nature portfolio

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

Supplementary information

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

In the format provided by the authors and unedited

Supplementary Information for:
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[3], 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]

Affiliations:

1. Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University Amsterdam, the Netherlands
2. Department of Clinical Genetics, Section Complex Trait Genetics, Neuroscience Campus Amsterdam, VU Medical Center, 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, 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 Center, Amsterdam, The Netherlands.

*Correspondence to: Danielle Posthuma, Department of Complex Trait Genetics, VU University, De Boelelaan 1085, 1081 HV, Amsterdam, The Netherlands. Phone: +31 20 5982823, Fax: +31 20 5986926, E-mail: d.posthuma@vu.nl

Table of Contents

Supplementary Note

1. Cohort description and phenotype definition

1.1 Sample cohorts

UK Biobank:

The UK Biobank (UKB) study is designed to curate genotype and phenotype information for ~500,000 adult participants in the UK[1]. All participants provided written informed consent (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. The current study was carried out using the UK Biobank resource under application number 16406. We used genotype data released in May 2018 by UK Biobank (UKB) in which genotyping was completed in 489,212 individuals, and 488,377 individuals passed genotype quality controls performed centrally by UKB. These genotypes were imputed based on the Haplotype Reference Consortium (HRC) reference panel[2] and a combined reference panel of UK10K[3] and 1000 Genome projects Phase 3 (1000G)[4] panels. The imputed genotypes were available for 487,422 individuals. Imputed variants with INFO score >0.9 were converted to hard-calls at a certainty threshold of >0.9. Although UKB provides self-reported ancestry for each participant, we determined European ancestry by projecting 1000G genetic principal components on the UKB genotypes and assigned ancestry based on the closest Mahalanobis distance from the 1000G population average (see ref[6] for more details). This resulted in 460,527 individuals assigned to the European population, of whom 387,614 unrelated individuals were included in the current analyses.

23andMe:

23andMe Inc. is a personal genetics company. Customers of 23andMe have the option to consent to participate in research and answer survey questions on-line about a wide variety of phenotypes. We obtained insomnia GWAS summary statistics from 23andMe Inc. research participants who are of European ancestry. All participants included in the study provided informed consent, and the research study and data collection procedures were approved by an AAHRPP-accredited private institutional review board, Ethical and Independent Review Service.

DNA was extracted from saliva samples and each sample was genotyped by either by Illumina HumanHap550+ BeadChip (~560k SNPs), Illumina OmniExpress+ BeadChip (~950 SNPs), Infinium Global Screening Array (~640 SNPs) or fully customized array containing

~560k SNPs. Samples that failed to reach 98.5% call rate were re-analyzed. Individuals whose analyses failed repeatedly were re-contacted by 23andMe to provide additional samples. Genotypes were phased by Finch (23andMe's inhouse tool modified from Beagle[5]) or Eagle2[6] (for samples genotyped by the custom array). Subsequently, genotypes were imputed against a combined panel of 1000G[4] and UK10K[3] merged by Minimac3[7].
Ancestry of individuals was determined by Ancestry Composition[8]. The reference population data was derived from public datasets (the Human Genome Diversity Project[9], HapMap[10] and 1000G[4]) and 23andMe customers who have reported having four grandparents from the same country.
A maximal set of unrelated European individuals was used for each association analyses, resulted in 1,038,003, 1,200,179 and 1,978,022 for male-only, female-only and sex-combined GWAS, respectively.

1.2 Phenotype assessment

UK Biobank:

Participants were asked the question 'Do you have trouble falling asleep at night or do you wake up in the middle of the night?' (Data-Field 1200) and answers were selected from the following 4 options; 'never/rarely', 'sometimes', 'usually' and 'prefer not to answer' on a touchscreen. If the participant opened the 'help', the message 'If this varies a lot, answer this question in relation to the last 4 weeks.' was displayed. There were 3 assessments available for insomnia (f.1200.0.0 - f.1200.0.2). For subjects with a missing value in the first assessment (f.1200.0.0), the second assessment (f.1200.0.1) was used if that contained a non-missing value. Subsequently, for subjects with NA in both first and second assessments, the third assessment (f.1200.0.2) was used, again only if there was a non-missing value. Note that for subjects with non-missing values in multiple assessments, only the first assessment was used. We excluded individuals who answered 'prefer not to answer' resulting in a total of 386,988 subjects. We then dichotomized the phenotype where subjects who answered 'usually' were considered as cases, otherwise they were considered controls, following previous strategy[6].

23andMe:

Participants were asked to answer one or more questions related to seven sleep-related traits. Participants with positive response to any of the following questions were considered as cases: 1) 'Have you ever been diagnosed with, or treated for, insomnia?', 2) 'Were you

diagnosed with insomnia?', 3) 'Have you ever been diagnosed by a doctor with any of the following neurological conditions?' (Sleep disturbance), 4) 'Do you routinely have trouble getting to sleep at night?', 5) 'What sleep disorders have you been diagnosed with? Please select all that apply.' (Insomnia, trouble falling or staying asleep), 6) 'Have you ever taken these medications?' (Prescription sleep aids) and 7) 'In the last 2 years, have you taken any of these medications?' (Prescription sleep aids). Participants who did not provide either positive or uncertain answer ('I don't know' or 'I am not sure') in any of the 7 questions nor the followings are considered as controls: 1) 'Have you ever been diagnosed with, or treated for, any of the following conditions?' (Insomnia; Narcolepsy; Sleep apnea; Restless leg syndrome), 2) 'In the past 12 months, have you been newly diagnosed with any of the following conditions by a medical professional?' (Insomnia; Sleep apnea; Migraines), 3) 'Have you ever been diagnosed with or treated for any of the following conditions?' (Post-traumatic stress disorder; Autism; Asperger's; Sleep disorder), 4) 'Have you ever been diagnosed with or treated for a sleep disorder?', and 5) 'Have you ever been diagnosed with or treated for any of the following conditions?' (A sleep disorder). A detailed flowchart of the case/control decisions is provided in the previous study[6].

2. GWAS results for UKB and 23andMe GWAS

We first performed GWAS in UKB (n=386,988) and 23andMe (n=1,978,022) separately. The GWAS on the UKB cohort identified 14 independent risk loci, the SNP heritability (h^2_{SNP}) was 8.23% (SE=0.36%), with a λ_{1000}=1.00, an intercept of 1.02 (SE=0.0086) and a ratio (LD score intercept-1 divided by mean χ^2-1) of 0.043 estimated with LD score regression (LDSC)[11] (**Method**). The 23andMe cohort GWAS identified 477 independent risk loci, with an h^2_{SNP} of 8.15% (SE=0.0021), a λ_{1000}=1.00, an intercept of 1.15 (SE=0.016) and a ratio of 0.084 (SE=0.009). h^2_{SNP} estimates agree well between the two cohorts, and both indicate as much as 95.7% and 91.6% of the observed inflation could be ascribed to true polygenicity and large sample size, respectively. The genetic correlation the GWAS in UKB and 23andMe was 0.66 (SE=0.0179, p=2.7×10^{-292}). This is relatively low, as was also noted in the previous GWAS for insomnia[12]. This between-cohort discrepancy may be due to the lower accuracy of the 23andMe insomnia phenotype as compared the UKB insomnia phenotype, which was also shown when both their measures were benchmarked to the same independent cohort (using the Netherlands Sleep Registry (N=1,918): where we previously showed sensitivity 98% and 84%, and specificity 96% and 80%, for the items used in UKB and 23andMe,

respectively[12,13]). These results suggest that the items used in both cohorts are good predictors of insomnia, although what they ascertain is slightly different phenotypically. In addition, insomnia is a heterogeneous trait and might consist of multiple subdomains of symptoms. In the meta-analysis, associations that are similar across the two cohorts will be amplified, while dissimilar results are less likely to show significant association, supported by both cohorts, in the meta-analysis. There, the relatively low genetic correlation is expected to decrease our statistical power, yet the large sample size partly counterbalances this.

3. UKB GWAS with related individuals

All analyses reported in the main text that concern the UKB sample are based on the UKB sample including only unrelated individuals. However, we also ran the same analyses including related individuals and efficiently correcting for relatedness using REGENIE[14]. We here provide the (slight) differences in results across these two analyses (with and without including related individuals for UKB sample).

By including related UKB EUR individuals, there are 131,177 cases and 328,590 controls in total, increasing 21,629 cases and 51,150 controls from the main UKB GWAS analysis with unrelated EUR individuals. Phenotype was defined same as described above, and the same sets of covariates (age, sex, genotyping array and the first 10 ancestry principal components) were used. We performed two-step REGENIE analysis with the following parameters. For step 1, we used a pruned dataset containing 142,007 variants (r^2<0.1 and MAF>0.01), with the block size of 100 and leave one chromosome out validation[14]. The step 2 was performed for 12,856,090 variants with minimum minor allele count of 100, with the default parameters. We identified 23 independent risk loci (**Methods**), 14 of them were significant in UKB GWAS without related EUR individuals (**Supplementary Table 2**). Of 9 loci that were only identified in UKB GWAS with related individuals, 2 loci showed significant signal in 23andMe GWAS and 4 loci (including those 2) showed significant signal in the meta-analysis (with UKB unrelated EUR and 23andMe, **Supplementary Table 2**). We did not observe an increase in genetic correlation between UKB GWAS and 23andMe GWAS (both UKB GWAS with and without related EUR individuals showed 0.66).

Given the large sample size of 23andMe and a relatively small increase in the sample size of UKB GWAS by including related individuals, this should not substantially affect the conclusions derived by the meta-analysis of UKB GWAS without related individuals and 23andMe GWAS.

4. SNP heritability estimate of insomnia stratified by age and psychiatric condition

To evaluate whether there are different genetic effects of insomnia depending on age or psychiatric condition, we performed a GWAS in the UKB cohort stratified by age groups and depression status.

For age, we performed a GWAS by splitting the sample by a median age 58 years (N=186,656 with age 59-73, and N=200,332 with age 38-58) and calculated the genetic correlations between the two GWAS using LD Score regression[11]. The genetic correlation was 0.96 (p=1.4e-104), showing that there is only a small difference in the genetic factors across age groups. The SNP heritability was 7.95% (SE=0.0056) in the first group (age > 58) and 9.09% (SE=0.0062) in the second group (age <= 58), which indicates that insomnia in the younger population is somewhat more explained by genetic factors compared to the older population.

For depression status, we used 2 UKB phenotypes, "Frequency of depressed mood in last 2 weeks" (field ID 2050) and "Frequency of unenthusiasm / disinterest in last 2 weeks" (field ID 2060) as previously done[13]. Subjects answered "Do not know" or "Prefer not to answer" in both questions were filtered out. We then defined subjects answered either "Several days", "More than half the days" or "Nearly every day" for at least one of the questions as having depressive symptoms and remaining as not having depressive symptoms. This resulted in 102,783 and 260,549 subjects with and without depressive symptoms, respectively. We performed the GWAS for each group separately and observed the genetic correlation of 1.02 (p=2.3e-44) and SNP heritability of 7.38% (SE=0.0081) and 7.67% (SE=0.0051) for with and without depressive symptoms, respectively. This suggests that there is little to no difference in insomnia genetic effects between groups with and without depressive symptoms in the UKB cohort.

5. Estimation of polygenicity and discoverability of insomnia

To estimate the polygenicity and discoverability of insomnia, we used the causal mixture model for GWAS summary statistics (univariate MiXeR) proposed by Holland *et al.*[15,16] (https://github.com/precimed/mixer). In this model, the distribution of SNP effect sizes is treated as a mixture of two gaussian distributions for causal and non-causal SNPs as the following[16]:

$$\beta = \pi N\left(0, \sigma_\beta^2\right) + (1-\pi)N(0,0)$$

where π is the proportion of (independent) causal SNPs and σ_β^2 is the variance of the effect sizes of causal SNPs. Therefore, π and σ_β^2 respectively represent polygenicity and discoverability of the trait. As recommended in the original study, we used 1000 Genome Phase 3 European subjects as a reference panel and restricted to HapMap 3 SNPs. SNPs with chi-square statistics >80 and the MHC region (chr6:26Mb-34Mb) were excluded.
We observed a total SNP heritability of 4.3% of which 17.4% (SE=8.29e-4) is explained by genome-wide significant SNPs.
MiXeR estimated polygenicity of π=4.85e-3 and discoverability of σ_β^2=4.34e-6. By comparing the polygenicity with previously reported traits[17], insomnia is the 3rd most polygenic trait following major depressive disorders (π=5.75e-4, σ_β^2=6.02e-6) and educational attainment (π=4.89e-4, σ_β^2=1.21e-5).
To estimate the sample size required to explain 90% of the $h2_{SNP}$, we used the output of GWAS power estimates calculated in the MiXeR software, which contains 51 data points of sample size and the proportion of SNP heritability explained. By using the *interp1* function from the pracma package in R, we estimated that ~57 million are needed to explain 90% of the $h2_{SNP}$ with GWS SNPs.

6. Polygenic score prediction

The semi-out of sample polygenic score (PGS) prediction was computed for three randomly selected 10,000 UKB target hold-out samples (**Methods**), which explained 2.46% of the phenotypic variance at most (**Extended Data Fig. 1** and **Supplementary Table 3**). This is slightly lower than previously reported (2.60%)[12]. To evaluate the predictive power of the discovery GWAS, we performed the same analyses with insomnia meta-analysis of the UKB GWAS of the training set and the previous 23andMe GWAS[12] (total ~1.3 million samples). We observed a higher predictive power with the larger discovery dataset for 2 out of 3 target sets. We also obtained 3 sets of 3,000 target sets used in the previous study, which showed generally increasing predictive power with the increasing sample size of the discovery GWAS. These results show that, even though the current sample may include more heterogeneity as indicated by a relatively decreased genetic correlation between the UKB and 23andMe GWAS (0.66) compared to the previous study (0.69), the current larger sample does provide slightly increased predictive power compared to Jansen *et al.* when using the same target sample. We note that we did not replicate 2.60% with the previous discovery GWAS and the target samples because the UKB GWAS was re-performed with newer

release of genotype datasets. In addition, we report R^2 adjusted for ascertainment with a population prevalence 0.3 which is recommended by Choi *et al.*[18] while it was not corrected in the previous study.

We further performed PGS prediction for an independent cohort consisting of 45,355 insomnia cases and 138,589 controls (the Million Veteran Program (MVP). To evaluate the predictive power of the insomnia GWAS meta-analysis, we computed PGS for an independent cohort, the Million Veteran Program (MVP)[19]. The MVP is a research program funded by the US Department of Veterans Affairs to learn how genes, lifestyle, and military exposures affect health and illness[51]. In the MVP cohort, insomnia cases were defined as individuals answering "Yes" to any of the following three questions : "Trouble falling asleep when you first go to bed", "Waking up during the night and not easily going back to sleep, or "Waking up in the morning earlier than planned or desired". To be defined as a case, individuals also had to respond "Yes" to the following: "Feeling unsatisfied or not rested by your night's sleep". Controls were defined as not responding "Yes" to any of the first three questions and responding "No' to the final question. This resulted in 45,355 cases and 138,589 controls. We then computed PGS using PRSice in the same way as described above, except the summary statistics of the full meta-analysis were used as a training set this time. Using the full meta-analysis results to compute PGS, at most 0.66% of the phenotypic variance was explained at (**Supplementary Table 3**).

Although the sample prevalence of the MVP cohort (24.7%) is somewhat close to the UKB and 23andMe combined samples (25.1%), the predictive power for the MVP cohort based on UKB and 23andMe meta-analysis was considerably lower than the subset of UKB samples based on remaining UKB and 23andMe meta-analysis. The higher prediction power of the latter analysis is likely because the training samples are from a relatively homogenic cohort and part of samples from the same cohort are included in the training sets, or because there is phenotypic heterogeneity. We note that MVP cohort consists of mainly male samples (93.7%), although the sex imbalance might be a cause of differential findings or lower predictive power, as previous studies have showed high concordance between MVP and other cohorts that had a more balanced sex-ratio for psychiatric and behavioral traits[20–24].

7. Suspicious loci

Using FUMA[25], we identified 558 risk loci from the sex-combined meta-analysis. Of those, 8 loci contained only a single SNP. We visually examined those loci and 4 loci were considered

likely false positives as none of the nearby SNPs showed any potential signals (**Supplementary Fig. 2**). These 4 loci were marked in **Supplementary Table 4** and were excluded from all analyses except MAGMA gene/gene-set analysis, because these analyses are based on aggregate association values across many SNPs in a gene and are therefore not strongly influenced by a single SNP.

8. P-value threshold for genome-wide significance

Four GWAS studies have been performed for insomnia to date[12,13,26,27], which allows to compare GWAS findings with different sample sizes. We noticed that 18 out of 202 loci (8.9%) identified from the previous GWAS meta-analysis were not replicated in the current study suggesting they may have been false positives[12]. The significance of those loci was mostly borderline in the previous GWAS; i.e. 14 loci with p<1e-8 and 4 with p<5e-10. It has been recently debated whether the 'golden-standard' of the genome-wide significance threshold, p<5e-8 is still applicable as the number of accessible SNPs increases with increasing sample size[28]. Although this threshold was preserved in the current study, we compared the number of independent loci with different P-value thresholds. We observed that 23.6% of loci are no-longer significant by decreasing the threshold to 1e-8 from 5e-8 and the number of loci almost exponentially decreased along with the decreasing P-value threshold (**Supplementary Fig. 3**). To evaluate whether this proportion (i.e. 23.6%) is specific to insomnia, we conducted similar analyses for three other traits with GWAS outcomes for multiple sample sizes: educational attainment[29] (https://www.thessgac.org/data), height[17] (https://atlas.ctglab.nl/traitDB/3187) and type 2 diabetes[30] (https://www.diagram-consortium.org/downloads.html). Risk loci were defined in the same way as for insomnia GWAS. In these three traits we observed 23.5%, 17.0% and 22.5% decrease of the number of risk loci for educational attainment, height and type 2 diabetes, respectively by decreasing the P-value threshold to 1e-8 from 5e-8. Thus all three traits showed very similar proportions of decreasing number of the loci along with the decreasing P-value threshold with insomnia GWAS (**Supplementary Fig. 3**).

9. Effects of increasing sample size in insomnia GWAS

Here we further sought to investigate changes in association P-values and effect sizes of SNPs in insomnia GWAS as a function of sample size from ~100k up to over 2 million. Using the 4 previously published GWAS studies as well as the current study, we compared

the sample size and number of detected risk loci from six GWAS cohorts: UK Biobank 1st release (UKB1, N=113,006), UK Biobank latest release (UKB2, N=386,988), 23andMe 1st GWAS (23andMe1, N=944,477), meta-analysis of UKB1 and 23andMe1 (meta1, N=1,331,010), 23andMe 2nd GWAS (23andMe2, N=1,978,022) and meta-analysis of UKB2 and 23andMe2 (meta2, N=2,365,010; main analysis of the current study).

Insomnia GWAS summary statistics for 6 different sample sizes were obtained as below.

- *UKB1* (N=113,006): The summary statistics were obtained from the study of Hammerschlag *et al.*[13] (https://ctg.cncr.nl/software/summary_statistics).
- *UKB2* (N=386,988): UKB only summary statistics from the current study.
- *23andMe1* (N=944,477): Only 23andMe cohort from the study of Jansen *et al.*[12] (summary statistics is not publicly available, only the number of risk loci was obtained).
- *Meta1* (N=1,331,010): The summary statistics of meta-analysis from the study of Jansen *et al.*[6] (summary statistics is not publicly available, only lead SNPs were obtained).
- *23andMe2* (N=1,978,022): 23andMe only summary statistics from the current study (not publicly available).
- *Meta2* (N=2,365,010): The summary statistics of the main meta-analysis from the current study (not publicly available).

For each GWAS, risk loci were defined as described above. For 23andMe1 and Meta1, the number of risk loci were obtained from the previous study[12] as the summary statistics are not available. For comparison of P-values of the same locus across GWAS with different sizes, we obtained lead SNPs for Meta1 from the previous study[6] while 23andMe1 was excluded from the analysis.

To obtain matched SNPs across GWAS to compare effect sizes, we first extracted genome-wide significant SNPs for all risk loci from 5 GWAS (23andMe1 was excluded from the analysis). For Meta1, the analysis was limited to lead SNPs provided by the previous study[12], therefore not all GWS SNPs in the risk loci were assessed. We then selected SNPs that reached genome-wide significance in at least 3 GWAS out of 5. From each locus, the single SNP with the minimum P-value was further selected (when there are multiple) for the analyses. Standardized effect sizes of those SNPs were computed based on summary statistics from each GWAS as described above. When Z-score were not available, P-values were converted to Z-score (two-sided). We do note that several of these cohorts overlap (i.e.,

UKB2 includes UKB1, 23andMe2 includes 23andMe2, Meta1 is UKB1+23andMe1, and Meta2 is UKB2+23andMe2).

As expected, the number of risk loci almost linearly increased with the sample size (**Fig. 1b**). As mentioned above, several risk loci identified in the previous insomnia meta-analysis (Meta1) were not replicated in this study (Meta2). To investigate whether the distributions of P-values differed between replicated and non-replicated risk loci, we categorized risk loci into 4 types: Type 1) loci that were only detected in a single GWAS Type 2) novel loci identified in the corresponding GWAS and replicated in the GWAS(s) with larger sample size but not seen in GWAS of smaller sample size, Type 3) loci identified in the GWAS with both smaller and larger sample sizes, and Type 4) loci identified in the GWAS(s) with smaller sample size(s) but not in the GWAS with larger sample size. Note that type 1 and 4 in meta2 are potential type 2 and 3 loci, respectively. As we expected, type 1 loci tended to show P-values just below the genome-wide significance threshold (p=5-e8), and showed significantly higher P-values compared to other loci (p=3.2e-19, Mann-Whitney U test, two-sided, excluding SNPs from UKB1 and meta2; **Supplementary Fig. 4a**). Novel (type 2) loci tend to show higher P-values compared to type 3 and 4 loci (p=5.2e-34), indicating, as expected for true positive findings, a decreasing P-value of the GWAS loci with increasing sample sizes (**Supplementary Fig. 4a**). However, type 4 loci did not show a significant difference of P-values compared to type 2 loci (p=0.24; **Methods**).

We further investigated the changes in effect sizes of SNPs with increasing sample size. We selected SNPs that reached genome-wide significance in at least 3 GWAS and we selected one SNP per locus with the minimum P-value in case there were multiple SNPs available (**Methods**). Thus, we note that the evaluated SNPs did not necessarily represent the most significant SNP in the corresponding locus in all GWAS. This resulted in 175 unique SNPs. From each GWAS, standardized effect sizes were computed for each SNP, and absolute effect sizes were regressed on the sample size per SNP (**Methods**). Our results did not show any significant increase or decrease in standardized effect sizes as a function of sample size (**Supplementary Fig. 4b**).

10. BUHMBOX analysis

To evaluate whether the most significantly associated gene *PTPRD*, which was previously reported to be associated with restless leg syndrome (RLS), is due to pleiotropy, or comorbidity or misclassification of RLS and/or insomnia, we performed BUHMBOX

(Breaking Up Heterogeneous Mixture Based On Cross-locus correlation; https://software.broadinstitute.org/mpg/buhmbox/)[31].

To perform BUHMBOX[31], from the largest RLS meta-analysis[32], we obtained 23 SNPs in Table 1 (excluding 2 SNPs that were not replicated). Of these 23 SNPs, 22 SNPs were present in our insomnia meta-analysis (**Supplementary Table 9**). We used genotype data of 109,548 cases and 277,440 controls from the UKB cohort to test the presence of a sub-group of RLS cases within the insomnia cases, possibly driving the observed association with PTPRD.

With 22 RLS SNPs[32], we observed a significant result (p=9.7e-4) indicating there may be a subgroup of insomnia cases that are RLS cases driving the association of these loci. However, the BUHMBOX result remained significant when removing 2 *PTPRD* loci (p=6.2e-3) while it was no longer significant by removing *MEIS1* locus (p=8.5e-2). These results suggest that the association of *PTPRD* is likely pleiotropic with insomnia and RLS while the association of *MEIS1* is more likely due to a subgroup of RLS cases within insomnia cases. We note that the analysis is limited to the UKB cohort as we did not have access to the individual level genotype data of the 23andMe cohort. We were also unable to perform reverse analysis (whether RLS associations were due to a sub-group of insomnia cases within RLS cases) as we did not have access to the individual level genotype data from Didriksen *et al*.

11. Results for sex specific meta-analyses

From the male specific meta-analysis (222,753 cases and 993,280 controls), 4,781 SNPs reached genome-wide significance resulting in 100 loci including 114 lead SNPs (**Supplementary Tables 8** and **9**). 90 loci overlapped with one of the 554 loci from the sex-combined meta-analysis. From the female specific meta-analysis (390,750 cases and 1,018,386 controls), 24,181 SNPs reached genome-wide significance resulting in 303 loci and 377 lead SNPs (**Supplementary Tables 11** and **12**). 275 loci overlapped with one of the 554 loci from sex-combined meta-analysis. The larger number of risk loci identified from the female meta-analysis is most likely due to the higher effective sample size (male N_{eff}=727,796 vs female N_{eff}=1,129,585). In total, there were 10 and 28 loci specifically identified for male and female meta-analyses, respectively (**Supplementary Fig. 5a**). Subsequently, MAGMA gene analyses were performed for the sex specific meta-analyses. We identified 278 (in males) and 871 (in females) significant genes (0.05/19751=2.5e-6) and 191 genes were significantly associated in both sexes (**Supplementary Table 7**). The

correlation of -log10 gene-based P-value between the sexes was considerably low (r=0.56, p<1e-323; **Supplementary Fig. 5b**), given high genetic correlation (r_g=0.92), while the correlation between sex-specific and sex-combined meta-analysis was relatively high, 0.78 (for males) and 0.92 (for females), both with p<1e-323 (**Supplementary Fig. 5c-d**). The higher correlation of sex-combined with female-only meta-analysis than male-only meta-analysis could again be due to the difference in the effective sample size, which is greater for females than males. The low correlation between sexes suggests there may be sex specific genetic causes to the insomnia.

12. Genetic overlap with multiple clusters of traits and heterogeneity of insomnia risk loci

It was previously shown that insomnia is genetically correlated with multiple psychiatric and metabolic traits[4]. To further investigate these associations, we used LDSC[11] to estimate genetic correlations (r_g) between insomnia and 551 traits with a SNP h^2_{SNP}>0.01 and Z-score>2 reported by Watanabe *et al.*[17], excluding insomnia and one depression item that indexes insomnia (trouble falling asleep; **Methods**). After Bonferroni correction, 350 traits showed a significant r_g with insomnia (0.05/551=9.1e-5), of which 270 traits showed positive and 80 traits showed negative correlations with insomnia (**Extended Data Fig. 2** and **Supplementary Table 15**). The strongest positive correlation was with major depressive disorder (MDD; r_g=0.65, p=2.7e-178), in line with previous findings[12]. The strongest negative correlation was with health satisfaction (r_g=-0.60, p=3.9e-110). We also observed that insomnia was positively correlated with multiple cardiovascular, metabolic, psychiatric diseases/disorders, in agreement with previous reports (**Extended Data Fig. 2** and **Supplementary Table 15**).

We next sought to identify insomnia risk loci sharing the same causal SNPs with the 350 significant genetically correlated traits by colocalizing each of the 554 risk loci with the GWAS summary statistics of the 350 traits (**Methods**). Of the 554 loci, 282 were colocalized with at least one of, in total, 231 traits (**Supplementary Table 16**). Body Mass Index showed the greatest number of colocalized loci with insomnia (41 loci), and 18 other metabolic traits colocalized with >20 loci (**Supplementary Table 17**). From other trait domains, height (27 loci), overall health rating (26 loci), educational attainment (24 loci) and neuroticism (23 loci) showed the largest number of colocalized loci among the others (**Supplementary Table 17**). Despite the high r_g with insomnia, MDD only colocalized with 5 loci.

To investigate colocalization patterns of insomnia risk loci across traits, we counted the number of shared colocalized loci between each pair of 231 traits and projected them onto a 2D-map using the t-distribution stochastic neighbor embedding (tSNE; **Methods**). Subsequently, we performed density-based spatial clustering of applications with noise (DBSCAN) on the tSNE 2D-map to identify clusters of traits.

To identify clusters of traits that are more likely to share colocalized insomnia loci, we tested whether the number of shared loci within each cluster was higher than the number of shared loci between clusters with Mann-Whitney U tests (one-sided, greater). Clusters with $p \geq 0.05$ were discarded (**Methods**). We identified 5 dense clusters of traits (**Supplementary Fig. 6a**). Cluster #1 contained the largest number of traits (82) and was dominated by metabolic traits (33) but also included traits from multiple domains such as height, intelligence, health rating, diabetes, high blood pressure and sleep duration (**Supplementary Tables 17** and **18**). Cluster #2 consisted of 29 traits, of which 24 were psychiatric traits including neuroticism, ever smoker, risk taking and broad depression (**Supplementary Tables 17** and **18**). Cluster #3 consisted of 21 traits, of which 7 were nutritional traits (e.g. intake of food and drinks), but the cluster also included multiple cardiovascular traits such as CAD and high blood pressure as well as body fat percentage (**Supplementary Tables 17** and **18**). Clusters #4 and #5 consisted of 7 and 5 traits, respectively. Both clusters included traits from multiple domains and the majority of traits showed less than 3 loci colocalized with insomnia.

Next, we evaluated whether specific insomnia loci show similar colocalization patterns across 231 traits. To do so, we counted the number of shared colocalized traits for each pair of 282 loci and projected onto a 2D-map using tSNE (**Methods**). In the same way as done for the trait clustering, loci were clustered based on DBSCAN. We identified 11 dense clusters of loci (**Supplementary Fig. 6b**). These clusters were not driven by the location of loci on the genome as loci on the same chromosomes did not form a cluster (**Supplementary Fig. 6c**). Additionally, by projecting loci which contained prioritized genes that were part of the significantly associated gene-set (regulation of nervous system development), we did not observe a cluster specific to the gene-set (**Supplementary Fig. 6d-g**). On the other hand, clusters of loci were likely to be representing the specificity of colocalized traits. Three clusters of loci were mainly colocalized with metabolic traits (cluster #1, 3, and 9), and 3 other clusters of loci mainly with psychiatric traits (cluster #2, 4, and 6; **Supplementary Fig. 6b**). Other clusters also aggregated in trait domains, e.g. reproduction (cluster #8), cardiovascular (cluster #10 and 11), while 2 clusters were colocalized with multiple traits from multiple domains (**Supplementary Tables 19** and **20**). Although the colocalized loci

suggest there is a shared causal SNP between a pair of traits, this does not infer a causal relationship between traits. There are two possible scenarios when loci are colocalized: i) the same causal SNP is pleiotropic and directly causes both traits, ii) trait A causes trait B and GWAS on trait B captures causal SNPs of trait A or iii) misclassification of phenotypes across samples. It is statistically challenging to infer a causal relationship (and direction) between traits, because existing methods, such as Mendelian randomization, require strict assumptions[33,34] which are often violated. Nonetheless, we observed independent clusters of loci specifically colocalized with metabolic and psychiatric traits.

To further investigate whether the two groups of loci clusters that were colocalized mainly with metabolic or psychiatric traits had differential predictive value, we computed polygenic risk scores (PRS) using SNPs within the loci in each of the clusters (cluster #1, 3 and 9 for metabolic and #2, 4, and 6 for psychiatric). PRS was computed with PRSice for 3 sets of 10,000 target samples from UKB cohort used for genome wide PRS analysis (**Methods**). For each of the 3 sets of the target samples, we obtained the top and bottom 1, 5 and 10% subjects ranked by the PRS, combined across 3 datasets. To assess whether subjects with high PRS show different health outcomes compared to subjects with low PRS for each of the clusters of metabolic and psychiatric loci, we assessed predictive power for 3 phenotypes; overall health rating (field ID 2178), depressive symptoms (scored as sum of field ID 2050 and 2060 where individuals are coded 1 if the answer was "Several days", "More than half the days" or "Nearly every day", 0 otherwise) and body fat percentage (field ID 23099). We corrected for age, sex, array and the first 10 PCs and used residuals to perform two-sided Mann-Whitney U tests. Multiple testing (3 tested phenotypes and 3 PRS thresholds) was corrected for across all metabolic and psychiatric loci (0.05/9 = 5.6e-3). For PRS based on metabolic loci, the top 5 and 10% subjects showed a significantly lower overall health rating compared to the bottom 5 and 10% (p=3.5e-4 and 2.2e-6, respectively) (**Extended Data Fig. 3 and Supplementary Table 21**). In addition, the top 1 and 10% subjects showed significantly higher body fat percentage compared to the bottom 1 and 10% (p=1.2e-3 and 7.0e-5, respectively) while no difference was seen for depressive symptoms (**Extended Data Fig. 3 and Supplementary Table 21**). For PRS based on psychiatric loci, the top 5 and 10% subjects showed significantly lower overall health rating (p=5.3e-7 and 9.3e-7) and a significantly higher depressive symptom score (p=2.8e-4 and 4.9e-6) compared to the bottom 5 and 10%, while no difference was seen for body fat percentage (**Extended Data Fig. 3 and Supplementary Table 21**). These results suggest that there might be independent pathogenic mechanisms underlying insomnia that are related either to metabolic or to psychiatric traits.

13. Tissue, cell type and gene-set association analyses using the full GWAS results

To assess whether insomnia associated genes converge into biological functions, tissue types or specific cell types, using the full GWAS results (as opposed to prioritizing genes), MAGMA gene-property and gene-set analysis[35] was performed on 54 tissue types from GTEx v8[36], 54 lower and 106 higher resolution of brain regions from the Allen Human Brain Atlas (AHBA)[37], 885 brain cell types from 9 single-cell RNA-sequencing (scRNA-seq) datasets[38–44], and 6,089 gene-sets with at least 20 genes from MsigDB v7.0[45] and SynGO v1.0[46] (**Methods**). Bonferroni correction was applied across all tested items (0.05/7188=7.0e-6). In each dataset, pairwise conditional analyses were performed for significantly associated items and confounders[47]; tissues, cell types or gene-sets whose associations were mostly explained by a more significantly associated item within the same dataset are not reported (see **Methods** for details, full results are available in **Supplementary Tables 22-33**). For scRNA-seq datasets, we also performed cross-datasets conditional analyses as proposed previously[29] (see **Supplementary Note 18**).

Tissue analyses reveal association with cerebral cortex

Tissue specificity analysis with GTEx showed two groups of significantly associated brain regions; cerebellar hemisphere/cerebellum and cortex. The most significant association was seen with the cerebellar hemisphere (p=2.7e-19) (**Fig. 2** and **Supplementary Table 22**) in line with previous findings[12]. Previously identified basal ganglia also showed a significant association in the current study, however, this association was almost completely explained by the association of cerebellum (**Supplementary Table 23**). We note that the specificity of these associated brain regions was defined relative to the average expression of genes across all available tissue types in GTEx (i.e. brain and non-brain), and thus may reflect a general effect of genes expressed in brain. Indeed, when we conditioned on the average expression of genes across 13 brain regions, the significance of specific brain regions was largely decreased (**Extended Data Fig. 4a** and **Supplementary Table 22**). Therefore, these results using the GTEx resource implied a general association of insomnia with brain-specific expression and did not reveal enrichment in specific brain areas.

To gain insight into more specific brain regions associated with insomnia, we tested enrichment in 54 brain regions from AHBA (**Methods**). We observed the most significant associations with regions from the cerebral cortex followed by the basal forebrain and

preoptic region (**Fig. 2** and **Supplementary Tables 24 and 25**). We then assessed enrichment in 106 more specific brain regions from AHBA (**Methods**). Out of 106 regions, 32 regions, all from the cerebral cortex, showed significant associations with insomnia (**Fig. 2** and **Supplementary Table 26**). These associations were not independent and indicated a general association with cerebral cortex (**Supplementary Table 27**).

Cell type analyses reveal associations with neuronal cell types in sub-cortical regions

Next, we aimed to identify cell-type specificity using brain specific scRNA-seq datasets. We used 9 datasets[38–44] consisting of 885 cell types (**Methods**). The most significant association was seen in neurons from the lateral geniculate nucleus (LGN, p=9.9e-16) from DropViz which was collinear with GABAergic neurons from embryonic mouse midbrain dataset (p=2.3e-6). In addition, neurons in habenula, ventral pallidum (VP, p=2.2e-13) and anterior pretectal nucleus (APN, p=2.5e-11), all from DropViz, showed (partially) independent associations from LGN (**Fig. 2** and **Supplementary Table 28**). There were four additional neuronal cell types that reached significance from other datasets than DropViz, including the previously identified hypothalamus Vglut 2 neurons[12], however, these associations were largely dependent on the association of LGN (**Supplementary Tables 28-31**). Claustrum neurons from DropViz also showed a significant association, in line with previous findings[4] while its association was almost completely explained by the association of LGN (**Supplementary Table 29**). Association of the medium spiny neurons (MSN) was not replicated in the current study, yet we believe this is due to methodological differences explained in reference[48].

We note that although both AHBA and DropViz datasets contain samples from cortical and subcortical regions, AHBA showed more significant associations for insomnia with cortex while DropViz is showed stronger associations with cell types from subcortical regions close to the thalamus and globus pallidus. There are two possible reasons for this discrepancy. First, in the AHBA datasets, associations of cortical regions might be confounded by the true causal cell types which are not available in AHBA dataset. This is difficult to test since we do not know the true causal cell types, yet we do know that there are multiple specific neuronal cell types from the frontal and posterior cortex in DropViz that showed significant associations with insomnia whose associations were mostly explained by the LGN, which is not available in AHBA (**Supplementary Table 27**). Second, AHBA is based on microarray data which normally captures a lower number of genes and is limited by the number of probes used in the array compared to more robust RNA-seq. There were ~13,000 genes

available in AHBA while there were ~15,000 genes presented in DropViz. Of these ~11,000 genes were available in both datasets. By limiting to the overlapping 11k genes in the DropViz dataset, the significance of associations and effect size of LGN, habenula, VP, and APN showed a notable decrease. The remaining 4k genes that are available in DropViz but not in AHBA showed stronger associations with insomnia than the other 11k genes (**Extended Data Fig. 4b** and **Supplementary Table 29**). On the other hand, for AHBA, ~11,000 genes mostly explained the marginal association of cortical regions (**Extended Data Fig. 4b** and **Supplementary Table 29**). Thus, we found that the genes that are not available in AHBA contributed strongly to the associations of subcortical regions in DropViz which might explain the discrepancies and in particular the absence of associations with subcortical regions in AHBA.

Gene-set analyses of functional categories reveals enrichment in synaptic and neuronal development pathways

Gene-set analysis showed independent association with five gene sets; the most significant was 'process in the synapse' (SynGO:BP), followed by 'behavior' (GO:BP), 'synapse organization' (SynGO:BP), 'synapse part' (GO:CC) and 'regulation of neuron differentiation' (GO:BP) (**Fig. 2, Supplementary Tables 32** and **33**). A significant association of 'behavior' has been previously reported[4], while the associations of gene-sets related to functions of synapse and neuronal development are novel (**Fig. 2**).

14. Results of Fine-mapping

Statistical fine-mapping was performed for 554 risk loci identified from the sex-combined meta-analysis (**Methods**). The number of causal SNPs (k) was optimized at k=1 in 401 loci (72.4%) and in total 525 loci (94.8%) were optimized at $k \leq 5$ (**Supplementary Fig. 7a**). There were 15 loci that reached k=10 indicating that either the locus contained ≥10 independent causal SNPs or it was not optimized, likely due to a too complex structure of the local LD or subtle effect sizes. Despite the fact that the majority of loci were likely to contain a single causal SNP (as the k was optimized at 1), 112 loci resulted in a set of ≤10 credible SNPs (that are part of 95% credible sets, see **Methods** for details), and 66 loci contained sets of >100 credible SNPs indicating average PIP<0.01in those loci (**Supplementary Fig. 7b**). In addition, only 91 out of 554 loci contained at least one credible SNP with PIP>0.8 (**Supplementary Fig. 7c**), which increased to 166 and 426 loci by decreasing the PIP

threshold to 0.5 and 0.1, respectively (**Supplementary Fig. 7d-e**). A large proportion of loci was thus unsolved, indicating that for those loci the probability for a SNP of being causal was distributed across a relatively large number of SNPs within those loci. It has been previously shown that fine-mapping resolution can be improved using functional annotations as a prior[30], although such methods currently only support k=1 and careful selection of annotations to compute priors is required as they can bias the posterior probability. It is also possible that actual causal SNPs are not tagged in the current GWAS. As insomnia is one of the most polygenic traits and effect sizes of single variants are very small, the distribution of effect sizes and LD structure alone might not be sufficient to solve the statistical fine-mapping in some loci. Although PIP>0.8 is a widely employed threshold to select likely causal SNPs from fine-mapping, due to a low coverage of loci, we considered credible SNPs with PIP>0.1 for prioritization of genes from insomnia risk loci, so that each locus is allowed maximum of 10 times k credible SNPs.

15. Characteristics of high confidence genes identified in insomnia risk loci

Here we assessed whether there are differences in local genomic features such as size, number of SNPs and genes between loci with and without high confidence (HC) genes. We observed that the 281 loci with at least one HC gene had significantly more SNPs and genes, and larger locus size compared to the remaining 273 loci (two-sided Mann-Whitney U test, p=5.4e-3, 2.5e-26 and 4.6e-4, respectively). We also observed that the number of SNPs in the loci and the size of loci were significantly lower in the loci with HC genes identified by credible SNPs compared to GWS SNPs (two-sided Mann-Whitney U test, p=3.3e-10 and 1.2e-5, respectively), whereas the number of genes was not different between these types of loci (p=0.07). The difference in the number of SNPs and size of loci is because FINEMAP is less likely to identify credible SNPs (with PIP>0.1) in the loci with a relatively higher number of SNPs, and because GWS SNPs were used to identify HC genes in those unresolved loci.

We further assessed whether the number of identified HC genes depends on the local genomic features such as the size, the number of SNPs and genes in the loci. Obviously, if a locus only includes one gene, that gene is the most likely causal gene (i.e. HC-1), but in cases there are >1 gene, the number of genes present in a locus is obviously expected to influence the number of HC genes that can be detected. We indeed observed a significant correlation between the number of HC genes and the size (ρ=0.22, p=1.1e-7), the number of SNPs

(ρ=0.14, p=4.7e-5) and genes in the loci (ρ=0.62, p=2.4e-54, **Supplementary Fig. 8**). This suggests that, when there are more SNPs or genes to start with, a larger number of genes will be prioritized. As expected, the number of genes present in a locus showed the strongest correlation with the number of HC genes detected. Therefore, it is more difficult to pinpoint single HC genes from larger loci or loci with higher gene density, however, we believe that the HC-1 genes can be used to further narrow down potential causal genes from loci with a higher number of HC genes as described in the main text.

We tested the same correlations by separating loci with the HC genes prioritized by credible SNPs and by GWS SNPs. For loci with the HC genes prioritized by credible SNPs, the size (ρ=0.15, p=1.3e-3) and the number of genes in the loci (ρ=0.55, p=7.9e-32) showed a significant correlation with the number of the HC genes while the number of SNPs did not (ρ=0.08, p=0.09, **Supplementary Fig. 8**). This is because, when FINEMAP successfully identifies credible SNPs with relatively high PIP (>0.1 in this study), the number of SNPs to start with for the prioritization is much less than other loci without credible SNPs.

For loci with the HC genes prioritized by GWS SNPs, the size (ρ=0.27, p=6.2e-8), the number of SNPs (ρ=0.23, p=9.6e-6) and genes in the loci (ρ=0.73, p=4.3e-55) showed significant correlation with the number of HC genes (**Supplementary Fig. 8**). These results show that the local genomic features are more strongly correlated with the number of HC genes identified by GWS compared to credible SNPs, indicating narrowing down the potential causal variants using FINEMAP is effective for pinpointing likely causal genes from the locus.

16. Validation of gene prioritization

To validate whether the gene prioritization approach used in this study can identify known causal genes, we used 4 GWAS of 3 molecular traits (Urate, IGF-1, Testosterone male and Testosterone female) from Sinnott-Armstring *et al.*[49], and used "core" genes reported in the study as positive controls (**Supplementary Table 38**). We then performed the gene prioritization strategy to each of 4 GWAS as described in the main text. However, to also evaluate different types of linking genes apart from PPI, we also obtained gene co-expression for 53 tissues using GTEx v8[36] dataset and obtained 54 sets of HCP genes (based on PPI + 53 tissue specific co-expression). Gene co-expression was obtained by downloading gene TPMs (transcripts per million) from GTEx v8 (https://gtexportal.org/home/datasets), log-transformed with pseudocount 1 (i.e. log2(TPM+1)), then computed pairwise correlation of

genes across samples for each of 53 tissue types. We then obtained pairs of genes with absolute correlation coefficient > 0.8 as the co-expressed pairs. Enrichment of core-genes in HCP genes was tested by one-sided Fisher's exact test. We defined significant enrichment after Bonferroni correction ($p < 0.05/54 = 9.26e\text{-}4$). HC and HCP genes for each GWAS are reported in **Supplementary Table 39-42** and the results of core-genes enrichment analysis is reported in **Supplementary Table 43**.

Urate

From 246 loci, 699 HC genes were identified. With PPI, 167 HCP genes were prioritized and with gene co-expression dataset, on average 191 HCP genes were prioritized with maximum 354 HCP genes using gene co-expression in the kidney cortex (**Supplementary Table 39**). Although the enrichment of core-gene did reach significance after Bonferroni correction ($p<9.26e\text{-}4$), the HCP genes based on gene co-expression in the kidney cortex showed the lowest p-value followed by PPI (**Supplementary Table 43**).

IGF-1

From 417 loci, 1,158 HC genes were identified. With PPI, 286 HCP genes were prioritized and with gene co-expression dataset, on average 338 HCP genes were prioritized with maximum 602 HCP genes using gene co-expression in the nucleus accumbens basal ganglia in the brain (**Supplementary Table 40**). The HCP genes based on PPI showed the most significant enrichment of the core-genes, followed by the minor salivary gland, breast mammary tissue and pancreas (**Supplementary Table 43**).

Testosterone male

From 99 loci, 338 HC genes were identified. With PPI, 45 HCP genes were prioritized and with gene co-expression dataset, on average 67 HCP genes were prioritized with maximum 170 HCP genes using gene co-expression in the putamen basal ganglia in the brain (**Supplementary Table 41**). The HCP genes based on gene co-expression in the spleen showed the most significant enrichment of the core genes followed by atrial appendage of the heart, small intestine terminal ileum and adipose visceral omentum, while HCP genes based on PPI did not show enrichment of the core-genes (**Supplementary Table 43**).

Testosterone female

From 72 loci, 227 HC genes were identified. With PPI, 30 HCP genes were prioritized and with gene co-expression dataset, on average 49 HCP genes were prioritized with maximum 103 HCP genes using gene co-expression in putamen basal ganglia in the brain (**Supplementary Table 42**). The HCP genes based on small intestine terminal ileum showed the most significant enrichment of the core-genes, followed by bladder, PPI and adrenal gland (**Supplementary Table 43**).

In summary, the gene prioritization approach used in this study was able to prioritize a list of genes enriched in known core-genes. We also showed that significant enrichment of core-genes is in the HCP genes using gene co-expression in the trait relevant tissue types. At the same time, the HCP genes based on PPI also tended to show strong enrichment. Therefore, it is effective to use trait relevant tissues to prioritize HCP genes when it is known, while it is often unknown especially for highly polygenic traits like insomnia. In those cases, PPI would be a good alternative solution.

17. Additional results for prioritized genes from insomnia risk loci

To obtain more insight into the distribution of the 'high confidence prioritized' (HCP) genes across insomnia risk loci, we defined 2 additional sets of genes. Of 3,526 genes mapped by FUMA using GWS SNPs, genes from the loci which had at least one of the HCP genes are grouped as 'excluded' and genes mapped from other loci are grouped as 'unsolved'. This resulted in 2,122 excluded and 1,116 unsolved genes. Note that HCP genes are not necessarily a subset of 3,526 genes as credible SNPs are not required to have a genome-wide significant P-value. Indeed 1 of the 289 HCP genes was not part of the 3,526 genes. To investigate the difference between significant associations observed by full GWAS results and the HCP genes, we conditioned three sets of genes (i.e. HCP, unsolved and excluded genes) on gene-set based on expression in several brain regions (top 5 associations for AHBA high and low resolution), cell types (4 independent associations from DropViz) and gene sets (5 independently associated sets) that showed significant associations based on full GWAS results, using MAGMA. For brain regions (AHBA high and low datasets) and cell types (DropViz), the HCP genes showed a slight increase of association P-values compared to the marginal P-value suggesting at least some contribution of the HCP genes (**Extended Data Fig. 5**). The unsolved genes showed greater contributions than HCP genes which may be due to the greater number of genes in the conditioned gene set. On the other hand, excluded genes

showed the least contribution to those associations, even though it is the largest set of genes. At the same time, after conditioning either HCP or unsolved genes, there is still a substantially strong signal remaining (**Extended Data Fig. 5**). These results suggest that the absence of specific brain region and cell type associations when using the HCP genes is likely because those associations are highly polygenic where genes from unsolved loci and outside of GWAS loci are contributing, and the 289 HCP genes are not sufficient to explain these associations. For gene sets, a similar pattern was observed as brain regions and cell types, except 'behavior' (GO:BP) which did not show a notable decrease of the signal by conditioning any of the HCP, unsolved or excluded genes (**Extended Data Fig. 5**). This result suggests the association of the gene set is mainly driven by genes outside of the GWAS loci. Since the prioritization of genes is limited to GWAS loci, associations identified by full GWAS (which rely on MAGMA using the full GWAS outcomes) are missed when relying on selecting credible SNPs and genes from risk loci when indicating HCP genes.

18. Conditional analysis for associated tissues, cell types and gene-set

To identify independent association signals in gene-set or gene-property associations, we performed conditional analyses for sets that were significantly associated with insomnia per dataset (i.e., GTEx, AHBA low/high resolution, scRNA and gene-sets) in both the analyses based on the full genome-wide distribution (using linear regression as implemented in MAGMA) and the set of prioritized genes, using linear regressions, where we used tissue or cell type specific gene expression value as an outcome of a linear regression and binary status of genes as a predictor while conditioning on average expression and gene size.

To systematically identify independent and dependent associations of gene-properties (or gene-sets), and group them into clusters of independent associations, we set thresholds for conditional P-value and proportional significance (PS) of conditional P-value relative to marginal P-value. To simplify, we use cell type A and B for examples hereafter, while the same applies to tissue type or gene-set. Here we denote, p_A and p_B as the marginal P-values of cell type A and B where $p_A<p_B$, $p_{A,B}$ as the P-value of cell type A conditioning on cell type B, and PS_A as the proportional significance of cell type A conditioning on cell type B (computed as $-\log10(p_{A,B})/-\log10(p_A)$). We then defined the relationship between cell types A and B based on the decision rules as below (ordered according to the priority of the rules).

- $PS_A \geq 0.8$ and $PS_B \geq 0.8$: associations are independent.

- $p_{A,B} \geq 0.05$ and $p_{B,A} \geq 0.05$: both cell types are jointly associated and both are equally likely true signals (grouped into a same cluster).
- $PS_A \geq 0.5$ and $p_{B,A} \geq 0.05$: the association of cell type B is confounded by cell type A and cell type B is discarded.
- $PS_A \geq 0.5$ and $PS_B < 0.2$: the association of cell type B is mostly dependent on cell type A and cell type B is discarded.
- $PS_A \geq 0.2$ and $p_{B,A} \geq 0.05$: the association of cell type A is partially jointly explained by cell type B but the association of cell type B is confounded by A, therefore cell type B is discarded.
- $PS_A \geq 0.2$ and $PS_B < 0.2$: the association of both cell types are partially jointly explained while the association of cell type B is largely depending on cell type A, therefore cell type B is discarded.
- $PS_A \geq 0.5$ and $PS_B < 0.5$: same as above.
- $PS_A < 0.2$ and $PS_B < 0.2$: while there are still remaining signals by conditioning on each other, both cell types are largely jointly explained (grouped into a cluster).
- $PS_A \geq 0.2$, $PS_B \geq 0.2$, $p_{A,B} \geq 0.01$ and $p_{B,A} \geq 0.01$: although there are still small signal remaining in both cell types, associations of both cell types are considered as jointly explained (grouped into a cluster).
- $PS_A \geq 0.2$ and $PS_B \geq 0.2$: both cell types are partially jointly explained while there are remaining signals, therefore cell type A and B are grouped into separate clusters.
- When cell type A and B are collinear (in a case when regression fails), they are considered as joint associations and grouped into a cluster.

The conditional analyses were performed with the most significant sets until all cell types are denoted independent, clustered or discarded.

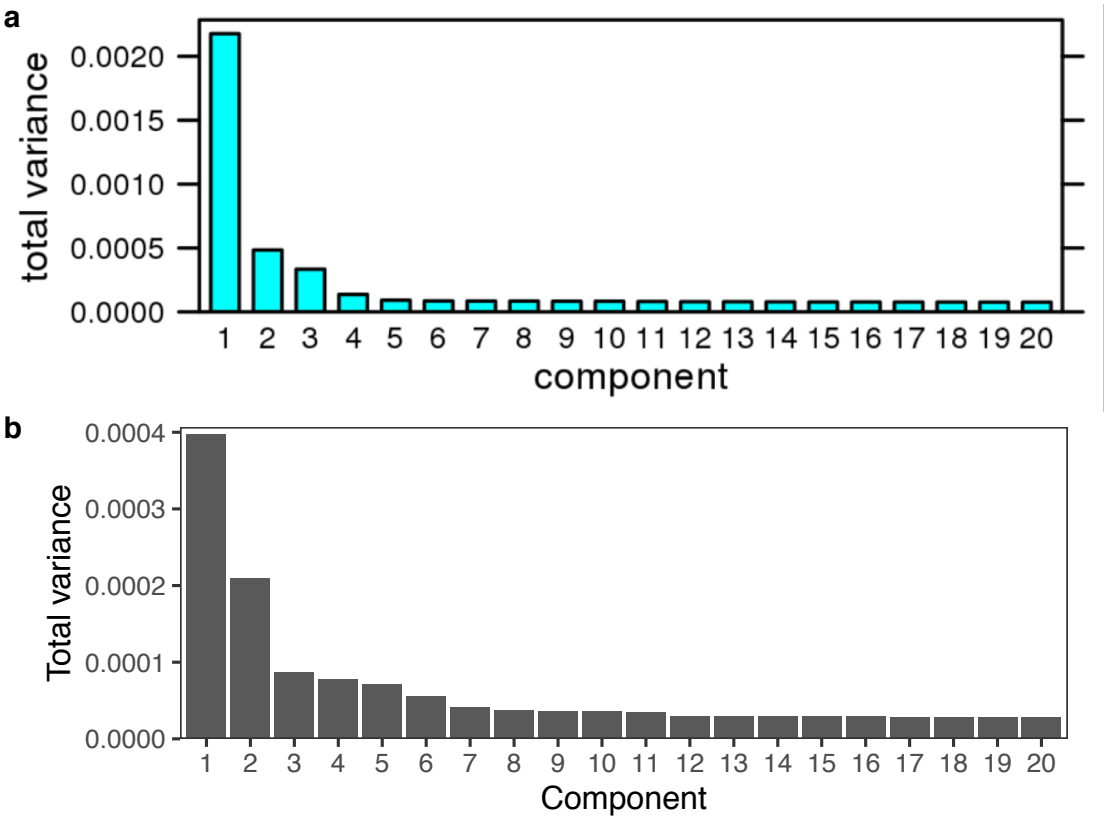
References

1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (2018).
2. Mccarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48,** 1279–1283 (2016).
3. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526,** 82–89 (2015).
4. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
5. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).
6. Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48,** 811–816 (2016).
7. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48,** 1284–1287 (2016).
8. Durand, E., Do, C., Mountain, J. & Macpherson, J. M. Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution. *bioRxiv* https://doi.org/10.1101/010512 (2014).
9. Cavalli-Sforza, L. L. The human genome diversity project: Past, present and future. *Nat. Rev. Genet.* **6,** 333–340 (2005).
10. International HapMap Consortium. International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
11. Bulik-sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).
12. 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).
13. 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).
14. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53,** 1097–1103 (2021).
15. Frei, O. *et al.* Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. *Nat. Commun.* **10,** 2417 (2019).
16. Holland, D. *et al.* Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLoS Genet.* **16,** e1008612 (2020).
17. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51,** 1339–1348 (2019).
18. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15,** 2759–2772 (2020).
19. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70,** 214–223 (2016).
20. Levey, D. *et al.* GWAS of Depression Phenotypes in the Million Veteran Program and Meta-analysis in More than 1.2 Million Participants Yields 178 Independent Risk Loci. *medRxiv* https://doi.org/10.1101/2020.0 (2020).
21. Stein, M. B. *et al.* Genome-wide association analyses of post-traumatic stress disorder and its symptom subdomains in the Million Veteran Program. *Nat. Genet.* **53,** 174–184 (2021).

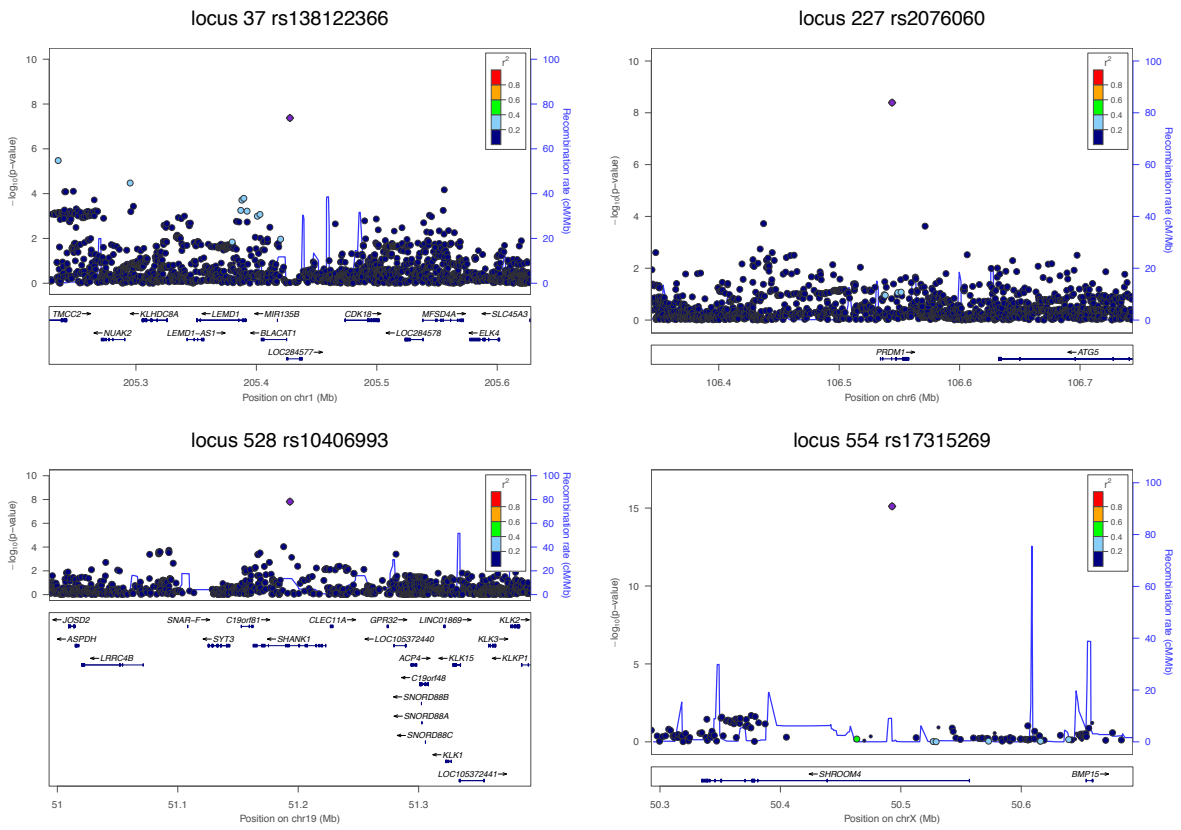
22. Zhou, H. *et al.* Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nat. Neurosci.* **23,** 809–818 (2020).
23. Nievergelt, C. M. *et al.* International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci. *Nat. Commun.* **10,** 4558 (2019).
24. Gelernter, J. *et al.* Genomewide Association Study of Maximum Habitual Alcohol Intake in >140,000 US European- and African-American Veterans Yields Novel Risk Loci Access. *Physiol. Behav.* **176,** 139–148 (2018).
25. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8,** 1826 (2017).
26. 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).
27. Lane, J. M. *et al.* Biological and clinical insights from genetics of insomnia symptoms. *Nat. Genet.* **51,** 387–393 (2019).
28. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet.* **24,** 1202–1205 (2016).
29. 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).
30. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50,** 1505–1513 (2018).
31. Han, B. *et al.* A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases. *Nat. Genet.* **48,** 803–810 (2016).
32. Didriksen, M. *et al.* Large genome-wide association study identifies three novel risk variants for restless legs syndrome. *Commun. Biol.* **3,** 703 (2020).
33. Evans, D. M. & Davey Smith, G. Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. *Annu. Rev. Genomics Hum. Genet.* **16,** 327–350 (2015).
34. Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *BMJ* **362,** k601 (2018).
35. 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).
36. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science (80-.).* **369,** 1318–1330 (2020).
37. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489,** 391–399 (2012).
38. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (80-.).* **348,** 1138–1142 (2015).
39. Romanov, R. A. *et al.* Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* **20,** 176–188 (2017).
40. Marques, S. *et al.* Oligodendrocyte heterogneity in the mouse juvenile and adult central nervous system. *Science (80-.).* **352,** 1326–1329 (2016).
41. La Manno, G. *et al.* Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167,** 566–580 (2016).
42. 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).
43. Saunders, A. *et al.* Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174,** 1015–1030 (2018).
44. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science (80-.).* **362,** eaat8464 (2018).
45. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27,** 1739–1740 (2011).
46. Koopmans, F. *et al.* SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron* **103,** 1–18 (2019).
47. de Leeuw, C. A., Stringer, S., Dekkers, I. A., Heskes, T. & Posthuma, D. Conditional and interaction gene-set analysis reveals novel functional pathways for blood pressure. *Nat. Commun.* **9,** 3768 (2018).
48. 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).
49. 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).

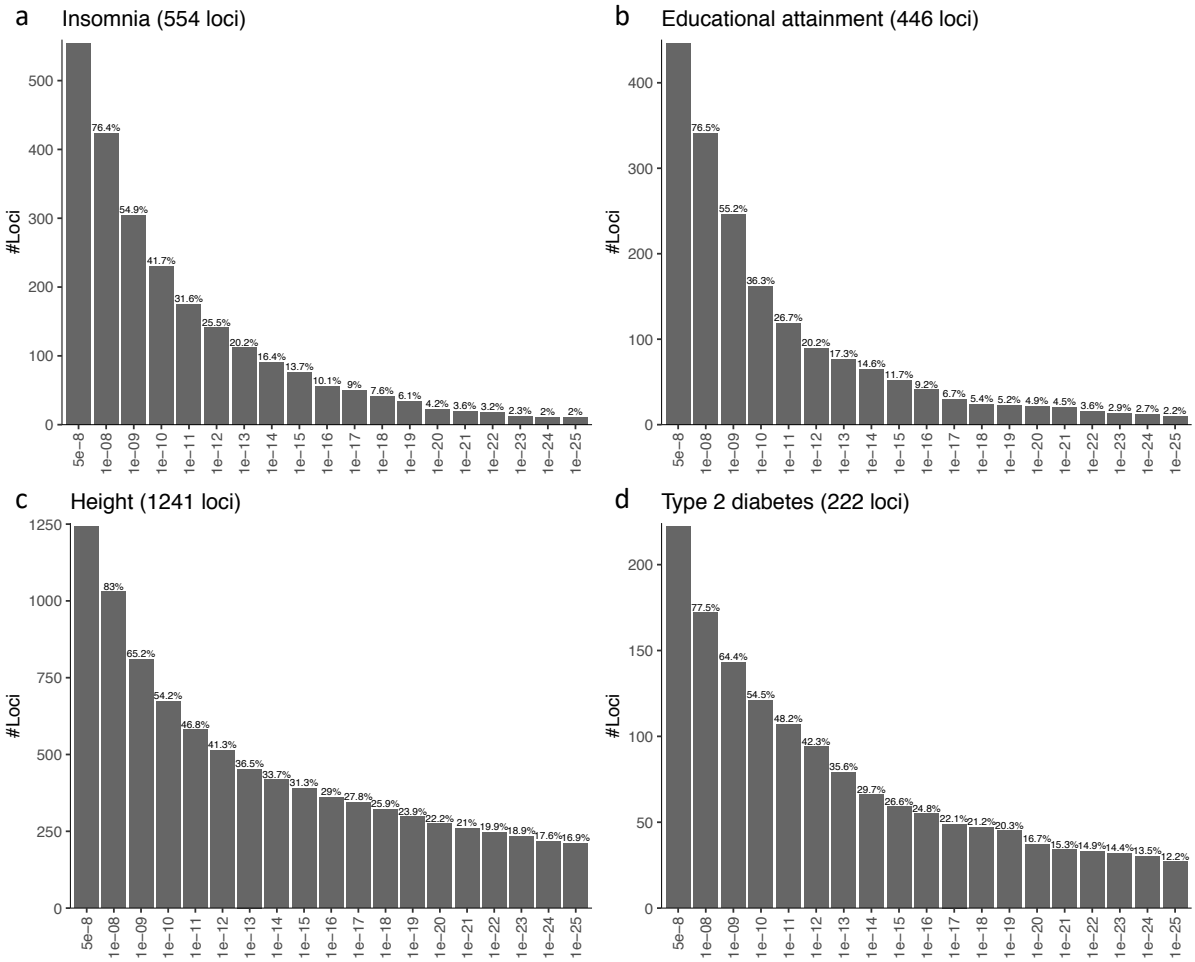
Supplementary Figures



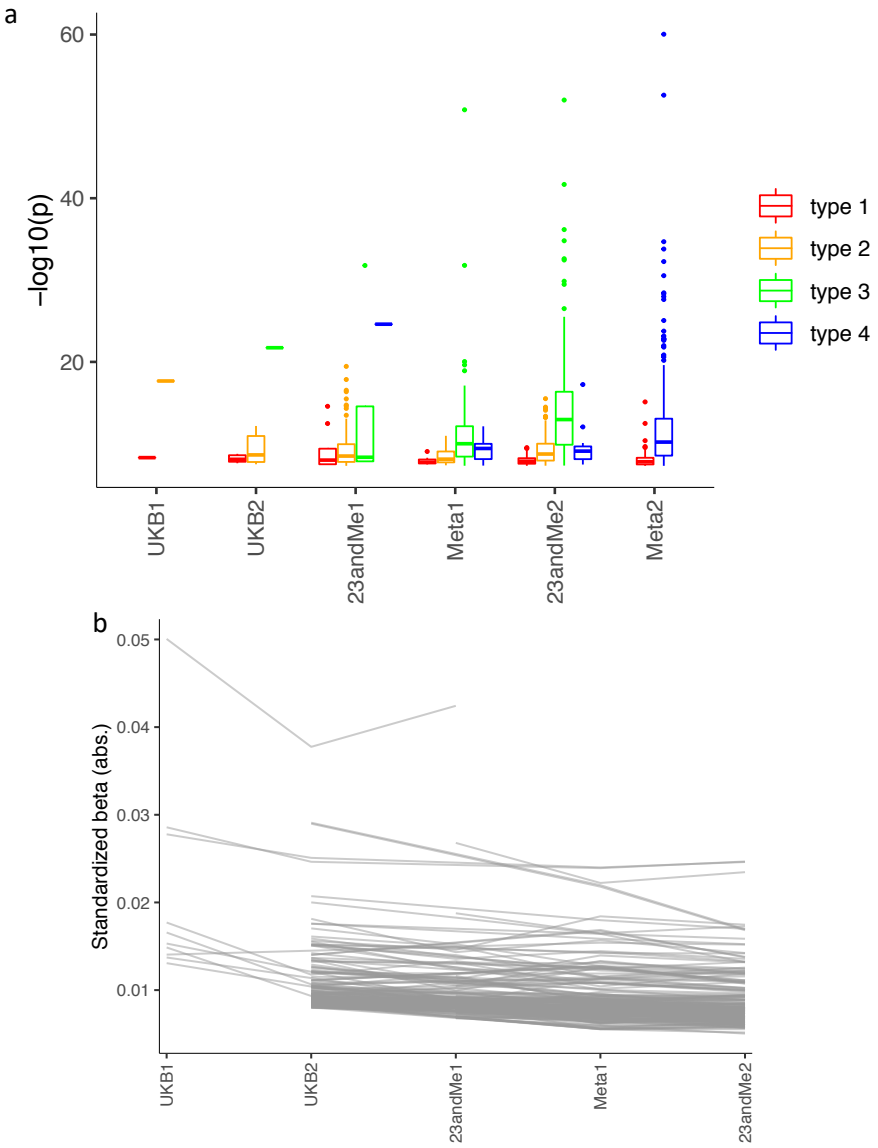
Supplementary Fig. 1. Total variance explained by genetic principal components. The principal components were computed for EUR samples of the 23andMe (a) and UKB (b).



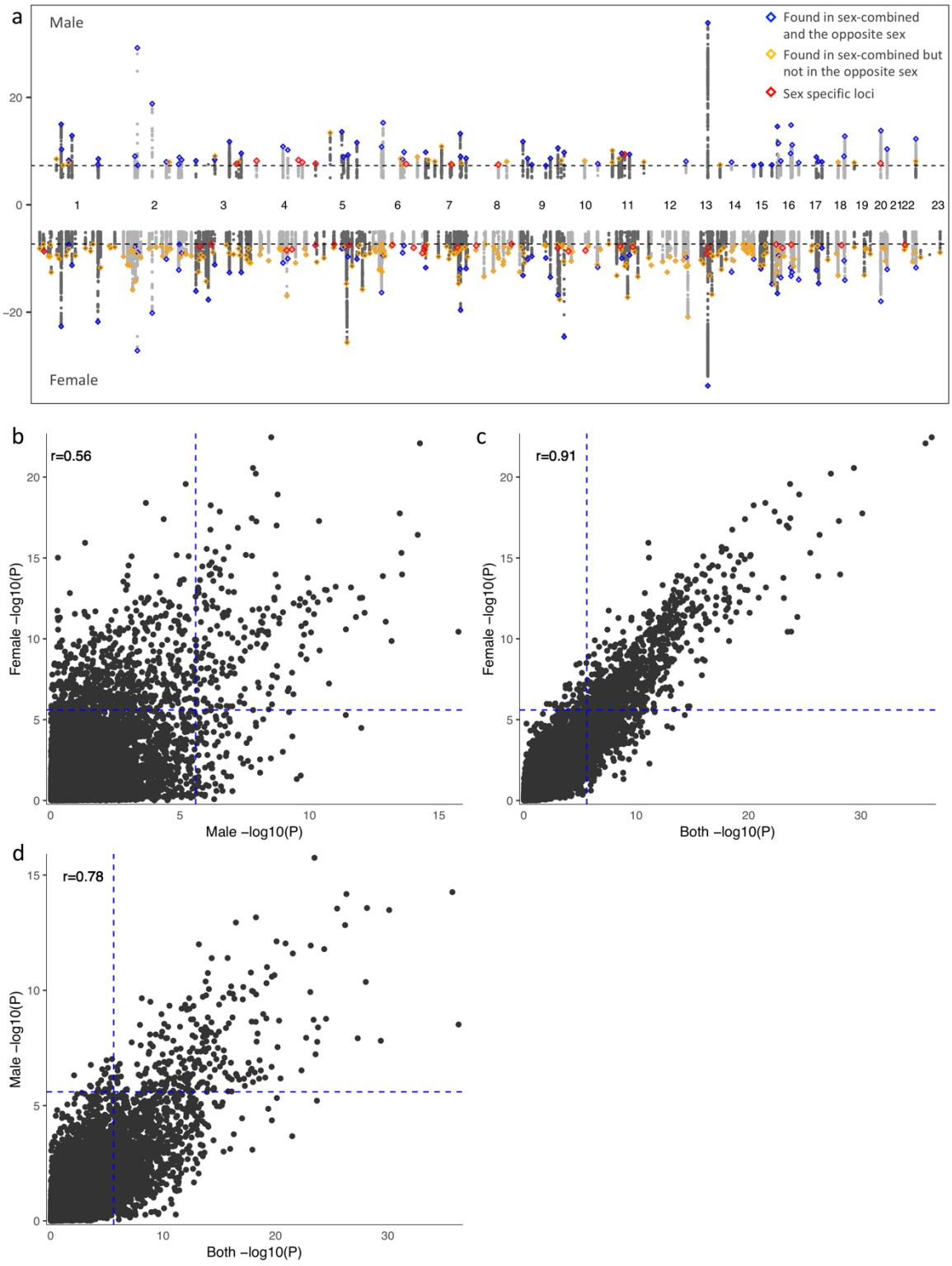
Supplementary Fig. 2. Locus Zoom plots for 4 suspicious loci.



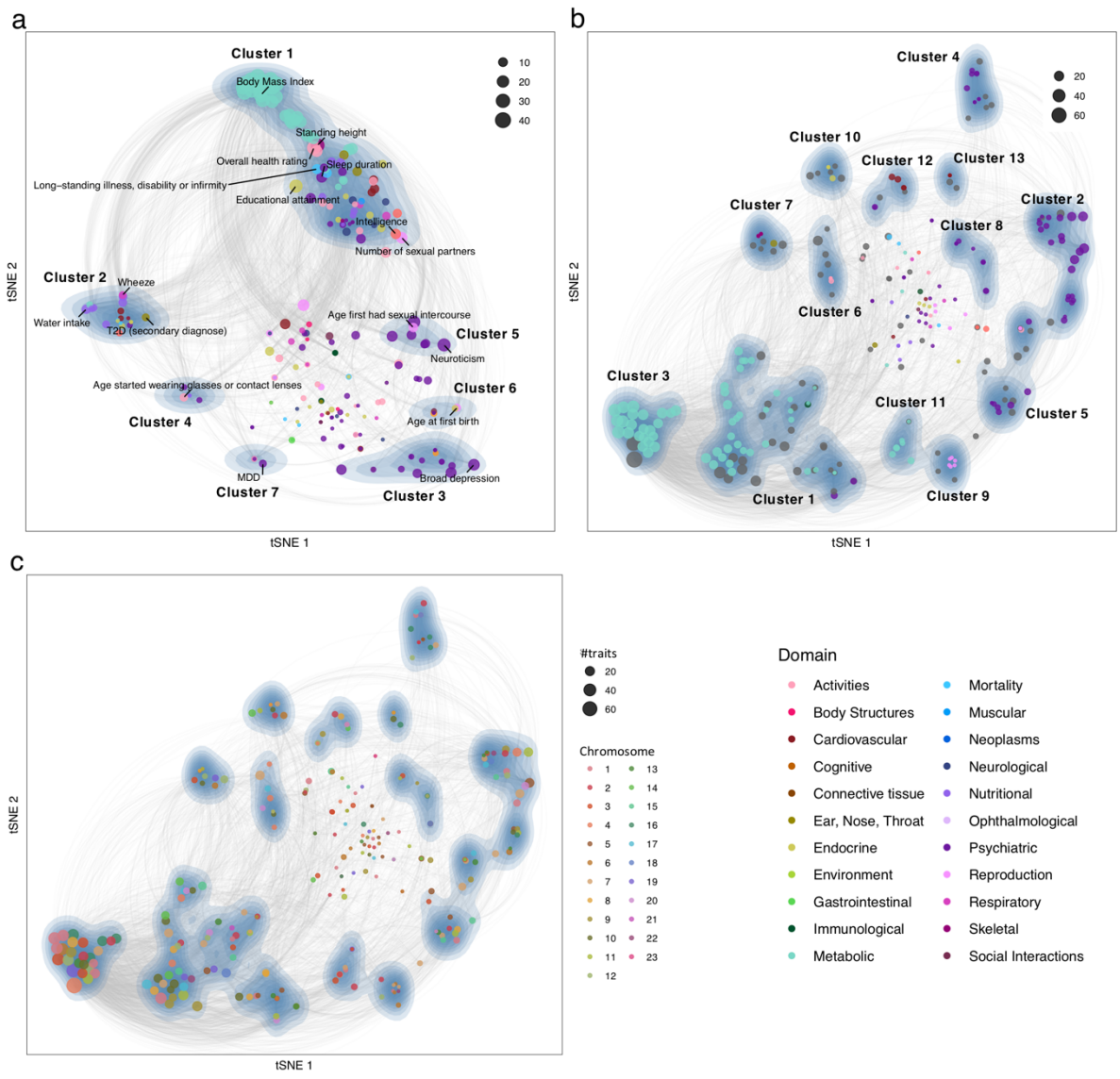
Supplementary Fig. 3. The number of risk loci with different P-value thresholds.

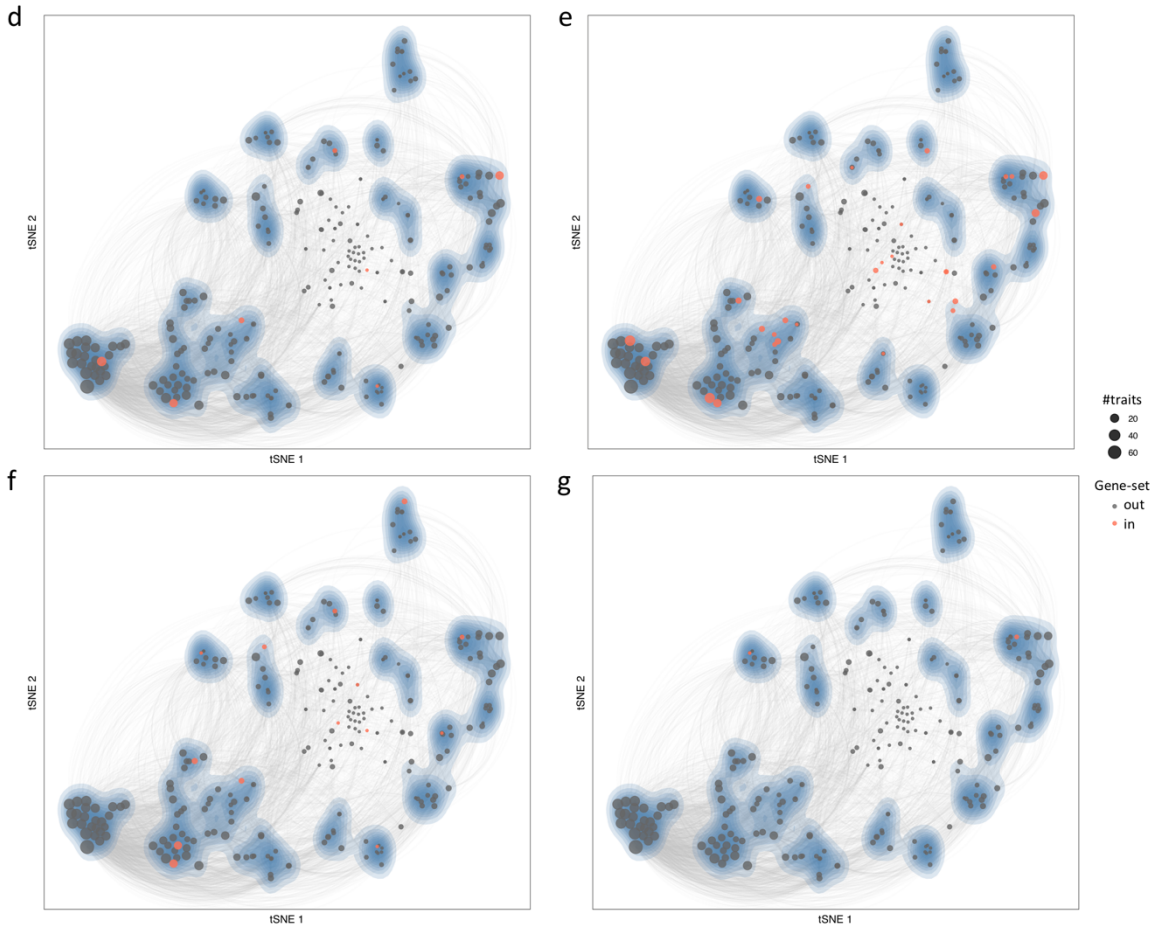


Supplementary Fig. 4. Comparison of significance and effect sizes across insomnia GWAS with different sample sizes. (**a**) Box and whiskers plot of -log10 P-value of the top SNPs from each risk locus from the corresponding GWAS (P-values are based on multiple linear regression for UKB1, UKB2, 23andMe1 and 23andMe2, two-sided Z-test for Meta1 and Meta2). Type 1: loci that are detected once but not replicated by any other GWAS (smaller or larger), type 2: loci that are detected first time and subsequently replicated by a larger GWAS, type 3: loci that were previously identified with smaller GWAS and replicated by a larger GWAS, type 4: loci that were previously identified with a smaller GWAS but not replicated by larger GWAS. 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 (loci) are 1 type 1 and 1 type 2 for UKB1, 5 type 1, 8 type 2 and 1 type 3 for UKB2, 8 type 1, 135 type 2, 4 type 3 and 1 type 4 for 23andMe1, 15 type 1, 75 type 2, 108 type 3 and 6 type 4 for Meta1, 52 type 1, 237 type 2, 177 type 2 and 11 type 4 for 23andMe2, 118 type 1 and 440 type 4 for Meta2. (**b**) Absolute standardized effect sizes of genome-wide significant SNPs. Only SNPs that are significant in at least 3 GWAS are displayed.

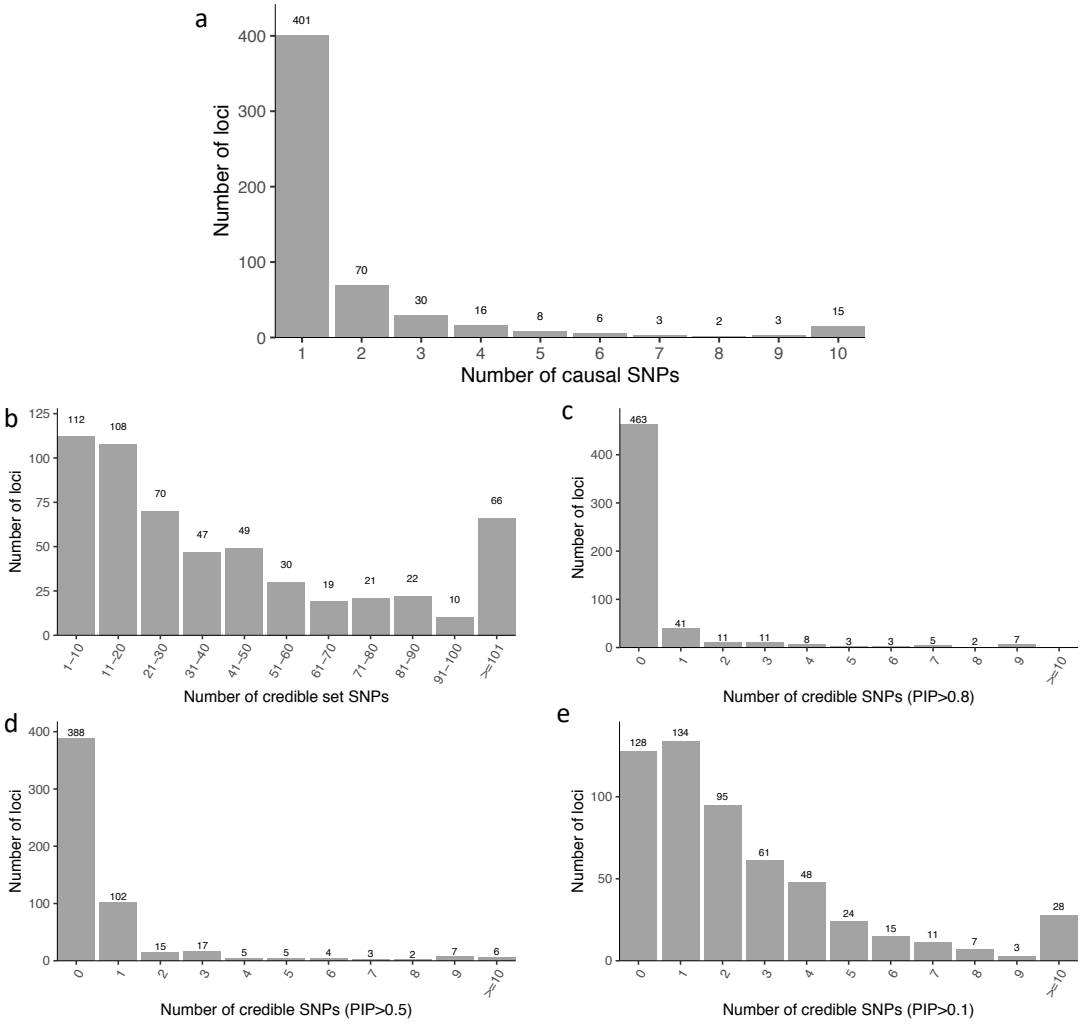


Supplementary Fig. 5. Comparison of sex-specific insomnia meta-analyses. (a) Manhattan plot for male only (top) and female only (top) summary statistics of meta-analysis where P-values were based on two-sided Z-test. SNPs with $p \geq$1e-5 are omitted. Horizontal dashed line represents genome-wide significance (p=5e-8). (**b-d**) Comparison of gene-based P-values computed by MAGMA gene analysis based on one-sided T-test for the regression coefficient of the gene expression. Blue dashed line represents the Bonferroni corrected P-value threshold (p=0.05/19751).

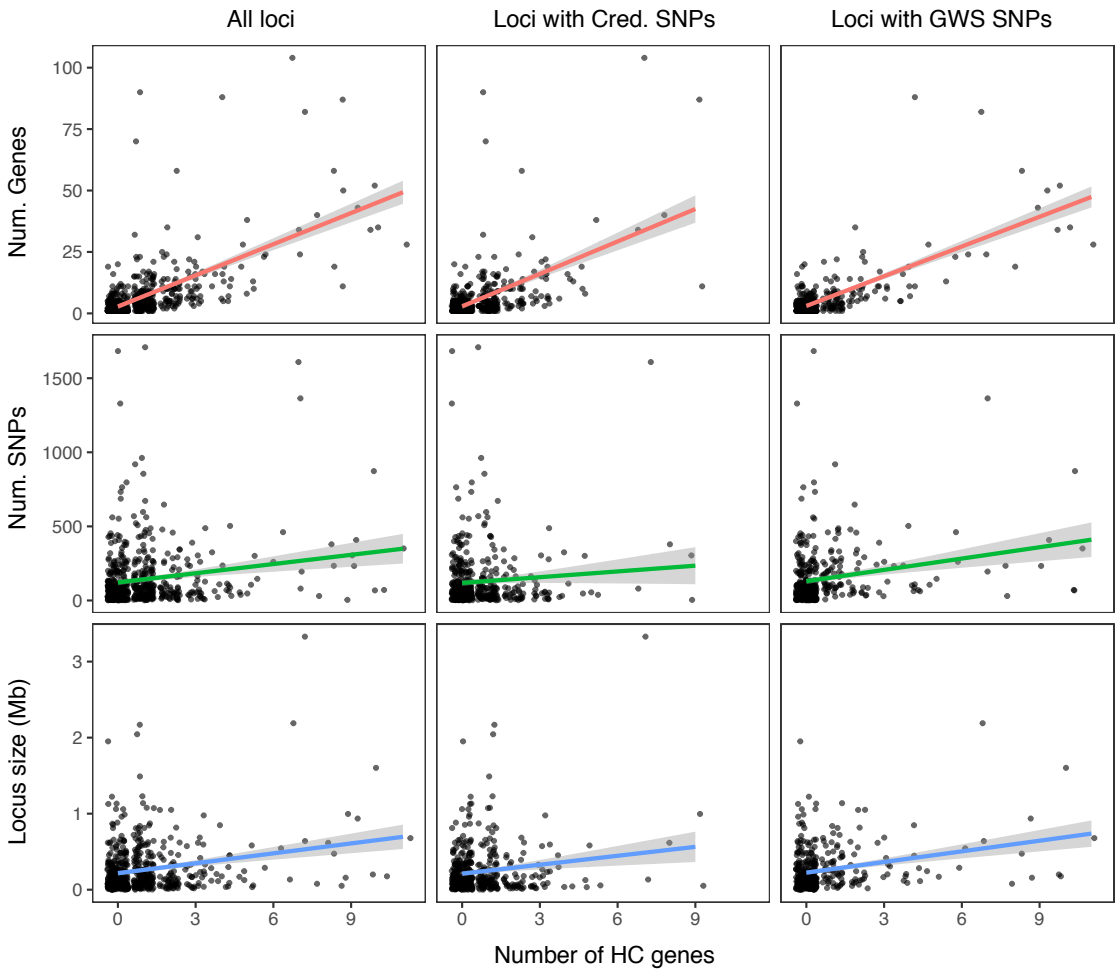

a
Cluster 1
Body Mass Index
Standing height
Overall health rating
Sleep duration
Long−standing illness, disability or infirmity
Educational attainment
Intelligence
Number of sexual partners
Cluster 2
Wheeze
Water intake
T2D (secondary diagnose)
Age first had sexual intercourse
Cluster 5
Neuroticism
Age started wearing glasses or contact lenses
Cluster 6
Cluster 4
Age at first birth
MDD
Broad depression
Cluster 7
Cluster 3
10
20
30
40
tSNE 2
tSNE 1
b
Cluster 4
Cluster 10
Cluster 12
Cluster 13
Cluster 7
Cluster 2
Cluster 8
Cluster 6
Cluster 3
Cluster 11
Cluster 5
Cluster 1
Cluster 9
20
40
60
tSNE 2
tSNE 1
c
tSNE 2
tSNE 1
#traits
20
40
60
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Domain
Activities
Body Structures
Cardiovascular
Cognitive
Connective tissue
Ear, Nose, Throat
Endocrine
Environment
Gastrointestinal
Immunological
Metabolic
Mortality
Muscular
Neoplasms
Neurological
Nutritional
Ophthalmological
Psychiatric
Reproduction
Respiratory
Skeletal
Social Interactions



Supplementary Fig. 6. Clusters of other traits and insomnia loci based on colocalization patterns. (**a**) t-SNE map of traits based on colocalization patterns across 554 insomnia risk loci. 227 traits that had at least one colocalized locus are displayed. Density maps indicate clusters of traits. Each data point represents a trait, colored by the domain and sized by the number of colocalized loci. Grey links between traits represent shared colocalized loci between traits from different clusters (within cluster links are omitted). (**b-g**) t-SNE map of insomnia loci based on colocalization patterns across 350 traits. 282 loci that were colocalized with at least one of the 350 traits are displayed. Density maps indicate clusters of loci. Each data point represents a locus and is sized by the number of colocalized traits. Each locus is colored by the domain of traits which account for >50% of traits colocalized with the locus, otherwise colored in grey (when none of the trait domain represent >50% of the colocalized traits) (**b**), chromosome (**c**) or significantly enriched gene-sets ((**d**) modulation of chemical synaptic transmission (SynGO:BP), (**e**) neuron differentiation (GO:BP), (**f**) regulation of trans synaptic signaling (GO:BP), (**g**) inclusion body (GO:CC)). Full results are available in **Supplementary Tables 15-19.**



Supplementary Fig. 7. Summary of fine-mapping results. (**a**) Distribution of the estimated number of causal SNPs per locus. (**b**) Distribution of the number of credible set SNPs (that are part of 95% credible sets). (**c-e**) Distribution of the number of credible SNPs with PIP>0.8 (**c**), >0.5 (**d**) and >0.1 (**e**).



Supplementary Fig. 8. Number of high confidence (HC) genes and local genomic features. "Loci with Cred. SNPs" include loci with no HC genes and loci with HC genes identified by credible SNPs (with PIP>0.1). "Loci with GWS SNPs" include loci with no HC genes and loci with HC genes identified only by genome-wide significant (GWS) SNPs. The linear lines indicate fitted linear model, grey shades indicate 95% confidence interval based on number of data points 544, 178 and 103 for all loci, loci with cred. SNPs and loci with GWS SNPs, respectively.