scientific reports

OPEN

Reproducibility in the UK biobank of genome-wide significant signals discovered in earlier genome-wide association studies

Jack W. O'Sullivan[1,2]✉ & John P. A. Ioannidis[2,3]

With the establishment of large biobanks, discovery of single nucleotide variants (SNVs, also known as single nucleotide polymorphisms (SNVs)) associated with various phenotypes has accelerated. An open question is whether genome-wide significant SNVs identified in earlier genome-wide association studies (GWAS) are replicated in later GWAS conducted in biobanks. To address this, we examined a publicly available GWAS database and identified two, independent GWAS on the same phenotype (an earlier, "discovery" GWAS and a later, "replication" GWAS done in the UK biobank). The analysis evaluated 136,318,924 SNVs (of which 6289 reached *P* < 5e−8 in the discovery GWAS) from 4,397,962 participants across nine phenotypes. The overall replication rate was 85.0%; although lower for binary than quantitative phenotypes (58.1% versus 94.8% respectively). There was a 18.0% decrease in SNV effect size for binary phenotypes, but a 12.0% increase for quantitative phenotypes. Using the discovery SNV effect size, phenotype trait (binary or quantitative), and discovery *P* value, we built and validated a model that predicted SNV replication with area under the Receiver Operator Curve = 0.90. While non-replication may reflect lack of power rather than genuine false-positives, these results provide insights about which discovered associations are likely to be replicated across subsequent GWAS.

Genome-wide association studies (GWAS) have resulted in the discovery of tens of thousands of genetic associations for various traits and phenotypes. Polygenic risk scores[1], innovative drug discovery[2], and gene-editing[3] have all been enhanced, or even based on, GWAS results. Genome-wide association studies investigate the association of individual single nucleotide variants (SNVs) on a phenotype of interest (for example coronary artery diseases)[4]. Most GWAS identify SNVs with, individually, small effects[4]. This supports the notion that most diseases are polygenic, rather than monogenic, in nature[5].

To observe the small effect of individual SNVs, GWAS have relied on increasingly larger sample sizes[4]. Recent advances have seen rapidly increasing sample sizes, particularly with the establishment of large biobanks. The most widely used and analyzed biobank in human genetics is the UK Biobank (UKBB)[6]. Analyses done in the UKBB and other similar biobanks have the opportunity not only to identify new associations but also to replicate previously proposed associations that arose from other GWAS investigations. It is not unexpected that some SNVs that were considered to be associated with a phenotype in an earlier GWAS may not be replicated in a subsequent GWAS. Even if they are replicated, their effect size may change, e.g. because of the winner's curse phenomenon[7], where early discoveries see attenuation of their effect size when they are replicated in subsequent studies. This has implications for all scientific progress, and even patient care, (i.e. polygenic risk scores) reliant on GWAS results, if these scores include variants that have null effects or effects that are smaller than those anticipated based on their earlier discovery profile.

Although several studies have looked at SNV replication for specific phenotypes, it remains broadly unclear across phenotypes how often SNVs replicate, how this varies between binary and quantitative traits, at different *P* values, across varying effect sizes, and how effect sizes change between earlier, smaller GWAS and later, larger GWAS examining the same phenotype. A most interesting comparison would be to contrast earlier GWAS versus

[1]Division of Cardiovascular Medicine, Stanford University, Stanford, CA, USA. [2]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA. [3]Departments of Medicine, of Epidemiology and Population Health, of Biomedical Data Science, and of Statistics, Stanford University, Stanford, CA, USA. ✉email: jackos@stanford.edu

the UKBB, which has become a standard, widely used resource. We set out to address these questions, and, from our results, built a model to predict SNV replication.

Methods

Data acquisition. To determine the reproducibility of SNVs between an earlier GWAS and the UKBB, we identified two, independent GWAS on the same trait, one without data from the UKBB and the second being done on UKBB data. To do this, we systematically searched a publically available database of genome-wide association studies (GWAS) (available at: https://atlas.ctglab.nl/)8 for GWAS that had been conducted for the same trait (e.g. systolic blood pressure) first using data independent of the UKBB and then a second, independent GWAS using exclusively UKBB data. Thus a trait was eligible if there were two independent GWAS available for it; one not using UKBB data (hereafter referred to as: discovery GWAS) and one using UKBB data (hereafter: replication GWAS). All discovery GWAS occured before the replication GWAS. Further inclusion criteria was GWAS conducted in European subjects (or results available for exclusively Europeans) and GWAS with more than 50 genome-wide significant SNVs, so as to allow having a meaningful number of discoveries to be assessed for replication. More information on the GWAS database we searched and its accompanying paper[8] are available in the appendix. Upon acceptance, we will make all the data available and its accompanying code (https://github.com/jackosullivanoxford, specifically: https://github.com/jackosullivanoxford/Repro_GWAS/blob/master/Data_cleaning_meta_analysis_regression_prediction).

Determination of reproducibility. To determine the reproducibility of SNVs in the discovery and replication GWAS we performed three broad steps: (1) Determined overlap of SNVs between discovery and replication GWAS (via rsID) and only included SNVs shared between two GWAS cohorts. We then identified the SNVs that reached genome-wide significance (defined using the accepted significant threshold for GWAS: $P<5e{-}8$, regardless of the threshold that the original authors might have used) in the discovery GWAS—these were the SNVs we determined the reproducibility of. (2) Aligned the effect allele between the discovery and replication GWAS, and consequently inverted the effect size if effect alleles did not originally match and (3) Classified SNVs as replicated if they reached genome-wide significant ($P<5e{-}8$) in both discovery and replication GWAS and had congruent effect directions in both GWAS (e.g. odds ratio (OR) above 1 in both GWAS). All SNV effect sizes were converted to OR before reproducibility was determined via the Chinn formula[9]. Thus, SNV effect sizes that were originally produced from linear models for quantitative (continuous) traits were converted to OR. Lastly, as a sensitivity analysis we explored the reproducibility of SNVs using the more lenient significance of $P<10e{-}6$. For this analysis we tested the reproducibility of SNVs that had a P value < 10e−6 in the discovery cohort, and used a reproducibility P value threshold of $P<10e{-}6$ in the reproducibility cohort. Further details appear in the appendix.

Calculating reproducibility. We calculated the replication rate for each included trait individually, for all traits collectively, and for binary (e.g. coronary artery disease) and quantitative (e.g. diastolic blood pressure) traits separately. To calculate replication rate for each individual trait we calculated a simple proportion (e.g. [number of SNVs replicated]/[number of SNVs shared between discovery and replication GWAS]). To calculate the replication rate for all traits collectively we constructed a inverse-variance meta-analysis[10] using fixed-effects. Further, we constructed similar inverse-variance meta-analysis[10] to determine the replication rate for binary and quantitative traits; including only traits recorded in a binary fashion (yes/no) or on a continuous scale, respectively. To explore the replication rate across P values and odd ratios, we also performed meta-analysis assessing the replication of SNVs with certain P value and OR characteristics (from the discovery GWAS). We calculated the reproducibility of SNVs across the following discovery GWAS P value categories: 5e−8 to 5e−9, 5e−9 to 5e−10, 5e−10 to 5e−11, and < 5e−11. We calculated the reproducibility of SNVs across the following discovery GWAS OR categories: 1–1.05, 1.05–1.1, 1.1–1.15, 1.15–1.2, 1.2–1.3, 1.3–1.4, > 1.4.

Quantifying the change in effect size between GWAS. To determine if a change in SNV effect size occured between the earlier, discovery GWAS and the later, replication GWAS in the UKBB we constructed a single variate linear model, with the discovery OR as the predictor variable and replication OR as the outcome variable. As stated above (see '*Determination of reproducibility*'), we converted all SNV effect sizes to an OR via the Chinn formula[9]. Then, to help interpret the output from this model, we converted all OR values to above 1 (using the formula 1/OR if the original SNV OR was < 1) Finally we combined SNVs across all traits for the model. From the regression model, we determined the regression coefficient for the discovery OR and interpreted this coefficient as the change in OR between GWAS (e.g. a regression coefficient of 0.80 would imply that 20% decrease in OR between discovery and replication GWAS). We only quantified the change in effect size of SNVs that were replicated, and also for all SNVs that had reached genome-wide significance in the discovery GWAS, regardless of whether they were replicated or not in the replication GWAS. We performed similar analyses for binary and quantitative traits individually.

Prediction model for SNV replication. First we constructed a multivariate logistic regression model to examine the association of our predictors (odds ratio, P value, P value category (as above), and trait characteristic (binary vs. quantitative) on replication. We initially split our data into test and train sets (split, randomly, by half). Using the train set, we constructed a logistic regression model using the following predictors: odds ratio (numeric, not category), P value category, trait characteristic (binary vs. quantitative), minor allele frequency (taken from the discovery cohort), INFO score (to reflect imputation quality—taken from replication cohort), and a sample size ratio (ratio of replication cohort sample size divided by discovery cohort sample size). We then

Disease	Total sample size	Number of genome-wide significant SNVs	Number of SNVs that are replicated (%)
Asthma	225,309	889	494 (56%)
SBP	430,797	110	107 (97%)
Eczema	330,142	640	337 (53%)
BMI	613,900	1835	1756 (96%)
Waist Circumference	618,033	937	827 (89%)
Hip circumference	598,925	1083	1043 (96%)
Coronary Artery Disease/IHD	387,786	159	149 (94%)
Resting Heart rate/Pulse Rate	447,198	549	547 (99%)
DBP	430,806	87	83 (95%)

Table 1. Replication across phenotypes. Total sample size is the sample size of the discovery and replication GWAS collectively.

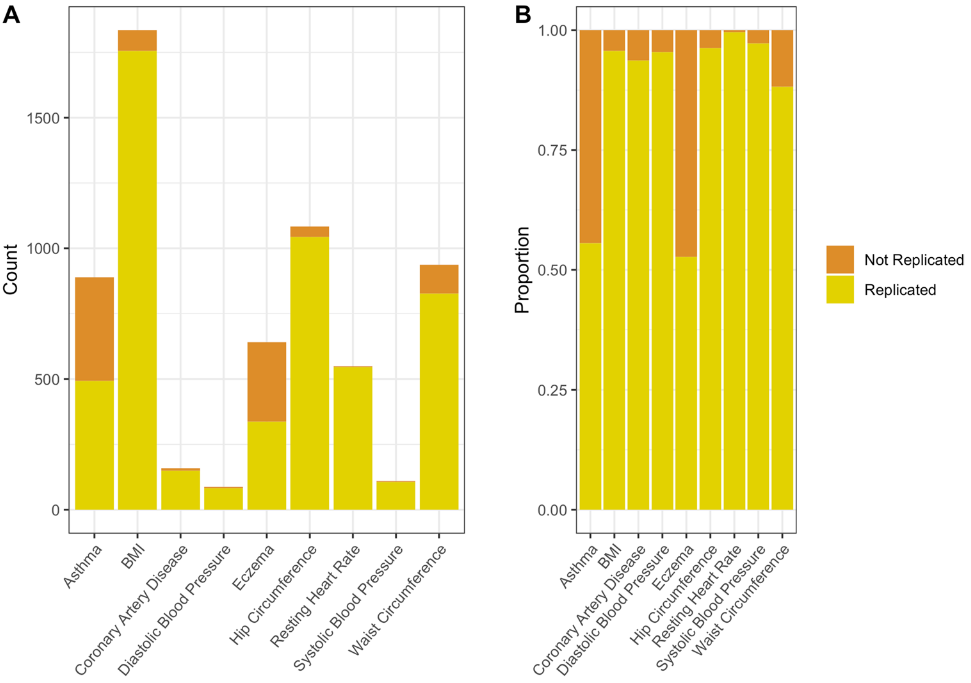

Figure 1. Replication of SNVs across traits.

tested the constructed model on the test set. We report the model's predictive accuracy via the following metrics: sensitivity, specificity, and area under the curve (AUC) all with 95% confidence intervals. We further assessed model fit via McFadden's R^2; a measure of explained variation in logistic regression models, defined as the difference between the the (maximized) likelihood value from the fitted model and that of the null model – the model with only an intercept and no covariates.

Results

We analysed 136,318,924 SNVs from 4,397,962 participants across nine different phenotypes (from 18 GWAS, 9 pairs) (Table 1). The traits included were: asthma, systolic blood pressure (SBP), eczema, body mass index (BMI), waist circumference, hip circumference, coronary artery disease (CAD), resting pulse rate, and diastolic blood pressure (DBP). Of the 136,318,924 included SNVs, 6,289 reached genome-wide significance ($P < 5e{-}8$) in the discovery GWAS (Table 1 and eTable 1).

Replication rate. Of the 6289 SNVs that were genome-wide significant in the discovery cohort, 5343 were replicated in the replication cohort (85.0%, 95% Confidence Interval (CI): 84.1% to 85.8%) (eFigure 1). Results varied substantially between binary and quantitative traits; the replication rate for exclusively binary phenotypes was 58.1% (95% CI 55.7% to 60.4%) (eFigure 2), compared with 94.8% (95% CI 94.2% to 95.4%) for quantitative traits (eFigure 3). The replication rate varied across the included phenotypes from 52.7 to 99.6% (Fig. 1). Using more lenient, Bonferroni-corrected P value thresholds (calculated by $P = 0.05/N$, where N is the number

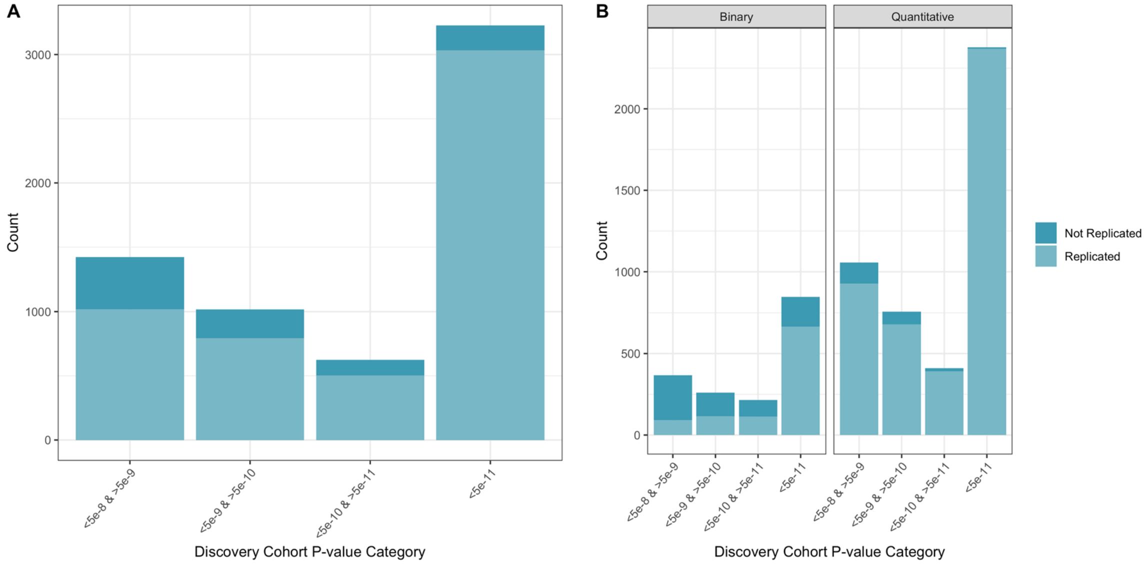

Figure 2. Replication of SNVs across *P* values.

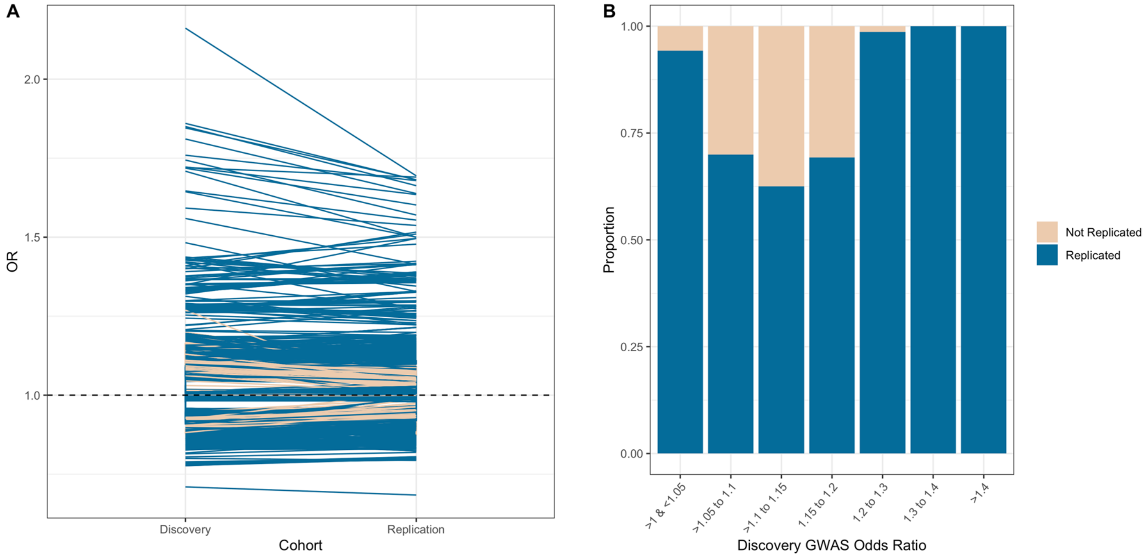

Figure 3. Replication of SNVs across odds ratios.

of significant SNVs in discovery GWAS, used of using 5e−8), the replication rate improved for all phenotypes, particularly the binary phenotypes (eTable 3). However, the opposite was appreciated when using the *P* value threshold of < 10e−6 (i.e. reproducibility decreased across all phenotypes (eTable 4).

Furthermore, the replication rate varied across discovery GWAS *P* values and OR (Figs. 2, 3, eFigure 4 and eFigure 5). As is expected, the replication rate increased as the discovery GWAS SNV *P* value decreased (Table 2); the highest replication was observed with a *P* value < 5e−11 (94% (95% CI 93% to 95%). A less consistent pattern was observed with discovery GWAS OR, almost all OR > / = 1.2 were replicated (Table 2), however a similarly large number of SNVs with a discovery OR of > 1 to < 1.05 were replicated (94.3% (95% CI 93.5% to 95.0%)). This is likely due to the fact that all SNVs > 1 to < 1.05 were for quantitative traits, with no SNVs corresponding to binary traits (Fig. 4).

Change in effect size between GWAS. When considering SNVs that were replicated in both cohorts, we found a 9.6% (95% CI 8.9% to 10.2%) decrease in replicated SNV OR between discovery and replication cohorts (Fig. 3), for all phenotypes collectively. This decrease in effect size was larger for binary traits (18.0% (95% CI 16.0% to 20.0%), eFigure 6), however for quantitative traits an increase in effect size was observed (12.0% (95% CI 11.0% to 13.0%), eFigure 6). The change in effect size varied substantially across phenotypes (eFigure 7).

Metric	Category	Replication rate (95% CI)
P value	5e−8 to > 5e−9	72% (69% to 74%)
	5e−9 to > 5e−10	78% (75% to 80%)
	5e−10 to > 5e−11	81% (77% to 83%)
	< 5e−11	94% (93% to 95%)
Odds ratio	1–1.05	94.3% (93.5% to 95.0%)
	1.05–1.1	70.0% (66.8% to 72.9%)
	1.1–1.15	62.5% (59.4% to 65.6%)
	1.15–1.2	69.3% (64.3% to 73.9%)
	1.2–1.3	98.7% (91.0% to 99.8%)
	1.3–1.4	100%*
	> 1.4	100%*

Table 2. Replication across *P* values and odds ratios. *Paucity of data prevented formal meta-analysis.

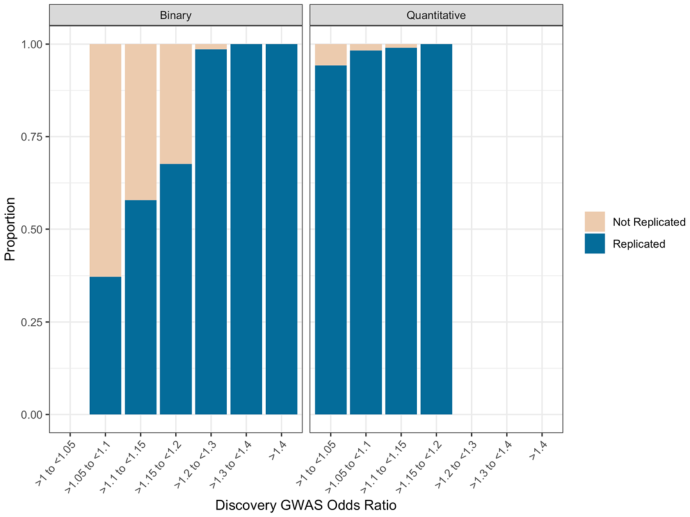

Figure 4. Replication of SNVs across odds ratios between Binary and Quantitative traits.

When considering SNVs that reached genome-wide significance in the discovery cohort (and weren't necessarily replicated), we found a 16.4% (95% CI 82.8% to 84.4%) decrease in SNV OR between discovery and replication cohorts, for all phenotypes collectively. For binary traits this decrease was 13.6% (95% CI 11.4% to 15.9%), whereas we observed a 10.9% (95% CI 9.9% to 11.9%) increase for quantitative traits.

Predicting SNV replication. First, from our training model the following predictors were significantly associated with SNV replication: discovery cohort SNV odds ratio (Fig. 4), discovery cohort trait (binary or quantitative), discovery cohort SNV *P* value < 5e−10 & > 5e−11, and discovery cohort SNV *P* value < 5e−11 (both categorical variables with *P* value < 5e−8 & > 5e−9 as reference) (eTable 2). *P* value as a continuous variable and *P* value < 5e−9 & > 5e−10 were not significant (eTable 2).

When we applied our training model to our test data set, we found an area under the Receiver Operator Curve (ROC) of 0.90 (95% CI 0.88 to 0.91) corresponding to a sensitivity and specificity of 82.1% (95% CI 77.5% to 93.8%) and 82.7% (95% CI 70.2% to 87.7%) respectively. We found a McFadden's R^2 of 0.36, reflecting a modest explanation of the variation.

Discussion

We analysed 136,318,924 SNVs from 4,397,962 participants across nine different phenotypes (18 GWAS). Of these 136,318,924 SNVs, 6,289 SNVs reached genome-wide significance in the respective discovery GWAS, of which 5,343 were replicated in their replication GWAS (85.0%, 95% Confidence Interval (CI): 84.1% to 85.8%). Replication rate varied substantially between binary and quantitative phenotypes and it was lower in the former. Further, replication rate varied across *P* value and OR of discovery GWAS SNV. We also found that SNV odds ratios (OR) decreased between discovery and replication GWAS for binary phenotypes, but increased for quantitative phenotypes. Lastly, we developed and then validated a model to predict SNV replication, and found it to be accurate (0.90 (95% CI 0.89 to 0.91)).

Implications. Our results have implications for the GWAS results. First, the SNV replication rate for quantitative phenotypes is very high; implying that quantitative GWAS in the UKBB had likely reached sufficient power to accurately detect all SNVs that were truly associated with a phenotype and that had been discovered by earlier GWAS efforts. We also quantified, using non-stimulated data, the concept of winner's curse; the change in effect size between our smaller discovery cohort and larger replication cohort may be a useful comparison for future studies that aim to quantify winner's curse. The high replication rate observed for quantitative traits may also reflect the precision and relative ease in which quantitative traits can be measured. The converse of this, the likely measurement error and ultimate definition heterogeneity of binary phenotypes, may be one explanation for the relatively low rate of replication in binary phenotypes. For instance, binary phenotypes often represent complex clinical diseases that can have (a) broad diagnostic criteria (e.g. angina, and myocardial infarction are often captured under "Coronary Artery Disease") and (b) are defined via an array of data sources, of varying quality. The UKBB, for instance, defines their phenotypes with ICD codes based on linked electronic health records (EHR)[6]. While this probably represents the best current method to define phenotypes in large cohorts, EHR data is messy and likely to include some administrative and clinical error[11]. An improvement in the phenotyping in data used for GWAS of binary phenotypes is likely to result in improved SNV replication. This may be even more crucial for phenotypes where we saw low replication rates, e.g. eczema.

On the one hand, it is encouraging that much scientific progress has been accomplished with current binary GWAS. For instance, polygenic risk scores based on current binary GWAS have been shown to accurately predict complex, common phenotypes[12,13]. With improved phenotyping, it seems plausible that these scores will continue to improve. Nevertheless, in the meantime there may be other ways to enhance current binary GWAS results for polygenic risk scores. First, our results clearly show a superior replication rate with quantitative phenotypes. These quantitative phenotypes are often more in line with physiological processes (e.g. systolic blood pressure) than clinical diseases (e.g. coronary artery disease). As such, future GWAS that directly use metabolomic data as outcomes (such as protein expression) are likely to, similarly, have higher accuracy than clinical disease phenotypes. Future research merging metabolomic outcomes and GWAS may be a useful addition to our scientific knowledge. For instance, some evidence suggests that the use of 'intermediate' phenotypes—between the genotype and the disease-based phenotype—may improve disease prediction[14]. For example, a 2021 study showed that the integration of polygenic risk scores for both disease-associated biomarkers and polygenic risk scores for the disease itself showed enhanced prediction over the polygenic risk score for the disease exclusively[14]. Second, almost all SNVs for binary traits with an OR > / = 1.2 were replicated, whereas the majority of SNVs with an OR below 1.2 were not replicated and this may reflect lack of power in the replication dataset. Of note, many of the replication UKBB datasets that we considered here did not use the full UKBB data, and power is likely to improve as complete biobank data are used and many biobanks are combined.

Limitations in comparison to previous literature. We were surprised to find only nine phenotypes where two GWAS had been conducted in truly independent participants and where inclusion or not of UKBB data was a distinguishing feature. It is plausible that further independent GWAS on the same traits exist, although this seems unlikely given the thorough and systematic search we performed of the GWAS atlas[8]. It is, however, likely that more GWAS are available, but they contain overlapping samples between GWAS (i.e. two GWAS of the same phenotype are not truly independent as they contain similar cohorts of participants), aren't of sufficient quality to be included in the GWAS Atlas, are conducted in a non-European population, or have not made their summary statistics available. An earlier study[15] reports building a model for SNV replication using GWAS for over 50 phenotypes, although it is unclear what, if any, measures were taken to determine if these numerous GWAS were truly independent i.e. did not include overlapping participants. Also, this study validated their model in two, small GWAS of one trait. Furthermore, this study didn't actually quantify a SNV replication rate, nor did they stratify their results by binary and quantitative phenotypes. A further limitation of our study is that we didn't include other SNV features, ideally we would have liked to include, for instance, LD as predictors in our model. However, this data was sparsely available. Lastly, it should be acknowledged that large disease-specific consortiums generally qualitatively describe the replication of SNVs as their consortium increases. Our study quantifies this formally and, importantly, quantifies replication across more than one phenotype.

Future research. We have identified a number of future research priorities. First, improving the phenotyping of binary phenotypes seems to be a priority for GWAS. Second, to facilitate an assessment of SNV replication, future independent cohorts are likely required. Many efforts to do this are already underway (e.g. AllofUs cohort and Millions Veteran Program).

Conclusions. The replication of SNVs discovered from GWAS was high for quantitative phenotypes. Genome-wide Association Studies appear to be entirely sufficient to detect SNVs associated with quantitative traits. For binary traits, however, the replication rate is modest. We have built a simple prediction model that can accurately ascertain SNV replication in later GWAS. It may be of use for researchers and clinicians that utilize GWAS results.

Data availability

All data used is publicly available from https://atlas.ctglab.nl/.

Received: 26 January 2021; Accepted: 31 August 2021
Published online: 20 September 2021

References

1. O'Sullivan, J. W. *et al.* Combining Clinical and Polygenic Risk Improves Stroke Prediction Among Individuals With Atrial Fibrillation. *Circ Genom Precis Med.* **14**(3), e003168. https://doi.org/10.1161/CIRCGEN.120.003168 (2020).
2. Shu, L., Blencowe, M. & Yang, X. Translating GWAS findings to novel therapeutic targets for coronary artery disease. *Front. Cardiovasc. Med.* **5**, 56 (2018).
3. Wu, S. *et al.* Genome-wide association studies and CRISPR/Cas9-mediated gene editing identify regulatory variants influencing eyebrow thickness in humans. *PLoS Genet.* **14**, e1007640 (2018).
4. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
5. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).
6. Sudlow, C. *et al.* UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
7. Xiao, R. & Boehnke, M. Quantifying and correcting for the winner's curse in genetic association studies. *Genet. Epidemiol.* **33**, 453–462 (2009).
8. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
9. Chinn, S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Stat. Med.* **19**, 3127–3131 (2000).
10. Barendregt, J. J., Doi, S. A., Lee, Y. Y., Norman, R. E. & Vos, T. Meta-analysis of prevalence. *J. Epidemiol. Commun. Health* **67**, 974–978 (2013).
11. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
12. Inouye, M. *et al.* Genomic risk prediction of coronary artery disease in 480,000 adults: Implications for primary prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).
13. Abraham, G. *et al.* Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat. Commun.* **10**, 5819 (2019).
14. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**, 185–194 (2021).
15. Gorlov, I. P. *et al.* SNP characteristics predict replication success in association studies. *Hum. Genet.* **133**, 1477–1486 (2014).

Author contributions

J.O.S. and J.I. conceptualized the study design, J.O.S. attained and curated the data curation, J.O.S. performed the formal analysis; J.I. supervised the study; J.O.S. and J.I. drafted and edited the manuscript.

Funding

There was no specific funding for this project. The lead author (JOS) was supported by an National Institutes of Health (NIH) T32 fellowship.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-97896-y.

Correspondence and requests for materials should be addressed to J.W.O.

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.