Genes Affecting Circadian Period in *Drosophila melanogaster* are Correlated with Latitude

Leon Zong, Emma Jia, Kunal Deb, Rohit Marol, Timothy Lai

Abstract

We selected single-nucleotide point mutations (SNPs) with a significant effect on circadian period based on prior research. We analyzed the genomes of 240 populations in Europe and North America for the frequency of the identified mutations. Using principal component analysis (PCA) and fixation index (FST), we develop a model to analyze the correlation of the mutations with latitude with minimal confounding factors. We discover a relationship in 47 SNPs, indicating potential genes selected for by latitude.

1 Introduction

The tilt of the Earth's axis results in differing variance in the day/night cycle across different latitudes. For an extreme example, winters in the polar circle are shrouded in constant darkness and summers enshrined in constant daylight, but winters and summers near the equator are a nearly consistent 12 hours of daylight throughout the year. Our study aims to report on how *Drosophila melanogaster* has adapted to this variance.

The circadian rhythm plays a key role in regulating locomotor behavior, the sleep-/wake cycle, as well as some development. Understanding how the circadian rhythm is itself regulated is necessary for understanding the way organisms have evolved and adapted to changing seasonal patterns. The *D. melanogaster* (Dmel) circadian rhythm pathway is extensively well-studied. Our molecular knowledge of the general circadian rhythm pathway is heavily based on findings in Dmel (Huang, 2018). Findings from our analysis on how Dmel flies may have evolved adaptations to varying day/night cycle lengths (Δdaylight) can provide a starting point for analyzing circadian rhythm anomalies in humans. One such anomaly is seasonal affective disorder (SAD), which is notable for being seasonal and directly correlated with shorter daylight periods (Magnusson and Boivin, 2003).

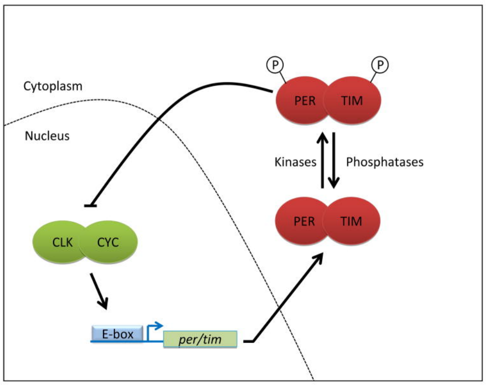


Fig. 1: Core Regulation Loop. The PER/TIM dimer accumulates during the night, until they are eventually phosphorylated during the day and degraded, releasing the CLK/CYC dimer to once again begin promoting PER/TIM transcription. (Tataroglu and Emery, 2014)

The key transcription factors involved in circadian rhythm regulation are Clock (CLK) and Cycle (CYC). These transcription factors promote the transcription of other protein factors – Period (PER) and Timeless (TIM). These factors promote the deactivation of CLK/CYC, forming a cyclical control loop, as shown in Figure 1 (Tataroglu and Emery, 2014).

Light entrainment

There are, of course, environmental inputs to this pathway. One of the strongest interactions is with blue light, the primary source of which is from sunlight (Wong and Bahmani, 2022). Blue light activates the Cryptochrome (CRY) protein upon photon absorption, which binds to Tim. This act promotes the degradation of TIM, which "resets" the clock (Zeng et al, 1996). Overall, it has been shown that the synchronization (sometimes called entrainment) of the Dmel circadian clock occurs as a highly-sensitive response to light (Dubruille and Emery, 2008).

Additional studies have begun to examine the exact mechanism of CRY-TIM degradation. The protein JETLAG (JET) performs a similar function to CRY, but JET mutants require an additional mutation in Tim for arhythmicity in low light conditions (Koh et al, 2006). That mutation – resulting in a variant called L-TIM – interacts less efficiently with CRY (Lin et al, 2001). Other kinases, like SGG, have been shown to bind to Cry and inhibit its function (Kumar et al, 2021).

Since the Dmel circadian clock is sensitive to light, we hypothesize that differences in changes in light length (Δdaylight) can be a selecting factor for mutations in genes in this pathway. In other words, Δdaylight affects the synchronization of the circadian clock, and we expect that there will be genetic adaptation of this synchronization to various Δdaylights (various latitudes).

Circadian effects

The circadian clock affects a large range of downstream pathology. Yoshii et al (2004) find a set of genes correlating with extended circadian periods. In addition, Dubowy and Sehgal (2017) find a mechanism in sleep-related genes which correlates with longer sleep periods. These effects of circadian rhythm are crucial, as we expect that larger variance in effective daylight will select for beneficial variations in these core pathologies.

2 Methods

Identification of circadian-related alleles

Our initial analysis began from a set of mutations identified to have a significant effect on circadian period. To accomplish this, we examined genome-wide association studies (GWASs) relating to various circadian factors. Harbison et al (2019) identified single nucleotide polymorphisms (SNPs) significantly affecting circadian period and rhythmicity in the *Drosophila* Genetic Reference Panel (DGRP) (Mackay et al, 2012). Kumar et al (2021) observed a particular population – DGRP_892 – with a significantly long period length (31 hours). They identified mutations in this population which had significant effect on period.

From these GWAS, we created a list of candidate mutations. Each candidate mutation entry originally contained the Flybase ID, gene name, as well as identified effect sizes (for each sex, sex-averaged and sex-difference). We used Python to query the public FlyBase API endpoint (v1.0) to acquire a summary paragraph for each gene (Gramates et al, 2022). We parsed this summary to extract biological and molecular function.

DEST data

The Drosophila Evolution over Space and Time (DEST) dataset is the current largest collection of pooled genomics data for Dmel (Kapun et al, 2021). DEST contains the union of genomics data for 102 populations across countries, continents, and across 7 years of collection. The total number of pooled population samples is 246. The full range of the DEST dataset is displayed in Figure 2.

The entire PoolSeq DEST dataset was downloaded onto the University of Rochester BlueHive compute cluster in compressed BCF format. It was then sorted and indexed using the `bcftools` library. No particular filtering was done upon initial import, as the DEST dataset has already been minimally filtered for minimum read depth and repetitive elements.

Variant extraction

We used a Python script to filter our GWAS-identified candidate mutations, all of which are single-nucleotide polymorphisms (SNPs), for only those which exist in the DEST dataset. Filtering eliminated 455 of 585 total SNPs (77.8%). After filtering, we used another Python script to extract the key information for each SNP, including

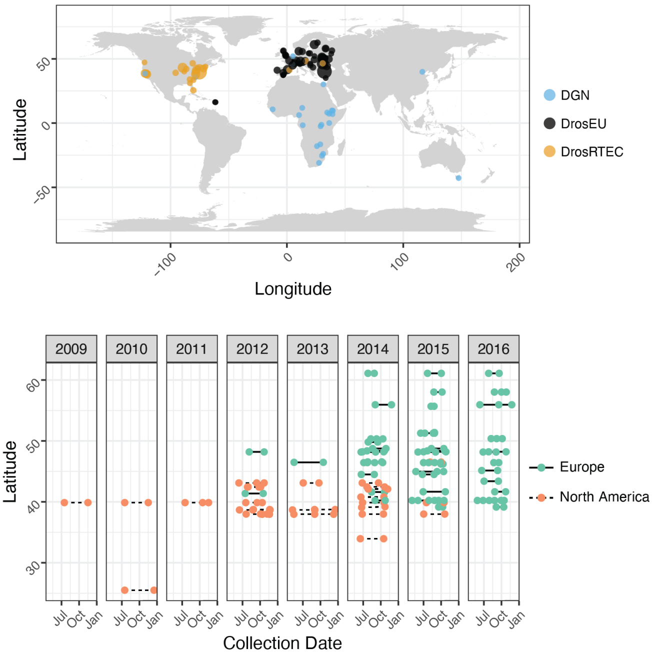


Fig. 2: DEST Dataset (Kapun et al, 2021)

position, reference allele, alternate allele, frequency, and raw read counts for each population. This gave us a dataset of over 30,000 individual frequency entries.

Environmental variables

For each population, we used the DEST metadata file to match population IDs (from the VCF data) to population metadata. While a large range of biological and climatic variables are provided in DEST for each population, we are only interested in latitude as our key environmental variable.

Principal component analysis (PCA)

PCA is a widely used dimensionality reduction tool in data science. In genomics, PCA is used to "identify structure in the distribution of genetic variation across geographical location and ethnic background" (McVean, 2009). In short, PCA reduces all of the variability across the whole genome to a few key components. PCA relies upon the assumption that features in the original data are independent.

To perform PCA, we used the powerful and widely-used PLINK 2.0 genomics library (Chang et al, 2015). First, we convert the entire DEST dataset into the PLINK binary format. To account for genetic linkage introducing dependencies in close SNPs, we use the '`--indep-pairwise`' filter with a window size of 50Kb, window step of 10 bases, and an R^2 threshold of 0.1 based on modifications to the recommended values for the human genome in the PLINK documentation.

Finally, we perform PCA using PLINK, extracting 10 principal components (PCs) based on the default value of PLINK. PLINK documentation states “in practice, 10 PCs has been effective across a wide range of studies.” We exclude the X chromosome as standard practice.

Statistical analysis

We built a linear model incorporating the results of PCA and the raw environmental metadata. For each analyzed SNP, we calculated a regression using the latitude and PCA features as the independent variable, and the frequency of the SNP as the dependent variable. To ensure no correlation between our input variables, we ran separate ordinary least-squares (OLS) regressions for latitude versus each of the PCs and found no significant correlation.

Once the regressions for each SNP were calculated, we used a two-tailed Student’s t-test to investigate statistical significance in the coefficient for latitude.

FST

As a separate analysis, we performed F_{ST} analysis on the data using the R package `poolfstat` (V. et al, 2018; M. et al, 2022). We first subset the regions of the genes of interest using `bcftools`. We take the pairwise F_{ST} for the union of these regions, and divide by the continental background for each value to derive a ratio of the level of F_{ST} at the GWAS-identified genes versus background. We generated a heatmap sorted by latitude using R, visualizing only values above the 95% quantile.

Gene interaction

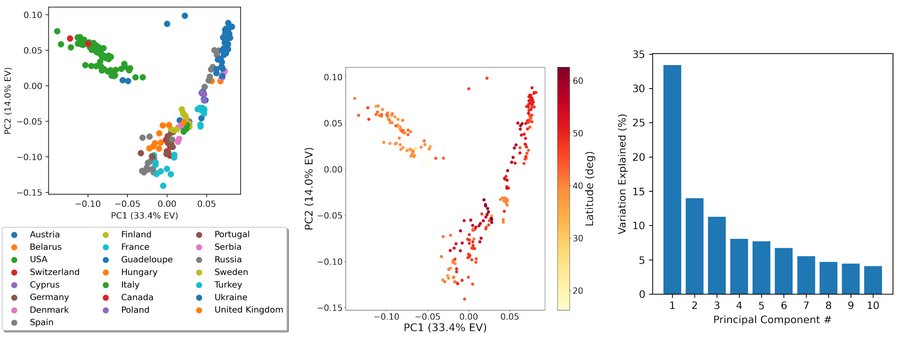
To visualize gene interactions, we used the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (Szklarczyk et al, 2022). We include 50 STRING-identified additional interactors, and limit interactions to those with medium confidence (> 0.4). We only keep genes with at least one interaction in the visualization. Nodes were colored manually.

3 Results

Genes associated with circadian period are correlated with latitude

Regression models are a simple yet powerful tool to study correlations in potentially correlated variables. We applied a combination of techniques to analyze the hypothesis that latitude and the corresponding Δdaylight are correlated with circadian period mutations on a dataset spanning 246 Dmel populations. We performed multiple linear regressions, one for each mutation, to determine correlation.

Since genetic diversity in Dmel is largely affected by population effects, like population structure, a simple analysis of only latitude versus frequency of mutation was likely to have significant omitted variable bias.



(a) (b) (c)

Fig. 3: Result of PCA. (a) demonstrates effectiveness of PCs in capturing population structure effect. (b) demonstrates an exemplar SNP (in CG10089) with high correlation with the PCs but a low correlation with latitude. (c) shows the percentage of genome-wide variance explained by each PC.

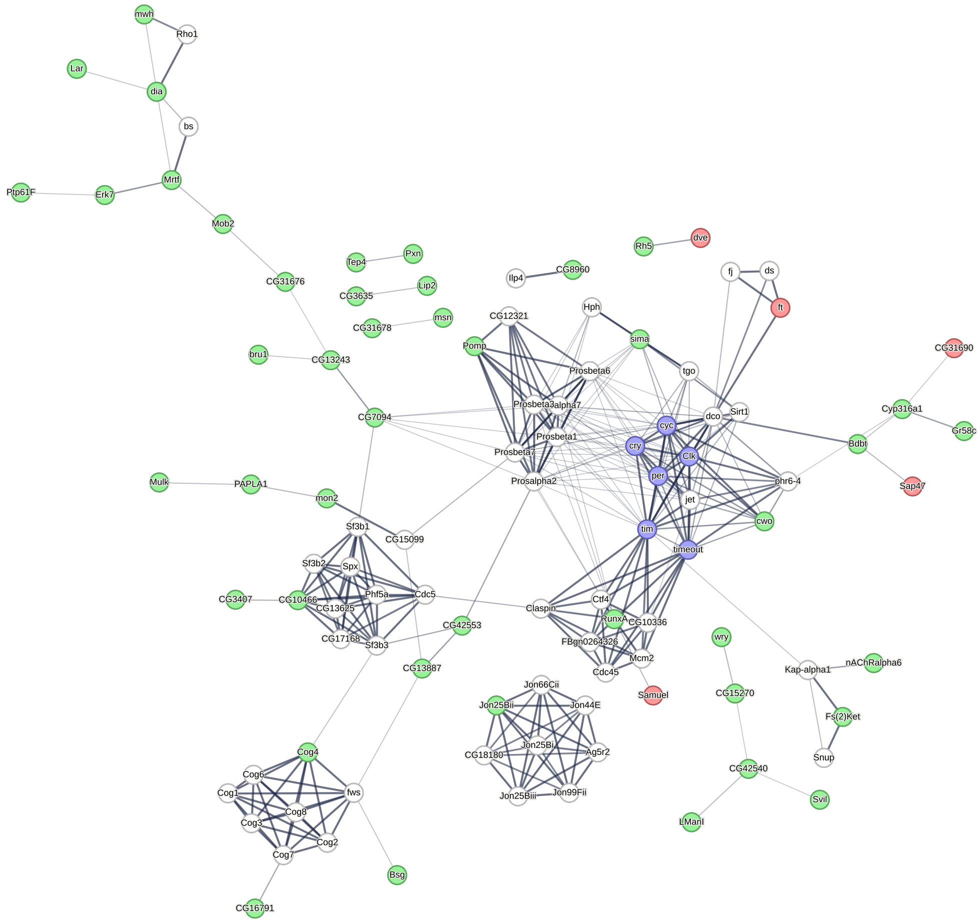


Fig. 4: Gene interaction network. Red genes are significant with respect to latitude. Green genes are non-significant with respect to latitude but are found in the GWASs. Blue genes are the core circadian regulators CLK, CYC, TIM, PER, CRY, and DBL. White genes are co-interactors generated by STRING.

For example, populations on the same continent may have similar mutation frequencies due to shared history, even if their latitudes are different. In the opposite example, populations on different continents may have significantly different mutation frequencies, even if their latitudes are the same.

We corrected for this factor by incorporating the PCA components as variables in our regression models. The PCs represent structural variation across the *whole genome* (Figure 3a, 3b). Including the PCs in the model allows it to assign variability in frequency to the PCs rather than latitude. If the coefficient for latitude becomes non-significant upon addition of the PCs to the model, no correlation between latitude and frequency can be detected.

Our initial analysis, not including PCs, found a total of 78 significantly correlated mutations. Only 47 mutations remained significantly correlated with latitude after including PCs in the regression (Figure 5). The coefficient of latitude in the significant mutations is small (all are ≤ 0.015) yet still significant.

For each mutation, we were interested in the pathway of the corresponding gene, and whether it was involved downstream of the circadian regulators. Viewing interactions of both significant mutations with respect to latitude and non-significant mutations with core circadian regulators (Figure 4), we notice an interesting relationship. Several of the significantly correlated genes are found to interact with the master circadian regulators, indicating that mutations in their interaction neighborhood are correlated with latitude. In general, we see that the genes associated with circadian period differences (in green) have interactions with the core circadian regulators.

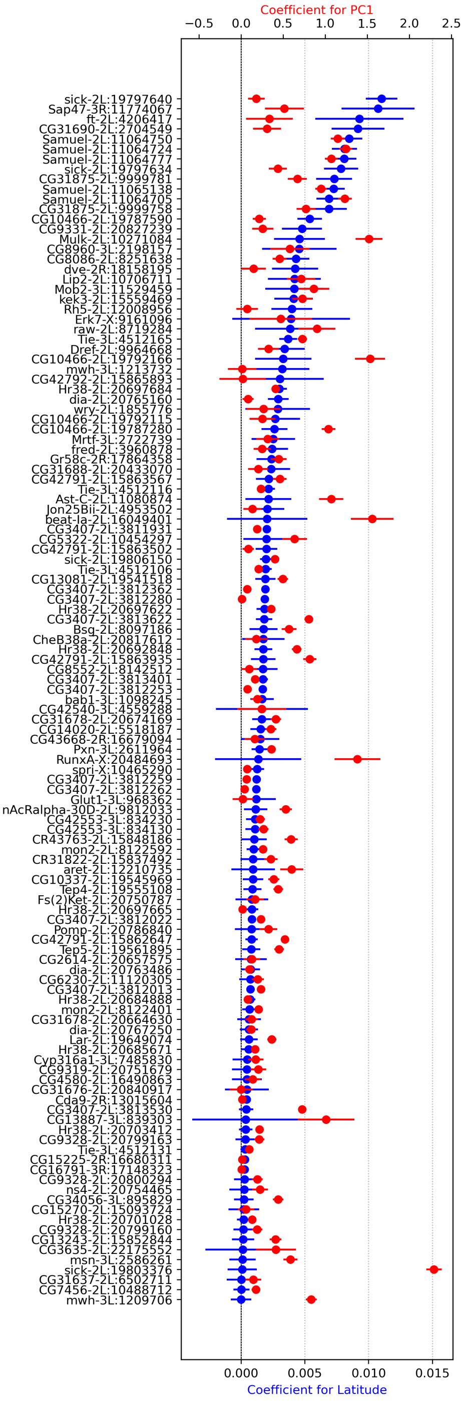


Fig. 5: SNPs, by gene and position, and their coefficient for latitude and PC1

Genes associated with circadian period show a weak signature of selection

F_{ST}, or fixation index, is a population genetics statistic often used to detect selection in a particular gene (Porto-Neto et al, 2013). High values of F_{ST} indicate a high degree of differentiation. While high differentiation does not necessarily mean that selection is occurring, it is a potential indicator.

We calculated the ratio of F_{ST} in our genes of interest to the background genomic F_{ST} for each pair of populations, which provides a reference value of differentiation versus background (Figure 6). Heat concentrated in the bottom right corner indicates higher F_{ST} values for populations at more distant latitudes. As the correlation is not strong, we infer that the genes are only potentially under higher selection.

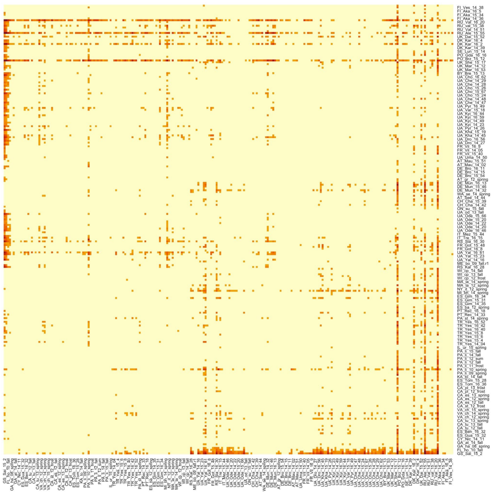


Fig. 6: Heat map of filtered 95% quantile F_{ST} Values in all populations, sorted by latitude. Higher latitudes are towards the top and towards the right.

4 Discussion

Overall, we find that a large percentage of genes determined via GWAS to have significant effects on circadian period and rhythmicity are correlated with latitude, even after accounting for population structure. We find a weak signature of selection via F_{ST} analysis of the genes of interest, noting that populations at farther latitudes are more differentiated compared to background. While we cannot outright say that these mutations are being *selected* for by latitude, the correlation and selection signature are still interesting.

In the field of circadian rhythm research, a molecular-level analysis can be performed based on the mutations detected to be significantly correlated with latitude. This might reveal the underlying mechanism which provide the beneficial phenotype being selected for. Should a mechanism be discovered, further research can examine homologous pathways in humans

Our study did have some limitations. Primarily, the underlying environmental variable of interest – Δdaylight – and potential confounders – Δtemperature – should be included in the model. These were left out in our study due to time constraints, and latitude was chosen as a directly correlated variable which would be easier to analyze. Including them would allow a direct correlation between the potential selecting variable and the selected-for mutation to be calculated.

Additionally, more robust statistical analyses can be performed. Notably, similar research by Forni et al (2014) uses non-parametric statistical methods like Kendall's rank correlation to avoid underlying assumptions about the distribution of the mutations. They also use correction factors like Bonferroni correction to test at multiple confidence levels.

Despite the limitations, the analysis of mutation frequency across a huge number of populations, alongside the population structure analyses, provide a useful model setup for analyzing any phenotypic effect across a wide variety of environmental conditions. By conditioning the model on the innate population structure, we take steps to isolate the effect of the environmental variable itself. We believe that our analysis is sound enough to detect the unbiased correlations between latitude and mutation frequency. Should that be true, we can conclude that the circadian rhythm pathway has adapted to environmental factors at different latitudes. We expect that **Δdaylight** is the environmental factor which selects for the circadian period associated mutations analyzed in this study.

Declarations

Availability of data, materials, and code

The DEST dataset is available online (Kapun et al, 2021). Code is available in the lab notebook or upon request.

Authors' contributions

LZ, EJ, KD, RM, and TL planned analysis and performed initial research. LZ and TL performed initial data extraction and regressions. LZ performed PCA and interaction query. EJ performed F_{ST} analysis. KD and RM analyzed results for specific genes of interest. KD kept materials organized. LZ wrote this paper. Notes on week-by-week contributions can be found in the lab notebook.

References

Chang CC, Chow CC, Tellier LC, et al (2015) Second-generation plink: rising to the challenge of larger and richer datasets. Gigascience 4:7. https://doi.org/10.1186/s13742-015-0047-8

Dubowy C, Sehgal A (2017) Circadian rhythms and sleep in drosophila melanogaster. Genetics 205(4):1373–1397. https://doi.org/10.1534/genetics.115.185157

Dubruille R, Emery P (2008) A plastic clock: How circadian rhythms respond to environmental cues in drosophila. Molecular Neurobiology 38(2):129–145. https://doi.org/10.1007/s12035-008-8035-y, URL https://doi.org/10.1007/s12035-008-8035-y

Forni D, Pozzoli U, Cagliani R, et al (2014) Genetic adaptation of the human circadian clock to day-length latitudinal variations and relevance for affective disorders. Genome Biology 15(10):499. https://doi.org/10.1186/s13059-014-0499-7, URL https://doi.org/10.1186/s13059-014-0499-7

Gramates LS, Agapite J, Attrill H, et al (2022) FlyBase: a guided tour of highlighted features. Genetics 220(4):iyac035. https://doi.org/10.1093/genetics/iyac035, URL https://doi.org/10.1093/genetics/iyac035, https://academic.oup.com/genetics/article-pdf/220/4/iyac035/43709097/iyac035.pdf

Harbison ST, Kumar S, Huang W, et al (2019) Genome-wide association study of circadian behavior in drosophila melanogaster. Behav Genet 49(1):60–82. https://doi.org/10.1007/s10519-018-9932-0

Huang RC (2018) The discoveries of molecular mechanisms for the circadian rhythm: The 2017 nobel prize in physiology or medicine. Biomed J 41(1):5–8. https://doi.org/10.1016/j.bj.2018.02.003

Kapun M, Nunez JCB, Bogaerts-Márquez M, et al (2021) Drosophila evolution over space and time (dest): A new population genomics resource. Molecular Biology and Evolution 38(12):5782–5805. https://doi.org/10.1093/molbev/msab259, URL https://doi.org/10.1093/molbev/msab259

Koh K, Zheng X, Sehgal A (2006) Jetlag resets the drosophila circadian clock by promoting light-induced degradation of timeless. Science 312(5781):1809–1812. https://doi.org/10.1126/science.1124951

Kumar S, Tunc I, Tansey TR, et al (2021) Identification of genes contributing to a long circadian period in drosophila melanogaster. J Biol Rhythms 36(3):239–253. https://doi.org/10.1177/0748730420975946

Lin FJ, Song W, Meyer-Bernstein E, et al (2001) Photic signaling by cryptochrome in the drosophila circadian system. Mol Cell Biol 21(21):7287–7294. https://doi.org/10.1128/MCB.21.21.7287-7294.2001

M. G, R. V, L. F, et al (2022) f-statistics estimation and admixture graph construction with pool-seq or allele count data using the r package poolfstat. Molecular Ecology Resources 22(4):1394–1416

Mackay TFC, Richards S, Stone EA, et al (2012) The drosophila melanogaster genetic reference panel. Nature 482(7384):173–178. https://doi.org/10.1038/nature10811, URL https://doi.org/10.1038/nature10811

Magnusson A, Boivin D (2003) Seasonal affective disorder: an overview. Chronobiology international 20(2):189–207

McVean G (2009) A genealogical interpretation of principal components analysis. PLoS Genet 5(10):e1000686. https://doi.org/10.1371/journal.pgen.1000686

Porto-Neto LR, Lee SH, Lee HK, et al (2013) Detection of signatures of selection using fst. Methods Mol Biol 1019:423–436. https://doi.org/10.1007/978-1-62703-447-0{_}19

Szklarczyk D, Kirsch R, Koutrouli M, et al (2022) The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research 51(D1):D638–D646. https://doi.org/10.1093/nar/gkac1000, URL https://doi.org/10.1093/nar/gkac1000, https://academic.oup.com/nar/article-pdf/51/D1/D638/48440966/gkac1000.pdf

Tataroglu O, Emery P (2014) Studying circadian rhythms in drosophila melanogaster. Methods 68(1):140–150. https://doi.org/10.1016/j.ymeth.2014.01.001

V. H, R. L, E.J. P, et al (2018) Measuring genetic differentiation from pool-seq data. Genetics 210(1):315–330

Wong NA, Bahmani H (2022) A review of the current state of research on artificial blue light safety as it applies to digital devices. Heliyon 8(8):e10282. https://doi.org/10.1016/j.heliyon.2022.e10282

Yoshii T, Funada Y, Ibuki-Ishibashi T, et al (2004) Drosophila cryb mutation reveals two circadian clocks that drive locomotor rhythm and have different responsiveness to light. J Insect Physiol 50(6):479–488. https://doi.org/10.1016/j.jinsphys.2004.02.011

Zeng H, Qian Z, Myers MP, et al (1996) A light-entrainment mechanism for the drosophila circadian clock. Nature 380(6570):129–135. https://doi.org/10.1038/380129a0