An oligogenic architecture underlying ecological and reproductive divergence in sympatric populations

Dušica Briševac[1], Carolina M. Peralta[1], Tobias S. Kaiser[1]*

[1] Max Planck Research Group *Biological Clocks*, Max Planck Institute for Evolutionary Biology, August-Thienemann-Strasse 2, 24306 Ploen, Germany

* kaiser@evolbio.mpg.de

Abstract

The evolutionary trajectories and genetic architectures underlying ecological divergence with gene flow are poorly understood. Sympatric timing types of the intertidal insect *Clunio marinus* (Diptera) from Roscoff (France) differ in lunar reproductive timing. One type reproduces at full moon, the other at new moon, controlled by a circalunar clock of yet unknown molecular nature. Lunar reproductive timing is a magic trait for a sympatric speciation process, as it is both ecologically relevant and entails assortative mating. Here we show that the difference in reproductive timing is controlled by at least four quantitative trait loci (QTL) on three different chromosomes. They are partly associated with complex inversions, but differentiation of the inversion haplotypes cannot explain the different phenotypes. The most differentiated locus in the entire genome, with QTL support, is the *period* locus, implying that this gene could not only be involved in circadian timing but also in lunar timing. Our data indicate that magic traits can be based on an oligogenic architecture and can be maintained by selection on several unlinked loci.

Introduction

After a long debate on whether speciation can only happen with geographic isolation (*allopatry*) or also with full range overlap (*sympatry*)[1-4], there is growing evidence for population divergence with gene flow[4,5]. The process is generally assumed to start with divergent ecological selection, but for speciation to complete, additional mechanisms for reproductive isolation must come into play[6,7]. Traits underlying ecological divergence and reproductive isolation must be genetically coupled, as otherwise they cannot withstand the homogenizing effects of gene flow and recombination[6]. One possibility to achieve such coupling are *magic traits*, i.e. traits that entail ecological divergence and reproductive isolation at the same time[8]. For example, divergent ecological adaptations can lead to differences in reproductive timing, i.e. *allochrony*[9] or *isolation by time*[10]. One example is found in the apple maggot *Rhagoletis*[11,12]. Magic traits imply that ecological divergence and reproduction are pleiotropically controlled by the same set of genes[13], usually involving only one or very few loci. As another possibility, genomic regions of low or suppressed recombination, such as chromosomal inversions, can entail a coupling of ecological and reproductive traits even if they are controlled by different sets of genes. Examples of inversions associated with ecologically relevant traits are found in *Littorina* snails[14,15] and *Heliconius* butterflies[16]. In this study, we assessed if these mechanisms play a role in sympatric population divergence in the marine midge *Clunio marinus* (Diptera: Chironomidae).

C. marinus is found in the intertidal zone of the European Atlantic coast. Larvae live at the lower levels of the intertidal, which are almost permanently submerged. For successful reproduction, the adults need these regions to be exposed by the low tide. The lowest tides predictably recur just after new moon and full moon. Therefore, *C. marinus* adults emerge

52 only during full moon or new moon low tides, reproduce immediately and die a few hours
53 later in the rising tide. This life cycle adaptation is based on a circalunar clock, i.e. an
54 endogenous time-keeping mechanism which synchronizes development and maturation with
55 lunar phase. Circalunar clocks are common in marine organisms, but their molecular basis is
56 still unknown[17]. Additionally, a circadian clock ensures that *C. marinus* adults only emerge
57 during the low tide[18]. As the pattern and amplitude of the tides differ dramatically along the
58 coastline, *C. marinus* populations from different geographic sites show various genetic
59 adaptations in circadian and circalunar timing[18-20].

60 Notably, full moon and new moon low tides are ecologically equally suitable for *C. marinus'*
61 reproduction, representing different *timing niches* that are occupied by different *timing types*
62 (for details and definitions see[21]). Some timing types of *C. marinus* use both niches (*"semi-*
63 *lunar rhythm"*; SL type), but there are also dedicated full moon (FM) or new moon (NM) timing
64 types, which occupy only one timing niche[21]. Divergence into FM and NM timing types
65 represents a magic trait, as it automatically entails assortative mating. We recently discovered
66 that in Roscoff (France) FM and NM timing types occur in sympatry, largely separated by
67 reproductive timing, but still connected by gene flow[21]. Here we present evidence for
68 polymorphic chromosomal inversions in all chromosome arms of the sympatric FM and NM
69 types. Quantitative trait loci (QTL) mapping for divergence in lunar reproductive timing
70 identifies four unlinked and additive QTL in different chromosomes and inversions. Individual
71 loci inside the inversions are more differentiated than the inversions, suggesting that the
72 inversions themselves are not directly required to maintain genetic linkage of adaptive loci
73 and that ecological divergence is kept up by permanent selection on unlinked loci on different
74 chromosomes.

Results

Genome sequencing confirms limited genetic differentiation between the sympatric FM and NM types. In order to gain insights into the loci and processes underlying sympatric divergence in reproductive timing, we sequenced 48 individual genomes for the FM and NM timing types (24 individuals each, defined by their lunar timing phenotype). The resulting set of 721,000 genetic variants (608,599 SNPs; 112,401 small indels) indicated limited genetic differentiation between the FM and NM types (Supplementary Fig. 1; weighted genome-wide F_{ST} = 0.028). Principal component analysis (PCA) and ADMIXTURE identify one individual as a *migrant in time*, which was caught at full moon among FM type individuals, but genetically clearly is of NM type (Supplementary Fig. 1A and B; blue circle and arrow). Many other individuals, particularly in the NM type, show ADMIXTURE fractions close to 0.5 or 0.25, suggesting they are F1 hybrids or backcrosses (Supplementary Fig. 1A and B; yellow circle and arrows). Finally, there are four FM individuals, which are genetically distinct along principal component 2 and in ADMIXTURE at K=4 (Supplementary Fig. 1A and B; red circle and arrows). These might constitute either a sub-lineage of the FM type or migrants from a different geographic location. Taken together, full genome resequencing confirms that there is strong gene flow between sympatric timing types.

A differentiated chromosomal inversion system on chromosome 1. Plotting genetic differentiation of the phenotypically defined FM and NM types along the genome (Supplementary Fig. 1C) revealed a region of elevated F_{ST} on the telocentric chromosome 1 (Fig. 1A). This region coincides with a block of long-range linkage-disequilibrium (LD) in the FM type (Fig. 1B; Supplementary Fig. 2). Structural variant (SV) calling based on additional

99 long-read sequencing data supports that this block represents a chromosomal inversion in
100 both timing types (Fig. 1C). Notably, the NM type shows a smaller LD block (Fig. 1B;
101 Supplementary Fig. 2), suggesting that a second structural variant occurs within the limits of
102 the detected large chromosomal inversion. While this second variant was not picked up in SV
103 calling, genetic linkage information obtained from crosses (described in detail below)
104 indicates that this is a second inversion (see double inverted marker order in Fig. 4A,B).

105 For a more detailed analysis of this inversion system, we subdivided it into three windows,
106 based on the inversion coordinates suggested by long-range LD. Two windows correspond to
107 those parts of the large inversion that do not overlap with the small inversion (roughly 2.7-
108 4.4 Mbp and 11.3-18.7 Mbp; Supplementary Table 1; Fig. 1A-C). The central window
109 corresponds to the overlap of both inversions (roughly 4.4-11.3 Mbp; Supplementary Table
110 4; Fig. 1A-C). In a Principal component analysis (PCA) on the genetic variants in these windows
111 (Supplementary Table 2), the four FM individuals that were already identified as genetically
112 distinct (Supplementary Fig. 1) always cluster separately and cannot be assessed (Fig. 1 D to
113 G, red question mark). The remaining 44 individuals are split into three genotype classes when
114 PCA is performed on the non-overlapping regions of the large inversion (Fig. 1D and F). These
115 classes correspond to individuals homozygous for the standard haplotype (SS), individuals
116 homozygous for an inverted haplotype (I_1I_1), and heterozygotes for the inversion (SI_1). The
117 haplotype with the higher observed heterozygosity in homozygotes is considered the
118 ancestral standard haplotype S (see Fig. 1H-J). Observed heterozygosity is clearly elevated in
119 individuals heterozygous for the inversion (Fig. 1H and J), suggesting there are polymorphisms
120 specific to the S or I_1 haplotypes. This is corroborated by the patterns of genetic
121 differentiation between SS and I_1I_1 homozygotes, which show F_{ST} values of up to 1
122 (Supplementary Figure 3).

123 In the region where both inversions overlap (Fig. 1E), the 44 individuals separate into six
124 genotype clusters, corresponding to three clusters of inversion haplotype homozygotes (SS,
125 I_1I_1, I_2I_2) and three clusters of heterozygotes (SI_1, SI_2, I_1I_2). Individuals that are I_1I_1 genotypes
126 in the 2.7-4.4 Mbp and 11.3-18.7 Mbp windows (Fig. 1D and F) now separate into the I_1I_1, I_1I_2
127 and I_2I_2 clusters (Fig. 1E). This indicates that the small inversion leading to haplotype I_2
128 happened in the already inverted haplotype I_1 of the large inversion (compare Fig. 1L). This is
129 backed up by genetic linkage data (Fig. 4A,B) and consistent with I_2I_2 homozygotes showing
130 even lower heterozygosity than I_1I_1 homozygotes (Fig. 1I). Again, heterozygosity is clearly
131 elevated in the inversion heterozygotes (Fig. 1I). A sliding-window PCA analysis along the
132 chromosomes (500kb windows, 100kb steps; Supplementary Figure 4) confirms these
133 patterns of segregation and the approximate genomic positions of inversion breakpoints.
134 Sliding-window PCA also indicated that there might be a third inversion, which ends at 24.8
135 Mbp (Fig. 1A and G; Supplementary Figure 4). Its start overlaps with the differentiated
136 inversion system and is therefore hard to identify, but according to observed heterozygosity
137 along the chromosome, it might lie at around 9 Mbp (Supplementary Fig. 5A). This inversion
138 is only clearly detected in three SI_3 heterozygotes (Fig. 1G and K). These three individuals are
139 found in the SS cluster in the other genomic windows, implying that the inversion leading to
140 haplotype I_3 happened in the standard haplotype S (Fig. 1L).
141 One peculiarity deserves further investigation: In the window from 2.7-4.4 Mbp some
142 individuals of the I_1I_1 genotype cluster close to the SI_1 heterozygotes. However, the pattern in
143 which these individuals segregate in the 4.4-11.3 Mbp window unequivocally indicates that
144 these must be I_1I_1 homozygotes. They also show much lower heterozygosity than the SI_1
145 heterozygotes. Their peculiar clustering may indicate a certain degree of gene conversion
146 between inversion haplotypes or a complex demographic history.

147 Taken together, our data support three inversion events, which we name In(1a), In(1b) and
148 In(1c), leading to four distinct haplotypes of chromosome 1 (S, I_1, I_2 and I_3; Fig. 1L). From the
149 PCA-derived genotypes (Fig. 1D-G), we can infer the frequency of each haplotype (Fig. 1L).
150 There is marked genetic differentiation between the sympatric timing types. The FM type
151 carries all haplotypes but is dominated by the S haplotype, while the NM type almost
152 exclusively carries the inverted I_1 and I_2 haplotypes.
153
154 **Non-differentiated inversions on chromosomes 2 and 3.** The analysis of long-range LD
155 also identifies one putative chromosomal inversion on each arm of the metacentric
156 chromosomes 2 and 3 (Fig. 2B,C; Supplementary Fig. 6). We call them In(2L), In(2R), In(3L) and
157 In(3R). In(2L) is confirmed by inverted marker order in linkage mapping (Fig. 4A,B) and In(2R)
158 is confirmed by SV calling. As above, windowed PCA along the chromosomes confirms the
159 approximate inversion breakpoints and allows to genotype the individuals for the inversion
160 haplotypes (Fig. 2D-G). Heterozygosity is clearly elevated in inversion heterozygotes (SI; Fig.
161 2H-K) and lowered in the inversion homozygotes (II; Fig. 2H-K). Notably, inversion
162 homozygotes are absent for two of these inversions (Fig. 2D and F) and very rare for the other
163 two (Fig. 2E and G), suggesting these inversions might impose some fitness constraints. These
164 four inversions are only weakly differentiated between populations (Fig. 2 L-O). Only In(2L)
165 shows up as a block of mildly elevated genetic differentiation between the strains (Fig. 2A).
166
167 **Crosses between the FM and NM types indicate that lunar reproductive timing is**
168 **heritable and controlled by four QTL.** Next, we tested for a genetic basis of lunar reproductive
169 timing by performing crossing experiments between the FM and NM types (Fig. 3). F1 hybrids
170 emerge intermediate between the FM and NM type, with a slight shift towards FM emergence

171 (Fig. 3C). In the F2 generation, the phenotype distribution is spread out, but does not fully
172 segregate into parental and F1 phenotype classes (Fig. 3D). This indicates that the difference
173 in lunar reproductive timing is controlled by more than one genetic locus. From the crossing
174 experiment, we picked a set of several F2 families that together comprised 158 individuals,
175 which all go back to a single parental pair – a FM type mother and a NM type father. Because
176 of the limited genetic differentiation between the timing types (see Supplementary Fig. 1C
177 and [21]), we sequenced the full genomes of the two parents in order to identify informative
178 genetic markers for linkage mapping and QTL mapping. We picked a set of 32 microsatellite
179 markers, 23 of which turned out to be reliably amplified and informative in the F2 families, as
180 well as four insertion-deletion polymorphisms (Supplementary Table 3).

181 Genetic linkage mapping confirmed the existence of the inversions In(1a) and In(1b), both
182 of which are supported by inverted marker order on the genetic linkage map (Fig. 4A). Marker
183 order is also inverted for In(2L), but not for the other inversions (Fig. 4A). We then performed
184 stepwise QTL identification and Multiple QTL Mapping (MQM) as implemented in R/qtl and
185 found four largely additive QTL for the difference in lunar reproductive timing (Fig. 4A and
186 Supplementary Table 4). A specific scan for epistatic effects only shows a weak interaction of
187 chromosome 1 with the QTL on chromosome 2 (Supplementary Figure 7). The four QTL model
188 explains 54% of the total variance in lunar reproductive timing and the estimated additive
189 effects of the QTL account for 6 days of the timing difference between the FM and NM type
190 (Fig. 4A, see Supplementary Table 4 for individual QTL effects and variance explained).

191 In both MQM and additional Composite Interval Mapping (CIM; Supplementary Figure 8)
192 we estimated confidence intervals for the QTL (Fig. 4A). There are two QTL on chromosome
193 1, one of which overlaps with the inversion system. While the inversions reduce
194 recombination and thus map length for the respective genomic region (Fig. 4A), there was

195 sufficient recombination to assign the QTL interval to only one end of In(1a). Based on the
196 first genetic marker outside the QTL's confidence intervals, we matched the QTL region with
197 the corresponding genomic reference sequence (Fig. 4B). The second QTL on chromosome 1
198 maps outside the inversion system and is not linked with the other QTL (Fig. 4A,B). The other
199 two QTL are found on chromosomes 2 and 3 and overlap with In(2R) and In(3L) respectively.
200 The genomic regions underlying these QTL are large and show several peaks in genetic
201 differentiation (Fig. 4B). Possibly, the QTL on chromosome 2 and 3 harbor a set of linked loci
202 influencing the lunar emergence phenotype. Taken together, there are at least four unlinked
203 loci which influence this magic trait, giving it a complex genetic architecture.
204
205 **Divergent alleles are not associated with specific inversion haplotypes.** While some
206 chromosomal inversions show up as genetically differentiated blocks, in all chromosomal
207 inversions there are loci that are more differentiated than the inversion itself (Fig. 4B). This
208 implies that the inversions are not required for protecting genetic differentiation. Instead,
209 there must be sufficient levels of recombination and gene conversion to place the divergent
210 alleles in all inversion haplotypes, so that individual loci can diverge, while the surrounding
211 inversions remain less differentiated. We corroborated this scenario for the most
212 differentiated locus in the genome (Fig. 4B, red asterisk), which is the *period* locus (Fig. 4C).
213 We identified an insertion-deletion (indel) mutation in the highly differentiated first intron of
214 the *period* gene (Fig. 4C, green arrows) and re-genotyped the 48 individuals for this indel (Fig.
215 4D; Supplementary Figure 9). This confirmed that *period* alleles indeed strongly differentiated
216 between the FM and NM types. However, we could not find a close association between the
217 *period* alleles and the inversion haplotypes of chromosome 1 (Fig. 4D). This confirms that
218 differentiation of the *period* locus is not protected by the inversions.

219

220 **Candidate loci.** The existence of individual divergence peaks within the QTL, despite the
221 overlap with chromosomal inversions, allows us to pinpoint candidate genes for controlling
222 the magic trait, i.e. phenotypic divergence in lunar reproductive timing (see Supplementary
223 Table 5 for an overview). Some loci are particularly conspicuous in that they are by far the
224 most divergent within a specific QTL. The QTL in the inversion system on chromosome 1 is
225 narrow and harbors a single strongly divergent locus, which is the *period* locus (Fig. 4B, red
226 asterisk). The *period* locus is the most differentiated locus in the entire genome (maximum
227 F_{ST} = 0.86) and *period* is a core circadian clock gene. The QTL at the end of chromosome 1,
228 outside the inversion system, also shows a single strong divergence peak (Fig. 4B, red circle).
229 This is the first intron of the *plum* gene (Supplementary Figure 10). The large intron likely
230 contains regulatory regions for both *plum* and the directly adjacent *cask* gene. On
231 chromosome arm 2L the QTL has several divergence peaks, but the strongest peak hits the
232 *stat1* gene (Fig 4B, red cross; Supplementary Figure 11). *Plum*, *cask* and *stat1* are all involved
233 in nervous system development in *Drosophila melanogaster*[22-24], suggesting that divergence
234 in lunar reproductive timing involves nervous system remodeling as well as circadian
235 timekeeping.

236 **Discussion**

237

238 We found multiple large and polymorphic chromosomal inversions in the sympatric FM
239 and NM strains, covering all chromosome arms of the *C. marinus* genome. This is consistent
240 with large non-pairing and inverted regions observed in *C. marinus* polytene chromsomes[25].
241 These inversions reduce recombination, as is observed in large LD blocks (Supplementary
242 Figures 2 and 6), and possibly lock several loci involved in divergent reproductive timing into
243 supergenes (Figure 1A and 2A). However, most of these inversions are not differentiated
244 between timing types and in all inversions, there are loci more divergent than the inversion
245 itself. We show that divergent alleles at the *period* locus are not associated with specific
246 inversion haplotypes or the standard haplotype (Figure 4D). There must be considerable
247 recombination and/or gene conversion inside the inversion, placing the *period* alleles in
248 different inversion haplotypes. This is likely also true for other divergent loci and forms the
249 basis for genetic differentiation at individual loci to exceed the inversion background. We
250 propose that natural selection on the divergent loci is more important for maintaining the
251 magic trait than the lowering of recombination rate by the inversions. Ecological divergence
252 rather drives differentiation of the inversion system on chromosome 1 than being protected
253 by it. Nevertheless, differentiation of the inversion system creates a genetic substrate for
254 further ecological and reproductive divergence.

255 At the same time, QTL mapping revealed at least four independent loci underlying lunar
256 reproductive timing. Therefore, we are neither dealing with a single pleiotropic locus that
257 affects the magic trait, nor with a single supergene that would combine genes underlying both
258 ecological divergence and reproductive isolation. Our data show that an oligogenic
259 architecture can underlie ecological and reproductive divergence in sympatric populations. In

260 addition, the four QTL we identified here for the difference in lunar reproductive timing
261 between the FM and NM types in Roscoff do not overlap with the two QTL we identified
262 previously for difference in lunar reproductive timing between Normandy (Por-1SL) and
263 Basque Coast (Jean-2NM) strains[26,27]. This indicates that within the species *C. marinus*,
264 different populations have found different genetic solutions for adapting lunar reproductive
265 timing to the local tidal regime. Thus, the genetic architecture underlying the evolution of this
266 magic trait is considerably more complex than what is generally assumed in models of
267 sympatric population divergence. Notably, the first records of *C. marinus* for Roscoff date to
268 1978 and 1989 and only mention the FM type[28,29]. The authors sampled larvae, which would
269 most likely have revealed the NM type as well, had it been present at the time. This sets a
270 very recent time frame to the establishment of sympatric timing types and the evolution of
271 the magic trait. Given the two timing niches imposed by the tides, divergent selection seems
272 to lead to rapid ecological and reproductive divergence. This may eventually lead to
273 speciation through a process of evolutionary branching[30,31].

274 Genetic differentiation between the inversion haplotypes is much higher than what could
275 evolve in such a recent time-frame (Supplementary Figure 3). We thus assume that the
276 inversions were already segregating in the ancestral populations or may have recently
277 introgressed from a different geographic site.

278 The detection of *period* as the most differentiated locus between NM and FM types,
279 overlapping with a QTL for lunar reproductive timing, suggests that this circadian clock gene
280 might also be involved in lunar time-keeping. The other prominent candidates, *plum*, *cask* and
281 *stat1* are involved in neuronal or synaptic development and plasticity. Notably, *cask* is a direct
282 interaction partner of *CaMKII* [23], which we previously found to be involved in circadian timing
283 differences in *Clunio*[26]. Hence, both QTL on chromosome 1 suggest a close connection of

284 circadian and circalunar time-keeping. This is backed by recent findings in Baltic and Arctic
285 *Clunio* strains, for which genomic analysis of the loss of lunar rhythms also implied circadian
286 GO terms as the top candidates for affecting circalunar time-keeping, followed by GO terms
287 involved in nervous system development[32]. Thus, our data not only call for a new evaluation
288 of the genetic architectures underlying magic traits, but also give hints to the genetic basis of
289 the yet enigmatic circalunar clock.

290 **Methods**

291

292 **Sampling and laboratory strains.** Laboratory strains for crossing experiments and field
293 samples for whole genome sequencing of 48 individuals were available from a previous
294 study[21]. Laboratory strains were kept in standard culture conditions[33] at 20°C and under a
295 light-dark cycle of 16:8. An artificial moonlight cycle with four nights of dim light every 30 days
296 served to synchronize reproduction in the laboratory strains.

297 **Sequencing, read mapping and genotype calling.** DNA from the 48 field caught individuals
298 were subject to whole genome sequencing in the Max Planck Sequencing Centre (Cologne)
299 on an Illumina HiSeq3000 according to standard protocols. Independent sequencing runs
300 were merged with the *cat* function. Adapters were trimmed with Trimmomatic [34] using the
301 following parameters: ILLUMINACLIP <Adapter file> :2:30:10:8:true, LEADING:20,
302 TRAILING:20, MINLEN:75. Overlapping read pairs were merged with PEAR [35] using -n 75 -c 20
303 -k and mapped to the Cluma_2.0 reference genome (available in the Open Research Data
304 Repository of the Max Planck Society under DOI 10.17617/3.42NMN2; manuscript in
305 preparation) with bwa mem [36] version 0.7.15-r1140. Mapped reads were merged into a single
306 file, filtered for -q 20 and sorted with samtools v1.9 [37]. SNPs and small indels were called using
307 GATK v3.7-0-gcfedb67 [38]. All reads in the q20 sorted file were assigned to a single new read-
308 group with 'AddOrReplaceReadGroups' script with LB=whatever PL=illumina PU=whatever
309 parameters. Genotype calling was then performed with HaplotypeCaller and parameters --
310 emitRefConfidence GVCF -stand_call_conf 30, recalibration of base qualities using GATK
311 BaseRecalibrator with '-knownSites'. Preparing recalibrated BAM files with GATK PrintReads
312 using -BQSR. Recalling of genotypes using GATK HaplotypeCaller with previously mentioned
313 parameters. Individual VCF files were combined into a single file using GATK GenotypeGVCFs.

314 **Genetic differentiation (F_{ST}).** The vcf file containing GATK-called SNPs and small indels
315 from 48 Ros-2FM and Ros-2NM males was filtered with vcftools version 0.1.14 [39] for minor
316 allele frequency of 0.05, maximum of 2 alleles, minimum quality (minQ) of 20, maximum
317 missing genotypes of 20%, which finally left 721.000 variants. Genetic differentiation between
318 the two populations (fst) was estimated using vcftools parameters --weir-fst-pop --fst-
319 window-size 1 --fst-window-step 1.

320 **Long-range linkage disequilibrium (LD)**. Linkage disequilibrium was calculated between all
321 variants along the three chromosomes in each of the populations in search of the signatures
322 of large genomic inversions. Vcf file containing GATK-called SNPs and small indels from 24
323 Ros-2FM males was filtered for minor allele frequency of 0.20, maximum of 2 alleles,
324 minimum quality (minQ) of 20, maximum missing genotypes per site of 20% leaving 344.331
325 variants. The same was done for the 24 Ros-2NM individuals resulting in 352.915 variants.
326 They were converted to plink input files and plink version 1.90 beta [40] was used to calculate
327 linkage disequilibrium between all variants with parameters --r2 --inter-chr. Large LD file was
328 then sorted into 3 files, one for each chromosome. Breakpoints of large chromosomal
329 inversions were identified by plotting r^2 and manually looking for obvious breaks in the r^2
330 scores (Supplementary Figure 2 and 6; Supplementary Table 1).

331 **Principal component analysis (PCA).** PCA was run for the entire genome, for sliding
332 windows along the chromosome ("windowed PCA"), and for windows corresponding to the
333 inversions ("local PCA"). The vcf file containing GATK-called SNPs and small indels from 48
334 Ros-2FM and Ros-2NM males was filtered with vcftools version 0.1.14 [39] for minor allele
335 frequency of 0.05, minimum quality (minQ) of 20, maximum missing genotypes per site of
336 20%, leaving 703.579 variants. Principle component analysis was calculated with plink version
337 1.90 beta [40] using flags --nonfounders --pca var-wts --chr-set 3 no-xy no-mt. Windowed PCA:

338 In order to investigate regions of the genome with unusual degree of variance, PCA was run
339 in windows along the three chromosomes. Vcf files were subdivided into vcf files containing
340 variants that belong to 500kb sliding windows with 100kb steps. They were further used to
341 calculate PCA values for each window as noted above. Local PCA: To get the genotype-
342 estimate of the identified inversions (see long-range disequilibrium section), we subdivided
343 vcf files according to the estimated inversion breakpoints (Supplementary Table 4), and
344 calculated principal components 1 and 2 as described above. Number of variants that belong
345 to each window are listed in Supplementary Table 2.

346 **Admixture.** Ancestry and relatedness of the 48 individuals of Ros-2FM and Ros-2NM
347 population were investigated with admixture version 1.3.0 [41]. Admixture was run for the
348 entire genome, for sliding windows along the chromosome ("windowed admixture"), and for
349 windows corresponding to the inversions ("local admixture"). We used vcf files previously
350 described in the PCA section as input. Admixture was calculated using k of 2 to 4. Windowed
351 admixture: In order to identify regions of the genome with ancestry different from the general
352 population, we ran admixture on vcf files containing variants that belong to 500kb sliding
353 windows with 100kb steps. Local admixture: In order to genotype the inversions (see long-
354 range disequilibrium section) we calculated admixture using vcf files containing only variants
355 from those regions of the genome (Supplementary Tables 1 and 2).

356 **Observed heterozygosity.** In order to complement the windowed and local PCA and
357 admixture analyses and further corroborate the indirect genotyping of the chromosomal
358 inversions, we calculated observed heterozygosity using vcftools version 0.1.14 [39] with --het
359 flag. Proportion of observed heterozygosity was calculated for each file by dividing the
360 number of observed heterozygotes with the number of variants. Patterns of heterozygosity
361 along the chromosomes were assessed based on the vcf files containing variants that belong

362 to 500kb sliding windows with 100kb steps (see PCA and admixture section). Finally, in order
363 to calculate observed heterozygosity of the chromosomal inversions, observed
364 heterozygosity was calculated per inversion window for each individual (Supplementary
365 Tables 1 and 2). Then observed heterozygosity was calculated and plotted for each inversion
366 genotype according to local PCA and admixture.

367 **Crosses, phenotyping and genotyping.** After synchronizing the NM and FM strains by
368 applying different moonlight regimes, single pair crosses were performed. F1 egg clutches
369 were reared individually and the emerging adults were allowed to mate freely within each F1
370 family, leading to several sets of F2 families that go back to a single pair of parents. One of
371 these sets was picked for QTL mapping. As the second peak is not under clock control, but a
372 direct response to moonlight[21], analysis was restricted to individuals emerging in the first
373 peak (n=158). DNA was extracted with a salting out method[42] and amplified with the *REPLI-g*
374 *Mini Kit* (Qiagen) according to the manufacturer's instructions. The two parents were subject
375 to whole genome sequencing with 2x150bp reads on an Illumina HiSeq2000 according to
376 standard protocols. Read mapping and genotype calling were performed as described above.
377 Microsatellite and indel genotypes were obtained by custom parsing of the vcf files. PCR
378 primers (Supplementary Table 3) for amplifying the microsatellite and indel regions were
379 designed with Primer3. Indels were PCR amplified and then run and scored on 1.5% agarose
380 gels. Microsatellites were amplified with HEX- and FAM-labeled primers and run on an ABI
381 PRISM 3100 Genetic Analyzer. Chromatograms were analyzed and scored with in R with the
382 *Fragman* package[43]. The resulting genotype matrix can be found in the R/qtl input file
383 (Supplementary File 1).

384 **Linkage & QTL mapping**. Linkage analysis and QTL mapping were performed in R/qtl
385 according to the script in Supplementary File 2. Briefly, the recombination fraction and

386 distribution of alleles were checked and the correct marker order was inferred by rippling.
387 Missing data and error load were assessed. Then the best model was obtained in a stepwise
388 selection procedure (*stepwiseqtl*). Additional interactions were checked for in a two QTL scan
389 (*scantwo*), but were negligible and not considered in the final model. Finally, a model with 4
390 additive QTLs was subject to model fitting for Multiple QTL Mapping (*fitqtl*) and QTL
391 confidence intervals were estimated (*bayesint*). Composite Interval mapping was performed
392 in QTLcartographer[44] with various selection procedures (forward, backward), exclusion
393 windows (10 cM and 20 cM) and covariates (3, 5, 10; Supplementary Figure 8). The LOD
394 significance threshold was estimated by running 1000 permutations and a p value of 0.05.
395 **PacBio Data and structural variant calling.** PacBio long read data was obtained for pools
396 of 300 to 500 individuals from laboratory strains of Ros-2NM and Ros-2FM (Roscoff, France).
397 DNA was extracted as above and sequenced with standard protocols on a PacBio Sequel II at
398 the Max Planck Sequencing Facility in Cologne, Germany. Raw long-reads of Ros-2FM and Ros-
399 2NM were mapped against the reference CLUMA 2.0 (publication in preparation) using NGM-
400 LR v.0.2.7 [45] with default settings. Alignments were sorted, filtered (q20) and indexed with
401 samtools v.1.9 [37]. Three different SV-calling tools were used per population to discover SVs.
402 Sniffles v.1.0.11 [45] was run with parameters -- min_het_af 0.1 and –genotype. SVIM v1.2.0 [46]
403 was run with default parameters (svim alignment). Finally, Delly v0.8.6 [47] was run with
404 parameters lr -y pb -q 20 --svtype (INS, insertions; INV, inversions; DUP, duplications; DEL,
405 deletions) and the output VCF files per SV type were merged with a custom script. The
406 resulting VCF files were then sorted using vcf-sort and filtered for quality "PASS" using a
407 custom bash script. SURVIVOR v1.0.6 [48] was used to filter variants for a minimum size of 300
408 bp and at least 5 reads supporting each variant (SURVIVOR filter NA 300 -1 0 5). BND variants
409 detected with Sniffles and SVIM calls were excluded with a custom bash script before using

410 SURVIVOR merge (options set to 50 1 1 0 0 300) on all VCFs produced per population. The
411 merged VCFs were then used as an input to reiterate SV calling with Sniffles using the same
412 parameters as above plus --Ivcf option and --min_support 5 (minimum number of reads
413 supporting a SV).

414 ***Period* genotyping.** An insertion-deletion (indel) mutation in the *period* locus was
415 genotyped. The fragment was PCR amplified (primers: 5'-GAATACTGAGTGTAAGACTTGGC and
416 5'-ACAACGTGACCTGTGACAAT) and run and scored on 1.5% agarose gels.

417

418

419 **Data availability**

420 Sequencing data was submitted to ENA under project number PRJEB54033. The CLUMA2.0
421 reference genome is available on the Open Research Data Repository of the Max Planck
422 Society (EDMOND) under DOI 10.17617/3.42NMN2.

423

424

425 **Materials & Correspondence**

426

427 Requests should be addressed to Tobias S. Kaiser (kaiser@evolbio.mpg.de).

428 **Author contributions**

429

430 TSK conceived, designed and supervised the study, performed crosses and QTL mapping,
431 interpreted data and wrote the manuscript. DB analyzed short-read sequencing data and
432 participated in QTL mapping. CMP analyzed long-read sequencing data. All authors read and
433 approved of the manuscript.

434

435

436 **Acknowledgements**

437

438 Tjorben Nawroth, Kerstin Schäfer, Susanne Mentz, Elke Bustorf and Sina Schirmer provided
439 laboratory assistance. Jürgen Reunert provided animal care. We thank all members of the
440 research group *Biological Clocks,* as well as Diethard Tautz for feedback and discussion. This
441 work was supported by the Max Planck Society via an independent Max Planck Research
442 Group and by an ERC Starting Grant (No 802923) awarded to TSK.

443

444

445 **Competing interests**

446

447 There are no competing interests.

References

1 Mayr, E. Ecological factors in speciation. *Evolution*, 263-288 (1947).

2 Maynard Smith, J. Sympatric speciation. *Amer Natur* **100**, 637-650 (1966).

3 Via, S. Sympatric speciation in animals: the ugly duckling grows up. *Trends in ecology & evolution* **16**, 381-390 (2001).

4 Foote, A. D. Sympatric speciation in the genomic era. *Trends in Ecology & Evolution* **33**, 85-95 (2018).

5 Richards, E. J., Servedio, M. R. & Martin, C. H. Searching for sympatric speciation in the genomic era. *BioEssays* **41**, 1900047 (2019).

6 Smadja, C. M. & Butlin, R. K. A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology* **20**, 5123-5140, doi:Doi 10.1111/J.1365-294x.2011.05350.X (2011).

7 Seehausen, O. *et al.* Genomics and the origin of species. *Nature Reviews Genetics* **15**, 176-192, doi:Doi 10.1038/Nrg3644 (2014).

8 Gavrilets, S. *Fitness Landscapes and the Origin of Species*. Vol. 41 476 (Princeton University Press, 2004).

9 Taylor, R. S. & Friesen, V. L. The role of allochrony in speciation. *Mol Ecol* **26**, 3330-3342, doi:10.1111/mec.14126 (2017).

10 Hendry, A. P. & Day, T. Population structure attributable to reproductive time: isolation by time and adaptation by time. *Molecular Ecology* **14**, 901-916 (2005).

11 Filchak, K. E., Roethele, J. B. & Feder, J. L. Natural selection and sympatric divergence in the apple maggot Rhagoletis pomonella. *Nature* **407**, 739-742, doi:10.1038/35037578 (2000).

12 Doellman, M. *et al.* Genomic differentiation during speciation-with-gene-flow: comparing geographic and host-related variation in divergent life history adaptation in Rhagoletis pomonella. *Genes* **9**, 262 (2018).

13 Servedio, M. R., Van Doorn, G. S., Kopp, M., Frame, A. M. & Nosil, P. Magic traits in speciation:'magic'but not rare? *Trends in ecology & evolution* **26**, 389-397 (2011).

14 Koch, E. L. *et al.* Genetic variation for adaptive traits is associated with polymorphic inversions in Littorina saxatilis. *Evol Lett* **5**, 196-213 (2021).

15 Faria, R. *et al.* Multiple chromosomal rearrangements in a hybrid zone between Littorina saxatilis ecotypes. *Molecular ecology* **28**, 1375-1393 (2019).

16 Joron, M. *et al.* Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**, 203-U102, doi:10.1038/nature10341 (2011).

17 Kaiser, T. S. & Neumann, J. Circalunar clocks—Old experiments for a new era. *BioEssays* **43**, 2100074 (2021).

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

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

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

21 Kaiser, T. S., von Haeseler, A., Tessmar-Raible, K. & Heckel, D. G. Timing strains of the marine insect Clunio marinus diverged and persist with gene flow. *Molecular Ecology* **30**, 1264-1280, doi:https://doi.org/10.1111/mec.15791 (2021).

22 Yu, X. M. *et al.* Plum, an immunoglobulin superfamily protein, regulates axon pruning by facilitating TGF-β signaling. *Neuron* **78**, 456-468 (2013).

23 Gillespie, J. M. & Hodge, J. J. CASK regulates CaMKII autophosphorylation in neuronal growth, calcium signaling, and learning. *Frontiers in molecular neuroscience* **6**, 27 (2013).

24 Ngo, K. T. *et al.* Concomitant requirement for Notch and Jak/Stat signaling during neuro-epithelial differentiation in the Drosophila optic lobe. *Developmental biology* **346**, 284-295 (2010).

25 Michailova, P. Comparative External Morphological and Karyological Characteristics of European Species of Genus Clunio HALIDAY 1855 *VII. International Symposium of Chironomidae*, 9-15 (1980).

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

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

28 Neumann, D. & Heimbach, F. in *Cyclic Phenomena in Marine Plants and Animals* (eds E. Naylor & R.G. Hartnoll) 423-433 (Pergamon Press, 1979).

29 Neumann, D. Circadian Components of Semilunar and Lunar Timing Mechanisms. *Journal of Biological Rhythms* **4**, 285-294 (1989).

30 Doebeli, M. & Dieckmann, U. Speciation along environmental gradients. *Nature* **421**, 259-264 (2003).

31 Doebeli, M. & Dieckmann, U. Evolutionary branching and sympatric speciation caused by different types of ecological interactions. *The american naturalist* **156**, S77-S101 (2000).

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

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

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

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

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

37 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

38 McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

538 39 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158
539 (2011).
540 40 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
541 datasets. *Gigascience* **4**, s13742-13015-10047-13748 (2015).
542 41 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in
543 unrelated individuals. *Genome research* **19**, 1655-1664 (2009).
544 42 Reineke, A., Karlovsky, P. & Zebitz, C. P. W. Preparation and purification of DNA from
545 insects for AFLP analysis. *Insect Molecular Biology* **7**, 95-99 (1998).
546 43 Covarrubias-Pazaran, G., Diaz-Garcia, L., Schlautman, B., Salazar, W. & Zalapa, J.
547 Fragman: an R package for fragment analysis. *BMC genetics* **17**, 1-8 (2016).
548 44 Windows QTL Cartographer (Department of Statistics, North Carolina State University,
549 Raleigh, NC, 2006).
550 45 Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-
551 molecule sequencing. *Nat Methods* **15**, 461-468, doi:10.1038/s41592-018-0001-7
552 (2018).
553 46 Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long
554 reads. *Bioinformatics* **35**, 2907-2915 (2019).
555 47 Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-
556 read analysis. *Bioinformatics* **28**, I333-I339, doi:10.1093/bioinformatics/bts378
557 (2012).
558 48 Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative
559 traits and reproductive isolation in fission yeast. *Nature communications* **8**, 1-11
560 (2017).
561

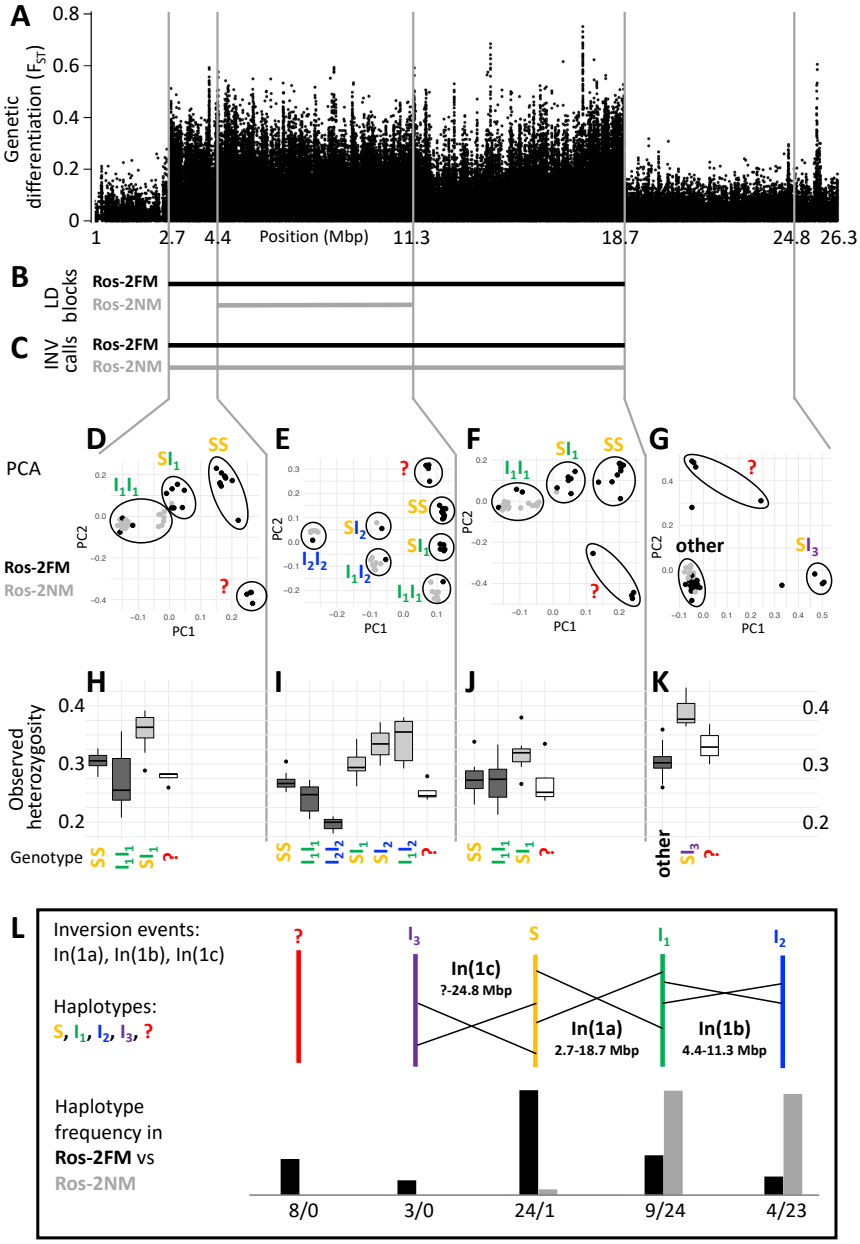


Figure 1 A complex inversion system on chromosome 1. (**A**) Chromosome 1 harbors a block of markedly elevated genetic differentiation. (**B**) This genomic block coincides with two windows of elevated long-range LD in the FM and NM strains. (**C**) Structural variant (SV) calling from long read sequencing data supports that the larger LD block is due to an inversion. (**D-G**) Principal component analysis (PCA) for chromosomal sub-windows corresponding to suggested inversions separates the individuals into clusters corresponding to standard haplotype homozygotes (SS), inversion homozygotes (I_1I_1, I_2I_2) and inversion heterozygotes (SI_1, SI_2, SI_3, I_1I_2). The four individuals which were already found distinct in whole genome analysis (red question mark) cannot be assessed. (**H-K**) Observed heterozygosity is clearly elevated in inversion heterozygotes, underpinning substantial genetic differentiation between the inversions. (**L**) A schematic overview of the sequence of inversion events and the resulting haplotypes. The frequencies of the haplotypes differ markedly between the FM and NM types.

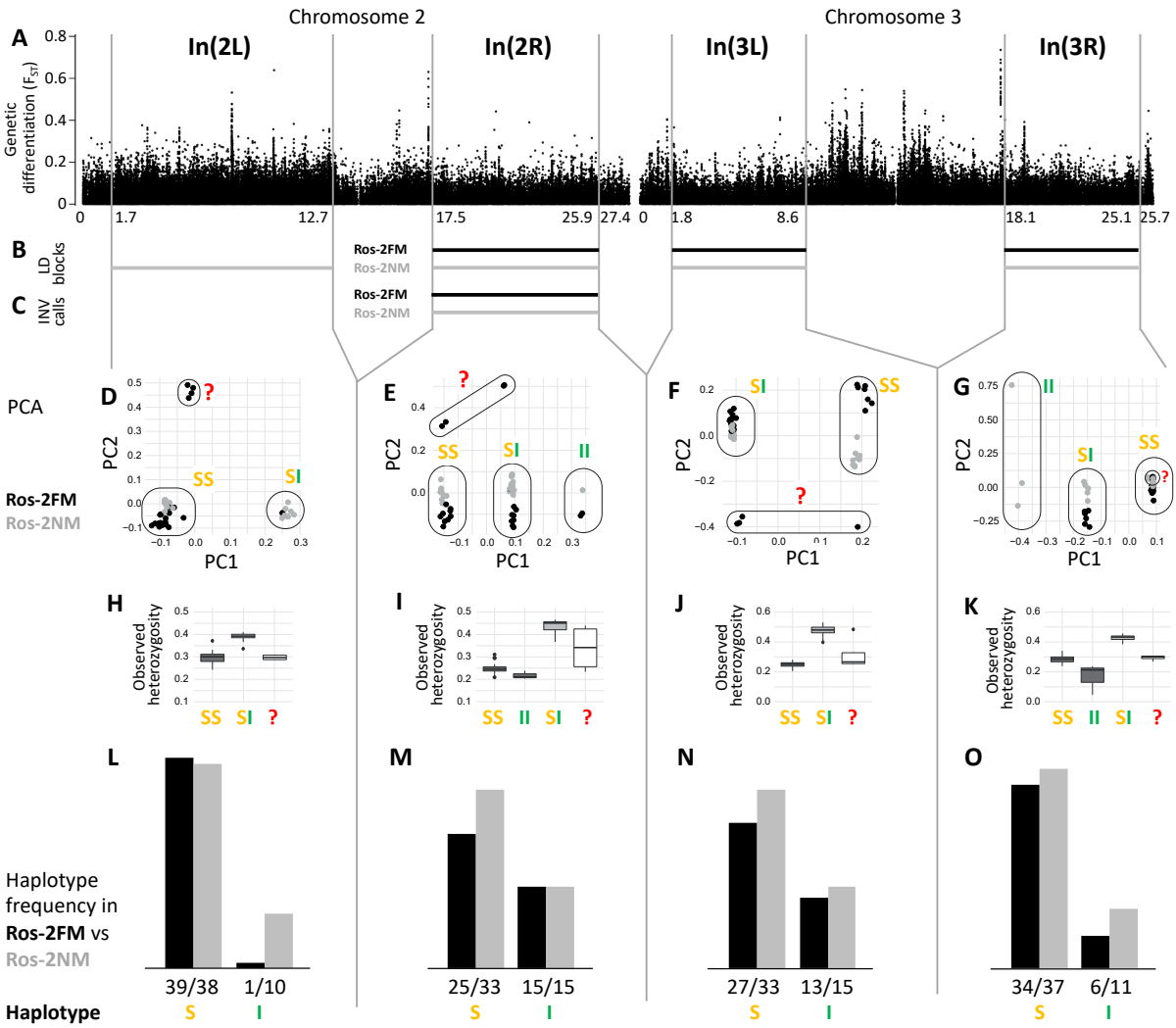
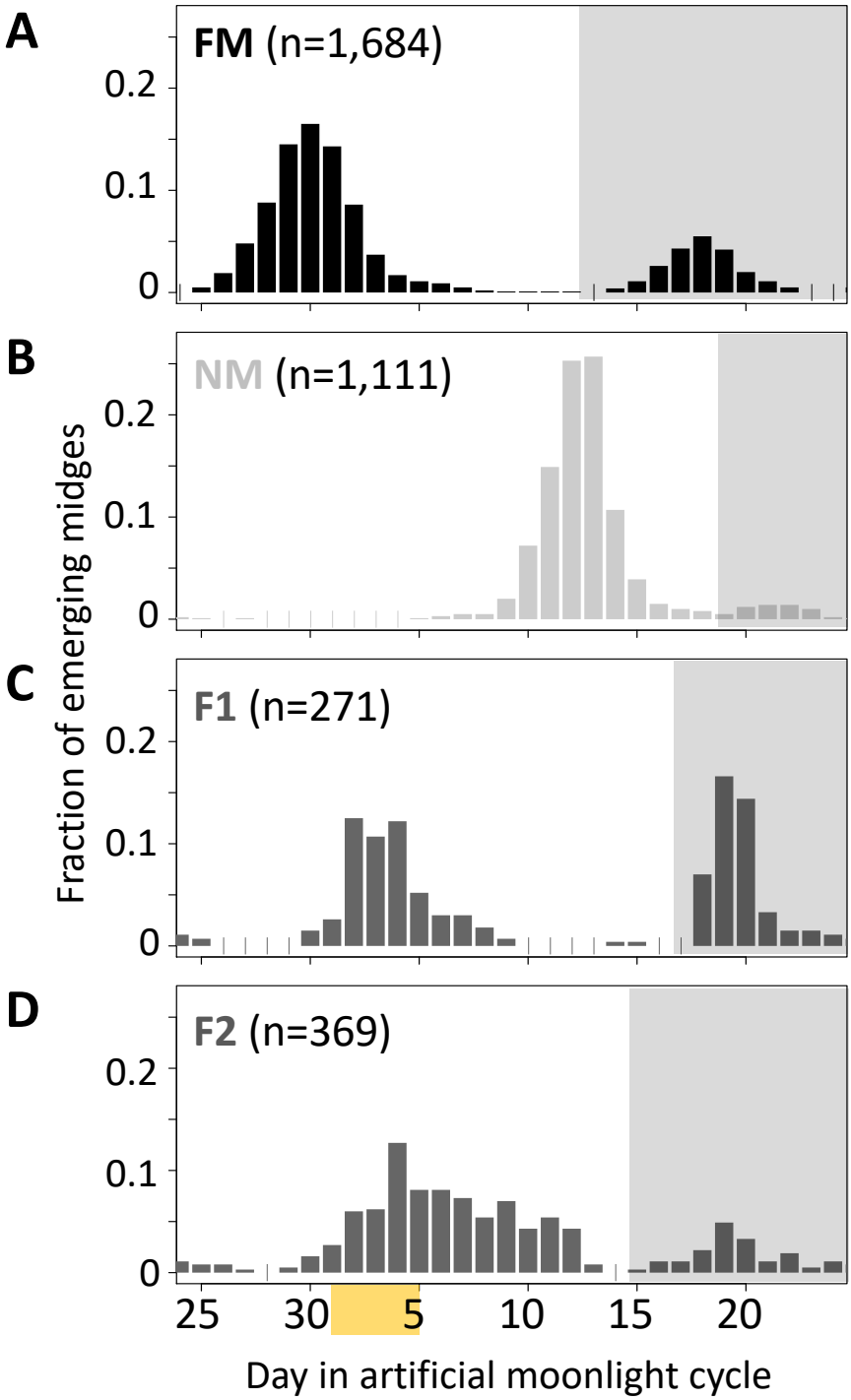


Figure 2 Inversions in chromosomes 2 and 3. (**A**) Chromosome arm 2L harbors a block of mildly elevated genetic differentiation. (**B**) Several blocks of long-range LD point to additional inversions in the FM and NM strains, one on each chromosome arm. (**C**) Structural variant (SV) calling from long read sequencing data supports the inversion on chromosome arm 2R. (**D-G**) Principal component analysis (PCA) for chromosomal sub-windows corresponding to suggested inversions separates the individuals into clusters corresponding to standard haplotype homozygotes (SS), inversion homozygotes (II) and inversion heterozygotes (SI). The four individuals which were already found distinct in whole genome analysis (red question mark) cannot be assessed. (**H-K**) Observed heterozygosity is clearly elevated in inversion heterozygotes, underpinning substantial genetic differentiation between the inversions. (**L-O**) The frequencies of the haplotypes do not differ much between the FM and NM types.



593

594 **Figure 3 Lunar emergence time is heritable.** In a cross between FM type (**A**) and NM type
595 (**B**), the F1 hybrids emerges intermediate between the parents (**C**). In the F2 (**D**), the
596 phenotypes spread out again, but do not segregate completely, suggesting that more than
597 one locus controls lunar emergence time. The peaks around day 20 (grey shading) are direct
598 responses to the artificial moonlight treatment (see [21]). They do not occur in the field and
599 were not considered in our analyses. Yellow shading indicates the days with artificial
600 moonlight treatment in the laboratory cultures.
601

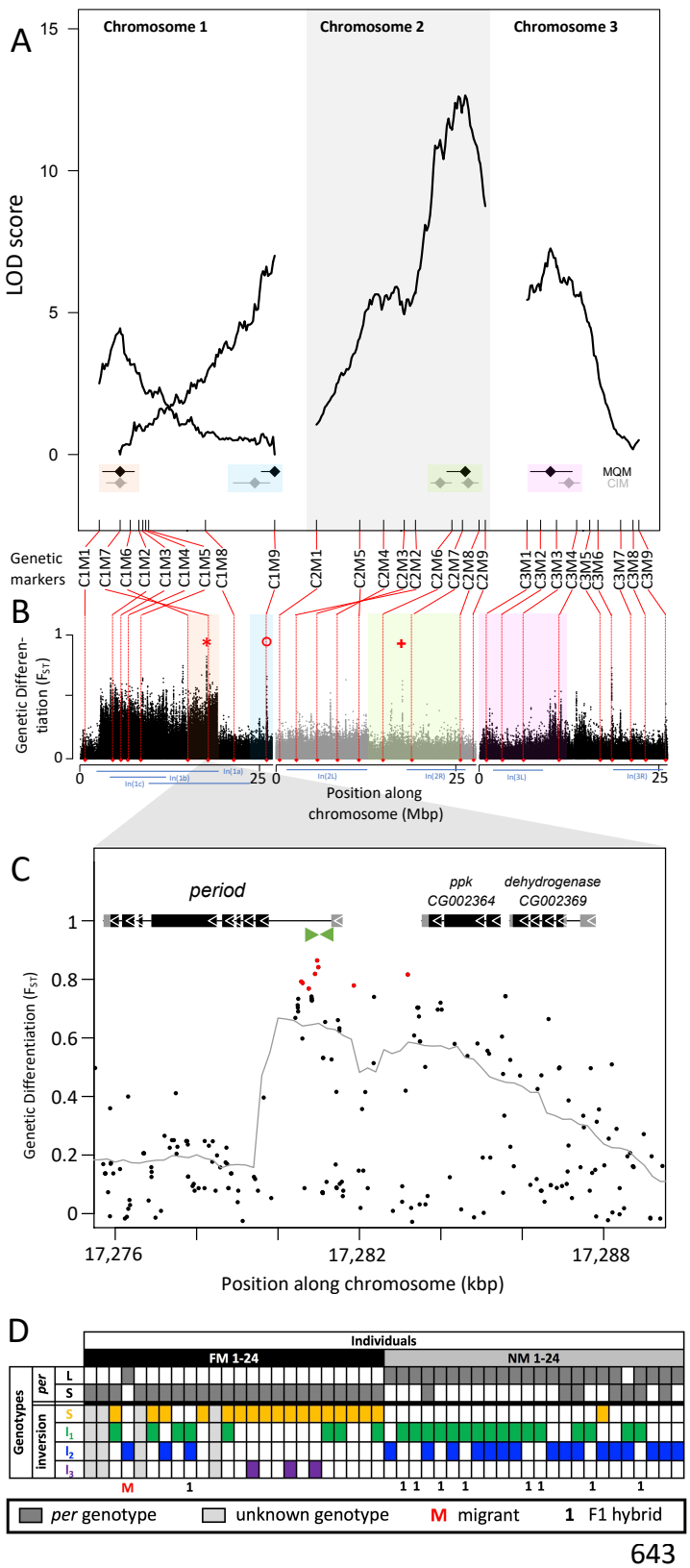


Figure 4 Quantitative trait loci (QTL) and genomic regions underlying FM vs NM emergence. (**A**) Multiple QTL Mapping (MQM) identified four significant QTL controlling lunar emergence time. Confidence intervals were determined in MQM and Composite interval mapping (CIM). The marker order supports In(1a) and the nested In(1b) on chromosome 1, as well as In(2L) on chromosome 2. (**B**) QTL intervals were colour-coded, transferred to the reference sequence and assessed for loci which are strongly differentiated (F_{ST}) between the FM and NM types. Inversions are represented by blue bars below the plot. The QTL in In(1a) overlaps with the most differentiated locus in the genome (red asterisk) – the *period* locus. Other markedly differentiated loci within the QTL intervals are the *plum/cask* locus (red circle) and the *stat1* locus (red cross). (**C**) Genetic differentiation is strongest in the first intron and the intergenic region just upstream of the *period* gene. (**D**) The *period* locus was re-genotyped based on an insertion-deletion mutation in the first intron (green arrows in (C)). The long (L) and short (S) *period* alleles are only loosely associated with inversion haplotypes.

643