PLOS GENETICS

RESEARCH ARTICLE

Automated feature extraction from population wearable device data identified novel loci associated with sleep and circadian rhythms

Xinyue Li[1], **Hongyu Zhao**[2,3,4]*

1 School of Data Science, City University of Hong Kong, Hong Kong, China, **2** Department of Biostatistics, Yale School of Public Health, New Haven, CT, United States of America, **3** Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, United States of America, **4** Department of Genetics, Yale University School of Medicine, New Haven, CT, United States of America

* hongyu.zhao@yale.edu

Citation: Li X, Zhao H (2020) Automated feature extraction from population wearable device data identified novel loci associated with sleep and circadian rhythms. PLoS Genet 16(10): e1009089. https://doi.org/10.1371/journal.pgen.1009089

Editor: Michael Snyder, Stanford University School of Medicine, UNITED STATES

Received: April 3, 2020

Accepted: August 31, 2020

Published: October 19, 2020

Data Availability Statement: Data are available in the UK Biobank project: www.ukbiobank.ac.uk. Access to data was obtained under application number 29900.

Funding: Supported in part by NIH grant R01GM122078 and NSF grants DMS 1713120 and DMS 1902903.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Wearable devices have been increasingly used in research to provide continuous physical activity monitoring, but how to effectively extract features remains challenging for researchers. To analyze the generated actigraphy data in large-scale population studies, we developed computationally efficient methods to derive sleep and activity features through a Hidden Markov Model-based sleep/wake identification algorithm, and circadian rhythm features through a Penalized Multi-band Learning approach adapted from machine learning. Unsupervised feature extraction is useful when labeled data are unavailable, especially in large-scale population studies. We applied these two methods to the UK Biobank wearable device data and used the derived sleep and circadian features as phenotypes in genome-wide association studies. We identified 53 genetic loci with $p<5\times10^{-8}$ including genes known to be associated with sleep disorders and circadian rhythms as well as novel loci associated with Body Mass Index, mental diseases and neurological disorders, which suggest shared genetic factors of sleep and circadian rhythms with physical and mental health. Further cross-tissue enrichment analysis highlights the important role of the central nervous system and the shared genetic architecture with metabolism-related traits and the metabolic system. Our study demonstrates the effectiveness of our unsupervised methods for wearable device data when additional training data cannot be easily acquired, and our study further expands the application of wearable devices in population studies and genetic studies to provide novel biological insights.

Author summary

While wearable devices have been increasingly used in research for objective and continuous activity monitoring, how to effectively extract sleep and rest-activity circadian rhythm features remains the major obstacle for researchers, especially in population studies where

labeled outcome data such as sleep diaries are unavailable and thus existing supervised methods cannot be applied. Here, we developed unsupervised feature extraction methods based on machine learning without the need for labeled outcome data. We applied the methods to population wearable device data to extract sleep and circadian features, and we further identified novel associated loci and the key roles of the central nervous system and the metabolic system. The findings are essential for understanding the underlying shared genetic architecture of sleep and circadian rhythms with physical and mental health, and the proposed methods can largely expand and promote the use of wearable device data in population and genetic studies.

Introduction

Sleep is essential for human health and well-being, and changes in sleeping patterns or habits can negatively affect health, leading to physical and mental disorders [1–4]. The corresponding sleep-wake circadian rhythm is also essential for human health. Circadian rhythms are endogenous biological processes that follow a period of approximately 24 hours and are entrained by environmental stimuli such as the light/dark cycle to adjust the 24-hour cycle [5]. The circadian system is important for sleep regulation, and dysregulated sleep-wake circadian rhythms can cause diseases including sleep disorders, metabolic syndrome, and psychiatric and neurodegenerative diseases [6–8]. While it is vital to obtain a thorough understanding of the important roles of sleep and circadian rhythms in human health, they still remain poorly understood.

Actigraphy has been increasingly used in sleep and circadian studies, as it can provide continuous and objective activity monitoring and is low-cost and easy to wear. Actigraphy can address some of the limitations with traditional sleep diaries, including subjectivity, bias, difficulty in completion by young children or patients, and extra manual work for caregivers. While polysomnography (PSG) as the "gold standard" in sleep studies does not have the same issues as sleep logs, it is limited by high costs, in-lab setting, intrusive measures, and difficulty in long-time monitoring. The continuous and objective measures provided by actigraphy can provide reliable information on sleep [9], and it is especially useful for studying long-duration circadian rhythms. In cases of large-scale epidemiological studies, the availability of actigraphy data provides excellent opportunities for studying population-level and individual-level sleep characteristics and circadian rhythm patterns. However, the analysis of actigraphy data remains a major obstacle for researchers.

One major challenge in the application of actigraphy is to extract sleep and circadian rhythm features without additional information, such as sleep diaries and/or PSG validation that are often unavailable or labor-intensive to collect. To extract sleep features such as sleep start, sleep end, and sleep duration, it is typical to either obtain the gold-standard PSG records or obtain sleep dairies on go-to-sleep time and wake-up time to build sleep identification algorithms [10–13]. To ease researcher's efforts in collecting additional data and still accurately infer sleep features, there is a need to develop necessary methodology to infer sleep parameters based on actigraphy in the absence of sleep records. The method will be useful in circadian studies where PSG for long-time monitoring cannot be acquired and continuous activity logs requires much manual work. It is also particularly useful in large-scale epidemiological studies where collecting PSG for all participants is unrealistic and recording sleep logs is labor-intensive.

In this paper, we analyzed data from the UK Biobank study, where accelerometer data from over 100,000 participants and genetic data from near 500,000 participants are available. We

applied novel data processing methods to the accelerometer data from 90,515 participants after quality control procedures to automatically extract sleep and circadian rhythm features. The features were then utilized in genome-wide association studies (GWAS). Our study differs from previous studies in that we take on unsupervised and individualized approaches for feature extraction and that we extracted new features for further genetic analysis, such as activity levels during sleep and periodic features. Among previous studies using UK Biobank wearable device data, Doherty et al. 2018 predicted activity types using a random forest algorithm trained with additional labelled data in a separate study, classified activities into sleep, sedentary, walking and moderate intensity activity behaviors, and further conducted GWAS [14, 15]. Dashti et al. 2019 and Jones et al. 2019 studied sleep related characteristics using self-reported measures and device-based measures with a heuristic algorithm inferring posture changes and possible sleep behaviors based on variance in the estimated z-axis angle, and information such as sleep duration, sleep midpoint and sleep quality were used as traits in GWAS [16, 17].

Our study proposed new unsupervised and individualized approaches for wearable device data analysis and extracted new features such as activity levels during sleep, chronotype-related features, and periodic features for further genetic analysis. Unsupervised approaches can be widely applied to population studies without the need of additional studies for collecting labeled data, and individualized approaches can well account for individual variations and capture individual characteristics. Using new features for GWAS and post-GWAS analysis, our study provides new insights into the molecular regulation and genetic basis of sleep and circadian rhythms.

The analysis pipeline for automatic sleep and circadian feature extraction from wearable device data and further GWAS is summarized in Fig 1. We identified 19 and 34 genetic loci associated with sleep traits and circadian rhythm traits at $p<5\times10^{-8}$ respectively, of which 5 and 13 loci reached the significance level $p<5\times10^{-9}$. Further tissue enrichment analysis highlights the important roles of the central nervous system and the metabolic system, thereby providing new insights into the molecular regulation and genetic basis of sleep and circadian rhythms.

Results

Loci associated with sleep-activity traits and circadian traits

The Manhattan plots of GWAS results for HMM inferred sleep and activity traits are shown in S1 Fig. The heritability estimates from LD score regression [18] for mean activity levels during sleep and during wake are ~5% and ~7%, respectively. For mean activity levels during sleep, four independent regions on chromosomes 2, 5, 6, and 14 contained significant SNPs with p-value $< 5\times10^{-8}$, two of which had p-values $< 5\times10^{-9}$ (Table 1). The strongest association signal was on chromosome 2 at SNPs within gene MEIS1, known for association with Restless Leg Syndrome and Insomnia [19–22]. The other three genetic loci were rs188904275 in JAKMIP2 (p-value = 3.7×10^{-8}), rs184670665 near IMPG1 (p-value = 2.4×10^{-10}), and rs73586669 near OR4E1 (p-value = 2.4×10^{-8}), with JAKMIP2 previously found to be associated with Body Mass Index (BMI) measurements and pulmonary diseases [23]. These novel loci were not previously associated with activity levels during sleep. SNP rs7087063 in gene CELF2 on chromosome 10 was found to be associated with the HMM estimated activity variability during wake (p-value = 3.0×10^{-8}), and we note that CELF2 was previously found to be associated with Alzheimer's disease [24] (Table 1). No association was detected for HMM estimated activity variability during sleep or mean activity levels during wake. The QQ-plots checking for population stratification indicate that the population structure was properly controlled for, with small values of the inflation factor λ under 1.04 (S2 Fig).

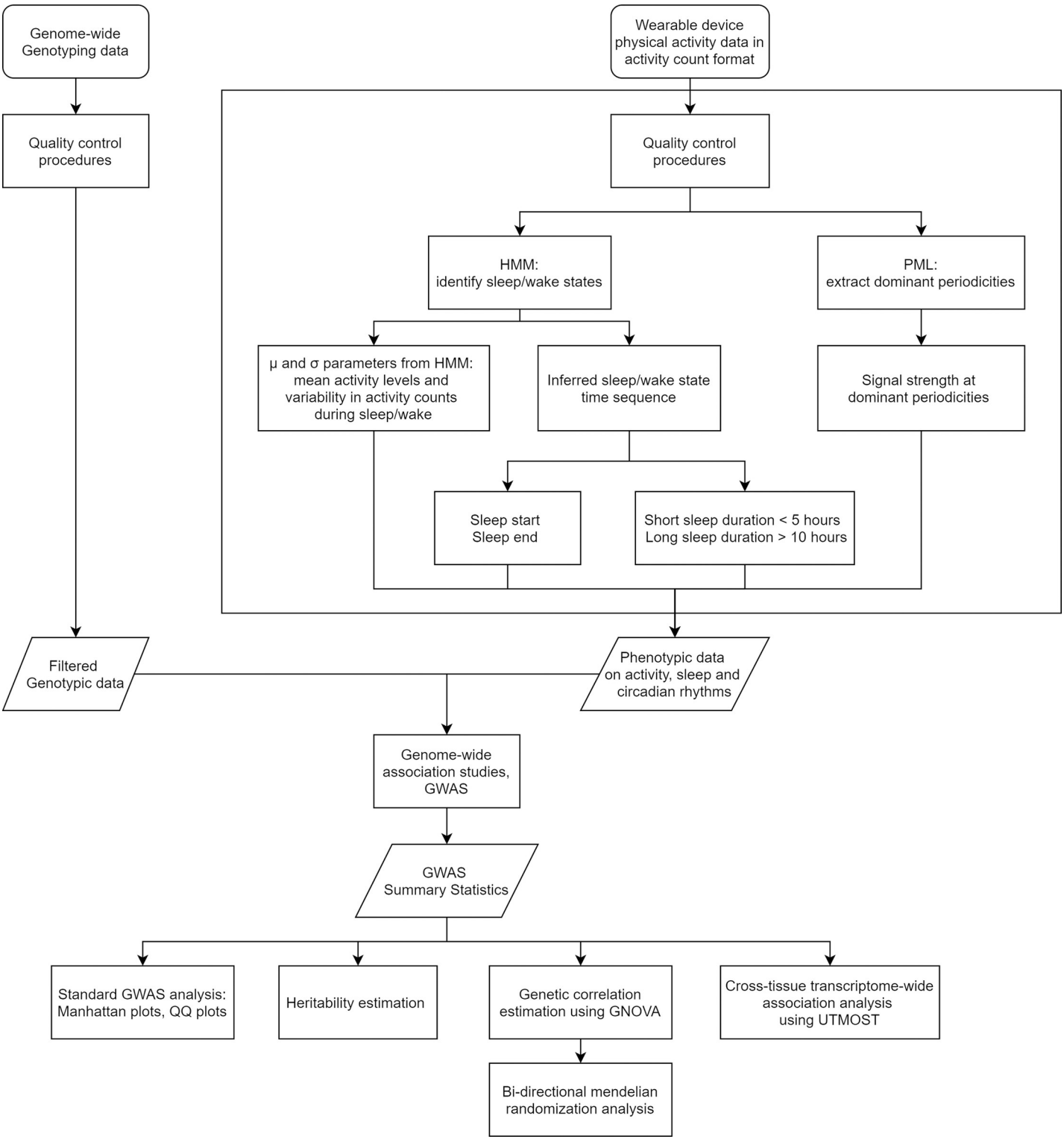

Fig 1. The pipeline of analyzing wearable device data and using extracted sleep and circadian features in genome-wide association studies.

https://doi.org/10.1371/journal.pgen.1009089.g001

Table 1. The SNPs identified in genome-wide association studies at the significance level of 5×10^{-8} that are associated with sleep and activity traits inferred from accelerometer-measured physical activity in 90,515 UK Biobank participants.

Trait	Chr	Position	ID	Novel	Function	Transcription Factor Binding Site	Nearest Gene	Risk Allele	BETA	SE	P
Mean Activity During Sleep	2	66745864	rs62144053	No	intronic	Yes	MEIS1	G	0.049	0.008	3.07E-09
	2	66747480	rs62144054	No	intronic	No	MEIS1	G	0.048	0.008	6.17E-09
	2	66750564	rs113851554	No	intronic	No	MEIS1	G	0.091	0.011	1.71E-17
	2	66785180	rs11679120	No	intronic	Yes	MEIS1	G	0.089	0.012	6.27E-14
	2	66799986	rs11693221	No	downstream	Yes	MEIS1(dist = 95)	C	0.089	0.012	3.91E-14
	5	147129599	rs188904275	Yes	intronic	Yes	JAKMIP2	A	-0.241	0.044	3.71E-08
	6	77084619	rs184670665	Yes	intergenic	Yes	IMPG1 (dist = 302224)	A	-0.518	0.082	2.42E-10
	14	22575973	rs73586669	Yes	intergenic	Yes	OR4E1 (dist = 436741)	T	-0.401	0.072	2.45E-08
Activity Variability During Wake	10	11355672	rs7087063	Yes	intronic	No	CELF2	G	0.027	0.005	2.96E-08

https://doi.org/10.1371/journal.pgen.1009089.t001

The Manhattan plots of GWAS results for sleep duration, sleep start and sleep end traits are shown in S1 Fig. The heritability estimates from LD score regression [18] for sleep start and sleep end are ~8% and ~7%, respectively. SNP rs573901234 near gene JHDM1D-AS1 on chromosome 7 was associated with short sleep duration < 5 hours (p-value = 1.4×10^{-8}), and five SNPs were associated with long sleep duration > 10 hours, among which rs74460673 at gene TMEM39B was previously found to be associated with BMI [25, 26] and rs573982927 at MYO3B was associated with obesity traits [27, 28] (Table 2).

Twenty seven SNPs were found to be associated with sleep start, including three SNPs at or near gene MEIS1 related to Restless Leg Syndrome and insomnia [19–22] and nineteen SNPs at gene BTBD9 also related to Restless Leg Syndrome [29] (S1 Table). Sleep start is also associated with one SNP at gene CYP7B1 related to BMI, two SNPs near gene HSD17B12 related to BMI [25, 30, 31], two SNPs near MIR129-2 also related to BMI [25, 26] and Alzheimer's Disease [32], and two SNPs near LOC101928944 related to schizophrenia [33, 34]. As for sleep end, association signals were found for five SNPs in the intergenic regions near LINC02260, which is known to be related to red blood cell measures such as cell counts and hemoglobin

Table 2. The SNPs identified in genome-wide association studies at the significance level of 5×10^{-8} that are associated with sleep duration traits inferred from accelerometer-measured physical activity in 90,515 UK Biobank participants.

Trait	Chr	Position	ID	Novel	Function	Transcription Factor Binding Site	Nearest Gene	Risk Allele	BETA	SE	P
Sleep Duration < 5h	7	139916399	rs573901234	Yes	intergenic	No	JHDM1D-AS1 (dist = 36959)	T	3.294	0.210	1.44E-08
Sleep Duration > 10h	1	32558622	rs74460673	Yes	intronic	Yes	TMEM39B	A	3.822	0.242	2.96E-08
	2	171183777	rs573982927	Yes	intronic	No	MYO3B	T	3.460	0.227	4.45E-08
	4	139129025	rs182651559	Yes	intronic	Yes	SLC7A11	T	4.989	0.280	9.72E-09
	8	131387296	rs571444813	Yes	intronic	No	ASAP1	T	5.568	0.314	4.62E-08
	22	24204438	rs138381486	Yes	intronic	Yes	SLC2A11	T	1.797	0.107	4.54E-08

https://doi.org/10.1371/journal.pgen.1009089.t002

content [35], and blood cell information is also known to be associated with sleep deprivation and sleep disorders [36–38]. Sleep end is also associated with one SNP at gene NTNG1 related to Restless Leg Syndrome [29] and BMI [25, 39], one SNP near LINC00963, and one SNP near GLRX3.

From the penalized multi-band learning approach, the most dominant periodicities are: 1-day, 1/2-day, and 1/3-day. The Manhattan plots for the GWAS results are shown in S1 Fig. The heritability estimate from LD score regression [18] for the circadian feature 1-day periodicity is 9%. For the strength of 1-day periodicity, five circadian SNPs were identified: rs189005747 at gene XKR4 (p-value = 9.9×10^{-9}), rs534035399 at LINC01508 (p-value = 1.8×10^{-9}), rs144874087 near LINC01935 (p-value = 1.9×10^{-8}), rs181820530 near LINC01935 (p-value = 4.5×10^{-8}), and rs554696049 at LINC01501 (p-value = 4.8×10^{-8}), in which XKR4 was previously found to be associated with thyroid stimulating hormone [40–42] and coronary artery disease, and LINC01508 and LINC01501 are RNA genes (shown in S2 Table).

1/2-day periodicity measures the strength of day-night rhythmicity [43] and the strongest association signals (p-value $< 5\times10^{-9}$) were detected in the intergenic regions near UBE2F-SCLY and FBXO15 and in the intronic regions at FYB1 and CFAP44, in which intergenic regions near FBXO15 were previously found to be associated with insomnia [44] and FYB1 was associated with depression [45, 46]. There were also association signals for SNP clusters in the intergenic regions near gene TNR as well as for SNPs at RASEF, TMEM132D, ERCC2, MPPED1 and near BRINP3, LINC01287, and MLYCD (p-value $< 5\times10^{-8}$; S2 Table).

1/3-day periodicity measures the 1/3-day rhythmicity that not only involves activities during the day but also captures activities during sleep [43]. The strongest signals (p-value $< 5\times10^{-9}$) were identified at SNP clusters in the intergenic regions near BRINP3, URB2, GRIA1 and LOC400682 and in the intronic regions at MGAT5, C3orf20, and LINC01861, where BRINP3 was associated with BMI measurement [25, 26, 31], depression [47], and rheumatoid arthritis [48], and GRIA1 was associated with schizophrenia [33, 49, 50] and circadian rhythm [51]. A cluster of five SNPs at CDH6 reached significance level at 5×10^{-8}, and CDH6 was associated with resting heart rates [52]. Details on other novel SNP associations are listed in S2 Table. QQ-plots show that the estimated values of the inflation factor λ are under 1.07 and that the population structure was properly controlled for (S2 Fig).

With respect to the quality control procedures for all GWAS, the inflation factor λ was under 1.12 for all traits considered, suggesting appropriate control of population structure in the analysis (S2 Fig). LD Score intercepts estimated using LD score regression [18] were in the range of 0.99–1.02, indicating that the genomic inflation was due to polygenic architectures rather than uncorrected population structure. Further, we estimated partitioned heritability [53] using ten broad tissue categories and no tissue was significantly enriched. For all sleep time-related traits, the adrenal/pancreatic tissues were relatively more enriched than the other tissues (Fig 2). For activity levels during sleep, the cardiovascular tissues were relatively more enriched compared to the other tissues (Fig 2).

For sleep time-related traits and 1-day periodicity, the adrenal/pancreatic tissues were relatively more enriched than the other tissues (Fig 2). It is known that cortisol released from the adrenal cortex exhibits a diurnal rhythm, with a steady rise during sleep and a peak in the morning to prepare for stresses associated with wakefulness and increased activity [54]. Insulin secreted from pancreas also exhibits a diurnal rhythm with a peak at around 5pm and a nadir at 4am in the morning, consistent with changes in nutrient storage in the awake/fed state and the sleep/fasted state [54]. For activity levels during sleep, the cardiovascular tissues were relatively more enriched compared to the other tissues (Fig 2). Clinical and laboratory studies have shown bidirectional effects between sleep and the cardiovascular system, and in particular, arousals from sleep, which are common in normal sleep, and body movements are

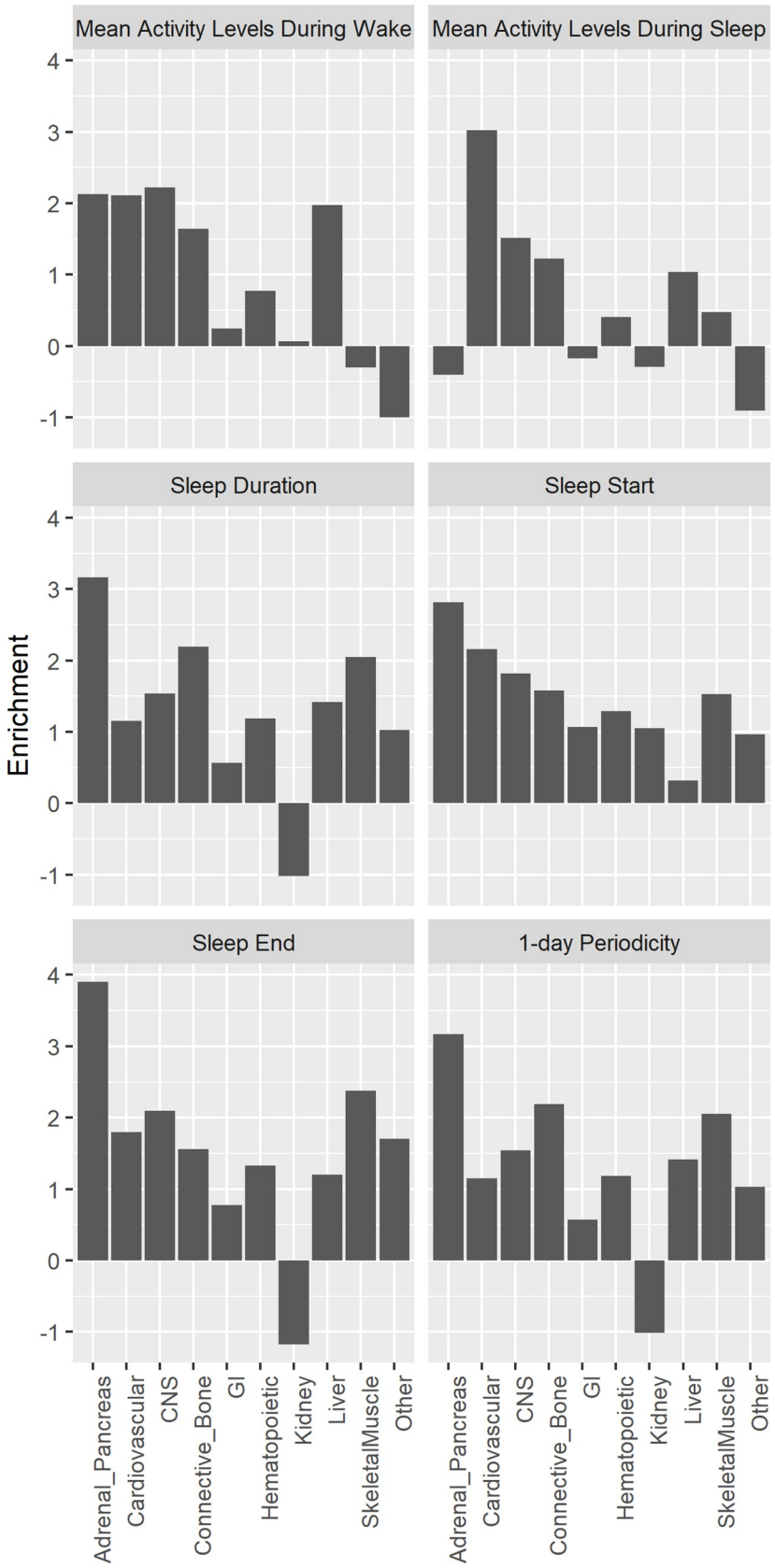
Mean Activity Levels During Wake
Mean Activity Levels During Sleep
Sleep Duration
Sleep Start
Sleep End
1-day Periodicity
Enrichment
4
3
2
1
0
-1
Adrenal_Pancreas
Cardiovascular
CNS
Connective_Bone
GI
Hematopoietic
Kidney
Liver
SkeletalMuscle
Other

Fig 2. Partitioned heritability enrichment analysis across 10 broad tissue types for activity, sleep, and circadian traits.

https://doi.org/10.1371/journal.pgen.1009089.g002

associated with vigorous responses in the cardiovascular system and marked changes in the sleep-related pattern of cardiovascular activities [55, 56].

Genetic correlation of sleep and daily rhythm with other traits

To further investigate the relationships of sleep and circadian traits with other complex traits, we estimated genetic correlation with a number of traits, including screen exposure, sleep, mental health, BMI and diet, alcohol consumption, shift-work, and diseases such as respiratory diseases and anemia (S3 Table) and we set the statistical significance threshold at 7.9×10^{-5} through Bonferroni correction (12 sleep and circadian traits × 53 traits). For sedentary and screen exposure traits, we observed significant negative genetic correlation of time spent watching TV or using computer with the strength of circadian rhythmicity (correlation = -0.300, $p = 2.1 \times 10^{-12}$ and correlation = -0.294, $p = 5.7 \times 10^{-10}$, respectively), and we also observed significant negative correlation of time spent using computer with sleep duration, sleep start (go to bed early), and activity levels during the wake status(correlation = -0.238, $p = 2.7 \times 10^{-7}$; correlation = -0.353, $p = 3.5 \times 10^{-5}$ and correlation = -0.200, $p = 1.8 \times 10^{-22}$ respectively). There was positive genetic correlation between the time spent watching TV and activity levels during sleep (correlation = 0.203, $p = 7.3 \times 10^{-5}$). For BMI and diet related traits, we observed significant negative genetic correlation between BMI and weight with the strength of circadian rhythm (1-day periodicity), activity levels when awake, and sleep duration (all negative correlations with p-values $< 7.9 \times 10^{-5}$). We did not observe significant genetic correlation with mental health traits, alcohol consumption, shift-work, and respiratory diseases either. While the genetic correlation with anemia was not significant, there was moderate positive genetic correlation between activity levels during sleep and iron deficiency anemia (correlation = 0.404, $p = 6.8 \times 10^{-3}$), and iron deficiency is known to be associated with poor sleep quality [57] and restless legs syndrome [58, 59].

We also observed significant positive genetic correlation between accelerometer-derived and self-reported sleep duration, between accelerometer-derived sleep start/end and self-reported chronotypes, and between accelerometer-derived sleep end and self-reported hypersomnia (all p-values $< 7.9 \times 10^{-5}$). These results suggest agreement between accelerometer-derived measures and self-reported measures for sleep time and are also consistent with previous studies [17]. We did not observe significant genetic correlation of accelerometer-derived measures with sleep disorders, possibly because the ambiguity in the definition of sleep disorders. We did not observe significant genetic correlation between accelerometer-derived activity level measures and self-reported physical activity measures.

For the significant trait-pair of BMI and 1-day periodicity denoting the strength of circadian rhythm, we further conducted two-sample Mendelian Randomization (MR) analysis using GWAS summary statistics from the GIANT study [31]. The estimated causal effects across different MR estimation methods are negative for both directions but were not statistically significant (S6 Table).

Cross-tissue transcriptome-wide association analysis

We applied UTMOST [60] to perform tissue enrichment analysis in 44 tissue types and identified single-tissue and cross-tissue gene-trait associations. For single-tissue tests, 124 gene-trait pairs were identified (S4 Table). There were 43 unique genes in total and 16 gene-trait pairs

were identified in more than one tissue type (p-values $< 3.3\times10^{-6}$ after Bonferroni correction for 15,000 genes [60]). Among them, the GLTP-sleep duration pair was identified in 38 tissue types, which is a novel association not reported in previous studies, and GLTP is related to Glycolipid Transfer Protein for protein binding and lipid binding [61]. L3MBTL2, Lethal (3) malignant brain tumor-like protein 2 related to protein binding and DNA regulation [62], was associated with more than one trait, including activity levels during sleep and wake as well as the circadian rhythm, and the three L3MBTL2-trait pairs were all significant in the subcutaneous adipose tissues. Most genes that appeared in more than one gene-trait pair have functions in protein binding, including CEP70, GIPC2, TRAF3, ABCD2, LIMS1, SEPN1. BRSK1 and LRFN4 are related to neurotransmitter activities [62]. TREH is related to digestion and galactose metabolism [62], and the TREH-sleep start pairs are significant in mucosa esophagus, stomach, and transverse colon, all of which belong to the digestive system. Among all 44 tissue types, subcutaneous adipose, brain anterior cingulate cortex, and skeletal muscles are most enriched with seven gene-trait pairs (Fig 3). Tissue enrichment analysis through UTMOST identified novel genes associated with sleep and circadian traits and highlighted the relevance of tissues in the central nervous system and the metabolic system in sleep and circadian regulation.

Joint tests for gene-trait associations across tissues identified 34 gene-trait pairs with p-values $< 3.3\times10^{-6}$ after Bonferroni correction for 15,000 genes [60] and 20 gene-trait pairs with p-values $< 3.3\times10^{-7}$ if Bonferroni correction further adjusts for 10 traits tested (S5

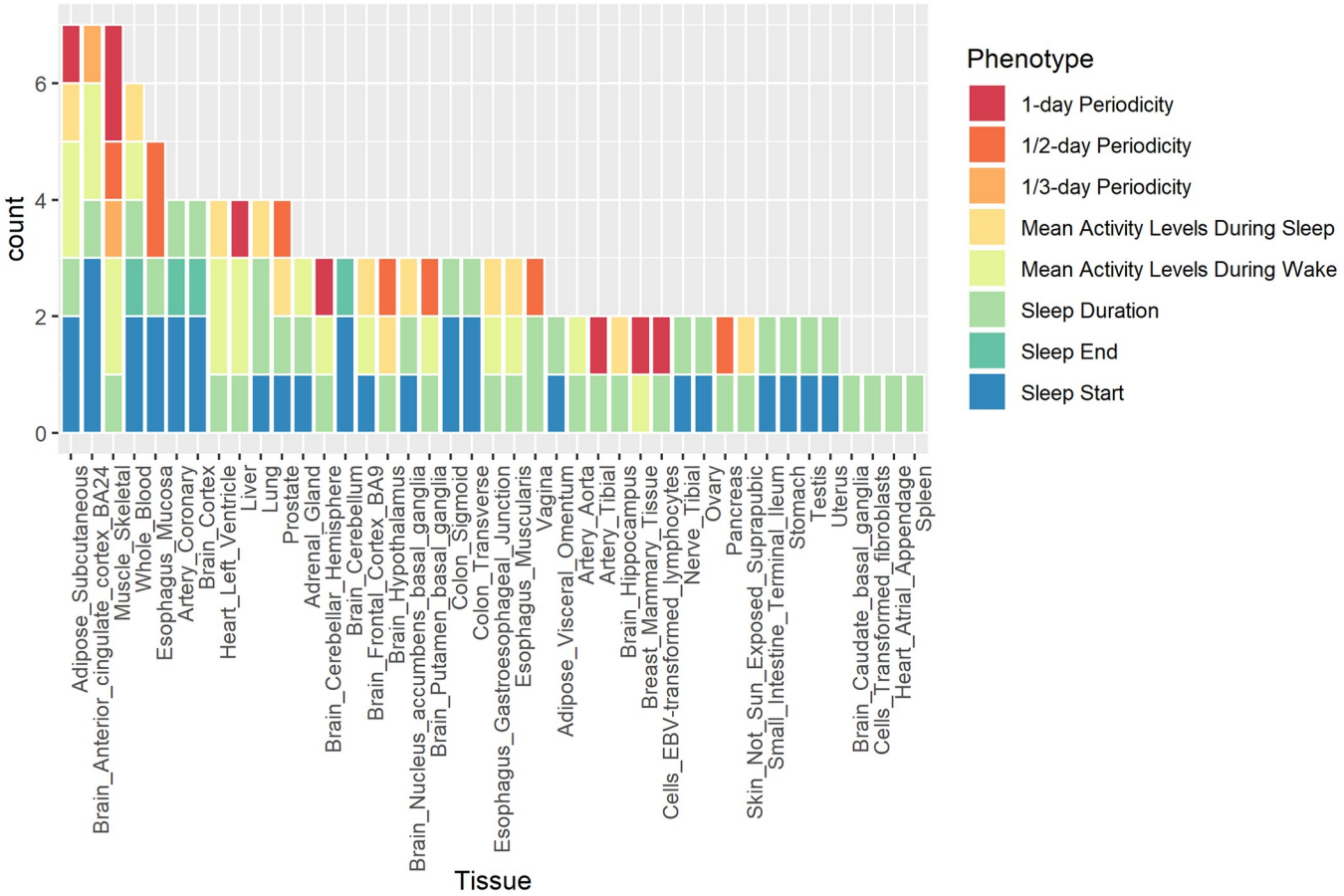

Fig 3. Tissue enrichment analysis across 44 tissue types for activity, sleep, and circadian rhythm traits.

https://doi.org/10.1371/journal.pgen.1009089.g003

Table). Three genes, CA7, DYNC1LI1, and ELMOD2, were associated with more than one trait. CA7 was associated with activity levels during sleep and activity levels during wake, and in previous studies its expression in brain was associated with neurological disorders [63]. DYNC1LI1 was associated with sleep duration and daily rhythmicity and its functions are RNA binding [64] and protein binding [65]. ELMOD2 was associated with activity variability during wake, sleep duration, sleep end and daily rhythmicity and its function may be related to respiratory diseases [66]. Among gene-trait pairs, GRIA1 was also identified in GWAS analysis and was associated with circadian rhythms and sleep traits in previous studies [44, 51], and its functions involve amyloid-beta binding in hippocampal neurons and ionotropic neurotransmitter receptor activities [67]. Gene Ontology (GO) enrichment analysis using DAVID [68] did not identify significant enrichment among different GO terms.

Discussion

Using information related to sleep and activity inferred from the UK Biobank accelerometer data, we identified five genetic loci associated with HMM derived sleep related traits. The association of activities during sleep with restless leg syndrome, obesity, and pulmonary diseases may be attributed to the fact that individuals with these traits may be less likely to sleep well at night and are prone to being restless. Sleep start and end are also strongly associated with Restless Leg Syndrome, as people with this disorder may have difficulty falling asleep or staying asleep, and because symptoms worsen at night only with a short period in early morning that is symptom-free, people with Restless Leg Syndrome tend to have late sleep onset [69]. Sleep duration, sleep start and sleep end are all associated with BMI, consistent with the established association between sleep and obesity. As previous studies [70] discussed how decreased sleep duration can elevate obesity risks and how sleep disorders can increase risks for chronic health conditions, our study also provides evidence of genetic correlations among sleep, metabolism and obesity. For sleep end, it is interesting to identify SNPs near LINC02260, which was previously found to be associated with red blood cell related measures [35]. It is known from previous studies that sleep deprivation and sleep disorders can lead to changes in metabolism with altered red blood cell measurements such as increased red blood cell counts [36–38]. Thus, our study also suggests the link between sleep and blood cell metabolism, and the complex relationship remains to be studied.

For circadian and daily rhythm analysis, we were able to identify 13 loci associated with periodic traits. The SNPs located at FBXO15 and GRIA1 were associated with circadian rhythms and sleep traits in previous studies [44, 51], and the associations of SNPs in XKR4 and CDH6 with thyroid stimulating hormones and resting heart rates can be explained by the circadian oscillation natures of hormones and heart rates [71, 72]. These association results suggest that the Penalized Multi-band Learning approach [43] is effective in extracting clinically meaningful circadian features from objective physical activity measures. Furthermore, we identified novel associations for SNPs at or near FYB1, GRIA1, CDH6 and BRINP3, which are associated with BMI and mental problems, suggesting shared genetics between sleep-wake circadian rhythms and multiple traits.

Our study found that sleep and physical activity are closely related to mental and neurological problems, as the identified SNPs/loci from our GWAS results were also associated with traits including depression, schizophrenia and Alzheimer's disease. Our cross-tissue transcriptome-wide association analysis also implied the important role of the central nervous system in physical activity, sleep, and circadian rhythm. Our results are consistent with previous UK Biobank studies [14, 73] and confirmed the shared genetic architectures of sleep and physical activity with mental health and neurological disorders.

Our study shows strong evidence for shared genetics of activity and circadian rhythm with metabolism-related traits and the metabolic system. In addition to GWAS results that suggest shared genetic architecture with BMI, genome-wide genetic correlation estimates also provide strong evidence that a higher BMI is associated with lower physical activity levels and weaker daily rhythmicity. Furthermore, our UTMOST cross-tissue transcriptome-wide association analyses also implicate that the adipose tissues and skeletal muscle tissues, which are mostly related to metabolism and physical activity, may have important roles as they together with the brain cortex are most enriched. Our study suggests the complex interplay among activity, circadian rhythm, metabolic phenotypes and the central nervous system.

We also note that the gene TREH, which was identified in UTMOST tissue enrichment analysis and whose function is related to digestion and galactose metabolism, is associated with sleep start in esophagus, stomach and colon that all belong to the digestive system, suggesting the link between digestion and sleep. From daily experience, a heavy dinner or an empty stomach may affect the ability to fall asleep, and on the other hand, insomnia may also affect the digestive system. It is known from previous studies [57, 74, 75] that sleep problems are associated with gastrointestinal problems and digestive disorders such as gastroesophageal reflux disease and irritable bowel syndrome. We are just beginning to dissect the underlying genetic architecture of sleep, and details on the complex relationships of sleep with the central nervous system and the metabolic system remain to be studied.

This study demonstrates the utility of device measures by presenting applications of population-level objective physical activity data in genetic studies, using novel methods to effectively extract sleep, activity and daily rhythm features. Our results show that accelerometer-derived sleep duration and sleep start/end correlate well with self-reported sleep duration, chronotypes, and hypersomnia. There are discrepancies between accelerometer-derived and self-reported physical activity measures, which may be due to commonly observed self-report bias, as people tend to overestimate their daily activity and underestimate sedentary behaviors [76]. The HMM-based algorithm can classify sleep/wake epochs, and the estimated HMM parameters can further be used as sleep and activity related features directly: mean activity levels and activity variability during sleep or wake can characterize individual sleep-activity patterns effectively [77]. In a similar manner, the Penalized Multi-band Learning approach [43] can identify the population-level dominant periodic information that depicts daily rhythms, and further, individual variations in the strength of periodic signals can be utilized as circadian and daily rhythm features in further analyses. Our study effectively extracted periodic features from objective physical activity data and utilized them in genetic studies to examine the genetic architecture of circadian and daily rhythms.

Our study promotes the utility of objective activity measures in sleep and physical activity studies when coupled with automated algorithms. The HMM-based sleep/wake identification algorithm and the Penalized Multi-band Learning approach are particularly useful in large-scale population studies where additional sleep validation data are unavailable and manual data examination is labor-intensive and not feasible. More statistical methods are needed to expand and promote the application of actigraphy in clinical and epidemiological studies. Our current circadian study focuses on periodicities, and we will develop new methodology such as functional analysis to extract new dimensions of information and fully exploit actigraphy data in our future work.

In summary, using large-scale population studies like the UK Biobank study with objective physical activity data and genetic data available, we were able to extract meaningful sleep, activity and circadian variables from time-series data using automated algorithms and further conduct genetic analysis to deepen our understanding of the underlying genetic structure. Our study demonstrates the effectiveness of our methods and the utility of device-based activity

data in sleep and circadian studies. Our methods can help expand the application of wearable device data in health studies and further provide novel insights into the shared genetic architectures of sleep, activity, and circadian rhythms with metabolic and neurological traits.

Methods

Data

Data were collected from the UK Biobank study [78], a longitudinal population-based study with around 500,000 participants living in the UK. Genetic data from 487,409 participants were available when we accessed the UK Biobank data [78, 79] in September, 2017. The dataset includes around 96 million single nucleotide polymorphisms (SNPs), including imputation based on the UK10K haplotype, 1000 Genomes Phase 3, and Haplotype Reference Consortium (HRC) reference panels [79]. We applied filters to exclude SNPs with Minor Allele Frequency $< 0.1\%$, Hardy-Weinberg equilibrium $<$ 1e-10 and imputation quality score (UKBB information score [79]) < 0.8. After these steps, a total of 11,024,754 SNPs remained in further analyses.

Besides genetic data, a subset of 103,712 UK Biobank participants agreed to have their objective physical activity data [80] collected, and they were asked to wear an Axivity AX3 wrist accelerometer for seven consecutive days from 2013 to 2015. The device has been demonstrated to provide equivalent output to the GENEActiv accelerometer, which has been validated against free-living energy expenditure assessment methods [81–83]. This study was covered by the general ethical approval for UK Biobank studies from the National Research Ethics Service by National Health Service on 17th June 2011 (Reference 11/NW/0382). The accelerometer dataset that we acquired in September of 2017 consists of data in the activity count format summarized every five-second epoch from 103,706 participants. We applied similar quality control procedures as other accelerometer studies [84]. We excluded individuals with flagged data problems, poor wear time, poor calibration, recorded interrupted periods, or inability to calibrate activity data on the device worn itself requiring the use of other data. We also excluded individuals if the number of data recording errors was greater than 3rd quartile + 1.5×IQR. After pre-processing, 92,631 individuals remained, and 90,515 of them had genotyping data that were available for further genetic analyses.

Identification of loci associated with sleep and circadian rhythms

Sleep and circadian rhythm phenotypes were derived from accelerometer data. The analysis pipeline is summarized in Fig 1. For sleep, because it is a large-scale population study and no validation sleep logs are available, we need unsupervised algorithms to extract sleep features from accelerometer data. Specifically, we developed an unsupervised sleep-wake identification algorithm [77] based on Hidden Markov Model (HMM) [85, 86] to infer the sequence of "hidden states" of sleep or wake for each individual. This HMM can be directly applied to summary activity count data, which are widely used activity data formats that can save storage space, lengthen the duration of use of wearable devices, and increase computational efficiency. HMM assumes that in the sleep state or the wake state, activity counts follow different distributions as activity counts tend to be fewer in the sleep state compared to those in the wake state [77]. We used log transformation of activity counts and assumed that they follow different Gaussian distributions to infer the sequence of "hidden states" [77].

In our model, we consider the log transformed data: log(count+1) as the observed data, because the large range of observed activity counts from zero to several thousand per epoch poses both statistical and computational challenges in data analysis. Our empirical results suggest that the HMM algorithm works well for the log transformed data. We observe activity

count data from time 1 to time T: $O^{(T)} = \{O_1, O_2, \ldots, O_T\}$. Let $X^{(T)} = \{X_1, X_2, \ldots, X_T\}$ denote the sequence of the corresponding hidden states across these T time points, where each X_i can be one of the two possible hidden states $S = \{s_1, s_2\}$ in each epoch, with s_1 denoting the sleep state and s_2 denoting the wake state. We assume that X_i follows a Markov model, that is the hidden state X_{t+1} at time t+1 solely depends on X_t, and the observation O_t at time t solely depends on the hidden state X_t.

A denotes the 2 by 2 transition probability matrix, in which a_{ij} represents the transition probability from state s_i to state s_j. The emission probability $P(O_t|X_t)$ denoted by B depends on the state of X_t. If $X_t = s_1$ in the sleep state, we assume that the log transformed count follows zero-inflated truncated Gaussian distribution, which is truncated from 0 to the left. It has a zero component because sleep is associated with rare movements and activity measurements during sleep often involve many zeros. Therefore, the emission probability $b_1(0)$ of observing 0 and the emission probability $b_1(k)$ of observing k are as follows:

$$b_1(0) = P(O_t = 0|X_t = s_1) = \alpha + (1-\alpha) \cdot \frac{\frac{1}{\sigma_1}\phi\left(\frac{0-\mu_1}{\sigma_1}\right)}{1 - \Phi\left(\frac{0-\mu_1}{\sigma_1}\right)}$$

$$b_1(k) = P(O_t = k|X_t = s_1) = (1-\alpha) \cdot \frac{\frac{1}{\sigma_1}\phi\left(\frac{k-\mu_1}{\sigma_1}\right)}{1 - \Phi\left(\frac{0-\mu_1}{\sigma_1}\right)}$$

where α is the probability of extra zeros, μ_1 is the mean, σ_1 is the standard deviation, $\phi(\cdot)$is the probability density function of the standard normal distribution, and $\Phi(\cdot)$ is its cumulative distribution function.

If $X_t = s_2$ in the wake state, we assume that the log transformed count follows the Gaussian distribution:

$$b_2(k) = P(O_t = k|X_t = s_2) = \frac{1}{\sigma_2}\phi\left(\frac{k-\mu_2}{\sigma_2}\right)$$

where μ_2 is the mean and σ_2 is the standard deviation of the Gaussian distribution.

Therefore, the set of parameters for the emission probability is $B = \{B_1, B_2\} = \{\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2\}$. To initiate the Markov chain, we also need the initial state probabilities $\Pi = \{\pi_0, \pi_1\}$. HMM can be fully specified by $\Theta = \{A, B, \Pi\}$. To obtain $\Theta^* = \text{argmax}_\Theta P\{O^{(T)}|\Theta\}$, we can use the Baum-Welch algorithm, and we further look for the optimal path of hidden states $X^{(T)^*} = \text{argmax}_{X^{(T)}} P\{X^{(T)}, O^{(T)}\Theta^*\}$ using the Viterbi algorithm [87]. The optimal hidden states $X^{(T)^*}$ are exactly the sequence of inferred sleep/wake states. R code for implementing the HMM algorithm is at https://github.com/xinyue-L/hmmacc.

Besides the inferred sequence of sleep/wake states, the HMM parameters estimated for each individual, including mean activity levels and variability (standard deviation) for sleep and wake states, can characterize individual sleep and activity behaviors and therefore were used as sleep and activity phenotypes. In addition, we inferred sleep duration, sleep start and end based on the inferred sequence of sleep/wake states. We created two categorical variables as sleep duration phenotypes indicating whether the sleep duration is < 5 hours or > 10 hours, the thresholds of which come from the National Sleep Foundation [88] and are not recommended for middle-age and older adults. The inferred sleep start time and sleep end time are used as phenotypes for the timing of sleep onset and wake-up. Sleep start and sleep end are related to but not exactly the same as chronotypes, which describe whether a person is a morning person, getting up early and remaining more active in the day, or a night person,

remaining more active later of the day and staying up late at night, and chronotypes are usually measured in self-reported questionnaires. Published work [89] using UK Biobank data has utilized midpoint of sleep, the least active 5 hours of the day, as sleep timing to be compared with self-reported chronotypes. Here we did not replicate the study but examined timing of sleep onset and wake-up, as for different types of sleep disorders some people have difficulty falling asleep while others find it problematic to wake up too early [90].

For circadian rhythm characteristics, we derived circadian features by utilizing the Penalized Multi-Band Learning (PML) approach [43], which extracts periodic information using Fast Fourier Transform (FFT) and then performs penalized selection based on regularization, a classic approach used in machine learning [91, 92], to identify population-level dominant periodicities such as 1-day, 1/2-day, and 1/3-day periodicities that can characterize daily activity rhythms. The strengths of FFT signals at dominant periodicities are then used as circadian phenotypes in genetic analysis.

The PML algorithm is briefly described as follows [43]. Let matrix $X \in R^{n \times p}$, where n denotes the number of individual observations, and p denotes the number of periodicities from FFT. Specifically, $X = (x_1, x_2, \cdots, x_p)$, where x_j is the vector of length n for the jth periodicity.

Let Θ be the diagonal matrix selecting columns from X such that $\hat{X} = X\Theta$:

$$\Theta = \begin{pmatrix} \theta_{1,1} & 0 & \cdots & 0 \\ 0 & \theta_{2,2} & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_{p,p} \end{pmatrix}$$

where $0 \leq \theta_{j,j} \leq 1$, $j = 1, \ldots, p$. Θ identifies columns of dominant periodicities from X in the way that dominant periodicities corresponding to nonzero $\theta_{j,j}$'s are selected. We minimize the squared Frobenius norm $||X - \hat{X}||_F^2$, and by using properties of the Frobenius norm, we can get:

$$\begin{aligned} ||X - \hat{X}||_F^2 &= ||X - X\Theta||_F^2 = \mathrm{tr}((X - X\Theta)^T(X - X\Theta)) \\ &= \mathrm{tr}(X^TX - X^TX\Theta - \Theta^TX^TX + \Theta^TX^TX\Theta) \end{aligned}$$

Because X^TX is fixed, it is equivalent to minimize:

$$\mathrm{tr}(X^TX\Theta - \Theta^TX^TX + \Theta^TX^TX\Theta) = -2\sum_j \theta_{j,j}||x_j||^2 + \sum_j \theta_{j,j}^2||x_j||^2$$

In order to estimate Θ and identify dominant periodicities, we use a penalized selection method similar to Lasso, a widely used method in shrinkage and feature selection in regression models that is most effective in selecting a few important features while suppressing other non-selected features to 0 [92]. In our case, Lasso penalty serves to select a few dominant periodicities through diagonal elements of Θ instead of regression coefficients. Further, we add an elastic-net like penalty term onto the Frobenius norm, namely a combination of L1 and L2 norms [91]:

$$g(\theta) = -2\sum_j \theta_{j,j}||x_j||^2 + \sum_j \theta_{j,j}^2||x_j||^2 + \lambda\left(\frac{1-\alpha}{2}\sum_j \theta_{j,j}^2 + \alpha\sum_j \theta_{j,j}\right)$$

where λ is the tuning parameter and α controls the balance between the L1 and L2 norms. Note that $\theta_{j,j}$'s are nonnegative and thus we do not need to take the absolute value for the L1

norm. By setting λ large enough, all diagonal elements of Θ, namely all $\theta_{j,j}'s$, are suppressed to zero and no periodicities are selected. As λ decreases, some $\theta_{j,j}'s$ become nonzero and they correspond to the most dominant periodicities that are selected sequentially according to how dominant they are.

To minimize $g(\theta)$, we take the partial derivative of $g(\theta)$ with respect to each $\theta_{k,k}$: $\frac{\partial g(\theta)}{\partial \theta_{k,k}} = -2||x_k||^2 + 2\theta_{k,k}||x_k||^2 + (1-\alpha)\lambda\theta_{k,k} + \alpha\lambda$, which is convex and also subject to the constraint $0 \leq \theta_{k,k} \leq 1$. Thus, we have:

$$\hat{\theta}_{k,k} = \arg\min g(\theta) = \max\left(\frac{2||x_k||^2 - \alpha\lambda}{2||x_k||^2 - (1-\alpha)\lambda}, 0\right)$$

If we only have the L1 penalty, $\alpha = 1$ and $\hat{\theta}_{k,k} = \max\left(\frac{2||x_k||^2 - \lambda}{2||x_k||^2}, 0\right)$. In our case, we use Lasso L1 penalty alone and train λ, because we want to select the most important periodicities while suppressing other periodicities to 0.

We use mean squared error (MSE), which is equivalent to the squared Frobenius norm $||X - \hat{X}||_F^2$, as the criterion for choosing λ and the number of nonzero $\theta_{j,j}'s$ (the number of dominant periodicities selected). We train λ from $2 \cdot \max_{1 \leq j \leq p}(||x_j||^2)$ to 0, as $\lambda = 2 \cdot \max_{1 \leq j \leq p}(||x_j||^2)$ suppresses all $\theta_{j,j}$'s to 0 and $\lambda = 0$ gives no penalty. By decreasing λ, we identify dominant periodicities sequentially to characterize the daily sleep-activity rhythm. An R package named PML has been developed (https://CRAN.R-project.org/package=PML) for the implementation of the PML algorithm [93].

To identify genetic loci associated with each sleep and circadian phenotype, we conducted genome-wide association analysis, using PLINK (version 1.9) [94]. We included age, sex, and the first 20 principal components as covariates when fitting linear models, and we report statistically significant loci using a traditional threshold of 5×10^{-8}. We also highlight results from a more stringent threshold of 5×10^{-9} to take into account multiple testing using Bonferroni correction, which is the same threshold suggested in previous studies [14, 95].

Genetic architecture of sleep and circadian rhythm

To estimate heritability [18], we applied LD score regression [18, 53] with the LDSC tool. We also conducted partitioned heritability analysis [53] across tissue categories using the same tool. Significant enrichments for individual traits were identified using the Bonferroni corrected threshold of $p < 5 \times 10^{-4}$ (10 traits ×10 tissue types).

Genetic correlation of sleep and daily rhythms with other traits

To examine the genetic correlation of sleep-activity traits and circadian rhythm traits with other traits and diseases, we downloaded GWAS summary statistics in the second round of results from the Neale lab released on August 1, 2018 (http://www.nealelab.is/uk-biobank/). Specifically, we chose activity related traits including time spent doing moderate or vigorous physical activity, screen exposure traits including time spent watching television, computer, or mobile phone, sleep traits such as self-reported sleep duration, chronotypes, and insomnia, mental health traits related to anxiety and depression, BMI and diet traits, alcohol consumption traits, shift-work traits, and respiratory disease treats. The detailed list of 53 traits are shown in S3 Table. We calculated cross-trait genetic correlation using GNOVA [96] and highlight the trait-pairs with Bonferroni corrected p-values $< 7.9 \times 10^{-5}$ (12 sleep and circadian traits and 53 other traits).

For significantly correlated trait-pairs, we further investigated causal relationships by conducting bi-directional Mendelian Randomization (MR) analyses. We did not analyze trait-pairs related to sleep and activity, which have been studied elsewhere [14, 17], but primarily focused on the circadian rhythm. We performed two-sample MR analysis using publicly available GWAS summary data extracted from the MR-Base web platform [97]. We used leave-one-out analysis and single-SNP analysis as sensitivity analyses and considered the inverse-variance weighted (IVW) approach, MR-Egger [98], weighted median estimation and weighted mode estimation methods to examine whether there are consistent MR results across methods.

Cross-tissue transcriptome-wide association analysis

To investigate functional and biological mechanisms underlying sleep-activity and circadian rhythms, we applied UTMOST [60] that utilizes GWAS summary statistics and integrates eQTL information to perform tissue enrichment analysis in 44 tissue types and identify single-tissue and cross-tissue gene-trait associations. The cross-tissue gene-trait association is evaluated via a joint test summarizing single-tissue association statistics and quantifying the overall gene-trait association. The UTMOST [60] p-value threshold after Bonferroni correction for 15,120 genes is 3.3×10^{-6}. Gene ontology enrichment analysis was further conducted using DAVID [68].

Supporting information

S1 Fig. Manhattan plots for genome-wide association studies of sleep and activity traits, derived sleep traits, and circadian traits.
(PDF)

S2 Fig. QQ-plots for checking population stratification in genome-wide association studies.
(DOCX)

S1 Table. The SNPs identified in genome-wide association studies at the significance level of 5×10^{-8} that are associated with sleep start and sleep end traits inferred from accelerometer-measured physical activity in 90,515 UK Biobank participants.
(DOCX)

S2 Table. The SNPs identified in genome-wide association studies at the significance level of 5×10^{-8} that are associated with dominant periodicities as circadian traits inferred from accelerometer-measured physical activity in 90,515 UK Biobank participants.
(DOCX)

S3 Table. Estimated genetic correlation of sleep and circadian traits with other complex traits and the corresponding p-values.
(DOCX)

S4 Table. Single-tissue gene-trait association test results from tissue enrichment analysis using UTMOST.
(DOCX)

S5 Table. Cross-tissue gene-trait association test results from tissue enrichment analysis using UTMOST.
(DOCX)

S6 Table. Two-sample Mendelian Randomization analysis for the strength of circadian rhythm using GWAS summary statistics from the GIANT study.
(DOCX)

Acknowledgments

This study was conducted using the UK Biobank resource under application number 29900. We would like to thank all individuals who participated in the UK Biobank study.

Author Contributions

Conceptualization: Xinyue Li, Hongyu Zhao.

Data curation: Xinyue Li.

Formal analysis: Xinyue Li.

Funding acquisition: Hongyu Zhao.

Investigation: Xinyue Li, Hongyu Zhao.

Methodology: Xinyue Li, Hongyu Zhao.

Project administration: Xinyue Li.

Resources: Xinyue Li, Hongyu Zhao.

Software: Xinyue Li.

Supervision: Hongyu Zhao.

Validation: Xinyue Li.

Visualization: Xinyue Li.

Writing – original draft: Xinyue Li, Hongyu Zhao.

Writing – review & editing: Xinyue Li, Hongyu Zhao.

References

1. Fernandez-Mendoza J. The insomnia with short sleep duration phenotype: an update on it's importance for health and prevention. Curr Opin Psychiatry. 2017; 30(1):56–63. https://doi.org/10.1097/YCO.0000000000000292 PMID: 27764017
2. Fernandez-Mendoza J, Vgontzas AN. Insomnia and its impact on physical and mental health. Curr Psychiatry Rep. 2013; 15(12):418. https://doi.org/10.1007/s11920-013-0418-8 PMID: 24189774
3. Luyster FS, Strollo PJ Jr, Zee PC, Walsh JK, Boards of Directors of the American Academy of Sleep M, the Sleep Research S. Sleep: a health imperative. Sleep. 2012; 35(6):727–34. https://doi.org/10.5665/sleep.1846 PMID: 22654183
4. Sterniczuk R, Theou O, Rusak B, Rockwood K. Sleep disturbance is associated with incident dementia and mortality. Curr Alzheimer Res. 2013; 10(7):767–75. https://doi.org/10.2174/15672050113109990134 PMID: 23905991
5. Zornoza-Moreno M, Fuentes-Hernandez S, Sanchez-Solis M, Rol MA, Larque E, Madrid JA. Assessment of circadian rhythms of both skin temperature and motor activity in infants during the first 6 months of life. Chronobiol Int. 2011; 28(4):330–7. https://doi.org/10.3109/07420528.2011.565895 PMID: 21539424
6. Zhu L, Zee PC. Circadian rhythm sleep disorders. Neurol Clin. 2012; 30(4):1167–91. https://doi.org/10.1016/j.ncl.2012.08.011 PMID: 23099133
7. Potter GD, Skene DJ, Arendt J, Cade JE, Grant PJ, Hardie LJ. Circadian Rhythm and Sleep Disruption: Causes, Metabolic Consequences, and Countermeasures. Endocr Rev. 2016; 37(6):584–608. https://doi.org/10.1210/er.2016-1083 PMID: 27763782

8. Baron KG, Reid KJ. Circadian misalignment and health. Int Rev Psychiatry. 2014; 26(2):139–54. https://doi.org/10.3109/09540261.2014.911149 PMID: 24892891

9. Smith MT, McCrae CS, Cheung J, Martin JL, Harrod CG, Heald JL, et al. Use of Actigraphy for the Evaluation of Sleep Disorders and Circadian Rhythm Sleep-Wake Disorders: An American Academy of Sleep Medicine Systematic Review, Meta-Analysis, and GRADE Assessment. J Clin Sleep Med. 2018; 14(7):1209–30. https://doi.org/10.5664/jcsm.7228 PMID: 29991438

10. van Hees VT, Sabia S, Anderson KN, Denton SJ, Oliver J, Catt M, et al. A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer. PLoS One. 2015; 10(11):e0142533. https://doi.org/10.1371/journal.pone.0142533 PMID: 26569414

11. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. Sleep. 1992; 15(5):461–9. https://doi.org/10.1093/sleep/15.5.461 PMID: 1455130

12. Sadeh A, Sharkey KM, Carskadon MA. Activity-based sleep-wake identification: an empirical test of methodological issues. Sleep. 1994; 17(3):201–7. https://doi.org/10.1093/sleep/17.3.201 PMID: 7939118

13. Tilmanne J, Urbain J, Kothare MV, Wouwer AV, Kothare SV. Algorithms for sleep–wake identification using actigraphy: a comparative study and new results. Journal of Sleep Research. 2009; 18(1):85–98. https://doi.org/10.1111/j.1365-2869.2008.00706.x PMID: 19250177

14. Doherty A, Smith-Byrne K, Ferreira T, Holmes MV, Holmes C, Pulit SL, et al. GWAS identifies 14 loci for device-measured physical activity and sleep duration. Nature Communications. 2018; 9(1):5257. https://doi.org/10.1038/s41467-018-07743-4 PMID: 30531941

15. Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. Sci Rep. 2018; 8(1):7961. https://doi.org/10.1038/s41598-018-26174-1 PMID: 29784928

16. Dashti HS, Jones SE, Wood AR, Lane JM, van Hees VT, Wang H, et al. Genome-wide association study identifies genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates. Nature Communications. 2019; 10(1):1100. https://doi.org/10.1038/s41467-019-08917-4 PMID: 30846698

17. Jones SE, van Hees VT, Mazzotti DR, Marques-Vidal P, Sabia S, van der Spek A, et al. Genetic studies of accelerometer-based sleep measures yield new insights into human sleep behaviour. Nat Commun. 2019; 10(1):1585. https://doi.org/10.1038/s41467-019-09576-1 PMID: 30952852

18. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015; 47 (3):291–5. https://doi.org/10.1038/ng.3211 PMID: 25642630

19. Lane JM, Liang J, Vlasac I, Anderson SG, Bechtold DA, Bowden J, et al. Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. Nature genetics. 2017; 49(2):274. https://doi.org/10.1038/ng.3749 PMID: 27992416

20. Winkelmann J, Czamara D, Schormair B, Knauf F, Schulte EC, Trenkwalder C, et al. Genome-wide association study identifies novel restless legs syndrome susceptibility loci on 2p14 and 16q12. 1. PLoS genetics. 2011; 7(7):e1002171. https://doi.org/10.1371/journal.pgen.1002171 PMID: 21779176

21. Jansen PR, Watanabe K, Stringer S, Skene N, Bryois J, Hammerschlag AR, et al. Genome-wide Analysis of Insomnia (N = 1,331,010) Identifies Novel Loci and Functional Pathways. bioRxiv. 2018:214973.

22. Lane JM, Vlasac I, Anderson SG, Kyle SD, Dixon WG, Bechtold DA, et al. Genome-wide association analysis identifies novel loci for chronotype in 100,420 individuals from the UK Biobank. Nature communications. 2016; 7:10889. https://doi.org/10.1038/ncomms10889 PMID: 26955885

23. McDonald MN, Won S, Mattheisen M, Castaldi PJ, Cho MH, Rutten E, et al. Body mass index change in gastrointestinal cancer and chronic obstructive pulmonary disease is associated with Dedicator of Cytokinesis 1. J Cachexia Sarcopenia Muscle. 2017; 8(3):428–36. https://doi.org/10.1002/jcsm.12171 PMID: 28044437

24. Wijsman EM, Pankratz ND, Choi Y, Rothstein JH, Faber KM, Cheng R, et al. Genome-wide association of familial late-onset Alzheimer's disease replicates BIN1 and CLU and nominates CUGBP2 in interaction with APOE. PLoS Genet. 2011; 7(2):e1001308. https://doi.org/10.1371/journal.pgen.1001308 PMID: 21379329

25. Hoffmann TJ, Choquet H, Yin J, Banda Y, Kvale MN, Glymour M, et al. A Large Multiethnic Genome-Wide Association Study of Adult Body Mass Index Identifies Novel Loci. Genetics. 2018; 210(2):499–515. https://doi.org/10.1534/genetics.118.301479 PMID: 30108127

26. Winkler TW, Justice AE, Graff M, Barata L, Feitosa MF, Chu S, et al. The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. PLoS Genet. 2015; 11(10):e1005378. https://doi.org/10.1371/journal.pgen.1005378 PMID: 26426971

27. Huffman JE, Albrecht E, Teumer A, Mangino M, Kapur K, Johnson T, et al. Modulation of genetic associations with serum urate levels by body-mass-index in humans. PLoS One. 2015; 10(3):e0119752. https://doi.org/10.1371/journal.pone.0119752 PMID: 25811787

28. Wu JH, Lemaitre RN, Manichaikul A, Guan W, Tanaka T, Foy M, et al. Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: results from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. Circ Cardiovasc Genet. 2013; 6(2):171–83. https://doi.org/10.1161/CIRCGENETICS.112.964619 PMID: 23362303

29. Schormair B, Zhao C, Bell S, Tilch E, Salminen AV, Putz B, et al. Identification of novel risk loci for restless legs syndrome in genome-wide association studies in individuals of European ancestry: a meta-analysis. Lancet Neurol. 2017; 16(11):898–907. https://doi.org/10.1016/S1474-4422(17)30327-7 PMID: 29029846

30. Wood AR, Tyrrell J, Beaumont R, Jones SE, Tuke MA, Ruth KS, et al. Variants in the FTO and CDKAL1 loci have recessive effects on risk of obesity and type 2 diabetes, respectively. Diabetologia. 2016; 59 (6):1214–21. https://doi.org/10.1007/s00125-016-3908-5 PMID: 26961502

31. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015; 518(7538):197–206. https://doi.org/10.1038/nature14177 PMID: 25673413

32. Herold C, Hooli BV, Mullin K, Liu T, Roehr JT, Mattheisen M, et al. Family-based association analyses of imputed genotypes reveal genome-wide significant association of Alzheimer's disease with OSBPL6, PTPRG, and PDCL3. Mol Psychiatry. 2016; 21(11):1608–12. https://doi.org/10.1038/mp.2015.218 PMID: 26830138

33. Goes FS, McGrath J, Avramopoulos D, Wolyniec P, Pirooznia M, Ruczinski I, et al. Genome-wide association study of schizophrenia in Ashkenazi Jews. Am J Med Genet B Neuropsychiatr Genet. 2015; 168 (8):649–59. https://doi.org/10.1002/ajmg.b.32349 PMID: 26198764

34. Autism Spectrum Disorders Working Group of The Psychiatric Genomics C. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. Mol Autism. 2017; 8:21. https://doi.org/10.1186/s13229-017-0137-9 PMID: 28540026

35. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016; 167(5):1415–29 e19. https://doi.org/10.1016/j.cell.2016.10.042 PMID: 27863252

36. Loprinzi PD. Sleep duration and sleep disorder with red blood cell distribution width. Am J Health Behav. 2015; 39(4):471–4. https://doi.org/10.5993/AJHB.39.4.3 PMID: 26018095

37. Liu H, Wang G, Luan G, Liu Q. Effects of sleep and sleep deprivation on blood cell count and hemostasis parameters in healthy humans. J Thromb Thrombolysis. 2009; 28(1):46–9. https://doi.org/10.1007/s11239-008-0240-z PMID: 18597046

38. Choi JB, Loredo JS, Norman D, Mills PJ, Ancoli-Israel S, Ziegler MG, et al. Does obstructive sleep apnea increase hematocrit? Sleep Breath. 2006; 10(3):155–60. https://doi.org/10.1007/s11325-006-0064-z PMID: 16770648

39. Akiyama M, Okada Y, Kanai M, Takahashi A, Momozawa Y, Ikeda M, et al. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat Genet. 2017; 49 (10):1458–67.

40. van der Harst P, Verweij N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circ Res. 2018; 122(3):433–43.

41. Chen C, Xia F, Chen Y, Zhang K, Cheng J, Li Q, et al. Association Between Thyroid-Stimulating Hormone and Renal Function: a Mendelian Randomization Study. Kidney Blood Press Res. 2018; 43 (4):1121–30. https://doi.org/10.1159/000491808 PMID: 30016786

42. Zhan M, Chen G, Pan CM, Gu ZH, Zhao SX, Liu W, et al. Genome-wide association study identifies a novel susceptibility gene for serum TSH levels in Chinese populations. Hum Mol Genet. 2014; 23 (20):5505–17. https://doi.org/10.1093/hmg/ddu250 PMID: 24852370

43. Li X, Kane M, Zhang Y, Sun W, Song Y, Dong S, et al. Penalized Selection of Periodicities Characterizes the Consolidation of Sleep-Wake Circadian Rhythms During Early Childhood Development. Submitted. 2019.

44. Byrne EM, Gehrman PR, Medland SE, Nyholt DR, Heath AC, Madden PA, et al. A genome-wide association study of sleep habits and insomnia. Am J Med Genet B Neuropsychiatr Genet. 2013; 162B (5):439–51. https://doi.org/10.1002/ajmg.b.32168 PMID: 23728906

45. Heinzman JT, Hoth KF, Cho MH, Sakornsakolpat P, Regan EA, Make BJ, et al. GWAS and systems biology analysis of depressive symptoms among smokers from the COPDGene cohort. J Affect Disord. 2019; 243:16–22. https://doi.org/10.1016/j.jad.2018.09.003 PMID: 30219690

46. Li QS, Tian C, Seabrook GR, Drevets WC, Narayan VA. Analysis of 23andMe antidepressant efficacy survey data: implication of circadian rhythm and neuroplasticity in bupropion response. Transl Psychiatry. 2016; 6(9):e889. https://doi.org/10.1038/tp.2016.171 PMID: 27622933

47. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 2018; 50(2):229–37.

48. Saxena R, Plenge RM, Bjonnes AC, Dashti HS, Okada Y, Gad El Haq W, et al. A Multinational Arab Genome-Wide Association Study Identifies New Genetic Associations for Rheumatoid Arthritis. Arthritis Rheumatol. 2017; 69(5):976–85. https://doi.org/10.1002/art.40051 PMID: 28118524

49. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511(7510):421–7. https://doi.org/10.1038/nature13595 PMID: 25056061

50. Pardinas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet. 2018; 50(3):381–9. https://doi.org/10.1038/s41588-018-0059-2 PMID: 29483656

51. Hu Y, Shmygelska A, Tran D, Eriksson N, Tung JY, Hinds DA. GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. Nat Commun. 2016; 7:10448. https://doi.org/10.1038/ncomms10448 PMID: 26835600

52. Eppinga RN, Hagemeijer Y, Burgess S, Hinds DA, Stefansson K, Gudbjartsson DF, et al. Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. Nat Genet. 2016; 48(12):1557–63. https://doi.org/10.1038/ng.3708 PMID: 27798624

53. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015; 47 (11):1228–35.

54. Gamble KL, Berry R, Frank SJ, Young ME. Circadian clock control of endocrine factors. Nat Rev Endocrinol. 2014; 10(8):466–75. https://doi.org/10.1038/nrendo.2014.78 PMID: 24863387

55. Trinder J, Waloszek J, Woods MJ, Jordan AS. Sleep and cardiovascular regulation. Pflugers Arch. 2012; 463(1):161–8. https://doi.org/10.1007/s00424-011-1041-3 PMID: 22038322

56. Benarroch EE. Control of the cardiovascular and respiratory systems during sleep. Auton Neurosci. 2019; 218:54–63. https://doi.org/10.1016/j.autneu.2019.01.007 PMID: 30890349

57. Hyun MK, Baek Y, Lee S. Association between digestive symptoms and sleep disturbance: a cross-sectional community-based study. BMC Gastroenterol. 2019; 19(1):34. https://doi.org/10.1186/s12876-019-0945-9 PMID: 30782128

58. Murat S, Ali U, Serdal K, Suleyman D, Ilknur P, Mehmet S, et al. Assessment of subjective sleep quality in iron deficiency anaemia. Afr Health Sci. 2015; 15(2):621–7. https://doi.org/10.4314/ahs.v15i2.40 PMID: 26124812

59. Li X, Allen RP, Earley CJ, Liu H, Cruz TE, Edden RAE, et al. Brain iron deficiency in idiopathic restless legs syndrome measured by quantitative magnetic susceptibility at 7 tesla. Sleep Med. 2016; 22:75–82. https://doi.org/10.1016/j.sleep.2016.05.001 PMID: 27544840

60. Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nat Genet. 2019; 51(3):568–76. https://doi.org/10.1038/s41588-019-0345-7 PMID: 30804563

61. Kjellberg MA, Backman AP, Ohvo-Rekila H, Mattjus P. Alternation in the glycolipid transfer protein expression causes changes in the cellular lipidome. PLoS One. 2014; 9(5):e97263.

62. Carbon S, Mungall C. Gene Ontology Data Archive. 2018.

63. Hokama M, Oka S, Leon J, Ninomiya T, Honda H, Sasaki K, et al. Altered expression of diabetes-related genes in Alzheimer's disease brains: the Hisayama study. Cereb Cortex. 2014; 24(9):2476–88. https://doi.org/10.1093/cercor/bht101 PMID: 23595620

64. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell. 2012; 149(6):1393–406. https://doi.org/10.1016/j.cell.2012.04.031 PMID: 22658674

65. Schroeder CM, Ostrem JM, Hertz NT, Vale RD. A Ras-like domain in the light intermediate chain bridges the dynein motor to a cargo-binding region. Elife. 2014; 3:e03351. https://doi.org/10.7554/eLife.03351 PMID: 25272277

66. Hodgson U, Pulkkinen V, Dixon M, Peyrard-Janvid M, Rehn M, Lahermo P, et al. ELMOD2 is a candidate gene for familial idiopathic pulmonary fibrosis. Am J Hum Genet. 2006; 79(1):149–54. https://doi.org/10.1086/504639 PMID: 16773575

67. Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res. 2009; 37(Database issue):D32–6. https://doi.org/10.1093/nar/gkn721 PMID: 18927115

68. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4(1):44–57. https://doi.org/10.1038/nprot.2008.211 PMID: 19131956

69. Bogan RK. Effects of restless legs syndrome (RLS) on sleep. Neuropsychiatr Dis Treat. 2006; 2 (4):513–9. https://doi.org/10.2147/nedt.2006.2.4.513 PMID: 19412499

70. Hargens TA, Kaleth AS, Edwards ES, Butner KL. Association between sleep disorders, obesity, and exercise: a review. Nat Sci Sleep. 2013; 5:27–35. https://doi.org/10.2147/NSS.S34838 PMID: 23620691

71. Philippe J, Dibner C. Thyroid circadian timing: roles in physiology and thyroid malignancies. J Biol Rhythms. 2015; 30(2):76–83.

72. Vandewalle G, Middleton B, Rajaratnam SM, Stone BM, Thorleifsdottir B, Arendt J, et al. Robust circadian rhythm in heart rate and its variability: influence of exogenous melatonin and photoperiod. J Sleep Res. 2007; 16(2):148–55.

73. Klimentidis YC, Raichlen DA, Bea J, Garcia DO, Wineinger NE, Mandarino LJ, et al. Genome-wide association study of habitual physical activity in over 377,000 UK Biobank participants identifies multiple variants including CADM2 and APOE. Int J Obes (Lond). 2018; 42(6):1161–76.

74. Oh JH. Gastroesophageal reflux disease: recent advances and its association with sleep. Ann N Y Acad Sci. 2016; 1380(1):195–203. https://doi.org/10.1111/nyas.13143 PMID: 27391766

75. Tu Q, Heitkemper MM, Jarrett ME, Buchanan DT. Sleep disturbances in irritable bowel syndrome: a systematic review. Neurogastroenterol Motil. 2017; 29(3).

76. Prince SA, Cardilli L, Reed JL, Saunders TJ, Kite C, Douillette K, et al. A comparison of self-reported and device measured sedentary behaviour in adults: a systematic review and meta-analysis. Int J Behav Nutr Phys Act. 2020; 17(1):31. https://doi.org/10.1186/s12966-020-00938-3 PMID: 32131845

77. Li X, Zhang Y, Jiang F, Zhao H. A novel machine learning unsupervised algorithm for sleep/wake identification using actigraphy. Chronobiology International. 2020;1–14.

78. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015; 12(3):e1001779. https://doi.org/10.1371/journal.pmed.1001779 PMID: 25826379

79. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. Genome-wide genetic data on~ 500,000 UK Biobank participants. BioRxiv. 2017:166298.

80. Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, et al. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. PLoS One. 2017; 12(2):e0169649. https://doi.org/10.1371/journal.pone.0169649 PMID: 28146576

81. White T, Westgate K, Wareham NJ, Brage S. Estimation of Physical Activity Energy Expenditure during Free-Living from Wrist Accelerometry in UK Adults. PLoS One. 2016; 11(12):e0167472. https://doi.org/10.1371/journal.pone.0167472 PMID: 27936024

82. Esliger DW, Rowlands AV, Hurst TL, Catt M, Murray P, Eston RG. Validation of the GENEA Accelerometer. Med Sci Sports Exerc. 2011; 43(6):1085–93. https://doi.org/10.1249/MSS.0b013e31820513be PMID: 21088628

83. Rowlands AV, Mirkes EM, Yates T, Clemes S, Davies M, Khunti K, et al. Accelerometer-assessed Physical Activity in Epidemiology: Are Monitors Equivalent? Med Sci Sports Exerc. 2018; 50(2):257–65. https://doi.org/10.1249/MSS.0000000000001435 PMID: 28976493

84. Jones SE, van Hees VT, Mazzotti DR, Marques-Vidal P, Sabia S, van der Spek A, et al. Genetic studies of accelerometer-based sleep measures in 85,670 individuals yield new insights into human sleep behaviour. bioRxiv. 2018:303925.

85. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. Annals of Mathematical Statistics. 1966; 37(6):1554–63.

86. Baum LE, Petrie T, Soules G, Weiss N. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. Annals of Mathematical Statistics. 1970; 41(1):164–71.

87. Ryan MS, Nudd GR. The Viterbi Algorithm. Proc IEEE. 1973; 61(5):268–78.

88. Hirshkowitz M, Whiton K, Albert SM, Alessi C, Bruni O, DonCarlos L, et al. National Sleep Foundation's sleep time duration recommendations: methodology and results summary. Sleep Health. 2015; 1 (1):40–3. https://doi.org/10.1016/j.sleh.2014.12.010 PMID: 29073412

89. Jones SE, Lane JM, Wood AR, Van Hees VT, Tyrrell J, Beaumont RN, et al. Genome-wide association analyses of chronotype in 697,828 individuals provides new insights into circadian rhythms in humans and links to disease. BioRxiv. 2018:303941.

90. Thorpy MJ. Classification of sleep disorders. Neurotherapeutics. 2012; 9(4):687–701. https://doi.org/10.1007/s13311-012-0145-6 PMID: 22976557

91. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67(2):301–20.

92. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996:267–88.

93. Li X, Kane M. PML: Penalized Multi-Band Learning for Circadian Rhythm Analysis Using Actigraphy 2019 [Available from: https://CRAN.R-project.org/package=PML.

94. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4:7. https://doi.org/10.1186/s13742-015-0047-8 PMID: 25722852

95. Pulit SL, de With SA, de Bakker PI. Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. Genet Epidemiol. 2017; 41(2):145–51. https://doi.org/10.1002/gepi.22032 PMID: 27990689

96. Lu Q, Li B, Ou D, Erlendsdottir M, Powles RL, Jiang T, et al. A Powerful Approach to Estimating Annotation-Stratified Genetic Covariance via GWAS Summary Statistics. Am J Hum Genet. 2017; 101(6):939–64. https://doi.org/10.1016/j.ajhg.2017.11.001 PMID: 29220677

97. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7.

98. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015; 44(2):512–25. https://doi.org/10.1093/ije/dyv080 PMID: 26050253