SLEEPJ, 2018, 1–3

doi: 10.1093/sleep/zsy211
Editorial

EDITORIAL

Power versus phenotyping precision of genome-wide association studies on sleep traits

Konrad Oexle*

Institute of Neurogenomics, Neurogenetic Systems Analysis Unit, Helmholtz Zentrum München, Neuherberg, Germany

*Corresponding author. Konrad Oexle, Institute of Neurogenomics, Neurogenetic Systems Analysis Unit, Helmholtz Zentrum München, Ingolstädter Landstr. 1, D-85764 Neuherberg, Germany. Email: konrad.oexle@helmholtz-muenchen.de.

Genome-wide association studies (GWAS) have provided unprecedented molecular insight into the polygenic architecture of multifactorial traits and diseases. GWAS results allow for disease-risk prediction that may be as good as in monogenic disorders [1], enable impartial cause–effect ordering of traits using Mendelian randomization, open unbiased and mostly novel perspectives on molecular physiology, and can be used for various correlation analyses, not only with other traits but also with specific expression patterns, regulatory pathways, or epigenetic states in development, for instance [2–4].

Individual genetic effects in multifactorial traits are usually small, however. Therefore, sample sizes have to be large [5, 6]. To reach sufficient size, early GWAS consortia combined their many small, precisely phenotyped and expensively genotyped samples in meta-analyses. However, with the rapidly decreasing cost of genome-wide SNP genotyping (currently about 30 USD per individual), very large samples ($N = 10^4$–10^6) have taken the stage, albeit with less precise phenotyping due to feasibility reasons. 23andme, a direct-to-consumer genotyping company that phenotypes its customers by web-based questionnaire only [7], soon achieved more GWAS power than international academic consortia, as the example of myopia has demonstrated [8, 9]. Indeed, a case selection error of up to 30% is compensated by only doubling the sample size (Figure 1a). Meanwhile, public- and charity-funded population databases such as UK Biobank (UKB) have joined in, providing genetic epidemiologists with genome-wide genotypes, questionnaire data, and, in subsets, health records, imaging and monitoring data.

Restless legs syndrome (RLS) was the first sleep-related trait to be analyzed successfully by GWAS [10]. Intronic variants of the TALE homeobox gene *MEIS1* reached genome-wide significance already in a relatively small sample of 401 clinically diagnosed cases. This was due to the unusually large effect size of *MEIS1* on RLS. Ten years later, two groups [11, 12] performed GWAS on insomnia complaints as assessed in UKB by a single query: "Do you have trouble falling asleep at night or do you wake up in the middle of the night?" The answer "Usually" to that query was considered as a proxy of chronic insomnia disorder (CID). Here also, *MEIS1* was identified as the most strongly associated locus, albeit with a much lower effect than in RLS (top signal at rs113851554 with *OR* = 1.2 in insomnia complaints when compared with *OR* = 2.0 in RLS [4]).

Its association with both RLS and insomnia complaints raised the question whether the effect of *MEIS1* on insomnia indicates true pleiotropy or whether it merely reflects contamination of the UKB insomnia cases with cases of RLS. Such enrichment of the case sample with another disease due to unspecific phenotyping may have consequences that are not compensated by a large sample size (Figure 1b). This reveals a problem of imprecise phenotyping in large databases which initially may not have been properly recognized. Hammerschlag et al. [11] tried to estimate the confounding of their insomnia GWAS by RLS: They calculated the RLS enrichment $P(RLS|ic)$ of usually having insomnia complaints (*ic*) as posterior probability by applying Bayes' formula on the RLS prevalence $P(RLS)$ as prior probability, using sensitivity $P(ic|RLS)$ and specificity $P(non\text{-}ic|nonRLS)$ in identifying RLS by the UKB query for the calculation. Sensitivity and specificity were estimated from independent databases containing information on both RLS and *ic*, as $P(ic|RLS) = 0.44$ and $P(non\text{-}ic|nonRLS) = 0.73$, respectively. $P(RLS|ic)$ was thus derived as 12%. After analogous calculation of the de-enrichment of RLS in UKB insomnia

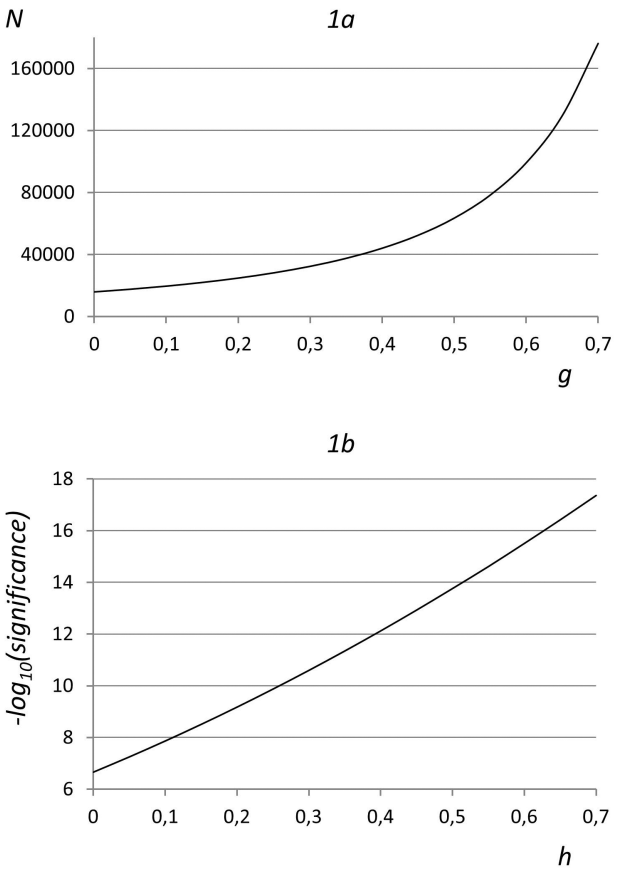

Figure 1. (a) Required sample size N in an association study on a genetic variant with low allele frequency (p = 0.03) and weak effect (odds ratio, OR = 1.2) as a function of the precision in phenotyping, that is, of the proportion g of confounding controls among cases. This shows that substantial imprecision can be compensated by moderate increase in sample size, e.g. by +23% for g = 10%. For comparison, a decrease of effect size from OR = 1.2 to OR = 1.1 requires a +300% increase in sample size. N is calculated at a power threshold of 80% ($Z_{1-\beta} = -0.84$) and a genome-wide significance threshold of 5×10^{-8} ($Z_{\alpha} = -5.33$). With $p^2 \to 0$, the power calculation is substantially simplified because the homozygotes can be ignored, the expected number μ of carriers among controls and cases can be approximated as $\mu_{ctrls} \approx 2Np$ and $\mu_{cases} \approx 2Np(OR(1-g)+g)$, respectively, and the standard deviation σ of μ is about $(2Np)^{1/2}$ in both samples if $OR \to 1$. For the statistical power to be sufficient, μ_{ctrls} and μ_{cases} must range sufficiently far apart from each other, requiring at least that $\mu_{ctrls} - \sigma Z_{\alpha} = \mu_{cases} + \sigma Z_{1-\beta}$, where the Zs are quantiles of a standard normal distribution. The latter equation yields $N \approx (Z_{\alpha} + Z_{1-\beta})^2/(2p(OR-1)^2(1-g)^2)$. (b) Expected association significance of a rare variant (p = 0.03) with weak effect (OR_1 = 1.2) as a function of a subset of proportion h which confounds the case sample by a second trait on which the variant has a stronger effect (OR_2 = 1.4). One-sided test with approximations as in (a), $\mu_{cases} \approx 2Np(OR_1(1-h)+OR_2 h)$, and N = 10,000 for both cases and controls, showing a nearly linear relation between h and the logarithm of the significance, and thereby indicating that little such confounding may already generate false-positive genome-wide significance beyond 5×10^{-8}.

controls, *P(RLS|non-ic)*, the ic GWAS signal at rs113851554 within the MEIS1 locus was estimated under the assumption that only the RLS enrichment in cases and de-enrichment in controls drive that signal. However, even with an assumed effect of rs113851554 on RLS being as large as OR = 2.6, the confidence interval of the estimation did not comprise the actually measured signal strength in insomnia, suggesting that there is an independent pleiotropic effect on CID.

In this issue of *SLEEP*, El Gewely et al. [13] now have examined the role of *MEIS1* in CID by analyzing three *MEIS1* variants (including rs113851554) in a sample of 646 people with clinically diagnosed primary CID. 26.5% of these patients also had RLS. Comparing the CID+RLS cases to the CID-only cases confirmed the association of RLS with *MEIS1*, whereas comparing the 476 CID-only cases to an RLS-free control sample of similar size did not show any *MEIS1* association at all. This result seems to suggest that the *MEIS1* signal in the large GWAS on insomnia symptoms discussed above was driven by confounding RLS only. However, the sample size of El Gewely et al. does not have sufficient statistical power to evaluate the effect size determined in the insomnia GWAS. The sample sizes of cases and controls would have to be three times as large (N = 1,460), if an CID association of rs113851554 at an effect size of OR = 1.2 was to be detected in a one-sided test with the usual power of 80% (as calculated using the approximation in Figure 1 with p = 0.053, α = 0.05, $Z_{\alpha} = -1.64$, $Z_{1-\beta} = -0.84$, g = 0, and ignoring that the control sample possibly contains some people with CID). Nonetheless, the analysis of Al Gewely et al. provides some first evidence. Moreover, they introduce their well-characterized CID sample to the research community. This sample should be useful for evaluating large-scale samples collected by simplified questionnaire phenotyping. Moreover, with genome-wide genotyping, it could be used to assess the genetic correlation between CID-only and RLS-only, and if joined meta-analytically with other such samples, it could be used for replication studies, scrutinizing the signals of large-scale GWAS on insomnia complaints.

Indeed, by combining the UKB and 23&me populations, GWAS on insomnia complaints has now arrived at samples sizes of 4×10^5 cases and 9×10^5 controls, yielding 202 genome-wide significant loci [3]. However, as discussed above, phenotyping is imprecise, especially in the 23&me dataset, where the sensitivity and specificity for CID were shown to be only 84% and 80%, respectively. These, together with the percentage of cases in 23&me (30%), imply that the proportion of CID among cases, *P(CID|case)*, is as low as 45%. Although this deficiency can be compensated by the very large sample size (Figure 1a), there was also strong evidence of confounding RLS: 10, that is, 50% of the 20 independent lead SNPs of the most recent RLS GWAS [4] reside within one of the insomnia risk loci although the latter together comprise less than 2% of the genome. In fact, 7 of these 10 RLS loci are identical to the corresponding insomnia loci. Moreover, the 10 loci include the first, second, third, fifth, sixth, and seventh most significant RLS loci, and the 10 insomnia loci driven by them include the first (*MEIS1*), sixth (*BTBD9*), and tenth (*PTPRD*) most significant of the 202 insomnia loci (see Supplementary Material of Ref. 3). Hence, although there are differences between the results of RLS GWAS and GWAS on insomnia complaints, an RLS subset appears to be a major driver of the insomnia GWAS signals. Unfortunately, according to their bioRxiv prepublication of February 2018 [3], the authors of that insomnia GWAS have not yet sufficiently addressed this point.

In summary, we are left with the question of whether the overlap between the RLS and insomnia GWAS signals is merely the result of confounding RLS or whether it reveals some pleiotropic genetic effects, too. In fact, the term "pleiotropy" deserves discussion itself. It might just reflect the existence of oligosymptomatic RLS, manifesting with little more than insomnia complaints. This may also be important, since such insomnia cases could be treatable by RLS-specific medication.

Conflict of interest statement. None declared.

References

1. Khera AV, *et al*. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;**50**(9):1219–1224.
2. Forrest MP, *et al*. Open chromatin profiling in hiPSC-derived neurons prioritizes functional noncoding psychiatric risk variants and highlights neurodevelopmental loci. *Cell Stem Cell*. 2018;**21**(3):305–318.e8.
3. Jansen PR, *et al*. Genome-wide analysis of insomnia (N=1,331,010) identifies novel loci and functional pathways. *bioRxiv*. February 2018. doi:10.1101/214973
4. Schormair B, *et al*. Identification of novel risk loci for restless legs syndrome in genome-wide association studies in individuals of European ancestry: a meta-analysis. *Lancet Neurol*. 2017;**16**(11):898–907.
5. Oexle K. A remark on rare variants. *J Hum Genet*. 2010;**55**(4): 219–226.
6. Risch N, *et al*. The future of genetic studies of complex human diseases. *Science*. 1996;**273**(5281):1516–1517.
7. Eriksson N, *et al*. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet*. 2010;**6**(6):e1000993.
8. Kiefer AK, *et al*. Genome-wide analysis points to roles for extracellular matrix remodeling, the visual cycle, and neuronal development in myopia. *PLoS Genet*. 2013;**9**(2):e1003299.
9. Verhoeven VJ, *et al*. . Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat Genet*. 2013;**45**(3): 314–318.
10. Winkelmann J, *et al*. Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. *Nat Genet*. 2007;**39**(8): 1000–1006.
11. Hammerschlag AR, *et al*. Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits. *Nat Genet*. 2017;**49**(11):1584–1592.
12. Lane JM, *et al*. Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. *Nat Genet*. 2017;**49**(2):274–281.
13. El Gewely, M, *et al*. Reassessing GWAS findings for the shared genetic basis of insomnia and restless legs syndrome. *Sleep*. 2018; **41**(11). doi:10.1093/sleep/zsy164