The loss of circalunar rhythms in arctic and tide-free habitats

Genomic investigations into lunar-arrhythmic populations of *Clunio marinus*

Dissertation

zur Erlangung des Doktorgrades

Doctor rerum naturalium (*Dr. rer. nat.*)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Nico Fuhrmann

Plön, Februar, 2022

First referee:	Prof. Dr. Eva H. Stukenbrock
Second referee:	Prof. Dr. Hinrich Schulenburg
Date of disputation:	21.01.2022

Contents

List of Figures

List of Tables

Summary of the dissertation

Biological rhythms are adaptations to periodically changing environmental conditions. The non-biting midge *Clunio marinus* (Diptera: Chironomidae) is known for the link between its reproduction and the tidal regime. The short-lived adults emerge when most of the intertidal habitat is exposed. The spring low tides occur at location specific times on days around the full moon and new moon. *C. marinus* populations at the European Atlantic coast are locally adapted to the day time and lunar phase of the spring low tides. This timing is achieved through the combination of an circadian and circalunar rhythm. While the circadian rhythm is controlled by a transcriptional-translational feedback loop, the molecular workings of the circalunar rhythm are not understood yet.

As tides are almost neglectable in the Baltic Sea, the local *Clunio* populations have adapted to lay the egg clutches in the open sea instead of an exposed intertidal substrate. This simultaneously removed the selective pressure to time the reproduction to the lunar phase and allowed for lunar-arrhythmic emergence throughout the entire mating season. In arctic habitats of the Atlantic coast tides are still present. During the mating season the sun illuminates the habitat around the clock, preventing the perception of moon light. *C. marinus* changed from circadian-circalunar-controlled emergence to circatidal rhythms in polar day conditions. The adults emerge every day at every low tide throughout the mating season. In my thesis, I investigated these cases of lunar-arrhythmicity in Northern European *Clunio* populations. By exploring the genetic features linked to the evolution of the here described ecotypes of *C. marinus*, we step further towards understanding the enigmatic circalunar rhythms. My investigations resulted in one published article, one published preprint and an additional chapter.

The first article had two aims: First, I investigated the ability of short mitochondrial fragments to recover the whole mitochondrial biogeography of geographically distinct populations. DNA barcodes are short, conserved genomic fragments and commonly used to reconstruct the biogeography of species. With my *Clunio* populations as example I wanted to point out what issues can arise from blindly using those highly conserved DNA fragments. The second aim was to get the basic mitochondrial biogeography of all distinct population as a foundation to the investigations into the evolution of lunar-arrhythmic ecotypes.

My second chapter is separated into two parts. At first I take a look at the evolution of lunar-arrhythmicity in the studied populations. Population structure and admixture analyses in addition to the mitochondrial biogeography were combined to identify the historical scenario which lead to the evolution of lunar-arrhythmic populations. Secondly, I used direct genomic comparisons to find differentiated regions and adaptive loci between rhythmic populations from the Atlantic coast and the arrhythmic populations from the Baltic Sea specifically. Established laboratory cultures of two sympatric populations were crossed for further insight into the nature of the maintenance of both populations under gene flow. In my article I identify genetic variants differentiated between lunar-rhythmic and lunar-arrhythmic populations. Genetic clusters affected by those genetic variants comprise genes for the control of circadian rhythms, neuronal development, mating behavior, responses to hypoxia and sodium ion transport.

In my third chapter I performed a crossing experiment to identify putative genotypes linked to lunar-rhythmic phenotypes. By crossing two sympatric populations with differing ecotypes, I was able to raise an F2 generation with a mix of rhythmic and arrhythmic phenotypes. With the use of PCR-primers designed specifically for differentiated regions between the grandparent genomes I obtained genotypes for six distinct loci per chromosome of 237 individuals. The QTL analysis revealed multiple significant loci on all chromosomes with nine investigated phenotypes linked to lunar-rhythmicity.

My thesis takes a large step towards the understanding of the circalunar rhythms in *C. marinus* by comparing rhythmic to naturally occurring arrhythmic populations. I generated a comprehensive genomic resource for geographically and ecologically distant populations of the same species. Genomic screens for ecotype-adaptive loci identified a putative involvement of circadian clock genes in circalunar rhythms of *C. marinus*. A crossing experiment between rhythmic and arrhythmic ecotypes of the sympatric Bergen populations hinted towards the involvement of multiple loci across the genome in lunar-rhythmicity. The addition of further genetic markers could identify a link of the circadian clock to circalunar rhythms as well as unravel the maintenance of sympatric ecotypes.

Zusammenfassung der Dissertation

Biologische Rhythmen sind Anpassungen an sich periodisch verändernde Umweltbedingungen. Die Zuckmücke *Clunio marinus* (Diptera: Chironomidae) ist bekannt für ihre an die Gezeiten gekoppelte Fortpflanzung. Wenn der Großteil der Gezeitenzone trockengefallen ist schlüpfen die kurzlebigen adulten Insekten. Zu ortsspezifischen Zeiten um die Vollmond- und Neumond-Tage kommen Springniedrigwasser vor. *C. marinus* Populationen der Europäischen Atlantikküste sind lokal an die Tageszeiten und die Mondphasen, in der die Springniedrigwasser vorkommen angepasst. Die Kombination von circadianen und circalunaren Rhythmen ermöglicht diese zeitliche Koordinierung. Während der circadiane Rhythmus durch eine Transkriptions-Translations-Rückkopplungsschleife kontrolliert wird, kann die molekulare Funktionsweise des circalunaren Rhythmus bisher nicht nachvollzogen werden.

Da der Tidenhub in der Ostsee zu vernachlässigen ist, legen dortige *Clunio*-Populationen als Anpassung die Gelege auf offener See anstatt dem Substrat der freigelegten Gezeitenzone ab. Gleichzeitig verschwindet der Selektionsdruck, das Fortpflanzen zeitlich an die Mondphasen anzupassen und so wird ein lunar-arrhythmisches Schlüpfen während der gesamten Paarungszeit möglich. In arktischen Gebieten des Atlantischen Ozeans sind die Gezeiten immer noch präsent. Mondlicht kann aber hier während der Paarungszeit nicht wahrgenommen werden, weil die Sonne das Habitat rund um die Uhr beleuchtet. Unter Polartag-Bedingungen hat *C. marinus* von einem circadian-circalunar-kontrollierten Schlüpfen zu einem circatidalen Rhythmus gewechselt. Die ausgewachsenen Tiere schlüpfen jeden Tag zu jedem Niedrigwasser während der Paarungszeit. Ich habe in meiner Dissertation diese Fälle lunarer Arrhythmizität in nordeuropäischen *Clunio*-Populationen untersucht. Durch die Erforschung genetischer Merkmale welche an die hier beschriebenen *C. marinus* Ökotypen gekoppelt sind, kommen wir dem Verständnis dieser enigmatischen circalunaren Rhythmen ein ganzes Stück näher. Meine Untersuchungen resultieren in einem publizierten Artikel, einem veröffentlichten Preprint und ein zusätzliches Kapitel.

Der erste Artikel hatte zwei Ziele: Zuerst untersuchte ich die Leistungsfähigkeit kurzer mitochondrialer Fragmente, die gesamt-mitochondriale Biogeographie geographisch unterscheidbarer Populationen wiederherzustellen. DNA-Barcodes sind kurze, konservierte genomische Fragmente und werden häufig verwendet um die Biogeographie einer Art zu rekonstruieren. Mit dem Beispiel meiner *Clunio*-Populationen wollte ich hervorheben, welche Probleme auftreten können, wenn diese hoch konservierten DNS Fragmente blin-

dlings verwendet werden. Das zweite Ziel war das Erlangen einer grundlegenden mitochondrialen Biogeographie aller unterscheidbarer Populationen als Fundament nachfolgender Untersuchungen in die Evolution lunar-arrhythmischer Ökotypen.

Mein zweites Kapitel ist unterteilt in zwei Abschnitte. Zuerst sehe ich mir die Evolution der lunaren Arrhythmizität in den untersuchten Populationen an. Die Struktur der Populationen und ADMIXTURE Analysen wurden zusätzlich zu der mitochondrialen Biogeographie kombiniert um das historische Szenario zu identifizieren, welches zur Evolution von lunar-arrhythmischen Populationen führte. Zweitens verwendete ich direkte genomische Vergleiche, um differenzierte Bereiche und angepasste Loci explizit zwischen rhythmischen Populationen der Atlantikküste und den arrhythmischen Populationen der Ostsee zu finden. Um ein näheres Verständnis für die Art der Aufrechterhaltung zweier Populationen unter Genfluss zu bekommen, wurden etablierte Laborkulturen von zwei sympatrischen Populationen gekreuzt. Ich identifiziere in meinem Artikel genetische Varianten, welche zwischen lunar-rhythmischen und lunar-arrhythmischen Populationen differenziert sind. Gen-Cluster, die von diesen genetischen Varianten betroffen sind, setzen sich zusammen aus Genen für die Kontrolle des circadianen Rhythmus, Entwicklung des Nervensystems, Paarungsverhalten, Reaktion auf Sauerstoffmangel und Natriumionentransport.

In meinem dritten Kapitel führte ich ein Kreuzungs Experiment durch um mögliche Verbindungen zwischen Genotypen und lunar-rhythmischen Phänotypen zu identifizieren. Durch die Verkreuzung zweier sympatrischer Populationen mit unterschiedlichen Ökotypen konnte ich eine F2 Generation mit einer Mischung aus rhythmischen und arrhythmischen Phänotypen heranziehen. Speziell für differenzierte Bereiche zwischen den Genomen der Großeltern entworfener PCR-Primer wurden verwendet um Genotypen für sechs unterscheidbare Loci je Chromosom von 237 Individuen zu erhalten. Die QTL Analyse ergab, dass mehrere signifikante Loci auf allen Chromosomen mit allen neun untersuchten Phänotypen eine Verbindung to lunarer Rhythmizität haben.

Meine Dissertation kommt dem Verstehen von lunaren Rhythmen in *C. marinus* einen großen Schritt näher, durch den Vergleich von rhythmischen zu natürlich vorkommenden arrhythmischen Populationen. So erstellte ich eine umfangreiche genomische Ressource von geographisch und ökologisch unterscheidbarer Populationen derselben Art. Genomische Suchen nach Ökotyp-angepassten Loci identifizierten eine mögliche Beteiligung von den Genen der circadianen Uhr an den circalunaren Rhythmen von *C. marinus*. Ein Kreuzungsversuch zwischen dem rhythmischen und arrhythmischen Ökotyp der sympa-

trischen Bergen Populationen lieferten Hinweise auf eine Beteiligung mehrerer Loci an lunarer Rhythmizität auf dem gesamten Genom. Zusätzliche genetische Marker könnten eine Verbingung der circadianen Uhr mit den circalunaren Rhythmen identifizieren sowie Hinweise zu der Aufrechterhaltung sympatrischer Ökotypen entschlüsseln.

1 General introduction

1.1 The origin of natural cycles in the environment

There are periodic environmental conditions on earth to which almost all life on earth is constantly exposed to. Prominent periods in nature are from short to long tidal cycles, day-night (diel) cycles, lunar (synodic) cycles and annual cycles. All these cycles originate from the planetary movements and relative positions of the earth, moon and sun (Brown, 1960). The rotation of the earth around its axis causes both the tidal and the diel cycle. One full rotation of the earth around its axis takes 24 hours resulting in the change from day to night (Fig. 1A:α) (Brown, 1960; Neumann, 2014; Palmer, 1995). The movement of the moon around the earth is much slower than the earth's rotation and one full cycle (synodic month) takes 29.53 days (Fig. 1A:β). The water masses of the oceans being constantly pulled toward the moon and the sun, while the earth keeps the 24-hour rotations, causing tidal waves along most ocean coastlines. As the moon's position is moving throughout the synodic month, the daytime of the tidal cycle takes 12.4 hours and occurs parallel to the synodic month.

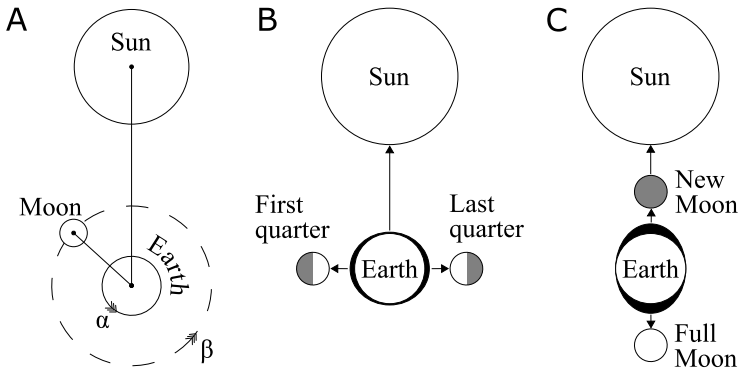

Figure 1: Natural cycles and their effects on Earth.

(**A**) Schematic origin of the two major geophysical cycles affecting intertidal zones. The moon's orbit is marked as a dashed line. The earth's (α) and moon's (β) rotation direction are shown as three arrowheads. Edited from Brown (1960). (**B,C**) Effects of the relative angle between the moon and the sun to the earth on the tidal amplitude at (B) the first (left side gray circle) and last quarter (right side gray circle) of the synodic month and at (C) days around the full moon (empty circle) and new moon (gray circle). The pulled water masses are depicted as black ovals around the schematic earth. The directions of gravitational pulls are shown as black arrows. Edited from Palmer (1974).

This periodicity results in the same daily timing of tidal waves every 14.77 days, on the full moon and the new moon days (Neumann, 2014). Finally, annual cycles are observed on earth as seasons and result from the position of the earth on its elliptic orbit around the sun, which takes 365.25 days for one full cycle (Brown, 1960).

1.2 Fundamentals of biological rhythms

Organism in periodic environmental conditions have to adapt by anticipating the timing of the cycles and prepare for them (Neumann, 2014). Biological rhythms are an observable response in a cell, organism or population and an evolutionary adaptation to periodically imposed changes of the environment (Aschoff, 1981a). Biological rhythms are synchronized or entrained by an external stimulus ("zeitgeber") which is part of a natural cycle. Rhythms are further distinguished between exogenous or endogenous by checking the presence of the "free-running" period when the zeitgeber has been removed after the initial entrainment (Aschoff, 1981b; Pittendrigh, 1960). An exogenous rhythm is only expressed as direct reaction to the zeitgeber. If the zeitgeber is not perceived by the organism, the observed rhythm stops after the last entrained cycle as it does not have a self-sustaining endogenous oscillator (Aschoff, 1981b; Daan, 1982; Pittendrigh, 1960). Without the zeitgeber, an endogenous rhythm will free-run over several cycles before it fades to arrhythmicity. This observation is the product of an endogenous oscillator called the "biological clock", which is entrained and synchronized by the zeitgeber (Aschoff, 1981b; Daan, 1982; Pittendrigh, 1960). The circadian clock is the most extensively studied and best understood endogenous oscillator to this day. The core oscillator is a genomic complex of multiple transcription-translation feedback loops, modulated by sunlight as zeitgeber through degradation of a core protein and following interruption of the loop during the day (Golombek and Rosenstein, 2010). Meanwhile, the basic molecular mechanism of other rhythms, like the circalunar rhythm is yet to be identified.

1.3 Two unique biological rhythms of intertidal organisms

Not just the periodic diel occurrence of the low and high tides is determined by the lunar phase as described in section 1.1, but also the water level amplitude between them. This has strong implications on coastal areas periodically exposed and submerged by the tidal regime (intertidal zone). The angle and the gravitational-pull-direction of the moon in relation to the earth and the sun changes throughout the synodic month (Fig. 1B,C) (Palmer, 1995). The direction of the gravitational pull affects the tidal amplitude in the

way, that in the first and third quarter of the synodic month the gravitation of the moon and sun work in different directions to each other. This causes the lowest tidal amplitudes, the neap tides, which at low tide keep most of the intertidal zone submerged (Fig. 1B). In contrast, the highest tidal amplitude occurs when the gravitational pull of the sun and moon work in the same plane around the full moon and new moon days (Fig. 1C) (Neumann, 2014; Palmer, 1995). Along the European Atlantic coast there are low tides occur every 12.4 hours. Depending on various geophysical factors, total tidal amplitudes reach from a few centimeters to more than 10 meters, around full moon and new moon days ("spring tides": e.g. Roscoff, France with $>$ 10 m tidal amplitude; Fig. 2; Neumann (2014)). Two clock systems are known as a specific evolutionary adaptation to the tidal regime. One is the circatidal clock, as a direct response to the changing sea level and is described to oscillate in a 12.4 hour time frame (Neumann, 2014). As an example, the crustacean *Eurydice pulchra* exhibits characteristic tidal swimming behavior related to the local tidal regime as well as the circadian rhythm (Zhang et al., 2013). It was shown, that the circatidal and circadian rhythms can be dissociated by environmental and molecular manipulation of the circadian clock. This implied the presence of an independent circatidal clock (Zhang et al., 2013). The second clock system known in intertidal organisms are circalunar clocks. They are entrained to the moon-dependent zeitgebers which are moonlight, tidal turbulence (Neumann, 1966) and tidal temperature fluctuations (Kaiser, 2014; Neumann and Honegger, 1969).

Figure 2: Photographs of the intertidal zone at the Marine Station Biologique de Roscoff (Brittany, France).

Both pictures were taken by Nico Fuhrmann on 25.04.2018, the left at 15:21 (CEST), the right at 19:18 (CEST).

Circalunar clocks oscillate in a 14.77- (circasemilunar) or 29.54-day (circalunar) cycle (Naylor, 1982; Neumann, 2014; Palmer, 1995). In *Platynereis dumerilii*, a lunar-rhythmic marine Polychaeta, the transcription of circadian clock genes was chemically inhibited to test the presence of an independent circalunar clock (Zantke et al., 2013). Interestingly when no circadian oscillation was observed the circalunar rhythm was still functioning. Both studies in *Eurydice pulchra* and *Platynereis dumerilii* proved the independence of the circatidal and circalunar clocks to the circadian clock. However, both circatidal and circalunar clocks were so far studied based solely on behavior and their link to the circadian clock (Neumann, 1967; Zantke et al., 2013), but on the cellular and molecular level they remain unknown.

1.4 The circalunar clock model *Clunio marinus*

The marine non-biting midge *Clunio marinus* (Diptera: Chironomidae) is extensively studied for their lunar-rhythmic adult emergence patterns. Depending on food supply, seasonal temperature or substrate location in the intertidal, even under common garden conditions the lifecycle takes from 45 to 95 days (Neumann, 1966, 1986). The longest life stage are the larvae, which build tubes from surrounding substrate in regularly submerged rocky areas. After a few days of pupation the short lived adult insects emerge to mate and deposit the egg clutches on exposed substrate (Kaiser, 2014; Neumann, 1966, 1986). The emergence of adults is strictly tied to few hours right before one of the two spring low tides, which occur twice in the synodic month (see section 1.1). The circadian clock synchronizes the adult emergence to the location-specific daytimes of the spring low tides. A circalunar rhythm narrows the days of emergence down to hit only few days of the synodic month when the spring tides occur (Kaiser, 2014; Neumann, 1966, 1986). As part of the population synchronization to few specific days of the synodic month, the last larval instar is capable to arrest its development until the timepoint approaches to pupate and emerge (Neumann, 1986). As mentioned above, *C. marinus* is population-specific genetically adapted to their local tidal regime in lunar and diel emergence time (Kaiser, 2014; Kaiser et al., 2011, 2010, 2021; Neumann, 1966, 1986). Kaiser et al. (2011) were able to show evidence for the genetic control of both the circadian and the circalunar rhythms. Hybridization of two Atlantic *C. marinus* populations differing in circadian and circalunar emergence resulted in an intermediate emergence distribution for both rhythms in the first generation offspring. The emergence distribution of the back crosses shifted towards the emergence distribution of the parental population. Further investigations

found with quantitative trait loci mapping (QTL) two genomic regions linked to circadian and circalunar timing (Kaiser et al., 2016).

1.5 Lunar-arrhythmic *Clunio* populations in Northern European habitats

The majority of described and characterized European *C. marinus* populations show the characteristics described previously in section 1.4. Specifically their larval substrate in rocky areas, the circadian and circalunar emergence of the adult insects and the egg deposition in the exposed intertidal zone (Kaiser, 2014; Neumann, 1966, 1986). Populations with these described ecological characteristics will further be summarized as *"Atlantic ecotype"*. Previous studies identified *Clunio* populations in other European coastal areas which form two distinct lunar-arrhythmic ecotypes. The first *"Baltic ecotype"* inhabits rocky areas of the Baltic Sea where the tidal regime is largely neglectable. Due to the geophysical properties of the connection between the Baltic Sea and the Atlantic ocean the tidal amplitude is too small to expose large parts of the substrate (Endraß, 1976a; Neumann, 1966; Palmén and Lindeberg, 1959; Remmert, 1955). The *Baltic ecotype* exhibits a lunar-independent diurnal emergence throughout the emergence seasons. Individuals emerge after the sunset throughout dusk into the night (Endraß, 1976a; Heimbach, 1978; Neumann, 1966). Additionally, the *Baltic ecotype's* egg clutch properties and females egg deposition behavior differs from the *Atlantic ecotype*. The female lays the egg clutch in the open water by pushing the tip of the abdomen through the water surface. The released egg clutch is able to sink through the water body to the submerged larval substrate (Endraß, 1976b; Neumann, 1966). Egg clutches of the Atlantic ecotype are deposited on exposed intertidal substrate and stick to it (Neumann, 1966). When females lay them on the water surface, the egg clutches tend to swim and not sink. Investigations into the egg clutch properties identified an ecotype-specific difference in the surface structure of the gelatinous mass which encloses the eggs as likely cause for swimming or sinking abilities (Endraß, 1976b). Finally, the *Baltic ecotype's* larvae are also found in deeper substrates as the larvae of the *Atlantic ecotype* (Neumann, 1966).

The first population of the second lunar-arrhythmic ecotype was discovered and field-described in Tromsø (Norway) (Neumann and Honegger, 1969; Remmert, 1965). Populations of this *"Arctic ecotype"* emerge in the Atlantic intertidal zone when the water temperatures rise above 5°C. In high arctic latitudes, this condition is met only during few months in summer (Pflüger, 1973). Moonlight is an unreliable zeitgeber in the arc-

tic habitat, because the Midnight Sun is constantly present throughout their emergence season (Neumann and Honegger, 1969; Pflüger, 1973; Remmert, 1965). The egg clutch deposition and their properties do not differ from the *Atlantic ecotype*. But in contrast to the *Atlantic ecotype* a circalunar or circadian rhythm is not observable in the *Arctic ecotype's* adult emergence. The emergence of adults is linked to the larval habitat falling dry during the low tides twice per day (Neumann and Honegger, 1969; Pflüger, 1973; Remmert, 1965). Previous studies on the emergence of the *Arctic ecotype* show that the populations most likely do not exhibit a circatidal clock. Most likely an hourglass-timer-system is entrained by the temperature fluctuations and light intensity changes resulting from the tidal regime. This system is responsible for the observable exogenous circatidal rhythm (Pflüger, 1973; Pflüger and Neumann, 1971). All individuals, which finished their pupal development will emerge with the current low tide. Individuals in development will postpone the emergence until the following ebb is expected in 11-13 hours. Additionally, the *Arctic ecotype* changed the larval substrate from rock settled algae patches to sandy mudflats (Neumann and Honegger, 1969; Pflüger, 1973; Remmert, 1965).

An interesting case is a *Clunio* population which was discovered in the Kviturdvikpollen close to Bergen (Norway). Their circalunar emergence distribution overall would appear lunar-arrhythmic but with emergence peaks in the first and last quarter of the synodic month when the neap tides occur (Koskinen, 1968; Neumann and Honegger, 1969). First investigations into the circadian emergence pattern revealed occasionally the emergence start in the late afternoon and the general emergence after dusk (Neumann and Honegger, 1969). It was later shown by laboratory investigations of the field samples, that at the location both an *Atlantic* and *Baltic ecotype* population existed in sympatry, separated by their circadian and circalunar emergence timing (Heimbach, 1978). It was also shown that hybridization between the *Atlantic* and *Baltic ecotype* under laboratory conditions is possible and resulted in fertile offspring (Heimbach, 1978; Neumann, 1966). Despite the lack of hybrid observations in the field at the expected intermediate times, gene flow between both the ecotypes is a possible scenario. Therefore, this site is a valuable location to study the genomic maintenance and the potential genetic interactions of two ecotypes.

Finally, the habitats of both lunar-arrhythmic ecotypes were inaccessible for intertidal organisms during the last glacial maximum (LGM). While Northern Europe was covered by massive glaciers, the Baltic Sea resulted from the landscape changing force of the ice (Patton et al., 2017). After the glaciers retreated at the end of the LGM, the Norwegian

coast was accessible while the Baltic Sea remained a massive freshwater reservoir for a few thousand years. About 8,000 years ago, it was connected to the Atlantic Ocean and became a gradually brackish habitat with varying salinity concentrations. Additionally, Chironomidae larvae head capsules characterized as *Clunio* species were found in 8,000 years old sediment samples from the German Baltic Sea basin (Hofmann and Winn, 2000). This gives an additional approximate timeframe for the establishment of both the *Arctic* and the *Baltic ecotype*.

1.6 Research aims of the thesis

While *C. marinus* is an established model for circalunar rhythms and local adaptation, the focus of all studies with an aim on the genetic or genomic elements of these characteristics remained limited to the ecologically uniform populations of the *Atlantic ecotype* differing in circadian and circalunar timing (Kaiser et al., 2021). While closely related populations isolated by distance allows to identify the minute differences which maintain local adapted rhythmicity, potentially strongly conserved genes for the function of a putative circalunar clock may stay hidden. This thesis is taking a different route to the genomic mechanics of lunar-rhythmicity away from local adaptation of circadian and circalunar emergence distributions. Instead the focus switches to identify the genomic differences between the *Atlantic ecotype* and the above described lunar-arrhythmic ecotypes.

The first objective and foundation of the thesis was to find and sample populations described in literature for whole genome sequencing and to establish laboratory strains. Three lunar-arrhythmic populations and both sympatric populations from Bergen (see section 1.5) were collected during several field trips. By adding two previously established *Atlantic ecotypes* the whole genome set was composed out of 168 resequenced individual nuclear and mitochondrial genomes. The second objective will take a look at the mitochondrial genomes of the geographically distinct populations of the data set. Focusing on small fragments and their ability to reconstruct the mitochondrial biogeography in *Clunio*. The third objective is a broad investigation in the evolutionary history, genomic differentiations and detailed scans of ecotype-associated regions of the nuclear genome of the lunar-arrhythmic populations. The fourth and final objective is an investigation into possible links between quantitative trait loci (QTL) and lunar-rhythmic phenotypes in the sympatric Bergen populations using low numbers of polymorphic markers from amplicon sequencing data.

1) Biogeographic signals of mitochondrial haplotypes in geographically isolated populations.

Previous studies on the geographic differentiation among geographically isolated *Clunio* populations had troubles resolving the biogeography when using a standardized DNA fragment for species identification (DNA barcode) (Kaiser et al., 2010). My first published article investigates the biogeography of five isolated populations (120 individuals) using mitochondrial sequences from the new compiled genomic set. Creating haplotype networks for non-overlapping windows and statistical analysis, I evaluated each by the resolution of the biogeographic pattern in comparison to the reconstruction of the entire mitochondrial genome. The comprehensive genomic data set compiled during my thesis is an excellent example to test the effectiveness in reconstructing the biogeography from small DNA fragments in comparison to the entire mitochondrial genome. Furthermore, the analysis of the entire mitochondrial genome provided insight into the ancestral distribution of haplotypes and the evolution of distinct mitochondrial lineages in the studied *Clunio* populations.

2) Genetic signals of maintained ecological adaptations in isolated and sympatric ecotypes.

The second published article is strongly focused on the nuclear genome divergence between the lunar-rhythmic and lunar-arrhythmic ecotypes. By first analyzing the population structure and admixture present in all 168 samples and adding the mitochondrial diversification gathered from the previous article, I aim to resolve the historical scenario which resulted in *Baltic* and *Arctic ecotypes*. The second part of the article aims to identify genes involved in the formation of the *Atlantic* and the *Baltic ecotype*. The starting point were the sympatric populations for the identification of the most divergent genotypes between the two ecotypes. Crossing both under controlled conditions, revealed the heterogeneous polygenic nature of the lunar-arrhythmic trait. The resulting crossing families provide additional resources for further mapping of quantitative traits. The sampled genomic data of 72 individual genomes per ecotype provided another valuable resource, a window into the diversity of the young *Baltic ecotype*. While overall genomic differentiation is high between geographically distant populations, I was able to detect multiple genotype variants in one differentiated genomic region associated with the presence or absence of the lunar rhythm. These results and further investigations into the identities of affected genes are described in the second article.

3) Putative genetic markers linked to lunar rhythmicity are identified by QTL mapping. By creating a crossing family from laboratory strains of two sympatric populations, I was able to perform a QTL mapping experiment and identify genomic regions linked to lunar-rhythmicity. The final dataset comprised 237 F2 individuals and four to six genetic markers from each chromosome. The polymorphic markers were designed from differentiated and ecotype-associated genetic variants between the parental populations. Multiple phenotypes linked to lunar-rhythmicity were investigated. The analyses identified multiple loci across the genome with interactive, additive and individual effects. Overall, multiple identified QTLs and their varying effects indicate lunar-rhythmicity may be a multi locus trait with causing QTLs on independent chromosomes.

The published articles and additional manuscript of my thesis took the first look into this comprehensive compilation of genomic resource from ecologically divergent and recently evolved populations from a single species of marine Chironomidae. These results provide genetic candidates and possible links to the loss of circalunar rhythms in *C. marinus*. The circatidal timer strategy of the *Arctic ecotype* was not further investigated in the presented results because only one population of this ecotype was available. The generated data could still contain candidates which lead to the discovery of timer controlling genomic regions or genes. Finally, the migration to habitats of low salinity, oxygen stress or lacking of diel cycles within few generations demonstrates the ability to rapidly adapt from standing genetic variation when facing changes in climate. QTL mapping provided conclusive results in favor of the hypothesis, that lunar-rhythmicity could be affected by multiple loci across the genome.

1.7 References

Aschoff, J. (1981a). Chapter 1 A Survey on Biological Rhythms. In Aschoff, J., editor, *Handbook of Behavioral Neurobiology: Volume 4 Biological Rhythms*, pages 3–10. Plenum Press, New York.

Aschoff, J. (1981b). Chapter 6 Freerunning and Entrained Circadian Rhythms. In Aschoff, J., editor, *Handbook of Behavioral Neurobiology: Volume 4 Biological Rhythms*, pages 81–94. Plenum Press, New York.

Brown, Frank A, J. (1960). Response to Pervasive Geophysical Factors and the Biological Clock Problem. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 25, pages 57–71. Cold Spring Harbor Laboratory Press.

Daan, S. (1982). 2 Circadian rhythms in animals and plants. In Brady, J., editor, *SEBS 14: Biological Timekeeping*, pages 11–32. Cambridge University Press, Cambridge.

Endraß, U. (1976a). Physiologische Anpassungen eines marinen Insekts. I. Die zeitliche Steuerung der Entwicklung. *Marine Biology*, 34(4):361–368.

Endraß, U. (1976b). Physiologische Anpassungen eines marinen Insekts. II. Die Eigenschaften von schwimmenden und absinkenden Eigelegen. *Marine Biology*, 36(1):47–60.

Golombek, D. A. and Rosenstein, R. E. (2010). Physiology of Circadian Entrainment. *Physiological Reviews*, 90(3):1063–1102.

Heimbach, F. (1978). Sympatric Species, *Clunio marinus* Hal. and *Cl. balticus* n. sp.(Dipt., Chironomidae), Isolated by Differences in Diel Emergence Time. *Oecologia*, 32(2):195–202.

Hofmann, W. and Winn, K. (2000). The Littorina Transgression in the Western Baltic Sea as Indicated by Subfossil Chironomidae (Diptera) and Cladocera (Crustacea). *International Review of Hydrobiology*, 85:267–291.

Kaiser, T. S. (2014). Local Adaptations of Circalunar and Circadian Clocks: The Case of *Clunio marinus*. In Numata, H. and Helm, B., editors, *Annual, Lunar, and Tidal Clocks*, pages 121–141. Springer.

Kaiser, T. S., Neumann, D., and Heckel, D. G. (2011). Timing the tides: Genetic control of diurnal and lunar emergence times is correlated in the marine midge *Clunio marinus*. *BMC Genetics*, 12(1):1–12.

Kaiser, T. S., Neumann, D., Heckel, D. G., and Berendonk, T. U. (2010). Strong genetic differentiation and postglacial origin of populations in the marine midge *Clunio marinus* (Chironomidae, Diptera). *Molecular Ecology*, 19(14):2845–2857.

Kaiser, T. S., Poehn, B., Szkiba, D., Preussner, M., Sedlazeck, F. J., Zrim, A., Neumann, T., Nguyen, L.-T., Betancourt, A. J., Hummel, T., Vogel, H., Dorner, S., Heyd, F., von Haeseler, A., and Tessmar-Raible, K. (2016). The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature*, 540(7631):69–73.

Kaiser, T. S., von Haeseler, A., Tessmar-Raible, K., and Heckel, D. G. (2021). Timing strains of the marine insect *Clunio marinus* diverged and persist with gene flow. *Molecular Ecology*, 30(5):1264–1280.

Koskinen, R. (1968). Seasonal emergence of *Clunio marinus* Haliday (Dipt., Chironomidae) in Western Norway. In *Annales Zoologici Fennici*, volume 5, pages 71–75. Societas Biologica Fennica Vanamo.

Naylor, E. (1982). 3 Tidal and lunar rhythms in animals and plants. In Brady, J., editor, *SEBS 14: Biological Timekeeping*, pages 33–48. Cambridge University Press, Cambridge.

Neumann, D. (1966). Die lunare und tägliche Schlüpfperiodik der Mücke *Clunio* - Steuerung und Abstimmung auf die Gezeitenperiodik. *Zeitschrift für Vergleichende Physiologie*, 53(1):1–61.

Neumann, D. (1967). Genetic adaptation in emergence time of *Clunio* populations to different tidal conditions. *Helgoländer wissenschaftliche Meeresuntersuchungen*, 15(1):163–171.

Neumann, D. (1986). Chapter 1 Life cycle strategies of an intertidal midge between subtropic and arctic latitudes. In Taylor, F. and Karban, R., editors, *The Evolution of Insect Life Cycles*, pages 3–19. Springer.

Neumann, D. (2014). Timing in Tidal, Semilunar, and Lunar Rhythms. In Numata, H. and Helm, B., editors, *Annual, Lunar, and Tidal Clocks*, pages 3–24. Springer.

Neumann, D. and Honegger, H. W. (1969). Adaptations of the intertidal midge *Clunio* to arctic conditions. *Oecologia*, 3:1–13.

Palmén, E. and Lindeberg, B. (1959). The marine midge, *Clunio marinus* Hal.(Dipt., Chironomidae), found in brackish water in the northern Baltic. *Internationale Revue der gesamten Hydrobiologie und Hydrographie*, 44(1-4):383–394.

Palmer, J. D. (1974). 1 Introduction. In Palmer, J. D., editor, *Biological Clocks in Marine Organisms: The Control of Physiological and Behavioral Tidal Rhythms*, page 9. Wiley.

Palmer, J. D. (1995). 1. Introduction to Organismic Rhythms and Tidal Cycles. In Palmer, J. D., editor, *The Biological Rhythms and Clocks of Intertidal Animals*, pages 3–13. Oxford University Press, New York.

Patton, H., Hubbard, A., Andreassen, K., Auriac, A., Whitehouse, P. L., Stroeven, A. P., Shackleton, C., Winsborrow, M., Heyman, J., and Hall, A. M. (2017). Deglaciation of the Eurasian ice sheet complex. *Quaternary Science Reviews*, 169:148–172.

Pflüger, W. (1973). Die Sanduhrsteuerung der gezeitensynchronen Schlüpfrhythmik der Mücke *Clunio marinus* im arktischen Mittsommer (Hour-Glass Control of the Tidal Rhythm of *Clunio marinus* (Chironomidae) in Adaptation to Arctic Conditions). *Oecologia*, pages 113–150.

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

Pittendrigh, C. S. (1960). Circadian Rhythms and the Circadian Organization of Living Systems. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 25, pages 159–184. Cold Spring Harbor Laboratory Press.

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

Remmert, H. (1965). Über den Tagesrhythmus arktischer Tiere. *Zeitschrift für Morphologie und Ökologie der Tiere*, 55(2):142–160.

Zantke, J., Ishikawa-Fujiwara, T., Arboleda, E., Lohs, C., Schipany, K., Hallay, N., Straw, A. D., Todo, T., and Tessmar-Raible, K. (2013). Circadian and Circalunar Clock Interactions in a Marine Annelid. *Cell Reports*, 5(1):99–113.

Zhang, L., Hastings, M. H., Green, E. W., Tauber, E., Sladek, M., Webster, S. G., Kyriacou, C. P., and Wilcockson, D. C. (2013). Dissociation of Circadian and Circatidal Timekeeping in the Marine Crustacean *Eurydice pulchra*. *Current Biology*, 23(19):1863–1873.

2 The importance of DNA barcode choice in biogeographic analyses – a case study on marine midges of the genus *Clunio*

Published article; *Genome, 64*; pp. 242-252; 2021

DOI: https://doi.org/10.1139/gen-2019-0191

Authors: Nico Fuhrmann[1] and Tobias S. Kaiser[1]

Affiliation: [1]Max Planck Institute for Evolutionary Biology, Max Planck Research Group "Biological Clocks", August-Thienemann-Strasse 2, 24306 Plön, Germany.

2.1 Abstract

DNA barcodes are widely used for species identification and biogeographic studies. Here we compare the use of full mitochondrial genomes vs. DNA barcodes and various other mitochondrial DNA fragments for biogeographic and ecological analyses. Our empirical test dataset comprised 120 mitochondrial genomes from the genus *Clunio* (Diptera: Chironomidae), comprising five populations from two closely related species (*Clunio marinus* and *Clunio balticus*) and three ecotypes. From this dataset we extracted cytochrome oxidase *c* subunit I (COI) barcodes (658 bp) and we partitioned the mitochondrial genomes into non-overlapping windows of 750 bp or 1,500 bp respectively. We calculated haplotype networks as well as several diversity indices and compared them to those resulting from full mitochondrial genomes (15.4 kb). Full mitochondrial genomes indicate complete geographic isolation between populations, but do not allow conclusions on the separation of ecotypes or species. COI barcodes have comparatively few polymorphisms, ideal for species identification, but do not resolve geographic isolation. Many of the similarly sized 750 bp windows have higher nucleotide and haplotype diversity than COI barcodes, but still none of them resolve biogeography. Only when increasing the window size to 1,500 bp two windows resolve biogeography reasonably well. Our results suggest that the design and use of DNA barcodes in biogeographic studies must be carefully evaluated for each investigated species.

Keywords: DNA barcoding, mitochondrial genome, cytochrome oxidase I, haplotype network, diversity, ecotype

2.2 Introduction

DNA barcoding is a powerful and cost-effective tool for species identification (Adamowicz et al., 2019; Čandek and Kuntner, 2015; Hebert et al., 2003). The formally accepted DNA barcode for animals is a 658 bp fragment at the 5'-end of mitochondrial cytochrome oxidase *c* subunit I (COI; Hebert et al. (2003)). Today, the generation of DNA barcodes is so easily done that the availability of reference libraries for matching the generated DNA barcodes has become the major limiting factor in DNA barcoding (Čandek and Kuntner, 2015; Ekrem et al., 2018). The use of DNA barcodes was soon extended to other purposes such as ecological or biogeographic studies (Can et al., 2018; Janssen et al., 2019; Kress et al., 2015; Lalonde and Marcus, 2019). However, DNA barcodes have known limitations in resolving species in secondary contact, frequent hybridization, recent radiation and incomplete lineage sorting (Kress et al., 2015; Moritz and Cicero, 2004). The development of high-throughput sequencing (HTS) has further expanded the possibilities of DNA barcode generation, allowing for sequencing large bulk samples of many species or individuals simultaneously. One major development was the use of HTS-based DNA metabarcoding for analyzing local biodiversity and monitoring species (Cristescu, 2014; Taberlet et al., 2012). Other studies utilized HTS to sequence full mitochondrial genomes for better resolution of species' or biogeographic relationships (Delsuc et al., 2019; Paijmans et al., 2018). Studies in fish (Chen et al., 2019) and insects (Djoumad et al., 2017; Gómez-Rodríguez et al., 2017) have found severe discrepancies between the results obtained from COI barcodes and full mitochondrial genomes, the latter clearly giving better resolution. However, the possibility to obtain full mitochondrial genomes is limited by complex library preparation procedures (for specifically targeting the mitochondrial genome) or high sequencing costs (in case the mitochondrial genome is sequenced together with the much larger nuclear genome). In this study we explore if it is possible to circumvent the sequencing of full mitochondrial genomes by developing specific mitochondrial DNA fragments as markers for the given research question and species. Clearly, conservation and diversity of mitochondrial and nuclear genomes varies along the sequence, as well as between populations, ecotypes and species. In the light of that, it might be possible and maybe often also necessary to pick the best suited mitochondrial DNA fragment or set of fragments for a specific question. This will require some initial effort to find suitable fragments other than the established DNA barcode, but may subsequently allow for combining the low costs of sequence generation with sufficient resolution for biogeographic and ecological analyses.

We explore the limitations and possibilities of this approach in a case study on an empirical dataset for intertidal midges of the genus *Clunio* (Diptera: Chironomidae). In Northern Europe, *Clunio* has three major ecotypes (Fig. 3A). Populations of *Clunio marinus* (Fig. 3A, blue, *Atlantic ecotype*) require the larval substrates to be exposed by the low tide for oviposition. They have a very short adult life (2-3 hours), which is timed to the lowest low tides by a combination of circadian and circalunar clocks (Kaiser et al., 2016; Neumann, 1986). Above the Arctic Circle, polar day interferes with the use of circadian and circalunar clocks. This is why the *Arctic ecotype* of *C. marinus* (Fig. 3A, yellow) has resorted to a tidal timing mechanism that makes midges of this ecotype emerge during every low tide (Pflüger, 1973). In the Baltic Sea the lack of tides renders circalunar clocks obsolete as well. Baltic *Clunio* midges (Fig. 3A, green) are considered a separate species *Clunio balticus* (Heimbach, 1978). They do not rely on exposed larval substrates, but oviposit through the water surface so that the eggs sink to the submerged substrates, independent of the tides (Endraß, 1976). Baltic *Clunio* midges do not have a circalunar rhythm and emerge and reproduce every day at dusk (Heimbach, 1978).

We sampled 120 *Clunio* individuals from the three ecotypes and five geographic locations, performed whole-genome sequencing and extracted the full mitochondrial genomes. First, we calculated a haplotype network for the full mitochondrial genomes to obtain a robust assessment of biogeographic and ecological relationships. Compared to this gold standard, we then assessed the performance of COI barcodes in resolving these relationships. Finally, we explored if sequences of other parts of the mitochondrial genome, sequences of other sizes or combinations of sequences can more efficiently resolve biogeographic relationships than COI barcodes do.

2.3 Materials and Methods

***Clunio* sampling and determination.** We analyzed field-caught individuals for five populations (Fig. 3A; Tab. S1). For the lunar-rhythmic *Atlantic ecotype C. marinus* we collected 23 individuals from Port-en-Bessin (Por-1SL; France, 2009) and 25 from Helgoland (He-2SL1SL; Germany, 2005). For the lunar-arrhythmic *Arctic ecotype* of *C. marinus* we collected 24 samples from Tromsø (Tro-tAR; Norway, 2018). For *C. balticus*, the *Baltic ecotype*, we collected 24 individuals each from Sehlendorf (Seh-2AR; Germany, 2017) and Ar (Ar-2AR; Gotland, Sweden, 2018). As species and ecotypes are morphologically indistinguishable, we assigned all individuals based on their behavior and emergence timing in the field. From each sampling campaign we also established laboratory strains, which allowed us to confirm the respective ecotype and species assignments. There are no females in the sample, as they are wingless and immobile and therefore almost invisible in the field. All animals were directly collected in 99.98% ethanol and stored at -20°C.

DNA extraction and amplification. DNA was extracted from whole individuals with a salting out method (Reineke et al., 1998). DNA amount and purity were assessed with a NanoDrop ND-1000 Spectrophotometer (PREQLAB Biotechnologie GmbH). In order to allow or all intended analyses, genomic DNA was amplified using the REPLI-g Mini Kit (QIAGEN) with volume modifications (Tab. S3). Resulting products were checked for successful amplification by gel electrophoresis. Library preparation and sequencing were performed by the Max Planck Genome Centre (Cologne, Germany) according to standard protocols. All samples were subject to whole genome shotgun sequencing at 15-20 x coverage on an Illumina HiSeq3000 with 150 bp paired-end reads.

Sequence preprocessing, mapping and transformation. Sequencing reads were trimmed for adapters and base quality using Trimmomatic (Bolger et al., 2014) 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 (Zhang et al., 2014), setting the minimum assembled sequence length to 75 bp and a capping quality score of 20. The processed reads were mapped to the whole mitochondrial reference genome of *C. marinus* (ENA accession CVRI01023763.1; Kaiser et al. (2016)). The mitochondrial reference genome does not contain the control region. The reads were aligned with BWA-MEM (Li, 2013). From the mapped reads in SAM format we produced mi-

tochondrial genome sequences in FASTQ format with BCFtools, SAMtools and the perl script vcfutils.pl (Li, 2010, 2011; Li et al., 2009). Finally, we transformed the FASTQ file to FASTA and then PHYLIP format and combined all sequences into an alignment with a custom script. All ambiguous bases were set to N, as most tools cannot handle degenerate base codes and we also do not expect degenerate bases in the haploid mitochondrion. The resulting 120 full mitochondrial genomes are found in Supplementary Data 1. As all individuals' reads were aligned to the same reference sequence, the output sequences were automatically aligned. Before analysis, multiple state positions (i.e. positions with at least three different bases scored in different individuals) were identified using the R package 'Biostrings' 2.52.0 (Pagès et al., 2019) and excluded from the alignment. For our dataset, we excluded five positions with multiple states (Fig. 4-6: "∧"-marked positions at 5,353 bp, 5,919 bp, 8,705 bp, 10,955 bp and 11,660 bp).

From the corrected PHYLIP alignment we extracted the accepted COI barcode sequence based on the conserved primer pair LCO1490 and HCO2198 (Folmer et al., 1994). We also extracted non-overlapping 750 bp or 1,500 bp windows. The remaining last window of 428 bp was excluded from the analysis due to its non-comparable size and because it contained a gap. Extraction of the COI barcode and other sequence fragments was done with a custom script.

Data analysis. Mitochondrial haplotype networks were calculated using the Median-Joining algorithm (Bandelt et al., 1999) with Network 5.0.1.0 (fluxus-engineering.com). Nucleotide and haplotype diversity, as well as the number of haplotypes, were calculated with the R package 'pegas' 0.11 (Paradis, 2010). In order to assess how well the populations were separated in a haplotype network, we defined a "Population Separation Index" as the fraction of individuals having a haplotype that is not shared between populations.

Figure preparation. Figures were prepared in R (Crawley, 2007). The map of Europe (Fig. 3A) was generated using the packages 'ggplot2' (Wickham, 2016), 'ggrepel' (Slowikowski, 2020), 'map' (Brownrigg et al., 2018) and 'mapdata' (Brownrigg, 2018). The map was taken from the CIA World DataBank II (http://www.evl.uic.edu/pape/data/WDB/). Schematic circular mitochondria (Fig. 4; Fig. 6) were generated using the R package 'circlize' (Gu et al., 2014). Histograms (Fig. 5) were created with the R package 'shape' (Soetaert, 2018). Pearson correlation tests (Fig. S1) were calculated and plotted using the R package 'ggpubr' (Kassambara, 2017).

2.4 Results and Discussion

Biogeography of Northern European *Clunio* populations. We first produced a haplotype network for the full mitochondrial genomes (Fig. 3B). All haplotypes form perfectly isolated clusters according to geographic location. The position of each cluster in the network does not reflect the geographic distance between the populations, but all clusters diverge from two hypothetical haplotypes in the center of the network. This suggests a single common origin of all populations, but perfect geographic isolation between the maternally inherited mitochondrial genomes. This finding is congruent with a previous study on European populations of *C. marinus*, which employed mitochondrial sequences from the 3'-end of COI (i.e. not the standard COI barcode; Kaiser et al. (2010)). In this study mitochondrial haplotypes were also perfectly private for each location. These findings likely reflect the strong sexual dimorphism of the species: Females are wingless and basically immobile, so that we must assume that their dispersal is very limited. As a consequence, mitochondrial genomes are fully geographically isolated and divergent in sequence already at the population level, i.e. on small geographic scales. This interferes with making statements on the broader relationship of ecotypes or species. For example, the mitochondrial genomes of *C. balticus* are different from those of *C. marinus*, but due to the strong geographic isolation of the mitochondrial genomes, divergence between populations of the same species is as pronounced as divergence between the two species (Fig. 3B). In order to assess differentiation or gene flow between species or ecotypes, additional nuclear sequence data is required. The biology of the genus *Clunio* per se restricts the informativeness of its mitochondrial DNA sequence.

Within each population, whole mitochondrial genomes are generally diverse. In our sample of 23 to 25 individuals per population we rarely pick up the same mitochondrial genome twice (Fig. 3B). Usually, within a population, the individuals' mitochondrial genomes differ by 1 to 20 SNPs. Notably, in the Seh-2AR population there is one mitochondrial haplotype shared by five individuals that differs in more than 25 SNPs and diverges from the other haplotypes at the center of the network. This suggests the local coexistence of two divergent mitochondrial lineages in Seh-2AR, which both seem to date back to the origin of the Northern European radiation of *Clunio*.

Assessment of the cytochrome oxidase *c* subunit I barcode. From the full mitochondrial genomes, we extracted the COI barcode (658 bp) and assessed its performance compared to the full mitochondrial genomes. First, we checked the obtained COI bar-

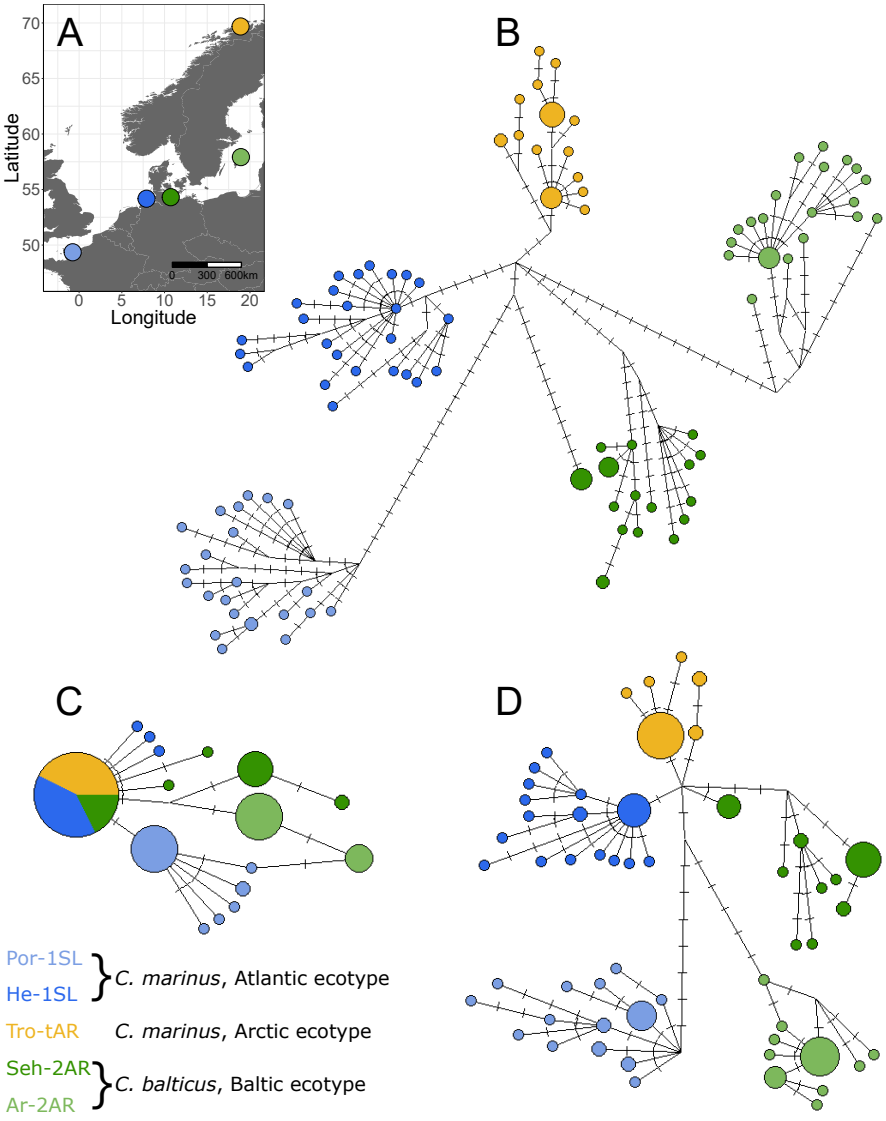

Figure 3: Mitochondrial haplotype networks for 120 individuals from Northern European *Clunio* populations.

(**A**) Geographic distribution of the sampled populations. (**B**) Haplotype network for full mitochondrial genomes (15.4 kb). Size of the circles is proportional to the number of individuals carrying a haplotype; color corresponds to population identity as in Fig. 3A. Short ticks on the lines connecting the haplotypes represent singe nucleotide polymorphism separating the haplotypes. (**C**) A network for the standard COI barcode (658 bp). (**D**) A network for a combination of two selected 1.5 kb windows, comprising the full COI and COIII as well as parts of ATP6 and ND3. The latter can resolve biogeographic signals as obtained from full mitochondrial genomes, whereas the standard COI barcode cannot.

codes against the Barcode Index Number (BIN) of the Barcode Of Life Data System v4 (Ratnasingham and Hebert, 2013). All barcodes matched the BIN of *Clunio marinus*. *Clunio balticus* is not represented in the database. Thus, in order to assess if species identification would be possible based on the COI barcode, we must consult the haplotype network (Fig. 3C). Given that the small COI barcode can of course not harbor as much variation as the entire mitochondrial genome, it is not surprising that the COI haplotype network collapses compared to the full mitochondrial genomes (compare Fig. 3B,C). There is a major COI haplotype shared by both species and all three ecotypes. Thus, species or ecotype identification is not possible. The signature of strong geographic isolation between populations which was picked up with the full mitochondrial genomes is also not visible any longer. Furthermore, the little variation left in the COI network suggests relationships that would not stand the test against full mitochondrial genomes, e.g. a connection of Ar-2AR and Por-1SL, which are again populations from different ecotypes and species. Thus, it becomes apparent that the COI barcode cannot be used for any statements on biogeography, ecotypes or the two closely related species.

COI barcode networks in comparison to 750 bp windows. Given the poor performance of COI barcodes, we wanted to find out if this is a general effect of the short length of the barcode, or if any other mitochondrial DNA sequence of similar length would be able to resolve the biogeographic relationships identified based on full mitochondrial genomes. We therefore partitioned the whole mitochondrial genome into 20 non-overlapping 750 bp windows and calculated haplotype networks for all windows (Fig. 4). The window from 1,501-2,250 bp comprises the entire COI barcode. All haplotype networks differ from each other and all are collapsed, usually showing one major haplotype that is shared by a majority of the populations (Fig. 4). Some windows have almost no sequence variation (e.g. 13,501-14,250 bp, 16S rDNA). Other windows show some resolution of geographic relationships, clearly better than the COI barcode (e.g. 2,251-3,000 bp, the 3' half of COI; 5,251-6,000 bp, COIII and ND3). However, none of these windows resolves the clear genetic isolation between all populations as observed in the analysis of full mitochondrial genomes. Again, there are misleading haplotype networks. For example, in the haplotype network for the window of 6,001-6,750 bp the haplotypes of the two different lunar-arrhythmic ecotypes (Seh-2AR and Tro-tAR; green and yellow) cluster together. This would suggest that the two arrhythmic ecotypes could be closely related and possibly derived from each other, inconsistent with the signal from the full mitochondrial genomes.

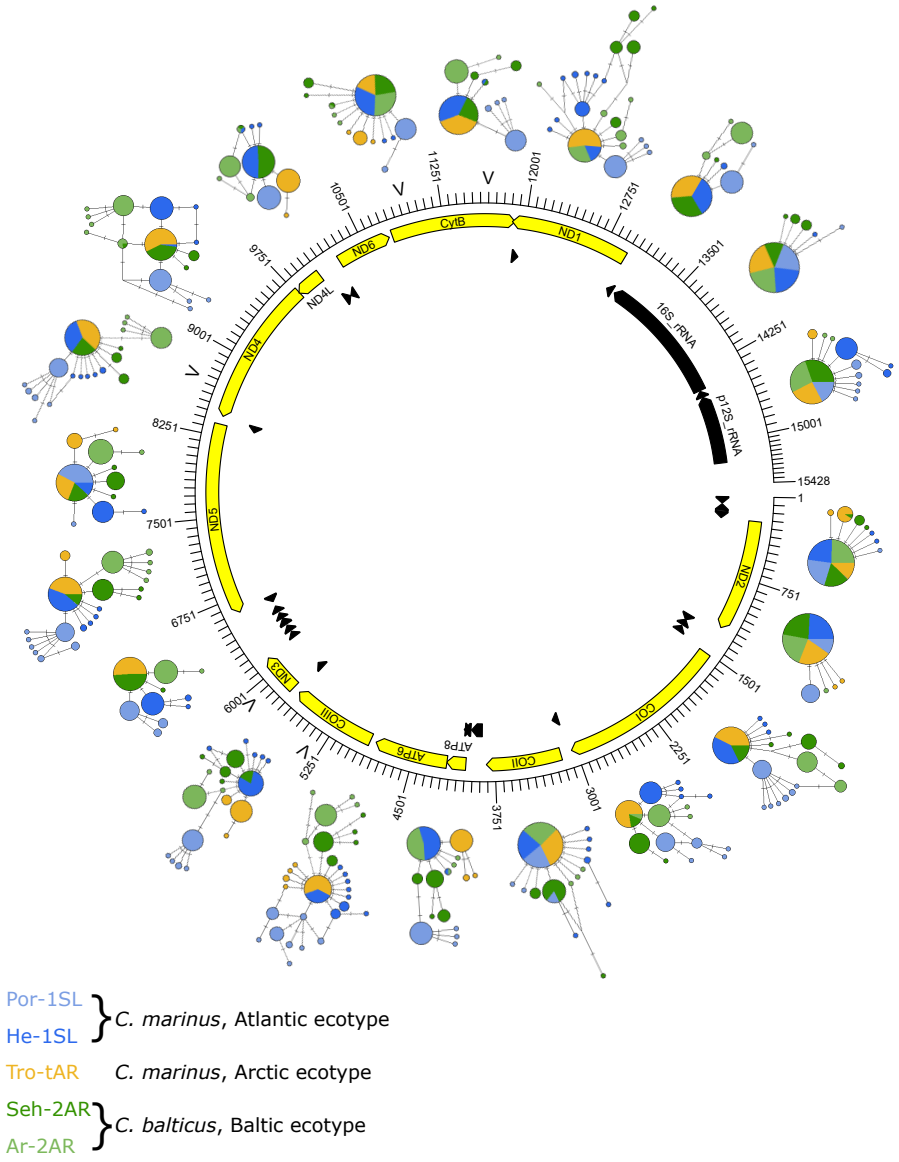

Figure 4: Mitochondrial haplotype networks for 750 bp windows along the *Clunio* mitochondrial genome.

The haplotype networks are based on the same data of 120 individuals as Fig. 3B, but only non-overlapping 750 bp windows along the mitochondrial were analyzed. Haplotype networks for each window are plotted around a schematic mitochondrion with genes (yellow arrows) and RNAs (black arrows). Colors of the haplotypes correspond to population identity as in Fig. 3A. The window of 1,501-2,250 bp comprises the entire COI barcode. Multiple state positions that were removed from analysis are indicated by small arrows on the schematic mitochondrion ("∧"; see 2.3).

In order to more systematically assess which mitochondrial windows are better suited for resolving the known complete geographic isolation between populations, we calculated a Population Separation Index (PSI; Fig. 5A). As clearly sequence variation is the limiting factor for resolution, we compared the PSI to nucleotide diversity (π), haplotype diversity and the total number of haplotypes for each window (Fig. 5B-D). All measures of diversity are correlated with the PSI (Fig. S1). The correlation is strongest for haplotype diversity, suggesting it can be used as a predictor for finding more suitable mitochondrial sequence windows for the design of diagnostic mitochondrial fragments. For example, the COI barcode has a haplotype diversity of 0.74 and windows with haplotype diversity > 0.8 show better geographic resolution of populations than the COI barcode (Fig. 4; Fig. 5A,C).

1,500 bp windows can resolve biogeography in the genus *Clunio*. Finally, we tested if larger mitochondrial fragments could give sufficient resolution for separating the populations in our sample. We partitioned the whole mitochondrial genome into 10 nonoverlapping 1,500 bp windows. This size was chosen because 1,500 bp are the limit of what can be sequenced on a Sanger sequencer if PCR amplicons are sequenced from both ends. Increased sequence length clearly increases haplotype diversity (compare Fig. 5C,H) and the resolution of the resulting haplotype networks (Fig. 6). There is one window in which all populations have separate haplotypes (4,501-6,000 bp, ATP6, COIII and ND3). For several other windows all populations but one pair are separated (e.g. 1,501-3,000 bp, COI; 9,001-10,500 bp, ND4, ND4L and ND6). Again, haplotype diversity is the best predictor of population separation (Fig. 5F-I; Fig. S1).

One imperfection that remains also with the 1,500 bp windows is that the center of the haplotype networks always is occupied by existing haplotypes, indicating that there are mitochondrial genomes that lack mutations relative to the ancestral haplotype. If the corresponding individuals are from a single population, as they are for window 4,510-6,000 bp, this would suggest that this population is ancestral to the others. This conclusion conflicts with the results from the full mitochondrial genomes. By combining the DNA fragments from window 1,501-3,000 bp (also containing the COI barcode) with window 4,510-6,000 bp, the resulting haplotype network (Fig. 3D) has no haplotype in its center and shows the same biogeographic relationships as the full mitochondrial genomes. This illustrates that a combination of two carefully picked diagnostic mitochondrial DNA fragments can give as much resolution as the sequencing of full mitochondrial genomes.

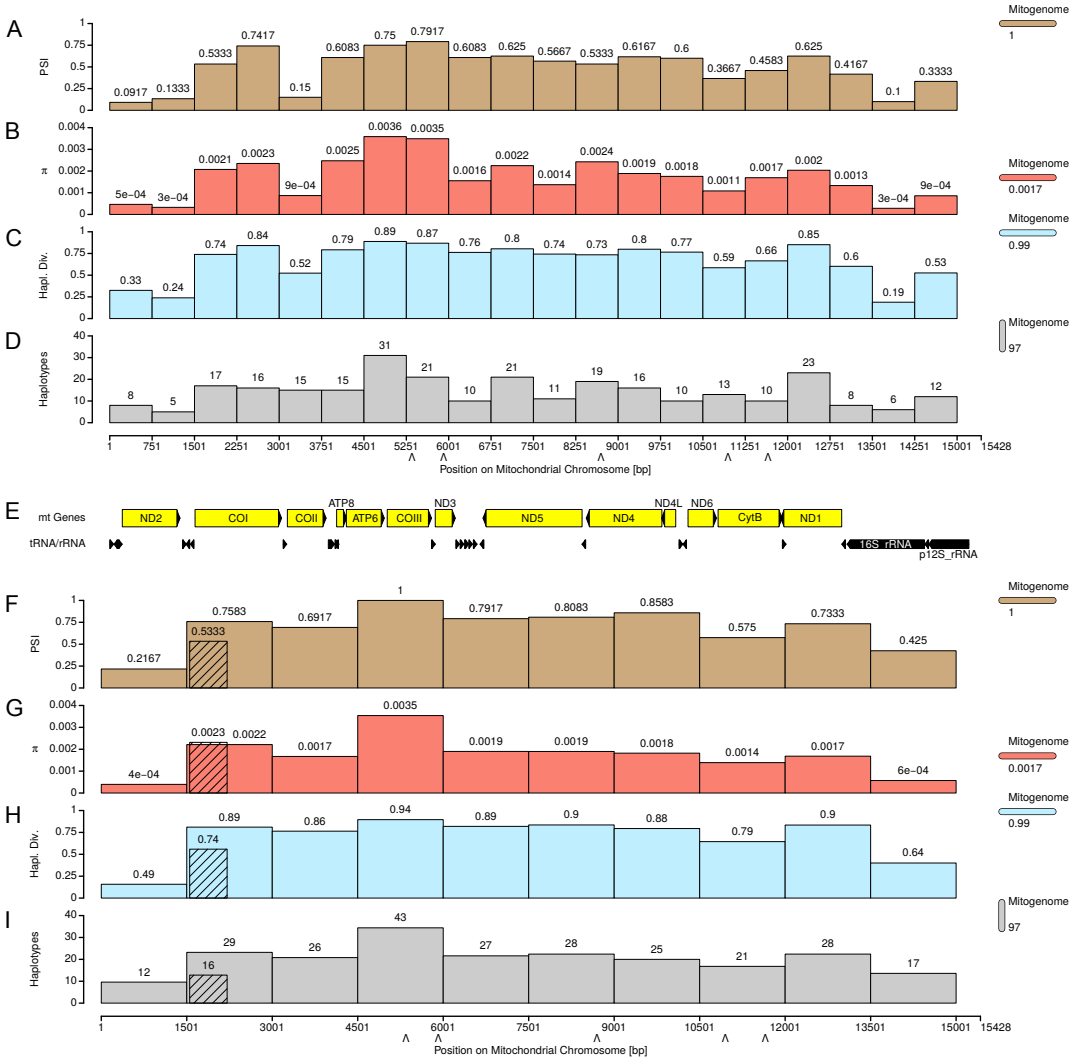

Figure 5: Population separation and measures of diversity along the mitochondrial genome in windows of 750 bp (A-D) and 1,500 bp (F-I).

Population separation and diversity are plotted relative to a schematic mitochondrial genome (**E**) with position of genes (yellow arrows) and RNAs (black arrows). Population Separation Index (PSI), i.e. the fraction of individuals carrying haplotypes that are not shared between populations, is plotted in light brown for 750 bp windows (**A**) and 1,500 bp windows (**F**). Nucleotide diversity (π) is given in red for 750 bp windows (**B**) and 1,500 bp windows (**G**). Haplotype Diversity is given in light blue for 750 bp windows (**C**) and 1,500 bp windows (**H**). The number of haplotypes is given in light gray for 750 bp windows (**D**) and 1,500 bp windows (**I**). In panels A-D the standard COI barcode is fully comprised in the window from 1,501-2,250 bp, whereas in panels F-I the COI barcode is represented by a hatched bar. On the right-hand side values for the full mitochondrial genome are given for comparison. Multiple state positions that were removed from analysis are indicated by small arrows on the schematic mitochondrion ("∧"; see 2.3).

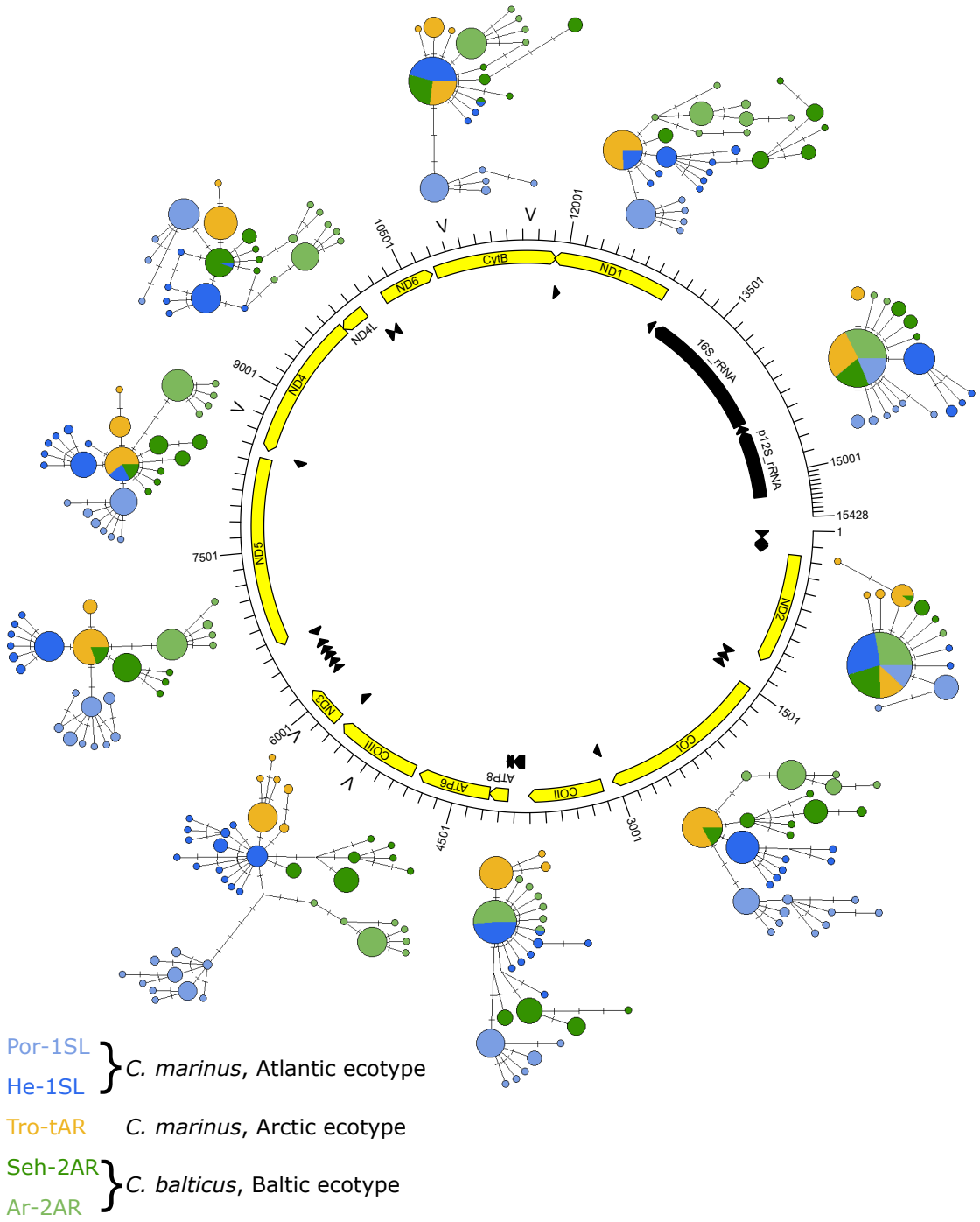

Figure 6: Mitochondrial haplotype networks for 1,500 bp windows along the *Clunio* mitochondrial genome.

The haplotype networks are based on the same data of 120 individuals as Fig. 3B, but only non-overlapping 1,500 bp windows along the mitochondrial were analyzed. Haplotype networks for each window are plotted around a schematic mitochondrion with genes (yellow arrows) and RNAs (black arrows). Colors of the haplotypes correspond to population identity as in Fig. 3A. Multiple state positions that were removed from analysis are indicated by small arrows on the schematic mitochondrion ("∧"; see 2.3).

2.5 Conclusions

We assessed the use of COI barcodes and other mitochondrial fragments in ecological and biogeographic studies based on a comprehensive set of 120 complete mitochondrial genomes from two species of the genus *Clunio*. These mitochondrial genomes carry a strong signal of complete geographic isolation of *Clunio* populations, likely due to the fact that *Clunio* females are wingless and therefore very limited in their dispersal. This signal severely limits all other conclusions that can be drawn from mitochondrial sequences. In particular, an assessment of the differentiation of ecotypes or species clearly requires additional nuclear sequence data. Our results underline that any statement based on mitochondrial sequence data must be carefully evaluated in the light of the given species' biology. Further, we systematically assessed if short mitochondrial sequences of different length and in different positions of the mitochondrion can capture the strong geographic signal obtained from full mitochondrial genomes. None of the short sequences gave sufficient resolution and some were severely misleading. These results illustrate the known caveats of drawing ecological, evolutionary or biogeographic conclusions from a single short DNA sequence.

For the case of *Clunio* in Northern Europe, a combination of two 1,500 bp sequences was able to recapitulate the full picture obtained from mitochondrial genomes, suggesting it can pay off to make an initial investment into genetic marker development for a specific question. After generating a limited set of whole mitochondrial (and possibly nuclear) genomes, suitable sequence markers can be identified based on diversity indices or haplotype networks. In our study a combination of two large sequences was required. Combining several sequences is costly when using Sanger sequencing, but easily feasible with HTS, where multiplex PCRs for many indexed individuals can be sequenced in a single run. The sequence length limitations imposed by Illumina sequencing, still the standard HTS technology for DNA metabarcoding, can be overcome in two ways: Long sequences can be subdivided into several small ones, requiring higher PCR multiplexing. Alternatively, sequencing libraries can be prepared by treating non-indexed large PCR products as if they were DNA samples. Tn5-based library preparation is available non-commercially (Hennig et al., 2018), reducing its costs but requiring molecular laboratory infrastructure for expressing and purifying the enzyme. Clearly, developing sequence markers requires more effort than obtaining standard DNA barcodes. At a certain point it may be more efficient to move to low-cost SNP genotyping (e.g. ddRAD-seq; Peterson et al. (2012)), or even whole genome sequencing.

In the light of the trade-off between investment and analytical power, the interesting open question becomes: Is it necessary to develop new sequence markers for all species and all questions? Or are there regions of the mitochondrion that generally work better as high-resolution markers than others? For *Clunio* only the window around ATP6 and ND3 separated all populations. ATP6 has also proven a good marker in certain moths (Djoumad et al., 2017), birds (Kerr, 2011) and fish (Perdices and Doadrio, 2001). A second promising window in *Clunio* contained the full COI gene, which would be particularly convenient, as it includes the standard COI barcode used for species identification in animals. Finally, as a high degree of genetic variation increases resolution for biogeographic studies, the control region of the mitochondrion is another candidate marker (Zhang and Hewitt, 1997). The control region was not assessed for *Clunio*, as the reference mitochondrial genome contains a gap in this position due to unresolved tandem repeats, immediately highlighting possible problems in assessing this region. Eventually, only a thorough evaluation of whole mitochondrial genome datasets for other species will reveal if there is a common "magic barcode" for biogeographic and ecological studies in animals.

2.6 Acknowledgements

For field work we obtained logistic support from the Ar Research Station (Uppsala University) and Even Jørgensen (The Arctic University of Norway, Tromsø). Sequencing was performed at the Max Planck Genome Center (Cologne) with financial support from the Max Planck Society. This work was supported through funds of the Max Planck Research Group *Biological Clocks*.

2.7 References

Adamowicz, S. J., Boatwright, J. S., Chain, F., Fisher, B. L., Hogg, I. D., Leese, F., Lijtmaer, D. A., Mwale, M., Naaum, A. M., Pochon, X., Steinke, D., Wilson, J.-J., Wood, S., Xu, J., Xu, S., Zhou, X., and van der Bank, M. (2019). Trends in DNA barcoding and metabarcoding. *Genome*, 62(3):v–viii.

Bandelt, H.-J., Forster, P., and Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, 16(1):37–48.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Brownrigg, R. (2018). mapdata: Extra Map Databases. R package version 2.3.0.

Brownrigg, R., Minka, T. P., and Deckmyn, A. (2018). maps: Draw Geographical Maps. R package version 3.3.0.

Can, F., Ulaşlı, B., and Hausmann, A. (2018). An Integrative Approach to Understand the Biogeography, Taxonomy and Ecology of the Macroheteroceran Fauna of the Amanos Mountains in Southern Turkey. *Journal of the Entomological Research Society*, 20(2):91–101.

Čandek, K. and Kuntner, M. (2015). DNA barcoding gap: reliable species identification over morphological and geographical scales. *Molecular Ecology Resources*, 15(2):268–277.

Chen, W., Yang, J., Li, Y., and Li, X. (2019). Exploring taxonomic diversity and biogeography of the family Nemacheilinae (Cypriniformes). *Ecology and Evolution*, 9(18):10343–10353.

Crawley, M. J. (2007). *The R Book*. Wiley.

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, 29(10):566–571.

Delsuc, F., Kuch, M., Gibb, G. C., Karpinski, E., Hackenberger, D., Szpak, P., Martínez, J. G., Mead, J. I., McDonald, H. G., MacPhee, R. D., Billet, G., Hautier, L., and Poinar, H. N. (2019). Ancient Mitogenomes Reveal the Evolutionary History and Biogeography of Sloths. *Current Biology*, 29(12):2031–2042.

Djoumad, A., Nisole, A., Zahiri, R., Freschi, L., Picq, S., Gundersen-Rindal, D. E., Sparks, M. E., Dewar, K., Stewart, D., Maaroufi, H., Levesque, R. C., Hamelin, R. C., and Cusson, M. (2017). Comparative analysis of mitochondrial genomes of geographic variants of the gypsy moth, *Lymantria dispar*, reveals a previously undescribed genotypic entity. *Scientific Reports*, 7(1):1–12.

Ekrem, T., Stur, E., Orton, M. G., and Adamowicz, S. J. (2018). DNA barcode data reveal biogeographic trends in Arctic non-biting midges. *Genome*, 61(11):787–796.

Endraß, U. (1976). Physiologische Anpassungen eines marinen Insekts. II. Die Eigenschaften von schwimmenden und absinkenden Eigelegen. *Marine Biology*, 36(1):47–60.

Folmer, O., Hoeh, W., Black, M., and Vrijenhoek, R. (1994). Conserved primers for PCR amplification of mitochondrial DNA from different invertebrate phyla. *Molecular Marine Biology and Biotechnology*, 3:294–299.

Gómez-Rodríguez, C., Timmermans, M. J., Crampton-Platt, A., and Vogler, A. P. (2017). Intraspecific genetic variation in complex assemblages from mitochondrial metagenomics: comparison with DNA barcodes. *Methods in Ecology and Evolution*, 8(2):248–256.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). *circlize* implements and enhances circular visualization in R. *Bioinformatics*, 30(19):2811–2812.

Hebert, P. D., Cywinska, A., Ball, S. L., and Dewaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1512):313–321.

Heimbach, F. (1978). Sympatric Species, *Clunio marinus* Hal. and *Cl. balticus* n. sp.(Dipt., Chironomidae), Isolated by Differences in Diel Emergence Time. *Oecologia*, 32(2):195–202.

Hennig, B. P., Velten, L., Racke, I., Tu, C. S., Thoms, M., Rybin, V., Besir, H., Remans, K., and Steinmetz, L. M. (2018). Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol. *G3: Genes, Genomes, Genetics*, 8(1):79–89.

Janssen, A., Stuckas, H., Vink, A., and Arbizu, P. M. (2019). Biogeography and population structure of predominant macrofaunal taxa (Annelida and Isopoda) in abyssal polymetallic nodule fields: implications for conservation and management. *Marine Biodiversity*, 49(6):2641–2658.

Kaiser, T. S., Neumann, D., Heckel, D. G., and Berendonk, T. U. (2010). Strong genetic differentiation and postglacial origin of populations in the marine midge *Clunio marinus* (Chironomidae, Diptera). *Molecular Ecology*, 19(14):2845–2857.

Kaiser, T. S., Poehn, B., Szkiba, D., Preussner, M., Sedlazeck, F. J., Zrim, A., Neumann, T., Nguyen, L.-T., Betancourt, A. J., Hummel, T., Vogel, H., Dorner, S., Heyd, F., von Haeseler, A., and Tessmar-Raible, K. (2016). The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature*, 540(7631):69–73.

Kassambara, A. (2017). ggpubr:"ggplot2" based publication ready plots. R package version 0.2.3.999.

Kerr, K. C. (2011). Searching for evidence of selection in avian DNA barcodes. *Molecular Ecology Resources*, 11(6):1045–1055.

Kress, W. J., García-Robledo, C., Uriarte, M., and Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evolution*, 30(1):25–35.

Lalonde, M. M. and Marcus, J. M. (2019). Getting western: biogeographical analysis of morphological variation, mitochondrial haplotypes and nuclear markers reveals cryptic species and hybrid zones in the *Junonia* butterflies of the American southwest and Mexico. *Systematic Entomology*, 44(3):465–489.

Li, H. (2010). Mathematical notes on samtools algorithms. *Computer Science*.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Moritz, C. and Cicero, C. (2004). DNA Barcoding: Promise and Pitfalls. *PLOS Biology*, 2(10).

Neumann, D. (1986). Life Cycle Strategies of an Intertidal Midge Between Subtropic and Arctic Latitudes. In Taylor, F. and Karban, R., editors, *The Evolution of Insect Life Cycles*, pages 3–19. Springer.

Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2019). Biostrings: efficient manipulation of biological strings. R package version 2.52.0.

Paijmans, J. L., Barlow, A., Förster, D. W., Henneberger, K., Meyer, M., Nickel, B., Nagel, D., Havmøller, R. W., Baryshnikov, G. F., Joger, U., Rosendahl, W., and Hofreiter, M. (2018). Historical biogeography of the leopard (*Panthera pardus*) and its extinct Eurasian populations. *BMC Evolutionary Biology*, 18(1):156.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, 26(3):419–420.

Perdices, A. and Doadrio, I. (2001). The Molecular Systematics and Biogeography of the European Cobitids Based on Mitochondrial DNA Sequences. *Molecular Phylogenetics and Evolution*, 19(3):468–478.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE*, 7(5).

Pflüger, W. (1973). Hour-Glass Control of Tidal Rhythm of *Clunio marinus* (Chironomidae) in Adaptation to Arctic Conditions. *Oecologia*, 11(2):113–150.

Ratnasingham, S. and Hebert, P. D. N. (2013). A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLOS ONE*, 8(7).

Reineke, A., Karlovsky, P., and Zebitz, C. P. W. (1998). Preparation and purification of DNA from insects for AFLP analysis. *Insect Molecular Biology*, 7(1):95–99.

Slowikowski, K. (2020). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.8.2.

Soetaert, K. (2018). shape: Functions for plotting graphical shapes, colors. R package version 1.4.4.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8):2045–2050.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Zhang, D.-X. and Hewitt, G. M. (1997). Insect mitochondrial control region: A review of its structure, evolution and usefulness in evolutionary studies. *Biochemical Systematics and Ecology*, 25(2):99–120.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5):614–620.

3 Polygenic adaptation from standing genetic variation allows rapid ecotype formation

Published preprint; *bioRxiv*; 2021

DOI: https://doi.org/10.1101/2021.04.16.440113

Authors: Nico Fuhrmann[1], Celine Prakash[1] and Tobias S. Kaiser[1]

Affiliation: [1]Max Planck Institute for Evolutionary Biology, Max Planck Research Group "Biological Clocks", August-Thienemann-Strasse 2, 24306 Plön, Germany.

3.1 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

3.2 Introduction

Understanding the processes underlying local adaptation and ecotype formation is a vital theme in evolutionary ecology (Kawecki and Ebert, 2004; Savolainen et al., 2013), but also increasingly important for conservation of biodiversity in the face of climate change and deterioration of natural habitats (Hoffmann et al., 2015). Adaptation in reproductive timing is at particular risk, as under rising temperatures it can be severely mismatched with the environment (Thackeray et al., 2016). 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 (Savolainen et al., 2013; Yuan and Stinchcombe, 2020). 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 (Neumann, 2014). Their molecular basis is unknown (Andreatta and Tessmar-Raible, 2020), 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 (Kaiser, 2014). 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 (Kaiser, 2014; Kaiser et al., 2011; Neumann, 1967). Analysis of these timing adaptations gave first insights into the genetic underpinnings of circalunar clocks (Kaiser and Heckel, 2012; Kaiser et al., 2016).

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 (Endraß, 1976a; Palmén and Lindeberg, 1959; Remmert, 1955) and in the high Arctic (Neumann and Honegger, 1969; Pflüger and Neumann, 1971) (see Fig. 7 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 (Endraß, 1976b; Remmert, 1955). Reproduction of the *Baltic ecotype* happens every day precisely at dusk under control of a circadian clock (Heimbach, 1978). There is no detectable circalunar rhythm (Endraß, 1976a). 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 (Heimbach, 1978). Therefore, the *Baltic ecotype* is considered a separate species – *C. balticus*. However, *C. balticus* and *C. marinus* can be successfully interbred in the laboratory (Heimbach, 1978). In the high Arctic there are normal tides and the *Arctic ecotype* of *C. marinus* is found in intertidal habitats (Neumann and Honegger, 1969). 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. It does not show circalunar or circadian rhythms (Pflüger and Neumann, 1971).

The geological history of Northern Europe (Patton et al., 2017) and subfossil *Clunio* head capsules in Baltic Sea sediment cores (Hofmann and Winn, 2000) 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.

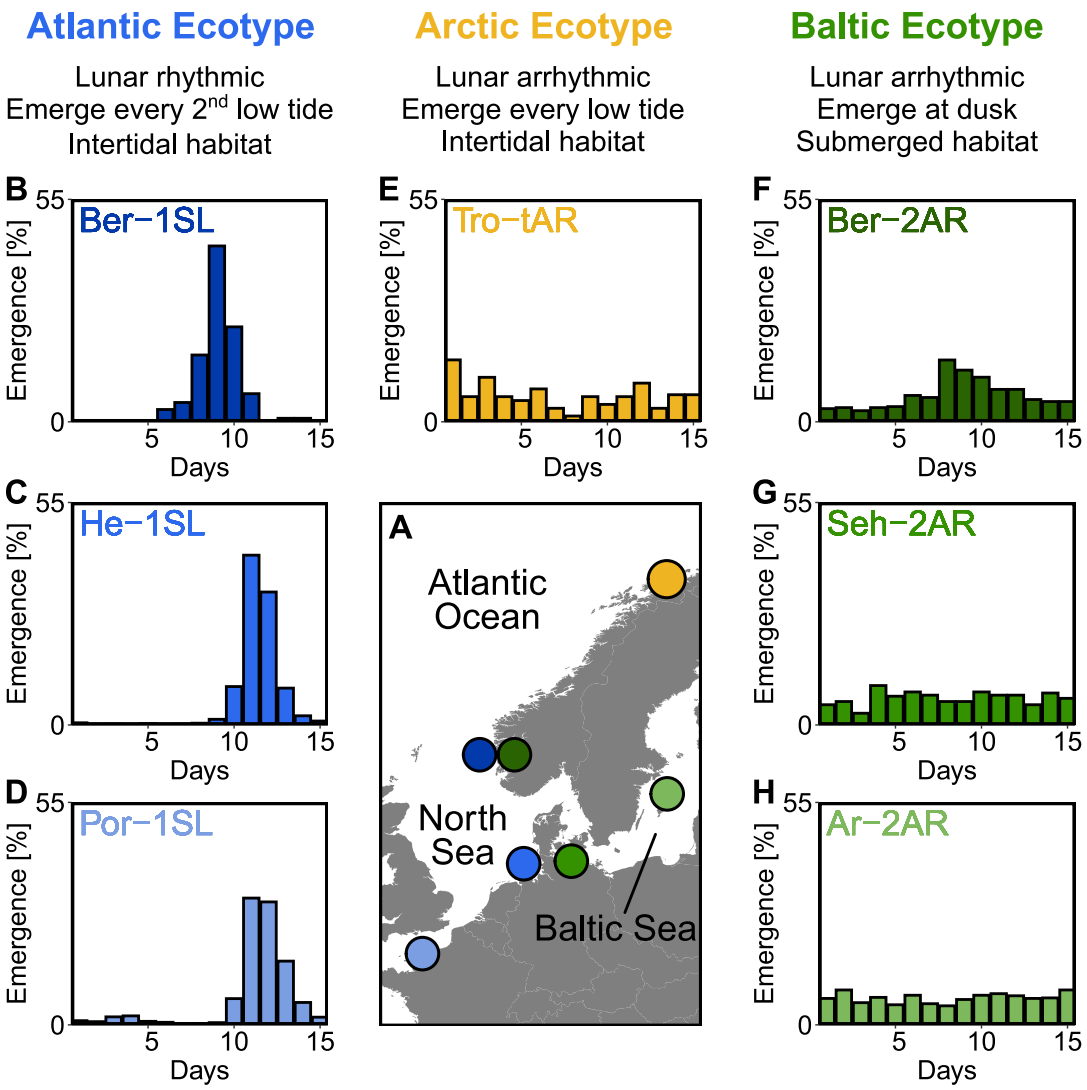

Figure 7: Northern European ecotypes of *Clunio* and their lunar rhythms. The *Atlantic*, *Arctic* and *Baltic ecotypes* of *Clunio* differ mainly in their lunar rhythms (**B-H**), circadian rhythms (Fig. S2), as well as their habitat and the resulting oviposition behavior (see Supplementary Note 5.3). (**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.

3.3 Results

***Clunio* ecotypes and their lunar rhythms.** Starting from field work in Northern Europe (Fig. 7A), we established one laboratory strain of the *Arctic ecotype* from Tromsø (Norway, Tro-tAR; see 3.5 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. 7B-H), their circadian rhythm (Fig. S2B-H) and their oviposition behavior (for details see Supplementary Note 5.3). The *Baltic ecotype* from Bergen (Ber-2AR, Fig. 7F) 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 (Fig. S3), 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 fieldcaught 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. 8A-B) and testing for genetic admixture (Fig. 8C). We also constructed a haplotype network of complete mitochondrial genomes (Fig. 8D). 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. 8A-B, Fig. S4). Mitochondrial haplotypes are not shared and are highly divergent between geographic sites (Fig. 8D). In ADMIXTURE, the optimal number of genetic groups is six (Fig. S5), corresponding to the number of geographic sites, and there is basically no mixing between the six clusters (Fig. 8C; K = 6).

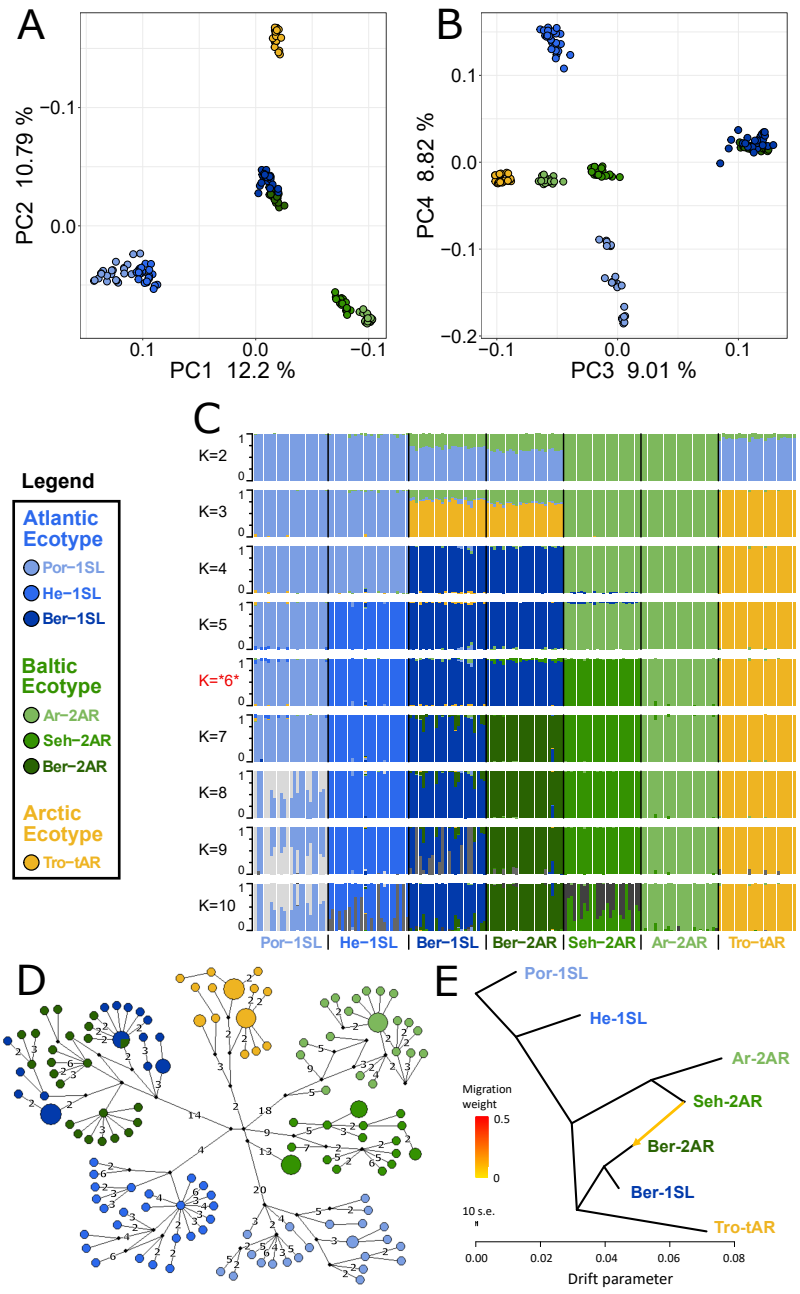

Figure 8: Genetic structure and evolutionary history of Northern European *Clunio* ecotypes.

(**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.

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. 8A-B) and they are the only populations that share mitochondrial haplotypes (Fig. 8D). 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. 8C; 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 (Heimbach, 1978). Given that genetic differentiation rather corresponds to geography than to ecotype (Fig. 8), we consider all three ecotypes part of a single species, *C. marinus*.

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. 8D). In line with this, 34% of polymorphic sites are polymorphic in all seven populations and 93% are polymorphic in at least two populations (Fig. S6). 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. 8A), suggests that *Clunio* expanded into the high Arctic and into the Baltic Sea independently. Congruently, nucleotide diversity significantly decreases towards both expansion fronts (Fig. S7; Tab. S4). 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. 8C). Congruently, TreeMix detects introgression from Seh-2AR into Ber-2AR (Fig. 8E), but no other introgression events (Fig. S8). 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. 8C, 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. 9A; Fig. S9; Supplementary Data 2). The tree topology obtained for the entire genome (Fig. S10) only dominates in few genomic windows (Fig. 9A, black bars "Orig."), while usually one or several other topologies account for more than 75% of the tree topologies (Fig. 9A, grey bars "Misc."). Hardly ever do all combinations of six individuals follow a single population tree topology (Fig. 9A, 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 both Bergen populations (Fig. 9A, yellow bars "Intr."; all topologies grouping Por-1SL and He-1SL vs. Ber-1SL, 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. 9A, red bars "Ecol.").

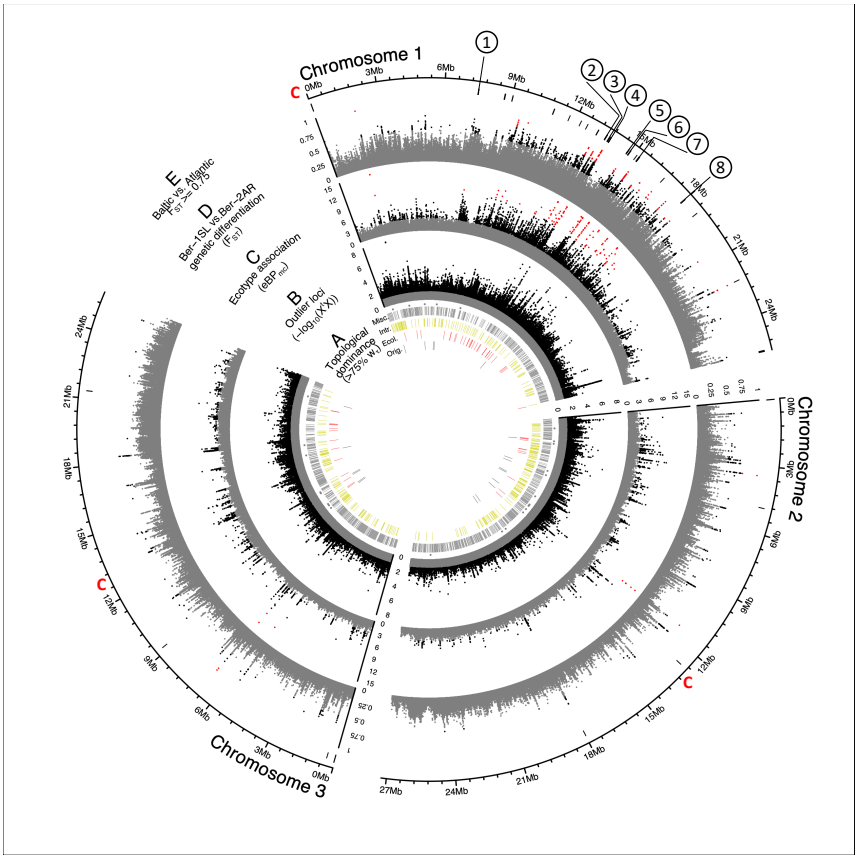

Figure 9: Genome screens for haplotype sharing and genotype-ecotype associations.

(**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 ($\omega_\tau > 75\%$). Most windows are not dominated by the consensus population topology ("Orig."; Fig. S10), 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 $F_{ST} \geq 0.75$ in the *Baltic* vs. *Atlantic ecotypes*. Circled numbers mark the location of the eight most differentiated loci (see Fig. 11). Centromeres of the chromosomes are marked by a red "C".

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 Ω matrix, Fig. S11A-B). Such variants may indicate local adaptation. At the same time, we tested for association of these variants with ecotype, as implemented in BayPass (Gautier, 2015). Overly differentiated variants (X^tX statistic; Fig. 9B) and ecotype-associated variants ecotype (eBP_{mc}; Fig. 9C) 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 (Fig. S11D-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 (Fig. S12; Fig. S13). Pairwise differentiation between Ber-1SL and Ber-2AR (Fig. 9D) 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. 9E; Fig. S14), 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. 9E). Genetic divergence (d_{xy}), nucleotide diversity (π) and local linkage disequilibrium (r^2) of the two Bergen populations do not show marked differences along or between chromosomes (Fig. S15). The cluster of ecotype-associated variants on chromosome 1 overlaps with three large blocks of long-range linkage disequilibrium (LD; Fig. S16).

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 (Fig. S17). 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. 9D). Taken together, numerous genomic loci – inside and outside the cluster on chromosome 1 – are associated with eco-

logical 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; Fig. S18A) 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) (Kaiser et al., 2016), 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; Fig. S18B) 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 3; 15,193 gene models). Based on this confidence gene set, we then assessed the location of variants relative to genes, as well as the resulting mutational effects (SnpEff (Cingolani et al., 2012); Fig. S19; statistics in Tab. S5). 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 4; Supplementary Data 5). We transferred GO terms from other species to the *Clunio* confidence annotations based on gene orthology (5,393 genes; see section 3.5 and Supplementary Data 6). GO term enrichment analysis suggests that ecological adaptation prominently involves the circadian clock, supported by three of the top four GO terms (Fig. 10A). 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. 10B;

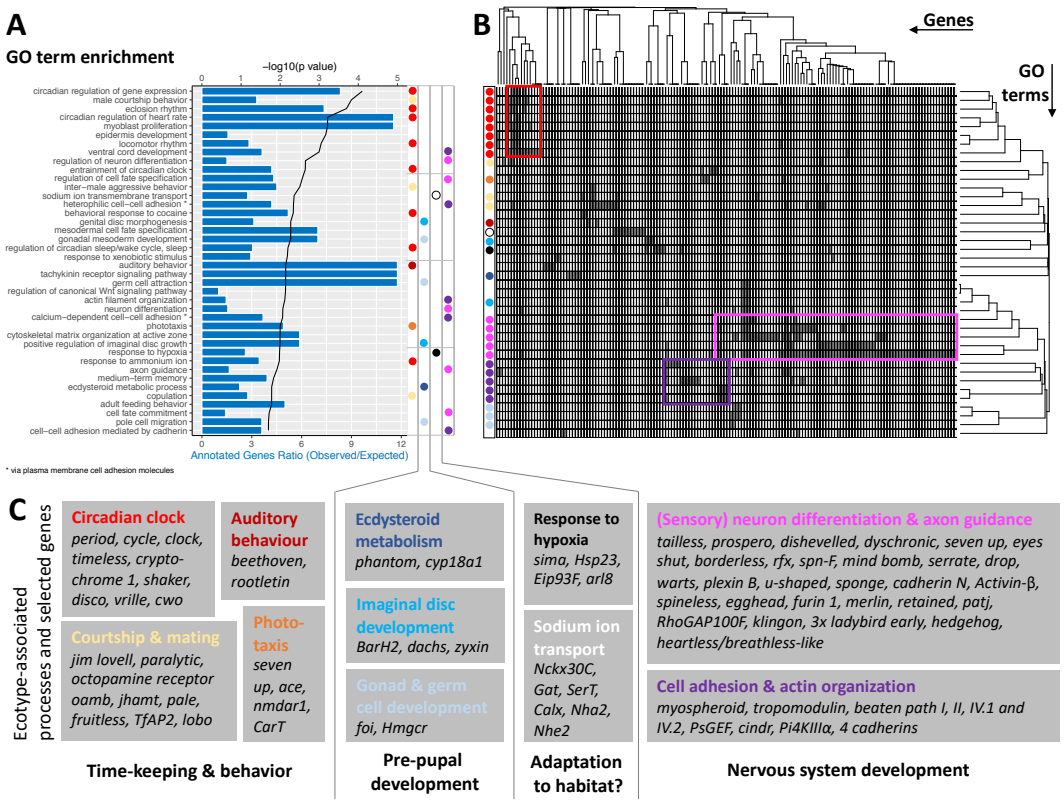

Figure 10: GO term analysis of ecotype associated SNPs.

(**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 Data 7). (**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.

Supplementary Data 7). We individually confirmed their gene annotations and associated GO terms. Clustering the resulting table by genes and GO terms reveals two dominant signatures (Fig. 10B). 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. 10B,C). As a second strong signal, almost half of the genes are annotated with biological processes involved in nervous system development (Fig. 10B,C). GO term enrichment is also found for ecdysteroid metabolism, imaginal disc development and gonad development (Fig. 10). 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 (Neumann, 2014) 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 6). 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. 10). 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 8; Supplementary Data 9). Additionally, we only considered genes with variants that are strongly differentiated between the ecotypes ($F_{ST} \geq 0.75$, compare Fig. 9E), leaving thirteen genes in eight distinct genomic regions (Fig. 11A; num-

bered in Fig. 9E). Two of these regions contain two genes each with no homology outside *C. marinus* (indicated by "NA", Fig. 11A), 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*) (Velarde et al., 2005), the transcription factor *longitudinals lacking* (*lola*; in fruit fly involved in axon guidance (Crowner et al., 2002) and photoreceptor determination (Zheng and Carthew, 2008))and the nuclear receptor *tailless* (*tll*; in fruit fly involved in development of brain and eye (Suzuki and Saigo, 2000)), underscoring that ecotype characteristics might involve differential light sensitivity.

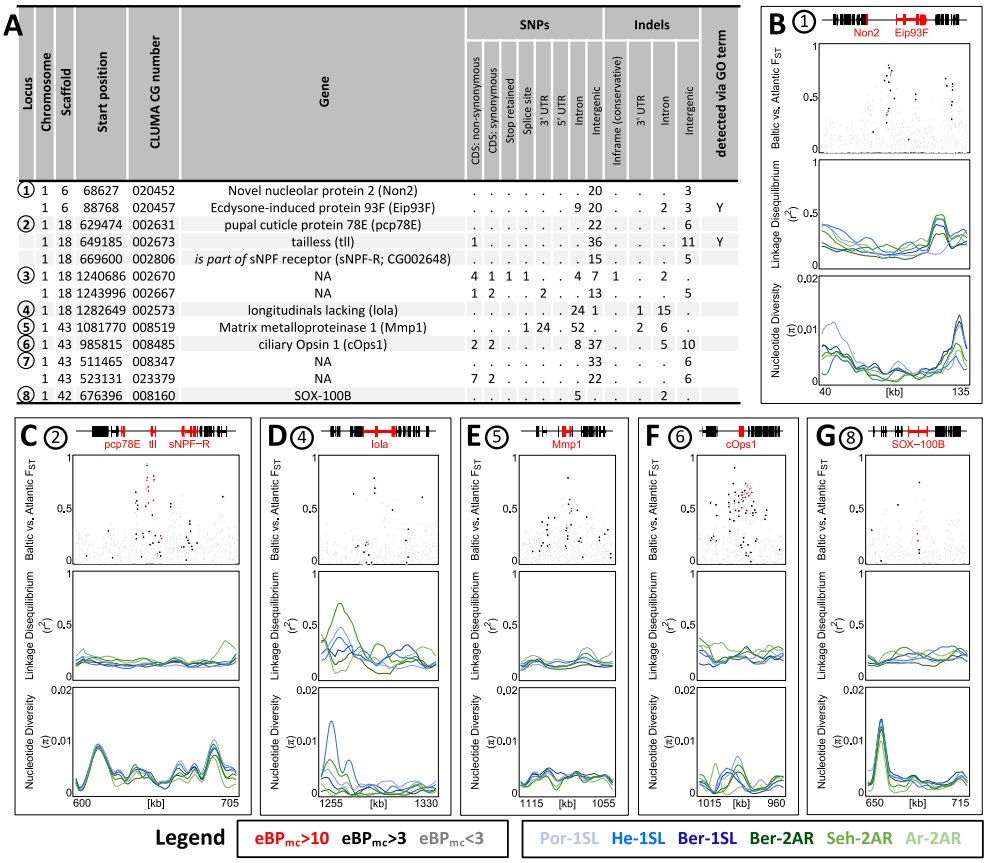

Figure 11: The 13 most differentiated ecotype-associated genes.

(**A**) Loci with highly ecotype-associated variants were selected based on $eBP_{mc} > 3$ 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 (π). 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. 7 and Fig. 8. The are no strong signatures of selection.

Interestingly, *tll* also affects development of the neuroendocrine centres involved in ecdysteroid production and adult emergence (de Velasco et al., 2007). 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 (Selcho et al., 2017). 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*) (Depetris-Chauvin et al., 2014). 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 (Nanda et al., 2009) and the *ecdysone-induced protein 93F* is involved in response to hypoxia in flies (Lee et al., 2008), but was recently found to also affect reproductive cycles in mosquitoes (Wang et al., 2021). In summary, only two out of the top 13 ecotype-associated genes were comprised in the top 40 GO terms (Fig. 11A). Nevertheless, all major biological processes detected in GO term analysis (Fig. 10) are also reflected in the strongly ecotype-associated loci (Fig. 11), 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. 11B-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. 11B-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.

3.4 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 (Chan et al., 2010; Kingman et al., 2020; Verta and Jones, 2019) or mimicry in *Heliconius* butterflies (Edelman et al., 2019). 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 (Kaiser, 2014; Kaiser et al., 2016, 2021; Neumann, 1967). 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 (Pavlidis et al., 2012). 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. (Zhang et al., 2021)). But QTL mapping with recently developed refined statistical algorithms for detection of polygenic signals may hold some promise (Wellenreuther and Hansson, 2016). Based on our genomic comparison of lunar-rhythmic and lunar-arrhythmic ecotypes, we propose three not mutually exclu-

sive 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 (Neumann and Spindler, 1991). 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 (Bünning and Müller, 1961) and such a mechanism has been experimentally confirmed in the midge *Pontomyia oceana* (Soong and Chang, 2012). 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 (Neumann, 2014). 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 (Kaiser, 2014). 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 (Joron et al., 2011) or reproductive morphs of the ruff (Küpper et al., 2016). 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 (Kaiser et al., 2016) 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.

3.5 Methods

Nomenclature of ecotypes. We expanded the existing naming convention of *C. marinus* timing types (Kaiser et al., 2021) 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* (Kaiser et al., 2021)).

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 (Tab. S2). 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 (Tab. S2). 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 (Kaiser et al., 2010, 2016, 2021), 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 (Neumann, 1966) 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) (Neumann, 1978; Neumann and Heimbach, 1979).

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 5.3). Daily emergence times were recorded in 1 h intervals by direct observation (Seh-2AR, Ar-2AR) or with the help of a fraction collector (Honegger, 1977) (Ber-1SL, Ber-2AR, Tro-tAR, Por-1SL, He-1SL; Fig. S2). 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 (Neumann, 1983), emergence data for Por-1SL was taken from a manuscript in preparation (Brišević et al.).

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 (Reineke et al., 1998) and amplified using the REPLI-g Mini Kit (QIAGEN) according to the manufacturer's protocol with volume modifications (Tab. S3). 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 (Bolger et al., 2014) 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 (Zhang et al., 2014), 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 (Li, 2013) to the nuclear reference genome (Kaiser et al., 2016) (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 (Li et al., 2009).

Read group information was added with the *AddOrReplaceReadGroups.jar* v.1.74 script from the Picard toolkit (http://picard.sourceforge.net/) (DePristo et al., 2011).

For the nuclear genome, SNPs and insertion-deletion (indel) genotypes were called using GATK v.3.8-0-ge9d806836 (McKenna et al., 2010). 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 (Danecek et al., 2011) 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 Fuhrmann and Kaiser (2021).

Population genomic analyses. Mitochondrial haplotype networks were calculated using the Median-Joining algorithm (Bandelt et al., 1999) 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 (Chang et al., 2015) 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 (Alexander and Lange, 2011), 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 (Ravinet, 2018) and running TreeMix v.1.13 (Pickrell and Pritchard, 2012) 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 (π) was calculated for SNPs with VCFtools '- -window-pi'. For the genome-wide average, calculations were repeated with 200 kb 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'. Pre-

liminary tests showed that local LD decays within a few hundred base pairs (Fig. S20). 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 Delmore et al. (2015).

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 (Ortiz, 2019) 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 (Nguyen et al., 2015) 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 (Martin and Van Belleghem, 2017) and the parameter '- -method complete'.

Association analysis. Population-based association between genetic variants (SNPs and indels) and ecotype, as well as environmental variables (Tab. S6) was assessed in BayPass v.2.2 (Gautier, 2015). Allele counts were obtained with VCFtools option '- -counts'. Analyzed covariates were ecotype, sea surface salinity (obtained from Hordoir et al. (2019)) and average water temperature of the year 2020 (obtained from weather-atlas.com, accessed 27.04.2020; 16:38), as given in Tab. S6. 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 (Kaiser et al., 2016) 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 Kaiser et al. (2016). 2) Overlap of either at least 20% with mapped transcript data or 40% with mapped protein data, as reported in Kaiser et al. (2016). 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 (Cingolani et al., 2012) (build 2020-04-15 22:26, non-default parameter '-ud 0'). Gene Ontology (GO) terms were annotated with emapper-2.0.1. (Huerta-Cepas et al., 2017) from the eggNOG 5.0 database (Huerta-Cepas et al., 2019), using DIAMOND (Buchfink et al., 2015), 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 (Alexa and Rahnenführer, 2010) (version 2.42.0, R version 4.0.3).

Figure preparation. Figures were prepared in R (Crawley, 2007). Data were handled with the 'data.table' (Dowle et al., 2020) and 'plyr' (Wickham, 2011) packages. The map of Europe was generated using the packages 'ggplot2' (Wickham, 2016) and 'ggrepel' (Slowikowski, 2020), 'maps' (Brownrigg et al., 2018) and 'mapdata' (Brownrigg, 2018). 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' (Gu et al., 2014). Multiple plots were combined in R using the package 'Rmisc' (Hope, 2013). The graphical editing of the whole genome phylogeny was done in Archeopteryx (http://www.phylosoft.org/archaeopteryx; Han and Zmasek (2009)). Final figure combination and graphical editing of the raw plot files was done in *Inkscape*. Neighbor Joining trees of the Ω statistic distances from BayPass were created with the R package 'ape' (Paradis and Schliep, 2019). In all plots the order and orientation of scaffolds within the chromosomes follows the published genetic linkage map (Kaiser et al., 2016).

3.6 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 Tobias S. Kaiser.

3.7 Author contributions

Nico Fuhrmann performed field work, laboratory work, population genomic analyses and association analysis, prepared the figures and participated in drafting the manuscript. Celine Prakash analyzed SNP effects and GO term enrichment and participated in figure preparation. Tobias S. Kaiser conceived, designed and supervised the study, participated in data analysis and figure preparation and wrote the manuscript. All authors approved of the final manuscript.

3.8 References

Alexa, A. and Rahnenführer, J. (2010). topGO: enrichment analysis for gene ontology. R package version 4.0.3.

Alexander, D. H. and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC bioinformatics*, 12(1):246.

Andreatta, G. and Tessmar-Raible, K. (2020). The Still Dark Side of the Moon: Molecular Mechanisms of Lunar-Controlled Rhythms and Clocks. *Journal of Molecular Biology*, 432(12):3525–3546.

Bandelt, H.-J., Forster, P., and Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, 16(1):37–48.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Brownrigg, R. (2018). mapdata: Extra Map Databases. R package version 2.3.0.

Brownrigg, R., Minka, T. P., and Deckmyn, A. (2018). maps: Draw Geographical Maps. R package version 3.3.0.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60.

Bünning, E. and Müller, D. (1961). Wie messen Organismen lunare Zyklen? *Zeitschrift für Naturforschung*, 16 b(6):391–395.

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Shapiro, M. D., Brady, S. D., Southwick, A. M., Absher, D. M., Grimwood, J., Schmutz, J., Myers, R. M., Petrov, D., Jónsson, B., Schluter, D., Bell, M. A., and Kingsley, D. M. (2010). Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a *Pitx1* Enhancer. *Science*, 327(5963):302–305.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):s13742–015.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). 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(2):80–92.

Crawley, M. J. (2007). *The R Book*. Wiley.

Crowner, D., Madden, K., Goeke, S., and Giniger, E. (2002). Lola regulates midline crossing of CNS axons in *Drosophila*. *Development*, 129(6):1317–1325.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and Group, . G. P. A. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

de Velasco, B., Erclik, T., Shy, D., Sclafani, J., Lipshitz, H., McInnes, R., and Hartenstein, V. (2007). Specification and development of the pars intercerebralis and pars lateralis, neuroendocrine command centers in the *Drosophila* brain. *Developmental biology*, 302(1):309–323.

Delmore, K. E., Hübner, S., Kane, N. C., Schuster, R., Andrew, R. L., Câmara, F., Guigó, R., and Irwin, D. E. (2015). Genomic analysis of a migratory divide reveals candidate genes for migration and implicates selective sweeps in generating islands of differentiation. *Molecular Ecology*, 24(8):1873–1888.

Depetris-Chauvin, A., Fernández-Gamba, Á., Gorostiza, E. A., Herrero, A., Castaño, E. M., and Ceriani, M. F. (2014). Mmp1 Processing of the PDF Neuropeptide Regulates Circadian Structural Plasticity of Pacemaker Neurons. *PLOS Genetics*, 10(10):e1004700.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491.

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., et al. (2020). Package 'data.table'. Extension of 'data.frame'. R package version 1.13.2.

Edelman, N. B., Frandsen, P. B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., García-Accinelli, G., Van Belleghem, S. M., Patterson, N., Neafsey, D. E., Challis, R., Kumar, S., Moreira, G. R. P., Salazar, C., Chouteau, M., Counterman, B. A., Papa, R., Blaxter, M., Reed, R. D., Dasmahapatra, K. K., Kronforst, M., Joron, M., Jiggins, C. D., McMillan, W. O., Di Palma, F., Blumberg, A. J., Wakeley, J., Jaffe, D., and Mallet, J. (2019). Genomic architecture and introgression shape a butterfly radiation. *Science*, 366(6465):594–599.

Endraß, U. (1976a). Physiologische Anpassungen eines marinen Insekts. I. Die zeitliche Steuerung der Entwicklung. *Marine Biology*, 34(4):361–368.

Endraß, U. (1976b). Physiologische Anpassungen eines marinen Insekts. II. Die Eigenschaften von schwimmenden und absinkenden Eigelegen. *Marine Biology*, 36(1):47–60.

Fuhrmann, N. and Kaiser, T. S. (2021). The importance of DNA barcode choice in biogeographic analyses — a case study on marine midges of the genus *Clunio*. *Genome*, 64(3):242–252.

Gautier, M. (2015). Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics*, 201(4):1555–1579.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). *circlize* implements and enhances circular visualization in R. *Bioinformatics*, 30(19):2811–2812.

Han, M. V. and Zmasek, C. M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, 10(1):1–6.

Heimbach, F. (1978). Sympatric Species, *Clunio marinus* Hal. and *Cl. balticus* n. sp.(Dipt., Chironomidae), Isolated by Differences in Diel Emergence Time. *Oecologia*, 32(2):195–202.

Hoffmann, A., Griffin, P., Dillon, S., Catullo, R., Rane, R., Byrne, M., Jordan, R., Oakeshott, J., Weeks, A., Joseph, L., Lockhart, P., Borevitz, J., and Sgrò, C. (2015). A framework for incorporating evolutionary genomics into biodiversity conservation and management. *Climate Change Responses*, 2(1):1–24.

Hofmann, W. and Winn, K. (2000). The Littorina Transgression in the Western Baltic Sea as Indicated by Subfossil Chironomidae (Diptera) and Cladocera (Crustacea). *International Review of Hydrobiology*, 85:267–291.

Honegger, H. W. (1977). An Automatic Device for the Investigation of the Rhythmic Emergence Pattern of *Clunio marinus*. *International Journal of Chronobiology*, 4:217–221.

Hope, R. M. (2013). Rmisc: Ryan miscellaneous. R package version 1.5.

Hordoir, R., Axell, L., Höglund, A., Dieterich, C., Fransner, F., Gröger, M., Liu, Y., Pemberton, P., Schimanke, S., Andersson, H., Ljungemyr, P., Nygren, P., Falahat, S., Nord, A., Jönsson, A., Lake, I., Döös, K., Hieronymus, M., Dietze, H., Löptien, U., Kuznetsov, I., Westerlund, A., Tuomi, L., and Haapala, J. (2019). Nemo-Nordic 1.0: a NEMO-based ocean model for the Baltic and North seas–research and operational applications. *Geoscientific Model Development*, 12(1):363–386.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., and Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, 34(8):2115–2122.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., and Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314.

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., Whibley, A., Becuwe, M., Baxter, S. W., Ferguson, L., Wilkinson, P. A., Salazar, C., Davidson, C., Clark, R., Quail, M. A., Beasley, H., Glithero, R., Lloyd, C., Sims, S., Jones, M. C., Rogers, J., Jiggins, C. D., and ffrench Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, 477(7363):203–206.

Kaiser, T. S. (2014). Local Adaptations of Circalunar and Circadian Clocks: The Case of *Clunio marinus*. In Numata, H. and Helm, B., editors, *Annual, Lunar, and Tidal Clocks*, pages 121–141. Springer.

Kaiser, T. S. and Heckel, D. G. (2012). Genetic Architecture of Local Adaptation in Lunar and Diurnal Emergence Times of the Marine Midge *Clunio marinus* (Chironomidae, Diptera). *PLoS ONE*, 7(2):e32092.

Kaiser, T. S., Neumann, D., and Heckel, D. G. (2011). Timing the tides: Genetic control of diurnal and lunar emergence times is correlated in the marine midge *Clunio marinus*. *BMC Genetics*, 12(1):1–12.

Kaiser, T. S., Neumann, D., Heckel, D. G., and Berendonk, T. U. (2010). Strong genetic differentiation and postglacial origin of populations in the marine midge *Clunio marinus* (Chironomidae, Diptera). *Molecular Ecology*, 19(14):2845–2857.

Kaiser, T. S., Poehn, B., Szkiba, D., Preussner, M., Sedlazeck, F. J., Zrim, A., Neumann, T., Nguyen, L.-T., Betancourt, A. J., Hummel, T., Vogel, H., Dorner, S., Heyd, F., von Haeseler, A., and Tessmar-Raible, K. (2016). The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature*, 540(7631):69–73.

Kaiser, T. S., von Haeseler, A., Tessmar-Raible, K., and Heckel, D. G. (2021). Timing strains of the marine insect *Clunio marinus* diverged and persist with gene flow. *Molecular Ecology*, 30(5):1264–1280.

Kawecki, T. J. and Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters*, 7(12):1225–1241.

Kingman, G. A. R., Vyas, D. N., Jones, F. C., Brady, S. D., Chen, H. I., Reid, K., Millhaven, M., Bertino, T. S., Aguirre, W. E., Heins, D. C., von Hippel, F. A., Park, P. J., Kirch, M., Absher, D. M., Myers, R. M., Di Palma, F., Bell, M. A., Kingsley, D. M., and Veeramah, K. R. (2020). Predicting future from past: The genomic basis of recurrent and rapid stickleback evolution. *bioRxiv*.

Küpper, C., Stocks, M., Risse, J. E., Dos Remedios, N., Farrell, L. L., McRae, S. B., Morgan, T. C., Karlionova, N., Pinchuk, P., Verkuil, Y. I., Kitaysky, A. S., Wingfield, J. C., Piersma, T., Zeng, K., Slate, J., Blaxter, M., Lank, D. B., and Burke, T. (2016). A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics*, 48(1):79–83.

Lee, S.-J., Feldman, R., and O'Farrell, P. H. (2008). 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(10):4051–4061.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Martin, S. H. and Van Belleghem, S. M. (2017). Exploring Evolutionary Relationships Across the Genome Using Topology Weighting. *Genetics*, 206(1):429–438.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

Nanda, S., DeFalco, T., Loh, S. H. Y., Phochanukul, N., Camara, N., Van Doren, M., and

Russell, S. (2009). *Sox100B*, a *Drosophila* Group E Sox-domain Gene, Is Required for Somatic Testis Differentiation. *Sexual Development*, 3(1):26–37.

Neumann, D. (1966). Die lunare und tägliche Schlüpfperiodik der Mücke *Clunio* - Steuerung und Abstimmung auf die Gezeitenperiodik. *Zeitschrift für Vergleichende Physiologie*, 53(1):1–61.

Neumann, D. (1967). Genetic adaptation in emergence time of *Clunio* populations to different tidal conditions. *Helgoländer wissenschaftliche Meeresuntersuchungen*, 15(1):163–171.

Neumann, D. (1978). Entrainment of a Semilunar Rhythm by Simulated Tidal Cycles of Mechanical Disturbance. *Journal of Experimental Marine Biology and Ecology*, 35(1):73–85.

Neumann, D. (1983). Die zeitliche Programmierung von Tieren auf periodische Umweltbedingungen. In *Rheinisch-Westfälische Akademie der Wissenschaften, Natur- Ingenieur- und Wirtschaftswissenschaften, Vorträge N 324*, pages 31–68. Westdeutscher Verlag.

Neumann, D. (2014). Timing in Tidal, Semilunar, and Lunar Rhythms. In Numata, H. and Helm, B., editors, *Annual, Lunar, and Tidal Clocks*, pages 3–24. Springer.

Neumann, D. and Heimbach, F. (1979). Time Cues for Semilunar Reproduction Rhythms in European Populations of *Clunio marinus*. I. The Influence of Tidal Cycles of Mechanical Disturbance. In Naylor, E. and Hartnoll, R. G., editors, *Cyclic Phenomena in Marine Plants and Animals: Proceedings of the 13th European Marine Biology Symposium, Isle of Man, 27 September - 4 October 1978*, pages 423–433. Pergamon Press.

Neumann, D. and Honegger, H. W. (1969). Adaptations of the intertidal midge *Clunio* to arctic conditions. *Oecologia*, 3:1–13.

Neumann, D. and Spindler, K.-D. (1991). Circasemilunar Control of Imaginal Disk Development in *Clunio marinus*: Temporal Switching Point, Temperature-Compensated Developmental Time and Ecdysteroid Profile. *Journal of Insect Physiology*, 37(2):101–109.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Ortiz, E. M. (2019). vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. 2540861.

Palmén, E. and Lindeberg, B. (1959). The marine midge, *Clunio marinus* Hal.(Dipt., Chironomidae), found in brackish water in the northern Baltic. *Internationale Revue der gesamten Hydrobiologie und Hydrographie*, 44(1-4):383–394.

Paradis, E. and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526–528.

Patton, H., Hubbard, A., Andreassen, K., Auriac, A., Whitehouse, P. L., Stroeven, A. P., Shackleton, C., Winsborrow, M., Heyman, J., and Hall, A. M. (2017). Deglaciation of the Eurasian ice sheet complex. *Quaternary Science Reviews*, 169:148–172.

Pavlidis, P., Jensen, J. D., Stephan, W., and Stamatakis, A. (2012). A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans. *Molecular Biology and Evolution*, 29(10):3237–3248.

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

Pickrell, J. and Pritchard, J. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *Nature Precedings*, pages 1–1.

Ravinet, M. (2018). vcf2treemix.sh. last accessed 16th April 2021.

Reineke, A., Karlovsky, P., and Zebitz, C. P. W. (1998). Preparation and purification of DNA from insects for AFLP analysis. *Insect Molecular Biology*, 7(1):95–99.

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

Savolainen, O., Lascoux, M., and Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews Genetics*, 14(11):807–820.

Selcho, M., Millán, C., Palacios-Muñoz, A., Ruf, F., Ubillo, L., Chen, J., Bergmann, G., Ito, C., Silva, V., Wegener, C., et al. (2017). Central and peripheral clocks are coupled by a neuropeptide pathway in *Drosophila*. *Nature communications*, 8(1):1–13.

Slowikowski, K. (2020). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.8.2.

Soong, K. and Chang, Y.-H. (2012). Counting Circadian Cycles to Determine the Period of a Circasemilunar Rhythm in a Marine Insect. *Chronobiology International*, 29(10):1329–1335.

Suzuki, T. and Saigo, K. (2000). 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(7):1531–1540.

Thackeray, S. J., Henrys, P. A., Hemming, D., Bell, J. R., Botham, M. S., Burthe, S., Helaouet, P., Johns, D. G., Jones, I. D., Leech, D. I., Mackay, E. B., Massimino, D., Atkinson, S., Bacon, P. J., Brereton, T. M., Carvalho, L., Clutton-Brock, T. H., Duck, C., Edwards, M., Elliott, J. M., Hall, S. J. G., Harrington, R., Pearce-Higgins, J. W., Høye, T. T., Kruuk, L. E. B., Pemberton, J. M., Sparks, T. H., Thompson, P. M., White, I., Winfield, I. J., and Wanless, S. (2016). Phenological sensitivity to climate across taxa and trophic levels. *Nature*, 535(7611):241–245.

Velarde, R. A., Sauer, C. D., Walden, K. K., Fahrbach, S. E., and Robertson, H. M. (2005). Pteropsin: A vertebrate-like non-visual opsin expressed in the honey bee brain. *Insect biochemistry and molecular biology*, 35(12):1367–1377.

Verta, J.-P. and Jones, F. C. (2019). Predominance of *cis*-regulatory changes in parallel expression divergence of sticklebacks. *Elife*, 8:e43785.

Wang, X., Ding, Y., Lu, X., Geng, D., Li, S., Raikhel, A. S., and Zou, Z. (2021). 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(8).

Wellenreuther, M. and Hansson, B. (2016). Detecting Polygenic Evolution: Problems, Pitfalls, and Promises. *Trends in Genetics*, 32(3):155–164.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, 40(1):1–29.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Yuan, M. and Stinchcombe, J. R. (2020). Population genomics of parallel adaptation. *Molecular Ecology*, 29(21):4033–4036.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5):614–620.

Zhang, W., Reeves, G. R., and Tautz, D. (2021). Testing Implications of the Omnigenic Model for the Genetic Analysis of Loci Identified through Genome-wide Association. *Current Biology*, 31(5):1092–1098.

Zheng, L. and Carthew, R. W. (2008). Lola regulates cell fate by antagonizing Notch induction in the *Drosophila* eye. *Mechanisms of development*, 125(1-2):18–29.

4 QTL mapping with limited genetic markers detect loci linked to lunar-rhythmicity in *Clunio marinus*

Manuscript

Authors: Nico Fuhrmann[1] and Tobias S. Kaiser[1]

Affiliation: [1]Max Planck Institute for Evolutionary Biology, Max Planck Research Group "Biological Clocks", August-Thienemann-Strasse 2, 24306 Plön, Germany.

4.1 Abstract

The mapping of quantitative trait loci (QTL) in the F2 offspring of phenotypically divergent lines is a powerful tool to identify linkage between the phenotypes and genomic loci. Sympatric populations provide the unique opportunity to study diverging genetic characteristics between individuals which are able to exchange genetic material and share the same ancestry. While most of the genome is mixed between the populations, differentiated loci stand out. The life-cycles of diverging *Clunio marinus* ecotypes are very unique, as adults emerge either synchronized around full moon and new moon days or every day. A crossing experiment was performed between the sympatric *Clunio* ecotypes Ber-1SL and Ber-2AR with the aim to identify QTLs linked to lunar-rhythmic emergence of the adult insects. As the degree of diversification between both populations is known from previous studies, in this approach a set of genetic markers is used to directly investigate divergent loci while still covering large stretches of the chromosomes. The F2 generation was sequenced using custom designed amplicon primers to cover differentiated genomic loci between the parental populations and to identify informative polymorphic markers in them. The resulting dataset after filtering is comprised of 237 F2 individuals and between four and six polymorphic markers on the three chromosomes. Even with this limited marker set it was possible to detect QTLs linked to lunar-rhythmic phenotypes. Multiple QTLs across the genome were discovered by varying phenotypes which are linked to lunar-rhythmicity. The promising results from the dataset shows that detecting significant QTLs with a low number of known polymorphic markers is possible in these sympatric populations. Further analyses with additional genetic markers should follow this study to obtain a higher resolution of the QTL numbers and positions linked to lunar-rhythmicity.

4.2 Introduction

In many cases phenotype variation between distinct populations is linked to specific genotypes. While we can use diversity statistics and genome comparisons to identify genetic variation, these methods cannot link genetic functionality to phenotypes (Stinchcombe and Hoekstra, 2008). The functional genetic variants underlying a phenotype in a population are often obscured by ecological adaptations, unique life histories and putatively admixed population structure (Broman and Sen, 2009; Stinchcombe and Hoekstra, 2008). Quantitative trait loci (QTL) mapping is a powerful method to detect linkage between a polymorphic marker and phenotypes. This is achieved in a QTL experiment through the observation of an appropriate number of crossing over events at genetic loci of interest. QTL mapping success depends on the number of observed recombination events in the F1 and F2 offspring. Detection power increases with offspring and genetic marker number. Through targeted crossings experiments between populations of distinct phenotypes, present genetic composition and population structure is mixed which allows to trace the causing genotypes of a heritable phenotype through the crossing over events in F1, F2 or backcross generations (Broman and Sen, 2009; Stinchcombe and Hoekstra, 2008).

Sympatric ecotypes, populations or species can experience similar ecological conditions and have opportunities to exchange genetic material through gene flow. Differing characteristics like behavior, emergence times or ecological niche occupation would stand out in an admixed population dynamic, if selected for. These are ideal conditions for a QTL study on diverging phenotypes. Such a sympatric populations scenario occurs between *Clunio* populations at the Kviturdvikpollen close to Bergen (Norway). *Clunio marinus* is a marine non-biting midge (Diptera: Chironomidae) with a peculiar adult emergence timing in diverging ecotypes (Fuhrmann et al., 2021). The *Atlantic ecotype* of *Clunio marinus* inhabits rocky intertidal zones and is perfectly adapted to the local tidal regime. The emergence of adults, mating and egg deposition is linked to the dry-falling of the larval substrate during spring low tides around full moon and new moon days (Kaiser et al., 2016; Neumann, 1986). The *Baltic ecotype* of *Clunio marinus* has lost the lunar-rhythmicity with the lack of strong tides in the Baltic Sea (Endraß, 1976; Palmén and Lindeberg, 1959; Remmert, 1955). In the Kviturdvikpollen near Bergen the lunar-rhythmic *Atlantic ecotype* Ber-1SL and the lunar-arrhythmic *Baltic ecotype* Ber-2AR coexist (Heimbach, 1978). The loss of lunar-rhythmicity in one of the sympatric populations and the maintenance of this ecotype under gene flow with an

Atlantic ecotype population (Fuhrmann et al., 2021) is the ideal initial condition for QTL mapping of the lunar-rhythmic emergence phenotype. Laboratory strains of both natural populations were established and numerous crossing experiments performed between the *Atlantic* and *Baltic ecotype*. For the presented QTL mapping analysis the F2 offspring of 12 different F1 siblings from one pair of grandparents were selected (Fig. 12).

In QTL mapping individual phenotypes are linked to the present genotypes (Broman and Sen, 2009). Since every individual can emerge only once in their life, the rhythm can only be observed on population level in the case of *Clunio* adult emergence. Depending on the overall population emergence distribution the individual diel and lunar emergence times can be used to distinguish between "rhythmic" and "arrhythmic" individuals. In lunar-rhythmic populations, emerging in the populations' main peak would qualify an individual for the "rhythmic" phenotype. However, lunar-arrhythmic populations have an equally distributed emergence of individuals. This makes phenotyping in crossing experiments between rhythmic and arrhythmic populations very difficult. Arrhythmic individ-

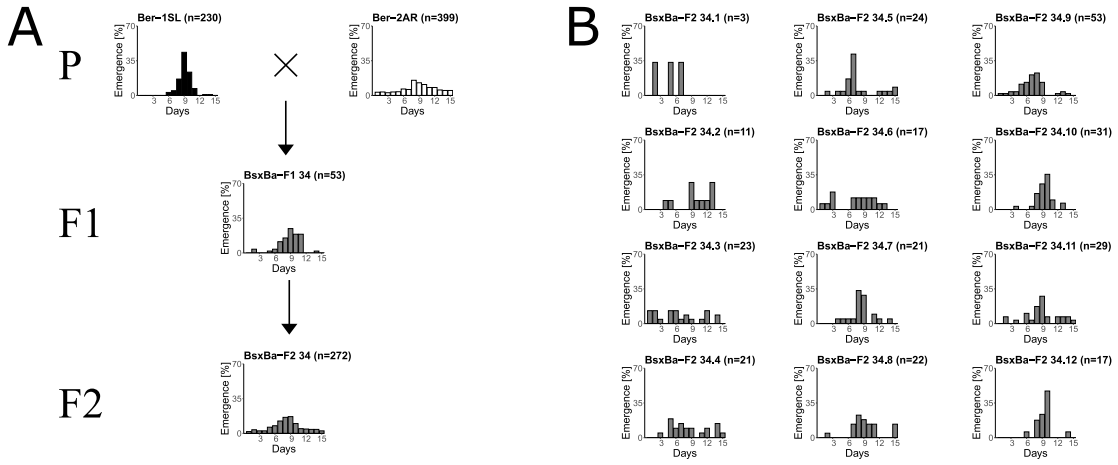

Figure 12: Lunar emergence patterns of the studied crossing family.
(**A**) One single pair cross between the sympatric Ber-1SL and Ber-2AR strain resulted in an phenotypically distinct F1 family. The F1 siblings could freely mate and only the fertilized eggs were raised as separate F2 families (see section 4.3 Materials and Methods). The overall F2 emergence distribution shows an arrhythmic emergence with a pronounced increase in emerging individuals at the time when the parental lunar-rhythmic Ber-1SL population is emerging as well. (**B**) When plotting the F2 families individually, it becomes apparent that there are highly rhythmic families (e.g. F2-34.10) and completely arrhythmic families (e.g. F2-34.3 and F2-34.4), suggesting that this F2 generation is segregating for rhythmicity alleles.

uals would emerge around the time of the rhythmic populations' peak as well as outside of the peak days. Therefore, it is impossible to assign the individuals within the peak to either phenotype with certainty when crossing arrhythmic and rhythmic phenotypes.

The aim of this study is to identify QTLs of lunar-rhythmic phenotypes in a crossing family of two sympatric-ecotype *Clunio* laboratory strains with a limited number of genetic markers. As the degree of differentiation between the parental strains is known from previous studies, the presented analyses attempts to investigate those loci directly for a potential link to lunar-rhythmicity. Therefore, few but precisely located and selected genetic markers were selected. From whole-genome resequencing of both grandparents, six specific genetic markers for each chromosome were selected and amplified in a multiplex PCR using F2 samples as template. The crossing-dataset is composed of 272 F2 individuals from a single crossing family. A possibility to get around the individual phenotype uncertainty described above is to utilize the differing phenotype ratios between the peak and every other emergence time. Rhythmic phenotypes should increase during and disappear outside of the peak. The contrast can be increased by removing flanking regions of the peak in which the ratio between the phenotypes is closer to be equal. By doing so one can create a binary phenotype which can be used to identify rhythmic QTLs based on phenotype ratio and remove phenotyping uncertainty. Three binary lunar-subset phenotypes were defined by excluding specific flanking days around the defined day of the rhythmic peak (Fig. 13).

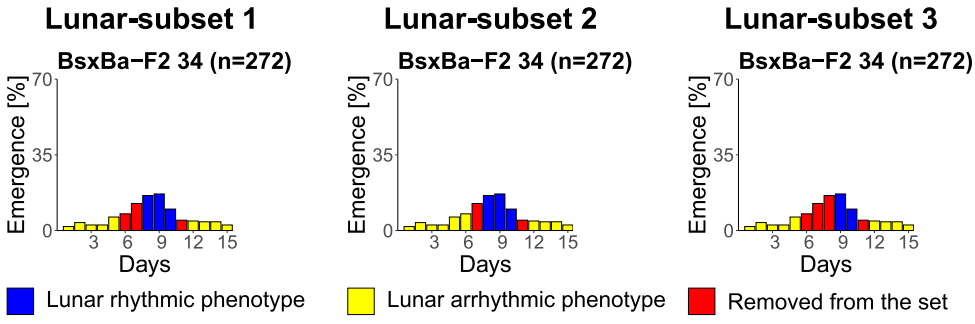

Figure 13: Selection of three binary lunar-subset phenotypes.
For all three subsets the selected (blue; yellow) and removed (red) individuals (see section 4.3 Materials and Methods) from the actual F2 emergence distribution are marked by the color of the emergence day in the turbulence cycle. Individuals which emerged in the range of the rhythmic peak (blue) and in the days of arrhythmic emergence (yellow) are included for the binary phenotype.

Additionally, six non-binary lunar-rhythmicity-linked phenotypes were selected. The phenotype "developmental time" refers to the number of days from fertilization of the eggs to the emergence of the adult insect. Directly linked to the lunar-rhythm are "exact emergence day", "emergence distance to day 9", "emergence distance to day 10" and "emergence distance to day 9/10" as they refer to the adult emergence day in the 15-day turbulence cycle the cultures were entrained to. Finally, the individuals were phenotypically scored by "chance of being arrhythmic", defined as the expected fraction of lunar-arrhythmic individuals at each day in the artificial lunar cycle. Several QTLs linked to lunar-rhythmicity were detected in almost all investigated phenotypes. QTLs identified with the binary lunar-rhythmicity subsets, interactive, additive or individual effects in multiple analytical runs were discovered with manually fitted QTL models. Ultimately the limited but precisely selected genetic markers were able to detect significant QTLs linked to lunar-rhythmicity in the investigated crossing family. More genetic markers should be added to obtain a higher resolution of the specific QTL positions which are linked to lunar-rhythmicity in *Clunio*.

4.3 Materials and Methods

Culture conditions, crossing experiment and sample collection. The laboratory strains Ber-1SL and Ber-2AR were used for QTL mapping experiments (see section 3.5). All laboratory cultures were kept in 18 hours of light and six hours of darkness, entrained by 12.4 hours 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) (Neumann, 1978; Neumann and Heimbach, 1979). Larvae were kept in natural seawater diluted 1:1 with deionized water. Ber-2AR was fed with diatoms (*Phaeodactylum tricornutum*) and powdered red algae (90% *Delesseria spp.*, 10% *Ceramium spp.*, obtained from F. Weinberger and N. Stärck, GEOMAR, Kiel). Ber-1SL and progeny from the crosses (F1 and F2 generations) were fed with diatoms and powdered nettles (*Urtica sp.*) (Neumann, 1978; Neumann and Heimbach, 1979).

The circadian emergence of Ber-2AR under the above described constant laboratory conditions is five hours later than the sympatric strain Ber-1SL. To cross the parental strains, the L:D regime of the Ber-2AR strain was reset five hours before the schedule of Ber-1SL. Because the F1 would emerge in an intermediate diel time to the parental strains (Heimbach (1978); Kaiser et al. (2011); Fuhrmann, unpublished data), immediate collection of the crosses was limited by the light shut down for the night time. Furthermore, the presented crossing family was one among 33 others which were emerging at the same time, adding another logistical difficulty for immediate crossing selection in the F1 generation. To obtain the F2 generation, all F1 individuals were allowed to freely mate within their own culture box. Adults and egg clutches were collected the next morning after diel emergence ended. Adult insects were preserved in 99.98% ethanol and stored at -20°C until DNA extraction. Egg clutches were only labeled as new F2 family, if fertilization was successful. For only two F2 families of the presented crossing experiment are both and for four more at least one parent clearly identified.

DNA extraction, sequencing and read preparation of the grandparents. Grandparents of one successful crossing family were selected for the QTL mapping experiment (Fig. 12) and prepared by Kerstin Schaefer. Family 34 has the most promising and diverse emergence distributions. The F1 appeared mostly rhythmic and the F2 was a mix between rhythmic and arrhythmic individuals with a high sample size of 272 F2 individuals. Genomic DNA of these two samples was extracted using a salting out method (Reineke et al., 1998) and due to low genomic DNA yield (female: 0.51 ng/µl; male:

0.39 ng/µl) amplified using the REPLI-g Mini Kit (QIAGEN) according to the manufacturer's protocol with volume modifications (Tab. S3). They were then subject to whole genome shotgun sequencing at 10-15x target coverage on an Illumina HiSeq3000 sequencer with 150 bp paired-end reads. Low-input library preparation and sequencing were performed by the Max Planck Genome Centre (Cologne, Germany) according to standard protocols.

Raw sequence reads were trimmed for adapters and base quality using Trimmomatic v.0.38 (Bolger et al., 2014) 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 (Zhang et al., 2014), 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 (Li, 2013) to the nuclear reference genome (Kaiser et al., 2016) (ENA accession GCA_900005825.1) of *C. marinus*. Mapped reads were sorted, indexed, filtered for mapping quality ('-q 20') and transformed to BAM format with SAMtools v.1.9 (Li et al., 2009). Read group information was added with the *AddOrReplaceReadGroups.jar* v.1.74 script from the Picard toolkit (http://picard.sourceforge.net/) (DePristo et al., 2011).

SNPs and insertion-deletion (indel) genotypes were called using GATK v.3.8-0-ge9d806836 (McKenna et al., 2010). 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 (Danecek et al., 2011) 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 656,368 variants.

Identifying potential genetic markers between sympatric populations. Low coverage whole genome sequencing would be the ideal approach for a similar QTL experiment. But the extremely low genomic DNA yield, which is also expected in the F2 individuals makes an amplification necessary in any case. Additionally, the chosen grandparent strains have besides distinct peaks an overall low genomic differentiation. Therefore, instead of using restriction site associated DNA markers on whole genome am-

plification products for each individual, genomic differentiation data from chapter 3 was used to identify potential candidate markers near the highest differentiated areas in equally spaced regions across the genome for amplicon sequencing (Fig. 9D; Fig. S12 "Ber-2AR vs. Ber-1SL"). Using a custom R-script (Crawley, 2007) with the packages 'data.table' (Dowle and Srinivasan, 2020), 'dplyr' (Wickham et al., 2020) and 'tidyr' (Wickham, 2020) the VCF of the grandparents was filtered for homozygous loci. The conditions were, first homozygous in both grandparents, and secondly, they had to be differentiated. This filtering reduced the entire genotype set to 53,097 population-unique homozygous alleles.

As equally spaced genetic markers across the genome of *Clunio* were needed, the reference genome was subdivided into six regions per chromosome. The summed length of the chromosomal scaffolds per region was set to be of comparatively equal base pair length within the chromosome boundaries. Potential candidate loci within these regions with respect to the highest FST values and the number of additional population-unique homozygous alleles in the upstream and downstream 500 base pair range were selected (Fig. 14). Based on the putative genetic-marker-loci, Kerstin Schaefer designed primer pairs which covered fragment lengths from 227 bp to 925 bp in approximately 100 bp steps per chromosome (Tab. S10). The primer pairs were tested in PCR experiments on field sample DNA from each sex and each population (Ber-2AR-female; Ber-2AR-male; Ber-1SL-female; Ber-1SL-male).

DNA extraction, sequencing, read preparation and genotype calling of the F2 generation. DNA extraction of the F2 individuals was performed using the QuickExtract DNA Extraction Solution (Lucigen, QE0905T) according to manufacturer's protocol with modifications. Instead of vortexing before incubation, an autoclaved pestle was used to grind the sample. The resulting genomic DNA concentration was as expected from the grandparents very low (between 0.28 ng/µl and 1.39 ng/µl in test measurements). All steps above were performed by Kerstin Schaefer. The extracted genomic DNA was finally used for a multiplex PCR with all previously generated and tested Primers (see section 4.3). Multiplex PCR was performed for each chromosome primer set (six primer pairs) separately. The QIAGEN Multiplex PCR Kit (206143) was used to prepare a plate-master-mix for each 96-well plate. The PCR program started with 15 minutes at 95°C followed by 40 cycles of denaturation at 94°C for 30 seconds, 90 seconds of annealing at 57°C, the elongation at 72°C for 90 seconds and closed with 72°C for 10 minutes. Library preparation and demultiplexing of the samples after sequencing was performed at the MPI

for Evolutionary Biology. All PCR products were first sequenced with 150 bp paired-end reads at the MPI on the MiSeq Micro and MiSeq 600 cycles (spike in 10%), and additionally on an Illumina NovaSeq 6000 each at the Competence Centre for Genomic Analysis (CCGA) Kiel. The approximated final coverage among the samples is between 300x to 400x.

Raw read preprocessing was performed in the same way until the finally merged VCF as for the grandparents (see section 4.3). The final VCF contained the grandparents and the entire F2 generation. SNP and indel genotypes were again filtered with VCFtools v.0.1.14 and similar parameters as in section 4.3, with the exceptions of a minimum minor allele frequency of 0.25 ('- -maf 0.25') and a maximum proportion of missing data per locus of 30% ('- -max-missing 0.7'), resulting in 57 remaining loci for QTL mapping.

Quality filtering of genetic markers, genotypes and individuals and selected quantitative traits. The genotype information was extracted from the VCF and summarized to represent only the available information about homo- or heterozygosity per individual and genotype (A/A; B/B; A/B; NA). The grandfather's alleles were set as 'A' and the grandmother's alleles as 'B'. Heterozygous sites are uninformative, since we need to see recombination between the grandparent genotypes to successfully identify a

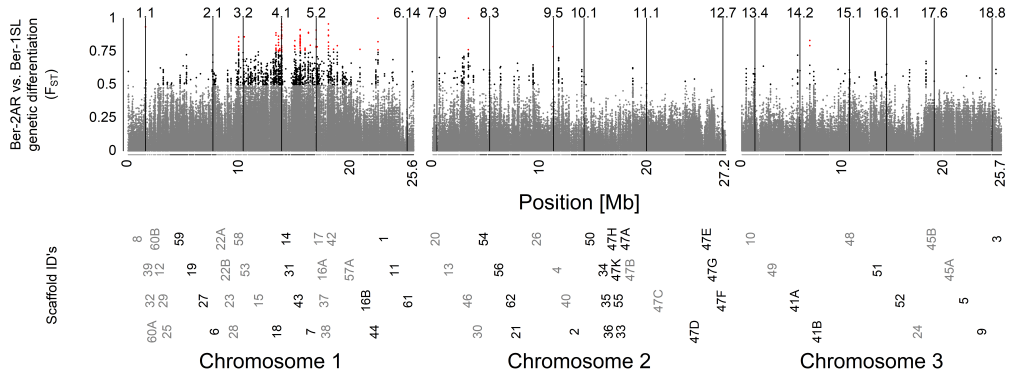

Figure 14: Location of the designed genetic markers across the genome of *Clunio*.

The Manhattan plot shows the genetic differentiation (F_{ST}) between the populations Ber-2AR and Ber-1SL. The F_{ST} values are color-coded in grey ($F_{ST} < 0.5$), black ($0.5 \leq F_{ST} < 0.75$) and red ($F_{ST} \geq 0.75$). The genetic markers are marked with black lines at their exact location on the chromosomes. Below the plot are the scaffold ID's, subdivided by color changes between gray and black to their respective assigned regions on the chromosomes.

QTL. Because there were almost no fully differentiated loci between the parental populations, there is a possibility that the grandparents share the same heterozygous or even homozygous genotypes at one or more loci. Therefore, loci were removed which were not differentiated between the grandparent or homozygous in one of them (see section 4.3). The sequenced PCR products varied in length (Tab. S10), which resulted in the presence of multiple genotypes within some of the markers. If more than one genotype remained at a marker location the majority consensus genotype state (homo- or heterozygous) was selected. After read preprocessing, mapping, genotype calling and initial filtering of the 18 genetic markers from each chromosome, 15 remained for the QTL analysis (chromosome 1: five markers; chromosome 2: four markers; chromosome 3: six markers). The next step was to filter individuals with more or equal to 50% missing genotypes per chromosome, reducing the initial 272 F2 individuals to 237 (Supplementary Data 10).

Nine different quantitative traits were selected for the presented QTL mapping approach. The binary traits were three different lunar-rhythmic subsets (Fig. 13). The subsets excluded individuals from the flanks of the emergence distribution and separated the remaining individuals in within the peak (rhythmic) and outside the peak (arrhythmic) (Fig. 13; Supplementary Data 11-13). The following non-binary traits were considered as well. "Developmental time" as the number of days from fertilization to emergence of each individual. The "exact emergence day" in the 15-day turbulence cycle. The "chance of being arrhythmic" is defined as the fraction of expected lunar-arrhythmic individuals on each day of the artificial lunar cycle. The number of individuals, which emerged from the parental lunar-arrhythmic Ber-2AR strain, was divided by the total number of individuals which emerged on the given day in Ber-2AR and Ber-1SL together.

The "emergence distance to day 9", "emergence distance to day 9/10" and "emergence distance to day 10" are defined as the distance of the individuals emergence day to the set emergence peak day (day 9, 9/10 or 10). As example, individuals emerging on day 6 or 12 are three days distant from day 9. Based on the emergence distribution of the rhythmic parental population Ber-1SL under turbulence entrainment, day 9, 9/10 and 10 were selected as the main peak days for rhythmic emergence. The Kolmogorov-Smirnov test and the Shapiro-Wilk test of normality performed in R resulted in a significant deviation from a normal distribution in the traits mentioned above (Fig. 16-17A). Hence these traits were treated as non-parametric in posterior analyses.

Statistical analysis and identification of quantitative traits using R/qtl. The previously mentioned input sample dataset (n=237) was analyzed using the R package 'qtl' (Broman et al., 2003). The pairwise recombination fractions were estimated, likely genotyping errors were identified and different marker orders for each chromosome were assessed. No apparent problems or errors were found, except for the marker order on chromosome 3. In all datasets the original fourth marker (marker number 16.1 on scaffold 52; see Fig. 14) was dropped as suggested by R/qtl because of a sudden switch from one to the other homozygous parental genotype and back in multiple samples (Fig. S21).

The genome scans with different QTL models followed the calculation of conditional genotype probabilities in 5 cM distances and simulated genotypes from 64 imputations with the same distance. The non-parametric phenotypes limited the options to genome scans with a single QTL model and the scanone (method = "em") function. Significance thresholds were always set at top 5% of the logarithm of the odds (LOD) scores for 1000 permutations. In cases where a locus crossed the set threshold in a single QTL scan, additionally estimated Bayesian credible intervals were used to identify the QTL interval.

For the binary lunar-rhythmicity phenotypes the three lunar-subsets (set 1: n=177; set 2: n=198; set 3: n=144) were quality checked in the same way as the non-parametric phenotypes before the QTL scans. Again, only the marker on chromosome 3 was dropped as in the non-binary phenotype analysis. The binary phenotypes allowed an additional single QTL scan using the Haley-Knott regression method, with multiple imputations in place of a maximum likelihood approach. On top, a two-dimensional genome scan for QTLs was possible using binary phenotypes. In the case of the lunar-subsets, based on the high LOD score results of the two-dimensional genome scans for QTLs, QTL subsets were created using the function makeqtl(). By doing so, only those QTLs were included which appear to be involved in the phenotype and perform an analysis of variance (ANOVA). Additive and interactive effects between individually selected QTL models were calculated using the fitqtl() function (Tab. S7-S9).

4.4 Results and Discussion

Binary "lunar-rhythmicity" phenotypes detect a single QTL and interactions between two chromosomes. As described in section 4.2 discrimination between rhythmic and arrhythmic individuals in a circalunar emergence distribution is not possible by the phenotype alone. Three binary phenotypes were used to cover the different ratios of rhythmic and arrhythmic genotypes between the emergence peak and the remaining emergence distribution. To keep the ratios different, the flanking days of the peak were excluded (Fig. 13). In the first and second (Fig. 15A,B) lunar-subset a QTL was identified in the single QTL model scan passing the LOD score threshold ("Lunar-subset 1": 2.41 LOD; "Lunar-subset 2": 2.4 LOD; "Lunar-subset 3": 2.44 LOD). The QTL spans the third genetic marker on chromosome 2 according to the Bayesian credible interval. In the third lunar-subset (Fig. 15C) a LOD score peak is visible, but does not pass the threshold. In the two-dimensional QTL model scans all sets show some evidence for interactive QTLs within and between all chromosomes (Fig. 15D-F, yellow areas in the upper left triangles).

In the area of the third genetic marker of chromosome 2, interaction is only detected with the entire chromosome 3 but not chromosome 1. A stronger signal is visible for an additive QTL (Fig. 15D-F, yellow areas in the lower right triangles). Evidence is visible for additive effects between this area and the entirety of every other chromosome, including the rest of chromosome 2. The subsequently performed analysis of variance (ANOVA) on selected subsets of QTLs with high interactive or additive LOD scores revealed a different picture (Tab. S7-S9). Lunar-subsets 2 and 3 are fairly similar. In the general additive model significant additive effects are present across the chromosomes in subset 2 (Tab. S8, Model: $y \sim Q1 + Q2 + Q3 + Q4 + Q5 + Q6$) and an additive interaction between chromosome 1 and 2 in subset 3 (Tab. S9, Model: $y \sim Q1 + Q2 + Q3 + Q4 + Q5 + Q6$). In the pairwise models QTL interactions between chromosome 1 and the end of chromosome 3 were identified in both subset 2 and 3 (Tab. S8-5.38, Model: $y \sim Q1 + Q3 + Q1 : Q3$). On the other hand, there is significant evidence for the stand-alone QTL on chromosome 2 in the other two pairwise models (Tab. S8-S9, Model: $y \sim Q1 + Q2 + Q1 : Q2$ and $y \sim Q2 + Q3 + Q2 : Q3$). Lunar-subset 1 did not show any significant additive effects between any QTLs on the chromosomes (Tab. S7, Model: $y \sim Q1 + Q2 + Q3 + Q4$).

Additionally, the interactive and additive effects between chromosome 1 and 3 do not reappear as significant, but a QTL on chromosome 2 and 3 show evidence for single and ad-

ditive effects. The lunar-subset 2 is the largest dataset with 198 individuals remaining from the whole dataset, has the most significant results and is at least in the two-dimensional QTL model scan and the ANOVA comparable to subset 3, which has the lowest remaining number of individuals with 147. In regards to ANOVA, subset 1 differs most from the others but has 177 individuals still remaining. The main difference to subset 2 is, that day 6 was excluded in addition to day 7, while in subset 3 day 6 to 8 were excluded (Fig. 13). By excluding day 6 it appears, that the composition of the genotypes changed so much, that the significant QTLs switched from chromosome 1 and 3 interactive to chromosome 2 and 3 additive effect. This change could be recovered by additionally excluding day 8, reducing the genotype ratio in the set. Evidence points to an involvement of the reoccurring high *LOD* score QTL on chromosome 2 in lunar-rhythmicity in single and two-dimensional QTL scans. An interaction between the QTLs on chromosome 1 and 3 was identified by ANOVA. While the range of the Bayesian credible interval is very broad and the high *LOD* scores are limited to chromosome 2, the two-dimensional QTL model scan and manual fitting of specific QTL models identify interactive and additive QTLs on chromosome 1 and 3 to be involved with the lunar-rhythmic trait as well. But only four genetic markers on chromosome 2 are not sufficient to pinpoint the exact location of this lunar-rhythmic QTL. But the detection of one significant QTL on chromosome 2 in every subset shows the effectiveness of the population-emergence oriented binary phenotyping and the presence of a strong genetic component with which the phenotype differences can be explained.

It also indicates the presence of a genotype characteristic in all subsets which distinguished lunar-rhythmic from lunar-arrhythmic population emergence. This is a different result from previous population genomic investigations, which discovered numerous loci predominantly concentrated across the entire chromosome 1 linked to the same ecotypes investigated in this study (Fuhrmann et al., 2021). These discrepancies could have originated from the design of the previous study in which multiple populations from diverse geographic locations were mixed and analyzed. These potentially strong differences in ecological adaptations to each geographic habitat could have overlayed the divergent loci responsible for lunar-arrhythmicity in one ecotype. Such divergent ecological adaptations are absent in the present QTL analyses of the sympatric populations. On the other hand could these discrepancies also indicate different genetic solutions to lunar-arrhythmicity within the same ecotype. This would identify the lunar rhythm as a multi loci trait with varying causing genetic variants.

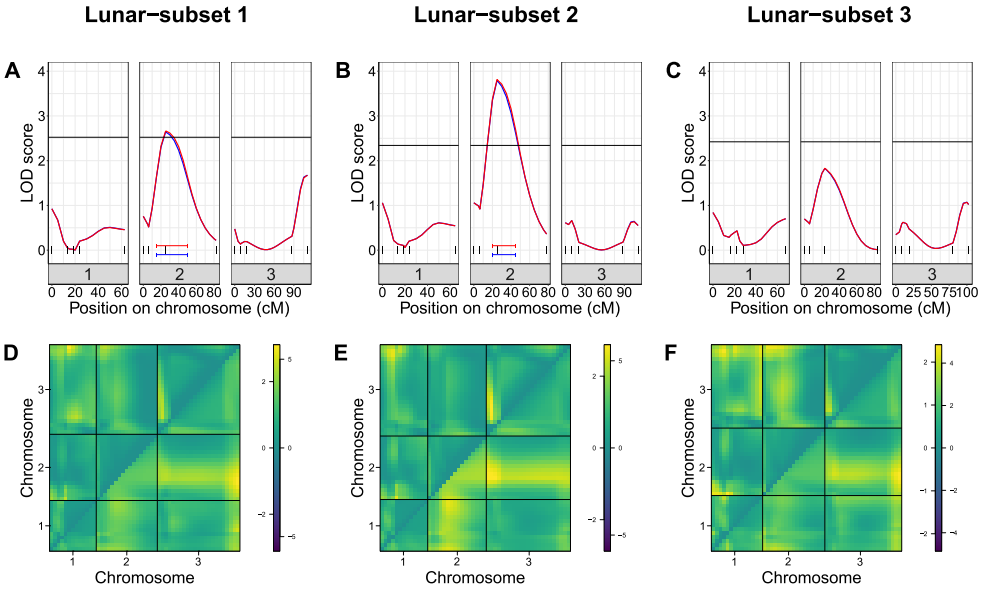

Figure 15: Results of the single and two-dimensional QTL model scans of all "lunar-subset" phenotypes.

Each column depicts the single and two-dimensional QTL model scans for lunar-subsets 1 (**A,D**; n=177), 2 (**B,E**; n=198) and 3 (**D,F**; n=144). In the first row (A-C) are the standard single QTL model scan across the three chromosomes. The *LOD* scores calculated with the scanone (method = "em") function (blue) and Haley-Knott regression method (red) are shown for each genetic marker (bottom strokes). The *LOD* score threshold ("lunar-subset 1": 2.41 *LOD*; "lunar-subset 2": 2.4 *LOD*; "lunar-subset 3": 2.44 *LOD*; top 5%, see section 4.3 Materials and Methods) for significant scores is marked as horizontal line. If a *LOD* score passed this threshold, the Bayesian credible interval was calculated and the range marked with a horizontal line/stroke at the bottom of the plot, colored like the respective algorithm used. In the second row (D-F) are two-dimensional QTL model scan as heatmap. In the upper left triangle of the heatmaps, the evidence for a local improvement of interactive over additive QTLs (LOD_i) is plotted. The evidence for a local improvement of the full two-locus QTL model (additive and interactive; LOD_f) over the null model is plotted in the lower right triangle of the heatmaps. The numbers on the left side of the scale corresponds to the LOD_i, numbers on the right correspond to the LOD_f values.

Individual phenotyping can identify significant QTLs linked to lunar-rhythmicity. While phenotyping of lunar-rhythmicity is difficult on an individual basis (see section 4.2 Introduction) some individually scorable phenotypes can be linked to rhythmic emergence of the adult insects despite being non-binary (e.g. "exact emergence day"). QTLs with significant *LOD* scores were found for many of those non-binary individual phenotypes through R/qtl single QTL model scans (Fig. 16-17B). The strong signal at the beginning of chromosome 1 in the phenotype "developmental time" is remarkable in several ways (Fig. 16B: "Developmental time"). First, it is the strongest signal among the non-binary phenotypes. Second, the phenotype is in rhythmic individuals affected by the circalunar rhythm. Their development will be arrested to time the emergence exactly for the next genetically controlled peak in the lunar month (Kaiser et al., 2011; Neumann, 1986). This links developmental time of the individuals to lunar-rhythmicity, enabling rhythmic individuals to "wait" for the right time to emerge. Arrhythmic individuals will emerge whenever they are ready. And third, all genetic markers but the first and last are significantly affected by the QTL, which is supported by the Bayesian credible interval (Fig. 16B: "Developmental time"). It spans all three markers and extends into the area before the last marker. Because many additional markers are missing in that area, the exact range of the QTL cannot be estimated, but it covered already the first half of the most differentiated area between the populations (Fig. 14). Therefore, the QTL of the "developmental time" phenotype could be spanning the entire differentiated and ecotype-associated area on chromosome 1 (see section 3.3; Fig. 9). By the evidence provided in the previous chapter on that exact region, this phenotype could be involved in lunar-rhythmic synchronization of *Clunio* populations.

The phenotype "exact emergence day" is indirectly associated with lunar-rhythmicity. While it is genetically determined when rhythmic individuals emerge (Kaiser et al., 2011), arrhythmic individuals will still emerge on any day. This causes the density of the appearing phenotypes to be similar to a normal distribution (Fig. 16A: "Exact emergence day"). The QTL scan is comparing all 15 possible emergence days in which the only difference is the fraction of the genotypes of rhythmic individuals among the arrhythmic. The identified QTL of this phenotype is located at the first marker of chromosome 1 (Fig. 16B: "Exact emergence day"). The Bayesian confidence interval stretches across the first three markers as well, which indicates that the end of the telocentric chromosome 1 could be involved as well since the beginning is not covered by genetic markers (Fig. 16B: "Exact emergence day"). The third phenotype "chance of being arrhythmic" has only one QTL

which reaches close to the *LOD* score significance threshold (Fig. 16B: "Chance of being arrhythmic"). The phenotype is calculated from the emergence distributions of the parental populations and is therefore not an individual phenotype, but imposed on the individuals. The resulting sample groups seem to be poorly composed to detect significant QTLs.

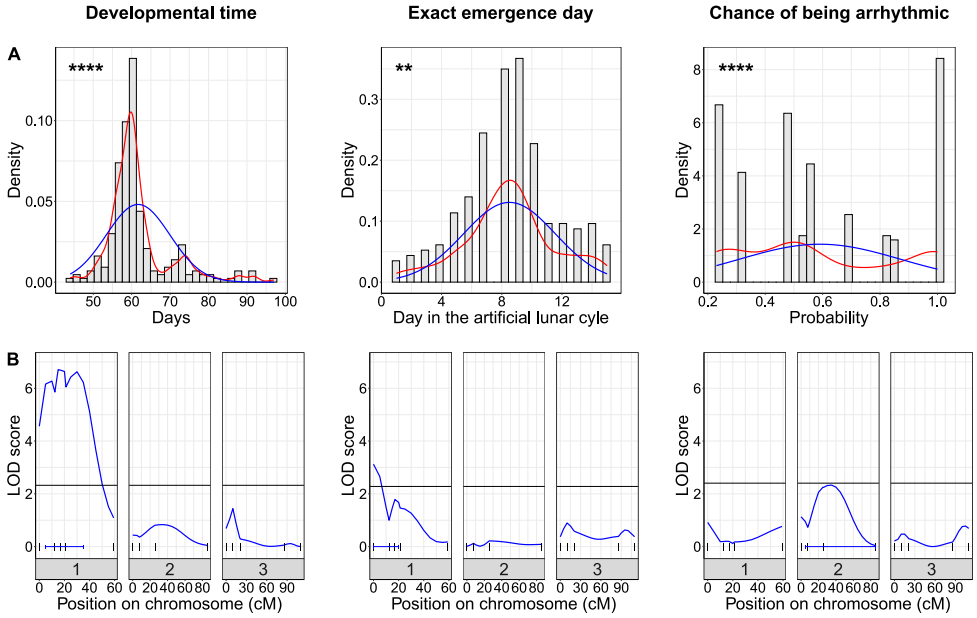

Figure 16: QTL mapping of the non-binary phenotypes "developmental time", "exact emergence day" and "chance of being arrhythmic" (left to right). There are three panels (**A-C**) for each phenotype. (**A**) The density histograms depict the distribution of individuals across the phenotype ranges. Kernel density estimate based on this distribution is plotted as red line, an estimated normal distributed curve based on the mean and standard deviation of the dataset is plotted as blue line. The significance level of the Kolmogorov-Smirnov test and the Shapiro-Wilk test of normality are shown in the upper right corner of each plot as p-value symbols. (**B**) Single QTL model scans across the three chromosomes. The *LOD* scores calculated with the scanone (method = "em") function are shown for each genetic marker (bottom strokes). The *LOD* score threshold ("developmental time": 2.35 *LOD*; "exact emergence day": 2.24 *LOD*; "chance of being arrhythmic": 2.31 *LOD*; top 5%, see section 4.3 Materials and Methods) for significant scores is marked as horizontal line. If a *LOD* score passed this threshold, the Bayesian credible interval was calculated and the range marked with a horizontal line at the bottom of the plot. All p-value symbols correspond to: **** - 0-0.0001; *** - 0.0001-0.001; ** - 0.001-0.01; * - 0.01-0.05; ns - 0.05-1.

Finally, the non-binary phenotypes “emergence distance to day 9”, “emergence distance to day 9/10” and “emergence distance to day 10” were investigated and are based on the emergence day of each individual. Instead of comparing 15 different phenotypes as with the “exact emergence day”, only eight are compared by merging the days with the same distance to the defined ’main peak’ of the distribution (Fig. 17B). This difference seems to dramatically shift the outcome of the single QTL model scan. With all the one-sided-distributed phenotypes “emergence distance to day X” a QTL with significant *LOD* score appears on chromosome 2 (Fig. 17B). The respective Bayesian credible interval spans almost the entire chromosome in the first two phenotypes, except for the first genetic marker (Fig. 17B: “Emergence distance to day 9”, “Emergence distance to day 9/10”). The last phenotype “emergence distance to day 10” could not identify a significant QTL, but indicates a similarly positioned and close to significance threshold QTL on chromosome 2 as before (Fig. 17B: “Emergence distance to day 10”).

Interestingly, the more abstract phenotypes “chance of being arrhythmic”, “emergence distance to day X” indicated a similarly positioned QTL, as found with the lunar-subsets on chromosome 2 in the single QTL model scans. This position appears to have some involvement in lunar-rhythmicity when investigating the individual phenotypes as well. However, the significant QTLs on chromosome 1 in “developmental time” and “exact emergence day” should be noticed. The detection of these QTLs could either hint towards additional causative genotypes for lunar-rhythmicity or indicate that developmental time as example is not linked to lunar-rhythmicity. Nonetheless, the reoccurring QTL on chromosome 2 detected via several lunar-emergence linked phenotypes highlights the presence of a potential lunar emergence controlling loci on this chromosome. The low number of investigated markers does not allow to conclude if lunar-rhythmicity is controlled mainly by a single genomic loci or it may be a multi loci or polygenic trait. More QTLs even on chromosome 2 could be involved in lunar-rhythmicity but more markers are needed.

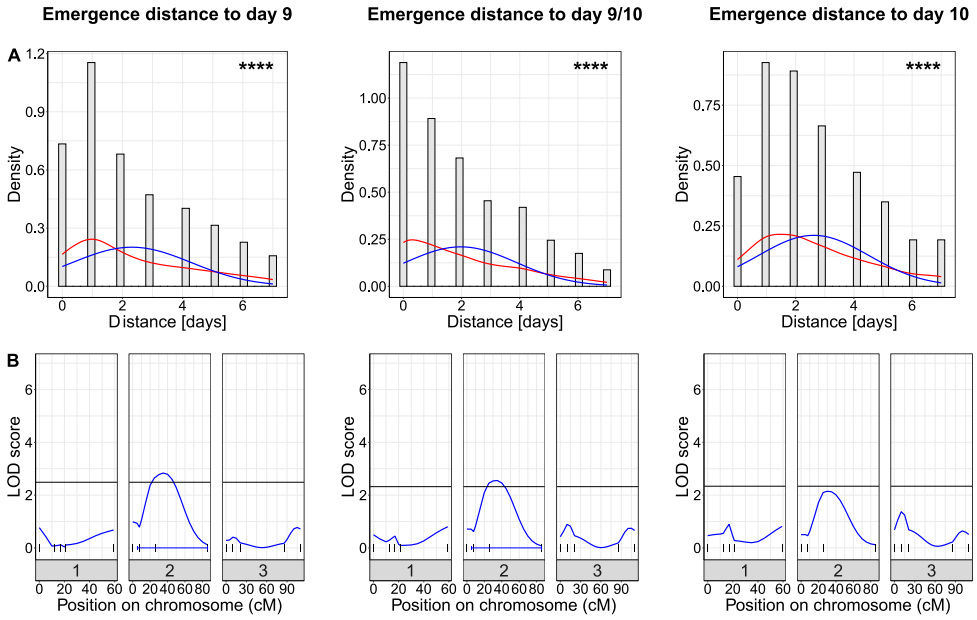

Figure 17: QTL mapping of the non-binary phenotypes "emergence distance to day 9", "emergence distance to day 9/10" and "emergence distance to day 10" (left to right).

(**A**) The density histograms depict the distribution of individuals across the phenotype ranges. Kernel density estimate based on this distribution is plotted as red line, an estimated normal distributed curve based on the mean and standard deviation of the dataset is plotted as blue line. The significance level of the Kolmogorov-Smirnov test and the Shapiro-Wilk test of normality are shown in the upper right corner of each plot as p-value symbols. (**B**) Single QTL model scans across the three chromosomes. The *LOD* scores calculated with the scanone (method = "em") function are shown for each genetic marker (bottom strokes). The *LOD* score threshold ("emergence distance from day9": 2.41 *LOD*; "emergence distance from day 9/10": 2.41 *LOD*; "emergence distance from day 10": 2.36 *LOD*; top 5%, see section 4.3 Materials and Methods) for significant scores is marked as horizontal line. If a *LOD* score passed this threshold, the Bayesian credible interval was calculated and the range marked with a horizontal line at the bottom of the plot. All p-value symbols correspond to: **** - 0-0.0001; *** - 0.0001-0.001; ** - 0.001-0.01; * - 0.01-0.05; ns - 0.05-1.

4.5 Conclusions

The presented QTL mapping in this study is the first attempt to identify significant QTLs linked to lunar-rhythmicity between the sympatric *Clunio* ecotypes from Bergen (Norway). Although the analyzed data is limited by only four to six genetic markers per chromosome in 237 F2 individuals, it is remarkable how three significant QTLs could be identified at different chromosomal positions. Three reduced datasets of 144 to 198 individuals detected one predominating reoccurring QTL on chromosome 2 using binary phenotyping. However, an exact location of the lunar-rhythmicity linked QTLs was not possible due to the low genomic coverage of the limited genetic marker set, leaving it open whether it is a single locus, or a cluster of multiple loci on chromosome 2.

As explained in section 4.2, it is difficult in the case of *Clunio* to determine discrete rhythmicity phenotypes. This is not trivial since the rhythmicity is an observation of the entire population, which increases resolution with increasing numbers of individuals. In this study different ways to obtain high ratios for a binary lunar-rhythmicity phenotype was tested, by excluding various flanking days of the main peak in the emergence distribution (Fig. 13). All three lunar-subsets were effective in detecting the same QTL on chromosome 2. This indicates, that the population-emergence based binary phenotyping and subsetting of the initial dataset is a successful way to identify significant QTLs. The only accurate individual phenotypes are "developmental time", "exact emergence day" and "emergence distance to day X". Although both "developmental time" and "exact emergence day" point to a completely different QTL on chromosome 1 instead of chromosome 2, they do as well detect significant QTLs. All phenotyping ultimately lead to significant QTLs indicating a multi loci phenotype in contrast to a genome wide regulated trait. Under a genome wide trait, no significant QTL is expected. If many loci all across the genome play a role in the phenotype, no QTL on their own would stick out like the here identified QTLs. But as mentioned, it is not possible with the low number of genetic markers in this study to determine how many QTLs are linked to lunar-rhythmicity or where they are exactly located on the chromosomes.

The here analyzed crossing family between the sympatric lunar-rhythmic Ber-1SL and the lunar-arrhythmic Ber-2AR laboratory strains is an optimal candidate for QTL mapping. The high number of 237 F2 individuals and their broad emergence distribution (Fig. 12) shows an intermediate phenotype between arrhythmic emergence and a concentrated peak in which arrhythmic and rhythmic individuals overlap. Non-binary and

binary phenotyping leads to the identification of few QTLs on different chromosomes, with individual and additive or interactive effects. The most prominent QTL reappears on chromosome 2 with multiple independent phenotyping approaches. Despite the fundamental link of several phenotypes with lunar-rhythmicity, the analyses detected multiple QTLs which hints towards a multi loci trait with few causative QTLs. These results are a further step towards the molecular understanding of lunar-rhythmicity in one of the few model organisms of the lunar chronobiology field and shows the strength of sympatric ecotypes in QTL mapping. The limitation in genetic markers may not affect the ability to detect significant QTLs, but it makes the identification of the exact position difficult. A higher genetic marker density would be needed to pinpoint QTL positions with greater certainty.

4.6 References

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Broman, K. W. and Sen, S. (2009). *A Guide to QTL Mapping with R/qtl*, volume 46. Springer.

Broman, K. W., Wu, H., Sen, Ś., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19(7):889–890.

Crawley, M. J. (2007). *The R Book*. Wiley.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and Group, . G. P. A. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491.

Dowle, M. and Srinivasan, A. (2020). *data.table: Extension of 'data.frame'*. R package version 1.13.2.

Endraß, U. (1976). Physiologische Anpassungen eines marinen Insekts. I. Die zeitliche Steuerung der Entwicklung. *Marine Biology*, 34(4):361–368.

Fuhrmann, N., Prakash, C., and Kaiser, T. S. (2021). Polygenic adaptation from standing genetic variation allows rapid ecotype formation. *bioRxiv*.

Heimbach, F. (1978). Sympatric Species, *Clunio marinus* Hal. and *Cl. balticus* n. sp.(Dipt., Chironomidae), Isolated by Differences in Diel Emergence Time. *Oecologia*, 32(2):195–202.

Kaiser, T. S., Neumann, D., and Heckel, D. G. (2011). Timing the tides: Genetic control of diurnal and lunar emergence times is correlated in the marine midge *Clunio marinus*. *BMC Genetics*, 12(1):1–12.

Kaiser, T. S., Poehn, B., Szkiba, D., Preussner, M., Sedlazeck, F. J., Zrim, A., Neumann, T., Nguyen, L.-T., Betancourt, A. J., Hummel, T., Vogel, H., Dorner, S., Heyd, F., von Haeseler, A., and Tessmar-Raible, K. (2016). The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature*, 540(7631):69–73.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

Neumann, D. (1978). Entrainment of a Semilunar Rhythm by Simulated Tidal Cycles of Mechanical Disturbance. *Journal of Experimental Marine Biology and Ecology*, 35(1):73–85.

Neumann, D. (1986). Chapter 1 Life cycle strategies of an intertidal midge between subtropic and arctic latitudes. In Taylor, F. and Karban, R., editors, *The Evolution of Insect Life Cycles*, pages 3–19. Springer.

Neumann, D. and Heimbach, F. (1979). Time Cues for Semilunar Reproduction Rhythms in European Populations of *Clunio marinus*. I. The Influence of Tidal Cycles of Mechanical Disturbance. In Naylor, E. and Hartnoll, R. G., editors, *Cyclic Phenomena in Marine Plants and Animals: Proceedings of the 13th European Marine Biology Symposium, Isle of Man, 27 September - 4 October 1978*, pages 423–433. Pergamon Press.

Palmén, E. and Lindeberg, B. (1959). The marine midge, *Clunio marinus* Hal.(Dipt., Chironomidae), found in brackish water in the northern Baltic. *Internationale Revue der gesamten Hydrobiologie und Hydrographie*, 44(1-4):383–394.

Reineke, A., Karlovsky, P., and Zebitz, C. P. W. (1998). Preparation and purification of DNA from insects for AFLP analysis. *Insect Molecular Biology*, 7(1):95–99.

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

Stinchcombe, J. R. and Hoekstra, H. E. (2008). Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, 100(2):158–170.

Wickham, H. (2020). *tidyr: Tidy Messy Data*. R package version 1.1.2.

Wickham, H., Francois, R., Henry, L., and Muller, K. (2020). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.2.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5):614–620.

5 General discussion

In section 1.6 I outlined the goals of my thesis, which were targeted in chapter 2, 3 and 4. I first investigated the abilities of short DNA fragments to recover the whole mitochondrial biogeography. This was followed by a detailed analysis of the studied *Clunio* populations, focusing on the genomic differences between the *Atlantic* and *Baltic ecotype* at the end. I identified ecotype-associated loci and the genes they affected. Finally, I conducted a QTL mapping to find phenotype-linked genetic markers in the genome of F2 offspring between the sympatric parental populations Ber-1SL and Ber-2AR. The results of the experiment pointed towards multiple significant loci linked to lunar-rhythmicity spread across different chromosomes. As all chapters extensively discuss their results, I will use the general discussion to focus on the compiled dataset, the putative molecular basis and the importance of arrhythmic strains in deciphering the enigmatic circalunar rhythms.

(5.1) Why the studied populations were chosen, what effects did this have on the results and what this dataset may have missed.

(5.2) Which implications have the identified ecotype-associated loci on the putative molecular function of circalunar rhythms in *Clunio* and are there alternative explanations for the identified loci.

(5.3) At the end, I want to provide conclusive remarks on the significance of the arrhythmic ecotypes and further research opportunities on lunar-arrhythmicity in *Clunio*.

5.1 Geographic range of lunar-arrhythmic ecotypes in the present dataset

The collected populations used for genetic and ecological evaluation in both articles represent a large geographic range covering distinct ecotypes. All five new populations were not discovered but established by myself. They were chosen from previous field excursions (Seh-2AR, first discovered by Tobias S. Kaiser in 2016) or from literature sources (Ar-2AR, found by Per Brink in Olander and Palmén (1968); Ber-2AR and Ber-1SL, described by Heimbach (1978); Tro-tAR, Neumann and Honegger (1969)). This decision was made to ensure the successful recovery and establishment of the targeted populations. In the course of the project I then established and raised all five strains to stable individual numbers for common garden experiments and different crossing experiments. To underline the general importance of my field work, two of the established populations are

currently also used by other projects in the research group. Previous studies found with larger geographic distance between populations genetic differentiation increases between the *Clunio* populations (Kaiser et al., 2010). As expected, strong geographic separation is present in both nuclear and mitochondrial genomes of all populations, as shown in section 2 and 3. This could cause problems with the discovery of ecotype-associated loci in a geographically distant dataset if not taken into account. Two valuable sample populations are the sympatric populations to identify the ecotype-specific genomic changes without the geographic background.

The overrepresentation of the *Baltic ecotype* in the present dataset is partially caused by the extensive study on Baltic Sea Chironomidae by previous authors, and the lack of collected material between the locations near Bergen and Tromsø. It is known that *Clunio* populations have a locally adapted reproductive timing (see section 1.3; Kaiser (2014); Kaiser et al. (2011, 2010); Neumann (1966, 1986))and populations are often found by chance in suitable locations. This becomes obvious from the vast variety of *Clunio* populations in the Brittany (Kaiser et al., 2021). Not just different timing strains, but two ecotypes can occupy the same location possibly by the larvae inhabiting different depths as the sympatric ecotypes from Bergen show (Heimbach, 1978; Neumann, 1966). In contrast to the *Atlantic* and *Baltic ecotype* larvae which inhabit algae patches on rocks (Neumann, 1966; Olander and Palmén, 1968; Remmert, 1955), the *Arctic ecotype* is predominantly found on sandy mud flats (Neumann and Honegger, 1969). Among the known *C. marinus* ecotypes, this substrate switch is unique.

Taken together, there is a realistic chance to find more *Arctic ecotypes* below the Arctic cycle in the same area as an *Atlantic ecotype* population, given that the *Baltic ecotype* could successfully migrate to the *Atlantic ecotype's* habitat as well. In fact, during my field trip in Tromsø, I took a two-day trip 320 km south to Bodø and discovered an undescribed *Clunio* population. But this population was not included in my thesis, as no laboratory strain could be established and the emergence could not be tracked properly in the field with insufficient emergence numbers. I discovered on a separate trip to Trondheim another *Clunio* population and was able to establish a laboratory culture. Due to time limitations, the population could not be integrated into the analyses. Emergence under laboratory conditions suggested an hourglass-timer-controlled circatidal emergence similar to the Tromsø population (Fuhrmann, unpublished).

The identified ecotype-associated loci from section 3 show not just a circadian clock involvement in the ecotype formation, but also direct responses to habitat adaptations (e.g. "Response to hypoxia" and "Sodium ion transport"; see section 3.3; see Fig. 10). As in contrast to the *Baltic ecotype* no circadian or circalunar rhythm is observable in the Tro-tAR population (see section 3.3; see Fig. S2E-H), ecotype-associated variants could identify additional or different regions correlating to lunar-arrhythmicity in the *Arctic ecotype*. But additional populations of the *Arctic ecotype* should be included to identify ecotype-associated loci and rule out locally adaptive loci of the Tro-tAR population.

Presented evidence from divergent population structure in PC1-4 (see Fig. 8A-B) and the local adaptive radiation of mitochondrial haplotypes from a uniform ancestral haplotype (see Fig. 3B and Fig. 8D) point towards an independent loss of lunar-arrhythmicity in *Baltic* and *Arctic ecotypes*. But population admixture with few genetic groups suggests a Baltic movement towards the Norwegian coastline. The question remains if the PCA and mitochondrial data are heavily influenced by the geographic separation, hiding a Baltic introgression or migration directed towards the Arctic habitat. As this scenario would question the independence of the *Arctic ecotype's* lunar-arrhythmicity, an additional population demography analysis was considered. Using the sequentially Markov coalescent (SMC) for example could infer the approximate time of the migration split towards the Norwegian coast and Baltic Sea before the ecotype formation. Then a potential later introgression from a *Baltic ecotype* into the lunar-rhythmic Ber-1SL to form the sympatric Ber-2AR there. And finally, a potential contact between the *Arctic* and the *Baltic ecotype* after the initial migration or alternatively the independent loss of lunar-rhythmicity in arctic habitats (Dutheil, 2020; Terhorst et al., 2017). However, as the correct recovery of the recombination breakpoints is critical, our scaffolded reference genome lacks long connected genomic stretches and could mislead the analysis or cause false positives. Furthermore, geological analyses (Patton et al., 2017) and sediment cores (Hofmann and Winn, 2000) suggest a migration event less than 10,000 years ago. As SMC analyses are more reliable and of a higher resolution, the further the analysis reaches back in time. Although the algorithm implemented in SMC++ (Terhorst et al., 2017) provided low error rates in more recent time ranges, both issues of the fragmented *Clunio* genome and the time frame we expect for the population split led to the decision to drop the reconstruction of the population demography in this thesis.

5.2 The putative molecular basis of circalunar clocks in *Clunio marinus*

While the results of this thesis could provide first candidates linked to lunar-arrhythmicity, the molecular function or existence of circalunar clocks in the studied organisms remains a mystery. This is a solid starting point for functional annotations and targeted knockouts. But here I want to discuss the implications and alternative explanations for the identified genes. Among the ecotype-associated hits from section 3.3 were circadian clock genes like *period*, *cycle*, *clock* and *timeless* and others. This suggests a scenario in which the circadian clock may be involved in the circalunar rhythm, as already discussed in section 3.4. The identified candidate genotypes could be targeted by genome editing (e.g. CRISPR gene editing; Bak et al. (2018)) or transcription silencing via RNAi (Cerutti and Casas-Mollano, 2006) to assess their involvement in the lunar-rhythmicity in *Clunio*.

Another possible and less exciting explanation to the ecotype-association of circadian clock genes could be the ecotype specific diversification rate of the circadian rhythm. As explained in section 1.4, the *Atlantic ecotype* is characteristic for the local adaptation of the lunar day and daytime emergence distributions. The daily emergence time is strongly linked and likely selected to one of the low tides on the circalunar determined days. The time of the low tides varies along the coastline with the progressing tidal waves (see Fig. S2B-D; Kaiser et al. (2011)). The *Baltic ecotype* on the other hand has a fixed circadian emergence around dusk and into the first hours of darkness (see Fig. S2F-H; Heimbach (1978)). Without largely exposed intertidal substrate and the ability to lay the egg clutches on the open water, the *Baltic ecotype* is not bound to emerge into the spring low tides. While adult insects can emerge every day and find a mating partner, the average life-span of few hours imposes quite a strong selection on a uniform circadian emergence time. The transition from high to low light intensities at dusk is a similar suitable synchronization time and is not locally specific as the diel occurrence of the spring low tides. This difference in circadian variability and fixation between the *Atlantic* and *Baltic ecotype* could possibly cause the ecotype-association of circadian control, while unidentified genes of the circalunar control could remain hidden as lesser associated. The major weak point of this hypothesis is that previous quantitative trait loci (QTL) mapping between different timing strains of the *Atlantic ecotype* did not pick up any of the core circadian clock genes identified in my thesis (Kaiser et al., 2016). Suggesting circadian

timing variation is maybe not linked to genetic variability in the core circadian clock genes.

Furthermore, an ecotype-association for genotypes between all identified lunar-arrhythmic populations would provide much more insight into the loss of circalunar rhythms in Northern European *C. marinus* ecotypes. Specifically interesting is the switch from a circalunar-circadian-emergence to a circatidal emergence pattern in the *Arctic ecotype* (see section 1.5). In contrast to the behavioral egg deposition and abiotic factor adaptations of the *Baltic ecotype*, the *Arctic ecotype* has a much more similar ecology than the *Atlantic ecotype*. The only differences being the circatidal emergence and the switch in larval substrate. In *Drosophila melanogaster*, pupation is strictly hormone regulated and timed by the decline of a transcription repressor early on in puparium formation (Akagi et al., 2016). A similar timer system is observable in *Clunio*. The pupation start is set by the circalunar rhythm 5-6 days before the emergence of the adult insects (Neumann, 1966, 1986). On the day of emergence a timer is set for the circadian emergence into the correct daytime (e.g. daytime of the low tide). This is an ancestral hourglass-timer-system linked to the circadian clock to time the emergence of the adult insects from the pupa.

It was proposed before, that the circatidal timer of the *Arctic ecotype* could be a fragment of the circadian oscillator, a downstream and independent timing process of the "bypassed" clock or a novel timing system evolved in the *Arctic ecotype* independently (Pflüger, 1973). A novel timing mechanism in this ecotype is highly unlikely, due to the possible timer regulated circadian emergence in the *Atlantic ecotype*. If the circatidal timer is a functional part of the circadian clock, all other circadian processes present in the other ecotypes could be missing or possibly distorted. However, it is known from other organisms under polar day conditions, that circadian rhythms can be maintained by other environmental cues than diel cycles (Ware et al., 2020; Williams et al., 2015). And that the light intensity in the habitat of *Clunio* still changes with the diel cycle between 1 and 100 kLux (Pflüger, 1973). Therefore, disentangling the circadian clock from the circatidal timer is possibly very difficult, as the only known circadian behavior in *Clunio* so far was the emergence of adult insects. Behavioral experiments are very limited without further investigations into circadian larval behavior (e.g. migration, feeding, sleep).

A good way to control the functionality of the circadian clock would be through the identification of cycling transcripts of core clock genes or known global neuronal modulators for sleep from *D. melanogaster* (e.g. *Shaker*, *sleepless*; Dubowy and Sehgal (2017)). Alternatively, if the clock itself is unlinked from the circatidal timer and has no function

in the *Arctic ecotype* at all, it is likely drifting unlinked from any selective force and the core genes would show high variability within populations or between them. For further investigations into the loss of circalunar rhythms in general and the circatidal rhythm specifically, additional *Arctic ecotype* populations should be added to the dataset.

5.3 An outlook to the research on lunar-arrhythmicity in *Clunio marinus*

There are two strategies to decode the molecular function of complex systems like biological rhythms. A reverse genetic approach works through the identification of core components and their effect on the whole system. This can mean, aiming at a specific target and taking it out or suppressing it, to break or alter the system. A prominent example is the discovery of the first core component of the circadian clock through mutants of the *period* gene in *D. melanogaster* (Hardin et al., 1990). However, rhythmic systems as biological clocks are nested within much more complex system networks and identifying specific putative targets can be a great challenge. Therefore, naturally occurring arrhythmic relatives of rhythmic organisms provide a point of comparison to identify these targets. Such a target-unspecific comparison of closely related organisms with different rhythmic phenotypes is the opposite strategy, a forward genetic approach. In chronobiology, the second discovered core circadian clock gene *timeless* was discovered through forward genetic screening (Sehgal et al., 1994).

In the presented work I used the latter strategy to tackle the capability of *C. marinus* to link their reproduction to the changing tidal regime over the synodic month. This first genomic comparisons between the lunar-rhythmic and lunar-arrhythmic ecotypes provided evidence for a circadian involvement in the formation of the studied ecotypes. While I discussed the implications of these results in section 5.2, it does visualize the strength of comparing naturally occurring opposite phenotypes to get substantially closer to the function of enigmatic systems. Nonetheless the presented work is only the starting point to identify what gave rise to the lunar-arrhythmicity both in the Baltic Sea and the high arctic latitudes.

The ecotype-associated hits from my analysis were containing neuronal development and responses to abiotic factors besides circadian control. Further loci related to lunar-arrhythmicity could likely be hidden under the genetic noise of the adaptation to the Baltic Sea habitat. Also, the BayPass (Gautier, 2015) thresholds were set higher to reduce the

results to the most significant hits. Additional genome screens should be considered to find other potential candidates. As mentioned in section 5.1 and 5.2, the ecotype associations have to be extended to incorporate the *Arctic ecotype* and its adaptations as well. More populations would need to be collected, sequenced and added to the genomic dataset. Only if all ecotypes are geographically diverse and represented in equal numbers, more specific candidates involved in the loss of lunar-rhythmicity can be reliably identified. To filter ecotype specific adaptations aside from the rhythmic control, several independent calculations with varying ecotype grouping could be considered (e.g. *Arctic* compared to *Atlantic* and *Baltic ecotype*).

Finally, one further experiment with the *Arctic* or *Baltic ecotype* alone is thinkable. A selection experiment for a specific circalunar emergence range on the arrhythmic populations over many generations could impose a selective pressure to regain lunar-rhythmicity. Ideally, the selection would be applied in combination with turbulence entrainment and due to the 15-day period of the turbulence cycle, also collected in circasemilunar periods. If imposed artificial rhythmic selection is able to turn an arrhythmic population back to the ancestral rhythmic state is an open question which was never tested before in *Clunio*.

It should be noted, that the establishment of the populations originated from field-caught copulating pairs in most cases over a short range of days or randomly drawn time points. But no initial periodicity was observed in any arrhythmic populations' first laboratory generations. As the establishment of the laboratory culture is a strong bottleneck on the gene pool, a scenario in which randomly drifting phases of lunar-rhythmicity are overlapping and mimic lunar-arrhythmicity in the field is unlikely. In the case of a successful selection of a lunar-rhythmic laboratory type from a lunar-arrhythmic ecotype, both the laboratory type and the original ecotype population could be used for QTL mapping and exactly pinpoint which genomic regions caused this evolution. And both geographic separation and other ecotype differences would be absolutely neglectable.

An arrhythmic phenotype in *D. melanogaster* played a key role to identify one of the core circadian clock gene *timeless* (Sehgal et al., 1994). In a similar forward genetic approach this thesis highlighted how naturally occurring lunar-arrhythmic *Clunio* populations set the path to understanding the enigmatic circalunar rhythms. The presented cases of lunar-arrhythmicity in arctic and tide-free habitats have potentially much more genomic adaptations to offer for future chronobiological research into *Clunio marinus*.

5.4 References

Akagi, K., Sarhan, M., Sultan, A.-R. S., Nishida, H., Koie, A., Nakayama, T., and Ueda, H. (2016). A biological timer in the fat body comprising Blimp-1, βFtz-f1 and Shade regulates pupation timing in *Drosophila melanogaster*. *Development*, 143(13):2410–2416.

Bak, R. O., Gomez-Ospina, N., and Porteus, M. H. (2018). Gene Editing on Center Stage. *Trends in Genetics*, 34(8):600–611.

Cerutti, H. and Casas-Mollano, J. A. (2006). On the origin and functions of RNA-mediated silencing: from protists to man. *Current Genetics*, 50(2):81–99.

Dubowy, C. and Sehgal, A. (2017). Circadian Rhythms and Sleep in *Drosophila melanogaster*. *Genetics*, 205(4):1373–1397.

Dutheil, J. Y. (2020). Towards more realistic models of genomes in populations: the Markov-modulated sequentially Markov coalescent. *arXiv preprint arXiv:2010.08359*.

Gautier, M. (2015). Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics*, 201(4):1555–1579.

Hardin, P. E., Hall, J. C., and Rosbash, M. (1990). Feedback of the *Drosophila* period gene product on circadian cycling of its messenger RNA levels. *Nature*, 343(6258):536–540.

Heimbach, F. (1978). Sympatric Species, *Clunio marinus* Hal. and *Cl. balticus* n. sp.(Dipt., Chironomidae), Isolated by Differences in Diel Emergence Time. *Oecologia*, 32(2):195–202.

Hofmann, W. and Winn, K. (2000). The Littorina Transgression in the Western Baltic Sea as Indicated by Subfossil Chironomidae (Diptera) and Cladocera (Crustacea). *International Review of Hydrobiology*, 85:267–291.

Kaiser, T. S. (2014). Local Adaptations of Circalunar and Circadian Clocks: The Case of *Clunio marinus*. In Numata, H. and Helm, B., editors, *Annual, Lunar, and Tidal Clocks*, pages 121–141. Springer.

Kaiser, T. S., Neumann, D., and Heckel, D. G. (2011). Timing the tides: Genetic control of diurnal and lunar emergence times is correlated in the marine midge *Clunio marinus*. *BMC Genetics*, 12(1):1–12.

Kaiser, T. S., Neumann, D., Heckel, D. G., and Berendonk, T. U. (2010). Strong genetic differentiation and postglacial origin of populations in the marine midge *Clunio marinus* (Chironomidae, Diptera). *Molecular Ecology*, 19(14):2845–2857.

Kaiser, T. S., Poehn, B., Szkiba, D., Preussner, M., Sedlazeck, F. J., Zrim, A., Neumann, T., Nguyen, L.-T., Betancourt, A. J., Hummel, T., Vogel, H., Dorner, S., Heyd, F., von Haeseler, A., and Tessmar-Raible, K. (2016). The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature*, 540(7631):69–73.

Kaiser, T. S., von Haeseler, A., Tessmar-Raible, K., and Heckel, D. G. (2021). Timing strains of the marine insect *Clunio marinus* diverged and persist with gene flow. *Molecular Ecology*, 30(5):1264–1280.

Neumann, D. (1966). Die lunare und tägliche Schlüpfperiodik der Mücke *Clunio* - Steuerung und Abstimmung auf die Gezeitenperiodik. *Zeitschrift für Vergleichende Physiologie*, 53(1):1–61.

Neumann, D. (1986). Chapter 1 Life cycle strategies of an intertidal midge between subtropic and arctic latitudes. In Taylor, F. and Karban, R., editors, *The Evolution of Insect Life Cycles*, pages 3–19. Springer.

Neumann, D. and Honegger, H. W. (1969). Adaptations of the intertidal midge *Clunio* to arctic conditions. *Oecologia*, 3:1–13.

Olander, R. and Palmén, E. (1968). Taxonomy, ecology and behaviour of the northern Baltic *Clunio marinus* Halid.(Dipt., Chironomidae). In *Annales Zoologici Fennici*, volume 5, pages 97–110. Societas Biologica Fennica Vanamo.

Patton, H., Hubbard, A., Andreassen, K., Auriac, A., Whitehouse, P. L., Stroeven, A. P., Shackleton, C., Winsborrow, M., Heyman, J., and Hall, A. M. (2017). Deglaciation of the Eurasian ice sheet complex. *Quaternary Science Reviews*, 169:148–172.

Pflüger, W. (1973). Die Sanduhrsteuerung der gezeitensynchronen Schlüpfrhythmik der Mücke *Clunio marinus* im arktischen Mittsommer (Hour-Glass Control of the Tidal Rhythm of *Clunio marinus* (Chironomidae) in Adaptation to Arctic Conditions). *Oecologia*, pages 113–150.

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

Sehgal, A., Price, J. L., Man, B., and Young, M. W. (1994). Loss of circadian behavioral rhythms and per RNA oscillations in the *Drosophila* mutant *timeless*. *Science*, 263(5153):1603–1606.

Terhorst, J., Kamm, J. A., and Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature genetics*, 49(2):303–309.

Ware, J. V., Rode, K. D., Robbins, C. T., Leise, T., Weil, C. R., and Jansen, H. T. (2020). The Clock Keeps Ticking: Circadian Rhythms of Free-Ranging Polar Bears. *Journal of Biological Rhythms*, 35(2):180–194.

Williams, C. T., Barnes, B. M., and Buck, C. L. (2015). Persistence, Entrainment, and Function of Circadian Rhythms in Polar Vertebrates. *Physiology*, 30(2):86–96.

Contribution to the thesis

2 The importance of DNA barcode choice in biogeographic analyses – a case study on marine midges of the genus *Clunio*

Nico Fuhrmann (NF) and Tobias S. Kaiser (TSK) conceived the study and wrote the article. NF performed all molecular work, statistical analyses and figure preparation.

3 Polygenic adaptation from standing genetic variation allows rapid ecotype formation

NF did all field work, phenotyping, molecular work, crossing experiments, population genomic analyses and figure preparation, unless stated otherwise. NF also contributed to experimental design and writing of the manuscript. Celine Prakash analyzed SNP effects, performed the GO term enrichment analyses and contributed to figure preparation and writing. TSK conceived and designed the study, contributed to data analysis and figure preparation and drafted the manuscript.

4 QTL mapping of sympatric *Clunio* ecotypes unravel putative causing genetic loci for lunar-rhythmicity

NF and TSK planned the experiment, sequencing strategy and QTL analysis workflow. Kerstin Schaefer performed the DNA extractions and sample preparation for sequencing. NF performed the crossing experiment and collection of samples, conducted the statistical and QTL analysis, preprocessed the sequencing results, prepared the figures and wrote the chapter.

Acknowledgements

At first, I want to thank my direct supervisor Tobias S. Kaiser. Tobias formed the core idea of this diverse project and saw in me the right person to work on the tasks and accomplish its goals. Tobias thought me the practical knowledge from fieldwork to *Clunio* culture establishment and sparked my interest in scripting. I highly valued the trust, encouragement, independence but also critique you offered on my scientific way to where I am now.

My deepest gratitude goes to Eva H. Stukenbrock as well, who voluntarily became my formal supervisor and first referee for my dissertation and examination. Thank you for the scientific support, discussions and advice which helped me a lot throughout my PhD, but also for the great support during my final phase.

I also want to thank the other members of my thesis advisory committee Thomas Roeder and Oscar Puebla for the fruitful discussions on the project and my scientific development. For many helpful comments, discussions and suggestions throughout the entire project including the writing process I want to thank Diethard Tautz, Guy R. Reeves, Miriam Liedvogel and Chaitanya Gokhale as well. My gratitude goes out to all members of the Max Planck Research Group "Biological Clocks" who helped to make this project and my thesis possible. Here a special thanks to Dušica Briševac for all the years of experiment share, scientific discussions, openness and friendship.

I further want to mention and thank everyone who significantly supported me on my travels and fieldwork. Jürgen Reunert was a great help and excellent road trip companion on my last trip to Roscoff. More thanks goes out to Even Jørgensen and Lars Folkow of the Department of Arctic and Marine Biology at the UiT. Even gave me full support during my one month fieldwork in Tromsø, helping me to find accommodation, supplies and get around in the city. Lars offered me office space and a workplace at the department during my stay. During a short trip to Bodø Galice Hoarau was so kind to help me find an accommodation on very short notice. And Marvin Choquet kindly to drove and joined me on my search for an undocumented *Clunio* population. I finally would like to thank the staff at the Marine Biological Station Espegrend of the University of Bergen, the Askö Laboratory of the Stockholm University and the Ar Research Station of the Uppsala University for the friendly welcome, accommodation and the provided supplies.

Finally I would like to thank my whole family for all the support throughout the years. And a great thanks to all my friends which were always there for listening, help and support. A special thanks goes to my parents and grandparents. I would never have made

it that far without their support and help. Further I want to thank my partner Christian Hanne who stayed with me without question and gave me strength throughout all those years. Loukas Theodosiou was among the first persons Tobias introduced me to and I am very glad we developed a strong bond and had so many fun social events together. Muhammad Bilal Haider is among my best friends and I cannot imagine doing my PhD in Plön without him. I will never forget our long parties and I am very grateful that we could share these years. But we were never alone and I am very happy that I could share so many fun social nights with good friends like Michael Barnett, Joanna Summers, Devika Bhave, Demetris Taliadoros and Elena Damm. Many more names should be mentioned and every single made their contribution to the great time I had and helped to close this chapter in my life.

Affidavit

I hereby declare that this thesis

- And the contents were written and designed by myself under the guidance of my supervisor. Detailed contributions of me and all other authors are listed in the "Contribution to the thesis" section (5.3) of the thesis;

- Has not been submitted elsewhere partially or wholly as a part of a doctoral degree and no other materials are published or submitted for publication than indicated in the thesis;

- The work and thesis has been performed and prepared following the "Rules of Good Scientific Practice" of the German Research Foundation (DFG).

I further declare, that no academic degree has ever been withdrawn from me.

Plön,

Nico Fuhrmann

Supplementing Notes

Oviposition behavior in the *Baltic ecotype*

Because of the lack of tides in the Baltic Sea, the *Baltic ecotype* cannot rely on the tides to expose the larval substrates for egg deposition. Therefore, in contrast to the *Atlantic* and *Arctic ecotypes*, the *Baltic ecotype* does not oviposit on exposed larval substrates, but on the water surface, from where the eggs sink to the bottom of the sea. This change in oviposition preference is accompanied by a specific behavioral change, namely the bending of the female's abdomen down through the water surface (Endraß, 1976), which ensures that the egg masses are not caught in the water's surface tension. Additionally, the *Baltic ecotype's* egg jelly is reported to have slightly differing properties (Endraß, 1976).

In our laboratory culture, the animals are kept in plastic boxes with a constant water level, so that larval substrates are never exposed. Under these conditions, most of the *Baltic ecotype's* egg masses will be found submerged at the bottom of the culture box, as expected. Very rarely egg masses are deposited on the walls or the lid of the box. The *Atlantic* and *Arctic ecotypes* will – in the absence of exposed larval substrates – also oviposit on the water surface. However, as they lack the characteristic banding of the female abdomen during oviposition, their egg masses will be trapped in the water's surface tension and cannot sink to the bottom of the culture box. Their egg masses will always float on the water surface, usually with the female still sticking to the egg jelly. These floating egg masses often form aggregations on the water surface.

We used these differences in where the egg clutches are found in our laboratory cultures to additionally confirm ecotype identity of our laboratory strains.

References

Endraß, U. (1976). Physiologische Anpassungen eines marinen Insekts. II. Die Eigenschaften von schwimmenden und absinkenden Eigelegen. *Marine Biology*, 36(1):47–60.

Supplementing Tables

Table S1: Sampling sites for all examined *Clunio* populations.

Population Code	Location Name	Latitude	Longitude	Year	Collector
Ar-2AR (Sweden)	Ar, Gotland	57° 55' 06.0" N	18° 56' 19.0" E	2018	Nico Fuhrmann
He-2SL (Germany)	Helgoland	54° 11' 19.2" N	7° 52' 10.3" E	2005	Tobias S. Kaiser
Por-1SL (France)	Port-en-Bessin	49° 21' 00.0" N	0° 45' 10.0" W	2009	Tobias S. Kaiser
Seh-2AR (Germany)	Sehlendorf	54° 18' 29.0" N	10° 43' 30.0" E	2017	Nico Fuhrmann, Tobias S. Kaiser
Tro-tAR (Norway)	Tromsø	69° 41' 04.0" N	18° 53' 59.0" E	2018	Nico Fuhrmann

Table S2: Sampling sites and sampling campaigns for the five newly established laboratory strains in this study.

On sampling dates in squared brackets egg clutches for setting up laboratory cultures were collected from copulating pairs on the water surface. Samples from underlined dates were used for genomic analyses of the wild populations.

Sampling Site	Coordinates		Ecotype	Laboratory Strain	Field Samples	Founding Egg Clutches
Sehlendorf (Germany)	54° 18' 29" N	10° 43' 30" E	Baltic	Seh-2AR	[05/2017], [06/2017]	69
Ar (Gotland, Sweden)	57° 55' 06" N	18° 56' 19" E	Baltic	Ar-2AR	[08/2018]	36
Kviturdvikpollen (Norway)	60° 15' 58" N	05° 14' 52" E	Baltic	Ber-2AR	06/2017, [08/2018]	40
Kviturdvikpollen (Norway)	60° 15' 58" N	05° 14' 52" E	Atlantic	Ber-1SL	06/2017, [08/2018]	60
Tromsø (Norway)	69° 41' 04" N	18° 53' 59" E	Arctic	Tro-tAR	[06/2018], [07/2018]	47

Table S3: Volume modifications for the REPLI-g Mini Kit QIAGEN 150025 whole genome amplification kit (QIAGEN).

MM - Master Mix.

Phases	template DNA	Buffer D1	Buffer N1	MM: Nuclease-free water	MM: REPLI-g Mini Reaction Buffer	MM: REPLI-g Mini DNA Poly-merase	MM: Adding total volume
Manu-facturers' Volume (suitable for 15 reactions)	2.5 µl	2.5 µl	5 µl	10 µl	29 µl	1 µl	40 µl
Edited Volume	7.5 µl	1 µl	1.5 µl	0 µl	14.5 µl	0.5 µl	15 µl

Table S4: Wilcoxon rank sum test with continuity correction for significant differences in nucleotide diversity (π) between populations, based on the arithmetic mean of 200 kb genomic windows.

The lower-left part of the table shows the p-values and the upper-right part the corresponding significance levels: **** - 0-0.0001; *** - 0.0001-0.001; ** - 0.001-0.01; * - 0.01-0.05; ns - 0.05-1.

	Por-1SL	He-1SL	Ber-1SL	Ber-2AR	Seh-2AR	Ar-2AR	Tro-tAR
Por-1SL	-	ns	ns	ns	****	****	****
He-1SL	0.06925	-	**	ns	****	****	****
Ber-1SL	0.3596	0.006378	-	ns	****	****	****
Ber-2AR	0.619	0.1703	0.1576	-	****	****	****
Seh-2AR	$5.63e^{-10}$	$4.87e^{-06}$	$1.74e^{-13}$	$1.93e^{-09}$	-	***	ns
Ar-2AR	$2.2e^{-16}$	$1.74e^{-15}$	$2.2e^{-16}$	$2.2e^{-16}$	0.0001517	-	*
Tro-tAR	$5.7e^{-15}$	$5.68e^{-10}$	$2.2e^{-16}$	$1.27e^{-14}$	0.06954	0.03893	-

Table S5: SnpEff analysis for selected and all variants.

SnpEff was run on the selected BayPass variants in association with the ecotype and all variants. For each analytical group (Region, Impact, Function, Variant) the total numbers for each subgroup and the fractions in respect to the entire analytical group are given. Additionally, the p-values for significant deviation between the selected and the entire genome variants were calculated using Fisher's exact test.

	Ecotype-associated variants (n=4,741)		All variants (n=948,128)		
	n	%	n	%	p-value
Region					
Exon	1061	19.06	239000	20.69	0.002538563
Intron	1821	32.71	389723	33.74	0.1053189
UTR	418	7.51	98365	8.52	0.0069941
Intergenic	2209	39.68	413274	35.78	$1.811437e^{-9}$
Other	58	1.04	14663	1.27	0.1490632
Impact					
High	31	0.56	10744	0.93	0.00251666
Low	612	10.99	141746	12.27	0.003404886
Moderate	479	8.6	101698	8.8	0.6185761
Modifier	4445	79.85	900837	77.99	0.0008337278
Function					
Missense	465	45.32	98560	43.06	0.1462238
Nonsense	2	0.19	1670	0.73	0.04045629
Silent	559	54.48	128647	56.21	0.2698506
Variant					
SNP	3890	82.05	792032	83.54	0.006363973
Indel	851	17.95	156096	16.46	0.006363973

Table S6: The three covariates used for association analysis in BayPass.

Covariates	Ar-2AR	Seh-2AR	Ber-2AR	Ber-1SL	He-1SL	Por-1SL
Ecotype (binary)	0	0	0	1	1	1
Sea Surface Salinity (PSU)	7	14.5	31.5	31.5	31.3	33.9
Average Water Temperature 2020 (°C)	7.5	9.1	8.9	8.9	10.1	11.8

Table S7: ANOVA table from manually fitted QTL model scans of lunar-subset 1.

Tests for additive effects are shown in one row per QTL. The test for two interactive QTLs shows the specified QTLs and their positions separated by a colon. p-value symbols: **** - 0-0.0001; *** - 0.0001-0.001; ** - 0.001-0.01; * - 0.01-0.05; ns - 0.05-1.

QTL Position(s)	*LOD* score	Variance explained (%)	p-value (χ^2)
Model: $y \sim Q1 + Q2 + Q3 + Q4$			
1@0.0	0.9243	2.108	0.1190 ; ns
2@30.0	0.6463	1.469	0.2258 ; ns
2@25.5	0.5442	1.235	0.2856 ; ns
3@110.0	1.2162	2.784	0.0608 ; ns
Model: $y \sim Q1 + Q2 + Q1 : Q2$			
1@50.0	2.019	4.780	0.1574 ; ns
2@25.0	4.017	9.762	0.0051 ; **
1@50.0:2@25.0	1.136	2.658	0.2644 ; ns
Model: $y \sim Q1 + Q3 + Q1 : Q3$			
1@0.0	2.533	6.079	0.0699 ; ns
3@105.0	3.454	8.391	0.0143 ; *
1@0.0:3@105.0	1.766	4.196	0.0868 ; ns
Model: $y \sim Q2 + Q3 + Q2 : Q3$			
2@30.0	4.151	9.802	0.00397 ; **
3@110.0	3.165	7.376	0.02384 ; *
2@30.0:3@110.0	1.740	3.979	0.09118 ; ns

Table S8: ANOVA table from manually fitted QTL model scans of lunar-subset 2.

Tests for additive effects are shown in one row per QTL. The test for two interactive QTLs shows the specified QTLs and their positions separated by a colon. p-value symbols: **** - 0-0.0001; *** - 0.0001-0.001; ** - 0.001-0.01; * - 0.01-0.05; ns - 0.05-1.

QTL Position(s)	*LOD* score	Variance explained (%)	p-value (χ^2)	
Model: $y \sim Q1 + Q2 + Q3 + Q4 + Q5 + Q6$				
1@0.0	1.5223	2.959	0.03004	*
1@60.0	0.9530	1.840	0.11144	ns
2@15.0	0.5104	0.980	0.30877	ns
2@24.9	2.1899	4.290	0.00646	**
3@15.0	1.6636	3.239	0.02170	*
3@95.0	1.3712	2.661	0.04254	*
Model: $y \sim Q1 + Q2 + Q1 : Q2$				
1@60.0	1.763	3.686	0.229649	ns
2@24.9	4.917	10.674	0.000924	***
1@60.0:2@24.9	1.014	2.102	0.322996	ns
Model: $y \sim Q1 + Q3 + Q1 : Q3$				
1@10.0	3.021	6.697	0.0306	*
3@100.0	3.302	7.343	0.0187	*
1@10.0:3@100.0	2.753	6.083	0.0130	*
Model: $y \sim Q2 + Q3 + Q2 : Q3$				
2@25.0	5.510	11.845	0.000291	***
3@5.0	2.479	5.141	0.076376	ns
2@25.0:3@5.0	1.553	3.186	0.128088	ns

Table S9: ANOVA table from manually fitted QTL model scans of lunar-subset 3.

Tests for additive effects are shown in one row per QTL. The test for two interactive QTLs shows the specified QTLs and their positions separated by a colon. p-value symbols: **** - 0-0.0001; *** - 0.0001-0.001; ** - 0.001-0.01; * - 0.01-0.05; ns - 0.05-1.

QTL Position(s)	*LOD* score	Variance explained (%)	p-value (χ^2)	
Model: $y \sim Q1 + Q2 + Q3$				
1@0.0	1.329	3.744	0.04689	*
2@23.3	2.437	6.987	0.00365	**
3@10.0	1.088	3.052	0.08173	ns
Model: $y \sim Q1 + Q2 + Q1 : Q2$				
1@0.0	2.115	6.060	0.1361	ns
2@20.0	3.074	8.944	0.0279	*
1@0.0:2@20.0	1.024	2.883	0.3180	ns
Model: $y \sim Q1 + Q3 + Q1 : Q3$				
1@65.0	2.903	8.383	0.0376	*
3@100.0	3.416	9.947	0.0153	*
1@65.0:3@100.0	2.390	6.847	0.0265	*
Model: $y \sim Q2 + Q3 + Q2 : Q3$				
2@23.3	3.932	11.363	0.00597	**
3@10.0	2.723	7.720	0.05092	ns
2@23.3:3@10.0	1.834	5.126	0.07657	ns

Table S10: Multiplex PCR primer pairs for QTL mapping between the sympatric populations Ber-1SL and Ber-2AR.

All primers were designed and tested by Kerstin Schaefer.

Chromosome	Scaffold	Primer name	Sequence 5'-3'	Product length
1	8	1.10_f	TTCCTGGAAATACTGCAGCAA	278 bp
		1.10_r	TCAACGAGATTTTCCGGTTTCA	
	6	2.1_f	ACACGAGATGAAACTCATGCG	909 bp
		2.1_r	CCAAGGTCTACGAAAAAGCAGC	
	53	3.2_f	CACGGTTAACATGCTTGAAGGA	484 bp
		3.2_r	TCAAGTTCCAGCGCATTCGAT	
	18	4.1_f	TGTGACTAACGGCAGGAAAGA	782 bp
		4.1_r	ACAAAACGATTGGAAGCGACG	
	17	5.2_f	AGGAAGTTTGCCACTCGGATA	378 bp
		5.2_r	TGCCGAACTCTTTATGGATCA	
	61	6.14_f	AGCGACACAGATGAAAGACGA	698 bp
		6.14_r	TGCGGCAATTTGTTTTAATGTCC	
2	20	7.9_f	ACCTCAATATGCACGTCCAGT	543 bp
		7.9_r	TCTGATATGGTCGCAAGTGGT	
	47C	8.3_f	CCCCAGCGCTTTTCACATTT	369 bp
		8.3_r	TTTCCTCGGTTCTTCGTCCC	
	4	9.5_f	AGCTCCACTATTCCAAACAGCT	727 bp
		9.5_r	GCCATCAGCCTTGAGAAATGC	
	50	10.1_f	TGCTTTAAGGGGTCGTGCAA	925 bp
		10.1_r	CACGGATGTCAGCTCAGTGA	
	47C	11.1_f	TGATGATGACGCTAGCAGCA	443 bp
		11.1_r	TGTGGTCATTTCAGAAACGGT	
	47F	12.7_f	GCCTTGCAATAGCGCATACG	849 bp
		12.7_r	TAGGCTCTGGGGTATGGCTT	
3	10	13.4_f	GTTCCTTGCCACTCATTTGCT	450 bp
		13.4_r	TGTCGCCAAACAGGATTTGTG	
	41A	14.2_f	TGGCAAATCAAGGACCAGACA	740 bp
		14.2_r	TCGACTTGAGGCATTATTGGGA	
	48	15.1_f	CTCGATTCAGACCGTCCGATT	639 bp
		15.1_r	GGCGTGACCTTCATGAAGAGA	
	52	16.1_f	AGGCAGCTCACTTAGAAGACG	499 bp
		16.1_r	TTGACTACCATGGCTGTGGAG	
	45B	17.6_f	TCTCACTTTTCACGACGCTTTT	308 bp
		17.6_r	TAGAAGCTGTTCCGCCTCAAG	
	3	18.8_f	CCACTGAAACCATCGGGACA	227 bp
		18.8_r	AGAATGTGTGACTGTTGCAAAGT	

Supplementing Figures

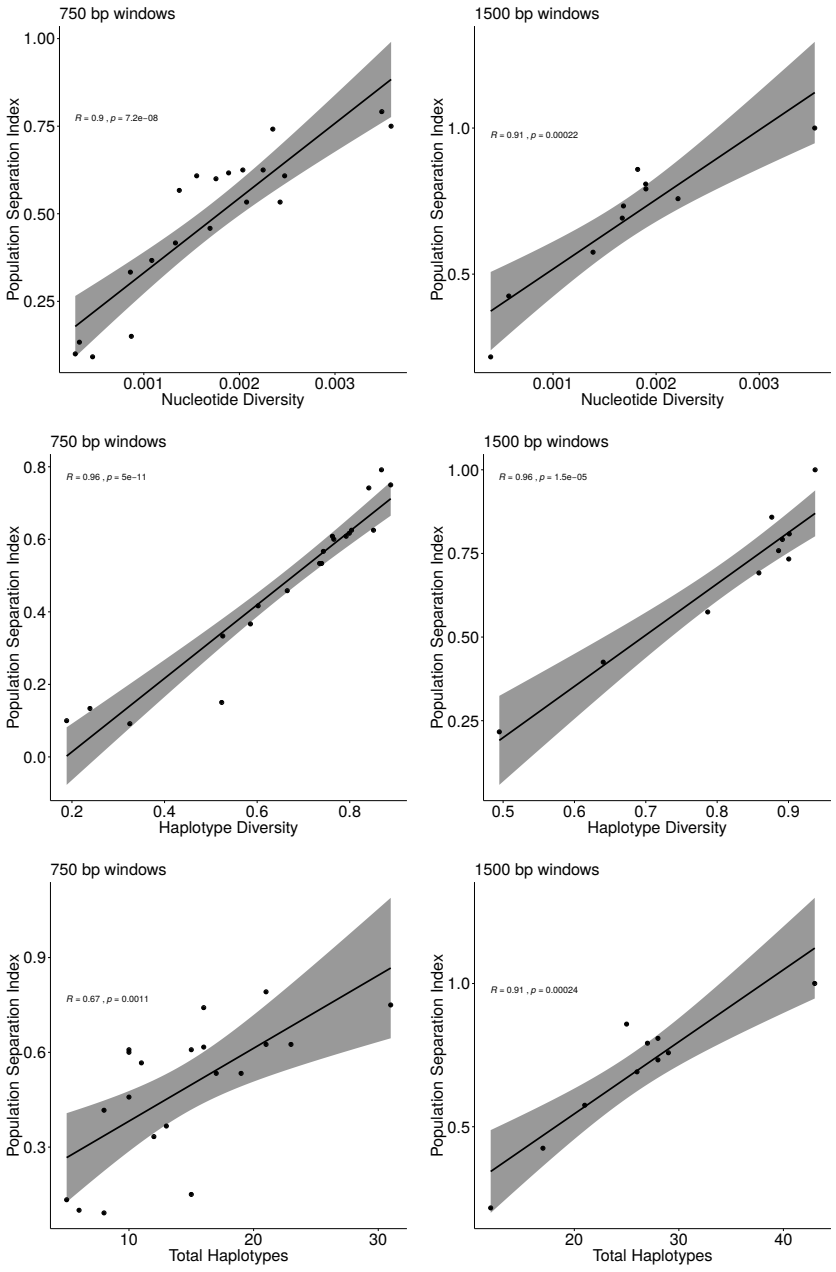

Figure S1: Correlation between measures of diversity and a Population Separation Index (PSI).

The PSI was calculated as the fraction of individuals in haplotypes that are not shared between populations. Correlations were calculated for 750 bp and 1,500 bp windows separately with Pearson's Product Moment correlation as implemented in R.

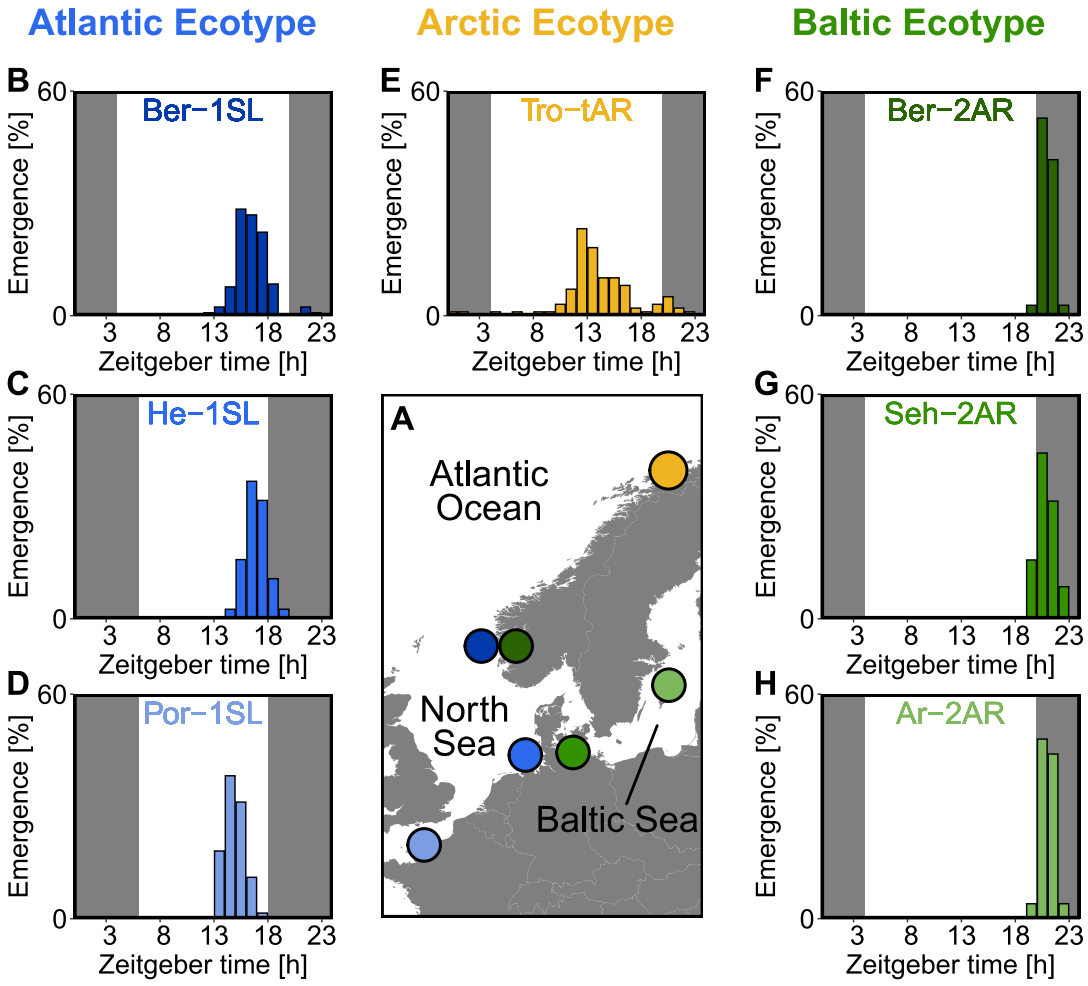

Figure S2: Circadian emergence rhythm of the studied *Clunio* strains under laboratory conditions.

(**A**) Geographic locations. (**B-H**) Circadian emergence rhythms of laboratory strains. Dark shading indicates the dark phase, the middle of dark phase is defined as zeitgeber time 0. Data for Ber-1SL (B; n=130), Ber-2AR (F; n=36) and Tro-tAR (E; n=99) was recorded with a custom-made fraction collector in 1-h intervals under an artificial light cycle with 16 h of light and 8 h of darkness (LD 16:8). Data for Seh-2AR (G; n=70) and Ar-2AR (H; n=25) was recorded manually while performing crosses. Data of He-1SL (C) and Por-1SL (D) was taken from Neumann (1966) and recorded under LD 12:12.

Figure S3: Lunar emergence patterns of crosses between the Ber-1SL and Ber-2AR strains.

Single pair crosses were set up between the Ber-1SL and Ber-2AR strains, resulting in several F1 families, two of which are shown here (F1-31 and F1-34). F2 families were obtained by letting the siblings within each F1 family mate freely with each other. Thus, from each F1 we obtained several F2 families, which all go back to a single pair of parents. F1-31 and F1-34 differ in the degree of lunar-rhythmicity, as do F2-31 and F2-34. When plotting the F2 families individually, it becomes apparent that in F2-31 all families are quite rhythmic. In contrast, in F2-34 there are highly rhythmic families (e.g. F2-34.10) and completely arrhythmic families (e.g. F2-34.3 and F2-34.4), suggesting that this F2 generation is segregating for rhythmicity alleles. The observation suggests that in the cross leading to F1-31/F2-31, the Ber-2AR parent carried a considerable fraction of rhythmic alleles (basically "Ber-1SL" alleles), which rendered the resulting F1 and F2 generations largely rhythmic. The Ber-2AR parent for the F1-34/F2-34 cross seemed to carry largely arrhythmic alleles, as expected, which then segregate in the F2. Overall, we conclude that the Ber-2AR strain seems to carry a certain fraction of lunar-rhythmic alleles. The segregation of lunar-rhythmicity not only within but also between F2 families suggests a heterogeneous polygenic architecture of the trait.

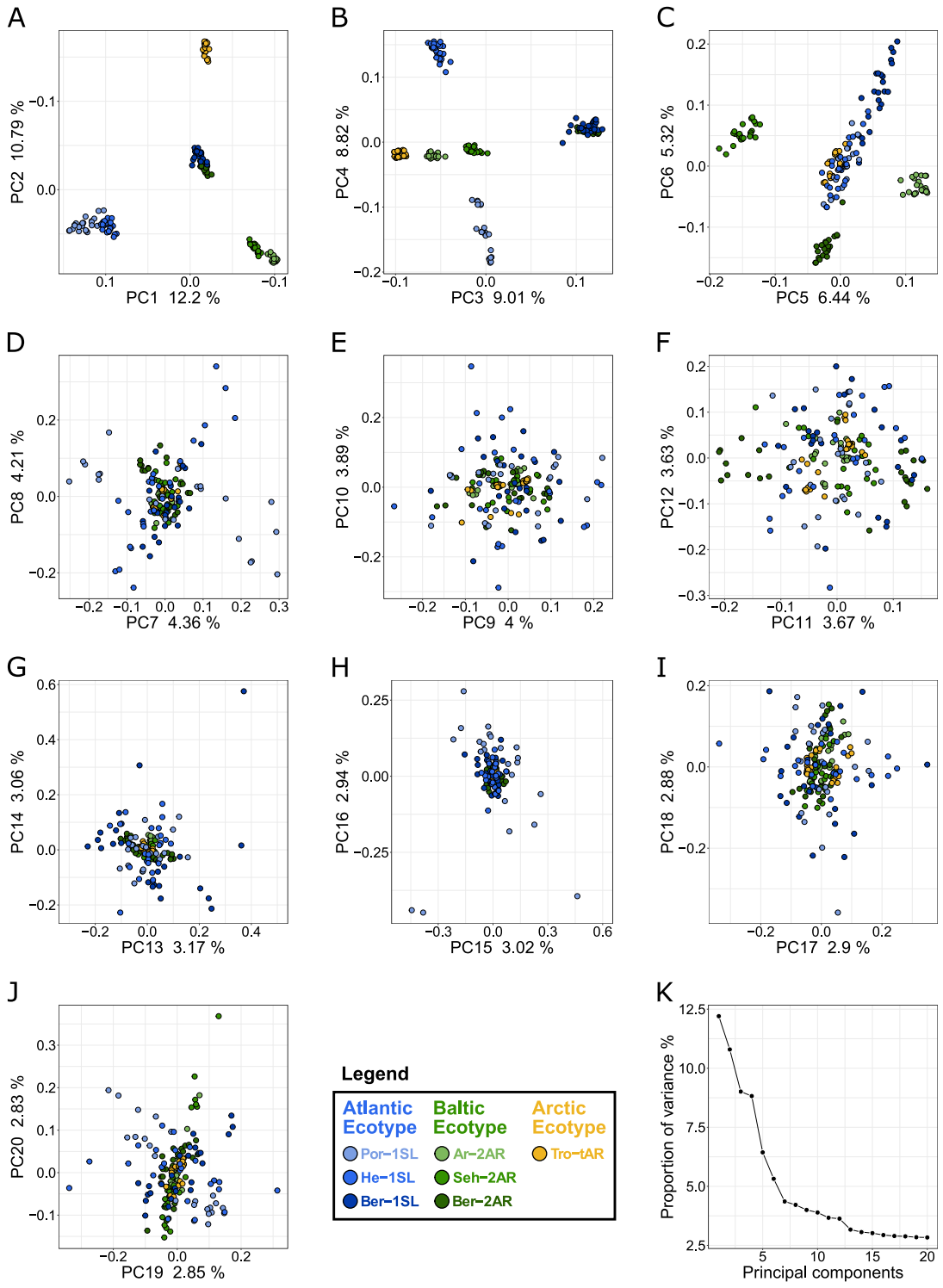

Figure S4: Principal component analysis (PCA) of all individuals for all seven populations.

(**A-J**) The principal components (PCs) are plotted in pairs from 1-2 (A) to 19-20 (J). The eigenvalue as fraction of total variance is given on the respective axes. The individuals are color coded according to population (see legend). (**K**) Overview of the eigenvalue as fraction of total variance for all 20 PCs.

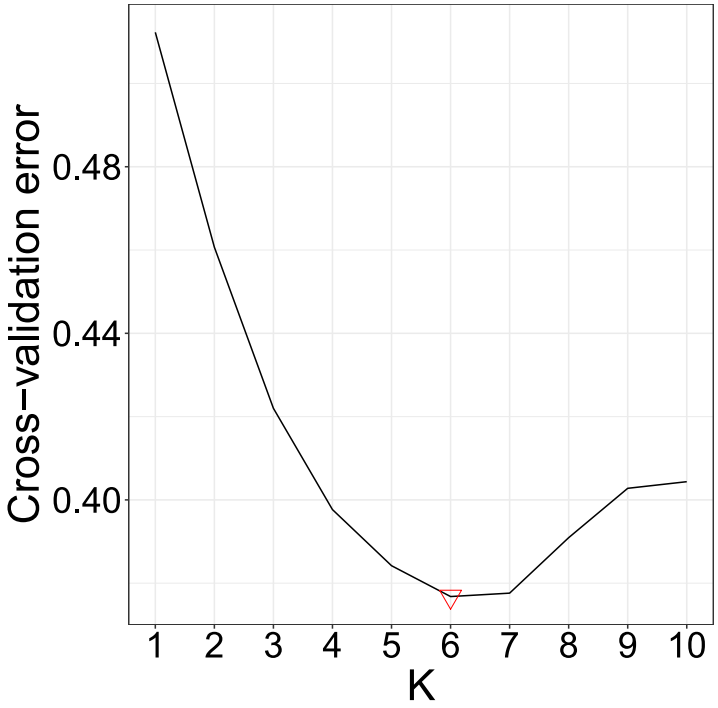

Figure S5: The results of the cross-validation test of the ADMIXTURE analysis.

The optimal K is marked with a red downward directed triangle.

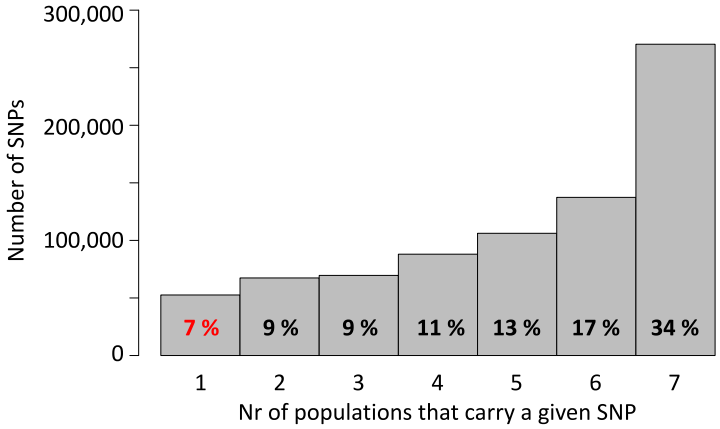

Figure S6: Detected single nucleotide polymorphisms (SNPs) are largely shared between the seven populations.

Of a total of 792,032 SNPs, 34% are found in all seven populations, whereas only 7% are private to one of the seven populations.

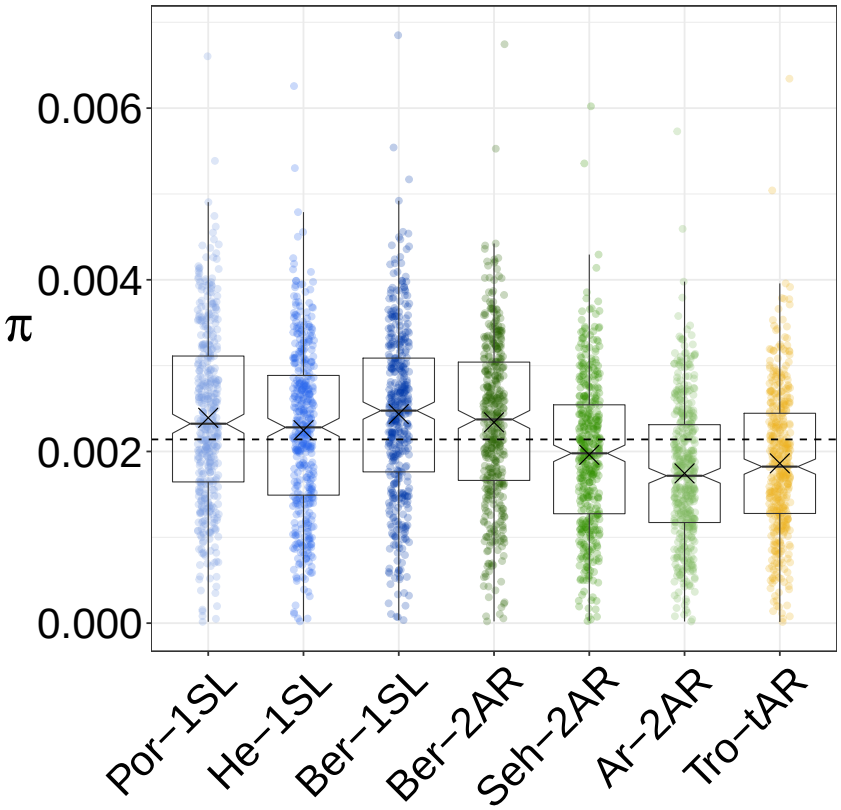

Figure S7: Nucleotide diversity π per population in 200 kb non-overlapping windows across the genome.

Populations are colour coded as in Fig. 7A and all individuals per population were used for the analysis. The boxplots show the median, 25^{th} and 75^{th} quantile, data maximum and minimum, and the outliers. The arithmetic mean is marked per boxplot as 'X' and the overall arithmetic mean as dashed line.

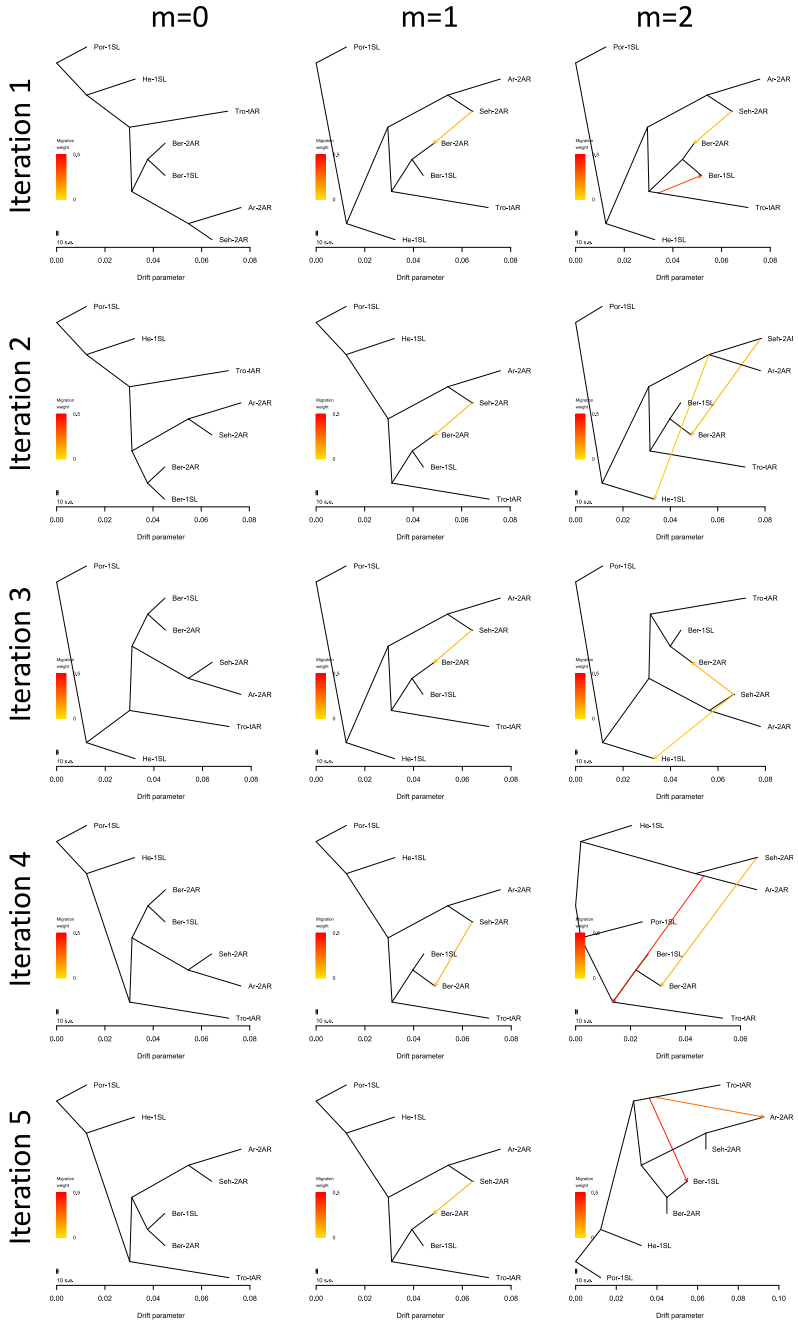

Figure S8: TreeMix analysis for introgression events.

TreeMix was run for no, one or two migration events with five iterations each. Analysis with zero migration events yields the same structure of historic population splits in all five iterations, congruent with the genome-wide consensus phylogeny (Fig. S10). With one migration event, the algorithm stably detects introgression from Seh-2AR into Ber-2AR. With two migration events, introgression from Seh-2AR into Ber-2AR is usually still detected (4 out of 5 iterations), but the second introgression event is random, suggesting there is no reliably detectable second introgression event.

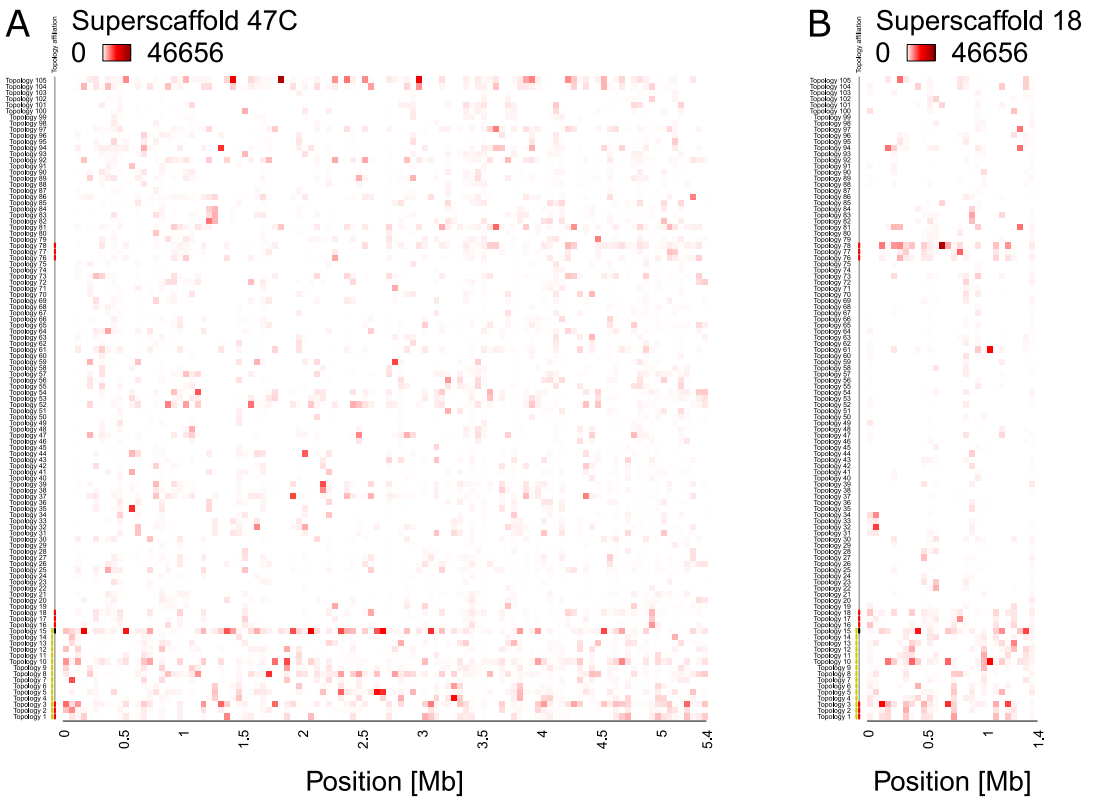

Figure S9: Incomplete lineage sorting, as illustrated by the relative support for 105 population topologies in 50 kb windows across superscaffolds 47C and 18.

The heatmap shows the assigned topology weight for each topology per window ranging from 0 (white, no support) to 46,656 (red, support by all trees). The supported topologies change quickly along the chromosome. In almost all windows, several topologies are supported, which implies that individuals do not cluster according to population, highlighting incomplete lineage sorting. Topology 15 (black mark on the left) is the whole-genome consensus population topology. Topologies marked in red are separating populations according to ecotypes. Topologies marked in yellow are consistent with introgression. (**A**) Superscaffold47C – the longest scaffold in the reference genome – gives an impression of the genomic background, with highest overall support for topology 15. (**B**) Superscaffold 18, from the middle of chromosome 1, is strongly enriched for ecotype-associated topologies (compare Fig. 9).

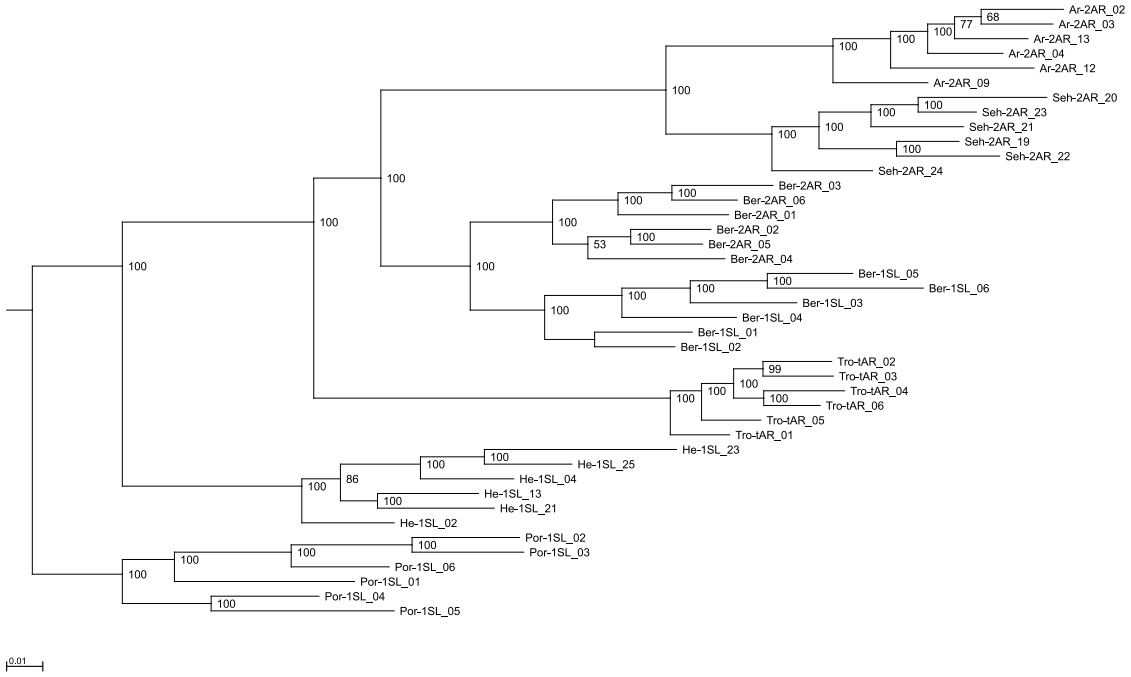

Figure S10: Genome-wide consensus genealogy for six individuals from each of the seven populations.

The same six individuals from each population were also used in the TWISST topology weighting analysis. The genealogy is based on the whole genome SNP sequences, composed of 792,032 positions. The SNPs were extracted from the VCF using the *vcf2phylip.py* v.2.3 script. The genealogy was calculated with IQ-TREE v.1.6.12 and graphically edited with Archaeopteryx v.0.968.

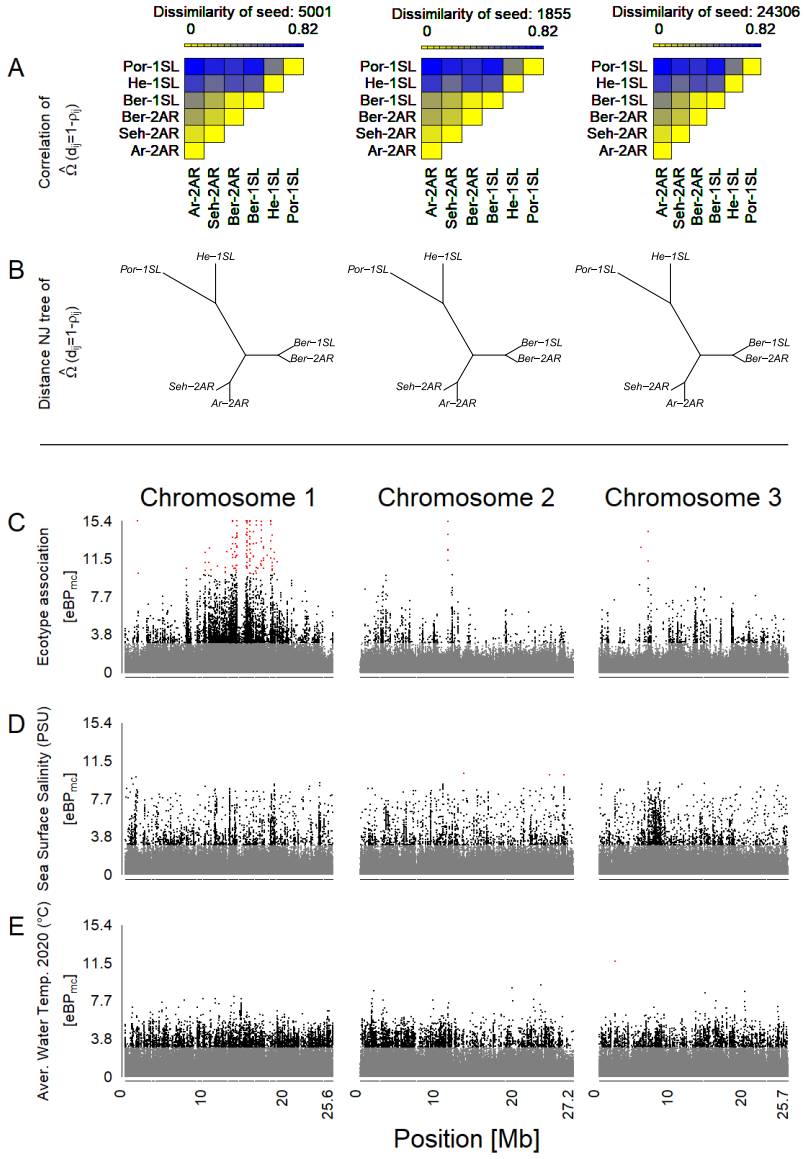

Figure S11: An overview of outlier and association analysis for ecotype (C), sea surface salinity (D) and water temperature (E).

(**A**) Before assessment of outlier loci and association, the dataset was corrected for population structure based on an Ω matrix, which is here visualized as a dissimilarity matrix. Three independent runs (seeds 5001, 1855 and 24306) converge. (**B**) Ω matrices can also be visualized as Neighbor Joining trees. (**C-E**) Association of genetic variants with ecotype (C), sea surface salinity (D) and water temperature (E) is assessed via the eBP_{mc} score and plotted along the three chromosomes, color-coded in grey ($eBP_{mc} < 3$), black ($3 \leq eBP_{mc} < 10$) and red ($eBP_{mc} \geq 10$).

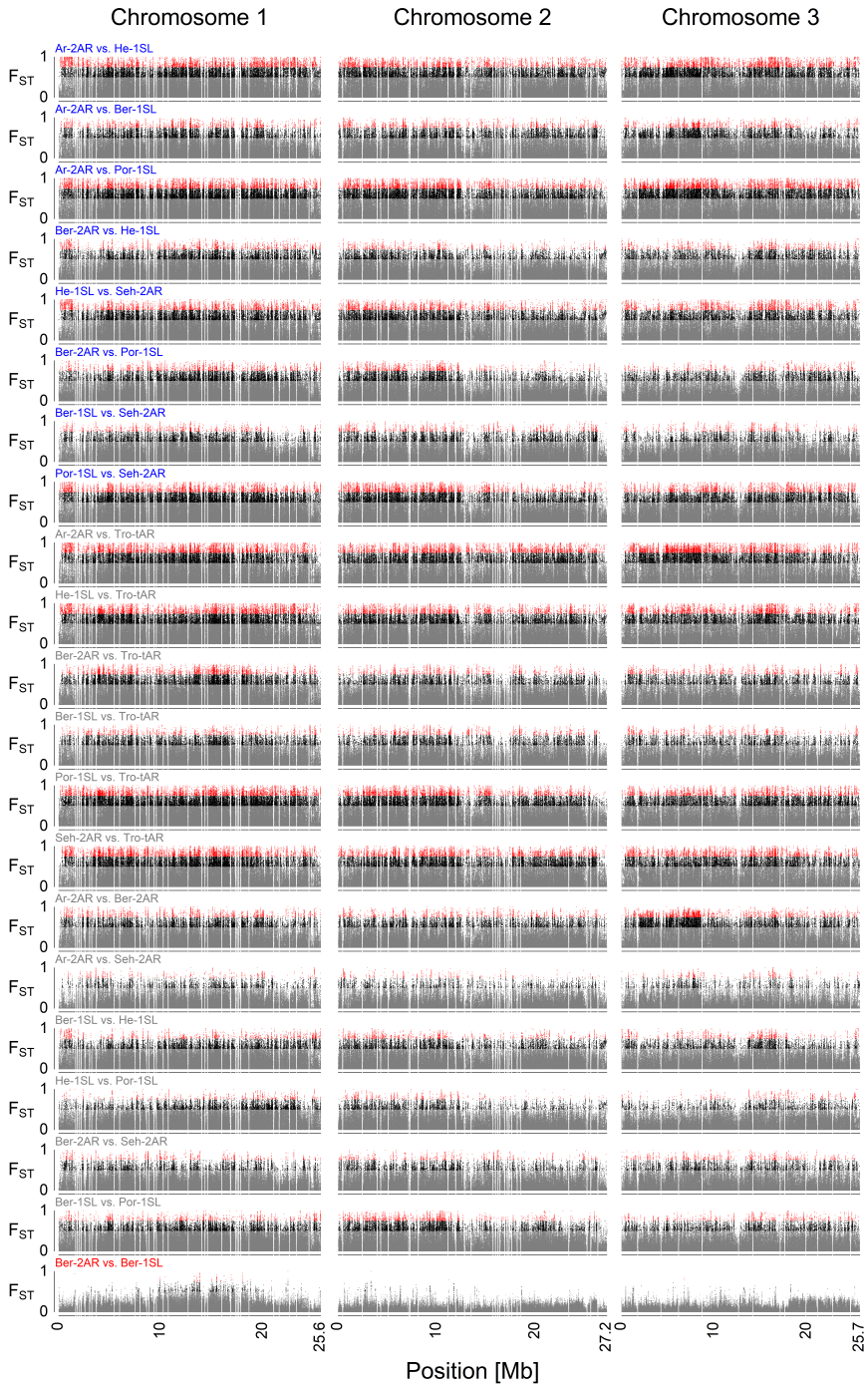

Figure S12: Pairwise F_{ST} comparisons between all seven populations.
The specific populations of the comparison are written above each plot, color-coded for specific comparisons. The comparison between the sympatric Ber-2AR and Ber-1SL ecotypes is plotted in red. Other comparisons of *Atlantic* vs. *Baltic ecotype* populations are plotted in blue. All other comparisons are plotted in grey. The F_{ST} values are colour coded in grey ($F_{ST} < 0.5$), black ($0.5 \leq F_{ST} < 0.75$) and red ($F_{ST} \geq 0.75$). The sympatric ecotypes from Bergen are much less differentiated than all other populations.

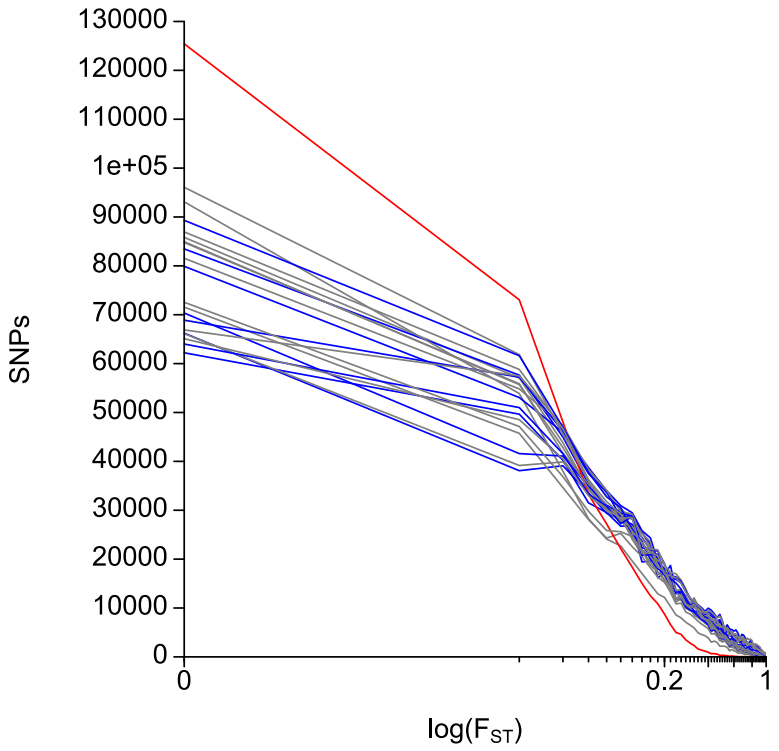

Figure S13: Distribution of F_{ST} values in all pairwise population comparisons. F_{ST} values for 792,032 SNPs were categorized in 0.02 bins. The comparison between the sympatric Ber-2AR and Ber-1SL ecotypes is plotted in red. Other comparisons of *Atlantic* vs. *Baltic ecotype* populations are plotted in blue. All other comparisons are plotted in grey.

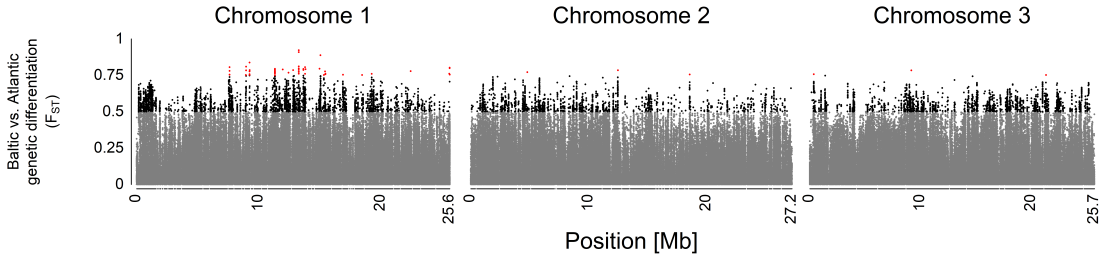

Figure S14: Genetic differentiation (F_{ST}) between *Atlantic* and *Baltic ecotype*. *Atlantic ecotype* populations (Por-1SL, He-1SL, Ber-1SL) and *Baltic ecotype* populations (Ar-2AR, Seh-2AR, Ber-2AR) were grouped. The F_{ST} values are color-coded in grey ($F_{ST} < 0.5$), black ($0.5 \leq F_{ST} < 0.75$) and red ($F_{ST} \geq 0.75$).

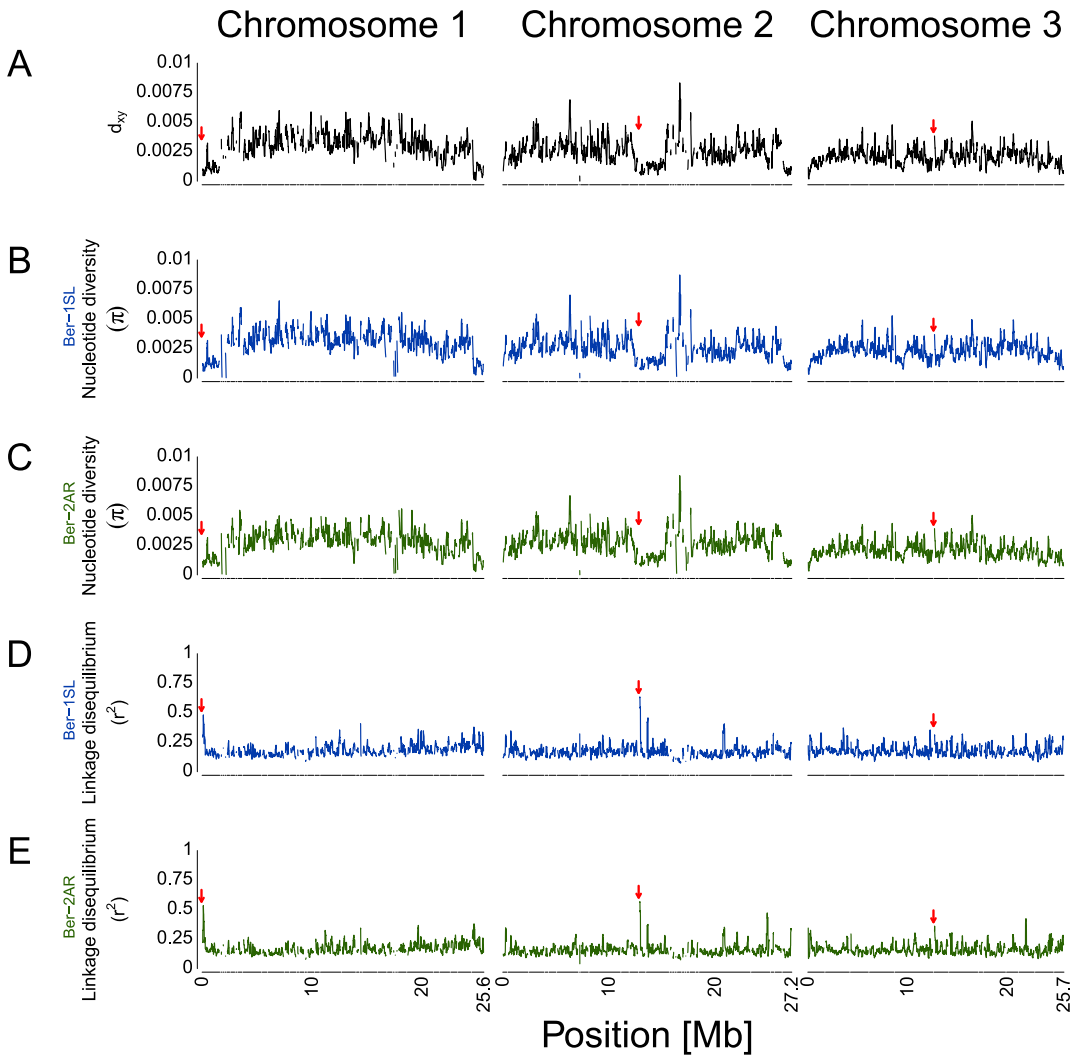

Figure S15: Genetic divergence (d_{xy}), nucleotide diversity (π) and short-range linkage disequilibrium (r^2) for the Ber-2AR vs. Ber-1SL comparison.
Summary statistics are based on 792,032 SNPs and were calculated in 100 kb sliding windows with 10 kb steps. (**A**) Genetic divergence (d_{xy}) based on allele frequencies. (**B**, **C**) Nucleotide diversity π. (**D**, **E**) Linkage disequilibrium measured as r^2. Ber-1SL and Ber-2AR are given in their respective population colors. The approximate position of the centromeres are marked on each chromosome with red arrows. The position is based on the data provided by Kaiser et al. (2016).

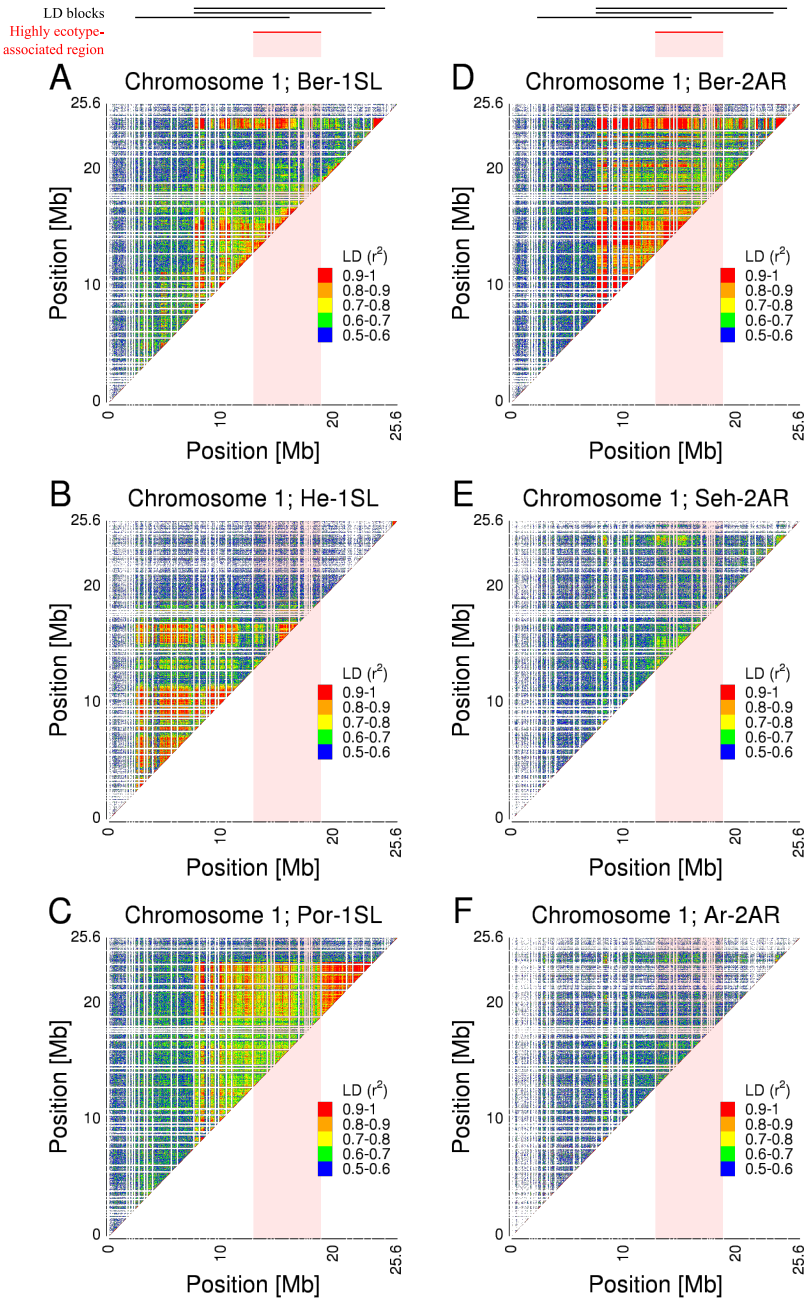

Figure S16: Pairwise linkage disequilibrium (LD) across chromosome 1 of all *Atlantic* and *Baltic ecotype* populations.

Pairwise LD was calculated as '- -geno-r2' in VCFtools with a '- -maf 0.2' filter and a LD threshold of 0.5 r^2. The r^2 values are color coded from 0.5 (blue) to 1 (red) in 0.1 steps. The panels show the chromosome 1 r^2 values for each population (A: Ber-1SL, B: He-1SL, C: Por-1SL, D: Ber-2AR, E: Seh-2AR, F: Ar-2AR).

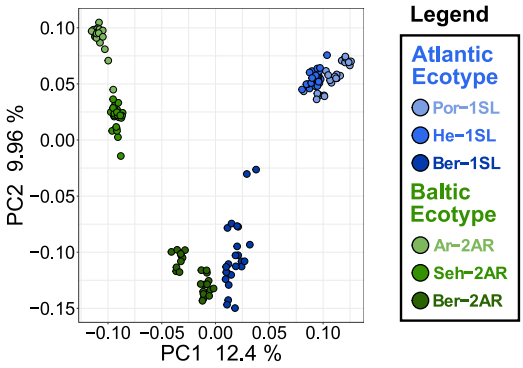

Figure S17: Principal component analysis (PCA) of the *Atlantic* and *Baltic ecotypes* for the highly ecotype-assocated region on chromosome 1

The highly ecotype-associated region on chromosome 1 ranges from superscaffolds 18 to 42. PCA for this region separates the ecotypes well, but does not show any patterns characteristic for a polymorphic structural variant (SV) separating the ecotypes. Only within the Por-1SL strain individuals cluster into three groups, suggestive of an overlapping polymorphic SV within that strain.

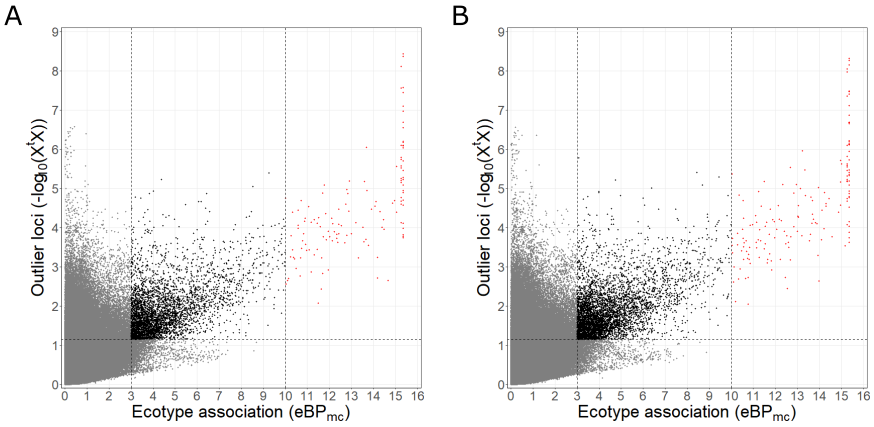

Figure S18: Selection of ecotype associated genetic variants based on X^tX and eBP_{mc} values.

(**A**) Ecotype-associated SNPs were selected based on an X^tX threshold of 1.152094 (as obtained from subsampling) and eBP_{mc} thresholds of 3 or 10 respectively (corresponding to p-values for association of 10^{-3} or 10^{-10}). SNPs above the X^tX threshold and with $eBP_{mc} \geq 3$ are colored black, those with $eBP_{mc} \geq 10$ are colored in red. (**B**) Ecotype-associated variants (SNPs and indels) were selected based on an X^tX threshold of 1.148764 and the same eBP_{mc} thresholds.

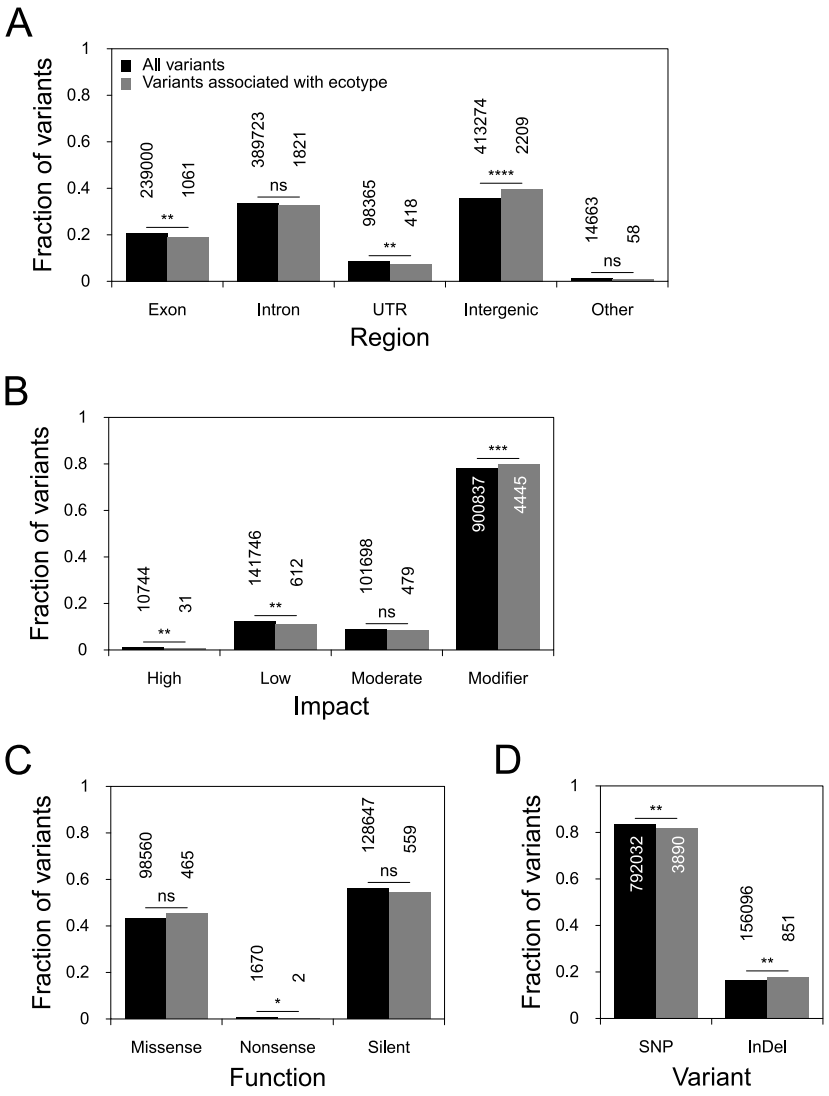

Figure S19: The effects of ecotype-associated genetic variants on genes, as compared to the genome-wide set of genetic variants.

SnpEff assesses genetic variants for their position relative to annotated gene models and infers the effect that these variants have on the genes. We compared the locations and effects of all variants in the dataset (n=948,128; black bars) to those of ecotype-associated variants (n=4,741; grey bars) and tested for significant differences with Fisher's exact test. The absolute numbers for each class of variants are given above or inside the bar. (**A**) The location of genetic variants relative to gene models. (**B**) The estimated impact of these variants on the gene models. (**C**) The effect of variants that are found in coding regions. (**D**) The proportions of SNPs vs. indels. p-value symbols: **** - 0-0.0001; *** - 0.0001-0.001; ** - 0.001-0.01; * - 0.01-0.05; ns - 0.05-1.

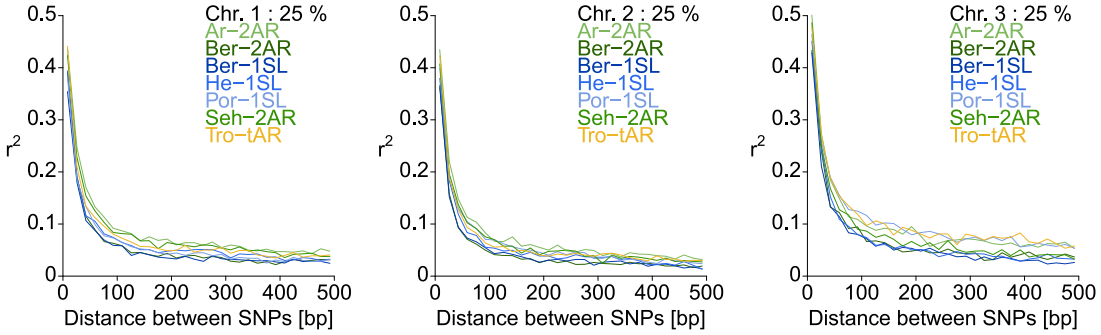

Figure S20: Linkage disequilibrium (LD) decay in the studied *C. marinus* populations.

Populations were analyzed separately and the corresponding lines in each plot are color-coded. To make a pairwise LD calculation measured as r^2 feasible, the SNP set was randomly subsampled to 25%. The distance between the SNPs was limited to 500 bp for LD calculation. The resulting r^2 values were then averaged within 30 average-distance-classes.

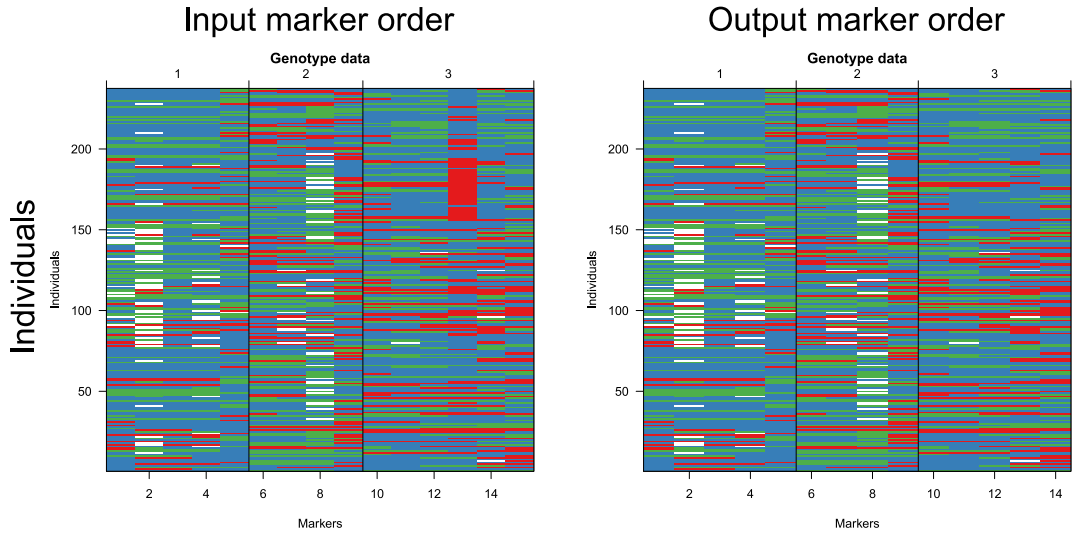

Figure S21: Order of the genetic markers and the present genotypes.

The input marker order corresponds to (Tab. S10) and (Fig. 14), but one excluded genotype on chromosome 1 and two on chromosome 2 as mentioned in section 4.3. Genotype color codes: AA - red; AB - blue; BB - green.

References

Kaiser, T. S., Poehn, B., Szkiba, D., Preussner, M., Sedlazeck, F. J., Zrim, A., Neumann, T., Nguyen, L.-T., Betancourt, A. J., Hummel, T., Vogel, H., Dorner, S., Heyd, F., von Haeseler, A., and Tessmar-Raible, K. (2016). The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature*, 540(7631):69–73.

Neumann, D. (1966). Die lunare und tägliche Schlüpfperiodik der Mücke *Clunio* - Steuerung und Abstimmung auf die Gezeitenperiodik. *Zeitschrift für Vergleichende Physiologie*, 53(1):1–61.