Published in final edited form as:

Hum Genet. 2014 December ; 133(12): 1477–1486. doi:10.1007/s00439-014-1493-6.

SNP CHARACTERISTICS PREDICT REPLICATION SUCCESS IN ASSOCIATION STUDIES

Ivan P. Gorlov^{1,*}, Jason H. Moore², Bo Peng³, Jennifer L. Jin⁴, Olga Y. Gorlova¹, and Christopher I. Amos¹

¹Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, 74 College Street Vail 7th Floor, HB 7260 Vail, Hanover, NH 03755

²The Geisel School of Medicine, Dartmouth College, HB7937, One Medical Center Dr., Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756

³Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, 1400 Pressler St, Unit 1410, Houston, TX, 77030

⁴Department of Mathematics, Dartmouth College, HB 1819, Hanover, NH 03755

Abstract

Successful independent replication is the most direct approach for distinguishing real genotype-disease associations from false discoveries in Genome Wide Association Studies (GWAS). Selecting SNPs for replication has been primarily based on p-values from the discovery stage, although additional characteristics of SNPs may be used to improve replication success.

We used disease-associated SNPs from more than 2,000 published GWASs to identify predictors of SNP reproducibility. SNP reproducibility was defined as a proportion of successful replications among all replication attempts. The study reporting association for the first time was considered to be discovery and all consequent studies targeting the same phenotype replications. We found that $-\text{Log}(P)$, where P is a p-value from the discovery study, is the strongest predictor of the SNP reproducibility. Other significant predictors include type of the SNP (e.g. missense vs intronic SNPs) and minor allele frequency. Features of the genes linked to the disease-associated SNP also predict SNP reproducibility.

Based on empirically defined rules, we developed a reproducibility score (RS) to predict SNP reproducibility independently of $-\text{Log}(P)$. We used data from two lung cancer GWAS studies as well as recently reported disease-associated SNPs to validate RS. Minus $\text{Log}(P)$ outperforms RS when the very top SNPs are selected, while RS works better with relaxed selection criteria. In conclusion, we propose an empirical model to predict SNP reproducibility, which can be used to select SNPs for validation and prioritization.

*Correspondence: Ivan Gorlov: Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, HB 7260 Vail, Hanover, NH 03755, Tel: 603-653-3524, ivan.p.gorlov@dartmouth.edu.

Introduction

Genome-wide association studies (GWAS) are extensively used to identify single-nucleotide polymorphisms (SNPs) associated with disease risk (Visscher, Brown, McCarthy, & Yang, 2012). Only a fraction (1–4%) of SNPs detected by candidate gene association studies has been replicated by GWAS (Ioannidis, Tarone, & McLaughlin, 2011; Siontis, Patsopoulos, & Ioannidis, 2010). A typical GWAS includes discovery and replication stage (Kraft & Cox, 2008). Replication is considered to be a gold standard for validation of GWAS detected associations (Greenwood, Rangrej, & Sun, 2007; Kraft, Zeggini, & Ioannidis, 2009). Typically SNPs are selected for replication based on P-values from the discovery phase (Hakonarson & Grant, 2011; Smith & Newton-Cheh, 2009). Usually only a fraction of SNPs from the discovery GWAS is replicated in the replication phase. Our analysis of GWAS studies published in Nature Genetics showed a typical inside study replication rate of 1–5% with the replication rate negatively correlated with the number of SNPs selected for replication. Identification of SNPs with a highest prior probability to be replicated would allow targeting a smaller set of SNPs for replication, reducing the number of statistical tests and increasing power. Because reproducibility rate for SNPs selected solely on p-values is low, additional predictors of SNP reproducibility are needed.

Published studies suggest that a SNP's characteristics influence its chance to be detected and replicated by GWAS. GWAS-detected SNPs tend influence quantitative trait loci (eQTLs) (Cookson, Liang, Abecasis, Moffatt, & Lathrop, 2009; Nicolae et al., 2010). Exonic SNPs and SNPs located in 5' gene regions are detected more often in GWA studies than expected by chance (Schork et al., 2013).

Thousands genome-wide association studies have been conducted and the results are reported in several databases. One of the most comprehensive databases is the catalogue of published GWASs (CPGWAS) (Hindorff et al., 2009) (<http://www.genome.gov/gwastudies/>). More than 7,000 SNP associated with nearly 5,000 genes have been reported in CPGWAS making it a valuable resource for studying SNP reproducibility. The goal of our study was to identify SNP features associated with SNP reproducibility.

MATERIALS AND METHODS

Data sources

Disease associated SNPs used in this analysis came from the Catalogue of Published Genome-Wide Association Studies (CPGWAS) (<http://www.genome.gov/26525384/>). The database was accessed on September 15, 2013. CPGWAS provides information on studied disease/phenotype. The database also lists SNPs identified by the study, as well as linked genes (Hindorff et al., 2009). In this study we focused on SNPs associated with human diseases. The major reason for this was practical significance. Another reason was that characteristics of the disease-associated SNPs may be different from those of the SNPs associated with morphological traits (Alfoldi & Lindblad-Toh, 2013) or quantitative traits. Only SNPs mapped to a single gene were used in the analysis. Supplementary Table S1 provides the list of the GWAS-detected disease risk associated SNPs from CPGWAS used in this study.

SNP reproducibility

In total, 106 different diseases targeted by GWASs are reported by CPGWAS. Different diseases are targeted by GWASs a different number of times: for example, there were 29 GWASs on the type 2 diabetes and one GWAS on neonatal lupus. Seventy four diseases were targeted at least twice. SNPs detected in those studies can be used to assess SNP reproducibility. The list of the human diseases targeted by GWASs and with at least two reported studies can be found in supplementary Table S2.

To estimate SNP reproducibility we first sorted GWASs targeting the same phenotype by publication date. A study reporting a SNP for the first time was considered as a discovery. Subsequent studies were considered as replications. We consider consequent studies as replication studies because subsequent studies are likely to include earlier reported SNPs on their genotyping platforms. They also tend to be larger, both in terms of sample size and the number of genotyped SNPs. SNP reproducibility was computed as the ratio of the number of successes (times the SNP was replicated) to the number of attempts at replication, with replication defined as obtaining a p-value less than 5×10^{-8} in the subsequent study. Therefore, SNP reproducibility varied from 0 to 1. SNP reproducibility was weighted by the number of replication studies: each case was repeated n times in analysis: where n is the number of validation studies.

Variables used as predictors of SNP reproducibility

Based on published evidence as well as results of our pilot analyses we preselected 13 variables that could influence SNP reproducibility. Table 1 shows the list of the variables. The variables can be divided into SNP-related and gene-related groups. SNP-related predictors included $-\log(P)$ from the discovery study, minor allele frequency (MAF), SNP type, e.g. intronic or missense mutations, and SNP eQTL status (whether the SNP is reported among eQTLs detected on HapMap samples). SNP type was defined as it was done in CPGWAS: (listed in alphabetic order): 3' UTR, 3' Downstream, 5' Upstream, 5' UTR, Coding nonsynonymous, Coding synonymous, Intergenic, Intronic, Non-coding, and Non-coding intronic. Details on the definition of SNP types can be found in the paper by Ullah et al. (Dayem Ullah, Lemoine, & Chelala, 2013). Gene/protein-related traits included gene size, cellular localization of the gene product, expression pattern (tissue specific versus not), type of encoded protein, e.g. transcription factor, and the level of evolutionary conservation. Data on gene sizes were retrieved from the Gene database (<http://www.ncbi.nlm.nih.gov/>). We used HomoloGene database (Coordinators, 2014) (<http://www.ncbi.nlm.nih.gov/homologene>) to assess the level of evolutionary conservation – conservation Index (CI). CI was defined based on the most distant detectable ortholog of the human gene. The CI changes from 1 (gene is unique to *Homo sapiens*) to 10 (ortholog of the human gene is detected in at least one plant species; see our recent paper (Gorlova, Fedorov, Logothetis, Amos, & Gorlov, 2014) for details). The CI we used was similar to the one used by Domazet-Loso and Tautz (Domazet-Loso & Tautz, 2008). The Tissue-specific Gene Expression and Regulation (TiGER) database (Liu, Yu, Zack, Zhu, & Qian, 2008): <http://bioinfo.wilmer.jhu.edu/tiger/> was used to identify genes having a tissue specific pattern of the gene expression. Genes reported as tissue specific in TiGER were also considered to be

tissue specific in this analysis. Gene categories, e.g. kinases, were retrieved from the Gene Ontology (GO) database (Balakrishnan, Harris, Huntley, Van Auken, & Cherry, 2013).

We estimated the association between SNP reproducibility and individual predictors listed in Table 1. One-way Analyse of Variance (ANOVA) was used in univariate analysis. Scatterplots were utilized to visualize association between individual predictors and SNP reproducibility. Main effects ANOVA was used to identify a subset of variables independently predicting SNP reproducibility.

An important goal of this study was to develop a simple, easily implemented predictor of SNP reproducibility. We constructed an empirical reproducibility score (RS) to simplify the prediction of reproducibility. We dichotomized each significant predictor into “0” or “1” with ones assigned to the categories associated with increased reproducibility. For example, for the conservation index (CI), which has 10 categories (Figure 2.10) one was assigned to category CI=1 and zero to all others because having CI=1 is associated with higher SNP reproducibility. The details of the scoring system we used are shown in last column of Table 3. To compute RS we summarized individual scores for the 9 significant predictors. For the SNP type individual score can take 3 values: 0, 1 and 2 (Table 3). Therefore RS is an integer between 0 and 10, including 0 and 10. RS can be easily computed for any SNP in the human genome. Detailed instructions on computing of RSs are shown in supplementary materials: “Computing Reproducibility Score”.

Validation of RS model

To validate the model we used lung cancer GWAS results generated at M. D. Anderson Cancer Center (MDACC) (Amos et al., 2008) and International Agency for Research on Cancer (IARC) (Hung et al., 2008). The results of these two studies were not used to build the prediction model. We assumed that if the model is effective there should be an inflation of $-\text{Log}(P)$ among SNPs with high RSs (6 and higher).

For additional validation of the model we used disease-associated SNPs that were reported in the catalogue of published GWASs after we completed the analysis. One can expect that those novel, GWAS-detected SNPs will have a higher reproducibility score.

RESULTS AND DISCUSSION

Univariate analysis

Figure 1 shows the relationship between $-\text{Log}(P)$ and SNP reproducibility. SNPs were categorized into 12 groups based on $-\text{Log}(P)$ reported for discovery studies. We observed a linear relationship between $-\text{Log}(P)$ in the discovery study and SNP reproducibility. Mean reproducibility rate for SNPs with $-\text{Log}(P)$ between 5 and 6 was 0.006 ± 0.001 while reproducibility rate for SNPs with $-\text{Log}(P) > 20$ was much higher: 0.109 ± 0.006 . Table 2 shows the results of statistical analyses for other predictors. The most significant association was detected for SNP type. Other significant predictors include linked gene being a kinase reported in OMIM database. Table 3 provides details on coding of predictors for main effects ANOVA and for computing of reproducibility score.

Figure 2 provides details on the shape of association between predictors and SNP reproducibility. For gene related traits we used three types of categorical values: “yes”, “no” and “NA”. “NA” was used for SNPs that cannot be linked to any specific gene - intergenic SNPs. We found that SNP reproducibility strongly depends on the SNP type (Figure 2.13). We observed a significant difference in reproducibility between synonymous and nonsynonymous SNPs: 0.043 ± 0.006 for synonymous and 0.091 ± 0.006 for nonsynonymous substitutions: Mann-Whitney test, $Z=3.06$, $P = 0.002$. It is generally accepted that the majority of the SNPs detected by GWAS are tagging SNPs. One can expect that tagging efficacy will be similar for missense and synonymous SNPs, suggesting that reproducibility for missense and synonymous SNPs will be similar if they are mostly tagging SNPs. Significantly higher reproducibility rate for missense SNPs suggests a higher proportion of causal variants among them compared to synonymous SNPs, making them a lucrative target to identify causal variants (Lewis & Tomlinson, 2012; Marjoram, Zubair, & Nuzhdin, 2014).

Selecting SNPs to construct Reproducibility Score (RS)

We used main effects ANOVA to identify variables predicting SNP reproducibility independently of $-\log(P)$. The results of the analysis are shown in Table 4. The most significant predictor of SNP reproducibility was $-\log(P)$. Nine out of 13 predictors were also significant: SNP Type, MAF, OMIM, Nuclear Localization, Gene Size (200kb increment), Kinases, Conservation Index, Growth Factors, and Tissue Specific.

We used two measures of SNP reproducibility. The first one – SNP reproducibility was estimated based on the results of published GWASs. SNP reproducibility is computed as a ratio of successful replications to the total number of replication attempts. SNP reproducibility is available only for GWAS-reported SNPs. We used SNP reproducibility to identify its predictors and to build the RS model. Unlike SNP reproducibility, RS can be computed for any SNP in the human genome by using the algorithm described in supplementary materials. Both SNP reproducibility and RS are available for the GWAS detected SNPs. Figure 3a shows the relationship between RS and SNP reproducibility. The dependence between the two variables is exponential: there is no difference between the group of RS 0–1 and the group of RS 2–3: Mann-Whitney U test = 1.34, $P = 0.18$. On the other hand, SNP reproducibility was significantly higher in the group of RS 4–5 and very high in the group RS 6–8. Interestingly, the dependence of SNP reproducibility on $-\log(P)$ is linear (Figure 3b). Both $-\log(P)$ and RS were significant in main effects ANOVA model with similar estimated P-values: $1.2 \cdot 10^{-12}$ for $-\log(P)$ and $3.6 \cdot 10^{-11}$ for RS.

Validation of RS-based prediction of SNP reproducibility

We used the results of lung cancer GWASs conducted at MD Anderson Cancer Center (MDACC) (Amos et al., 2008) and International Agency for Research on Cancer (IARC) (Hung et al., 2008). Adenocarcinoma and squamous cell carcinoma were analyzed separately. As a result we had 4 datasets: IARC_ADENO, IARC_SQUAMOUS, MDACC_ADENO and MDACC_SQUAMOUS. For all genotyped SNPs we computed RS and subdivided SNPs into 4 groups: RS = 0–1, RS = 2–3, RS = 4–5, and RS = 6–8 (Figure 4.1). Minus Log(P) does not change much in the interval from 0 to 5 and increases for the

SNPs with reproducibility score of 6 and higher. The increase in $-\text{Log}(P)$ was significant in all groups except MDACC_SQUAMOUS.

To check if high $-\text{Log}(P)$ in one (discovery) study predicts high $-\text{Log}(P)$ in the other (replication) study we subdivided SNPs into 4 sets based on p-values, as follows: (1) SNPs with $-\text{LOG}(P)$ between 0 and 1.3, which corresponds to P-values between 1 and 0.05; (2) SNPs with $-\text{LOG}(P)$ 1.3–3, corresponding to P-values between 0.05 and 0.001, (3) SNPs with $-\text{LOG}(P)$ 3–4, corresponding to P-values between 0.001 and 0.0001, and (4) SNPs with $-\text{LOG}(P)$ 4 and higher, corresponding to P-values of 0.0001. The results of the analysis are shown on Figure 4.2-5. IARC study was used as discovery and MDACC as replication, and then the other way around. Therefore we found that selection of the top SNPs based on $-\text{Log}(P)$ was associated with a significant enrichment of the $-\text{Log}(P)$ in one out of 4 datasets, while selection of SNPs based on RS (regardless of $-\text{LOG}(P)$) led to a significant inflation of $-\text{Log}(P)$ in 3 out of 4 studies. The result suggests that selection of SNPs based on RS may be more effective than selection of SNPs based on $-\text{Log}(P)$.

We hypothesized that $-\text{Log}(P)$ -based SNP selection is more efficient for selection of most significant SNPs in the discovery study, while RS-based approach is more efficient for selection of a larger number of SNPs with smaller effect size. We estimated $-\text{Log}(P)$ in MDACC sample among SNPs that were: (1) nominally significant in IARC; and (2) top hundred most significant IARC SNPs. We also conducted the analysis the other way around. Table 5 shows the results of the analysis. The 100 most significant SNPs in the discovery study tend to have a significantly higher $-\text{Log}(P)$ in validation, however, SNPs nominally significant in the discovery study ($p < 0.05$; $n = 19,000$) do not show $-\text{Log}(P)$ inflation. The result was different for RS-selected SNPs. The top 100 SNPs selected based on RS do not show a $-\text{Log}(P)$ inflation, while the top 19,000 SNPs selected based on RS (the number was chosen to match the number of nominally significant SNPs) show a significant inflation of $-\text{Log}(P)$ in 3 out of 4 samples.

RS among disease-associated SNPs reported in recently published GWASs

When we revisited the Catalogue of Published GWAS 6 month after our first visit we found 580 new disease-associated SNPs added to the catalog. Those SNPs were not used to build RS model and therefore can be used for validation. We computed the average RS for the novel SNPs and compared it to the average for all SNPs on the Illumina 500k array (reference group). The average RS for the reference group was 1.43 ± 0.01 while the average RS in the group of the new disease-associated SNPs was significantly higher: 1.89 ± 0.08 . Figure 5 shows the distributions of SNPs by RSs in the group of novel disease associated SNPs and in the reference group. In the group of the novel disease-associated SNPs the distribution is shifted to the right: $\chi^2 = 65.3$, $df = 9$, $P = 6.5 \cdot 10^{-17}$.

Interactions between RS predictors

Factorial regression was used to explore second and third levels of interactions between predictors. Table 6 shows the results of the analysis.

As an example of interactions between predictors Figure 6 shows SNP reproducibility in four groups defined by OMIM status and SNP type. We found that missense mutations in

OMIM genes have higher reproducibility than that expected from the additive model. The analysis of the third-order interactions found that SNPs that are (1) relatively rare (less than 10% - but previously SNPs with higher MAFs were more reproducible), (2) missense mutations, in (3) OMIM genes have about tenfold higher reproducibility rate compared to the average reproducibility rate for the GWAS-detected SNPs.

Conclusions

SNP reproducibility is important in distinguishing true and false positives from genome-wide association studies. In this study we used the results of ~2,000 published GWASs to identify predictors of SNP reproducibility. We constructed a reproducibility score (RS) for *a priori* assessment of SNP reproducibility. RS is based on SNP characteristics only and does not use prior knowledge from GWAS.

In most diseases GWAS-detected SNPs explain only a fraction of estimated heritability – the problem of missing heritability (Koch, 2014; Shen, 2013; Zuk et al., 2014). Missing heritability may be attributable to SNPs with small effect sizes. Such SNPs are difficult to detect since a very large sample size is required. Preselecting SNPs for association study based on RS may improve detection of SNPs with small effect size because more focused approach would relax correction of the type I error for multiple testing. We believe that RS-based SNP prioritization may provide guidance for more targeted and powered approach to detecting the disease-associated SNPs with small effect size.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported in part by the National Institutes of Health U19 CA148127 grant and the National Institutes of Health Grants 5 P30 CA016672, LM009012, LM010098 and GM103534. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Literature Cited

- Alfoldi J, Lindblad-Toh K. Comparative genomics as a tool to understand evolution and disease. *Genome Res.* 2013; 23(7):1063–1068. [PubMed: 23817047]
- Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Houlston RS. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet.* 2008; 40(5):616–622. [PubMed: 18385676]
- Balakrishnan R, Harris MA, Huntley R, Van Auken K, Cherry JM. A guide to best practices for Gene Ontology (GO) manual annotation. *Database (Oxford).* 2013; 2013:bat054. [PubMed: 23842463]
- Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nat Rev Genet.* 2009; 10(3):184–194. [PubMed: 19223927]
- Coordinators NR. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2014; 42(Database issue):D7–D17. [PubMed: 24259429]
- Dayem Ullah AZ, Lemoine NR, Chelala C. A practical guide for the functional annotation of genetic variations using SNPnexus. *Brief Bioinform.* 2013; 14(4):437–447. [PubMed: 23395730]
- Domazet-Lošo T, Tautz D. An ancient evolutionary origin of genes associated with human genetic diseases. *Mol Biol Evol.* 2008; 25(12):2699–2707. [PubMed: 18820252]

- Gorlova O, Fedorov A, Logothetis C, Amos C, Gorlov I. Genes with a large intronic burden show greater evolutionary conservation on the protein level. *BMC Evol Biol.* 2014; 14(1):50. [PubMed: 24629165]
- Greenwood CM, Rangrej J, Sun L. Optimal selection of markers for validation or replication from genome-wide association studies. *Genet Epidemiol.* 2007; 31(5):396–407. [PubMed: 17410553]
- Hakonarson H, Grant SF. Planning a genome-wide association study: points to consider. *Ann Med.* 2011; 43(6):451–460. [PubMed: 21595511]
- Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A.* 2009; 106(23):9362–9367. [PubMed: 19474294]
- Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Brennan P. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature.* 2008; 452(7187):633–637. [PubMed: 18385738]
- Ioannidis JP, Tarone R, McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology.* 2011; 22(4):450–456. [PubMed: 21490505]
- Koch L. Disease genetics: Insights into missing heritability. *Nat Rev Genet.* 2014; 15(4):218.
- Kraft P, Cox DG. Study designs for genome-wide association studies. *Adv Genet.* 2008; 60:465–504. [PubMed: 18358330]
- Kraft P, Zeggini E, Ioannidis JP. Replication in genome-wide association studies. *Stat Sci.* 2009; 24(4):561–573. [PubMed: 20454541]
- Lewis A, Tomlinson I. Cancer. The utility of mouse models in post-GWAS research. *Science.* 2012; 338(6112):1301–1302. [PubMed: 23224545]
- Liu X, Yu X, Zack DJ, Zhu H, Qian J. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics.* 2008; 9:271. [PubMed: 18541026]
- Marjoram P, Zubair A, Nuzhdin SV. Post-GWAS: where next? More samples, more SNPs or more biology? *Heredity (Edinb).* 2014; 112(1):79–88. [PubMed: 23759726]
- Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 2010; 6(4):e1000888. [PubMed: 20369019]
- Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, Dale AM. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet.* 2013; 9(4):e1003449. [PubMed: 23637621]
- Shen X. The curse of the missing heritability. *Front Genet.* 2013; 4:225. [PubMed: 24204378]
- Siontis KC, Patsopoulos NA, Ioannidis JP. Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *Eur J Hum Genet.* 2010; 18(7):832–837. [PubMed: 20234392]
- Smith JG, Newton-Cheh C. Genome-wide association study in humans. *Methods Mol Biol.* 2009; 573:231–258. [PubMed: 19763932]
- Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet.* 2012; 90(1):7–24. [PubMed: 22243964]
- Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Lander ES. Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A.* 2014; 111(4):E455–E464. [PubMed: 24443550]

Figure 1.

The association between $-\text{Log}(P)$ and reproducibility rate. SNPs were categorized into 12 groups based on the $-\text{Log}(P)$ in the discovery study. Vertical bars represent standard errors

Figure 2.
The association between SNP reproducibility and predictors

Figure 3. Reproducibility among SNPs grouped by the reproducibility score (RS) (a) and by $-\log(P)$ (b)

Figure 4.

Minus Log(P) in different groups of SNPs. For the plot 1 SNPs we categorized into 4 groups based on their RS. For the plots 2 and 4 SNPs were divided into 4 groups based on P-values from IARC study, while for the plots 3 and 5 SNPs were categorized into 4 groups based on MDACC study and the results of IARC study were used to check for inflation of $-\text{Log}(P)$ in the four groups.

Figure 5.

The distribution of the novel disease-associated SNPs by the reproducibility score compared to the distribution of SNPs in the reference group

Figure 6.
SNP reproducibility in 4 groups of SNPs defined by OMIM status and SNP type

Table 1

Variables used to predict SNP reproducibility

Name	Description	Source of the data
Conservation index	The level of evolutionary conservation of the protein based on the most distant homolog of the human gene	NCBI HomoloGene database: http://www.ncbi.nlm.nih.gov/homologene
eQTL	SNP reported as an eQTL for HapMap Data	eQTL SNPs identified in lymphoblastic cell lines from HapMap project [16]
Gene Size	Size of the gene region in nucleotides	NCBI RefSeq database: http://www.ncbi.nlm.nih.gov/refseq/
Growth factor	The protein encoded by the linked gene is a growth factor	Gene Ontology (GO) database http://geneontology.org/
Kinase	The protein encoded by the linked gene is a kinase	Gene Ontology (GO) database http://geneontology.org/
–Log(P)	Minus LOG(P) where P is the discovery P-value reported in CPG	Catalog of Published GWAS (CPG): http://www.genome.gov/26525384
MAF	Minor allele frequency	Catalog of Published GWAS (CPG): http://www.genome.gov/26525384 . MAF reported in control group were used.
Nuclear Localization	The protein encoded by the linked gene is localized in the nucleus	Gene Ontology (GO) database http://geneontology.org/
OMIM	We checked if the linked gene is listed in Online Mendelian Inheritance in Man (OMIM) database	http://www.ncbi.nlm.nih.gov/omim
Plasma Membrane	The protein encoded by the linked gene is localized in plasma membrane	Gene Ontology (GO) database http://geneontology.org/
Receptor	The protein encoded by the linked gene is a receptor	Gene Ontology (GO) database http://geneontology.org/
SNP type	Type of the SNP. For the details see Materials and Methods	Catalog of the Published GWAS (CPG): http://www.genome.gov/26525384
Tissue specific	The expression of the linked gene is tissue specific	Tissue specific Gene Expression and Regulation (TiGER) database: http://bioinfo.wilmer.jhu.edu/tiger/
Transcription factor	The protein encoded by the linked gene is a transcription factor	Gene Ontology (GO) database http://geneontology.org/

Table 2

The results of univariate analysis of SNP reproducibility

Predictor	DF*	One-way ANOVA F-test	P-value
Conservation Index	2	31.1	$9.7 \cdot 10^{-4}$
eQTL	2	10.2	$7.7 \cdot 10^{-3}$
Gene Size	12	87.8	$6.0 \cdot 10^{-11}$
Growth Factor	2	40.8	$5.7 \cdot 10^{-4}$
Kinase	2	81.2	$1.5 \cdot 10^{-4}$
–Log(P)	11	182.1	$2.7 \cdot 10^{-12}$
MAF	5	62.8	$6.4 \cdot 10^{-6}$
Nuclear Localization	2	61.1	$2.7 \cdot 10^{-4}$
OMIM	1	304.8	$6.0 \cdot 10^{-5}$
Plasma Membrane	2	37.5	$6.7 \cdot 10^{-4}$
Receptor	2	44.1	$4.9 \cdot 10^{-4}$
SNP Type	8	64.6	$1.1 \cdot 10^{-7}$
Tissue Specific	2	46.6	$4.4 \cdot 10^{-4}$
Transcription Factor	2	39.7	$6.0 \cdot 10^{-4}$

*
degree of freedom

Table 3

Coding of the predictors for main effects ANOVA

Variable	Categories			Coding of the variables for Reproducibility Score(RS)
	Yes	No	NA	
OMIM	yes	no	na	One for “yes” 0 others
Receptors	yes	no	na	One for “yes” 0 others
Kinases	yes	no	na	One for “yes” 0 others
Growth Factors	yes	no	na	One for “yes” 0 others
Transcription Factors	yes	no	na	One for “yes” 0 others
Tissue specific	yes	no	na	One for “yes” 0 others
Plasma Membrane Localization	yes	no	na	One for “yes” 0 others
Nuclear localization	yes	no	na	One for “yes” 0 others
Conservation Index (CI)	9 CI categories. NA when alignments are not available			One if CI=1, 0 others
Gene Size	Categories are based on 200k increment.			Zero for the two longest categories 5 and 6, one others
HapMap eQTL	yes/no			One for “yes” zero for “no”
MAF Category	5 categories based on 10% increment			One for MAF 0.3, 0 for others
SNP Type	Ten SNP type categories			2 for coding nonsynonymous and 5'UTR. 1 for coding synonymous, 3' downstream and non-coding. Zero for others.

Table 4

The results of main effects ANOVA to predict SNP reproducibility

Variable	D. of F.*	F test	P
Intercept	1	16.2	$2.6 \cdot 10^{-5}$
–Log(P)	11	30.5	$7.1 \cdot 10^{-4}$
SNP Type	2	13.1	$1.7 \cdot 10^{-3}$
MAF	5	19.4	$4.8 \cdot 10^{-3}$
OMIM	1	9.2	$7.3 \cdot 10^{-3}$
Nuclear Localization	2	8.6	$8.8 \cdot 10^{-3}$
Gene Size (200kb increment)	2	6.6	$9.7 \cdot 10^{-3}$
Kinases	2	6.4	$1.2 \cdot 10^{-2}$
Conservation Index	1	6.1	$2.9 \cdot 10^{-2}$
Growth Factors	1	5.2	$4.1 \cdot 10^{-2}$
Tissue Specific	2	3.8	$4.8 \cdot 10^{-2}$
Transcription Factors	2	2.8	$5.9 \cdot 10^{-2}$
Plasma Membrane Localization	2	2.2	$6.7 \cdot 10^{-2}$
eQTL (HapMap)	1	1.3	$7.4 \cdot 10^{-2}$
Receptors	2	0.2	$4.6 \cdot 10^{-1}$

*
D. of F. – degree of freedom

Variables shown in bold are used to compute reproducibility score (RS)

Table 5

Minus Log(P) in replication set for the group of SNPs selected from a discovery set as the top 100 most significant or nominally significant

Type of SNP	Inflation of $-\log(P)$			
	IARC_ADENO	MDACC_ADENO	IARC_SQUAMOS	MDACC_SQUAMOS
Top 100 from IARC_ADENO		0.66±0.08 vs 0.45±0.01; T-test = 4.8, P = 2.31·10 ⁻⁶		
Top 100 from MDACC_ADENO	0.49±0.08 vs 0.45±0.01; T-test = 0.8, P = 0.32			
Top 100 from IARC_SQUAMOS				0.59±0.08 vs 0.44±0.01; T-test = 2.0, P = 0.04
Top 100 from MDACC_SQUAMOS			0.57±0.08 vs 0.45±0.01; T-test = 1.7, P = 0.09	
Top 100 from reproducibility score	0.47±0.08 vs 0.45±0.01; T-test = 0.5, P = 0.59	0.48±0.08 vs 0.45±0.01; T-test = 0.5, P = 0.60	0.52±0.08 vs 0.45±0.01; T-test = 1.76, P = 0.08	0.45±0.08 vs 0.44±0.01; T-test = 0.9, P = 0.91
Nominally significant from IARC_ADENO		0.45±0.08 vs 0.44±0.01; T-test = 1.8, P = 0.07		
Nominally significant from MDACC_ADENO	0.45±0.08 vs 0.44±0.01; T-test = 0.8, P = 0.46			
Nominally significant from IARC_SQUAMOS				0.44±0.08 vs 0.44±0.01; T-test = 0.4, P = 0.68
Nominally significant from MDACC_SQUAMOS			0.45±0.08 vs 0.45±0.01; T-test = 1.4, P = 0.15	
Nominally significant RS	0.46±0.02 vs 0.45±0.01; T-test = 1.4, P = 0.31	0.53±0.08 vs 0.45±0.01; T-test = 1.9, P = 0.04	0.46±0.01 vs 0.44±0.01; T-test = 2.73, P = 6.12·10 ⁻³	0.47±0.01 vs 0.44±0.01; T-test = 6.4, P = 1.33·10 ⁻⁶

Table 8

Significant interactions between predictors of SNP reproducibility

First predictor	Second predictor	Third predictor	F	P	Combination of trait with highest reproducibility	Highest group reproducibility
OMIM	SNP Type		129.25	0.0002	Missense or 5'UTR, OMIM gene	0.17±0.01
MAF	Kinases		46.03	0.0009	MAF <10%, Kinase gene	0.13±0.02
MAF	SNP Type		38.8	0.001	MAF <10%, Missense or 5'UTR	0.17±0.02
MAF	SNP Type	OMIM	14.65	0.01	MAF <10%, Missense or 5'UTR, OMIM gene	0.21±0.02