Human Genetics (2022) 141:229–238
https://doi.org/10.1007/s00439-021-02407-8

ORIGINAL INVESTIGATION

SNP characteristics and validation success in genome wide association studies

Olga Y. Gorlova[1] · Xiangjun Xiao[1] · Spiridon Tsavachidis[1] · Christopher I. Amos[1] · Ivan P. Gorlov[1]

Received: 6 August 2021 / Accepted: 27 November 2021 / Published online: 4 January 2022

Abstract

Genome wide association studies (GWASs) have identified tens of thousands of single nucleotide polymorphisms (SNPs) associated with human diseases and characteristics. A significant fraction of GWAS findings can be false positives. The gold standard for true positives is an independent validation. The goal of this study was to identify SNP features associated with validation success. Summary statistics from the Catalog of Published GWASs were used in the analysis. Since our goal was an analysis of reproducibility, we focused on the diseases/phenotypes targeted by at least 10 GWASs. GWASs were arranged in discovery-validation pairs based on the time of publication, with the discovery GWAS published before validation. We used four definitions of the validation success that differ by stringency. Associations of SNP features with validation success were consistent across the definitions. The strongest predictor of SNP validation was the level of statistical significance in the discovery GWAS. The magnitude of the effect size was associated with validation success in a non-linear manner. SNPs with risk allele frequencies in the range 30–70% showed a higher validation success rate compared to rarer or more common SNPs. Missense, 5'UTR, stop gained, and SNPs located in transcription factor binding sites had a higher validation success rate compared to intergenic, intronic and synonymous SNPs. There was a positive association between validation success and the level of evolutionary conservation of the sites. In addition, validation success was higher when discovery and validation GWASs targeted the same ethnicity. All predictors of validation success remained significant in a multivariate logistic regression model indicating their independent contribution. To conclude, we identified SNP features predicting validation success of GWAS hits. These features can be used to select SNPs for validation and downstream functional studies.

Introduction

Genome-wide association studies (GWASs) revolutionized the study of genetic control of human phenotypes and diseases (Tam et al. 2019; Visscher et al. 2012). GWASs test millions of SNPs in phenotypically different individuals to identify genotype–phenotype associations. Thousands of associations between SNPs and diseases/traits have been detected (Bosse and Amos 2018; Gallagher and Chen-Plotkin 2018; Horwitz et al. 2019; Liang et al. 2020). Despite using the strict genome-wide threshold for statistical significance ($p < 5 \times 10^{-8}$ or equivalently $-\log(p) > 7.3$), a considerable number of detected SNP-phenotype associations fail independent validation (Brzyski et al. 2017; Marigorta et al. 2018). Identifying SNP characteristics predicting validation success (true positives) is important for prioritizing SNPs for targeted validation and downstream functional studies. We and others identified a number of SNP characteristics associated with the validation success (Gorlov et al. 2014; Merelli et al. 2013; Xu and Taylor 2009).

Here we present results of an updated analysis of associations between SNP characteristics and validation success.

Materials and methods

Data used

We used data from the Catalog of Published GWASs (CPG) (Buniello et al. 2019). The catalog was accessed on May 12, 2021. We retrieved summary statistics for SNPs with the genome-wide level of statistical significance ($p < 5 \times 10^{-8}$) and gray zone SNPs ($10^{-5} < p < 5 \times 10^{-8}$). The latter were

✉ Olga Y. Gorlova
olga.gorlova@bcm.edu

[1] Department of Medicine, Institute for Clinical and Translational Research, Baylor College of Medicine, One Baylor Plaza, Mailstop: BCM451, Houston, TX 77030, USA

included to test if they are enriched by true positives. We focused on diseases/traits that were targeted by at least 10 studies. The results of the published GWASs were included regardless of the sample size used in the study. A total of 40 diseases/traits were analyzed in the study (Table 1). Diseases/traits' labels were used exactly how they were reported in the CPG.

Validation attempts

For each disease/trait, GWASs were arranged into pairs according to the publication date. Each pair was considered to be a validation attempt, with the earlier GWAS considered the discovery and the later, validation. The complete list of discovery-validation pairs can be found in Supplementary Table S1. The supplementary table also includes pairwise linkage disequilibrium (LD) for three major ethnic groups: Europeans, Africans and Asians.

Definitions of successful validation

We used four definitions of successful validation that differ by the stringency. Under the *strict* definition, a SNP was considered validated when the validation GWAS detected the same SNP at the genome-wide level of significance ($p < 5 \times 10^{-8}$). Under the *relaxed* definition, a SNP was considered validated when the same or a linked SNP ($r^2 > 0.8$ in the validation population) was detected at the genome-wide level of significance. LD data were downloaded from LDLink database (Myers et al. 2020). Under the *soft* definition, a SNP was considered validated if the original SNP or a SNP in tight LD with it was detected in the validation GWAS at the liberal level of significance of $p < 10^{-5}$. Finally, under the ultra-soft definition of validation success, a SNP was considered validated if the original SNP or a tightly linked SNP reached the GWAS level of significance in at least one out of at least three subsequent GWASs (attempts). Therefore, the principal difference of the ultra-soft definition of the validation success from the other three definitions is that for the latter, the validation success was per single attempt, while under the ultra-soft definition of the validation success, at least three validation attempts are required and the SNP is considered validated if at least one attempt is successful.

Predictors of the validation success

The following SNP characteristics were used as predictors of the validation success: (1) the level of statistical significance in the discovery GWAS expressed as $-\log(p)$, where p is the p value; (2) the effect size (either original odds ratios (OR) or transformed to 1/OR for ORs < 1 to keep them on same scale as ORs > 1); (3) risk allele frequency; (4) the type of the SNP (see below); (5) the level of evolutionary conservation of the site estimated by the PhyloP method (Pollard et al. 2010).

The PhyloP uses the distribution of nucleotide substitutions in an evolutionary tree of 44 vertebrate species to estimate the expected number of substitutions per site under the

Table 1 Diseases/traits with the corresponding numbers of conducted GWASs

N	Disease/trait	Number of studies
1	Type 2 diabetes	69
2	Body mass index	51
3	Breast cancer	39
4	HDL cholesterol levels	38
5	Schizophrenia	36
6	Colorectal cancer	33
7	Prostate cancer	32
8	Height	31
9	Diastolic blood pressure	29
10	Alzheimer's disease	25
11	Asthma	25
12	Rheumatoid arthritis	24
13	Parkinson's disease	23
14	Crohn's disease	22
15	Systemic lupus erythematosus	22
16	Bipolar disorder	21
17	Multiple sclerosis	20
18	Amyotrophic lateral sclerosis	19
19	Major depressive disorder	19
20	Ulcerative colitis	19
21	Hypertension	18
22	Coronary heart disease	16
23	Alcohol dependence	15
24	Glaucoma (primary open-angle)	15
25	Psoriasis	15
26	Type 1 diabetes	15
27	Bone mineral density	14
28	Intraocular pressure	14
29	Lung cancer	14
30	Telomere length	13
31	Adiponectin levels	12
32	Attention deficit hyperactivity disorder	12
33	Fasting plasma glucose	12
34	Glycated hemoglobin levels	12
35	Age-related macular degeneration	11
36	Atrial fibrillation	11
37	Bilirubin levels	11
38	QT interval	11
39	Venous thromboembolism	11
40	Pancreatic cancer	10

assumption of neutral evolution. The observed number of substitutions at the site is compared to the expected under the assumption of selective neutrality. A higher PhyloP score means a higher level of evolutionary conservation.

Statistical analysis and visualization

To visualize the associations of quantitative features, e.g. $-\log(p)$, with validation success, we stratified predictors by deciles. First we ranked SNPs by the corresponding characteristic and then stratified them into ten groups. Validation success rate was estimated for each group separately. To estimate and compare different types of SNPs by validation success we used SNP types reported by CPG. The list of the most frequent SNP types used in the analysis is as follows: "intron variant", "intergenic variant", "missense variant", "non-coding transcript exon variant", "3′UTR variant", "TF binding site variant", "5′UTR variant", and "synonymous variant". To estimate the effect of the same/different ethnicities in the discovery and the validation GWASs we used the CPG data. The most frequently reported ethnicities are Europeans, East Asians, African American, Hispanic/Latino and Ashkenazi Jews.

Initially associations were estimated using univariate analyses. Features significant in the univariate analyses were included in multivariable logistic regression. Validation status was treated as the outcome—validated/not-validated. All significant predictors were included into the model, to evaluate their independent effects. We present the two extreme definitions of validation success: ultra-soft and strict. The results of two other models were similar to the strict and ultra-soft models of validation success. All statistical analyses were performed using STATISTICA (TIBCO Software Inc.) and Origin (OriginLab Corporation, Northampton, MA, USA).

Results

Diseases/traits differ by the average validation success

Figure 1 shows validation success rates across diseases/traits. We observed a higher than an order of magnitude variation among the phenotypes. Those with lowest validation success rates included "Major depressive disorder", "Attention deficit hyperactivity disorder", "Bone mineral density", "Alcohol dependence", "Coronary heart disease" and "Bipolar disorder". Diseases/traits with highest validation success rate included "Breast cancer", "Asthma", "Venous thromboembolism", "QT interval', "Atrial fibrillation", and "Age-related macular degeneration".

Validation success rates for different definitions

The overall average validation success rate for SNPs across all phenotypes varied depending on the definition of the validation success: 6.42 ± 0.07% under the strict definition, 6.66 ± 0.07% under the relaxed definition, 7.87 ± 0.08% under the soft definition, and 50.87 ± 0.16 under the ultra-soft definition.

The level of statistical significance in the discovery GWAS is positively associated with validation success

We observed a strong positive association of $-\log(p)$ in the discovery GWAS with the validation success under all 4 definitions of the validations success. Figure 2 shows mean validation success rate for SNPs categorized by the deciles of $-\log(p)$ in the discovery study. The proportion of validated SNPs is positively associated with the $-\log(p)$ in the range of $-\log(p)$ 5–7.5; for higher $-\log(p)$ deciles the slope is less steep. Similar shapes were observed for all definitions of validation success including the ultra-soft definition that dwarfs validation rate for a single validation attempt (Figs. 2b, 3).

Odds ratios in the discovery GWAS and validation success

Overall negative correlations of the OR with SNP validation success were detected under strict, relaxed, soft, and ultra-soft definitions of the validation success (corresponding Spearman rank order correlations were $\rho = -0.03$, $N = 60{,}166$, $p = 7.1 \times 10^{-15}$, $\rho = -0.01$, $N = 60{,}166$, $p = 2.8 \times 10^{-3}$, $\rho = -0.02$, $N = 60{,}166$, $p = 5.1 \times 10^{-5}$, $\rho = -0.1$, $N = 57{,}352$, $p = 5.6 \times 10^{-25}$). The association between OR and validation success using decile stratification shows a more complex relationship. Highest validation success was for the SNPs with ORs in the range of 1.06–1.3, while the SNPs with reported ORs < 1.06 or > 1.3 had lower validation success. Details on the ranges of ORs for each of the ten groups defined by the deciles are shown in Supplementary Table S2.

Association between the risk allele frequency and validation success

Under strict, relaxed, and soft definition of the validation success there is a tendency for common risk-associated alleles (allelic frequency close to 0.5) to have a higher validation success rate (Fig. 4a). The association is more evident under the ultra-soft definition of the validation success (Fig. 4b). When we used MAF of the reported SNP (Fig. 5), we found that SNPs with MAF 0.3–0.5 tended to

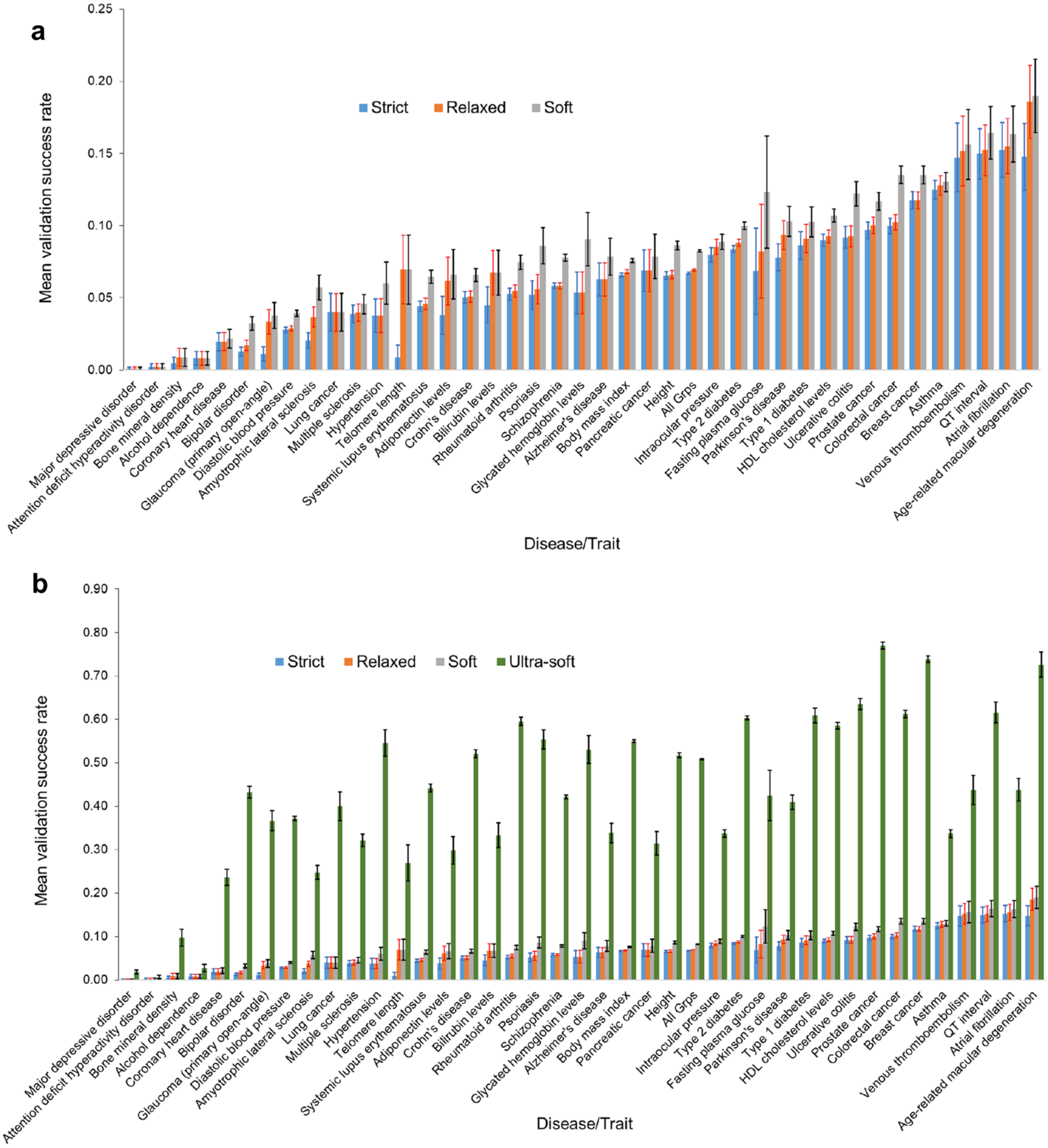

Fig. 1 Validation rate for different diseases/traits across all SNPs. Vertical black bars show 95% confidence intervals. Phenotypes are arranged from lowest to highest rate under strict, relaxed, and soft definitions of the validation success (**a**). **b** The same plus ultra-soft definition of validation success

be validated more often compared to rarer or more common SNPs.

Different types of SNPs differ by validation success

We compared different types of SNPs by validation success (Fig. 6). Intergenic and intron variants had the lowest validation success rate. Validation success rate of SNPs producing missense mutations, stop gained, located in TF binding sites or 5' UTR were the highest.

Validation success is higher for SNPs located in evolutionary conserved sites

We found positive correlations between level of evolutionary conservation—PhyloP score and validation success under strict, relaxed, soft, and ultra-soft definitions of the validation success (corresponding Spearman rank order correlations were $\rho = 0.04$, $N = 125{,}087$, $p < 10^{-25}$; $\rho = 0.04$, $N = 125{,}087$, $p < 10^{-25}$; $\rho = 0.04$, $N = 125{,}087$, $p < 10^{-25}$; $\rho = 0.07$, $N = 117{,}643$, $p < 10^{-25}$). Figure 7 shows

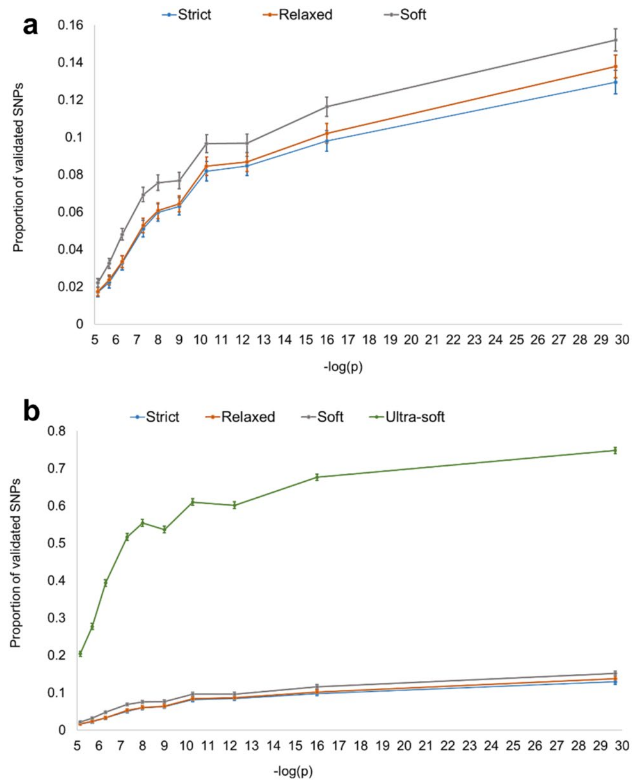

Fig. 2 **a** The proportion of validated SNPs in categories of $-\log(p)$ in the discovery GWAS stratified by deciles. *X*-coordinates of the dots represent the median $-\log(p)$ in each group. Bars represent 95% CI for the proportion of validated SNPs. **a** Strict, relaxed, and soft definition of the validation success. **b** The same as **a** plus the ultra-soft definition of the validation success

that SNPs with evidence of evolutionary conservation are more likely to be validated.

SNPs are more likely to be validated when the same race/ethnicity is targeted by discovery and validation GWASs

When the discovery and the validation GWASs target the same race/ethnicity, the validation success rate is higher compared to the situation when the ethnicities in the discovery and validation GWASs are different. This is true regardless of the definition of the validation success (Fig. 8).

Multivariate logistic regression analysis

We analyzed the predictors simultaneously using binary logistic regression model with validation status as the outcome. All predictors remained significant, for both the strict and ultra-soft definitions of validation (Tables 2, 3).

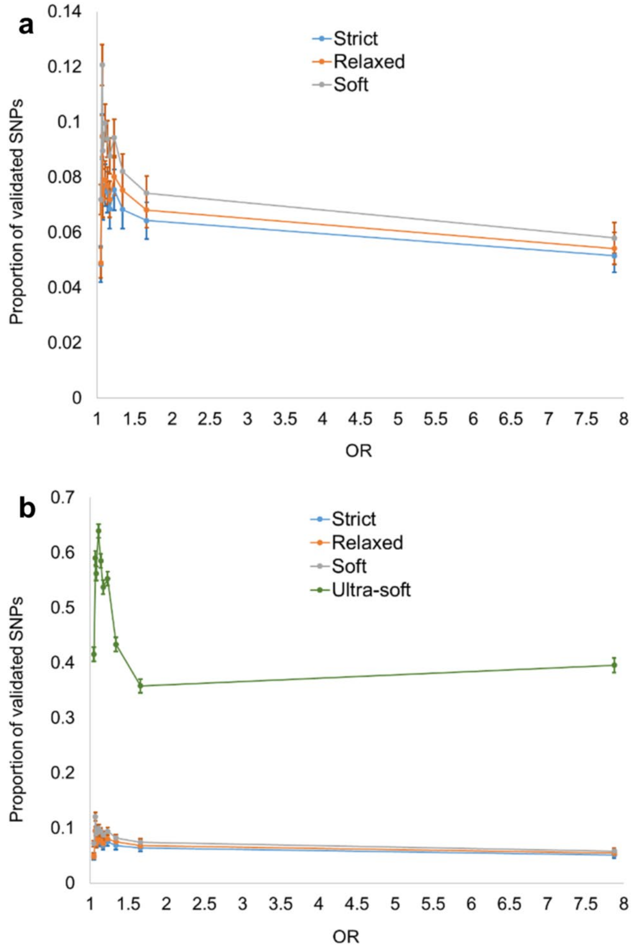

Fig. 3 **a** The proportion of validated SNPs in categories of ORs in the discovery GWAS stratified by deciles (Supplementary Table S2). *X*-coordinates of the dots represent the median OR in each group. The bars represent 95% CI for the proportion of validated SNPs. **a** Strict, relaxed, and soft definition of the validation success. **b** The same as **a** plus ultra-soft definition of the validation success

Discussion

Compared to our previous study (Gorlov et al. 2014), the current analysis is based on a larger sample size and includes more predictors. We confirmed the previous associations and added new ones. Validation success rate per single validation attempt was similar for the strict, relaxed and soft definitions in the range of 6–8%. One of the possible reasons for the low validation rate for SNPs could be that our analysis included gray zone SNPs. However, when such SNPs were excluded from the analysis, the validation success increased only marginally for strict definition, from 6.42 ± 0.07% to 6.46 ± 0.08%. A similarly slight increase in validation success after removing gray zone SNPs was observed for the relaxed and soft definitions of validation. It is unlikely, also, that differences in genotyping platforms are a major contributor to the low validation success. By

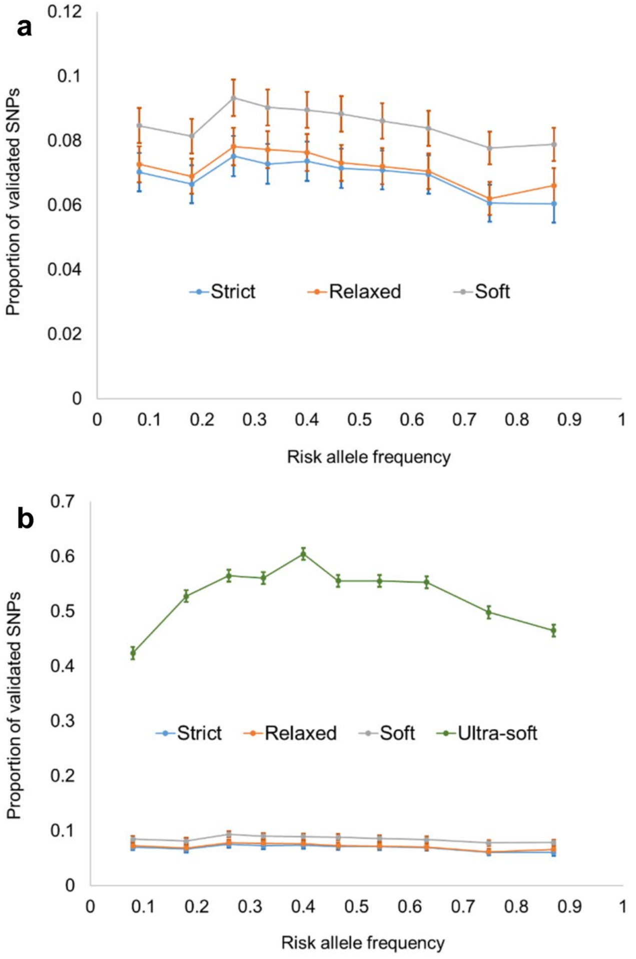

Fig. 4 **a** The proportion of validated SNPs in groups of risk allele frequency in the discovery GWAS stratified by deciles. *X*-coordinates of the dots represent the median risk allele frequency in each group. The bars represent 95% CI for the proportion of validated SNPs. **a** Strict, relaxed, and soft definition of the validation success. **b** The same as **a** plus ultra-soft definition of the validation success

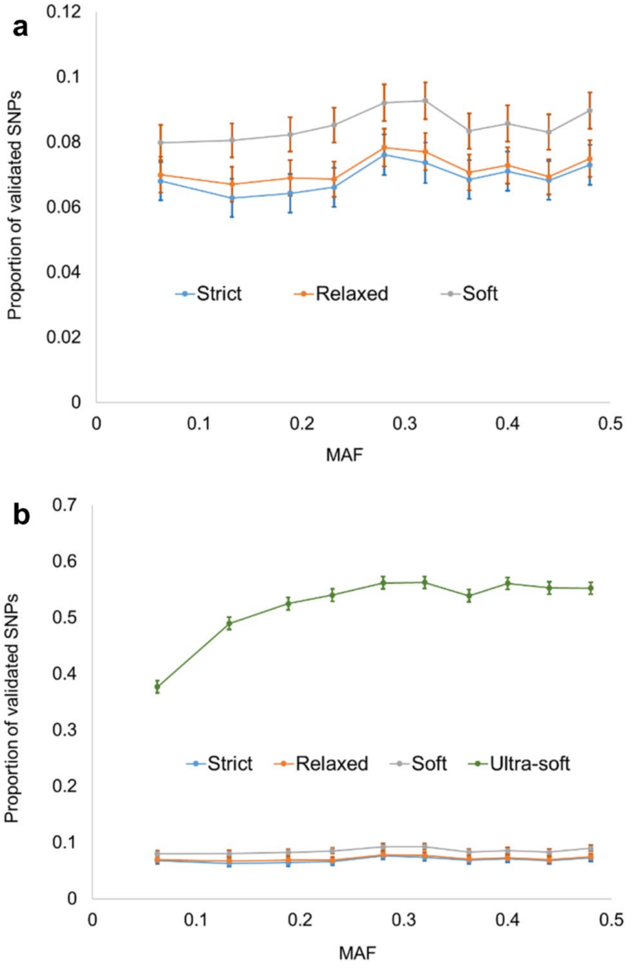

Fig. 5 **a** The proportion of validated SNPs in categories of minor risk allele frequency in the discovery GWAS stratified by deciles. *X*-coordinates of the dots represent the median MAF in each group. The bars represent 95% CI for the proportion of validated SNPs. **a** Strict, relaxed, and soft definition of the validation success. **b** The same as **a** plus ultra-soft definition of the validation success

definition, validation GWASs are performed later than discovery, and later GWASs tend to use denser genotyping platforms. Besides, the investigators usually impute SNPs that were detected as significant earlier if they were not on the genotyping platform (Li et al. 2009; Shi et al. 2018). We found that having the same genotyping platform increased chances of validation only by 1.1%.

Ethnicity is an important factor to consider in GWASs since many SNPs show significant population variation. However, to assess the external validity of the associations of SNPs with diseases/traits, we used validation studies that did not exactly match the characteristics of the corresponding discovery studies. Thus, the effect of race on SNP reproducibility was one of the factors we wanted to explore. We showed that targeting the same ethnicity in the discovery and validation GWASs has a profound effect on validation success (about two-fold, Fig. 8). This indicates that targeting genetically similar populations is important for successful validation. Note that we have used only major population categories: Europeans, East Asians, African American, Hispanic/Latino and Ashkenazi Jews. The major population groups are genetically heterogeneous. There are, for example, significant genetic differences among European subpopulations, which also can impact reproducibility (Lao et al. 2008).

Not surprisingly, the level of statistical significance in the discovery GWAS was the strongest predictor of the validation success. The association between validation success and OR was markedly nonlinear. The highest validation success rate was in the group of SNPs with ORs in the range from 1.1 to 1.3, suggesting that “real” ORs tend to be within this range. Compared to these, SNPs with ORs > 2 in the

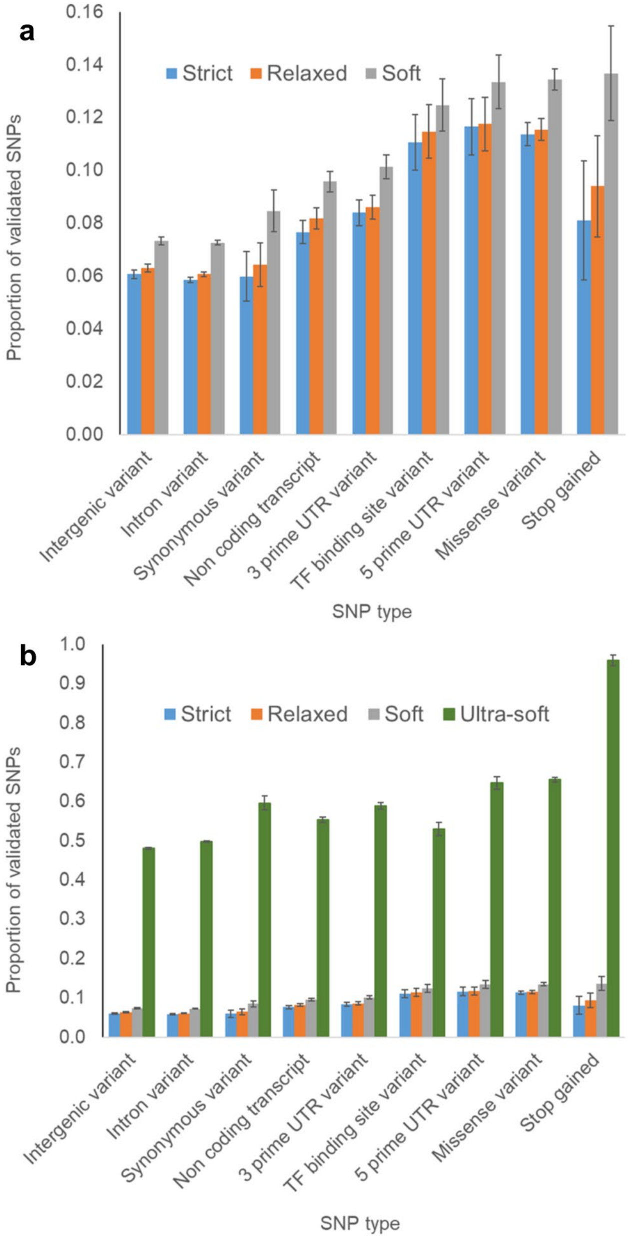

Fig. 6 Validation rate for different types of SNPs. **a** Strict, relaxed, and soft definition of the validation success. **b** The same as **a** plus ultra-soft definition of the validation success. Vertical bars show 95% CI

discovery GWAS are validated 40% less likely. This can be because the initial discoveries tend to overestimate the effect sizes—a "winner's curse" (Lohmueller et al. 2003; Shi et al. 2016; Xiao and Boehnke 2011). Validation success rate was highest for most polymorphic SNPs, likely because statistical power is the highest for the SNPs with a frequency close to 0.5 (Hong and Park 2012; Sham and Purcell 2014).

Intronic, intergenic and synonymous SNPs showed lower validation rates compared to the missense SNPs, SNPs located in TF binding sites or in 5'UTR regions. The

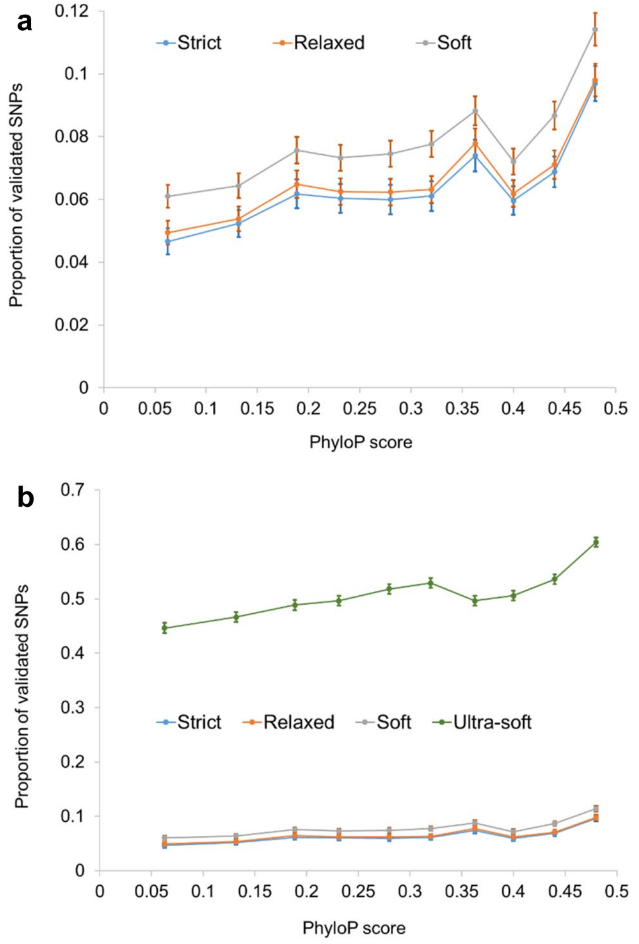

Fig. 7 **a** The proportion of validated SNPs in categories of PhyloP score stratified by deciles. *X* coordinates of the dots represent the median PhyloP score in each group. The bars represent 95% CI for the proportion of validated SNPs. **a** Strict, relaxed, and soft definition of the validation success. **b** The same as **a** plus ultra-soft definition of the validation success

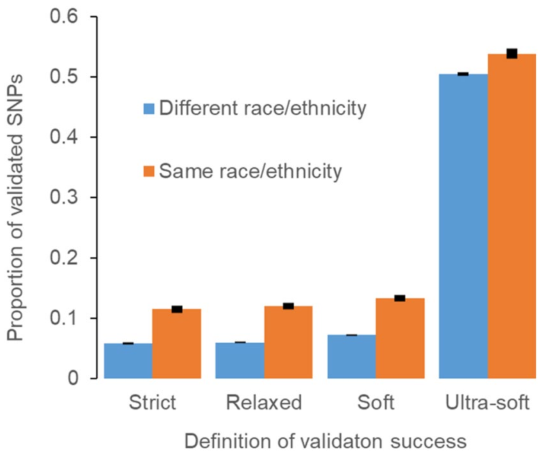

Fig. 8 The proportion of validated SNPs under strict, relaxed, soft and ultra-soft definitions of the validation success. Vertical bars show 95% CI

Table 2 Multivariable prediction of SNP validation success in GWASs (strict definition of validation)

Predictor	*p* value	OR	95% CI	
			Lower	Upper
SNP MAF	0.01233	1.494	1.091	2.046
− log *p*-value at discovery	1.67×10^{-60}	1.013	1.012	1.015
Different population in validation and discovery[a]	1.061×10^{-93}	0.390	0.356	0.427
PhyloP score	5.647×10^{-41}	1.202	1.170	1.235
OR groups stratified by deciles[b] 1 (1.05) reference				
2 (1.07)	4.699×10^{-4}	0.745	0.632	0.879
3 (1.08)	0.01359	0.833	0.721	0.963
4 (1.11)	3.456×10^{-3}	0.778	0.657	0.921
5 (1.14)	8.721×10^{-3}	0.558	0.472	0.660
6 (1.17)	8.752×10^{-5}	0.718	0.608	0.847
7 (1.23)	2.305×10^{-7}	0.665	0.569	0.776
8 (1.34)	3.642×10^{-8}	0.627	0.531	0.741
9 (1.66)	6.647×10^{-7}	0.652	0.551	0.772
10 (7.87)	2.022×10^{-15}	0.474	0.395	0.570
SNP type categories[c]: likely non-functional, reference				
Other	0.1435	1.089	0.971	1.222
Likely functional	4.233×10^{-22}	1.854	1.636	2.101

[a]Reference, the same population in the discovery and validation

[b]Median ORs for each group are shown in parenthesis

[c]Non-functional: intergenic, synonymous, intronic; functional: 5′ UTR, missense, nonsense, located in transcription factor binding sites; other—non-coding exonic, 3′ UTR

Table 3 Multivariable prediction of SNP validation success in GWASs (ultra-soft definition of validation)

Predictor	*p* value	OR	95% CI	
			Lower	Upper
SNP MAF	1.76×10^{-20}	2.415	2.005	2.91
− log *p* value at discovery	1.81×10^{-188}	1.078	1.073	1.082
Different population in validation and discovery[a]	9.34×10^{-11}	0.808	0.757	0.862
PhyloP score	$3.00\text{E} \times 10^{-54}$	1.164	1.142	1.187
OR groups stratified by deciles[b] 1 (1.05), reference				
2 (1.07)	1.1×10^{-3}	1.172	1.063	1.292
3 (1.08)	1.74×10^{-17}	1.491	1.36	1.635
4 (1.11)	0.011	1.146	1.032	1.273
5 (1.14)	3.82×10^{-8}	0.771	0.702	0.846
6 (1.17)	0.75	1.017	0.92	1.123
7 (1.23)	3.317×10^{-10}	0.741	0.674	0.813
8 (1.34)	6.999×10^{-91}	0.481	0.437	0.529
9 (1.66)	8.602×10^{-118}	0.294	0.265	0.326
10 (7.87)	1.848×10^{-43}	0.488	0.441	0.541
SNP type categories[c]: likely non-functional, reference				
Other	1.170×10^{-43}	0.776	0.726	0.83
Likely functional	2.165×10^{-54}	2.204	1.994	2.435

[a]Reference, the same population in the discovery and validation

[b] Median ORs for each group are shown in parenthesis

[c]Non-functional: intergenic, synonymous, intronic; functional: 5′ UTR, missense, nonsense, located in transcription factor binding sites; other—non-coding exonic, 3′ UTR

most likely explanation for this can be that some GWAS-detected SNPs are causal (Caballero et al. 2015; Schaid et al. 2018; Wang et al. 2020). Functional SNPs affect the level of expression and/or protein function, including protein folding. Missense SNPs and SNPs located in TF binding sites or 5’UTR regions (often loaded with regulatory elements) are likely to be functional (Buroker 2014; Huo et al. 2019; Lou et al. 2017).

It is accepted that the level of evolutionary conservation of the site reflects its functional importance (O'Connor et al. 2019; Zeng et al. 2018) suggesting that the positive association between the level of evolutionary conservation of the site and replication success that we found is due to the presence of functional causal SNPs among GWAS top hits.

All predictors of validation success detected in univariate analysis remained significant in the multivariate logistic regression analysis. The most significant predictors in multivariate analysis were the level of statistical significance in the discovery, followed by SNP type and PhyloP score (Tables 2, 3).

The results of this study suggest that SNP features may help to select SNPs with highest chances to be validated. Indeed, when we selected SNPs based on the five major predictors of validation success, as follows: (1) the SNP is genome-wide significant in the discovery GWAS; (2) the risk allele frequency is between 0.1 and 0.9; (3) the SNP is missense, or is located in a TF binding site or in 5’UTR region; (4) the SNP has a high level of evolutionary conservation, and (5) the same ethnicity in the discovery and validation GWASs, the resulting SNPs showed validation success rate of 32.6% ± 5.8 under the strict definition, which is much higher than the overall average.

Surprisingly, we found that the validation success rate of the gray zone SNPs ($10^{-5} < p < 5 \times 10^{-8}$) was inferior but still comparable to that of SNPs with genome-wide level of statistical significance in the discovery: 4.19 ± 0.09% versus 12.26 ± 0.08% under the strict definition of the validation success (Fig. 2 first 4 points versus other points). This indicates that gray SNPs are enriched by true positives.

Limitations of the study

Subsequent GWASs targeting the same phenotype were considered in this study as an independent validation. That is not always the case. In some cases the subsequent GWASs include a subset of samples already used in an earlier GWAS, which is likely to inflate the validation success rate. We do not think, however, that this issue substantially affects the findings on associations between the SNP characteristics and validation rate. Besides, based on our experience with lung cancer GWASs and a limited review of published GWASs, we believe that a typical overlap (if exists) does not exceed 20%. We found that the associations were very similar across different definitions of validation success. Another limitation is that we did not handle the meta-analysis studies any differently from standard two-phase GWASs. We formally followed the classification adopted by the Catalog of the Published GWASs because it reflects the current state of knowledge of disease etiology. We acknowledge that disease classification is a moving target and a disease once considered genetically homogeneous may be later reclassified into several distinct diseases as it becomes studied better.

When assessing the effect of different ethnicity in the discovery and validation GWASs on validation success, we did not take into account possible effects of differences in allele frequency between discovery and validation GWAS on statistical power. For non-validated SNPs risk allele frequency in validation GWASs are not available. This precluded us from using differences in MAFs between discovery and validation GWASs as a predictor in the multivariate model, which could have helped to decide whether the effect of ethnicity on validation success is due to the differences in allele frequencies between the discovery and validation populations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00439-021-02407-8.

Funding Partial financial support was received from National Institutes of Health Grants U19CA203654, U19CA203654S1, R01CA231141, and P01 CA206980-01A1, Cancer Prevention and Research Institute of Texas Grant RR170048. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials The data from A Catalog of Published Genome-Wide Association Studies https://www.genome.gov/catalog-of-published-genomewide-association-studies, UCSC Human Genome Browser https://genome.ucsc.edu, The Ensembl Regulatory Build http://useast.ensembl.org/info/genome/funcgen/regulatory_build.html, and ENCODE https://www.encodeproject.org/, all in the public domain, were used in this project.

Code availability Not applicable.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics approval Not applicable: the study used aggregate statistics from datasets in the public domain.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Bosse Y, Amos CI (2018) A decade of GWAS results in lung cancer. Cancer Epidemiol Biomark Prev 27(4):363–379. https://doi.org/10.1158/1055-9965.EPI-16-0794

Brzyski D, Peterson CB, Sobczyk P, Candes EJ, Bogdan M, Sabatti C (2017) Controlling the rate of GWAS false discoveries. Genetics 205(1):61–75. https://doi.org/10.1534/genetics.116.193987

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, Parkinson H et al (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47(D1):D1005–D1012. https://doi.org/10.1093/nar/gky1120

Buroker NE (2014) Regulatory SNPs and transcriptional factor binding sites in ADRBK1, AKT3, ATF3, DIO2, TBXA2R and VEGFA. Transcription 5(4):e964559. https://doi.org/10.4161/21541264.2014.964559

Caballero A, Tenesa A, Keightley PD (2015) The nature of genetic variation for complex traits revealed by GWAS and regional heritability mapping analyses. Genetics 201(4):1601–1613. https://doi.org/10.1534/genetics.115.177220

Gallagher MD, Chen-Plotkin AS (2018) The post-GWAS era: from association to function. Am J Hum Genet 102(5):717–730. https://doi.org/10.1016/j.ajhg.2018.04.002

Gorlov IP, Moore JH, Peng B, Jin JL, Gorlova OY, Amos CI (2014) SNP characteristics predict replication success in association studies. Hum Genet 133(12):1477–1486. https://doi.org/10.1007/s00439-014-1493-6

Hong EP, Park JW (2012) Sample size and statistical power calculation in genetic association studies. Genomics Inform 10(2):117–122. https://doi.org/10.5808/GI.2012.10.2.117

Horwitz T, Lam K, Chen Y, Xia Y, Liu C (2019) A decade in psychiatric GWAS research. Mol Psychiatry 24(3):378–389. https://doi.org/10.1038/s41380-018-0055-z

Huo Y, Li S, Liu J, Li X, Luo XJ (2019) Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. Nat Commun 10(1):670. https://doi.org/10.1038/s41467-019-08666-4

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Kayser M et al (2008) Correlation between genetic and geographic structure in Europe. Curr Biol 18(16):1241–1248. https://doi.org/10.1016/j.cub.2008.07.049

Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242

Liang B, Ding H, Huang L, Luo H, Zhu X (2020) GWAS in cancer: progress and challenges. Mol Genet Genomics 295(3):537–561. https://doi.org/10.1007/s00438-020-01647-z

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33(2):177–182. https://doi.org/10.1038/ng1071

Lou J, Gong J, Ke J, Tian J, Zhang Y, Li J, Miao X et al (2017) A functional polymorphism located at transcription factor binding sites, rs6695837 near LAMC1 gene, confers risk of colorectal cancer in Chinese populations. Carcinogenesis 38(2):177–183. https://doi.org/10.1093/carcin/bgw204

Marigorta UM, Rodriguez JA, Gibson G, Navarro A (2018) Replicability and prediction: lessons and challenges from GWAS. Trends Genet 34(7):504–517. https://doi.org/10.1016/j.tig.2018.03.005

Merelli I, Calabria A, Cozzi P, Viti F, Mosca E, Milanesi L (2013) SNPranker 2.0: a gene-centric data mining tool for diseases associated SNP prioritization in GWAS. BMC Bioinform 14(Suppl 1):S9. https://doi.org/10.1186/1471-2105-14-S1-S9

Myers TA, Chanock SJ, Machiela MJ (2020) LDlinkR: an R package for rapidly calculating linkage disequilibrium statistics in diverse populations. Front Genet 11:157. https://doi.org/10.3389/fgene.2020.00157

O'Connor LJ, Schoech AP, Hormozdiari F, Gazal S, Patterson N, Price AL (2019) Extreme polygenicity of complex traits is explained by negative selection. Am J Hum Genet 105(3):456–476. https://doi.org/10.1016/j.ajhg.2019.07.003

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20(1):110–121. https://doi.org/10.1101/gr.097857.109

Schaid DJ, Chen W, Larson NB (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet 19(8):491–504. https://doi.org/10.1038/s41576-018-0016-z

Sham PC, Purcell SM (2014) Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet 15(5):335–346. https://doi.org/10.1038/nrg3706

Shi J, Park JH, Duan J, Berndt ST, Moy W, Yu K, Chatterjee N et al (2016) Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. PLoS Genet 12(12):e1006493. https://doi.org/10.1371/journal.pgen.1006493

Shi S, Yuan N, Yang M, Du Z, Wang J, Sheng X, Xiao J et al (2018) Comprehensive assessment of genotype imputation performance. Hum Hered 83(3):107–116. https://doi.org/10.1159/000489758

Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D (2019) Benefits and limitations of genome-wide association studies. Nat Rev Genet 20(8):467–484. https://doi.org/10.1038/s41576-019-0127-1

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90(1):7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

Wang J, Huang D, Zhou Y, Yao H, Liu H, Zhai S, Li MJ et al (2020) CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies. Nucleic Acids Res 48(D1):D807–D816. https://doi.org/10.1093/nar/gkz1026

Xiao R, Boehnke M (2011) Quantifying and correcting for the winner's curse in quantitative-trait association studies. Genet Epidemiol 35(3):133–138. https://doi.org/10.1002/gepi.20551

Xu Z, Taylor JA (2009) SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic Acids Res 37(Web Server issue):W600-605. https://doi.org/10.1093/nar/gkp290

Zeng J, de Vlaming R, Wu Y, Robinson MR, Lloyd-Jones LR, Yengo L, Yang J et al (2018) Signatures of negative selection in the genetic architecture of human complex traits. Nat Genet 50(5):746–753. https://doi.org/10.1038/s41588-018-0101-4