Genome-wide meta-analysis of insomnia prioritizes genes associated with metabolic and psychiatric pathways

23Andme Research Team

published in
Nature genetics
2022

DOI (link to publisher)
10.1038/s41588-022-01124-w

document version
Publisher's PDF, also known as Version of record

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-022-01124-w

Check for updates

Genome-wide meta-analysis of insomnia prioritizes genes associated with metabolic and psychiatric pathways

Kyoko Watanabe[1], Philip R. Jansen[1,2], Jeanne E. Savage[1], Priyanka Nandakumar[3], Xin Wang[3], 23andMe Research Team*, David A. Hinds[3], Joel Gelernter[4,5], Daniel F. Levey[4,5], Renato Polimanti[4,5], Murray B. Stein[6,7], Eus J. W. Van Someren[8,9], August B. Smit[10] and Danielle Posthuma[1,11] ✉

Insomnia is a heritable, highly prevalent sleep disorder for which no sufficient treatment currently exists. Previous genome-wide association studies with up to 1.3 million subjects identified over 200 associated loci. This extreme polygenicity suggested that many more loci remain to be discovered. The current study almost doubled the sample size to 593,724 cases and 1,771,286 controls, thereby increasing statistical power, and identified 554 risk loci (including 364 novel loci). To capitalize on this large number of loci, we propose a novel strategy to prioritize genes using external biological resources and functional interactions between genes across risk loci. Of all 3,898 genes naively implicated from the risk loci, we prioritize 289 and find brain-tissue expression specificity and enrichment in specific gene sets of synaptic signaling functions and neuronal differentiation. We show that this novel gene prioritization strategy yields specific hypotheses on underlying mechanisms of insomnia that would have been missed by traditional approaches.

Insomnia is a sleep disorder characterized by difficulty in falling or remaining asleep. It is highly prevalent in the population[1] and is associated with high morbidity, mortality[2] and societal costs[3]. It is moderately heritable (twin-based heritability 38–59% (ref. [4]), and single-nucleotide polymorphism (SNP)-based heritability 7% (ref. [5])) and genome-wide association studies (GWAS) have improved understanding of the complex polygenic etiology of insomnia[5–7]. A recent GWAS in over 1.3 million individuals reported >200 genomic loci linked to insomnia, in which the polygenic risk score (PRS) explained a quarter of the estimated heritability[5], implicated several neurobiological processes, cell types, brain areas and circuitries and showed considerable overlap with genetic risk for psychiatric disorders[5,7].

Previous genetic studies have, however, also shown that insomnia is among the most polygenic traits[8], predicted to require at least 50 million individuals for detection of SNPs at the level of genome-wide significance ($P < 5 \times 10^{-8}$) to explain 90% of the genetic variance (SNP heritability from GWAS summary statistics)[8]. With the current rapid expansion of sample sizes, we may expect to reach levels of 50 million in the next decade. Nevertheless, even when this is achieved it will be far from straightforward to separate true causal variants and genes from those that are statistically associated due to linkage disequilibrium (LD) with the true causal ones ('LD byproducts'). Efficient separation requires in silico post-GWAS analyses followed by wet-lab functional experimentation to advance our understanding of how the combined effects of truly causal variants disrupt biological systems and ultimately lead to insomnia.

In silico strategies to prioritize causal variants may focus on improving LD resolution by comparing results from cohorts with different LD patterns (that is, due to ancestry)[9], while wet-lab strategies could involve large-scale screening of candidate variants using CRISPR–Cas9 technology within a locus[10]. These strategies, however, are not always feasible due to a lack of data accessibility or informative readouts for the functional experiments. We propose an alternative in silico approach in which we combine statistical fine-mapping with cross-locus linking of genes for which external data are available, aimed at more specific hypotheses and targets for wet-lab experiments. This approach is especially suited for GWAS's that identify hundreds of loci.

Here we performed a meta-analysis of insomnia GWAS in the UK Biobank and 23andMe, Inc. cohorts, including 593,724 cases and 1,771,286 controls. We found 554 loci, implicating 3,898 genes using standard functional annotation and gene-based methods.

[1]Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University Amsterdam, Amsterdam, the Netherlands. [2]Department of Human Genetics, Section Clinical Genetics, Amsterdam University Medical Centers, Amsterdam, the Netherlands. [3]23andMe, Inc., Sunnyvale, CA, USA. [4]Department of Psychiatry, Yale University School of Medicine, West Haven, CT, USA. [5]Department of Psychiatry, Veterans Affairs Connecticut Healthcare System, West Haven, CT, USA. [6]Department of Psychiatry, University of California San Diego, La Jolla, CA, USA. [7]Psychiatry Service, Veterans Affairs San Diego Healthcare System, San Diego, CA, USA. [8]Departments of Integrative Neurophysiology and Psychiatry InGeest, Amsterdam Neuroscience, VU University and Medical Center, Amsterdam, the Netherlands. [9]Department of Sleep and Cognition, Netherlands Institute for Neuroscience, an institute of the Royal Netherlands Academy of Arts and Sciences, Amsterdam, the Netherlands. [10]Department of Molecular and Cellular Neurobiology, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University Amsterdam, Amsterdam, the Netherlands. [11]Department of Child and Adolescent Psychiatry and Pediatric Psychology, Section Complex Trait Genetics, Amsterdam Neuroscience, Vrije Universiteit Medical Center, Amsterdam University Medical Centers, Amsterdam, the Netherlands. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: d.posthuma@vu.nl

Using per-locus fine-mapping and cross-locus linking of genes, we prioritized 289 genes from 239 loci and show the association with brain-specific gene expression and enrichment in gene sets related to synaptic signaling and neuronal development.

Results

We conducted a GWAS meta-analysis of insomnia in 593,724 cases and 1,771,286 controls, including data from the UK Biobank study (UKB) and 23andMe, Inc. (23andMe). In UKB, insomnia was assessed using a single question that was dichotomized following previous studies[5,11] (Supplementary Note) and previously validated as being a reliable proxy of insomnia disorder[11]. We performed a GWAS on 386,988 unrelated European UKB subjects (109,548 cases and 277,440 controls). In 23andMe, insomnia was defined based on multiple questions completed by online surveys (Supplementary Note), and the GWAS in this cohort was performed on 1,978,022 unrelated participants of European ancestry (484,176 cases, 1,493,846 controls; Methods and Supplementary Fig. 1). Details regarding sample size, prevalence and SNP heritability for each cohort (and sex-specific samples) are provided in Supplementary Table 1; Supplementary Note and Supplementary Table 2 show cohort-specific GWAS results.

The combined meta-analysis included in total 593,724 cases and 1,771,286 controls and was performed using a fixed-effect model in METAL[12] (Methods). The genetic correlation between the cohort-specific GWAS was 0.66 (s.e.m. = 0.0179, $P = 2.7 \times 10^{-292}$), in line with previous work[5] (Supplementary Note).

For the meta-analysis, LD score regression (LDSC)[13] estimated h^2_{SNP} at 7.2% (s.e.m. = 0.17%) (Supplementary Note), $\lambda_{1,000}$ at 1.00, an intercept of 1.15 (s.e.m. = 0.0158) and a ratio of 0.079 (s.e.m. = 0.08). The latter indicates that, at most, 92.1% of the observed inflation is due to the high polygenicity of insomnia. We estimated that current genome-wide significant (GWS) SNPs explain 17.3% of total h^2_{SNP} (Supplementary Note). Insomnia is currently estimated to be the third most polygenic trait, following major depressive disorder (MDD) and educational attainment[8] (Supplementary Note).

We found that a PRS calculated for 10,000 individuals from the UKB samples (Methods) explained 2.46% of the phenotypic variation at most (Supplementary Note, Extended Data Fig. 1 and Supplementary Table 3). We observed a marked decrease in predictive power using a non-UKB independent cohort as the target sample (the Million Veteran Program (MVP) cohort, $n = 183,944$), in which PRS explained 0.66% of the phenotypic variation at most (Supplementary Note and Supplementary Table 3). This discrepancy between the UKB holdout and MVP samples may be due to dissimilarity in the phenotype as measured in the MVP cohort compared with the combined UKB and 23andme cohorts, or to ancestral mismatches.

In addition, we performed sex-specific meta-analyses to evaluate whether results may differ between males and females. Using LDSC[13], we estimated the genetic correlation (r_g) between GWAS results for males and females to be 0.92 ($P < 1 \times 10^{-323}$), 0.85 ($P = 3.2 \times 10^{-64}$) and 0.91 ($P < 1 \times 10^{-323}$) in the meta-analysis and UKB and 23andMe cohort-specific GWAS, respectively, consistent with a previous report on partly overlapping data[5]. Sex-specific results are available in the Supplementary Information; here we will focus on results obtained from the sex-combined meta-analysis.

SNP and gene-based findings from the meta-analysis. The meta-analysis yielded 51,876 genome-wide significant SNPs residing in 554 distinct loci containing 791 independent lead SNPs ($r^2 < 0.1$; Methods, Supplementary Note, Supplementary Data, Supplementary Fig. 2 and Supplementary Tables 4 and 5). Of the top 554 SNPs (that is, SNPs with the lowest P value in each locus), 97.1% showed concordance in effect direction between the two cohorts. Out of 554 loci, 11 were genome-wide significant in the UKB and 419 in the 23andMe cohort-specific GWAS, while nine loci were identified in both cohorts (Supplementary Table 5). The total summed length of risk loci was 145.2 Mb, which represents 4.9% of the genomic regions containing known SNPs in the entire genome. Of these 554 loci, 190 overlapped with previously identified risk loci[5–7,11] and 364 were novel (Methods, Fig. 1a and Supplementary Table 5). Of the loci reported in Hammerschlag et al.[11], Lane et al.[6,7] and Jansen et al.[5], 1/2, 3/5, 25/57 and 18/202 loci, respectively, were no longer significant in the meta-analyses (Supplementary Note, Supplementary Table 5 and Supplementary Fig. 3; note that the meta-analysis in this study includes samples from Hammerschlag et al.[11] and Jansen et al.[5]). We show that the number of risk loci increases almost linearly as a function of sample size (Fig. 1b), and that both newly identified and unreplicated risk loci from the previous GWAS showed significantly higher P values compared with other risk loci, as expected (Supplementary Note and Supplementary Fig. 4).

The 51,876 genome-wide significant SNPs showed enrichment in intronic, intergenic and 3' untranslated regions, while they were depleted in exonic regions compared with all analyzed SNPs (Fig. 1c and Supplementary Table 6). Stratified heritability analyses[14] with 28 annotations (Methods) showed that SNP heritability was most strongly enriched in conserved regions, followed by multiple chromatin modification markers, consistent with a previous meta-analysis[5] (Fig. 1d and Supplementary Table 7).

Next, we performed a gene-based association test using MAGMA. Of 19,751 protein-coding genes analyzed, 1,429 reached genome-wide significance ($0.05/19,751 = 2.53 \times 10^{-6}$; Fig. 1a and Supplementary Table 8). The most significant gene was *PTPRD* ($P = 7.2 \times 10^{-37}$), which has been associated with insomnia[5], restless leg syndrome (RLS)[15], type 2 diabetes[16] and coronary artery disease[17]. Results show that the association of *PTPRD* is unlikely to be driven by a misclassification or comorbidity of RLS within insomnia cases (Supplementary Note and Supplementary Table 9). The second-most significant gene was *LSAMP* ($P = 2.8 \times 10^{-36}$), which was not significant in the previous insomnia GWAS but has been associated with MDD[18] and suicidal behavior[19], which are highly genetically correlated to insomnia. The most significantly associated genes from the previous insomnia GWAS, *MEIS1* (ref. [11]) and *BTBD9* (ref. [5]), were also supported in the current study ($P = 1.2 \times 10^{-14}$ and 4.8×10^{-24}, respectively). Risk loci and gene analyses for sex-specific meta-analyses are summarized in Supplementary Note, Supplementary Tables 8 and 10–13 and Supplementary Fig. 5.

The 51,876 GWS SNPs were mapped to 3,526 genes (of which 1,455 are located within the risk loci) using positional, expression quantitative trait loci (eQTL) and chromatin interaction mapping strategies[20] (Methods and Supplementary Table 14). Together with genes significantly associated in gene-based tests, 3,898 unique genes were implicated.

We observed significant genetic correlations with 350 traits out of 551 tested, including multiple cardiovascular, metabolic and psychiatric traits, in agreement with previous reports[5] (Methods, Supplementary Note, Supplementary Table 15 and Extended Data Fig. 2). Of 554 insomnia risk loci, 282 were colocalized with one of the 350 traits, indicating shared causal variants among traits (Supplementary Note). In addition, a clustering of these 282 loci based on the colocalization pattern across 350 traits suggested the presence of locus heterogeneity where we observed distinct clusters of loci: one was mainly colocalized with metabolic traits and the other mainly with psychiatric traits (Supplementary Note, Supplementary Tables 16–21, Supplementary Fig. 6 and Extended Data Fig. 3).

We performed extensive post-GWAS analyses to test for convergence of genetic association signals in tissue types, brain regions, cell types and biological pathways associated with insomnia, based on the total genome-wide distribution of genetic associations and

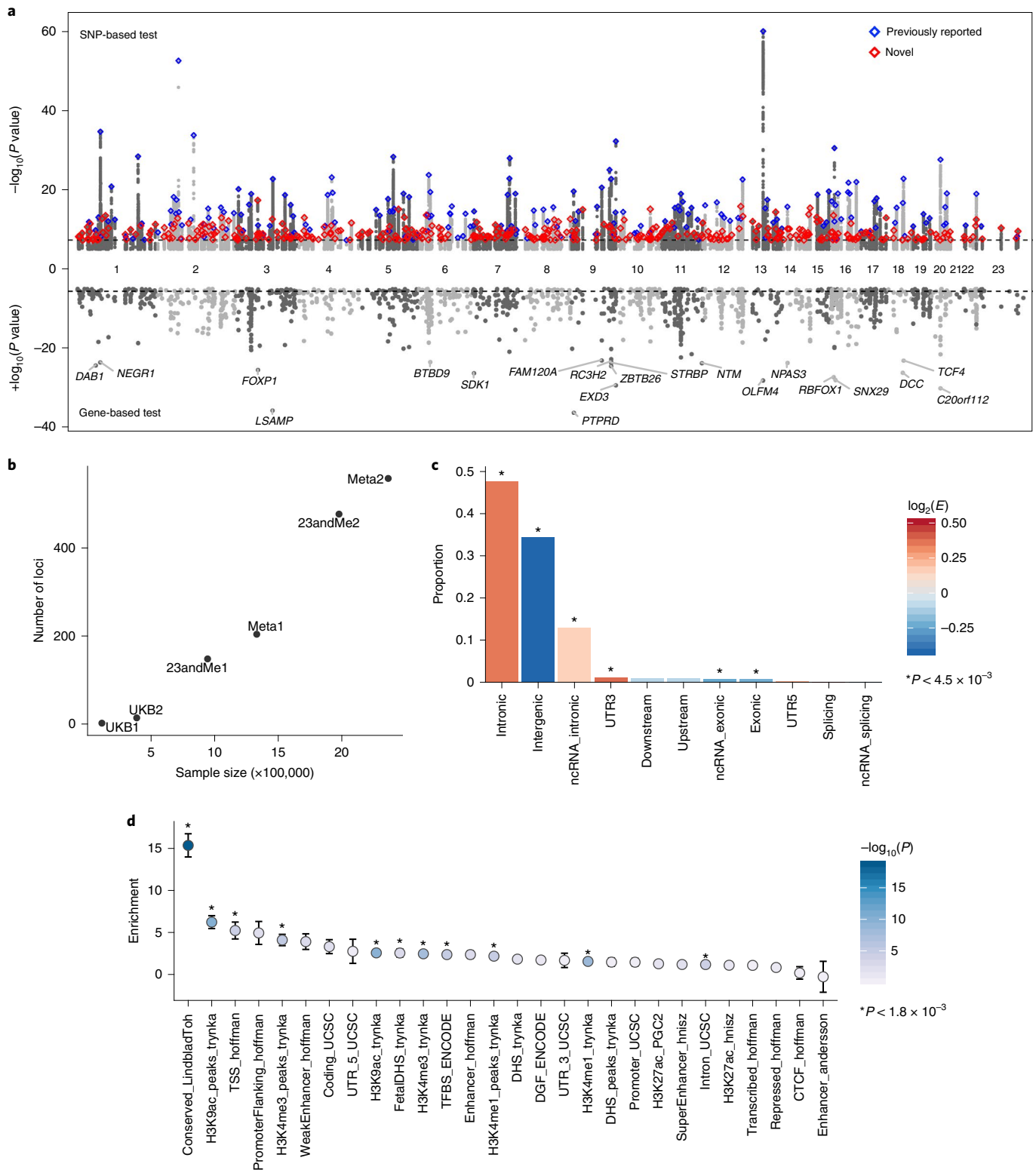


Fig. 1 | GWAS meta-analysis of insomnia in 2,365,010 individuals. a, Manhattan plot for SNPs (top) and genes (bottom) based on MAGMA gene analysis. The top SNPs of previously identified loci are labeled in blue, novel loci in red. **b**, Number of risk loci identified by insomnia GWAS with different sample sizes. **c**, Proportion of GWS SNPs in each functional category. Bars are colored by $\log_2$-transformed enrichment value (E; the proportion of GWS SNPs in a category divided by the proportion of all analyzed SNPs in the same category). Asterisks denote significant enrichment or depletion in comparison with all analyzed SNPs based on Fisher's exact test (two-sided). **d**, Enrichment of SNP heritability in 28 annotation categories computed by LD score regression. Error bars indicate 95% confidence intervals based on block jackknife over SNPs with 200 equally sized blocks of adjacent SNPs[13]; center data points indicate the point estimate. P values based on two-sided z-test. Data points are colored by $-\log_{10}(P\text{ value})$. Asterisks denote significantly enriched annotations.

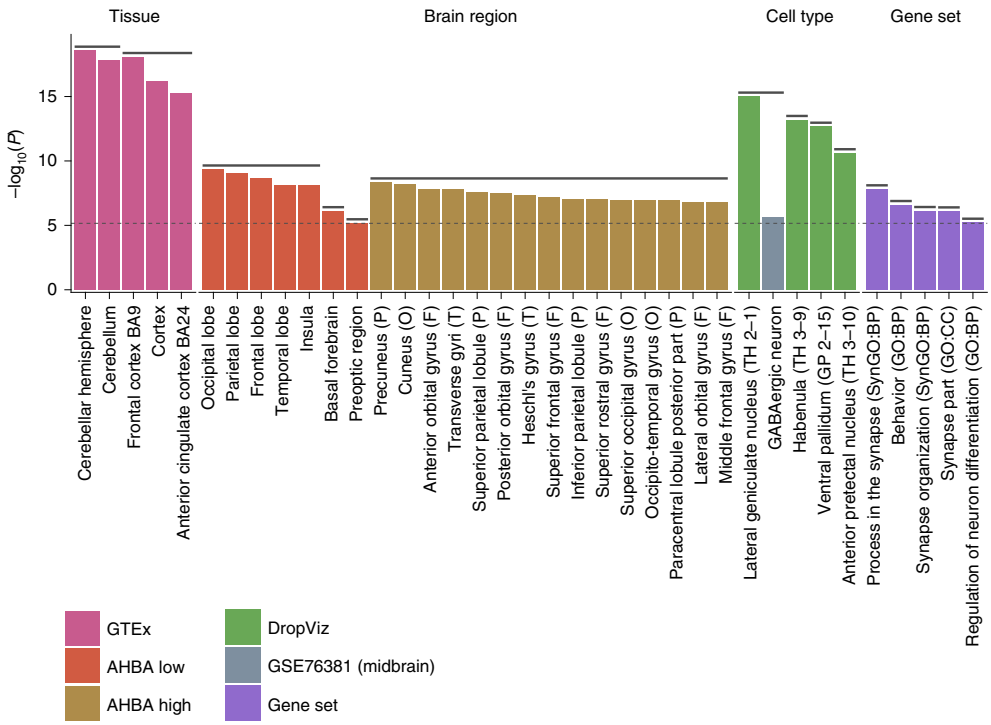


Fig. 2 | Tissues, brain regions, cell types and gene sets associated with insomnia based on MAGMA analyses. *P* values of significantly associated tissues, brain regions, cell types and gene sets based on one-sided *t*-test for the regression coefficient of gene expression. Confounders after pairwise conditional analyses per dataset are not displayed (full results are available in Supplementary Tables 22–33). Horizontal solid lines above bars represent clusters of items that are either collinear or jointly associated with insomnia. Bars are colored by datasets. Letters in parentheses for brain regions from the AHBA high dataset refer to parent brain regions from the AHBA low dataset (F, frontal lobe; O, occipital lobe; P, parietal lobe; T, temporal lobe). Labels in parentheses for cell types from the DropViz dataset refer to the cluster label from the original study. BP, biological process; CC, cellular components. The dashed line indicates the Bonferroni-corrected threshold for statistical significance ($P = 0.05/5974$).

weighted by statistical significance using MAGMA gene set analyses[21] (Methods and Supplementary Note). Based on the full GWAS distribution, we found evidence for enrichment in brain tissue (specifically the cerebral cortex), neurons in four specific brain areas (lateral geniculate nucleus (LGN), habenula, ventral pallidum and anterior pretectal nucleus), GABAergic neurons and biological pathways involved in synaptic functions, behavior and neuron differentiation (Fig. 2, Extended Data Fig. 4, Supplementary Note and Supplementary Tables 22–33).

Prioritization of high-confidence genes using multilocus information. Because of LD, many noncausal SNPs will show statistical association with a trait simply because they are correlated with causal SNPs. We call these noncausal significant SNPs LD byproducts. Because we do not know which SNPs are causal and which are LD byproducts, post-GWAS annotation provides information on many SNPs that are probably not causally related to the trait, and therefore conventional post-GWAS analysis testing for convergence may still contain considerable noise from these LD byproducts. Multiple studies have been proposed to conduct fine-mapping per locus and prioritize credible SNPs and genes from each locus before testing for convergence[22–25]. Here we propose to prioritize genes additionally based on cross-locus connections.

We assume the following: (1) credible SNPs can be indicated using in silico fine-mapping strategies; (2) SNPs that have a structural or regulatory effect on a gene product are more likely to be causal than those that have no such effect; (3) genes that are implicated as the only gene in a locus are likely to be the gene responsible for the statistical association, and are therefore probably a true causal gene; and (4) if insomnia is influenced by hundreds of genes, at least some of those are functionally related (Fig. 3).

We first defined credible SNPs by performing statistical fine-mapping for each of the 554 insomnia risk loci using FINEMAP[23] (Methods). For each locus, the 95% credible sets were extracted resulting in a total of 26,016 unique SNPs (Supplementary Note and Supplementary Fig. 7 show detailed results). More than 94.5% of the SNPs in 95% credible sets had a posterior inclusion probability (PIP) ≤ 0.1, suggesting that those SNPs were 'unsolved' by FINEMAP, which could be due to many variants being highly correlated with each other within the loci, and to small effect sizes. We retained only 1,423 credible SNPs with PIP > 0.1 distributed over 429 loci (Supplementary Table 34). Credible SNPs were then mapped to genes if they were either deleterious coding SNPs (nonsynonymous, stop-gain, stop-loss or splicing SNPs) (Supplementary Table 35) or colocalized with eQTLs from GTEx v.8 (ref. [26]), PsychENCODE[27] and eQTLGen[28] (Methods and Supplementary Table 36). This resulted in labeling 314 genes from 178 loci as having a credible SNP that was either nonsynonymous coding or colocalized with eQTLs (Supplementary Table 37).

For 376 of the remaining insomnia loci, no genes were present with the above criteria. This was due to a lack of either credible SNPs or credible SNPs that were deleterious coding or colocalized eQTLs. For these loci, we used GWS SNPs rather than credible SNPs and again evaluated whether they were deleterious coding SNPs or colocalized with eQTLs. This resulted in an additional 257 genes from 103 loci (Supplementary Table 37). Together, these 571 genes from 281 loci were termed high-confidence (HC) genes (Supplementary Note and Supplementary Fig. 8). Of these 571 HC genes, 216 (37.8%) were those closest to one of the index SNPs.

A single locus could contain multiple HC genes, but some were mapped to a single gene. In the case of a single HC gene, that gene was considered the only probable causal candidate from the

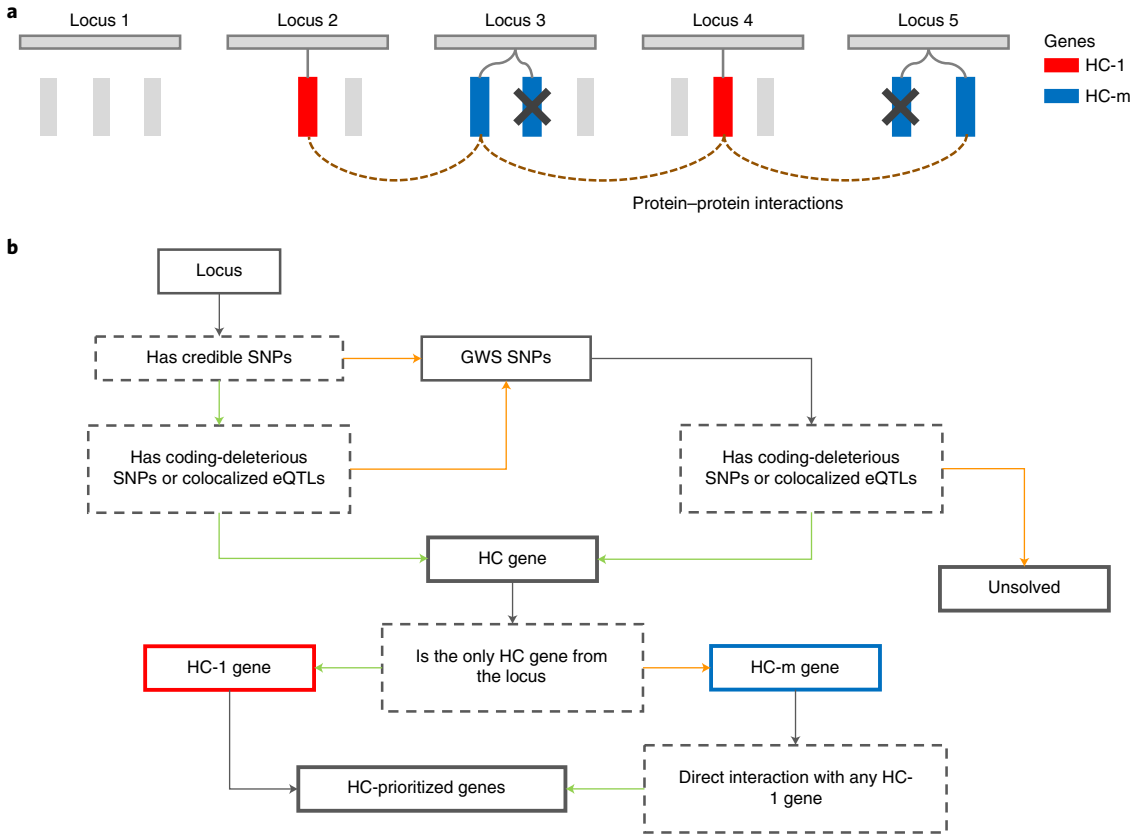


Fig. 3 | Schematic overview of gene prioritization strategy from risk loci. a, Conceptual illustration of gene prioritization based on multiloci information. Rectangles represent genes colored by prioritization status. Genes colored gray are located within or close to insomnia risk loci but were not implicated either due to a lack of functional evidence or because other genes in the same locus had higher priority. Genes that do not have a direct protein–protein interaction to HC-1 genes are filtered out (indicated by a cross). **b**, Flowchart of gene prioritization using credible SNPs and GWS SNPs. Solid boxes represent inputs or outputs, dashed gray boxes represent conditions, green arrows represent paths where an originated condition is positive, orange arrows represent paths where an originated condition is negative, gray arrows represent paths that do not have any other option.

locus. We labeled these single HC genes as high-confidence single (HC-1) genes (highest confidence) and the remaining genes as high-confidence genes with multiple genes (each of which could be the causal gene(s)) in the same locus (HC-m genes) (Fig. 3). We obtained 164 HC-1 genes from 166 loci (two genes were mapped from two loci due to distal eQTLs) and 407 HC-m genes from 116 loci (Supplementary Table 37).

Assuming that HC-1 genes are the most probable causal genes, and that functional relationships exist among the set of causal genes, we then used the set of HC-1 genes to select the most probable causal genes from the set of HC-m genes. To assess functional relations we used protein–protein interactions (PPI) using InWeb[29] and identified HC-m genes that have a direct interaction with HC-1 genes (Methods). We chose to use PPI for this and not, for example, for coexpression, because we aimed to identify genes whose products are known to form a protein complex, increasing the likelihood that they have a common function. Of 407 HC-m genes, 125 from 74 loci were found to have a direct PPI with HC-1 genes. We then defined 289 genes (164 HC-1 and 125 HC-m genes connected to HC-1) from 239 loci as the 'high-confidence prioritized' (HCP) genes (Supplementary Table 37). HCP genes are those we believe to be most probably functionally associated with insomnia based on the associated loci. Out of 239 loci, 202 were linked to single HCP genes and the maximum number of genes from a single locus was five. We consider these 239 loci to be 'resolved' and the remaining 315 to be unsolved, due to lack of biological evidence or information at present. Although we still cannot identify genes from 56.7% of the insomnia loci, we assume that these loci are randomly distributed and thus we expect HCP genes to explain some of the underlying biological mechanisms of insomnia, because these are more refined and likely to contain fewer false positives compared with the 3,526 genes mapped by all GWS SNPs.

To validate whether the prioritization we propose here can identify known causal genes, we used GWAS of three molecular traits (urate, IGF-1 and testosterone) in which causal biological mechanisms are well known[30]. The results showed that HCP genes were generally enriched in the core genes reported by Sinnott-Armstrong et al.[30], supporting the effectiveness of our gene prioritization method (Supplementary Note and Supplementary Tables 38-43).

The HCP genes we identify here for insomnia include *NEGR1*, which has been reported in multiple traits such as body mass index[31], MDD[32] and cognitive traits[33,34] as well as in a previous insomnia GWAS[5], and its effect on neuronal growth and behavior has been suggested[35,36]. HCP genes also include *EP300*, known to be involved in the circadian rhythm[37] that regulates sleep timing (see Supplementary Table 36 for a list of HCP genes).

To look for convergence in biological functions of HCP genes we assessed associations with tissues, brain regions, cell types and biological pathways (Methods and Supplementary Note).

The HCP genes showed significant joint associations with frontal cortex BA9 and anterior cingulate cortex BA24 (Fig. 3 and Supplementary Tables 44 and 45). These tissue enrichments seen with HCP genes were also detected when using the full GWAS results (Supplementary Tables 46–48 and Supplementary Note).

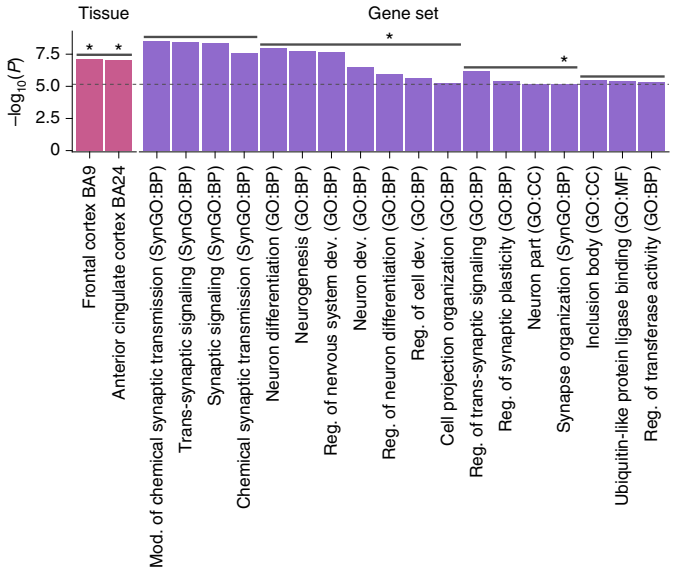


Fig. 4 | Tissues and gene sets associated with the HCP genes. *P* values of significantly associated tissues, brain regions, cell types and gene sets based on one-sided *t*-test for the regression coefficient of gene expression. Confounders after pairwise conditional analyses per dataset are not displayed. Horizontal solid lines above bars represent clusters of items that are either collinear or jointly associated with insomnia. Asterisks represent items that also showed significant association with insomnia when using the full GWAS results. Bars are colored by datasets. MF, molecular function; Mod., modulation; Reg., regulation.

On the other hand, some tissue enrichments that were significant when using the full GWAS were no longer detected when using the HCP gene list. This is because genes driving these associations in the full GWAS were not prioritized, which suggests that these initial enrichments were less likely to be based on true causal genes. (Supplementary Note and Extended Data Fig. 5).

From the gene set enrichment analyses, HCP genes showed significant enrichment in 18 gene sets clustered into four independent groups (Fig. 4, Extended Data Fig. 6 and Supplementary Table 49). The most significantly enriched gene sets from each cluster were 'Modulation of chemical synaptic transmission' (SynGO:BP), 'neuron differentiation' (GO:BP), 'regulation of trans-synaptic signaling' (GO:BP) and 'inclusion body' (GO:CC). We identified additional, less redundant gene sets when using the prioritized genes than when using the full GWAS results. This may indicate that, by filtering potential noise out of LD byproducts, we were able to identify gene sets specific to the most probable causal genes prioritized from insomnia-associated loci that were missed using the full GWAS results.

Discussion

In this study we performed a GWAS meta-analysis for insomnia that included >2.3 million subjects. We identified 554 insomnia risk loci, doubling the number identified by the largest previous study, which included 1.3 million subjects[5].

We observed clusters of insomnia loci based on colocalization patterns across multiple traits, indicating potential locus heterogeneity. In particular, separation of locus clusters that are colocalized with either metabolic or psychiatric traits is clinically relevant. This suggests that insomnia is a genetically heterogeneous phenotype consisting of different genetic subtypes—for example, insomnia symptoms that are related more to either metabolic disturbances or other factors in the brain that may require different treatment approaches. Indeed, metabolic disturbances have been found to contribute to hyperarousal in insomniacs compared with controls, such as increased whole-body and brain metabolism, altered hormone secretion and sympathetic activation[38].

Using multiple cell-specific gene expression datasets, we identified novel associations of insomnia with neuronal cells including habenular, LGN and GABAergic neurons, among others. These findings are supported by previous evidence of involvement in sleep regulation, but have not been linked by GWAS until now. The habenular nuclei have reciprocal connections with the pineal gland along which it coevolved (together forming the epithalamus)[39,40], and its activity follows a strong circadian pattern[40,41]. Among its hypothesized functions are sleep and circadian rhythm regulation through production of melatonin[39,40] and maintaining rapid-eye-movement (REM) sleep, as evidenced by REM disturbances induced by habenular lesions[42,43]. The LGN is part of the visual system, which relays retinal information to cortical brain areas. In addition, the LGN is involved in circadian rhythm regulation through its intrinsic timekeeping properties[44] and indirect interactions with the suprachiasmatic nucleus (SCN) through neuropeptide-Y[45,46]. Lesions in the LGN indeed have shown to affect circadian activity in animal models via disturbed processing of environmental cues[46]. GABA is among the most abundant neurotransmitters in the brain and is the main neurotransmitter of the circadian system[47] whose inhibitory action induces a sleep state[48–50], and the SCN consists almost entirely of GABAergic neurons[51]. Interestingly, the GABAergic system is the mechanism of action of drugs such as benzodiazepines that are often used to treat insomnia[49]. These observations point to several different but related mechanisms in the brain that may provide a basis for further study by experimental designs.

We also observed that, in spite of almost doubling our sample size, SNP-based heritability did not notably increase (7–8%). Because SNP-based heritability sets an upper limit to the prediction power, the increased accuracy of effects sizes also did not lead to an improved prediction. These results support the extreme polygenicity of insomnia as it is operationalized in the current (and previous) GWAS. Our current results, however, do hint at heterogeneous forms of insomnia, one that is due to a metabolic–genetic pathway and a second due to a psychiatric–genetic pathway. Future studies aimed at increasing prediction may benefit from the collection of deep phenotyping data on insomnia patients and identify subtypes of insomnia.

We demonstrated a novel strategy using known biological functions of SNPs and multilocus functional relations of genes to prioritize the most probable causal genes, and based post-GWAS analyses for convergence on these genes. Applying this strategy, we identified 289 HCP genes from 239 loci and compared associated tissue and cell types, as well as gene sets based both on the set of prioritized genes and all genes implicated in the GWAS. We found that the former is less likely to contain LD byproducts and provided more specific results. Indeed, the gene set showed that the most significant enrichment in HCP genes, Modulation of chemical synaptic transmission (SynGO:BP), is at the lowest hierarchy of the gene ontology tree in the SynGO dataset (Modulation of chemical synaptic transmission < Chemical synaptic transmission < Trans-synaptic signaling < Synaptic signaling < Process in the synapse), while we only identified the broadest ontology, Process in the synapse, by using the full GWAS with MAGMA.

We identified enrichment of HCP genes in gene sets related to synaptic and neuronal processes, including neurogenesis and differentiation, which were not previously observed. Evidence of synaptic transmission of neurotransmitters in insomnia has previously been found in imaging studies demonstrating imbalance of neurotransmitters in the brain of insomniacs[52,53], including altered levels of GABA and glutamate[54]. In addition, the observed neuronal processes could point towards developmental mechanisms that predispose the brain to insomnia. Alternatively, neurogenesis

and neuron differentiation were recently reported to occur in the hypothalamus[51], a major regulator of circadian rhythm where new neurons support and maintain its normal functioning. It is hypothesized that (age-related) decline in neurogenesis may contribute to impaired sleep–wake regulation in humans (a review is provided by Kostin et al.[51]). The ultimate test of whether HCP genes are actually causally involved still lies in functional follow-up experiments.

There are several limitations of the gene prioritization strategy proposed in this study. First, the strategy using multilocus information is feasible only for polygenic traits with a reasonable number of independent risk loci identified by GWAS and a reasonable number of loci with single implicated genes. Second, the prioritization procedure depends on the availability and accuracy of functional annotations of SNPs and genes. For example, we defined an HC-1 gene as the only gene from a single locus with high-confidence biological evidence. However, in future more SNPs may be found to be deleterious (by increasing accessible (rare) SNPs in GWAS) or colocalized eQTLs (by increasing statistical power to detect eQTLs and their availability in specific cell types), which may change the current results and allow us to identify additional high-confidence genes from loci. Third, cross-locus linking of genes depends on the availability and reliability of biological information (PPI, coexpression networks or any other gene-correlation matrix deemed relevant), which is currently not abundantly available. We do believe that the use of cross-locus information greatly aids in making sense of the multitude of associated genes, and the current study shows that this strategy indicates a role for more specific biological functions in insomnia.

In conclusion we show that, for extremely polygenic traits such as insomnia, increasing sample size does lead to an increase in detected SNPs, loci, genes and pathways, providing more confidence in existing and novel mechanisms. We also show that increasing sample size in this case does not lead to increasing predictive power, and provide some suggestions of why this might be the case (for example, genetic subtypes, extreme polygenicity and phenotype operationalization). In addition, we provide a novel gene prioritization method that relies on the large number of detected loci, using a small percentage of those loci with clearly identifiable probable causal genes to prioritize genes from the remaining loci, which aids in generating hypotheses about biological processes underlying insomnia that can be tested in functional experiments.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01124-w.

Received: 2 January 2021; Accepted: 6 June 2022;
Published online: 14 July 2022

References

1. Roth, T. Insomnia: definition, prevalence, etiology, and consquences. *J. Clin. Sleep Med.* **3**, S7–S10 (2007).
2. Kripke, D. F., Garfinkel, L., Wingard, D. L., Klauber, M. R. & Marler, M. R. Mortality associated with sleep duration and insomnia. *Arch. Gen. Psychiatry* **59**, 131–136 (2002).
3. Daley, M., Morin, C. M., Leblanc, M., Grégoire, J. & Savard, J. The economic burden of insomnia: direct and indirect costs for individuals with insomnia. *Sleep* **32**, 55–64 (2009).
4. 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).
5. 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).
6. 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).
7. Lane, J. M. et al. Biological and clinical insights from genetics of insomnia symptoms. *Nat. Genet.* **51**, 387–393 (2019).
8. Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
9. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E. & Halperin, E. Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.* **86**, 23–33 (2010).
10. Findlay, G. M. et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
11. 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).
12. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
13. Bulik-sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
14. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
15. 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).
16. Tsai, F. J. et al. A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLoS Genet.* **6**, e1000847 (2010).
17. Schunkert, H. et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–340 (2011).
18. Koido, K. et al. Associations between LSAMP gene polymorphisms and major depressive disorder and panic disorder. *Transl. Psychiatry* **2**, e152 (2012).
19. Must, A. et al. Association of limbic system-associated membrane protein (LSAMP) to male completed suicide. *BMC Med. Genet.* **9**, 34 (2008).
20. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
21. 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).
22. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, 2–3 (2014).
23. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
24. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
25. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
26. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
27. Wang, D. et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
28. Võsa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
29. Li, T. et al. A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14**, 61–64 (2016).
30. Sinnott-Armstrong, N., Naqvi, S., Rivas, M. & Pritchard, J. K. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife* **10**, e58615 (2021).
31. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
32. 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).
33. Savage, J. E. et al. GWAS meta-analysis (N=279,930) identifies new genes and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
34. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
35. Singh, K. et al. Neuronal growth and behavioral alterations in mice deficient for the psychiatric disease-associated *negr1* gene. *Front. Mol. Neurosci.* **11**, 30 (2018).
36. Singh, K. et al. Neural cell adhesion molecule Negr1 deficiency in mouse results in structural brain endophenotypes and behavioral deviations related to psychiatric disorders. *Sci. Rep.* **9**, 5457 (2019).
37. Koike, N. et al. Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science* **338**, 349–354 (2012).

38. Bonnet, M. H. & Arand, D. L. Hyperarousal and insomnia: state of the science. *Sleep Med. Rev.* **14**, 9–15 (2010).
39. Hikosaka, O. The habenula: from stress evasion to value-based decision-making. *Nat. Rev. Neurosci.* **11**, 503–513 (2010).
40. Benarroch, E. E. Habenula: recently recognized functions and potential clinical relevance. *Neurology* **58**, 992–1000 (2015).
41. Zhao, H. & Rusak, B. Circadian firing-rate rhythms and light responses of rat habenular nucleus neurons in vivo and in vitro. *Neuroscience* **132**, 519–528 (2005).
42. Haun, F., Eckenrode, T. C. & Murray, M. Habenula and thalamus cell transplants restore normal sleep behaviors disrupted by denervation of the interpeduncular nucleus. *J. Neurosci.* **12**, 3282–3290 (1992).
43. Bianco, I. H. & Wilson, S. W. The habenular nuclei: a conserved asymmetric relay station in the vertebrate brain. *Philos. Trans. R. Soc. B Biol. Sci.* **364**, 1005–1020 (2009).
44. Chrobok, L. et al. Intrinsic circadian timekeeping properties of the thalamic lateral geniculate nucleus. *J. Neurosci. Res.* **99**, 3306–3324 (2021).
45. Harrington, M. E. The ventral lateral geniculate nucleus and the intergeniculate leaflet: interrelated structures in the visual and circadian systems. *Neurosci. Biobehav. Rev.* **21**, 705–727 (1997).
46. Johnson, R. F., Moore, R. Y. & Morin, L. P. Lateral geniculate lesions alter circadian activity rhythms in the hamster. *Brain Res. Bull.* **22**, 411–422 (1989).
47. Moore, R. Y. & Speh, J. C. GABA is the principal neurotransmitter of the circadian system. *Neurosci. Lett.* **150**, 112–116 (1993).
48. Melzer, S. & Monyer, H. Diversity and function of corticopetal and corticofugal GABAergic projection neurons. *Nat. Rev. Neurosci.* **21**, 499–515 (2020).
49. España, R. A. & Scammell, T. E. Sleep neurobiology from a clinical perspective. *Sleep* **34**, 845–858 (2011).
50. Gottesmann, C. GABA mechanisms and sleep. *Neuroscience* **111**, 231–239 (2002).
51. Kostin, A., Alam, M. A., McGinty, D. & Alam, M. N. Adult hypothalamic neurogenesis and sleep-wake dysfunction in aging. *Sleep* **44**, zsaa173 (2021).
52. Levenson, J. C., Kay, D. B. & Buysse, D. J. The pathophysiology of insomnia. *Chest* **147**, 1179–1192 (2015).
53. Spiegelhalder, K. et al. Neuroimaging insights into insomnia. *Curr. Neurol. Neurosci. Rep.* **15**, 9 (2015).
54. Kay, D. B. & Buysse, D. J. Hyperarousal and beyond: new insights to the pathophysiology of insomnia disorder through functional neuroimaging studies. *Brain Sci.* **7**, brainsci7030023 (2017).

23andMe Research Team

Michelle Agee[3], Stella Aslibekyan[3], Adam Auton[3], Robert K. Bell[3], Katarzyna Bryc[3], Sarah K. Clark[3], Sarah L. Elson[3], Kipper Fletez-Brant[3], Pierre Fontanillas[3], Nicholas A. Furlotte[3], Pooja M. Gandhi[3], Karl Heilbron[3], Barry Hicks[3], Karen E. Huber[3], Ethan M. Jewett[3], Yunxuan Jiang[3], Aaron Kleinman[3], Keng-Han Lin[3], Nadia K. Litterman[3], Jennifer C. McCreight[3], Matthew H. McIntyre[3], Kimberly F. McManus[3], Joanna L. Mountain[3], Sahar V. Mozaffari[3], Elizabeth S. Noblin[3], Carrie A. M. Northover[3], Jared O'Connell[3], Steven J. Pitts[3], G. David Poznik[3], J. Fah Sathirapongsasuti[3], Janie F. Shelton[3], Jing Shi[3], Suyash Shringarpure[3], Chao Tian[3], Joyce Y. Tung[3], Robert J. Tunney[3], Vladimir Vacic[3] and Wei Wang[3]

Methods

Genome-wide association analysis. *UKB.* We performed genome-wide association analysis for insomnia with PLINK 2.0 (ref. [55]), using logistic regression with age, sex, genotype array and ten genetic principal components (PCs) computed with unrelated European subjects defined above, based on 145,432 independent SNPs ($r^2 < 0.1$, minor allele frequency (MAF) > 0.01, imputation quality score (INFO) = 1) using FlashPCA[56] as covariates. We analyzed autosomal, pseudo-autosomal and X chromosomes. For X chromosomes we used a model where genotype was coded [0, 2] for males (with a PLINK flag –xchr-model 2) for consistency with the 23andMe cohort. SNPs were limited to those with a minor allele count (MAC) > 100. For sex-specific GWAS we followed the same criteria except that sex was excluded from the covariates. The numbers of analyzed SNPs and sample sizes are summarized in Supplementary Table 1.

23andMe. Summary statistics were obtained from 23andMe, Inc. based on logistic regression with age, sex, genotype array and the first five genetic ancestry PCs. 23andMe included the first five PCs compared with ten in UKB, because the first PCs of 23andMe explain more variance than in the UKB cohort: the variance is flat after the fifth PC in 23andMe while this plateau was reached after the tenth PC in UKB (Supplementary Fig. 1). We first extracted SNPs that passed quality control by 23andMe. When there were both genotyped and imputed genotypes available for a single SNP, the imputed SNP was retained. We then further extracted SNPs with MAC (MAF × sample size × 2) > 100. The analyses were performed for unrelated European subjects for sex-combined (n = 1,978,022), male-only (n = 1,038,003) and female-only (n = 1,200,179) separately.

GWAS meta-analysis. A meta-analysis of GWAS summary statistics in UKB and 23andMe was performed using METAL software[12], based on the fixed-effect model with SNP P values weighted by sample size. The meta-analysis was performed separately for the sex-combined and each sex-specific GWAS. We used a unique marker ID consisting of chromosome, position and alphabetically ordered alleles to match SNPs between cohorts. Because indels were coded as I/D in the 23andMe cohort, exact alleles were assigned for indels based on information from the UKB cohort. We assigned alleles to indels in 23andMe only when there were biallelic indels on the same position in the UKB cohort. The numbers of analyzed SNPs and sample sizes are summarized in Supplementary Table 1.

We converted z-statistics to standardized effect sizes (and their standard error) as a function of MAF and sample size as below[57]:

$$\beta = \frac{z}{\sqrt{2p(1-p)(n+z^2)}}, \text{s.e.m.} = \frac{1}{\sqrt{2p(1-p)(n+z^2)}}$$

where p is MAF and n is sample size for a given SNP. Log odds ratio (OR) was approximated using the fraction of cases:

$$\log \text{OR} = \frac{\beta}{u(1-u)}$$

where u is a case fraction.

SNP heritability and stratified heritability. SNP heritability was estimated for sex-combined and sex-specific GWAS on UKB, 23andMe and meta-analyses using LDSC[13]. Precomputed LD scores for 1,000 Genome Phase 1 European subjects were downloaded from https://data.broadinstitute.org/alkesgroup/LDSCORE/. The analyses were limited to HapMap3 SNPs, with the MHC region (chr6: 26–34 Mb) excluded. In addition, SNPs with chi-square statistics >80 were excluded. To compute SNP heritability on the liability scale we provided a population prevalence of 30% (ref. [1]). Because the sample size of this study is large, the genomic inflation factor λ_{GC} was scaled for 1,000 cases and 1,000 controls as $\lambda_{1,000} = 1 + (\lambda_{GC} - 1) \times (1/n_{cases} + 1/n_{controls}) \times 500$.

To test whether SNP heritability was enriched in a specific category of functional annotations, we partitioned it for 28 binary SNP annotations[14]. Enrichment was computed as the proportion of SNP heritability explained by SNPs with annotations divided by the proportion of SNPs with annotations. We obtained 28 functional annotations from https://data.broadinstitute.org/alkesgroup/LDSCORE/.

Polygenic risk scoring. Phenotypic variance explained by our meta-analysis was estimated using polygenic scores (PGS) that were computed based on SNP effect size using PRSice v.2.2.1 (ref. [58]) (http://www.prsice.info/) with default parameters. PRSice performs clumping of SNPS at $r^2 = 0.1$. We set P value thresholds of input SNPs at $P < 1.0$, 0.5, 0.05, 0.01, 5×10^{-3}, 1×10^{-3} and 1×10^{-5}. We computed PGS for three randomly selected sets of 10,000 UKB subjects, with summary statistics recalculated excluding those 10,000 subjects each time for the UKB cohort. We then meta-analyzed with the 23andMe GWAS that we used as training data. Variants with MAF < 0.01, missing rate > 0.05 or Hardy–Weinberg equilibrium $P < 1 \times 10^{-6}$ were filtered out from the target samples. We then report R^2 adjusted for ascertainment with population prevalence of 0.3. To evaluate the predictive power of insomnia GWAS meta-analysis with a different sample size, we also meta-analyzed GWAS of the UKB training dataset and the 23andMe GWAS from a previous study[5] (~1.3 million samples in total), then performed prediction of the same target samples. PRS were also carried out with the Million Veteran Program as an independent cohort (Supplementary Note).

Definition of risk loci. Genomic risk loci were defined within Functional Mapping and Annotation (FUMA; https://fuma.ctglab.nl) as previously described[20], by first clumping SNPs with $P < 5 \times 10^{-8}$ at $r^2 = 0.6$ using SNPs with $P < 1 \times 10^{-5}$ to define independent significant SNPs. Those independent significant SNPs were further clumped at $r^2 = 0.1$ to define lead SNPs. Independent significant SNPs that were in LD with the same lead SNPs ($r^2 > 0.1$) and LD blocks closer than 250 kb were merged into a single locus. Each locus is represented by the most significant (top) SNP. A risk locus can contain multiple lead and independent significant SNPs. We manually excluded suspicious loci (for example, a single SNP reaching genome-wide significance with no SNPs with $P < 1 \times 10^{-5}$ and $r^2 > 0.6$) by examining LocusZoom plots (Supplementary Note 5).

MAGMA gene, gene property and gene set analysis. Gene-based testing was performed using MAGMA v.1.07 (ref. [21]) to obtain gene P values using summary statistics of sex-combined meta-analysis. From 20,260 protein-coding genes, SNPs were assigned to one of the 19,751 genes within 2-kb upstream and 1-kb downstream windows, based on the location obtained from Ensembl v.92 GRCh37 using BioMart. We used the SNP-wise mean model and randomly selected 10,000 unrelated European subjects from the UKB cohort as a reference panel.

For tissue specificity analyses we obtained RNA sequencing (RNA-seq) data for 54 tissue types from GTEx v.8 (ref. [26]). Reads per kilobase per million were $\log_2$ transformed with pseudocount 1 followed by winsorization at 50, and average per tissue type was computed for each gene. In a gene property analysis, the average across 54 tissue types was conditioned and a one-sided test was performed to identify positive associations of tissue-specific gene expression with insomnia.

For gene expression of specific brain regions we obtained normalized microarray data for 3,702 samples from six healthy donors[59] from the Allen Human Brain Atlas (AHBA; http://human.brain-map.org/static/download). From 58,692 probes, 31,098 with missing values in <20% of samples were first extracted. When there were multiple probes per gene, genes with the highest variance across 3,702 samples were selected, resulting in 17,916 unique genes. Of those, 13,943 were mapped to a unique Ensembl gene ID (v.92 GRCh37). Each of 3,702 samples was assigned to the structural ID of the brain where there were multiple layers of hierarchical structure[59]. We assigned the annotation of brain regions at the fourth and fifth layers of the hierarchical structure where the top of the tree ('brain') was considered as layer 0. We limited consideration to brain regions with at least five samples, resulting in 54 and 106 brain regions for layers 4 and 5, respectively. For each brain region, average expression values were computed for each gene. We then performed gene property analysis conditioning on average across brain regions, and a one-sided test was performed to identify positive associations of brain region-specific gene expression with insomnia.

For cell type specificity analysis we used five datasets from Linnarsson's group from a previous insomnia meta-analysis study: mouse samples from cortex and hippocampus (GSE60361, level 2 neurons)[60], hypothalamus (GSE74672, level 2 neurons)[61], oligodendrocytes (GSE75330)[62], midbrain (GSE76381)[63] and striatum (GSE97478)[64]. We additionally tested one of the most comprehensive small cytoplasmic RNA (scRNA)-seq datasets for mouse brain, DropViz[65] (http://dropviz.org/; 565 subclusters from nine brain regions). In total, we obtained 728 cell types from six datasets. Each dataset was preprocessed as described in a previous study[66]. We performed a three-step workflow (per dataset analysis and within- and cross-dataset conditional analyses)[66] to determine significant associations supported by multiple independent datasets.

Gene sets for Gene Ontology terms and canonical pathways were obtained from MsigDB v.6.2 (ref. [67]) (http://software.broadinstitute.org/gsea/msigdb/index.jsp). We performed gene set analyses for 5,033 gene sets with at least 20 genes.

We evaluated all 5,974 tested items against the Bonferroni-corrected threshold ($0.05/5{,}974 = 8.4 \times 10^{-6}$).

Fine-mapping and credible SNPs. We performed fine-mapping of 554 risk loci using FINEMAP with the shotgun stochastic search algorithm[23] (http://www.christianbenner.com/). For each risk locus, SNPs within 50 kb of the top SNP (with the minimum P value) or locus boundary, whichever was larger, and with $P < 0.05$ were used for fine-mapping. The pairwise LD matrix of SNPs was estimated based on a randomly selected group of 100,000 unrelated European individuals from UKB using LDstore[68] (http://www.christianbenner.com/). The maximum number of causal SNPs (k) was set to 10; an exception was made for locus no. 4, which contained seven SNPs, where k was set to 5.

FINEMAP outputs a set of models (all possible combinations of k causal SNPs in a locus) with posterior probability (PP) of being a causal model. A 95% credible set was defined by taking models from the highest PP until the cumulative sum of PP reached 0.95 for each locus. The 95% credible set of SNPs were, therefore, defined by taking unique SNPs from 95% credible sets. In addition, PIP was calculated for each SNP as the sum of PPs of credible sets containing that SNP. In

this study, although we used only credible SNPs with PIP > 0.1, the results of all credible set SNPs are described in Supplementary Note and Supplementary Fig. 6.

Annotation of SNPs with FUMA. All GWS SNPs were annotated with their functions using FUMA. Functional consequences of SNPs on genes were obtained by performing ANNOVAR[69] gene-based annotation using Ensembl genes. Enrichments of GWS SNPs in each annotation were tested by Fisher's exact test (two-sided) by comparing them with the annotations of all SNPs analyzed in the meta-analysis. To determine the deleteriousness of SNPs, a CADD v.1.4 (ref. [70]) score was annotated for each SNP. In addition, a RegulomeDB score (categorical score indicating the likelihood of SNPs being involved in regulatory elements)[71] and a 15-core chromatin state for 127 tissue/cell types obtained from Roadmap[72,73] were annotated.

eQTL colocalization. We colocalized insomnia summary statistics from the sex-combined meta-analysis within 554 risk loci with eQTL summary statistics using the coloc R package[74]. We tested only those genes whose significant eQTLs were overlapping with at least one GWS SNP in the insomnia GWAS, and colocalization was performed for each gene for each of 51 eQTL datasets (that is, 49 tissues from GTEx v.8 (ref. [26]), meta-analysis of blood samples from the eQTLGen consortium[28] and prefrontal cortex from PsychENCODE[27]). For each colocalization, we extracted SNPs available in both eQTL summary statistics of a testing gene and within 10 kb of insomnia risk loci. We then used the *coloc.abf* function. We did not perform colocalization when there were fewer than ten SNPs overlapping between insomnia and eQTL summary statistics.

The *coloc.abf* function assumes a single causal SNP for each trait and estimates the PP of the following five scenarios for each testing region: H_0, neither trait has a genetic association; H_1, only trait 1 has a genetic association; H_2, only trait 2 has a genetic association; H_3, traits 1 and 2 are both associated but with different causal SNPs; and H_4, traits 1 and 2 are both associated with the same single causal SNP. In the case of our study, trait 1 is insomnia and trait 2 is expression of a tested gene. Because we limited the analyses to genes where there was at least one overlap of significant SNPs with insomnia GWS SNPs, this discards scenarios H_0–H_2 and we are thus interested only in whether H_4 is most likely. We therefore defined eQTLs of a testing gene as colocalized with insomnia summary statistics when $H_4 > 0.9$. It is possible that genomic regions outside of the predefined risk loci could also be colocalized with eQTLs. However, we limited the analyses to the risk loci in this study because the primary aim was to prioritize genes linked from the insomnia risk loci.

Gene mapping with FUMA. We used FUMA to map SNPs to genes using three criteria: positional, eQTL and chromatin interaction mapping[20].

Positional mapping. SNPs were mapped to one of 20,260 protein-coding genes with 10-kb windows on both sides.

eQTL mapping. Significant eQTLs in 49 tissue types from GTEx v.8 (ref. [26]), blood samples from the eQTLGen consortium[28] and prefrontal cortex samples from PsychENCODE[27] were used for mapping. FUMA annotates those significant eQTLs with candidate SNPs, and these SNPs are mapped to the gene whose expression is potentially affected by the SNPs.

Chromatin interaction mapping. Significant chromatin loops (false discovery rate $< 1 \times 10^{-6}$) in 14 tissues, defined based on HiC (high-throughput chromosome conformation capture) data from Schmitt et al.[75], and preprocessed chromatin loops based on HiC of prefrontal cortex from PsychENCODE[27] were used for mapping. In FUMA, candidate SNPs are required to be overlapped with one end of the loop, and transcription start sites (TSSs) of genes (500 base pairs upstream and 250 base pairs downstream of the TSS) are required to be overlapped with the other end of the loop to be mapped. Because HiC is designed to measure physical interactions of two genomic regions, not all significant loops necessarily contain functional interactions. We therefore further limited chromatin interaction mapping to those where SNPs were overlapping with enhancer regions and gene TSSs were overlapping with promoter regions predicted by the Roadmap consortium[72] (http://egg2.wustl.edu/roadmap/data/byDataType/dnase/). We used all available 113 cell types for enhancers and promoters.

We performed gene mapping for all genome-wide significant SNPs and credible SNPs separately. We further performed filtering outside of FUMA, as described above. Protein–protein interaction was obtained from InWeb InBio Map[29].

Tissue and cell type association, and gene set enrichment tests with prioritized genes. Associations of tissue and cell-type-specific gene expression with prioritized genes were tested with a linear regression model using the *lm* function in R. We defined the model as follows to correct for average expression across tissues or cell types within a dataset and gene size:

$$E \sim \beta_G G + \beta_A A + \beta_S \log(S)$$

where G is a binary status reflecting whether the gene was prioritized (1) or not (0), E is a tissue/cell-type-specific expression value, A is the average expression of the gene across all available tissues/cell types in a dataset and S is gene length. We performed a one-sided test ($\beta_G > 0$) to evaluate how well the prioritized gene status predicts the specificity of gene expression in a testing tissue/cell type. For tissue/cell types significantly associated with the prioritized genes, we performed conditional analyses as below and performed a one-sided test ($\beta_{G\_1} > 0$ and $\beta_{G\_2} > 0$):

$$E_1 \sim \beta_{G_1} G + \beta_{E_2} E_2 + \beta_A A + \beta_S \log(S)$$

$$E_2 \sim \beta_{G_2} G + \beta_{E_1} E_1 + \beta_A A + \beta_S \log(S)$$

For gene set enrichment analyses, a one-sided hypergeometric test (greater) was performed. We tested the same datasets as those in the MAGMA gene set analysis (53 tissues, 54 and 106 brain regions, 728 cell types and 5,033 gene sets). As done in MAGMA v.1.07, gene property (gene expression value) was truncated when the value was above or below 5 s.e.m. The analyses were limited to genes available in each dataset out of 20,260 protein-coding genes based on Ensembl v.92 GRCh 37.

We performed Bonferroni correction across all 5,974 tested items ($0.05/5974 = 8.4 \times 10^{-6}$).

Genetic correlation. We first selected 551 GWAS (with 551 unique traits) that showed $h^2_{SNP} > 0.01$ and z-score > 2 from 558 GWAS analyzed previously in the study of Watanabe et al.[8], excluding insomnia and trouble falling asleep (depression item). We estimated genetic correlations of insomnia sex-combined meta-analysis with 551 traits using LDSC[13]. Precomputed LD scores for 1000 Genome Phase 1 European subjects were downloaded from https://data.broadinstitute.org/alkesgroup/LDSCORE/. The analyses were limited to HapMap3 SNPs, and the MHC region was excluded. Additionally, SNPs with chi-square statistics >80 were excluded. We defined genetic correlation as significant after Bonferroni correction ($P < 0.05/551 = 9.07 \times 10^{-5}$).

Colocalization of risk loci and clustering. Colocalization of 554 insomnia risk loci was performed using the *coloc* package in R, as described in eQTL colocalization. Risk loci of length <10 kb were expanded to 10 kb by centering the top SNP. Colocalization of each of 554 insomnia risk loci was tested with GWAS summary statistics of 350 traits by selection of overlapping SNPs, and loci were considered colocalized when $H_4 > 0.9$.

We performed *t*-distributed stochastic neighbor embedding (t-SNE)[76] 100 times, with the optimal solution being obtained by minimization of Kullback–Leibler divergence. The clustering of traits was performed on a t-SNE two-dimensional (2D) matrix using DBSCAN (dbscan function from the R package fpc), and the clustering cutoff was optimized by maximization of silhouette score. One percent of data points was allowed to be 'unclustered'. Since t-SNE projects data into a certain number of dimensions based on the similarity of data points, traits that do not share any colocalized loci or share fewer than other traits can form a cluster. To distinguish this from a cluster where traits within the cluster share more colocalized loci than others, we tested for each cluster to determine whether the number of shared colocalized loci within clusters was larger than between clusters using the Mann–Whitney *U*-test (one-sided, greater). Clusters that did not show significant difference ($P \geq 0.05$) were discarded. In the same way, insomnia risk loci were projected onto the 2D map with t-SNE and dense clusters were identified based on colocalization patterns across traits.

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

Data availability

The full GWAS summary statistics for UKB and the top 10,000 SNPs for 23andMe are available at https://ctg.cncr.nl/software/summary_statistics/. The full GWAS summary statistics for the 23andMe dataset will be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of 23andMe participants. Please visit https://research.23andme.com/collaborate/#publication for more information and to apply to access the data. The following publicly available datasets were used in this manuscript: GTEx v.8 (https://gtexportal.org/home/datasets), Allen Human Brain Atlas (http://human.brain-map.org/static/download), scRNA-seq from Linnerson's laboratory (http://linnarssonlab.org/data/; GSE60361, GSE74672, GSE75330, GSE76381, GSE97478), DropViz (http://dropviz.org/), MsigDB v.6.2 (http://software.broadinstitute.org/gsea/msigdb/index.jsp), InWeb protein–protein interaction (https://inbio-discover.com/download), eQTLGen (https://www.eqtlgen.org/) and PsychEncode (http://resource.psychencode.org/).

Code availability

The R script used to perform gene prioritization approach proposed in this manuscript is available at https://doi.org/10.5281/zenodo.6598552 (ref. [77]). The following software and packages were used for data analysis: PLINK 2.0 (https://www.cog-genomics.org/plink/2.0/), METAL (http://csg.sph.umich.edu/abecasis/Metal/download/), MAGMA v.1.07 (https://ctg.cncr.nl/software/magma), FUMA

(https://fuma.ctglab.nl/), LDscore (https://github.com/bulik/ldsc), LDstore v.1.1 (http://www.christianbenner.com/#), FINEMAP v.1.3.1 (http://www.christianbenner.com/#), PRSice v.2.2.1 (https://www.prsice.info/), Eagle2 (https://alkesgroup.broadinstitute.org/Eagle/downloads/), Minimac3 (https://genome.sph.umich.edu/wiki/Minimac3), REGENIE v.2.0.1 (https://rgcgithub.github.io/regenie/), MiXeR (https://github.com/precimed/mixer), BUHMBOX (https://software.broadinstitute.org/mpg/buhmbox/) and R v.3.6.0 (https://www.r-project.org/) with packages data.table v.1.12.2, GenomicRegion v.1.36.0, stats v.3.6.3, fpc v.2.2-3, coloc v.3.2-1, Rtsne v.0.15 and ggplot2 v.3.2.0.

References

55. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
56. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
57. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
58. Euesden, J., Lewis, C. M. & Reilly, P. F. O. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).
59. Hawrylycz, M. J. et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).
60. Zeisel, A. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **348**, 1138–1142 (2015).
61. Romanov, R. A. et al. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* **20**, 176–188 (2017).
62. Marques, S. et al. Oligodendrocyte heterogneity in the mouse juvenile and adult central nervous system. *Science* **352**, 1326–1329 (2016).
63. La Manno, G. et al. Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167**, 566–580 (2016).
64. Muñoz-Manchado, A. B. et al. Diversity of interneurons in the dorsal atriatum revealed by single-cell RNA sequencing and PatchSeq. *Cell Rep.* **24**, 2179–2190 (2018).
65. Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030 (2018).
66. Watanabe, K., Umićević Mirkov, M., de Leeuw, C. A., van den Heuvel, M. P. & Posthuma, D. Genetic mapping of cell type specificity for complex traits. *Nat. Commun.* **10**, 3222 (2019).
67. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
68. Benner, C. et al. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017).
69. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
70. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
71. Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
72. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
73. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
74. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
75. 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).
76. Maaten, L. VanDer & Hinton, G. Visualizing high dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
77. Watanabe, K. Gene prioritization using multi-loci information for insomnia meta analysis. https://doi.org/10.5281/zenodo.6598552

Acknowledgements

We thank both UKB and 23andMe participants who consented to participate in research, and researchers who collected and contributed the data. D.P. was funded by The Netherlands Organization for Scientific Research (no. NWO VICI 453-14-005), NWO Gravitation: BRAINSCAPES: A Roadmap from Neurogenetics to Neurobiology (grant no. 024.004.012) and a European Research Council advanced grant (no. ERC-2018-AdG GWAS2FUNC 834057). E.J.W.V.S. was funded by the European Research Council (no. ERC-ADG-2014-671084 INSOMNIA) and P.R.J. was funded by the Netherlands Organization for Scientific Research (no. ZonMW VENI-09150162010138). The research was conducted using the UK Biobank Resource (application no. 16406). Analyses were carried out on the Genetic Cluster Computer hosted by the Dutch National Computing and Networking Services, SurfSARA. We additionally thank the GTEx Portal for providing RNA-seq data. The research was based in part on data from the Million Veteran Program – Office of Research and Development, Veterans Health Administration, supported by award nos. CSP 575B and Merit 1I01CX001849.e.

Author contributions

D.P. conceived the study. K.W. performed analyses. J.F.S. performed quality control on the UKB data and wrote the analysis pipeline. P.N., D.A.H., X.W. and the 23andMe Research Team contributed and analyzed the 23andMe cohort data. J.G., D.F.L., R.P. and M.B.S. performed PGS analysis for the MVP cohort. E.J.W.V.S and A.B.S provided valuable discussions. K.W., P.R.J. and D.P. wrote the paper.

Competing interests

P.N., X.W., D.A.H. and members of the 23andMe research team are employees of 23andMe, Inc. and hold stock or stock options in 23andMe, Inc. K.W. is a current employee of Regeneron Pharmaceuticals and holds stock and stock options in Regeneron Pharmaceuticals. The remaining authors declare no competing interests.

Additional information

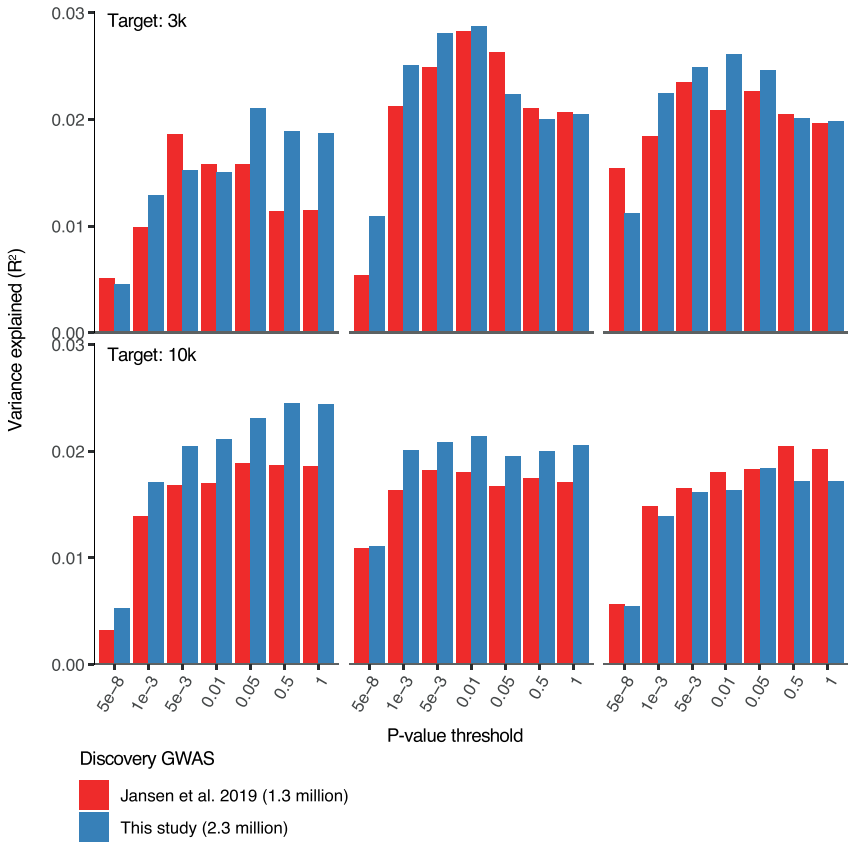
Extended data is available for this paper at https://doi.org/10.1038/s41588-022-01124-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01124-w.

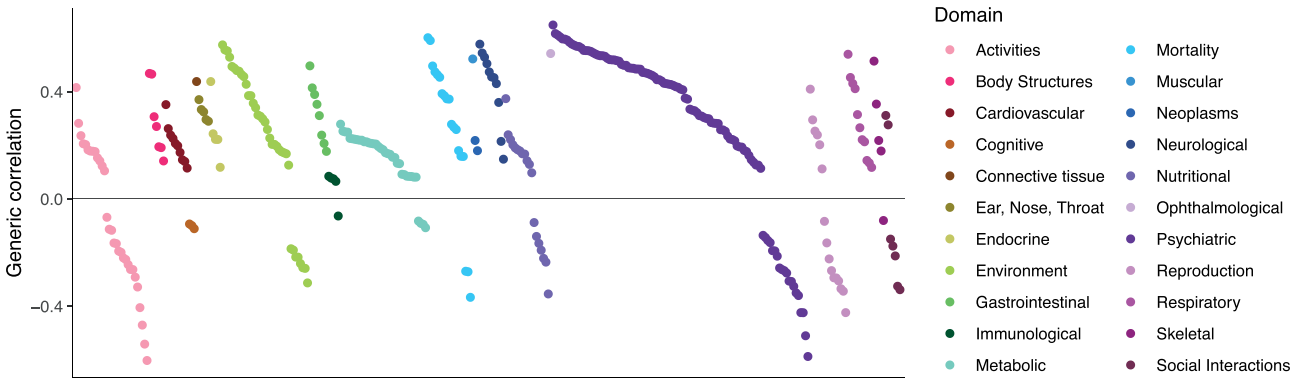
Correspondence and requests for materials should be addressed to Danielle Posthuma.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

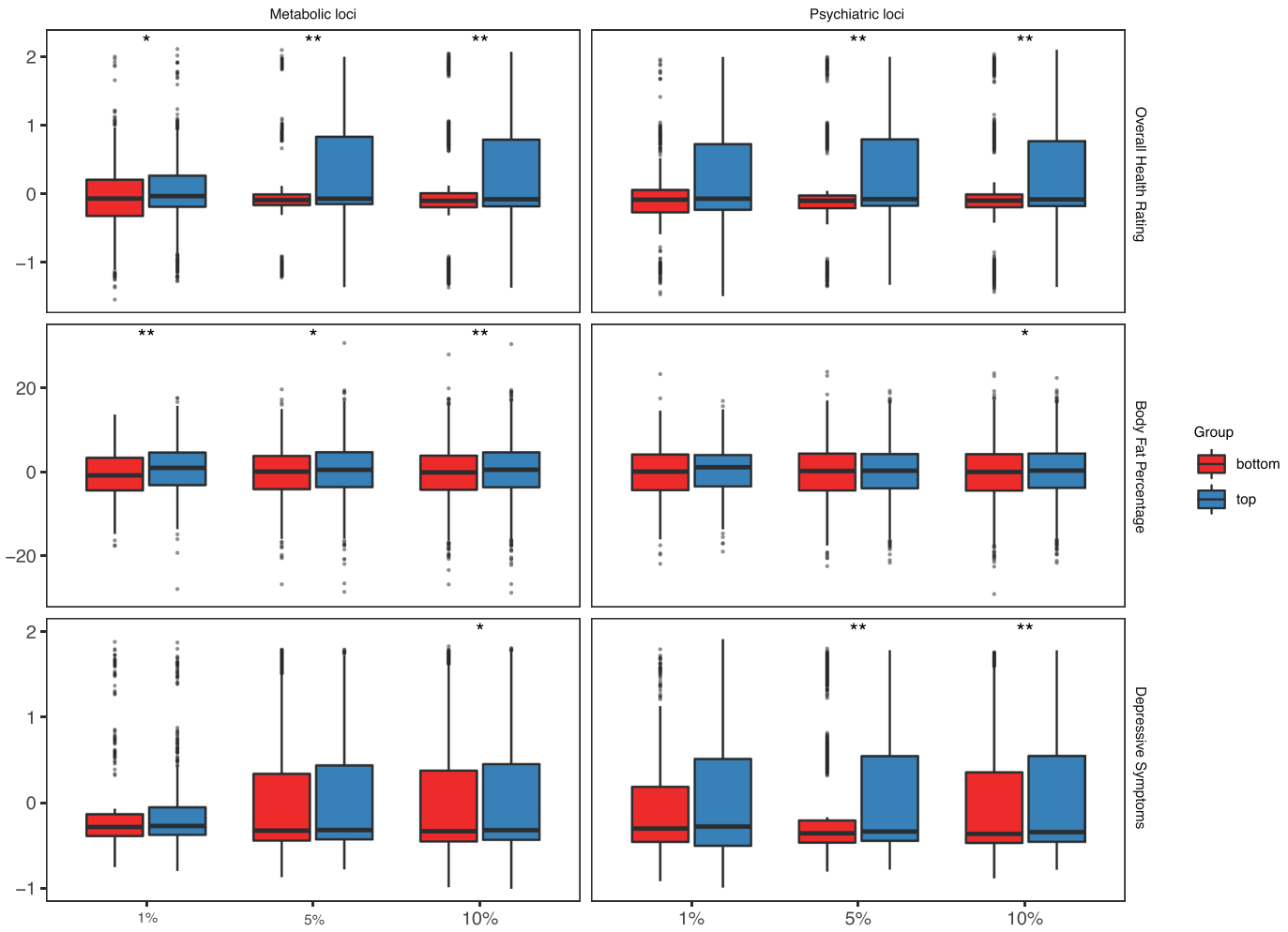
Reprints and permissions information is available at www.nature.com/reprints.



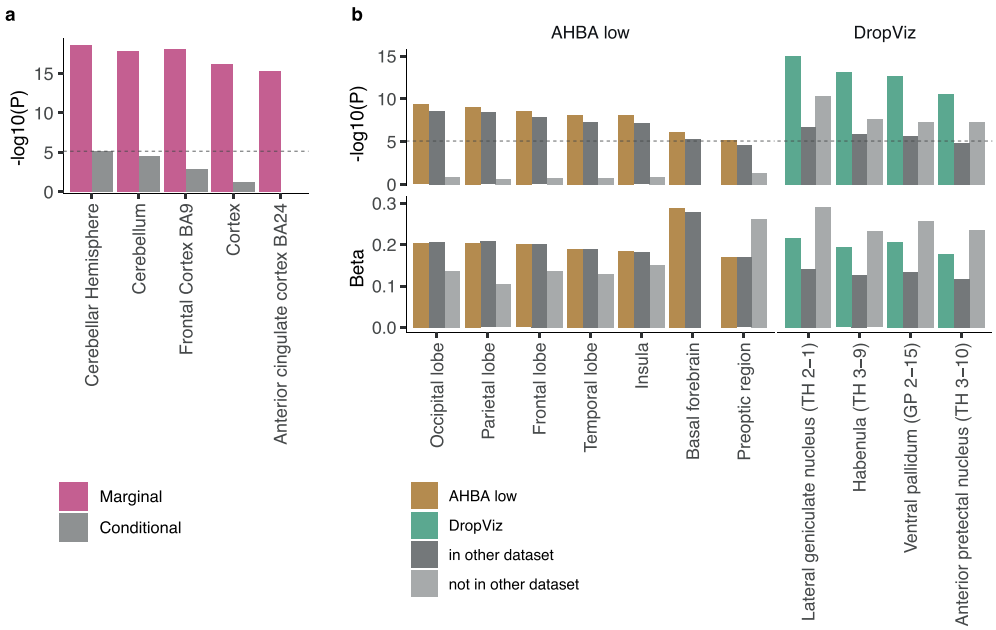
Extended Data Fig. 1 | Phenotypic variance explained by polygenic risk scoring. Bars are colored by P-value threshold of SNPs used to compute the polygenic risk score.



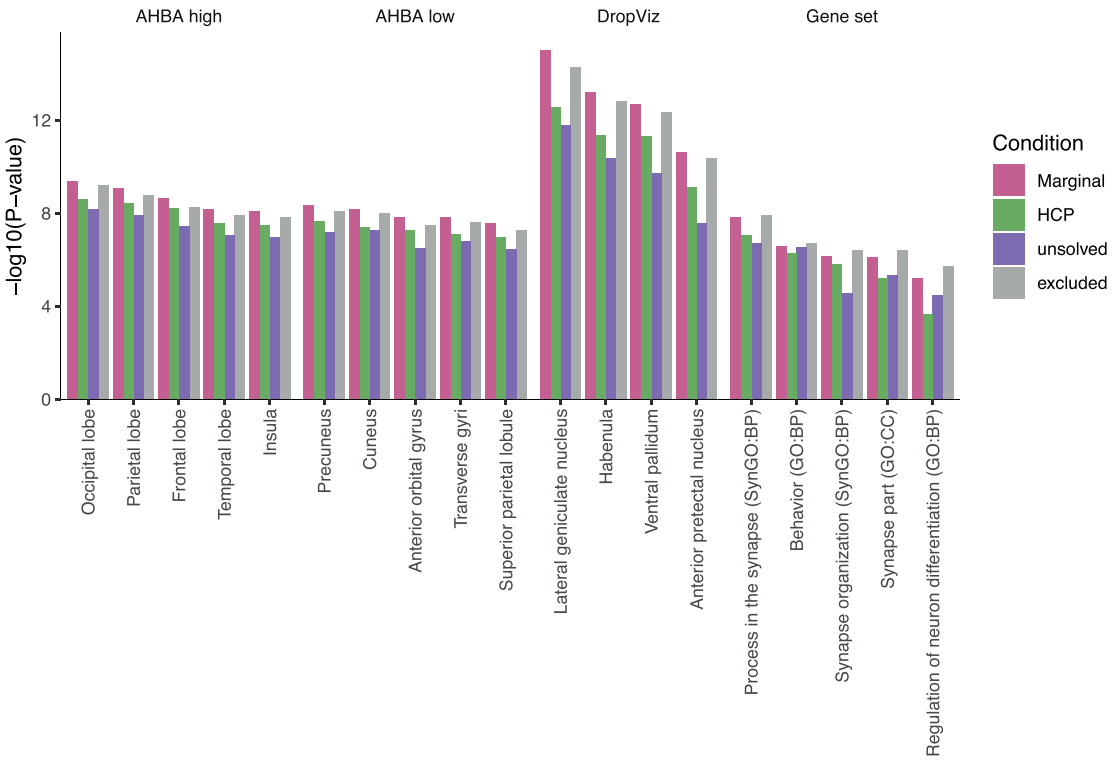
Extended Data Fig. 2 | Genetic overlap between insomnia and 350 traits. Significant genetic correlations of insomnia with 350 traits after Bonferroni correction ($p < 9.07e\text{-}5$). P-values were based on two-sided Z-test. Each data point represents a trait and is colored by the domain category.



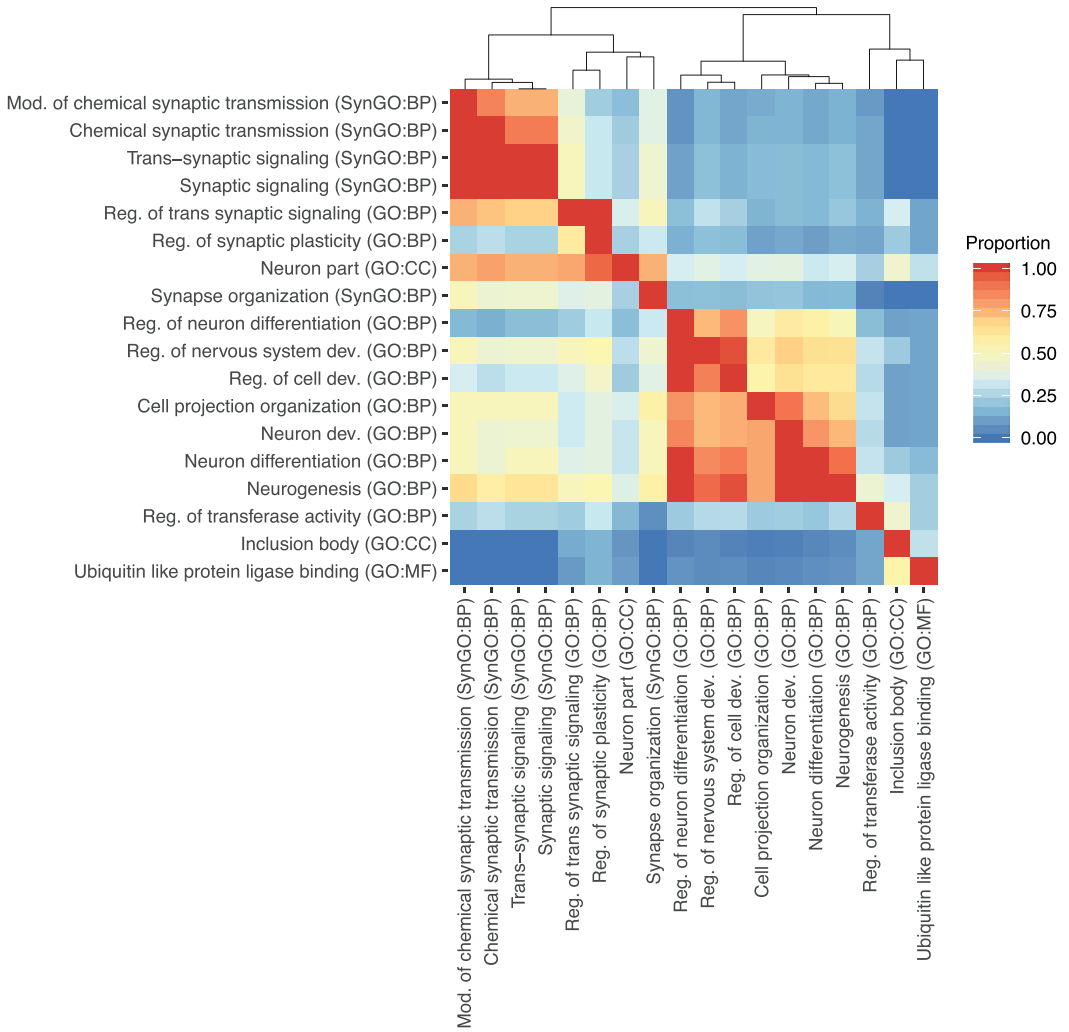
Extended Data Fig. 3 | Distribution of PRS based on metabolic and psychiatric loci. A single star represents nominal significant ($p<0.05$) and double star represents significant after Bonferroni correction ($p<0.05/9$) of two-sided Mann-Whitney U test (see Supplementary Table 21 for full results). The boxes indicate 25% (Q1) and 75% (Q3) quantiles and horizontal black lines indelicate median. The minimum and maximum of the whisker are Q1-1.5*IQR and Q3+1.5*IQR where IQR is Q3-Q1. Data points which do not fall within the whisker's interval are displayed as dots. Number of data points (individuals) are: for column 1 (based on metabolic loci) 300 top and 299 bottom 1%, 1495 top and bottom 5%, 2986 top and 2984 bottom 10% for overall health rating, 297 top and bottom 1%, 1475 top and 1471 bottom 5%, 2950 top and 2948 bottom 10% for body fat percentage, 281 top and 283 bottom 1%, 1405 top and 1402 bottom 5%, 2812 top and 2815 bottom 10% for depressive symptoms, for column 2 (based on psychiatric loci) 299 top and 298 bottom 1%, 1490 top and 1493 bottom 5%, 2986 top and 2990 bottom 10% for overall health rating, 297 top and 294 bottom 1%, 1470 top and 1477 bottom 5%, 2941 top and 2959 bottom 10% for body fat percentage, 287 top and 285 bottom 1%, 1403 top and 1418 bottom 5%, 2809 top and 2844 bottom 10% for depressive symptoms.



Extended Data Fig. 4 | Additional conditional analyses for MAGMA tissue and brain region association analyses. P-values were computed by MAGMA gene analysis based on one-sided T-test for the regression coefficient of the gene expression. (**a**) P-values of brain regions from GTEx, with (Conditional) and without (Marginal) conditioning on the average expression across 13 brain regions. (**b**) Comparison of AHBA (low resolution) and DropViz datasets with MAGMA gene-property analysis. P-values (top) and standardized effect size (Beta, bottom) of brain regions from the AHBA low dataset and cell types from the DropViz dataset. The most left bar indicates the marginal association statistics for each item. The middle bar indicates the association statistics based only on genes present in both datasets (~11,000 genes). The most right bar indicates the association statistics based only on genes that are not available in the other dataset (~2,000 for AHBA low and ~4,000 for DropViz). The horizontal dashed line indicates the Bonferroni corrected threshold for statistical significance (p=0.05/5974).



Extended Data Fig. 5 | MAGMA gene-property and gene-set analyses conditioning on sets of genes from insomnia risk loci. The top (most significantly associated) 5 brain regions/cell types/gene-sets (referred to as gene-sets hereafter) were selected for each dataset, except for DropViz where 4 independently associated cell types were selected. For each gene-set, MAGMA was performed while conditioning on 3 sets of genes; high-confidence prioritized (HCP), unsolved and excluded genes.



Extended Data Fig. 6 | Heatmap of the overlap of genes across significantly enriched gene-sets. The displayed 18 gene-sets showed significant enrichment with 289 HCP genes. The heatmap is asymmetric. A cell of row *i* and column *j* represents the proportion of the prioritized genes in the gene-set *i* relative to the number of prioritized genes in the gene-set *j*.

nature research

Corresponding author(s): Danielle Posthuma

Last updated by author(s): May 31th, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ 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 all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection purpose
Data analysis	R script to perform gene prioritization approach proposed in this manuscript is available at https://doi.org/10.5281/zenodo.6598552. The following software and packages were used for data analysis: PLINK 2.0 (https://www.cog-genomics.org/plink/2.0/), METAL (http://csg.sph.umich.edu/abecasis/Metal/download/), MAGMA v1.07 (https://ctg.cncr.nl/software/magma), FUMA (https://fuma.ctglab.nl/), LDscore (https://github.com/bulik/ldsc), LDstore v1.1 (http://www.christianbenner.com/#), FINEMAP v1.3.1 (http://www.christianbenner.com/#), PRSice v2.2.1 (https://www.prsice.info/), Eagle2 (https://alkesgroup.broadinstitute.org/Eagle/downloads/), Minimac3 (https://genome.sph.umich.edu/wiki/Minimac3), REGENIE v2.0.1 (https://rgcgithub.github.io/regenie/), MiXeR (https://github.com/precimed/mixer), BUHMBOX (https://software.broadinstitute.org/mpg/buhmbox/), R v3.6.0 (https://www.r-project.org/) with packages data.table v1.12.2, GenomicRegion v1.36.0, stats v3.6.3, fpc v2.2-3, coloc v3.2-1, Rtsne v0.15 and ggplot2 v3.2.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The full GWAS summary statistics for UKB and top 10,000 SNPs for 23andMe are available at https://ctg.cncr.nl/software/summary_statistics/. The full GWAS

summary statistics for the 23andMe data set will be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Please visit https://research.23andme.com/collaborate/#publication for more information and to apply to access the data. The following publicly available datasets were used in this manuscript: GTEx v8 (https://gtexportal.org/home/datasets), Allen Human Brain Atlas (http://human.brain-map.org/static/download), scRNA-seq from Linnerson's lab (http://linnarssonlab.org/data/; GSE60361, GSE74672, GSE75330, GSE76381, GSE97478), DropViz (http://dropviz.org/), MsigDB v6.2 (http://software.broadinstitute.org/gsea/msigdb/index.jsp, InWeb protein-protein interaction (https://www.intomics.com/inbio/map/api/get_data?file=InBio_Map_core_2016_09_12.tar.gz), eQTLGen (https://www.eqtlgen.org/), PsychEncode (http://resource.psychencode.org/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For all samples the sample size consists of all individuals that remain after quality control of the data and exclusion of withdrawn subjects. Detailed information on the samples used, as well as the exclusion/inclusion criteria, are provided in the Method (Sample cohorts section).
Data exclusions	See "Sample cohorts" section in the Method. All individuals of non-European ancestry, individuals who failed quality control, or individuals who asked to withdraw from the study were removed.
Replication	Both UKB and 23andMe datasets were included in the primary meta-analysis. We explicitly examined the concordance of effects within the novel loci between these two parts of our data. In addition, we checked how our results replicated previously reported loci. This information is available in Supplementary Table 5. There were 18 out of 202 loci from the previous insomnia meta-analysis (Jansen et al. Nat. Genet. 2019) that were no longer significant in this study. These loci are most likely false positive in the previous study as they showed significantly higher P-value compared to replicated loci (details are discussed in Supplementary Note).
Randomization	Covariates were used to control for differences in age, ancestry, and sex.
Blinding	This analysis was exploratory so blinding was not applicable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	We selected UKB 386,989 unrelated participants who had insomnia phenotype and assigned to European ancestry (208,958 females and 178,030 males). Participants are aged between 38 and 73 years old. All individuals included in the study provided informed consent.
Recruitment	Participants were recruited based on participation in biobanks
Ethics oversight	The UKB received ethical approval from the National Research Ethics Service Committee North West-Haydock), and all study procedures were in accordance with the World Medical Association for medical research.

Note that full information on the approval of the study protocol must also be provided in the manuscript.