ARTICLES
https://doi.org/10.1038/s41588-018-0333-3
nature genetics

Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways

Philip R. Jansen[1,2], Kyoko Watanabe[1], Sven Stringer[1], Nathan Skene[3,4], Julien Bryois[5], Anke R. Hammerschlag[1], Christiaan A. de Leeuw[1], Jeroen S. Benjamins[6,7], Ana B. Muñoz-Manchado[3], Mats Nagel[1,8], Jeanne E. Savage[1], Henning Tiemeier[2,9], Tonya White[2], The 23andMe Research Team[10], Joyce Y. Tung[11], David A. Hinds[11], Vladimir Vacic[11], Xin Wang[11], Patrick F. Sullivan[4,12,13], Sophie van der Sluis[1,8], Tinca J. C. Polderman[1], August B. Smit[14], Jens Hjerling-Leffler[3], Eus J. W. Van Someren[15,16,17] and Danielle Posthuma[1,8,17]*

Insomnia is the second most prevalent mental disorder, with no sufficient treatment available. Despite substantial heritability, insight into the associated genes and neurobiological pathways remains limited. Here, we use a large genetic association sample ($n = 1,331,010$) to detect novel loci and gain insight into the pathways, tissue and cell types involved in insomnia complaints. We identify 202 loci implicating 956 genes through positional, expression quantitative trait loci, and chromatin mapping. The meta-analysis explained 2.6% of the variance. We show gene set enrichments for the axonal part of neurons, cortical and subcortical tissues, and specific cell types, including striatal, hypothalamic, and claustrum neurons. We found considerable genetic correlations with psychiatric traits and sleep duration, and modest correlations with other sleep-related traits. Mendelian randomization identified the causal effects of insomnia on depression, diabetes, and cardiovascular disease, and the protective effects of educational attainment and intracranial volume. Our findings highlight key brain areas and cell types implicated in insomnia, and provide new treatment targets.

Insomnia is the second most prevalent mental disorder[1]. One-third of the general population reports insomnia complaints. The diagnostic criteria for insomnia disorder[2] (that is, difficulties with initiating or maintaining sleep with accompanying daytime complaints at least three times a week for at least three months, which cannot be attributed to inadequate circumstances for sleep[3]) are met by 10% of individuals, and up to one-third of older age individuals[4]. Insomnia contributes significantly to the risk and severity of cardiovascular, metabolic, mood, and neurodegenerative disorders[2]. Despite evidence of a considerable genetic component (heritability 38–59%[5]), only a small number of genetic loci moderating the risk of insomnia have been identified thus far. Recent genome-wide association studies (GWAS)[6,7] for insomnia complaints ($n = 113,006$) demonstrated its polygenic architecture and implicated three genome-wide significant (GWS) loci and seven genes. A prominent role was reported for *MEIS1*, which is associated with insomnia complaints[6,7] and restless legs syndrome (RLS)[8] through pleiotropy and phenotypic overlap; yet, the role of other genes was not unambiguously shown. We set out to substantially increase the sample size to allow the detection of more genetic risk variants for insomnia complaints, which may aid in understanding its neurobiological mechanisms. By combining data collected in the UK Biobank (UKB) version 2[9] ($n = 386,533$) and 23andMe, a privately held personal genomics and biotechnology company[10,11] ($n = 944,477$), we obtained an unprecedented sample size of 1,331,010 individuals. Insomnia complaints were measured using questionnaire data; an independent sample (the Netherlands Sleep Register)[12], which gives access to similar question data, as well as clinical interviews assessing insomnia disorder (see Supplementary Note), was used to validate the specific questions so that they were good proxies of insomnia disorder.

[1]Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University Amsterdam, Amsterdam, the Netherlands. [2]Department of Child and Adolescent Psychiatry, Erasmus University Medical Center, Rotterdam, the Netherlands. [3]Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. [4]UCL Institute of Neurology, Queen Square, London, UK. [5]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [6]Department of Social, Health and Organisational Psychology, Utrecht University, Utrecht, the Netherlands. [7]Department of Experimental Psychology, Helmholtz Institute, Utrecht University, Utrecht, the Netherlands. [8]Department of Clinical Genetics, Section of Complex Trait Genetics, Amsterdam Neuroscience, VU University Medical Center, Amsterdam, The Netherlands. [9]Department of Psychiatry, Erasmus University Medical Center, Rotterdam, the Netherlands. [10]A list of members and affiliations appears at the end of the paper.[11]23andMe, Inc., Mountain View, CA, USA. [12]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [13]Department of Psychiatry, University of North Carolina, Chapel Hill, NC, USA. [14]Department of Molecular and Cellular Neurobiology, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University Amsterdam, Amsterdam, The Netherlands. [15]Department of Sleep and Cognition, Netherlands Institute for Neuroscience (an institute of the Royal Netherlands Academy of Arts and Sciences), Amsterdam, The Netherlands. [16]Departments of Psychiatry and Integrative Neurophysiology, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University, Amsterdam University Medical Center, Amsterdam, The Netherlands. [17]These authors contributed equally: Eus J.W. Van Someren, Danielle Posthuma. *e-mail: d.posthuma@vu.nl

We found 202 risk loci for insomnia; extensive functional in silico analyses showed the involvement of specific tissue and cell types. Mendelian randomization identified causal effects of insomnia on metabolic and psychiatric traits.

Results

Meta-analysis yields 202 risk loci. The UKB assessed insomnia complaints (hereafter referred to as 'insomnia') with a touchscreen device, whereas 23andMe research participants completed online surveys (Supplementary Tables 1 and 2). The assessment of insomnia in both samples shows high accuracy for insomnia disorder in the UKB and somewhat lower accuracy in 23andMe (sensitivity/specificity: UKB = 98/96%; 23andMe = 84/80%) (see Supplementary Note). The prevalence of insomnia was 28.3% in the UKB version 2 sample, 30.5% in the 23andMe sample, and 29.9% in the combined sample, which is in keeping with previous estimates for people of advanced age in the UK[4] and elsewhere[13,14]. Older people dominate the UKB (mean age = 56.7, s.d. = 8.0) and 23andMe (two-thirds of the sample older than 45, one-third older than 60 years of age) samples. Prevalence was higher in females (34.6%) than males (24.5%), yielding an odds ratio (OR) of 1.6, which is close to the 1.4 OR reported in a meta-analysis[15].

Quality control was conducted separately per sample, following standardized, stringent protocols (see Methods). The GWAS was run separately per sample (UKB: $n = 386{,}533$; 23andMe: $n = 944{,}477$) (Supplementary Fig. 1), and then meta-analyzed with METAL[16] by weighing the single nucleotide polymorphism (SNP) effect by sample size (see Methods). We first analyzed males and females separately (Supplementary Fig. 2) and observed a high genetic correlation between the sexes ($r_g = 0.92$, s.e.m. = 0.02; Supplementary Table 3), indicating strong overlap of genetic effects. Owing to the large sample size, the r_g of 0.92 was significantly different from 1 (one-sided Wald test, $P = 2.54 \times 10^{-6}$), suggesting a small role for sex-specific genetic risk factors, consistent with our previous study[6]. However, since sex-specific effects were relatively small, we focused on identifying genetic effects important in both sexes and continued with the combined sample. (Supplementary Tables 4 and 5 and the Supplementary Note provide sex-specific results.) The genetic correlation of insomnia between the full UKB and 23andMe results was $r_g = 0.69$ (s.e.m. = 0.02).

We observed a significant polygenic signal in the GWAS (lambda inflation factor = 1.808), which could not be ascribed to spurious association (linkage disequilibrium score intercept = 1.075)[17] (Supplementary Fig. 3a). Meta-analysis identified 11,990 GWS SNPs ($P < 5 \times 10^{-8}$), represented by 248 independent lead SNPs ($r^2 < 0.1$), located in 202 genomic risk loci (Fig. 1a, Supplementary Data Set 1, and Supplementary Tables 6 and 7). All lead SNPs showed concordant signs of effect in both samples (Supplementary Fig. 3b). We confirmed two (chr2:66,785,180 and chr5:135,393,752) out of six previously reported loci for insomnia[6,7] (Supplementary Table 8). Polygenic score (PGS) prediction in three randomly selected hold-out samples ($n = 3 \times 3{,}000$) estimated the current results to explain up to 2.6% of the variance in insomnia (Fig. 1b, Supplementary Fig. 4, and Supplementary Table 9).

The SNP-based heritability (h^2_{SNP}) was estimated at 7.0% (s.e.m. = 0.002). Partitioning the heritability by functional categories of SNPs (see Methods) showed the strongest enrichment of h^2_{SNP} in conserved regions (enrichment = 15.8, $P = 1.57 \times 10^{-14}$). In addition, h^2_{SNP} was enriched in common SNPs (minor allele frequency (MAF) > 0.3) and depleted in rarer SNPs (MAF < 0.01; Fig. 1c and Supplementary Table 10).

We used FUMA[18] to functionally annotate all SNPs in the risk loci that were in linkage disequilibrium ($r^2 \geq 0.6$) with one of the independent significant SNPs (see Methods). The majority of the 22,068 annotated SNPs (76.8%) were in open chromatin regions[19] as indicated by a minimum chromatin state of 1–7 (Fig. 1d and Supplementary Table 11). In line with findings for other traits[6,20], about half of these SNPs were in intergenic (35.5%) or non-coding RNA (13.0%) regions (Fig. 1e); of these, 0.72% were highly likely to have a regulatory function as indicated by a RegulomeDB score < 2 (see Methods). However, of these, 51.5% were located inside a protein-coding gene and 0.81% were exonic. Of the 177 exonic SNPs, 71 were exonic non-synonymous (Supplementary Table 12 and Supplementary Note). *WDR90* included four exonic non-synonymous SNPs (rs7190775, rs4984906, rs3752493, and rs3803697) all in high linkage disequilibrium with the same independent significant SNP (rs3184470). There were two exonic non-synonymous SNPs with extremely high combined annotation-dependent depletion (CADD) scores[21], suggesting a strong deleterious effect on protein function: rs13107325 in *SLC39A8* (locus 56, $P = 8.31 \times 10^{-16}$) with the derived allele T (MAF = 0.03), associated with an increased risk of insomnia; and rs35713889 in *LAMB2* (locus 42, $P = 1.77 \times 10^{-7}$), where the derived allele T of rs35713889 (MAF = 0.11) was also associated with an increased risk of insomnia complaints. Supplementary Table 13 provides a detailed overview of the functional impact of all variants in the genomic risk loci.

Genes implicated in insomnia. To obtain an insight into the (functional) consequences of individual GWS SNPs, we used FUMA[18] to apply three strategies to map associated variants to genes (see Methods). Positional gene mapping aligned SNPs to 412 genes by location. Expression quantitative trait loci (eQTL) gene mapping matched cis-eQTL SNPs to 594 genes whose expression levels they influence. Chromatin interaction mapping annotated SNPs to 159 genes based on three-dimensional DNA–DNA interactions between genomic regions of the GWS SNPs and nearby or distant genes (Supplementary Data Set 2, Supplementary Table 14, and Supplementary Note). Ninety-two genes were mapped by all three strategies (Supplementary Table 15), and 336 genes were physically located outside the risk loci but were implicated by eQTL associations (306 genes), chromatin interactions (16 genes), or both (14 genes). Several genes were implicated by GWS SNPs originating from two distinct risk loci on the same chromosome (Fig. 2a,b): *MEIS1*, located on chromosome 2 in the strongest associated locus (locus 20), was positionally mapped by 51 SNPs and mapped by chromatin interactions in 10 tissue types, including cross-loci interactions from locus 21, and is a known gene involved in insomnia[6]; and *LRGUK*, located on chromosome 7 in locus 106, was positionally mapped by 22 SNPs and chromatin interactions in 3 tissue types, including cross-loci interactions from locus 105. *LRGUK* was also implicated by eQTL associations of 125 SNPs in 14 general tissue types. *LRGUK* was previously implicated in type 2 diabetes[22] and autism spectrum disorder[23] (disorders with prominent insomnia). However, it is not yet directly implicated in sleep-related phenotype, and is the most likely candidate to explain the observed association at loci 105 and 106.

Apart from linking individually associated genetic variants to genes, we conducted a genome-wide gene association analysis (GWGAS) using MAGMA[24]. GWGAS provides aggregate association *P* values based on all variants located in a gene, and complements the three FUMA mapping strategies (see Methods). GWGAS identified 517 associated genes (Fig. 2c and Supplementary Table 16). The top gene *BTBD9* ($P = 8.51 \times 10^{-23}$) on chromosome 6 in locus 81 was also mapped using positional and eQTL mapping (tissue type: left ventricle of the heart), and is part of a pathway that regulates circadian rhythms. *BTBD9* has been associated with RLS, periodic limb movement disorder[25,26], and Tourette syndrome[27]. Involvement in sleep regulation was shown in *Drosophila*[28]; mouse mutants show fragmented sleep[29] and increased levels of dynamin 1[30], a protein that mediates the increased sleep onset latency that follows presleep arousal[31].

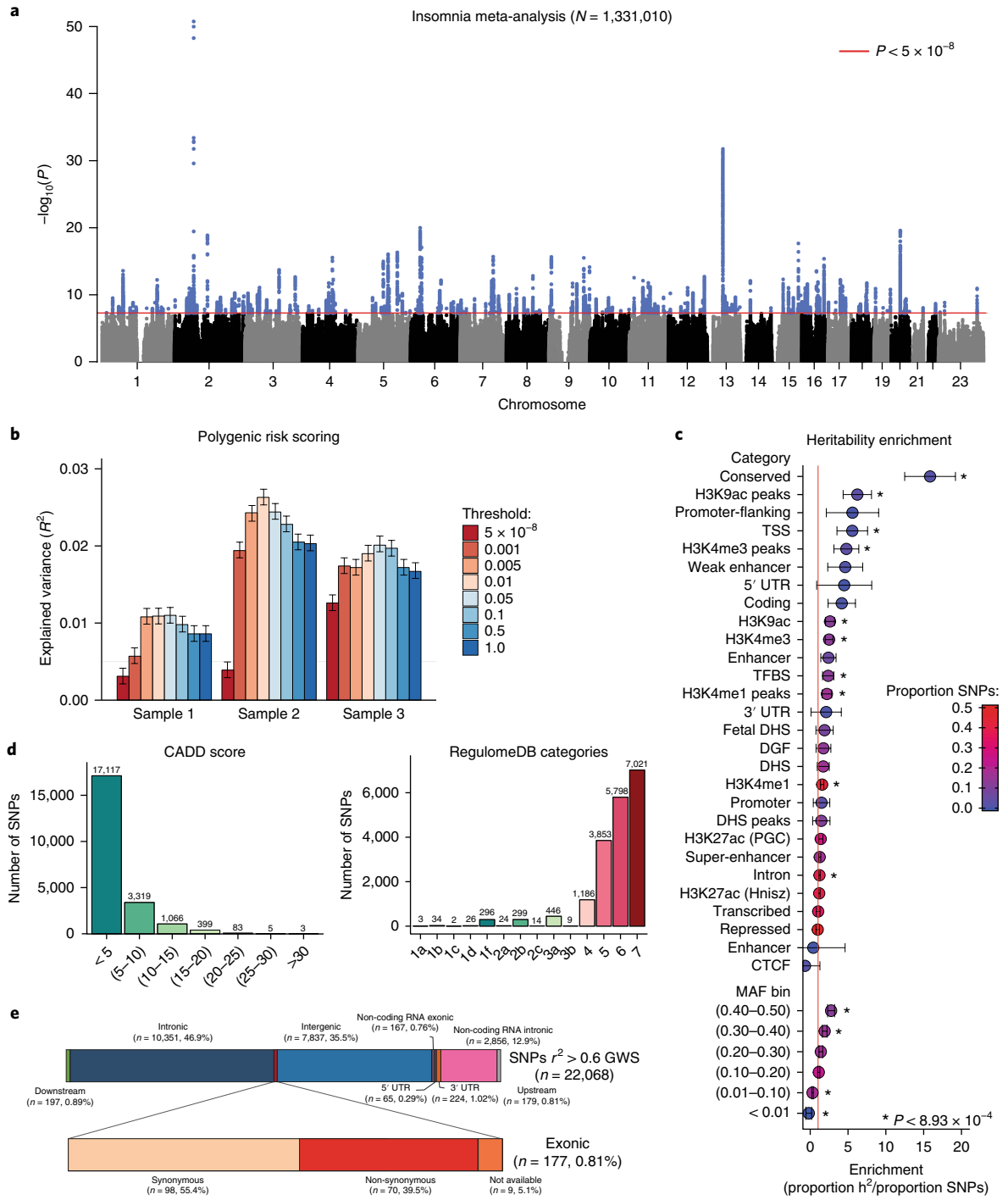

Fig. 1 | SNP-based results from the GWAS meta-analysis on insomnia in 1,331,010 individuals. **a**, Manhattan plot of the GWAS meta-analysis of insomnia in the UKB and 23andMe cohorts, showing the negative $\log_{10}$-transformed *P* value for each SNP. SNP two-sided *P* values from a linear model were calculated using METAL, weighting SNP associations by sample size. **b**, PGS prediction in three hold-out samples ($n = 3{,}000$), showing the increase in explained variance in insomnia (Nagelkerke's pseudo R^2) in a logistic regression model and 95% confidence intervals for each *P* value threshold. All *P* value thresholds were statistically significant. **c**, Heritability enrichment for functional SNP categories and MAF bins. Enrichment was calculated by dividing the proportion of heritability for each category by the proportion of SNPs in that category. The error bars show the 95% confidence interval around the estimate. Significant enrichments after Bonferroni correction (28 functional categories + 6 MAF bins + 22 chromosomes) are indicated by an asterisk ($P < 0.05/56$ categories $= 8.93 \times 10^{-4}$). TFBS, transcription factor binding site; DHS, DNase I hypersensitive site; DGF, digital genomic footprint; PGC, Psychiatric Genomics Consortium; Hnisz, as reported in Hnisz et al.; CTCF, CCCTC-binding factor. **d**, Distribution of CADD scores and RegulomeDB categories of all annotated SNPs in linkage disequilibrium ($r^2 \geq 0.6$) with one of the GWS SNPs ($n = 22{,}068$). **e**, Functional consequences of these annotated SNPs.

Of the 517 MAGMA-based associated genes, 222 were outside of the GWAS risk loci, and 309 were also mapped by FUMA. In total, 956 unique genes were mapped by at least one of the three FUMA gene mapping strategies or by MAGMA (Supplementary Fig. 5). Of these, *MEIS1*, *MED27*, *IPO7*, and *ACBD4* confirmed previous results[6,7] (Supplementary Table 17). Sixty-two genes were implicated

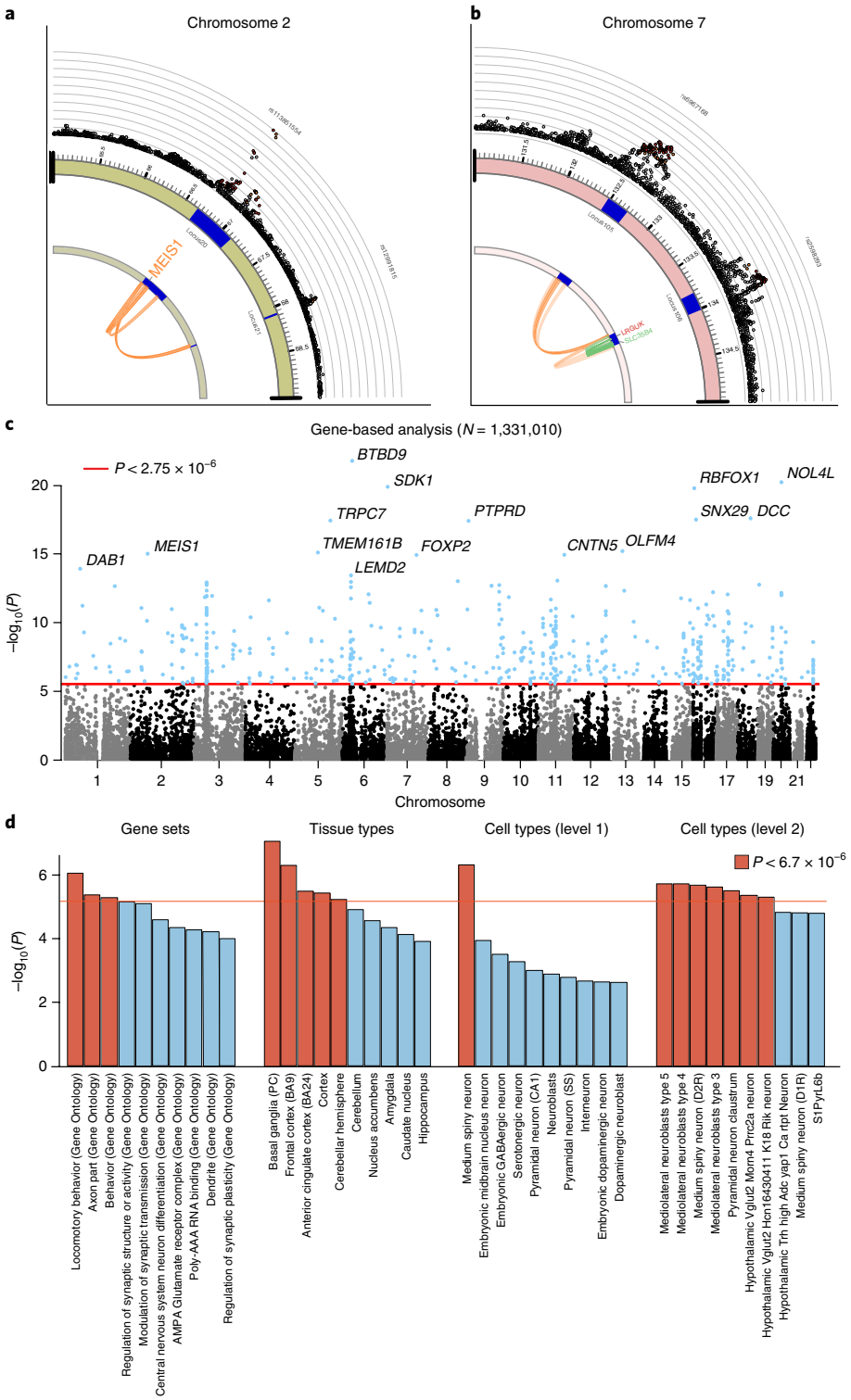

Fig. 2 | Gene-based and gene set analyses of insomnia in 1,331,010 individuals. **a**,**b**, Zoomed-in circos plots showing the genes implicated by two genomic risk loci on chromosome 2 (**a**) and chromosome 7 (**b**), with the genomic risk loci indicated as blue areas, eQTL associations in green, and chromatin interactions in orange. Genes mapped by both eQTL and chromatin interactions are red. The outer layer shows a Manhattan plot containing the negative $\log_{10}$-transformed *P* value of each SNP in the GWAS meta-analysis of insomnia in the UKB and 23andMe cohorts. Full circos plots of all autosomal chromosomes are provided in Supplementary Data Set 2. **c**, Genome-wide gene-based analysis (GWAS) of 18,185 genes that were tested for association with insomnia in MAGMA. The *y* axis shows the negative $\log_{10}$-transformed two-sided *P* value of the gene-based test, and the *x* axis shows the starting position on the chromosome. Gene-based two-sided *P* values were calculated with MAGMA. The red line indicates the Bonferroni-corrected threshold for genome-wide significance ($P = 0.05/18{,}185$ genes $= 2.75 \times 10^{-6}$). The top 15 most significant genes are highlighted. **d**, Gene set analysis of the top 10 for each of the MSigDB pathways, tissue expression of GTEx tissue types, and cell types from single-cell RNA sequencing. Gene set analyses were performed with MAGMA. The red line shows the Bonferroni significance threshold ($P < 0.05/7{,}473$ gene sets $= 6.7 \times 10^{-6}$), correcting for the total number of tested gene sets. The red bars indicate the significant gene sets.

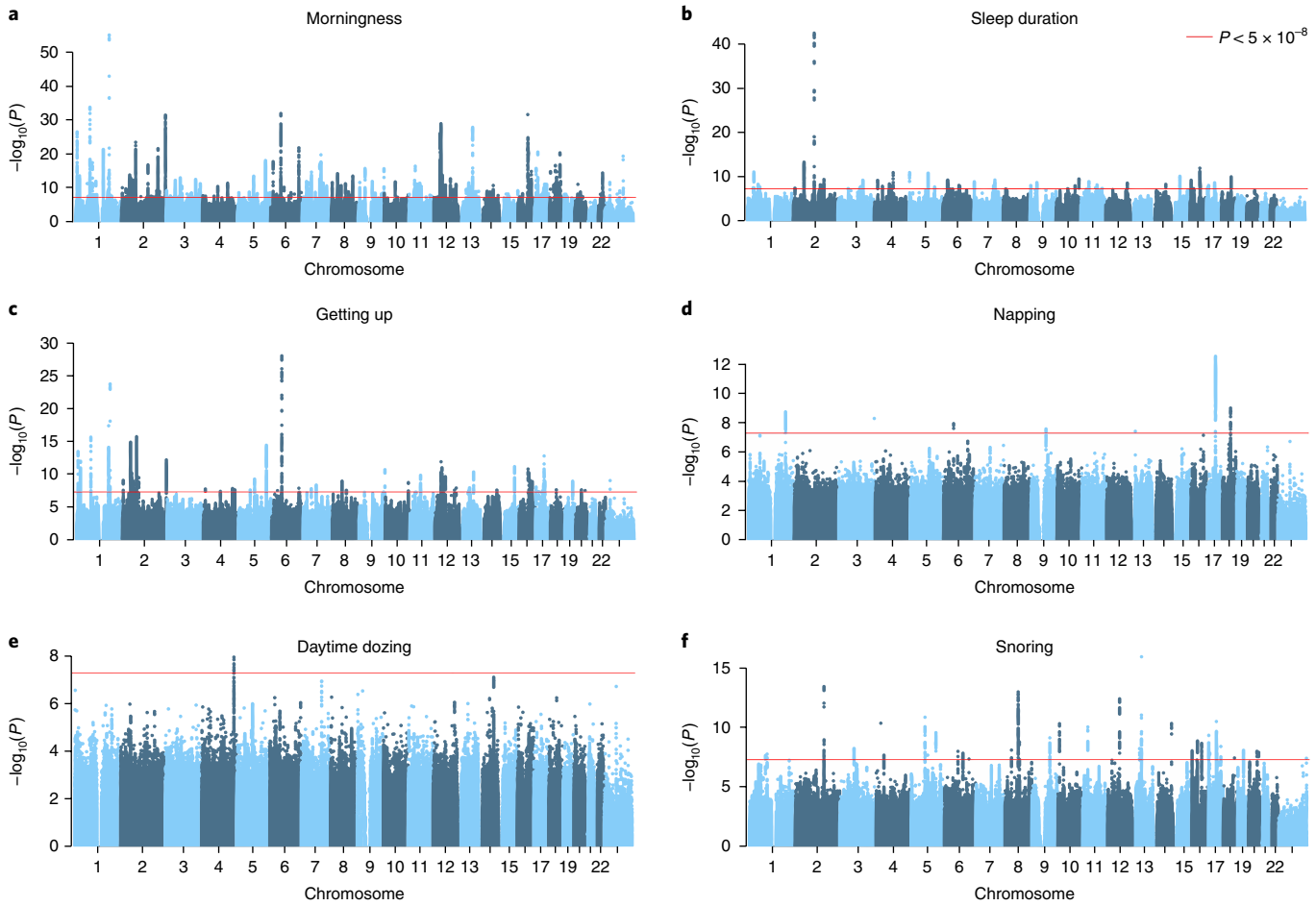

Fig. 3 | Genome-wide analyses of six sleep-related traits. a–f, Manhattan plots of the genome-wide analyses of (**a**) morningness (UKB and 23andMe cohorts, $n = 434{,}835$), (**b**) sleep duration (UKB, $n = 384{,}317$), (**c**) ease of getting up (UKB, $n = 385{,}949$), (**d**) napping (UKB, $n = 386{,}577$), (**e**) daytime dozing (UKB, $n = 386{,}548$), and (**f**) snoring (UKB, $n = 359{,}916$). The *y* axis shows the negative $\log_{10}$-transformed SNP two-sided *P* value from a linear or logistic regression model, and the *x* axis the base-pair position of the SNPs on each chromosome. The red line indicates the Bonferroni-corrected significance threshold ($P < 5 \times 10^{-8}$).

by all four mapping strategies, indicating that, apart from a GWS gene-based *P* value, there were: (1) GWS SNPs located in proximity of or inside these genes; (2) GWS SNPs associated with differential expression of these genes; and (3) GWS SNPs involved in genomic regions interacting with these genes. We note that genes that were indicated by positional mapping and GWS gene-based *P* values, but not via eQTL or chromatin interaction mapping ($n = 54$ genes), may be of equal importance; yet, they are more likely to exert their influence on insomnia via structural changes in gene products (that is, at the protein level) and not via quantitative changes in the availability of gene products.

Implicated pathways, tissues, and cell types. To test whether GWS genes converged in functional gene sets and pathways, we conducted gene-set analyses using MAGMA (see Methods). We tested the associations of 7,473 gene sets: 7,246 sets derived from the MSigDB[32]; gene expression values from 54 tissues from the GTEx database[33]; and cell-specific gene expression in 173 types of brain cells (Fig. 2d and Supplementary Table 18). Competitive testing was used and a Bonferroni-corrected threshold of $P < 6.7 \times 10^{-6}$ (0.05/7,473) to correct for multiple testing. Of the MSigDB gene sets, three Gene Ontology gene sets survived multiple testing: Gene Ontology:locomotory behavior ($P = 8.95 \times 10^{-7}$); Gene Ontology:behavior ($P = 5.23 \times 10^{-6}$); and Gene Ontology:axon part ($P = 4.25 \times 10^{-6}$). Twelve genes (*LRRK2*, *CRH*, *DLG4*, *DNM1*, *DRD1*, *DRD2*, *DRD4*, *GRIN1*, *NTSR1*, *SNCA*, *CNTN2*, and *CALB1*) were included in all of these gene sets, and two of these (*SNCA* and *DNM1*) had a GWS gene-based *P* value (Supplementary Table 19). *SNCA* encodes α-synuclein and has been implicated in rapid eye movement (REM) sleep behavior disorder[34] and Parkinson's disease[35]. Altered expression in mice changes sleep and wake electroencephalogram spectra[36] along the same dimensions that have been implicated in insomnia disorder[37]. *DNM1* encodes the synaptic neuronal protein dynamin 1, which is increased in *BTBD9* mutant mice[30] and mediates the sleep-disruptive effect of presleep arousal (see earlier; *BTBD9* is the top associated gene). Conditional gene-set analyses suggested that the association with the gene-set behavior is almost completely explained by the association of locomotory behavior, and that the effect of axon part is independent of this (Supplementary Note). Gene Ontology:locomotory behavior includes 175 genes involved in stimulus-evoked movement[38]. This set includes 16 GWS genes: *BTBD9*, *MEIS1*, *DAB1*, *SNCA*, *GNAO1*, *ATP2B2*, *NEGR1*, *SLC4A10*, *GIP*, *DNM1*, *GPRC5B*, *GRM5*, *NRG1*, *PARK2*, *TAL1*, and *OXR1*. Gene Ontology:axon part reflects a very general cellular component representing 219 genes, of which 14 were GWS (*KIF3B*, *SNCA*, *GRIA1*, *CDH8*, *ROBO2*, *DNM1*, *RANGAP1*, *GABBR1*, *P2RX3*, *NRG1*, *POLG*, *DAG1*, *NFASC*, and *CALB2*).

Tissue specific gene-set analyses showed strong enrichment of genetic signal in genes expressed in the brain. Correcting for overall expression, four specific brain tissues reached the threshold for

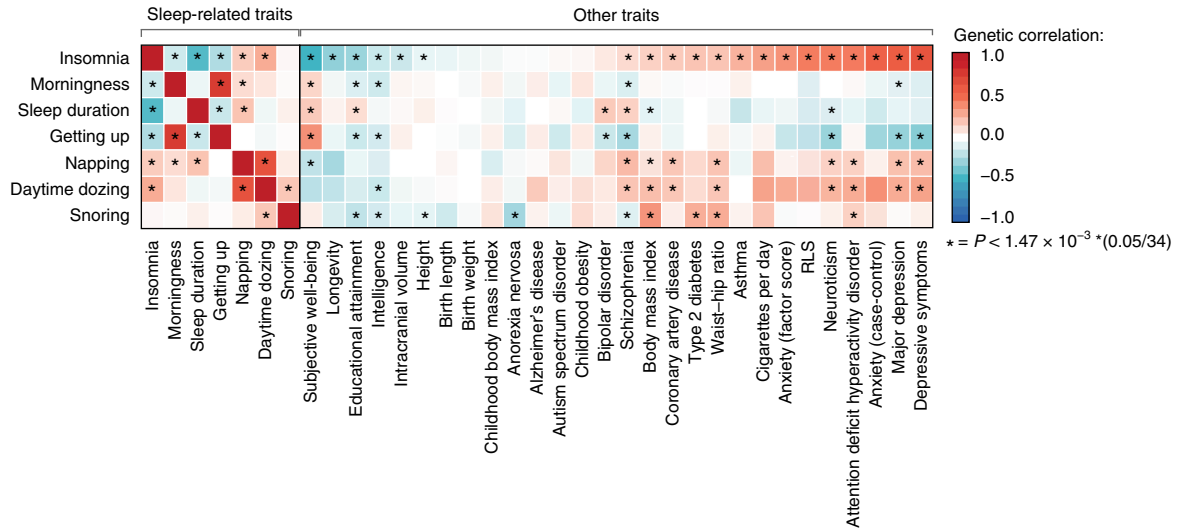

Fig. 4 | Genetic overlap of insomnia with other sleep-related traits and psychiatric and metabolic traits. Heatmap of genetic correlations between the insomnia GWAS meta-analysis, sleep-related phenotypes, and neuropsychiatric and metabolic traits studies. Genetic correlations and two-sided *P* values were calculated using linkage disequilibrium score regression. Red indicates a positive r_g, whereas green indicates a negative r_g. Correlations that were significant after Bonferroni correction ($P < 0.05/34$ traits $= 1.47 \times 10^{-3}$) are indicated with an asterisk (see also Supplementary Tables 21 and 29).

significance: the overall cerebral cortex ($P = 3.68 \times 10^{-6}$); Brodmann area 9 of the frontal cortex ($P = 5.04 \times 10^{-7}$); BA24 of the anterior cingulate cortex ($P = 3.25 \times 10^{-6}$); and the cerebellar hemisphere ($P = 5.93 \times 10^{-6}$). Several other brain tissues also showed strong enrichment just below the threshold, including three striatal basal ganglia structures (nucleus accumbens, caudate nucleus, putamen). To test whether genes expressed in all three basal ganglia structures together would show significant enrichment of low *P* values, we used the first principal component (BG_{PC}) of these basal ganglia structures (Methods) and found significant enrichment ($P = 8.33 \times 10^{-8}$). When conditioning the three top cortical structures on the BG_{PC}, they were no longer significantly associated after multiple testing correction (minimum $P = 0.03$), which was expected given that the BG_{PC} correlated strongly with gene expression in cortical (and other) areas ($r > 0.96$). Similar results were obtained vice versa; that is, using the first principal component of all cortical areas and conditioning the three basal ganglia structures on this resulted in no evidence of enrichment of low *P* values for basal ganglia structures (minimum $P = 0.53$). These results show that (1) genes expressed in the brain are important in insomnia, (2) genes expressed in cortical areas are more strongly associated than genes expressed in basal ganglia, and (3) there is a strong correlation between gene expression patterns across brain tissues, which suggests involvement of general cellular signatures rather than specific brain tissue structures.

Brain cell type-specific gene-set analyses were first carried out on 24 broad, cell-type categories. Cell type-specific gene expression was quantified using single-cell RNA sequencing of disassociated cells from the somatosensory cortex, hippocampus, hypothalamus, striatum, and midbrain from the mouse (see Methods), which closely resembles gene expression in humans[39]. Results indicated that genes expressed specifically in the medium spiny neurons (MSNs, $P = 4.83 \times 10^{-7}$) were associated with insomnia; no other broad, cell type-specific gene set survived our strict threshold of $P < 6.7 \times 10^{-6}$. MSNs represent 95% of neurons within the human striatum, which is one of the four major nuclei of the subcortical basal ganglia. Specifically, the striatum consists of the ventral (nucleus accumbens and olfactory tubercle) and dorsal (caudate nucleus and putamen) subdivisions. The association with MSNs thus likely explains the observed association of the basal ganglia striatal structures (nucleus accumbens, caudate nucleus, putamen).

Using broad cell classes risks not detecting associations that are specific to distinctive yet rare cell types. To account for this, we then tested 149 specific brain cell-type categories and found significant associations with 7 specific cell types: mediolateral neuroblasts type 3, 4, and 5 ($P = 2.36 \times 10^{-6}$, $P = 1.88 \times 10^{-6}$, and $P = 1.87 \times 10^{-6}$, respectively); D2-type MSNs ($P = 2.12 \times 10^{-6}$); claustrum pyramidal neurons ($P = 3.09 \times 10^{-6}$); hypothalamic Vglut2 Morn4 Prrc2a neurons ($P = 4.36 \times 10^{-6}$); and hypothalamic Vglut2 Hcn16430411 K18 Rik neurons ($P = 4.98 \times 10^{-6}$). The hypothalamus contains multiple nuclei that are key to the control of sleep and arousal, including the suprachiasmatic nucleus, which accommodates the biological clock of the brain[40]. These results suggest a role of distinct mature and developing cell types in the midbrain and hypothalamus.

Modest genetic overlap with sleep traits. Other sleep-related traits may easily be confounded with specific symptoms of insomnia, like early morning awakening, and difficulties maintaining sleep. The most recent genome-wide studies for other sleep-related traits included 59,128–128,266 individuals and assessed genetic effects on morningness[41–43] (that is, being a morning person), sleep duration[7,43], and daytime sleepiness/dozing[7]. Using increased sample sizes for each of these sleep-related traits (maximum $n = 434{,}835$), we investigated to what extent insomnia and other sleep-related traits are genetically distinct or overlapping. We performed GWAS analyses for the following six sleep-related traits: morningness; sleep duration; ease of getting up in the morning; taking naps during the day; daytime dozing; and snoring (Supplementary Note and Supplementary Figs. 6 and 7). Of the 202 risk loci for insomnia, 39 were also GWS in at least one of the other sleep-related traits (Fig. 3 and Supplementary Table 20). The strongest overlap in loci was found with sleep duration; 14 out of 49 sleep duration loci overlapped with insomnia. Insomnia showed the highest genetic correlation with sleep duration (−0.47, s.e.m. = 0.02; Supplementary Table 21) compared to other sleep-related traits; this was not surprising given that insomnia also shared the largest number of risk loci with sleep duration (see further discussion of results for sleep phenotypes in the Supplementary Note).

Gene mapping of SNP associations of sleep-related traits resulted in 973 unique genes (Supplementary Fig. 8 and Supplementary Tables 22–26). Gene-based analysis showed that, of the 517 GWS

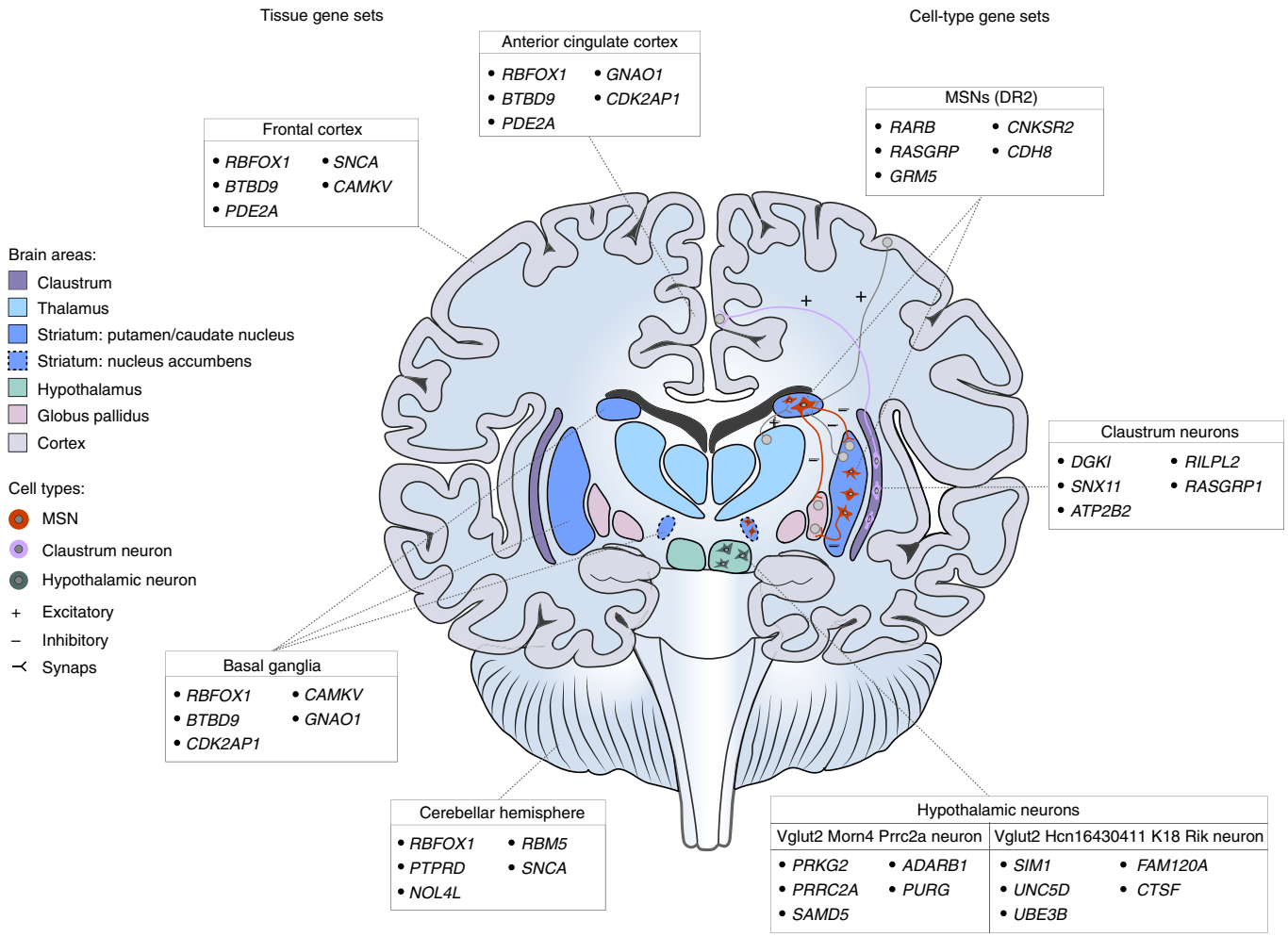

Fig. 5 | Overview of brain tissues and cell types associated with insomnia based on the GWAS results in 1,331,010 individuals. For each associated gene set, the top five genes driving the association are reported for each brain area and cell type. The results for the GTEx brain tissue type gene expression are shown on the left side, whereas the results from the level 2 single-cell gene expression are shown on the right.

genes for insomnia, 120 were GWS in at least one of the other sleep-related traits, and one gene (*RBFOX1*) was GWS in all traits except napping and daytime dozing (Supplementary Table 27). The largest proportion of overlap in GWS genes for insomnia was again with sleep duration, with 37 of the 134(27%) GWS genes for sleep duration being GWS for insomnia also. There was overlap in tissue enrichment in cortical structures and basal ganglia between insomnia and both morningness and sleep duration. At the single-cell level, MSNs were also implicated for morningness and sleep duration, but not for the other sleep-related traits (Supplementary Table 28). Taken together, these results suggest that, at a genetic level, insomnia shows considerable genetic overlap with sleep duration, and modest overlap with other sleep-related traits.

Strong overlap between insomnia and psychiatric traits. We confirm previously reported genetic correlations between insomnia and neuropsychiatric and metabolic traits, including type 2 diabetes, waist–hip ratio, and body mass index[6,41] (Supplementary Table 29), and also identify several GWS SNPs for insomnia that have previously been associated with these traits (Supplementary Table 30).

The strongest correlations were with depressive symptoms ($r_g = 0.64$, s.e.m. = 0.04, $P = 1.21 \times 10^{-71}$), followed by anxiety disorder ($r_g = 0.56$, s.e.m. = 0.11, $P = 1.40 \times 10^{-7}$), subjective well-being ($r_g = -0.51$, s.e.m. = 0.03, $P = 4.93 \times 10^{-52}$), major depression ($r_g = 0.50$, s.e.m. = 0.07, $P = 8.08 \times 10^{-12}$), and neuroticism ($r_g = 0.48$, s.e.m. = 0.02, $P = 8.72 \times 10^{-80}$). Genetic correlations with metabolic traits ranged between 0.09 and 0.20. Notably, we observed a positive correlation with RLS ($r_g = 0.44$, s.e.m. = 0.07, $P = 4.36 \times 10^{-10}$), a trait that shares phenotypic characteristics with insomnia[6]. This suggests a partial genetic overlap, which we discuss in more detail in the Supplementary Note and Supplementary Tables 31 and 32. In this study, we show that although insomnia lead SNPs are enriched in RLS, there is only a partial genome-wide overlap between insomnia and RLS, in line with previous analyses[6]. The genetic correlations between insomnia and anxiety and depression-related traits (anxiety, neuroticism, major depression, and depressive symptoms) were also stronger than the correlations between insomnia and the other sleep-related traits (Mann–Whitney *U*-test *Z* score = −2.56, $P = 0.01$). Since a similar high reliability has been reported for both sleep and psychiatric phenotypes, the findings suggest that genetically insomnia more closely resembles neuropsychiatric traits than other sleep-related traits (Fig. 4). These genetic correlations were consistent within the two meta-analyzed samples separately (Pearson's $r^2 = 0.98$; Supplementary Fig. 9). To infer directional associations between insomnia and these correlated traits, we performed bidirectional multi-SNP Mendelian randomization analysis[44] (see Methods). The results support a direct risk effect of insomnia on metabolic syndrome phenotypes including body mass index ($b_{xy} = 0.36$, s.e.m. = 0.05, $P = 1.25 \times 10^{-12}$), type 2 diabetes ($b_{xy} = 0.62$, s.e.m. = 0.11, $P = 2.29 \times 10^{-8}$), and coronary artery disease

(b_{xy} = 0.61, s.e.m. = 0.09, $P = 2.88 \times 10^{-12}$). We also found risk effects of insomnia on several psychiatric traits, including major depression (b_{xy} = 1.57, s.e.m. = 0.07, $P = 1.73 \times 10^{-111}$), schizophrenia ($b_{xy}$ = 0.68, s.e.m. = 0.10, $P = 5.12 \times 10^{-11}$), attention deficit hyperactivity disorder (b_{xy} = 0.77, s.e.m. = 0.06, $P = 2.50 \times 10^{-45}$), neuroticism (b_{xy} = 0.45, s.e.m. = 0.02, $P = 3.56 \times 10^{-92}$), and anxiety disorder (b_{xy} = 0.47, s.e.m. = 0.10, $P = 4.11 \times 10^{-6}$), with evidence of a reverse risk effect of major depression (b_{xy} = 0.06, s.e.m. = 0.003, $P = 6.93 \times 10^{-99}$) and neuroticism (b_{xy} = 0.24, s.e.m. = 0.01, $P = 7.90 \times 10^{-157}$) on insomnia. In addition, insomnia was bidirectionally associated with educational attainment (b_{xy} = −0.32, s.e.m. = 0.02, $P = 4.12 \times 10^{-45}$) and vice versa (b_{xy} = −0.10, s.e.m. = 0.01, $P = 2.27 \times 10^{-23}$); the same bidirectional pattern was observed for intelligence. Unidirectional protective effects were only observed for height (b_{xy} = −0.03, s.e.m. = 0.02, $P = 1.68 \times 10^{-77}$) and intracranial volume (b_{xy} = −0.03, s.e.m. = 0.01, $P = 3.72 \times 10^{-16}$). Using GWAS results from RLS in the 23andMe cohort, we observed patterns of bidirectional effects of insomnia on RLS (b_{xy} = 0.35, s.e.m. = 0.05, $P = 2.53 \times 10^{-12}$) and vice versa (b_{xy} = 0.12, s.e.m. = 0.01, $P = 1.21 \times 10^{-35}$). Overall, only a small proportion of SNPs showed pleiotropy between insomnia and other traits (Supplementary Table 33 and Supplementary Note).

Discussion

In the largest GWAS study to date of 1,331,010 participants, we identified 202 genomic risk loci for insomnia. Using extensive functional annotation of associated genetic variants, we demonstrated that the genetic component of insomnia points toward a role of genes enriched in locomotory behavior, and enriched in specific cell types from the claustrum, hypothalamus, and striatum, specifically in MSNs (Fig. 5). MSNs are γ-aminobutyric acid (GABA)ergic inhibitory cells and represent 95% of neurons in the human striatum, one of the four major nuclei of the basal ganglia (for reviews, see Vetrivelan et al.[45], Lazarus et al.[46], and Swardfager et al.[47]). MSNs were the first neurons in which the up and down states characteristic of slow-wave sleep were described[48]. Cell body-specific striatal lesions of the rostral striatum induce profound sleep fragmentation, which is highly characteristic of insomnia[45,49]. As discussed more extensively in the Supplementary Note, fragmented REM sleep is highly characteristic of insomnia and related to the ongoing thought-like mental content that makes patients with insomnia underestimate sleep duration[50–52]. Consistently short objective sleep across nights occurs only in a minority of patients with insomnia[53].

A role for the basal ganglia in sleep regulation is also suggested by the high prevalence of insomnia in neurodegenerative disorders, such as Parkinson's disease and Huntington's disease, where the basal ganglia are affected. Vetrivelan et al.[45] proposed a cortex-striatum-globus pallidus$_{external}$-cortex network involved in the control of sleep–wake behavior and cortical activation, where midbrain dopamine disinhibits the globus pallidus$_{external}$ and promotes sleep through the activation of D2 receptors in this network. Furthermore, brain imaging studies have suggested that the caudate nucleus of the striatum is a key node in the neuronal network imbalance of insomnia[54]; they also reported abnormal function in the cortical areas we found to be most enriched (BA9[55], BA24[56]). Our results support the involvement of the striato-cortical network in insomnia, by showing enrichment of risk genes for insomnia in cortical areas as well as the striatum, and specifically in MSNs. We recently showed that, along with several other cell types, MSNs mediate the risk for mood disorders[57] and schizophrenia[39]. MSNs are strongly implicated in reward processing; future work should address whether the genetic overlap between insomnia and mood disorders is mediated by gene function in MSNs.

Our results also showed enrichment of insomnia genes in the pyramidal neurons of the claustrum. This subcortical brain region is structurally closely associated with the amygdala and has been implicated in salience coding of incoming stimuli and binding of multisensory information into conscious percepts[58]. These functions are highly relevant to insomnia because the disorder is characterized by increased processing of incoming stimuli[59]. Claustrum activity during REM sleep is moreover key to activation of the anterior cingulate cortex that was also enriched for insomnia gene expression[60].

We found enrichment of insomnia genes in mediolateral neuroblasts from the embryonic midbrain and in two hypothalamic cell types. The role of the mediolateral neuroblasts is less clear; although they were obtained from the embryonic midbrain, at present it is unknown what type of mature neurons they differentiate into. We note that the midbrain is similar on a bulk transcriptomic level to the pons[61], and lacking cells from that region we cannot conclusively say that midbrain cell types are enriched.

The current findings provide an insight into the causal mechanism of insomnia, showing enrichment in specific cell types, brain areas, and biological functions. These findings are starting points for the development of new therapeutic targets for insomnia and may also provide valuable insights into other genetically related disorders.

URLs. GWAS Summary Statistics, https://ctg.cncr.nl/software/summary_statistics MAGMA, http://ctg.cncr.nl/software/magma FUMA GWAS, http://fuma.ctglab.nl PLINK 1.90 beta, https://www.cog-genomics.org/plink2 LD Hub, http://ldsc.broadinstitute.org/ldhub MSigDB Collections, http://software.broadinstitute.org/gsea/msigdb/collections.jsp METAL, http://genome.sph.umich.edu/wiki/METAL_Program LDSC (LD SCore), https://github.com/bulik/ldsc gsmr R-package, http://cnsgenomics.com/software/gsmr/ GTEx Portal, https://www.gtexportal.org/home/ BUHMBOX, http://software.broadinstitute.org/mpg/buhmbox/.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0333-3.

Received: 6 February 2018; Accepted: 13 December 2018;
Published online: 25 February 2019

References

1. Wittchen, H. U. et al. The size and burden of mental disorders and other disorders of the brain in Europe 2010. *Eur. Neuropsychopharmacol.* **21**, 655–679 (2011).
2. Morin, C. M. et al. Insomnia disorder. *Nat. Rev. Dis. Primers* **1**, 15026 (2015).
3. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* 5th edn (American Psychiatric Association Publishing, Washington, DC, 2013).
4. Morphy, H., Dunn, K. M., Lewis, M., Boardman, H. F. & Croft, P. R. Epidemiology of insomnia: a longitudinal study in a UK population. *Sleep* **30**, 274–280 (2007).
5. Lind, M. J., Aggen, S. H., Kirkpatrick, R. M., Kendler, K. S. & Amstadter, A. B. A longitudinal twin study of insomnia symptoms in adults. *Sleep* **38**, 1423–1430 (2015).
6. Hammerschlag, A. R. et al. Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits. *Nat. Genet.* **49**, 1584–1592 (2017).
7. 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).
8. Schormair, 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.* **16**, 898–907 (2017).
9. 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).
10. Eriksson, N. et al. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* **6**, e1000993 (2010).

11. Tung, J. Y. et al. Efficient replication of over 180 genetic associations with self-reported medical data. *PLoS ONE* **6**, e23473 (2011).
12. Benjamins, J. S. et al. Insomnia heterogeneity: characteristics to consider for data-driven multivariate subtyping. *Sleep Med. Rev.* **36**, 71–81 (2017).
13. Paparrigopoulos, T. et al. Insomnia and its correlates in a representative sample of the Greek population. *BMC Public Health* **10**, 531 (2010).
14. Cho, Y. W. et al. Epidemiology of insomnia in Korean adults: prevalence and associated factors. *J. Clin. Neurol.* **5**, 20–23 (2009).
15. Zhang, B. & Wing, Y.-K. Sex differences in insomnia: a meta-analysis. *Sleep* **29**, 85–93 (2006).
16. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
17. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
18. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
19. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
20. Sniekers, S. et al. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).
21. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
22. Laramie, J. M. et al. Polymorphisms near EXOC4 and LRGUK on chromosome 7q32 are associated with type 2 Diabetes and fasting glucose; the NHLBI Family Heart Study. *BMC Med. Genet.* **9**, 46 (2008).
23. Butler, M. G., Rafi, S. K. & Manzardo, A. M. High-resolution chromosome ideogram representation of currently recognized genes for autism spectrum disorders. *Int. J. Mol. Sci.* **16**, 6464–6495 (2015).
24. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
25. Kripke, D. F. et al. Genetic variants associated with sleep disorders. *Sleep Med.* **16**, 217–224 (2015).
26. Stefansson, H. et al. A genetic risk factor for periodic limb movements in sleep. *N. Engl. J. Med.* **357**, 639–647 (2007).
27. Janik, P., Berdyński, M., Safranow, K. & Żekanowski, C. The BTBD9 gene polymorphisms in Polish patients with Gilles de la Tourette syndrome. *Acta Neurobiol Exp (Wars)* **74**, 218–226 (2014).
28. Freeman, A. et al. Sleep fragmentation and motor restlessness in a *Drosophila* model of Restless Legs Syndrome. *Curr. Biol.* **22**, 1142–1148 (2012).
29. DeAndrade, M. P. et al. Motor restlessness, sleep disturbances, thermal sensory alterations and elevated serum iron levels in Btbd9 mutant mice. *Hum. Mol. Genet.* **21**, 3984–3992 (2012).
30. DeAndrade, M. P. et al. Enhanced hippocampal long-term potentiation and fear memory in Btbd9 mutant mice. *PLoS ONE* **7**, e35518 (2012).
31. Suzuki, A., Sinton, C. M., Greene, R. W. & Yanagisawa, M. Behavioral and biochemical dissociation of arousal and homeostatic sleep need influenced by prior wakeful experience in mice. *Proc. Natl Acad. Sci. USA* **110**, 10288–10293 (2013).
32. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
33. Ardlie, K. G. et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
34. Toffoli, M. et al. SNCA 3′UTR genetic variants in patients with Parkinson's disease and REM sleep behavior disorder. *Neurol. Sci.* **38**, 1233–1240 (2017).
35. Edwards, T. L. et al. Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. *Ann. Hum. Genet.* **74**, 97–109 (2010).
36. McDowell, K. A., Shin, D., Roos, K. P. & Chesselet, M.-F. Sleep dysfunction and EEG alterations in mice overexpressing alpha-synuclein. *J. Parkinsons Dis.* **4**, 531–539 (2014).
37. Colombo, M. A. et al. Wake high-density electroencephalographic spatiospectral signatures of insomnia. *Sleep* **39**, 1015–1027 (2016).
38. Harris, M. A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
39. Skene, N. G. et al. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
40. Saper, C. B., Scammell, T. E. & Lu, J. Hypothalamic regulation of sleep and circadian rhythms. *Nature* **437**, 1257–1263 (2005).
41. Lane, J. M. et al. Genome-wide association analysis identifies novel loci for chronotype in 100,420 individuals from the UK Biobank. *Nat. Commun.* **7**, 10889 (2016).
42. Hu, Y. et al. GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. *Nat. Commun.* **7**, 10448 (2016).
43. 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).
44. Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
45. Vetrivelan, R., Qiu, M.-H., Chang, C. & Lu, J. Role of basal ganglia in sleep–wake regulation: neural circuitry and clinical significance. *Front. Neuroanat.* **4**, 145 (2010).
46. Lazarus, M., Huang, Z.-L., Lu, J., Urade, Y. & Chen, J.-F. How do the basal ganglia regulate sleep–wake behavior? *Trends Neurosci.* **35**, 723–732 (2012).
47. Swardfager, W., Rosenblat, J. D., Benlamri, M. & McIntyre, R. S. Mapping inflammation onto mood: inflammatory mediators of anhedonia. *Neurosci. Biobehav. Rev.* **64**, 148–166 (2016).
48. Wilson, C. J. & Groves, P. M. Spontaneous firing patterns of identified spiny neurons in the rat neostriatum. *Brain Res.* **220**, 67–80 (1981).
49. Qiu, M., Vetrivelan, R., Fuller, P. M. & Lu, J. Basal ganglia control of sleep–wake behavior and cortical activation. *Eur. J. Neurosci.* **31**, 499–507 (2010).
50. Wassing, R. et al. Slow dissolving of emotional distress contributes to hyperarousal. *Proc. Natl Acad. Sci. USA* **113**, 2538–2543 (2016).
51. Feige, B. et al. Insomnia—perchance a dream? Results from a NREM/REM sleep awakening study in good sleepers and patients with insomnia. *Sleep* **41**, 10.1093/sleep/zsy032 (2018).
52. Krystal, A. D., Edinger, J. D., Wohlgemuth, W. K. & Marsh, G. R. NREM sleep EEG frequency spectral correlates of sleep complaints in primary insomnia subtypes. *Sleep* **25**, 630–640 (2002).
53. Johann, A. F. et al. Insomnia with objective short sleep duration is associated with longer duration of insomnia in the Freiburg Insomnia Cohort compared to insomnia with normal sleep duration, but not with hypertension. *PLoS ONE* **12**, e0180339 (2017).
54. Stoffers, D. et al. The caudate: a key node in the neuronal network imbalance of insomnia? *Brain* **137**, 610–620 (2014).
55. Altena, E. et al. Prefrontal hypoactivation and recovery in insomnia. *Sleep* **31**, 1271–1276 (2008).
56. Dai, X.-J. et al. Altered intrinsic regional brain spontaneous activity and subjective sleep quality in patients with chronic primary insomnia: a resting-state fMRI study. *Neuropsychiatr. Dis. Treat.* **10**, 2163–2175 (2014).
57. 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).
58. Mathur, B. N. The claustrum in review. *Front. Syst. Neurosci.* **8**, 48 (2014).
59. Wei, Y. et al. I keep a close watch on this heart of mine: increased interoception in insomnia. *Sleep* **39**, 2113–2124 (2016).
60. Renouard, L. et al. The supramammillary nucleus and the claustrum activate the cortex during REM sleep. *Sci. Adv.* **1**, e1400177 (2015).
61. Hawrylycz, M. et al. Canonical genetic signatures of the adult human brain. *Nat. Neurosci.* **18**, 1832–1844 (2015).

Acknowledgements

This work was funded by The Netherlands Organization for Scientific Research (NWO Brain and Cognition 433-09-228, NWO MagW VIDI 452-12-014, NWO VICI 435-13-005 and 453-07-001, and NWO 645-000-003). P.R.J. was funded by the Sophia Foundation for Scientific Research (S14-27), E.J.W.V.S. was funded by the European Research Council (grant no. ERC-ADG-2014-671084 INSOMNIA), and J.B. was funded by the Swiss National Science Foundation (grant no. P2GEP3_165049). N.G.S. was supported by the Wellcome Trust (grant no. 108726/Z/15/Z). J.H.L. was funded by the Swedish Research Council (Vetenskapsrådet, award no. 2014-3863), the Swedish Brain Foundation (Hjärnfonden) and the Wellcome Trust (grant no. 108726/Z/15/Z). Analyses were carried out on the Genetic Cluster Computer, which is financed by the NWO (480-05-003), by the VU University, Amsterdam, and by the Dutch Brain Foundation, and is hosted by the Dutch National Computing and Networking Services SurfSARA. This research has been conducted using the UK Biobank Resource (application no. 16406). We thank the UK Biobank and 23andMe research participants and employees for making this work possible.

Author contributions

D.P. and E.J.W.V.S. conceived the study. D.P. supervised the pre- and post-GWAS analysis pipeline. P.R.J. and K.W. performed the analyses. S.St. performed the quality control on the UKB data and wrote the analysis pipeline. K.W. wrote the online platform (FUMA) that was used for the follow-up analyses. C.d.L. conducted conditional gene-set analyses. J.B., N.S., A.M.M., and J.H.L. contributed single-cell RNA sequencing information. J.Y.T., D.A.H., V.V., X.W., and the 23andMe Research

Team contributed and analyzed the 23andMe cohort data. D.P., E.J.W.V.S., and P.R.J. wrote the paper. A.R.H., J.S.B., M.N., J.E.S., P.F.S., S.v.d.S., T.J.C.P. conducted part of the analyses. A.B.S. interpreted the findings in biological context and commented on the manuscript. H.T. and T.W. read the manuscript. All authors discussed the results and approved the final version of the paper.

Competing interests

P.F.S. is a grant recipient and advisor to Lundbeck A/G. D.A.H., J.Y.T., V.V., and X.W. are employees of 23andMe.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0333-3.

Correspondence and requests for materials should be addressed to D.P.

The 23andMe Research Team

Michelle Agee[11], Babak Alipanahi[11], Adam Auton[11], Robert K. Bell[11], Katarzyna Bryc[11], Sarah L. Elson[11], Pierre Fontanillas[11], Nicholas A. Furlotte[11], David A. Hinds[11], Karen E. Huber[11], Aaron Kleinman[11], Nadia K. Litterman[11], Jennifer C. McCreight[11], Matthew H. McIntyre[11], Joanna L. Mountain[11], Elizabeth S. Noblin[11], Carrie A. M. Northover[11], Steven J. Pitts[11], J. Fah Sathirapongsasuti[11], Olga V. Sazonova[11], Janie F. Shelton[11], Suyash Shringarpure[11], Chao Tian[11] and Catherine H. Wilson[11]

Methods

Meta-analysis. A meta-analysis of the GWAS results of insomnia and morningness in the UKB and 23andMe cohorts was performed using fixed-effects meta-analysis METAL[16], using SNP *P* values weighted by sample size. To investigate sex-specific genetic effects, we ran the meta-analysis between the UKB and 23andMe datasets for males and females separately.

Genomic risk loci definition. We used FUMA[18] version 1.2.4 (see URLs), an online platform for functional mapping and annotation of genetic variants, to define genomic risk loci and obtain functional information of the relevant SNPs in these loci. FUMA provides comprehensive annotation information by combining several external data sources. We first identified independent significant SNPs that had a GWS *P* value ($< 5 \times 10^{-8}$) and were independent from each other at $r^2 < 0.6$. These SNPs were further represented by lead SNPs, a subset of the independent significant SNPs that were in approximate linkage equilibrium with each other at $r^2 < 0.1$. We then defined independent genomic risk loci by identifying physical regions in linkage disequilibrium with these lead SNPs that were > 250 kilobases (kb) apart from each other. The borders of the genomic risk loci were defined by identifying all SNPs in linkage disequilibrium ($r^2 \geq 0.6$) with one of the independent significant SNPs in the locus; the region containing all these candidate SNPs was considered to be a single independent genomic risk locus. Linkage disequilibrium information was calculated using the UKB genotype data as a reference. Risk loci were defined based on evidence from independent significant SNPs available in both 23andMe and UKB datasets.

SNPs that were GWS but only available in the 23andMe dataset were not included when defining genomic risk loci and were not included in any follow-up annotations or analyses because there was no external replication in the UKB sample. If such SNPs were located in a risk locus, they are displayed in LocusZoom plots (gray, as there is no linkage disequilibrium information in the UKB). When risk loci contained GWS SNPs based solely on the 23andMe dataset, we did not count that risk locus because there were no other SNPs available in both samples that supported these GWS SNPs.

Gene-based analysis. SNP-based *P* values from the meta-analysis were used as input for the GWGAS; 18,182–18,185 protein-coding genes (each containing at least one SNP in the GWAS, the total number of tested genes can thus be slightly different across phenotypes) from the NCBI 37.3 gene definitions were used as the basis for the GWGAS in MAGMA[24]. Bonferroni correction was applied to correct for multiple testing ($P < 2.75 \times 10^{-6}$).

Gene-set analysis. Results from the GWGAS analyses were used to test for association in three types of 7,473 predefined gene sets:

1. 7,246 curated gene sets representing known biological and metabolic pathways derived from 9 data resources, cataloged by and obtained from the MSigDB version 6.0 (ref. [62], see URLs).
2. Gene expression values from 53 tissues obtained from GTEx[33], $\log_2$-transformed with pseudocount 1 after winsorization at 50 and averaged per tissue (+1 combined gene expression in the basal ganglia by taking the first principal component from principal component analysis of gene expression in three basal ganglia structures). We caution that only a limited set of brain tissues were included; thus, we cannot rule out associations with many important areas such as the pons, midbrain, or thalamus based on this analysis.
3. Cell type-specific expression in 173 types of brain cells (24 broad categories of cell types 'level 1', and 149 specific categories of cell types 'level 2'), which were calculated following the method described by Skene et al.[39]. Briefly, brain cell-type expression data was drawn from single-cell RNA sequencing data from mouse brains. For each gene, the value for each cell type was calculated by dividing the mean unique molecular identifier counts for the given cell type by the summed mean unique molecular identifier counts across all cell types. Single-cell gene sets were derived by grouping genes into 40 equal bins based on specificity of expression. Mouse cell gene expression was shown to closely approximate gene expression in postmortem human tissue[39].

These gene sets were tested using MAGMA. MAGMA uses a continuous measure of association (gene-based *P* value) of all genes that could be mapped by at least one SNP in the gene-based test and can perform gene-set analysis based on dichotomous gene sets (genes present in a gene set or not) or continuous values of gene expression in tissues and cells. We computed competitive *P* values, which represent the test of association for a specific gene set compared with genes not in the gene set to correct for the baseline level of genetic association in the data[63]. The Bonferroni-corrected significance threshold was $P = 0.05/7{,}473$ gene sets $= 6.7 \times 10^{-6}$. Conditional analyses were performed as a follow-up using MAGMA to test whether each significant association observed was independent of all others. The association between each gene set in each of the three categories was tested conditional on the most strongly associated set, and then, if any substantial ($P < 0.05$/number of gene sets) associations remained, by conditioning on the first and second most strongly associated set, and so on until no associations remained. Gene sets that retained their association after correcting for other sets were considered to represent independent signals. We note that this is not a test of association per se, but rather a strategy to identify, among gene sets with known significant associations and overlap in genes, which set(s) are responsible for driving the observed association.

SNP-based heritability and genetic correlation. Linkage disequilibrium score regression[17] was used to estimate genomic inflation and SNP-based heritability of the phenotypes, and to estimate the cross-cohort genetic correlations. Precalculated linkage disequilibrium scores from the 1000 Genomes European reference population were obtained from https://data.broadinstitute.org/alkesgroup/LDSCORE/.

Genetic correlations. Genetic correlations between sleep-related traits, and between sleep-related traits and previously published GWAS studies of sufficient sample size were calculated using linkage disequilibrium score regression on HapMap 3 SNPs only. Genetic correlations were corrected for multiple testing based on the total number of correlations (between 6 sleep-related phenotypes and 28 previous GWAS studies) by applying a Bonferroni-corrected threshold of $P < 0.05/34 = 1.47 \times 10^{-3}$.

Stratified heritability. To test whether specific categories of SNP annotations were enriched for heritability, we partitioned SNP heritability for binary annotations using stratified linkage disequilibrium score regression[64]. Heritability enrichment was calculated as the proportion of heritability explained by an SNP category divided by the proportion of SNPs that are in that category. Partitioned heritability was computed by 28 functional annotation categories, by MAF in six percentile bins, and by 22 chromosomes. Annotations for binary categories of functional genomic characteristics (for example, coding or regulatory regions) were obtained from the LD Score website (see URLs). The Bonferroni-corrected significance threshold for 56 annotations was set at $P < 0.05/56 = 8.93 \times 10^{-4}$.

Functional annotation of SNPs. Functional annotation of SNPs implicated in the meta-analysis was performed using FUMA[17]. We selected all candidate SNPs in genomic risk loci having an $r^2 \geq 0.6$ with one of the independent significant SNPs (see above), a *P* value ($P < 1 \times 10^{-5}$), a MAF > 0.0001 for annotations, and availability in both UKB and 23andMe datasets. The functional consequences for these SNPs were obtained by matching each SNP's chromosome location, base-pair position, reference, and alternate alleles to databases containing known functional annotations, including ANNOVAR[65] categories, CADD scores, RegulomeDB[20] scores, and chromatin state[66]. ANNOVAR categories identify the SNP's genic position (for example, intron, exon, intergenic) and associated function. CADD scores predict how deleterious the effect of an SNP is likely to be for a protein structure/function, with higher scores representing higher deleteriousness. A CADD score > 12.37 is potentially pathogenic[21]. The RegulomeDB score is a categorical score based on information from eQTLs and chromatin marks, which ranges from 1a to 7 with lower scores indicating an increased likelihood of having a regulatory function. Scores are as follows: 1a = eQTL + transcription factor binding + matched transcription factor motif + matched DNase footprint + DNase peak; 1b = eQTL + transcription factor binding + any motif + DNase footprint + DNase peak; 1c = eQTL + transcription factor binding + matched transcription factor motif + DNase peak; 1d = eQTL + transcription factor binding + any motif + DNase peak; 1e = eQTL + transcription factor binding + matched transcription factor motif; 1f = eQTL + transcription factor binding/DNase peak; 2a = transcription factor binding + matched transcription factor motif + matched DNase footprint + DNase peak; 2b = transcription factor binding + any motif + DNase footprint + DNase peak; 2c = transcription factor binding + matched transcription factor motif + DNase peak; 3a = transcription factor binding + any motif + DNase peak; 3b = transcription factor binding + matched transcription factor motif; 4 = transcription factor binding + DNase peak; 5 = transcription factor binding or DNase peak; 6 = other; 7 = not available. The chromatin state represents the accessibility of genomic regions (every 200 base pairs (bp)) with 15 categorical states predicted by a hidden Markov model based on 5 chromatin marks for 127 epigenomes in the Roadmap Epigenomics Project[67]. A lower state indicates higher accessibility, with states 1–7 referring to open chromatin states. We annotated the minimum chromatin state across tissues to SNPs. The 15 core chromatin states as suggested by the Roadmap Epigenomics Project are as follows: 1 = active transcription start site (TSS); 2 = flanking active TSS; 3 = transcription at gene 5′ and 3′; 4 = strong transcription; 5 = weak transcription; 6 = genic enhancers; 7 = enhancers; 8 = zinc finger genes and repeats; 9 = heterochromatic; 10 = bivalent/poised TSS; 11 = flanking bivalent/poised TSS/enhancer; 12 = bivalent enhancer; 13 = repressed polycomb; 14 = weak repressed polycomb; 15 = quiescent/low.

Gene mapping. GWS loci obtained by GWAS were mapped to genes in FUMA[18] using three strategies:

1. Positional mapping maps SNPs to genes based on physical distance (within a 10-kb window) from known protein-coding genes in the human reference assembly (GRCh37/hg19).

2. eQTL mapping maps SNPs to genes with which they show a significant eQTL association (that is, allelic variation at the SNP is associated with the

expression level of that gene). eQTL mapping uses information from 45 tissue types in 3 data repositories (GTEx[33], Blood eQTL browser[66], BIOS QTL browser[68]), and is based on cis-eQTLs that can map SNPs to genes up to 1 megabase apart. We used a false discovery rate of 0.05 to define significant eQTL associations.

3. Chromatin interaction mapping was performed to map SNPs to genes when there is a three-dimensional DNA–DNA interaction between the SNP region and another gene region. Chromatin interaction mapping can involve long-range interactions since it does not have a distance boundary. FUMA currently contains Hi-C data of 14 tissue types from the study of Schmitt et al.[69]. Since chromatin interactions are often defined in a certain resolution, such as 40 kb, an interacting region can span multiple genes. If an SNP is located in a region that interacts with a region containing multiple genes, it will be mapped to each of those genes. To further prioritize candidate genes, we selected only interaction-mapped genes where one region involved in the interaction overlaps with a predicted enhancer region in any of the 111 tissue/cell types from the Roadmap Epigenomics Project[67], and the other region is located in a gene promoter region (250 bp upstream and 500 bp downstream of the TSS and also predicted by the Roadmap Epigenomics Project to be a promoter region). This method reduces the number of genes mapped but increases the likelihood that those identified will indeed have a plausible biological function. We used a P false discovery rate $< 1 \times 10^{-5}$ to define significant interactions, based on previous recommendations[69] and modified to account for the differences in the cell lines used in this study.

GWAS catalog lookup. We used FUMA to identify SNPs with previously reported ($P < 5 \times 10^{-5}$) phenotypic associations in published GWAS listed in the NHGRI-EBI catalog[70], which matched with SNPs in linkage disequilibrium with one of the independent significant SNPs identified in the meta-analysis.

Polygenic risk scoring. To calculate the explained variance in insomnia by our GWAS results, we calculated PGS based on the SNP effect sizes in the meta-analysis. The PGS were calculated using two methods: LDpred[71] and PRSice[72], a script for calculating P value thresholded PGS in PLINK (see URLs). PGS were calculated using a leave-one-out method, where summary statistics were recalculated each time with one sample of $n = 3{,}000$ from the UKB dataset excluded from the analysis. This sample was then used as a target sample for estimating the explained variance in insomnia by the PGS.

Mendelian randomization. To investigate causal associations between insomnia and genetically correlated traits, we analyzed the direction of effects using generalized summary-data-based Mendelian randomization[44] (see URLs). This method uses effect sizes from GWAS summary statistics (standardized betas or log-transformed ORs) to infer causality of effects between two traits based on GWS SNPs. Built-in HEIDI outlier detection was applied to remove SNPs with pleiotropic effects on both traits, since these may bias the results. We tested for causal associations between insomnia and traits that were significantly genetically correlated with insomnia (b_{zx}). In addition, we tested for bidirectional associations by using other traits as the determinant and insomnia as the outcome (b_{zy}). We selected independent ($r^2 < 0.1$) lead SNPs with a GWS $P < 5 \times 10^{-8}$ as instrumental variables in the analyses. For traits with < 10 lead SNPs (that is, the minimum number of SNPs on which generalized summary-data-based Mendelian randomization can perform a reliable analysis) we selected independent SNPs ($r^2 < 0.1$), with a $P < 1 \times 10^{-5}$. If the outcome trait is binary, the estimated b_{zx} and b_{zy} are approximately equal to the natural log of the OR. An OR of 2 can be interpreted as a doubled risk compared to the population prevalence of a binary trait for every s.d. increase in the exposure trait. For quantitative traits, b_{zx} and b_{zy} can be interpreted as a 1 s.d. increase explained in the outcome trait for every s.d. increase in the exposure trait.

Statistical analysis. SNP associations were tested using linear or logistic regression models depending on the sleep phenotype. We report two-sided P values of each statistical test unless otherwise specified.

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

Data availability

The data analyzed in the current study were partly provided by the UK Biobank Study (www.ukbiobank.ac.uk), received under UK Biobank application no. 16406. Our policy is to make genome-wide summary statistics (sumstats) publicly available. Sumstats from the GWAS conducted are available for download from the CNCR Complex Trait Genetics lab at https://ctg.cncr.nl/; see also https://ctg.cncr.nl/software/summary_statistics. Note that our freely available meta-analytic sumstats (insomnia and morningness) represent results excluding the 23andMe sample. This is a non-negotiable clause in the 23andMe data transfer agreement, intended to protect the privacy of the 23andMe research participants. To fully recreate our meta-analytic results for insomnia and morningness: (1) obtain insomnia and morningness sumstats from 23andMe; (2) conduct a meta-analysis of our sumstats with the 23andMe sumstats. 23andMe participant data are shared according to community standards that have been developed to protect against breaches of privacy. Currently, these standards allow for the sharing of summary statistics for at most 10,000 SNPs. The full set of summary statistics can be made available to qualified investigators who enter into an agreement with 23andMe that protects participants' confidentiality. Interested investigators should email dataset-request@23andme.com for more information.

References

62. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
63. de Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* **17**, 353–364 (2016).
64. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
65. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
66. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
67. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
68. Zhernakova, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
69. Schmitt, A. D. et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17**, 2042–2059 (2016).
70. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
71. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
72. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).

nature research

Corresponding author(s): Danielle Posthuma

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.

For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

We included a sample size of 1,331,010 individuals, by combining results from UK Biobank and 23andMe, in order to maximize power for the detection of SNPs with low individual effects.

2. Data exclusions

Describe any data exclusions.

We excluded individuals based on non-European ancestry, low genotype quality, relatedness and missing phenotype

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

We provide replication results between the two large datasets (23andMe and UKB). To support external replication, we choose to make our UK Biobank GWAS results publicly available upon publication for the scientific community. These summary statistics, however, only include results from the UK Biobank study, as restrictions prohibit the publication of results in the 23andMe sample. In our methods section, we describe the steps other researchers need to take to obtain the same results.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No randomization procedures were used. We controlled for possible population stratification by including genetic principal components in all genome-wide analyses after selection of European individuals

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Not applicable. Group allocation was based on answers to sleep-related questions, assessing insomnia complaints. There was no treatment or intervention so blinded allocation was not applicable

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

Plink (version 1.9), MAGMA (version 1.05b), FLASHPCA (version 2.0), LD Score regression (version 1.1.0), LDpred (0.9.09), GSMR (version 1.0.5), METAL (version 2011-03-25), R (version 3.3.1), FUMA (online platform, fuma.ctglab.nl), BUHMBOX (version 0.38)

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

UK Biobank data has been provided under an approved application (Application number: 16406). 23andMe summary statistics can be accessed after applying to 23andMe.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used in the study

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used in the study

b. Describe the method of cell line authentication used.

No cell lines were used in the study

c. Report whether the cell lines were tested for mycoplasma contamination.

No cell lines were used in the study

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No cell lines were used in the study

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

No animals or animal-derived material were used in the study

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

We used data of participants of the UK Biobank Study. Data were previously collected at one of the UK Biobank research centers in the UK, and included participants between the age of 40 and 69 years old. The 23andMe study collected online questionnaire data of customers that had previously been genotyped. Genome-wide association analysis was corrected for age, sex and genetic principal components.