of chromosomal rearrangements and breakage (Bovero *et al.* 2002), we propose that the heterozygous *Cla-element* pattern leads to incorrect synapsis of homologous chromosomes as an obvious evidence of DMI on the level of chromosome structure. Even if polytene chromosomes are rather special, the distribution of genomic elements and intrinsic consequences correspond. Correct synapsis of homologous chromosomes is an important component of meiosis, during which chromosome structural axes develop between homologs, forming the synaptonemal complex that enables recombination (Tsai & McKee 2011; Zickler & Kleckner 2015).

Crossing of Chironomus riparius populations affects hybrid fitness

Whether the detected chromosome aberrations might actually directly influence the fitness of individuals is difficult to test with our data, and we cannot provide a causal connection between the different levels of biological organization that we included in our analysis. Nevertheless, crossing of the three different populations hints at an early stage of postzygotic isolation. We can exclude any mechanism of prezygotic isolation because mating and fertilization among the different populations was successful as numbers of fertile clutches produced during hybridization (F_0) did not differ (Fig. 5). Our observation of reduced fertility of hybrid clutches in the subsequent generation, that is the first hybrid generation (Fig. 5), is in line with the observations of reduced egg hatch in hybrids of different *C. riparius* strains that suffered from the HLE dysgenesis syndrome (Hägele 1995). Impairment of clutch fertility was especially prominent in crosses of the Hessen population (MG) with the Andalusian population (SS), suggesting a population-specific dysgenesis potential (Hägele 1995). The hybrid crosses between MG × SS and SI × SS show reduced clutch fertility as well as an increased number of clutches per female, potentially as a compensatory effect. As the experimental design did not allow for multiple reproduction cages, fitness parameters are based on unreplicated data and only allow a limited assessment of hybrid fitness. Nevertheless, as a standard test organism in ecotoxicological research, life-trait parameters are well known for *C. riparius* and accordingly the here-observed effects of significantly decreased fertility of clutches and more than a single egg clutch per female are not common for *C. riparius* when kept under standard conditions (OECD 2004; Weltje & Seliger 2010). Future experiments will be necessary to investigate if hybrid individuals show reduced fitness in regard to the respective environmental conditions of the natural pure populations, for example temperature regimes. However, as an intrinsic genomic turnover component (Dover 1982), TEs drive population divergence independently of the local environment (Bierne *et al.* 2011). If this results in negative fitness effects, hybrids are not necessarily expected to show intermediate phenotypes between clines. For the present study, we documented that fertility of offspring is reduced when populations of differing genetic backgrounds are hybridized, which is not a proof but a necessary prerequisite for a role of the *Cla-element* in RI. In

Drosophila, the dysgenesis syndromes involving the P-, I- or the *hobo* element are known to be restricted to one crossing direction (transposon-carrying males × transposon-free females, Bingham *et al.* 1982; Bucheton *et al.* 1984; Galindo *et al.* 1995). Lack of such a bias in our study indicates that the underlying mechanism may be different in *C. riparius* and can be better explained with the general concept of DMI that results from the presence–absence pattern of *Cla-elements* throughout the genome, irrespective of specific loci.

Conclusion

The improved and annotated *C. riparius* draft genome offers a valuable genomic resource for studying genome evolution. We show that TEs might be an underestimated source of genomic differentiation and, hence, the divergence of conspecific populations. The suggested negative selection against heterozygous *Cla-element* insertions, which inevitably occur in interpopulation hybrids, creates a yet-undescribed endogenous genetic barrier. Only the observed RI among populations allows to tenure the hypothesis that *Cla-element* divergence patterns are involved in it; however, causality remains to be established.

Acknowledgements

We thank Steffen Lemke (University of Heidelberg, Germany) and Tobias Kaiser (Max Planck Institute of Evolutionary Biology, Plön, Germany) providing initial support and advice for the genome assembly. The 1KITE Antliophora group (BioProject ID 183205, www.1kite.org) gratefully provided us with additional transcriptome data for the genome annotation process. We here thank Alexander Donath and Karen Meusemann for 1KITE submission and data management, as well as the 1KITE speakers Xin Zhou, Karl Kjer and Bernhard Misof. We also thank Ingo Ebersberger (Goethe University Frankfurt, Germany) for advice and fruitful discussions, as well as Daniel Barbash (Cornell University, USA) for helpful comments. We thank Matthew Forrest (Senckenberg BIK-F, Germany) for English proof-reading. Funding was provided by DFG (PF390/8-1; Ha2103/7-1). Parts of this research were conducted using the supercomputer Mogon and advisory services offered by Johannes Gutenberg University Mainz (www.hpc.uni-mainz.de), which is a member of the AHRP and the Gauss Alliance e.V. T.H. gratefully acknowledges support by the Centre for Computational Sciences, Johannes Gutenberg University Mainz (CSM/SRFN). Robert Kofler acknowledges funding within the ERC Grant 'ArchAdapt'.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Barron MG, Fiston-Lavier AS, Petrov DA, Gonzalez J (2014) Population genomics of transposable elements in *Drosophila*. *Annual Review of Genetics*, **48**, 561–581.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Bikard D, Loot C, Baharoglu Z, Mazel D (2010) Folded DNA in action: hairpin formation and biological functions in prokaryotes. *Microbiology and Molecular Biology Reviews*, **74**, 570–588.

Bingham PM, Kidwell MG, Rubin GM (1982) The molecular-basis of P-M hybrid dysgenesis - the role of the P-element, a P-strain-specific transposon family. *Cell*, **29**, 995–1004.

Blumenstiel JP, Hartl DL (2005) Evidence for maternally transmitted small interfering RNA in the repression of transposition in *Drosophila virilis*. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15965–15970.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, **27**, 578–579.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Boussy IA, Periquet G (1993) The transposable element hobo in *Drosophila melanogaster* and related species. In: *Transposable Elements and Evolution* (ed. McDonald JF), pp. 192–200. Springer Netherlands, Dordrecht.

Bovero S, Hankeln T, Michailova P, Schmidt E, Sella G (2002) Nonrandom chromosomal distribution of spontaneous breakpoints and satellite DNA clusters in two geographically distant populations of *Chironomus riparius* (Diptera: Chironomidae). *Genetica*, **115**, 273–281.

Brown JD, O'Neill RJ (2010) Chromosomes, conflict, and epigenetics: chromosomal speciation revisited. *Annual Review of Genomics and Human Genetics*, **11**, 291–316.

Bucheton A, Paro R, Sang HM, Pelisson A, Finnegan DJ (1984) The molecular-basis of I-R hybrid dysgenesis in *Drosophila melanogaster* - identification, cloning, and properties of the I-factor. *Cell*, **38**, 153–163.

Cantarel BL, Korf I, Robb SMC *et al.* (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, **18**, 188–196.

Castillo DM, Moyle LC (2012) Evolutionary implications of mechanistic models of TE-mediated hybrid incompatibility. *International Journal of Evolutionary Biology*, **2012**, 698198.

Catasti P, Gupta G, Garcia AE *et al.* (1994) Unusual structures of the tandem repetitive DNA-sequences located at human centromeres. *Biochemistry*, **33**, 3819–3830.

Charlesworth B, Charlesworth D (1983) The population-dynamics of transposable elements. *Genetical Research*, **42**, 1–27.

Crespi B, Nosil P (2013) Conflictual speciation: species formation via genomic conflict. *Trends in Ecology & Evolution*, **28**, 48–57.

Cridland JM, Macdonald SJ, Long AD, Thornton KR (2013) Abundance and distribution of transposable elements in two *Drosophila* QTL mapping resources. *Molecular Biology and Evolution*, **30**, 2311–2327.

Dion-Cote AM, Renaut S, Normandeau E, Bernatchez L (2014) RNA-seq reveals transcriptomic shock involving

transposable elements reactivation in hybrids of young lake whitefish species. *Molecular Biology and Evolution*, **31**, 1188–1199.

Dobzhansky T (1937) *Genetics and the Origin of Species Columbia*. University Press, West Sussex.

Dover G (1982) Molecular drive: a cohesive mode of species evolution. *Nature*, **299**, 111–117.

Edelaar P, Björklund M (2011) If F_{ST} does not measure neutral genetic differentiation, then comparing it with Q_{ST} is misleading. Or is it? *Molecular Ecology*, **20**, 1805–1812.

Ewing AD (2015) Transposable element detection from whole genome sequence data. *Mobile DNA*, **6**, 1–9.

Fontdevila A (1992) Genetic instability and rapid speciation - are they coupled. *Genetica*, **86**, 247–258.

Franssen SU, Nolte V, Tobler R, Schlotterer C (2015) Patterns of linkage disequilibrium and long range hitchhiking in evolving experimental *Drosophila melanogaster* populations. *Molecular Biology and Evolution*, **32**, 495–509.

Futschik A, Schlötterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

Galindo MI, Ladeveze V, Lemeunier F *et al.* (1995) Spread of the autonomous transposable element hobo in the genome of *Drosophila melanogaster*. *Molecular Biology and Evolution*, **12**, 723–734.

Hägele K (1984) Different hybrid effects in reciprocal crosses between *Chironomus thummi thummi* and *Ch. th. piger* including spontaneous chromosome-aberrations and sterility. *Genetica*, **63**, 105–111.

Hägele K (1985) Identification of a polytene chromosome band containing a male sex determiner of *Chironomus thummi thummi*. *Chromosoma*, **91**, 167–171.

Hägele K (1987) Nonreciprocal gonadal-dysgenesis in *Chironomus thummi* Hybrids - temperature sensitivity of female sterility. *Developmental Genetics*, **8**, 17–26.

Hägele K (1995) The hle hybrid dysgenesis syndrome in the midge *Chironomus thummi* - differences in the dysgenic potential between strains. *Genetica*, **96**, 239–245.

Hägele K, Kasper-Sonnenberg M (2000) Egg size and hybrid syndrome-dependent embryo mortality in *Chironomus* hybrids (Diptera: Chironomidae). *European Journal of Entomology*, **97**, 1–6.

Hägele K, Lachmann J (1992) In the Rud hybrid dysgenesis syndrome of *Chironomus* hybrids maternal chromosomes condense earlier than paternal chromosomes. *Chromatin*, **1**, 59–67.

Hägele K, Oschmann B (1987) Nonreciprocal gonadal-dysgenesis in hybrids of the chironomid midge *Chironomus thummi*. 3 Germ line specific abnormalities. *Chromosoma*, **96**, 50–54.

Hägele K, Nyhuis E, Oschmann B (1995) Heterogeneous development of hybrids of the midge *Chironomus thummi* - a 3rd trait of the Hle syndrome contributing to reproductive isolation. *Biologisches Zentralblatt*, **114**, 243–251.

Hailu C, Pingel H, Saar W (1999) Impact of chromosome aberrations on the embryonic mortality of hybrid ducks (*Cairina moschata* x *Anas platyrhynchos*). *Archiv Fur Geflugelkunde*, **63**, 174–181.

Hankeln T (1990) *Molekulare Analyse phylogenetisch bedeutsamer repetitiver DNA bei Chironomiden* PhD Thesis, Ruhr University.

Hankeln T, Schmidt ER (1987) Cotransposition of a highly repetitive DNA element with flanking sequences in the genome of the midge *Chironomus thummi*. *Journal of Molecular Evolution*, **26**, 311–319.

Hankeln T, Rohwedder A, Weich B, Schmidt ER (1994) Transposition of minisatellite-like DNA in *Chironomus* midges. *Genome*, **37**, 542–549.

Hernandez-Hernandez A, Rincon-Arano H, Recillas-Targa F *et al.* (2008) Differential distribution and association of repeat DNA sequences in the lateral element of the synaptonemal complex in rat spermatocytes. *Chromosoma*, **117**, 77–87.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, **12**, 1–14.

Hunt M, Kikuchi T, Sanders M *et al.* (2013) REAPR: a universal tool for genome assembly evaluation. *Genome Biology*, **14**, 1–10.

Ilkova J, Cervella P, Zampicinini G, Sella G, Michailova P (2013) Chromosomal breakpoints and transposable-element-insertion sites in salivary gland chromosomes of *Chironomus riparius* meigen (Diptera, Chironomidae) from trace metal polluted stations. *Acta Zoologica Bulgarica*, **65**, 59–73.

Israelewski N (1983) Structure and function of an at-rich, interspersed repetitive sequence from *Chironomus thummi* - solenoidal DNA, 142 Bp palindrome-frame and homologies with the sequence for site-specific recombination of bacterial transposons. *Nucleic Acids Research*, **11**, 6985–6996.

Jurka J, Bao WD, Kojima KK (2011) Families of transposable elements, population structure and the origin of species. *Biology Direct*, **6**, 1–16.

Kajitani R, Toshimoto K, Noguchi H *et al.* (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, **24**, 1384–1395.

Kavi HH, Fernandez HR, Xie WW, Birchler JA (2005) RNA silencing in *Drosophila*. *FEBS Letters*, **579**, 5940–5949.

Kazazian HH (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.

Keyl HG (1965) A demonstrable local and geometric increase in the chromosomal DNA of *Chironomus*. *Experientia*, **21**, 191–193.

Kidwell MG (1985) Hybrid dysgenesis in *Drosophila-melanogaster* - nature and inheritance of P-element regulation. *Genetics*, **111**, 337–350.

Kofler R, Orozco-terWengel P, De Maio N *et al.* (2011a) POPOOLATION: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE*, **6**, 1–9.

Kofler R, Pandey RV, Schlotterer C (2011b) POPOOLATION2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, **27**, 3435–3436.

Kofler R, Betancourt AJ, Schlotterer C (2012) Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. *Plos Genetics*, **8**, 1–16.

Kofler R, Hill T, Nolte V, Betancourt AJ, Schlotterer C (2015) The recent invasion of natural *Drosophila simulans*

populations by the P-element. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 6659–6663.

Kutsenko A, Svensson T, Nystedt B *et al.* (2014) The *Chironomus tentans* genome sequence and the organization of the Balbiani ring genes. *BMC Genomics*, **15**, 819.

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, **4**, 203–221.

Levinson G, Marsh JL, Epplen JT, Gutman GA (1985) Cross-hybridizing snake satellite, *Drosophila*, and mouse DNA-sequences may have arisen independently. *Molecular Biology and Evolution*, **2**, 494–504.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Maheshwari S, Barbash DA (2011) The genetics of hybrid Incompatibilities. *Annual Review of Genetics*, **45**, 331–355.

Mariappan SVS, Garcia AE, Gupta G (1996) Structure and dynamics of the DNA hairpins formed by tandemly repeated CTG triplets associated with myotonic dystrophy. *Nucleic Acids Research*, **24**, 775–783.

Martienssen RA (2010) Heterochromatin, small RNA and post-fertilization dysgenesis in allopolyploid and interploid hybrids of *Arabidopsis*. *New Phytologist*, **186**, 46–53.

Martinez-Guitarte JL, Planello R, Morcillo G (2012) Overexpression of long non-coding RNAs following exposure to xenobiotics in the aquatic midge *Chironomus riparius*. *Aquatic Toxicology*, **110**, 84–90.

Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, **28**, 659–669.

Montgomery EA, Huang SM, Langley CH, Judd BH (1991) Chromosome Rearrangement by Ectopic Recombination in *Drosophila melanogaster* - Genome Structure and Evolution. *Genetics*, **129**, 1085–1098.

Moschetti R, Dimitri P, Caizzi R, Junakovic N (2010) Genomic instability of I elements of *Drosophila melanogaster* in absence of dysgenic crosses. *PLoS ONE*, **5**, e13142.

Muller HJ (1942) Isolating mechanisms, evolution and temperature. *Biological Symposia*, **6**, 71–125.

OECD (2004) *Test No. 219: Sediment-Water Chironomid Toxicity Using Spiked Water*. OECD Publishing, Paris.

Okonechnikov K, Conesa A, Garcia-Alcalde F (2016) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, **32**, 292–294.

Oleszczuk S, Lukaszewski AJ (2014) The origin of unusual chromosome constitutions among newly formed allopolyploids. *American Journal of Botany*, **101**, 318–326.

Oppold AM, Pedrosa JA, Balint M *et al.* (2016) Support for the evolutionary speed hypothesis from intraspecific population genetic data in the non-biting midge *Chironomus riparius*. *Proceedings. Biological sciences*, **283**, 20152413.

Pasyukova EG, Nuzhdin SV, Morozova TV, Mackay TFC (2004) Accumulation of transposable elements in the genome of *Drosophila melanogaster* is associated with a decrease in fitness. *Journal of Heredity*, **95**, 284–290.

Petrov DA, Fiston-Lavier AS, Lipatov M, Lenkov K, Gonzalez J (2011) Population Genomics of Transposable Elements in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **28**, 1633–1644.

Pezer Z, Ugarkovic D (2012) Satellite DNA-associated siRNAs as mediators of heat shock response in insects. *RNA Biology*, **9**, 587–595.

Pfenninger M, Nowak C, Kley C, Steinke D, Streit B (2007) Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic *Chironomus* (Diptera) species. *Molecular Ecology*, **16**, 1957–1968.

Radic MZ, Lundgren K, Hamkalo BA (1987) Curvature of mouse satellite DNA and condensation of heterochromatin. *Cell*, **50**, 1101–1108.

R-Core T (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ribet D, Dewannieux M, Heidmann T (2004) An active murine transposon family pair: retrotransposition of "master" MusD copies and ETn trans-mobilization. *Genome Research*, **14**, 2261–2267.

Rio DC (1990) Molecular mechanisms regulating *Drosophila*-P element transposition. *Annual Review of Genetics*, **24**, 543–578.

Rizzon C, Marais G, Gouy M, Biemont C (2002) Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Research*, **12**, 400–407.

Robinson JT, Thorvaldsdottir H, Winckler W *et al.* (2011) Integrative genomics viewer. *Nature Biotechnology*, **29**, 24–26.

Ross R, Hankeln T, Schmidt ER (1997) Complex evolution of tandem-repetitive DNA in *Chironomus*. *Journal of Molecular Evolution*, **44**, 321–326.

dos Santos G, Schroeder AJ, Goodman JL *et al.* (2015) FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, **43**, D690–D697.

Schmidt ER (1981) The development of a 120 basepair repetitive DNA-sequence in *Chironomus thummi* is correlated to the duplication of defined chromosomal segments. *FEBS Letters*, **129**, 21–24.

Schmidt ER (1984) Clustered and interspersed repetitive DNA-sequence family of *Chironomus* - the nucleotide-sequence of the Cla-elements and of various flanking sequences. *Journal of Molecular Biology*, **178**, 1–15.

Schmidt ER, Keyl HG, Hankeln T (1988) In situ localization of 2 hemoglobin gene clusters in the chromosomes of 13 species of *Chironomus*. *Chromosoma*, **96**, 353–359.

Schmidt H, Greshake B, Feldmeyer B, Hankeln T, Pfenninger M (2013) Genomic basis of ecological niche divergence among cryptic sister species of non-biting midges. *BMC Genomics*, **14**, 1–11.

Schmidt-Ott U, Rafiqi AM, Sander K, Johnston JS (2009) Extremely small genomes in two unrelated dipteran insects with shared early developmental traits. *Development Genes and Evolution*, **219**, 207–210.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Simkin A, Wong A, Poh YP, Theurkauf WE, Jensen JD (2013) Recurrent and recent selective sweeps in the Pirna pathway. *Evolution*, **67**, 1081–1090.

Simmons MJ, Thorp MW, Buschette JT, Becker JR (2015) Transposon regulation in *Drosophila*: piRNA-producing P elements facilitate repression of hybrid dysgenesis by a P element that encodes a repressor polypeptide. *Molecular Genetics and Genomics*, **290**, 127–140.

Slatkin M (1987) Gene flow and the geographic structure of natural-populations. *Science*, **236**, 787–792.

Smit AFA, Hubley R, Green P (2013-2015) RepeatMasker Open-4.0; repeatmasker.org.

Talbert PB, Henikoff S (2010) Centromeres convert but don't cross. *PLoS Biology*, **8**, e1000326.

Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, **14**, 178–192.

Tsai JH, McKee BD (2011) Homologous pairing and the role of pairing centers in meiosis. *Journal of Cell Science*, **124**, 1955–1963.

Vicoso B, Bachtrog D (2015) Numerous transitions of sex chromosomes in Diptera. *PLoS Biology*, **13**, 1–22.

Weltje L, Seliger R (2010) *Improving Life-Cycle Testing with the Non-Biting Midge Chironomus riparius Through Adult Feeding*. SETAC EU, Seville, Spain.

Werren JH (2011) Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 10863–10870.

Wicker T, Sabot F, Hua-Van A *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, **8**, 973–982.

Yannopoulos G, Stamatis N, Monastirioti M, Hatzopoulos P, Louis C (1987) Hobo is responsible for the induction of hybrid dysgenesis by strains of *Drosophila-melanogaster* bearing the male recombination factor 23.5mrf. *Cell*, **49**, 487–495.

Zampicinini G, Cervella P, Biemont C, Sella G (2011) Insertional variability of four transposable elements and population structure of the midge *Chironomus riparius* (Diptera). *Molecular Genetics and Genomics*, **286**, 293–305.

Zickler D, Kleckner N (2015) Recombination, pairing, and synapsis of homologs during meiosis. *Cold Spring Harbor Perspectives in Biology*, **7**, 1–26.

Data accessibility

Genome sequencing raw reads are deposited in the EMBL-EBI European Nucleotide Archive (study Accession no.: PRJEB15223, raw reads: ERS1460012-ERS1460016). Transcriptome sequencing raw reads are deposited in the EMBL-EBI European Nucleotide Archive (ERS1472439-ERS1472442). The final DNA sequence assembly is deposited in the EMBL-EBI European Nucleotide Archive (ERS1474014). The genome annotation file is available from hankeln@uni-mainz.de. Pool-Seq raw reads are deposited in the EMBL-EBI ENA (study Accession no.: PRJEB19848, raw reads: ERS1590058-ERS1590062). Mapping data in *bam format are deposited at Dryad (DOI: https://doi.org/10.5061/dryad.c8h50/4; DOI: https://doi.org/10.5061/dryad.c8h50/5; DOI: https://doi.org/10.5061/dryad.c8h50/6; DOI: https://doi.org/10.5061/dryad.c8h50/7; DOI: https://doi.org/10.5061/dryad.c8h50/8). Coordinates of sampling locations are available in the Supplementary Material.

A.-M.O., T.H. and M.P. conceived the study; T.H., F.R. and U.S.-O. performed genome sequencing; H.S, S.L.H. and F.R. assembled the draft genome; H.S., F.D. and S.L.H. annotated the draft genome; A.-M.O. performed sequencing of Pool-Seq genome scans, A.-M.O., R.K. and M.P. performed population genomic analyses, M.R. and A.-M.O. performed the hybridization experiment, M.R., B.W., T.H. and E.S. performed and analysed FISH of polytene chromosomes, and A.-M.O., H.S., S.L.H. and M.P. drafted the manuscript.

Supporting information

Additional supporting information may be found in the online version of this article.

Methods S1 Sequencing, assembly, scaffolding and annotation of the *C. riparius* draft genome.

Table S1 Information about *C. riparius* populations that were used for Pool-Seq genome scans and all experiments of this study.

Table S2 Sequence data used for genome assembly and scaffolding.

Table S3 Assembly statistics.

Table S4 Back-mapping of reads used to obtain the assembly onto the *draft genome*.

Table S5 Gene content evaluation.

Table S6 Datasets used for the assembly of a *C. riparius* reference transcriptome.

Table S7 Content of protein-coding genes in *C. riparius* draft genome compared to *C. tentans* and *D. melanogaster*.

Table S8 TEs inserted close to or into protein-coding regions.

Table S9 Statistics of TE distribution in the context of genome annotation.

Table S10 Bioinformatic analysis of *Cla-element* transcription.

Table S11 Number of successful FISH documentations for each chromosome of the different populations.

Table S12 Comparison of the presented *draft genome* with other published *draft genomes* across the nematoceran infraorder Culicomorpha.

Figure S1 Exponential one-phase decay model of increasing window-size with the number of (A) TE-insertions (exclusively *Cla-element* insertions) and (B) *Cla-element* insertions.

Figure S2 Test design of the hybridisation experiment with three *C. riparius* populations (MG – Hessen, SI – Piemonte, SS – Andalucia).

Figure S3 Frequency distribution of all transposable elements (TE) identified in different *C. riparius* populations.

Figure S4 Frequency of *Cla-element* insertions across the genome (absent insertions not taken into account) per population.

Figure S5 Venn diagram of genome-wide *Cla-element* insertions in *C. riparius* including insertions at all frequencies.

Figure S6 Relation of the annual mean temperature at the population site with population-unique *Cla-element* insertions described as exponential one-phase decay function (given regression r^2).

Figure S7 Distribution of F_{ST} calculated from genome-wide SNPs (black), TEs (excluding Cla-elements, grey), and exclusively Cla-elements (red).

Figure S8 (A) Mean of heterozygosity on private *Cla-element* insertions, TE insertions excluding *Cla-elements*, and genome-wide SNPs, respectively in all *C. riparius* populations.

Figure S9 Fluorescent in-situ hybridisation (FISH) micrographs of the *Cla-element* dimer clone *pAD* 2603 to polytene chromosomes of different *C. riparius* pure populations (MG – Hesse, SI – Piemont, SS – Andalusia).

Figure S10 Life-cycle parameters in two subsequent generations (F_0 and F_1) of the hybridisation experiment: (A) mortality, (B) time when 50% of all imagoes emerged (EmT_{50}), (C) population growth rate as integrated measure of all parameters (calculation according to Vogt *et al*. 2007), and (D) number of eggs per clutch.

Appendix S1 Custom Chironomus TE-library.

Appendix S2 Chironomus TE-hierarchy.

Appendix S3 TE-merger.