# Polygenic adaptation from standing genetic variation allows rapid ecotype formation

Nico Fuhrmann, Celine Prakash, Tobias S. Kaiser*

Max Planck Institute for Evolutionary Biology,

Max Planck Research Group *Biological Clocks*,

August-Thienemann-Strasse 2, 24306 Plön, Germany

*kaiser@evolbio.mpg.de

**Abstract**

Adaptive ecotype formation is the first step to speciation, but the genetic underpinnings of this process are poorly understood. While in marine midges of the genus *Clunio* (Diptera) reproduction generally follows a lunar rhythm, we here characterize two lunar-arrhythmic ecotypes. Analysis of 168 genomes reveals a recent establishment of these ecotypes, reflected in massive haplotype sharing between ecotypes, irrespective of whether there is ongoing gene flow or geographic isolation. Genetic analysis and genome screens reveal patterns of polygenic adaptation from standing genetic variation. Ecotype-associated loci prominently include circadian clock genes, as well as genes affecting sensory perception and nervous system development, hinting to a central role of these processes in lunar time-keeping. Our data show that adaptive ecotype formation can occur rapidly, with ongoing gene flow and largely based on a re-assortment of existing and potentially co-adapted alleles.

**Keywords**: Local adaptation, reproductive timing, lunar rhythm, biological clocks, sympatric speciation, gene flow, Chironomidae, marine ecology

## Introduction

Understanding the processes underlying local adaptation and ecotype formation is a vital theme in evolutionary ecology[1,2], but also increasingly important for conservation of biodiversity in the face of climate change and deterioration of natural habitats[3]. Adaptation in reproductive timing is at particular risk, as under rising temperatures it can be severely mismatched with the environment[4]. Major open questions for understanding the process of adaptation are to what extent it requires novel mutations or reuses existing genetic variation, and how these different paths of adaptation are constrained by population history and genome architecture[2,5]. Answering these questions generally requires identification of the adaptive genetic loci. For obtaining a broad understanding, this tedious endeavor must be undertaken in a diverse array of model and non-model organisms. Here we present a study on the recent evolution of ecotypes in marine midges of the genus *Clunio* (Diptera: Chironomidae), which in adaptation to their habitat differ in oviposition behavior and reproductive timing, involving both circadian and circalunar clocks.

Circalunar clocks are biological time-keeping mechanisms that allow organisms to anticipate lunar phase[6]. Their molecular basis is unknown[7], making identification of adaptive loci for lunar timing both particularly challenging and interesting. In many marine organisms, circalunar clocks synchronize reproduction within a population. In *Clunio marinus* they have additional ecological relevance[8]. Living in the intertidal zone of the European Atlantic coasts, *C. marinus* requires the habitat to be exposed by the tide for successful oviposition. The habitat is maximally exposed during the low waters of spring tide days around full moon and new moon. Adult emergence is restricted to these occasions by a circalunar clock, which tightly regulates development and maturation. Additionally, a circadian clock ensures emergence during only one of the two daily low tides. The adults reproduce immediately after emergence and die few hours later. As tidal regimes vary dramatically along the coastline, *C. marinus* populations have evolved various local timing adaptations[8-10]. Analysis of these timing adaptations gave first insights into the genetic underpinnings of circalunar clocks[11,12].

In addition to the above-described lunar-rhythmic *Atlantic ecotype* of *C. marinus*, literature reports two lunar-arrhythmic ecotypes of *Clunio* in the Baltic Sea[13-15] and in the high Arctic[16,17] (see Fig. 1 for a summary of defining characteristics of the three ecotypes). In the Baltic Sea the tides are negligible and the *Baltic ecotype* oviposits on the open water, from where the eggs quickly sink to the submerged larval habitat at water depths of up to 20 metres[13,18]. Reproduction of the *Baltic ecotype* happens every day precisely at dusk under control of a circadian clock[19]. There is no detectable circalunar rhythm[14]. Near Bergen (Norway) the *Baltic* and *Atlantic ecotypes* were reported to co-occur in sympatry, but in temporal reproductive isolation. The *Baltic ecotype* reproduces at dusk, the *Atlantic ecotype* reproduces during the afternoon low tide[19]. Therefore, the *Baltic ecotype* is considered a separate species – *C. balticus*. However, *C. balticus* and *C. marinus* can be successfully interbred in the laboratory[19].

In the high Arctic there are normal tides and the *Arctic ecotype* of *C. marinus* is found in intertidal habitats[16]. During its reproductive season, the permanent light of polar day precludes synchronization of the circadian and circalunar clocks with the environment. Thus, the *Arctic ecotype* relies on a so-called tidal hourglass timer, which allows it to emerge and reproduce during every low tide[17]. It does not show circalunar or circadian rhythms[17].

The geological history of Northern Europe[20] and subfossil *Clunio* head capsules in Baltic Sea sediment cores[21] suggest that the Baltic Sea and the high Arctic were colonized by *Clunio* after the last ice age, setting a time frame of less than 10,000 years for formation of the lunar-arrhythmic ecotypes. In this study, we confirmed and characterised these ecotypes in field work and laboratory experiments. Sequencing 168 individual genomes highlighted the evolutionary history of the three ecotypes, the processes underlying ecotype formation and major molecular pathways determining their ecotype characteristics.

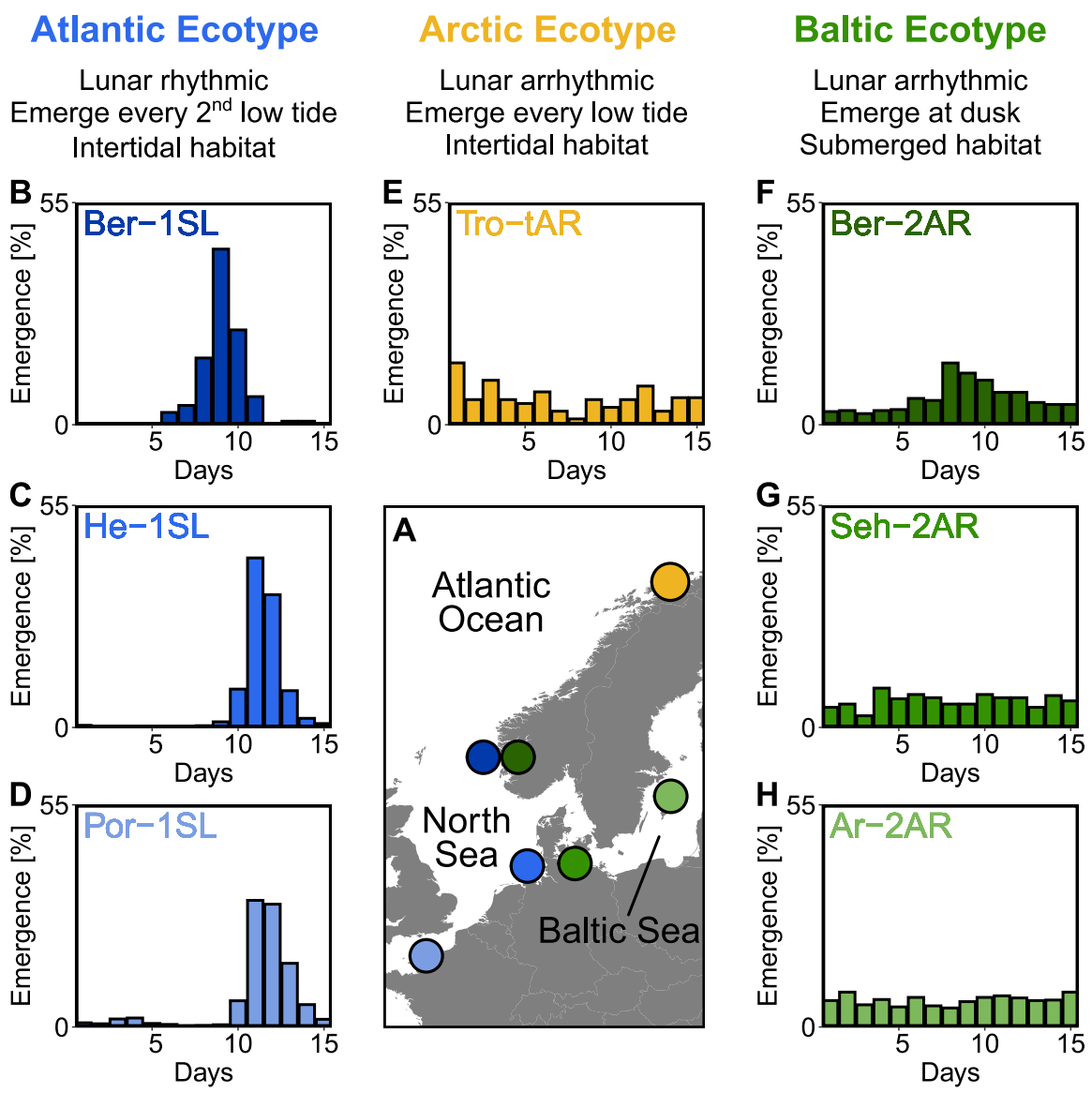



**Figure 1** **Northern European ecotypes of *Clunio* and their lunar rhythms**

The *Atlantic*, *Arctic* and *Baltic ecotypes* of *Clunio* differ mainly in their lunar rhythms (B-H), circadian rhythms (Suppl. Fig. 1), as well as their habitat and the resulting oviposition behavior (Suppl. Note 1). **(A)** Sampling sites for this study. **(B-H)** Lunar rhythms of adult emergence in corresponding laboratory strains under common garden conditions, with 16 hours of daylight and simulated tidal turbulence cycles to synchronize the lunar rhythm. In *Arctic* and *Baltic ecotypes* the lunar rhythm is absent (E,G,H) or very weak (F). Por-1SL: n=1,263; He-1SL: n=2,075; Ber-1SL: n=230; Tro-tAR: n=209; Ber-2AR: n=399; Seh-2AR: n= 380; Ar-2AR: n=765.

## Results

***Clunio* ecotypes and their lunar rhythms.** Starting from field work in Northern Europe (Fig. 1A), we established one laboratory strain of the *Arctic ecotype* from Tromsø (Norway, Tro-tAR; see methods for strain nomenclature) and three laboratory strains of the *Baltic ecotype*, from Bergen (Norway, Ber-2AR), Sehlendorf (Germany; Seh-2AR) and Ar (Sweden; Ar-2AR). We also established a strain of the *Atlantic ecotype* from Bergen (Ber-1SL, sympatric with Ber-2AR) and used two existing *Atlantic ecotype* laboratory strains from Helgoland (Germany; He-1SL) and Port-en-Bessin (France; Por-1SL). We confirmed the identity of the ecotypes in the laboratory by the absence of a lunar rhythm in the *Baltic* and *Arctic ecotypes* (Fig. 1B-H), their circadian rhythm (Supplementary Fig. 1B-H) and their oviposition behavior (for details see Supplementary Note 1). The *Baltic ecotype* from Bergen (Ber-2AR, Fig. 1F) was found weakly lunar-rhythmic. In crossing experiments between the Ber-2AR and Ber-1SL laboratory strains, the degree of lunar rhythmicity segregates within and between crossing families (Supplementary Fig. 2), suggesting a heterogeneous polygenic basis of lunar arrhythmicity. Genetic segregation implies that the weak rhythm in Ber-2AR is due to genetic polymorphism. The Ber-2AR strain seems to carry some lunar-rhythmic alleles, likely due to gene flow from the sympatric *Atlantic ecotype* (see results below).

**Evolutionary history and species status.** We sequenced the full nuclear and mitochondrial genomes of 168 field-caught individuals, 24 from each population (23 for Por-1SL, 25 for He-1SL). Based on a set of 792,032 single nucleotide polymorphisms (SNPs), we first investigated population structure and evolutionary history by performing a principal component analysis (PCA; Fig. 2A-B) and testing for genetic admixture (Fig. 2C). We also constructed a haplotype network of complete mitochondrial genomes (Fig. 2D). There are four major observations.

First, there is strong geographic isolation between populations from different sites. In PCA, clusters are formed according to geography (Fig. 2A-B, Supplementary Fig. 3). Mitochondrial haplotypes are not shared and are highly divergent between geographic sites (Fig. 2D). In ADMIXTURE, the optimal number of genetic groups is six (Supplementary Fig. 4), corresponding

to the number of geographic sites, and there is basically no mixing between the six clusters (Fig. 2C; K=6).

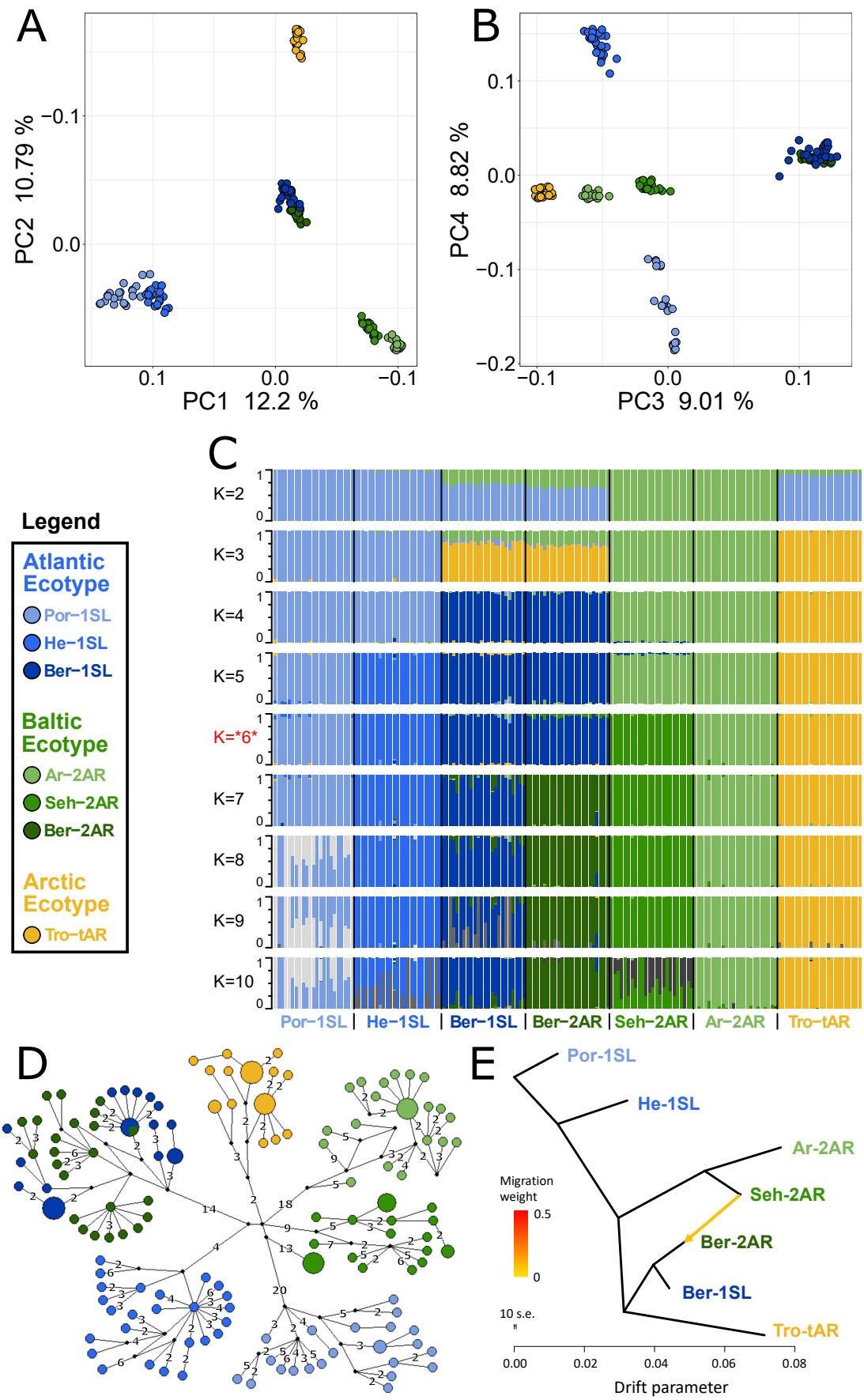
Second, and much in contrast to the above, the sympatric ecotypes in Bergen are genetically very similar. In PCA they are not separated in the first four principal components (Fig. 2A-B) and they are the only populations that share mitochondrial haplotypes (Fig. 2D). In the ADMIXTURE analysis, they are only distinguished at K=7, a value larger than the optimal K. As soon as the two populations are distinguished, some individuals show signals of admixed origin (Fig. 2C; K=7), indicative of ongoing gene flow and incomplete reproductive isolation. These observations question the species status of *C. balticus*, which was based on the assumption of temporal isolation between these two populations[19]. Given that genetic differentiation rather corresponds to geography than to ecotype (Fig. 2), we consider all three ecotypes part of a single species, *C. marinus*.

---

▶

**Figure 2 Genetic structure and evolutionary history of Northern European *Clunio* ecotypes**

**(A,B)** Principal component analysis (PCA) based on 792,032 SNPs separates populations by geographic location rather than ecotype. **(C)** ADMIXTURE analysis supports strong differentiation by geographic site (best K=6), but a notable genetic component from the Baltic Sea in the Bergen populations (see K=2 and 3). The Bergen populations are only separated at K=7 and then show a number of admixed individuals. **(D)** Haplotype network of full mitochondrial genomes reveals highly divergent clusters according to geographic site, but haplotype sharing between Ber-1SL and Ber-2AR. **(E)** Correlated allele frequencies indicate introgression from Seh-2AR into Ber-2AR.


A
PC2 10.79 %
PC1 12.2 %
B
PC4 8.82 %
PC3 9.01 %
C
Legend
Atlantic Ecotype
Por−1SL
He−1SL
Ber−1SL
Baltic Ecotype
Ar−2AR
Seh−2AR
Ber−2AR
Arctic Ecotype
Tro−tAR
K=2
K=3
K=4
K=5
K=*6*
K=7
K=8
K=9
K=10
Por−1SL | He−1SL | Ber−1SL | Ber−2AR | Seh−2AR | Ar−2AR | Tro−tAR
D
E
Por-1SL
He-1SL
Ar-2AR
Seh-2AR
Ber-2AR
Ber-1SL
Tro-tAR
Migration weight
0.5
0
10 s.e.
0.00
0.02
0.04
0.06
0.08
Drift parameter

Third, the data suggest that after the ice age *Clunio* colonized northern Europe from a single source and expanded along two fronts into the Baltic Sea and into the high Arctic. The mitochondrial haplotype network expands from a single center, which implies a quick radiation from a single ancestral haplotype (Fig. 2D). In line with this, 34% of polymorphic sites are polymorphic in all seven populations and 93% are polymorphic in at least two populations (Supplementary Fig. 5). In the light of the detected strong geographic isolation, this reflects a large amount of shared ancestral polymorphism. Separation of the Baltic Sea populations along PC1 and the Arctic population along PC2 (Fig. 2A), suggests that *Clunio* expanded into the high Arctic and into the Baltic Sea independently. Congruently, nucleotide diversity significantly decreases towards both expansion fronts (Supplementary Fig. 6, Supplementary Tab. 1). Postglacial establishment from a common source indicates that the lunar-arrhythmic *Baltic* and *Arctic ecotypes* must be derived from the lunar-rhythmic *Atlantic ecotype*.

Fourth, ADMIXTURE analysis reveals that sympatric co-existence of the *Atlantic* and *Baltic ecotypes* in Bergen likely results from introgression of *Baltic ecotype* individuals into an existing *Atlantic ecotype* population. At K=2 and K=3 the two Baltic Sea populations Seh-2AR and Ar-2AR are separated from all other populations and the two Bergen populations Ber-2AR and Ber-1SL show a marked genetic component coming from these Baltic Sea populations (Fig. 2C). Congruently, TreeMix detects introgression from Seh-2AR into Ber-2AR (Fig. 2E), but no other introgression events (Supplementary Fig. 7). The genetic component from the Baltic is largely shared between the two Bergen populations, underscoring again that the *Baltic* and *Atlantic ecotypes* in Bergen are not fully reproductively isolated. However, the Baltic genetic component is slightly larger for the *Baltic ecotype* Ber-2AR population than for the *Atlantic ecotype* Ber-1SL population. The small fraction of introgressed alleles by which the Bergen populations differ might determine *Baltic ecotype* characteristics. Interestingly, the *Arctic ecotype* also shares a small genetic component with the Baltic populations (Fig. 2C, K=2), leaving open whether it evolved lunar arrhythmicity independently from the *Baltic ecotype* or whether arrhythmicity alleles from the Baltic were carried all the way north.

**Incomplete lineage sorting and introgression.** All subsequent analyses of the evolutionary processes and genomic loci underlying ecological adaptation were focussed on the *Atlantic* and *Baltic ecotypes*, represented by three populations each. First, we reconstructed the genealogical relationship between 36 individuals (six from each population) in 50 kb windows (n=1,607) along the genome, followed by topology weighting. There are 105 possible unrooted tree topologies for six populations, and 46,656 possibilities to pick one individual from each population out of the set of 36. For each window along the genome, we assessed the relative support of each of the 105 population tree topologies by all 46,656 combinations of six individuals. We found that tree topologies change rapidly along the chromosome (Fig. 3A; Supplementary Fig. 8; Supplementary Data 1). The tree topology obtained for the entire genome (Supplementary Fig. 9) only dominates in few genomic windows (Fig. 3A, black bars "Orig."), while usually one or several other topologies account for more than 75% of the tree topologies (Fig. 3A, grey bars "Misc."). Hardly ever do all combinations of six individuals follow a single population tree topology (Fig. 3A, stars), which implies that in most genomic windows some individuals do not group with their population. Taken together, this indicates a massive sharing of haplotypes across populations and high levels of incomplete lineage sorting. In such a highly mixed genomic landscape, it is close to impossible to separate signals of introgression from incomplete lineage sorting. Still, we highlighted genomic windows that are consistent with the detected introgression from the *Baltic ecotype* into the Bergen populations (Fig. 3A, yellow bars "Intr."; all topologies grouping Por-1SL and He-1SL vs Ber1SL, Ber-2AR, Seh-2AR and Ar-2AR). Regions consistent with introgression are scattered over the entire genome.

**Genomic regions associated with ecotype formation.** Next, we applied three approaches to identify genomic regions associated with divergence between *Atlantic* and *Baltic ecotypes*. First, genomic windows which are dominated by tree topologies that group populations according to ecotype were highlighted (Fig. 3A, red bars "Ecol."). Second, we screened all genetic variants (SNPs and indels; n = 948,128) for those that are overly differentiated between the six populations after correcting for the neutral covariance structure across population allele frequencies (see $\Omega$ matrix, Supplementary Fig. 10A-B). Such variants may indicate local

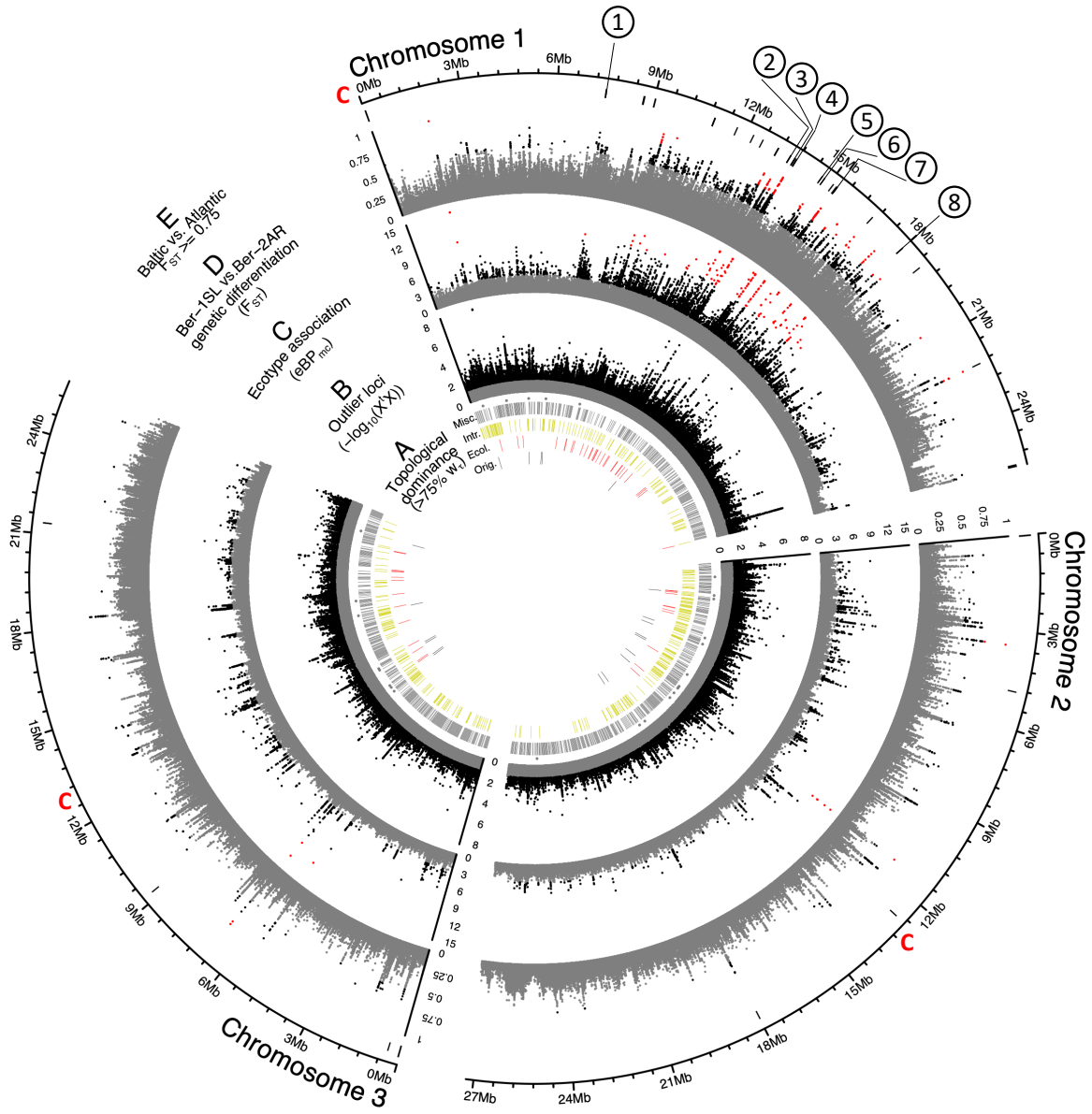



**Figure 3** **Genome screens for haplotype sharing and genotype-ecotype associations**

(**A**) Topology weighting of phylogenetic trees for 36 individuals from the *Baltic* and *Atlantic ecotypes*, as obtained from 50 kb non-overlapping genomic windows. Windows were marked by a bar if they were dominated by one kind of topology ($w_\tau > 75\%$). Most windows are not dominated by the consensus population topology ("Orig."; Suppl. Fig. 9), but by combinations of other topologies ("Misc."). Windows dominated by topologies that separate the *Baltic* and *Atlantic ecotypes* ("Ecol.") are mostly on chromosome 1. Windows consistent with introgression are found all over the genome ("Intr."). (**B**) Distribution of outlier variants (SNPs and Indels) between the six Baltic and *Atlantic ecotype* populations, after global correction for population structure ($X^tX$ statistic). Values below the significance threshold (as obtained by subsampling) are plotted in grey. (**C**) Association of variant frequencies with *Baltic* vs. *Atlantic ecotype* ($eBP_{mc}$). Values below the threshold of 3 (corresponding to $p = 10^{-3}$) are given in grey, values above 10 are given in red. (**D**) Genetic differentiation ($F_{ST}$) between the sympatric ecotypes in Bergen. Values above 0.5 are given in black, values above 0.75 in red. (**E**) The distribution of SNPs with $F_{ST} >= 0.75$ in the *Baltic* vs. *Atlantic ecotypes*. Circled numbers mark the location of the eight most differentiated loci (see Fig. 5). Centromeres of the chromosomes are marked by a red "C".

adaptation. At the same time, we tested for association of these variants with ecotype, as implemented in BayPass[22]. Overly differentiated variants ($X^tX$ statistic; Fig. 3B) and ecotype-associated variants ecotype ($eBP_{mc}$; Fig. 3C) were detected all over the genome, but many were concentrated in the middle of the telocentric chromosome 1. Tests for association of variants with environmental variables such as sea surface temperature or salinity find fewer associated SNPs and no concentration on chromosome 1 (Supplementary Fig. 10D-E), confirming that the detected signals are not due to general genome properties, but are specific to the ecotypes. Third, we expected that gene flow between the sympatric Ber-1SL and Ber-2AR populations would largely homogenize their genomes except for regions involved in ecological adaptation, which would be highlighted as peaks of genetic differentiation. The distributions of $F_{ST}$ values in all pairwise population comparisons confirmed that genetic differentiation was particularly low in the Ber-1SL vs. Ber-2AR comparison (Supplementary Fig. 11 and 12). Pairwise differentiation between Ber-1SL and Ber-2AR (Fig. 3D) shows marked peaks on chromosome 1, most of which coincide with peaks in $X^tX$ and $eBP_{mc}$. Notably, when assessing genetic differentiation of *Baltic* vs *Atlantic ecotype* (72 vs 72 individuals; Fig. 3E; Supplementary Fig. 13), there is not a single diagnostic variant ($F_{ST} = 1$), and even variants with $F_{ST} \geq 0.75$ are very rare (n=63; Fig. 3E).

Genetic divergence ($d_{xy}$), nucleotide diversity ($\pi$) and local linkage disequilibrium ($r^2$) of the two Bergen populations do not show marked differences along or between chromosomes (Supplementary Fig. 14). The cluster of ecotype-associated variants on chromosome 1 overlaps with three large blocks of long-range linkage disequilibrium (LD; Supplementary Fig. 15). However, the boundaries of the LD blocks do not correspond to the ecotype-associated region and differ between populations. LD blocks are not ecotype-specific. Local PCA of the strongly ecotype associated region does not reveal patterns consistent with a chromosomal inversion or another segregating structural variant (Supplementary Fig. 16). Thus, there is no obvious link between the clustering of ecotype-associated loci and structural variation. Notably, genetic differentiation is not generally elevated in the ecotype-associated cluster on chromosome 1, as would be expected for a segregating structural variant, but drops to baseline levels in between ecotype-associated loci (Fig. 3D). Taken together, numerous genomic loci – inside

and outside the cluster on chromosome 1 – are associated with ecological adaptation and none of these are differentially fixed between ecotypes, suggesting that ecotype formation relies on a complex polygenic architecture.

**Adaptation from standing genetic variation.** We next investigated whether adaptive alleles underlying ecotype formation rather represent de novo mutations or standing genetic variation. We selected highly ecotype-associated SNPs ($X^tX$ > 1.152094, threshold obtained from randomized subsampling; $eBP_{mc}$ > 3; n = 3,976; Supplementary Fig. 17A) and assessed to which degree these alleles are shared between the studied populations and other populations across Europe. Allele sharing between the Bergen populations is likely due to ongoing gene flow, and hence Bergen populations were excluded from the analysis. In turn, allele sharing between the geographically isolated Seh-2AR, Ar-2AR, Por-1SL and He-1SL populations likely represents shared ancient polymorphism. Based on this comparison, we found that 82% of the ecotype-associated SNPs are polymorphic in both *Atlantic* and *Baltic ecotypes*, suggesting that the largest part of ecotype-associated alleles originates from standing genetic variation. We then retrieved the same genomic positions from published population resequencing data for *Atlantic ecotype* populations from Vigo (Spain) and St. Jean-de-Luz (Jean, southern France)[11], an area that is potentially the source of postglacial colonization of all locations in this study. We found that 90% of the alleles associated with the Northern European ecotypes are also segregating in at least one of these southern populations, underscoring that adaptation in the North involves a re-assortment of existing standing genetic variation.

**Ecotypes differ mainly in the circadian clock and nervous system development.** We then assessed how all ecotype-associated variants (SNPs and indels; $X^tX$ > 1.148764; $eBP_{mc}$ > 3, n = 4,741; Supplementary Fig. 17B) may affect *C. marinus'* genes. In a first step, we filtered the existing gene models in the CLUMA1.0 reference genome to those that are supported by transcript or protein evidence, have known homologues or PFAM domains, or were manually curated (filtered annotations provided in Supplementary Data 2; 15,193 gene models). Based on this confidence gene set, we then assessed the location of variants relative to genes, as well

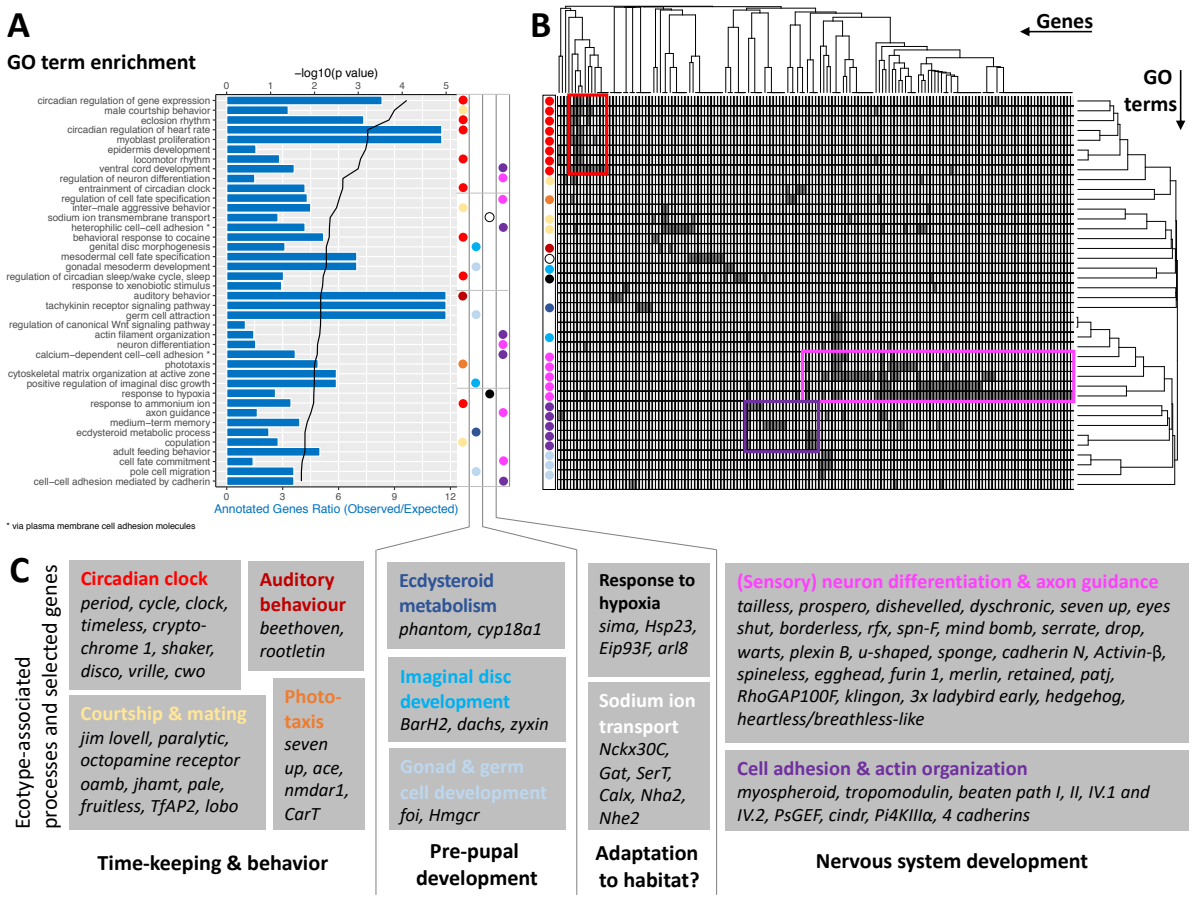



**Figure 4** **GO term analysis of ecotype associated SNPs**

(**A**) The top 40 enriched GO terms are listed for the 1,400 genes that are found to be affected by ecotype-associated genetic variants ($eBP_{mc} > 3$). For each GO term the significance level (black line, top y-axis) and the observed-expected ratio of genes annotated to the respective GO term (blue bars, bottom y-axis) are given. (**B**) The top 40 GO terms are driven by 168 genes. Hierarchical clustering of genes and GO terms reveals major signals in the circadian clock and nervous system development (more details in Supplementary Table 3). (**C**) Most GO terms are consistent with the known ecotype differences and selected genes are highlighted for all of them. Notably, basically all core circadian clock genes are affected.

as the resulting mutational effects (SNPeff[23]; Supplementary Fig. 18; statistics in Supplementary Tab. 2). The vast majority of ecotype-specific variants are classified as intergenic modifier variants, suggesting that ecotype formation might primarily rely on regulatory mutations.

The ecotype-specific SNPs are found in and around 1,400 genes (Supplementary Data 3 and 4). We transferred GO terms from other species to the *Clunio* confidence annotations based on gene orthology (5,393 genes; see methods and Supplementary Data 5). GO term enrichment analysis suggests that ecological adaptation prominently involves the circadian clock, supported by three of the top four GO terms (Fig. 4A). In order to identify which genes drive GO term enrichment in the top 40 GO terms, we extracted the genes that harbour ecotype-associated SNPs (168 genes; Fig. 4B; Supplementary Table 3). We individually confirmed their gene annotations and associated GO terms. Clustering the resulting table by genes and GO terms reveals two dominant signatures (Fig. 4B). Many GO terms are associated with circadian timing and are driven by a small number of genes, which include almost all core circadian clock genes (Fig. 4B and C). As a second strong signal, almost half of the genes are annotated with biological processes involved in nervous system development (Fig. 4B and C). GO term enrichment is also found for ecdysteroid metabolism, imaginal disc development and gonad development (Fig. 4). These processes of pre-pupal development are expected to be under circalunar clock control. The fact that circalunar clocks are responsive to moonlight and water turbulence[6] renders the finding of GO term enrichment for “auditory behaviour” and “phototaxis” interesting. Furthermore, many of the genes involved in nervous system development and sodium ion transport, also have GO terms that implicate them in light- and mechanoreceptor development, wiring or sensitivity (Supplementary Data 5). With the exception of “response to hypoxia” and possibly “sodium ion transmembrane transport”, there are very few GO terms that can be linked to the submerged larval habitat of the *Baltic ecotype*, which is usually low in salinity and can turn hypoxic in summer. There is a striking absence of GO terms involved in metabolic processes or immune response.

Taken together, the detected GO terms are highly consistent with the known ecotype differences and suggest that ecotypes are mainly defined by changes in the circadian clock and nervous system development. A previously unknown aspect of *Clunio* ecotype formation is

highlighted by the GO terms “male courtship behaviour”, “inter-male aggression” and “copulation” (Fig. 4). These processes are subject to sexual selection and considered to evolve fast. They could in the long term entail assortative mating between ecotypes.

**Strongly differentiated loci correspond to GO-term enriched biological processes.** While GO term analysis gives a broad picture of which processes have many genes affected by ecotype-associated SNPs, this does not necessarily imply that these genes and processes also show the strongest association with ecotype. Additionally, major genes might be missed because they were not assigned GO terms. As a second line of evidence, we therefore selected variants with the highest ecotype-association by increasing the $eBP_{mc}$ cut-off to 10. This reduced the set of affected genes from 1,400 to 69 (Supplementary Data 6 and 7). Additionally, we only considered genes with variants that are strongly differentiated between the ecotypes ($F_{ST} > 0.75$, compare Fig. 3E), leaving thirteen genes in eight distinct genomic regions (Fig. 5A; numbered in Fig 3E). Two of these regions contain two genes each with no homology outside *C. marinus* (indicated by “NA”, Fig. 5A), confirming that GO term analysis missed major loci because of a lack of annotation. Three other regions contain the – likely non-visual – photoreceptor *ciliary Opsin 1* (*cOps1*)[24], the transcription factor *longitudinals lacking* (*lola;* in fruit fly involved in axon guidance[25] and photoreceptor determination[26]) and the nuclear receptor *tailless* (*tll;* in fruit fly involved in development of brain and eye[27]), underscoring that ecotype characteristics might involve differential light sensitivity. Interestingly, *tll* also affects development of the neuroendocrine centres involved in ecdysteroid production and adult emergence[28]. Even more, re-annotation of this genomic locus revealed that the neighbouring gene, which is also affected by ecotype specific variants, is the *short neuropeptide F receptor* (*sNPF-R*) gene. Among other functions, sNPF-R is involved in coupling adult emergence to the circadian clock[29]. Similarly, only 100 kb from *cOps1* there is the differentiated locus of *matrix metalloprotease 1* (*Mmp1*), which is known to regulate circadian clock outputs via processing of the neuropeptide *pigment dispersing factor* (PDF)[30]. In both cases, the close genetic linkage could possibly form pre-adapted haplotypes and entail a concerted alteration of sensory and circadian functions in the formation of ecotypes. In the remaining two loci, *sox100B* is known

to affect male gonad development[31] and the *ecdysone-induced protein 93F* is involved in response to hypoxia in flies[32], but was recently found to also affect reproductive cycles in mosquitoes[33]. In summary, only two out of the top 13 ecotype-associated genes were comprised in the top 40 GO terms (Fig. 5A). Nevertheless, all major biological processes detected in GO term analysis (Fig. 4) are also reflected in the strongly ecotype-associated loci (Fig. 5), giving a robust signal that circadian timing, sensory perception and nervous system development are underlying ecotype formation in *C. marinus*.

Finally, we assessed the top 13 strongly ecotype-associated loci for signatures of selective sweeps in genetic diversity and LD (Fig. 5B-G). Despite these loci being the most differentiated between ecotypes in the entire genome, there is at best a mild reduction in genetic diversity and a mild increase in LD (Fig. 5B-G). If selection acted on these loci, it must have been very soft, underscoring a history of polygenic adaptation from standing genetic variation and continued recombination.

**A**

| Locus | Chromosome | Scaffold | Start position | CLUMA CG number | Gene | SNPs: CDS: non-synonymous | SNPs: CDS: synonymous | SNPs: Stop retained | SNPs: Splice site | SNPs: 3' UTR | SNPs: 5' UTR | SNPs: Intron | SNPs: Intergenic | Indels: Inframe (conservative) | Indels: 3' UTR | Indels: Intron | Indels: Intergenic | detected via GO term |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 6 | 68627 | 020452 | Novel nucleolar protein 2 (Non2) | . | . | . | . | . | . | . | 20 | . | . | . | 3 | |
| | 1 | 6 | 88768 | 020457 | Ecdysone-induced protein 93F (Eip93F) | . | . | . | . | . | . | 9 | 20 | . | . | 2 | 3 | Y |
| 2 | 1 | 18 | 629474 | 002631 | pupal cuticle protein 78E (pcp78E) | . | . | . | . | . | . | . | 22 | . | . | . | 6 | |
| | 1 | 18 | 649185 | 002673 | tailless (tll) | 1 | . | . | . | . | . | . | 36 | . | . | . | 11 | Y |
| | 1 | 18 | 669600 | 002806 | *is part of* sNPF receptor (sNPF-R; CG002648) | . | . | . | . | . | . | . | 15 | . | . | . | 5 | |
| 3 | 1 | 18 | 1240686 | 002670 | NA | 4 | 1 | 1 | 1 | . | . | 4 | 7 | 1 | . | 2 | . | |
| | 1 | 18 | 1243996 | 002667 | NA | 1 | 2 | . | . | 2 | . | . | 13 | . | . | . | 5 | |
| 4 | 1 | 18 | 1282649 | 002573 | longitudinals lacking (lola) | . | . | . | . | . | . | 24 | 1 | . | 1 | 15 | . | |
| 5 | 1 | 43 | 1081770 | 008519 | Matrix metalloproteinase 1 (Mmp1) | . | . | . | 1 | 24 | . | 52 | . | . | 2 | 6 | . | |
| 6 | 1 | 43 | 985815 | 008485 | ciliary Opsin 1 (cOps1) | 2 | 2 | . | . | . | . | 8 | 37 | . | . | 5 | 10 | |
| 7 | 1 | 43 | 511465 | 008347 | NA | . | . | . | . | . | . | . | 33 | . | . | . | 6 | |
| | 1 | 43 | 523131 | 023379 | NA | 7 | 2 | . | . | . | . | . | 22 | . | . | . | 6 | |
| 8 | 1 | 42 | 676396 | 008160 | SOX-100B | . | . | . | . | . | . | 5 | . | . | . | 2 | . | |

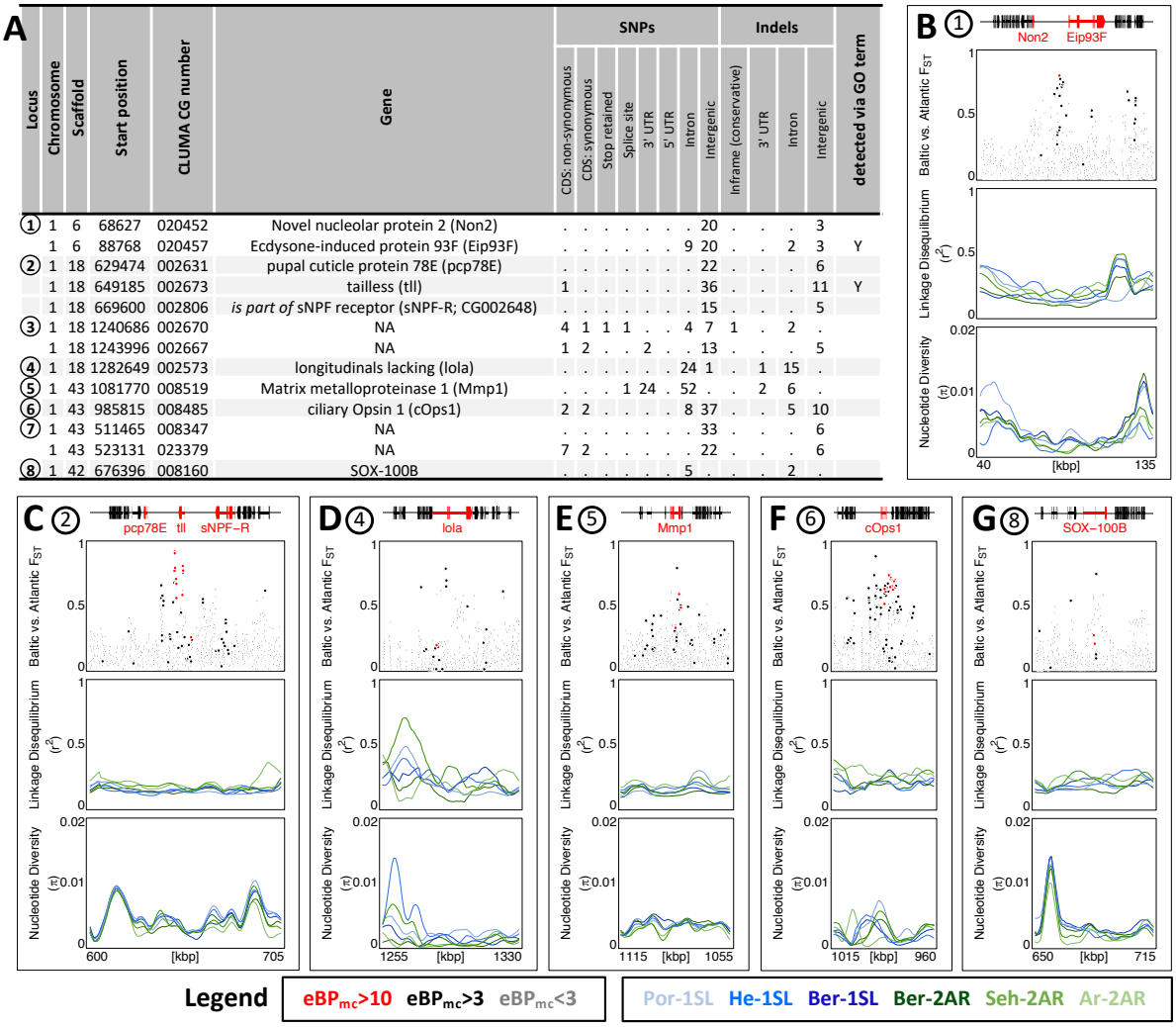



**Figure 5** **The 13 most differentiated ecotype-associated genes**

(**A**) Loci with highly ecotype-associated variants were selected based on eBPmc > 3 and $F_{ST}$(Baltic-Atlantic) > 0.75. There are 13 genes in eight distinct genomic loci. (**B-G**) An overview is given for the six loci with identified genes. In each panel, from top to bottom the sub-panels show the gene models, $F_{ST}$ values of genetic variants in the region, local linkage disequilibrium (LD) and genetic diversity ($\pi$). $F_{ST}$ values are coloured by ecotype association of the variant (red: $eBP_{mc}$ > 10; black: 10 > $eBP_{mc}$ > 3; grey: $eBP_{mc}$ < 3). LD and genetic diversity are shown for the six populations independently, coloured-coded as in Fig 1 and 2. The are no strong signatures of selection.

## Discussion

Inspired by classic literature, we confirmed the existence of three distinct ecotypes of *C. marinus* in Northern Europe. Based on the analysis of 168 genomes, these ecotypes form a single genetic species, exchange genetic material where they occur in sympatry, and established very recently from a common ancestor. While ecotype-associated alleles differ in allele frequency, they are largely shared between ecotypes, which suggests that adaptation primarily involves standing genetic variation from many different loci. A similar re-use of existing regulatory variation has been found in ecotype formation in sticklebacks[34-36] or mimicry in *Heliconius* butterflies[37]. However, while in *Heliconius* alleles are shared over large evolutionary distances via introgression, *Clunio* ecotypes diverged recently from a common source, as is illustrated by massive and genome-wide shared polymorphism. Combined with the observation that many genes from the same biological processes have ecotype-associated alleles, this draws a picture of polygenic adaptation, involving many pre-existing alleles with probably small phenotypic effects. Particularly for adaptation in circadian timing this scenario is highly plausible. The ancestral *Atlantic ecotype* comprises many genetically determined circadian timing types that are adapted to the local tides[8,9,11,38]. Existing genetic variants conveying emergence at dusk were likely selected or re-assorted to form the *Baltic ecotype*'s highly concentrated emergence at dusk.

Besides circadian timing, the ecotypes differ in circalunar timing and oviposition behavior. In our study the vast majority of GO terms and candidate genes is consistent with these functions, leaving little risk for evolutionary "story-telling" based on individual genes or GO terms[39]. We propose that good congruence between known phenotypic differences and detected biological processes could be a hallmark of polygenic adaptation, as only polygenic adaptation is expected to leave a footprint in many genes of the same ecologically relevant biological process. In turn, because of the polygenic architecture, pinpointing individual genes' contributions to a specific phenotype will require additional experiments. Genetic manipulation may not be very informative when assessing highly polygenic traits (see e.g. [40]). But QTL

mapping with recently developed refined statistical algorithms for detection of polygenic signals may hold some promise[41].

Based on our genomic comparison of lunar-rhythmic and lunar-arrhythmic ecotypes, we propose three not mutually exclusive hypotheses on molecular pathways involved in the unknown circalunar clock. Firstly, *Clunio*'s circalunar clock is known to tightly regulate ecdysteroid-dependent development and maturation just prior to pupation[42]. Congruently, our screen identified ecotype-associated genes in the development of imaginal discs and genital discs, and in ecdysteroid metabolism. Lunar arrhythmicity may rely on an escape of these processes from circalunar clock control. Secondly, it has been hypothesized that circalunar clocks involve a circadian clock[43] and such a mechanism has been experimentally confirmed in the midge *Pontomyia oceana*[44]. Thus, the overwhelming circadian signal in our data might be responsible for both circadian timing adaptations and the loss of circalunar rhythms. Thirdly, *Clunio*'s circalunar clock is synchronized with the external lunar cycle via moonlight, as well as tidal cycles of water turbulence and temperature[6]. Our data suggests that sensory receptor development, wiring or sensitivity might differ between ecotypes. Interestingly, some *Atlantic ecotype* populations are insensitive to specific lunar time cues, either moonlight or mechanical stimulation by the tides[8]. These pre-existing insensitivities may have been combined to form completely insensitive and hence lunar-arrhythmic ecotypes. This scenario would fit the general pattern of polygenic adaptation through a re-assortment of standing genetic variation, which emerges from our study.

In several species, genes involved in complex behavioral or ecological syndromes were found to be locked into supergenes by chromosomal inversions, e.g. in *Heliconius* butterfly mimicry[45] or reproductive morphs of the ruff[46]. While we observe a clustering of ecotype-associated alleles in *Clunio*, there is no obvious connection to an underlying structural variant (SV). Possibly, the SV is so complex that it did not leave an interpretable genomic signal. Alternatively, *Clunio*'s long history of genome rearrangements[11] may have resulted in a clustering of ecologically relevant loci without locking them into a single SV. Clustering could be stabilized by low recombination, consistent with the observed three LD blocks, which – while not ecotype-specific – all overlap with the differentiated region. Epistatic interactions between

the clustered loci and co-adaptation of alleles might further reduce the fitness of recombinants and lead to a concerted response to selection. Such an interconnected adaptive cluster might allow for more flexible evolutionary responses than a single, completely linked supergene. Further studies will have to show whether such a genome architecture exists, whether it facilitates adaptation and whether it might itself be selected for.

## Methods

**Nomenclature of ecotypes.** We expanded the existing naming convention of *C. marinus* timing types[38] to also include *Baltic* and *Arctic ecotypes*. Names of populations and corresponding laboratory strains consist of an abbreviation for geographic origin followed by a code for the daily and lunar timing phenotypes. Daily phenotypes in this study are emergence during the first 12 hours after sunrise ("1") or, emergence during the second 12 hours after sunrise ("2") or emergence during every low tide ("t" for tidal rhythm). Lunar phenotypes in this study are either emergence during full moon and new moon low tides ("SL" for semi-lunar) or arrhythmic emergence ("AR"). As a consequence, the *Arctic ecotype* is "tAR", the *Baltic ecotype* is "2AR" and the *Atlantic ecotype* populations in this study are all of timing type "1SL" (while other timing types exist within the *Atlantic ecotype*[38]).

**Fieldwork and sample collection.** Field samples for genetic analysis and establishment of laboratory strains were collected in Sehlendorf (Seh, Germany), Ar (Sweden), Tromsø (Tro, Norway) and Bergen (Ber, Norway) during eight field trips in 2017 and 2018 (Supplementary Tab. 4). Field caught adult males for DNA extraction were directly collected in 99.98 % ethanol and stored at -20°C. Females are immobile and basically invisible in the field, unless found in copulation. Laboratory strains were established by catching copulating pairs in the field and transferring multiple fertilized egg clutches to the laboratory (Supplementary Tab. 4). Samples and laboratory strains of the sympatric ecotypes in Bergen were collected at the same location but at different daytime. Additional samples and laboratory strains from Helgoland (He, Germany) and Port-en-Bessin (Por, France) were collected and described earlier[11,38,47], but had previously not been subject to whole genome sequencing of individuals.

**Laboratory culture and phenotyping of ecotypes.** Laboratory strains were reared under standard conditions[48] at 20°C with 16 h of light and 8 h of darkness. *Atlantic* and *Arctic ecotype* strains were kept in natural seawater diluted 1:1 with deionized water and fed with diatoms (*Phaeodactylum tricornutum*) and powdered nettles (*Urtica sp.*). The *Baltic ecotype* was kept

in natural Sea water diluted 1:2 and fed with diatoms and powdered red algae (90%, *Delesseria spp.*, 10% *Ceramium spp.*, obtained from F. Weinberger and N. Stärck, GEOMAR, Kiel). For entrainment of the lunar rhythm all strains were provided with 12.4 h tidal cycles of water turbulence (mechanically induced vibrations produced by an unbalanced motor, 50 Hz, roughly 30 dB above background noise, 6.2 h on, 6.2 h off)[49,50].

Assignment of strains to ecotypes was confirmed based on their phenotypes as recorded in laboratory culture. Oviposition behavior was assessed during standard culture maintenance: *Baltic ecotype* eggs are generally found submerged at the bottom of the culture vessel, *Atlantic* and *Arctic ecotype* eggs are always found floating on the water surface or on the walls of the culture vessel (see Supplementary Note 1). Daily emergence times were recorded in 1h intervals by direct observation (Seh-2AR, Ar-2AR) or with the help of a fraction collector[51] (Ber-1SL, Ber-2AR, Tro-tAR, Por-1SL, He-1SL; Supplementary Fig. 1). Lunar emergence times were recorded by counting the number of emerged midges in the laboratory cultures every day over several months and summing them up over several tidal turbulence cycles. Emergence data for He-1SL was taken from[52], emergence data for Por-1SL was taken from a manuscript in preparation (D Briševac, C Prakash, TS Kaiser).

**DNA extraction and whole genome sequencing.** For each of the seven populations, 24 field caught males (23 for Por-1SL, 25 for He-1SL) were subject to whole genome sequencing. DNA was extracted from entire individuals with a salting out method[53] and amplified using the REPLI-g Mini Kit (QIAGEN) according to the manufacturer's protocol with volume modifications (Supplementary Tab. 5). All samples were subject to whole genome shotgun sequencing at 15-20x target coverage on an Illumina HiSeq3000 sequencer with 150 bp paired-end reads. Library preparation and sequencing were performed by the Max Planck Genome Centre (Cologne, Germany) according to standard protocols. Raw sequence reads are deposited at ENA under Accession PRJEB43766.

**Sequence data processing, genotyping and SNP filtering.** Raw sequence reads were trimmed for adapters and base quality using Trimmomatic v.0.38[54] with parameters

'ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:8:true', 'LEADING:20', 'TRAILING:20', 'MINLEN:75'. Overlapping paired end reads were merged with PEAR v.0.9.10[55], setting the minimum assembled sequence length to 75 bp and a capping quality score of 20. Assembled and unassembled reads were mapped with BWA-MEM[56] to the nuclear reference genome[11] (ENA accession GCA_900005825.1) and the mitochondrial reference genome (ENA accession CVRI01023763.1) of *C. marinus*. Mapped reads were sorted, indexed, filtered for mapping quality (-q 20) and transformed to BAM format with SAMtools v.1.9[57]. Read group information was added with the AddOrReplaceReadGroups.jar v.1.74 script from the Picard toolkit (http://picard.sourceforge.net/) [58].

For the nuclear genome, SNPs and insertion-deletion (indel) genotypes were called using GATK v.3.8-0-ge9d806836[59]. After initial genotype calling with the GATK HaplotypeCaller and the parameter '-stand_call_conf 30', base qualities were recalibrated with the GATK BaseRecalibrator with '-knownSites' and genotype calling was repeated on the recalibrated BAM files to obtain the final individual VCF files. Individual VCF files were combined using GATK GenotypeGVCFs. SNP and indel genotypes were filtered with VCFtools v.0.1.14 [60] to keep only biallelic polymorphisms (--max-alleles 2), with a minimum minor allele frequency of 0.02 (--maf 0.02), a minimum genotype quality of 20 (--minQ 20) and a maximum proportion of missing data per locus of 40% (--max-missing 0.6), resulting in 792,032 SNPs and 156,096 indels over the entire set of 168 individuals. For certain analyses indels were excluded with VCFtools ('--remove-indels').

Reads mapped to the mitochondrial genome were transformed into mitochondrial haplotypes as described in [61].

**Population genomic analyses.** Mitochondrial haplotype networks were calculated using the Median-Joining algorithm[62] with Network v.10.1.0.0 (fluxus-engineering.com).

Nuclear SNP genotypes were converted to PLINK format with VCFtools. SNPs were LD pruned with PLINK v.1.90b4[63] and parameters '--indep-pairwise 50 10 0.5' as well as '--chr-set 3 no-xy no-mt --nonfounders'. Principal component analysis (PCA) was performed in PLINK using the option '--pca' with the options default settings. The pruned BED file from PLINK was

used as input to ADMIXTURE v.1.3.0[64], with which we assessed a series of models for K=1 to K=10 genetic components, as well as the corresponding the cross-validation error ('--cv'). Migration was further tested by converting the SNP data to TreeMix format with the *vcf2treemix.sh* script[65] and running *TreeMix 1.13*[66] with default parameters and the southernmost Por-1SL as root population.

Population estimates along the chromosomes were calculated in 100 kb overlapping sliding-windows with 10 kb steps. Nucleotide diversity ($\pi$) was calculated for SNPs with VCFtools '--window-pi'. For the genome-wide average, calculations were repeated with 200kb non-overlapping windows. Linkage disequilibrium (LD; as $r^2$) was calculated in VCFtools with '--geno-r2'. Local LD was calculated with '--ld-window-bp 500'. Preliminary tests showed that local LD decays within a few hundred base pairs (Supplementary Fig. 19). For long range LD minor allele frequency was filtered to 0.2 ('--maf 0.2', resulting in 335,800 SNPs), only values larger 0.5 were allowed with '--min-r2 0.5' and the '--ld-window-bp 500' filter was removed. Pairwise $F_{ST}$ was calculated with VCFtools '--weir-fst-pop' option per SNP and in sliding windows. For calculation of genetic divergence ($d_{xy}$), allele frequencies were extracted with VCFtools '--freq' and $d_{xy}$ was estimated from allele frequencies according to [67].

**Phylogenomics and topology weighting.** Nuclear genome phylogeny was calculated for a random set of six individuals from each population, without Tro-tAR (n=36). For windowed phylogenies, the VCF file was subset into non-overlapping 50 kb windows using VCFtools '--from-bp --to-bp'. SNP genotypes were transformed into FASTA alignments of only informative sites with the vcf2phylip.py v.2.3 script[68] and parameters '-m 1 -p -f'. Heterozygous genotypes were represented by the respective IUPAC code for both bases. Whole genome and windowed phylogenies were calculated with IQ-TREE v.1.6.12 [69] using the parameters '-st DNA -m MFP -keep-ident -redo' for the windowed and '-st DNA -m MFP -keep-ident -bb 1000 -bnni -nt 10 -redo' for the whole genome phylogenies. Topology weighting was performed on the windowed phylogenies with TWISST [70] and the parameter '--method complete'.

**Association analysis.** Population-based association between genetic variants (SNPs and Indels) and ecotype, as well as environmental variables (Supplementary Tab. 6) was assessed in BayPass v.2.2 [22]. Allele counts were obtained with VCFtools option '--counts'. Analyzed covariates were ecotype, sea surface salinity (obtained from [71]) and average water temperature of the year 2020 (obtained from weather-atlas.com, accessed 27.04.2020; 16:38), as given in Supplementary Tab. 6. BayPass was run with the MCMC covariate model. BayPass corrects for population structure via Ω dissimilarity matrices, then calculates the $X^tX$ statistics and finally assesses the approximate Bayesian p value of association ($eBP_{mc}$). To obtain a significance threshold for $X^tX$ values, the data was randomly subsampled (100,000 genetic variants) and re-analyzed with the standard covariate model, as implemented in baypass_utils.R. All analyses we performed in three replicates (starting seeds 5001, 24306 and 1855) and the median is shown.

**SNP effects and GO term enrichment analysis.** Gene annotations to the CLUMA1.0 reference genome [11] were considered reliable if they fulfilled one of three criteria: 1) Identified ortholog in UniProtKB/Swiss-Prot or non-redundant protein sequences (nr) at NCBI or PFAM domain, as reported in [11]. 2) Overlap of either at least 20% with mapped transcript data or 40% with mapped protein data, as reported in [11]. 3) Manually annotated. This resulted in a 15,193 confidence genes models. The location and putative effects of the SNPs and indels relative to these confidence gene models were annotated using SnpEff 4.5 [23] (build 2020-04-15 22:26, non-default parameter `-ud 0'). Gene Ontology (GO) terms were annotated with emapper-2.0.1. [72] from the eggNOG 5.0 database[73], using DIAMOND [74], BLASTP e-value $<1e^{-10}$ and subject-query alignment coverage of >60%. Conservatively, we only transferred GO terms with "non-electronic" GO evidence from best-hit orthologs restricted to an automatically adjusted per-query taxonomic scope, resulting in 5,393 *C. marinus* gene models with GO term annotations. Enrichment of "Biological Process" GO terms in the genes associated with ecotype-specific polymorphisms was assessed with the weight01 Fisher's exact test implemented in topGO [75] (version 2.42.0, R version 4.0.3).

**Figure preparation.** Figures were prepared in R[76]. Data were handled with the 'data.table'[77] and 'plyr'[78] packages. The map of Europe was generated using the packages 'ggplot2'[79] and 'ggrepel' [80], 'maps'[81] and 'mapdata'[82]. The map was taken from the CIA World DataBank II (http://www.evl.uic.edu/pape/data/WDB/). Circular plots were prepared using the R package 'circlize'[83]. Multiple plots were combined in R using the package 'Rmisc'[84]. The graphical editing of the whole genome phylogeny was done in Archeopteryx (http://www.phylosoft.org/archaeopteryx)[85]. Final figure combination and graphical editing of the raw plot files was done in *Inkscape*. Neighbor Joining trees of the omega statistic distances from BayPass were created with the R package 'ape'.[86] In all plots the order and orientation of scaffolds within the chromosomes follows the published genetic linkage map[11].

## Acknowledgements

For field work we obtained logistic support from the Ar Research Station (Uppsala University), the Marine Biological Station Espegrend (University of Bergen), Even Jørgensen (The Arctic University of Norway, Tromsø) and Florian Weinberger and Nadja Stärck (GEOMAR Helmholtz Centre for Ocean Research Kiel). We thank Jürgen Reunert, Kerstin Schäfer and Susanne Mentz for technical assistance, as well as all members of the MPRG "Biological Clocks" for discussion and support. Diethard Tautz and Julien Dutheil critically read the manuscript. Whole genome sequencing was performed at the Max Planck Genome Center (Cologne) with financial support from the Max Planck Society. This work was funded by the Max Planck Society through the Max Planck Research Group "Biological Clocks" and a sequencing grant. The work was further funded by the European Research Council (ERC) under the Horizon 2020 research and innovation program with an ERC Starting Grant (Grant agreement 802923) awarded to TSK.

## Author contributions

NF performed field work, laboratory work, population genomic analyses and association analysis, prepared the figures and participated in drafting the manuscript. CP analyzed SNP effects and GO term enrichment and participated in figure preparation. TSK conceived, designed and supervised the study, participated in data analysis and figure preparation and wrote the manuscript. All authors approved of the final manuscript.

## References


1 Kawecki, T. J. & Ebert, D. Conceptual issues in local adaptation. *Ecology letters* **7**, 1225-1241 (2004).

2 Savolainen, O., Lascoux, M. & Merilä, J. Ecological genomics of local adaptation. *Nature Reviews Genetics* **14**, 807-820 (2013).

3 Hoffmann, A. *et al.* A framework for incorporating evolutionary genomics into biodiversity conservation and management. *Climate Change Responses* **2**, 1 (2015).

4 Thackeray, S. J. *et al.* Phenological sensitivity to climate across taxa and trophic levels. *Nature* **535**, 241-245 (2016).

5 Yuan, M. & Stinchcombe, J. R. Population genomics of parallel adaptation. *Molecular Ecology* **29**, 4033-4036 (2020).

6 Neumann, D. in *Annual, Lunar, and Tidal Clocks* (eds Hideharu Numata & Barbara Helm) Ch. 1, 3-24 (Springer Japan, 2014).

7 Andreatta, G. & Tessmar-Raible, K. The still dark side of the moon: molecular mechanisms of lunar-controlled rhythms and clocks. *J Mol Biol* **432**, 3525-3546 (2020).

8 Kaiser, T. S. in *Annual, Lunar, and Tidal Clocks* (eds Hideharu Numata & Barbara Helm) Ch. 7, 121-141 (Springer Japan, 2014).

9 Neumann, D. Genetic adaptation in emergence time of *Clunio* populations to different tidal conditions. *Helgoländer wissenschaftliche Meeresuntersuchungen* **15**, 163-171 (1967).

10 Kaiser, T. S., Neumann, D. & Heckel, D. G. Timing the tides: Genetic control of diurnal and lunar emergence times is correlated in the marine midge *Clunio marinus*. *BMC Genetics* **12**, 49, doi:10.1186/1471-2156-12-49 (2011).

11 Kaiser, T. S. *et al.* The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature* **540**, 69-73, doi:10.1038/nature20151 (2016).

12 Kaiser, T. S. & Heckel, D. G. Genetic Architecture of Local Adaptation in Lunar and Diurnal Emergence Times of the Marine Midge *Clunio marinus* (Chironomidae, Diptera). *PLoS ONE* **7**, e32092, doi:10.1371/journal.pone.0032092 (2012).

13 Remmert, H. Ökologische Untersuchungen über die Dipteren der Nord- und Ostsee. *Archiv für Hydrobiologie* **51**, 1-53 (1955).

14 Endraß, U. Physiologische Anpassungen eines marinen Insekts. I. Die zeitliche Steuerung der Entwicklung. *Marine Biology* **34**, 361-368 (1976).

15 Palmén, E. & Lindeberg, B. The marine midge, Clunio marinus Hal. (Dipt., Chironomidae), found in brackish water in the northern Baltic. *Internationale Revue der gesamten Hydrobiologie und Hydrogeographie* **44**, 383-393 (1959).

16 Neumann, D. & Honegger, H. W. Adaptations of the Intertidal Midge Clunio to Arctic Conditions. *Oecologia* **3**, 1-13 (1969).

17 Pflüger, W. & Neumann, D. Die Steuerung einer gezeitenparallelen Schlüpfrhythmik nach dem Sanduhr-Prinzip. *Oecologia* **7**, 262-266 (1971).

18 Endraß, U. Physiologische Anpassungen eines marinen Insekts. II. Die Eigenschaften von schwimmenden und absinkenden Eigelegen. *Marine Biology* **36**, 47-60 (1976).

19 Heimbach, F. Sympatric species, Clunio marinus Hal. and Cl. balticus n. sp. (Dipt., Chironomidae), isolated by differences in diel emergence time. *Oecologia* **32**, 195-202 (1978).

20 Patton, H. *et al.* Deglaciation of the Eurasian ice sheet complex. *Quaternary Science Reviews* **169**, 148-172 (2017).

21 Hofmann, W. & Winn, K. The littorina transgression in the Western Baltic Sea as indicated by subfossil Chironomidae (Diptera) and Cladocera (Crustacea). *International Review of Hydrobiology* **85**, 267-291 (2000).
22 Gautier, M. Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* **201**, 1555-1579 (2015).
23 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain $w^{1118}$; *iso*-2; *iso*-3. *Fly* **6**, 80-92, doi:10.4161/fly.19695 (2012).
24 Velarde, R. A., Sauer, C. D., Walden, K. K. O., Fahrbach, S. E. & Robertson, H. M. Pteropsin: A vertebrate-like non-visual opsin expressed in the honey bee brain. *Insect Biochemistry and Molecular Biology* **35**, 1367-1377 (2005).
25 Crowner, D., Madden, K., Goeke, S. & Giniger, E. Lola regulates midline crossing of CNS axons in Drosophila. *Development* **129**, 1317-1325 (2002).
26 Zheng, L. & Carthew, R. W. Lola regulates cell fate by antagonizing Notch induction in the Drosophila eye. *Mechanisms of Development* **125**, 18-29 (2008).
27 Suzuki, T. & Saigo, K. Transcriptional regulation of atonal required for Drosophila larval eye development by concerted action of eyes absent, sine oculis and hedgehog signaling independent of fused kinase and cubitus interruptus. *Development* **127**, 1531-1540 (2000).
28 de Velasco, B. *et al.* Specification and development of the pars intercerebralis and pars lateralis, neuroendocrine command centers in the Drosophila brain. *Developmental Biology* **302**, 309-323 (2007).
29 Selcho, M. *et al.* Central and peripheral clocks are coupled by a neuropeptide pathway in Drosophila. *Nature Communications* **8**, 15563, doi:10.1038/ncomms15563 (2017).
30 Depetris-Chauvin, A. *et al.* Mmp1 Processing of the PDF Neuropeptide Regulates Circadian Structural Plasticity of Pacemaker Neurons. *PLOS Genetics* **10**, e1004700, doi:10.1371/journal.pgen.1004700 (2014).
31 Nanda, S. *et al.* Sox100B, a Drosophila group E Sox-domain gene, is required for somatic testis differentiation. *Sexual Development* **3**, 26-37 (2009).
32 Lee, S.-J., Feldman, R. & O'Farrell, P. H. An RNA interference screen identifies a novel regulator of target of rapamycin that mediates hypoxia suppression of translation in Drosophila S2 cells. *Molecular Biology of the Cell* **19**, 4051-4061 (2008).
33 Wang, X. *et al.* The ecdysone-induced protein 93 is a key factor regulating gonadotrophic cycles in the adult female mosquito Aedes aegypti. *Proceedings of the National Academy of Sciences* **118** (2021).
34 Chan, Y. F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**, 302-305 (2010).
35 Kingman, G. A. R. *et al.* Predicting future from past: The genomic basis of recurrent and rapid stickleback evolution. *bioRxiv* (2020).
36 Verta, J.-P. & Jones, F. C. Predominance of cis-regulatory changes in parallel expression divergence of sticklebacks. *Elife* **8**, e43785 (2019).
37 Edelman, N. B. *et al.* Genomic architecture and introgression shape a butterfly radiation. *Science* **366**, 594-599, doi:10.1126/science.aaw2090 (2019).
38 Kaiser, T. S., von Haeseler, A., Tessmar-Raible, K. & Heckel, D. G. Timing strains of the marine insect Clunio marinus diverged and persist with gene flow. *Molecular Ecology* **30**, 1264-1280, doi:https://doi.org/10.1111/mec.15791 (2021).
39 Pavlidis, P., Jensen, J. D., Stephan, W. & Stamatakis, A. A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic

Scans. *Molecular Biology and Evolution* **29**, 3237-3248, doi:Doi 10.1093/Molbev/Mss136 (2012).
40 Zhang, W., Reeves, G. R. & Tautz, D. Testing implications of the omnigenic model for the genetic analysis of loci identified through genome-wide association. *Current Biology* **31**, 1092-1098. e1096 (2021).
41 Wellenreuther, M. & Hansson, B. Detecting polygenic evolution: problems, pitfalls, and promises. *Trends in Genetics* **32**, 155-164 (2016).
42 Neumann, D. & Spindler, K. D. Circasemilunar Control of Imaginal Disk Development in Clunio marinus - Temporal Switching Point, Temperature-Compensated Developmental Time and Ecdysteroid Profile. *Journal of Insect Physiology* **37**, 101-109 (1991).
43 Bünning, E. & Müller, D. Wie messen Organismen lunare Zyklen? *Zeitschrift für Naturforschung* **16**, 391-395 (1962).
44 Soong, K. & Chang, Y.-H. Counting Circadian Cycles to Determine the Period of a Circasemilunar Rhythm in a Marine Insect. *Chronobiology International* **29**, 1329-1335, doi:10.3109/07420528.2012.728548 (2012).
45 Joron, M. *et al.* Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**, 203-U102, doi:10.1038/nature10341 (2011).
46 Küpper, C. *et al.* A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics* **48**, 79-83 (2016).
47 Kaiser, T. S., Neumann, D., Heckel, D. G. & Berendonk, T. U. Strong genetic differentiation and postglacial origin of populations in the marine midge *Clunio marinus* (Chironomidae, Diptera). *Molecular Ecology* **19**, 2845-2857, doi:10.1111/j.1365-294X.2010.04706.x (2010).
48 Neumann, D. Die lunare und tägliche Schlüpfperiodik der Mücke *Clunio* - Steuerung und Abstimmung auf die Gezeitenperiodik. *Zeitschrift für Vergleichende Physiologie* **53**, 1-61 (1966).
49 Neumann, D. & Heimbach, F. in *Cyclic Phenomena in Marine Plants and Animals* (eds E. Naylor & R.G. Hartnoll) 423-433 (Pergamon Press, 1979).
50 Neumann, D. Entrainment of a Semilunar Rhythm by Simulated Tidal Cycles of Mechanical Disturbance. *Journal of Experimental Marine Biology and Ecology* **35**, 73-85 (1978).
51 Honegger, H. W. An automatic device for the investigation of the rhythmic emergence pattern of Clunio marinus. *International Journal of Chronobiology* **4**, 217-221 (1977).
52 Neumann, D. Die zeitliche Programmierung von Tieren auf periodische Umweltbedingungen. *Rheinisch-Westfälische Akademie der Wissenschaften, Natur-Ingenieur- und Wirtschaftswissenschaften, Vortraege*, 31-62 (1983).
53 Reineke, A., Karlovsky, P. & Zebitz, C. P. W. Preparation and purification of DNA from insects for AFLP analysis. *Insect Molecular Biology* **7**, 95-99 (1998).
54 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
55 Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614-620 (2014).
56 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
57 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

58 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491 (2011).
59 McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
60 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
61 Fuhrmann, N. & Kaiser, T. S. The importance of DNA barcode choice in biogeographic analyses—a case study on marine midges of the genus Clunio. *Genome*, 1-11 (2020).
62 Bandelt, H. J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* **16**, 37-48 (1999).
63 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-13015-10047-13748 (2015).
64 Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246, doi:10.1186/1471-2105-12-246 (2011).
65 Ravinet, M. <https://github.com/speciationgenomics/scripts/blob/master/vcf2treemix.sh> (last accessed 16th April 2021).
66 Pickrell, J. & Pritchard, J. Inference of population splits and mixtures from genome-wide allele frequency data. *Nature Precedings*, 1-1 (2012).
67 Delmore, K. E. *et al.* Genomic analysis of a migratory divide reveals candidate genes for migration and implicates selective sweeps in generating islands of differentiation. *Molecular Ecology* **24**, 1873-1888 (2015).
68 Ortiz, E. vcf2phylip v2. 0: convert a VCF matrix into several matrix formats for phylogenetic analysis. *URL https://doi. org/10.5281/zenodo* **2540861** (2019).
69 Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**, 268-274 (2015).
70 Martin, S. H. & Van Belleghem, S. M. Exploring evolutionary relationships across the genome using topology weighting. *Genetics* **206**, 429-438 (2017).
71 Hordoir, R. *et al.* Nemo-Nordic 1.0: a NEMO-based ocean model for the Baltic and North seas–research and operational applications. *Geoscientific Model Development* **12**, 363-386 (2019).
72 Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution* **34**, 2115-2122 (2017).
73 Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309-D314 (2019).
74 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59-60 (2015).
75 Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for gene ontology. *R package version* **2**, 2010 (2010).
76 Crawley, M. J. *The R Book*. (John Wiley & Sons Ltd., 2007).
77 Dowle, M. *et al.* Package 'data.table'. *Extension of 'data. frame* (2019).
78 Wickham, H. The split-apply-combine strategy for data analysis. *Journal of Statistical Software* **40**, 1-29 (2011).

79 Wickham, H. *ggplot2: elegant graphics for data analysis*. (springer, 2016).
80 Slowikowski, K. *et al.* Package ggrepel. *Automatically Position Non-Overlapping Text Labels with 'ggplot2* (2018).
81 Brownrigg, R., Minka, T. & Deckmyn, A. maps: Draw Geographical Maps. R package version 3.3.0. (2018).
82 Becker, R., Wilks, A. & Brownrigg, R. mapdata: Extra map databases. R package version 2.3.0. (2018).
83 Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2812 (2014).
84 Hope, R. M. 'Rmisc: Ryan miscellaneous'. R package version 1.5. 2 (2013).
85 Han, M. V. & Zmasek, C. M. phyloXML: XML for evolutionary biology and comparative genomics. *BMC bioinformatics* **10**, 1-6 (2009).
86 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528 (2019).