Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies

John P.A. Ioannidis, Konstantinos C.M. Siontis, Nikolaos A. Patsopoulos

Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies

Konstantinos C. M. Siontis[1], Nikolaos A. Patsopoulos[1], John P. A. Ioannidis[1,2,3]

[1]Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece; [2]Center for Genetic Epidemiology and Modeling, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Department of Medicine, Tufts University School of Medicine, Boston, USA [3]Biomedical Research Institute, Foundation for Research and Technology-Hellas, Ioannina, Greece

Address correspondence to: John PA Ioannidis, MD, Professor and Chairman, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece. Tel: +302651097807; Fax: +302651097867; e-mail: jioannid@cc.uoi.gr

ABSTRACT

Genome-wide association studies (GWAS) have created a paradigm shift in discovering genetic associations for common diseases and phenotypes, but it is unclear whether the thousands of candidate genetic association studies performed in the pre-GWAS era had found any reliable associations for common diseases and phenotypes. We aimed to systematically evaluate whether loci proposed to harbor candidate associations before the advent of GWAS are replicated in GWAS. GWAS published through 8/2008 and included in the NHGRI catalog were screened and variants in candidate loci were selected on the basis of statistical significance ($p<0.05$) to create a list of independent, non-redundant associations. Altogether 159 articles on GWAS were evaluated, 100 of which addressed past proposed candidate loci. A total of 291 independent, nominally significant ($p<0.05$) candidate gene associations were assembled after keeping only the lowest p-value SNP for each locus and each phenotype; 108 of those had $p<10^{-3}$ for association and 41 had $p<10^{-7}$. Twenty-two of these 41 candidate gene associations pertained to binary phenotypes with a median odds ratio 2.91 (IQR 1.82-4.6) and median minor allele frequency 0.17 (IQR 0.12-0.29) in Caucasians; for comparison, 60 new associations of binary outcomes with $p<10^{-7}$ discovered in the same GWAS had much smaller effects (median odds ratio 1.30, IQR 1.18-1.58) and modestly larger minor allele frequencies (median 0.27, IQR 0.15-0.43). Overall, few of the numerous genetic associations proposed in the candidate gene era have been replicated in GWAS, but those that have been conclusively replicated have large genetic effects that should not be discarded.

Keywords: genome-wide association studies; candidate loci; single nucleotide polymorphisms; common diseases/phenotypes; replication

INTRODUCTION

The search for common genetic variants influencing the risk of common diseases and phenotypes of medical interest has undergone a major paradigm shift. Until some years ago, the effort of discovering new genetic associations was dominated by targeted approaches where specific genes and variants were chosen based on known or suspected biological considerations, or at best by perusal of selected areas of the genome (e.g. those giving strong signals in linkage scans).[1] These approaches have had limited success in yielding conclusive results for the proposed "candidate gene" associations.[2] Nevertheless, a large literature of candidate gene associations was generated and continues to be published, with over 7000 articles annually.[3] Given a relatively poor replication record, the credibility of most of these associations has been questioned.[4,5]

In the meanwhile, genome-wide association studies (GWAS) rapidly evaluate hundreds of thousands of single nucleotide polymorphisms (SNPs) across the whole genome, in an agnostic fashion, i.e. without any prior predilection for specific loci.[6,7] GWAS have markedly accelerated the pace of discovery of associations with very strong statistical support.[8] The enthusiasm about newly discovered loci has left the previously proposed candidate variants in a state of uncertainty. Should we just disregard these candidate associations that formed the corpus of genetic epidemiology until recently and that continue to be studied in thousands of papers?

In theory, well-conducted GWAS offer an excellent opportunity to evaluate systematically and often with very good coverage[9] genetic loci that were previously proposed as candidates in the older literature. Here we aimed to record systematically and evaluate previously proposed candidate loci that have been replicated in GWAS.

MATERIALS AND METHODS

Definitions

We used a broad definition of a "candidate" locus (gene/region) as any gene or specific region that has been proposed to be potentially associated with any phenotype before being proposed by any agnostic GWAS. We accepted associations regardless of whether the impetus to study them had been derived from biological reasoning, functional data, in vitro or animal work, linkage signals, or other types of research. We also considered associations regardless of whether the same exact SNPs had been evaluated in the candidate gene era studies and in the GWAS, provided that the same gene or region was involved and the GWAS investigators acknowledged that this was a locus already proposed in the candidate literature. Additionally we accepted situations where a SNP belonged to a gene other than a candidate one, but it was in linkage disequilibrium with polymorphisms of a candidate gene, as reported by the authors. Genes whose putative associations originated only from evidence other than human population association studies (e.g. animals studies or functional in vitro data) were accepted only when there was an a priori plan to look at them before obtaining the GWAS results. We excluded gene variants where the animal or functional data were invoked only after they had been discovered to be associated in GWAS. In addition, we focused only on common variants, excluding rare variants. We accepted the definition of each of the GWAS articles on what are considered to be common variants and we recorded for each variant the minor allele frequency (MAF) according to HapMap (release 27) and NCBI dbSNP data for Caucasian populations.

Study selection and eligibility criteria for GWA studies

The online Catalog of Published Genome-Wide Association Studies of NHGRI (www.genome.gov/gwastudies) was searched for eligible studies published until August 01, 2008 (last update access September 15, 2008). Furthermore,

references of the eligible studies were screened for any other studies meeting the eligibility criteria that were not listed in the catalog.

GWAS articles were eligible if they genotyped more than 100,000 SNPs spanning across the whole genome in the first stage in at least one human population, in pools or individuals, and had analyzed at least one phenotype. Some of the eligible included studies eventually ended with less than 100,000 successfully genotyped SNPs after data quality surveillance procedures, but this was not considered a reason for excluding them. We also included follow-up publications and meta-analyses of GWAS that reported genotyping data on candidate variants from the stage 1 of GWAS that had not been reported in the primary GWAS publications. We excluded genome-wide studies on copy-number variants and studies that included only family-based designs in the first stage.

Availability and selection of data for variants in candidate loci

We scrutinized both the published articles and the online supplements of all eligible studies for any mention of candidate genes/regions. When any such mention was made, we perused the text and the corresponding references, if any, to ensure that this was not an association that had first appeared in other GWAS before any candidate work had been performed. Additionally, we queried the NHGRI catalog (www.genome.gov/gwastudies) and the HuGE Navigator database[10] to exclude genes that had been first proposed by GWAS. Whenever any mention was made in the GWAS articles on the past candidate association(s), we examined whether this was just a simple reference without providing any data, or whether any kind of data were also given. We noted in particular whether there was a preformed list of candidate genes; whether the stage 1 platform had been specifically enriched to add genotyping for variants considered to represent candidate genes; and whether the threshold used

for reporting data on candidate genes/regions was different compared to the one used for the other loci.

Data extraction for quantitative information

We considered only variants that were related to a specific candidate gene or particular regions, such as a specific intergenic region, that were highlighted by previous studies, rather than a region spanning many genes. Exception to the above was specific clusters (e.g. *HLA, APOE, APOA*, beta globin, *CYP2C),* even though these encompass several genes. Twelve GWAS presented data only on candidate variants belonging to large non-specific chromosomal regions (e.g. regions 2q31, 20q, etc.) that showed linkage in previous studies.

For those GWAS with stage 1 numerical data for at least one candidate variant, we identified for each variant with a stage 1 p-value<0.05, the gene locus, the SNP, and the p-value. Four GWAS reported no nominally significant associations for candidate loci (all p-values>0.05). From the others, we isolated one SNP per locus with the lowest p-value. When more than one phenotype had been probed for association with a single candidate variant in a study, each phenotype was accounted for separately. We considered uncorrected for multiple comparisons p-values. When both unadjusted and adjusted (for covariates) analyses were presented, we preferred the former. In addition, genotypic p-values were preferred over trend ones. Finally, for the studies that genotyped two or more distinct populations in the first stage we considered the combined stage 1 results, if available. If not, the lowest p-value for each SNP across the different cohorts was recorded.

A number of further steps were taken to create a list of independent, non-redundant associations, free of duplicates consisting of the same candidate locus and

the same or similar/related phenotype (described in detail in the Supplementary Methods).

The resulting list was examined to identify whether the candidate loci had evidence from at least one previous human population study on the same or some related phenotype(s) preceding the GWAS. When no such evidence was found, candidate status had been assigned apparently based on other (animal, functional, etc) considerations. For previous human population studies we queried the HuGE Navigator database,[10] Pubmed (www.pubmed.gov), PharmGKB (www.pharmgkb.org), AlzGene[11] and SzGene database[12].

For each of the replicated candidate associations with p-value<10^{-7}, we searched the HuGE Navigator database[10] to record the number of studies published on the association of the specific gene and the same or a similar/related phenotype until the year before the publication of the GWAS. We also assessed whether the specific gene-phenotype association had been initially derived from a linkage study or whether there was at least suggestive evidence for them in linkage studies. Finally, we recorded whether Mendelian mutations of the specific gene have been reported in association with the same or a similar/related phenotype, according to the Online Mendelian Inheritance in Man (OMIM) database (www.ncbi.nlm.nih.gov/omim).

Data extraction process

The series of actions taken for the selection and extraction of quantitative information is shown in **Supplementary Figure 1**. Two investigators (KCMS, NAP) perused the studies for eligibility and extracted the data. Discrepancies were resolved by a third investigator (JPAI).

Analyses

We present descriptives on the availability, selection rules employed, and reporting of candidate loci in the eligible GWAS. We present the distribution of p-values for the accrued list of independent associations of candidate loci.

For the SNPs that pertained to binary outcomes, we also recorded or calculated the odds ratio and 95% confidence interval and a Bayesian credibility method was applied[13,14] (details in the Supplementary Methods).

Using the NHGRI catalog, we also recorded the GWAS-discovered associations for binary phenotypes with robust statistical support ($p<10^{-7}$) that were observed in the 100 GWAS that had also addressed candidate gene associations (described in detail in the Supplementary Methods). Finally, we obtained data from HapMap on the minor allele frequencies in Caucasians (CEU) of these newly discovered associations for comparison with the minor allele frequencies of candidate loci with p-values$<10^{-7}$ using the Mann-Whitney U test.

RESULTS

Eligible studies and data on candidates

We identified 173 potentially eligible articles on GWAS through the NHGRI list, 159 of which were eligible for our analyses (**Supplementary Fig. 2**). Of those, in 32 (20%) no mention of past candidate variants was made and in another 27 (17%) the authors commented on the existence of previously proposed associations, but no GWAS-derived data were given. Of the remaining 100 GWAS (**Supplementary references**) with data on candidates, 2 provided non-numerical comments and quantitative data on candidate loci were given in 98 studies (62%).

In 52 studies results on candidate loci were reported according to less strict statistical significance thresholds compared to those applied for other loci. The authors had selected the candidate variants to report based on a clearly stated

preformed list in 37 (37%) studies. In 12 of these 37 studies results were presented for all candidates considered, in another 22 results were reported according to specific thresholds, while the selection on what to present was unclear in the remaining 3 studies. In 4 studies additional SNPs, apart from those present on the main platform, were genotyped to enhance coverage of some candidate loci.

Statistically significant SNPs in candidate loci

Some GWAS reported on candidate gene associations for many SNPs in the same locus and/or for several similar or related phenotypes, and some associations and loci had been targeted by two or more GWAS. In these cases, we selected the single lowest presented p-value for any related phenotype on the same candidate locus. The distribution of nominally significant p-values (<0.05) in the GWAS for the compiled 291 independent, non-redundant associations is shown in **Figure 1** and details appear in **Supplementary Table 1**. Of the 291 associations, 108 had $p<10^{-3}$ (**Table 1)** and 41 of the 108 had $p<10^{-7}$. Of all SNPs, 77.4% had MAF>0.10, 14.6% had MAF ranging from 0.05 to 0.10, and 8% had MAF<0.05. The 291 independent associations pertained to 233 different loci plus the HLA region and a wide variety of different types of phenotypes (**Supplementary Table 1**). For 32 genes plus the HLA region, nominally significant associations (p<0.05) were recorded on more than one type of phenotype, suggesting potential pleiotropic effects. Besides HLA that was associated with 11 different types of phenotypes, another 3 gene loci (*ADRB2, APOE, ESR1)* had nominally significant associations with 4 different types of phenotypes each.

For 32 of the associations with p-value$<10^{-7}$ (not including the HLA and beta-globin region variants), the median number of pre-GWAS publications per each gene-phenotype association was 4 (IQR 2.75-20). However, there was large variability and

while 6 associations only had a single previous candidate study, there were 797 publications on *APOE* and Alzheimer's disease. Six of the 32 associations referred to situations where variants in a gene seemed to regulate directly the levels of the protein produced by that gene (ICAM-1, CRP, YKL-40, cystatin C, factor VII, sIL-6R). Eight associations (**Supplementary Table 2**) were originally discovered through linkage studies. Another one (*APOE*/Alzheimer's disease) would have modest/suggestive linkage in its chromosomal locus in genome linkage scans[15] although it was originally discovered through association analyses. Similarly, *PTPN22* was initially found to be associated with type 1 diabetes in an association study; then association was found to exist also with other autoimmune diseases (rheumatoid arthritis, and systemic lupus erythematosus) for which retrospectively modest linkage signals had been observed in the respective chromosomal area (1p13).[16,17] Mendelian effects have been reported for 16 of the 32 associations (**Supplementary Table 2**). Only one of the 32 associations (regulation of CRP levels by an APOE variant) had no precedent of a Mendelian effect or linkage signal and referred to the regulation of the levels of a different protein than the protein produced directly by the gene of interest.

Magnitude of effects and Bayes factors for odds ratios

Figure 2a shows the distribution of 70 odds ratios and their 95% confidence intervals for the subset of independent binary-phenotype associations with previous human population studies on the same gene-phenotype pair (**Supplementary Table 3)**. The median odds ratio was 1.50 (IQR 1.28-2.38). Seven SNPs had an OR above 5 and another 21 above 2. In a sensitivity analysis excluding the 7 ORs above 5, the median OR was still 1.45 (IQR 1.27-1.95).

Bayes factors under different prior assumptions are also shown for the associations listed in **Supplementary Table 3**. Of the 70 listed associations, the

Bayes factor was <0.10 for 40 of them under at least one set of assumptions, suggesting that for those associations the odds of the association being true increased over 10-fold by the results of the genome-wide investigation, compared to what one thought before that study. Conversely for 30 associations (including the majority of the nominally statistically significant candidate associations on obesity, coronary heart disease, and Alzheimer's disease), the Bayes factor was unimpressive (>0.1) under all assumptions, suggesting poor credibility of the proposed associations.

Genetic effects in robustly replicated candidate and new GWAS-discovered loci

Twenty-two of the 70 candidate associations with binary outcomes had $p<10^{-7}$ in the GWAS. Examination of the NHGRI catalog showed that the 100 GWAS that provided results on past candidate associations had led to the discovery of 60 independent, non-redundant associations with binary phenotypes with equally robust statistical support ($p<10^{-7}$) (**Supplementary Table 4)**.

While the newly discovered loci overall far outnumbered the previously proposed candidate ones by 3 to 1, there were differences in the relative preponderance of candidate versus novel loci for various disease phenotypes (**Table 2**). For cancer phenotypes, coronary artery disease, restless leg syndrome, bipolar disorder, and gallstone disease, all the loci were newly discovered, with no variants in previously proposed candidate loci reaching $p<10^{-7}$. In inflammatory bowel disease and type 2 diabetes, there was a strong preponderance of newly-discovered loci, with few validated candidate genes. Conversely, there was a more balanced picture with both candidate and newly-discovered loci for pigmentation phenotypes and in most autoimmune diseases. Finally, for Alzheimer's disease and statin-induced myopathy, the sole locus with strong support had already been proposed in the candidate era.

Among associations with robust statistical support, the magnitude of the effects was on average much larger for the 22 candidates than for the 60 GWAS-discovered loci ($p<0.00001$) (**Figure 2b)**. The median odds ratio was 2.91 (IQR 1.82-4.6) for the candidate versus only 1.30 (IQR 1.18-1.58) for the GWAS-discovered associations. When we examined all 77 non-redundant independent associations for binary phenotypes discovered with documented $p<10^{-7}$ across all the 159 GWAS (including also those that did not address candidate loci at all), the median odds ratio was 1.32 (IQR, 1.19 to 1.59) which was still much smaller than the magnitude of the effects for the 22 associations from past candidate loci.

The minor allele frequency in Caucasians was smaller for the 22 associations than for the 60 GWAS-discovered associations, but even though the difference was nominally significant ($p=0.008$), the absolute difference was not impressive (median 0.17 [IQR 0.12-0.29] versus 0.27 [IQR 0.15-0.43]). Minor allele frequencies of 0.05 or less were seen only in 1 of the 22 SNPs representing candidate loci and 4 of the 60 SNPs representing new GWAS-derived discoveries.

DISCUSSION

We have accumulated data from 100 GWAS that addressed previously proposed candidate gene loci. Even though the reporting of candidate loci in these GWAS was not always systematic or comprehensive, we have catalogued a substantial number of candidate gene associations with considerable support for association in datasets of GWAS.

This catalogue is definitely not complete. Each of the evaluated GWAS used different criteria and thresholds for reporting on previously proposed associations. Furthermore some associations may not be replicated in GWAS due to suboptimal representation and coverage of the culprit candidate variants among the tag-SNPs

used in the high-throughput platforms. In addition, most associations of common genetic variants with complex phenotypes have weak effects and a GWAS may be underpowered to replicate them. For example, a GWAS with 1,000 cases and 1,000 controls has 12% power to detect a per-allele OR of 1.5 at $\alpha=10^{-7}$ for minor allele frequency of 10%, and the power increases to 85% for a minor allele frequency of 40%. Power would be negligible for detecting candidate gene associations with ORs of 1.2 or less, even for very common variants. Therefore, the replicated candidate variants with $p<10^{-7}$ are likely to be heavily selected in favor of those with the largest effect sizes and substantial minor allele frequencies. Finally, almost all GWAS analyzed have been performed in Caucasian populations and candidate gene associations that are relatively specific to non-Caucasian ancestry may have been missed.

Future GWAS may benefit from examining previously proposed candidate gene loci in a more systematic fashion, since the replication status of some of these may still be open to question and debate. Moreover, even for loci that are generally accepted, their exact genetic architecture may still be unknown and warrant further replication and detailed study. Detailed fine mapping and resequencing of discovered loci has suggested that in many cases one can identify multiple independent markers.[18,19,20] Systematic databases such as the HuGE Navigator[10] are available that can help create comprehensive lists of previously proposed loci and synopses of the genetic association literature may also be helpful to keep track of the evidence.[11,12,21]

The thresholds where past candidate loci should be claimed to be robustly replicated in GWAS platforms can be debated. Some may argue that similar stringent thresholds such as those proposed for newly discovered variants may be needed, e.g. $p<10^{-7}$ or even lower.[22,23] However, this may be too stringent a threshold for loci that

have already been proposed and tested for association in the past, even if not the same exact SNPs have been assessed. At the other end of the spectrum, a very lenient threshold, e.g. $p<0.05$, for isolated replication, will probably result in many false positives. Also one should caution that whenever associations are selected based on statistical significance thresholds, the effect sizes of the selected associations that pass the required threshold may be inflated compared with the true effect.[24,25] However, this is likely to affect both candidate and newly-discovered associations and is unlikely to invalidate the observation that validated candidate gene variants had much larger odds ratios than the newly-discovered variants. In the new wave of discoveries, currently emerging through meta-analyses of multiple GWAS, effects may be even smaller.[26,27] Future analyses of rare variants might, nevertheless, produce stronger signals with considerable effect sizes. With the currently available GWAS-derived data, the impact of candidate gene variants on the proportion of variance explained may be larger, yet still limited in average, than the respective impact of newly discovered GWAS signals.

We noted that half of the robustly replicated candidate associations were in genes that have known mutations producing relevant phenotypes. This may suggest that genes with known important mutations need to be screened with more in-depth sequencing for the recognition of additional common or rare variants that may affect the relevant phenotypes. Moreover, a considerable number of robustly replicated candidate associations are in areas that have given strong signals in linkage scans. It has been proposed that one may use linkage information to pre-weigh favorably the respective areas in GWAS analyses.[28] In some cases, we found pleiotropy with effects on several diseases with similar pathogenesis. Pleiotropic effects also need further study by examining systematically related phenotypes once an association has been

strongly replicated with one particular phenotype. Finally, it is not surprising that the list of robustly replicated associations should contain some situations where a gene variant directly regulates the levels of the protein produced by that gene. Otherwise, proposed candidate associations without such Mendelian or linkage precedent evidence may have low credibility.

The relative importance of previously proposed candidate loci differs depending on the phenotype. Despite a huge literature on cancer candidate genes,[29,30] candidate associations with highly definitive evidence are sparse, while there is a flurry of newly discovered loci. For coronary artery disease,[31,32] a huge candidate literature left hardly any strongly credible signals. Conversely, the picture is more balanced for autoimmune diseases, where candidate genes have strong documented effects, mainly represented by the MHC region. Finally, for some phenotypes such as Alzheimer's disease and pharmacogenetic associations (e.g. statin-induced myopathy or anticoagulant dosage and bleeding risk),[33] GWAS are still unable to produce additional associations with the robustness of those proposed already in the candidate era. Moreover, in the current efforts of full sequencing and with increasing emphasis placed on rare variants, candidate genes may also find some rekindled interest, where focused evaluation of specific genes may be one option to reduce the multiplicity of analyses for rare variants and where otherwise power to detect association is more limited.[34] Finally, both candidate and agnostic-derived genes may contribute to understanding of pathogenesis pathways but it should be acknowledged that the identification of the true culprits and their biological function is often very difficult both in the candidate-gene approach and in the agnostic GWAS setting.[35,36]

We should acknowledge that here we made no effort to select functional variants from each locus, since this would have been usually futile given the limited

information available in each of the GWAS that we analyzed and the difficulty and subjectivity in prioritizing functional importance. Another limitation is that for each candidate locus, it is possible that there may be several recombination hotspots defining different haplotype blocks and more than one independent signal may exist in the same locus. Also, the catalogue of replicated candidate loci would be larger, if all GWAS systematically reported on candidate loci and data were meta-analyzed across several GWAS.[37,38] What we have catalogued probably underestimates the number of GWAS-replicated candidate loci, but offers an indicative sample of replicated signals. On the other end on the spectrum, when GWAS' results are considered, numerous proposed candidate associations turn out to be false positives, but the evaluation of this large volume of non-replicated associations was beyond the scope of this study.

Overall, while GWAS have unquestionably led to a dramatic paradigm shift in discovering genetic associations, there is still some useful evidence to be gleaned from previously proposed candidate associations. Thousands of studies are still performed on past candidate loci, and unfortunately much of this research may be chasing futile, non-validated associations. Focusing candidate gene research efforts on those loci that are also systematically validated in GWAS platforms may improve the efficiency of this huge research agenda and help expedite the successful translation of this accumulating information.

Supplementary information is available at European Journal of Human Genetics' website

Acknowledgements

We are grateful to Patrick Gaffney, Sagiv Shifman, David van Heel, and Rachel Gibson for offering helpful clarifications on their data.

Funding: Nikolaos A Patsopoulos is funded by a PENED grant from the European Union and the General Secretariat for Research and Technology, Greece (PI: John Ioannidis).

Conflicts of Interest statement: None declared

REFERENCES

1 Cordell HJ, Clayton DG: Genetic association studies. *Lancet* 2005; **366**: 1121-1131.

2 Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: A comprehensive review of genetic association studies. *Genet Med* 2002; **4**: 45-61.

3 Lin BK, Clyne M, Walsh M *et al*: Tracking the epidemiology of human genes in the literature: the HuGE Published Literature database. *Am J Epidemiol* 2006; **164**: 1-4.

4 Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG: Replication validity of genetic association studies. *Nat Genet* 2001; **29**: 306-309.

5 Ioannidis JP: Genetic associations: false or true? *Trends Mol Med* 2003; **9**: 135-138.

6 Cardon LR: Genetics. Delivering new disease genes. *Science* 2006; **314**: 1403-1405.

7 McCarthy MI, Abecasis GR, Cardon LR *et al*: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008; **9**: 356-369.

8 Manolio TA, Brooks LD, Collins FS: A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008; **118**: 1590-1605.

9 Barrett JC, Cardon LR: Evaluating coverage of genome-wide association studies. *Nat Genet* 2006; **38**: 659-662.

10 Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ: A navigator for human genome epidemiology. *Nat Genet* 2008; **40**: 124-125.

11 Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE: Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet* 2007; **39**: 17-23.

12 Allen NC, Bagade S, McQueen MB *et al*: Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nat Genet* 2008; **40**: 827-834.

13 Ioannidis JP: Calibration of credibility of agnostic genome-wide associations. *Am J Med Genet B Neuropsychiatr Genet* 2008; **147B**: 964-972.

14 Ioannidis JP: Effect of formal statistical significance on the credibility of observational associations. *Am J Epidemiol* 2008; **168**: 374-383.

15 Kehoe P, Wavrant-De Vrieze F, Crook R *et al*: A full genome scan for late onset Alzheimer's disease. *Hum Mol Genet*. 1999; **8**: 237-245.

16 Gaffney PM, Kearns GM, Shark KB *et al*: A genome-wide search for susceptibility genes in human systemic lupus erythematosus sib-pair families. *Proc Natl Acad Sci U S A*. 1998; **95**: 14875-14879.

17 Jawaheer D, Seldin MF, Amos CI *et al*: Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families. *Arthritis Rheum*. 2003; **48**: 906-916.

18 Gudbjartsson DF, Arnar DO, Helgadottir A *et al*: Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* 2007; **448**: 353-357.

19 Haiman CA, Patterson N, Freedman ML *et al*: Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet* 2007; **39**: 638-644.

20 Graham RR, Kyogoku C, Sigurdsson S *et al*: Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc Natl Acad Sci U S A* 2007; **104**: 6758-6763.

21 Ioannidis JP, Gwinn M, Little J *et al*: A road map for efficient and reliable human genome epidemiology. *Nat Genet* 2006; **38**: 3-5.

22 Pe'er I, Yelensky R, Altshuler D, Daly MJ: Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* 2008; **32**: 381-385.

23 Hoggart CJ, Clark TG, De Iorio M, Whittaker JC, Balding DJ: Genome-wide significance for dense SNP and resequencing data. *Genet Epidemiol* 2008; **32**: 179-185.

24 Ioannidis JP: Why most discovered true associations are inflated. *Epidemiology* 2008; **19**: 640-648.

25 Zollner S, Pritchard JK: Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am J Hum Genet* 2007; **80**: 605-615.

26 Zeggini E, Weedon MN, Lindgren CM *et al*: Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007; **316**: 1336-1341.

27 Zeggini E, Scott LJ, Saxena R *et al*: Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008; **40**: 638-645.

28 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747-753.

29 Dong LM, Potter JD, White E, Ulrich CM, Cardon LR, Peters U: Genetic susceptibility to cancer: the role of polymorphisms in candidate genes. *JAMA* 2008; **299**: 2423-2436.

30 Vineis P, Manuguerra M, Kavvoura FK *et al*: A field synopsis on low-penetrance variants in DNA repair genes and cancer susceptibility. *J Natl Cancer Inst* 2009; **101**: 24-36.

31 Ntzani EE, Rizos EC, Ioannidis JP: Genetic effects versus bias for candidate polymorphisms in myocardial infarction: case study and overview of large-scale evidence. *Am J Epidemiol* 2007; **165**: 973-984.

32 Morgan TM, Krumholz HM, Lifton RP, Spertus JA: Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. *JAMA* 2007; **297**: 1551-1561.

33 Wang L, Weinshilboum RM: Pharmacogenomics: candidate gene identification, functional validation and mechanisms. *Hum Mol Genet* 2008; **17**: R174-179.

34 Bodmer W, Bonilla C: Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008; **40**: 695-701.

35 Ioannidis JP, Thomas T, Daly MJ: Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet* 2009; **10**: 318-329.

36 McCarthy MI, Hirschhorn JN: Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet* 2008; **17(R2)**: R156-165.

37 Zeggini E, Ioannidis JP: Meta-analysis in genome-wide association studies. *Pharmacogenomics* 2009; **10**: 191-201.

38 Richards JB, Kavvoura FK, Rivadeneira F *et al*: Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture. *Ann Intern Med*. 2009; **151**: 528-537.

Titles and legends to figures

Figure 1 Distribution of the P-values of nominally statistically significant independent, non-redundant associations for variants from candidate loci across 100 genome-wide association studies.

Figure 2 Magnitude of genetic effects in 100 genome-wide association studies: **(a)** Odds ratios and 95% confidence intervals for independent, non-redundant associations of candidate loci with binary phenotypes for which previous human population studies had been performed. Odds ratios could be obtained for 70 of 84 eligible associations; **(b)** Comparison of odds ratios of new GWAS-discovered versus candidate associations with p-values $< 10^{-7}$ for binary phenotypes.

Table 1. List of independent, non-redundant associations with variants from candidate loci where p<0.001 in the genome-wide association study

PMID of GWAS	Phenotype	Locus	SNP	p-value	Past evidence on human populations	Number of related phenotypes
17903297	Factor 3: attention and executive function-Trails A and B	A2M	rs2889717	9.30E-05	(+)	
18193044	HDL-C levels	ABCA1	rs3890182	3.00E-05	+	
17505501	Hepatic adverse events during ximelagatran treatment	ABCG1	rs7281093	8.70E-05	-	
17903303	Ankle brachial index	ADM	rs10500724	2.00E-04	(+)	
18445777	Spine bone mineral density	AHSG[1]	rs9898	7.60E-04	+	1
17903297	Factor 1:Verbal Memory	APBB2	rs10517001	1.40E-04	(+)	
18193044	Triglyceride levels	APOA1-C3-A4-A5	rs28927680	6.00E-05	+	
18179892	Triglyceride levels	APOA5	rs6589566	2.89E-11	+	1
17505501	Hepatic adverse events during ximelagatran treatment	APOB[1]	rs676210	4.40E-04	-	
18262040	LDL-C levels	APOB[1]	rs562338	1.40E-09	+	2
17463246	LDL-C levels	APOE cluster[1]	rs4420638	3.40E-13	+	
17998437	Alzheimer's disease	APOE[1]	rs4420638	2.30E-44	+	
18439548	CRP levels	APOE[1]	rs769449	8.90E-21	+	
17903297	Boston Naming Test	BACE2	rs10483073	1.70E-04	(+)	1
18245381	HbF levels	b-globin cluster	rs11886868	6.74E-35	+	
17435756	Crohn disease	CARD15/NOD2	rs2076756	7.01E-14	+	
17463246	HDL-C levels	CETP	rs1800775	2.50E-13	+	1
18403759	YKL-40 levels	CHI3L1	rs4950928	1.10E-13	+	
17505501	Hepatic adverse events during ximelagatran treatment	CHN2	rs4722897	5.10E-05	-	
17158188	Nicotine dependence	CHRNB3	rs13277254	6.54E-05	+	
17903297	Temporal Brain volume (MRI)	CNTN1	rs10506176	8.80E-04	-	
18439548	CRP levels	CRP[1]	rs3091244	6.16E-28	+	1
17903292	CysC levels	CST3[1]	rs1158167	8.00E-09	+	
17903297	Occipital Brain Volume (MRI)	CST3[1]	rs1158167	6.50E-04	(+)	
17554300	Type 1 diabetes	CTLA4[1]	rs3087243	1.80E-05	+	

17505501	Hepatic adverse events during ximelagatran treatment	CUTL2	rs2157876	1.90E-04	-	
18535201	Warfarin maintenance dose	CYP2C18-CYP2C19-CYP2C8-CYP2C9	rs7896133	3.60E-05	+	
18535201	Warfarin maintenance dose	CYP2C9[1]	rs4917639	9.70E-05	+	
17903297	Total Cerebral Brain Volume (MRI)	DCDC2	rs10484657	6.00E-04	(+)	
18282107	Schizophrenia	DGCR2	NA	1.00E-04	+	
17447842	Crohn disease	DLG5	NA	1.00E-04	+	
17903305	Breast cancer	ERBB4	rs905883	2.00E-04	+	
18445777	Spine bone mineral density	ESR1[1]	rs2504063	5.68E-08	+	2
17903294	Factor VII levels	F7	rs561241	4.50E-16	+	
18204446	Systemic lupus erythematosus	FCGR2A	rs1801274	6.78E-07	+	
17903295	Age at death	FOXO1a	rs10507486	1.30E-04	+	
18521090	Response to iloperidone treatment	GFRA2	rs7837682	2.10E-04	-	
18445777	Spine bone mineral density	GnRH	rs12549314	5.40E-04	+	1
18521090	Response to iloperidone treatment	GRIA4	rs2513265	1.70E-04	(+)	
17903307	Mean FVC from two exams	GSTO2	rs156697	9.80E-06	-	
17767159	HbF levels	HBS1L-c-MYB intergenic region	rs9399137	2.80E-27	+	
18445777	Hip bone mineral density	HDC	rs4390539	3.90E-04	-	
17632545	Type 1 diabetes	HLA region	rs2647044	5.18E-142	+	
17505501	Hepatic adverse events during ximelagatran treatment	HLA region	rs2858869	6.00E-06	+	
17558408	Celiac disease	HLA region	rs2187668	1.00E-19	+	
17641165	HIV viral load	HLA region	rs2395029	9.36E-12	+	
17660530	Multiple sclerosis	HLA region	rs3135388	8.94E-81	+	
17804836	Rheumatoid arthritis	HLA region	rs2395175	8.00E-108	+	
18204446	Systemic lupus erythematosus	HLA region	rs3131379	1.7E-52[2]	+	
18364390	Psoriasis	HLA region	rs3134792	1.00E-09	+	
18369459	Psoriatic arthritis	HLA region	rs10484554	5.79E-07	+	
18193044	LDL-C levels	HMGCR	rs12654264	4.00E-04	+	
18604267	Soluble ICAM-1 levels	ICAM1	rs1799969	2.10E-28	+	
17554300	Type 1 diabetes	IL2RA	rs2104286	4.32E-05	+	
18464913	sIL-6R levels	IL6R[1]	rs4129267	1.82E-57	+	

17632545	Type 1 diabetes	INS	rs1004446	4.38E-09	+	
18445777	Spine bone mineral density	IRAK1	rs4898457	3.80E-04	+	1
18677312	Systemic lupus erythematosus	IRF5	rs12531711	4.03E-12	+	
16648850	QT interval	KCNK1	rs2282428	1.00E-04	+	
18193044	LDL-C levels	LDLR[1]	rs6511720	9.00E-07	+	
18193044	HDL-C levels	LIPC[1]	rs1800588	3.00E-05	+	
18179892	Triglyceride levels	LPL[1]	rs17482753	1.17E-09	+	1
18455228	Bone mineral density	LRP5	rs3736228	1.90E-05	+	2
17903297	Factor 1:Verbal Memory	LRRK2	rs7975693	8.10E-05	(+)	3
17903297	Similarities	LTA4H[1]	rs10492226	6.70E-04	(+)	
17903297	Parietal Brain Volume (MRI)	LTB4R2	rs724165	4.50E-04	-	
17999355	Skin pigmentation	MATP/SLC45A2	rs16891982	3.21E-11	+	2
18483556	Red vs non-red hair	MC1R	rs258322	5.00E-27	+	1
17903305	Prostate cancer	MSR1	rs9325782	8.20E-04	+	
17903297	Hippocampal volume (MRI)	NGFB	rs10489531	9.70E-04	(+)	
18521090	Response to iloperidone treatment	NPAS3	rs11851892	8.60E-05	(+)	
18445777	Hip bone mineral density	NR3C1	rs258799	1.90E-04	(+)	1
17903297	White Matter Hyperintensity Volume (MRI)	NRG1	rs10503926	8.70E-05	+	3
17903297	Factor 1:Verbal Memory	NTRK2[1]	rs10512152	7.60E-05	(+)	
17903297	Lateral Ventricular Volume (MRI)	NTRK3	rs10520671	6.20E-05	(+)	
18521090	Response to iloperidone treatment	NUDT9P1	rs4528226	2.80E-04	-	
17952075	Blue vs brown eyes	OCA2	rs1667394	1.40E-124	+	5
18445777	Spine bone mineral density	OSCAR	rs12150965	6.80E-04	+	
17903297	Similarities	PDE4D[1]	rs10514882	6.40E-05	(+)	1
17053108	Wet age-related macular degeneration	PLEKHA1-HTRA1 intergenic region	rs10490924	4.08E-12	+	
17505501	Hepatic adverse events during ximelagatran treatment	PON1[1]	rs2299257	1.10E-05	(+)	
17903297	Parietal Brain Volume (MRI)	PRNP	rs2326510	7.18E-04	+	1
17903297	Total Cerebral Brain Volume (MRI)	PRSS25	rs363685	5.70E-05	(+)	1
17554300	Rheumatoid arthritis	PTPN22[1]	rs6679677	5.55E-25	+	
17554300	Type 1 diabetes	PTPN22[1]	rs6679677	5.43E-26	+	

18204446	Systemic lupus erythematosus	PTPN22[1]	rs2476601	5.20E-06	+	
18445777	Hip bone mineral density	RANK	rs3018362	4.25E-05	+	1
17767159	HbF levels	Region 11p15.4 (beta globin locus)	XmnI-Gg	2.00E-30	+	
15761122	Age-related macular degeneration	Region 1q31 (CFH)[3]	rs380390	4.10E-08	+	
17505501	Hepatic adverse events during ximelagatran treatment	RXRG	rs17469292	7.30E-05	-	
18464913	SHBG levels	SHBG	rs6761	3.08E-07	+	
18650507	Statin-Induced Myopathy	SLCO1B1	rs4149056	1.00E-08	+	
17903297	Hippocampal volume (MRI)	SNCA	rs7678651	1.30E-04	+	2
18445777	Hip bone mineral density	SOST	rs1107748	1.10E-04	+	1
18204098	Systemic lupus erythematosus	STAT4	rs7574865	9.00E-14	+	
17293876	Type 2 diabetes	TCF7L2	rs7903146	3.20E-17	+	
17903297	Temporal Brain volume (MRI)	TEK	rs628873	4.90E-04	-	
17903297	Temporal Brain volume (MRI)	THBS2	rs6937001	6.00E-04	-	
18445777	Spine bone mineral density	TNFRSF11B/OPG[1]	rs6469804	8.53E-06	+	2
18445777	Spine bone mineral density	TNFSF11/RANKL	rs9594759	1.17E-08	+	1
18521090	Response to iloperidone treatment	TNR	rs875326	1.30E-04	-	
17999355	Skin pigmentation	TYR	rs1042602	4.48E-10	+	
18488028	Blue vs non-blue eyes	TYRP1	rs1408799	1.50E-09	+	
18535201	Warfarin maintenance dose	VKORC1	rs10871454	6.20E-13	+	
17903297	Total Cerebral Brain Volume (MRI)	VLDLR	rs502309	5.40E-06	(+)	1
17505501	Hepatic adverse events during ximelagatran treatment	WNT7A	rs1368576	2.50E-04	-	
17903295	Age at death	WRN	rs2543600	4.20E-06	+	
18521090	Response to iloperidone treatment	XKR4	rs9643483	1.30E-04	-	

"+" and "(+)" indicate the existence of previous human population study on the same phenotype or a related phenotype respectively, whereas "-" means that no previous human population study was identified (the candidate locus had been proposed based on other considerations only). The number of related phenotypes refers to the related phenotypes that had nominally statistically significant associations at $p<0.05$ (see **Supplementary Table 1** for details).

[1] Different types of phenotypes were found to be nominally associated with these genes ($p<0.05$, see **Supplementary Table 1** for details) suggesting the potential for pleiotropic effects

[2] p-value = 10^{-25} in the strict definition of stage 1 discovery

[3] The first genome-wide association study and the candidate-gene study (which renders CFH a candidate gene) were published concomitantly in the same journal issue of Science

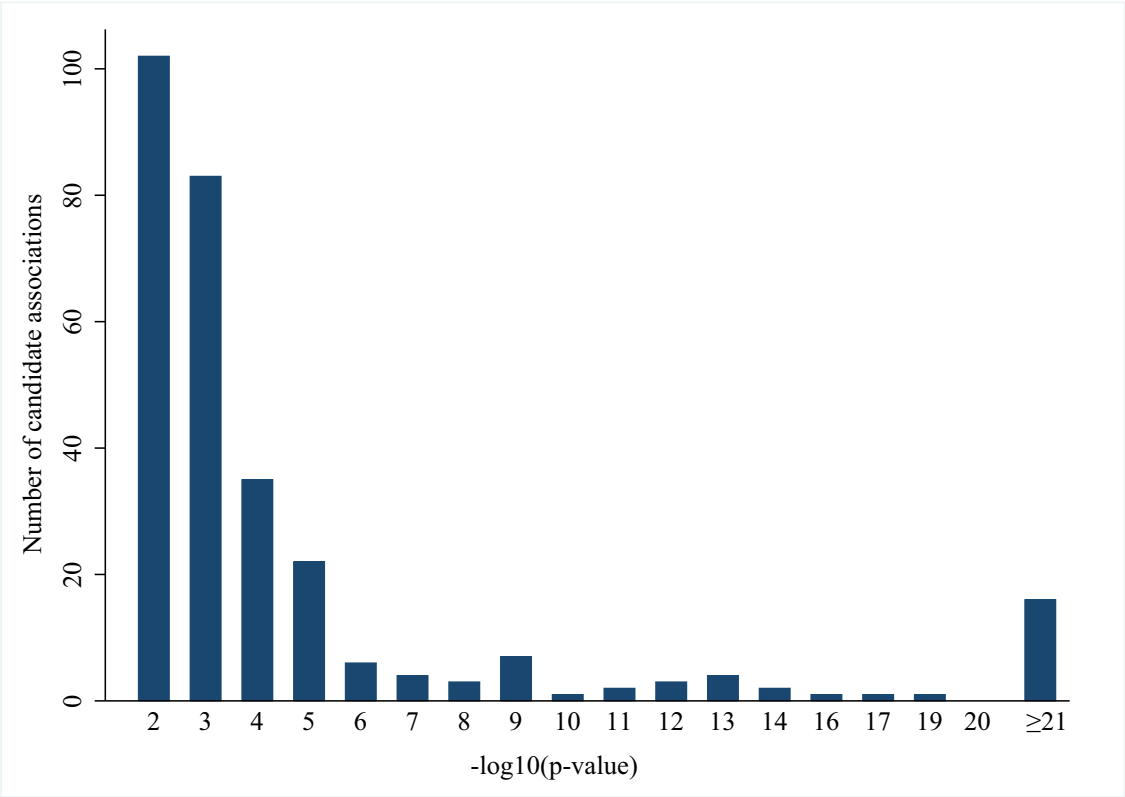
PMID: PubMed ID; LDL-C: Low-density lipoprotein cholesterol; HDL-C: High-density lipoprotein cholesterol; HbF: Fetal hemoglobin; CysC: Cystatin-C; FEV: Forced expiratory volume; FVC: Forced vital capacity; FEF: Forced expiratory flow; CRP: C-Reactive protein; sIL-6R: Soluble interleukin-6 receptor; SHBG: Sex-hormone binding globulin; NA: Not available

Table 2. Summary of binary phenotype associations with p-value < 10^{-7} for novel GWAS-derived SNPs versus variants in candidate loci in the 100 GWAS that reported also on candidate loci

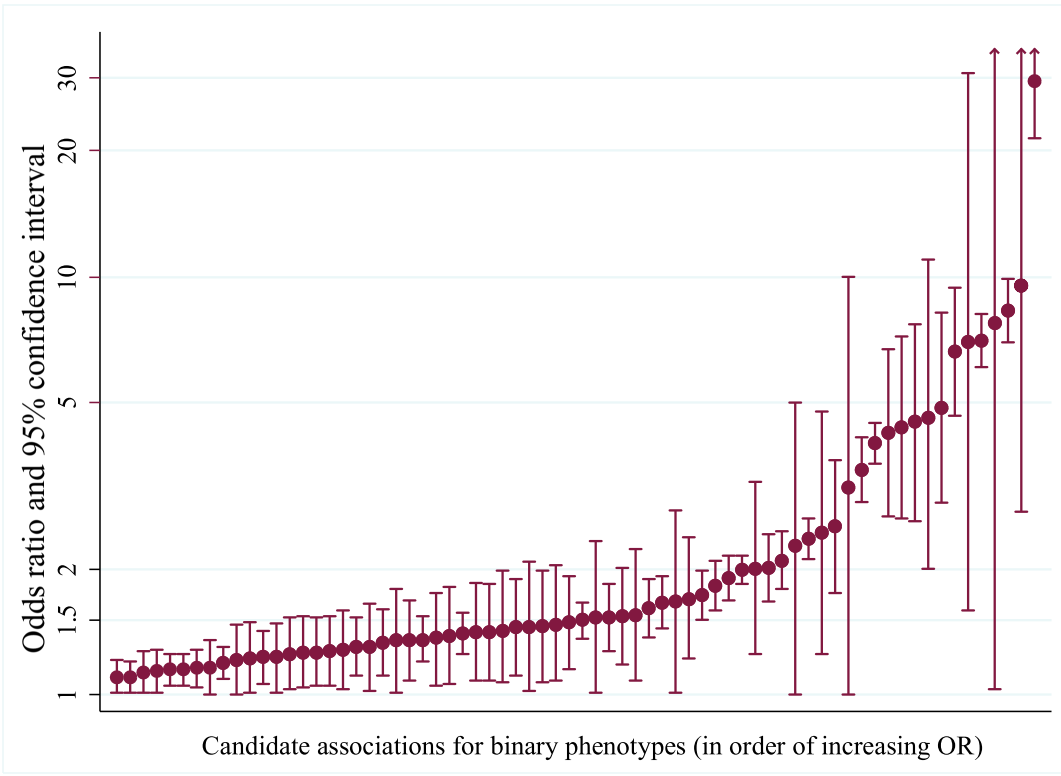
		No of SNPs	
		GWAS-derived (n=60)	**Candidates (n=22)**
Phenotypes	Breast cancer	6	0
	Prostate cancer	9	0
	Coronary artery disease	2	0
	Restless leg syndrome	1	0
	Bipolar disorder	2	0
	Gallstone disease	1	0
	Inflammatory bowel disease	10	1
	Type 2 diabetes	6	1
	Pigmentation	9	5
	Age-related macular degeneration	1	2
	Celiac disease	1	1
	Multiple sclerosis	1	1
	Psoriasis	1	1
	Rheumatoid arthritis	1	2
	Systemic lupus erythematosus	6	3
	Type 1 diabetes	3	3
	Alzheimer's disease	0	1
	Statin-induced myopathy	0	1

Associations are limited to those described in the 100 articles that did provide information on candidates. Otherwise, the number of credible associations is larger than what is shown here. Under "pigmentation" are included hair, skin and eye colour comparisons, freckles, skin sensitivity and tanning ability.

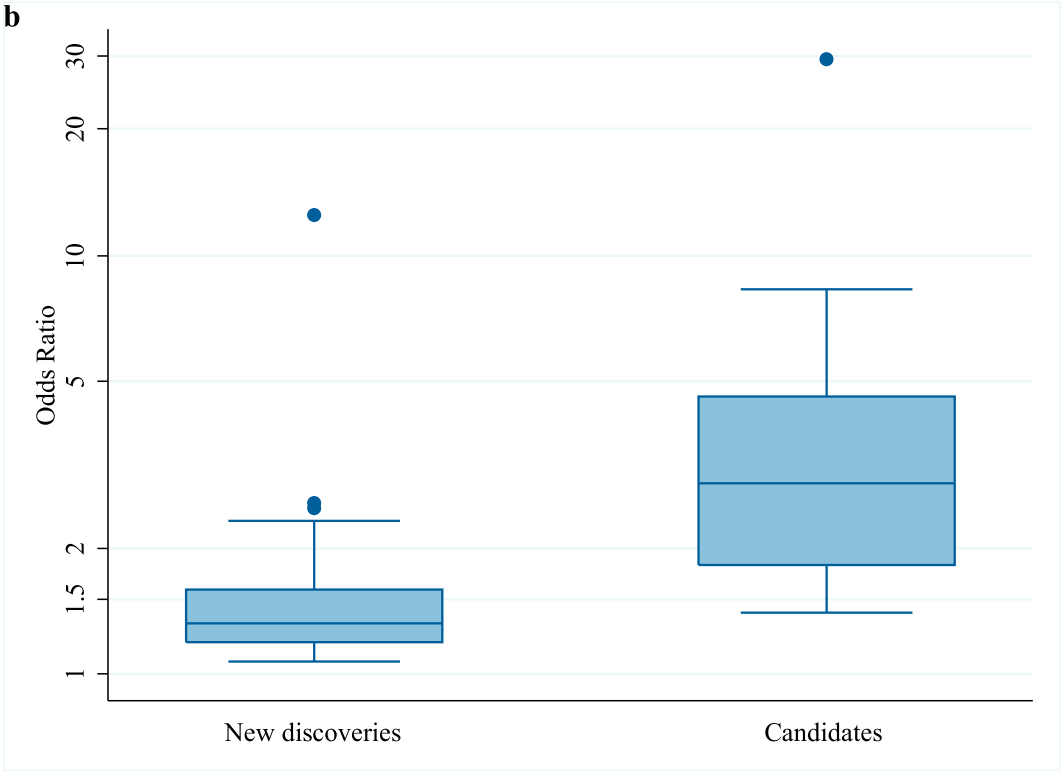
"Age-related macular degeneration" includes both dry and wet forms.


Number of candidate associations
0
20
40
60
80
100
2
3
4
5
6
7
8
9
10
11
12
13
14
16
17
19
20
≥21
-log10(p-value)

a


Odds ratio and 95% confidence interval
1
1.5
2
5
10
20
30
Candidate associations for binary phenotypes (in order of increasing OR)

b


Odds Ratio
1
1.5
2
5
10
20
30
New discoveries
Candidates