PNAS
www.pnas.org

Supplementary Information for

Population-level rhythms in human skin with implications for circadian medicine

Gang Wu[a], Marc D. Ruben[a], Robert E. Schmidt[a], Lauren J. Francey[a], David F. Smith[b,c], Ron C. Anafi[d], Jacob J. Hughey[e], Ryan Tasseff[f], Joseph D. Sherrill[f], John E. Oblong[f], Kevin J. Mills[f], John B. Hogenesch[a,1]

John B. Hogenesch
Email: john.hogenesch@cchmc.org

This PDF file includes:

Supplementary text
Figs. S1 to S10
Tables S1
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Datasets S1to S2

www.pnas.org/cgi/doi/10.1073/pnas.1809442115

SI Methods

Subject inclusion criteria

Those subjects recruited for collecting ordered skin samples are free of skin disease, and with a normal sleeping pattern (6–8 h/night). All subjects were provided with regularly scheduled meals without specific dietary restrictions and were not allowed to perform any exhaustive physical activities. No food was served after 10 PM. Subjects with systemic diseases (e.g., diabetes) or taking a medication (e.g., hormone replacement therapy, anti-inflammatory agents, immunosuppressants) that could alter clock function were not included in collecting unordered samples.

Salivary melatonin and cortisol

Melatonin (N-acetyl-5-hydroxytryptamine) is a hormone produced by the pineal gland that helps orchestrate the sleep-wake cycle. Light input to the retina inhibits melatonin synthesis while darkness promotes it, therefore, melatonin production has a pronounced circadian rhythm. In general, in people with normal sleep patterns, melatonin secretion begins to increase in the evening and reach a peak during hours that most people tend to sleep, which is the pattern we observed in this study (Fig. S8).

Cortisol, also known as hydrocortisone, is a glucocorticoid that is produced in the adrenal cortex and is released in response to stress, as well as, a low level of circulating glucocorticoids. Its primary functions are to increase blood sugar through gluconeogenesis; suppress the immune system; and aid in fat, protein and carbohydrate metabolism. In most healthy individuals, cortisol shows a pronounced circadian rhythm, with circulating levels beginning to increase in the morning and then tapering off later in the day (Fig. S9). However, there are many stressors and disease states that can cause chronically elevated cortisol levels and this can have detrimental effects on a number of tissues, including the skin. In this cohort we identified a cortisol rhythm that aligns with the published literature, albeit with large interindividual variability. The rhythms of melatonin and cortisol add a robust, internal control to this study allowing for meaningful detection of circadian transcriptional rhythms.

LCM protocol

Samples were frozen in optimum cutting temperature compound (OTC) (Sakura Finetek, Torrance, CA). Embedded tissue blocks were cut into 14 μm sections and mounted on polyethylene naphthalate slides (Life Technologies, Grand Island, NY) and kept at -80°C until

use. Sections were immediately fixed in 95% ethanol for 1 min then rinsed in deionized water. After water rinse, the sections were stained (30 seconds) with cresyl violet (Sigma Aldrich, St. Louis, MO) and eosin Y (Sigma Aldrich, St. Louis, MO) and subsequently rinsed with deionized water. The sections were then dehydrated in 95% and 100% ethanol, cleared in xylene and allowed to air dry at room temperature. After drying, epidermal skin cells were isolated from the sections by LCM with the PALM Microbeam system (Carl Zeiss MicroImaging, Munchen, Germany) for transcriptome analysis.

mRNA target labeling and processing steps

Total RNA from laser capture microdissection samples was isolated using the Pico Pure RNA Isolation Kit (Life Technologies) according to the manufacturer's recommendations. The quality and concentration of the isolated RNA were analyzer using a 2100 Bioanalyzer and RNA 6000 Pico Kit (Agilent Technologies) according to the manufacturer's recommendations. 25 ng of total RNA was reverse-transcribed into cDNA copies using oligo-dT primers and reverse transcriptase followed by second strand synthesis using DNA polymerase I. In vitro transcription synthesizes cRNA that is biotin incorporated and purified using the Affymetrix HT 3' IVT Express kit (Cat. #901253), as executed on a Beckman Biomek FXp Laboratory Automation Workstation (Beckman Cat. #A31842). Biotinylated cRNA was fragmented by limited alkaline hydrolysis and then hybridized overnight to Affymetrix GeneTitan U219 array plates using the Affymetrix GeneTitan instrument.

Extracting expression profiles and running MetaCycle on time series datasets

We used the RMA algorithm in the affy R package (1) to extract expression profiles from the raw CEL files of time-series human epidermis samples. Raw CEL files of mouse anagen and telogen skin were downloaded from the GEO database. The RMA algorithm in the oligo R package (2) was applied to each tissue separately to get the expression profile. The RMA normalized expression profiles of 12 mouse tissues (3) were downloaded from the GEO database. The time-series human epidermis samples were analyzed with meta3d using the default settings in addition to cycMethodOne = "ARS" and period length being set to 24 for both 'minper' and 'maxper'. MetaCycle's meta3d function uses ARSER to analyze the time-series data individual by individual, then uses Fisher's method to integrate the *P* values from 19 subjects and gets the average value of other parameters (e.g. phase) from 19 subjects. Transcripts with meta3d_Base (mean baseline expression of the 19 time-series profiles) less than 16 were filtered out. The probe sets were annotated with gene symbols and one representative transcript was selected for each gene based on the following criteria: the lowest *P* value or the largest rAMP (multiple transcripts

have equal *P* value).

For the mouse skin (4) and 12 other tissues (3), MetaCycle::meta2d was used to detect circadian transcripts with the default settings and period length set to 24 for both 'minper' and 'maxper'. The cycMethod = c ("JTK", "LS") was used for analyzing those 12 mouse tissues. Annotation files for the human and mouse array were downloaded from Affymetrix (version date Mar 2016 and October 2014 for human and mouse respectively; http://www.affymetrix.com/support/technical/byproduct.affx?product). The probe sets were annotated with gene symbols, and one representative probe set was selected for each gene with the similar procedure as mentioned above, with the meta2d_Base (baseline expression of time-series profiles) cut-off at 64 for mouse telogen and anagen, and 63.1 for other 12 mouse tissues. The circadian genes identified in mouse anagen and telogen skin were linked to human homolog genes using the NCBI homologene file (Build 68).

Batch effect adjustment

All 298 human epidermis samples listed in Table S1 were combined for the hybrid design. The RMA algorithm was used to extract expression profiles of raw CEL files from these 298 human epidermis samples. PCA analysis using the ggfortify package (5) detected a strong batch effect. To correct for these batch effects, we applied the ComBat function in the sva package (6, 7). The ComBat adjusted expression profiles were annotated with gene symbols. Probe sets with no annotated gene symbols or stable expression in all samples were filtered out. The remaining probe sets were grouped by gene symbols, and one probe set with the maximum median absolute deviation value was selected for each gene.

Testing the revised CYCLOPS pipeline with mouse time-series skin samples

To improve the ordering of human samples with CYCLOPS (8), we added three new steps to the pipeline: i) assess the robustness of the clock in the data set using an expression correlation matrix (ECM) of clock and clock-associated genes (CCGs), ii) order samples with selected eigengenes from Oscope (9), and iii) validate the best ordering result using samples with time information. The ECM of CCGs in mouse anagen and telogen (Fig. S10 A and C) are similar with the ECM obtained from other mouse tissues (Fig. 4A) indicating that there is a functional clock in these samples. If the dataset is noisy and/or there is weak circadian clock structure (for example human blood samples; Fig. 4A and Fig. S6), the ECM will correlate poorly with benchmark mouse ECM.

To sharpen CYCLOPS ordering, we selected a set of seed circadian genes that are conserved between human and mouse skin tissues, or robustly cycling in two mouse skin tissues. To

increase the weight of seed genes from human epidermal tissue, we set a less strict cut-off ($P <$ 0.1) which identified 233 circadian genes (Fig. S1A). A total of 120 genes (Dataset S2) met our criteria for seed genes. The expression profiles of these seed genes were inputs to CYCLOPS to obtain a set of eigengenes.

Oscope was then used to select the best eigengenes using the pair-sine model, referred to as the eigengene cluster. In addition to eigengene clusters automatically selected by Oscope, we also included eigengene pairs with dissimilarity values ≤ the 20th percentile of the dissimilarity matrix given by the OscopeSine function. Only the first two eigengenes extracted from mouse anagen and telogen data sets show periodic expression patterns (Fig. S10 B and D), which were also selected by Oscope. Using the first two eigengenes, CYCLOPS gave the order for mouse anagen and telogen samples. Lastly, we verified a linear relationship between CYCLOPS-sample order and actual sampling time (Fig. S10E), which indicates the CYCLOPS order is correct.

Ordering human epidermis samples with the revised CYCLOPS pipeline

The revised CYCLOPS pipeline was used to order all 298 human epidermis samples using seed genes (Dataset S2) mentioned above. CYCLOPS produced sample orderings based on the relationship between multiple eigengenes. Default settings were used to run both CYCLOPS and Oscope, except 'Frac_Var', 'DFrac_Var' and 'Seed_MinMean' were set at 0.99, 0.01 and 16, respectively for CYCLOPS, and 'quan' and 'maxK' were set at 0.5 and 10 for OscopeKM function. In addition to the default output parameters, the false discovery rate (FDR) was calculated using the p.adjust function in R (3.3.2), and the relative amplitude (rAMP) was calculated by dividing the amplitude (amp) by fitted expression baseline (fitmean).

Evaluation of CYCLOPS output

A total of 10 eigengene clusters were selected by Oscope and were used for CYCLOPS ordering. From these results, we selected the best sample phase ordering based on two evaluation criteria. First, the sample phase given by CYCLOPS was compared to sampling time, using samples with time information, as mentioned above. The sample phase orders that reflected external sampling time were considered valid. Second, we required that CCGs were detected as rhythmic, with phase relationships highly correlated to benchmark mouse CCGs. The best ordering was selected by visual inspection and Fisher's circular correlation coefficient.

Comparison of circadian genes identified from human epidermis and mouse telogen

188 circadian genes identified from modified cosinor regression analysis on the best CYCLOPS sample ordering were separated into five groups based on the following comparison with

circadian genes identified from mouse telogen: i) The NoProbe group includes 15 genes without mouse homolog genes or with mouse homolog genes not represented on the mouse affymetrix array platform. Based on the rAMP and *P* value from the MetaCycle analysis of mouse telogen, the remaining 173 circadian genes were separated into additional other four groups: ii) rAMP ≤ 0.1 and *P* value ≥ 0.05, iii) rAMP > 0.1 and *P* value ≥ 0.05, iv) rAMP ≤ 0.1 and *P* value < 0.05, and v) rAMP > 0.1 and *P* value < 0.05. The 46 circadian genes with mouse homolog genes and robustly cycle in mouse telogen (rAMP > 0.1, *P* value < 0.05) were selected for phase comparison. Circadian gene phases were adjusted to *ARNTL*/*Arntl* phase and converted to a positive value between 0 and 2π.

Phase set enrichment analysis of circadian genes identified from mammalian skin

Before running PSEA (10) on circadian genes from mouse telogen ($P < 0.05$), mouse gene symbols were aligned to homologous human genes, and MetaCycle gene phases were adjusted with *ARNTL*/*Arntl* phase and converted to a positive number between 0 and 2π. The gene set file (c2.cp.v5.0.symbols.gmt) was downloaded from MSigDB. The following parameters were used to run PSEA on circadian genes identified from CYCLOPS (230 genes at $P < 0.05$ were used for PSEA analysis; Fig. S4): Min items/set is 5, Max sims/test is 1,000,000, Domain is from 0 to 2π, q value is smaller than 0.05. Similar parameters were used for running PSEA on circadian genes identified from mouse telogen, except Min items/set is 15 and Max sims/test is 100, 000. Output results in the vsUniform folder were used in this study.

Comparative analysis of circadian clock structure between time-series multiple mouse tissues, human whole blood, and human epidermis

The human whole blood samples were from a research study investigating the effects of insufficient sleep on circadian rhythmicity (11). 26 young and healthy subjects were recruited to take part in a 6 h sleep-restriction condition and a 10 h control condition for one week. Following the last day of sleep-restriction or control condition, blood samples were collected at 3 h intervals for 27 hours. The quantile normalized signal intensities of all blood samples and associated annotation files were downloaded from GEO database (GSE39445), and 201 samples collected from 21 subjects (each subject has at least seven samples) under sleep control condition were used for determining the ECM of core clock genes (validated locomotor activity rhythm changes in knockout mice) (12–15) in human blood. Samples taken from human epidermis with known sampling time information (n = 79) from this study were used to determine the ECM of core clock genes in human skin. In order to have an equal comparison, samples collected from each mouse tissue were restricted to a time-window between ZT24 and ZT46. Spearman's rank

correlation was used to calculate the ECM of core clock genes. Mantel test ('mantel.test' function implemented in the ape R package) was used to test the correlation between ECMs.

Test for functional clocks in defined human blood cell types

For the array data, quantile normalized expression profiles of peripheral $CD14^+$ monocytes (GSE56045) and $CD4^+$ T cells (GSE56580) isolated from human blood samples (16) and their associated annotation files were downloaded from the GEO database. Only those samples collected from subjects younger than 60 years old (593 monocytes samples and 121 T-cell samples) were used for calculating ECM of CCGs. For the RNA-seq data, normalized time-series expression profiles of $CD14^+$ monocytes collected from 10 young subjects (17) were used for calculating ECM of CCGs.

Circadian biomarker selection with ZeitZeiger

The default settings were used to run ZeitZeiger (18) on CYCLOPS ordered samples, with two modifications. First, we normalized the expression value of each gene with the average expression value of three non-cyclers (*GPKOW*, *BMS1* and *NAPG*) in human epidermis. The three non-cyclers were selected by their average expression value (median-high expression genes with intensity value between 60 and 120) and standard deviation (top 10 median-high expression genes ranked by their standard deviation) among all epidermis samples. Second, we filtered out those genes with an rsq value ≤ 0.1. This reduced the number of genes used to search for candidate biomarkers. We ran ZeitZeiger using the training data set with the sumabsv = 3. We then filtered those genes in the first two SPCs with an absolute coefficient value ≥ 0.05 to obtain circadian marker genes. Then expression profiles of these selected circadian marker genes in the same training set were used for a second run of ZeitZeiger with sumabsv = 3, and used the first two SPCs to predict the sample phase in the testing data set.

Statistical analysis

The following statistical methods were used in this study: Fisher's method in MetaCycle, Kuiper's test in PSEA, circular node autoencoders, modified cosinor regression, and the F test as implemented in CYCLOPS, K-medoids clustering in Oscope, maximum-likelihood estimation in ZeitZeiger, Spearman's rank correlation, and Mantel test. Depending on the experimental design and statistical method *P* value cutoffs of < 0.05 and < 0.01 were used for time-series analysis and population level analysis.

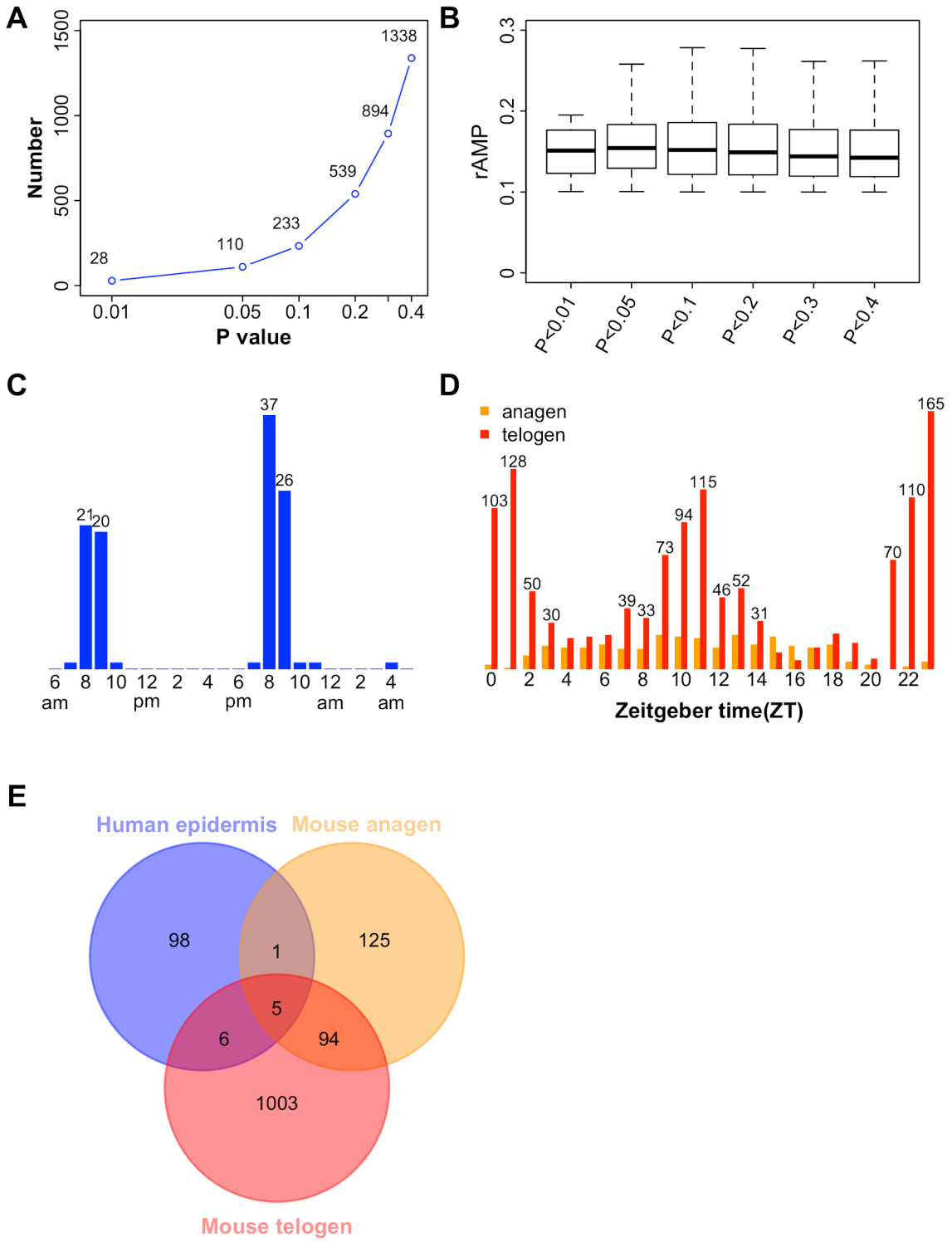

Fig. S1. Comparison of circadian genes identified from time-series analyses of human epidermis and mouse skin tissues. Gene expression data were analyzed using meta3d (human) and meta2d (mouse) using the MetaCycle package. (A) Number of circadian genes identified from time-series

analysis of human epidermis at a series of *P* value cut-offs. (B) The relative amplitude (rAMP) distributions of circadian genes at a series of *P* value cutoffs. (C) Phase distribution of circadian genes ($P < 0.05$) identified in human epidermis. (D) Phase distribution of circadian genes identified in mouse anagen and telogen skin ($P < 0.05$). (E) Overlap of circadian genes between human epidermis (blue) and human homologs from mouse anagen (orange) and telogen (red).

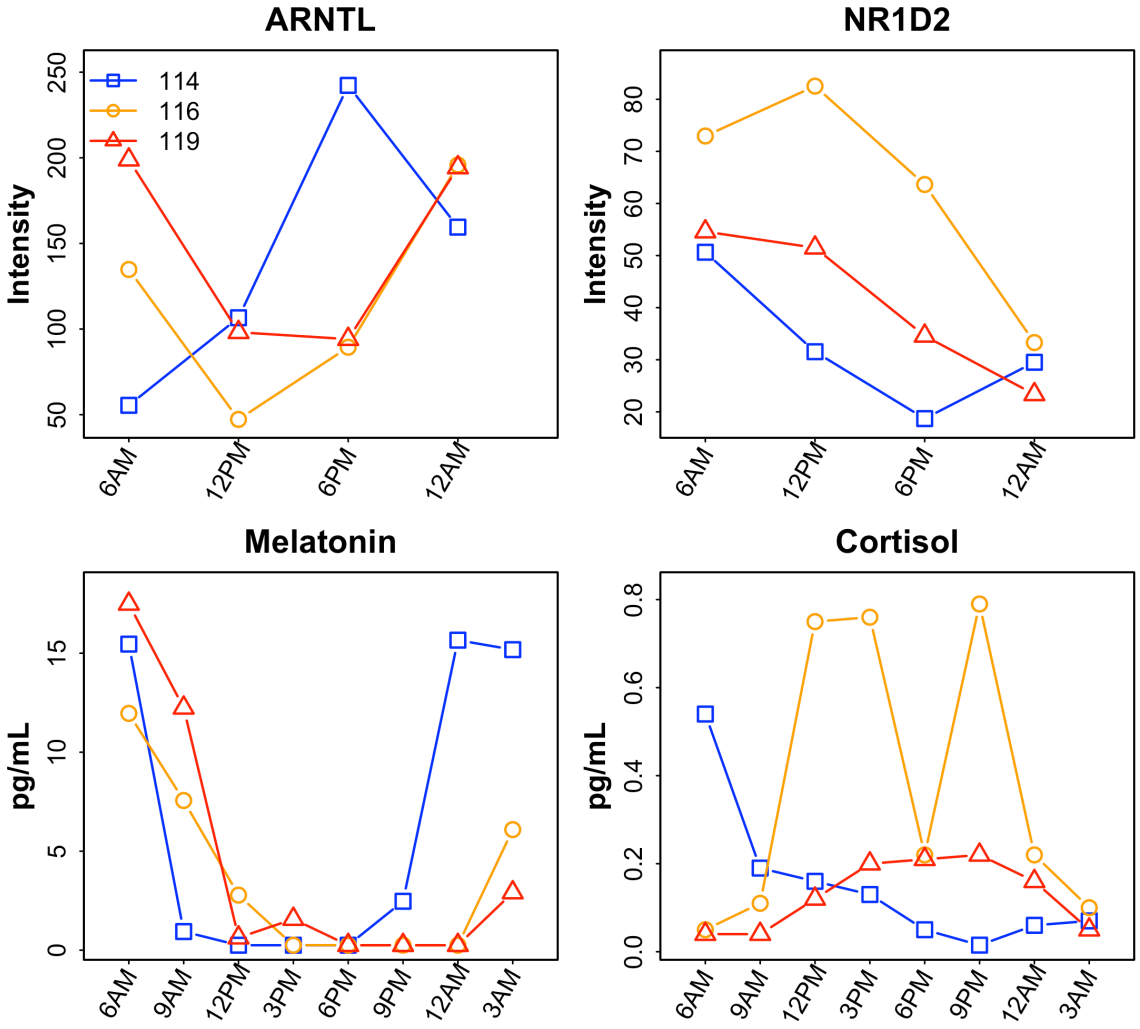

Fig. S2. Salivary melatonin and cortisol levels and clock gene expression profiles in subjects 114, 116, and 119. Subjects indicated with different colors. Melatonin and cortisol both have well-characterized circadian rhythms. In people with normal sleep patterns, melatonin begins to increase in the evening and peaks during hours that most people tend to sleep, which is the pattern we observed in this study. Conversely, cortisol levels begin to increase in the morning and then taper off later in the day.

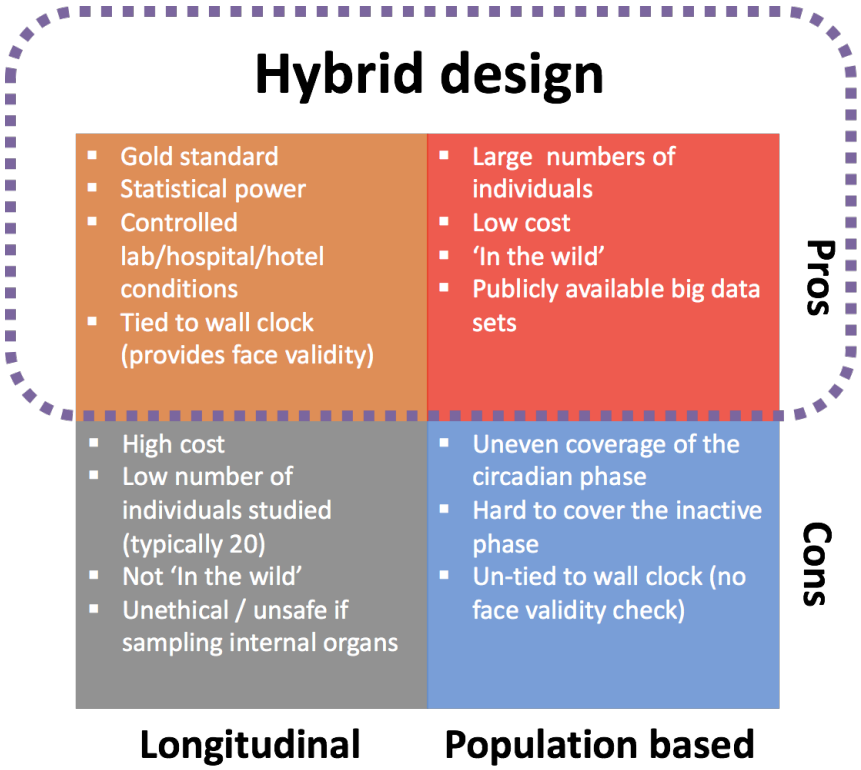

Fig. S3. Pros and cons of experimental designs for human circadian biology. The hybrid method exploits the advantages of both longitudinal and population based designs.

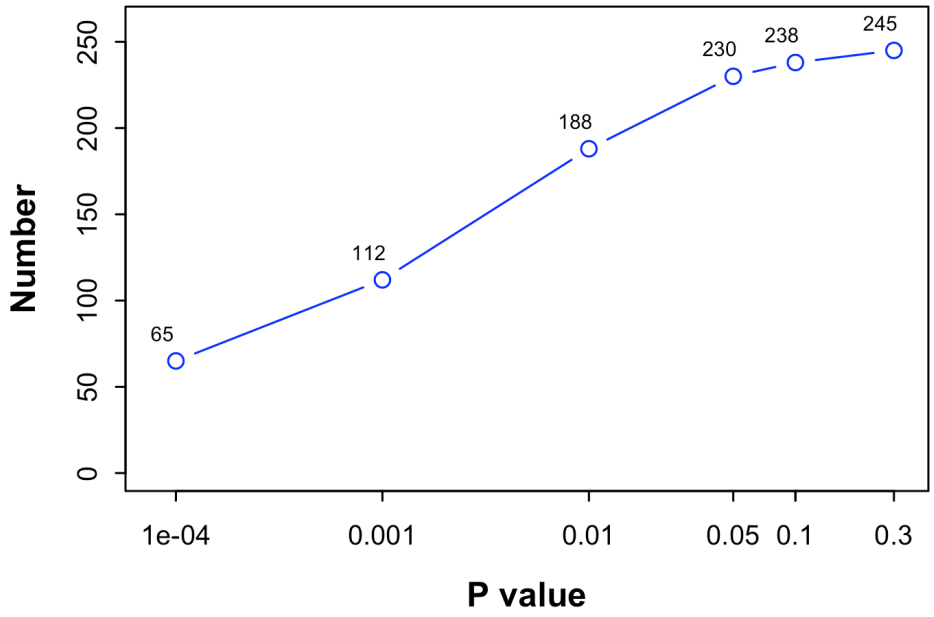

Fig. S4. Number of CYCLOPS-predicted circadian genes identified in human-population level epidermis samples at a series of *P* value cut-offs.

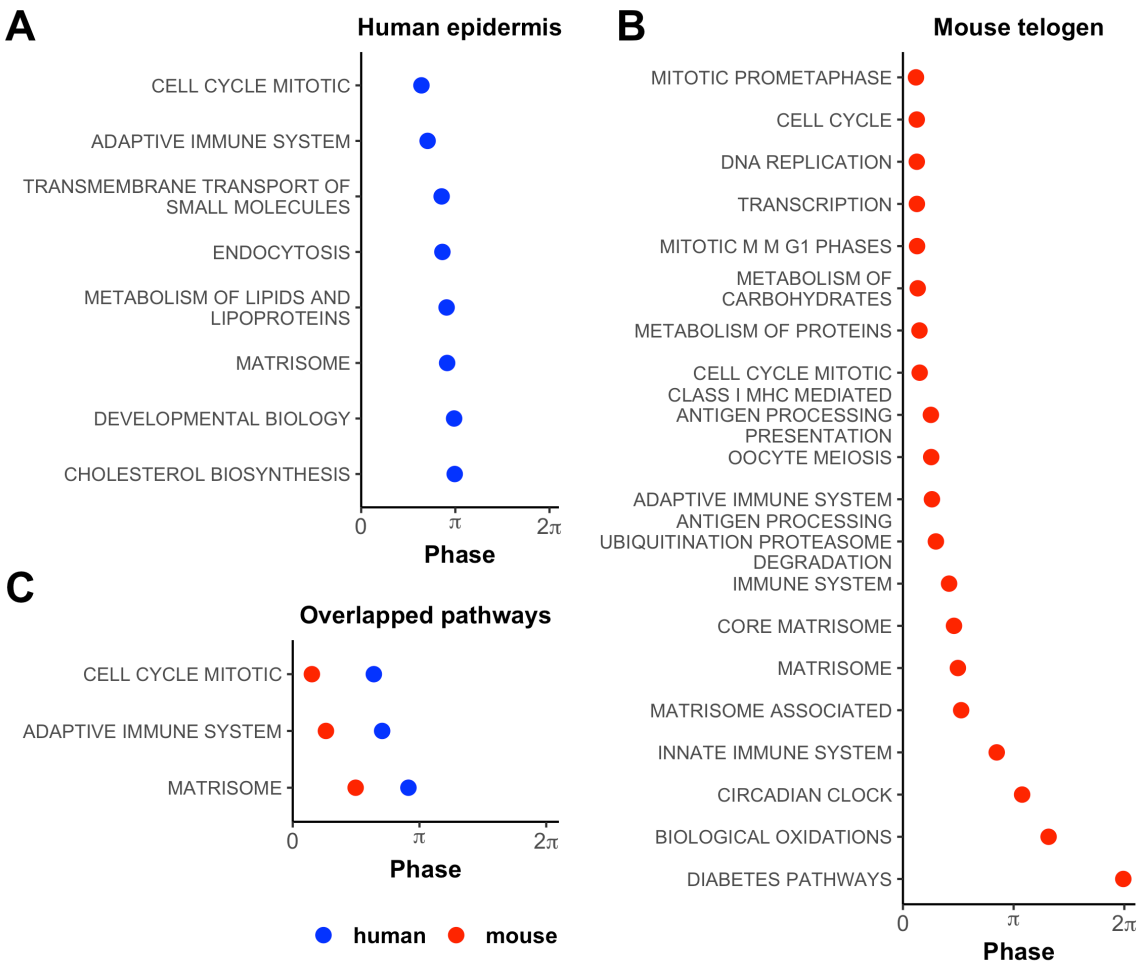

Fig. S5. PSEA analysis of circadian genes identified in human-population level epidermis samples (A) and mouse telogen (B). A less significant cut-off ($P < 0.05$) was used to select genes for performing PSEA analysis in human epidermis. (C) The phase of three overlapping time-dependent pathways in human epidermis (blue) and mouse telogen (red). The x-axis in (A), (B) and (C) depict the circular average phase of circadian genes mapped in the same pathway, which is *ARNTL*/*Arntl* adjusted. Points indicate different pathways. The x-axis depicts circadian gene phase, which is *ARNTL*/*Arntl* adjusted.

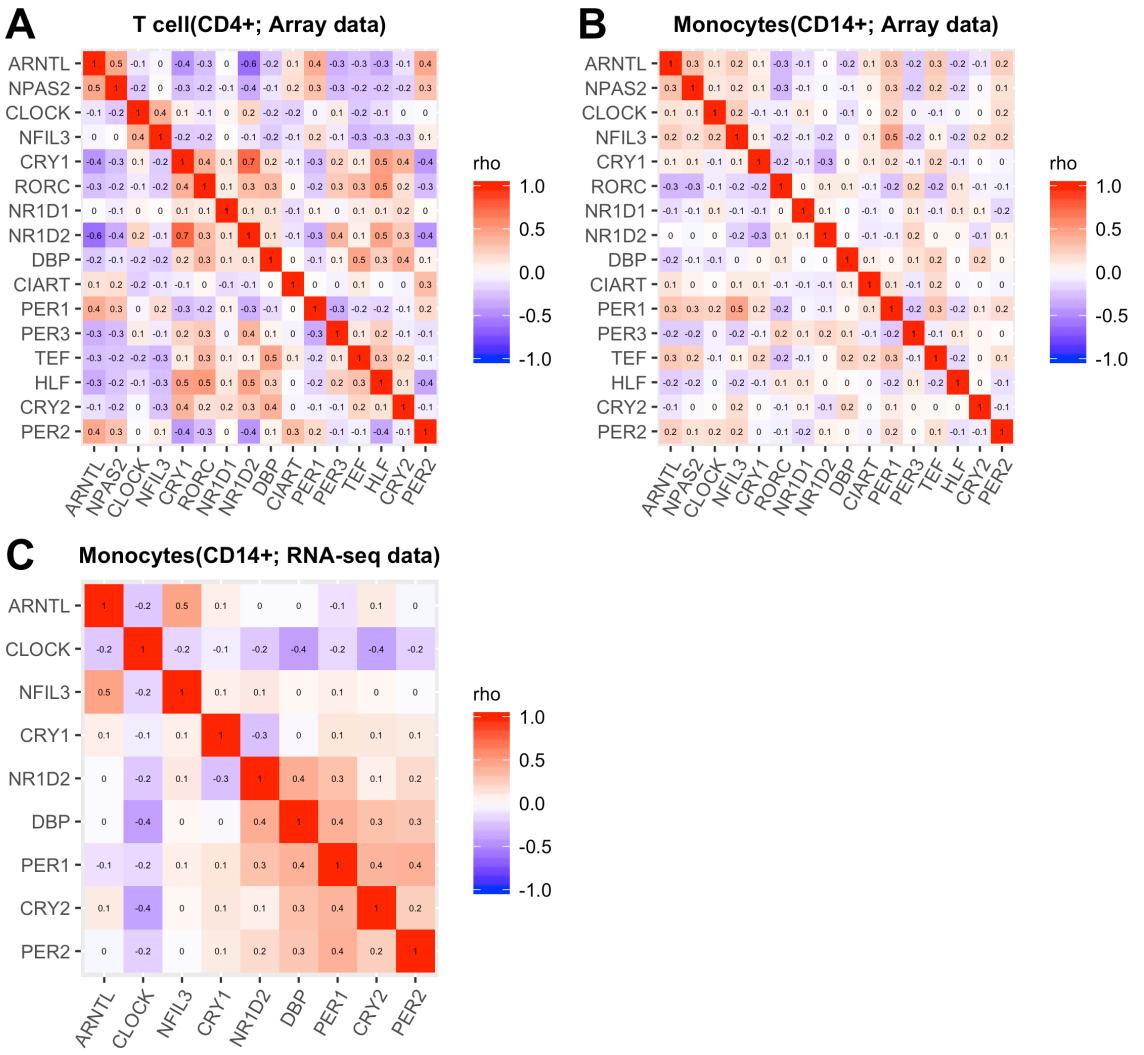

Fig. S6. Evaluation of circadian clock robustness in isolated cell types from human blood. The heat maps of Spearman's rho for CCGs were drawn using array data from isolated $CD4^+$ T-cell (A) and $CD14^+$ monocytes (B) from hundreds of human subjects (16), or using time-series RNA-seq data of isolated $CD14^+$ monocytes (C) from 10 human subjects (17). Note: the clock in monocytes and T-cells is weaker than in mouse tissues and human epidermis shown in Figure 4A. Red and blue indicate positive and negative Spearman's rho values, respectively.

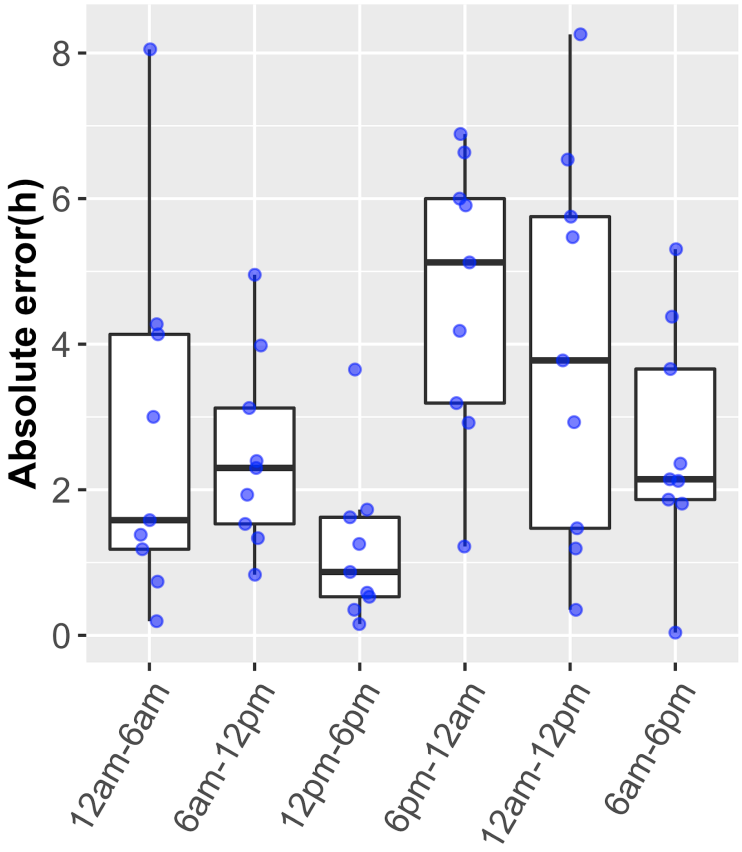

Fig. S7. Validation of circadian markers. Absolute errors are plotted for nine subjects at six time windows.

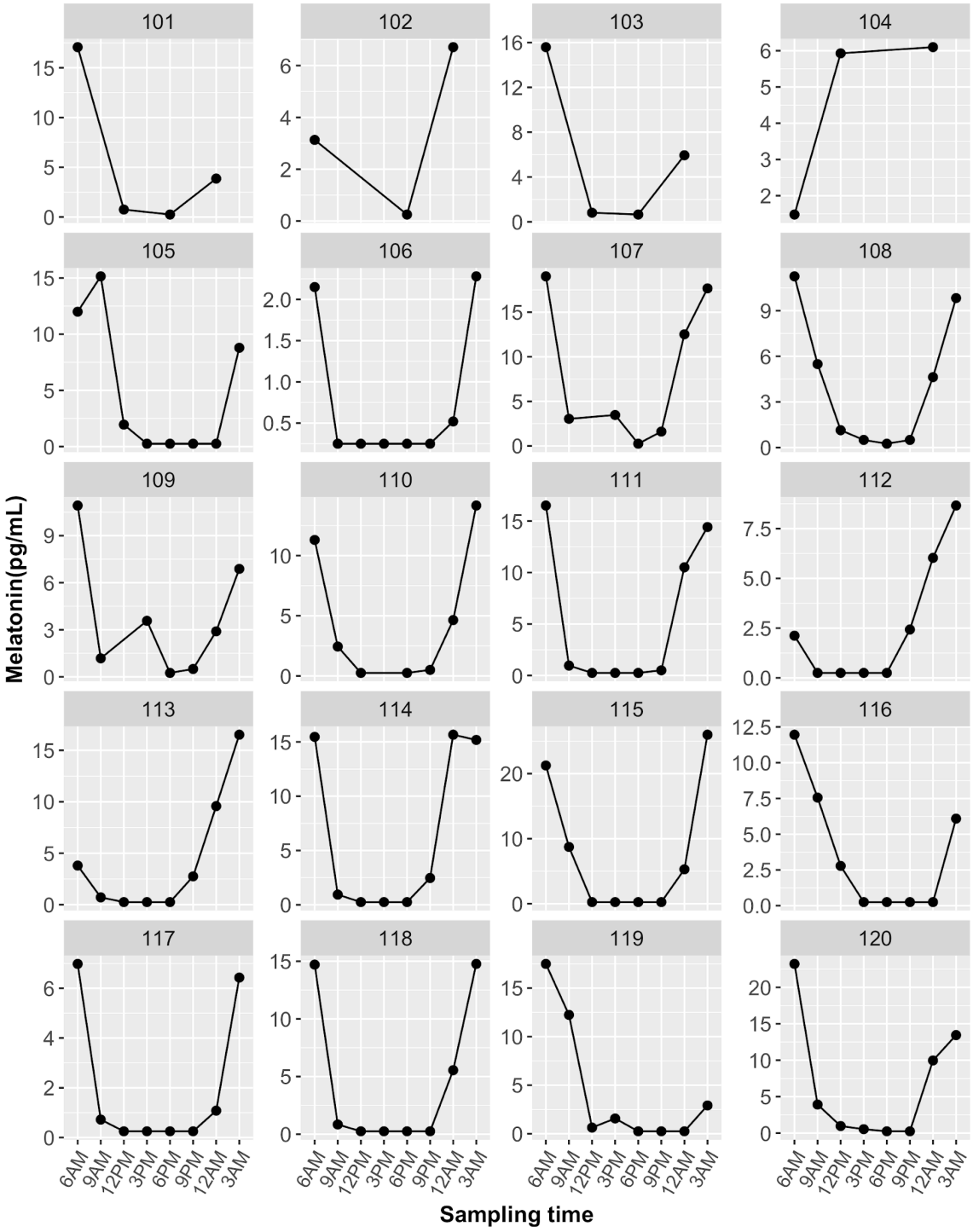

Fig. S8. Daily salivary melatonin levels for 20 subjects. Salivary samples were collected every 3 h for 24 h. Subjects 101, 102, 103, 104, 107, 109 and 110 have some missing values.

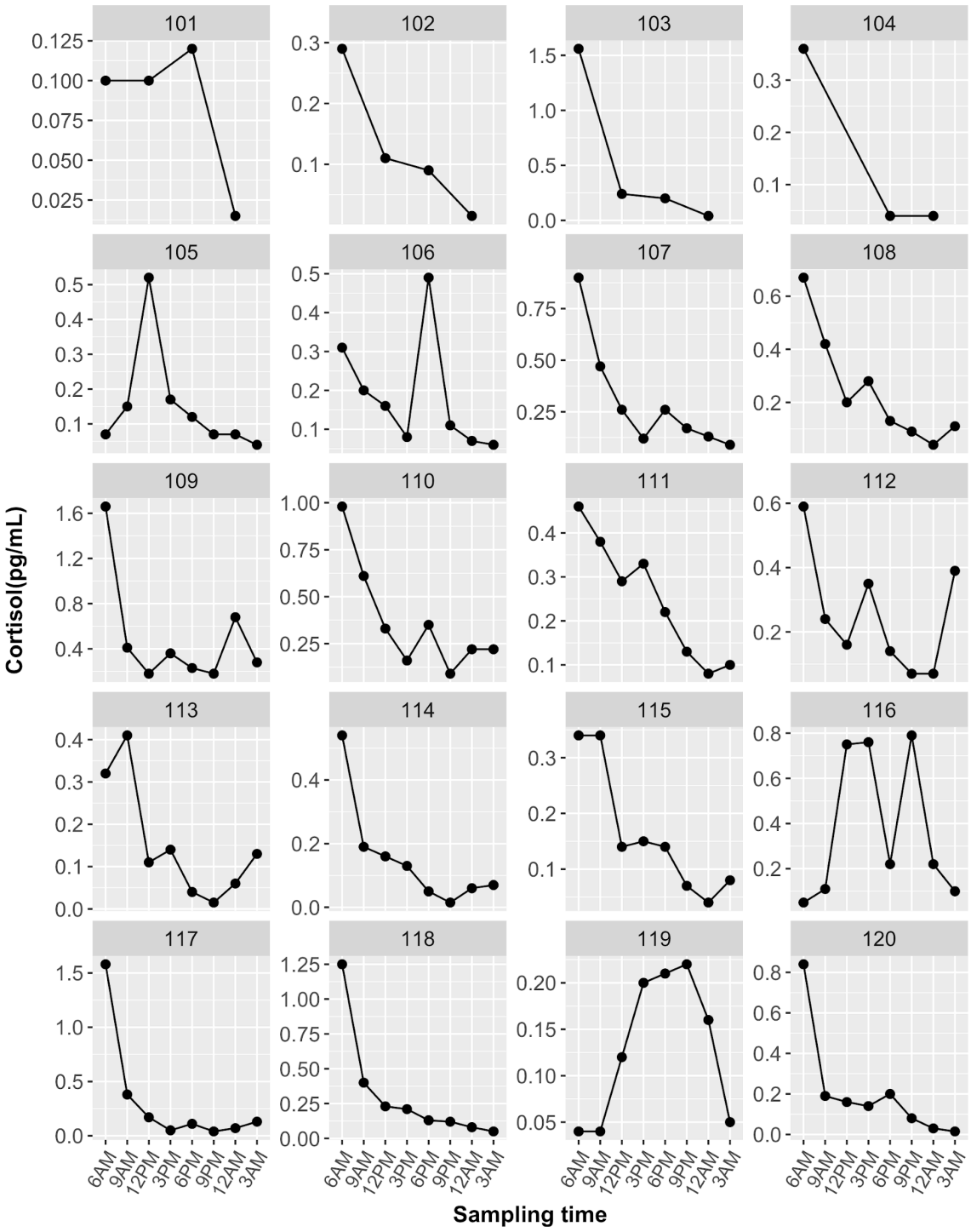

Fig. S9. Daily salivary cortisol levels for 20 subjects. Salivary samples were collected every 3 h for 24 h. Subjects 101, 102, 103 and 104 have some missing values.

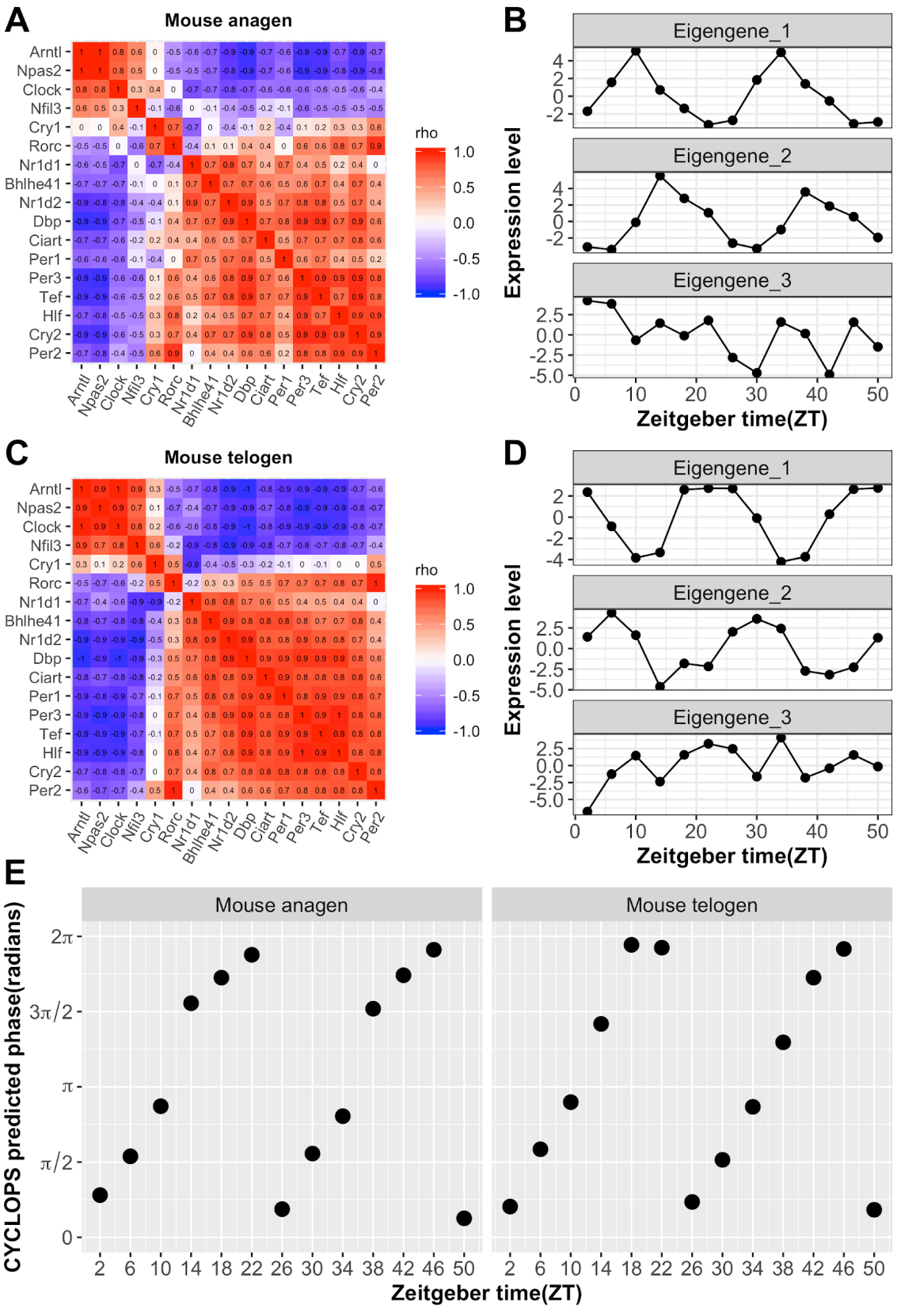
A
Mouse anagen
Arntl
Npas2
Clock
Nfil3
Cry1
Rorc
Nr1d1
Bhlhe41
Nr1d2
Dbp
Ciart
Per1
Per3
Tef
Hlf
Cry2
Per2
rho
1.0
0.5
0.0
-0.5
-1.0
B
Eigengene_1
Eigengene_2
Eigengene_3
Expression level
Zeitgeber time(ZT)
C
Mouse telogen
D
E
Mouse anagen
Mouse telogen
CYCLOPS predicted phase(radians)
2π
3π/2
π
π/2
0
Zeitgeber time(ZT)

Fig. S10. CYCLOPS accurately predicts sample orders from mouse anagen and telogen datasets. A heat map of Spearman's rho for CCGs from ordered mouse anagen (A) and telogen data (C) shows conserved correlation. The expression profiles of the top 3 eigengenes from mouse anagen (B) and telogen (D). Only Eigengenes 1 and 2 in mouse anagen and telogen show periodic profiles, which were selected by Oscope and used for ordering. (E) CYCLOPS accurately re-ordered the mouse anagen and telogen samples. Sampling time and the CYCLOPS predicted sample phase are shown on the x and y axes, respectively.

Table S1. List of datasets used in this study

Datasets	Reference	Platform	Site	#Samples	Experimental design
Mouse anagen	Geyfman M., et al., 2012, PNAS. GSE38623	MoGene-1_0-st	anagen	13	C57BL/6CR mice were housed under 12 h:12 h LD cycles with food and water ad libitum. Whole skin was collected at 4-h intervals for 48 h. Anagen samples were collected from P30 mice. Equal amounts of RNA from the three mice for each time point were pooled.
Mouse telogen	Geyfman M., et al., 2012, PNAS. GSE38622	MoGene-1_0-st	telogen	13	C57BL/6CR mice were housed under 12 h:12 h LD cycles with food and water ad libitum. Whole skin was collected at 4-h intervals for 48 h. Telogen samples were collected from P46 mice. Equal amounts of RNA from the three mice for each time point were pooled.
Caucasian epidermal skin (ordered)	This study. GSE112660	HG-219 array	forearm	79	Four forearm skin samples were collected at 4 time points (12am, 6am, 6pm, 12pm) for each of 20 Caucasian male subjects, except one missing sample from subject 115. The ages of these 20 subjects are between 21 and 49 years old. LCM was performed to separate dermis from epidermis.
Caucasian epidermal skin (unordered)	Kimball A. B., et al., 2018, J Am Acad Dermatol. GSE112660	HG-219 array	forearm	152	One forearm skin sample was taken from each of 152 subjects in a Caucasian female population, aged between 20 and 74 years old. Samples were designed to collect during the working hours, between 9am to 5pm. LCM was performed to separate dermis from epidermis.
African-Americans epidermal skin (unordered)	This study. GSE112660	HG-219 array	forearm	67	One forearm skin sample was taken from each of 67 subjects in a African-Americans female population, aged between 20 and 65 years old. Samples were designed to collect during the working hours, between 9am to 5pm. LCM was performed to separate dermis from epidermis.

Additional data table S1 (separate file)
List of genes cycling in human epidermis at population level.

Additional data table S2 (separate file)
List of seed genes used for CYCLOPS.

References

1. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20(3):307–315.

2. Carvalho BS, Irizarry RA (2010) A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 26(19):2363–2367.

3. Zhang R, Lahens NF, Ballance HI, Hughes ME, Hogenesch JB (2014) A circadian gene expression atlas in mammals: implications for biology and medicine. *Proc Natl Acad Sci U S A* 111(45):16219–16224.

4. Geyfman M, et al. (2012) Brain and muscle Arnt-like protein-1 (BMAL1) controls circadian cell proliferation and susceptibility to UVB-induced DNA damage in the epidermis. *Proc Natl Acad Sci U S A* 109(29):11758–11763.

5. Tang Y, Horikoshi M, Li W (2016) ggfortify: Unified interface to visualize statistical results of popular R Packages. *R J*. Available at: https://journal.r-project.org/archive/2016/RJ-2016-060/RJ-2016-060.pdf.

6. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28(6):882–883.

7. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1):118–127.

8. Anafi RC, Francey LJ, Hogenesch JB, Kim J (2017) CYCLOPS reveals human transcriptional rhythms in health and disease. *Proc Natl Acad Sci U S A* 114(20):5312–5317.

9. Leng N, et al. (2015) Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat Methods* 12(10):947–950.

10. Zhang R, Podtelezhnikov AA, Hogenesch JB, Anafi RC (2016) Discovering Biology in Periodic Data through Phase Set Enrichment Analysis (PSEA). *J Biol Rhythms* 31(3):244–257.

11. Möller-Levet CS, et al. (2013) Effects of insufficient sleep on circadian rhythmicity and expression amplitude of the human blood transcriptome. *Proceedings of the National Academy of Sciences* 110(12):E1132–E1141.

12. Ripperger JA, Jud C, Albrecht U (2011) The daily rhythm of mice. *FEBS Lett* 585(10):1384–1392.

13. Anafi RC, et al. (2014) Machine learning helps identify CHRONO as a circadian clock component. *PLoS Biol* 12(4):e1001840.

14. Goriki A, et al. (2014) A novel protein, CHRONO, functions as a core component of the mammalian circadian clock. *PLoS Biol* 12(4):e1001839.

15. Annayev Y, et al. (2014) Gene model 129 (Gm129) encodes a novel transcriptional repressor that modulates circadian gene expression. *J Biol Chem* 289(8):5013–5024.

16. Reynolds LM, et al. (2014) Age-related variations in the methylome associated with gene expression in human monocytes and T cells. *Nat Commun* 5:5366.

17. Wittenbrink N, et al. (2018) High-accuracy determination of internal circadian time from a single blood sample. *J Clin Invest*. doi:10.1172/JCI120874.

18. Hughey JJ, Hastie T, Butte AJ (2016) ZeitZeiger: supervised learning for high-dimensional data from an oscillatory system. *Nucleic Acids Res* 44(8):e80.