Figures and figure supplements

Blood transcriptome based biomarkers for human circadian phase

Emma E Laing et al
Mapping of $x$ on to $y$ using $n$ features via:

1. Molecular timetable
2. Zeitzeiger
3. Partial Least Squares Regression

with $n$ specific parameters optimised via cross-validation.

Application of models to predict $y$ from $x$

Validation: Model testing

Training: Model building

Figure 1. Development and validation of models for predicting melatonin phase from transcriptomic samples. mRNA abundance and melatonin data from 53 participants collected in four conditions (panels c, d, g and h) were partitioned into two groups: a training set (329 mRNA samples from 26 participants) to build the model and a validation set (349 mRNA samples from 27 participant). For each participant in each condition, hourly melatonin samples (closed circles in plots represent the average melatonin profile of participant in a given experimental condition) and multiple transcriptomic samples (colored triangles) separated by 3 (conditions i and ii) or 4 hr (conditions ii and iv) were collected. Based on a participants’ melatonin profile, each transcriptomic sample for that participant was assigned a circadian phase and plotted separately for each of the conditions (panels c, g). The coloring of circadian bars in panels c and g correspond to the coloring of the transcriptomic sampling time-point triangles in panels d, h. For example, the third mRNA sample (blue triangle) in condition iv was on average taken just before the onset of melatonin (dark green melatonin profile) and the blue horizontal bars in condition iv panel of panel c therefore are located around 300 degrees. This shows the distribution of circadian phases obtained at each sampling point across participants in the training and validation set. These data were used to construct and evaluate the Molecular timetable, Zeitzeiger and Partial Least Square Regression (PLSR) based models. We mapped the melatonin phase $y$ (panels c, g) onto the transcriptomic profile comprising $x$ features in the training set (panels a, e). Parameters to build the model were selected by performing leave-one-participant-out cross-validation of the training set (panel b). The final trained model was applied to the validation set for assessing the overall performance of the model (panel f).

DOI: 10.7554/eLife.20214.002
Figure 2. Rhythmity in the conditions and identification of molecular timetable genes. (a, d) Square of correlation value ($r^2$) vs. rank of correlation as a measure of overall 24 hr rhythmity in the transcriptome, separately and across conditions. For each transcript, the correlation with a cosine wave was
computed, squared and the entire transcriptome was then rank-ordered separately for each conditions and across conditions. (b, e) Rank of the correlation of phase marker genes from different phase marker lists across conditions (indicated by color panel to the left) for participants in the training (b) and validation set (e). First, correlations of all ~26K features were ranked separately for conditions and training and validation set and then the rank of identified phase marker genes was identified and plotted. (c,f) Heatmap of r values of correlation between a 24 hr cosine wave and mRNA abundance profiles of a feature(s) targeting a gene identified as a phase marker by our molecular timetable, and the circadian phase marker genes published in Lech et al. (2016) and Hughey et al. (2016). Correlations were computed separately for each of the four conditions (indicated by the color panel at the top of the heatmap) and all conditions combined for the training (c) or validation set (f). Rows in the heatmap are ordered based on the correlation column for the ‘all conditions’ dataset, that is, the column indicated with black. Data source files: Figure 2—source data 1.

DOI: 10.7554/eLife.20214.003

The following source data is available for figure 2:

Source data 1. Source data file for generating panels in Figure 2.
DOI: 10.7554/eLife.20214.004
Figure 2—figure supplement 1. Identification of correlation cutoff threshold for the construction of a molecular timetable. (a) Effect of cutoff threshold on performance of molecular timetable models constructed using genes meeting a particular threshold. Performance assessed by $R^2$ values following leave-one-participant-out cross-validation of the training set. (b) Performance quantified as the proportion of samples vs. cumulative error for each constructed model following leave-one-participant-out cross-validation of the training set (colors correspond to the threshold colors in panel a).

DOI: 10.7554/eLife.20214.005
Figure 2—figure supplement 2. mRNA abundance profiles for genes belonging to the phase marker gene list generated by our implementation of the molecular timetable model. Data are shown separately for the four conditions. Color of lines correspond to that in Figure 1 of the main text. Light blue: ‘sleep in phase with melatonin’, Dark blue: ‘sleep out of phase with melatonin’, Light green: ‘total sleep deprivation, no prior sleep debt’, and Dark green: ‘total sleep deprivation, prior sleep debt’. Title of each sub-plot denotes the probe and gene name for which the temporal profile is shown. x-axis of each sub-plot corresponds to circadian phase in degrees, where zero degrees corresponds to Dim Light Melatonin Onset (DLMO). y-axis indicates the Z-scored mRNA abundance. Average time course calculated from aligned (by DLMO) participant mRNA abundance profiles, with error bars indicating Standard Error. Data shown for participants belonging to the training set only.

DOI: 10.7554/eLife.20214.006
Figure 2—figure supplement 3. mRNA abundance profiles for genes belonging to the phase marker gene list published in (Lech et al., 2016). Data are shown separately for the four conditions. Color of lines correspond to that in Figure 1 of the main text; Light blue: ‘sleep in phase with melatonin’, Dark blue: ‘sleep out of phase with melatonin’, Light green: ‘total sleep deprivation, no prior sleep debt’, and Dark green: ‘total sleep deprivation, prior sleep debt’. Title of each sub-plot denotes the probe and gene name for which the temporal profile is shown. x-axis of each sub-plot corresponds to circadian phase in degrees, where 0 degrees corresponds to Dim Light Melatonin Onset (DLMO). y-axis the Z-scored mRNA abundance. Average time course calculated from aligned (by DLMO) participant mRNA abundance profiles, with error bars indicating Standard Error. Data shown for participants belonging to the training set only.
DOI: 10.7554/eLife.20214.007
Figure 2—figure supplement 4. mRNA abundance profiles for genes belonging to the phase marker gene list published in (Hughey et al., 2016). Data are shown separately for the four conditions. Color of lines correspond to that in Figure 1 of the main text; Light blue: ‘sleep in phase with melatonin’, Dark blue: ‘sleep out of phase with melatonin’, Light green: ‘total sleep deprivation, no prior sleep debt’, and Dark green: ‘total sleep deprivation, prior sleep debt’. Title of each sub-plot denotes the probe and gene name for which the temporal profile is shown. x-axis of each sub-plot corresponds to circadian phase in degrees, where zero degrees corresponds to Dim Light Melatonin Onset (DLMO). y-axis the Z-scored mRNA abundance. Average time course calculated from aligned (by DLMO) participant mRNA abundance profiles, with error bars indicating Standard Error. Data shown for participants belonging to the training set only.

DOI: 10.7554/eLife.20214.008
Figure 3. Performance of one-sample models derived from the training set when used to predict circadian phase in the validation set. (a, d and g) Predicted circadian phase of a blood sample vs. the observed melatonin phase for each sample in the validation set for the one-sample molecular timetable model (a), the one-sample Zeitzeiger model (d) and the one-sample PLSR-based model (g). Circadian phase 0 corresponds to the onset of the melatonin rhythm, which across conditions occurred on average at 23:30 ± 15 min (across all conditions and the training and validation data sets). Please note that the data are double-plotted. The line of unity represents perfect prediction. (b, e and h) Average (mean) error and 95% confidence intervals of the mean error per 30°, that is 120 min, for each 30° circadian phase bins across the circadian cycle, for each of the models. The grey area plot represents the average melatonin profile. Panels (c, f and i) frequency distribution of errors. Errors were sorted into 15°, that is, 60 min bins for each model. For each of the models, the mode of the distribution is near 0, but the width of the distribution varies across models. (j) Proportion of predictions based on one sample vs. cumulative error as a measure of overall performance. (k) Proportion of predictions based on two consecutive samples vs. cumulative two sample average error as a measure of overall performance. (l) Proportion of predictions based on three consecutive samples vs. cumulative three sample average error as a measure of overall performance. Data source file Figure 3—source data 1.

DOI: 10.7554/eLife.20214.009

The following source data is available for figure 3:

Source data 1. Source data file for generating panels in Figure 3.

DOI: 10.7554/eLife.20214.010
**Figure 3—figure supplement 1.** Performance of one-sample molecular timetable models when used to predict the circadian phase of samples in the validation set. Color of lines correspond to that in Figure 1 of the main text; Light blue: ‘sleep in phase with melatonin’, Dark blue: ‘sleep out of phase with melatonin’, Light green: ‘total sleep deprivation, no prior sleep debt’, and Dark green: ‘total sleep deprivation, prior sleep debt’. Panels (a, d and g) predicted circadian phase vs. experimentally observed circadian phase for each sample in the validation set for our one-sample molecular timetable model (a), molecular timetable constructed from the genes published in Lech et al. (2016) (d), molecular timetable constructed from the genes published in Hughey et al. (2016) (g). Data are double plotted. (b, e and h) Average (mean) error and 95% confidence intervals of the mean error per 30°, that is, 120 min, for each 30°Circadian phase bins across the circadian cycle, for each of the models. The grey area plot represents the average melatonin profile. (c, f and i) Frequency distribution of errors. Errors were sorted into 15°, that is, 60 min bins, for each model. For each of the models, the mode of the distribution is near 0 but the width of the distribution varies across models. **Figure 3—figure supplement 1 continued on next page**
Figure 3—figure supplement 2. Parameter selection for constructing a Zeitzeiger based model. To build a Zeitzeiger predictor, given a training data set, there are two parameters that need to be optimized: ‘sumabsv’, which controls the number of features to be used, and ‘nSPC’, which controls the number of principal components that best describe (explain) the variance in mRNA abundance, measured by features, over time. The performance of differing combinations of sumabsv and nSPC were assessed using cross-validation (see Materials and methods). Panel a shows the change in performance, measured by mean average error (MAE), between the predicted and true circadian phase of samples in the training set, of the sumabsv parameter selection.
input parameter of the Zeitzeiger algorithm across different nSPC values. The sumabsv value (4) resulting in the smallest MAE is optimal. Panel b shows the proportion of variance explained of differing values of the nSPC input parameter when sumabsv = 4. The nSPC value (3) giving the highest proportion of variance explained is optimal. (c) mRNA abundance features identified for the construction of the final Zeitzeiger model, using the parameters sumabsv = 4 and nSPC = 3. Bar chart represents the contribution (i.e. weight, [coefficient on x-axis]) of a feature to each of the three components spc1, spc2 and spc2. Note that there are multiple, different features (probes) that target a gene on the microarray platform we have used (e.g. multiple CPID probes targeting PER3, see GEO [Barrett et al., 2013; Accession number: GPL15331]).

DOI: 10.7554/eLife.20214.012
Figure 3—figure supplement 3. Phase profiles of genes in the phase marker lists for the molecular timetable and Zeitzeiger models. Grey lines indicate the circadian phase of a gene’s maxima when Z-scored, where the number of grey lines equals the number of genes (features) used to construct the model, either the molecular timetable (a) or the Zeitzeiger model (b). Colored lines represent cosine waves with a given (gene’s) phase, with a gradient of red to blue via yellow that is cosine waves with lower phases are colored in red, cosine waves with higher phases are colored in blue. DOI: 10.7554/eLife.20214.013
Figure 3—figure supplement 4. Selection of number of abundance features and latent factors and pseudo time-course of latent factor scores for Partial Least Squares Regression (PLSR)-based models. (a) Leave-one-participant-out cross-validation performance of one-sample PLSR models applied.
Figure 3—figure supplement 4 continued

to training set when using different combinations of n mRNA abundance features and T latent factors. (b) Pseudo-time course of latent factor scores for each of the five latent factors used to build the one-sample PLSR-based model; x-axis comprises the 329 samples in the training set ordered by their true circadian phase vs. y-axis of corresponding z-scored latent factor score (a linear combination of the 100 feature values used to generate the model multiplied by the corresponding latent factor loadings [weights]) for the sample. Black line represents the fitted cosine wave to the pseudo time-course. (c) Leave-one-participant-out cross-validation performance of two-sample differential mRNA abundance PLSR models applied to training set when using different combinations of n mRNA abundance features and T latent factors. (d) Pseudo-time course of factor scores for each of the five latent factors used to build the two-sample differential PLSR-based model; x-axis comprises the 163 paired samples in the training set ordered by their true circadian phase vs y-axis of corresponding z-scored latent factor score (a linear combination of the 100 feature values used to generate the model multiplied by the corresponding latent factor loadings [weights]) for the sample. Black line represents the fitted cosine wave to the pseudo time-course.

DOI: 10.7554/eLife.20214.014
Figure 3—figure supplement 5. Comparison of the number of samples used as input vs. accuracy for each circadian phase prediction method. Difference in model performance, as measured by the proportion of predictions vs. cumulative error, when using a given number of samples (one sample, differential two samples, two samples, three samples).
sample, two consecutive samples, three consecutive samples, two sample differential, see Materials and methods), as input for each circadian phase prediction method. (a) Molecular timetable method using the genes published in (Lech et al., 2016); (b) Molecular timetable method using the genes published in (Hughey et al., 2016); (c) Molecular timetable method using the set of 73 features derived from this study; (d) Zeitzeiger method using the set of 107 features derived from this study; (e) PLSR method using the 100 features identified by our one-sample approach (for the one, two and three samples) and 100 features identified by our two-sample differential approach (for the two-sample differential).

DOI: 10.7554/eLife.20214.015
Figure 4. Performance of two-sample differential mRNA abundance-based models when used to predict the circadian phase in the validation set. (a and d) Predicted circadian phase of a blood sample vs. the observed circadian melatonin phase for each sample in the validation set for the two-sample differential molecular timetable model (a) and the two-sample differential PLSR-based model (d). Data are double-plotted. (b and e) Average (mean) error and 95% confidence intervals of the mean error per 30°, that is, 120 min, for each 30° circadian phase bins across the circadian cycle, for Figure 4 continued on next page
each of the models. The grey area plot represents the average melatonin profile. (c and f) Frequency distribution of errors. Errors were sorted into 15°, that is, 60 min bins, for each model. For each of the models, the mode of the distribution is near 0, but the width of the distribution varies across models. (g) Performance of the differential PLSR (orange), and the differential molecular timetable method (black) quantified by proportion of samples vs. cumulative error. Data source file Figure 4—source data 1.
DOI: 10.7554/eLife.20214.017
The following source data is available for figure 4:

Source data 1. Source data file for generating panels in Figure 4.
DOI: 10.7554/eLife.20214.018
Figure 5. Circos plot for visual comparison of lists of biomarkers forming the generated circadian phase prediction models. Circular tracks from outside in: (1) Name of model being compared (Molecular timetable, Zeitgeber, PLSR one-sample, PLSR two-sample differential), where all models are.

Figure 5 continued on next page.
constructed from the same training set. The color of the model name corresponds to the color of the bar in track 3. (2) Unique HGNC Gene symbols or probe identifiers for features from which the specified model (adjacent track 1 and 3) are created; (3) Color bar indicating the model (Black = Molecular timetable, Orange = PLSR one-sample, Dark orange = PLSR two-sample differential, Blue = Zeitzeiger). Inner arcs represent the overlap between models. Genes present in two models are linked by yellow arcs; genes present in three models are linked by blue arcs; and genes present in all models are linked by red arcs. Figure generated using the R package OmicCircos (Hu et al., 2014).

DOI: 10.7554/eLife.20214.019
A glucocorticoid-driven network links many of the top-ranked PLSR genes. Twelve of the top-25 ranked genes in the PLSR onesample and PLSR two-sample differential models (indicated by a red asterisk) are linked in a network that is driven by glucocorticoid signaling. Glucocorticoid (GC) acting via glucocorticoid receptors (GR) has been shown to stimulate the expression of thousands of genes in human peripheral blood (Menke et al., 2012). GR is inactive when it is bound by a group of chaperone molecules that includes FKBP5 (Park et al., 2007). Active GR binds to GC response elements (GRE) in the promoters and induces the expression of GC-regulated genes, which include peripheral circadian clock genes such as PER1 and PER2 (Cuesta et al., 2015; So et al., 2009; Leliavski et al., 2015), but also in many other genes such as FKBP5 (Park et al., 2007). GR can also induce the rhythmic expression of NR1D2 (So et al., 2009). GC induces the expression of the phosphatase DUSP1 (Shah et al., 2014), which suppresses the inflammatory response by inhibiting MAP kinases (Abraham and Clark, 2006). GC down-regulates the expression of the tyrosine kinase TXK, which regulates T-cell development (Petrillo et al., 2014). GR increases the activity of CEBPβ (Berg et al., 2005), which is a transcription factor whose expression is increased during macrophage differentiation and mediates the expression of inflammatory mediators including CLECA4 (Akagi et al., 2010). GC reduces the expression of the metalloproteinase ADAMTS1 that regulates extracellular matrix function (Wong et al., 2016). GC inhibits WNT signaling (Naito et al., 2015). DAAM2 binds to and stabilizes DVL enhancing WNT signaling during cell development (Lee and Deneen, 2012). The mTOR signaling pathway senses nutrient stress and interacts with components of the WNT signaling pathway (Inoki et al., 2006). DDIT4 is a repressor of mTOR (Gordon et al., 2015) and is transcriptionally up-regulated by GC in response to stress and fasting (Polman et al., 2012; McGhee et al., 2009). FLT3 is a tyrosine kinase that binds to GR and potentiates its signaling (Asadi et al., 2008). FLT3 also regulates dendritic cell development by activating mTOR signaling (Chen et al., 2010). ZBTB16 is a transcriptional repressor that is enhanced by GC and regulates GC response genes and affects GC-induced apoptosis (Wasim et al., 2010, 2012).

DOI: 10.7554/eLife.20214.020