

# Potential etiologic and functional implications of genome-wide association loci for human diseases and traits

Lucia A. Hindorf<sup>a,1</sup>, Praveen Sethupathy<sup>b,1</sup>, Heather A. Junkins<sup>a</sup>, Erin M. Ramos<sup>a</sup>, Jayashri P. Mehta<sup>c</sup>, Francis S. Collins<sup>b,2</sup>, and Teri A. Manolio<sup>a,2</sup>

<sup>a</sup>Office of Population Genomics, <sup>b</sup>Genome Technology Branch, National Human Genome Research Institute, and <sup>c</sup>National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20892

Contributed by Francis S. Collins, March 28, 2009 (sent for review February 19, 2009)

We have developed an online catalog of SNP-trait associations from published genome-wide association studies for use in investigating genomic characteristics of trait/disease-associated SNPs (TASs). Reported TASs were common [median risk allele frequency 36%, interquartile range (IQR) 21%–53%] and were associated with modest effect sizes [median odds ratio (OR) 1.33, IQR 1.20–1.61]. Among 20 genomic annotation sets, reported TASs were significantly overrepresented only in nonsynonymous sites [OR = 3.9 (2.2–7.0),  $p = 3.5 \times 10^{-7}$ ] and 5kb-promoter regions [OR = 2.3 (1.5–3.6),  $p = 3 \times 10^{-4}$ ] compared to SNPs randomly selected from genotyping arrays. Although 88% of TASs were intronic (45%) or intergenic (43%), TASs were not overrepresented in introns and were significantly depleted in intergenic regions [OR = 0.44 (0.34–0.58),  $p = 2.0 \times 10^{-9}$ ]. Only slightly more TASs than expected by chance were predicted to be in regions under positive selection [OR = 1.3 (0.8–2.1),  $p = 0.2$ ]. This new online resource, together with bioinformatic predictions of the underlying functionality at trait/disease-associated loci, is well-suited to guide future investigations of the role of common variants in complex disease etiology.

catalog | evolution | GWAS | polymorphism | disorders

In the past 3 years, genome-wide association studies (GWAS) assaying hundreds of thousands of SNPs in thousands of individuals have reproducibly identified hundreds of associations of common genetic variants with over 80 diseases and traits (<http://www.genome.gov/gwastudies>). These studies have progressed from assaying fewer than 100,000 SNPs to more than one million, and sample sizes have increased dramatically as the search for variants that explain more of the disease/trait heritability has intensified (1). Important insights from these studies thus far include generally small effect sizes (odds ratios often <1.5), putative risk loci in or near genes not previously suspected of being involved in the etiology of a particular disease/trait, associated loci in common among diseases not previously thought to share etiologic pathways, and associations in many chromosomal regions currently annotated as gene poor (1).

The rapid increase in the number of GWAS provides an unprecedented opportunity to examine the potential impact of common genetic variants on complex diseases by systematically cataloging and summarizing key characteristics of the observed associations and the trait/disease associated SNPs (TASs) underlying them. Although some of these aspects have been examined on a smaller scale for individual diseases such as type 2 diabetes (2), inflammatory bowel disease (3), and cancer (4), a comprehensive genome-wide analysis across all GWAS published to date has not been conducted.

Identifying published GWAS can be challenging. For example, a simple PubMed search using the words “genome wide association studies” produced over 2,000 citations through December 2008, most of which are not actual GWAS. With this in mind, we developed the manually curated National Human Genome

Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies (<http://www.genome.gov/gwastudies>), an online, regularly updated database of SNP-trait associations extracted from published GWAS. Here, we (i) describe the features of this resource and the methods we have used to produce it, (ii) provide and examine key descriptive characteristics of reported TASs such as estimated risk allele frequencies and odds ratios, (iii) examine the underlying functionality of reported risk loci by mapping them to genomic annotation sets and assessing overrepresentation via Monte Carlo simulations and (iv) investigate the relationship between recent human evolution and human disease phenotypes. There are several challenges in conducting these analyses within the present context wherein the actual functional variant is often unknown. Specifically, for a given TAS, the causative variant may be: (i) the TAS itself, (ii) a known common SNP in strong linkage disequilibrium (LD) with the TAS, (iii) an unknown common SNP or rare single nucleotide variant tagged by a haplotype on which the TAS occurs, or (iv) a linked copy number variant. Due to the limited annotation for categories (iii) and (iv), our analyses focus on the role of reported TASs and their LD partners (TASPs) only. Further, the power of analyses to detect overrepresentation of causative variants in specific functional categories may be weakened by the need to consider all of the TASPs within the LD block tagged by a TAS (we will refer to this as a TAS block). To circumvent this issue we use a strategy that calculates the overrepresentation of unique TAS blocks in specific categories (i.e., for a particular category, a TAS block is counted once if one or more unique member TASPs map to the category) rather than individual TASPs (*Methods*).

## Results

**Descriptive and Association Data.** We examined associations from a total of 151 (of 237) published GWAS (through December, 2008) reporting at least one TAS at  $p < 5 \times 10^{-8}$  ([supporting information \(SI\) Table S1](#)). Including replication sample sizes reported in 130 (86%) of these 151 studies, the median total sample size for initial + replication studies was 7,858 participants (range, 146–91,749). From the 151 studies, we extracted information on 531 SNP-trait associations (limiting to one TAS per gene region and trait, as described in *SI Text*).

Author contributions: L.A.H., P.S., F.S.C., and T.A.M. designed research; L.A.H., P.S., H.A.J., E.M.R., J.P.M., and T.A.M. performed research; L.A.H., P.S., H.A.J., F.S.C., and T.A.M. contributed new reagents/analytic tools; L.A.H. and P.S. analyzed data; and L.A.H., P.S., H.A.J., E.M.R., J.P.M., F.S.C., and T.A.M. wrote the paper.

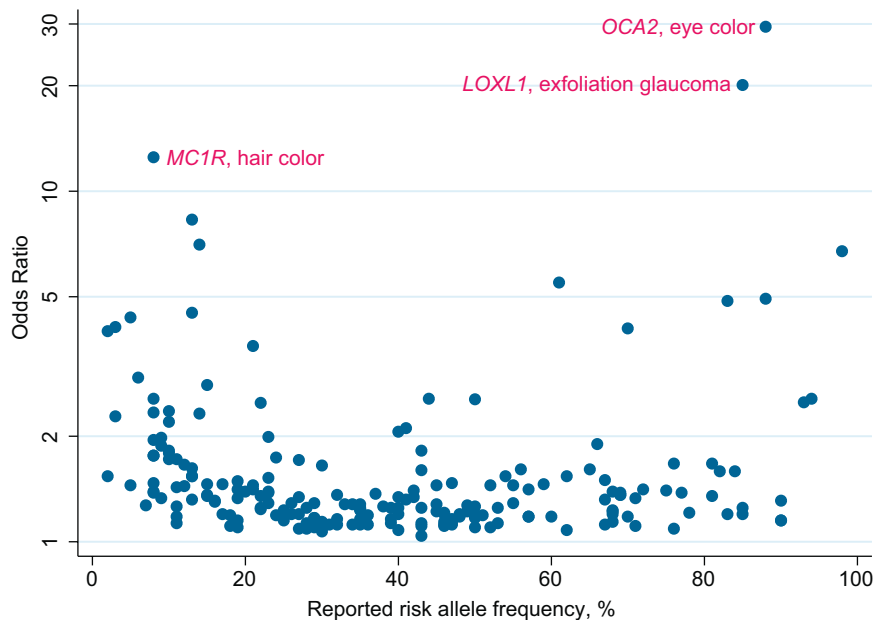
The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

<sup>1</sup>L.A.H. and P.S. contributed equally toward this work.

<sup>2</sup>To whom correspondence may be addressed. E-mail: francis.collins2@nih.gov or manolio@nih.gov.

This article contains supporting information online at [www.pnas.org/cgi/content/full/0903103106/DCSupplemental](http://www.pnas.org/cgi/content/full/0903103106/DCSupplemental).



**Fig. 1.** Published odds ratios for discrete traits by reported risk allele frequencies. Labeled SNP-trait associations are those with the highest ORs. Note that the y axis is on the log scale.

Unsurprisingly, since the GWAS method is primarily powered for common alleles, risk allele frequencies were well above 5% (median risk allele frequency 36%, interquartile range, IQR, 21%–53%) in the populations analyzed as well as in the HapMap populations (CEU: 37%, 21–54%; YRI: 33%, 13–65%; combined JPT+CHB 32%, 13–58%; Fig. S1).

The 531 reported SNP-trait associations represented 465 unique TASs; 43% ( $n = 199$ ) of which were located in an intergenic region, 45% ( $n = 208$ ) were intronic, 9% ( $n = 41$ ) were nonsynonymous, 2% ( $n = 10$ ) were in a 5' or 3' untranslated region, and 2% ( $n = 7$ ) were synonymous, according to the University of California Santa Cruz Genome Browser (5). Discrete traits were the focus of 227 (43%) of the 531 SNP-trait associations, which had associated odds ratios (ORs) ranging from 1.04 to 29.4 (median 1.33, IQR 1.20–1.61; Fig. S2). Among the discrete traits, the range of ORs was similar between nonsynonymous and other TASs; however, the right tail of the OR distribution for nonsynonymous TASs was slightly skewed toward higher values. The highest ORs were reported for pigmentation traits [Fig. 1; *MC1R* and hair color (6) and *OCA2* and eye color (6)]. SNP-trait associations were also distributed widely across diseases of high population prevalence, including heart disease, obesity, diabetes, and cancer (Table S2). Trait prevalence was not associated with the magnitude of ORs and risk allele frequencies, which were similar between the 10 most prevalent traits and all others combined (median ORs 1.26 and 1.29, respectively; median risk allele frequencies, 40% and 35%, respectively).

Among genes or regions harboring TASs that were reported in multiple studies of discrete traits, 18 were associated with seemingly distinct traits that may suggest clues toward common etiologic pathways (Table 1). Several TASs were located in previously characterized candidate genes, such as *APOE*, *HLA*, *KCNJ11*, *PPARG*, and *CARD15*, and were detected through GWAS at comparable effect sizes and stronger levels of statistical significance (Table S3). In these instances, GWAS-identified SNPs served as reasonable positive controls for known disease-associated genetic variants.

**Functional Analysis.** To assess the underlying functionality at the trait/disease-associated genetic loci, we systematically mapped

all TASPs (reported index TASs with an association  $p$  value  $< 5.0 \times 10^{-8}$  and all HapMap phase II CEU SNPs in LD [ $r^2 > 0.9$ ]) to 20 nonmutually exclusive genomic annotation sets (Table S4). For each annotation set, we did the following. For every unique TAS block, we determined whether any TASPs mapped to the annotation set. If none mapped, we did not count the block. However, if one or more TASPs mapped, then we counted 1 per block. To compute the odds of a TAS block mapping to the annotation set, we divided the number of unique TAS blocks that were counted in the annotation set ( $n$ ) by the number of TAS blocks that were not counted ( $N-n$ ). To evaluate whether any annotation set was significantly enriched or depleted for TAS blocks, we compared the observed odds with the expected odds calculated from 100 control datasets comprised of randomly selected SNPs and their LD partners. Importantly, the mapping and counting strategies were consistent across both the test and the control datasets to ensure a fair comparison. Further, the generation of the control datasets took into account the representation biases on the genotyping arrays that were used to identify the TASs (SI Text).

For 9 annotation sets (nonsynonymous sites, 1kb promoters, 5kb promoters, most conserved sequences (MCSs), 3' UTRs, microRNA target sites, Introns, CpG islands and experimentally validated regulatory regions from ORegAnno), the 95% confidence interval (CI) of the OR excluded 1.0 and the enrichment  $p$  values were  $< 0.05$  (Fig. S3), indicating that these categories may be significantly enriched for TAS blocks. Nonsynonymous sites had the strongest signal for enrichment (OR = 3.9 [2.2–7.0],  $p = 3.5 \times 10^{-7}$ ). After restricting the analysis to only those nonsynonymous SNPs predicted by PolyPhen (7) to be potentially deleterious (which reduces the sample size by approximately 65%), TAS blocks were even more strongly enriched (OR = 5.2 [1.8–15.3],  $p = 0.001$ ). Thirty nonsynonymous TASPs that are predicted to be potentially deleterious [by PolyPhen and an unpublished method, CDPred (P. Cherukuri and J. Mullikin, personal communication)] were identified as attractive candidates for functional follow-up (Table 2).

To examine the possibility that signals in other annotation sets might not represent bona fide TAS block enrichment, but rather a “hitchhiking” effect whereby TASPs closely linked with non-

**Table 1. Reported TASs associated with two or more distinct traits**

Chromosomal region	Rs number(s)	Attributed genes	Associated traits reported in catalog
1p13.2	rs2476601, rs6679677	<i>PTPN22</i>	Crohn's disease, type 1 diabetes, rheumatoid arthritis
1q23.2	rs2251746, rs2494250	<i>FCER1A</i>	Serum IgE levels, select biomarker traits (MCP1)
2p15	rs1186868, rs1427407	<i>BCL11A</i>	Fetal hemoglobin, F-cell distribution
2p23.3	rs780094	<i>GCKR</i>	CRP, lipids, waist circumference
6p21.33	rs3131379, rs3117582	<i>HLA / MHC region</i>	Systemic lupus erythematosus, lung cancer, psoriasis, inflammatory bowel disease, ulcerative colitis, celiac disease, rheumatoid arthritis, juvenile idiopathic arthritis, multiple sclerosis, type 1 diabetes
6p22.3	rs6908425, rs7756992, rs7754840, rs10946398, rs6931514	<i>CDKAL1</i>	Crohn's disease, type 2 diabetes
6p25.3	rs1540771, rs12203592, rs872071	<i>IRF4</i>	Freckles, hair color, chronic lymphocytic leukemia
6q23.3	rs5029939, rs10499194	<i>TNFAIP3</i>	Systemic lupus erythematosus, rheumatoid arthritis
7p15.1	rs1635852, rs864745	<i>JAZF1</i>	Height, type 2 diabetes*
8q24.21	rs6983267	<i>Intergenic</i>	Prostate or colorectal cancer, breast cancer
9p21.3	rs10811661, rs1333040, rs10811661, rs10757278, rs1333049	<i>CDKN2A, CDKN2B</i>	Type 2 diabetes, intracranial aneurysm, myocardial infarction
9q34.2	rs505922, rs507666, rs657152	<i>ABO</i>	Protein quantitative trait loci (TNF- $\alpha$ ), soluble ICAM-1, plasma levels of liver enzymes (alkaline phosphatase)
12q24	rs1169313, rs7310409, rs1169310, rs2650000	<i>HNF1A</i>	Plasma levels of liver enzyme (GGT), C-reactive protein, LDL cholesterol
16q12.2	rs8050136, rs9930506, rs6499640, rs9939609, rs1121980	<i>FTO</i>	Type 2 diabetes, body mass index or weight
17q12	rs7216389, rs2872507	<i>ORMDL3</i>	Asthma, Crohn's disease
17q12	rs4430796	<i>TCF2</i>	Prostate cancer, type 2 diabetes
18p11.21	rs2542151	<i>PTPN2</i>	Type 1 diabetes, Crohn's disease
19q13.32	rs4420638	<i>APOE, APOC1, APOC4</i>	Alzheimer's disease, lipids

\* The well known association of JAZF1 with prostate cancer was reported with a  $p$  value of  $2 \times 10^{-6}$  (18), which did not meet the threshold of  $5 \times 10^{-8}$  for this analysis.

synonymous SNPs map to nearby annotation sets and artificially increase their ORs, we removed all TASP blocks having  $r^2 > 0.6$  with any nonsynonymous HapMap CEU SNP and repeated the test (Fig. 2). Only 2 categories retained a clear signal for enrichment—1-kb promoters (OR = 3.0 [1.4 – 6.5],  $p = 0.005$ ) and 5-kb promoters (OR = 2.3 [1.5 – 3.6],  $p = 0.0003$ ). After Bonferroni correction for 20 comparisons, only the enrichment signal from 5-kb promoters remained significant, although this may reflect greater power relative to 1-kb promoters due to a larger number of mapped TASP blocks. Although no other category had even an uncorrected enrichment  $p$  value  $< 0.05$ , a few showed nonsignificant trends toward enrichment, such as the ORegAnno elements (OR = 2.0 [0.95 – 4.4],  $p = 0.09$ ). We note here that the number of TAS blocks mapping to 1kb promoter regions was similar to or smaller than several other categories (Table S4), so power is unlikely to explain the strong enrichment signal in promoter regions and lack thereof elsewhere.

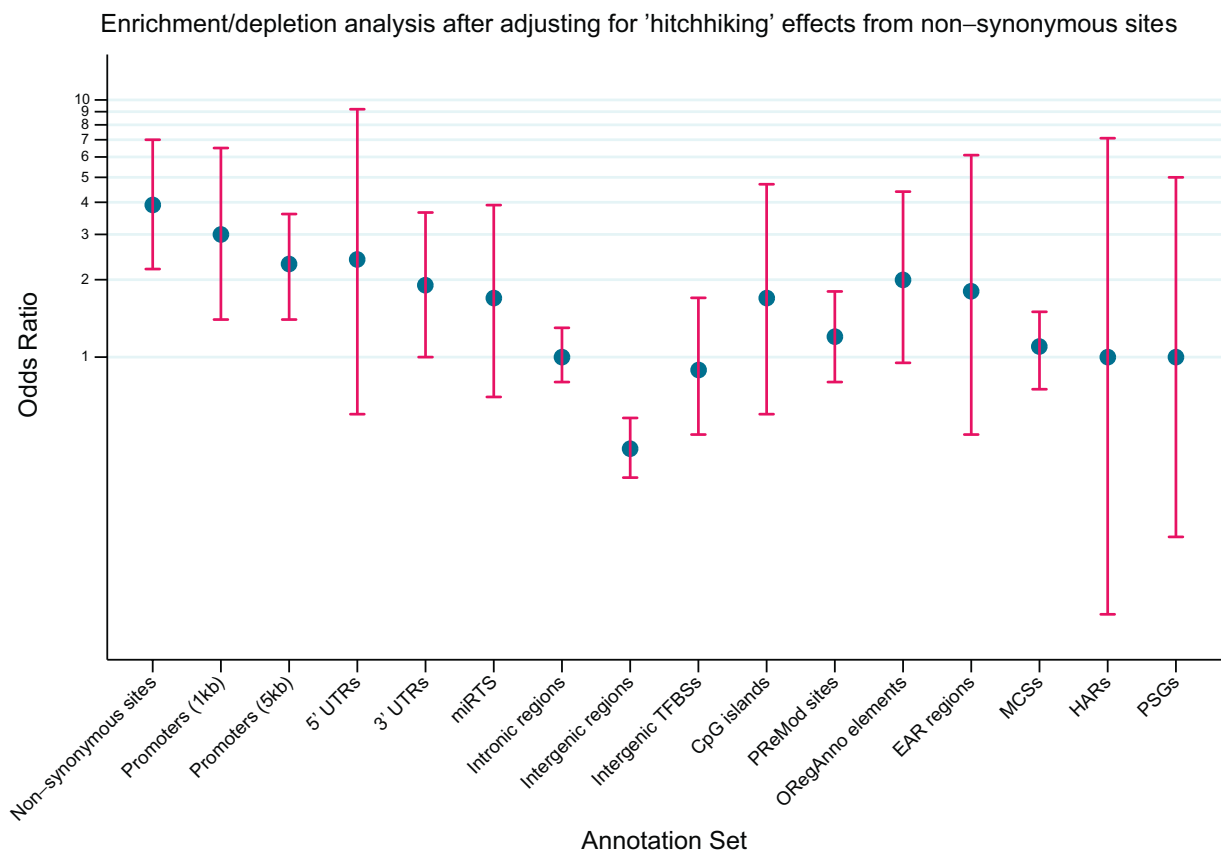
Using previous predictions of high confidence transcription factor binding sites in human 1-kb promoter regions, 4 TASP blocks were predicted to have strong allele-specific TF binding affinities (Table S5). These predictions may lead to compelling hypotheses for trait/disease etiology, as in the case of the protective allele [G] of SNP rs1077834, which is predicted to abolish a partially conserved binding site for HNF4 $\alpha$  696 base pairs upstream of the *LIPC* transcription start site. HNF4 $\alpha$  is an essential hepatic transcriptional activator that has been associated with metabolic pathways (8) and has been shown to activate *LIPC*, an important enzyme in lipid metabolism (9). One could thus hypothesize that the loss of HNF4 $\alpha$  binding in the presence of the protective allele may lower *LIPC* expression, leading to increased plasma HDL levels.

Intergenic regions, despite harboring the largest fraction of TAS blocks, were significantly depleted for TAS blocks (OR = 0.44 [0.34 – 0.58],  $p = 2.0 \times 10^{-9}$ ). This is consistent with the assumption that intergenic regions, although containing important regulatory sequences, have the smallest ratio of functional to total DNA. Intronic regions and several putative functional categories within intergenic regions (such as predicted intergenic transcription factor binding sites, experimentally supported enhancer regions, noncoding RNAs, and regions of conserved RNA secondary structures) did not show evidence for enrichment or depletion (Fig. 2). However, a definitive interpretation of this result is hindered by the current lack of extensive experimental annotation and noisy computational predictions of functional elements within intergenic regions.

Although conservation across species is a popular proxy for important functionality, enrichment analysis in the mammalian Most Conserved Sequences (MCSs) revealed no TAS block enrichment signal (OR = 1.07 [0.75 – 1.5],  $p = 0.79$ ). However, ORegAnno sites—experimentally supported regulatory elements of which many are nonconserved—did show a trend toward TAS block enrichment (Fig. 2; OR = 2.0 [0.95–4.4],  $p = 0.09$ ). This affirms the need for more experimental investigation into the architecture of noncoding regulatory elements (such as enhancers and microRNAs) to decrease the reliance on conservation and guide more integrative computational prediction methodologies.

To ensure the robustness of our results, we repeated the analyses using different  $r^2$  thresholds for defining LD partners ( $r^2 = 1.0$  and  $r^2 > 0.8$ ) and the results were essentially unchanged (data not shown). Although even lower  $r^2$  thresholds are reasonable for capturing more of the possible causative variants,





**Fig. 2.** Odds ratios for TAS block enrichment/depletion analysis after adjusting for “hitchhiking” effects from nonsynonymous sites. Four annotation sets (Splice sites, Validated enhancers, EvoFold elements, and noncoding RNAs) are not represented here because no TAS blocks mapped to these annotation sets. The blue circle represents the point estimate of the odds ratio (OR) and the red lines represent the 95% CI. Possible “hitchhiking” effects from nonsynonymous sites are reduced by discarding any TASP/control SNP in  $r^2 > 0.6$  with a nonsynonymous SNP. For an explanation of the annotation sets on the x axis, we refer the reader to Table S4. Note that the y axis is on the log scale. Nonsynonymous OR computation is not adjusted for “hitchhiking” effects.

provide important clues about the evolutionary history and underlying molecular mechanisms of certain TASP.

Several limitations of the underlying catalog data should be noted. We extracted all eligible associations from published articles and *SI Text*, but the number and quality of reported SNP associations is dependent upon the preferences of the individual author and journal. Also, the studies within the catalog generally test only those SNPs that are detectable via commonly used genotyping platforms in participants who tend to be from European-descent populations. The GWAS data are likely to be subject to varying degrees of upward bias in effect size estimates (the “winner’s curse” phenomenon), particularly to the extent that estimates from the GWAS discovery population, who may be less representative of the general population, influence those reported in our catalog. Nonetheless, in several instances in which known candidate SNPs have been previously identified, GWAS of the same trait tended to confirm these findings with similar effect sizes and stronger levels of statistical significance. Finally, TASP reported in published GWAS suffer from “lead TAS bias”; generally 1 or 2 TASP out of a cluster are selected from the initial study, often based on likely functional significance such as a conserved nonsynonymous site, for association analysis in the replication sample. To minimize the effect of this bias, we analyzed TAS blocks, which include the lead SNPs and their known LD partners based on HapMap phase II data. However, the true impact of the bias is difficult to quantify and it may still exert a slight effect on the enrichment/depletion signals especially for categories such as nonsynonymous sites.

An important question is to what extent GWAS have identified genetic variants likely to be of clinical or public health importance, particularly for developing preventive or therapeutic interventions. Answering this question must await better functional characterization of TASP or the true causative variants they may be tagging, evidence of effective interventions, and identification of potential modifiers of SNP-trait associations (1). However, the current study contributes empiric bounds on the expectations for the effect sizes and allele frequencies of TASP that can be identified from GWAS. It also highlights the distribution of promising SNP-trait associations across a wide variety of traits of substantial public health interest, such as obesity, hypertension, coronary artery disease, and cancer. Our results may guide future studies by highlighting genetic variants that are of particular interest from a descriptive, association, evolutionary, or functional perspective (such as predictions of TASP-mediated allele-specific transcription factor binding sites) and suggesting hypotheses for future study. Our description of GWAS-identified variants builds upon the important work previously targeted toward candidate genes, adding to a more complete picture of the contribution of common genetic variation to common diseases. It is clear, however, that the proportion of heritability explained by common variation for most common diseases to date is modest at best (17). As the power of the GWAS approach increases with access to more samples, and as the types of methods to test for genetic associations expand to include copy number variants and rarer alleles, more associations will likely be identified and timely analyses similar to those presented here will continue to update our knowledge of the influence of genomic structure and function on complex diseases.

