Supporting Information

Hindorff et al. 10.1073/pnas.0903103106

SI Text

Catalog Curation. Weekly PubMed searches were done using the terms "'genome-wide' OR 'genome AND identification' OR 'genome AND association'," with limits on the current year and human status. Articles and all available supporting information were downloaded. Studies focusing on copy number variants (CNV) were included in the catalog but known to be incomplete. Future efforts will focus on more completely ascertaining studies of CNVs.

Information was extracted on each study and each SNP as described in *Summary of Data Available in the NHGRI GWAS Catalog*. If the p value, OR, and 95% CI fields were not available for the combined population, we extracted estimates from the population group with the largest sample size. In extracting information, we followed these additional guidelines: missing or nonapplicable fields were denoted as follows: ?, allele not reported; NS, not significant (no associations at $p < 1 \times 10^{-5}$ were identified); NR, not reported. Where multiple genetic models were available, we prioritized effect sizes (ORs, variance proportions, increments) as follows: (*i*) genotypic model, per allele estimate; (*ii*) genotypic model, heterozygote estimate, (*iii*) allelic model, allelic estimate. Focusing on risk alleles, we inverted ORs <1 and their associated confidence intervals, and reported the opposite allele if available. If 95% CI were not published, we estimated them using standard errors where available (1). If more than one TAS within a gene met the above criteria, we reported one TAS unless there was evidence for an independent association. Associations attributed to a combination of one or more genetic variants were denoted as such in the rs number column (e.g., "rs1015362-G + rs4911414-T," "3-SNP haplotype 1"). If available, rs numbers for SNPs comprising the haplotype were indexed so that they would be searchable using the SNP search features described below. Genes attributed to a TAS were extracted verbatim from the published report; "intergenic" and "NR" (not reported) were used to denote a location which was not attributed to a particular gene (if it appeared that a gene was sought) or an absence of reporting on location information, respectively. The term "pending" was used to identify an eligible GWAS for which TAS information had not yet been extracted; studies of CNVs, which are known to be incompletely ascertained, are also noted as pending. We characterized the strength of the associations by per allele (additive) ORs for discrete traits and percent of variance explained or standard deviation increment per risk allele for quantitative traits.

To facilitate use of the catalog, we implemented several search features including journal title, first author (last name), disease/trait (string search or multiple option search), chromosomal region, reported gene name, SNP (rs number), OR (greater than a user-defined threshold), and p value (less than a user-defined threshold). The catalog data can also be downloaded as an Excel file.

Descriptive and Association Analyses. SNP allele frequencies for the three original HapMap populations were extracted from HapMart (http://www.hapmap.org/hapmart.html.en). Genomic annotations were extracted using the University of California Santa Cruz Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway). For the 10 associations in 5 papers that involved haplotypes or combinations of SNPs, we reported allele frequency and functional information only for the SNP with the most compelling functional annotation (nonsynonymous > synonymous/UTR > intronic > intergenic/near gene). Reported genes were extracted solely from the authors' published report; no attempt was made to standardize this information across papers using standard annotations or databases.

A trait was defined as discrete if the study design recruited participants or reported results based on presence or absence of the trait (e.g., case control status) or if a quantitative trait was dichotomized into 2 categories (e.g., skin pigmentation score above/below a certain threshold). For SNP-trait association analyses where the same SNP-trait combination was reported in multiple publications, we only included effect size and allele frequency information from the publication with the largest total sample size (85 associations excluded). We also identified instances in which reported genes harboring one or more TASs significant at $p < 5 \times 10^{-8}$ were observed in multiple reports or for multiple traits within a single report and determined whether the traits were very similar (such as body mass index and waist circumference) or seemingly distinct (such as type 1 diabetes and multiple sclerosis) based on our own judgment.

Analysis of Enrichment/Depletion in Annotation Sets. To compute the odds of a TAS block mapping to a particular annotation set (i.e., odds of at least one TASP within an LD block occurring in a particular annotation set), we mapped all TASPs onto the annotation set, counted the number of unique LD blocks (defined by the chosen r^2 threshold) with at least one mapped TASP, and divided by the number of LD blocks without any mapped TASPs. To assay for depletion or enrichment of TAS blocks in a particular annotation set, we first computed the odds of a randomly selected LD block mapping to the annotation set (i.e., odds of at least one SNP from a randomly selected LD block occurring in the annotation set). Specifically, we generated 100 random collections of SNPs where each collection's size was equal to the number of TASs and performed the following for each collection: expanded the collection by including LD partners from HapMap phase II data, mapped them onto the annotation set, counted the number of unique LD blocks with at least one mapped SNP and divided by the number of LD blocks without any mapped SNPs. We computed the expected odds by averaging across the 100 collections. We then computed the significance of the observed odds relative to this expected odds (through OR and two-tailed Fisher's Exact Test p value calculations). As the SNP arrays used in the various published GWAS may harbor substantial representational biases (e.g., the Illumina HumanHap300 platform includes a specific bias toward nonsynonymous sites), we generated the random collections of SNPs for the control dataset by drawing from SNP genotyping arrays according to the same distribution from which the TASs were identified. The genotyping platforms used in published GWAS and the percentage of TASs with $p < 5 \times 10^{-8}$ from each are provided here: Affymetrix 100K (4%), Affymetrix 250K (0.2%), Affymetrix 500K (20%), Affymetrix 5.0 (1%), Affymetrix 10K + 500K (0.9%), Affymetrix 100K + 500K (0.2%), Affymetrix 5.0 + 500K (0.4%), Affymetrix 6.0 + 500K (0.2%), Illumina 300K (33%), Illumina 550K (14%), Illumina 300K + 550K (0.9%), Affymetrix 500K + Illumina 300K (7%), Affymetrix 500K + Illumina 550K (0.6%), HapMap (16%) and Perlegen (2%).

Analysis of Deleterious Nonsynonymous TASs. The program PolyPhen (2) was used to determine whether a nonsynonymous TASP or control SNP was likely deleterious using predictions of whether one of the alleles has a benign, unknown, possibly damaging or probably damaging effect on protein structure. We repeated the enrichment analysis using only nonsynonymous SNPs predicted by PolyPhen to be possibly or probably damaging. To provide a list of all possible deleterious nonsynonymous TASPs, we combined those predicted by PolyPhen with those predicted by a novel, unpublished method, CDPred (P. Cherukuri and J. Mullikin, personal communication). CDPred assigns a "d-score" (deleterious score) for each nonsynonymous SNP. The d-score ranges from $+20$ (completely benign) to -30 (nonsense or frameshift) and $< = -3$ is considered deleterious.

Positive Selection Analysis via Integrated Haplotype Scoring (iHS). Integrated haplotype scores for all HapMap Phase II CEU SNPs were downloaded from http://hg-wen.uchicago.edu/selection/haplotter.htm. This method assigns an iHS for every HapMap Phase II SNP by measuring the differential extent of regional LD between the 2 alleles. To identify reported TASs under positive selection, we first selected one TAS per $r^2 > 0.6$ LD block. We then computed the number of such TASs with an iHS > 1.635 (which corresponds to the 90th percentile among HapMap Phase II CEU SNPs). We compared this number with the "expected" number (computed by averaging the number of randomly selected TASs with an iHS > 1.635 in each of the 100 control sets previously described) to compute the OR and p value.

Summary of Data Available in the NHGRI GWAS Catalog (www.genome.gov/gwastudies). Studies eligible for inclusion attempted to assay at least 100,000 SNPs in the initial design, excluding studies focusing only on candidate genes.

Study Level Information.

- Citation [last name of first author, title, journal, online or in print publication date (whichever was first), and HTML link to PubMed record]
- Trait/disease
- Initial sample size (summing across multiple Stage 1 populations, if applicable)
- Replication sample size (summing across all reported replication attempts)
- Genotyping platform manufacturer
- Number of SNPs passing quality control filters [using "up to (maximum number of SNPs)" if multiple platforms were used without imputation, the total number of imputed SNPs, or "pooled" to denote studies of pooled DNA, as applicable]
- Copy number variant study (initially excluded; additional studies to be added)

SNP-Trait Association Level Information.

- dbSNP reference number (rs number)
- Chromosomal region (extracted from University of California Santa Cruz Genome Browser)
- Gene(s) (as reported by authors)
- Risk allele
- Risk allele frequency in controls (if not available among all controls, among the control group with the largest sample size)
- p value and any relevant text (e.g., subgroups where applicable)
- Odds ratio, percent of variance explained, or increment size associated with risk allele, where specified, and 95% CI

Search Features.

- Journal title
- First author (last name)
- Disease/trait (2 options—string search or select multiple terms)
- Chromosomal region
- Reported gene name
- SNP (rs number)
- Odds ratio (greater than a user-defined threshold)
- p value (less than a user-defined threshold)

1. Rosner B, ed. (1995) *Fundamentals of Biostatistics* (Wadsworth Publishing Company, Belmont, CA).

2. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: Server and survey. *Nucleic Acids Res* 30:3894–3900.

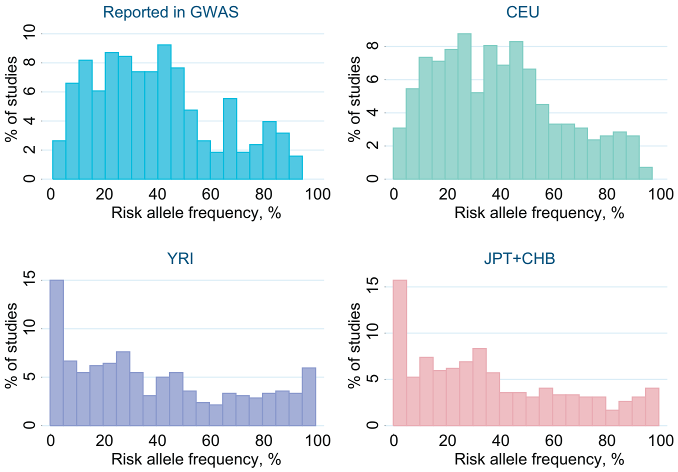

Fig. S1. Risk allele frequencies in published reports and HapMap populations.

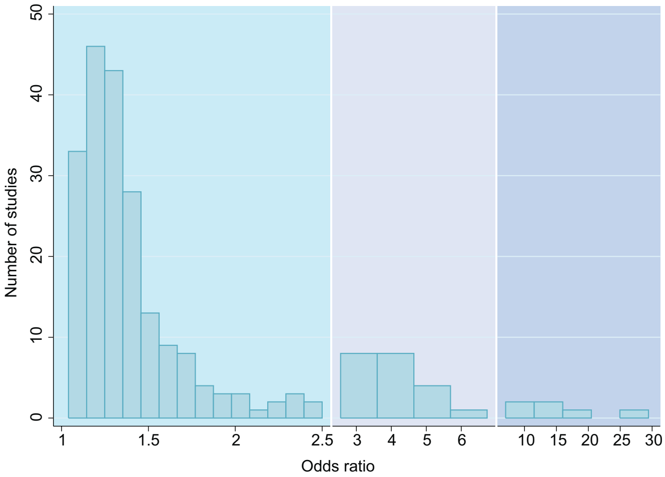

Fig. S2. Distribution of OR for discrete traits. Odds ratio thresholds indicate inclusive upper bound of each interval. Note the discontinuous *x* axis resulting from the juxtaposition of histograms of the following distributions: light blue, $1 < \text{OR} \leq 2.5$; lavender, $2.5 < \text{OR} \leq 7$; purple, $\text{OR} > 7$.

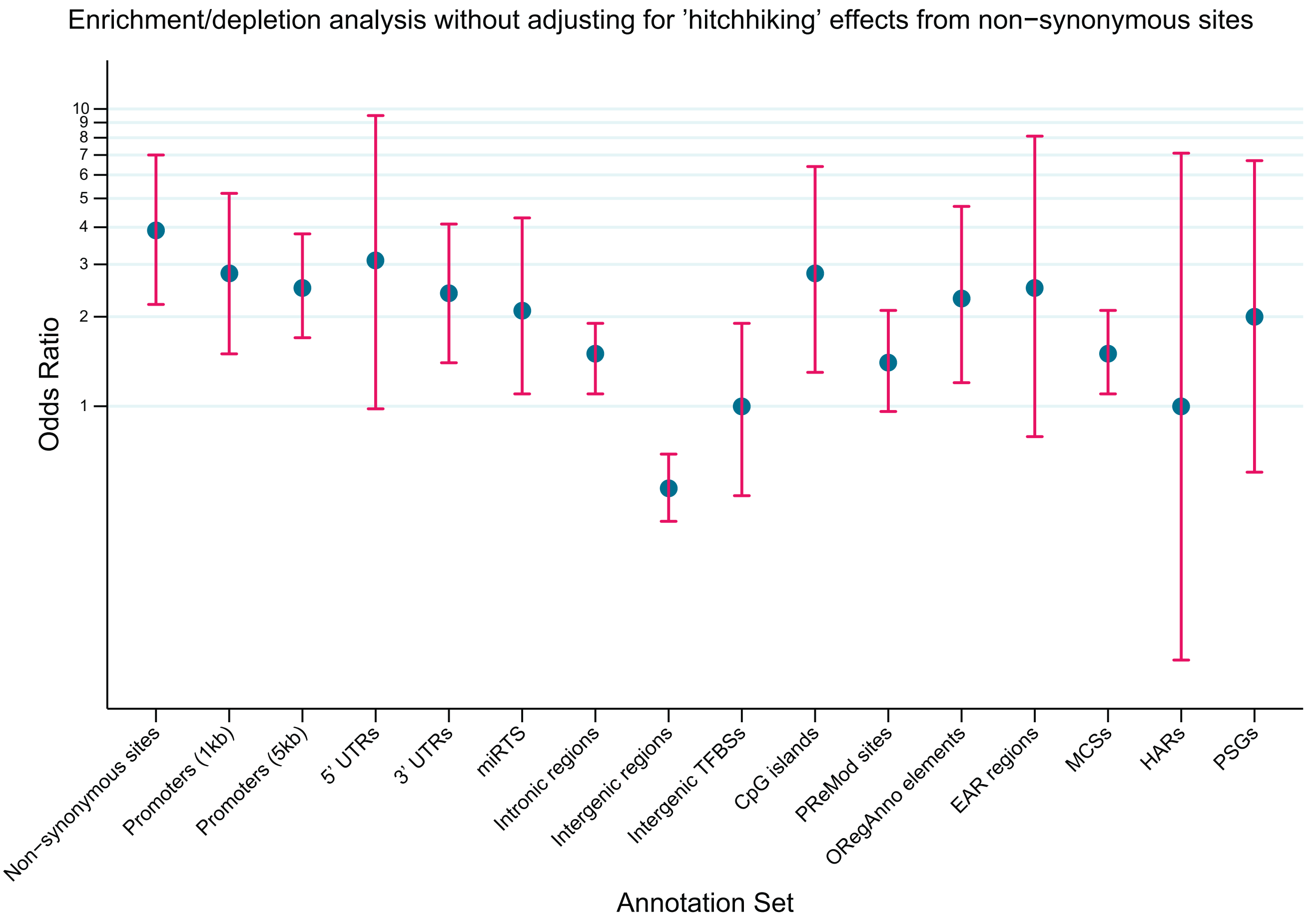

Fig. S3. ORs for TAS block enrichment/depletion analysis without adjusting for ''hitchhiking'' effects from nonsynonymous sites. Four annotation sets (Splice sites, Validated enhancers, EvoFold elements, and noncoding RNAs) are not represented here as their ORs are zero. The blue circle represents the point estimate of the OR and the red lines represent the 95% CI. For an explanation of each of the annotation sets on the *x* axis, please see Table S3. This analysis does not exclude any of the ''hitchhiking'' TASPs (those that are in $r^2 > 0.6$ with any nonsynonymous HapMap phase II CEU SNPs).

Table S1. Descriptive characteristics of the 151 GWAS publications and 531 associations included in the analysis

	Characteristic	N (%)[†]
Study publications (*N* = 151)[*]	Median combined[‡] sample size (range)	7,858 (146–91,479)
	Publication year	
	2005	1 (1)
	2006	3 (2)
	2007	54 (36)
	2008	93 (62)
	Replication sample reported	130 (86)
SNP-trait associations (*N* = 531)[*]	Median risk allele frequency, % (IQR)	36 (21–53)
	Median odds ratio, % (IQR)	1.33 (1.20–1.61)
	Associated with discrete outcome[§]	227 (43)
	Associated with quantitative trait[4]	304 (57)
	p-value	
	$<5 \times 10^{-8}$ and $\geq 10^{-9}$	136 (26)
	$<10^{-9}$ and $\geq 10^{-10}$	64 (12)
	$<10^{-10}$ and $\geq 10^{-20}$	215 (40)
	$<10^{-20}$	116 (22)

[*] Meeting a threshold of $p < 5 \times 10^{-8}$.

[†] Characteristics are reported as numbers (%), unless otherwise noted.

[‡] Combined across initial and replication sample sizes.

[§] A single TAS could be associated with both a quantitative and discrete trait.

Table S2. Number of independent reported SNP-trait associations at $p < 5 \times 10^{-8}$, most prevalent diseases. Only traits for which prevalence data for adults or an age-standardized population were available are reported. Traits for which prevalence data were only available in a limited subset of individuals (e.g., >65 years old) were excluded. Unless otherwise noted, prevalence rates are given for the adult (>18 years old) population. Number of independent associations refers to the number of SNP associations with each trait across all loci and publications (includes multiple SNPs published within the same paper and the same SNP reported in multiple publications); SNPs with $r^2 > 0.8$ were not considered independent. ″Obesity″ includes studies of BMI, weight and obesity

Disease/trait	Prevalence (per 10,000)	Number of independent reported associations	Prevalence source
Obesity	3,140*	19	National Institute of Diabetes and Digestive and Kidney Diseases weight control information network: http://www.win.niddk.nih.gov/statistics/#preval
Coronary disease	730	4	American Heart Association: http://www.americanheart.org/downloadable/heart/1200078608862HS_Stats%202008.final.pdf
Gallstones	710	1	(1)
Restless legs syndrome	700	6	(2)
Type 2 diabetes	530	21	Centers for Disease Control: http://www.cdc.gov/diabetes/statistics/prev/national/figage.htm
Myocardial infarction	400	2	Centers for Disease Control: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5606a2.htm#tab1
Bipolar disorder	260	2	National Institute of Mental Health: http://www.nimh.nih.gov/health/publications/the-numbers-count-mental-disorders-in-america.shtml#Bipolar
Stroke	260	1	American Heart Association, 2008 update. http://www.americanheart.org/downloadable/heart/1200078608862HS_Stats%202008.final.pdf
Psoriasis	220	3	National Psoriasis Foundation: http://www.psoriasis.org/about/stats/
Prostate cancer	163	18	Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov): Nov. 2007 data submission
Age-related macular degeneration	147	1	(3)
Breast cancer	140	8	Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov): Nov. 2007 data submission
Rheumatoid arthritis	60	10	Centers for Disease Control: http://www.cdc.gov/arthritis/arthritis/rheumatoid.htm#2
Male-pattern baldness	54	3	(4)

*Among adults > 20 years old.

1. Everhart JE, Khare M, Hill M, Maurer KR (1999) Prevalence and ethnic differences in gallbladder disease in the United States. *Gastroenterology* 117:632–639.
2. Zucconi M, Ferini-Strambi L (2004) Epidemiology and clinical findings of restless legs syndrome. *Sleep Medicine* 5:293–299.
3. Friedman DS, et al. (2004) Prevalence of age-related macular degeneration in the United States. *Arch Ophthalmol* 122:564–572.
4. Otberg N, Finner AM, Shapiro J (2007) Androgenetic alopecia. *Endocrinol Metab Clin North Am* 36:379–398.

Table S3. Comparison of selected associations reported using candidate gene and genome-wide association methods. Examples of well-characterized candidate loci were identified from Hirschhorn et al. (1); Glazier et al. (2); Ioannidis et al. (3); McCarthy and Zeggini (4). Estimates from candidate gene studies were preferentially extracted from meta-analyses where available and are presented in terms of the risk allele. If more than one candidate SNP within the specified gene was well-characterized, the association with the lowest p-value is presented

		Candidate gene association			Genome-wide association		
Candidate locus	Trait	Estimate† (95% CI)	p-value	Reference	Estimate† (95% CI)	p-value	Reference*
ADAM33	Asthma	1.46 [1.21–1.76]	3×10^{-4}	(5)	No reported associations.		
APOE	Alzheimer's disease	1.43 [1.3–1.57]	$<1 \times 10^{-8}$	(6)	4.01 (NR)	1×10^{-39}	(7)
					NR	2×10^{-44}	(8)
					NR	1×10^{-39}	(9)
	Lipids‡	44.0 [33.6–51.1] mg/dL higher LDL-C	NR	(10)	0.19% [0.15% - 0.23%] SD higher LDL -C	1×10^{-60}	(11)
		31% [23%–38%] higher LDL-C			6.61 (NR) mg/dl higher LDL-C	3×10^{-43}	(12)
CARD15 / NOD2	Crohn's Disease§	2.4 [1.4–4.3]	3.8×10^{-4}	(13)	3.99 [NR]	3×10^{-24}	(14)
					1.46 [1.29–1.64]	4×10^{-10}	(15)
CCR5	HIV progression	1.35 [1.03–1.79]	NR	(16)	No reported associations.		
CTLA4	Type 1 diabetes	1.45 [1.28–1.65]	< 0.001	(17)	NR	8 × 10-^{11}	(18)
F5	Venous thrombosis	4.24 [3.42–5.26]	< 0.001	(19)	Trait not in GWAS catalog		
GSTM1	Lung cancer	1.18 [1.14–1.23]	<0.01	(20)	No reported associations.		
HLA/MHC region	Type 1 diabetes	4.0 [NR]	< 0.0001	(21)	8.30 [6.97 - 9.89]	1×10^{-16}	(22)
					5.49 [4.83 - 6.24]	5×10^{-134}	(23)
KCNJ11	Type 2 diabetes	1.23 [1.12–1.36]	1.5×10^{-5}	(24)	1.16 [1.09–1.23]	4×10^{-7}	(25)
					1.14 [1.10–1.19]	7×10^{-11}	FUSION/WTCCC/DGI, 2007 (26–28)
MTHFR	Colorectal cancer	1.20 [1.08–1.33] (homozygote)	0.001	(29)	No reported associations.		
PPARG	Type 2 diabetes	1.27 [NR]	$< 2 \times 10^{-8}$	(30)	1.15 [1.10–1.21]	2×10^{-7}	(25)
					1.14 [1.08–1.20]	2×10^{-6}	FUSION/WTCCC/DGI, 2007 (26–28)
PRNP	Creutzfeldt-Jakob disease	2.86 [1.10–7.48]	0.03	(31)	Trait not in GWAS catalog		

NR, not reported; LDL-C, low density lipoprotein cholesterol.

*For ease of presentation, only selected GWAS findings are presented. A full listing can be found at www.genome.gov/gwastudies.

†Unless otherwise reported, effect sizes are odds ratios.

‡Includes triglycerides, LDL cholesterol

§Also includes irritable bowel syndrome, inflammatory bowel disease.

1. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. *Genet Med* 4: 45–61.
2. Glazier AM, Nadeau JH, Aitman TJ (2002) Finding genes that underlie complex traits. *Science* 298: 2345–2349.
3. Ioannidis JP, et al. (2006) A road map for efficient and reliable human genome epidemiology. *Nat Genet* 38: 3–5.
4. McCarthy MI, Zeggini E (2006) Genetics of type 2 diabetes. *Curr Diab Rep* 6: 147–154.
5. Contopoulos-Ioannidis DG, Kouri IN, Ioannidis JP (2007) Genetic predisposition to asthma and atopy. *Respiration* 74: 8–12.
6. Grupe A, et al. (2007) Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. *Hum Mol Genet* 16: 865–873.
7. Coon KD, et al. (2007) A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J Clin Psychiatry* 68: 613–618.
8. Li H, et al. (2008) Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol* 65: 45–53.
9. Webster JA, et al. (2008) Sorl1 as an Alzheimer's disease predisposition gene? *Neurodegener Dis* 5: 60–64.
10. Bennet AM, et al. (2007) Association of apolipoprotein E genotypes with lipid levels and coronary risk. *J Am Med Assoc* 298: 1300–1311.
11. Kathiresan S, et al. (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* 40:189–197.
12. Willer CJ, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* 40:161–169.
13. Oostenbrug LE, et al. (2006) CARD15 in inflammatory bowel disease and Crohn's disease phenotypes: an association study and pooled analysis. *Dig Liver Dis* 38: 834–845.
14. Barrett JC, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 40: 955–962.
15. Kugathasan S, et al. (2008) Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat Genet* 40: 1211–1215.
16. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. *Nat Genet* 29: 306–309.
17. Kavvoura FK, Ioannidis JP (2005) CTLA-4 gene polymorphisms and susceptibility to type 1 diabetes mellitus: A HuGE Review and meta-analysis. *Am J Epidemiol* 162: 3–16.
18. Cooper JD, et al. (2008) Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet* 40: 1399–1401.
19. Bezemer ID, et al. (2008) Gene variants associated with deep vein thrombosis. *J Am Med Assoc* 299: 1306–1314.
20. Ye Z, et al. (2006) Seven haemostatic gene polymorphisms in coronary disease: Meta-analysis of 66,155 cases and 91,307 controls. *Lancet* 367: 651–658.
21. Dorman JS, Bunker CH (2000) HLA-DQ locus of the human leukocyte antigen complex and type 1 diabetes mellitus: A HuGE review. *Epidemiol Rev* 22: 218–227.
22. Hakonarson H, et al. (2007) A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* 448: 591–594.
23. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661–678.
24. Gloyn AL, et al. (2003) Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes* 52: 568–572.
25. Zeggini E, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 40: 638–645.
26. Saxena R, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316: 1331–1336.
27. Scott LJ, et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316: 1341–1345.
28. Zeggini E, et al. (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316: 1336–1341.
29. Hubner RA, Houlston RS (2007) MTHFR C677T and colorectal cancer risk: A meta-analysis of 25 populations. *Int J Cancer* 120: 1027–1035.
30. Florez JC, Hirschhorn J, Altshuler D (2003) The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits. *Annu Rev Genomics Hum Genet* 4: 257–291.
31. Croes EA, et al. (2004) Polymorphisms in the prion protein gene and in the doppel gene increase susceptibility for Creutzfeldt-Jakob disease. *Eur J Hum Genet* 12: 389–394.

Table S4. Description of annotation sets and frequency of TAS blocks mapping to them

Annotation set	Description	Source	Frequency (n, %) of TAS blocks*	Frequency of Random LD blocks*
Non-synonymous	Genomic positions wherein a nucleotide substitution would cause an amino acid replacement	dbSNP version 129	57, 12.2%	15.8, 3.5%
Promoters (1kb)	1kb regions upstream of annotated transcription start sites	Ensembl v49 release at www.ensembl.org	26, 5.5%	8.8, 1.9%
Promoters (5kb)	5kb regions upstream of annotated transcription start sites	Ensembl v49 release at www.ensembl.org	65, 13.9%	30.3, 6.7%
5′ UTR	5′ untranslated regions	dbSNP version 129	7, 1.5%	2.8, 0.6%
3′ UTR	3′ untranslated regions	dbSNP version 129	28, 6.0%	14.6, 3.2%
miRTS	Predicted microRNA target sites (conserved and non-conserved) within 3′ UTRs	Predicted according to the TargetScan 4.2 algorithm (Grimson et al, 2007). Perl script downloaded from www.targetscan.org/.	15, 3.2%	8.1, 1.8%
Intronic	Non-coding regions within a gene	dbSNP version 129	189, 40.3%	183.2, 40.3%
Splice sites	Intronic regions that allow for splicing (removing introns and joining exons)	dbSNP version 129	0, 0%	0.1, 0.02%
Intergenic	Non-coding regions outside of genes	dbSNP version 129	186, 40.0%	276.0, 60.7%
Intergenic TFBSs	Predicted Human-Mouse-Rat conserved transcription factor binding sites in intergenic regions of the genome	UCSC Table Browser	21, 4.5%	20.3, 4.5%
CpG islands	Genomics regions that contain a high frequency of CG dinucleotides	UCSC Table Browser	10, 2.1%	5.1, 1.1%
Enhancers	Experimentally supported enhancer elements	Vista Enhancer Browser at enhancer.lbl.gov	0, 0%	0.42, 0.09%
PReMod	Predicted cis-regulatory modules	PReMod database at genomequebec.mcgill.ca/PReMod/	55, 11.7%	46.4, 10.2%
ORegAnno	Open source for Regulatory Annotation (experimentally supported regulatory regions)	UCSC Table Browser	18, 4.3%	9.2, 2.0%
EAR	Encode region Ancestral Repeats	UCSC Table Browser	7, 1.5%	3.0, 0.7%
EvoFold	Conserved RNA secondary structure	UCSC Table Browser	0, 0%	1.0, 0.2%
ncRNA	All types of experimentally supported non-coding RNA	RNAdb at research.imb.uq.edu.au/rnadb/	0, 0%	0.5, 0.10%
MCSs	Most Conserved Sequences across mammalian species	UCSC Table Browser	74, 15.8%	69.7, 15.3%
HAR	Regions under accelerated rates of substitution in the human genome	Bird et al, 2007, Pollard et al, 2006 and Prabhakar et al, 2006	2, 0.4%	1.1, 0.24%
PSG	Gene regions undergoing strong positive selection	UCSC Table Browser (derived from Kosiol et al, 2008)	3, 0.6%	2.4, 0.5%

* A TAS block is counted when at least one TASP within the LD block maps to the annotation set. Also, for all annotation sets (except for nonsynonymous) TASPs in $r^2 > 0.6$ with any nonsynonymous HapMap phase II CEU SNP are excluded from this count.

Table S5. TASPs in 1-kb promoter regions with putative allele-specific TF binding affinities. The 4 TASPs in this table are within human proximal promoters (defined as 1 kb upstream of every annotated transcription start site) and are not in even moderate LD [$r^2 > 0.6$) with any nonsynonymous SNP. The allele-specific binding affinities are derived from previous predictions from human promoters (1)

TASP	TF(s) predicted to bind reference allele	TF(s) predicted to bind non-reference allele	Downstream gene	Trait/Disease
rs1077834	HNF4	—	*LIPC*	HDL
rs573225	DBP	—	*G6PC2*	Fasting plasma glucose
rs7848647	CBF	—	*TNFSF15*	Inflammatory bowel disease
rs1420106	–	PAX-2,GATA-1	*IL18RAP*	Celiac disease

1. Sethupathy P, Giang H, Plotkin JB, Hannenhalli S (2008) *PLoS ONE* 3: e3137.

Table S6. Reported TASs potentially under positive selection. Positive selection is assessed according to the integrated haplotype score (iHS). Only reported TASs with $P < 5 \times 10^{-8}$ and only one TAS per $r^2 > 0.6$ LD block are included

TAS	Reported gene	Trait/disease	Putative selected allele (selection level)*	Risk allele
Metabolic disorders				
rs10923931	*NOTCH2*	T2D[†]	Ancestral (strong)	Derived
rs4402960	*IGF2BP2*	T2D	Ancestral (moderate)	Derived
rs492602	*FUT2*	Vitamin B12	Ancestral (strong)	Ancestral
rs4149268	*ABCA1*	HDL[‡] levels	Derived (strong)	Derived
rs173539	*CETP*	HDL levels	Ancestral (moderate)	Ancestral
rs7395662	*MADD, FOLH1*	HDL levels	Ancestral (strong)	Ancestral
rs10889353	*DOCK7*	Total cholesterol levels	Ancestral (moderate)	Derived
rs3846662	*HMGCR*	Total cholesterol levels	Derived (moderate)	Ancestral
rs4939883	*LIPG*	Total cholesterol levels	Ancestral (strong)	Derived
rs10913469	*SEC16B, RASAL2*	Weight	Ancestral (strong)	Ancestral
rs10838738	*MTCH2*	BMI	Derived (moderate)	Derived
rs1121980	*FTO*	BMI/obesity	Ancestral (moderate)	Ancestral
Autoimmune disorders				
rs6822844	Unknown	Celiac Disease	Derived (moderate)	Ancestral
rs660895	*HLA-DRB1*	Rheumatoid arthritis	Derived (strong)	Unknown
rs6920220	Unknown	Rheumatoid arthritis	Derived (moderate)	Unknown
rs12722489	*IL2RA*	Multiple sclerosis	Derived (moderate)	Ancestral
rs3129934	*HLA-DRB1*	Multiple sclerosis	Derived (strong)	Unknown
rs744166	*STAT3*	Crohn's Disease	Derived (moderate)	Derived
rs12708716	*CLEC16A*	T1D[†]	Ancestral (strong)	Ancestral
rs2647044	*HLA-E*	T1D	Ancestral (strong)	Ancestral
rs2188962	*LOC441108*	Crohn's Disease	Derived (strong)	Derived
rs17696736	*C12orf30*	T1D	Derived (strong)	Derived
rs13015714	*IL18R1*	Celiac Disease	Derived (strong)	Derived
rs10210302	*ATG16L1*	Crohn's Disease	Ancestral (moderate)	Ancestral
rs17810546	*IL12A*	Celiac Disease	Derived (strong)	Derived
rs5743289	*NOD2*	Inflammatory bowel disease	Ancestral (moderate)	Ancestral
Melanin synthesis				
rs11855019	*OCA2*	Blond hair color	Derived (strong)	Unknown
rs12913832	*HERC2*	Blond hair color	Derived (moderate)	Ancestral
rs1408799	*TYRP1*	Blue eye color	Derived (strong)	Derived
rs1042602	*TYR*	Freckles	Derived (moderate)	Derived
rs916977	*HERC2*	Iris color	Derived (strong)	Unknown
Cancer				
rs721048	*EHBP1*	Prostate cancer	Derived (moderate)	Derived
rs11083846	*PRKD2*	CLL[§]	Derived (moderate)	Derived
rs735665	*GRAMD1B*	CLL	Derived (moderate)	Derived
rs872071	*IRF4*	CLL	Derived (moderate)	Derived
rs7538876	*PADI6*	Cutaneous basal cell carcinoma	Ancestral (strong)	Ancestral
rs3117582	*BAT3MSH5*	Lung cancer	Derived (strong)	Unknown
rs10411210	*RHPN2*	Colorectal cancer	Ancestral (moderate)	Ancestral
Height				
rs798544	*GNA12*	Height	Ancestral (moderate)	Derived
rs1635852	*JAZF1*	Height	Derived (moderate)	Ancestral
rs6060373	*GDF5*	Height	Ancestral (moderate)	Derived
rs4533267	*ADAMTS17*	Height	Ancestral (moderate)	Ancestral
Other				
rs4128725	*OR10J1*	MCP1[¶] levels	Derived (strong)	Unknown
rs10494366	*NOS1AP*	QT interval	Ancestral (strong)	Unknown
rs10496265	Unknown	Aging	Derived (moderate)	Unknown
rs4355801	*TNFRSF11B*	Bone Mineral Density	Derived (strong)	Ancestral
rs10778213	Unknown	C-reactive protein	Derived (moderate)	Ancestral
rs1970546	*CDH4*	Volumetric brain MRI	Ancestral (strong)	Unknown
rs10958409	*SOX17*	Intracranial aneurysm	Ancestral (moderate)	Derived
rs2373115	*GAB2*	Late-onset Alzheimer's in APOE*e4 carriers	Derived (moderate)	Derived

*Moderate selection is defined as 1.635 < iHS < 2.0 and strong selection is defined as iHS > 2.0.

[†] T1D and T2D: Type 1 and Type 2 Diabetes, respectively

[‡] HDL: High Density Lipoprotein

[§] CLL: Chronic Lymphocytic Leukemia

[¶] MCP1: Monocyte Chemotactic Protein-1 which is involved in the immune response to injury and infection