UC San Diego
UC San Diego Previously Published Works

Title

Genome-wide association studies and cross-population meta-analyses investigating short and long sleep duration.

Permalink

Journal

Authors

Austin-Zimmerman, Isabelle
Levey, Daniel
Giannakopoulou, Olga
et al.

Publication Date

DOI

Peer reviewed

Genome-wide association studies and cross-population meta-analyses investigating short and long sleep duration

Isabelle Austin-Zimmerman [1,2], Daniel F. Levey [3,4], Olga Giannakopoulou [1,5], Joseph D. Deak [3,4], Marco Galimberti [3,4], Keyrun Adhikari [3,4], Hang Zhou [3,4], Spiros Denaxas[6], Haritz Irizar[1,7], Karoline Kuchenbaecker [1,5], Andrew McQuillin [1], the Million Veteran Program*, John Concato[8,9], Daniel J. Buysse[15], J. Michael Gaziano[16,17], Daniel J. Gottlieb[10,11,12], Renato Polimanti [3,4], Murray B. Stein [13,14], Elvira Bramon [1,18] & Joel Gelernter [3,4] ✉

Sleep duration has been linked to a wide range of negative health outcomes and to reduced life expectancy. We present genome-wide association studies of short (≤ 5 h) and long (≥ 10 h) sleep duration in adults of European (N = 445,966), African (N = 27,785), East Asian (N = 3141), and admixed-American (N = 16,250) ancestry from UK Biobank and the Million Veteran Programme. In a cross-population meta-analysis, we identify 84 independent loci for short sleep and 1 for long sleep. We estimate SNP-based heritability for both sleep traits in each ancestry based on population derived linkage disequilibrium (LD) scores using cov-LDSC. We identify positive genetic correlation between short and long sleep traits ($r_g = 0.16 \pm 0.04$; p = 0.0002), as well as similar patterns of genetic correlation with other psychiatric and cardiometabolic phenotypes. Mendelian randomisation reveals a directional causal relationship between short sleep and depression, and a bidirectional causal relationship between long sleep and depression.

Sleep is one of the most highly conserved traits across the animal kingdom, indicating a strong evolutionary requirement. It is an essential and fundamental property of neurons and networks across the brain[1,2]. Sleep occurs in any organism with even a very simple neuronal/glial network (e.g. *Cassiopeia*, *C. elegans*), and is preserved in subjects surviving lesions in any brain region[1,3–5]. However, many of the molecular processes underlying sleep remain unclear.

Human sleep can be characterised along dimensions such as duration, timing, efficiency, and regularity, each sometimes associated with adverse health outcomes. However, sleep duration has been most widely studied, and relates to outcomes including obesity, cardiovascular disease, and mortality[6]. Both unusually long and unusually short sleep duration have been related to multiple psychiatric conditions,

including major depressive disorder (MDD), anxiety, and psychosis, though a causal relationship between sleep duration and these disorders is not established[7–10].

Genetic research, and in particular genome-wide association studies (GWAS), may help elucidate some of the biological processes that underlie variability in sleep across individuals, by identifying risk loci associated with higher or lower-than-average sleep duration. Self-reported sleep duration is a complex trait, with a genetic component established through twin and family studies as well as several GWAS[8,9,11–15]. A recent GWAS in 446,118 European-ancestry (EUR) UK Biobank participants identified over 70 independent genetic loci associated with habitual, self-reported sleep duration (measured as a continuous trait reported in hour increments), as well as several linked

specifically to unusually long (nine hours or more) and short (six hours or less) sleep duration[7]. SNP-based heritability of sleep duration was reported to be 9.8%. This study, and several others, identified common variants at or near the *VRK2* and *PAX8* genes[8,16,17]. *VRK2* encodes a serine/threonine kinase protein which is essential to multiple signal transduction pathways[7,14,18]. Single nucleotide polymorphisms (SNPs) within this gene have been associated with a range of psychiatric disorders, such as schizophrenia and depression, as well as epilepsy and some cardiometabolic traits. *PAX8* is a transcription factor important in the development and function of the thyroid.

Like many GWAS, these studies have been conducted primarily in EUR participants. Replicating these findings in other populations, or identifying ancestry-specific risk loci, is essential for furthering our understanding of the biological mechanisms behind sleep, and the effects of sleep on biology.

The UK Biobank and Million Veteran Programme (MVP) represent two of the world's largest biobanks, both containing genetic data and a wide range of environmental and medical information. UK Biobank is a population-based study, including over 500,000 UK-based participants[19,20]. MVP is a US military sample[21], having recruited so far over 825,000 veterans. We conducted a cross-population meta-analysis of short and long sleep duration using GWAS results from UK Biobank and MVP. This study aimed to build on the existing understanding of the genetics of sleep duration and to take advantage of the diverse populations included in UK Biobank and MVP to consider risk loci across multiple populations, as well as ancestry-specific regions of interest.

Results
Sample
Table 1 outlines the age and sex distribution in each ancestry group across the two cohorts. UK Biobank has a higher proportion of women than men. As a US military veteran sample, men are heavily over-represented in MVP, especially in EUR and somewhat less so in the AFR, EAS and AMR samples. Both cohorts are adult samples, with UK Biobank specifically recruiting adults aged 40–70 years. Though there was no age restriction for recruitment to MVP, the median age in MVP is higher at 66 years versus 58 years in UK Biobank. AFR participants make up a higher percentage of the overall sample in MVP than UKB. Figure 1 summarises the distribution of sleep hours in each cohort.

Hours of daylight exposure
In the UK Biobank samples, higher levels of solar irradiation were significantly associated with shorter reported sleep duration, though the effect size is small (estimate $= -4.8 \times 10^{-4} \pm 5 \times 10^{-5}$ h, $p < 2 \times 10^{-16}$). In the MVP sample, based on annual irradiation data rather than monthly data, increased solar irradiation was not significantly associated with sleep duration (estimate $= -4.0 \times 10^{-3} \pm 3.0 \times 10^{-3}$ h, $p = 0.157$).

Ancestry-specific meta-analyses
A meta-analysis comparing short ($n = 47,054$) versus normal ($n = 382,950$) sleep duration EUR individuals from both cohorts identified 46 genomic risk loci that reached genome-wide significance (GWS) (Supplementary Data 1–6, Supplementary Figs. 17 and 19). Of these, 19 were previously associated with a variety of sleep-related phenotypes, including two (rs11693221 on chromosome 2 and rs6466488 on chromosome 7) which have been identified in GWAS on sleep-related phenotypes in independent samples (i.e., samples not including UK Biobank or MVP subjects[22,23] (Supplementary Data 1, 2).

A meta-analysis comparing long ($n = 15,962$) versus normal sleep duration in the EUR participants from both cohorts identified one genome-wide significant locus on chromosome 2 (rs62158206, OR $= 0.93 \pm 0.01$, $p = 3.6 \times 10^{-8}$) near the *PAX8* gene (Supplementary Figs. 17 and 20). This SNP has previously been identified in several GWAS of insomnia and sleep duration, including one study in an independent sample[8]. We conducted sensitivity analyses to consider the impact of sex and shift work patterns. We see highly consistent results when comparing these to our primary EUR analyses, although the reduced sample size limits the power to define genome-wide significant loci. In all cases, rg is close to 1 (Supplementary Material Sections 3, 4, and 5; Supplementary Figs. 28–31).

SNP-based heritability (h^2) was estimated to be 11.9% ($p = 2.45 \times 10^{-115}$) for short sleep, and 7.8% ($p = 1.61 \times 10^{-20}$) for long sleep. Inflation was within the expected range given the sample sizes and polygenicity of the traits in question, with LD intercept close to one (intercept short sleep $= 1.017 \pm 0.01$; intercept long sleep $= 0.99 \pm 0.01$) (Supplementary Data 11 and 12).

A meta-analysis comparing short ($n = 11,352$) versus normal ($n = 15,305$) sleep duration in AFR participants of both cohorts did not

Table 1 | Sample demographics and case/control status

Population[a]		UK Biobank	MVP
Total $N = 493,142$	N	293,037	200,100
	N_{short} (≤5 h) (%)	21,086 (7.2)	41,425 (20.7)
	N_{medium} (7–8 h) (%)	264,982 (90.4)	147,962 (73.9)
	N_{long} (≥10 h) (%)	6969 (2.4)	10,713 (5.4)
European (EUR) $N = 445,966$ (90.4%)	N (% female)	278,003 (54.1)	167,963 (7.4)
	Age (years)	Mean (SD): 56.8 (8.0) Median: 58	Mean (SD): 66.8 (11.6) Median: 67
	Sleep duration (hours)	Mean (SD): 7.3 (1.0) Median: 7	Mean: 7.1 (1.3) Median: 7
	N_{short} (≤5 h) (%)	18,915 (6.8)	28,139 (16.8)
	N_{medium} (7–8 h) (%)	252,567 (90.9)	130,383 (77.6)
	N_{long} (≥10 h) (%)	6521 (2.3)	9441 (5.6)
African (AFR) $N = 27,785$ (5.6%)	N (% female)	5657 (59.0)	22,128 (13.5)
	Age (years)	Mean (SD): 51.86 (8.1) Median: 50	Mean (SD): 60.7 (10.9) Median: 61
	Sleep duration (hours)	Mean (SD): 6.9 (1.6) Median: 7	Mean (SD): 6.4 (1.3) Median: 6
	N_{short} (≤5 h) (%)	1396 (24.7)	9956 (45.0)
	N_{medium} (7–8 h) (%)	4017 (71.0)	11,288 (51.0)
	N_{long} (≥10 h) (%)	244 (4.3)	884 (4.0)
Admixed American (AMR) $N = 16,250$ (3.3%)	N (% female)	7712 (56.7)	8538 (9.4)
	Age (years)	Mean (SD): 55.4 (8.3) Median: 56	Mean (SD): 60.6 (13.2) Median: 63
	Sleep duration (h)	Mean (SD): 7.2 (1.0) Median: 7	Mean (SD): 6.7 (1.3) Median: 7
	N_{short} (≤5 h) (%)	628 (8.1)	2824 (33.1)
	N_{medium} (7–8 h) (%)	6928 (89.8)	5371 (62.9)
	N_{long} (≥10 h) (%)	156 (2.0)	343 (4.0)
East Asian (EAS) $N = 3141$ (0.6%)	N (% female)	1670 (67.1)	1471 (10.8)
	Age (years)	Mean (SD): 51.9 (7.8) Median: 51	Mean (SD): 59.9 (15.8) Median: 63
	Sleep duration (h)	Mean (SD): 7.3 (1.1) Median: 7	Mean (SD): 6.5 (1.2) Median: 6
	N_{short} (≤5 h) (%)	152 (9.1)	506 (34.4)
	N_{medium} (7–8 h) (%)	1470 (88.0)	920 (62.5)
	N_{long} (≥10 h) (%)	48 (2.9)	45 (3.1)

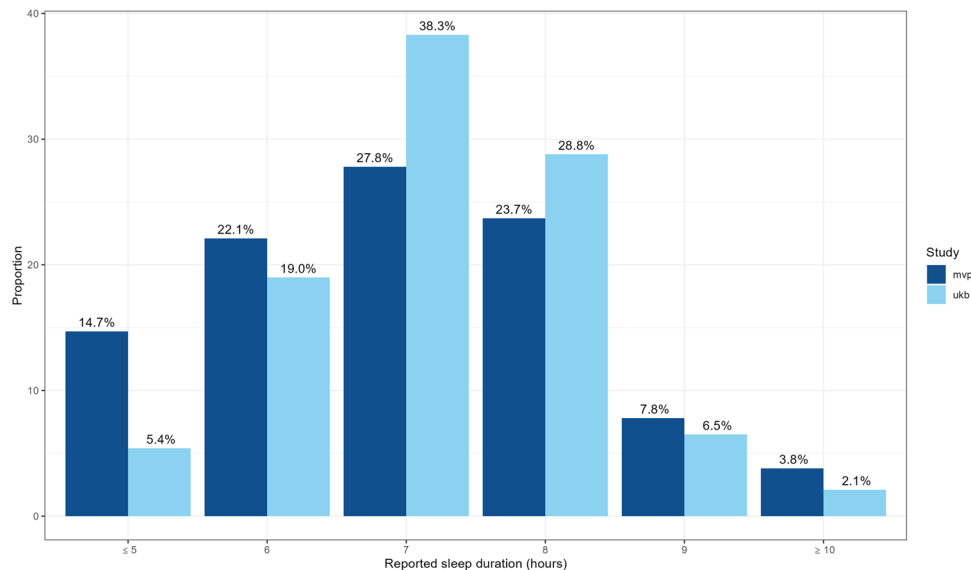[a]All groups defined based on reference panel[21].

Fig. 1 | Distribution of self-reported sleep duration in UK Biobank and MVP samples. Bars highlight the proportion of participants who reported sleeping each hour as a percentage of the total sample. UK Biobank sample represented in light blue, MVP in dark blue.

Table 2 | Summary of SNP-heritability estimates (liability scale) for the AFR population, derived using LD scores calculated using cov-LDSC, based on either UKB or MVP samples

SNP-h2 (SE); *p*-value	Short sleep using UKB-derived cov-LDSC scores	Short sleep using MVP-derived cov-LDSC scores	Long sleep using UKB-derived cov-LDSC scores	Long sleep using MVP-derived cov-LDSC scores
UK Biobank only	0.018 (0.23); 0.94	0.050 (0.158); 0.75	−0.47 (0.58); 0.42	−0.94 (0.41); 0.022
MVP only	0.14 (0.05); 0.012	0.077 (0.031); 0.012	−0.12 (0.30); 0.69	−0.048 (0.20); 0.81
UKB + MVP meta-analysis	0.088 (0.04); 0.039	0.075 (0.026); 0.0042	0.012 (0.19); 0.95	0.040 (0.12); 0.74

identify any GWS loci, though 42 loci reached a suggestive threshold of $p \leq 1 \times 10^{-5}$, with the strongest association at rs1412139 on chromosome 1 (OR = 1.1 ± 0.02, $p = 1.9 \times 10^{-7}$) (Supplementary Data 7–10, Supplementary Figs. 21 and 23).

A meta-analysis comparing long ($n = 1128$) versus normal sleep duration in AFR participants from both cohorts identified one GWS locus, rs148926968 on chromosome 13 (OR = 0.43 ± 0.1, $p = 2.6 \times 10^{-8}$) (Supplementary Figs. 21 and 24). SNP-based inflation was within the expected range given the sample sizes and polygenicity of the traits in question.

With LD scores calculated from UK Biobank data, we estimate the SNP-based heritability of short sleep duration in AFR to be 8.8 (intercept = 1.00 ± 0.01, $p = 0.04$), and using MVP LD score data, we estimate the SNP-based heritability of short sleep duration in AFR to be 7.5% (intercept = 0.99 ± 0.01, $p = 4.0 \times 10^{-3}$). SNP-based heritability for long sleep was not significant in this sample using either set of LD scores (Table 2 and Supplementary Data 11 and 12).

Detailed results for all primary analyses can be found in the supplementary material (Supplementary Material Section 1, Supplementary Figs. 1–16).

Cross-population meta-analyses
We conducted a cross-population meta-analysis incorporating both the EUR and AFR GWAS described previously, as well as data from smaller GWAS of EAS and AMR participants from both cohorts (see supplementary material for results of these primary GWAS). The results of both short versus normal and long versus normal sleep duration were filtered to remove any variant that was not present across all four population groups in at least one of the primary cohorts.

After filtering, the analysis of short ($n = 62,516$) versus normal ($n = 412,944$) sleep duration meta-analysis resulted in a total of

7,574,717 imputed genetic variants, among which we identified 84 independent GWS risk loci (Fig. 2, Supplementary Data 13). Of these 84, 13 have been previously associated with sleep-related phenotypes, including two in independent samples (rs62158206 on chromosome 2 and rs1989903 on chromosome 7)[8,23] (Supplementary Data 15).

After filtering, the meta-analysis of long ($n = 15,962$) versus normal sleep duration resulted in a total of 7,282,278 imputed genetic variants and revealed one genome-wide significant association on chromosome 3 (rs9810253, OR = 1.11 ± 0.02, $p = 1.24 \times 10^{-8}$) (Fig. 2, Supplementary Data 14). A further 64 independent loci reached a suggestive threshold of 1×10^{-5}. Of these 65 loci, seven have been previously associated with sleep-related phenotypes, including one in an independent sample (rs62158206 on chromosome 2)[8] (Supplementary Data 16).

Gene-based tests
A gene-based test mapped input SNPs for the short vs. normal meta-analysis in EUR subjects to 18,565 protein-coding genes. Of these, 54 reached a Bonferroni-adjusted significance threshold of 2.69×10^{-6} (Fig. 3). The top gene identified was *FOXP2* ($p = 2.29 \times 10^{-13}$). A gene-based test mapped input SNPs for the long vs. normal meta-analysis in EUR subjects to 18,460 protein-coding genes; none reached the Bonferroni-adjusted significance threshold of 2.65×10^{-6} (Supplementary Fig. 18).

In the AFR gene-based test, input SNPs for the short vs. normal meta-analysis mapped to 18,292 protein-coding genes, none of which reached a Bonferroni-adjusted significance threshold of 2.73×10^{-6} (Supplementary Fig. 22). Input SNPs for the long vs. normal meta-analysis mapped to 19,074 protein-coding genes; none reached a Bonferroni-adjusted significance threshold of 2.62×10^{-6}.

For the cross-population meta-analysis of short sleep duration, in a gene-based test, input SNPs for the short vs. normal analysis mapped
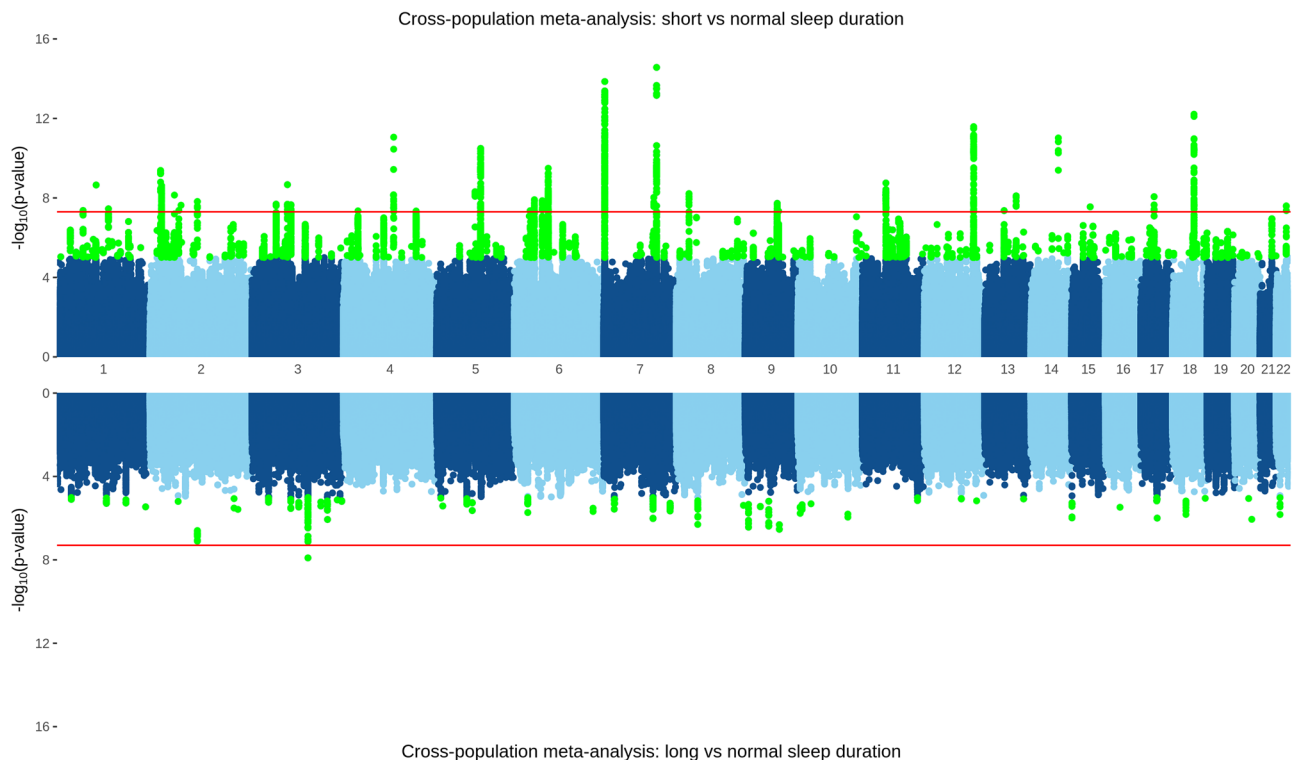
Fig. 2 | A cross-population meta-analysis of UK Biobank and MVP, with loci reaching a nominal threshold of 1×10^{-5} highlighted in green, and the red line indicating genome-wide significance threshold of 5×10^{-8}. TOP: short (≤5 h, $n = 62,516$) versus normal (7–8 h, $n = 412,944$) sleep duration, with 28 independent genetic-risk loci reaching genome-wide significance. BOTTOM: long (≥10 h, $n = 17,682$) versus normal sleep duration with one genome-wide significant locus.

Fig. 3 | Gene-based test for short (top) and long (bottom) sleep duration in cross-population meta-analysis of UK Biobank and MVP cohorts. Bonferroni-adjusted significance threshold $p < 2.7 \times 10^{-6}$ shown as the red line.

Short sleep duration

Fig. 4 | Cross-population replication analyses. Scatter plot for the *z*-score effect sizes (error band representing 95% confidence intervals) for 46 genome-wide significant loci from the EUR meta-analysis on short sleep on the *y*-axis against the *z*-score effect sizes for the same loci i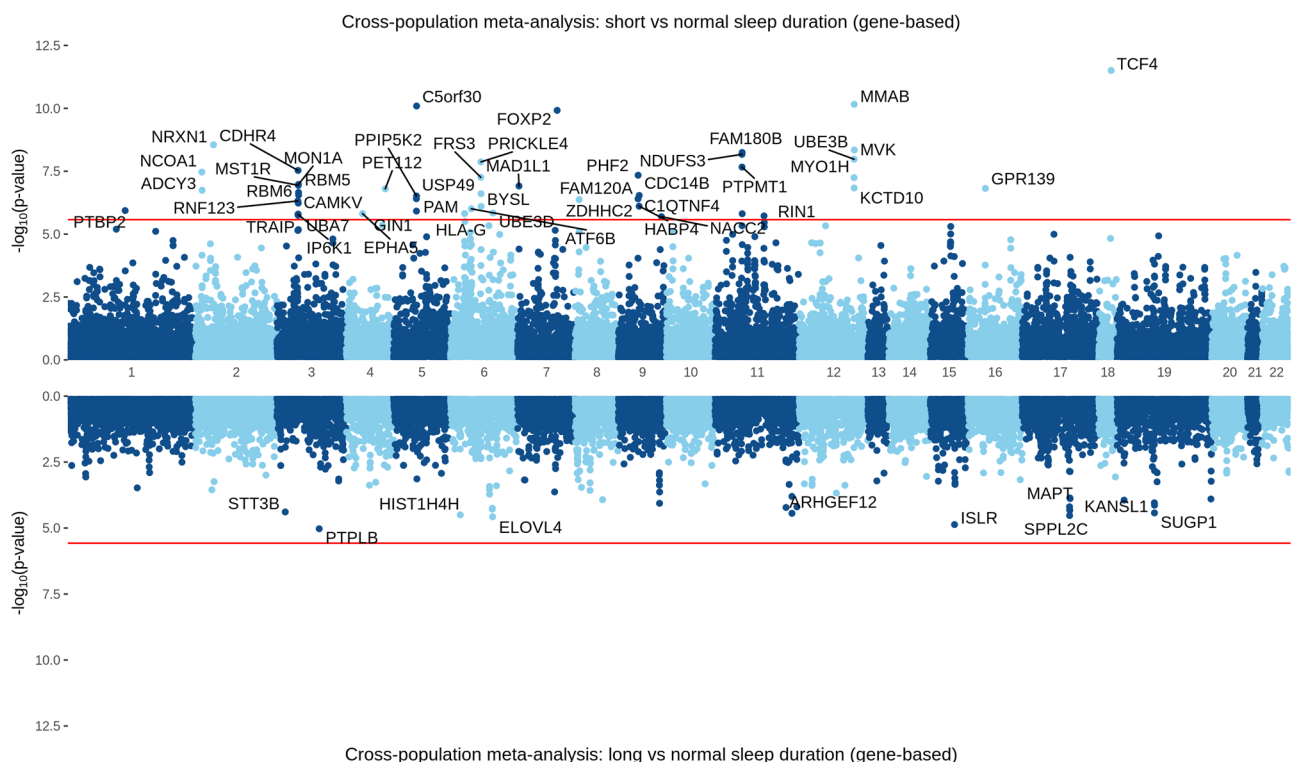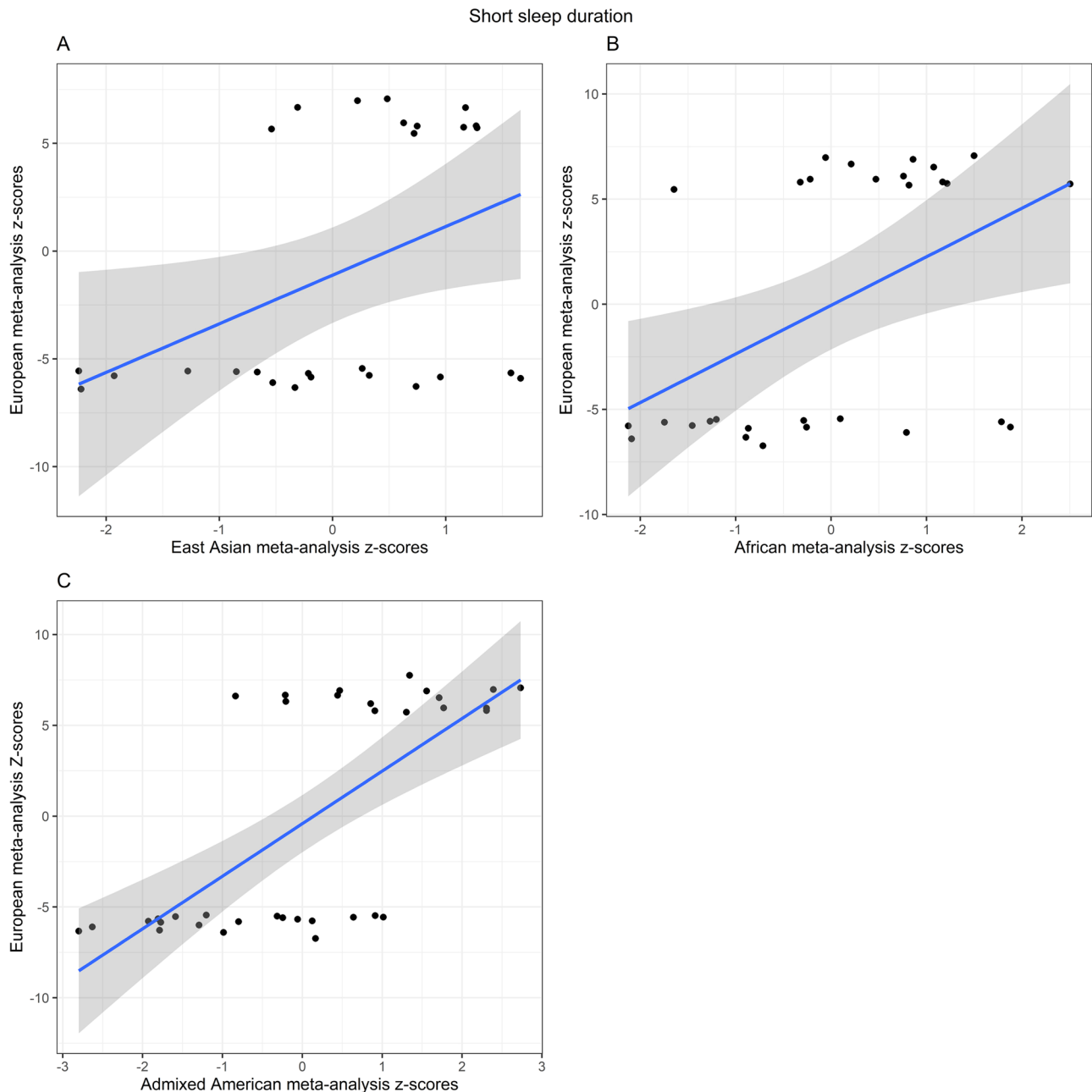n **A** EAS-only meta-analysis ($\rho = 0.290$), **B** AFR-only meta-analysis ($\rho = 0.293$), and **C** AMR-only meta-analysis ($\rho = 0.398$).

to 18,914 protein-coding genes. Of these, 47 reached a Bonferroni-adjusted significance threshold of 2.66×10^{-6}. The top gene identified was *TCF4* ($p = 53.11 \times 10^{-12}$) (Fig. 3, Supplementary Data 18). For long sleep duration, input SNPs for the long vs. normal analysis mapped to 18,903 protein-coding genes; none reached a Bonferroni-adjusted significance threshold of 2.67×10^{-6} (Fig. 3, Supplementary Data 19).

Cross-population transferability of loci

Of the nominally significant ($p < 1 \times 10^{-5}$) loci in the short and long sleep duration analyses in the AFR GWAS, none reached even a nominal significance threshold of $p < 0.05$ in the EUR analysis, and none demonstrate a lower p-value in the cross-population analysis. Of the 46 independent loci for a short sleep in the EUR-only analysis, 41 remain significant in the cross-population analysis and 30 are significant with a smaller *p*-value. The five loci that are no longer

genome-wide significant have a maximum *p*-value of 4.98×10^{-6} (Supplementary Data 17).

We performed cross-population lookups for the identified ancestry-specific GWS loci across the summary statistics for EUR, AFR, and EAS populations (Fig. 4, Supplementary Data 20). Of the 46 SNPs that reached genome-wide significance in the EUR meta-analysis of short sleep duration, 29 were present in the AFR summary statistics and 19 of them had a direction of effect consistent with the EUR results. Three of these loci reached nominal significance of $p < 0.05$ (rs12705972 on chromosome 7, $p = 0.012$; rs2111216 on chromosome 12, $p = 0.03$; rs7313797 on chromosome 12, $p = 0.04$).

In the EAS summary statistics, 27 of the EUR GWS loci were present and 19 had a consistent direction of effect. Two of these loci reached a nominal significance of $p < 0.05$ (rs146618518 on chromosome 5, $p = 0.025$; rs7313797 on chromosome 12, $p = 0.026$).
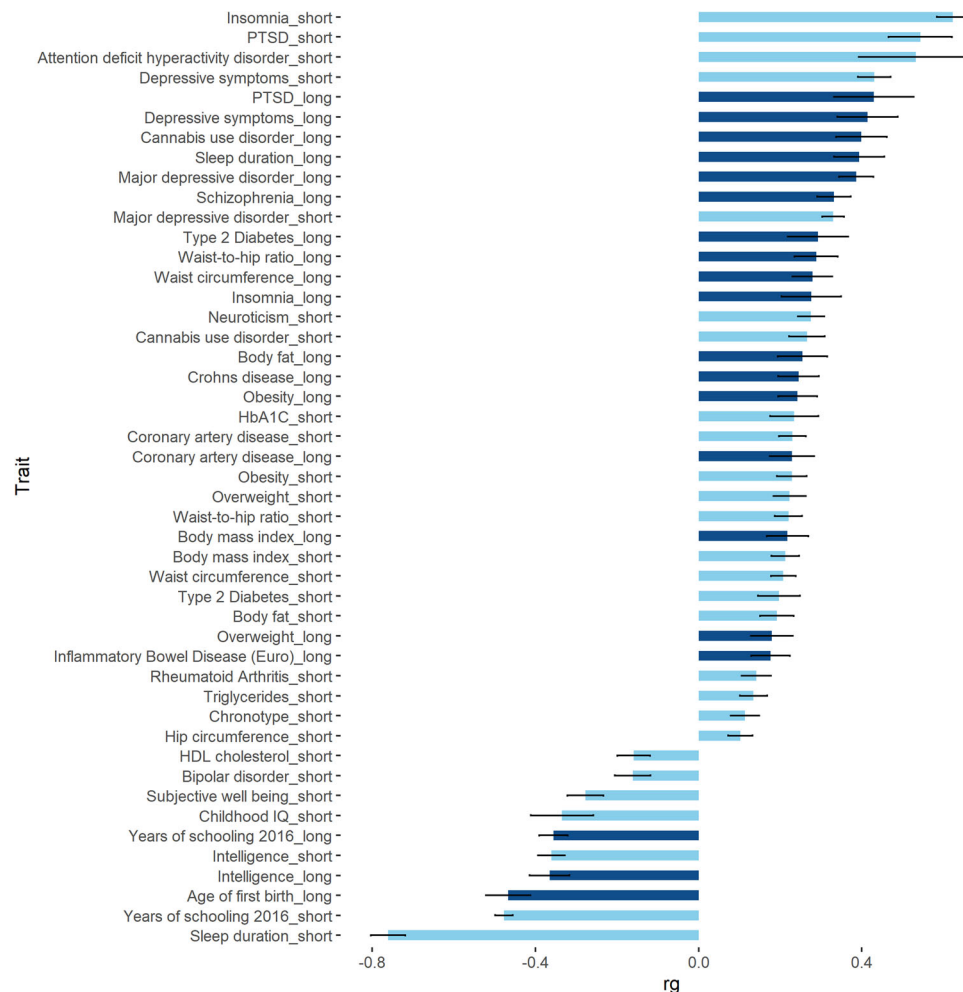
Fig. 5 | Summary of significant genetic correlations. Genetic correlations performed between short (light blue) and long (dark blue) sleep duration and previously published GWAS results (EUR only). Data presented as r_g (genetic correlation) estimates and standard error. Significance determined following Bonferroni correction for 40 tests; $p < 1.25 \times 10^{-3}$.

Twenty-two of the 46 EUR GWS loci were present in both AFR and EAS, and 10 of these had a consistent direction of effect across populations. Of these 10, only one SNP reached nominal significance in all three studies (rs7313797 on chromosome 12).

Rs62158206, the only SNP reaching genome-wide significance for long sleep duration in the EUR meta-analysis, was present in both the AFR (same direction of effect) and EAS summary statistics (opposite direction of effect), but $p > 0.05$ in both cases.

Given the known differences in LD structure and allele frequency across the population groups, we also considered all SNPs in high LD ($r^2 > 0.6$) with the GWS loci from the EUR meta-analysis of short sleep. One SNP was successfully matched to three SNPs in high LD that reached nominal significance in the AFR ancestry results (query SNP rs62144584, matched to a total of 14 SNPs in the AFR population with $p < 0.05$). The majority (11 out of 14) of these matched SNPs show a consistent direction of effect with the EUR results. Another SNP, rs201640077, was matched to three SNPs in the EAS population that reached nominal significance, with a consistent direction of effect (Supplementary Data 21).

Genetic correlation analysis

The within-phenotype genetic correlation between EUR participants in UK Biobank and MVP cohorts is 0.84 (± 0.05, $p = 3.74 \times 10^{-87}$) for short sleep and 1.106 (± 0.17, $p = 1.31 \times 10^{-10}$) for long sleep. For AFR, the correlations were non-significant for short sleep and long sleep could not be calculated for long sleep due to low sample sizes

(Supplementary Data 29). There was no significant cross-ancestry genetic correlation between the EUR and AFR participants for either short or long sleep (Supplementary Data 29).

We estimated a r_g between short and long sleep of 0.16 (± 0.04), $p = 2.0 \times 10^{-4}$ in the EUR-only meta-analyses and -0.09 (± 0.1), $p = 0.72$ in the AFR-only meta-analyses. The weak genetic correlation in the EUR sample suggests a distinct genetic architecture underlying the two traits. The lack of significant genetic correlation between these traits in the AFR sample may be due to lack of power, given the lower sample size in this analysis (Supplementary Data 30).

LDSC analysis was conducted in the EUR sample only (Fig. 5), due to a lack of available comparator data for AFR participants. The traits most strongly associated with both short and long sleep duration were years of schooling (short: $r_g = -0.48$ (0.02), $p = 2.41 \times 10^{-105}$; long: $r_g = -0.36$ (0.03), $p = 2.18 \times 10^{-24}$) and sleep duration (as a continuous trait), with directions of effect consistent with the definitions of short and long sleep duration (short: $r_g = -0.76$ (0.04), $p = 6.05 \times 10^{-72}$; long: $r_g = 0.39$ (0.06), $p = 1.99 \times 10^{-11}$). In addition, short sleep duration was positively correlated with attention deficit hyperactivity disorder ($r_g = 0.53$ (0.14), $p = 1.0 \times 10^{-4}$) and negatively correlated to bipolar disorder ($r_g = -0.16$ (0.04), $p = 2.0 \times 10^{-4}$). Long sleep duration was also positively correlated to schizophrenia ($r_g = 0.30$ (0.05), $p = 1.58 \times 10^{-10}$).

We observed several significant correlations between sleep duration and cardiometabolic traits. For example, both short and long sleep were positively correlated with coronary artery disease (short:
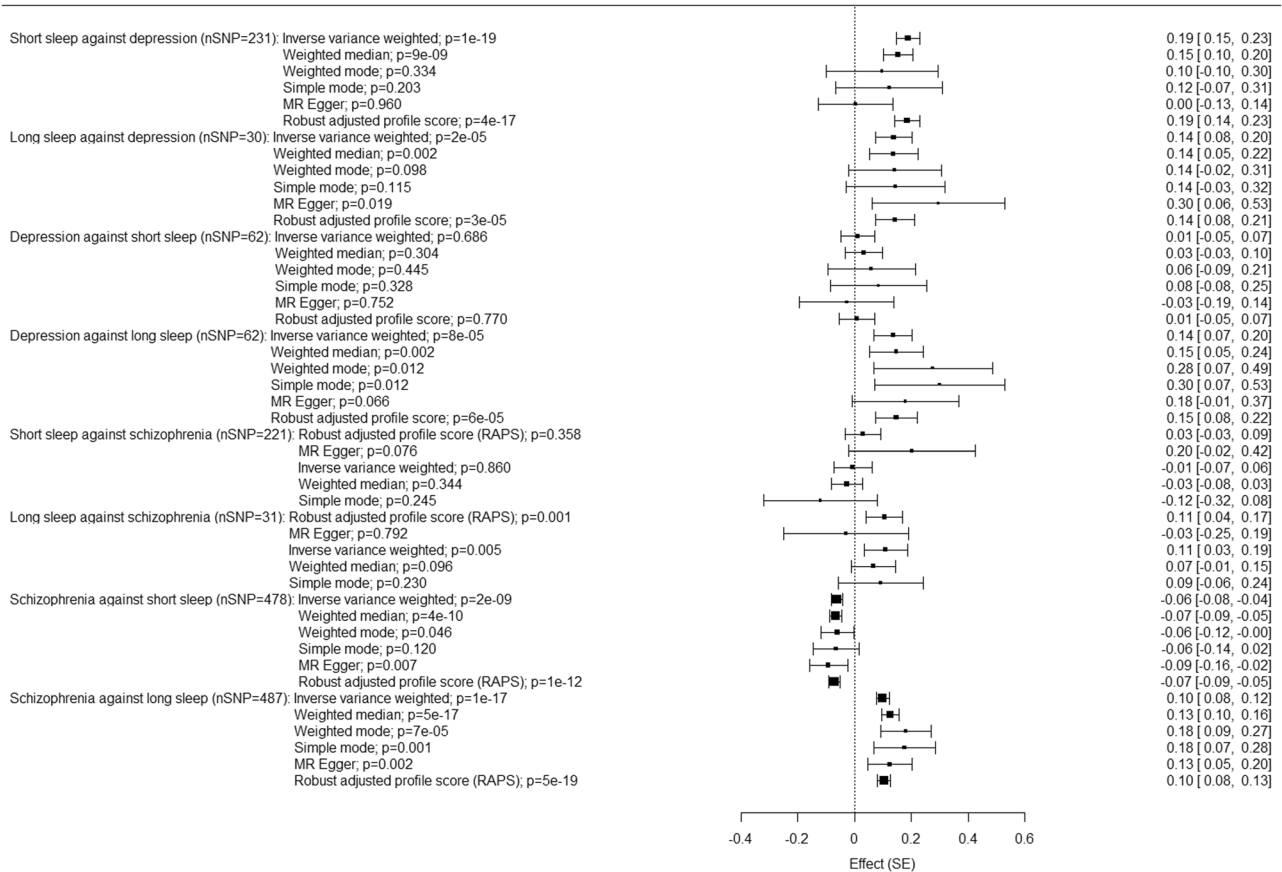
Fig. 6 | Effect estimates for the bi-directional Mendelian randomisation analyses between short and long sleep and MDD. The forest plots show the beta coefficient and 95% confidence intervals for the association between exposures and outcomes using different MR methods.

$r_g = 0.23$ (0.03), $p = 1.8 \times 10^{-12}$; long: $r_g = 0.23$ (0.05), $p = 2.6 \times 10^{-5}$), obesity (short: $r_g = 0.23$ (0.04), $p = 3.2 \times 10^{-10}$; long: $r_g = 0.24$ (0.05), $p = 4.2 \times 10^{-7}$), and type 2 diabetes (short: $r_g = 0.20$ (0.05), $p = 1.0 \times 10^{-4}$; long: $r_g = 0.29$ (0.07), $p = 7.8 \times 10^{-5}$). Additional significant (Bonferroni correction for 40 independent traits; $p < 1.25 \times 10^{-3}$) correlations are summarised in Fig. 5 and in Supplementary Data 31.

Mendelian randomisation

Exposure variables selected for MR analyses are listed in Supplementary Data 32. Results that reached a significance threshold of $p < 0.0125$ (0.05 divided by four independent tests) were considered significant.

MR analysis investigating the causal influence of short sleep on MDD supported a positive causal association between short sleep and increased risk of MDD, using the inverse variance weighted method ($\beta = 0.19$ (0.02) $p = 1.5 \times 10^{-19}$, Fig. 6, Supplementary data 33). Conversely, MR on the impact of MDD on short sleep did not support a causal association ($\beta = 0.01$ (0.03), $p = 0.69$). MR investigating the causal influence of long sleep on MDD reveals a positive causal association between long sleep and increased risk of MDD, using the inverse variance weighted method ($\beta = 0.14$ (0.03), $p = 1.64 \times 10^{-5}$). This was a bidirectional effect, with an analysis of the impact of MDD on long sleep also revealing evidence of a causal association ($\beta = 0.14$ (0.04), $p = 7.6 \times 10^{-5}$).

MR analysis investigating the causal influence of short sleep on schizophrenia showed no evidence of a causal association (Fig. 6, Supplementary Data 33). Conversely, MR on the impact of schizophrenia on short sleep did support a negative causal association ($\beta = -0.061$ (0.01), $p = 1.7 \times 10^{-9}$). MR investigating the causal influence of long sleep on schizophrenia reveals a positive causal association between long sleep and increased risk of schizophrenia, using the

inverse variance weighted method ($\beta = 0.14$ (0.03), $p = 7.58 \times 10^{-5}$). This was a bidirectional effect, with an analysis of the impact of schizophrenia on long sleep also revealing evidence of a causal association ($\beta = 0.10$ (0.01), $p = 1.10 \times 10^{-7}$).

The Egger-intercept was non-significant in three of four analyses (long sleep against MDD: Egger intercept −0.01 (0.009), $p = 0.18$; MDD against short sleep: Egger intercept 0.003 (0.005), $p = 0.62$; MDD against long sleep: Egger intercept −0.003 (0.006), $p = 0.63$). The Egger intercept for short sleep against MDD was significant, (0.009 (0.003), $p = 0.004$), suggesting the inverse variance weighted estimate may be biased. We therefore repeated this analysis using a more stringent p-value threshold for the exposure SNPs of 5×10^{-8}. In this case, the Egger intercept was non-significant (0.005 (0.012), $p = 0.65$); the inverse variance weighted estimate remained significant (0.23 (0.046), $p = 4.0 \times 10^{-7}$). The Egger intercept was non-significant in all four analyses of sleep and schizophrenia (short sleep against schizophrenia: Egger intercept −0.01 (0.005), $p = 0.055$; long sleep against schizophrenia: Egger intercept −0.01 (0.009), $p = 0.194$; schizophrenia against short sleep: Egger intercept 0.002 (0.002), $p = 0.35$; schizophrenia against long sleep: Egger intercept −0.002 (0.003), $p = 0.50$) (Supplementary data 33).

Weighted median, MR-Egger, and MR-RAPS analyses were conducted as sensitivity analyses. The weighted median regression and MR-RAPS analyses were significant for MDD against long sleep, and both short and long sleep against depression, following the same pattern as observed using the inverse-variance weighted method. These methods also yielded significant results for schizophrenia against short and long sleep, but this was a uni-directional finding. In the analyses with MDD, no MR-Egger tests survived multiple testing corrections. In the analyses with schizophrenia, MR-Egger analyses

were significant for schizophrenia against short sleep and against long sleep.

Discussion

The question of how to improve sleep quality and optimise sleep duration is of constant interest, with the global market for sleep aids and technologies exceeding 80 billion US dollars per year[24]. Along with genetic influences, a wide variety of demographic, social, and environmental factors can impact the quality and duration of sleep, including socioeconomic status, stressful life events, home and neighbourhood characteristics, work and school schedules, medication and substance use, and mental and physical health conditions. Indeed, sleep quality and duration can be considered both a risk factor for and symptom of many health conditions.

We present a large GWAS of self-reported sleep duration conducted for the first time in a diverse population. In addition to replicating associations with many genes previously linked to sleep traits, both in studies with overlapping samples and fully independent cohorts, our analyses expand on the previous understanding of the genetic architecture of sleep through the identification of numerous novel risk loci for short sleep duration and one novel locus for long sleep duration. We identify genes of interest in EUR, AFR, and cross-population analyses. Our findings add to the existing knowledge of the genetic basis of sleep duration, as well as highlighting at least one ancestry-specific risk locus and shared genetic risk with a variety of cognitive, neuropsychiatric, and metabolic traits.

We conducted a cross-population meta-analysis including all EUR, AFR, EAS, and AMR cohorts from UK Biobank and MVP. We identified 84 independent GWS risk loci for short sleep duration. The strongest associations were on chromosomes 7 (rs1989903, near *FOXP2*, and 7:2054314:C:CG, near *MAD1L1*) and 18 (rs11152363, near *TCF4*). *FOXP2* is a transcription factor that has been implicated in GWAS of insomnia, BMI, cannabis use disorder, and risk-taking, as well as short sleep duration[7,10,16,25]. *MAD1L1* is a member of a family of genes that encode proteins important in the mitotic checkpoint. This gene has previously been associated with sleep and several psychiatric traits in previous GWAS, including bipolar disorder, anxiety, PTSD, major depressive disorder, and schizophrenia[7,26–28]. *TCF4* is a transcription factor that has previously been associated with cognitive traits, educational attainment, alcohol consumption, autism spectrum disorder, schizophrenia, depression, lung function and BMI, as well as sleep duration[16,29–34].

The gene-based test identified 47 significant genes, with the strongest association with *TCF4*. Several of these significant loci and genes are the same as highlighted in EUR-only meta-analysis, which is unsurprising given the relative sample sizes of each population group. However, the increased number of genome-wide significant associations shows how the addition of these samples increased power and added valuable information.

In the cross-population analysis of long sleep duration, we identified one GWS locus at rs9810253 on chromosome 3 ($p = 1.24 \times 10^{-8}$), near *PTPLB*. This locus has a consistent direction of effect across all primary studies, with $p = 7.36 \times 10^{-7}$ in the EUR-only analysis and 0.005 in the AMR analysis.

We conducted GWAS separately for individuals of EUR and AFR ancestry before conducting a cross-population meta-analysis that also incorporated EAS and AMR samples. In a GWAS of short vs. normal sleep duration in the UK Biobank and MVP EUR samples, we identified 46 independent GWS risk loci. The strongest associations can be found on chromosome 7 (rs6466488, near *FOXP2* and rs111595851, near *MAD1L1*) and chromosome 4 (rs13107325, near *SLC39A8*). In EUR, we identified one locus with compelling evidence that it results in long sleep duration. The strongest association was at rs62158206 on chromosome 2. This SNP is intergenic to *PAX8*, which encodes a transcription factor that has previously been associated

with several sleep-related phenotypes, including insomnia (with a consistent and opposite direction of effect)[16,17] and sleep disturbance in depression[35]. In addition, this gene has been highlighted in several previous GWAS on sleep duration, both in fully independent and overlapping samples[7,8].

SNP-based heritability of short sleep duration was 11.9% ($p = 2.45 \times 10^{-115}$). Long sleep duration appeared less heritable, with an SNP-based heritability of 7.8% ($p = 1.61 \times 10^{-20}$). Both estimates are broadly consistent with previously published results[7,8]. The significant difference in the number of significant loci for long sleep compared to short sleep may be in part a result of this lower heritability, perhaps suggesting a greater influence of environmental factors on longer sleep duration; but lower power based on sample size may have been decisive for both measures (risk loci and observed heritability). There were also fewer long sleepers (≥10 h) in both the UK Biobank and MVP cohorts compared to short sleepers (≤5 h) and the effect estimates for the top associations were of comparable magnitude, indicating decreased power for the analysis.

Amongst AFR participants, we identified one risk locus, at rs148926968 on chromosome 13 associated with long sleep. This locus is intergenic for *LMO7*, which has been previously identified in GWAS of obsessive-compulsive disorder[36], several age-related diseases including hearing loss[37] and Alzheimer's disease[38], and several cross-population GWAS on eyesight-related traits[39,40]. We did not observe any genome-wide significant loci for short sleep duration in the AFR sample. Additional analyses in larger cohorts are required to understand the extent to which this might reflect differing genetic architecture in AFR compared to EUR, or if this is more a result of reduced power given the smaller sample size. In AFR, we found significant SNP-based heritability for short sleep duration (8.2%, $p = 0.04$), but not for long sleep duration. There are no prior published h^2 estimates for sleep duration in individuals of AFR. Given the lower observed heritability of long sleep in EUR, we suspect that the sample size included here is too small to confirm heritability in AFR.

Due to the limited GWS findings in the non-EUR analyses, we performed cross-population lookups in the EUR meta-analysis of UK Biobank and MVP cohorts. Of the 46 GWS loci in this EUR study, 22 were also present in the AFR and EAS meta-analysis, and 10 had a consistent direction of effect across all three populations. Of these ten, only one reached at least nominal significance in all populations: rs7313797 on chromosome 12. This SNP is intronic to *KCTD10*, which has previously been identified in GWAS of neurotic disorder, well-being, coronary artery disease, and HDL cholesterol levels[41–44].

Our results do not support a significant genetic correlation between the EUR and AFR samples for either short or long sleep, though these analyses are hindered by small sample sizes in the AFR population. We do find some evidence for portability across populations, as highlighted in Fig. 4, but future analysis in additional non-EUR datasets and in a larger AFR sample should help clarify the extent of genetic overlap and aid the identification of truly causal variants.

The genetic correlation between short and long sleep was low, with $r_g = 0.16$ ($p = 0.0002$) in EUR and a non-significant correlation in AFR, due to the smaller sample size. Larger samples are necessary to further evaluate the relationship of these sleep traits in non-European populations. Nevertheless, the result of the EUR analysis confirms our hypothesis that these traits, though clearly phenotypically related, possess distinct underlying genetic architecture and biology. There are notable similarities for some traits significantly correlated with both short and long sleep, such as depression, cannabis use disorder, PTSD, coronary artery disease, obesity, and diabetes. This indicates that an increased genetic risk for a range of traits is associated with an increased risk of sleep disturbance at either end of the spectrum, perhaps depending on the specific variants at play. In addition, short and long sleep were both significantly positively correlated with insomnia. Localised genetic correlation analyses may

be valuable in establishing if these shared patterns of r_g are in fact driven by the same loci[45].

There is significant evidence, genetic and otherwise, of high comorbidity between disturbed sleeping and a variety of neuropsychiatric, cognitive, and metabolic traits. Short and long sleep duration has been associated with all-cause mortality and decreased life expectancy. The results of the genetic correlation analyses in LDSC provide support for these observations and confirm that some of this comorbidity can be explained by shared genetic risk.

The MR analyses found a directional causal influence of short sleep duration resulting in an increased risk of MDD. We also identified evidence of a bi-directional causal relationship between long sleep duration and MDD. Many previous studies have identified phenotypic associations between sleep duration and MDD or depressive symptoms[46–50], but the causal direction of these associations has not always been clear. The findings from the present analyses, that both short and long sleep duration can cause an increased risk of MDD, support existing theories on the importance of healthy sleep patterns for mood regulation and highlight shared genetic risk factors between the two extremes. We demonstrate a negative causal association between schizophrenia and short sleep and a positive causal association between schizophrenia and long sleep duration. There is evidence of a bidirectional causal association between schizophrenia and long sleep, but not short sleep.

We considered hours of daylight exposure as a potential environmental factor impacting variance in sleep duration. In the UK Biobank sample, where we were able to consider monthly solar irradiation data, higher levels of estimated solar irradiation were significantly associated with shorter reported sleep duration. In the MVP sample, where we assessed annual irradiation only, increased solar irradiation was not significantly associated with sleep duration. Independently of solar irradiation levels, we found that the month of recruitment was significantly associated with sleep duration.

It is possible that the use of annual data only in the MVP sample may be limiting our ability to detect an effect; the large variation in daylight exposure across the year cannot be captured by annual data. However, there are other significant differences between the cohorts that could contribute to these findings, and the question on sleep duration in both UK Biobank and MVP was phrased such as to ask for habitual sleep patterns. Our data suggests solar irradiation is a significant influence, however, longitudinal data recording sleep patterns over the course of several years will be valuable in determining the full extent of the impact of hours of daylight on sleep duration.

Reliance on self-reported sleep data is a limitation. However, a previous study in the UK Biobank demonstrated a high level of consistency between the top variants identified in a GWAS of self-report sleep duration data and those seen in a study of 85,499 subjects using wrist-worn accelerometer data[7]. Both cohorts included predominantly older adults, where sleep disturbance (particularly short sleep) is more common and while this increases the power of the analyses intended, it limits our ability to generalise findings to younger groups. A further limitation is the underrepresentation of non-European populations. Although we were able to include four population groups in the cross-population meta-analysis, many post-GWAS analyses were conducted in EUR and AFR ancestry groups only, and some of the methods used are less reliable in population groups with higher levels of admixture.

In summary, we have identified multiple novel variants of GWS for short sleep duration, and one novel locus for long sleep duration, among UK Biobank and MVP cohorts. Several of these loci and genes warrant future investigation in populations of greater age, sex, and ancestral diversity. Genetic correlations provided support for shared genetic risk between sleep duration and a range of comorbid traits and MR analysis supports a causal association between sleep duration and depression. These findings highlight the value of understanding the genetic basis of sleep patterns in order to improve public health.

Methods

Inclusion and ethics statement
This research was not restricted or prohibited in the setting of any of the included researchers. UK Biobank was approved by a UK ethics review committee. MVP was approved by the Veterans Affairs central IRB. We do not believe our results will result in stigmatisation, incrimination, discrimination, or personal risk to participants.

Participants
The UK Biobank and MVP cohorts are described in detail in refs. [19–21]. The UK Biobank study was approved by the North-West Research Ethics Committee (ref 06/MREC08/65) in accordance with the Declaration of Helsinki. Research involving the MVP in general is approved by the VA Central Institutional Review Board. All participants in both cohorts provided written informed consent.

Genotyping, imputation, and quality control
UK Biobank. Genotyping and imputation of UK Biobank subjects are described in detail in ref. [19]. Briefly, genotyping for UK Biobank participants was undertaken using the Affymetrix UK BiLEVE Axiom array (used for the first ~50,000 participants) and the Affymetrix UK Biobank Axiom Array (~450,000 participants)[19]. These arrays are >95% similar and include ~820,000 SNP and indel markers (http://www.ukbiobank.ac.uk/). Quality control and imputation of over 90 million SNPs, indels and large structural variants were performed centrally[19]. Samples identified as outliers for heterozygosity and/or missingness were removed, leaving a total sample of 487,411. The fully imputed genetic data used in this study, with basic sample and variant level quality control as reported in ref. [19], were made available in March 2018.

Additional local post-imputation SNP-level quality control was conducted to remove SNPs with an imputation INFO score <0.3 or those with minor allele frequency (MAF) < 0.01. This filtering was done separately in each ancestry group to ensure that population-specific variants were not removed. Further individual-level quality control was conducted locally to remove samples with mismatch between reported sex and genetically inferred sex (due to risk of sample processing errors) and those with excessive genetic relatedness (>10 third-degree relatives based on kinship calculations provided centrally by UK Biobank). Individuals missing either sleep or essential quality data were excluded. The final list was then checked to remove those who had withdrawn consent.

Genetic ancestry of the UK Biobank sample was assessed using principal component analysis (PCA) in combination with self-reported ethnicity data. A list of 409,728 EUR individuals was identified centrally by the UK Biobank[19]. Further local analysis was conducted to delineate the ancestry of another 77,683 participants from diverse populations, applying the same thresholds as described in ref. [19]. Two rounds of PCA were performed using the PC-AiR algorithm[51], which captures population structure. Relatedness in this sample was assessed using PC-Relate and the ancestry representative PCs[52]. Of the samples that passed the QC procedures described here, over 99% provided self-reported sleep duration data.

MVP. Genotyping and imputation of MVP participants have been described previously[21]. Briefly, MVP subjects were genotyped using a customised Affymetrix Axiom Array, similar to the UK Biobank array. MVP genotype data were imputed using Minimac4 and a reference panel from the African Genome Resources (AGR) panel by the Sanger Institute. Indels and complex variants were imputed independently from the 1000 Genomes phase 3 panel (G1K) and merged in a similar approach to UKB HRC + UK10K. SNPs with an imputation info score < 0.3, estimated genotype hard call missingness rate of >0.2, or an MAF < 0.001 were removed. PCA was conducted using Eigensoft[53]. Genetic ancestry of the MVP participants was assigned separately within each data tranche, based upon the first 10 principal

components, with 1000 Genomes Project (phase 3) EUR, African (AFR), Admixed American (AMR) and East Asian (EAS) data as reference samples[21].

Phenotypic assessment and covariate measures
In both cohorts, we selected participants who had provided self-reported data on sleep duration. In UK Biobank, participants were asked "About how many hours sleep do you get in every 24 h? (please include naps)" as part of the baseline assessment. Responses were given in hour increments and participants who claimed to sleep less than three hours or more than 12 were prompted to confirm their answer. In the MVP, data on sleep duration was collected from the MVP lifestyle questionnaire, where participants were asked "How many hours do you usually sleep each day (24-h period)?". The response options were multiple choice: 5 or less, 6, 7, 8, 9 or 10 or more.

We did not wish to assume that short and long sleep are necessarily on the same biological continuum, and therefore defined separate phenotypes rather than reconciling the differing UK Biobank and MVP ordinal traits. We defined 'short' sleep duration as ≤5 h sleep, 'normal' as 7–8 h and 'long' as ≥10 h[54].

In both UK Biobank and MVP samples we included sex, age at recruitment, and the first 10 principal components as covariates in the GWAS. In the UK Biobank sample, the genotyping array used was included as an additional covariate. We found that individuals who reported being diagnosed with obstructive sleep apnoea were significantly over-represented in both the short and long-sleep groups, and we, therefore, excluded these participants from our analysis in both the UK Biobank and MVP cohorts.

Although the 'normal' or medium sleepers (7–8 h) represent the largest group in both samples and the distribution of responses is similar, the MVP sample has a significantly greater proportion of both short and long sleepers (Fig. 1). UK Biobank is a population-based study, and differences between the sample and the UK population have been described in ref. 55; this could also relate to how the phenotype was elicited. MVP recruits through the US Department of Veterans Affairs (VA) Healthcare System, meaning these participants can be considered a patient population with a broad range of potential health conditions, some of which can be expected to impact sleep duration.

Hours of daylight exposure
We evaluated the possible effects of hours of daylight on sleep hours, as this is known to have a significant impact on reported sleep duration[56–59]. We calculated the estimated solar irradiance for each participant based on the location of their recruitment site. We downloaded monthly direct normal solar irradiation data from the European Commission Photovoltaic Geographical Information System[60] and the National Solar Radiation Database[61]. For UK Biobank subjects, solar irradiation indices were based on the recruitment site and the month of their recruitment. For MVP subjects, only average annual data were available. We conducted a linear regression analysis to examine the effects of solar irradiance on sleep hours, including age and sex as covariates. This analysis was conducted using R version 3.5.0 (2018-04-23)[62].

Statistical analyses
We conducted GWAS on two separate phenotypes, short sleep duration vs. normal and long sleep duration vs. normal, in each of the independent primary samples (UKB-EUR, UKB-AFR, UKB-EAS, UKB-AMR, MVP-EUR, MVP-AFR, MVP-AMR. GWAS analysis was conducted by logistic regression using PLINK 2.0 on genotype dosage data, including age, sex, and first 10 principal components as covariates, and in the case of UK Biobank the genotype array as covariates[63]. Where kinship scores showed a relatedness between a pair as closer than the 2nd degree, one of each of the pairs was excluded. Where both pair members were cases or both were controls, the excluded participant was chosen at random. Where one pair member was a case and one was

a control, the control participant was excluded to maximise the case sample size.

We used METAL[64] to conduct independent fixed effect meta-analyses in ancestry-specific samples and to conduct a cross-population meta-analysis using all primary GWAS from UK Biobank and MVP. The resulting summary statistics were filtered to remove any SNP that did not appear across all four population groups, in either UK Biobank or MVP data. For each primary GWAS, meta-analysis and cross-population meta-analysis, we calculated LD intercept to assess genomic inflation due to sample size and polygenicity of trait[65]. Manhattan and quantile–quantile (Q–Q) plots were created using the R packages ggplot2 and Hudson[62,66]. Independent GWAS signals were identified through clumping of results with an r^2 of 0.6. A second clumping of the independent SNPs was performed with r^2 of 0.1 to identify lead SNPs[67].

In addition to these primary studies, we repeated our analyses of long and short sleep duration in a sex-stratified UKB cohort, and in a sub-sample of the UKB cohort excluding nightshift workers. Finally, we conducted a case-case comparison comparing long vs. short sleep in both UKB and MVP. Further detail on these analyses is provided in the Supplementary material sections 2, 3, 4 and Supplementary Data 34 and 35. We used the EUR MVP GWAS to provide an independent replication sample for previously published GWAS assessing self-reported sleep duration as both a binary and quantitative trait in 446,118 European UK Biobank subjects[7]. See Supplementary Material Section 7 and Supplementary Data 36–41.

Functional annotation and gene-based tests
We uploaded summary statistics from the primary GWAS and the meta-analyses into the functional mapping and annotation (FUMA) GWAS platform version 1.3.7 to annotate GWAS data[67]. Default settings were used, including using all 1000 genome project (1KG) reference populations for the cross-population meta-analyses (1KG EUR and AFR used for the EUR and AFR GWAS, respectively). SNPs were mapped according to chromosomal position based on ANNOVAR annotations, with a maximum distance of 10 kb between SNPs.

We examined gene-level associations using Multi-Marker Analysis of GenoMic Annotation (MAGMA) version 1.6, using default settings[67]. SNPs were assigned to genes based on Ensembl build 85, and the association with each sleep phenotype was calculated as a combined gene test statistic based on the individual p-values of the SNPs mapped to a given gene. The significance threshold was calculated using a Bonferroni multiple testing correction to account for the specific number of protein-coding genes in each gene-based test.

Transcriptome-wide association study and fine-mapping
SNP-level fine-mapping was conducted using PolyFun (POLYgenic FUNctionally-informed fine-mapping)[68]. PolyFun calculates per-SNP heritability for each variant in provided summary statistics, which is proportional to prior causal probability. We used these per-SNP heritability estimates along with downloaded functional annotations from The Broad Institute (functional annotation for ~19 million UK Biobank SNPs with MAF > 0.1%, based on the baseline-LF model described in ref. 69) to conduct functionally-informed fine-mapping using Sum of Single Effects (SuSiE)[70]. We extracted annotations for all SNPs with posterior inclusion probability (PIP) ≥ 0.95. We then estimated functional enrichment using S-LDSC (stratified LD-score regression)[65,71] using pre-calculated weights, and the resulting annotations were ranked (see Supplementary Data 22, 23).

We conducted a transcriptome-wide association study (TWAS) using FUSION[72], which integrates GWAS summary statistics and gene-expression data to identify gene expression patterns associated with long or short sleep duration. We performed expression imputation for autosomes using GTEx v8 multi-tissue expression weights from 49 tissues. As loci could be associated with multiple features, we identified genes that were conditionally independent, using the

'FUSION.post_process.R' script provided, which reads expression weights for selected genes and consolidates them into overlapping loci (described further at http://gusevlab.org/projects/fusion/)[72]. Based on 49 tissues and 27,977 Ensembl Gene IDs (representing genes, non-coding transcripts, and pseudogenes) we applied a multiple testing correction for significance at p-value $\leq 3.65 \times 10^{-8}$ (0.05/(49*27,977). Where a gene was significant and expressed in two tissues, we selected the gene expression with the lowest p-value (see Supplementary data 24, 25).

We then used Fine-mapping Of CaUsal gene sets (FOCUS)[73] to fine-map genomic risk regions identified through the TWAS using pre-computed expression quantitative trait loci (eQTL) weights from a multi-tissue, multiple eQTL reference database which combines GTEx v7 weights from PrediXcan, Metabolic Syndrome in Men Study (METSIM), Netherlands Twin Registry (NTR), Young Finns Study (YFS), and CommonMind Consortium (CMC). LD scores were obtained from 1000 Genome Phase 3. We filtered results based on a PIP threshold of ≥ 0.7[74] (see Supplementary data 26, 27).

Ingenuity pathways analysis

We performed a Core Analysis using Ingenuity Pathway Analysis software[75]. Gene lists came from the MAGMA analysis. A 0.05 false discovery rate was applied to the MAGMA output from the cross-population meta-analysis for short and long sleep, yielding lists of 380 and 0 genes, respectively. Because no genes survived the initial correction a second cut-off of 0.10 was applied to the long sleep analysis, with 18 genes surviving this less restrictive cut-off (see Supplementary Data 28).

Cross-population transferability of loci

There is often only limited overlap in genome-wide significant (GWS) risk loci across population groups. Differences in LD structure and allele frequency among population groups make it difficult to determine if an observed association from a primarily European study is replicated in other populations, as the truly causal variant is often unknown[76]. Using the R package LDlinkR[77] we developed 'credible sets' of SNPs in EAS and AFR populations that are in high LD ($r^2 > 0.6$) with the GWS loci from the European analysis of short and long sleep duration. We then searched the UK Biobank/MVP short and long sleep meta-analyses results for EAS and AFR populations for evidence of association ($p < 0.05$).

Genetic correlation and SNP-based heritability

We used linkage disequilibrium score regression (LDSC) to estimate SNP-based heritability. For the EUR sample, we used reference LD scores provided by the 1000 Genomes Project[78]. The high admixture in the AFR sample means that the reference panel data may be unreliable. We therefore used LD scores calculated from the primary genotype data with principal components included as covariates. For the UK Biobank sample, we used the scores published by the Pan-UKBB group. For the MVP sample, we used the equivalent cov-LDSC method (described in ref. 79, https://github.com/immunogenomics/cov-ldsc).

Converting the observed SNP-heritability to the liability-scale estimates presented here required an estimation of the population prevalence of 'cases' (i.e., short or long sleepers). We based these estimates on the proportion of short and long sleepers in the meta-analysis of UK Biobank and MVP data. These estimates were weighted based on the sample size of the primary studies. Due to observed differences in the distribution of sleep duration, we calculated these estimates separately in the EUR and AFR populations. A higher proportion of AFR subjects report sleeping six hours or less, and a smaller proportion of AFR subjects report sleeping seven, eight, or nine hours, compared to EUR. In the EUR cohort, this resulted in an estimated population prevalence (K) of 0.11 for short and $K = 0.04$ for long sleep. In the AFR cohort, this resulted in a $K = 0.43$ for short and $K = 0.05$ for long sleep.

For the EUR cohort, we used LDSC to calculate the genetic correlation between sleep duration and a range of cognitive, neuropsychiatric, cardiac, and metabolic traits[65,80,81]. We used LDHub to assess the genetic correlation of short and long sleep to all available traits. In addition, we used LDSC to assess the genetic correlation to traits based upon GWAS summary statistics downloaded from the Psychiatric Genomics Consortium (PGC) website (https://www.med.unc.edu/pgc/results-and-downloads/) and from the Sleep Disorder Knowledge Portal (http://www.kp4cd.org/dataset_downloads/sleep).

Finally, we assessed the genetic correlation between short and long sleep duration based upon the summary statistics for the meta-analyses as described above, using LDSC for the within-ancestry analyses[65] and Popcorn (version 0.9.6: https://github.com/brielin/Popcorn) for the cross-population analyses[82].

Mendelian randomisation (MR)

We conducted two-sample MR of both short and long sleep duration using summary statistics for MDD and schizophrenia from the Psychiatric Genomics Consortium[29,30,83,84], using the TwoSampleMR package in R[30,85]. The genetic instruments for all traits were defined as the independent variants that reached a significance threshold of $p < 1 \times 10^{-5}$. Independent associations were identified by LD clumping with $r^2 = 0.6$ and a window of 250 kb. To avoid sample overlap, we used PGC data excluding UK Biobank samples.

We used the inverse-variance weighted (IVW) method as our primary MR model. We also conducted MR-Egger, weighted median, and weighted mode analyses to test for horizontal pleiotropy and potentially invalid genetic instruments[86,87]. MR-robust associated profile score (MR-RAPS) was conducted as a further sensitivity analysis to account for potential weak instrument bias or extreme outliers[88]. MR analyses were conducted in EUR samples only.

Calculation of the Egger intercept can identify directional pleiotropy which can bias the inverse variance estimates. Where there is directional pleiotropy, the MR-Egger analysis may provide a more reliable effect estimate. Where the Egger-intercept is non-significant, this demonstrates a lack of directional horizontal pleiotropy and provides confidence in the estimates using the inverse variance method. If we identified a significant Egger intercept, we repeated the analyses using only genome-wide significant SNPs ($p < 5 \times 10^{-8}$).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The summary statistics for the GWAS and meta-analyses generated in this study have been deposited in dbGAP under accession number phs001672.v1.p1 and are also available on the Gelernter Lab website (https://medicine.yale.edu/lab/gelernter/stats/). The raw genotype data is available through UK Biobank (http://biobank.ndph.ox.ac.uk/showcase/). Data from the European Commission Photovoltaic Geographical Information System can be accessed here: https://re.jrc.ec.europa.eu/pvg_tools/en/. Data from the National Solar Radiation Database can be accessed here: https://nsrdb.nrel.gov/data-sets/how-to-access-data.

References

1. Krueger, J. M., Huang, Y. H., Rector, D. M. & Buysse, D. J. Sleep: a synchrony of cell activity-driven small network states. *Eur. J. Neurosci.* **38**, 2199–2209 (2013).
2. Krueger, J. M., Frank, M. G., Wisor, J. P. & Roy, S. Sleep function: toward elucidating an enigma. *Sleep Med. Rev.* **28**, 46–54 (2016).
3. Armstrong, T. S. et al. Sleep–wake disturbance in patients with brain tumors. *Neuro-Oncology* **19**, 323–335 (2017).

4.   Markand, O. N. & Dyken M. L. Sleep abnormalities in patients with brain stem lesions. *Neurology* **26**, 769–776 (1976).
5.   Krueger, J. M. Sleep and circadian rhythms: evolutionary entanglement and local regulation. *Neurobiol. Sleep Circadian Rhythm.* **9**, 100052 (2020).
6.   Buysse, D. J. Sleep health: can we define It? Does it matter? *Sleep* **37**, 9–17 (2014).
7.   Dashti, H. S. et al. Genome-wide association study identifies genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates. *Nat. Commun.* **10**, 1100 (2019).
8.   Gottlieb, D. J. et al. Novel loci associated with usual sleep duration: the CHARGE Consortium Genome-Wide Association Study. *Mol. Psychiatry* **20**, 1232–1239 (2015).
9.   Lane, J. M. et al. Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. *Nat. Genet.* **49**, 274–281 (2017).
10. Lane, J. M. et al. Biological and clinical insights from genetics of insomnia symptoms. *Nat. Genet.* **51**, 387–393 (2019).
11. Heath, A. C., Kendler, K. S., Eaves, L. J. & Martin, N. G. Evidence for genetic influences on sleep disturbance and sleep pattern in twins. *Sleep* **13**, 318–335 (1990).
12. De Castro, J. M. The influence of heredity on self-reported sleep patterns in free-living humans. *Physiol. Behav.* **76**, 479–486 (2002).
13. Byrne, E. M. et al. A genome-wide association study of sleep habits and insomnia. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **169**, 439–451 (2013).
14. Jones, S. E. et al. Genome-wide association analyses in 128,266 individuals identifies new morningness and sleep duration loci. *PLoS Genet.* **12**, e1006125 (2016).
15. Allebrandt, K. V. et al. A KATP channel gene effect on sleep duration: from genome-wide association studies to function in *Drosophila*. *Mol. Psychiatry* **18**, 122–132 (2013).
16. Jansen, P. R. et al. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat. Genet.* **51**, 394–403, (2019).
17. Watanabe K. et al. Genome-wide meta-analysis of insomnia in over 2.3 million individuals implicates involvement of specific biological pathways through gene-prioritization. *Psychiatry Clin. Psychol.* http://medrxiv.org/lookup/doi/10.1101/2020.12.07.20245209 (2020).
18. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, 1005–D1012 (2019).
19. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
20. Sudlow, C. 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.* **12**, e1001779 (2015).
21. Gaziano, J. M. et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
22. Kaprio, J. et al. Genetic influences on use and abuse of alcohol: a study of 5638 adult Finnish twin brothers. *Alcohol Clin. Exp. Res.* **11**, 349–356 (1987).
23. Song, W. et al. Genome-wide association analysis of insomnia using data from Partners Biobank. *Sci. Rep.* **10**, 6928 (2020).
24. *Sleeping Aids Market Share, Size, Trends, Industry Analysis Report, By Product (Mattresses & Pillows, Sleep Laboratories, Medications, Sleep Apnea Devices); By Sleep Disorders; By Region; Segment Forecast, 2022–2030.* Report No. PM2287 (Polaris, 2022) https://www.researchandmarkets.com/reports/5569731/sleeping-aids-market-share-size-trends?utm_source=BW&utm_medium=PressRelease&utm_code=pbg5gh&utm_campaign=1699073+-+Global+Sleeping+Aids+Market+(2022+to+2030)+-+Share%2c+Size%2c+Trends%2c+Industry+Analysis+Report&utm_exec=jamu273prd.
25. Johnson, E. C. et al. A large-scale genome-wide association study meta-analysis of cannabis use disorder. *Lancet Psychiatry* **7**, 1032–1045 (2020).
26. Levey, D. F. et al. Reproducible genetic risk loci for anxiety: results from ~200,000 participants in the million veteran program. *Am. J. Psychiatry* **177**, 223–232 (2020).
27. Stein, M. B. et al. Genome-wide association analyses of post-traumatic stress disorder and its symptom subdomains in the Million Veteran Program. *Nat. Genet.* **53**, 174–184 (2021).
28. Levey, D. F. et al. Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat. Neurosci.* **24**, 954–963 (2021).
29. Lam, M. et al. Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* **51**, 1670–1678 (2019).
30. Howard, D. M. et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).
31. Aberg, K. A. et al. A comprehensive family-based replication study of schizophrenia genes. *JAMA Psychiatry* **70**, 573–81 (2013).
32. Ripke, S. et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
33. Lam, M. et al. Pleiotropic meta-analysis of cognition, education, and schizophrenia differentiates roles of early neurodevelopmental and adult synaptic pathways. *Am. J. Hum. Genet.* **105**, 334–350 (2019).
34. Smoller, J. W. et al. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet* **381**, 1360 (2013).
35. Thorp, J. G. et al. Genetic heterogeneity in self-reported depressive symptoms identified through genetic analyses of the PHQ-9. *Psychol. Med.* **50**, 2385–2396 (2020).
36. Khramtsova, E. A. et al. Sex differences in the genetic architecture of obsessive-compulsive disorder. *Am. J. Med. Genet. Part B* **180**, 351–364 (2019).
37. Ivarsdottir, E. V. et al. The genetic architecture of age-related hearing impairment revealed by genome-wide association analysis. *Commun. Biol.* **4**, 706 (2021).
38. Wang, H. et al. Genome-wide interaction analysis of pathological hallmarks in Alzheimer's disease. *Neurobiol. Aging* **93**, 61–68 (2020).
39. Lencer, R. et al. Genome-wide association studies of smooth pursuit and antisaccade eye movements in psychotic disorders: findings from the B-SNIP study. *Transl. Psychiatry* **7**, e1249 (2017).
40. Gharahkhani, P. et al. Genome-wide meta-analysis identifies 127 open-angle glaucoma loci with consistent effect across ancestries. *Nat. Commun.* **12**, 1258 (2021).
41. Nagel, M. et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet.* **50**, 920–927 (2018).
42. Koyama, S. et al. Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nat. Genet.* **52**, 1169–1177 (2020).
43. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
44. Baselmans, B. M. L. et al. Multivariate genome-wide analyses of the well-being spectrum. *Nat. Genet.* **51**, 445–451 (2019).
45. Werme, J., van der Sluis, S., Posthuma, D. & de Leeuw, C. A. An integrated framework for local genetic correlation analysis. *Nat. Genet.* **54**, 274–282 (2022).
46. Li, L., Wu, C., Gan, Y., Qu, X. & Lu, Z. Insomnia and the risk of depression: A meta-analysis of prospective cohort studies. *BMC Psychiatry* **16**, 375 (2016).
47. Nutt, D., Wilson, S. & Paterson, L. Sleep disorders as core symptoms of depression. *Dialogues Clin. Neurosci.* **10**, 329–336 (2008).

48. Gahr, M., Connemann, B. J., Zeiss, R. & Fröhlich, A. Schlafstörungen und Beeinträchtigungen des Schlafs als Nebenwirkungen von Psychopharmaka: eine Bewertung der Daten aus Fachinformationen. *Fortschr. Neurol. · Psychiatr.* **86**, 410–421 (2018).

49. Murphy, M. J. & Peterson, M. J. Sleep disturbances in depression. *Sleep Med. Clin.* **10**, 17–23 (2015).

50. Wichniak, A., Wierzbicka, A. & Jernajczyk, W. Sleep and antidepressant treatment. *Curr. Pharm. Des.* **18**, 5802–5817 (2012).

51. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* **39**, 276–293 (2015).

52. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**, 127–148 (2016).

53. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

54. Watson, N. F. et al. Recommended amount of sleep for a healthy adult: a joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society. *Sleep* https://academic.oup.com/sleep/article-lookup/doi/10.5665/sleep.4716 (2015).

55. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

56. Czeisler, C. A. et al. Exposure to bright light and darkness to treat physiologic maladaptation to night work. *N. Engl. J. Med.* **322**, 1253–1259 (1990).

57. Czeisler, C. A. & Gooley, J. J. Sleep and circadian rhythms in humans. *Cold Spring Harb. Symposia Quant. Biol.* **72**, 579–597 (2007).

58. Sivertsen, B., Overland, S., Krokstad, S. & Mykletun, A. Seasonal variations in sleep problems at latitude 63–65 in Norway: the Nord-Trondelag Health Study, 1995–1997. *Am. J. Epidemiol.* **174**, 147–153 (2011).

59. Leger, D., Bayon, V., Elbaz, M., Philip, P. & Choudat, D. Underexposure to light at work and its association to insomnia and sleepiness. A cross-sectional study of 13296 workers of one transportation company. *J. Psychosom. Res.* **70**, 29–36 (2011).

60. JRC Photovoltaic Geographical Information System (PVGIS)—European Commission. https://re.jrc.ec.europa.eu/pvg_tools/en/#PVP

61. National Solar Radiation Database. The National Renewable Energy Laboratory. https://nsrdb.nrel.gov

62. Core Team R. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2018).

63. Purcell, S. et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

64. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

65. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

66. Wickham H. *ggplot2: Elegant Graphics for Data Analysis. Media* (Springer-Verlag, New York, 2016).

67. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

68. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).

69. Gazal, S. et al. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).

70. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

71. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

72. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

73. Mancuso, N. et al. Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* **51**, 675–682 (2019).

74. Mullins, N. et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat. Genet.* **53**, 817–829 (2021).

75. Krämer, A., Green, J., Pollard, J. & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30**, 523–530 (2014).

76. Kuchenbaecker, K. et al. The transferability of lipid loci across African, Asian and European cohorts. *Nat. Commun.* **10**, 4330 (2019).

77. Myers, T. A., Chanock, S. J. & Machiela, M. J. LDlinkR: an R package for rapidly calculating linkage disequilibrium statistics in diverse populations. *Front. Genet.* **11**, 157 (2020).

78. Genomes Project, C. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

79. Luo, Y. et al. Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. *Hum. Mol. Genet.* **30**, 1521–1534 (2021).

80. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–41 (2015).

81. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).

82. Brown, B. C., Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).

83. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).

84. Adams, M. J. Consortium, Major Depressive Disorder Working Group of the psychiatric genomics. MDD2 (MDD2018) GWAS sumstats w/o UKBB. figshare p. 297769060 Bytes. https://figshare.com/articles/dataset/MDD2_MDD2018_GWAS_sumstats_w_o_UKBB/21655784/3 (2023).

85. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).

86. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).

87. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).

88. Zhao, Q., Wang, J., Hemani, G., Bowden, J. & Small, D. S. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Ann. Statist.* **48**, 1742–1769 (2020).

Acknowledgements

Author contributions

I.A.Z., D.F.L. and J.G. had primary responsibility for the design of the study. D.F.L., E.B. and J.G. supervised the study. I.A.Z. and D.F.L. had primary responsibility for the genetic and bioinformatics analyses, with support from O.G., J.D. M.G., K.A., H.Z., H.I., K.K., A.M., R.P. and J.G. O.G., K.K., H.I. and S.D. contributed to the initial quality control and data management of the UK Biobank data. The initial manuscript was drafted by I.A.Z., D.F.L. and J.G. Manuscript contributions and interpretation of results was provided by I.A.Z., D.F.J., D.B., R.P., M.B.S., E.B. and J.G. J.C., J.M.G. and D.G. contributed to other organisational and data-processing components of the study. All authors reviewed and approved the final manuscript.

Competing interests

D.J.B. has served as a paid or unpaid consultant to the National Cancer Institute, Pear Therapeutics, Sleep Number, Idorsia, Eisai, and Weight Watchers International. All consulting agreements have been for a total of <$5000 per year from any single entity. D.J.B. is an author of the Pittsburgh Sleep Quality Index, Pittsburgh Sleep Quality Index Addendum for PTSD (PSQI-A), Brief Pittsburgh Sleep Quality Index (B-PSQI), Daytime Insomnia Symptoms Scale, Pittsburgh Sleep Diary, Insomnia Symptom Questionnaire, and RU_SATED (copyrights held by University of Pittsburgh). These instruments have been licensed to commercial entities for fees. He is also co-author of the Consensus Sleep Diary (copyright held by Ryerson University), which is licensed to commercial entities for a fee. He has received grant support from NIH, PCORI, AHRQ, VA and Sleep Number. D.J.G. reports participation on Scientific Advisory Boards for Signifier Medical Technologies, Inc., Wesper, Inc., and Powell-Mansfield, Inc., outside of the submitted work. R.P. is paid for his editorial work in the journal Complex Psychiatry and received a research grant outside the scope of this study from Alkermes. M.B.S. has in the past 3 years received consulting income from Acadia Pharmaceuticals, Aptinyx, atai Life Sciences, BigHealth, Biogen, Bionomics, BioXcel Therapeutics, Boehringer Ingelheim, Clexio, Eisai, EmpowerPharm, Engrail Therapeutics, Janssen, Jazz Pharmaceuticals, NeuroTrauma Sciences, PureTech Health, Sage Therapeutics, Sumitomo Pharma, and Roche/Genentech. Dr. Stein has stock options in Oxeia Biopharmaceuticals and EpiVario. He has been paid for his editorial work on Depression and Anxiety (Editor-in-Chief), Biological Psychiatry (Deputy Editor), and UpToDate (Co-Editor-in-Chief for Psychiatry). He has also received research support from NIH, the Department of Veterans Affairs, and the Department of Defense. He is on the scientific advisory board for the Brain and Behavior Research Foundation and the Anxiety and Depression Association of America. J.G. is paid for editorial work for the journal Complex Psychiatry. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-41249-y.

Correspondence and requests for materials should be addressed to Joel Gelernter.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Department of Mental Health Neuroscience, Division of Psychiatry, University College London, London W1T 7BN, UK. [2]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London SE5 8AF, UK. [3]Division of Human Genetics, Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA. [4]Department of Psychiatry, Veterans Affairs Connecticut Healthcare Center, West Haven, CT, USA. [5]UCL Genetics Institute, Division of Biosciences, University College London, London WC1E 6BT, UK. [6]Health Data Research UK, Institute of Health Informatics, University College London, London NW1 2DA, UK. [7]Department of Genetics & Genomic Sciences and Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [8]School of Medicine, Yale University, New Haven, CT 06511, USA. [9]Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration (FDA), Silver Spring, MD, USA. [10]VA Boston Healthcare System, 1400 VFW Parkway (111PI), West Roxbury, MA 02132, USA. [11]Division of Sleep and Circadian Disorders, Brigham & Women's Hospital, Boston, MA, USA. [12]Division of Sleep Medicine, Harvard Medical School, Boston, MA, USA. [13]Psychiatry Service, VA San Diego Healthcare System, San Diego, CA, USA. [14]Departments of Psychiatry and Herbert Wertheim School of Public Health, University of California, San Diego, La Jolla, CA, USA. [15]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA. [16]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA 02130, USA. [17]Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, MA 02115, USA. [18]Institute of Cognitive Neuroscience, University College London, London, UK. ✉e-mail: joel.gelernter@yale.edu

the Million Veteran Program

Daniel F. Levey [3,4], **Joseph D. Deak** [3,4], **Marco Galimberti** [3,4], **Keyrun Adhikari** [3,4], **Hang Zhou** [3,4], **John Concato**[8,9],
J. Michael Gaziano[16,17], **Daniel J. Gottlieb**[10,11,12], **Renato Polimanti** [3,4], **Murray B. Stein** [13,14] & **Joel Gelernter** [3,4]✉