natureresearch

Peer Review Information

Journal: Nature Genetics
Manuscript Title: Genome-wide meta-analysis of insomnia prioritizes genes associated with metabolic and psychiatric pathways
Corresponding author name(s): Professor Danielle Posthuma

Transferred manuscripts - This manuscript has been previously reviewed at another journal. This document only contains reviewer comments, rebuttal and decision letters for versions considered at Nature Genetics.

Reviewer Comments & Decisions:

Decision Letter, initial version:

3rd Feb 2021

Dear Professor Posthuma,

Your Article, "Genome-wide meta-analysis of insomnia in over 2.3 million subjects indicates a role for specific biological pathways through gene-prioritization" has now been seen by 2 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

As you will see from these comments, both referees think that this is impressive work; but the biological insight would need to be improved. You may explore new concepts or develop new analytic methods to try to get new insight into the insomnia subtypes/heterogeneity. We hope that you will find the prioritized set of referee points to be useful when revising your study.

We therefore invite you to revise your manuscript taking into account all reviewer comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a “Response to referees” document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
<a href="http://www.nature.com/ng/authors/article_types/index.html">here</a>.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our <a href="https://www.nature.com/nature-research/editorial-policies/image-integrity">guidelines on digital image standards.</a>

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

<strong>Note:</strong> This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within twelve to sixteen weeks. We will be highly flexible about the timelines. If you need more time to prepare the revisions, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as ‘corresponding author’ on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on ‘Modify my Springer Nature account’. For more information please visit please visit <a href="http://www.springernature.com/orcid">www.springernature.com/orcid</a>.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Wei

Wei Li, PhD
Senior Editor
Nature Genetics
1 New York Plaza, 47th Fl.
New York, NY 10004, USA
www.nature.com/ng

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
In this manuscript Watanabe et al. performed a meta-analysis of genetic variants associated with insomnia using data from 23andMe and the UK Biobank (N~2.3 mil). They identify 554 risk loci and prioritize 289 genes in those loci. Overall the same biology as previously identified in insomnia GWAS remains supported and some new genes are raised as candidates for functional studies. The methods are appropriate for this study. At times the manuscript reads a bit dense and could be slightly better organized (I think a few section headers are missing??). My major concern is the missing discussion surrounding the phenotypic measurement heterogeneity between 23andMe and UKBB. Despite an increase of 1mil samples, the % of phenotypic variance and % of heritability explained both appear to be less than the 1.3 mil GWAS. It appears an important opportunity to discuss phenotypic measurement heterogeneity vs. sample size has been missed [noise vs. power].

Comments:
1. For the UK Biobank based GWAS, unrelated individuals only were used. Can the authors please explain why this subset only was used?
2. To address the heterogeneity in the phenotype definition of insomnia (low genetic correlation between UKBB and 23andMe GWAS (0.66)), it would be interesting to see the genetic correlation changes using different definitions of insomnia (eg. only including cases from single questions in 23andMe or the Lane et al. definition of insomnia in UKBB). "These results suggest that the items used in both cohorts are good predictors of insomnia, although what they ascertain is slightly different phenotypically." can be addressed with this line of investigation (and perhaps the aspect of insomnia captured is not different just more heterogenous in ascertainment using a variety of questions in 23andMe to define a case).
3. In the intro the authors state "Together, these loci explained around a quarter of the estimated heritability of insomnia [Jansen et al.]" and later "We estimated that current GWS SNPs explain 17.3% of the total h2 136 SNP". As well the % variation explained in this META analysis appears to be lower than the Jansen et al. study. Could the authors please discuss why this might be?
4. In the 23andMe GWAS only 5 PCs of ancestry are used as compared to the 10 PCs of ancestry in the UK Biobank. Do the PCs in 23andMe capture more study variation than those in the UK Biobank?
5. Seeing as an RLS gene (PTPRD) is the most significantly associated gene - did you perform any sensitivity analyses to address the RLS/Insomnia overlap seen in the current and prior GWAS of insomnia?
6. From the clustering of colocalized loci (line 209), did you observe any enrichment for SNPs identified in the UKBB GWAS vs. the 23andMe GWAS suggesting different components of insomnia

captured by the two studies?
7. For the prioritization of high confidence genes, as a benchmark, how often is the high confidence gene the nearest gene to the SNP? And is the # of SNPs in the locus related to the probability of identifying a high confidence gene?
8. Does utilizing PPI data to identify HC-m genes related to HC-1 genes introduce a source of bias, as this clearly won't identify genes that haven't been extensively studied? How many genes have PPI data available?
9. Through the manuscript, please be consistent and use "GWAS" as opposed to "GWASs".
10. The MVP cohort is largely male, perhaps this is a source of different between the cohorts and should be discussed.
11. In Supp Table 3 it would be helpful to report the effect as an odds ratio given these are case control GWAS.
12. In Supp Table 4 it is unclear what is meant by "suspicious loci".
13. When discussing replication of previous loci (lines 167-170) it should be noted that prior studies are NOT independent of the current study - therefore it is unsurprising that Jansen et al. SNPs replicate in a study that expands on the prior sample.
14. The paragraph on enrichment of SNPs be genomic region (lines 175-187) can be condensed and supp tables referred to, as these results are common for nearly all GWAS.
15. The discussion section could elaborate more on how the insights into insomnia biology in this study related to prior knowledge (do we see the same, what's a novel biological insight??).

Reviewer #2:
Remarks to the Author:
In this manuscript, Watanabe et al. describes the genetic architecture of insomnia in 2.3 million individuals and propose a specific pipeline for gene prioritization, identifying enrichment in genes involved in synaptic function and neuronal differentiation.

The effort is impressive, and the catalogue of genes identified is large. The statistics are state of the art. The findings are solid but at the end, rather disappointing (involvement of neuronal genes in the phenotype, high correlation with depression) and a mere extension of prior studies. It is a technical paper and not much novel insight is gained in term of understanding the neurobiology or phenotype of insomnia.

For example, with such a huge sample, one wonder why the authors did not attempt to look at different aspects of insomnia, such as difficulties falling asleep versus waking up, or look at age dependency of genetic effects, or try to look at more focused hypothesis with regard to the potential role of circadian factors, depression, anxiety or other sleep disorders in insomnia.

There is a brief mention that clustering of the loci based on co-localization across traits suggest locus heterogeneity and distinct clusters of loci, one linked with metabolic traits the other with psychiatric traits, but it is not further discussed; there is no attempt at furthering the significance of this statement. As it is, it is a straightforward analysis showing limited creativity.

The pipeline is well explained and generally solid. The use of protein-protein interactions (PPI) (in Web database) of HCPm with HCP1 as the main method used to prioritize higher confidence gene within the HC-m is however also not really well justified. The author mentioned it would also be possible to use other prioritization methods, such as spatial or temporal co-expression, but they do not show that these methods yield similar results (line 272). It seems that the validation of using PPI should be done first in diseases where pathophysiology is better understood than insomnia, such as autoimmune diseases, specific neurological disorders or cardiovascular diseases. Is there evidence of this?

In brief, it is a nice technical paper but it is hard to find it an original and exciting study.

Minor

Issues re reproducibility across subsamples, heritability and inflation/polygenicity are well-discussed, but at time this information is repetitive (lines 118, 199, 135).

Author Rebuttal to Initial comments

Below we provide a point-by-point response indicating the changes made to the manuscript. For clarity, our response is written in blue.

In addition to modifications on the manuscript regarding the reviewers' comments, we also updated the results of the colocalization analysis (insomnia loci with 350 other traits) as we noted that some loci were unintentionally left out of the analysis due to a small bug. We have updated Supplementary Fig. 8 and Supplementary Tables 14-18. The conclusions remain unchanged. .

While revising our manuscript we also noted that the target samples used for the polygenic risk score analyses had some variants included that should have been filtered out. We have now excluded variants with MAF < 0.01, missing rate > 0.1 and Hardy-Weinberg equilibrium P-value < 1e-6. We updated Supplementary Fig. 2. The result now shows that the insomnia GWAS meta-analysis explains at most 2.46% of the phenotypic variance (previously this was 2.03%).

We thank the reviewers for evaluating our manuscript.

Reviewer #1:
Remarks to the Author:
In this manuscript Watanabe et al. performed a meta-analysis of genetic variants associated with insomnia using data from 23andMe and the UK Biobank (N~2.3 mil). They identify 554 risk loci and prioritize 289 genes in those loci. Overall the same biology as previously identified in insomnia GWAS remains supported and some new genes are raised as candidates for functional studies. The methods are appropriate for this study. At times the manuscript reads a bit dense and could be slightly better organized (I think a few section headers are missing??). My major concern is the missing discussion surrounding the phenotypic measurement heterogeneity between 23andMe and UKBB. Despite an increase of 1mil samples, the % of phenotypic variance and % of heritability explained both appear to be less than the 1.3 mil GWAS. It appears an important opportunity to discuss phenotypic measurement heterogeneity vs. sample size has been missed [noise vs. power].

We thank the reviewer for taking the time to evaluate this manuscript and providing valuable comments. We have updated the manuscript regarding the reviewer's concern and added a discussion on the possible heterogeneity between samples. Please see below for the detailed responses to this and other remarks.

Comments:
1. For the UK Biobank based GWAS, unrelated individuals only were used. Can the authors please explain why this subset only was used?

We used only unrelated individuals for the UKB cohort mostly because of practical reasons; analysing related individuals requires a different kind of quality control and analysis than used for unrelated individuals, and since the number of related individuals was relatively small compared to the total sample size, we assumed excluding related individuals would lead to significantly different results.

However, to test this assumption, we have now also performed the GWAS for the UKB cohort including related samples using REGENIE (Mbatchou et al. 2021). This added 21,629 cases and 51,150 controls from related individuals resulting in a total 131,177 cases and 328,590 controls for the UKB cohort (although the effective sample size will be lower due to the relatedness).
We found that the genetic correlation between the GWAS of the UKB with and without related samples (taking into account the sample overlap) was 1.00, the genetic correlation between UKB with related samples and 23andme was had the same point estimate (0.66) as between UKB without related samples and 23andme.
The meta-analysis of 23andMe and UKB including related samples resulted in 573 loci , of which 56 loci did not overlap with the 554 loci that were found when not including the related samples. These 56 loci were just below threshold in the sample including related samples and just above threshold in the sample not including related samples. Given these minor changes in overall results, we believe including the related samples at this point would not significantly change our main results and conclusions, while it would entail updating all numbers, tables and figures, with just slightly altered values. We have therefore decided not to rerun all post-GWAS analyses and currently do not include the results based on the GWAS that includes related individuals. However, we are willing to do so in a next revision if the reviewer thinks this is necessary.

2. To address the heterogeneity in the phenotype definition of insomnia (low genetic correlation between UKBB and 23andMe GWAS (0.66)), it would be interesting to see the genetic correlation changes using different definitions of insomnia (eg. only including cases from single questions in 23andMe or the Lane et al. definition of insomnia in UKBB). "These results suggest that the items used in both cohorts are good predictors of insomnia, although what they ascertain is slightly different phenotypically." can be addressed with this line of investigation (and perhaps the aspect of insomnia captured is not different just more heterogenous in ascertainment using a variety of questions in 23andMe to define a case).

This is a nice suggestion. First to clarify - the definition of insomnia as used in the Lane et al. paper includes dichotomizing the question ‘Do you have trouble falling asleep at night or do you wake up in the middle of the night?’. Possible answers were ‘usually’, ‘sometimes’, ‘never/rarely’, and ‘prefer not to answer’. Lane et al. defined cases as those who answered ‘usually’, and controls as those who answered ‘never/rarely’, other answers were excluded. In our analyses, those with ‘sometimes’ were set to controls, as these participants do not seem to experience regular and chronic complaints. We chose this definition based on the maximal discriminative accuracy in an independent dataset of individuals diagnosed with insomnia disorder (Dutch Sleep Registry, see Hammerschlag et al. 2017).

To address this comment, we used a phenotype definition where individuals who answered ‘sometimes’ were excluded from the analyses (e.g. control subjects experiencing no complaint at all). We found that the rg using this definition in UKB with 23andMe was 0.6595 compared to an rg of 0.6552 in our original analysis. This suggests that excluding individuals who answered ‘sometimes’ in the UKB cohort does not account for explaining heterogeneity between the cohorts.

In addition, we calculated the predictive accuracy of this phenotype definition in the Dutch Sleep Registry. This showed an increase of sensitivity (1.00) and a drop in specificity to 0.92, compared to a sensitivity of 98% and specificity of 96% in our previous phenotype definition (see Hammerschlag et al. 2017). In brief, the criterion for including controls only marginally affects accuracy, and is slightly lower (accuracy: 0.96) compared to our previous definition (accuracy: 0.97).

Although we performed GWAS using alternative definitions as requested by the reviewer, for the main analyses we feel we have to stick to our originally selected definition which was based on results from an independent sample as described above, and was not driven by any knowledge of current results.

3. In the intro the authors state "Together, these loci explained around a quarter of the estimated heritability of insomnia [Jansen et al.]" and later "We estimated that current GWS SNPs explain 17.3% of the total h2 136 SNP". As well the % variation explained in this META analysis appears to be lower than the Jansen et al. study. Could the authors please discuss why this might be?

We acknowledge that this sentence was badly formatted and misleading. The first sentence in the introduction was supposed to describe the fact that the phenotypic variance explained by polygenic risk scoring (at most 2.6%, on average 1.9% across 3 target sets) is about a quarter of the total SNP heritability (7%). The second sentence is a result based on MiXeR which estimates the proportion of total SNP heritability (7.2%) that is explained by just the genome-wide significant SNPs (17.3%). Therefore these facts are not comparable. We have updated the introduction to avoid this confusion.

Introduction (line 58-)

A recent GWAS based on a sample size of >1.3 million individuals reported over 200 genomic loci linked to the risk of insomnia in which the polygenic risk score explained about a quarter of the estimated heritability[6]*, implicated the involvement of several neurobiological processes, cell types, brain areas, and circuitries in insomnia, and showed considerable overlap with genetic risk factors for psychiatric disorders*[6,7]*.*

The reviewer was however correct in noting that the % variation explained in this META analysis (2.5%) appears to be lower than as reported in the Jansen et al. study (2.6%). To sort out the reason for this, we reran PRS analyses using the current and Jansen meta-analysis as discovery and using the target samples from the current and the Jansen’s study. Results are summarized in the table below (full results are available in **Supplementary Table 2**). We observed that, regardless of the target set, the prediction power generally increased with the increased sample size of the discovery GWAS, and that the observed difference in explained variance was due to fluctuations as a result of the target sample. These results show that, even though the current sample may include more heterogeneity as pointed out by this reviewer and 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 for prediction.

Note that we updated the results of the current PRSice analysis as we noted the genotype dataset for the target samples were not property QCed (i.e. we accidentally did not filter on MAF, missing rate, and HWE). We also note that the results based on the target samples used in the previous study (Jansen et al.) are different from reported. This is mainly due to the different release of the UKB genotype dataset since the analysis was performed. In addition, we report R^2 adjusted for ascertainment with population prevalence of 0.3 which is recommended by Choi et al. while it was not corrected in the previous study.

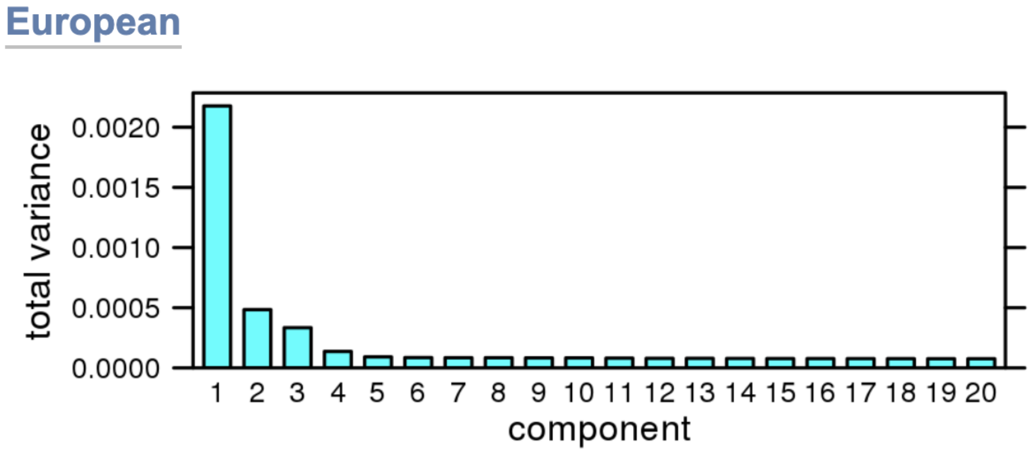
	Discovery	
Target	Jansen et al. ~1.3 million	This study ~2.3 million
Jansen et al. 3k	2.83%	2.87%
This study 10 k	2.05%	2.46%

Supplementary Note 2 (line 84-)
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[4] (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.[5] while it was not corrected in the previous study.

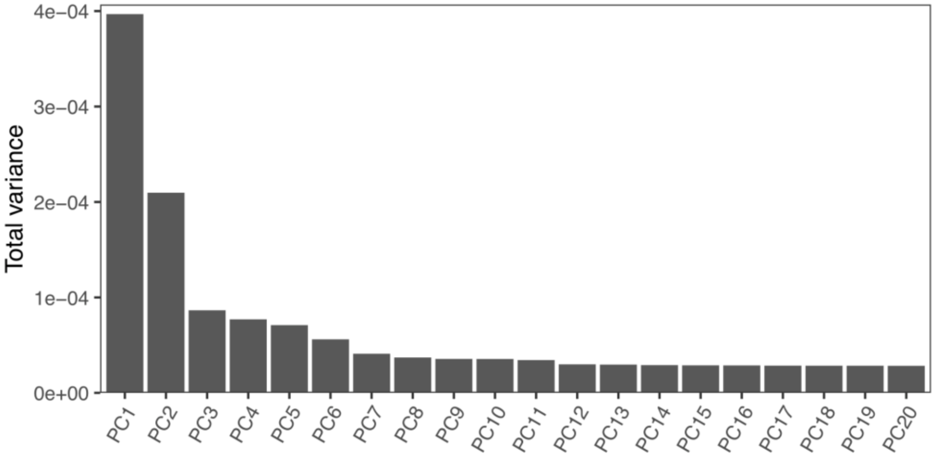
4. In the 23andMe GWAS only 5 PCs of ancestry are used as compared to the 10 PCs of ancestry in the UK Biobank. Do the PCs in 23andMe capture more study variation than those in the UK Biobank?

Past internal analyses for 23andMe have indicated that the first 5 PCs are sufficient as additional PCs did not result in a further reduction of genomic control inflation. As shown in the plots, the 1st PC of 23andMe explains more than 5 times of the variance explained by the 1st PC of the UKB. In addition, the variance is almost flat after the 5th PC for the 23andMe while the plateau is reached after 10th PC in the UKB.

Variance explained by PCs for 23andMe



Variance explained by PCs for UKB



We added these figures to **Supplementary Fig. 1**.

5. Seeing as an RLS gene (PTPRD) is the most significantly associated gene - did you perform any sensitivity analyses to address the RLS/Insomnia overlap seen in the current and prior GWAS of insomnia?

We performed BUHNMBOX (Breaking Up Heterogeneous Mixture Based On Cross-locus correlation) to test whether association of PTPRD and other RLS associated genes identified in this study are identified due to subset of RLS cases within insomnia cases. Results suggest that the association of the most significant gene PTPRD is not likely to be driven by the presence of RLS cases within insomnia cases.

Main text (line 189-)
*Results show that the association of PTPRD is not likely to be driven by a misclassification or comorbidity of RLS within the insomnia cases (**Methods**, **Supplementary Table 8** and **Supplementary Note 6**).*

Methods (line 660-)

*To perform BUHMBOX [ref], from the largest RLS meta-analysis [Didriksen et al. 2020], 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 8**). 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.*

Supplementary Note 6 (line 184-)

*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)[18] (**Methods**). With 22 RLS SNPs[19], 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.*

6. From the clustering of colocalized loci (line 209), did you observe any enrichment for SNPs identified in the UKBB GWAS vs. the 23andMe GWAS suggesting different components of insomnia captured by the two studies?

This is a valuable suggestion to evaluate whether the colocalization pattern of loci can indicate different components captured by the two cohorts.
However, there are very few loci that are specifically significant to UKB GWAS (there were 14 loci significant in UKB, of which 11 were also significant in meta-analysis but only 2 loci that were significant in UKB GWAS but not in 23andMe GWAS). The suggested comparison of enrichment would therefore not be statistically powerful enough when comparing UKB versus 23andme, thus we decided not to run this.

7. For the prioritization of high confidence genes, as a benchmark, how often is the high confidence gene the nearest gene to the SNP? And is the # of SNPs in the locus related to the probability of identifying a high confidence gene?

We have added this information to the manuscript:

Main text (line 257-)

Together, these 571 genes from 281 loci were called high confidence (HC) genes. Of the 571 HC genes, 216 (37.8%) genes were the closest genes of one of the index SNPs. The remaining 273 loci had significantly more SNPs and genes in the loci compared to 281 loci where at least one HC gene was identified (two-sided Mann-Whitney U test, p=5.4e-3 and p=2.5e-26, respectively).

8. Does utilizing PPI data to identify HC-m genes related to HC-1 genes introduce a source of bias, as this clearly won't identify genes that haven't been extensively studied? How many genes have PPI data available?

This is a valid remark; relying on PPI is presented in our manuscript as one way to link genes across loci, but alternatives, such as gene co-expression, can be used as well. The potential usage of co-expression networks instead of PPI (please also refer to the reviewer #2 point 3 for more details) is discussed in the manuscript (see below). PPI might indeed be biased by how extensively genes have been studied in this context. However, of the 20,260 protein-coding genes we used in this study, 80.6% had at least one PPI entry in the InWeb dataset, and of the 571 HC genes, 90.4% had at least one PPI in the InWeb. Therefore, the loss of genes due to lack of data is likely to be minimal, although we acknowledge that genes that have been studied more extensively are more likely to have multiple PPIs listed. We have also added a comparison between using PPI and using co-expression data (see revisions in text as listed at reviewer 2 comment 3)

Main text (line 390-)
Third, cross-locus linking of genes depends on the availability and reliability of biological information (PPI, co-expression networks, or any other gene-correlation matrix deemed relevant), which is currently not abundantly available and still imperfect.

9. Through the manuscript, please be consistent and use "GWAS" as opposed to "GWASs".

We have updated the manuscript accordingly.

10. The MVP cohort is largely male, perhaps this is a source of different between the cohorts and should be discussed.

The MVP indeed includes mainly male participants (>90% in this study). Although this could contribute to differences in gene discovery between MVP and other cohorts where sexes are more equally represented, previous studies showed that high concordance between MVP and other datasets for genetic effects related to psychiatric and behavioral traits with sex differences in their incidence and prevalence (PMID: 33510476, 32451486, 31906708, 31594949, 31151762l medRxiv 2020.05.18.20100685). We clarified this point in the revised manuscript.

Supplementary Note 2 (line 109-)
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[6–10]*.*

11. In Supp Table 3 it would be helpful to report the effect as an odds ratio given these are case control GWAS.

We added the Odds Ratio column in **Supplementary Table 3**.

12. In Supp Table 4 it is unclear what is meant by "suspicious loci".

We have clarified in the legend of **Supplementary Table 4**.

Suspicious loci: 1 if the locus was excluded from the analyses for suspicion of being a false positive, because there was only a single genome-wide significant variant and no SNPs in high LD, 0 otherwise

13. When discussing replication of previous loci (lines 167-170) it should be noted that prior studies are NOT independent of the current study - therefore it is unsurprising that Jansen et al. SNPs replicate in a study that expands on the prior sample.

We clarified this point in the manuscript and Supplementary Note.

Main text (line 172-)
note the meta-analysis in this study include samples from Hammerschlag et al. and Jansen et al.

Supplementary Note 5 (line 154-)
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).

14. The paragraph on enrichment of SNPs be genomic region (lines 175-187) can be condensed and supp tables referred to, as these results are common for nearly all GWAS.

As advised, we have shortened the section in the main text and referred to Fig. 1c and Supplementary Table 5.

Main text (line 178-)
*The 51,876 genome-wide significant SNPs showed enrichment in intronic, intergenic and 3' untranslated regions while they were depleted in exonic regions compared to all analysed SNPs (**Fig. 1c** and **Supplementary Table 5**).*

15. The discussion section could elaborate more on how the insights into insomnia biology in this study related to prior knowledge (do we see the same, what's a novel biological insight??).

The current study found enrichment of gene sets related to synaptic transmission/signalling and neuron differentiation based on insomnia GWAS results, which has not been reported previously. We also included a large set of diverse single cell gene-expression data, which allowed us to identify several new cell-types related to insomnia (e.g. habenular nuclei, lateral geniculate nuclei, GABAergic neurons). In addition, the identification of the potential locus heterogeneity based on metabolic and psychiatric traits is a novel finding. We have updated the discussion to emphasize these novel findings.

Main text (line 328-)
We observed clusters of insomnia loci based on colocalization patterns across multiple traits, indicating potential locus heterogeneity. In particular, a separation of the locus clusters that are either colocalized with metabolic or psychiatric traits is clinically relevant. This suggests that insomnia is a genetically heterogeneous phenotype consisting of different genetic subtypes, e.g. insomnia symptoms that are more related to metabolic disturbances or to other factors in the brain, and these genetic subtypes may require different treatment approaches. Indeed, metabolic disturbances have been found to contribute to hyperarousal in insomniacs compared to controls, such as increased whole body and brain metabolism, altered hormone secretion and sympathetic activation[51].

Using multiple cell-specific gene expression datasets, we identified novel associations of insomnia with neuronal cells, including habenular, lateral geniculate nuclei (LGN) and GABAergic neurons, among others. These findings are supported by prior 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 co-evolved (together forming the epithalamus)[38,39] and its activity follows a strong circadian pattern[39,40]. Among its hypothesized functions are sleep and circadian rhythm regulation through production of melatonin[38,39] and by maintaining REM-sleep, as evidenced by REM disturbances induced by habenular lesions[41,42]. The LGN is part of the visual system, and relays retinal information to cortical brain areas. In addition, the LGN is involved in circadian rhythm regulation through its intrinsic timekeeping properties[43] and indirect interactions with the supra-chiasmatic nucleus (SCN) through neuropeptide-Y[44,45]. Lesions in the LGN indeed have shown to affect circadian activity in animal models via disturbed processing of environmental cues[45]. GABA is among the most abundant neurotransmitters in the brain and is the main neurotransmitter of the circadian system[46] which inhibitory action induces a sleep-state[47–49], and the SCN consist almost entirely of GABAergic neurons[50]. Interestingly, the GABAergic system is the mechanism of action of drugs such as benzodiazepines that are often used to treat insomnia[48]. These observations point to several different but related mechanisms in the brain that may provide a basis for further study by experimental designs.

We demonstrated a novel strategy using known biological functions of SNPs and multi-locus functional relations of genes to prioritize the most likely 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 by-products, as it provided more specific results. Indeed, the gene set showed the most significant enrichment in the HCP genes, "Modulation of chemical synaptic transmission" (SynGO: BP), is at the lowest hierarchy of the gene ontology tree in the SynGO dataset (i.e. 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 the HCP genes in gene sets related to synaptic and neuronal processes, including neurogenesis and differentiation. This is the first time to link these gene sets from GWAS results. Evidence of synaptic transmission of neurotransmitters in insomnia has previously been found in imaging studies that demonstrates 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 are recently reported to occur in the hypothalamus[50], 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 an impaired sleep-wake regulation in humans (a review is provided by Kostin et al.[50]). The ultimate test of whether the HCP genes are actually causally involved still lies in functional follow-up experiments.

Reviewer #2:
Remarks to the Author:
In this manuscript, Watanabe et al. describes the genetic architecture of insomnia in 2.3 million individuals and propose a specific pipeline for gene prioritization, identifying enrichment in genes involved in synaptic function and neuronal differentiation.

The effort is impressive, and the catalogue of genes identified is large. The statistics are state of the art. The findings are solid but at the end, rather disappointing (involvement of neuronal genes in the phenotype, high correlation with depression) and a mere extension of prior studies. It is a technical paper and not much novel insight is gained in term of understanding the neurobiology or phenotype of insomnia.

We thank the reviewer for reviewing this manuscript and his/her valuable feedback. We believe this manuscript presents a novel approach to prioritize genes from GWAS that contain hundreds of significant loci and at least some loci that yield high confidence genesi. We believe it is timely as more studies are being published with >1 million individuals yielding an increasing number of risk loci. Apart from the technical part, we do also report novel findings based on the prioritized genes (please refer to reviewer #1 comment 15) for insomnia. Please see below for more detailed responses to each comment.

Point 1
For example, with such a huge sample, one wonder why the authors did not attempt to look at different aspects of insomnia, such as difficulties falling asleep versus waking up, or look at age dependency of genetic effects, or try to look at more focused hypothesis with regard to the potential role of circadian factors, depression, anxiety or other sleep disorders in insomnia.

This is an interesting point. However, the phenotype definition in UKB does not distinguish between symptoms of initiating sleep vs maintaining sleep (i.e. it is one question: 'Do you have trouble falling or staying asleep?'), which limits our possibilities to carry out these subtype analyses. There are other sleep related phenotypes available, such as duration of sleep or morningness, but we have shown in Jansen *et al.* that these correlate very low both at a phenotypic and genetic level with insomnia.

With regard to age-specific effects, we have repeated GWAS analyses in UKB while 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. The genetic correlation was 0.96, showing that there is only a little difference in the genetic factors across age groups. The SNP heritability was 7.95% in the first group (age > 58) and 9.09% 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.

To test whether genetic effects differ among individuals experiencing psychiatric symptoms, we performed additional GWASs by splitting the sample in those that report depressive symptoms (N=102,783) and those who did not (N=260,549). Here, we observe the genetic correlation of 1.02 and SNP heritability of 7.38% and 7.67%, 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.

We note that these analyses are limited to the UKB cohort, as we currently do not have access or an approved proposal to run these analyses on the phenotypic and genotype data of the 23andMe cohort. Due to the underpowered analysis compared to the main meta-analysis in this manuscript, we did not include these additional results in the manuscript as it is difficult to draw conclusions from our findings without replication in an independent cohort.

Point 2
There is a brief mention that clustering of the loci based on co-localization across traits suggest locus heterogeneity and distinct clusters of loci, one linked with metabolic traits the other with psychiatric traits, but it is not further discussed; there is no attempt at furthering the significance of this statement. As it is, it is a straightforward analysis showing limited creativity.

We thank the reviewer for this relevant point. We have now expanded on this:

Main text (line 328-)
We observed clusters of insomnia loci based on colocalization patterns across multiple traits, indicating potential locus heterogeneity. In particular, a separation of the locus clusters that are either colocalized with metabolic or psychiatric traits is clinically relevant. This suggests that insomnia is a genetically heterogeneous phenotype consisting of different genetic subtypes, e.g. insomnia symptoms that are more related to metabolic disturbances or to other factors in the brain, and these genetic subtypes may require different treatment approaches. Indeed, metabolic disturbances have been found to contribute to hyperarousal in insomniacs compared to controls, such as increased whole body and brain metabolism, altered hormone secretion and sympathetic activation[51].

Point 3
The pipeline is well explained and generally solid. The use of protein-protein interactions (PPI) (in Web database) of HCPm with HCP1 as the main method used to prioritize higher confidence gene within the HC-m is however also not really well justified. The author mentioned it would also be possible to use other prioritization methods, such as spatial or temporal co-expression, but they do not show that these methods yield similar results (line 272). It seems that the validation of using PPI should be done first in diseases where pathophysiology is better understood than insomnia, such as autoimmune diseases, specific neurological disorders or cardiovascular diseases. Is there evidence of this?

We agree that validation of the novel prioritization method is important. To address this, we have now used 4 GWAS of 3 molecular traits (Urate, IGF-1 and Testosterone males/females) from Sinnott-Armstrong et al. 2021 where the authors identified so-called core-genes that are thought to directly influence the trait of interest. We used the core-genes from these GWAS as positive controls and tested whether HCP genes are enriched for the core-genes.

Methods (line 745-)
Validation of gene prioritization
*We obtained the list of identified "core" genes for 3 molecular traits, Urate , IGF-1 and Testosterone from Sinnott-Armstrong et al. 2021 [ref] as a positive control to evaluate whether the prioritization approach proposed in this study can identify known causal genes. The core-genes were obtained from Supplementary File 8-10 from the study [ref] for Urate, IGF-1 and Testosterone, respectively (**Supplementary Table 36**). We used 4 GWAS (Urate, IGF-1, Testosterone male and Testosterone female) based on the UKB cohort used in the study [ref] and performed the gene prioritization strategy to link genes across loci. However, to also evaluate different types of linking genes apart from PPI, we also obtained gene co-expression for 53 tissues using GTEx v8 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-4).*

Main text (line 290-)
To validate whether the prioritization approach we propose here can identify known causal genes, we used GWAS of 3 molecular traits (Urate, IGF-1 and Testosterone) in which causal biological mechanisms are greatly 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 (**Methods**, **Supplementary Tables 36-41** and **Supplementary Note 11**).*

Supplementary Note 11 (line 405-)

*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.[37], and used "core" genes reported in the study as positive controls (**Methods** and **Supplementary Table 36**). We defined HC genes following the approach described in the main text, and identified 54 sets of HCP genes with PPI and 53 tissue specific gene co-expression datasets (**Methods**). HC and HCP genes for each GWAS are reported in **Supplementary Table 37-40** and the results of core-genes enrichment analysis is reported in **Supplementary Table 41**.*

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 37**). Although the enrichment of core-gene did reach significance after Bonferroni correction (p<9.26e-4), the HCP genes based on gene co-expression in the kidney cortex showed the lowest p-value followed by PPI (**Supplementary Table 41**).*

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 38**). 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 41**).*

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 39**). 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 41**).*

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 40**). 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 41**).*

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 are 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 tissus 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.*

In brief, it is a nice technical paper but it is hard to find it an original and exciting study.

We hope to have convinced the reviewer that the presented strategy is original as well as some of the novel insights this strategy provided for insomnia

Minor

Issues re reproducibility across subsamples, heritability and inflation/polygenicity are well-discussed, but at time this information is repetitive (lines 118, 199, 135).

Thank you for pointing this out. However, line 118-119 is discussing the inflation of the GWAS of the UKB and 23andMe separately while line 135 is for meta-analysis.

Decision Letter, first revision:

8th Oct 2021

Dear Professor Posthuma,

Your Article, "Genome-wide meta-analysis of insomnia in over 2.3 million subjects indicates a role for specific biological pathways through gene-prioritization" has now been seen by 2 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, with a view to identifying key priorities that should be addressed in revision. In this case, reviewer #1 has identified important issues in the GWAS and analyses that need to be addressed or clarified. In addition, both reviewers have concerns regarding the overall level of novelty and biological insight (in confidential comments), which should be improved, if possible.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
<a href="http://www.nature.com/ng/authors/article_types/index.html">here</a>.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our <a href="https://www.nature.com/nature-research/editorial-policies/image-integrity">guidelines on digital image standards.</a>

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

<strong>Note:</strong> This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within eight to twelve weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit <a href="http://www.springernature.com/orcid">www.springernature.com/orcid</a>.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Wei

Wei Li, PhD
Senior Editor
Nature Genetics
New York, NY 10004, USA
www.nature.com/ng

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Please find my further comments in relation to each rebuttal point.

R1, Comment 1. Given the relative ease of performing GWAS in related samples (BOLT-LMM and now REGENIE) it was a significant oversight of the research team to omit the related individuals from the beginning, but I am sensitive to the amount of work necessary to update all the tables. It would be helpful, however, if a supplementary table could be added detailing the additional loci in the related sample GWAS.

R1, Comment 3. In the edited introduction, it would be helpful to use "snp-based heritability" in line 1. I also think you have a typo where you write 1.3 instead of 2.3 million.

What I'm still missing from this discussion of variance explained, is the authors proposed explanation of why, despite a gain of 1 million people, the prediction power doesn't not seem to really increase all that much. Are we at a plateau for what we can achieve with such relatively broad insomnia Q's? I'm missing a little of the context of what would be the next steps for understanding the genetics of insomnia given sample size alone doesn't seem like it will yield much more.

R1, Comment 7. In many ways it is unsurprising that it is harder to identify an HC gene from a more complex region, but it would be helpful to compare the # of HC-1/HC-M genes IDed as a function of region complexity (#genes, #snps, size of region) to determine if the method has differential performance across a variety of loci. As well, it would be helpful to compare what is happening in the HC genes found via the credible set pathway vs. the GWS pathway (laid out in the flowchart in Fig 3b).

R1, Comment 12. Can you please specify the LD threshold used for "high LD"?

R2, Point 1: I disagree with the decision not to include these results, as these are key insights into the aetiology of insomnia and the relationship to age and underlying psychiatric disorders. Although the stratification by depressive symptoms has already been somewhat explored by sensitivity analysis in "healthy subsets" free of chronic diseases in prior insomnia GWAS.

R2, Point 2. I think this could be taken even further if not in this manuscript as future analysis – can you create a PRS for these clustered loci and look at health outcomes of the upper and lower quintiles of genetic risk.

Reviewer #2:
Remarks to the Author:
The authors have addressed my 2 major comments well: validation of the PPI interaction methods (also discussing in response to reviewer 1 regarding potential bias) and increasing discussion about improved biological insight. The revised version is addressing these points very well. Obviously it is highly frustrating sub-analyses cannot be done with the 23andme sample.

Re “Issues re reproducibility across subsamples, heritability and inflation/polygenicity are well-discussed, but at time this information is repetitive (lines 118, 199, 135)”, I was not implying that the information was technically identical, but that the same point could be made more briefly and some of the info in suppl data, and using more space for aspects that could be more interesting to the general readership such as phenotype heterogeneity, relation to RLS.

Author Rebuttal, first revision:

Below we provide a point-by-point response indicating the changes made to the manuscript. For clarity, our response is written in blue.

Editor’s comments:

In this case, reviewer #1 has identified important issues in the GWAS and analyses that need to be addressed or clarified.

We have taken up all suggested improvements of the reviewer #1.

In addition, both reviewers have concerns regarding the overall level of novelty and biological insight (in confidential comments), which should be improved, if possible.

We believe the overall novelty of the study can be summarized in three points:

- We show that for an extremely highly polygenic trait such as insomnia, increasing N leads to an increase in the detected SNPs, loci, genes and pathways, providing more confidence in existing and some novel pathways and cell types.
- We also show that increasing N in this case does not lead to increasing predictive power, and provide some suggestions of why this might be the case (i.e. genetic subtypes, extreme polygenicity, operationalization of the phenotype)
- 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 likely causal genes to prioritize genes from the remaining loci

This has now been added to the last paragraph of the discussion.

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:
Please find my further comments in relation to each rebuttal point.

R1, Comment 1. Given the relative ease of performing GWAS in related samples (BOLT-LMM and now REGENIE) it was a significant oversight of the research team to omit the related individuals from the beginning, but I am sensitive to the amount of work necessary to update all the tables. It would be helpful, however, if a supplementary table could be added detailing the additional loci in the related sample GWAS.

Following this reviewer's suggestion, we added the results of UKB European GWAS with related individuals using REGENIE, and highlighted the novel loci identified in the Supplementary Information (Supplementary Note 1 and Supplementary Table 2).

Supplementary Note 1

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[1]. 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 in the Methods, 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[1]. 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.

R1, Comment 3. In the edited introduction, it would be helpful to use "snp-based heritability" in line 1. I also think you have a typo where you write 1.3 instead of 2.3 million.

We have updated the sentence in the introduction to include SNP-based heritability.

Introduction (line 57)

It is moderately heritable (total heritability of 38-59%[4] and SNP-based heritability of 7%[5]), and recent genome-wide association studies (GWAS) have led to improved understanding of the complex polygenic etiology of insomnia[5–7].

We double-checked the use of 1.3 and 2.3 million, but we believe this is correct: 1.3 million refers to the previously published GWAS (Jansen et al 2019) whereas 2.3 million refers to the current GWAS.

What I'm still missing from this discussion of variance explained, is the authors proposed explanation of why, despite a gain of 1 million people, the prediction power doesn't not seem to really increase all that much. Are we at a plateau for what we can achieve with such relatively broad insomnia Q's? I'm missing a little of the context of what would be the next steps for understanding the genetics of insomnia given sample size alone doesn't seem like it will yield much more.

We added the following to the discussion:

Discussion (line 372)

We also observed that in spite of almost doubling our sample size, the SNP-based heritability did not notably increase (7-8%). Current genome-wide significant SNPs (i.e. p<5e-8) explain 17.3% of the total h^2_{SNP} and we estimated that over 57 million subjects are required to detect SNPs at genome-wide significance that explain at least 90% of the SNP heritability. As the 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, at least in the way it was operationalized in the current (and previous) GWAS. Our current results however do hint towards heterogenous forms of insomnia, one that is due to a metabo-genetic pathway and a second due to a psychiatric-genetic pathway. Future studies aimed at increasing prediction may benefit from collecting deep phenotyping data on insomnia patients and identify subtypes of insomnia.

Discussion (line 424)

In conclusion, in the current study we show that for an extremely polygenic trait such as insomnia, increasing sample size does lead to an increase in the detected SNPs, loci, genes and pathways, providing more confidence in existing and some novel pathways. 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 (i.e. genetic subtypes, extreme polygenicity and operationalization of the phenotype). 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 likely 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.

R1, Comment 7. In many ways it is unsurprising that it is harder to identify an HC gene from a more complex region, but it would be helpful to compare the # of HC-1/HC-M genes IDed as a function of region complexity (#genes, #snps, size of region) to determine if the method has differential performance across a variety of loci. As well, it would be helpful to compare what is happening in the HC genes found via the credible set pathway vs. the GWS pathway (laid out in the flowchart in Fig 3b).

Following this reviewer's suggestion, we assessed whether the number of HC genes is associated with the size, the number of SNPs and genes in the loci. We also compared across loci from which credible SNPs were identified and those where this was not the case.

Main text (line 266)

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.

Main text (line 279)

We observed that the number of HC genes in each locus is significantly and positively associated with the size, the number of SNPs and genes in the loci as expected (*p*=4.7e-5, 2.4e-54 and 1.1e-7, respectively, **Supplementary Fig. 13** and **Supplementary Note 13**). 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 next step.

Supplementary Note 13

Here we 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 of > 1 gene the number of genes present in a locus is 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. 13**). 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. 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. 13**). 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. 13**). 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.

R1, Comment 12. Can you please specify the LD threshold used for "high LD"?

We have updated the text to clarify this sentence.

Main text (line 245)

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 can be due to many variants being highly correlated with each other within the loci and small effect sizes.

R2, Point 1: I disagree with the decision not to include these results, as these are key insights into the aetiology of insomnia and the relationship to age and underlying psychiatric disorders. Although the stratification by depressive symptoms has already been somewhat explored by sensitivity analysis in "healthy subsets" free of chronic diseases in prior insomnia GWAS.

We have now added these results to Supplementary information.

Supplementary Note 2

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[5]. 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[6]. 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.

R2, Point 2. I think this could be taken even further if not in this manuscript as future analysis – can you create a PRS for these clustered loci and look at health outcomes of the upper and lower quintiles of genetic risk.

Following this suggestion, we grouped clusters of loci into two groups, one consisting of loci mainly colocalized with metabolic traits (cluster #1, 3 and 11) and the second consisting of psychiatric traits (cluster #2, 5 and 8; Supplementary Fig. 9b). We then extracted variants in those loci and computed PRS using PRSice for 3 x 10k samples used for the PRS analysis, for metabolic and psychiatric loci separately. For each of 3 sets of target samples, we obtained top and bottom 1, 5 and 10% subjects ranked by the PRS, and combined across 3 datasets. To assess whether subjects with high PRS show different health outcome from ones with low PRS for each metabolic and psychiatric loci, we obtained 3 phenotypes, overall health rating (field ID 2178), depressive symptom (scored based on 2050 and 2060, as described in the response above and Supple note 2) and body fat percentage (field ID 23099). We corrected for age, sex, array and first 10 PCs and used residuals to perform two Mann-Whitney U test. Multiple testing was corrected for metabolic and psychiatric loci with Bonferroni correction (0.05/9 = 5.6e-3 (3 phenotypes x 3 percentile)).

For PRS based on metabolic loci, top 5 and 10% subjects showed significantly lower overall health rating compared to bottom 5 and 10% (*p*=3.5e-4 and 2.2e-6, respectively). In addition, 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.

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.

We added this to Supplementary Note 10 and supplementary Fig 10.

Supplementary Note 10

To further investigate whether the two groups of 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) (**Supplementary Fig. 10**). 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 (**Supplementary Fig. 10**). 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 (**Supplementary Fig. 10**). These results suggest that there might be independent pathogenic mechanisms underlying insomnia that are related either to metabolic or to psychiatric traits.

Reviewer #2:
Remarks to the Author:
The authors have addressed my 2 major comments well: validation of the PPI interaction methods (also discussing in response to reviewer 1 regarding potential bias) and increasing discussion about improved biological insight. The revised version is addressing these points very well. Obviously it is highly frustrating sub-analyses cannot be done with the 23andme sample.

Re "Issues re reproducibility across subsamples, heritability and inflation/polygenicity are well-discussed, but at time this information is repetitive (lines 118, 199, 135)", I was not implying that the information was technically identical, but that the same point could be made more briefly and some of the info in suppl data, and using more space for aspects that could be more interesting to the general readership such as phenotype heterogeneity, relation to RLS.

We have now added a paragraph in the discussion that discusses some of the suggested topics of the reviewer #1 and that also briefly mentions the need for deep phenotyping.

Decision Letter, second revision:

Our ref: NG-A56654R1

26th Jan 2022

Dear Dr. Posthuma,

Thank you for submitting your revised manuscript "Genome-wide meta-analysis of insomnia in over 2.3 million subjects indicates a role for specific biological pathways through gene-prioritization" (NG-A56654R1). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics Please do not hesitate to contact me if you have any questions.

Sincerely,
Wei

Wei Li, PhD
Senior Editor
Nature Genetics
New York, NY 10004, USA
www.nature.com/ng

Reviewer #1 (Remarks to the Author):

The manuscript has been greatly improved during these rounds of revision. The reviewers have satisfactorily answered all of my comments.

Final Decision Letter:

In reply please quote: NG-A56654R2 Posthuma

6th Jun 2022

Dear Dr. Posthuma,

I am delighted to say that your manuscript "Genome-wide meta-analysis of insomnia prioritizes genes associated with metabolic and psychiatric pathways" has been accepted for publication in an upcoming issue of Nature Genetics.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Genetics style. Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

You will not receive your proofs until the publishing agreement has been received through our system.

Due to the importance of these deadlines, we ask that you please let us know now whether you will be difficult to contact over the next month. If this is the case, we ask you provide us with the contact information (email, phone and fax) of someone who will be able to check the proofs on your behalf, and who will be available to address any last-minute problems.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number (NG-A56654R2) and the name of the journal, which they will need when they contact our Press Office.

Before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that <i>Nature Genetics</i> is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. <a href="https://www.springernature.com/gp/open-research/transformative-journals"> Find out more about Transformative Journals</a>

Authors may need to take specific actions to achieve <a href="https://www.springernature.com/gp/open-research/funding/policy-compliance-faqs"> compliance</a> with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to <a href="https://www.springernature.com/gp/open-research/plan-s-compliance">Plan S principles</a>) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including <a href="https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Please note that Nature Portfolio offers an immediate open access option only for papers that were first submitted after 1 January, 2021.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at <a href="https://www.nature.com/reprints/author-reprints.html">https://www.nature.com/reprints/author-reprints.html</a>. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we invite you to upload the step-by-step protocols used in this manuscript to the Protocols Exchange, part of our on-line web resource, natureprotocols.com. If you complete the upload by the time you receive your manuscript proofs, we can insert links in your article that lead directly to the protocol details. Your protocol will be made freely available upon publication of your paper. By participating in natureprotocols.com, you are enabling researchers to more readily reproduce or adapt the methodology you use. Natureprotocols.com is fully searchable, providing your protocols and paper with increased utility and visibility. Please submit your protocol to https://protocolexchange.researchsquare.com/. After entering your nature.com username and password you will need to enter your manuscript number (NG-A56654R2). Further information can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#protocols

Sincerely,
Wei

Wei Li, PhD

Senior Editor
Nature Genetics
New York, NY 10004, USA
www.nature.com/ng