The genomic basis of circadian and circalunar timing adaptations in a midge

Tobias S. Kaiser\textsuperscript{1,2,3,9}, Birgit Poehn\textsuperscript{1,3}, David Szkiba\textsuperscript{2}, Marco Preussner\textsuperscript{4}, Fritz J. Sedlazeck\textsuperscript{2,10}, Alexander Zrim\textsuperscript{2}, Tobias Neumann\textsuperscript{1,2}, Lam-Tung Nguyen\textsuperscript{2,5}, Andrea J. Betancourt\textsuperscript{6}, Thomas Hummel\textsuperscript{3,7}, Heiko Vogel\textsuperscript{8}, Silke Dorner\textsuperscript{1}, Florian Heyd\textsuperscript{4}, Arndt von Haeseler\textsuperscript{2,3,5}, Kristin Tessmar-Raible\textsuperscript{4,3}

\textsuperscript{1} Max F. Perutz Laboratories, University of Vienna, Campus Vienna Biocenter, Dr. Bohr-Gasse 9/4, 1030 Vienna, Austria

\textsuperscript{2} Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna and Medical University of Vienna, Dr. Bohr-Gasse 9, 1030 Vienna, Austria

\textsuperscript{3} Research Platform “Rhythms of Life,” University of Vienna, 1030 Vienna, Austria

\textsuperscript{4} Department of Biology, Chemistry, Pharmacy, Institute of Chemistry and Biochemistry, FU Berlin, 14195 Berlin, Germany

\textsuperscript{5} Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria

\textsuperscript{6} Institute of Population Genetics, Department of Biomedical Sciences, University of Veterinary Medicine Vienna, Josef-Baumann-Gasse 1, 1210 Vienna, Austria

\textsuperscript{7} Dept. of Neurobiology, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

\textsuperscript{8} Department of Entomology, Max Planck Institute for Chemical Ecology, Hans-Knöll-Straße 8, 07745 Jena, Germany
9 current address: Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306 Plön, Germany

10 current address: Department of Computer Science, Johns Hopkins University, Baltimore, MD 21211, USA

Contact Information

kristin.tessmar@mfpl.ac.at,
kaiser@evolbio.mpg.de
Good timing is crucial in life. Thus, organisms use endogenous clocks to anticipate regular environmental cycles, such as days and tides. Natural variants resulting in differently timed behavior or physiology, known as chronotypes in humans, are little understood at the molecular level. We generated a high-quality reference genome of *Clunio marinus*, a marine midge whose reproduction is timed by circadian and circalunar clocks. Midges from different locations show strain-specific genetic timing adaptations. By studying genomic sequence variations in five *C. marinus* strains we identified genes associates with circalunar and circadian chronotypes, none encoding core circadian clock genes. The region most strongly associated with circadian chronotypes generates strain-specific abundance differences of Ca$^{2+}$/Calmodulin-dependent kinase II.1 (*CaMKII.1*) splice variants. As equivalent variants were shown to alter CaMKII activity in *D.melanogaster*, and Cma-CaMKII.1 increases the transcriptional activity of the *Cma-Clock/Cma-Cycle*, we suggest alternative splicing modulation as a mechanism for natural adaptation in circadian timing.

Around new or full moon, during a few specific hours surrounding low tide, millions of non-biting midges of the species *Clunio marinus* emerge out of the sea to perform their nuptial dance. Adults live only a few hours, during which they mate and oviposit. They must therefore emerge synchronously and – given that embryonic, larval and pupal development take place in the sea – at a time when the most extreme tides reliably expose the larval habitat. The lowest low tides occur predictably during specific days of the lunar month at a specific time of day. Consequently, adult emergence in *C. marinus* is under the control of circalunar and circadian clocks$^{1,2}$. Importantly, while the lowest low tides recur
invariably at a given location, their timing differs between geographic locations\(^3\). Congruently, \textit{C. marinus} strains from different locations (Extended Data Fig. 1a) show local adaptation in circadian and circalunar emergence times (Extended Data Fig. 1b,c). Crosses between the \textit{Jean} and \textit{Por} strains showed that the differences in circadian and circalunar timing are genetically determined\(^4,5\) and largely explained by two circadian and two circalunar quantitative trait loci (QTLs)\(^6\).

Studies on timing variation or chronotypes in animals and humans have often focused on candidate genes from the circadian transcription-translational oscillator: In \textit{D. melanogaster}, polymorphisms in the core circadian clock genes \textit{period}, \textit{timeless} and \textit{cryptochrome} are associated with adaptive differences in temperature compensation\(^7\), photo-responsiveness of the circadian clock\(^8\) and emergence rhythms\(^9\). While these studies offer insights into how evolution can tinker with known circadian clock molecules, genome-wide association studies\(^10,11\) and other forward genetic approaches (reviewed in\(^12\)) are essential to provide a comprehensive, unbiased assessment of natural timing variation, for instance underlying human sleep-phase disorders. While the alleles underlying human sleep disorders embody disease states and the adaptive nature of human chronotypes is obscure, the chronotypes of \textit{C. marinus} clearly represent evolutionary adaptations to their habitat. Our study aims to illustrate by which genes and mechanisms an organism can successfully adapt to its specific ecological “timing niche”. In addition, the genetic dissection of adaptive natural variants of non-circadian rhythms\(^13\), as also present in \textit{C. marinus}, may provide an entry point into their unknown molecular mechanisms.
As a starting point for these analyses, we sequenced, assembled, mapped and annotated a *C. marinus* reference genome.

**The Clunio genome and QTLs for timing**

Our reference genome CLUMA_1.0 from the *Jean* laboratory strain contains 85.6 Mb of sequence (Table I), close to the previous flow-cytometry estimate of 95 Mb\(^6\), underlining that chironomids have generally small genomes\(^{14-16}\). The final assembly has a scaffold N50 of 1.9 Mb. Genome-wide genotyping of a mapping family with Restriction-site Associated DNA (RAD) sequencing allowed anchoring of 92% of the reference sequence consistently along a genetic linkage map (Fig. 1a, Extended Data Fig. 2), improving the original linkage map (Supplementary Method 5). Automated genome annotation resulted in 21,672 gene models. Protein similarity and available transcripts support 14,041 gene models (Table S1), within the range of gene counts for *Drosophila melanogaster* (15,507) and *Anopheles gambiae* (13,460). Thus, the very small *C. marinus* genome appears to be complete (Table I; Extended Data Figure 2a; Supplementary Note 1; Table S2). The *C. marinus* reference genome makes chironomids the third dipteran subfamily with an annotated genome reconstructed to chromosome-scale (Fig. 1a, Extended Data Fig. 2, 3b-f) and thus represents a valuable resource for comparative genomics.

To test the quality of our reference genome, we performed a basic genome characterization and comparison to other dipterans. We delineated the five *C. marinus* chromosome arms (Supplementary Note 2; Extended Data Fig. 3c; Table S3), homologized them to *D. melanogaster* and *A. gambiae* by synteny comparisons (Extended Data Fig. 3 and 4, Supplementary Note 2; Table S3),
found the ZW-like sex-linked locus in *C. marinus* outside the X chromosome homolog (Supplementary Note 2), and detected an elevated rate of chromosomal re-arrangements (Fig. 1a; Supplementary Note 3; Extended Data Fig. 2, 3b-f, 4). Taken together, the *C. marinus* reference genome appears well assembled.

As the next step towards identifying the molecular basis of circadian and circalunar timing adaptations in *C. marinus*, we refined the previously identified timing QTL positions based on the new high-density RAD markers (Table S4; Supplementary Note 4) and determined the reference sequence corresponding to the QTL confidence intervals (Fig. 1, orange and cyan bars; Table S4). None of the core circadian clock genes locates within the QTLs (Fig. 1a). Only *timeout/timeless2*, a *timeless* homolog with a minor role in circadian clock resetting, is located within the QTLs.

**Genetic variation in *Clunio* timing strains**

We then re-sequenced the *Por* and *Jean* strains (Extended Data Fig. 1), for which the initial QTL analysis was performed. Two pools of 300 field-caught individuals were sequenced at >240x coverage (Table S5). Mapping reads against the reference genome identified 1,010,052 single nucleotide polymorphisms (SNPs), 72% of them being present in both the *Por* and *Jean* strains. Based on all SNPs we determined genetic differentiation (*F*$_{ST}$), genetic diversity (*θ*), and short-range linkage disequilibrium (LD; measured as *r*$_{2}$) (Fig. 1b; Extended Data Fig. 3c and 5a,b).

Genome-wide genetic differentiation between the *Por* and *Jean* strains is moderate (*F*$_{ST}$ = 0.11), providing a good basis for screening the genome for local timing adaptation based on genetic divergence. According to QTL analysis, the
two circadian QTLs explain 85% of the daily timing difference, and the two
circlunar QTLs explain the entire monthly timing difference (Table S4 and 6). As
each locus therefore has a strong effect on timing, selection against maladapted
alleles must be strong and timing loci should be strongly differentiated.

Within the QTLs’ confidence intervals, 158 SNPs and 106 indels are strongly
differentiated (F<sub>ST</sub>≥0.8; Fig. 1b; Extended Data Fig. 5; SNPs: red dots in F<sub>ST</sub> panels,
for genome-wide comparison see Supplementary Note 5). We compiled a list of
candidate genes for circadian and circlunar timing adaptations based on their
proximity to differentiated SNPs and indels in the QTLs (Table S6). The
candidate genes do neither comprise core circadian clock genes
(*timeless2*/timeout: max. F<sub>ST</sub> ≤ 0.5; average F<sub>ST</sub> = 0.07), nor are they enriched for
any particular pathway (GO-term analysis; Table S7).

Timing phenotype with genotype correlation

Given that the alleles responsible for timing adaptation likely originated from
standing genetic variation (Supplementary Note 5), genetic variation at timing
loci should not vary freely between strains, but rather strains with similar timing
should share functionally relevant alleles. To identify such loci, we extended the
genomic screen to three additional strains: Vigo, Helgoland (He) and Bergen (Ber;
see Extended Data Fig. 1; Table S5 and S8). We then tested all five sequenced
strains for correlations between genetic differentiation (F<sub>ST</sub>) and timing
differences, or geographic distances as a null model (Table S8).

Overall, genome-wide genetic differentiation is not correlated with circadian
(r = 0.10, p = 0.31) or circlunar (r = 0.56, p = 0.12) timing differences, but with
geographic distance (“isolation by distance”; r = 0.88, p = 0.008). Against this
genomic background signal of isolation by distance, we screened the genome in 5kb sliding windows for peaks of correlation between genetic differentiation and timing, resulting in a correlation score (Fig. 1b and Extended Data Fig. 5a,b, CS panels; 0 to 5; for details see Methods). Combining the evidence from the Por vs. Jean strain F_{ST} screen (Table S6) with these patterns of correlation between timing and genetic divergence reduced the candidate gene list to 49 genes (Table S9).

Particularly noteworthy, a single region in circadian QTL C2 is strikingly differentiated (Fig. 1b). In this region, LD in the Por strain is significantly elevated (permutation test; p = 0.002), and diversity significantly decreased in some stretches (permutation test; p = 0.037 and 0.020), compared to the Por genome average. This may indicate a recent episode of selection in Por, potentially during timing adaptation, as this region is also strongly enriched for timing-correlated polymorphisms (Fig. 1b, CS panel). The most extreme values of genetic differentiation, genetic diversity and timing correlation localize to the Ca2+/Calmodulin dependent Kinase II.1 (CaMKII.1) locus and the anterior section of a gene homologous to the big bang (bbg) gene.

CaMKII affects the circadian core clock

The CaMKII.1 locus not only harbors the highest number of differentiated polymorphisms (Table S9), but CaMKII has been shown to affect circadian timing. Mouse CaMKIIα phosphorylates CLOCK and facilitates its dimerization with BMAL in vivo^{18}. An inactive CaMKIIα enzyme (“kinase-dead”-mutation; K42R) leads to dampened circadian rhythms, and a lengthened circadian free-running period^{18}. CaMKII in Drosophila S2 cells also phosphorylates the CLOCK
protein\textsuperscript{19}, and inhibition of \textit{Dme-CaMKII} in a sensitized background with reduced [Ca\textsuperscript{2+}] levels lengthens the circadian free-running period\textsuperscript{20}, suggesting that the role of CAMKII in circadian timing is conserved across animals.

To verify if CaMKII can also affect the circadian core clock in \textit{C. marinus}, we tested the effect of \textit{Cma-CaMKII.1} in a S2 cell-based assay\textsuperscript{19,21}. We repeated previous experiments\textsuperscript{19} showing that the chemical inhibition of endogenous \textit{Drosophila} CaMKII reduces the amount of generated luciferase (Extended Data Fig. 6a), while addition of a [Ca\textsuperscript{2+}]-independent variant of CaMKII (mouse T286D) increases luciferase amounts (Extended Data Fig. 6b). Then we generated constructs for \textit{C. marinus clock}, \textit{C. marinus cycle}, as well as mutated kinase-dead (K42R) and [Ca\textsuperscript{2+}]-independent (T286D) versions of \textit{Cma-CaMKII.1}.

Transfection of \textit{Cma-clock} and \textit{Cma-cycle} into S2 cells leads to luciferase activity driven from the 3x69 per-promoter (Fig. 2a). The addition of [Ca\textsuperscript{2+}]-independent \textit{Cma-CaMKII.1} leads to a significant increase in the luciferase signal (Fig. 2a), whereas addition of the kinase-dead \textit{Cma-CaMKII.1} does not enhance luciferase activity (Fig. 2a). This set of experiments strongly suggests that CaMKII kinase activity enhances E-box dependent transcription via the CLOCK/CYCLE dimer in \textit{C. marinus}.

\textit{CaMKII.1} splicing correlates with timing

But how can the polymorphisms in the \textit{Cma-CaMKII.1} locus affect the enzyme? We found two \textit{CaMKII.1} alleles: one in the early emerging \textit{Por}, \textit{He} and \textit{Ber} strains, and another in the late emerging \textit{Jean} and \textit{Vigo} strains. Most strain-specific polymorphisms are located in introns (Fig. 2b,c; TableS9). If they are meaningful, they should affect \textit{CaMKII.1} expression and/or splicing. \textit{Cma-CaMKII.1} has four functional domains (Fig. 2b)\textsuperscript{22}. The majority of differentiated polymorphisms
cluster in the region of the variable linker domain (compare Fig. 2b,c), including a 125bp insertion (red dot in Fig. 2c; Extended Data Fig. 7). We identified four alternatively spliced full-length transcripts of *C. marinus* CaMKII.1 (RA-RD), which differ in the linker length (Fig. 2b). High-coverage RNA sequencing gave evidence for differential exon usage between the *Jean* and *Por* strains, as well as for previously non-annotated exons within the variable linker region (Extended Data Fig. 6c). PCR and Sanger sequencing confirmed several partial transcripts of additional splice variants of the linker region (RE to RO; Fig. 2b). We used transcript-specific qPCR to quantify all transcripts. Generally, transcripts RE to RO are very lowly expressed. Of those, only RO showed quantifiable expression differences between the *Jean* vs. *Por* strains (Fig. 3a, Extended Data Fig. 6d).

Importantly, transcript-specific qPCR confirmed significant differential expression of the major transcripts in the *Jean* vs. *Por* strains (Fig. 3a, Extended Data Fig. 6d), matching the RNAseq data (Extended Data Fig. 6c). Consistently, variants with long linkers (RA, RB) are higher expressed in the *Por* strain and shorter variants (RD, RO) are higher expressed in the *Jean* strain (Fig. 3a, Extended Data Fig. 6c,d).

If the detected differences in *CaMKII.1* splice variant abundance are associated with the timing differences, they should be directly caused by the strain-specific polymorphisms at the *CaMKII.1* locus. In order to test this, we generated minigenes that contained the alternatively spliced linker region of the *CaMKII.1* locus from either the *Jean* or the *Por* strain. The two minigenes were transfected into *Drosophila* S2R+ cells and expression of splice variants was analyzed by radioactive RT-PCR (Fig. 3b,c). We detected four variants, corresponding to splice variants RB, RC, RD and RO. All variants show the same strain-specific
abundance differences in the S2R+ cell assay and in *C. marinus in vivo* (Fig. 3a,b).

Since the cellular context is the same for both the *Jean* and *Por* minigenes in the S2R+ assay, *trans*-acting elements can be excluded as the cause of differential splicing, implying that it is a direct result of the genomic sequence differences at the *Cma-CaMKII.1* locus. While splice variants RB, RC and RD and their constituting exons are conserved in *D. melanogaster* (see Flybase annotations and 23), a *D. melanogaster* RA counterpart does not exist. This may explain why this variant is undetectable in S2R+ cells.

**From splice variants to timing differences**

CaMKII linker-length variants have been investigated in several species. *D. melanogaster* CaMKII isoforms corresponding to the RB, RC and RD variants of *C. marinus*, have different substrate affinities and rates of target phosphorylation23. These activity differences are explained by the fact that CaMKII functions as a dodecamer, and the linker length determines the compactness and thus the substrate accessibility of the holoenzyme – enzymes with long linkers have higher activity. This structure-functional relationship is likely universal, as it is conserved between humans and *C. elegans*22,24.

Inactivation or inhibition of CaMKII lengthens circadian period in mouse and fruit flies18,20. A connection between circadian period length and phase of activity in light/dark cycles is known from *per* mutations in *D. melanogaster*25 and human chronotypes26. These findings imply that in *C. marinus* the more active and more readily [Ca$^{2+}$]-activated long-linker *CaMKII.1* variants should advance adult emergence by shortening the circadian clock period. Indeed, we find that the early emerging *Por* and *He* strains, which possess the same long-linker biased
CaMKII.1 alleles, have shorter free-running circadian clock periods than the late emerging Jean strain (Fig. 3d).

Integrating our results with those from the aforementioned literature, the scenario emerges that regulating the ratio of CaMKII.1 splice variants constitutes an evolutionary mechanism to adapt circadian timing (Extended Data Fig.8): CaMKII.1 genomic sequence differences lead to differential CaMKII.1 splicing and activity. Among a number of possible targets this impacts on CLOCK/CYCLE dimer-dependent transcription, which in turn affects circadian period length and ultimately results in adult emergence time differences.
Annual, lunar, and tidal rhythms, as well as natural timing variation between individuals, are important and widespread, yet poorly understood, phenomena. The comprehensively mapped *C. marinus* reference genome and the genetic variation panel for five strains with differing circadian and circalunar timing establish new resources to gain insight into these topics.

We identified *C. marinus* orthologs for all core circadian clock genes, none of which appears to be involved in circadian or circalunar timing adaptations. For circalunar timing, this supports the molecular independence of the circalunar clock from the circadian clock as reported for *Platynereis dumerilii*.

For circadian timing, strain-specific modulation in alternative splicing of *CaMKII.1* emerges as a likely mechanism for natural adaptation. In the light of previous experiments in *Drosophila* and mouse, it seems most likely that differences in CaMKII activity of the different splice forms lead to circadian timing differences via phosphorylation of CLOCK/CYCLE (Extended Data Fig. 8).

It is also conceivable that CaMKII affects circadian timing via other targets. For example, CaMKII is known to phosphorylate the cAMP response element binding protein (CREB). CREB is linked to the circadian clock by cAMP response elements (CRE) in the promoters of the *period* and *timeless* genes, and by physical interaction of the CREB binding protein (CBP) with CREB, CLOCK and CYCLE. Furthermore, one of CaMKII’s best-studied roles is the morphological modulation of neuronal plasticity and connectivity. Such changes in connectivity have been increasingly implicated as part of the circadian timing mechanism in *Drosophila* and mammals. Interestingly,
CaMKII’s role in shaping neuronal connectivity has also been suggested to link to several neuropsychiatric diseases\textsuperscript{38}, which often co-occur with chronobiological disruptions\textsuperscript{39-42}. Could the modulation of CaMKII activity constitute a molecular link between these phenomena?
References


Supplementary Information is linked to the online version of the paper at www.nature.com/nature.
Acknowledgments

We thank the members of the Tessmar-Raible, Raible, von Haeseler groups for discussions, S. Bannister and F. Raible for comments on the manuscript. The research leading to these results has received funding from the research platform “Rhythms of Life” of the University of Vienna to KT-R, TH and AvH, the FWF (http://www.fwf.ac.at/) START award (#AY0041321), the HFSP (http://www.hfsp.org/) research grant (#RGY0082/2010), the ERC (FP7/2007-2013)/ERC Grant Agreement 337011 to KT-R and the DFG grant (#HE5398/4-1) to FH. TSK was supported by the Vienna International PostDoctoral Program for Molecular Life Sciences (VIPS), L-TN and AvH by the University of Vienna Initiativkolleg I059-N. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Author Contributions

Conceived and designed the study, interpreted the data: TSK, KT-R, AvH; field sampling, chronobiological experiments, (super)scaffolding, genetic/QTL mapping, population genomics: TSK; assembly to contigs: L-TN, TN, TSK; assembly filtering: TSK, TN; gap closing and repeated edge removal: TSK, FJS; RNAseq: TSK, FJS, HV; cDNA library: HV; genome annotation: TSK, DS; analysis of genome completeness: TSK, TN; chromosome homology and synteny comparisons: TSK, AZ; estimation of linkage disequilibrium: AJB; SNP effects and GO term analysis: DS, TSK; Cma-CaMKII.1 analyses: BP, TSK, SD; minigene assay: MP, FH; contributed material: TH; wrote the manuscript: TSK, KT-R.
Author Information

All sequence data are deposited in the European Nucleotide Archive (ENA) under PRJEB8339. The reference genome is also on ClunioBase (http://cluniobase.cibiv.univie.ac.at). Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials: K.T-R. (kristin.tessmar@mfpl.ac.at) and T.S.K. (kaiser@evolbio.mpg.de).
Fig. 1 Identification of candidate regions in the timing QTLs by combined genetic and molecular maps.

(a) The three linkage groups of *C. marinus* with reference scaffolds (right) anchored on a genetic linkage map (left). Scaffolds ordered and oriented: black bars; not oriented: grey bars; neither ordered nor oriented: white bars. Grey shadings: large non-recombining regions. QTLs: circadian (orange), circalunar (cyan). One circadian and circalunar QTL overlap, resulting in three physical QTL regions. (b) Population genomic analysis of QTL-C2. Panels 1-3: *Por* vs. *Jean* strains (blue vs. red in panel 2,3). *Panel 1*: Genetic differentiation, *Panel 2*: Genetic diversity (*θ*) in 20-kb (thin line) and 200-kb (thick line) windows. *Panel 3*: Linkage disequilibrium (*r*²). *Panel 4*: Correlation Score (CS; 0 to 5) for genetic differentiation with circadian timing (top), circalunar timing (middle) and geographic distance (bottom) for *Vigo, Jean, Por, He, Ber* strains. Bottom numbers: scaffold IDs. For further details, including QTLs C1/L1 and L2, see Extended Data Fig. 5a,b.
Fig. 2 CaMKII.1 regulates CLK/CYC transcriptional activity and exhibits strain specific splice variants

(a) Additional *C. marinus* CaMKII.1 increases the transcriptional activity of *C. marinus* Clk and Cyc in a S2 cell luciferase assay using the 3x69 E-box containing enhancer. (means; error bars: S.E.M.; two-sided Welch Two Sample t-test; biological replicates (BR): n=5, except for no clk control n=3; each BR represents the average of three prep replicates, *** p<0.0005). (b) Exons of full (RA-RD) and partial (RE-RO) *Cma-CaMKII.1* transcripts. (c) Distribution of SNPs (black), indels (orange) and a 125bp-insertion (red dot), all with $F_{ST} \geq 0.8$.  

---

26
Fig. 3  *CaMKII.1* splicing depends on the genomic sequence variations and correlate with endogenous circadian period lengths

(a) qPCR values for *CaMKII.1* splice variants from *Por* vs. *Jean*, normalized to *Por* (non-normalized: Extended Data Fig. 6d); BR: *Por* n=9, *Jean* n=10; RO: *Por* n=3, *Jean* n=8, since RO was not detectable in six *Por* BRs, suggesting an even larger expression difference; means; error bars: S.E.M.; two-sided Wilcoxon rank sum test; *p<0.05; **p<0.005; ***p<0.0005, “ns” non-significant; Holm correction for multiple testing. (quantitative RNAsequencing of further BRs: Extended Data Fig. 6c)  (b) Differential splicing of the *CaMKII.1* linker region in S2R+ cells, normalized to *Por*, BR: n=7; two-sided Two Sample T-test, otherwise as (a).  (c) Representative phosphorimaging gel sections as quantified for (b), two separate lanes from the same gel (full gel: source data).  (d) Free-running rhythm of adult emergence under constant dim white light (∼100 lux). *He* and *Por* share *CaMKII.1* alleles, while *Jean* posses the other. Free-running period: time between subsequent emergence peaks was averaged, weighting each peak by the number of individuals.
Table I

<table>
<thead>
<tr>
<th></th>
<th>Clunio marinus</th>
<th>Danaus plexippus&lt;sup&gt;43&lt;/sup&gt;</th>
<th>Tribolium castaneum&lt;sup&gt;44&lt;/sup&gt;</th>
<th>Apis mellifera&lt;sup&gt;45&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total bases (Mbp)</td>
<td>86</td>
<td>278</td>
<td>160</td>
<td>236</td>
</tr>
<tr>
<td>Mapped sequence (%)</td>
<td>92</td>
<td>NA</td>
<td>90</td>
<td>79</td>
</tr>
<tr>
<td>Scaffold N50 (Mbp)</td>
<td>1.9</td>
<td>0.2</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>Contig N50 (kbp)</td>
<td>79</td>
<td>50</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>AT content (%)</td>
<td>68.3</td>
<td>68.4</td>
<td>67</td>
<td>66</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>98.0</td>
<td>98.5</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Platform</td>
<td>Illumina</td>
<td>Illumina + 454</td>
<td>Sanger + BAC</td>
<td>Sanger + BAC</td>
</tr>
</tbody>
</table>

Comparison of the *C. marinus* reference assembly with published model insect genomes

Machine readable superscaffolding data and the computer source code for the removal of repeated edges are supplied as source data files.
Online Methods

Animal culture and light regimes

The *Clunio marinus* laboratory stocks were bred according to Neumann\(^1\), care was provided by the MFPL aquatic facility. Briefly, they were kept in 20x20x5cm plastic containers with sand and natural seawater diluted to 15\(^\circ\)o with desalted water, fed diatoms (*Phaeodactylum tricornutum*, strain UTEX 646) in early larval stages and nettle powder in later stages. Temperature in the climate chambers was set to 20\(^\circ\)C and the light dark cycle (LD) was 12:12 (except where noted differently). Moonlight was simulated with an incandescent flashlight bulb (about 1 Lux), which was switched on all night for four successive nights every 30 days.

Genome assembly

The genome assembly process (Extended Data Fig. 9a) was based on three sequencing libraries (Table S10): A 0.2kb insert library was prepared from a single adult male of the *Jean* laboratory strain (established from field samples taken at *St. Jean-de-Luz*, France, in 2007; >12 generations in the laboratory), which was starved and kept in seawater with Penicillin (60 units/ml), Streptomycin (60 \(\mu\)g/ml) and Neomycin (120 \(\mu\)g/ml) during the last 2 weeks of development. DNA was extracted with a salting out method\(^6\), sheared on a Covaris S2 sonicator (frequency sweeping mode; 4\(^\circ\)C; duty cycle: 10\%; intensity: 7; cycles/burst: 300; microTUBE AFA Fiber 6x16 mm; 30 s) and prepared for Illumina sequencing with standard protocols. A 2.2kb and a 7.6kb insert library were prepared from a polymorphic DNA pool of >300 field-caught *Jean* adult males by Eurofins MWG Operon (Ebersberg, Germany) according to their
proprietary protocol. Each library was sequenced in one lane of an Illumina HiSeq2000 with 100bp paired-end reads at the Next Generation Sequencing unit of the Vienna Biocenter Core Facilities (VBCF; http://vbcf.ac.at).

Reads were filtered for read quality, adapter and spacer sequences with cutadapt\textsuperscript{17} (-b 4 -n 3 -e 0.1 -O 8 -q 20 -m 13) and deduplicated with fastq-mcf from ea-utils\textsuperscript{18} (-D 70). Read pairs were interleaved with ngl-utils\textsuperscript{19}, leaving only paired reads. Contamination with human DNA found in the 0.2kb library was removed by deleting reads matching the human genome at a phred-scaled quality score ≥20 (alignment with BWA\textsuperscript{50}).

Assembly into contigs with Velvet\textsuperscript{51} (scaffolding disabled; 57bp kmers as determined by VelvetOptimiser\textsuperscript{52}) was based solely on the less polymorphic 0.2kb library. About 600 remaining adaptor sequences at the ends of assembled contigs were trimmed with cutadapt (-O 8 -e 0.1 -n 3). For assembly statistics see Table S11.

Scaffolding of the contigs was based on all three libraries and performed with SSPACE\textsuperscript{53} in two iterations, i.e. scaffolds from the first round were scaffolded again. Using different parameters in the iterations (Table S12) allowed different connections to be made and thus increased scaffold connectivity (Table S13). The effect is likely due to the polymorphic nature of the 2.2kb and 7.6kb libraries; it results in a “population-consensus most common arrangement of the scaffolds”. The iterative scaffolding process was performed with and without applying a size cutoff excluding contigs <1kb, resulting in two independent assemblies (CLUMA_0.3 and CLUMA_0.4; see Extended Data Fig. 9a), which differed in overall connectivity and sequence content (Table S11), but also in the identity and structure of the large scaffolds. In order to combine both
connectivity and sequence content, and in order to resolve the contradictions in
the structure of the largest scaffolds, the two assemblies were compared and
reconciled in a manual super-scaffolding process, as detailed in Supplementary
Method 1. Briefly, the overlap of scaffolds from the two assemblies was tested
with BLAST searches and represented in a graphical network structure. Scaffolds
with congruent sequence content in both assemblies would result in a linear
network, whereas scaffolds with contradictory sequence content would result in
branching networks. At the same time, both assemblies were subject to genetic
linkage mapping based on genotypes obtained from Restriction-site Associated
DNA sequencing (RAD sequencing) of a published mapping family6
(Supplementary Method 2). The resulting genetic linkage information served to
resolve the branching networks into the longest possible unambiguous linear
sub-networks with consistent genetic linkage information (see scheme A in
Supplementary Method 1). Finally, the structure of the resulting super-scaffolds
was coded in YAML format and translated into DNA sequence with Scaffolder54,
resulting in 75 mapped super-scaffolds.

The remaining small and unmapped scaffolds were filtered for fragments of the
mitochondrial genome, the histone gene cluster and 18S/28S ribosomal rDNA
gene cluster, which were assembled separately (Supplementary Method 3;
Extended Data Fig. 10). Unmapped scaffolds were also filtered for obvious
contamination from other species (Supplementary Method 3). The degree to
which the remaining unmapped scaffolds are leftover polymorphic variants of
parts of the mapped super-scaffolds was estimated by blasting the former
against the latter (Supplementary Method 3; Table S14).
All scaffolds were subject to gap closing with *GapFiller*\textsuperscript{55} and repeated edges, i.e. gaps with repetitive sequence at both sides that are generally due to genetic polymorphism, were assessed and if possible removed with a custom script (Supplementary Method 4; code available supplied as Source Data File).

The final assembly CLUMA\_1.0 was submitted under project PRJEB8339 (75 mapped scaffolds; 23,687 unmapped scaffolds $\geq$100bp). The assembly and further information can also be obtained from ClunioBase (http://cluniobase.cibiv.univie.ac.at)

**Reconstruction of chromosomes and QTL analysis**

Genetic linkage information for the final 75 super-scaffolds was obtained by repeating read mapping to genotype calling for the RAD sequencing experiment as described above (Supplementary Method 2), but now with assembly CLUMA\_1.0 as a reference. This allowed to place and orient super-scaffolds along the genetic linkage map (Fig. 1a, Extended Data Fig.2). The positions of the recombination events within a scaffold were approximated as the middle between the positions of the two RAD markers between which the marker pattern changed from one map location to the next. The published genetic linkage map was refined and revised (Supplementary Method 5; Extended Data Fig. 2). Based on the refined linkage map, QTL analysis of the published mapping family was repeated as described\textsuperscript{6} (Table S4; Supplementary Note 5). Using the correspondence between the reference assembly and the genetic linkage map, we were able to directly identify the genomic regions corresponding to the QTLs’ confidence intervals (Fig. 1, Extended Data Fig. 5a,b).

**Transcript sequencing**
From previous experiments assembled transcripts were available from a normalized cDNA library of all life stages and various *C. marinus* strains (454 sequencing) and RNA sequencing data was available for *Jean* strain adults (Illumina sequencing). Furthermore, specifically for genome annotation, RNA from 80 third instar larvae each from the *Jean* and *Por* laboratory strains was prepared for RNA sequencing according to standard protocols (Supplementary Method 6). Each sample was sequenced on a single lane of an Illumina HiSeq 2000. All transcript reads were submitted to the European Nucleotide Archive (ENA) under project PRJEB8339.

For the adult and larval RNA sequencing data, raw reads were quality checked with *fastqc*\(^66\), trimmed for adaptors quality with *cutadapt*\(^47\) and filtered to contain only read pairs using the interleave command in *ngm-utils*\(^49\). Reads were assembled separately for larvae and adults with *Trinity*\(^57\) (path_reinforcement_distance: 25; maximum paired-end insert size: 1500 bp; otherwise default parameters).

**Genome annotation**

Automated annotation was performed with MAKER2\(^58\). Repeats were masked based on all available databases in *repeatmasker*. MAKER2 combined evidence from assembled transcripts (see above), mapped protein datasets from *Culex quinquefasciatus* (Cpip1), *Anopheles gambiae* (AgamP3), *Drosophila melanogaster* (BDGP5), *Danaus plexippus* (DanPle_1.0), *Apis mellifera* (Amel4.0), *Tribolium castaneum* (Tcas3), *Strigamia maritima* (Smar1) and *Daphnia pulex* (Dappu1) and *ab initio* gene predictions with AUGUSTUS\(^59\) and SNAP\(^60\) into gene models. AUGUSTUS was trained for *C. marinus* based on assembled transcripts from the normalized cDNA library. SNAP was run with parameters for *A.*
mellifera, which had the highest congruence with known C. marinus genes in preliminary trials (Supplementary Method 7). MAKER was set to infer gene models from all evidence combined (not transcripts only) and gene predictions without transcript evidence were allowed. Splice variant detection was enabled, single-exon genes had to be larger than 250bp and intron size was limited to a maximum of 10 kb.

All gene models within the QTL confidence intervals, as well as all putative circadian clock genes and light receptor genes were manually curated: Exon-intron boundaries were corrected according to transcript evidence (~500 gene models), chimeric gene models were separated into the underlying individual genes (~100 gene models separated into ~300 gene models) and erroneously split gene models were joined (~15 gene models). Finally, this resulted in 21,672 gene models, which were given IDs from CLUMA.CG000001 to CLUMA.CG021672 ("CLUMA" for Clunio marinus, following the controlled vocabulary of species from the UniProt Knowledgebase; CG for "computed gene"). Splice variants of the same gene (detected in 752 gene models) were identified by the suffix "-RA", "-RB", etc. and the corresponding proteins by the suffix "-PA", "-PB", etc.

Gene models were considered as supported if they overlapped with mapped transcripts or protein data (Table S1). Gene counts for Drosophila melanogaster were retrieved from BDGP 5, version 75.546 and for Anopheles gambiae from AgamP3, version 75.3. The putative identities of the C. marinus gene models were determined in reciprocal BLAST searches, first against UniProtKB/Swiss-Prot (8,379 gene models assigned) and if no hit was found against nr at NCBI (1,802 additional genes assigned). Reciprocal best hits at an e-value < 1*10^-10
were considered putative orthologs (termed “putative gene X”), non-reciprocal
hits at the same e-value were considered paralogs (termed “similar to”). All
remaining gene models were searched against the PFAM database of protein
domains (111 gene models assigned; termed “gene containing domain X”). If still
no hit was found, the gene models were left unassigned (“NA”).

**Synteny comparisons**

Genome-wide synteny between the *C. marinus, D. melanogaster* and *A. gambiae*
genomes was assessed based on reciprocal best BLAST hits (e-value < 10*-10)
between the three protein datasets (Ensembl Genomes, Release 22, for *D.
melanogaster* and *A. gambiae*). Positions of pairwise orthologous genes were
retrieved from the reference genomes (BDGP 5, AgamP3 and CLUMA_1.0) and
plotted with Circos61. *C. marinus* chromosome arms were delimited based on
centromeric and telomeric signatures in genetic diversity and linkage
disequilibrium (Extended Data Fig. 3c; Table S3; for data source see “strain re-
sequencing” below). Homologies for *C. marinus* chromosome arms were assigned
based on enrichment with putative orthologous genes from specific chromosome
arms in *D. melanogaster* and *A. gambiae* (Extended Data Figures 3,4; Table S3).
Additionally, for the 5,388 detected putative 1:1:1 orthologs, microsynteny was
assessed by testing if all pairs of directly adjacent genes in one species were also
directly adjacent in the other species. The degree of microsynteny was then
calculated as the fraction of conserved adjacencies among all pairs of adjacent
genes. From this fraction the relative levels of chromosomal rearrangements in
the evolutionary lineage leading to *C. marinus* were estimated (Supplementary
Note 2; Extended Data Fig. 4).

**Strain re-sequencing**
Genetic variation in five *C. marinus* strains (Extended Data Fig. 1) was assessed based on pooled-sequencing data from field-caught males from the strains of St. Jean-de-Luz (*Jean*, Basque Coast, France; sampled in 2007; n=300), Port-en-Bessin (*Por*, Normandie, France; 2007; n=300), as well as Vigo (Spain; 2005; n=100), Helgoland (*He*, Germany; 2005; n=300) and Bergen (*Ber*, Norway; 2005; n=100). Samples from Vigo and Bergen, were provided by Dietrich Neumann and Christina Augustin, respectively. For each strain we chose the largest available number of individuals to get the best possible resolution of allele frequencies. Females are not available, because they are virtually invisible in the field. For an overview of the experimental procedure, see Extended Data Fig. 9b. DNA was extracted with a salting out method\(^{46}\) from sub-pools of 50 males, the DNA pools were mixed at equal DNA amounts, sheared and prepared as described above and sequenced on four lanes of an Illumina HiSeq2000 with paired-end 100 bp reads (*Ber* and *Vigo* combined in one lane, distinguished by index reads). All reads were submitted to the European Nucleotide Archive (ENA) under project PRJEB8339. Sequencing reads were filtered for read quality and adapter sequences with *cutadapt*\(^{47}\) (-b -n 2 -e 0.1 -O 8 -q 13 -m 15), interleaved with *ngm-utils*\(^{49}\) and deduplicated with *fastq-mcf* from *ea-utils*\(^{48}\) (-D 70). Reads were aligned to the mapped super-scaffolds of assembly CLUMA_1.0 with *BWA*\(^{50}\) (*aln* and *sampe*; maximal insert size (bp): –a 1500).

**Detection of re-arrangements**

Based on the unfiltered alignments, the samples from *Por* and *Jean* were screened for genomic inversions and insertion-deletions relative to the reference sequence with the multi-sample version of *DELLY*\(^{62}\). Paired-end information was
only considered if the mapping quality was high (q ≥ 20) (see also Supplementary Note 4).

**Population genomic analysis of the timing strains**

For population genomic analysis (Extended Data Fig. 9b), the alignments of the pool-seq data from *Vigo, Jean, Por, He* and *Ber* were filtered for mapping quality (q ≥ 20), sorted, merged and indexed with SAMtools. Reads were re-aligned around indels with the RealignerTargetCreator and the IndelRealigner in GATK. The resulting coverage per strain is given in Table S5.

For identification of single nucleotide polymorphisms (SNPs), a pileup file was created with the mpileup command of SAMtools. Base Alignment Quality (BAQ) computation was disabled (--B); instead, after creating a synchronized file with the mpileup2sync script in PoPulation2, indels that occurred more than ten times were masked (including 3bp upstream and downstream) with PoPoolations2's identify-indel-regions and filter-sync-by-gtf scripts. FST values were determined with the fst-sliding script of PoPolation2, applying a minimum allele count of 10 (so that any false-positive SNPs resulting from the remaining unmasked indels were effectively excluded) and a minimum coverage of 40x for the *Por vs. Jean* comparison or 10x for the comparison of all five strains. FST was calculated at single base resolution, as well as in windows of 5kb (step size: 1kb). Individual SNPs were only considered for further analyses or plotted if they were significantly differentiated as assessed by Fisher's exact test (*fisher-test* in PoPolation2).

Average genome-wide genetic differentiation between timing strains, as obtained by averaging over 5kb sliding-windows, was compared to the respective timing differences and geographic distances (see Table S8) in Mantel...
tests (Pearson's product moment correlation; 9,999 permutations), as implemented in the *vegan* package in the R statistical programming environment R<sup>66</sup>. Geographic distances and circadian timing differences were determined as described previously<sup>67</sup> (see Table S8). For determination of lunar timing differences when comparing lunar with semilunar rhythms see Supplementary Note 6. In order to find genomic regions for which genetic differentiation is correlated with the timing differences between strains, the Mantel test was then applied to 5kb genomic windows every 1kb along the reference sequence. 5kb is roughly the average size of a gene locus in *C. marinus*. Windows with a correlation coefficient of \( r \geq 0.5 \) were tested for significance (999 permutations). For each genomic position the number of overlapping significantly correlated 5kb windows was enumerated, resulting in a correlation score (CS; ranging from 0 to 5).

Genetic diversity, measured as Watterson's theta \( (\theta_W) \), for each strain was assessed with *PoPoolation1.1.2*<sup>68</sup> in 20kb windows at 10kb steps. In order to save computing time, the pileup files of *Jean*, *Por* and *He* were linearly downscaled to 100x coverage with the *subsample-pileup* script ("fraction" option), positions below 100x coverage being discarded. Indel regions were excluded (default in *PoPoolation 1.1.2*) and a minimum of 66% of a sliding window needed to be covered. SNPs were only considered in \( \theta_W \) calculations if present \( \geq 2 \) times, leading to slight inconsistencies in \( \theta_W \) estimates between strains due to differing coverage, but not affecting diversity comparisons within strains.

Linkage disequilibrium between the SNPs was determined for the *Por* and *Jean* strains with LDx<sup>69</sup>, assuming physical linkage between alleles on the same read or read pairs. \( r^2 \) was determined by a maximum likelihood estimator, minimum
and maximum read depths corresponded to the 2.5% and 97.5% coverage depths for each population (Jean: 111 to 315, Por: 98 to 319), total insert distance was limited to 600bp, minimum phred-scaled base quality was 20, minimum allele frequency was 0.1 and a minimum coverage per pair of SNPs was 11. SNPs were binned by their physical distance for the plots (0-200bp, 200-400bp, 400-600bp), with the mean value plotted.

Finally, small indels (<30bp) in the Por and Jean strains were detected with the UnifiedGenotyper (glm INDEL) in GATK54 for positions with more than 20x coverage. Genetic differentiation for indels was calculated with the classical formula $F_{ST} = (H_T-H_S) / H_T$, where $H_S$ is the average expected heterozygosity according to Hardy-Weinberg Equilibrium (HWE) in the two subpopulations and $H_T$ is the expected heterozygosity in HWE of the hypothetical combined total population. If more than two alleles were present, only the two most abundant alleles were considered in the calculation of $F_{ST}$.

**Assessment of candidate genes**

Gene models from the automated annotation were considered candidate genes, if they fulfilled the following criteria: (1) The gene was located within the reference sequence corresponding to the QTL confidence intervals as determined for the Por and Jean strains. (2) The gene contained a strongly differentiated SNP or small indel or they were directly adjacent to such a SNP or small indel ($F_{ST} \geq 0.8$ for Por vs. Jean, i.e. the strains used in QTL mapping). This resulted in a preliminary list of 133 genes based on the Por vs. Jean comparison (Table S6). These candidate genes were narrowed down based on their overlap with genomic 5kb windows, for which genetic differentiation between five
European timing strains correlated with their timing differences (Fig. 1a; Extended Data Fig. 5a,b; Table S9).

The location and putative effects of the SNPs and indels relative to the gene models were assessed with SNPeff\textsuperscript{70} (\textsuperscript{ud} 0, otherwise default parameters; Extended Data Fig. 5c,d; Table S6 and S9).

For Gene Ontology (GO) term analysis, all \textit{C. marinus} gene models with putative orthologs in the UniProtKB/Swiss-Prot and nr databases based on reciprocal best BLAST hits (see above) were annotated with the GO terms of their detected orthologs (6,837 gene models). Paralogs were not annotated. The enrichment of candidate SNPs and indels (\textit{F}_{\text{ST}}\geq 0.8 between \textit{Por} and \textit{Jean}) in specific GO terms was tested with SNP2GO\textsuperscript{71} (\textit{min.regions}=1, otherwise default parameters). Hyper-geometric sampling was applied to test if individual genes of a GO term or a whole pathway of genes are enriched for SNPs (Table S7).

**Molecular characterization of CaMKII.1**

RNAseq data of the \textit{Por} and \textit{Jean} strains for \textit{CaMKII.1} were obtained from the larval RNA sequencing experiment described above. Besides four assembled full-length transcripts (RA to RD) from RNAseq and assembled EST libraries, additional partial transcripts (RE to RO) were identified by PCR amplification (for PCR primers see Table S15), gel extraction (QIAquick Gel Extraction Kit, Qiagen), cloning with the CloneJET PCR Cloning Kit (Thermo Scientific) and Sanger sequencing with pJET1.2 primers (LGC Genomics & Microsynth). cDNA was prepared from RNA extracted from LIII larvae of the \textit{Por} and \textit{Jean} laboratory strains (RNA extraction with RNeasy Plus Mini Kit, Qiagen; reverse transcription with QuantiTect Reverse Transcription Kit, Qiagen).
qPCR was performed with variant-specific primers and actin as control gene (Table S16). cDNA was obtained from independent pools of 20 third instar larvae of the Por and Jean strains. Sample size was ten per strain to cover different time points during the day and to test for reproducibility (two samples each at Zeitgeber times 0, 4, 8, 16 and 20; for one Por sample extraction failed; RNA extraction and reverse transcription as above). qPCR was performed with Power SYBR Green PCR Master Mix on a StepOnePlus Real Time System (both Applied Biosystems). Fold-changes were calculated according to \(^2\) in a custom excel sheet. The assumption of equal variance was violated for the RD comparison (F-Test) and the assumption of normal distribution was violated for the data of RA and RC in the Por strain (Shapiro-Wilk normality test), possibly reflecting circadian effects in the samples from different times of day. Thus, expression differences were assessed for significance in a two-tailed Wilcoxon Rank Sum Test (wilcox.test in R\(^6\)). Holm correction\(^7\) was used for multiple testing (default in p.adjust function of R).

**CaMKII.1 minigenes**

PCR fragments containing the CaMKII.1 linker region (exons 10 to 15) were amplified from genomic Por or Jean DNA respectively with primers CaMKII-Sc61-F-344112 and CaMKII-Sc61-R-351298 (Table S15), cloned with the CloneJET PCR Cloning Kit (Thermo Scientific), transferred into the pcDNA3.1+ vector using NotI and XbaI (Thermo Scientific). These constructs were transfected into Drosophila S2R+ cells and RNA was prepared 48h post transfection. After DNAse digestion, isoform expression was analyzed by radioactive, splicing-sensitive RT-PCR (primers in Table S17) and Phosphorimager quantification as described\(^7\). Identity of isoforms is based on size and sequencing of PCR products. To test for
reproducibility, there were seven biological replicates (raw data in Table S18).

As the assumptions of equal variance (F-Test) and normal distribution of data
(Shapiro-Wilk normality test) were not violated, the significance of expression
differences was assessed in unpaired, two-sided two-sample t-tests. Holm
correction\textsuperscript{73} was used for multiple testing (default in \textit{p.adjust} function of R).

S2R+ cells were obtained from the lab of S. Sigrist, regularly authenticated by
morphology and routinely tested for absence of mycoplasma contamination. The
entire experiment was reproduced several months later with three biological
replicates (raw data in Table S18).

\textbf{S2 cell luciferase assay}

Firefly luciferase is driven from a \textit{period} 3x69 promoter under control of the
CLOCK and CYCLE protein\textsuperscript{19,21}. The \textit{Drosophila pAc-clk} construct was obtained
from F. Rouyer, \textit{pCopia-Renilla luciferase} and \textit{per3x69-luc} reporter constructs
from M. Rosbash, a [Ca\textsuperscript{2+}] independent mouse \textit{CaMKII} (T286D) was provided by
M. Mayford. The CaMKII inhibitor KN-93 was purchased from Abcam
(#ab120980).

\textit{C. marinus Cyc}, \textit{C. marinus Clk} and \textit{C. marinus CaMKII.1-RD} were cloned into the
\textit{pAc5.1/V5-His A} plasmid (Invitrogen) with stop codons before the tag. The \textit{Q5\textsuperscript{®}}
Site-Directed Mutagenesis Kit (NEB) was used to make kinase dead and [Ca\textsuperscript{2+}]
independent versions of \textit{C. marinus CaMKII.1-RD} (primers see Table S17).

\textit{Drosophila S2} cells (Invitrogen) were cultured at 25° C in Schneider's \textit{Drosophila}
medium (Lonza) supplemented with FBS (10\%, heat-inactivated, penicillin (100
U/ml), streptomycin (100 μg/ml) and 2 mM L-glutamine; Sigma). Cells were
seeded into 24 well plates (800,000 cells/well) and transfected with Effectene
transfection reagent (Qiagen) according to the manufacturer’s instructions.
Experiment with mouse [Ca^{2+}] independent CaMKII: 25ng pCopia-Renilla, 10ng per3x69-luc, 0.5ng Drosophila pAc-clk, 200ng mouse pAc-CaMKII-T286D.

Experiment with CaMKII inhibitor KN-93: 25ng pCopia-Renilla, 10ng per3x69-luc, Drosophila 0.5ng pAc-clk, various amounts of KN-93. Experiment with C. marinus genes: 25ng pCopia-Renilla, 10ng per3x69-luc, 100ng C. marinus pAc-cyc, 100ng C. marinus pAc-clk, 200ng C. marinus CaMKII.1-RD-K42R or 200ng C. marinus CaMKII.1-RD-T286D. In all experiments, the transfection mix was filled up to a total of 435ng DNA with empty pAc5.1/V5-His A vector per well. After 48 hours, cells were washed with PBS and lysed with Passive Lysis Buffer (Promega). Luciferase activities were determined on a Synergy H1 plate reader (Biotek) using a Dual-Luciferase Reporter Assay System (Promega). For each biological replicate three independent cell lysates were measured and their mean value determined. Firefly luciferase activity was normalized to Renilla luciferase activity and values were normalized to controls transfected with Drosophila pAc-clk or C. marinus pAc-clk and C. marinus pAc-cyc, respectively. S2 cells (Invitrogen/Life Technologies, Cat.no. R690-07) were regularly authenticated by morphology and routinely tested for absence of mycoplasma contamination (Lonza MycoAlert). Sample size was chosen to test for reproducibility.

Circadian free-run experiments

For circadian free-run experiments, culture boxes of the Por, He and Jean strains were transferred from standard LD (16:8) to constant dim light (LL; about 100 lux). Emerging adults were collected in 1-hour intervals by a custom made C. marinus fraction collector (similar to 75) and counted once a day. Because collection was automated, the experimenter had no influence on the results and blinding was not necessary. As the circalunar clock restricts adult emergence to
few days, the circadian emergence rhythm can only be assessed over few days.

Several culture boxes were transferred to LL at different time points. The
resulting emergence data were combined for each strain using the switch to LL
as a common reference point. We used the maximum number of available
individuals. Free-running period was calculated as the mean interval between
subsequent emergence peaks, weighing each peak by the number of individuals.
Extended Data Figure 1  The biology of *Clunio marinus*

(a) *C. marinus* is restricted to rocky shores (black lines), the localities differing in tidal regime (adapted from<sup>67</sup>). (b, c) Local strains show corresponding genetic adaptations in their circadian (<sup>b</sup><sup>67</sup>) and circalunar rhythms (c; He<sup>1</sup>, Jean<sup>5</sup>). Timing was measured in the laboratory under artificial moonlight (arrows in c) in a 30-day cycle and LD 12:12 (He, Por, Jean, Vigo) or 16:8 (Ber). Seasonal differences in daily illumination duration do not affect circadian emergence peaks<sup>1,76</sup>. Historically, for *C. marinus* “Zeitgeber time 0” is defined as the middle of dark phase.

Extended Data Figure 2  The reconstructed chromosomes of *C. marinus* based on the genetic linkage map

Left map: male informative markers. Right map: female informative markers. See Fig. 1a legend for further details.

Extended Data Figure 3  *C. marinus* genome characterization

(a) Representative genomic region with densely packed gene models (superscaffold1, from 535kb to 565kb). Gene models are given in blue on turquoise background. Gene predictions (SNAP) are purple. Transcript evidence is yellow. (b) Phylogenetic relationships of *C. marinus* to other Diptera (according to<sup>77</sup>). (c) Genetic diversity (θ; red) and linkage disequilibrium (r<sup>2</sup>; blue) of the Jean strain plotted for the three *C. marinus* linkage groups, revealing characteristic signatures of telomeres and centromeres. (d-f) Synteny comparisons among the genomes of *C. marinus*, *A. gambiae* and *D. melanogaster* based on 5,388 1:1:1 orthologs.

Extended Data Figure 4  Synteny analyses of *C. marinus* chromosome arms
Gene content of the *C. marinus* chromosome arms relative to the chromosome arms of *D. melanogaster* (black bars) and *A. gambiae* (grey bars). The very small chromosome 4 of *D. melanogaster* is neglected. Chromosome arms of *D. melanogaster* and *A. gambiae* are paired according to their published homology (Zdobnov et al. 2002). For four of the chromosome arms of *C. marinus* the homologous arms in *D. melanogaster* and *A. gambiae* are identified (grey shading). For comparison, the conservation of the identified *D. melanogaster* and *A. gambiae* homologs to each other is given by plotting the gene content of the homologous *D. melanogaster* chromosome arm relative to the different chromosome arms of *A. gambiae* (white bars). The numbers of orthologous genes considered in each comparison are given above the bars. For chromosome arm 2R of *C. marinus* the homologies are unclear. Possibly, chromosome arm 2R of *C. marinus* has undergone so many re-arrangements with other chromosome arms that it is no longer recognizable, which is consistent with complex polymorphic re-arrangements in this chromosome arm of *C. marinus* (see Supplementary Note 3). **(b)** Microsynteny is analyzed relative to *D. melanogaster* and *A. gambiae*, based on 5.388 1:1:1 orthologs. The fraction of genes in conserved microsynteny blocks is calculated and distributed along the phylogenetic tree. **(c, d)** A simulation was used to estimate how many chromosomal re-arrangements are required to produce the observed degree of microsyntenic conservation (for details see Supplementary Note 2).

**Extended Data Figure 5** Population genomic analysis of QTLs C1/L1 and C2 and genome-wide analysis of locations and putative effects of SNPs and indels

**(a, b)** Population genomic analysis of QTLs C1/L1 and C2.
Panels 1-3: *Por vs. Jean* strains (blue vs. red in panel 2,3). Panel 1: Genetic differentiation (red dots: SNPs with $F_{ST} \geq 0.8$; grey dots: $F_{ST} < 0.8$; back line: average $F_{ST}$ in 5-kb sliding windows). Panel 2: Genetic diversity (θ) in 20-kb (thin line) and 200-kb (thick line) windows. Panel 3: Linkage disequilibrium ($r^2$) for SNP pairs from 0-600 bp apart in 100-kb windows (step size: 5kb). Panel 4: Correlation Score (CS; 0 to 5) for genetic differentiation with circadian timing (top), circalunar timing (middle) and geographic distance (bottom) for five European *C. marinus* strains (*Vigo, Jean, Por, He, Ber*). Bottom numbers: scaffold IDs. See also Fig. 1. (c,d) Locations and putative effects of SNPs (c) and indels (d) with respect to the annotated gene models. The fractions of SNPs or indels in each category are compared for all SNPs and indels (black bars) vs. differentiated SNPs and indels ($FST \geq 0.8$ for *Por vs. Jean* strain; grey bars). Absolute numbers are given above the bars. In gene models with several splice forms, SNPs and indels can have different effects, e.g. “CDS: non-synonymous” for one splice form and “intronic” for another splice form. Therefore, the sum across locations is slightly larger than the actual numbers of SNPs and indels. “Codon changes” are all codon insertions or deletions that do not result in frame shifts beyond the actual insertion/deletion site. CDS = coding sequence; syn. = synonymous; non-syn. = non-synonymous; UTR = untranslated region.

Extended Data Figure 6 CaMKII regulates CLK/CYC transcriptional activity and exhibits strain specific splice variants

(a) Quantification of luciferase activity under the control of an artificial 3x69 E-box containing enhancer in S2 cells. Increasing amounts of the CaMKII inhibitor KN-93 decrease luciferase activity in a concentration-dependent manner, evidencing that endogenous CaMKII activity regulates the transcriptional activity
of the transfected CLOCK/CYCLE. (b) Without co-transfection of Drosophila clock, there is no detectable luciferase activity. The constitutively active form of CaMKII (mouse T286D) increases luciferase activity (normalised to CLOCK+; means; error bars: S.E.M.; biological replicates: n=4). (c) RNA sequencing reads mapped to the CaMKII.1 genomic locus. Arrows: major differences between the strains. (d) Relative expression levels of the four major CaMKII.1 transcripts (RA to RD) and the minor variant RO in the Por and Jean strains of C. marinus, as measured by qPCR (mean values; error bars: S.E.M.; two-sided Wilcoxon rank sum test; *** p<0.0005; * p<0.05; “ns” is not significant; Holm correction for multiple testing; biological replicates: Por n=9, Jean n=10; except for RO: Por n=3, Jean n=8). RO was not detectable in six additional biological replicates of the Por strain, suggesting that the expression differences are even greater than currently estimated. Fig. 3a shows the same data, normalized to the respective Por strain variants.

Extended Data Figure 7  A differentiated 125bp insertion in the CaMKII locus

(a) Alignment of the part of the CaMKII locus of the Por and Jean strains that carries a 125bp insertion in the Por strain. (b) Pool-Seq reads (>150x coverage) of this position for Por and Jean, as shown in the Integrated Genome Viewer (IGV). The reference does not have the 125bp insertion. At the position marked by the red box, the Jean strain has a 4bp polymorphic indel (ATAC; frequently misaligned due to a SNP 8bp downstream), whereas the Por strain has the 125bp insertion (but not the 4bp ATAC insertion). In Jean basically all reads span the indel, suggesting that if the 125bp insertion is present in Jean at all, its frequency
is very low. In contrast, in Por all reads but one end at this position, suggesting
the frequency of the 125bp insertion in Por is 154 of 155 reads or >0.99.

Extended Data Figure 8 Model of circadian timing adaptation via
sequence differences in the CaMKII.1 locus
Exon coloration as in Figure 4b. The arrows with question marks indicate
possible pathways that alone or in combination could mediate the effect of
CaMKII.1 on timing. Dotted lines: indirect effects.

Extended Data Figure 9 Analyses overview
(a) Overview of the genome assembly process. (b) Overview of the population
genomic analyses.

Extended Data Figure 10 Arrangement of the mitochondrial genome (a)
and of the histone gene cluster (b) of C. marinus.
Protein coding genes are given in yellow, tRNAs and rRNAs in red.
References for Online Methods and Extended Data Figures


Crawley, M. J. *The R Book*. (John Wiley & Sons Ltd., 2007).


**kinase domain and regulatory region**

<table>
<thead>
<tr>
<th>Compound</th>
<th>0</th>
<th>100</th>
<th>100</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cma-Clk (ng)</td>
<td>-</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Cma-Cyc (ng)</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Cma-CaMKII.1 K42R (ng)</td>
<td>-</td>
<td>-</td>
<td>200</td>
<td>-</td>
</tr>
<tr>
<td>Cma-CaMKII.1 T286D (ng)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>200</td>
</tr>
</tbody>
</table>

**Cma-Clk (ng)**

**Cma-Cyc (ng)**

**Cma-CaMKII.1 K42R (ng)**

**Cma-CaMKII.1 T286D (ng)**

---

**association domain**

<table>
<thead>
<tr>
<th>Compound</th>
<th>RA</th>
<th>RB</th>
<th>RC</th>
<th>RD</th>
<th>RE</th>
<th>RF</th>
<th>RG</th>
<th>RH</th>
<th>RI</th>
<th>RJ</th>
<th>RK</th>
<th>RL</th>
<th>RM</th>
<th>RN</th>
<th>RO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cma-Clk (ng)</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cma-Cyc (ng)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cma-CaMKII.1 K42R (ng)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cma-CaMKII.1 T286D (ng)</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**FST ≥ 0.8 SNPs**

Indels

**position (bp)**

340000 345000 350000
Splice variant fold change

<table>
<thead>
<tr>
<th></th>
<th>RA</th>
<th>RB</th>
<th>RC</th>
<th>RD</th>
<th>RO</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. marinus qPCR</strong></td>
<td>***</td>
<td>***</td>
<td>ns</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Splicing in S2R+ cells

<table>
<thead>
<tr>
<th></th>
<th>RA</th>
<th>RB</th>
<th>RC</th>
<th>RD</th>
<th>RO</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Por</strong></td>
<td>*</td>
<td>**</td>
<td>*</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td><strong>Jean</strong></td>
<td>not detected</td>
<td>not detected</td>
<td>not detected</td>
<td>not detected</td>
<td>not detected</td>
</tr>
</tbody>
</table>

Time (h) number of emerged midges

<table>
<thead>
<tr>
<th></th>
<th>He</th>
<th>Por</th>
<th>Jean</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>18</td>
<td>24</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>12</td>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>20</td>
<td>12</td>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\( \tau = 24.2 \) n = 155

\( \tau = 24.83 \) n = 103

\( \tau = 25.61 \) n = 251