SLEEP, 2023, 1–15
https://doi.org/10.1093/sleep/zsac279
Advance access publication 3 December 2022
Original Article

Original Article

Integration of genome-scale data identifies candidate sleep regulators

Yin Yeng Lee[1,2], Mehari Endale[3], Gang Wu[1], Marc D. Ruben[1], Lauren J. Francey[1], Andrew R. Morris[3], Natalie Y. Choo[4], Ron C. Anafi[5], David F. Smith[4,6,7,8], Andrew C. Liu[3] and John B. Hogenesch[1,7,*]

[1]Divisions of Human Genetics and Immunobiology, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 45229, USA,
[2]Department of Pharmacology and Systems Physiology, University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA,
[3]Department of Physiology and Aging, University of Florida College of Medicine, Gainesville, FL 32610, USA,
[4]Division of Pediatric Otolaryngology-Head and Neck Surgery, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA,
[5]Department of Medicine, Chronobiology and Sleep Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[6]Division of Pulmonary Medicine and the Sleep Center, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA,
[7]Center for Circadian Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA and
[8]Department of Otolaryngology - Head and Neck Surgery, University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA

[*]Corresponding author. John B. Hogenesch, Divisions of Human Genetics and Immunobiology, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 45229, USA, Email: john.hogenesch@cchmc.org.

Abstract

Study Objectives: Genetics impacts sleep, yet, the molecular mechanisms underlying sleep regulation remain elusive. In this study, we built machine learning models to predict sleep genes based on their similarity to genes that are known to regulate sleep.

Methods: We trained a prediction model on thousands of published datasets, representing circadian, immune, sleep deprivation, and many other processes, using a manually curated list of 109 sleep genes.

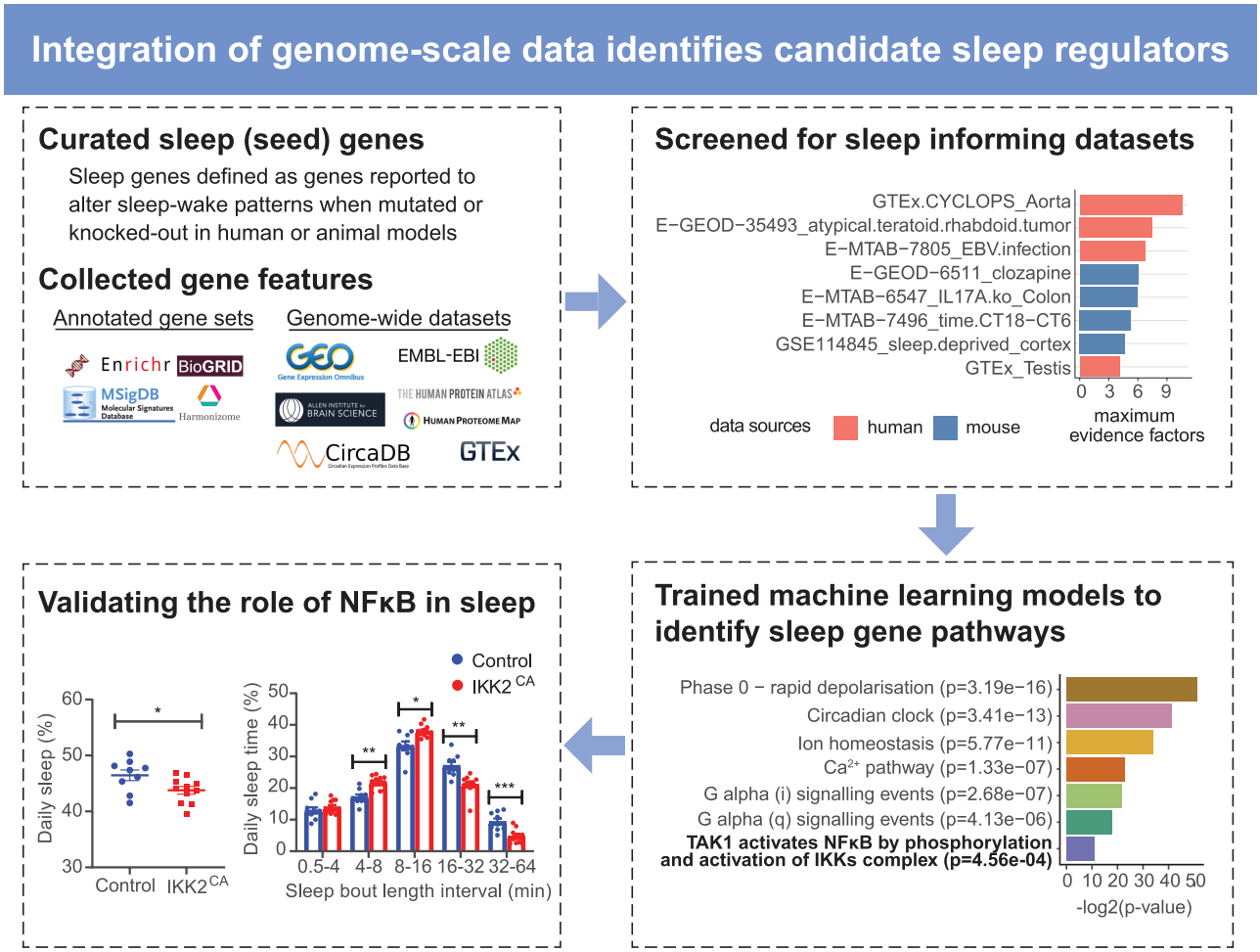
Results: Our predictions fit with prior knowledge of sleep regulation and identified key genes and pathways to pursue in follow-up studies. As an example, we focused on the NF-κB pathway and showed that chronic activation of NF-κB in a genetic mouse model impacted the sleep-wake patterns.

Conclusion: Our study highlights the power of machine learning in integrating prior knowledge and genome-wide data to study genetic regulation of complex behaviors such as sleep.

Key words: sleep regulation; genetics; machine learning; genome-scale data integration

Graphical Abstract

Submitted for publication: June 21, 2022; **Revised:** September 2, 2022



Statement of Significance

We applied computational approaches to integrate the wealth of publicly available genome-scale data to study the genetics of sleep regulation. We built machine learning models to predict clusters of genes based on similarity to known sleep genes. Our models identify a number of biological pathways involved in sleep and provide novel candidates for future research. We tested one of these predictions and showed that chronic activation of the NF-κB pathway alters sleep amount and leads to sleep fragmentation in mice. Thus, we present a framework for integrating large-scale genome data to discover genetic entities that underlie sleep and, potentially, other complex traits.

Introduction

Genetics impacts sleep. In humans, a handful of alleles are known to cause familial sleep disorders [1–8]. However, most of these alleles are rare and have not been broadly implicated in sleep regulation in human populations. Genome-wide association studies (GWAS) identified more sleep-trait-associated genes, but SNP-based heritability estimates are small [9–12], and few of these genes have been functionally validated. Many key features of sleep are conserved from invertebrates to vertebrates [13]. Large-scale forward genetics screens in flies [14–18] and mice [19, 20] have identified several genes whose alterations impacted sleep regulation. The two-process model proposed that both circadian clocks and sleep homeostasis drive the sleep-wake cycle [21]. Multiple genome-wide -omics studies have sought to identify key genes and proteins that respond to sleep homeostasis [22–26]. These efforts have built rich resources on data-driven research to identify genes and pathways that regulate sleep.

Computational approaches such as mathematical modeling can identify genes that regulate sleep [27, 28]. However, these models require detailed prior knowledge of mechanism and the computational cost increased drastically to explore models incorporating multiple pathways. Machine learning models, specifically discriminative models, have predictive power to classify genes based on hidden patterns in large datasets [29–33]. These models have been utilized to identify novel risk genes in complex diseases, such as autism and Alzheimer's disease [34, 35]. Advances in "omics" technology have led to increasingly large amounts of data generated each year. To date, the wealth of genome-wide datasets available in the public database has not been integrated to study the genetic regulation of complex physiology and

behavior like sleep. In prior work, we applied machine learning to identify novel circadian genes by incorporating five genome-wide datasets [36]. In this study, we applied machine learning to thousands of datasets with the goal of identifying genes and pathways involved in sleep regulation.

Using a manually curated list of 109 known sleep genes, we trained a prediction model on large-scale published datasets, representing circadian regulation, immune response, sleep deprivation, and many other processes. Our model predicted a list of top candidate sleep genes. Pathway enrichment analysis of these top candidates revealed the NF-κB pathway as a key sleep regulator. We validated that activation of the NF-κB pathway in neurons leads to sleep fragmentation in mice. Further exploration proposed that the NF-κB pathway affects sleep through calcium signaling. In summary, we present an integrative in silico approach with the potential to identify novel genetic regulators of complex physiology and behavior.

Methods

Data curation and preprocessing

Gene name conversion.

Gene name conversion between species was done using the homologene function (v1.5.68) (homologeneData2) [37]. Gene aliases (human and mouse) were converted to official gene symbols according to gene info downloaded from NCBI on 04/01/2020 (hereafter referred to as "gene_info_04012020") [38].

Annotated gene sets.

GMT files, representing gene and protein knowledge from annotate gene set collections, were downloaded from MSigDB [39] ($n = 11$) and Harmonizome [40] ($n = 7$). Protein–protein interaction information was downloaded from BioGRID [41, 42] ($n = 1$). The GMT file was created by including interaction types—colocalization, direct interaction, and association and physical associations from human data. Protein names were matched with the official gene name using gene_info_04012020.

Genome-wide profiling data.

Genome-wide profiling data were downloaded from multiple public repositories. In total, 7195 data metrics were processed, described below.

TISSUE-SPECIFIC TRANSCRIPT ABUNDANCE (n = 595).

Microarray data from human tissues were downloaded from BioGPS—GSE1133 [43, 44]. Average values from each tissue were transformed with log2 to create data metrics ($n = 84$). RNA-seq quantifications from human tissues were downloaded from GTEx [45]. Average TPM from the same tissues were transformed with log2 to create tissue-specific transcript abundance data metrics ($n = 54$). Additional RNA-seq data, transcript expression summarized at per gene (protein) level, were downloaded from Human Protein Atlas (HPA) [46]. Log2 protein-transcripts per million (pTPM) were used to create data metrics (n = 43). Brain region-specific transcripts quantifications (log2 transformed) from Allen Brain Map [47] were downloaded from Harmonizome. The mRNA expression data representing brain structures' specific transcript abundances were used to create data metrics ($n = 414$).

TISSUE-SPECIFIC PROTEIN ABUNDANCE (n = 30).

Mass spectrometry-based proteomics data from human adult and fetal tissue samples were downloaded from the Human Proteome Map [48]. Normalized quantifications from the gene-level expression matrix were transformed with log2 to create data matrices ($n = 30$).

SIGNIFICANCE OF CIRCADIAN EXPRESSION ($n = 25$).

Time-series data from mouse tissues were downloaded from GSE54652 [49] and rhythmic signals were detected using Meta2D-JTK in MetaCycle [50]. Transformed significant value, -log2(p-value), was used to create data metrics ($n = 12$) that represent the significance of circadian expression in mouse tissues. Circadian expressions from human populations, ordered by CYCLOPS [51], were downloaded [52, 53]. Transformed significance values, -log2(p-value) were used to create data metrics ($n = 13$).

TRANSCRIPTIONAL PROFILES UNDER PERTURBATIONS OR DIFFERENT PHYSIOLOGICAL/PATHOLOGICAL CONDITIONS ($n = 6540$).

15 datasets from Gene Expression Omnibus (GEO) [54] were downloaded and preprocessed manually. Absolute log2 fold-changes for each tested condition were used to create data metrics ($n = 46$). In addition, 2459 human and mouse-processed datasets were downloaded from EBI expression atlas [55]. Data metrics ($n = 6494$) were created using absolute log2 fold-changes for each tested condition.

MISCELLANEOUS ($n = 5$).

Phosphorylation site information was downloaded from qPhos [56]. The number of tyrosine, serine/threonine, tyrosine, and serine/threonine phosphorylation sites in each protein were used to create data metrics ($n = 3$). Vertebrate homology information from 10 vertebrates, including human, chimpanzee, rhesus macaque, dog, cattle, rat, mouse, chicken, western clawed frog, and zebrafish were downloaded from Mouse Genome Informatics (MGI) [57]. The number of vertebrates that share a homolog gene with humans was used to create a data metric ($n = 1$) representing conservation of genes. Transcriptomics profiles from HeLa cells enriched for different phases of the cell cycle were downloaded from GSE26922 [58]. Rhythmic genes were detected using Meta2D-LS in MetaCycle [50]. Transformed significant value, -log2(p-value), was used to create data metrics ($n = 1$) that represent the significance of cell-cycle rhythmicity in the cell line.

Preparing input for prediction models

Samples.

All human genes (61 527 unique genes) from gene_info_04012020 were used to create the gene list.

Labels.

Labels (sleep genes) were manually curated by detailed literature review. The initial set of sleep genes was collected from a review paper [59]. Additional sleep genes were searched with the keyword "sleep" in title, and "gene" AND "model" in the main text from PubMed and Scopus databases. A sleep gene was defined as a gene that has been reported to alter sleep traits in at least one animal model (flies or mammals) by genetic approaches. Altered sleep traits include changes in sleep timing (sleep phase), sleep duration, and other measurements of sleep quality from EEG (e.g. slow wave activity, NREM/REM ratio, number of sleep bouts, and sleep latency). We divided the list into 3 tiers. Tier I include "bona fide" sleep genes that harbored a causal mutation in any human sleep traits and were validated in animal models. Their roles are conserved across species. Tier II genes have evidence from any

non-human mammalian model system. Tier III genes were discovered to change sleep traits in *Drosophila* but not in vertebrates yet. All sleep genes from Tier I, II, and III are weighted equally in the feature selections and model training process. Sleep genes were updated till 8/13/2020 for this analysis.

Features.

Sleep gene-associated functional features are used to represent the similarity of a gene to the known sleep gene. We built the sleep gene-associated features using two lines of information.

GENE SET COLLECTIONS.

Gene set collections were downloaded and prepared as mentioned above. One feature was created from each gene set collection. In each feature, genes were scored by the level of similarities (Jaccard Index, JI) to the list of curated sleep genes under the biological context. We applied a two-step process to calculate the gene-level score.

$$Jaccard\ Index\ (JI)_{term} = \frac{|A \cap B|}{|A \cup B|}$$

$$Score_{gene_X} = \sum_{i=1}^{n} JI_{term_i} \quad , \quad gene_x \text{ is an element of } term_{i \in \{1,2,3,\ldots n\}}$$

We first calculated JI for all terms in a gene set collection. Assuming that set A is the curated sleep genes and set B is the genes in an assigned term, JI was calculated using the number of overlapped genes between A and B divided by the number of unique genes in A and B. Next, the JI score for each term is added to the genes assigned to this term. To reduce over-contribution of redundant terms, we calculated JI and overlap coefficient (OC) for all pairs of terms in a gene set collection. Terms with JI > 0.5 or OC > 0.9 were grouped into a cluster. Only the maximum JI from each of these clusters was added to the genes. These two steps were repeated for every term in each gene set collection. By the end of these calculations, we obtained a numeric vector with the sum of JI score for each gene. This created a feature representing the overall similarities of a gene with the labeled sleep genes, under the biological context of the gene set collection information.

GENOME-WIDE PROFILING DATASETS.

Genome-wide profiling datasets were prepared as mentioned in the Data Curation and Preprocessing section. Evidence factors were used to evaluate the degree of sleep genes over-representation in each of these data metrics. To calculate the evidence factor, we first split the samples (genes), using the sleep genes to form a sleep gene distribution, and the remaining genes to form the non-sleep gene distribution. Evidence factors were calculated by comparing the proportion of genes in these two distributions, within a bin (between D_1 and D_2). As the distribution of genes was sparse, there were chances that a bin was empty. To solve this issue, we used binning by minimum percentage of genes. We first split the data metrics by 100 equal breakpoints. We repeatedly merged the neighboring bins until a bin reaches (1) at least 10% sleep genes and at least 1% non-sleep genes, or (2) at least 10% non-sleep genes and at least 1% sleep genes. For each (merged) bin, we calculated evidence factors by dividing the proportion of genes in sleep genes by the proportion of genes in non-sleep genes. We used maximum evidence factors (maxEF) from a dataset as an index to select sleep gene-relevant data. For each data metric, we set a cutoff of at least 25% labels that must present with a real value to ensure sufficient sleep genes used to form the sleep gene distribution; and the range of the data metrics must have more than 3 steps to ensure sufficient resolution. Else, we skipped the maxEF calculation for this data metric. Data metrics that show positive evidence (maxEF>=3) were selected and used as features to train the prediction models. To remove features that were highly correlated, we ran pairwise correlation coefficients of all data pairs. If two data metrics had a correlation coefficient higher than 0.8, the data metric with lower evidence factors were excluded.

Building machine learning models to predict sleep genes

Input.

Samples and features were prepared as mentioned above. We then filtered out samples (genes) with > 50% missing values. One of the labels (sleep gene), *NPSR1*, was removed. In summary, an input table with 17 853 rows (genes) and 91 columns (gene-associated features), with 108 labels (sleep gene) was used to build the prediction models.

Data Preprocessing.

Data preprocessing was done using Python—sklearn.impute and sklearn.preprocessing package. Missing values from the input were imputed with mean value and rescaled with standard score (z-score). Curated labels (sleep genes) were replaced by "1" and the remaining samples(genes) were replaced by "0".

Model Architecture.

Our model had 108 sleep genes (positive labels), but with no information or confidence on genes that do not regulate sleep (no negative labels). This raised the problem of learning from Positive and Unlabeled data (PU learning). We applied a biased learning method [60, 61] to solve this problem. To do this, all non-labeled genes were treated as negative labels during the training process, and prediction results were made based on ensembles of numerous of these weak classifiers. In detail, we first subsampled our samples (genes) into a smaller subset, with the same proportion of positive, and unlabeled samples in the training and prediction sets. Sleep genes were marked as positive labels and all other genes in the training sets were marked as negative labels. In this case, the negatively labeled samples were expected to contain a mixture of true or false negative labels, hence, resulting in weak classifiers. We repeated this process 100 times and made the final predictions based on average performance from all cycles.

We have only a small number of labels (sleep genes, $n = 108$) in comparison to other samples (non-labeled genes, $n = 17\ 745$). Yet, these labels were not found completely at random. Most sleep genes were identified based on our existing knowledge of sleep regulations. Therefore, we expected these genes not to be distributed randomly (or equally) in all sleep-relevant pathways. As an example, 13 out of 108 sleep genes were parts of the circadian clock pathway. To increase the randomness of subsampling labels, as well as maintaining the best performance, we trained the machine learning model with different proportions of training input, ranging from 0.2 to 0.8, with the same parameter. By doing this, we increased the combinations of samples used in the training and prediction sets, and therefore expected to have more robust predictions.

Prediction Models.

Eight supervised classifier algorithms were built to find the best-supervised classifier that fits our prediction. The evaluated

classifiers included probabilistic models (naive Bayes), linear regressions models (logistics and linear SVM), decision trees (decision tree, random forests, and adaptive boosting), and neural networks (neural networks and ensemble neural networks). All machine learning models, except neural network and ensemble neural network, were built using scikit-learn (v0.22.2) [62]. Neural networks and ensemble neural networks were built using Tensorflow (v2.2.0) [63] and Keras (v.2.4.3) (https://github.com/fchollet/keras).

Default parameters were used, except those mentioned below.

classifier algorithms	parameters
logistic regression	max_iter = 1000
decision tree	max_leaf_nodes = 12
random forest	max_leaf_nodes = 12
adaptive boosting	max_leaf_nodes = 4, algorithm="SAMME", *n*_estimator = 200
neural networks	sequential models, 2 hidden layers (12 & 6 nodes each, "relu" activation function), final output activated by "sigmoid" function.
ensemble neural networks	Repeatedly run neural networks 20 times, used major voting as cutoff for final output (>50%)

Input to train prediction models was prepared as mentioned above. For naive Bayes classifiers, we transformed the input with principal component analysis to ensure the conditional independence between features applied. The ratio of labeled and unlabeled samples was skewed. To balance the weights of positive and negative labels to roughly 1:1, we assumed the total number of genes was around 20 000. The weight was calculated by using half of the total number of genes (10 000) divided by the number of labels. Samples (genes) were randomly split into training and testing sets. Model architecture was as described above. Training set was used to train the model, whereas the testing set fit into the trained models. The classes predicted by the "predict_classes" function were used to calculate the confusion matrix.

Model Evaluation.

Models were evaluated with sensitivity (recall, *r*) and PU-adjusted F-measure (F-score^{PU}, $\frac{r^2}{Pr(\hat{Y}=1)}$). Raw prediction scores from "predict_proba" function, or the probability of a gene predicted as sleep gene, were recorded for genes assigned to the prediction set (not including linear SVM, as the raw prediction scores from linear SVM were discordant with the binary predictions). The average prediction score from all iterations was calculated and used to plot the sensitivity-precision plots.

Random Labels.

To avoid a model that makes predictions based on random noise, we removed all existing labels and randomly assigned the same number of labels to the remaining samples (only samples that were not originally labeled). As features were built based on sleep genes labels but not the random labels, the randomly assigned labels are not likely to be recalled using these sets of features, unless called by random noise. Therefore, we selected models that have the lowest sensitivity and F-score^{PU} with the random labels input.

Final Prediction Model.

The final prediction model was built with random forest classifiers using scikit-learn [62] (v0.22.2). Default parameters were used, except max_leaf_nodes are set to 12 based on the lowest out-of-bag (OOB) error and highest sensitivity. As described in the Model Evaluation section, the weights of the labels were calculated using 10 000 divided by the number of labels. Samples (genes) were randomly split into training and prediction sets. Model architecture was described above. For each iteration, the raw prediction score was recorded for genes assigned to the prediction set. A prediction score of less than 0.1 was set to 0 to reduce noise. The average prediction score from all iterations was calculated. Given the same maximum leaf node (max_leaf_nodes = 12), fewer genes were predicted as sleep genes when smaller training samples were used. For this reason, we weighted the final ranking of candidate sleep genes with the minimum training ratio that leads to a positive prediction (min[r]). The final prediction score is calculated as:

$$\textit{final prediction score} = 10 * \textit{min}(r) + \textit{average prediction score}$$

Exploring sleep traits GWAS data

Summary statistics from 4 self-reported UK Biobank sleep traits GWAS (n = 453 379), including chronotype [9], overall sleep durations [12], daytime sleepiness, [64] and insomnia [11], were downloaded from Sleep Disorder Knowledge Portal (SDKP). FUMA's SNP2GENE process [65] is used to run gene annotations. SNPs were mapped to genes using the posMap, eqtlMap, and ciMap methods, with default parameters. Mapped GWAS genes overlapped with the top predicted sleep genes were marked in Dataset S2.

Pathway enrichment analysis

The top 495 predicted genes were used for pathway enrichment analyses using DAVID [66]—Reactome pathway database. Pathways were then clustered using kappa similarity in DAVID (kappa similarity threshold > 0.5). We filtered out pathways with less than 5 genes or Bonferroni-adjusted p-value larger than .1.

Model perturbation analysis

Pathway enrichment analysis was run using the 109 known sleep genes in DAVID [66]. Five terms were selected, based on their similarity to the top predicted pathways of the original models, and with the largest number of sleep genes included. New prediction model was built for each of these five terms. Each new model was trained only with a subset of the sleep genes. Sleep genes enriched to the term were removed from labels. For example, the masked-BioRhythm model was trained using only 91 sleep genes; the 18 genes enriched to "biological rhythm" were labeled as unknown for the input. Changes in prediction scores between the new and the original models were calculated. The average changes in the prediction score (mean ΔPredScore), for genes in each pathway, excluding genes in the masked gene list, were used as the measurement to represent the level of alterations of the pathway.

Animal models

The R26-stop^{FL}Ikk2^{CA} transgenic mice (Stock No: 008242) and Camk2a^{CreER} transgenic mice (Stock No: 012362) were both obtained from The Jackson Laboratory. The Camk2a^{CreER} and R26-stop^{FL}Ikk2^{CA} mice were crossed and housed under 12h light/12h dark (LD) cycle within the University of Florida communicore facility and fed and watered ad libitum. Animal care

and experimental procedures were approved by the Institutional Animal Care and Use Committee at University of Florida following the Guide for Care and Use of Laboratory Animals of the National Institute of Health (IACUC# 202110057).

Tamoxifen Injection

For tamoxifen-inducible Ikk2^{ca} knock-in activation, Camk2a^{CreER}::R26-Stop^{FL}Ikk2^{ca} transgenic mice were generated by crossing Camk2a^{CreER} mice with R26-stop^{FL}Ikk2^{CA} mice. Tamoxifen (TAM) (#T5648; Sigma-Aldrich, St. Louis, MO) was dissolved in corn oil (#C8267, Sigma-Aldrich, St. Louis, MO) at a concentration of 20mg/mL. 10–12 weeks-old male mice ($n = 5$ for each group) were dosed at 75 mg/kg body weight (TAM or corn oil) intraperitoneally once every 24 h for a total of five consecutive days. The sleep assay began 4-weeks after tamoxifen injections when constitutively active IKK2 expression was induced in this model.

Western blot

Brain tissue lysate preparation and immunoblotting analysis were performed using anti-Flag (65, Sigma-Aldrich, St. Louis, MO) antibodies. Briefly, brain tissue was snapped frozen, and lysed in the RIPA lysis buffer containing cocktails of proteases inhibitors (Roche) and phosphatase inhibitors (Sigma). Western blot was performed to determine Flag-tagged Ikk2^{CA} activation.

Sleep assay

The piezoelectric sleep monitoring system (PiezoSleep version 2.11, Signal Solutions, Lexington, KY), is a highly sensitive, non-invasive, high throughput, and automated piezoelectric system, which detects breathing and gross body movements to characterize sleep patterns in unsupervised sleep/wake recordings [67, 68].

For each experiment, 11 tamoxifen or 9 corn-oil-injected Camk2a^{CreER}::R26-stop^{FL}Ikk2^{CA} mice were individually housed in PiezoSleep cages with a sensor inside a temperature, humidity, and light-controlled box. The first 3–5 days of recording were considered as the acclimation period to the piezo device. The 12h light/12h dark (LD) cycle (light on at 07:00 to 19:00; 250 lux) was performed for the 15-days LD followed by the next 15-days of 12h dark/12h dark (DD) with ad libitum access to food, water, and nesting material. Sleep data were analyzed for multiple sleep traits of individual mice using sleepstats2p18 (Signal Solutions, Lexington, KY).

Results

Defining sleep gene features

Our goal was to build a machine-learning model to predict sleep-regulating genes based on functional features of known sleep genes. The hypothesis is that genes with similar functions to known sleep genes are more likely to play a role in sleep regulation. As a first step, we manually curated a list of known sleep genes (Supplementary Table S1, hereafter referred to as "sleep genes") through literature mining from the PubMed and Scopus databases. Sleep genes were defined as genes reported to alter sleep traits, including sleep timing, sleep duration, and measurements of sleep quality from EEG, in at least one animal model (flies or mammals).

Next, we identified functional features associated with these sleep genes. Features in machine learning models are the measurable variables that are useful to discriminate the characterized properties, in this case, to classify sleep genes from non-sleep genes. The lack of a strong molecular understanding of sleep regulation makes it difficult to know what information is useful in predicting sleep genes. To address this issue, we built sleep-associated features based on the sleep genes we have curated, using two sources of information. The first source includes gene and protein knowledge from annotated gene set collections, including canonical pathways, gene ontology, transcription factor target genes, and protein–protein interactions. We applied the Jaccard index (JI), or the Jaccard similarity coefficient [69], to quantify the similarity of a gene to the exemplar sleep genes in the context of a given gene set collection (Figure 1A). Using the JI scoring method, we generated 19 features representing the similarity of a gene to sleep genes in various biological contexts (Dataset S1).

The second source of information we used to define sleep gene-associated functional features includes genome-wide profiling datasets. We used evidence factors [70, 71] to select genome-wide datasets most likely to be informative for the prediction model. In prior work, we applied evidence factors to identify a novel circadian transcriptional repressor in mice [36]. We modified the application here to screen for datasets that show evidence of sleep genes (Figure 1B). To validate the concept, we tested three datasets: (1) a time-series transcriptomics profile of mouse suprachiasmatic nucleus (SCN) across a 48h time-span (GSE70392), (2) a transcriptomics profile of mouse cortex after sleep deprivation (GSE114845), and (3) a time-series transcriptomics profile of HeLa cells at different cell-cycle stages (GSE26922). We expected datasets (1) and (2) to show positive evidence for sleep genes, as the two-process model suggests roles for circadian rhythm (process C) and sleep homeostasis (Process S) in sleep regulation [21]. Dataset (3) was selected as a negative control because the cell-cycle stage is not expected to be predictive of sleep.

For the two time-series datasets, each gene is assigned a significance score for rhythmic expression using the published -log2 transformed p-value. For the sleep deprivation dataset, each gene is assigned a differential expressed score using the log2 transformed absolute fold change. For each dataset, we built two distributions using this score. The 109 known sleep genes are used to form a sleep gene distribution. All remaining genes are used to form a non-sleep gene distribution. The evidence factors are computed by comparing the proportion of genes in these two distributions. If the two distributions are similar, the maximum evidence factors (maxEF) are close to 1, which would indicate that there is no sleep gene over-representation in the dataset. In contrast, if the sleep gene distribution is different from the non-sleep gene distribution, maxEF would be much greater than 1. Evidence factors greater than 3 suggest positive evidence [70]. Therefore, a cutoff of maxEF larger than 3 was selected as an indicator of sleep gene over-representation. As expected, we found no evidence of sleep gene over-representation in the cell-cycle time-series dataset (maxEF = 1.3). Conversely, genes rhythmically expressed in mouse SCN or genes with expression altered after sleep deprivation in mouse cortex are more likely to be sleep genes (maxEFs are 4.9 and 4.7 respectively), suggesting that circadian expression and sleep homeostasis are sleep-gene-associated features and should be incorporated in our machine learning model (Figure 1B).

Using this method, we screened through 7195 genome-wide datasets and found 94 of them with positive evidence for sleep (maxEF > 3) (Supplementary Figure S1, Dataset S1). Datasets with the highest maxEF included circadian expression of genes in multiple tissues and altered gene expression in several brain diseases. Sleep genes were also over-represented in datasets pertaining to Epstein-Barr viral (EBV) infection, IL17A knockout in colon, clozapine (a hypnotic) treatment, sex or age differences, and anatomically specific datasets, including testis and human brain (Figure 1C).

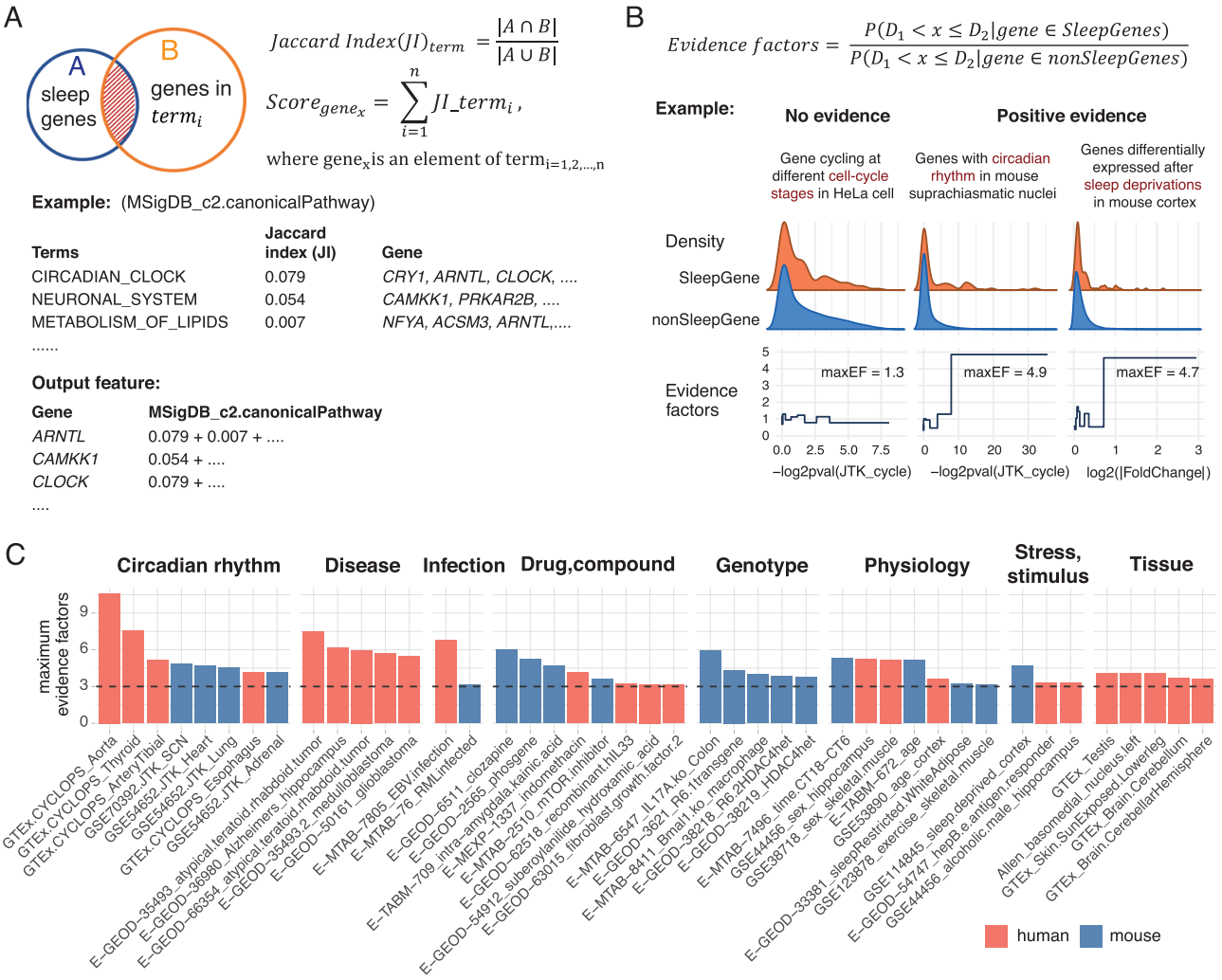


Figure 1. Defining sleep gene-associated features. Two lines of information are used to define the sleep gene-associated functional features. (A) Gene and protein knowledge from annotated gene set collections. An example of how Jaccard index is used to create a feature that represents the similarity of a gene to sleep genes under the biological context of the canonical pathways. (B) Genome-wide datasets. Evidence factors are used to screen for datasets that show an over-representation of sleep genes. Three transcriptomics datasets are shown as negative (cell-cycle stages) and positive controls (circadian rhythm and sleep deprivation) for this screening method. (C) Genome-wide datasets enriched for sleep genes. The top eight datasets with maxEF larger than three in each group (if available) are shown in the figure. Y-axis of the bar plots shows the maximum evidence factors for each dataset. Datasets from human and mouse samples are colored in red and blue, respectively.

Despite starting with over 7000 datasets, informative ones were limited to less than 100. We were particularly interested in the datasets from transcriptional studies that could inform sleep (Supplementary Figure S2). These datasets tell us the most about sleep-impacting pathways and genes. For example, acute viral or bacterial infections are known to cause sleepiness [72, 73]. We found datasets associated with EBV, SARS, and *Salmonella* were all enriched with sleep-regulating genes (Supplementary Figure S2). Other informative datasets include brain cancer, aging, hypnotic drugs, and exercise; which align with prior knowledge of factors that impact sleep patterns. These results support that the screening method enables us to identify biological factors that impact sleep, without making presumptions. Alterations in pathways or genes shared by these datasets enable testing their role in sleep regulation.

Selecting a classification algorithm for the prediction models

We selected 94 genome-wide datasets that are informative to predict sleep (maxEF > 3). We did pairwise correlation to filter out 23 datasets that are highly correlated (Pearson correlation > 0.8) (Supplementary Figure S3). The remaining 72 datasets and the 19 features from the JI scoring method are used to represent the sleep gene features. Genes with > 50% missing values from these 91 features were filtered out. One of the sleep genes, *NPSR1*, was excluded as it had missing values in more than half of the selected datasets. With this information, we generated an input table with 17 853 genes (samples), including 108 sleep genes (labels), and 91 features for training the machine learning models.

We applied a biased learning method to solve the problem of learning from Positive and Unlabeled(PU) data (Supplementary Figure S4), given that we have positive labels (sleep genes) but with no known negative labels (information or confidence on which genes do not regulate sleep) [60, 61]. We evaluated eight supervised classifiers seeking a model to maximize PU-adjusted F measures (F-score^{PU}) [61, 74]. The tested classification models included probabilistic (naive Bayes), linear regression (logistic and linear support vector machines), decision tree-based (decision tree, random forests, and adaptive boosting), and neural networks

(neural networks and ensemble neural networks). Models were built using Python packages scikit-learn [62] and Tensorflow [63].

Random forests had the highest F-score^{PU} (24.85) and second highest AUC (0.9783) in the sleep genes trained model. For purposes of follow-up biological studies, false positives are worse than false negatives, as the validation experiments are time-consuming and costly. As part of our evaluation, we re-trained all models with randomly-shuffled labels as inputs to minimize the chance that predictions are made based on noise. In these tests, random forests consistently outperformed other classifiers with the lowest sensitivity (0) and F-score^{PU} (0.007), suggesting that random forest predictions have the lowest false positive rates (Supplementary Figure S5).

Identify novel genes and pathways relevant to sleep regulation

We trained our prediction models using random forests as the classification algorithm. All genes were ranked based on the final prediction score. In total, 3373 out of 17 853 genes were predicted as sleep genes with 95% of the known sleep genes being recalled (Figure 2A& Supplementary Figure S6, Dataset S2). Overall, sleep genes identified from human samples ranked higher in comparison to sleep genes identified from other mammals or flies, despite the fact that all sleep genes were weighted equally during the feature selection and model training steps. This suggested that our models' predictions can detect human sleep genes and provide strong candidates for future study.

We intended to build a prediction model to reveal molecular mechanisms or pathways that may be involved in the regulation of the sleep-wake cycle. For this purpose, we use the optimal cutoff of selecting the top 495 genes for enrichment analysis (Figure 2A). The cutoff was determined based on a compromise of the highest sensitivity and the least number of novel genes being predicted as sleep genes. Of these top 495 genes, 86 were known sleep genes, and 409 were novel predicted sleep-regulating genes. We found 64 out of the 409 novel sleep genes are annotated genes from four GWAS studies pertaining to chronotype [9], overall sleep duration [12], insomnia, [11] and daytime sleepiness [64](Supplementary Figure S7).

Pathways enrichment analysis was run using DAVID [66] (Reactome). We identified 7 pathway clusters with at least 3 genes overlapping with the GWAS annotated genes (Figure 2B, 2C& Supplementary Figure S8). The top enriched pathways included those related to neuron activity, Phase 0 depolarization, ion homeostasis, and Ca^{2+} signaling. These findings are in line with a number of studies that reported the involvement of the neuronal synapse in the transition between sleep-wake states [23, 75–77].

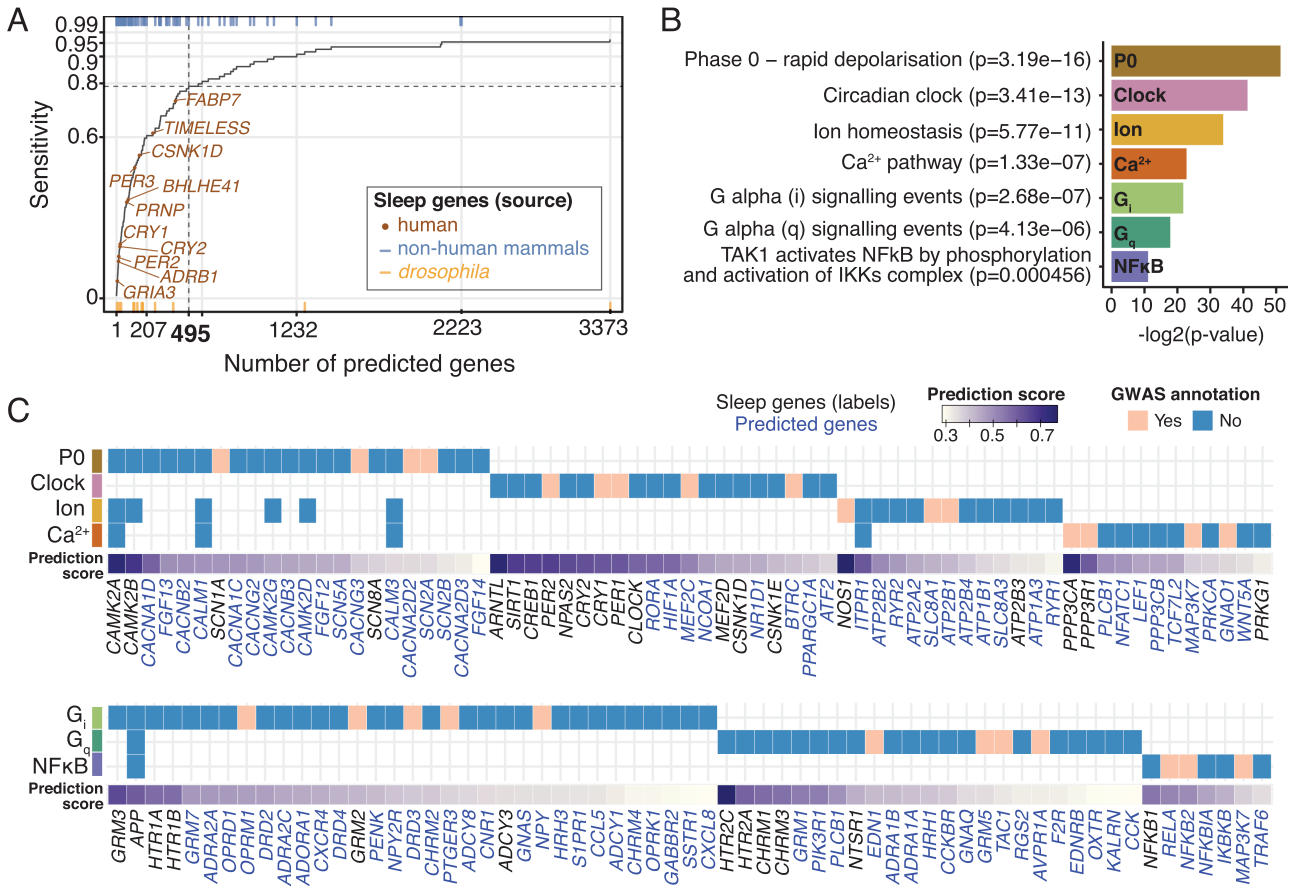


Figure 2. Predictions of novel sleep genes using random forests model. (A) Prediction result of the random forests model. X-axis shows the number of predicted genes. Y-axis shows the proportion of recall sleep genes (sensitivity). Sleep genes with evidence supported by humans are labeled in brown; sleep genes with evidence supported by non-human mammals and drosophila are marked by blue (top) and yellow (bottom) rugs, respectively. (B) Top enriched pathways with at least three genes overlapping with sleep traits GWAS are shown. (C) Genes in the top enriched pathway. Predicted genes overlapped to sleep traits GWAS are colored in orange and those that did not overlap are colored in blue. Sleep genes that train the machine learning model are labeled in black and the novel predicted sleep genes are labeled in blue. Within each pathway, genes are ordered by their predicted rankings, from left to right.

Previous GWAS for sleep traits have reported enrichment of circadian rhythm and G-protein-relevant pathways [9, 12]. Accordingly, the circadian clock and G alpha signaling (G_i and G_q) are among the top enriched pathways from our prediction models. Interestingly, *OPRD1* and *OPRM1* encoding two opioid receptors, and *PENK* encoding an endogenous opioid peptide, are among the top candidate genes in G_i signaling. Opioids are well-known sedatives. Clinical studies have shown that opioid medication impacts sleep architecture in healthy adults [78]. Our predictions suggest that among the three opioid receptors, the mu-, and delta- receptors are more likely to play key roles in sleep regulation at the molecular level. This is in agreement with an in vivo study using opioid receptor agonists in feline models [79].

Validation of a role for NF-κB in sleep regulation

Inflammation has been linked to sleep quality in healthy individuals without sleep disorders [80]. Our top predictions include a group of immune-related genes, and, in particular, the components of the NF-κB pathway (Figure 2C). NF-κB transcription factors play critical roles in inflammation and immunity, as well as cell proliferation, differentiation, and survival [81]. *RELA* is an NF-κB subunit and *IKBKB* (aka *IKKβ* and *IKK2*) is an upstream activator of NF-κB activation. The direct and indirect triggers of NF-κB activation have been reported to cause circadian disruption [82, 83]. Sleep loss alters immune function, and immune challenges alter sleep at least in part through regulation of several endogenous somnogens [84]. Previous studies reported that *Nfkb1*(p50) knockout mice showed increased duration of slow-wave and rapid eye movement (REM) sleep [85]. However, little is known about the direct effect of NF-κB activation on sleep.

We used Camk2a^{CreER}::R26-stop^{FL}Ikk2^{CA} mice to test the effect of NF-κB activation on sleep (Supplementary Figure S9). In this mouse model, the stop^{FL} cassette prevents expression of the constitutively active *Ikk2* (Ikk2^{CA}) [86]. Tamoxifen-inducible neuronal specific Camk2a^{CreER} recombinase [87] induces deletion of the stop^{FL} and expression of Ikk2^{CA}. IKK2 is a key component of the IKK complex that phosphorylates IkBα, leading to IkBα ubiquitination and proteasomal degradation [88]. Upon degradation of IkBα, NF-κB is free to translocate to the nucleus to induce transcription of target genes. Therefore, Ikk2^{CA} expression leads to constitutive NF-κB activation, and these mice represent a genetic model of chronic NF-κB activation. In a recent study, we used this model to assess the NF-κB effect on circadian behavioral rhythms [83].

Here, we used the piezoelectric sleep monitoring system (piezo) [67, 68] to assess sleep-wake phenotypes in mice with chronic NF-κB activation. The gold standard to determine sleep-wake states is electroencephalogram (EEG) and electromyogram (EMG), which are based on the brain and muscle activities. In contrast, piezo determines sleep-wake states based on breathing regularity. Statistical validation of these two techniques (EEG/EMG vs piezo) have shown a strong and significant positive correlation in measuring distribution and amount of total sleep time. The number of brief awakenings and short sleep episodes was higher when counted with the piezo system. However, measurements between the two techniques are positively correlated [68]. Therefore, the use of the piezo system to compare short and long sleep bouts between the two conditions is a valid, and convenient, initial in vivo approach to evaluate the effect of candidate sleep genes.

We show that, compared to control mice, Ikk2^{CA} mice had a reduced total sleep duration (Figure 3A). The reduction of sleep duration was only observed in the light (sleeping) phase but not in the dark (activity) phase. Sleep bout duration has been used as an indicator of sleep consolidation versus fragmentation [89]. Compared to controls, Ikk2^{CA} mice displayed increased sleep at shorter bout lengths and decreased sleep at longer bout lengths (Figure 3B), indicative of sleep fragmentation. Taken together, constitutive NF-κB pathway activation led to increased sleep fragmentation during the sleep phase.

Network of sleep regulation pathways

Our models predicted seven pathways enriched for sleep regulation, including the circadian clock, NF-κB, G-protein signaling, and multiple pathways involving neuronal activities (Phase 0—depolarization, ion homeostasis, and Ca^{2+} pathway) (Figure 2B). Sleep-wake transition is known to be achieved at the neuronal level. It is not clear how pathways not directly regulating neuronal activity (eg. NF-κB) impact sleep. We would like to explore if and how these pathways function in a network to regulate sleep through in-silico perturbation. The rationale here is that if pathway A has strong interaction with pathway B, sleep genes in pathway A will strongly influence the prediction of genes in pathway B. By this reasoning, when sleep genes in pathway A were removed from positive labels during the model training process, we would expect to observe large changes in the prediction score for genes in pathway B. Similar mask learning approach is established in explaining black box models, for example, to find the part of an image most responsible for a classifier decision [90, 91]. We implement the concept here to explain the relatedness of the top enriched pathways by sequentially removing them from the model, and evaluating the effects of these have on the detection effects of the other categories.

To do this, we first ran an enrichment analysis using the 109 known sleep genes and selected 5 terms, including genes enriched for the keywords "transport" (Transport), "biological rhythms" (BioRhythm), "G-protein coupled receptor" (GPCR), "calcium signaling pathway" (Calcium), and "inflammatory response" (Inflammation) (Supplementary Figure S10). These five terms were selected based on their similarity to the predicted pathways in Figure 2B and having the largest number of known sleep genes included. We built new prediction models for each of these five terms, using only sleep genes (positive labels) excluding the representing genes enriched to that term. For example, the masked-BioRhythm model was trained using only 91 sleep genes; the 18 genes enriched to "biological rhythm" (Supplementary Figure S10) were removed. Finally, we calculated the changes in prediction scores between the new and the initial models. Average changes in the prediction score (mean ΔPredScore) were used as the measurement to represent the level of alterations of a pathway.

Our approach revealed previously unappreciated relationships between the clock and NF-κB and Ca^{2+} signaling. Likewise, there is a strong relationship between the NF-κB and Ca^{2+} pathways. Not surprisingly, strong relationships between G-protein, Phase 0—depolarization, and ion channel signaling were observed. Ca^{2+} signaling was found as the key node connecting the circadian clock and NF-κB to these neuronal-related pathways (Figure 4A).

We further evaluated the interactions of the circadian clock, Ca^{2+} and NF-κB pathways at the gene level (Figure 4B). Among the top predicted genes in the NF-κB pathways, *RELA* showed the largest alteration in the masked-BioRhythm model, suggesting *RELA* as the key point of the interaction between the NF-κB and the clock pathways. This prediction aligns with the recent finding that RELA directly binds to BMAL1 at the CRY1 binding site and results in E-box transcriptional repression [83]. Among the top predicted genes in the Ca^{2+} pathways,

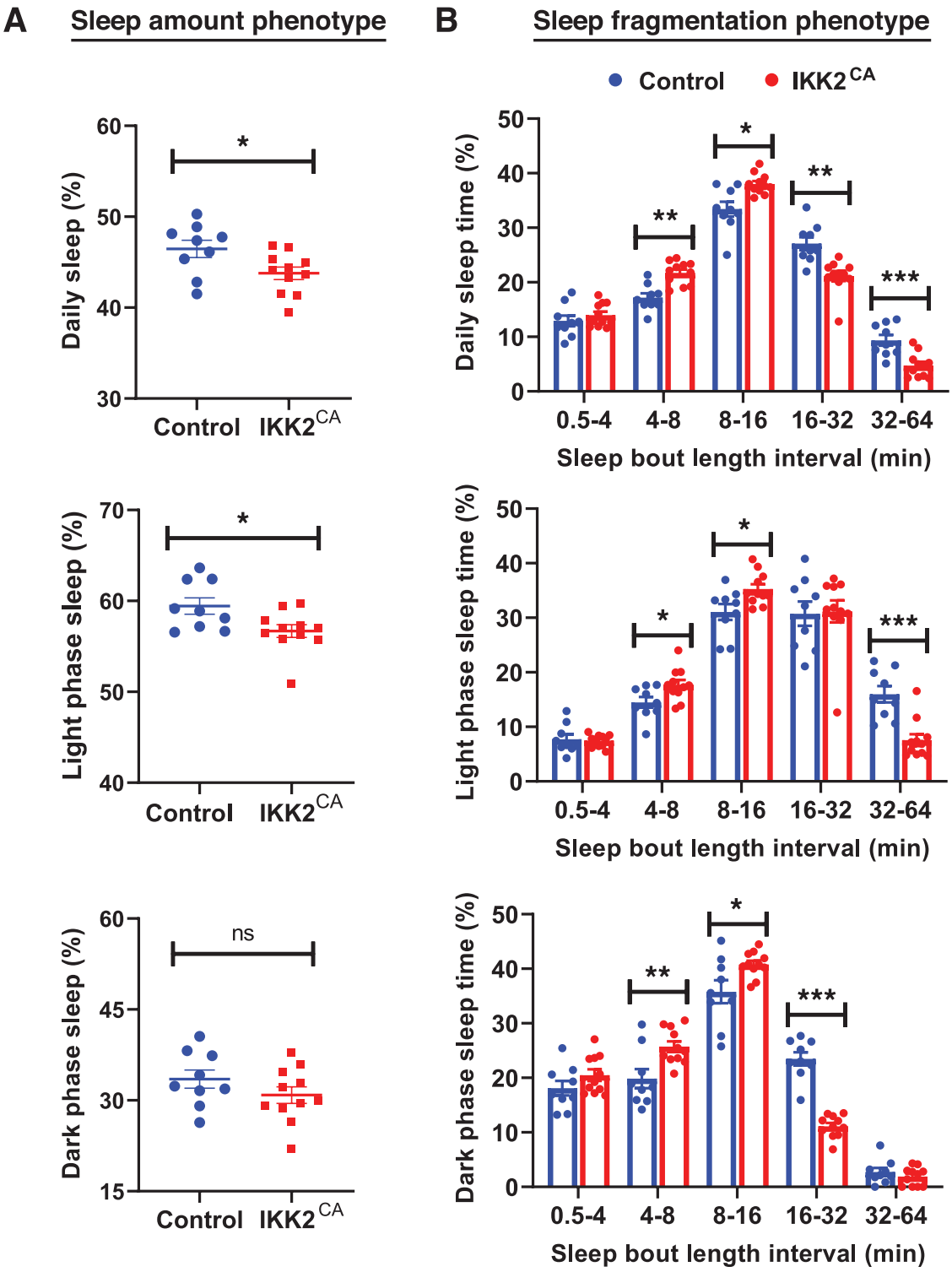


Figure 3. Sleep phenotyping in Ikk2^{CA} mice and control using PiezoSleep system. (A) Sleep amount phenotype. Percentages of total sleep (top), light phase sleep (middle), and dark phase sleep (bottom) between Ikk2^{CA} and control mice, recorded in LD 12:12, are shown in the figure. (B) Sleep fragmentation phenotype. Percentage of total sleep (top), light phase sleep (middle), and dark phase sleep (bottom) for different sleep bout length intervals, between Ikk2^{CA} and control mice, are shown in the figure. n = 9 for control and *n* = 11 for Ikk2^{CA} mice. All results are shown as mean ± SEM. *P*-value *<.05, **<.01, ***<.001, ns—not significant (Student's *t*-test).

LEF1, *TCF7L2*, and *PLCB1* are found as the top altered genes in the masked-BioRhythm model, whereas *TCF7L2*, *MAP3K7*, and *WNT5A* are the top altered genes in the masked-Inflammation model. Interestingly, these five genes were annotated to the Wnt/Ca^{2+} signaling pathway. *TCF7L2* is the gene with alteration observed in both masked-BioRhythm and masked-Inflammation models, suggesting that it might play a network-level connection to both the clock and inflammation pathways, making it a particularly attractive candidate gene for follow-up functional studies.

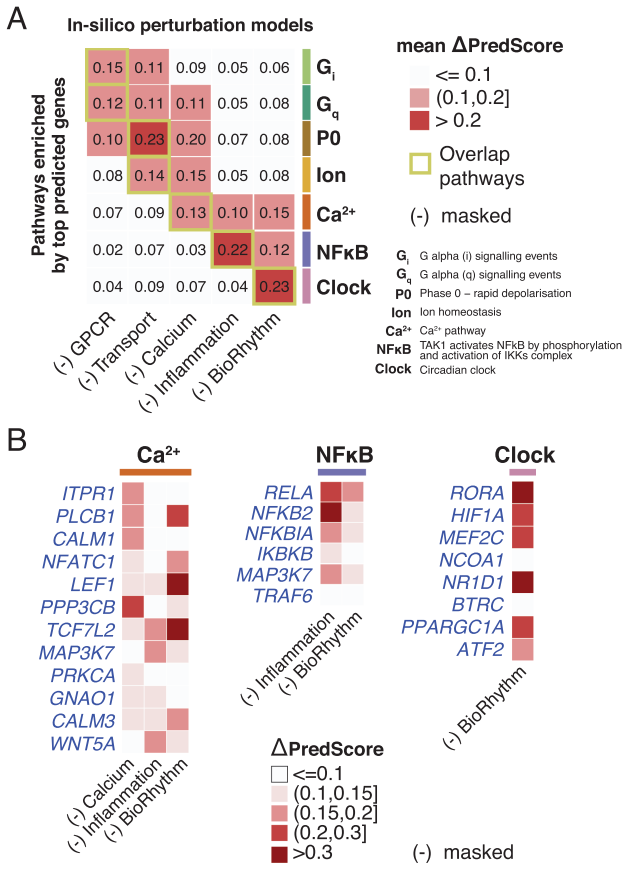


Figure 4. Perturbation analysis to identify sleep regulatory networks. (A) Heatmap represents the relatedness between top enriched pathways. X-axis represents the newly trained models by masking the genes linked to relevant terms from positive labels. Y-axis represents the enriched pathways from top predicted genes in Figure 2B. Mean ΔPredScore is the average change of the prediction scores between the original and the newly trained model, using only the novel predicted genes in each enriched pathway. (B) Gene-level changes of prediction score for Ca^{2+} (left), NF-κB (middle), and circadian clock (right) pathways. X-axis shows the newly trained models by masking the genes linked to relevant terms from positive labels. Y-axis shows the novel predicted genes for each pathway. Changes in prediction score for each gene between the original and newly trained model are shown in the heatmap.

Links between the circadian clock and sleep

Circadian rhythms are known to play a major role in sleep regulation. We applied the masked-BioRhythm model to explore how well-established circadian rhythm-related genes impact overall predictions at the genome-wide level (Supplementary Figure S11). As expected, most core-clock genes (e.g. *ARNTL*, *CLOCK*, *NPAS2*, *CRY1*, *CRY2*, and *PER3*) were absent in the masked-BioRhythm model (prediction score = 0). *PER1* and *PER2* are exceptions, which were still predicted at the 90-95% sensitivity range, suggesting that PERs play direct roles in regulating sleep.

TWIST1 is one of the non-clock genes shown to have a large alteration of prediction score in the masked-BioRhythm model, suggesting strong functional relationship to the clock (Supplementary Figure S11). *TWIST1* is a member of the basic helix-loop-helix (bHLH) transcription factor family that forms homo- or hetero-dimers and binds the same E-box target sites that BMAL1/CLOCK binds [92]. In mice, *Twist1* and *Twist2* are induced by TNFα through the canonical IKKβ-dependent NF-κB signaling pathway [93, 94]. Conversely, TWIST proteins inhibit cytokine genes through physical interactions with the RELA subunit of NF-κB complex at the transactivation domain [95, 96]. *Twist1*, but not *Twist2*, is found to mediate the TNF-induced repression of *Per* and *Dbp* genes in vitro through competing for E-box binding of BMAL1/CLOCK [97]. The impact of TWIST1 on sleep has not been reported. However, *Twist1* activation was observed in mouse models of Huntington's Diseases and found functionally important for mutant *Huntingtin*-induced neurotoxicity [98]. Common sleep disorders are reported in Huntington's disease patients, including insomnia, increased sleep onset latency, decrease in total sleep time, and frequent nocturnal arousals [99]. Taken together, TWIST1 may play a role in regulating sleep by acting downstream of TNFα-induced NF-κB activation and possibly through regulation of the circadian clock. This indicates *Twist1* may act similarly to *Dec2*, a known sleep regulator, in humans [4, 100].

Discussion

The immune system has a known role in sleep regulation. Sleep changes in response to infection, and inflammatory mediators such as IL-1, TNF, and prostaglandins have sleep-regulatory properties [101]. Recent findings from mouse models suggested mutual regulation between NF-κB and circadian rhythm pathways [83]. Our machine learning model suggested that the NF-κB activation, specifically through the phosphorylation of the IkBα complex, is a key regulator of sleep. We validated one of the predicted genes (*Ikbkb*) using a neuron-specific, constitutively activated IKK2 ($Ikk2^{CA}$) mouse model and found that $Ikk2^{CA}$ mice have reduced sleep duration and more fragmented sleep compared to controls. The decrease in bout length and reduced sleep duration during the inactivity (sleep) phase suggests disruption in sleep consolidation and increased sleep fragmentation. Sleep perturbations including fragmented sleep with frequent nighttime awakenings and excessive daytime sleepiness are common in humans, especially those with neurodegenerative diseases or cancer, and these sleep disruptions are a comorbidity for many sleep disorders [102, 103].

Machine learning has been widely applied to integrate biological data in recent years. Multiple studies reported the use of gene prioritization tools [33, 34, 104], but most are built on the hypothesis that causal variants or driver genes and pathways exist and thus may not be ideal for the understanding genetic regulation in complex traits. We sought to identify candidate sleep genes that share similar functional features to our defined set of known sleep genes. The carefully selected features are well-representing sleep genes, and therefore, these features are useful to identify genes that are similar to the sleep genes. In addition to our validation of the NF-κB pathway, a few of the top candidate genes, *Mef2c* [105], *GRM1* [106], and *Tac1* [107] were recently independently validated.

A key to our approach is the ability to define a comprehensive yet predictive set of features. In addition to the annotation resources (e.g. GO terms, MSigDB, and GWAS catalog), which are commonly used in other gene prioritization tools [30, 33], we applied a modified probabilistic regression method to screen and select sleep-relevant features. This step allowed us to include only a small but informative set of features from thousands of genome-scale studies, with the ability to identify unappreciated factors that might impact sleep. Although the model was initially built on a common understanding of known sleep genes, it also incorporates "hidden" information about these sleep genes that we didn't necessarily know a priori. As an example, of the 7000 + datasets, we found several sleep-related

factors we expected (e.g. EBV infection, clozapine), and dozens we didn't previously appreciate were related to sleep (e.g. polycystic ovary syndrome, suberoylanilide hydroxamic acid). A similar framework is applicable for integrating large amounts of genome-scale data to predict genetic regulators of many other complex traits.

Model evaluation results indicated that incorporated ensemble neural networks (F-score^{PU} = 24.71) perform much better than the single-run neural network (F-score^{PU} = 12.36). This suggested that bootstrap aggregation (bagging) methods significantly reduced false positive predictions. We have chosen random forests based on their performance and high efficiency, however, ensemble neural networks have higher sensitivity and might be useful to extend the candidate gene list.

We applied perturbation approaches, similar to the concept of a masked learning, to explore relatedness between top predicted sleep-regulating genes and pathways. The set of genes being masked in each model was chosen based on biological knowledge (pathway), but not permutation, for computational efficiency. Strategies to systematically mask or add gene(s) to the model, without prior assumptions, would likely help to infer an under-appreciated connection between genes involved in sleep regulation. Nonetheless, we would like to emphasize that our proposed method is able to reveal relatedness of pathways, or to prioritize possible genes as a key nodes between pathways. However, we note that these results provide no information about causal-response relationships.

Predictions of this machine learning model are limited to the availability and quality of the training labels (e.g. sleep genes), the relevance of features to labels, and the amount of information available per sample (e.g. genes). Missing information reduces performance of the models. For example, the expression of *NPSR1* is low or unmeasurable in most genome-wide studies. Therefore, despite evidence from human studies that *NPSR1* is a sleep gene, it is not likely to be identified by our prediction model nor useful to improve model predictions. New information, whenever available, will improve performance of the models.

This model is built on existing knowledge regarding the genetics of sleep. Therefore, other suitable uses of this approach include complex diseases and behaviors that have sufficient genetic knowledge but lack system-level understanding. To facilitate the use of these approaches, we have made the code available on GitHub https://github.com/yyenglee/ml-sleep. All necessary inputs are available on figshare https://doi.org/10.6084/m9.figshare.20517951.

Supplementary Material

Supplementary material is available at *SLEEP* online.

Acknowledgments

We thank Andrew I. Su and Casey S. Greene for technical assistance; Ying-Hui Fu, Leah C. Kottyan, Christian I. Hong, Tongli Zhang, Rochelle M. Witt, and Jiffin K. Paulose for thoughtful discussions.

Funding

This work was supported by the National Institute of Neurological Disorders and Stroke (5R01NS054794-13 to J.B.H. and A.C.L.).

Author Contributions

Y.Y.L., G.W., M.D.R., L.J.F., R.C.A., D.F.S., A.C.L., J.B.H. designed the research; Y.Y.L., G.W., L.J.F., N.Y.C. collected and preprocessed data; Y.Y.L., G.W., M.D.R., R.C.A., J.B.H. contributed the analytic tools; M.E. and A.R.M. performed experiments and interpreted data; Y.Y.L., M.E., G.W., M.D.R., L.J.F., A.C.L., J.B.H. wrote the manuscript.

Data and materials availability

The data underlying this article are available as mentioned below. Gene set collections is available at MSigDB http://www.gsea-msigdb.org/gsea/msigdb/index.jsp, harmonizome https://maayanlab.cloud/Harmonizome/, protein–protein interactions information is available at BioGRID https://downloads.thebiogrid.org; tissue-specific expression data is available at BioGPS http://biogps.org/downloads/, GTEx https://gtexportal.org/home/datasets, HPA https://www.proteinatlas.org/about/download, Allen Brain Map https://human.brain-map.org; protein abundances measurements is available at http://www.humanproteomemap.org/index.php; circadian expression is available at http://circadb.hogeneschlab.org; transcriptomics profile is available at GEO https://www.ncbi.nlm.nih.gov/geo/, EBI ArrayExpress https://www.ebi.ac.uk/arrayexpress/; phosphorylation sites information is available at http://qphos.cancerbio.info; statistical summary from sleep traits GWAS is available at https://sleep.hugeamp.org; GWAS gene annotation is run using webtools FUMA https://fuma.ctglab.nl; pathway enrichment analysis is run using https://david.ncifcrf.gov. Code and pipelines for the prediction models can be found at https://github.com/yyenglee/ml-sleep. Necessary inputs and references can be found at https://doi.org/10.6084/m9.figshare.20517951.

Disclosure Statement

The authors have no other interests that could be perceived as conflicts of interest to declare. This manuscript was deposited on bioRxiv: https://www.biorxiv.org/content/10.1101/2021.04.10.439249v2.full.

References

1. Medori R, *et al*. Fatal familial insomnia, a prion disease with a mutation at codon 178 of the prion protein gene. *N Engl J Med*. 1992;**326**(7):444–449. doi: 10.1056/NEJM199202133260704.
2. Toh KL, *et al*. An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome. *Science*. 2001;**291**(5506):1040–1043. doi: 10.1126/science.1057499.
3. Xu Y, *et al*. Functional consequences of a CKIdelta mutation causing familial advanced sleep phase syndrome. *Nature*. 2005;**434**(7033):640–644. doi: 10.1038/nature03453.
4. He Y, *et al*. The transcriptional repressor DEC2 regulates sleep length in mammals. *Science*. 2009;**325**(5942):866–870. doi: 10.1126/science.1174443.
5. Zhang L, *et al*. A PERIOD3 variant causes a circadian phenotype and is associated with a seasonal mood trait. *Proc Natl Acad Sci USA*. 2016;**113**(11):E1536–E1544. doi: 10.1073/pnas.1600039113.
6. Patke A, *et al*. Mutation of the human circadian clock gene CRY1 in familial delayed sleep phase Disorder. *Cell*. 2017;**169**(2):203–215.e13. doi: 10.1016/j.cell.2017.03.027.
7. Xing L, *et al*. Mutant neuropeptide S receptor reduces sleep duration with preserved memory consolidation. *Sci Transl Med*. 2019;**11**(514):eaax2014. doi: 10.1126/scitranslmed.aax2014.

8. Shi G, *et al*. A rare mutation of β1-adrenergic receptor affects sleep/wake behaviors. *Neuron*. 2019;**103**(6):1044–1055.e7. doi: 10.1016/j.neuron.2019.07.026.
9. Jones SE, *et al*. Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms. *Nat Commun*. 2019;**10**(1):343. doi: 10.1038/s41467-018-08259-7.
10. Jansen PR, *et al*.; 23andMe Research Team. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat Genet*. 2019;**51**(3):394–403. doi: 10.1038/s41588-018-0333-3.
11. Lane JM, *et al*.; HUNT All In Sleep. Biological and clinical insights from genetics of insomnia symptoms. *Nat Genet*. 2019;**51**(3):387–393. doi: 10.1038/s41588-019-0361-7.
12. Dashti HS, *et al*. Genome-wide association study identifies genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates. *Nat Commun*. 2019;**10**(1):1100. doi: 10.1038/s41467-019-08917-4.
13. Sehgal A, *et al*. Genetics of sleep and sleep disorders. *Cell*. 2011;**146**(2):194–207. doi: 10.1016/j.cell.2011.07.004.
14. Axelrod S, Saez L, Young MW. Chapter one - studying circadian rhythm and sleep using genetic screens in Drosophila. In: Sehgal A, ed. *Methods in Enzymology*. Vol **551**. Cambridge, MA: Academic Press; 2015:3–27. doi: 10.1016/bs.mie.2014.10.026.
15. Cirelli C, *et al*. Reduced sleep in Drosophila Shaker mutants. *Nature*. 2005;**434**(7037):1087–1092. doi: 10.1038/nature03486.
16. Koh K, *et al*. Identification of SLEEPLESS, a sleep-promoting factor. *Science*. 2008;**321**(5887):372–376. doi: 10.1126/science.1155942.
17. Harbison ST, *et al*. Selection for long and short sleep duration in Drosophila melanogaster reveals the complex genetic network underlying natural variation in sleep. *PLoS Genet*. 2017;**13**(12):e1007098e1007098. doi: 10.1371/journal.pgen.1007098.
18. Toda H, *et al*. A sleep-inducing gene, nemuri, links sleep and immune function in Drosophila. *Science*. 2019;**363**(6426):509–515. doi: 10.1126/science.aat1650.
19. Funato H, *et al*. Forward-genetics analysis of sleep in randomly mutagenized mice. *Nature*. 2016;**539**(7629):378–383. doi: 10.1038/nature20142.
20. Banks GT, *et al*. Forward genetics identifies a novel sleep mutant with sleep state inertia and REM sleep deficits. *Sci Adv*. 2020;**6**(33):eabb3567. doi: 10.1126/sciadv.abb3567.
21. Borbély AA, *et al*. The two-process model of sleep regulation: a reappraisal. *J Sleep Res*. 2016;**25**(2):131–143. doi: 10.1111/jsr.12371.
22. Diessler S, *et al*. A systems genetics resource and analysis of sleep regulation in the mouse. *PLoS Biol*. 2018;**16**(8):e2005750. doi: 10.1371/journal.pbio.2005750.
23. Noya SB, *et al*. The forebrain synaptic transcriptome is organized by clocks but its proteome is driven by sleep. *Science*. 2019;**366**(6462):eaav2642. doi: 10.1126/science.aav2642.
24. Hor CN, *et al*. Sleep-wake-driven and circadian contributions to daily rhythms in gene expression and chromatin accessibility in the murine cortex. *Proc Natl Acad Sci USA*. 2019;**116**(51):25773–25783. doi: 10.1073/pnas.1910590116.
25. Wang Z, *et al*. Quantitative phosphoproteomic analysis of the molecular substrates of sleep need. *Nature*. 2018;**558**(7710):435–439. doi: 10.1038/s41586-018-0218-8.
26. Brüning F, *et al*. Sleep-wake cycles drive daily dynamics of synaptic phosphorylation. *Science*. 2019;**366**(6462):eaav3617. doi: 10.1126/science.aav3617.
27. Tatsuki F, *et al*. Involvement of Ca(2+)-dependent hyperpolarization in sleep duration in mammals. *Neuron*. 2016;**90**(1):70–85. doi: 10.1016/j.neuron.2016.02.032.
28. Niwa Y, *et al*. Muscarinic acetylcholine receptors chrm1 and chrm3 are essential for REM sleep. *Cell Rep*. 2018;**24**(9):2231–2247.e7. doi: 10.1016/j.celrep.2018.07.082.
29. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;**16**(6):321–32. doi: 10.1038/nrg3920.
30. Tranchevent LC, *et al*. Candidate gene prioritization with Endeavour. *Nucleic Acids Res*. 2016;**44**(W1):W117–W121. doi: 10.1093/nar/gkw365.
31. Zitnik M, *et al*. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Inf Fusion*. 2019;**50**:71–91. doi: 10.1016/j.inffus.2018.09.012.
32. Nicholls HL, *et al*. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. *Front Genet*. 2020;**11**:350. doi: 10.3389/fgene.2020.00350.
33. Vitsios D, *et al*. Mantis-ml: disease-agnostic gene prioritization from high-throughput genomic screens by stochastic semi-supervised learning. *Am J Hum Genet*. 2020;**106**(5):659–678. doi: 10.1016/j.ajhg.2020.03.012.
34. Brueggeman L, *et al*. Forecasting risk gene discovery in autism with machine learning and genome-scale data. *Sci Rep*. 2020;**10**(1):4569. doi: 10.1038/s41598-020-61288-5.
35. Binder J, *et al*. Machine learning prediction and tau-based screening identifies potential Alzheimer's disease genes relevant to immunity. *Commun Biol*. 2022;**5**(1):125. doi: 10.1038/s42003-022-03068-7.
36. Anafi RC, *et al*. Machine learning helps identify CHRONO as a circadian clock component. Schibler U, ed. *PLoS Biol*. 2014;**12**(4):e1001840. doi: 10.1371/journal.pbio.1001840.
37. Ritchie ME, *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;**43**(7):e47. doi: 10.1093/\nar/gkv007.
38. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2014;**42**(Database issue):D7–17. doi: 10.1093/nar/gkt1146.
39. Liberzon A, *et al*. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;**27**(12):1739–1740. doi: 10.1093/bioinformatics/btr260.
40. Rouillard AD, *et al*. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*. 2016;**2016**:baw100. doi: 10.1093/database/baw100.
41. Stark C, *et al*. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*. 2006;**34**(Database issue):D535–D539. doi: 10.1093/nar/gkj109.
42. Oughtred R, *et al*. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci*. 2021;**30**(1):187–200. doi: 10.1002/pro.3978.
43. Su AI, *et al*. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA*. 2004;**101**(16):6062–6067. doi: 10.1073/pnas.0400782101.
44. Wu C, *et al*. BioGPS: building your own mash-up of gene annotations and expression profiles. *Nucleic Acids Res*. 2016;**44**(D1):D313–D316. doi: 10.1093/nar/gkv1104.
45. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group. *et al*.; Genetic effects on gene expression across human tissues. *Nature*. 2017;**550**(7675):204–213. doi: 10.1038/nature24277.
46. Uhlén M, *et al*. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;**347**(6220):1260419. doi: 10.1126/science.1260419.

47. Sunkin SM, *et al*. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res*. 2013;**41**(Database issue):D996–D1008. doi: 10.1093/nar/gks1042.
48. Kim MS, *et al*. A draft map of the human proteome. *Nature*. 2014;**509**(7502):575–581. doi: 10.1038/nature13302.
49. Zhang R, *et al*. A circadian gene expression atlas in mammals: implications for biology and medicine. *Proc Natl Acad Sci USA*. 2014;**111**(45):16219–16224. doi: 10.1073/pnas.1408886111.
50. Wu G, *et al*. MetaCycle: an integrated R package to evaluate periodicity in large scale data. *Bioinformatics*. 2016;**32**(21):3351–3353. doi: 10.1093/bioinformatics/btw405.
51. Anafi RC, *et al*. CYCLOPS reveals human transcriptional rhythms in health and disease. *Proc Natl Acad Sci USA*. 2017;**114**(20):5312–5317. doi: 10.1073/pnas.1619320114.
52. Ruben MD, *et al*. A database of tissue-specific rhythmically expressed human genes has potential applications in circadian medicine. *Sci Transl Med*. 2018;**10**(458):eaat8806. doi: 10.1126/scitranslmed.aat8806.
53. Pizarro A, *et al*. CircaDB: a database of mammalian circadian gene expression profiles. *Nucleic Acids Res*. 2013;**41**(Database issue):D1009–D1013. doi: 10.1093/nar/gks1161.
54. Edgar R, *et al*. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;**30**(1):207–210. doi: 10.1093/nar/30.1.207.
55. Papatheodorou I, *et al*. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res*. 2020;**48**(D1):D77–D83. doi: 10.1093/nar/gkz947.
56. Yu K, *et al*. qPhos: a database of protein phosphorylation dynamics in humans. *Nucleic Acids Res*. 2019;**47**(D1):D451–D458. doi: 10.1093/nar/gky1052.
57. Drabkin HJ, *et al*. Mouse genome informatics database. manual gene ontology annotation workflow at the mouse genome informatics database. *Database*. 2012;**2012**:bas045–bas045. doi: 10.1093/database/bas045.
58. Sadasivam S, *et al*. The MuvB complex sequentially recruits B-Myb and FoxM1 to promote mitotic gene expression. *Genes Dev*. 2012;**26**(5):474–489. doi: 10.1101/gad.181933.111.
59. Cirelli C. The genetic and molecular regulation of sleep: from fruit flies to humans. *Nat Rev Neurosci*. 2009;**10**(8):549–560. doi: 10.1038/nrn2683.
60. Li XL, Liu B. Learning from positive and unlabeled examples with different data distributions. In: *Machine Learning: ECML 2005. Lecture notes in computer science*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005:218–229. doi: 10.1007/11564096_24.
61. Bekker J, Davis J. Learning from positive and unlabeled data: a survey. *Mach Learn*. 2020;**109**(4):719–760. doi: 10.1007/s10994-020-05877-5.
62. Pedregosa F, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;**12**(85):2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.
63. Developers T. *TensorFlow*.; 2022. doi: 10.5281/zenodo.6574269.
64. Wang H, *et al*. Genome-wide association analysis of self-reported daytime sleepiness identifies 42 loci that suggest biological subtypes. *Nat Commun*. 2019;**10**(1):3503. doi: 10.1038/s41467-019-11456-7.
65. Watanabe K, *et al*. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun*. 2017;**8**(1):1826. doi: 10.1038/s41467-017-01261-5.
66. Huang DW, *et al*. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;**4**(1):44–57. doi: 10.1038/nprot.2008.211.
67. Donohue KD, *et al*. Assessment of a non-invasive high-throughput classifier for behaviours associated with sleep and wake in mice. *Biomed Eng Online*. 2008;**7**:14. doi: 10.1186/1475-925X-7-14.
68. Mang GM, *et al*. Evaluation of a piezoelectric system as an alternative to electroencephalogram/ electromyogram recordings in mouse sleep studies. *Sleep*. 2014;**37**(8):1383–1392. doi: 10.5665/sleep.3936.
69. Levandowsky M, *et al*. Distance between Sets. *Nature*. 1971;**234**(5323):34–35. doi: 10.1038/234034a0.
70. Kass RE, *et al*. Bayes factors. *J Am Stat Assoc*. 1995;**90**(430):773–795. doi: 10.1080/01621459.1995.10476572.
71. Harold Jeffreys. *The Theory of Probability*. 3rd ed. Oxford, United Kingdom: Oxford University Press; 1998:432.
72. Ibarra-Coronado EG, *et al*. The bidirectional relationship between sleep and immunity against infections. *J Immunol Res*. 2015;**2015**:678164. doi: 10.1155/2015/678164.
73. Lasselin J, *et al*. Sleep during naturally occurring respiratory infections: a pilot study. *Brain Behav Immun*. 2019;**79**:236–243. doi: 10.1016/j.bbi.2019.02.006.
74. Lee WS, Liu B. Learning with positive and unlabeled examples using weighted logistic regression. In: Proceedings of the Twentieth International Conference on International Conference on Machine Learning. Washington, DC: ICML'03. AAAI Press; 2003:448–455.
75. Gilestro GF, *et al*. Widespread changes in synaptic markers as a function of sleep and wakefulness in Drosophila. *Science*. 2009;**324**(5923):109–112. doi: 10.1126/science.1166673.
76. Weber F, Dan Y. Circuit-based interrogation of sleep control. *Nature*. 2016;**538**(7623):51–59 doi: 10.1038/nature19773.
77. Scammell TE, *et al*. Neural circuitry of wakefulness and sleep. *Neuron*. 2017;**93**(4):747–765. doi: 10.1016/j.neuron.2017.01.014.
78. Dimsdale JE, *et al*. The effect of opioids on sleep architecture. *J Clin Sleep Med*. 2007;**3**(1):33–36. doi: 10.5664/jcsm.26742.
79. Reinoso-Barbero F, *et al*. Effects of opioid microinjections in the nucleus of the solitary tract on the sleep-wakefulness cycle states in cats. *Anesthesiology*. 1995;**82**(1):144–152. doi: 10.1097/00000542-199501000-00019.
80. Mills PJ, *et al*. Inflammation and sleep in healthy individuals. *Sleep*. 2007;**30**(6):729–735. doi: 10.1093/sleep/30.6.729.
81. Zhang Q, *et al*. 30 Years of NF-κB: a blossoming of relevance to human pathobiology. *Cell*. 2017;**168**(1-2):37–57. doi: 10.1016/j.cell.2016.12.012.
82. Hong HK, *et al*. Requirement for NF-κB in maintenance of molecular and behavioral circadian rhythms in mice. *Genes Dev*. 2018;**32**(21-22):1367–1379. doi: 10.1101/gad.319228.118.
83. Shen Y, *et al*. NF-κB modifies the mammalian circadian clock through interaction with the core clock protein BMAL1. *PLoS Genet*. 2021;**17**(11):e1009933e1009933. doi: 10.1371/journal.pgen.1009933.
84. Irwin MR, *et al*. Sleep Health: Reciprocal Regulation of Sleep and Innate Immunity. *Neuropsychopharmacology*. 2017;**42**(1):129–155. doi: 10.1038/npp.2016.148.
85. Jhaveri KA, Ramkumar V, Trammell RA, Toth LA. Spontaneous, homeostatic, and inflammation-induced sleep in NF-κB p50 knockout mice. *Am J Physiol Regul Integr Comp Physiol*. November 2006;**291**(5):R1516–R1526. doi: 10.1152/ajpregu.00262.2006.
86. Sasaki Y, *et al*. Canonical NF-kappaB activity, dispensable for B cell development, replaces BAFF-receptor signals and promotes B cell proliferation upon activation. *Immunity*. 2006;**24**(6):729–739. doi: 10.1016/j.immuni.2006.04.005.
87. Madisen L, *et al*. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat Neurosci*. 2010;**13**(1):133–140. doi: 10.1038/nn.2467.

88. Hayden MS, *et al*. Shared principles in NF-kappaB signaling. *Cell*. 2008;**132**(3):344–362. doi: 10.1016/j.cell.2008.01.020.
89. Sethi M, *et al*. Increased fragmentation of sleep-wake cycles in the 5XFAD mouse model of Alzheimer's disease. *Neuroscience*. 2015;**290**:80–89. doi: 10.1016/j.neuroscience.2015.01.035.
90. Fong RC, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE; 2017. doi: 10.1109/iccv.2017.371.
91. Azodi CB, *et al*. Opening the black box: interpretable machine learning for geneticists. *Trends Genet*. 2020;**36**(6):442–455. doi: 10.1016/j.tig.2020.03.005.
92. Franco HL, *et al*. Redundant or separate entities?--roles of Twist1 and Twist2 as molecular switches during gene transcription. *Nucleic Acids Res*. 2011;**39**(4):1177–1186. doi: 10.1093/nar/gkq890.
93. Šošić D, *et al*. Twist regulates cytokine gene expression through a negative feedback loop that represses NF-kappaB activity. *Cell*. 2003;**112**(2):169–180. doi: 10.1016/s0092-8674(03)00002-3.
94. Li CW, *et al*. Epithelial-mesenchymal transition induced by TNF-α requires NF-κB-mediated transcriptional upregulation of Twist1. *Cancer Res*. 2012;**72**(5):1290–1300. doi: 10.1158/0008-5472.CAN-11-3123.
95. Li S, *et al*. TWIST1 associates with NF-κB subunit RELA via carboxyl-terminal WR domain to promote cell autonomous invasion through IL8 production. *BMC Biol*. 2012;**10**:73. doi: 10.1186/1741-7007-10-73.
96. Roberts CM, *et al*. Disruption of TWIST1-RELA binding by mutation and competitive inhibition to validate the TWIST1 WR domain as a therapeutic target. *BMC Cancer*. 2017;**17**(1):184. doi: 10.1186/s12885-017-3169-9.
97. Meier D, *et al*. Twist1 Is a TNF-inducible inhibitor of clock mediated activation of period genes. *PLoS One*. 2015;**10**(9):e0137229. doi: 10.1371/journal.pone.0137229.
98. Pan Y, *et al*. The role of Twist1 in mutant huntingtin–induced transcriptional alterations and neurotoxicity. *J Biol Chem*. 2018;**293**(30):11850–11866. doi: 10.1074/jbc.RA117.001211.
99. Herzog–Krzywoszanska R, *et al*. Sleep Disorders in Huntington's Disease. *Front Psychiatry*. 2019;**10**:221. doi: 10.3389/fpsyt.2019.00221.
100. Pellegrino R, *et al*. A novel BHLHE41 variant is associated with short sleep and resistance to sleep deprivation in humans. *Sleep*. 2014;**37**(8):1327–1336. doi: 10.5665/sleep.3924.
101. Besedovsky L, *et al*. The sleep-immune crosstalk in health and disease. *Physiol Rev*. 2019;**99**(3):1325–1380. doi: 10.1152/physrev.00010.2018.
102. Fiorentino L, *et al*. Sleep dysfunction in patients with cancer. *Curr Treat Options Neurol*. 2007;**9**(5):337–346. doi: 10.1007/s11940-007-0019-0.
103. Peter-Derex L, *et al*. Sleep and Alzheimer's disease. *Sleep Med Rev*. 2015;**19**:29–38. doi: 10.1016/j.smrv.2014.03.007.
104. Aerts S, *et al*. Gene prioritization through genomic data fusion. *Nat Biotechnol*. 2006;**24**(5):537–544. doi: 10.1038/nbt1203.
105. Bjorness TE, *et al*. An essential role for MEF2C in the cortical response to loss of sleep in mice. *Elife*. 2020;**9**:e58331. doi: 10.7554/eLife.58331.
106. Shi G, *et al*. Mutations in metabotropic glutamate receptor 1 contribute to natural short sleep trait. *Curr Biol*. 2021;**31**(1):13–24.e4. doi: 10.1016/j.cub.2020.09.071.
107. Reitz SL, *et al*. Activation of preoptic tachykinin 1 neurons promotes wakefulness over sleep and volatile anesthetic-induced unconsciousness. *Curr Biol*. 2021;**31**(2):394–405.e4. doi: 10.1016/j.cub.2020.10.050.