Chromosomal rearrangements represent modular cassettes for local adaptation across different geographic scales

Claire Mérot[1*], Emma Berdan[2], Hugo Cayuela[1,3], Haig Djambazian[4], Anne-Laure Ferchaud[1], Martin Laporte[1], Eric Normandeau[1], Jiannis Ragoussis[4], Maren Wellenreuther[5,6], Louis Bernatchez[1]

[1] Département de biologie, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, Canada

[2]Department of Ecology, Environment and Plant Sciences, Science for Life Laboratory, Stockholm University, Stockholm, Sweden SE-10691

[3] Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne Switzerland

[4] McGill University, Montréal, Canada

[5]Seafood Research Unit, Plant & Food Research, 300 Wakefield Quay, Port Nelson, Nelson, 7010, New Zealand

[6]School of Biological Sciences, University of Auckland, Auckland, New Zealand

* Claire Mérot claire.merot@gmail.com (corresponding author)

Abstract

Across a species range, spatially-varying environments can drive the evolution of local adaptation. Multiples sources of environmental heterogeneity, at small and large scales, draw complex landscapes of selection which may challenge adaptation, particularly when gene flow is high. Because linkage opposes gene flow but also limits the efficiency of natural selection by contrasting pressures, the key to multidimensional adaptation may reside in the heterogeneity of recombination along the genome. Structural variants like chromosomal inversions are important recombination modifiers that form massive co-segregating genomic blocks linking together alleles at numerous genes. In this study, we investigate the influence of chromosomal rearrangements on genetic variation to ask how their contribution to adaptation with gene flow varies across geographic scales. We sampled the seaweed fly *Coelopa frigida* along a bioclimatic gradient of 10° of latitude, a salinity gradient and across a range of heterogeneous, patchy habitats. We assembled a high-quality genome to analyse 1,446 low-coverage whole-genome sequences, and we found large non-recombining genomic regions, including putative inversions. In contrast to the collinear regions depicting extensive gene flow, inversions and low-recombining regions differentiated populations more strongly, either along an ecogeographic cline or at a fine-grained scale. Those genomic regions were disproportionately involved in associations with environmental factors and adaptive phenotypes, albeit with contrasting patterns between the different recombination modifiers. Altogether, our results highlight the importance of recombination in shaping the selection-migration balance and show that a set of several inversions behave as modular cassettes facilitating adaptation to environmental heterogeneity at local and large scales.

Introduction

Environmental variation is widespread and, across its range, a species experiences variable conditions across both small and large geographic scales. With various sources and various scales of environmental heterogeneity, local adaptation is a complex process driven by multiple dimensions of selection but constrained by the distribution of genetic diversity within the genome and the intensity of gene flow acting on it (Savolainen et al. 2013; Tigano and Friesen 2016). Gene flow is also a multifarious factor depending not only on connectivity, which varies across geographical scales, but also genomic recombination, which varies along the genome (Tigano and Friesen 2016; Semenov et al. 2019). Recombination determines whether alleles at nearby loci remain co-associated or are shuffled to different genomic combinations from one generation to the other; and thus whether selection and drift act on long haplotypes composed of several loci or on shorter fragments (Samuk et al. 2017; Martin et al. 2019; Semenov et al. 2019). Hence, the landscape of recombination influences adaptive trajectories since a linked genetic architecture is predicted to favor adaptation in a context of high gene flow (Yeaman 2013) while independent groups of genes may be necessary to adapt to several axes of environmental variation (Lotterhos et al. 2018). Accordingly, the relative scale of connectivity and environmental variation shapes the distribution of genetic variation across a species' range from widespread to spatially structured polymorphism (Tigano and Friesen 2016; McDonald and Yeaman 2018). However, while patterns of genetic diversity and divergence are generally heterogeneous along the genome, it is yet unclear to what extent such heterogeneity relates to variability in recombination landscape (Ortiz-Barrientos and James 2017; Stevison and McGaugh 2020).

Chromosomal inversions are major modifiers of the recombination landscape because recombination is reduced in heterozygotes bearing a derived and ancestral arrangement (Hoffmann et al. 2004). Furthermore, they modify recombination along large fractions of the genome since the same species can have multiple polymorphic inversions, each of them covering hundreds of kilobases or megabases (Wellenreuther and Bernatchez 2018). For instance, five polymorphic inversions are shared worldwide in *Drosophila melanogaster* (Kapun and Flatt 2019) and the maize *(Zea mays)* bear an inversion of 100Mb (Fang et al. 2012). The last decade has shown that such inversion polymorphisms are more common than previously thought in a wide range of species and has brought important insights into the role of inversions in shaping adaptive variation (reviewed in (Hoffmann and Rieseberg 2008; Wellenreuther and Bernatchez 2018; Mérot, Oomen, et al. 2020). Inversions with large-effect on complex multi-trait phenotypes, such as life-history, behaviour, and colour patterns, confirm that arrangements can behave as alleles of a "supergene", linking together combinations of alleles within each arrangement (Joron et al. 2011; Schwander et al. 2014; Kirubakaran et al. 2016; Wellenreuther and Bernatchez 2018; Yan et al. 2020). Likewise, covariation between inversion frequencies and environmental variables (Kapun et al. 2016; Kirubakaran et al. 2016; Faria, Chaube, et al. 2019; Huang and Rieseberg 2020) support the prediction that locally-adaptive loci clustered within an inversion is a favourable architecture for adaptation with gene flow (Kirkpatrick and Barton 2006; Yeaman 2013). However, most of

our knowledge about large inversions remains restricted to a few classic examples found by contrasting strikingly-different phenotypes or ecotypes and is hence limited to a subset of adaptive inversions along one dimension of selection (Joron et al. 2011; Lindtke et al. 2017; Wellenreuther and Bernatchez 2018; Faria, Chaube, et al. 2019; Mérot, Oomen, et al. 2020). Contrary to smaller structural variants which are increasingly catalogued by naive bioinformatic approaches (Ho et al. 2019), very few studies have finely scanned genomes to search for large chromosomal rearrangements without *a priori* grouping (but see (Huang et al. 2020; Todesco et al. 2020). Such a bottom-up approach is now needed to document the diversity of structural polymorphism and to integrate multiple large rearrangements in population genomics (Mérot, Oomen, et al. 2020). This will help to make sense of heterogeneous effects from drift, migration, and selection on genetic variation along the genome, as well as to understand the role of inversions in adaptation at multiple scales and to multiples sources of environmental variation.

Coelopa frigida is a seaweed fly that inhabits piles of rotting seaweed, so-called wrackbeds, on the coasts of the Northern Atlantic (Fig. 1). It provides an exemplar system to investigate how several chromosomal inversions contribute to genetic variation across space and environments at different scales. *C. frigida* is known to harbour one large inversion on chromosome I (hereafter called *Cf-Inv(1)*) that is polymorphic in Europe and America (Butlin, Collins, et al. 1982; Mérot et al. 2018), as well as four additional large polymorphic inversions described in one British population (Aziz 1975). The largest of these inversions is *Cf-Inv(1)*, encapsulates 10% of the genome and has two arrangements: α and β. These alternative *Cf-Inv(1)* arrangements have opposing effects on body size, fertility and development time, a combination of traits which results in different fitness depending on the local characteristics of the wrackbed (Butlin, Read, et al. 1982; Day et al. 1983; Butlin and Day 1985; Edward and Gilburn 2013; Wellenreuther et al. 2017; Berdan et al. 2018; Mérot et al. 2018; Mérot, Llaurens, et al. 2020). Almost nothing is known about the other inversions but, given that a large fraction of the *C. frigida* genome is impacted by polymorphic inversions, one can expect that these rearrangements play a significant role in structuring genetic variation and in enabling local adaptation. Spatial genetic structure and connectivity in *C. frigida* remain poorly-described although occasional long-distance migration bursts have been documented and regular dispersal is expected between nearby subpopulations occupying discrete patches of wrackbed (Egglishaw 1960; Dobson 1974). *Coelopa frigida* occupies a wide climatic range, being present from temperate to subarctic zones; it occurs along salinity gradients in the Baltic or St Lawrence R. Estuary, and also copes with variability in the quality and the composition of its wrackbed habitat (Egglishaw 1960; Dobson 1974). *Coelopa frigida* thus faces several sources of habitat heterogeneity, varying at both large and local scales, for which, depending on the scale of dispersal, a tightened genomic architecture may be favourable.

In the present study, we investigated how chromosomal inversions contribute to local adaptation across different scales of environmental heterogeneity, and how such

recombination modifiers shape the distribution of genetic diversity. To address this issue, using the seaweed fly *Coleopa frigida* as a biological model, we strived to adopt a systematic approach for localizing multiple chromosomal rearrangements, and we analysed genetic variation across several dimensions of environmental variation including a 1,500 km climatic gradient, a salinity gradient, and fine-scale, patchy habitat variation. We built the first reference genome assembly for *C. frigida* and leveraged the power derived from whole-genome sequencing of 1,446 flies. First, we analysed patterns of genetic polymorphism along the genome to identify putative recombination modifiers such as inversions. We expected to find several megabase-wide regions characterized by high linkage disequilibrium and strong differentiation between haplotypic clusters. Second, we examined the geographic genetic structure to assess the scale of connectivity across the landscape covered by our study. We tested the hypothesis that this species is characterised by high gene flow possibly opposed by recombination suppression, spatially-varying selection, and their interaction. Third, we tested genotype-environment and genotype-phenotype associations to ask what is the genetic architecture underlying adaptation to various sources of environmental variation acting at different geographic scales. We tested the prediction that recombination modifiers represent modular cassettes of adaptation with gene flow and uncovered contrasted dynamics between the different inversions, related to the modularity and geographic scale of adaptation.

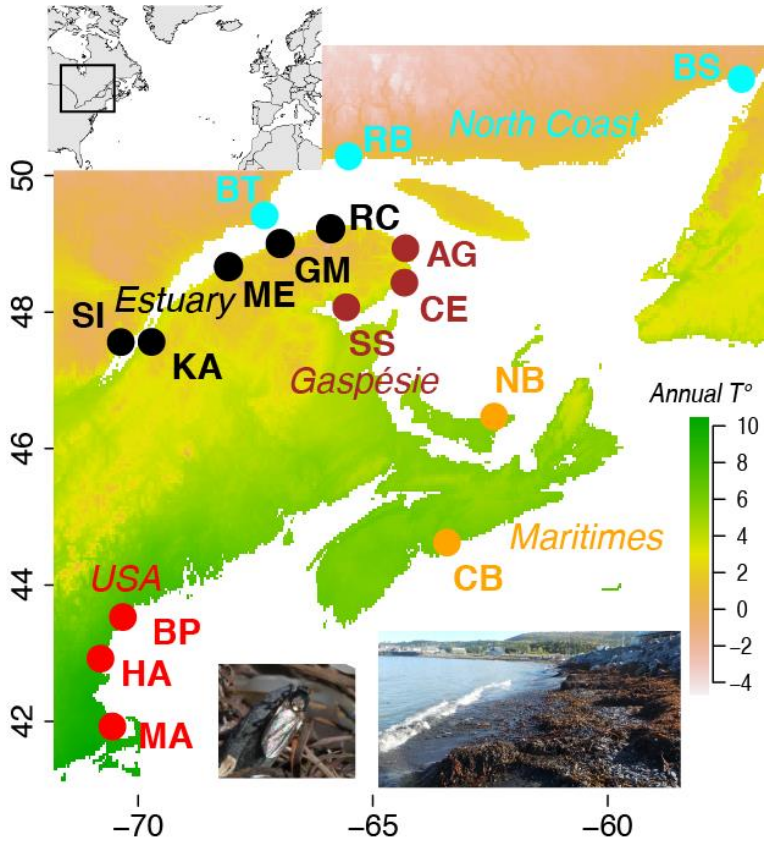

Figure 1: *Coelopa frigida* sampling across an environmental gradient
Map of the 16 sampling sites, coloured by geographic region. The background of the map displays the gradient of annual mean air temperature. The insert shows the location of the study area at a wider scale. Photos show *Coelopa frigida* and its habitat of seaweed beds.

Results

To facilitate the analysis of large chromosomal rearrangements and genome-wide variation, we built the first reference genome assembly for *Coelopa frigida* using a combination of long-read sequencing (PacBio) and linked-reads from 10xGenomics technology. A high-density linkage map (28,639 markers segregating across 6 linkage groups) allowed us to anchor and orientate more than 81% of the genome into 5 large chromosomes (LG1, LG2, LG3, LG4, LG5) and one small sex chromosome (LG6). This karyotype was consistent with previous cytogenetic work on *C. frigida* (Aziz 1975) and with the ancestral state in Diptera (Vicoso and Bachtrog 2015; Schaeffer 2018). The final assembly produced 6 chromosomes and 1832 unanchored scaffolds with a N50 of 37.7 Mb for a total genome size of 239.7 Mb. This reference had a high level of completeness, with 96% (metazoa) and 92% (arthropods) of universal single-copy orthologous genes completely assembled. It was annotated with a highly complete transcriptome (87% complete BUSCOs in the arthropods) based on RNA-sequencing of several ontogenetic stages and including 35,999 transcripts.

To analyse genomic variation at the population-scale, we used low-coverage (~1.4X) whole-genome sequencing of 1,446 flies from 16 locations along the North American Atlantic coast (88-94 adult flies/location). Sampled locations spanned a North-South gradient of 1,500km, over 10° of latitude, a pronounced salinity gradient in the St Lawrence Estuary, and a range of habitats with variable seaweed composition and wrackbed characteristics (Fig. 1, Table S1). After alignment of the 1,446 sequenced individuals to the reference genome, we filtered for quality and coverage and reported 2.83 million single-nucleotide polymorphisms (SNPs) with minor allelic frequency (MAF) higher than 5% for differentiation analyses.

- **Two large chromosomal inversions structure intraspecific genetic variation**

Decomposing whole-genome variation through a principal component analysis (PCA) revealed that the 1st and 2nd principal components (PCs) contained a large fraction of genetic variance, respectively 21.6 % and 3.9 %, and partitioned the 1,446 flies into 9 discrete groups (Fig 2A). Along PC1, the three groups corresponded to three genotypes of the inversion *Cf-Inv(1)* (αα, αβ, ββ), as identified with two diagnostic SNPs (Mérot et al. 2018) with respectively 100% and 98.3% concordance (Table S2). Along PC2, three distinct groups were identified that corresponded neither to sex nor geographic origins, and thus possibly represented three genotypes for another polymorphic inversion.

Admixture and clustering analyses supported the same strong structure as seen in the PCA (Fig. 2B). The two major genetic groups corresponded to the homokaryotypes of *Cf-Inv(1) (*αα and ββ), and the admixed individuals to the heterokaryotypes αβ. When increasing the number of genetic groups (K=4), we detected two additional genetic groups corresponding to the extreme clusters on PC2, with admixed individuals being the intermediate group.

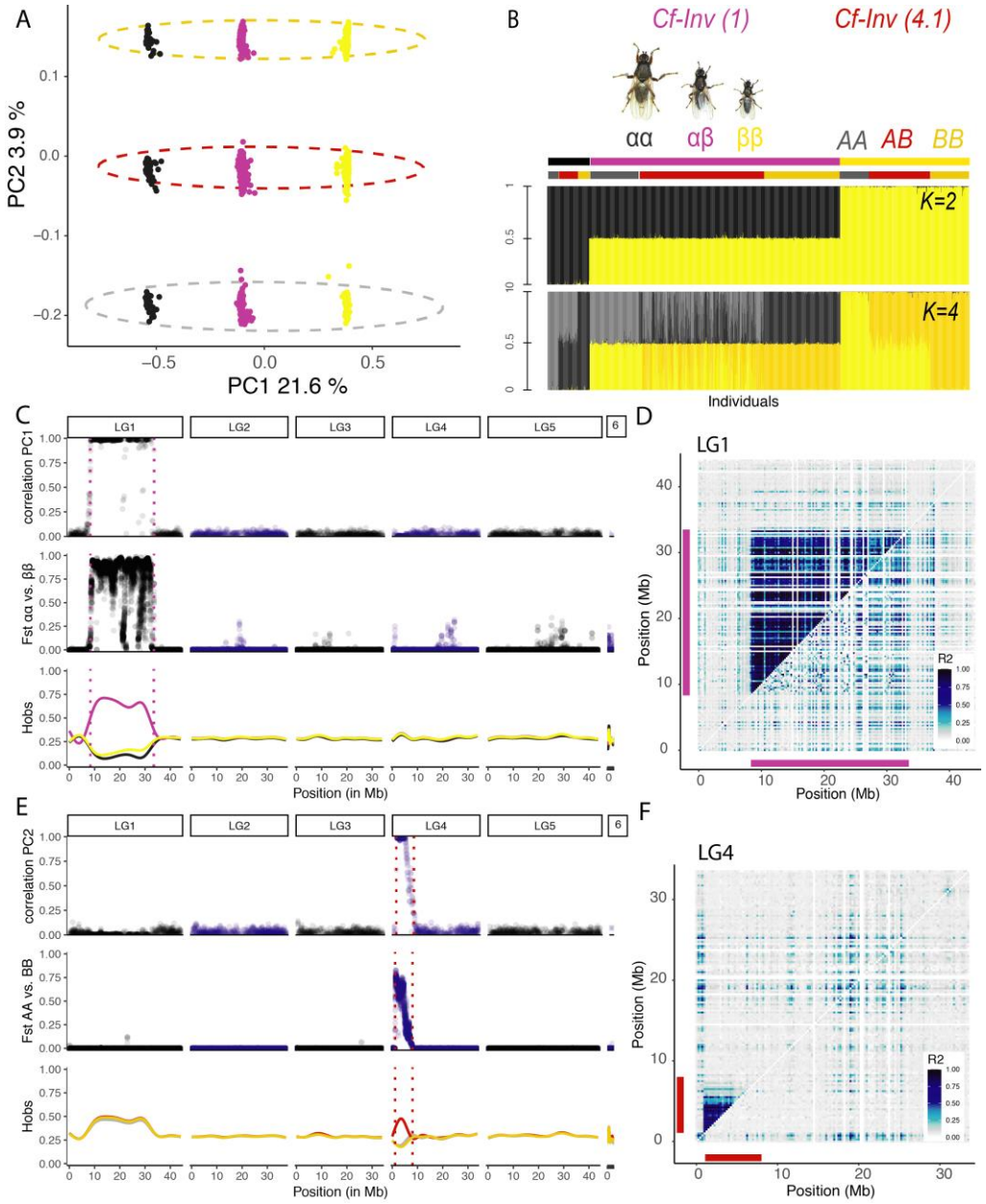

Figure 2: Two large chromosomal inversions structure within-species genetic variability

(A) Principal component analysis (PCA) of whole-genome variation. Individuals are coloured by karyotypes at the inversion *Cf-Inv(1),* as determined previously with a SNP marker (Mérot et al. 2018). Ellipse indicates secondary grouping along PC2. **(B)** Analyses of population structure and admixture based on Bayesian clustering with K=2 clusters and K=4 clusters. Each column represents the individual probability of belonging to a given cluster. Individuals are ordered based on their karyotype at *Cf-Inv(1),* and then based on karyotype at *Cf-Inv(4.1)* inferred from PC2 scores. **(C)** Along the genome, correlation between PC1 scores of local PCAs performed on windows of 100SNPs and PC1 scores of the PCA performed on the whole genome; FST differentiation between the two homokaryotypes of *Cf-Inv(1)* in sliding-windows of 25kb; and observed heterozygosity in the three karyotypic groups of *Cf-Inv(1)* smoothened for visualization. Dashed lines represent the inferred boundaries of the inversion *Cf-Inv(1)* **(D)** Linkage disequilibrium (LD) in LG1. The upper triangle includes all individuals and the lower triangle include homokaryotes for the most common arrangement. Bars represent the position of the inversion. **(E)** Correlation between PC1 scores of local PCAs performed on windows of 100SNPs and PC2 scores of the PCA performed on the whole genome; FST differentiation between the two homokaryotypes of *Cf-Inv(4.1)* in sliding-windows of 25kb; and observed heterozygosity in the three karyotypic groups of *Cf-Inv(4.1)* smoothened for visualization. Dashed lines represent the inferred boundaries of the inversion *Cf-Inv(4.1)* **(F)** Linkage disequilibrium (LD) in LG4. In both LD plots, the colour scale shows the 2nd higher percentile of the R^2 value between SNPs summarized by windows of 250kb.

To assess which regions of the genome reflected the patterns observed in the global PCA, we performed local PCA on windows of 100 SNPs along all the genome and evaluated the correlation between PC1 scores of each local PCA and PCs scores of the global PCA (Fig. 2C). PC1 was highly-correlated with a region of 25.1 Mb on LG 1, indicating the genomic position of the large *Cf-Inv(1)* inversion (Table 1). PC2 was highly correlated with a smaller region of 6.9 Mb on LG4 (Fig. 2G), consistent with the hypothesis of an inversion, hereafter called *Cf-Inv(4.1)*.

The two regions *Cf-Inv(1)* and *Cf-Inv(4.1)* presented several other characteristics typical of large polymorphic inversions with non-recombining haplotypic arrangements. First, differentiation was very high in the inverted region (Fig. 2D-F), reaching F_{ST} values up to 1 between the two homokaryotypes, and intermediate between the heterokaryotypes and homokaryotypes (Fig. S1). Almost no differentiation was observed between the karyotypes outside the inverted region. Second, the intermediate group on the PCA (heterokaryotypes) was characterized by high heterozygosity for all SNPs in the inverted region while extreme groups (homokaryotypes) showed a lack of heterozygosity (Fig 2E-G). Third, throughout the inverted region, linkage disequilibrium (LD) was very high when considering all individuals, but low within each group of homokaryotypes (Fig 2H-I), meaning that recombination is limited between the arrangements but occurs freely in homokaryotypes bearing the same arrangement.

- ***C. frigida*** **exhibit other regions affected by recombination modifiers including putative chromosomal rearrangements**

Looking for other putative polymorphic inversions, we re-analysed the local PCA performed along the genome and used a method based on multidimensional scaling (MDS) to identify clusters of PCA windows displaying a similar pattern (Li and Ralph 2019; Huang et al. 2020). Besides the aforementioned *Cf-Inv(1)* and *Cf-Inv(4.1)* inversions, that drove the 1st and 2nd axis of the MDS, we identified five outlier genomic regions across the different MDS axes (Fig.3, Fig. S2). In all five regions, a large proportion of variance was captured along the 1st PC (>50%), and linkage disequilibrium was high (Fig. 3A-B).

Two regions on LG4 represented convincing putative inversions of 2.7Mb and 1.4Mb, respectively. In both regions, the PCA displayed three groups of individuals with high clustering confidence, the central group was highly-heterozygous and the extreme groups were very divergent (Fig. 3E, Fig. S3). Genetic diversity was also higher or at the same level as the rest of the genome (pi > 0.01, Fig. S4). Karyotype assignment was the same between the two putative inversions, indicating that they are either tightly-linked or belong to a single inversion. The hypothesis of two linked inversions seemed more plausible because the high density of linkage map markers and the non-null recombination rate across this area of 50 cM provided confidence in the genome assembly and supported a gap of 5 Mb between them. Moreover, previous cytogenetic work showed that one chromosome of *C. frigida* exhibits a

polymorphic inversion on one arm (possibly *Cf-Inv(4.1)*) and, on the other arm, two polymorphic inversions which rarely recombine (Aziz 1975). Both inversions were subsequently analysed together and called *Cf-Inv(4.2)* and *Cf-Inv(4.3).*

The other three regions, spanning 6.8 Mb on LG2, 6.3 Mb on LG3 and 16.7 Mb on LG5, represented complex areas that behaved differently than the rest of the genome. Recombination was locally reduced, both in the linkage map and in wild populations, as indicated by the linkage disequilibrium (Fig. 3A-C). Those regions included several clusters of outlier windows, supporting non-recombining haplotypic blocks of medium size (<1Mb or < 10 000 SNPs) that appeared partially linked (Fig. S5-S7). These blocks, characterised by high diversity, were interspersed by regions of low diversity (Fig. S4). Hence, while the whole area or some clusters may correspond to structural rearrangements, possibly nested, complex or misassembled, we conservatively chose, in the absence of more information, to consider those three portions of the genome as "low-recombining regions".

Accordingly, the fraction of the genome subsequently called "collinear" excluded the seven regions identified as "recombination modifiers", the four inversions (*Cf-Inv(1)*, *Cf-Inv(4.1),* and the linked *Cf-Inv(4.2) Cf-Inv(4.3)*) as well as the three low-recombining regions (subsequently called *Cf-Lrr(2)*, *Cf-Lrr(3)*, *Cf-Lrr(5)*)

Table 1: Name, position and characteristics of the putative inversions and regions appearing as cluster of outlier windows in the local PCA analysis.

PC1 var. indicate the variance explained by a PCA run on SNPs within the target region, and Sum of squares (bet/tot) indicates the proportions of between-cluster sum of squares in k-means clustering.

Name	Status	Chr.	start	stop	size (MB)	Number of SNPs	PC1 var. (%)	Sum of Squares (bet/tot)
Cf-Inv(1)	*Known inversion*	LG1	8342182	33487673	25.1	441000	75	0.99
Cf-Inv(4.1)	*Probable inversion*	LG4	1088816	7995568	6.9	137700	57	0.99
Cf-Inv(4.2)	*Probably two linked*	LG4	22421881	25145365	2.7	17300	23	0.95
Cf-Inv(4.3)	*inversions*	LG4	30622035	31991919	1.4	23100	28	0.99
Cf-Lrr(2)	*Low-recombination region*	LG2	14083320	20869940	6.8	41000	17	0.88
Cf-Lrr(3)	*Low-recombination region*	LG3	7486933	13829649	6.3	38600	27	0.92
Cf-Lrr(5)	*Low-recombination region*	LG5	15940464	32665323	16.7	134400	13	0.82

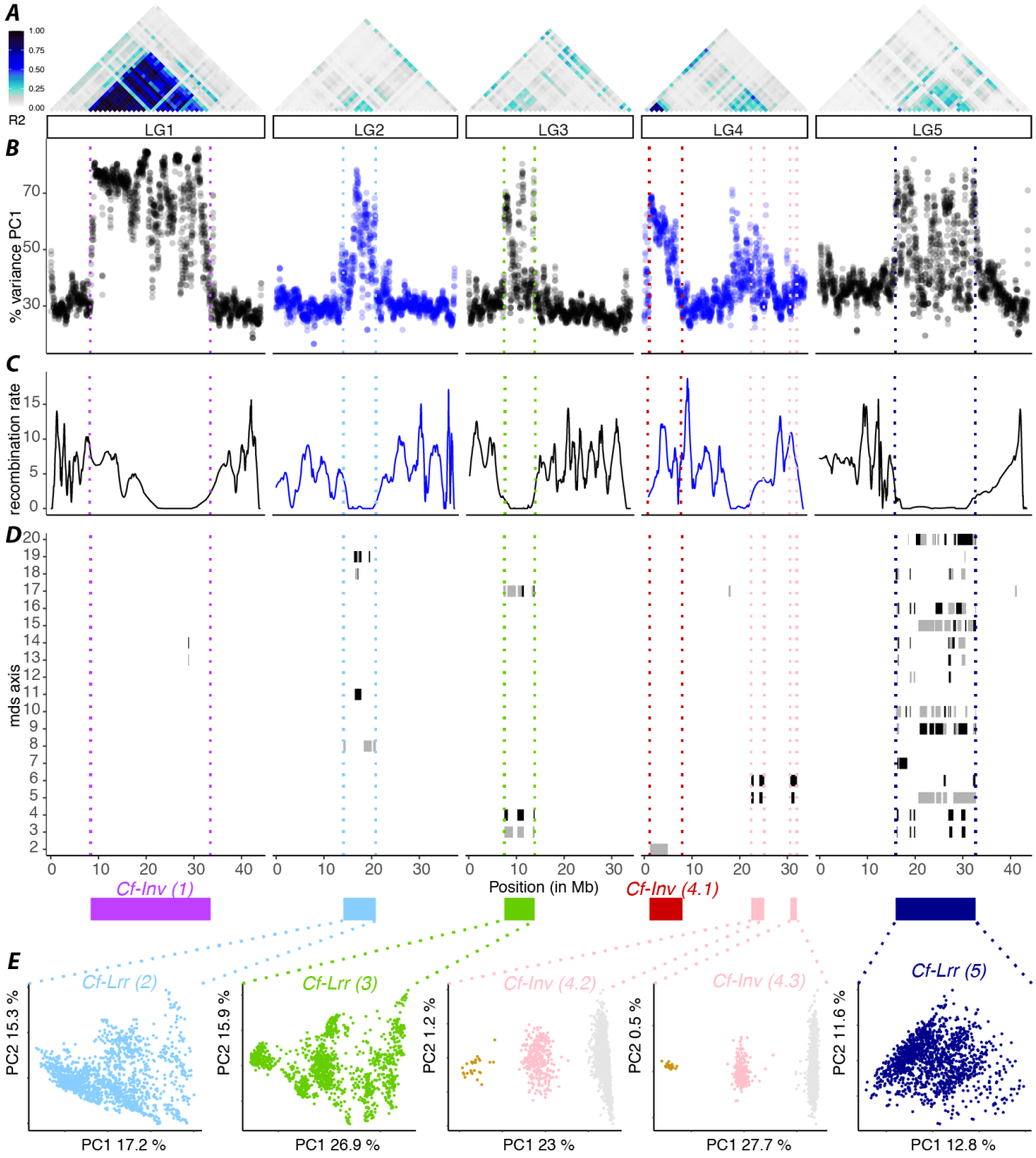

Figure 3: Detecting other regions affected by recombination modifiers

(A) LD across the 5 major chromosomes expressed as the 2nd higher percentile of the R^2 value between SNPs summarized by windows of 1Mb. **(B)** Proportion of variance explained by the 1st PC in local PCAs performed on windows of 100 SNPs along the genome, average by sliding-windows of 100kb (step 20kb). **(C)** Recombination rate (in cM/Mb) inferred from the linkage map, smoothened with a loess function accounting for 10% of the markers. **(D)** Position along the genome of clusters of local PCA windows scored as outliers (>4sd) along each axis of the MDS, at the upper end in black, and the lower end in grey. Coloured rectangles indicate the position of the inversions and the regions of interest gathering outlier clusters or putative inversions. Dashed lines represent their inferred boundaries across all plots. **(E)** PCA performed on SNPs within each region of interest. For the two regions on LG4 that appear as two linked putative inversions (*Cf-Inv(4.2)* and *Cf-Inv(4.3)*), three clusters were identified with high-confidence and coloured as putative homokaryotes and heterokaryotes. The same colours are used in both regions since karyotyping was consistent across all individuals.

- **Geographic genetic structure show contrasting signals within inversions and low-recombining regions**

Geography also played a major role in structuring genetic variation. On the PCA, the 3rd PC, which explained 1.4% of variance, spread genetic variation between individuals along the North-South gradient (Fig. 4A). Differentiation between pairs of populations, measured as F_{ST} on a subset of LD-pruned SNPs, also followed the North-South gradient but was globally weak (F_{ST} = 0.003 to 0.016, Fig. 4B)), corresponding to an estimation of about 15 to 80 N_em (migrants) per generation. Gene flow was even more pronounced at small scale, as we detected a strong signal of Isolation-By-Distance (IBD) when examining the correlation between genetic distances and Euclidean distances among the 16 populations (R^2=0.45, F=97, p<0.001, Table S3). We also highlighted a pattern of Isolation-By-Resistance (IBR) since the model including least-cost distances along the shoreline (R^2=0.63, F=199, p<0.001, Table 2) was better supported by the data than the model including Euclidean distances (ΔAIC=47, Table S3).

These IBD and IBR patterns varied significantly along the genome indicating a strong role for recombination in modulating them. When considering all SNPs, pairwise differentiation was more heterogeneous (F_{ST}=0.002 to 0.021, Fig. 4B) and IBR was much weaker, albeit significant (R^2=0.19, F=29, p<0.001) than when considering LD-pruned SNPs or collinear SNPs. We thus calculated pairwise F_{ST} between pairs of populations based on different subsets of SNPs, either from each recombination modifier or from the collinear genome.

While all recombination modifiers affected the geographic distribution of polymorphism they did so in different ways. The inversions were the most differentiated genomic regions between populations in comparison to the collinear genome (Table S3, Fig. S8). Within the *Cf-Inv(1)* inversion, there was no association between genetic and geographic distances (Fig. 4C, Table 2). In contrast with the collinear genome, genetic differentiation within the inversion *Cf-Inv(1)* was very variable both between nearby or distant populations. Conversely, variation within the LG4 inversions showed significant IBD/IBR patterns with a slope of correlation between genetic and geographic distances significantly steeper than in collinear regions (Fig. 4C-D, Table 2, Table S4, Fig. S9). The strong divergence between northern and southern populations was mirrored by a sharp latitudinal cline of inversion frequencies, ranging from 0.27 to 0.75 for *Cf-Inv(4.1)* and from 0.02 to 0.26 for *Cf-Inv(4.2/4.3)*, much steeper than random SNPs with similar average frequency (Fig 4E, Fig. S10). Within the three low-recombining regions, we also observed significant IBD/IBR. Compared to collinear regions of the same size, the slope of the correlation between genetic and geographic distances was significantly steeper for *Cf-Lrr(2)* and *Cf-Lrr(5)* but not for *Cf-Lrr(3)* (Fig. 4D, Table 2, Table S4, Fig. S9). Overall, the different recombination modifiers showed significantly different patterns, indicating that genetic differentiation was modulated by processes other than the migration-drift balance, possibly at different geographic scales for *Cf-In(1R)* vs. the other modifiers.

Table 2: Association between genetic distance and geographic distances measured as least-cost distances along the shoreline (Isolation-by-resistance) for the different fractions of the genome. Numbers between brackets indicate the limits of the 95% distribution of the slope coefficient. The comparison to collinear regions displays the output of a full model comparing each region to the collinear genome, providing the direction and the significance (*) of the interaction term.

SNP subset	R^2 adjusted	F	p-value	intercept	slope coefficient	Comparison
All	0.19	29.3	<0.001	0.0085	0.0020 [0.0013-0.0027]	
Collinear	0.54	138.6	<0.001	0.0062	0.0019 [0.0015-0.0022]	
LD pruned	0.63	199.5	<0.001	0.0057	0.0021 [0.0018-0.0024]	
Cf-Inv(1)	-0.01	0.3	0.59	0.0137	-0.0006 [-0.0032-0.0018]	- *
Cf-Inv(4.1)	0.29	49.4	<0.001	0.0172	0.0134 [0.0096-0.0172]	+ *
Cf-Inv(4.2/4.3)	0.50	121.5	<0.001	0.0075	0.0030 [0.0025-0.0036]	+ *
Cf-Lrr(2)	0.44	95.4	<0.001	0.0074	0.0028 [0.0023-0.0034]	+ *
Cf-Lrr(3)	0.49	113.1	<0.001	0.0066	0.0019 [0.0016-0.0023]	n.s.
Cf-Lrr(5)	0.55	147.2	<0.001	0.0080	0.0033 [0.0028-0.0038]	+ *

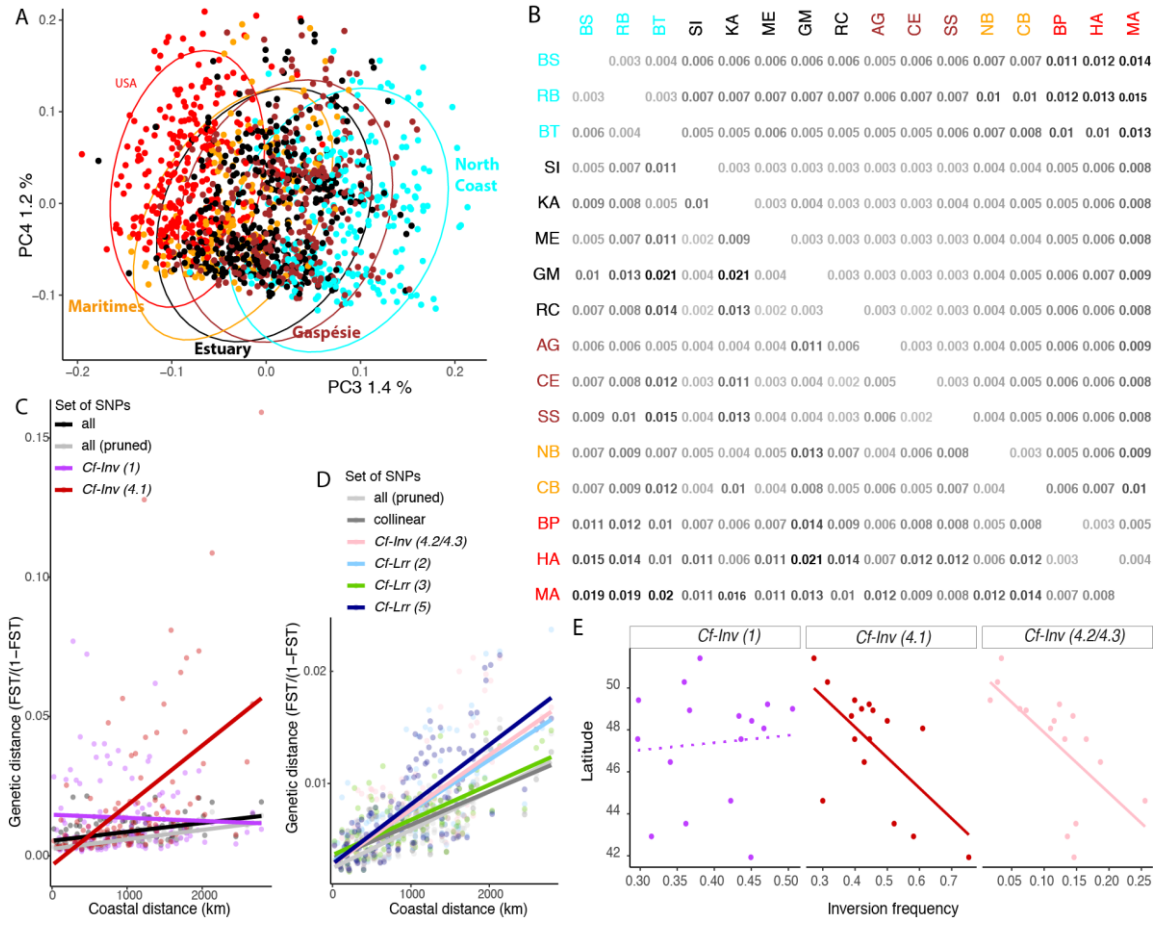

Figure 4: Genetic variation is geographically structured along a North-South gradient and display isolation-by-distance

(A) 3rd and 4th principal component of a PCA on whole-genome variation. Individuals are coloured by their geographic region, as on figure 1 **(B)** Pairwise FST between all population pairs, ordered by proximity from North to South and coloured by geographic regions. The values above main diagonal shows FST based on LD-pruned SNPs and those below main diagonal show FST based on all SNPs **(C-D)** Isolation by distance/Isolation by resistance displayed as the association between genetic distance (FST/(1-FST) and the distance by the least-cost path following the coast. Colours denote the subset of SNPs used for the calculation of the FST. The results are displayed in two panels with different y scale to better display the lower values. **(E)** Latitudinal variation of inversion frequencies.

- **Putative adaptive diversity co-localizes with regions affected by recombination modifiers**

To investigate patterns of adaptive variation *C. frigida*, we analyzed the association between SNP frequencies and environmental variables at local (abiotic and biotic characteristics of the wrackbed habitat) and large (thermal latitudinal gradient and salinity gradient in the St. Lawrence R. Estuary) spatial scales (Fig. 1, Fig. S11, Table S1). Analyses with two different GEA methods (latent factor mixed models and Bayesian models) showed consistent results highlighting high peaks of environmental associations and large clusters of outlier SNPs within the inversions or low-recombining regions, yet varying between the different environmental factors and spatial scales (Fig. 5A-E, Table 3, Table S5, Fig. S12-13). We considered SNPs consistently identified across analyses to be putatively adaptive.

At a large geographic scale, association with climatic variation along the latitudinal gradient showed a strong excess of outlier SNPs in the four inversions and the low-recombining regions of LG2 and LG5. Those regions exhibited particularly strong peaks of association (BF >50, Fig. 5A) and 2 to 5 times more outliers than expected by chance (Table 3). However, this was not the case for *Cf-Lrr(3)*. These results were consistent whether or not the model was controlled by the geographic population structure (Fig. S12-S13). Association with thermal variation only (without accounting for precipitation) was extremely strong in inversion *Cf-Inv(4.1)* which contained 36% of outliers (odds ratio of 7, Fig. S14). Variation along the salinity gradient, which also spanned variation in tidal amplitude, was significantly associated with a more limited number of SNPs but a large excess of such outliers were found in *Cf-Lrr(3)* and *Cf-Lrr(5)* (Tab.3).

Table 3: Genomic repartition of candidate SNPs associated with environmental variables
Repartition of the candidate SNPs associated with each environmental variation using the combination of two GEA methods. N is the number of outliers SNPs within a given region, % is the proportion of the outliers found in this region and OR indicate the odd-ratio. Values in bold with a star indicate significant excess of candidate SNPs in a Fisher exact test. Results obtained for each GEA method are presented in Table SX.

	Tested SNPs		**Climate**			**Salinity**			**Bed abiotic characteristics**			**Algal composition (PC1: Laminaria/Fucus)**			**Algal composition (PC2)**		
	N	%	N	%	OR	N	%	OR	N	%	OR	N	%	OR	N	%	OR
All	1155978		3635			509			780			372			2740		
Collinear	814279	70%	556	15%	0.2	301	59%	0.8	163	21%	0.3	254	68%	1.0	390	14%	0.2
Cf-Inv (1)	176963	15%	1474	41%	**2.6***	64	13%	0.8	584	75%	**4.9***	77	21%	**1.4***	1494	55%	**3.6***
Cf-Inv (4.1)	57323	5.0%	480	13%	**2.7***	15	2.9%	0.6	11	1.4%	0.3	14	3.8%	0.8	33	1.2%	0.2
Cf-Inv (4.2/4.3)	17019	1.5%	111	3.1%	**2.1***	8	1.6%	1.1	8	1.0%	0.7	3	0.8%	0.5	26	0.9%	0.6
Cf-Lrr(2)	20458	1.8%	93	2.6%	**1.4***	6	1.2%	0.7	9	1.2%	0.7	3	0.8%	0.5	15	0.5%	0.3
Cf-Lrr(3)	16313	1.4%	11	0.3%	0.2	28	5.5%	**3.9***	0	0.0%	0.0	3	0.8%	0.6	7	0.3%	0.2
Cf-Lrr(5)	53623	4.6%	910	25%	**5.4***	87	17%	**3.7***	5	0.6%	0.1	18	4.8%	1.0	775	28%	**6.1***

At finer geographic scale, outlier SNPs associated with wrackbed abiotic characteristics (depth, temperature and salinity) were strongly enriched within the inversion *Cf-Inv(1)* with an odds-ratio of 5, including outliers with very strong support (BF >20, Fig. 4C). No other recombination modifiers showed such enrichment or association with wrackbed abiotic characteristics. Variation in algal composition of the wrackbed, driven by the relative abundance of two dominant seaweeds, Fucaceae or Laminariaceae, was significantly associated with outliers SNPs quite widespread in the genome although they were overrepresented in the inversion *Cf-Inv(1)* by an odds-ratio of 1.4. Variation in secondary components of the substrate were more difficult to interpret as they co-varied with latitude and temperature (Fig. S11), but it was also associated with a large number of SNPs in the inversion *Cf-Inv(1)* and in *Cf-Lrr(5)* with odds-ratio of 3.6 to 6 (Fig. 5E).

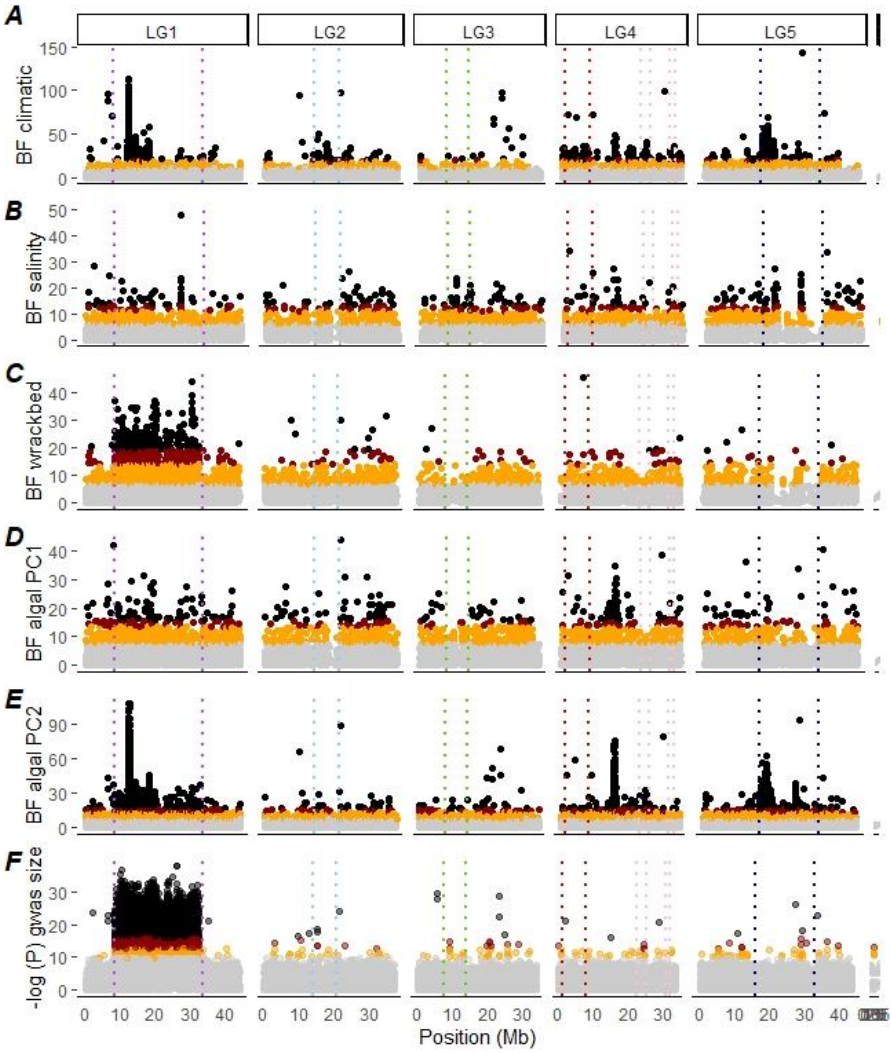

Figure 5: Environmental and phenotypic associations

Candidate SNPs associated with **(A)** climatic variation along the North-South gradient, **(B)** salinity variation along the Estuarian gradient, **(C)** variations in abiotic characteristics of the wrackbed habitat, **(E-F)** variation in wrackbed algal composition. The manhattan plot shows the Bayesian factor from the environmental association analysis performed in Baypass, controlling for population structure. **(G)** Candidate SNPs associated with wing size. The manhattan plot shows the pvalues from the GWAS. Points are coloured according to false-discovery rate (black: <0.00001, red: <0.0001, orange: <0.001). Dashed lines represent the inferred boundaries of all regions affected by recombination modifiers.

- **Genotype-phenotype association**

Variation in wing size (a proxy for body size) was strongly associated to the inversion *Cf-Inv(1)* using GWAS. Among the 124,701 candidate SNPs identified by the GWAS, more than 99.8% were located within the inverted region (Fig. 5F). In contrast, when running the same analysis on each group of homokaryotypes, we found almost no candidate SNPs associated with size variation (0 on the group of αα, and up to 3 SNPs when lowering the FDR to p=0.01 for the group ββ, Fig. S15). Gene ontology analysis among the SNPs significantly associated with size or among the genes present in the inversion *Cf-Inv(1)* unveiled an enrichment in several biological processes (Table S6-S7) among which morphogenesis, muscle development or neural system development, all consistent with large differences in wing size and life-history.

As a measure of thermal adaptation, we evaluated the recovery time after a chill coma in the fly progeny genotyped to build the linkage map. Cold-shock resistance localized to a QTL on LG4, which explained about 13% of the variation (Fig. S16). The main peak was located on LG4 around 25-28Mb. This broad QTL encompassed multiple outliers SNPs associated with climatic variation, and multiple annotated genes, among them two heat-shock proteins, which may represent relevant candidates for thermal adaptation (Uniprot P61604 at position 25,128,992 and P29844 at position 26,816,283). The peak was also between the two putative inversions *Cf-Inv(4.2)* and *Cf-Inv(4.3),* and there was a secondary peak at 8MB, around the breakpoint of *Cf-Inv(4.1).* It is noteworthy that the whole chromosome LG4 showed a high LOD value above 2, higher than any markers on another chromosomes (Fig. S16), possibly due to non-recombining paternal markers showing partial association with cold-shock resistance.

Discussion

Analysing more than a thousand whole genomes of *C. frigida* flies revealed that a large fraction of the genome is affected by recombination modifiers, including four large chromosomal inversions. These megabase-long stretches of the genome appear to play a predominant role in shaping genetic variation across two large-scale environmental gradients as well as heterogeneous patchy habitats. Yet, the different inversions showed contrasting patterns, which may be related to different selective forces acting on them since they both were disproportionately enriched in putatively adaptive variants strongly associated with non-overlapping ecological factors or phenotypes. In particular, the largest inversion *Cf-Inv(1)* was associated with body size and co-varied at a fine-scale with wrackbed habitat characteristics while inversion on LG4 displayed clinal variation along a geo-climatic gradient. Below, we discuss how our results provide new insights into the evolutionary role played by recombination modifiers such as inversions, and how our data suggest that those regions represent modular cassettes for local adaptation at different geographic scales in the face of high gene flow.

- **Low-coverage sequencing of a large fly dataset provides deep insights into genetic variation across species range and individual genomes**

Studying genetic variation across a species range is more accurate and powerful when sampling encompasses both fine and coarse geographical scales and multiple environmental conditions. When searching for signatures of adaptation or putative rearrangements an additional high density of genetic markers is required (Fuentes-Pardo and Ruzzante 2017). This creates the need to balance effort across the number of samples, the portion of the genome sequenced (i.e. reduced representation or whole-genome re-sequencing), and the depth of sequencing. To maximise insights, we sequenced the whole genome of 1,446 wild-collected flies but reduced individual coverage to about 1.4X. This strategy has been used efficiently in few pioneer studies in human genomics (Martin et al. 2020), conservation genomics (Therkildsen et al. 2019) and population genomics (Clucas et al. 2019). Simulations confirmed that sequencing many samples at low depth (1X) provides more robust estimates of allele frequencies and accurate population parameters than sequencing few samples at high depth (Alex Buerkle and Gompert 2013), particularly so with high coverage at the population level, as in this study. Additionally, thanks to a low-cost barcoding library preparation (Therkildsen and Palumbi 2017), individual information was retained, which allowed parameters that require individual information (LD, Hobs) to be accurately calculated and to perform individual analysis such as phenotypic associations. Importantly, allelic frequencies were also unbiased by *a priori* or unbalanced pooling as it may happen in pool-seq (Fuentes-Pardo and Ruzzante 2017), and any grouping could be subsequently chosen for the analyses.

Individual whole-genome sequencing at low coverage allowed us to uncover the genetic structure associated with structural rearrangements in *C. frigida* and to analyze environmental parameters and phenotypes potentially associated with those rearrangements. First, the large sample size brought power to make the most of a recently developed method of indirect inversion detection (Li and Ralph 2019; Huang et al. 2020). For instance, we would likely have missed the inversion(s) *Cf-Inv(4.2/4.3)* with smaller sample size, since the rare homokaryotype frequency was below 2% (26/1446 individuals). Second, the extensive density of markers along the genome provided accurate locations for the major inversions although characterizing the exact breakpoints remained challenging without long-read sequencing (Ho et al. 2019). Third, the retention of individual information allowed us to split the dataset into sub-groups of karyotypes as determined from the analyses of sequences and to characterize LD, heterozygosity, nucleotide diversity and the differentiation within and between karyotypes for all inversions. In fact, unlike systems in which several inversions co-vary along eco-geographic gradients (Berg et al. 2017; Christmas et al. 2018; Faria, Chaube, et al. 2019; Huang et al. 2020), or appear to do so given the sampling or sequencing design, we observed independence between the different inversions in *C. frigida*, and found contrasting dynamics from a geographic and ecological point of view.

- **Polymorphic inversions structure within-species genetic diversity**

Contrary to our expectations when studying such a wide geographic area for a small insect, we uncovered that the major factor explaining genetic variation in *C. frigida* was not geographic distances but structural rearrangements. Despite more than 1,500km (or 3,000 km of coastal distance) between the most distant populations, geographic genetic differentiation was very weak (Maximal F_{ST} <0.02). This is much lower than other coastal specialised insects such as the saltmarsh beetle *Pogonus chalceus* (F_{ST} ~0.2, (Van Belleghem et al. 2018) but comparable to small Diptera with large distributions like *Drosophila melanogaster* or *D. simulans* which typically exhibit F_{st} around 0.01-0.03, likely resulting from both high migration rate and large effective population size *Ne* (Machado et al. 2016; Kapun et al. 2020). Despite this weak genetic structure, we detected a strong signal of isolation-by-distance indicating that dispersal among populations and subsequent gene flow decreases with distance. Furthermore, our analyses also showed that the least-cost distance of coastline better explained genetic variation than Euclidean distance. This isolation-by-resistance pattern likely results from a stepping stone dispersal process (Gandon and Rousset 1999) where the absence of suitable habitat patches in mainland and marine areas drives gene flow along the coastline and constraints genetic connectivity.

In contrast with the overall weak geographic genetic structure, the haplotypic groups for the different inversions were highly differentiated, with fixed allelic differences between inverted sequences. For the inversion *Cf-Inv(1),* such genetic divergence was related to extreme body size differences, with αα males being three times bigger than ββ males (Butlin and Day 1985; Mérot et al. 2018). Such a high phenotypic and genotypic divergence between alternative arrangements are comparable to many other old inversions (Hoffmann and Rieseberg 2008; Wellenreuther and Bernatchez 2018). With a mean F_{ST} across the entire inversion of 0.27, co-occurring αα and ββ karyotypes also appear as differentiated as closely-related species (Roux et al. 2016), albeit with a very different landscape of genetic differentiation. With a high level of genetic and phenotypic divergence, often associated to phenotypic differences, inversion polymorphisms thus challenge the view of species as homogeneous units. Each non-recombining arrangement protects standing haplotypic variation that structures biological diversity at the intra-specific level.

- **Chromosomal inversions represent modular cassettes for adaptation to heterogeneous environments**

Across geographic and ecological gradients, inversions may contribute strongly to genetic differentiation and often appear as islands of differentiation (Hoffmann et al. 2004). For instance, in the mosquito *Anopheles gambiae,* genetic differentiation along a latitudinal cline is almost entirely concentrated in two inversions (Cheng et al. 2012). In the marine snails *Littorina saxatilis,* genetic variation between habitats is largely driven by several inverted regions (Morales et al. 2019). *Coelopa frigida* follows this trend since pairwise F_{ST} between

populations was higher within the inversions than within collinear regions, albeit at a different geographic scale for the different inversions. Along the North-South gradient, differentiation between populations was higher and isolation-by-distance was stronger within the inversions *Cf-Inv(4.1)* and *Cf-Inv(4.2/4.3)* than in collinear regions. F_{ST} based on SNPs within an inversion combined two levels of genetic variation because differentiation between populations was driven by frequency variation at each highly differentiated arrangement. Such frequencies showed strong latitudinal clines, resembling the clines observed for several inversions in *Drosophila* that are maintained by selection-migration balance (Kapun et al. 2016). In sharp contrast, the genetic differentiation within the inversion *Cf-Inv(1)* did not depend on geographic distances among populations. This pattern was related to the heterogeneous frequency of the α/β arrangements, which vary at a fine spatial scale but did not vary clinally. Yet, both the clines of *Cf-Inv(4.1)/ Cf-Inv(4.2/4.3)* and the heterogeneity of *Cf-Inv(1)* contrasted with the homogeneous frequency of collinear variants, supporting the hypothesis that ecological factors rather than the mere geographic distance determine inversion distribution, which subsequently modulates the genomic landscape of population differentiation at small and large geographic scales.

Genotype-environment associations (GEA) suggest a potential evolutionary role of inversions in adaptation to ecological conditions in *C. frigida*. Our analyses showed that inversion frequencies correlated to environmental variation and candidate SNPs disproportionately occurred in inverted regions. Here, one question that may arise is whether GEA can be biased by LD and whether SNPs comprised within a given inversion are more likely to be detected as outliers. We avoided such artefact by following the guidelines and best practices tested with simulations including neutral inversions or low-recombining regions (Lotterhos 2019). However, it remains that genome scan analyses are more likely to detect regions with strong divergence that are resistant to swamping by migration, while dispersed, transient or small-effect alleles are harder to detect (Yeaman 2015). Moreover, because of the high linkage within an inversion, several SNPs may not be causative but simply linked to an adaptive variant. Hence, the high density of outlier SNPs within inversions means neither that they are full of adaptive alleles, nor that they are the only variants relevant for local adaptation. Nevertheless, this high number of outliers combined with some of the strongest association statistics point to inversions as major and true candidates for adaptation to heterogeneous environments in *C. frigida*. This follows the prediction that genomic architectures like inversions, which can increase linkage disequilibrium between adaptive alleles, is likely to preserve clusters of adaptive alleles compared to regions characterized by a less-clustered architecture (Kirkpatrick 2010; Yeaman 2013). As such, the seaweed fly *C. frigida* joins an accumulating number of examples of species carrying ecologically-relevant inversions that are involved in local adaptation despite high gene flow (Joron et al. 2011; Lindtke et al. 2017; Wellenreuther and Bernatchez 2018; Todesco et al. 2019; Huang et al. 2020).

In many empirical cases, when several inversions are found in the same species, they tend to vary along the same environmental axis. For instance, in the silverside fish *Menidia menidia,*

several inverted haploblocks covary along a latitudinal gradient (Tigano et al. 2020; Wilder et al. 2020). The same tendency is observed for several inversions differentiating mountain and plain African honeybees *Apis mellifera scutellata* (Christmas et al. 2018), or dune and non-dune ecotypes of sunflower *Helianthus petiolaris* (Todesco et al. 2019; Huang et al. 2020). In contrast, for *C. frigida,* we observed two contrasting evolutionary patterns: The inversion *Cf-Inv(1)* was associated with wrackbed characteristics and composition which represents patchy habitats at a fine geographic scale. It also functions as a supergene for body size, a trait which is usually polygenic yet appears in *C. frigida* to be controlled largely, if not entirely, by this inversion. The ecological and phenotypic associations are consistent with previous work on European and American populations (Day et al. 1983; Butlin and Day 1985; Berdan et al. 2018; Mérot et al. 2018). They reflect how the quality, composition and depth of the wrackbed, possibly reflecting its stability, differently favour the opposite life-history strategies associated the inversion. The β arrangement provides quick growth and smaller size while the α arrangement provides high reproductive success linked to a larger size but at the expense of longer development time. This ecologically-related trade-off combined with heterozygote advantage results in strong balancing selection, confirmed in our data by strong heterozygote excess (Mérot et al. 2018; Mérot, Llaurens, et al. 2020). Conversely, the inversions *Cf-Inv(4.1)* and *Cf-Inv(4.2/4.3)* show no deviation from Hardy-Weinberg disequilibrium and a widespread polymorphism displaying geographic structure along a latitudinal cline. As *Cf-Inv(4.1)* and *Cf-Inv(4.2/4.3)* are associated with climatic variables, located nearby a QTL for recovery after chill coma, we suggest that they likely play a significant role in thermal adaptation.

Taken together, these results support the hypothesis that these different inversions may favour local adaptation along different axes of the ecological niche and at different scales of local adaptation, thus representing modular adaptive cassettes. Second, these inversions follow different evolutionary dynamics driven by different shapes of selection, *Cf-Inv(1)* being a cosmopolitan polymorphism, likely to be maintained over evolutionary times by balancing selection, while *Cf-Inv(4.1)* and *Cf-Inv(4.2/4.3)* represent divergent polymorphism, likely to transition towards fixed differences between lineages under directional or disruptive selection (Faria, Johannesson, et al. 2019).

- **Exploring low-recombination regions: what are they and why do they matter?**

Beyond the aforementioned inversions, analyzing PCA along the chromosomes also identified additional regions that are characterized by distinct haploblocks and high LD. Such outlier regions may result from introgression, structural variants, and linked selection (Li and Ralph 2019). Introgression seems unlikely in our case given the absence of any sympatric sister species in our study area. The genetic map confirmed that those outlier regions are characterized by low recombination. Yet, current data does not allow us to determine whether this is due to structural rearrangements or other recombination modifiers (e. g. centromeres). Given the size of those regions (from 6 to 16MB), it could also due to a combination of

misassembled structural rearrangements embedded in a low-recombining region; haploblocks that are seemingly separated could be adjacent. Our reference genome was scaffolded and ordered based on a linkage map from one family. Hence, inversions that are heterozygous in the mother, as well as any low-recombining regions, possibly cluster into flat and long portions of the map, where marker ordering may be less accurate. The landscape of nucleotide diversity was also very heterogeneous: low-recombining regions showed diversity peaks, comparable to that recorded in inversions, separated by deserts of diversity, as expected under linked selection in low-recombining areas. Additional data such as long-reads or connected molecule like Hi-C would be needed to improve the quality of the assembly in those specific areas and better characterize their DNA content. Despite these cautionary notes, our analysis provides an early annotation of putative structural rearrangements, or at least of regions affected by recombination modifiers, that do not behave like the rest of the genome in terms of geographic genetic structure and association with environmental factors.

Like inversions, the low-recombining regions differentiated populations more strongly than collinear regions and possibly included haplotypic variants involved in local adaptation. They were enriched in candidate SNPs, being associated with climatic variation for the *Cf-Lrr(2)* and *Cf-Lrr(5)*, and with the salinity cline for *Cf-Lrr(3)*. Without certainty about the mechanisms behind the reduced recombination, we can only propose alternative hypotheses about the evolutionary processes at play. First, if these regions are complex or misassembled structural variants, they would represent additional adaptive rearrangements contributing to modular adaptation in *C. frigida*, with different haplotypes bearing one or several locally-adapted alleles. If those regions are centromeric, or simply rarely recombining, they would highlight the importance of linked selection in structuring intra-specific variation and conversely the relevance of low-recombining regions in protecting locally adapted alleles. Evidence for an important evolutionary role of low-recombining regions is increasingly reported now that we can analyse genomic landscapes in the light of recombination. For instance, in three-spine stickleback (*Gasterosteus aculeatus*), putatively adaptive alleles tend to occur more often in regions of low recombination in populations facing divergent selection pressures and high gene flow (Samuk et al. 2017). Similarly, regions of low recombination are enriched in loci involved in parallel adaptation to alpine habitat in the Brassicaceae *Arabidopsis lyrata* (Hämälä and Savolainen 2019). While further work is needed to characterize the complex low recombining regions in *C. frigida,* identifying several of these regions as outliers of differentiation highlights how recombination may matter more than geographic distances in structuring intra-specific variation and in modulating the selection-migration balance.

Conclusions

Our findings support the growing evidence that large chromosomal inversions play a predominant evolutionary role in organisms characterised by extensive connectivity across a large geographical range. In those widespread flying insects, like in several marine species, migratory birds, and plants, rearrangements strongly structure genetic diversity and represent a key genetic architecture for adaptation in the face of gene flow. More importantly, perhaps,

the different haplotypic blocks appear adaptive for different selective constraints across a range of geographical scales, hence forming a modular architecture allowing not only to cope with gene flow but also with various sources and scales of environmental heterogeneity. More broadly, our analysis also highlights the importance of regions of low recombination in structuring intra-specific genetic variation and favouring environmental adaptation. With recombination varying both along the genome and between individuals or haplotypes, inversions may represent only the simplest aspect of a complex relationship between recombination, selection and gene flow that we are just starting to uncover through the prism of structural variants. By optimising whole-genome re-sequencing to include many individuals across a species range as done here, future work will have the possibility to better understand how the interplay between structural variation and recombination may matter for the evolution of diversity and facilitate adaptation to a broad range of environmental heterogeneity.

Methods

A reference genome assembly for *Coelopa frigida*

To generate a reference genome, we sequenced sibling *Coelopa frigida* females from a 3-generation inbred family homozygous for the α arrangement at the inversion *Cf-Inv(1R)*, obtained by crossing wild *C. frigida* collected in St Irénée (QC, Canada). A pool of DNA from three female siblings was sequenced on 4 cells of Pacific Biosystems Sequel sequencer at McGill University and produced a total of 16.1 Gbp (~64x coverage) of long-read sequencing data. One additional female from the same inbred family was sequenced following the 10xGenomics Chromium on one lane of an Illumina HiSeqXTen sequencer at McGill University, yielding 82 Gbp (~300x of coverage) of paired-end linked-reads of 150bp.

An initial assembly was carried out on the PacBio long reads using the Smrt Analysis v3.0 pbsmrtpipe analysis workflow and FALCON (Chin et al 2013), resulting in 2959 contigs (N50 = 320 kb), for a total assembly size of 233.7 Mbp. This assembly was subsequently polished by using the linked reads from the 10Xgenomics sequencing, first by correcting for sequence errors with Pilon (Walker et al. 2014) and, second, by correcting for misassemblies with Tigmint (Jackman et al. 2018). The resulting assembly accounted 3096 contigs with a N50 of 320 kb.

To scaffold the genome assembly, we used the program ARKS and LINKS (Coombe et al. 2018; Yeo et al. 2018) which relies on linked-reads to scaffold contigs. The resulting assembly accounted 2539 scaffolds with a N50 of 735 kb. Then, scaffolds were assembled into chromosomes using *Chromonomer* (Catchen et al. 2020), which anchors and orientates scaffolds based on the order of markers in a linkage map (see below). The final assembly accounted 6 chromosomes and 1832 unanchored scaffolds with a N50 of 37.7Mb for a total of 239.7Mb (195.4 Mb into chromosomes). The completeness of this reference was assessed with BUSCO version 3.0.1 (Simão et al. 2015).

The genome was annotated by mapping a transcriptome assembled from RNA sequences obtained from 8 adults (split by sex and by karyotype at the *Cf(Inv(1)* inversions), 4 pools of 3 larvae and pools of *C. frigida* at different stages (eggs and larvae) and annotated using the Triannotate pipeline. More details about the genome and transcriptome assemblies are provided as supplementary methods.

A high-density linkage map and QTL analyses

- **Sequencing and genotyping**

We generate an outbred F2 family of 136 progenies by crossing two F1 individuals of *Coelopa frigida* from different crosses obtained from wild individuals collected in Gaspésie (QC, Canada). The mother of the F2 family was genotyped homozygous for the α arrangement at the inversion *Cf-Inv(1).* The progeny, both parents, and two paternal grandparents were sequenced using a double-digest restriction library preparation (ddRAD-seq) on IonProton (ThermoFisher). Parents and grandparents were sequenced at greater depth than progeny to make an accurate catalogue of diploid genotypes possible in the cross. Reads for the 136 offspring and their parents were trimmed using cutadapt, split per sample using process_radtags and aligned on the scaffolded assembly with bwa-mem. Genotype likelihoods were obtained with SAMtools mpileup. Only markers with at least 3X of coverage in all individuals were kept. We explored more stringent filtering such as 6X and 10X, which led to very similar and colinear maps albeit with less marker density, an aspect which was the priority for efficient scaffolding. More details are provided in supplementary methods

- **Building the map**

Linkage map was built using *Lep-MAP3* (Rastas 2017) following a pipeline available at https://github.com/clairemerot/lepmap3_pipeline. With the *Filtering* module, markers with more than 30% of missing data, non-informative markers, and extreme segregation distortion (χ^2 test, $P < 0.001$) were excluded. Markers were assigned to linkage groups (LGs) using the *SeparateChromosomes* module with a logarithm of odds (LOD) of 15, a minimum size of 5 and assuming no recombination in males. This assigned 28,615 markers into 5 large LGs, as expected given previous karyotyping work on *Coelopa frigida* (Aziz 1975), and 25 sex-linked markers into 2 small LGs than were subsequently merged as one. Within each LG, markers were ordered with 5 iterations of the *OrderMarker* module. The marker order from the run with the best likelihood was retained and refined 3 times with the *evaluateOrder* flag with 5 iterations each. When more than 1000 markers were plateauing at the same position, usually at the beginning, the end or the middle of a LG, all markers at that position were removed. Exploration for more stringent filtering for missing data, different values of LOD or allowing recombination in males resulted in very consistent and collinear maps.

- **Estimating recombination rate**

To estimate recombination rate across the genome, we compared position of the markers along the genetic map with their position along the genome assembly with MAREYMAP (Rezvoy et al. 2007). The Loess method was used to estimate Local recombination rates were

estimated with a Loess method including 10% of the markers for fitting the local polynomial curve.

- **QTL analysis**

We used the linkage map for QTL (quantitative trait locus) analysis. All individuals used to build the map were scored for recovery from chill coma induced by putting them 10 minutes at -20°C and by reporting behavior when transferred at room temperature. We distinguished three categories: "0", the fly stands immediately or in less than 5 minutes; "1", the fly recovers with difficulties after 5 to 15 minutes; "2", the fly has not recovered after more than 15 minutes. A phased map was obtained by performing an additional iteration of the *OrderMarker* module and the option "outputPhasedData=1". QTL analysis was carried out using R/qtl (Broman et al. 2003). LOD scores correspond to the $-\log_{10}$ of the associated probabilities between genotype and phenotype with the Haley-knot method. The LOD threshold for significance was calculated using 1000 permutations.

Population-level re-sequencing

- **Sampling and characterisation of sex, size and karyotype**

We analysed 1446 *C. frigida,* sampled at 16 locations spanning over 10° of latitude (Fig.1A) in September/October 2016. Sampling, genotyping and phenotyping are described in details in (Mérot et al. 2018). Briefly, adult flies were examined under a binocular magnifier (Zeiss Stemi 2000C) to confirm species identification and sex. For 1426 flies with wings in good conditions, the size was estimated using wing length as a proxy. Genomic DNA was extracted from adult flies using a salt-extraction protocol (Aljanabi and Martinez 1997) with a RNase A treatment (Quiagen). 1438 flies were successfully genotyped at the inversion *Cf-Inv(1)* as homokaryotypes for each of arrangement or heterokaryotypes (αα, αβ, ββ) using a molecular marker developed in (Mérot et al. 2018).

- **Library preparation and sequencing**

DNA quality of each extract was evaluated with nanodrop and on a 1% agarose gel electrophoresis. Only samples with acceptable ratios that showed clear high molecular weight bands were retained for library preparation. Following (Therkildsen and Palumbi 2017), we remove DNA fragments shorter than 1kb by treating each extract with Axygen magnetic beads in a 0.4:1 ratio, and eluted the DNA in 10mM Tris-Cl, pH 8.5. We measured DNA concentrations with QuantiT Picogreen dsDNA Assay Kit (Invitrogen) and normalised all samples at a concentration of 5ng/µL. Then, sample DNA extracts were randomized, distributed in 17 plates (96-well) and re-normalised at 1ng/µL.

Whole-genome high-quality libraries were prepared for each fly sample according to the protocol described in (Baym et al. 2015; Therkildsen and Palumbi 2017). Briefly, a tagmentation reaction using enzyme from the Nextera kit, which simultaneously fragments the DNA and incorporates partial adapters, was carried out in a 2.5 µl volume with approximately 1 ng of input DNA. Then, we used a two-step PCR procedure with a total of 12 cycles (8+4) to add the remaining Illumina adapter sequence with dual index barcodes and

amplify the libraries. The PCR was conducted with the KAPA Library Amplification Kit and custom primers derived from Nextera XT set of barcodes A,B,C and D (total 384 combinations). Amplification products were purified from primers and size-selected with a two-steps Axygen magnetic beads cleaning protocol, first with a ratio 0.35:1, keeping the supernatant (medium and short DNA fragments), second with a ratio 0.7:1, keeping the beads (medium fragments). Final concentration of the libraries were quantified with QuantiT Picogreen dsDNA Assay Kit (Invitrogen) and fragment size distribution was estimated with an Agilent BioAnalyzer for a subset of 10 to 20 samples per plate.

Equimolar amount of 293 to 296 libraries were combined into 5 separate pools for sequencing on 5 lanes of paired-end 150bp reads on an Illumina HiSeq 4000 at the Norwegian Sequencing Center at the University of Oslo.

- **Sequence filtering and processing**

Raw reads were trimmed and filtered for quality with FastP (Chen et al. 2018). Reads were aligned to the reference genome with BWA-MEM (Li and Durbin 2009) and filtered with samtools v1.8 (Li et al. 2009) to keep only unpaired, orphaned, and concordantly paired reads with a mapping quality over 10. Duplicate reads were removed with the MarkDuplicates module of Picard Tools v1.119. Then, we realigned reads around indels with the GATK IndelRealigner (McKenna et al. 2010). Finally, to avoid double-counting the sequencing support during SNP calling, we used the clipOverlap program in the bamUtil package v1.0.14 (Breese and Liu 2013) to soft clip overlapping read ends and we kept only the read with the highest quality score in overlapping regions. This pipeline was inspired by (Therkildsen and Palumbi 2017) and is available at https://github.com/enormandeau/wgs_sample_preparation.

For most of the analysis, we used the program ANGSD v0.931 (Korneliussen et al. 2014), a software specifically designed to take genotype uncertainty into account instead of basing the analysis on called genotypes, which was appropriated for the low coverage of our data. The pipeline of analysis is available at https://github.com/clairemerot/angsd_pipeline. For all analysis, input reads were filtered to remove reads with a samtools flag above 255 (not primary, failure and duplicate reads, tag -remove_bads = 1), with mapping quality below 30 (-minMapQ = 30) and to remove bases with quality below 20 (-minQ 20). Note that to reduce the computational and analytic burden due to small scaffolds, all analyses were performed on a reduced genome including the 6 chromosomes and only 135 unanchored scaffolds, selected because they were longer than 25kb and bear more than 100 SNPs/scaffold. This reduced genome represents more than 89% of the total reference and more than 98.5% of all SNPs.

As a first step, we ran ANGSD to estimate the spectrum of allele frequency, minor allele frequency, depth and genotype likelihoods on the whole dataset (-doSaf -doMaf -doDepth -doCounts - doGlf). Genotype likelihoods were estimated with the GATK method (-GL 2). The major allele was based on the genotype likelihood and was the most frequent allele (-doMajorMinor 1). We filtered to keep only SNPs covered by at least one read in at least 50%

of the individuals, with a total coverage below 4338 (3 times the number of individuals) to avoid including repeated regions in the analysis, and with minor allele frequency above 5%. The list of SNPs passing those filters and their respective major and minor alleles was subsequently used as the SNPs list for most analysis (-sites). Using PLINK 1.9, we produced a subset of SNP pruned for high physical linkage using a sliding-window approach where SNPs with a variance inflation factor greater than two (VIF > 2) were removed from 100 SNP windows shifted by 5 SNPs after each iteration.

- **Genetic structure, PCA and inversion detection**

Genetic structure in the whole dataset was analysed with NGSadmix (Skotte et al. 2013), which uses genotype likelihoods in beagle format from low coverage data to infer putative clustering and admixture. We explored a set of clusters from K=2 to K=10. Genetic variation was next analysed by extracting an individual covariance matrix with PCAngsd (Meisner and Albrechtsen 2018) and decomposed into principal component analysis (PCA) with R, using a scaling 2 transformation, which add an eigenvalues correction, to obtain the individuals PC scores (Legendre and Legendre 1998).

To analyse the genetic structure along the genome, we next run PCAangsd (Meisner and Albrechtsen 2018) on genotype likelihoods in non-overlapping windows of 100 SNPs to extract local covariance matrices, and obtained local PCAs of genetic variation. For each local PCA, we analysed the correlation between individuals PC1 scores and PC scores in the PCA performed on all the genome. This allowed to locate two (inversions) regions underlying the structure observed on PC1 and PC2 (Fig2A). We set the boundaries of those regions as windows with a coefficient of correlation above one standard deviation. We also recorded how much variance is explained by the first and second component of the local PCAs.

To scan the genome for other putative inversions or regions structuring the dataset into groups of haplotypes, we used the R package *Lostruct* (Li and Ralph 2019). The previously-performed local PCAs (including PC1 and PC2) were analysed together to measure the similarity of patterns between windows using Euclidian distances. Similarity was then mapped using multidimensional scaling (MDS) up to 50 axes. Inspired by the systematic procedure developed by (Huang et al. 2020), clusters of outlier similar windows were defined along each MDS axis as those with either values greater than 4 standard deviations above the mean or values lower than 4 standard deviations below the mean. Adjacent clusters with less than 20 windows between them were pooled, and clusters with less than 5 windows were not considered. Different window sizes (100 to 1000), different subset of PCs (including 1 to 3 PCs) and different thresholds for defining clusters yielded consistent results.

While inversions are frequent structure underlying MDS outliers detected by a local PCA analysis, similar patterns can be generated by any process locally limiting random mixing of alleles, such as strong geographic/species structure, linked selection (REF) or any feature limiting recombination like centromeres, TE accumulation or more complex structural variants. A typical signature of a simple polymorphic inversion is to appear on a PCA as three

groups of individuals, the two homokaryotypes for the alternative arrangement and as an intermediate group, the heterokaryotypes. Therefore, all regions of the genome including clusters of outlier windows and each cluster detected were further examined either by a PCA as single blocks, or divided into several blocks when discontinuous. We then used K-means clustering on the first and second PC to identify putative groups of haplotypes. Clustering accuracy was maximised by rotation and the discreteness was evaluated by the proportion of the between-cluster sum of squares over the total.

- **Inversion analysis**

For the four putative inversions (*Cf-Inv(1), Cf-Inv(4.1)* and the two related *Cf-Inv(4.2/4.3)*), K-means assignment on PC scores was used as the karyotype of the sample. Differentiation among karyotypes was measured with Fst statistics, using ANGSD to estimate joint allele-frequency spectrum, realSFS functions to compute FST in sliding-windows of 25KB with a step of 5KB, and subsampling largest groups to balance sample size. Observed heterozygosity was calculated for each karyotype and each SNP using the function -doHWE in ANGSD, and then averaged across sliding-windows of 25KB with a step of 5KB, using the R package *windowscanr*.

- **Linkage disequilibrium**

Intrachromosomal linkage disequilibrium was calculated among a reduced number of SNPs, filtered with more stringent criteria (MAF > 10%, at least one read in 75% of the samples, less than 3 times the expected coverage). Pairwise R^2 values were calculated with NGS-LD (Fox et al. 2019) based on genotype likelihood in beagle format obtained by ANGSD, and grouped into windows of 1MB. Plots display the 2nd percentile of R^2 values per paired of windows. For the chromosome including inversions (LG1, LG4), R^2 was calculated first, within all samples, and second, within individuals homozygous for the most common orientation, controlling for sample size by subsampling the largest group, and plotted by windows of 250kb.

- **Geographic structure**

After considering the whole dataset or the groups of haplotypes, we grouped individuals by geographic sampling site, hereafter called populations. Allele frequency spectrum and minor allele frequency was calculated for each population using the -doMaf function and the previously obtained list of polymorphic SNPs and their polarisation as major or minor allele (options –sites and –doMajorMinor 3). Positions which were not covered by at least one read in 50% of the samples in a given population were filtered out. Pairwise F_{ST} differentiation was estimated using the realSFS function in ANGDS between all pairs of populations, subsampled to a similar size of 88 individuals. The weighted F_{ST} between pairs of population were computed by including either all SNPs, LD-pruned SNPs, or SNPs from a region of interest (inversions/low-recombining regions) or SNPs outside those regions (collinear SNPs).

Isolation-By-Distance (IBD) was tested for each subset of SNPs using a linear model in which pairwise genetic distance (F_{ST}/(1-F_{ST})) was included as the response variable and geographic

Euclidian distance was incorporated as an explanatory term. Isolation-by-Resistance was tested in the same way, except that physical distances were calculated along the shoreline, by inferring the least-cost path through areas of the map between -40 meters of depth and 20 meters of altitude using the R package *marmap*. Both models of IBD and IBR were compared to a null model using an ANOVA F-test, and to each other using adjusted $R2$ and AIC. To compare IBD and IBR patterns in each inversion/low-recombining region to the collinear genome, we built a full model explaining pairwise genetic distances by physical distances and genomic region (collinear vs. inversion) as a co-factor, and assessed the significance of the interaction term as well as the direction of the interaction slope coefficient. Since the collinear genome include more, and more dispersed, SNPs, we repeated this analysis 100 times with randomly-chosen collinear regions including the same number of contiguous SNPs as each inversion/low-recombining regions. This provided a distribution of the significance of the interaction term and its slope coefficient (Fig. S9). For the inversion *Cf-Inv(1),* no contiguous block with the same number of SNPs could be found in the genome, hence we gathered 3 blocks of 1/3 the number of SNPs in each of the 100 random replicates.

- **Environmental associations**

Environment at each location was described by three categories of variables: large-scale climatic/abiotic conditions, local wrackbed abiotic characteristics, and local wrackbed algal composition (Table S1). Large-scale climatic/abiotic conditions were extracted for each location from public databases and included annual means in precipitations, air temperature, sea surface temperature, sea surface salinity and tidal amplitude. Wrackbed abiotic characteristics included an estimation of the surface and a measure of depth, internal temperature and salinity. Wrackbed composition was an estimation of the relative proportions of Laminariaceae, Fucaceae, Zoosteraceae, plant debris and other seaweed species. Details about how environmental variables were measured or drawn from databases can be found in (Mérot et al. 2018). Correlation between environmental variables was tested with a Pearson correlation test, and variation was reduced by drawing a summary variable for each group of correlated environmental variables (climatic, salinity/tidal amplitude, abiotic characteristic of the wrackbed, algal composition) by retaining the first significant PC of a principal component analysis (PCA) on original variables relying on the Kaiser-Guttman and Broken Stick criteria (Borcard et al. 2011) (see Fig. S11).

After filtering for SNPs covered by at least 50% of the individuals in each population (so about a coverage of 50X), the matrix of allelic frequencies obtained for the 16 populations accounted 1,155,978 SNPs. A genetic-environment association (GEA) which evaluate SNPs frequencies as function of environmental variables was performed through a combination of two methods as recommended by (Villemereuil et al. 2014): (i) latent factor mixed models (LFMM2; (Frichot et al. 2013; Caye et al. 2019)), (ii) Bayes factor (BAYPASS) (Gautier 2015). Those three methods had also been shown to be robust to the presence of large inversions (Lotterhos 2019).

LFMM was run with the R package *lfmm2* (Caye et al. 2019), using a ridge regression which performed better in simulations including inversions (Lotterhos 2019), and parametrized using a K-value of 4 latent-factors based on the number of principal components that explain variation in population frequencies. P-values were calibrated following the recommendations of (François et al. 2016), using a Benjamini-Hochberg correction with a false-discovery-rate (fdr) of 0.05.

Using Baypass v2.2 (Gautier 2015), a Bayes factor (BP), which evaluate for each SNP the strength of an association with an environmental variable, was computed as the median of three run under the standard model using the default importance sampling estimator approach. Environmental variables were scaled using the -scalecov option. We run this analysis twice, first, without controlling for population structure and, second, by controlling with a covariance matrix extracted from an initial BayPass model run on the subset of LD-pruned SNPs without environmental covariables. To calculate a significance threshold for the BP factor, we simulated pseudo-observed data with 10,000 SNPs using the "simulate.baypass" function, analysed the pseudo-observed matrix of frequencies for each environmental variable as described above, and kept the 0.1% quantile as the significance threshold.

For each GEA method, and the combination of the two, the repartition of candidate SNPs for association with environment within and outside inversions/low-recombining regions was compared to the original repartition of SNPs. Deviation from this original repartition was tested with a Fisher's exact test, and the magnitude of the excess/deficit of outlier SNPs within each region of the genome was reported as the odd-ratio.

- **Phenotypic associations and gene ontology analysis**

We performed a genome-wide association study (GWAS) to detect SNPs associated with wing size variation. We used the GWAS implemented in ANGSD program that accounts for the genotype uncertainty in low depth NGS data and uses the genotypes likelihood in Beagle format (Jørsboe and Albrechtsen 2020). We used the latent genotype model (EM algorithm, -doAssso=4) where genotype is introduced as a latent variable and then the likelihood is maximized using weighted least squares regression. We considered a false discovery rate (FDR) of 0.001. The GWAS was applied on the whole dataset (1,426 flies with size information) and then on each subset of homaryotes at the inversion *Cf-Inv(1)* (140 αα and 436 ββ flies with size information).

Using BEDtools, we extracted the list of genes overlapping with significantly-associated SNPs, or within a window of 5kb upstream or downstream a gene. We then tested for the presence of over-represented GO terms using GOAtools (v0.6.1, pval = 0.05) and filtered the outputs of GOAtools to keep only GO terms for biological processes of levels 3 or more, and with an FDR value equal below 0.1. We performed the same GO enrichment analysis for the list of genes found in the two largest inversions (*Cf-Inv(1)* and *Cf-Inv(4.1)*).

Acknowledgments

We are very grateful to M. Lionard who sampled in Blanc-Sablon and to L. Johnson, E. Tamigneaux, D. Malloch for their advice during fieldwork. We thank C. Babin and P. Berube for their support in the lab. B. Boyle and N. Therkildsen provided key advice about sequencing and libraries preparation. We thank Y. Dorant, M. Leitwein, and C. Rougeux for their help and advice with analyses.

The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities, by McGill Sequencing Platform and by the genomic platform at IBIS (University Laval http://www.ibis.ulaval.ca/).

This research was supported by a Discovery research grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to L.B., by the Canadian Research Chair in genomics and conservation of aquatic resources held by L.B. and by the Swedish Research Council grant 2012-3996 to M.W. The genome assembly was supported by the Canada 150 Sequencing Initiative (CanSeq150). C.M. was supported by a postdoctoral fellowship from the Fonds de Recherche Québec (FRQNT FRQS) and a Banting Postdoctoral Fellowship from the NSERC.

References

Alex Buerkle C, Gompert Z. 2013. Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology* 22:3028–3035.

Aljanabi SM, Martinez I. 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic acids research* 25:4692–4693.

Aziz JB. 1975. Investigations into chromosomes 1, 2 and 3 of Coelopa frigida (Fab.). *Ph.D.Thesis, University of Newcastle upon Tyne*.

Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. 2015. Inexpensive multiplexed library preparation for megabase-sized genomes. *PloS one* 10:e0128036.

Berdan E, Rosenquist H, Larson K, Wellenreuther M. 2018. Inversion frequencies and phenotypic effects are modulated by the environment: insights from a reciprocal transplant study in Coelopa frigida. *Evolutionary ecology* 32:683–698.

Berg PR, Star B, Pampoulie C, Bradbury IR, Bentzen P, Hutchings J, Jentoft S, Jakobsen KS. 2017. Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity* 119:418–428.

Borcard D, Gillet F, Legendre P. 2011. Numerical ecology with R. Springer Science & Business Media

Breese MR, Liu Y. 2013. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics* 29:494–496.

Broman KW, Wu H, Sen Ś, Churchill GA. 2003. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19:889–890.

Butlin R, Collins P, Skevington S, Day T. 1982. Genetic variation at the alcohol dehydrogenase locus in natural populations of the seaweed fly, Coelopa frigida. *Heredity* 48:45–55.

Butlin R, Day T. 1985. Adult size, longevity and fecundity in the seaweed fly, Coelopa frigida. *Heredity* 54:107–110.

Butlin R, Read I, Day T. 1982. The effects of a chromosomal inversion on adult size and male mating success in the seaweed fly, Coelopa frigida. *Heredity* 49:51–62.

Catchen J, Amores A, Bassham S. 2020. Chromonomer: a tool set for repairing and enhancing assembled genomes through integration of genetic maps and conserved synteny. *bioRxiv*.

Caye K, Jumentier B, Lepeule J, François O. 2019. LFMM 2: Fast and accurate inference of gene-environment associations in genome-wide studies. *Molecular biology and evolution* 36:852–860.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890.

Cheng C, White BJ, Kamdem C, Mockaitis K, Costantini C, Hahn MW, Besansky NJ. 2012. Ecological genomics of Anopheles gambiae along a latitudinal cline: a population-resequencing approach. *Genetics* 190:1417–1432.

Christmas MJ, Wallberg A, Bunikis I, Olsson A, Wallerman O, Webster MT. 2018. Chromosomal inversions associated with environmental adaptation in honeybees. *Molecular Ecology* [Internet] 0. Available from: https://doi.org/10.1111/mec.14944

Clucas GV, Lou RN, Therkildsen NO, Kovach AI. 2019. Novel signals of adaptive genetic variation in northwestern Atlantic cod revealed by whole-genome sequencing. *Evolutionary applications* 12:1971–1987.

Coombe L, Zhang J, Vandervalk BP, Chu J, Jackman SD, Birol I, Warren RL. 2018. ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC bioinformatics* 19:234.

Day T, Dawe C, Dobson T, Hillier P. 1983. A chromosomal inversion polymorphism in Scandinavian populations of the seaweed fly, Coelopa frigida. *Hereditas* 99:135–145.

Dobson T. 1974. Studies on the biology of the kelp-fly Coelopa in Great Britain. *Journal of Natural History* 8:155–177.

Edward DA, Gilburn AS. 2013. Male-specific genotype by environment interactions influence viability selection acting on a sexually selected inversion system in the seaweed fly, Coelopa frigida. *Evolution* 67:295–302.

Egglishaw HJ. 1960. Studies on the family Coelopidae (Diptera). *Ecological Entomology* 112:109–140.

Fang Z, Pyhäjärvi T, Weber AL, Dawe RK, Glaubitz JC, González J de JS, Ross-Ibarra C, Doebley J, Morrell PL, Ross-Ibarra J. 2012. Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics* 191:883–894.

Faria R, Chaube P, Morales HE, Larsson T, Lemmon AR, Lemmon EM, Rafajlović M, Panova M, Ravinet M, Johannesson K. 2019. Multiple chromosomal rearrangements in a hybrid zone between Littorina saxatilis ecotypes. *Molecular ecology* 28:1375–1393.

Faria R, Johannesson K, Butlin RK, Westram AM. 2019. Evolving Inversions. *Trends in ecology & evolution*.

Fox EA, Wright AE, Fumagalli M, Vieira FG. 2019. ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics* 35:3855–3856.

François O, Martins H, Caye K, Schoville SD. 2016. Controlling false discoveries in genome scans for selection. *Molecular ecology* 25:454–469.

Frichot E, Schoville SD, Bouchard G, François O. 2013. Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular biology and evolution* 30:1687–1699.

Fuentes-Pardo AP, Ruzzante DE. 2017. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular ecology* 26:5369–5406.

Gandon S, Rousset F. 1999. Evolution of stepping-stone dispersal rates. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 266:2507–2513.

Gautier M. 2015. Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* 201:1555–1579.

Hämälä T, Savolainen O. 2019. Genomic patterns of local adaptation under gene flow in Arabidopsis lyrata. *Molecular Biology and Evolution* 36:2557–2571.

Ho SS, Urban AE, Mills RE. 2019. Structural variation in the sequencing era. *Nature Reviews Genetics*:1–19.

Hoffmann AA, Rieseberg LH. 2008. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annual review of ecology, evolution, and systematics* 39:21–42.

Hoffmann AA, Sgrò CM, Weeks AR. 2004. Chromosomal inversion polymorphisms and adaptation. *Trends in Ecology & Evolution* 19:482–488.

Huang K, Andrew RL, Owens GL, Ostevik KL, Rieseberg LH. 2020. Multiple chromosomal inversions contribute to adaptive divergence of a dune sunflower ecotype. *Molecular Ecology* XX:XX–XX.

Huang K, Rieseberg LH. 2020. Frequency, Origins, and Evolutionary Role of Chromosomal Inversions in Plants. *Frontiers in Plant Science* 11:296.

Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, Yeo S, Xue Z, Mohamadi H, Bohlmann J, Jones SJ. 2018. Tigmint: correcting assembly errors using linked reads from large molecules. *BMC bioinformatics* 19:1–10.

Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* 477:203–206.

Jørsboe E, Albrechtsen A. 2020. Efficient approaches for large scale GWAS studies with genotype uncertainty. *bioRxiv*:786384.

Kapun M, Barrón MG, Staubach F, Obbard DJ, Wiberg RAW, Vieira J, Goubert C, Rota-Stabelli O, Kankare M, Bogaerts-Márquez M, et al. 2020. Genomic Analysis of European Drosophila melanogaster Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses. *Molecular Biology and Evolution* [Internet]. Available from: https://doi.org/10.1093/molbev/msaa120

Kapun M, Fabian DK, Goudet J, Flatt T. 2016. Genomic Evidence for Adaptive Inversion Clines in Drosophila melanogaster. *Molecular Biology and Evolution* 33:1317–1336.

Kapun M, Flatt T. 2019. The adaptive significance of chromosomal inversion polymorphisms in Drosophila melanogaster. *Molecular ecology* 28:1263–1282.

Kirkpatrick M. 2010. How and why chromosome inversions evolve. *PLoS biology* 8:e1000501.

Kirkpatrick M, Barton N. 2006. Chromosome inversions, local adaptation and speciation. *Genetics* 173:419–434.

Kirubakaran TG, Grove H, Kent MP, Sandve SR, Baranski M, Nome T, De Rosa MC, Righino B, Johansen T, Otterå H, et al. 2016. Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Mol Ecol* 25:2130–2143.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: analysis of next generation sequencing data. *BMC bioinformatics* 15:356.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.

Li H, Ralph P. 2019. Local PCA shows how the effect of population structure differs along the genome. *Genetics* 211:289–304.

Lindtke D, Lucek K, Soria-Carrasco V, Villoutreix R, Farkas TE, Riesch R, Dennis SR, Gompert Z, Nosil P. 2017. Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect. *Molecular Ecology* 26:6189–6205.

Lotterhos KE. 2019. The effect of neutral recombination variation on genome scans for selection. *G3: Genes, Genomes, Genetics* 9:1851–1867.

Lotterhos KE, Yeaman S, Degner J, Aitken S, Hodgins KA. 2018. Modularity of genes involved in local adaptation to climate despite physical linkage. *Genome Biology* 19:157.

Machado HE, Bergland AO, O'Brien KR, Behrman EL, Schmidt PS, Petrov DA. 2016. Comparative population genomics of latitudinal variation in Drosophila simulans and Drosophila melanogaster. *Molecular ecology* 25:723–740.

Martin AR, Atkinson EG, Chapman SB, Stevenson A, Stroud RE, Abebe T, Akena D, Alemayehu M, Ashaba FK, Atwoli L, et al. 2020. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *bioRxiv*:2020.04.27.064832.

Martin SH, Davey JW, Salazar C, Jiggins CD. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biology* 17:e2006288.

McDonald TK, Yeaman S. 2018. Effect of migration and environmental heterogeneity on the maintenance of quantitative genetic variation: a simulation study. *Journal of evolutionary biology* 31:1386–1399.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20:1297–1303.

Meisner J, Albrechtsen A. 2018. Inferring population structure and admixture proportions in low-depth NGS data. *Genetics* 210:719–731.

Mérot C, Berdan EL, Babin C, Normandeau E, Wellenreuther M, Bernatchez L. 2018. Intercontinental karyotype–environment parallelism supports a role for a chromosomal inversion in local adaptation in a seaweed fly. *Proc Biol Sci* [Internet] 285. Available from: http://rspb.royalsocietypublishing.org/content/285/1881/20180519.abstract

Mérot C, Llaurens V, Normandeau E, Bernatchez L, Wellenreuther M. 2020. Balancing selection via life-history trade-offs maintains an inversion polymorphism in a seaweed fly. *Nature Communications* 11:1–11.

Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*.

Morales HE, Faria R, Johannesson K, Larsson T, Panova M, Westram AM, Butlin RK. 2019. Genomic architecture of parallel ecological divergence: beyond a single environmental contrast. *Science advances* 5:eaav9963.

Ortiz-Barrientos D, James M. 2017. Evolution of recombination rates and the genomic landscape of speciation. *Journal of evolutionary biology* 30:1519–1521.

Rastas P. 2017. Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics* 33:3726–3732.

Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. 2016. Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. *PLOS Biology* 14:e2000234.

Samuk K, Owens GL, Delmore KE, Miller SE, Rennison DJ, Schluter D. 2017. Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Molecular Ecology* 26:4378–4390.

Savolainen O, Lascoux M, Merilä J. 2013. Ecological genomics of local adaptation. *Nature Reviews Genetics* 14:807.

Schaeffer SW. 2018. Muller "Elements" in Drosophila: how the search for the genetic basis for speciation led to the birth of comparative genomics. *Genetics* 210:3–13.

Schwander T, Libbrecht R, Keller L. 2014. Supergenes and Complex Phenotypes. *Current Biology* 24:R288–R294.

Semenov GA, Safran RJ, Smith CC, Turbek SP, Mullen SP, Flaxman SM. 2019. Unifying Theoretical and Empirical Perspectives on Genomic Differentiation. *Trends in ecology & evolution*.

Simão 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.

Skotte L, Korneliussen TS, Albrechtsen A. 2013. Estimating individual admixture proportions from next generation sequencing data. *Genetics* 195:693–702.

Stevison LS, McGaugh SE. 2020. It's time to stop sweeping recombination rate under the genome scan rug.

Therkildsen NO, Palumbi SR. 2017. Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular ecology resources* 17:194–208.

Therkildsen NO, Wilder AP, Conover DO, Munch SB, Baumann H, Palumbi SR. 2019. Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing. *Science* 365:487–490.

Tigano A, Friesen VL. 2016. Genomics of local adaptation with gene flow. *Molecular ecology* 25:2144–2164.

Tigano A, Jacobs A, Wylder AP, Nand A, Zhan Y, Dekker J, Therkildsen NO. 2020. Chromosome-level assembly of the Atlantic silverside genome reveals extreme levels of sequence diversity and structural genetic variation. *bioRxiv*.

Todesco M, Owens GL, Bercovich N, Légaré J-S, Soudi S, Burge DO, Huang K, Ostevik KL, Drummond EB, Imerovski I. 2019. Massive haplotypes underlie ecotypic differentiation in sunflowers. *bioRxiv*:790279.

Todesco M, Owens GL, Bercovich N, Légaré J-S, Soudi S, Burge DO, Huang K, Ostevik KL, Drummond EB, Imerovski I. 2020. Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*:1–6.

Van Belleghem SM, Vangestel C, De Wolf K, De Corte Z, Möst M, Rastas P, De Meester L, Hendrickx F. 2018. Evolution at two time frames: Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLoS Genetics* 14:e1007796.

Vicoso B, Bachtrog D. 2015. Numerous transitions of sex chromosomes in Diptera. *PLoS biology* 13.

Villemereuil P, Frichot É, Bazin É, François O, Gaggiotti OE. 2014. Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology* 23:2006–2019.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* 9.

Wellenreuther M, Bernatchez L. 2018. Eco-evolutionary genomics of chromosomal inversions. *Trends in ecology & evolution*.

Wellenreuther M, Rosenquist H, Jaksons P, Larson KW. 2017. Local adaptation along an environmental cline in a species with an inversion polymorphism. *Journal of Evolutionary Biology* 30:1068–1077.

Wilder AP, Palumbi SR, Conover DO, Therkildsen NO. 2020. Footprints of local adaptation span hundreds of linked genes in the Atlantic silverside genome. *Evolution letters* 4:430–443.

Yan Z, Martin SH, Gotzek D, Arsenault SV, Duchen P, Helleu Q, Riba-Grognuz O, Hunt BG, Salamin N, Shoemaker D. 2020. Evolution of a supergene that regulates a trans-species social polymorphism. *Nature Ecology & Evolution* 4:240–249.

Yeaman S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences* 110:E1743–E1751.

Yeaman S. 2015. Local adaptation by alleles of small effect. *The American Naturalist* 186:S74–S89.

Yeo S, Coombe L, Warren RL, Chu J, Birol I. 2018. ARCS: scaffolding genome drafts with linked reads. *Bioinformatics* 34:725–731.