



Systematic Localization of Common Disease-Associated Variation in Regulatory DNA

Matthew T. Maurano *et al.*
Science **337**, 1190 (2012);
DOI: 10.1126/science.1222794

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of November 13, 2013):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/content/337/6099/1190.full.html>

Supporting Online Material can be found at:

<http://www.sciencemag.org/content/suppl/2012/09/05/science.1222794.DC2.html>

<http://www.sciencemag.org/content/suppl/2012/09/04/science.1222794.DC1.html>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/content/337/6099/1190.full.html#related>

This article **cites 45 articles**, 13 of which can be accessed free:

<http://www.sciencemag.org/content/337/6099/1190.full.html#ref-list-1>

This article has been **cited by** 32 articles hosted by HighWire Press; see:

<http://www.sciencemag.org/content/337/6099/1190.full.html#related-urls>

This article appears in the following **subject collections**:

Genetics

<http://www.sciencemag.org/cgi/collection/genetics>

Systematic Localization of Common Disease-Associated Variation in Regulatory DNA

Matthew T. Maurano,^{1*} Richard Humbert,^{1*} Eric Rynes,^{1*} Robert E. Thurman,¹ Eric Haugen,¹ Hao Wang,¹ Alex P. Reynolds,¹ Richard Sandstrom,¹ Hongzhu Qu,^{1,2} Jennifer Brody,³ Anthony Shafer,¹ Fidencio Neri,¹ Kristen Lee,¹ Tanya Kutayin,¹ Sandra Stehling-Sun,¹ Audra K. Johnson,¹ Theresa K. Canfield,¹ Erika Giste,¹ Morgan Diegel,¹ Daniel Bates,¹ R. Scott Hansen,⁴ Shane Neph,¹ Peter J. Sabo,¹ Shelly Heimfeld,⁵ Antony Raubitschek,⁶ Steven Ziegler,⁶ Chris Cotsapas,^{7,8} Nona Sotoodehnia,^{3,9} Ian Glass,¹⁰ Shamil R. Sunyaev,¹¹ Rajinder Kaul,⁴ John A. Stamatoyannopoulos^{1,12†}

Genome-wide association studies have identified many noncoding variants associated with common diseases and traits. We show that these variants are concentrated in regulatory DNA marked by deoxyribonuclease I (DNase I) hypersensitive sites (DHSs). Eighty-eight percent of such DHSs are active during fetal development and are enriched in variants associated with gestational exposure-related phenotypes. We identified distant gene targets for hundreds of variant-containing DHSs that may explain phenotype associations. Disease-associated variants systematically perturb transcription factor recognition sequences, frequently alter allelic chromatin states, and form regulatory networks. We also demonstrated tissue-selective enrichment of more weakly disease-associated variants within DHSs and the de novo identification of pathogenic cell types for Crohn's disease, multiple sclerosis, and an electrocardiogram trait, without prior knowledge of physiological mechanisms. Our results suggest pervasive involvement of regulatory DNA variation in common human disease and provide pathogenic insights into diverse disorders.

Disease- and trait-associated genetic variants are rapidly being identified with genome-wide association studies (GWAS) and related strategies (1). To date, hundreds of GWAS have been conducted, spanning diverse diseases and quantitative phenotypes (2) (fig. S1A). However, the majority (~93%) of disease- and trait-associated variants emerging from these studies lie within noncoding sequence (fig. S1B), complicating their functional evaluation. Several lines of evidence suggest the involvement of a proportion of such variants in transcriptional regulatory mechanisms, including modulation of promoter

and enhancer elements (3–6) and enrichment within expression quantitative trait loci (eQTL) (3, 7, 8).

Human regulatory DNA encompasses a variety of cis-regulatory elements within which the cooperative binding of transcription factors creates focal alterations in chromatin structure. Deoxyribonuclease I (DNase I) hypersensitive sites (DHSs) are sensitive and precise markers of this actuated regulatory DNA, and DNase I mapping has been instrumental in the discovery and census of human cis-regulatory elements (9). We performed DNase I mapping genome-wide (10) in 349 cell and tissue samples, including 85 cell types studied under the ENCODE Project (10) and 264 samples studied under the Roadmap Epigenomics Program (11). These encompass several classes of cell types, including cultured primary cells with limited proliferative potential ($n = 55$); cultured immortalized ($n = 6$), malignancy-derived ($n = 18$), or pluripotent ($n = 2$) cell lines; and primary hematopoietic cells ($n = 4$) as well as purified differentiated hematopoietic cells ($n = 11$), and a variety of multipotent progenitor and pluripotent cells ($n = 19$). We also surveyed regulatory DNA by generating DHS maps from 233 diverse fetal tissue samples across days ~60 to 160 after conception (late-first to late-second trimester of gestation). We used a uniform processing algorithm to identify DHSs and the surrounding boundaries of DNase I accessibility (i.e., the nucleosome-free region harboring regulatory factors) (12). We defined an average of 198,180 DHSs per cell type (range 89,526 to 369,920; table S1) spanning on average ~2.1% of the ge-

nome. In total, we identified 3,899,693 distinct DHS positions along the genome (collectively spanning 42.2%), each of which was detected in one or more cell or tissue types (median = 5).

Disease- and trait-associated variants are concentrated in regulatory DNA. We examined the distribution of 5654 noncoding genome-wide significant associations [5134 unique single-nucleotide polymorphisms (SNPs); fig. S1 and table S2] for 207 diseases and 447 quantitative traits (2) with the deep genome-scale maps of regulatory DNA marked by DHSs. This revealed a collective 40% enrichment of GWAS SNPs in DHSs (fig. S1C, $P < 10^{-55}$, binomial, compared to the distribution of HapMap SNPs). Fully 76.6% of all noncoding GWAS SNPs either lie within a DHS (57.1%, 2931 SNPs) or are in complete linkage disequilibrium (LD) with SNPs in a nearby DHS (19.5%, 999 SNPs) (Fig. 1A) (12). To confirm this enrichment, we sampled variants from the 1000 Genomes Project (13) with the same genomic feature localization (intronic versus intergenic), distance from the nearest transcriptional start site, and allele frequency in individuals of European ancestry. We confirmed significant enrichment both for SNPs within DHSs ($P < 10^{-59}$, simulation) and also including variants in complete LD ($r^2 = 1$) with SNPs in DHSs ($P < 10^{-37}$, simulation) (fig. S2).

In total, 47.5% of GWAS SNPs fall within gene bodies (fig. S1B); however, only 10.9% of intronic GWAS SNPs within DHSs are in strong LD ($r^2 \geq 0.8$) with a coding SNP, indicating that the vast majority of noncoding genic variants are not simply tagging coding sequence. Analogously, only 16.3% of GWAS variants within coding sequences are in strong LD with variants in DHSs. SNPs on widely used genotyping arrays (e.g., Affymetrix) were modestly enriched within DHSs (fig. S2), possibly due to selection of SNPs with robust experimental performance in genotyping assays. However, we found no evidence for sequence composition bias (table S3).

To further examine the enrichment of GWAS SNPs in regulatory DNA, we systematically classified all noncoding GWAS SNPs by the quality of their experimental replication. This disclosed 2436 unreplicated SNPs; 2374 “internally replicated” SNPs (confirmed in a second population in the initial publication); and 324 “externally replicated” SNPs (confirmed in an independent study) (table S2) (12). We observed a monotonic increase in the proportion of disease or trait variants localizing in DHSs with increasing quality of GWAS SNP experimental replication (Fig. 1B), as well as with increasing strength of association and study sample size (fig. S3). For externally replicated noncoding SNPs, 69.8% lie within a DHS ($n = 226$, $P < 10^{-14}$, simulation, fig. S2). To exclude the influence of population stratification, we compared the fixation index in African and European populations between GWAS SNPs in DHSs and matched SNPs not in DHSs and found them to be nearly identical ($F_{ST} = 0.0843$ versus 0.0847,

¹Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA. ²Laboratory of Disease Genomics and Individualized Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, 100029, China. ³Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA 98195, USA. ⁴Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA 98195, USA. ⁵Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. ⁶Immunology Program, Benaroya Research Institute, Seattle, WA 98101, USA. ⁷Department of Neurology, Yale University School of Medicine, New Haven, CT 06520, USA. ⁸Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA. ⁹Division of Cardiology, Department of Medicine, University of Washington, Seattle, WA 98195, USA. ¹⁰Division of Genetic Medicine, Department of Pediatrics, University of Washington, Seattle, WA 98195, USA. ¹¹Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA. ¹²Department of Medicine, Division of Oncology, University of Washington, Seattle, WA 98195, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: jstam@uw.edu

respectively) (12). The monotonic relationship between evidence for association and SNP concentration in DHSs strongly suggests that many variants are functional and that unreplicated or weaker associations may obscure the true degree of enrichment in DHSs.

GWAS variants localize in cell- and developmental stage-selective regulatory DNA. We observed selective localization within physiologically or pathogenically relevant specific cell or tissue types, including affected tissues or known or posited effector cell types (Fig. 1C). For a given disorder, cell-selective localization within physiologically or pathogenically relevant cell types was repeatedly observed for multiple independently associated SNPs distributed widely around the genome (fig. S4). These results suggest a tissue-specific regulatory role for many common variants, as well as the potential for comprehensive regulatory DNA maps to illuminate associations within disease-relevant cell types.

Many common disorders have been linked with early gestational exposures or environmental insults (14). Because of the known role of the chromatin accessibility landscape in mediating responses

to cellular exposures such as hormones (15), we examined if DHSs harboring GWAS variants were active during fetal developmental stages. Of 2931 noncoding disease- and trait-associated SNPs within DHSs globally, 88.1% (2583) lie within DHSs active in fetal cells and tissues. Of DHSs containing disease-associated variation, 57.8% are first detected in fetal cells and tissues and persist in adult cells ("fetal origin" DHSs), whereas 30.3% are fetal stage-specific DHSs (Fig. 1D). GWAS variants in adult stage-specific DHSs localize chiefly in mature hematopoietic cells, connective tissue, endothelial cells, and malignant cells (fig. S6).

We next analyzed the enrichment or depletion of replicated disease-specific GWAS variants in fetal-stage DHSs relative to the proportion of total GWAS SNPs in these DHSs. We found the greatest enrichment in phenotypes for which gestational exposures or growth trajectory have been shown to play major roles, including menarche, cardiovascular disease, and body mass index (Fig. 1E) (14, 16). By contrast, we observed relative depletion in fetal DHSs of aging-related diseases, cancer, and inflammatory disorders with presumed (postnatal) environmental triggers. These

findings suggest a recurring connection between an exposure-responsive gestational chromatin landscape, regulatory genotype, and risk for specific classes of adult diseases and traits.

DHSs harboring GWAS variants control distant phenotype-relevant genes. Enhancers may lie at great distances from the gene(s) they control (17) and function through long-range regulatory interactions (4, 18), complicating the identification of target genes of regulatory GWAS variants. Most DHSs display quantitative, cell-selective DNase I hypersensitivity patterns, which may be systematically correlated with DNase I sensitivity patterns at cis-linked promoters. DHSs that are strongly correlated ($r > 0.7$) with specific promoters function as enhancers that physically interact with their target promoter as detected by chromosome conformation capture methods, including chromosome conformation capture carbon copy (5C) and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) (10).

To systematically identify the genic targets of DHSs harboring GWAS variants and thereby gain insights into disease mechanisms, we applied the approach described in (10) to the much broader

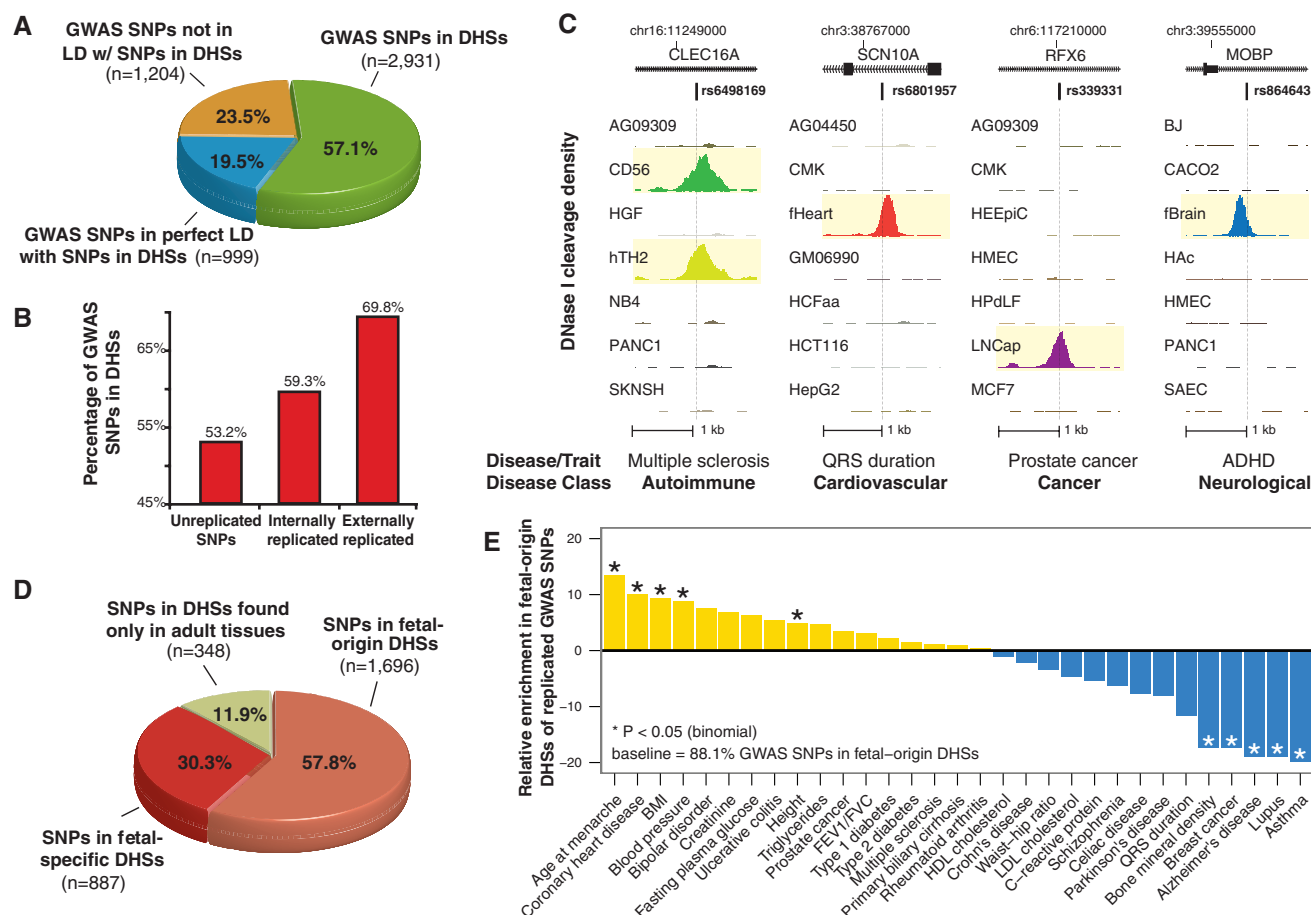


Fig. 1. Disease-associated variation is concentrated in DNase I hypersensitive sites. **(A)** Proportions of noncoding GWAS SNPs localizing within DHSs (green); in complete linkage disequilibrium ($r^2 = 1$) with a SNP in a DHS (blue); or neither (yellow). Note that 76.5% of GWAS SNPs are either within or in perfect LD with DHSs. **(B)** Proportions of GWAS SNPs overlapping DHSs after partition-

ing by degree of replication. **(C)** Representative DNase I hypersensitivity (tag density) patterns at diverse disease-associated variants. **(D)** Proportion of GWAS SNPs localizing in DHSs active in fetal tissues that persist in adult cells (salmon); fetal stage-specific DHSs (red); and adult stage DHSs (green). **(E)** GWAS SNPs in DHSs show phenotype-specific enrichment for fetal regulatory elements.

range of cell and tissue types in the present study (12) and intersected the result sets with GWAS data. This analysis revealed 419 DHSs harboring GWAS variants that were strongly correlated ($r > 0.7$) with the promoter of a specific target gene within ± 500 kb of the DHS (tables S6 and S7). These include numerous examples of target genes that plausibly explain the disease or trait association (Table 1 and fig. S7). For example, a SNP (rs385893) associated with platelet count (19) lies in a DHS tightly correlated ($r = 0.97$) and physically interacts with the 222-kb distant promoter of JAK2, a cytokine-activated signal transducer

linked with platelet production and myeloproliferative disorders (Fig. 2A). Fully 40.8% of correlated DHS-gene pairs span >250 kb (Fig. 2B), and 79% represent pairings with distant promoters versus those of the nearest gene (tables S6 and S7). Notably, these interactions typically extend beyond the range of LD (mean $r^2 = 0.06$; table S6).

GWAS variants in DHSs frequently alter allelic chromatin state. Next, we examined how GWAS variants in DHSs were distributed with respect to transcription factor recognition sequences, defined with a scan for known motif models at a stringency of $P < 10^{-4}$ (12). Of GWAS SNPs in DHSs, 93.2%

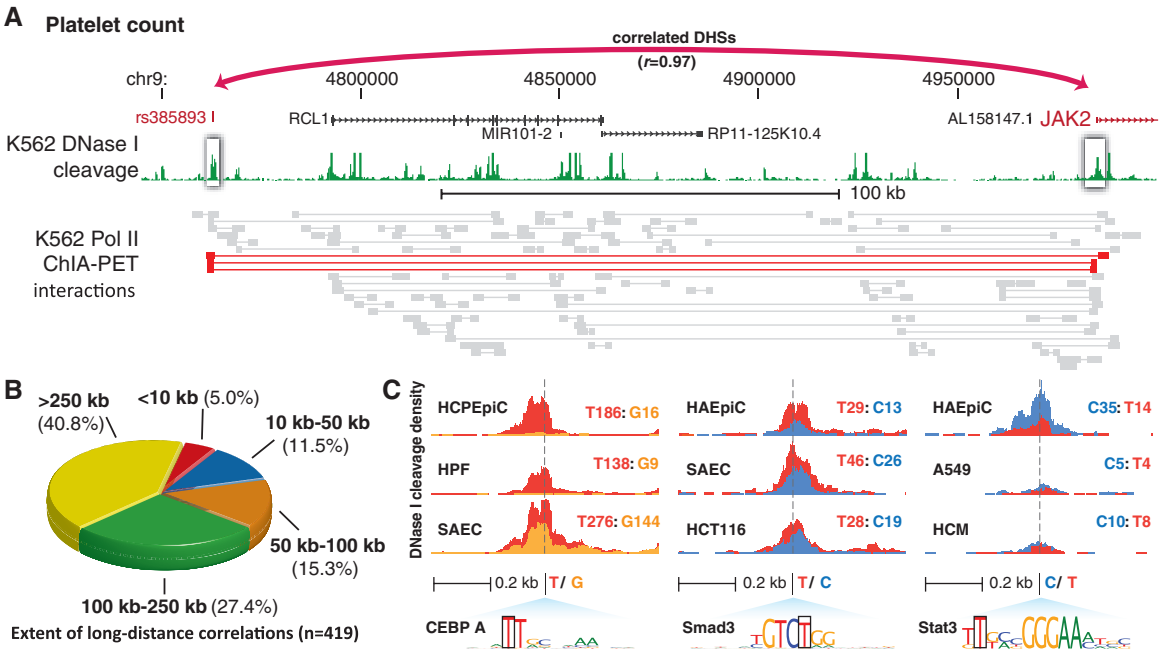
(2874) overlap a transcription factor recognition sequence. We partitioned GWAS variants into 10 disease or trait classes, and then determined the frequency of GWAS variants associated with a particular disease or trait class that localized within sites for transcription factors independently partitioned into the same classes on the basis of gene ontology annotations (fig. S8) (12). This analysis revealed that common variants associated with specific diseases or trait classes were systematically enriched in the recognition sequences of transcription factors governing physiological processes relevant to the same classes.

Table 1. Highest correlated genes of distal DHSs harboring GWAS variants. Examples of distal DHSs-to-promoter connections that highlight candidate genes potentially underlying the association. An asterisk (*) indicates the highest-correlated gene is not the nearest gene. r , Pearson's correlation coefficient; ADHD, attention deficit hyperactivity disorder; NMDA, N-methyl-D-aspartate.

indicates the highest-correlated gene is not the nearest gene. r , Pearson's correlation coefficient; ADHD, attention deficit hyperactivity disorder; NMDA, N-methyl-D-aspartate.

Disease or trait	r	Target gene	Function	Distance (kb)
Amyotrophic lateral sclerosis	1	SYNGAP1*	Axon formation; component of NMDA complex	411
Crohn's disease	1	TRIB1*	NF- κ B regulation	95
Time to first primary tooth	0.99	PRDM1*	Craniofacial development	452
C-reactive protein	0.99	NLRP3	Response to bacterial pathogens	20
Multiple sclerosis	0.98	AHI1*	White matter abnormalities	149
QRS duration	0.96	SCN10A*	Sodium channel involved in cardiac conduction	181
Breast cancer	0.96	TACC2*	Tumor suppressor	411
Schizophrenia/brain imaging	0.95	KIF1A*	Neuron-specific kinesin involved in axonal transport	428
Brain structure	0.94	CXCR6*	Chemokine receptor involved in glial migration	357
Rheumatoid arthritis	0.94	CTSB*	Cysteine proteinase linked to articular erosion	359
Ovarian cancer	0.93	HSPG2*	Ovarian tumor suppressor	268
Multiple sclerosis	0.93	ZP1*	Known autoantigen	153
ADHD	0.93	PDLIM5*	Neuronal calcium signaling	328
Breast cancer	0.88	MAP3K1*	Response to growth factors	158
Amyotrophic lateral sclerosis	0.88	CNTN4	Neuronal cell adhesion	306
Schizophrenia	0.81	FXR1*	Cognitive function	120
Type 1 diabetes	0.75	ACAD10*	Mitochondrial oxidation of fatty acids	343
Lupus	0.74	STAT4	Mediates IL-12 immune response and T _H 1 differentiation	113

Fig. 2. Candidate regulatory roles for GWAS SNPs. (A) GWAS variant associated with platelet count is connected with the JAK2 gene (myeloproliferative disorders) 222 kb away. Below, ChIA-PET tags (36) validate direct chromatin interactions; red tags demonstrate an interaction between this DHS and the JAK2 promoter (red). (B) Proportion of DHSs harboring GWAS variants that can be linked to target promoters at the indicated distance. (C) Examples of allele-specific DNase I sensitivity in cell types derived from heterozygous individuals for GWAS variants that alter transcription factor recognition motifs within DHSs (also see table S9). Each cell type track shows DNase I cleavage density scaled by allelic imbalance at the GWAS variant and colored by variant nucleotide: blue (C), green (A), yellow (G), red (T). Total reads from each allele are also shown.



Functional variants that alter transcription factor recognition sequences frequently affect local chromatin structure. At heterozygous SNPs alter-

ing transcription factor recognition sequences, altered nuclease accessibility of the chromatin template manifests as an imbalance in the frac-

tion of reads obtained from each allele (20, 21). As the concentration of sequence reads and highly overlapping read coverage results in an effective resequencing of DHSs, we could both detect cell types heterozygous for common SNPs and quantify the relative proportions of reads from each allele across all cell types (12). This imbalance is indicative of the functional effect of a particular allele on local chromatin state. We detected 584 heterozygous GWAS SNPs with sufficient sequencing coverage, of which 120 showed significant allelic imbalance in chromatin state [at a false discovery rate (FDR) of 5%]. We identified sites where regulatory variants were associated with allelic chromatin states, with the predicted higher-affinity allele exhibiting higher accessibility (Fig. 2C). In nearly 50% of cases, the magnitude of imbalance was $>2:1$ (fig. S9). The GWAS SNPs were the sole local sequence difference between haplotypes, indicating that disease-associated variants are responsible for modulating local chromatin accessibility. Further, at sites with very high sequencing depth ($>200\times$), 38.7% (53/137) show significant allelic imbalance (FDR $< 5\%$). As sensitivity to detect allelic imbalance is governed by sequencing depth, this suggests that nearly 40% of GWAS variants in similarly sequenced DHSs would be expected to show allelic imbalance.

Disease-associated variants cluster in transcriptional regulatory pathways. Transcriptional control of glucose homeostasis and beta cell genesis and function is mediated by a closely knit transcriptional regulatory pathway defined by specific transcription factors. The Mendelian phenotypes of maturity-onset diabetes of the young (MODY) are caused by separate lesions disrupting the coding sequences of each of these transcription factors (22). We observed clustering of common noncoding variants associated with abnormal glucose homeostasis, insulin and glycohemoglobin levels, and diabetic complications within recognition sites for the same six transcription factors ($P < 0.029$, binomial; 48% enrichment over random SNPs; Fig. 3A). This suggests that noncoding variants that predispose to dysregulation of glucose homeostasis perturb peripheral nodes of the same regulatory network responsible for Mendelian forms of type 2 diabetes.

Using known interacting sets of transcription factors, we identified related disease-associated variants in the recognition sequences of a central target factor and its interacting partners (Fig. 3B and figs. S11 and S12) for factors involved in autoimmune disease, cancer, and neurological development. IRF9 is a transcription factor associated with type I interferon induction (23) and is a critical component of the JAK/STAT signaling pathway. Of 26 transcription factors in the IRF9-centered interaction network, 15 represent transcription factors with recognition sequences in multiple distinct DHSs that contain GWAS variants associated with a wide variety of autoimmune disorders ($P < 1.6 \times 10^{-13}$, binomial; 2.8-fold enrichment versus random SNPs, Fig. 3B) (12).

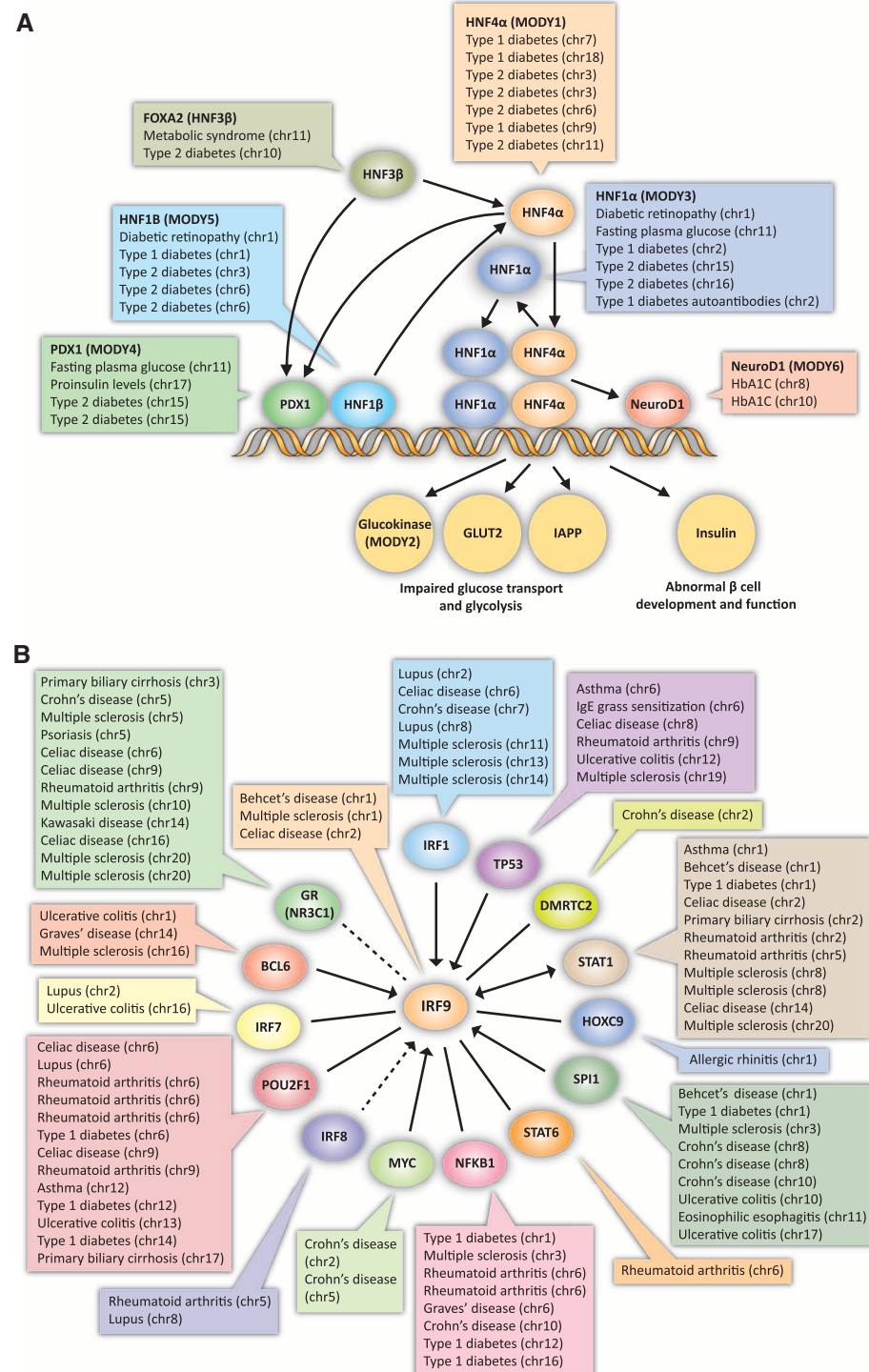


Fig. 3. Common disease-associated variants cluster in regulatory pathways. (A) SNPs in DHSs associated with diabetes (type I and type II), diabetic complications, and glucose homeostasis localize in recognition sites of transcriptional regulators (labeled ellipses) controlling glucose transport, glycolysis, and beta cell function that are structurally disrupted in the Mendelian phenotypes of maturity-onset diabetes of the young (MODY). Chromosome of each SNP associated with the indicated phenotype is listed (see table S2). **(B)** Of SNPs associated with autoimmune disorders that fall within DHSs, 24.4% localize in recognition sequences of TFs that interact with IRF9. Arrows indicate directionality of relationship, dotted lines represent indirect interactions (12). The complete network is shown in fig. S10.

Notably, 24.4% (64/262) of GWAS SNPs within DHSs of immune cells and associated with autoimmune disease alter one or more of the 15 transcription factor motifs from the IRF9-centered network, suggesting that JAK/STAT-mediated type I interferon responses may have important roles in mediating diverse common inflammatory disorders. This example and those in figs. S11 and S12 illustrate that disease-associated variants from the same or related disorders and traits repeatedly localize

within the recognition sequences of transcription factors that form interacting regulatory networks. **Common networks for common diseases.** The observation that GWAS variants associated with multiple distinct diseases within the same broader disease class (e.g., inflammation, cancer) repeatedly localize within the recognition sites of interacting transcription factors suggested that cohorts of such transcription factors might form shared regulatory architectures. To

explore whether noncoding GWAS SNPs from related diseases perturb different recognition sequences of a common set of transcription factors, we tabulated all transcription factors for which at least eight recognition sequences in DHSs were perturbed by GWAS SNPs associated with autoimmune diseases (Fig. 4A). Among the 22 factors identified were canonical immune signaling regulators, such as signal transducer and activators of transcription 1 (STAT1) and STAT3, nuclear

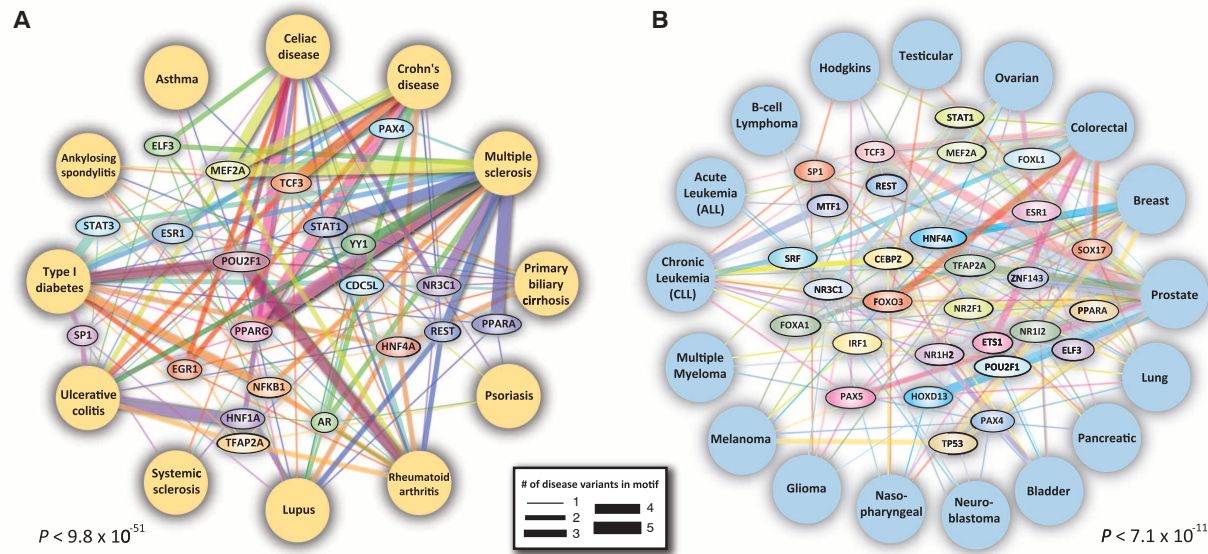


Fig. 4. Common disease networks. GWAS SNPs from related diseases repeatedly perturb recognition sequences of common transcription factors. Shown are factors whose recognition sequences harbor ≥ 8 or ≥ 6 GWAS SNPs in inflammatory or autoimmune diseases (A) and cancer (B), respectively.

Edge thickness represents number of associations between transcription factor and disease in DHSs in relevant tissues. Both networks are significantly enriched for overlap with disease-relevant GWAS SNPs and include many well-studied regulators.

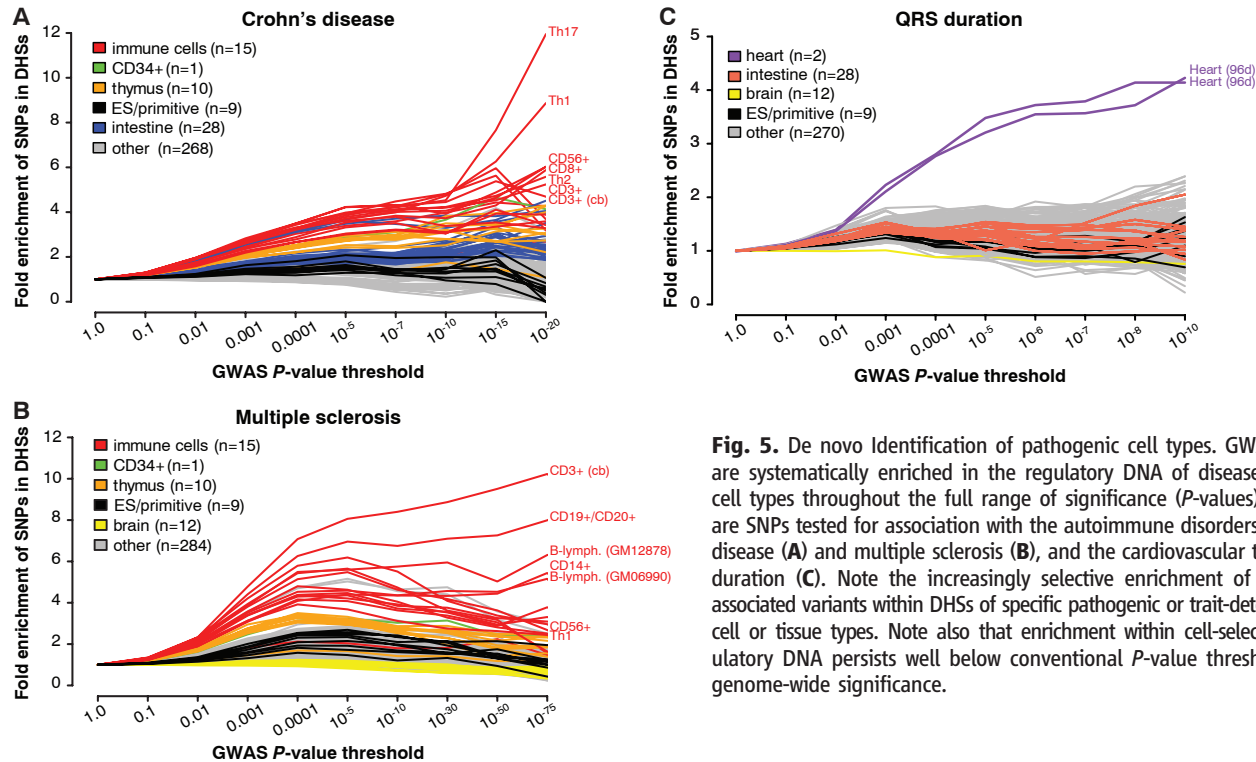


Fig. 5. De novo Identification of pathogenic cell types. GWAS SNPs are systematically enriched in the regulatory DNA of disease-specific cell types throughout the full range of significance (P -values). Shown are SNPs tested for association with the autoimmune disorders Crohn's disease (A) and multiple sclerosis (B), and the cardiovascular trait QRS duration (C). Note the increasingly selective enrichment of disease-associated variants within DHSs of specific pathogenic or trait-determining cell or tissue types. Note also that enrichment within cell-selective regulatory DNA persists well below conventional P -value thresholds for genome-wide significance.

factor κB (NF- κB), and peroxisome proliferator-activated receptor α (PPAR- α) and PPAR- γ . These 22 transcription factors comprise a highly significant [$P < 9.8 \times 10^{-51}$, simulation versus number of factors for random SNPs (12)], shared regulatory architecture that is repeatedly perturbed in a wide range of autoimmune disorders (Fig. 4A).

The same analysis in the context of 17 different malignancies exposed a very different network of transcription factors connecting seemingly disparate cancer types [$P < 7.1 \times 10^{-11}$, simulation (12)] including neoplastic regulatory relationships linking FOXA1 and breast cancer, FOXO3 and colorectal cancer, and TP53 and melanoma, breast, and prostate cancer (Fig. 4B). We also analyzed six neuropsychiatric disorders and identified 23 transcription factors whose recognition sequences were perturbed by at least three disease-associated variants (fig. S13). Collectively, these results support the hypothesis that shared genetic liability may underlie many related categories of disease (24, 25).

De novo identification of pathogenic cell types.

To provide insights into the cellular structure of disease and potentially highlight pathogenic cell types, we explored the selective localization of GWAS SNPs within the regulatory DNA of individual cell types. We further considered the enrichment of all tested variants, not just those with genome-wide significance, and performed serial determination of the cell- and tissue-selective enrichment patterns of progressively more strongly associated variants to expose collective localization within specific lineages or cell types. We used all SNPs tested in GWAS meta-analyses of two common autoimmune disorders, Crohn's disease (26) and multiple sclerosis (MS) (27), and a common continuous physiological trait, cardiac conduction measured by the electrocardiogram QRS duration (28) ($n = 938,703$, 2,465,832, and 2,828,796 SNPs, respectively). For SNPs meeting increasingly significant P -value cutoffs, we compared the proportion of SNPs in DHSs of each cell type to the proportion of all SNPs in DHSs of the same cell type (Fig. 5). For all three studies, we observed enrichment of more weakly associated variants in regulatory DNA. This enrichment suggests that a large number of functional variants of small quantitative effect act through modulation of regulatory DNA. Additionally, it suggests that conditioning association analyses on regulatory DNA may ameliorate the stringent statistical correction for multiple testing required for genome-wide testing of unselected SNPs.

Furthermore, with progressively stringent P -value thresholds, we observed increasingly selective enrichment of disease-associated variants within specific cell types (Fig. 5). In the case of Crohn's disease, the T helper 17 (T_H17) (12.0-fold enriched) and T_H1 (8.87-fold enriched) T cell subtypes have a concentration of the most-significant GWAS variants in their DHSs (Fig. 5A). Although Crohn's pathology has classically been associated with T_H1 cytokine responses, an emerging consensus points to a defining role for IL-17-producing T_H17 cells (29). Notably, this analysis was ac-

complished without any prior knowledge about Crohn's disease pathology.

In the case of MS, sequential cell-selective enrichment analysis highlighted two cell types: CD3⁺ T cells from cord blood and CD19⁺/CD20⁺ B cells (Fig. 5B). Although MS has long been thought to be T cell mediated, a critical role for B cells has only recently been recognized and has major therapeutic implications (30). It is notable that cord blood CD3⁺ cells—essentially a naïve population—garner the most highly selective enrichment, particularly in comparison with total adult CD3⁺ cells or other T cell subsets, suggesting a role for variants influencing immune education. Also of note, DHSs active in brain tissue were moderately depleted (~10%) for MS-associated variants, suggesting that neural gene expression regulatory elements do not play a substantial role in MS pathogenesis, as proposed (31). Analogously, analysis of variants associated with the continuously varying trait of QRS duration revealed similarly specific enrichment within fetal heart DHSs (Fig. 5C). In all three cases, the results were obtained without any prior knowledge of physiological mechanisms. These data suggest a generally applicable approach and highlight the value of extensive maps of regulatory DNA for gaining insights into disease physiology and pathogenesis.

Discussion. Despite a long appreciation of the involvement of regulatory variants in human disease (32–34), difficulty in delineating regulatory DNA regions, particularly in a cell-specific context, has until now prevented comprehensive assessment of the relationship between gene regulation and common phenotypes. Our results indicating widespread and systematic localization of variants associated with a wide spectrum of common diseases and traits in regulatory DNA marked by DHSs have many implications for interpreting diverse genotype-phenotype association studies. The connection of numerous DHSs harboring GWAS SNPs with promoters of distant genes expands the genomic horizon of disease and trait associations and provides a trove of plausible causal genes to explain those associations. The data also unify seemingly unconnected variants associated with related diseases by virtue of their convergent perturbation of common transcription factor networks. Tissue-selective enrichment of phenotype-associated variants raises the possibility of more focused genetic association studies that prioritize the regulatory DNA of a known or hypothesized target tissue type. Further, selective enrichment of many more weakly associated variants within regulatory DNA of pathogenic cell types points to the quantitative contribution of hundreds of variants of small effect size that modulate transcription factor binding characteristics, in contrast to Mendelian variants in transcription factor genes that may perturb entire networks. The results thus highlight a continuous quantitative spectrum of disordered gene regulation between common disease and Mendelian traits and lend a new perspective on the

genetic architecture and heritability of common human diseases and phenotypic traits (35).

References and Notes

1. B. E. Stranger, E. A. Stahl, T. Raj, *Genetics* **187**, 367 (2011).
2. L. A. Hindorf *et al.*, A Catalog of Published Genome-Wide Association Studies; available at www.genome.gov/gwastudies.
3. W. Cookson, L. Liang, G. Abecasis, M. Moffatt, M. Lathrop, *Nat. Rev. Genet.* **10**, 184 (2009).
4. M. M. Pomerantz *et al.*, *Nat. Genet.* **41**, 882 (2009).
5. K. Musunuru *et al.*, *Nature* **466**, 714 (2010).
6. O. Harismendy *et al.*, *Nature* **470**, 264 (2011).
7. D. L. Nicolae *et al.*, *PLoS Genet.* **6**, e1000888 (2010).
8. J. F. Degner *et al.*, *Nature* **482**, 390 (2012).
9. D. S. Gross, W. T. Garrard, *Annu. Rev. Biochem.* **57**, 159 (1988).
10. R. E. Thurman *et al.*, *Nature* **10.1038/nature11232** (2012).
11. B. E. Bernstein *et al.*, *Nat. Biotechnol.* **28**, 1045 (2010).
12. Detailed information on methods and analyses can be found in the supplementary materials available on Science Online.
13. The 1000 Genomes Project Consortium, *Nature* **467**, 1061 (2010).
14. M. E. Symonds, S. P. Sebert, M. A. Hyatt, H. Budge, *Nat. Rev. Endocrinol.* **5**, 604 (2009).
15. S. John *et al.*, *Nat. Genet.* **43**, 264 (2011).
16. D. J. Barker *et al.*, *Lancet* **341**, 938 (1993).
17. G. A. Maston, S. K. Evans, M. R. Green, *Annu. Rev. Genomics Hum. Genet.* **7**, 29 (2006).
18. L. A. Lettice *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7548 (2002).
19. N. Soranzo *et al.*, *Nat. Genet.* **41**, 1182 (2009).
20. J. C. Knight, B. J. Keating, K. A. Rockett, D. P. Kwiatkowski, *Nat. Genet.* **33**, 469 (2003).
21. M. T. Maurano, H. Wang, T. Kutayavin, J. A. Stamatojanopoulos, *PLoS Genet.* **8**, e1002599 (2012).
22. S. S. Fajans, G. I. Bell, K. S. Polonsky, *N. Engl. J. Med.* **345**, 971 (2001).
23. T. Tamura, H. Yanai, D. Savitsky, T. Taniguchi, *Annu. Rev. Immunol.* **26**, 535 (2008).
24. International Schizophrenia Consortium *et al.*, *Nature* **460**, 748 (2009).
25. C. Cotsapas *et al.*, *PLoS Genet.* **7**, e1002254 (2011).
26. A. Franke *et al.*, *Nat. Genet.* **42**, 1118 (2010).
27. N. A. Patsopoulos, *Ann. Neurol.* **70**, 897 (2011).
28. N. Sotoodehnia *et al.*, *Nat. Genet.* **42**, 1068 (2010).
29. S. Brand, *Gut* **58**, 1152 (2009).
30. H.-C. von Büdingen, A. Bar-Or, S. S. Zamvil, *Curr. Opin. Immunol.* **23**, 713 (2011).
31. International Multiple Sclerosis Genetics Consortium (IMSGC) *et al.*, *Nat. Genet.* **42**, 469, author reply 470 (2010).
32. M. C. King, A. C. Wilson, *Science* **188**, 107 (1975).
33. F. S. Collins *et al.*, *Nature* **313**, 325 (1985).
34. M. V. Rockman, G. A. Wray, *Mol. Biol. Evol.* **19**, 1991 (2002).
35. G. Gibson, *Nat. Rev. Genet.* **13**, 135 (2012).
36. G. Li *et al.*, *Cell* **148**, 84 (2012).

Acknowledgments: We thank many colleagues for their insightful comments and critical readings of the manuscript. We also thank many colleagues who provided individual cell samples for DNase I analysis. This work was supported by NIH grants U54HG004592 and U01ES01156 (J.A.S.), P30 DK056465 (S.H.), R01HL088456 (N.S.), and R24HD000836-47 (I.G.). We acknowledge the generous sharing of results from the International MS Genetics, International IBD Genetics, and CHARGE QRS Consortia. DNase I data have been deposited in Gene Expression Omnibus (GEO) under accession nos. GSE29692 and GSE18927 and can be viewed and downloaded at genome.ucsc.edu, www.uwencode.org/data, and www.epigenomebrowser.org.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1222794/DC1
Materials and Methods
Figs. S1 to S13
Tables S1 to S12
References (37–49)

3 April 2012; accepted 13 July 2012
Published online 5 September 2012;
10.1126/science.1222794