Genetic and Epigenetic Fine-Mapping of Causal Variants in Autoimmune Disease

Supplemenatary Methods

Supplementary References

Supplemental Tables

Supplemenatary Methods

Cell Isolation and Culture

Purification and Culture of Human $CD4^+$ T-cell Subsets

Cells were obtained from the peripheral blood of pooled healthy subjects in compliance with institutional review board protocols. Untouched $CD4^+$ T cells were isolated by gradient centrifugation (Ficoll-Hypaque; GE Healthcare) using the RosetteSep Human $CD4^+$ T cell Enrichment kit (Stemcell Technologies). $CD4^+$ T cells were next subjected to anti-CD25 magnetic bead labeling (Miltenyi Biotech), to allow magnetic cell separation (MACS) of $CD25^+$ and $CD25^-$ cells. Subsequently $CD25^+$ cells were stained with fluorescence-labeled monoclonal antibodies to CD4, CD25 and CD127 (BD Pharmingen), and sorted using a FACS ARIA (BD Biosciences) for $CD25^{high}CD127^{low/-}$ T_{reg} cells, which express FOXP3 (Biolegend) as confirmed by intracellular post-sort analysis by FACS **(Fig. S1a)**. Dead cells were excluded by Propidium iodide (BD). An aliquot of $CD25^-$ cells was labeled with fluorescence-labeled monoclonal antibodies to CD4, CD45RA and CD45RO (BD Pharmingen), and sorted on a FACS ARIA to isolate $CD45RO^+CD45RA^-$ memory (T_{mem}) and $CD45RO^-CD45RA^+$ naïve (T_{naive}) $CD4^+$ T-cell populations **(Fig. S1b)**. Dead cells were excluded by Propidium iodide. Highly pure human Th17 cells were isolated with modifications as previously described[1]. In brief, $CD25^-$ cells were stimulated in serum-free X-VIVO15 medium (BioWhittaker) with PMA (50ng ml^{-1}) and ionomycin (250ng ml^{-1}; both from Sigma-Aldrich) for 8 hours and sorted by a combined MACS and FACS cell sorting strategy based on surface expression of IL-17A. Stimulated cells were stained with anti-IL-17A-PE (Miltenyi) and labeled with anti-PE microbeads (Miltenyi) and subsequently pre-enriched over an LS column (Miltenyi). The IL-17A negative fraction was used as control population (Th_{stim}). MACS-enriched Th17 cells were further sorted on a FACS ARIA (BD) for highly pure $IL\text{-}17A^+$ cells (Th17) **(Fig. S1c)**.

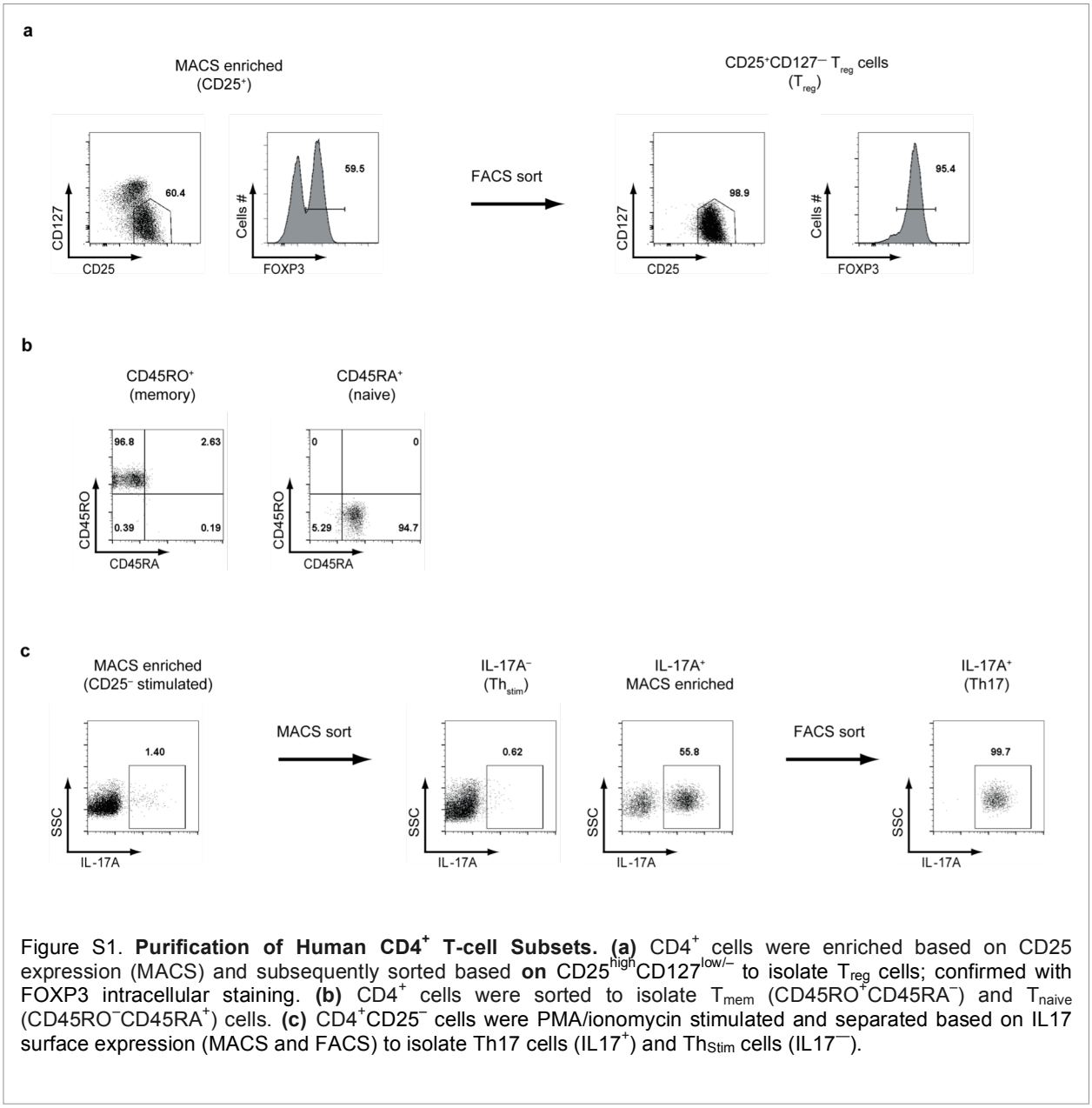

Figure S1. **Purification of Human CD4$^+$ T-cell Subsets. (a)** CD4$^+$ cells were enriched based on CD25 expression (MACS) and subsequently sorted based **on** CD25^{high}CD127$^{low/-}$ to isolate T_{reg} cells; confirmed with FOXP3 intracellular staining. **(b)** CD4$^+$ cells were sorted to isolate T_{mem} (CD45RO$^+$CD45RA$^-$) and T_{naive} (CD45RO$^-$CD45RA$^+$) cells. **(c)** CD4$^+$CD25$^-$ cells were PMA/ionomycin stimulated and separated based on IL17 surface expression (MACS and FACS) to isolate Th17 cells (IL17$^+$) and Th_{Stim} cells (IL17$^-$).

Purification of Human Naïve and Memory CD8$^+$ T cells

Leukocyte-enriched fractions of peripheral blood (byproduct of Trima platelet collection) from anonymous healthy donors were obtained from the Kraft Family Blood Donor Center (DFCI, Boston, MA) in compliance with the institutional Investigational Review Board protocol. For two independent purifications of each cell subset, blood fractions from 7 and 8 donors were pooled. Total T cells were isolated by immunodensity negative selection using the RosetteSep Human T-cell Enrichment Cocktail (STEMCELL Technologies, Vancouver, Canada) and gradient centrifugation on Ficoll-Paque PLUS (GE Healthcare, Pittsburgh, PA), according to the manufacturer's instructions. Subsequently, T cells were

stained at 4°C for 30 min using fluorescently-labeled monoclonal anti-human CD8 (FITC, 2.5 μg/ml, clone RPA-T8, Biolegend, San Diego, CA), CD4 (PE, 1.25 μg/ml, clone RPA-T4, Biolegend), CD45RA (PerCP-Cy5.5, 2.4 μg/ml, clone HI100, eBioscience, San Diego, CA) and CD45RO (APC, 0.6 μg/ml, clone UCHL1, eBioscience) antibodies diluted in staining buffer (PBS supplemented with 2% fetal bovine serum, FBS). 4',6-diamidino-2-phenylindole (DAPI, 2.5 μg/ml, Life Technologies, Grand Island, NY) was also included to stain for dead cells. After washing with staining buffer, naïve ($CD45RA^{+}CD45RO^{-}$) and memory ($CD45RA^{-}CD45RO^{+}$) $CD8^{+}$ or $CD4^{+}$ were isolated using a BD FACSAria 4-way cell sorter (BD Biosciences, San Jose, CA). Cell subsets were identified using a BD FACSDiva Software (BD Biosciences) after gating on lymphocytes (by plotting forward versus side scatters) and excluding aggregated (by plotting forward scatter pulse height versus pulse area), dead ($DAPI^{+}$), and CD8/CD4 double positive cells (**Fig. S2**). Cell purity was 90-94% $CD8^{+}$ or 97-99% $CD4^{+}$, and >99% naïve or memory.

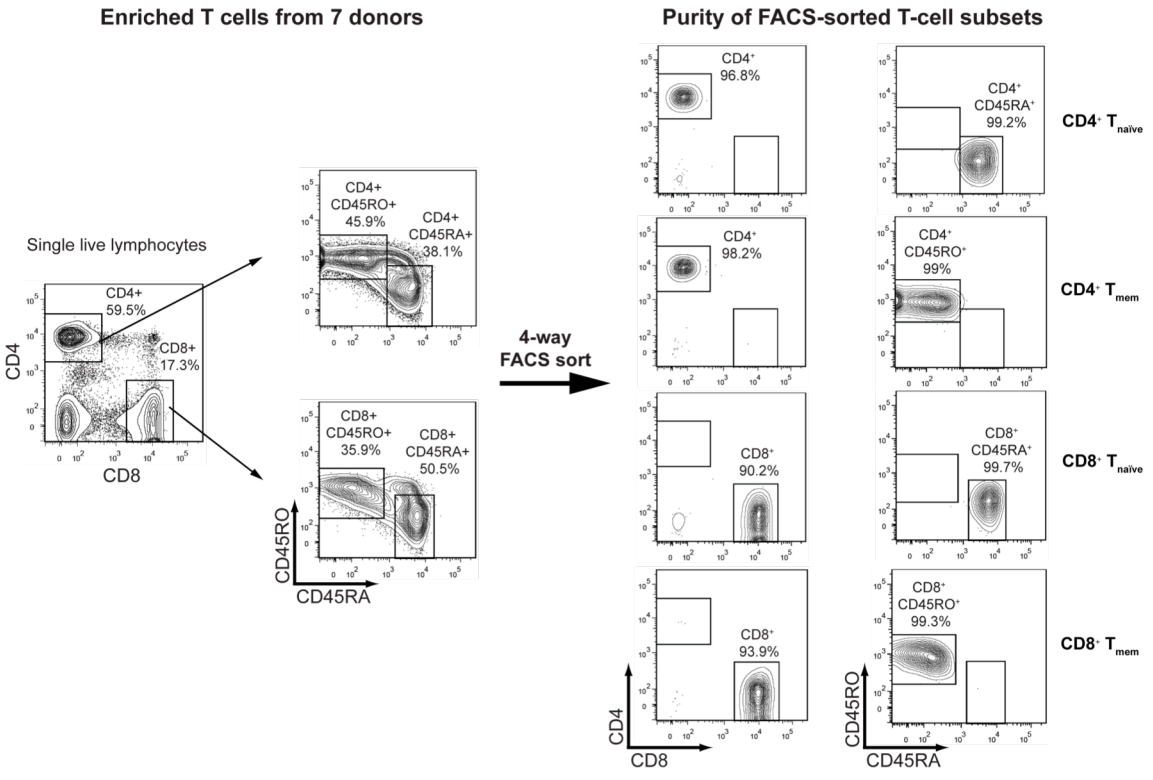

Figure S2. **Purification of Human Naïve and Memory $CD8^{+}$ T cells.** Naïve ($CD45RA^{+}CD45RO^{-}$) and memory ($CD45RA^{-}CD45RO^{+}$) $CD8^{+}$ T cells were isolated using a BD FACSAria 4-way cell sorter. Results are shown from one of two large-scale sorts.

Purification of Human B Centroblasts

For purification of human centroblasts, bulk mononuclear cells were isolated from fresh pediatric tonsillectomy specimens by mechanical disaggregation and Ficoll-Paque centrifugation[2]. MACS enrichment of germinal center cells was performed using anti-CD10-PE-Cy7 (BD Biosciences), and anti-PE microbeads (Miltenyi Biotec). Centroblasts[3] ($CD19^+CD10^+CXCR4^+CD44^-CD3^-$) were purified from the enriched germinal center cells by FACS antibodies for CD19 (APC, clone SJ25C1, BD), CD3 (BV606, clone OKT3, Biolegend), CD10 (PE-Cy7, clone HI10A, BD), CD44 (FITC, clone L178, BD) and CXCR4 (PE, clone 12G5, eBioscience) (**Fig. S3**).

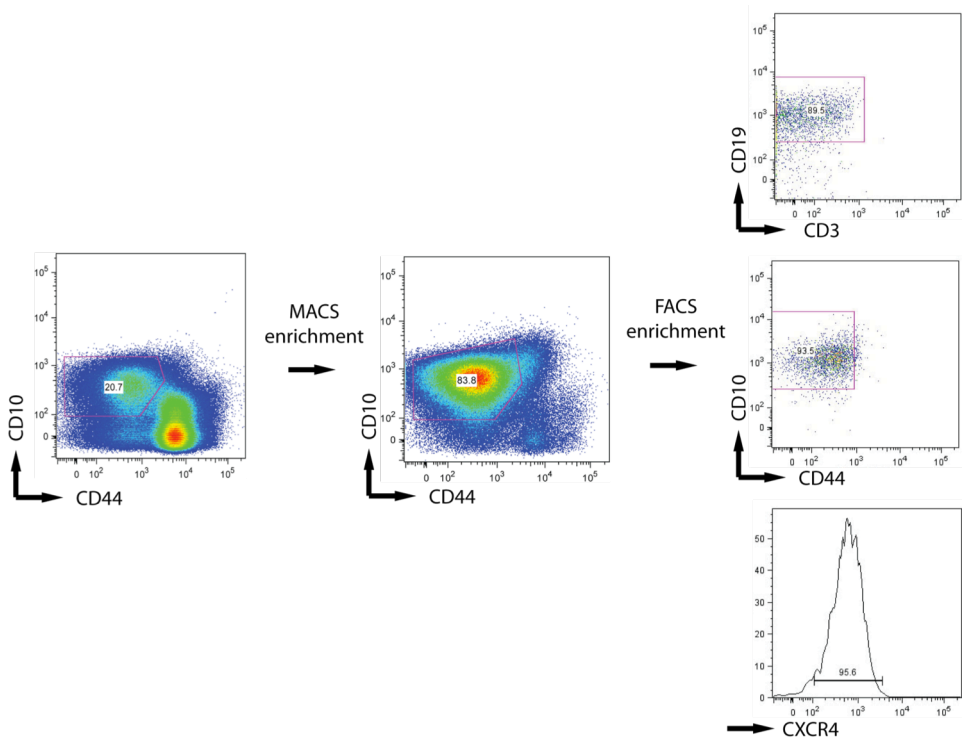

Figure S3. **Purification of Human B Centroblasts.** Mononuclear cells were isolated from pediatric tonsils. Following CD10 enrichment (MACS), B centroblasts ($CD19^+CD10^+CXCR4^+CD44^-CD3^-$) were purified by FACS.

Purification of Adult Human Peripheral Blood B Cells and Monocytes

Human peripheral B cells and monocytes were provided as a service by the S. Heimfeld Laboratory at the Fred Hutchinson Cancer Research Center. The cells were obtained from human leukapheresis product using standard procedures. Briefly, Peripheral B cells ($CD20^+CD19^+$) and monocytes ($CD14^+$) were isolated by immunomagnetic separation using the CliniMACS affinity-based technology (Miltenyi Biotec GmbH, Bergisch Gladbach, Germany) according to the manufacturer's recommendation. Reagents, tubing sets, and buffers are purchased from Miltenyi Biotec.

ChIP-Seq

Following isolation (+/- *ex vivo* stimulation), cells were cross-linked in 1% formaldehyde at room temperature or 37°C for 10 minutes in preparation for ChIP. Chromatin immunoprecipitation and sequencing were performed as previously described[4].

RNA-Seq

RNA was extracted from $CD4^+$ T-cell subsets with Trizol. Briefly, polyadenylated RNA was isolated using Oligo dT beads (Invitrogen) and fragmented to 200–600 base pairs and then ligated to RNA adaptors using T4 RNA Ligase (NEB), preserving strand of origin information as previously described[5,6].

Enhancer Annotation and Clustering

ChIP-seq data were processed as previously described[4]. Briefly, ChIP-seq reads of 36 bp were aligned to the reference genome (hg19) using the Burroughs-Wheeler Alignment tool (BWA)[7]. Reads aligned to the same position and strand were only counted once. Aligned reads were extended by 250 bp to approximate fragment sizes and then a 25-bp resolution chromatin map was derived by counting the number of fragments overlapping each position. H3K27ac and H3K4me1 peaks were identified by scanning the genome for enriched 1kb windows and then merging all enriched windows within 1kb, as described[4]. H3K27ac peaks that do not overlap ±2.5 kb region of an annotated TSS were defined as candidate distal regulatory elements. In order to define the cell-specific H3K27ac peaks, we calculated the mean signal in 5Kb regions centered at distal H3K27ac peaks and sorted the peaks by the ratio of signal in one cell type to all remaining cell types. For each cell type, the top 1000 distal H3K27ac peaks with highest ratio were cataloged as the cell-specific distal H3K27ac peaks. The heatmaps for H3K27ac and H3K4me1 signal were plotted over 10kb regions surrounding all distal cell-specific H3K27ac peaks.

The distal H3K27ac peaks were assigned to their potential target genes if they locate in the gene body or within 100kb regions upstream the TSS. The expression level the target genes were analyzed by RNA-seq data. Paired-end RNA-seq reads were aligned to the RefSeq transcripts using Bowtie2[8]. RNA-seq data for B cells, B centroblast, Macrophages, Th1, Th2 and Th0 are retrieved from NCBI GEO and SRA database (Bnaive: GSE45982; BgerminalCenter: GSE45982[9]; Macrophages: GSE36952[10]; Th0, Th1 and Th2: SRA082670[11]). RNA-seq data for lymphoblastoid (GM12878) was retrieved from ENCODE project[12]. The number of reads per kilobase per million reads (*RPKM*) was calculated for each gene locus. Heatmap of RNA-seq data shows the relative expression level of all potential target genes for each cluster of cell-specific regulatory elements.

Shared Genetic Loci for Common Human Diseases

Publicly available GWAS catalog data were obtained from the NHGRI website, http://www.genome.gov/gwastudies/, current as of July 2013[13,14]. Studies were included based on the criteria that they had at least 6 hits at the genome-wide significant level of $p \leq 5 \times 10^{-8}$. From a set of 21 autoimmune diseases and 18 representative non-autoimmune diseases/traits, we included index SNPs with significance $p \leq 10^{-6}$ for downstream analysis.

In some cases, the same disease had multiple index SNPs mapping to the same locus (defined as within 500kb of each other), due to independently conducted GWAS studies identifying different lead SNPs within the same region. For these loci, only the most significant GWAS index SNP was kept for downstream analysis, resulting in 1170 GWAS index SNPs for 39 diseases/traits. For each pair of diseases/traits, we compared their respective lists of index SNPs to find instances of common genetic loci (defined as the two diseases sharing index SNPs within 500kb of each other). The number of overlapping loci was calculated for each disease pair. To measure the genetic similarity between two diseases/traits, a disease-by-disease correlation matrix was calculated based on the number of overlapping loci for each disease/trait with each of the other diseases, and the results are shown in **Fig. 1A**.

Sources of Immunochip and Non-Immunochip GWAS Data

Summary statistics for published Immunochip studies of celiac disease[15], autoimmune thyroiditis[16], primary biliary cirrhosis[17], and rheumatoid arthritis[18] were downloaded from the Immunobase website, http://www.immunobase.org/. Full genotype data and PCA analysis for the multiple sclerosis Immunochip GWAS study[19] was provided by the International Multiple Sclerosis Genetics Consortium. For ankylosing spondylitis[20], atopic dermatitis[21], primary sclerosing cholangitis[22], juvenile idiopathic arthritis[23], and psoriasis[24], Immunochip studies had been previously been published, but only the lead SNPs from associated Immunochip regions were available. We also included GWAS of autoimmune diseases that had not been studied using Immunochip, including asthma, Kawasaki disease, Behcet's disease, vitiligo, alopecia arreata, systemic lupus erythematosis, systemic sclerosis, type 1 diabetes, Crohn's disease, and ulcerative colitis. For these diseases and the 18 representative non-immune diseases, index SNPs from the GWAS catalog were used[14]. In addition, full genotype data and PCA analysis for the inflammatory bowel disease Immunochip GWAS study were provided by the International Inflammatory Bowel Diseases Genetics Consortium for purposes of calculating the statistical models used in PICS. Because the results for the IBD Immunochip analysis are unpublished, we used the previously published index SNP results for inflammatory bowel disease from the GWAS catalog.

Probabilistic Identification of Causal SNPs (PICS)

Haplotype blocks are genomic regions with low recombination rates within which variants in linkage disequilibrium (LD) are inherited together and share similar association signal, making it challenging to distinguish causal variants from neutral SNPs in LD[25,26]. However, rare recombination events within haplotypes give rise to a small population of recombinant individuals that can provide information on the identity of the causal SNP, provided sufficient genotyping density and adequate sample size.

We developed a fine-mapping algorithm, which we call Probabilistic Identification of Causal SNPs (PICS), that makes use of densely-mapped genotyping data to estimate each SNP's probability of being a causal variant, given the observed pattern of association at the locus. We developed PICS on large cohorts of multiple sclerosis (MS) (14277 cases, 23605 controls[27]) and inflammatory bowel disease (IBD) (34594 cases, 28999 controls; unpublished data) association studies that were genotyped using the Immunochip, a targeted ultra-dense genotyping array with comprehensive coverage of 1000 Genomes Project SNPs within 189 autoimmune disease-associated loci[28].

Analysis of IBD risk associated with SNPS at the *IL23R* locus presents an illustrative example of the LD problem and the potential for PICS to overcome this challenge (**Fig. S4**). The most strongly associated SNP is rs11209026, a loss of function missense variant that changes a conserved arginine to glutamine at amino acid position 381 (R381Q) and decreases downstream signaling through the STAT3 pathway[29,30]. Association with IBD decreases with physical distance along the chromosome (**Fig. S4a**), due to rare recombination events that break up the haplotype and distinguish the causal missense mutation from other tightly linked neutral variants. These rare informative recombination events would be missed by standard genotyping arrays with probes spread thinly across the entire genome.

For neutral SNPs whose association signal is only due to being in LD with a causal SNP, the strength of association, as measured by chi-square (or log-pvalue, since chi-square and log-pvalue are asymptotically linear) scales linearly with their r^2 to the causal SNP. This is because strength of association is linear with r^2 by the formula for the Armitage Trend Test[31]:

$$\chi^2 = (n - 1)r^2$$

where χ^2 is the chi-square association test statistic, n is the sample size, and r^2 is the square of the correlation coefficient.

This linear trend is observed at the *IL23R* locus, consistent with a model where R381Q is the causal variant, and neutral SNPs demonstrate association signal in proportion to their LD to the causal variant (**Fig. S4b**). In addition, SNPs in linkage to R381Q do not perfectly fall on the expected line, due to statistical fluctuations. Independent association studies for the same disease tend to nominate different SNPs within a given locus as their best association, due to statistical fluctuation pushing a different SNP to the forefront in each subsequent study[32-35]. Note that a group of SNPs that are strongly associated to disease but are not in linkage with rs11209026 (R381Q) represent independent association signals at the locus.

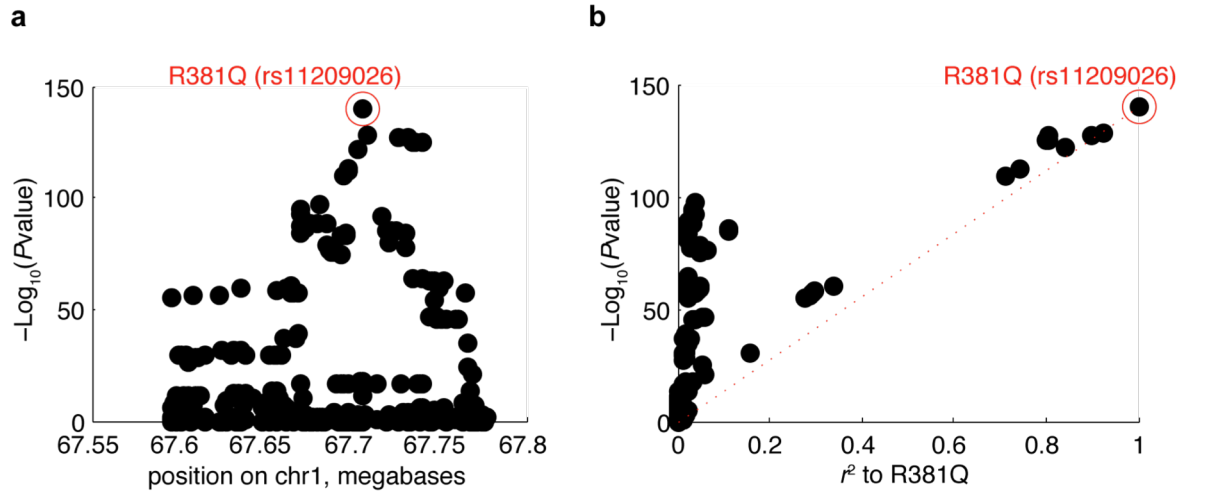

Figure S4. **GWAS Result for IBD Immunochip data at *IL23R* locus. (a)** Each of the 500 SNPs contained in the *IL23R* densely genotyped locus is plotted according to its association signal and position along the chromosome. The R381Q missense variant is circled in red. **(b)** Each of the 500 SNPs in the *IL23R* densely genotyped locus is plotted according to its association signal and r^2 linkage to R381Q.

Although we know from functional studies that R381Q is the likely causal variant, we sought additional statistical evidence to support R381Q as the causal variant, and to refute the null hypothesis that the prominent association of R381Q (compared to other SNPs in the haplotype) is due to chance. We simulated 1000 permutations with an association signal at R381Q, with all other SNPs being neutral, while preserving the LD relationships between SNPs in the locus. An odds ratio of 1.2 was used rather than the ~2-fold odds ratio naturally observed at R381Q, because this was more representative of the modest association signal strengths observed at other GWAS loci. For each round of permutation, we obtained the association signal at all SNPs in the locus. Because only the association signal at R381Q is fixed, the signal at the remaining neutral SNPs in the locus are free to vary due to statistical fluctuations; four typical examples of simulated association results at the R381Q locus are shown (**Fig. S5**), including one example where the causal variant is not the most strongly associated SNP in the locus. From these

1000 iterations, we calculated the standard deviation in the association signal for each of the SNPs in the *IL23R* locus (**Fig. S6**). We show that the distribution of association signals for each SNP approximates a normal distribution, centered at the expected value based on that SNP's r^2 to the causal variant.

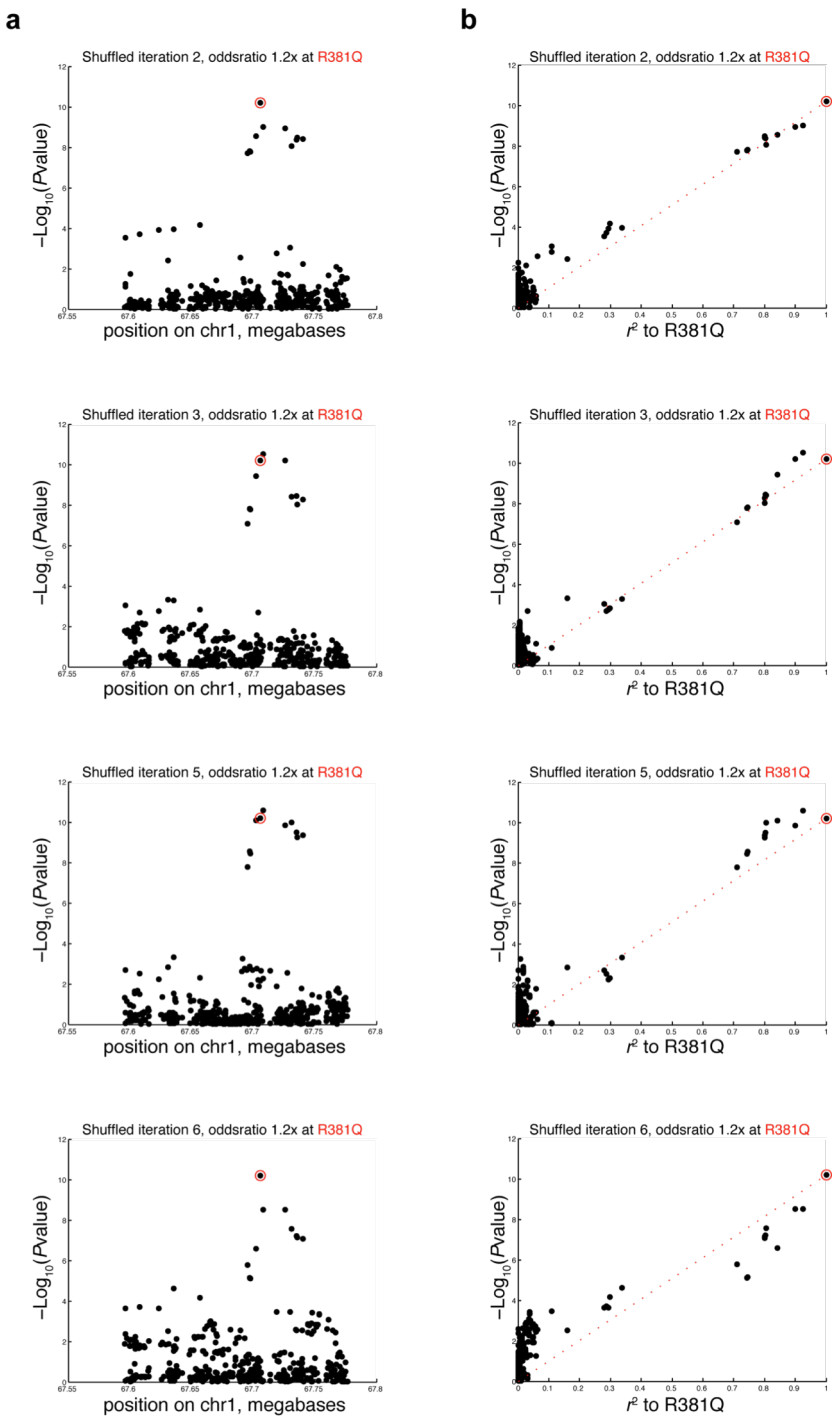

Figure S5. **Simulated permutation analysis of signal at *IL23R* locus**. 1.2-fold odds ratio signal was simulated at the R381Q SNP, with the association signal at R381Q fixed, and all other SNPs neutral and varying only with statistical noise. Four representative results from the simulations are shown, with the panels on the left (**a**) showing the association signal in genomic space, and the panels on the right (**b**) showing the association signal for each SNP in relation to r^2.

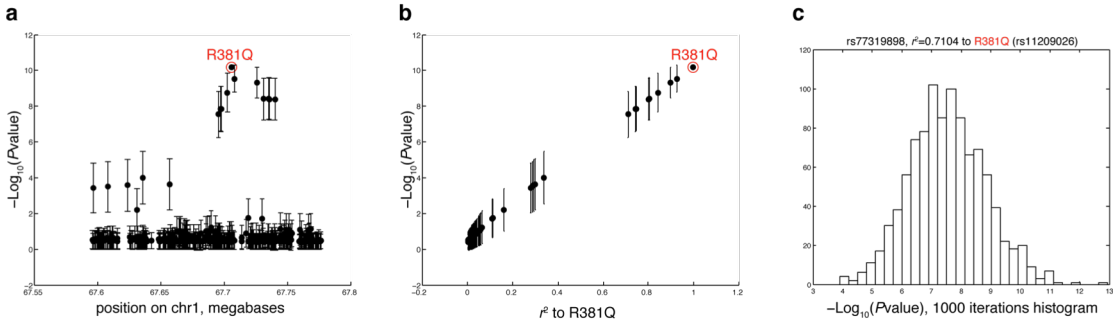

Figure S6. **Standard Deviations in the Association Signals for SNPs in the *IL23R* locus**. **(a)** For each SNP in the *IL23R* locus, the mean association signal and the standard deviation, calculated across 1000 permutations (using a 1.2-fold odds ratio at the R381Q SNP), are shown in genomic space and **(b)** in terms of each SNP's r^2 linkage disequilibrium to the causal R381Q variant. **(c)** As an example, the distribution of association signal values at rs77319898 (r^2 = 0.71 to the causal variant) for 1000 permutations is shown. The distribution of association signal values at each SNP approximated a normal distribution.

These permutations demonstrate that the causal variant need not be the most strongly associated SNP within the locus, due to statistical fluctuations. Rather, given the observed pattern of association at a locus, we are interested in knowing the probability of each SNP within the locus to be the causal variant. We can use Bayes' theorem to infer the probability of each SNP being the causal variant, by using information derived from the permutations. Since the prior probability of each SNP to be the causal variant is equal, the SNP most likely to be causal variant is therefore the SNP whose simulated signal most closely approximates the observed association at the locus. By performing permutations of a simulated association signal at each SNP within the locus, we can estimate the probability that the SNP could lead to the observed association at the locus.

For example, consider a two SNP example where SNP A and SNP B are in LD, and SNP A is the lead SNP in the locus (**Fig. S7**). If we are interested in knowing P(B|A), the probability that SNP B is the causal variant, given that SNP A is the top signal in the locus:

$$P(B^{causal}|A^{lead}) = P(A^{lead}|B^{causal}) \times P(A^{lead}) / P(B^{causal})$$

Where $P(A^{lead}|B^{causal})$ is the probability of SNP A being the top signal in the locus, given that SNP B is the causal variant. $P(A^{lead}|B^{causal})$ is straightforward to calculate by performing permutations with a simulated signal at SNP B, and measuring the number of permutations where SNP A emerges as the top signal in the locus despite SNP B being the actual causal variant. We have assumed that the prior probability of each SNP to be the causal variant or the lead SNP is equal, although this could be adjusted based on external information, such as functional annotation of the SNP to be a coding variant.

Using the formula above, we calculate both $P(B^{causal}|A^{lead})$ and $P(A^{causal}|A^{lead})$, and then normalize both of these probabilities so that $P(B^{causal}|A^{lead}) + P(A^{causal}|A^{lead}) = 1$. In cases where there are more than two SNPs to consider, we similarly normalize the probabilities so that they sum to 1. Probabilities were calculated for all SNPs with $r^2 > 0.5$ to the lead SNP. **Fig. S7** displays the example of R381Q (as SNP A) and rs77319898 (as SNP B), which have $r^2 = 0.71$, to illustrate this analysis.

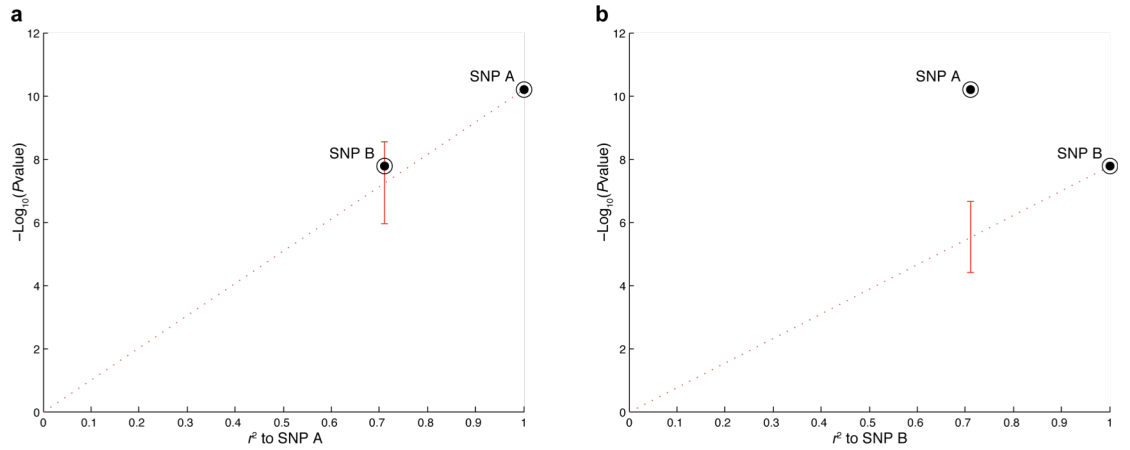

Figure S7. **PICS analysis of a Two SNP Case to Determine Which is More Likely to Explain the Pattern of Association at the Locus**. The SNPs represented here are R381Q (SNP A) and rs77319898 (SNP B), which has an $r^2 = 0.71$ to R381Q. **(a)** The signal at SNP B is well-explained by LD to SNP A, in a model where SNP A is treated as the putative causal variant. The error bars indicate the standard deviation in the association signal expected for SNP B, given that SNP A is postulated to be causal. (**b**) The signal at SNP A is poorly explained by LD to SNP B, in a model where SNP B is treated as the putative causal variant. The error bars indicate the standard deviation in the association signal expected for SNP A, given that SNP B is postulated to be causal.

Because the calculation of thousands of permutations is computationally expensive and requires full genotype data, we sought to generalize the results of the permutation-based method in order to extend it to the analysis of autoimmune diseases for which Immunochip data were not available, or only the identity of the lead index SNPs was reported, such as from the GWAS catalog. We developed a general model, where PICS was able to calculate $P(B^{causal}|A^{lead})$, where B is a SNP within a locus, and A is the lead SNP in the locus, by using LD relationships from the Immunochip where these were available, and from the 1000 Genomes Project otherwise. Since the distribution of association signal at neutral SNPs in the locus approximates a normal distribution, given the lead SNP in the locus, we need to be able to estimate the mean expected association for a neutral SNP in LD with the lead SNP, and the standard deviation for that SNP.

The expected mean association signal for SNPs in the locus scales linearly with r^2 to the lead SNP in the locus. We derived an approximation for the standard deviation for each SNP in the locus based on the

results of empiric testing. We picked 30,000 random SNPs from densely-mapped Immunochip loci, with half coming from the MS Immunochip data, and half coming from the IBD Immunochip data. For each SNP, we simulated 100 permutations with that SNP being the causal variant. SNPs selected had minor allele frequency above 0.05, and the odds ratio used varied from 1.1 fold to 2.0 fold. The number of cases and controls and total sample size were also allowed to randomly vary from 1%-100% of the total number of samples in the original studies. These results indicated that the standard deviation for the association signal at a SNP in LD (with $r^2>0.5$) to a causal variant in the locus was approximately:

$$s = \text{sqrt}(1-r^k) \times \text{sqrt}(\text{indexpval}) / 2$$

$$m = r^2 \times \text{indexpval}$$

where *s* is the standard deviation of the association signal at the SNP, *m* is the expected mean of the association signal at the SNP, *indexpval* is the –log10(p-value) of the lead SNP in the locus, r^2 is the square of the correlation coefficient (a measure of LD) between the SNP and the lead SNP in the locus, and *k* is an empiric constant that can be adjusted to fit the curve; in practice, we found that choosing *k* from a wide range of values between 6 and 8 had little measurable effect on the candidate causal SNPs selected, and we used a value of k=6.4. The results of the 30,000 simulated iterations and the empiric curve fitted using the above equation is shown in **Fig. S8**.

For each SNP in the locus, we used the estimated mean and standard deviation of the association signal at each neutral SNP in LD ($r^2 > 0.5$) to the lead SNP in the locus to calculate the probability of each SNP to be the causal variant relative to the lead SNP. We then normalized the probabilities so that the total of their probabilities summed to 1.

For diseases where summary SNP information was available, but the r^2 relationships between SNPs was unknown, the r^2 relationship was estimated based on the ratio between the association signal at the lead SNP versus the SNP in question. For diseases where only the lead SNP was known, r^2 values were drawn from the LD relationships from the MS Immunochip study if the SNP was from an Immunochip, or from the 1000 Genomes Project otherwise. 1000 Genomes European LD relationships were used for diseases, except for Kawasaki disease, for which 1000 Genomes East Asian LD relationships were used. For diseases which had both GWAS catalog results and Immunochip results, we used Immunochip results whenever possible, and GWAS catalog results in regions outside Immunochip dense-mapping coverage.

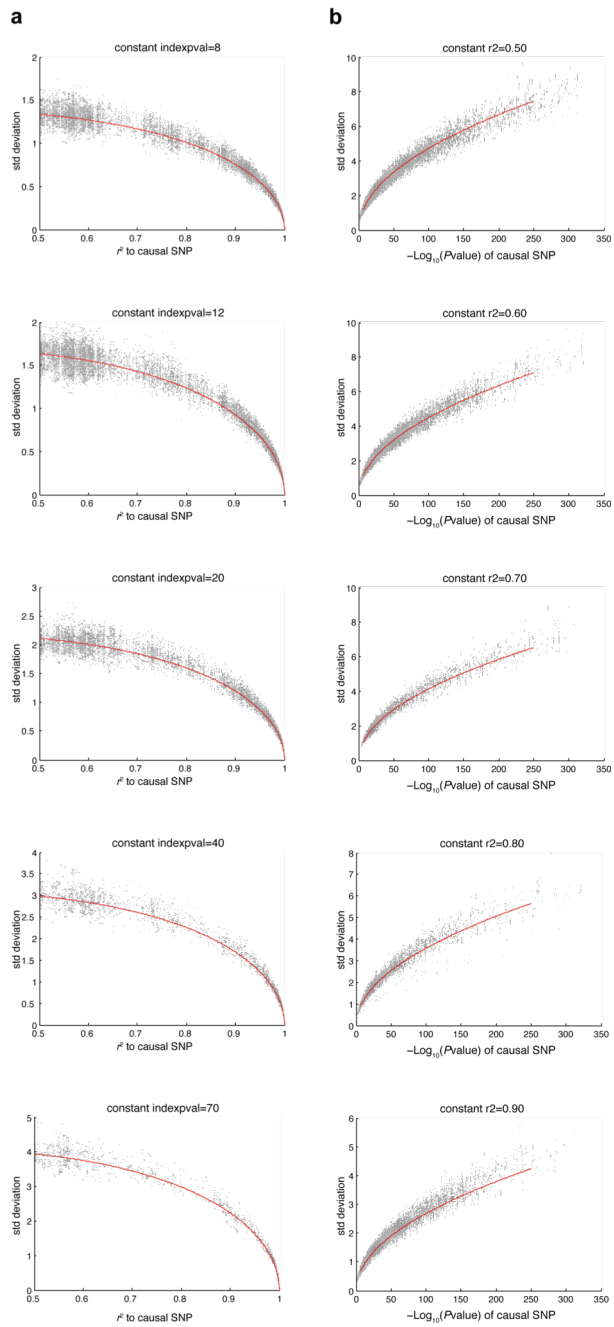

Figure S8. **Simulated Permutations and Empiric Curve Fitting for 30,000 GWAS Signals at Immunochip loci. (a)** For 30,000 simulated GWAS signals in densely mapped Immunochip regions, we considered neutral SNPs within $r^2 > 0.5$ of the simulated causal variant, and plotted the relationship between standard deviation at these neutral SNPs, their r^2 to the lead SNP, given a constant association signal strength for the causal SNP. The red line indicates the expected values derived from the empiric equation for the standard deviation of the association signal at neutral SNPs in LD with the lead SNP in the locus. **(b)** Analogous to **a**, we plotted the relationship between standard deviation at neutral SNPs within $r^2 > 0.5$ of the simulated causal variant and the signal strength at the lead SNP, given a constant r^2 to the lead SNP.

Multiple Independent Association Signals

For the MS data, we were able to use full genotyping information to distinguish multiple independent signals. We used stepwise regression to condition away SNPs one at a time until no associations remain at

the p<10-6 level, which is an effective method for separating independent signals, when LD between the independent causal variants is low. We then treated each independent signal separately for the purpose of using PICS to derive the likely causal variants.

Missing Immunochip Data

For the minority of SNPs that were missing from the Immunochip, we used 1000 Genomes SNPs LD relations to the index SNP to estimate their probability of being the causal SNP. For the diseases with only Immunochip summary statistic data, we could not be certain of the LD relationships, and therefore we estimated the LD to the index SNP from the difference between the association at the lead SNP and the SNP in question, since these follow a linear relationship. For the diseases that only had Immunochip index SNP data, we used Immunochip LD relationships where available from the MS data, and 1000 Genomes SNPs LD relations to the index SNP where these were not available.

Distance between GWAS Catalog SNPs and Lead SNPs

For Immunochip regions that were previously studied by non-Immunochip studies, we examined the performance of prior non-fine-mapped studies at correctly determining the lead SNP. GWAS catalog SNPs within 200kb of Immunochip regions were considered, and the LD and genomic distance between the catalog SNP and any Immunochip lead SNPs for that disease in the Immunochip region were measured and reported in the histograms in **Figs. 1D** and **1E**. PICS was also used to calculate the probability of GWAS catalog SNPs; the probability was 5.5% on average.

Number of Candidate Causal SNPs per GWAS Signal

For each GWAS signal, we obtained a set of candidate causal SNPs, each with a probability of being the causal variant. For each signal, we asked what was the minimum number of candidate causal SNPs required to to cover at least 75% of the probability .

Distribution of GWAS Signals in Functional Genomic Elements: Signal to Background

For downstream analyses, we considered the set of 4905 candidate causal SNPs which had mean probability of greater than 10% of being the candidate causal SNP (the cutoff was probability > 0.0275). We performed 1000 iterations, picking 4905 mean-allele-frequency-matched random SNPs from the same loci (from genomic regions within 50kb of the candidate causal SNPs and excluding the actual causal SNPs). It was necessary to match for mean-allele-frequency because lower MAF SNPs are far more likely to be coding variants. Furthermore, it was necessary to match for locus, because GWAS SNPs are greatly enriched at gene bodies, and using a background of random 1000 genome SNPs for comparison results in

massive nonspecific enrichment of all functional elements. Because we are comparing the candidate causal SNPs to a background set of control SNPs from the same regions, the observed enrichment at functional elements strongly argues that PICS is identifying the the correct causal variants within the locus. For each functional category (missense, nonsense, and frameshift were merged), we calculated the number of actual candidate causal SNPs above mean background (mean of 1000 random iterations), divided by the total number of GWAS signals represented (635), and used these results to populate the pie chart indicating the approximate percentage of GWAS signals that can be attributed to each assessed functional category.

Analysis of *ex vivo* Stimulation-dependent Enhancers

We compared the pattern of H3K27ac at each enhancer in PMA/Ionomycin stimulated T cells vs unstimulated T cells, and in the anti-CD3/CD28 stimulated T cells vs the unstimulated T cells. Because enhancer regions varied greatly in length, enhancers 2kb or longer were segmented and treated as multiple enhancers 1kb in size. We examined the differences in genomic DNA underlying the 15% of enhancers that showed the greatest increase in H3K27ac vs the 15% enhancers that showed the greatest decrease in H3K27ac, using the motif finding program HOMER (http://homer.salk.edu/homer/)[36]. AP1 was by far the most strongly enriched motif in enhancers that gained H3K27ac in both stimulation conditions (**Fig. 2D**). Additional motifs that were enriched in the stimulation-dependent enhancers included NFAT, NFKB, STAT.

Enhancer Signal-to-noise Analysis

For the remainder of enhancer analysis, we focused on 14 immune cell types (8 $CD4^+$ T-cell subsets, 2 $CD8^+$ T-cell subsets, $CD14^+$ monocytes, and 3 B-cell subsets) and 19 representative non-immune cell/tissue types from the Roadmap Epigenome project. Enhancers were broken up into 1kb segments and immune specific enhancers were identified based on the following criteria: (1) # of normalized mean H3K27ac ChIP-seq reads/base > 4, and (2) mean H3K27ac in the top 15th percentile when comparing immune cells to non-immune cells/tissues. We measured the percentage of PICS SNPs (with different probability cutoffs) that either map to an immune enhancer or cause an amino acid coding change. The results are plotted in **Fig. 3A**.

We next considered the 4300 candidate causal SNPs that were not associated with protein coding changes, and compared them against 1000 iterations of frequency and locus matched controls (again, picked from genomic regions within 50kb of the candidate causal SNPs and excluding the actual candidate causal SNPs) (see discussion of signal to background calculations above). Enhancers were

enriched approximately 2:1 above background. We also measured the signal to backround ratio for GWAS signals that had been attributed to coding variants; these produced a much lower signal to background ratio for immune enhancers, as would be anticipated by the fact that most of these are acting on coding regions rather than enhancers (**Fig. S9**). The mean signal above background was shown in a pie chart.

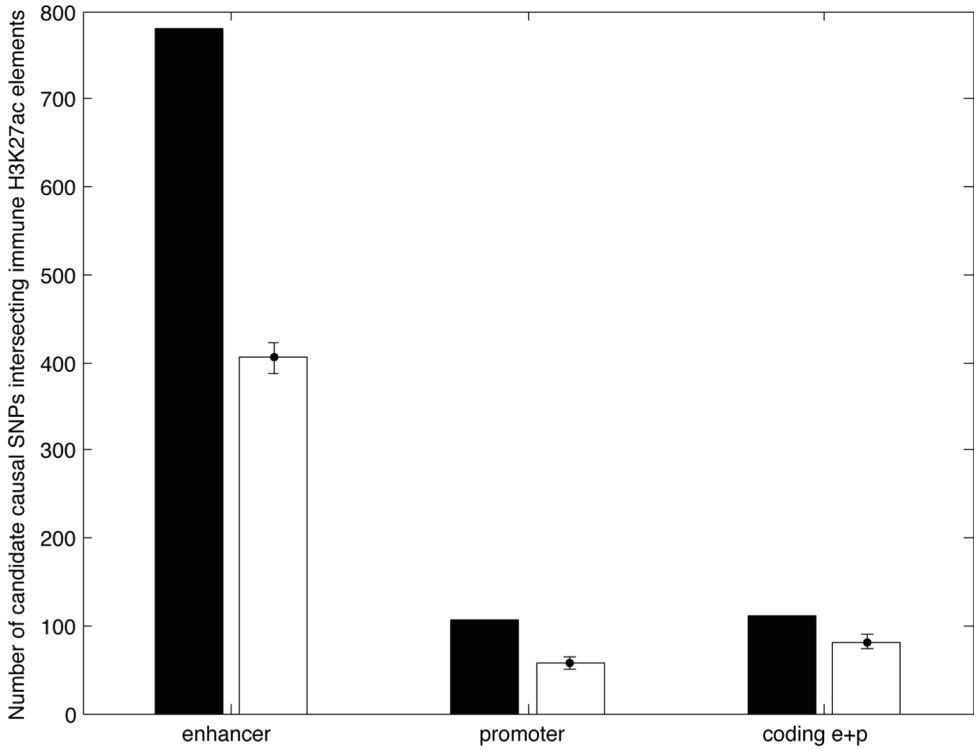

Figure S9. **Localization of PICS Autoimmunity SNPs in Immune Enhancers**. Enrichment of noncoding autoimmune disease candidate causal SNPs within immune enhancers and promoters (black bars) compared to background expectation (white bars). The background expectation represents the overlap between frequency-matched control SNPs drawn from the same locus (within 50kb of the candidate causal SNPs) with immune enhancers and promoters. Candidate causal SNPs whose signal could already be explained by amino acid coding changes (paired bar graphs on the right) showed a much smaller degree of enrichment in immune enhancers and promoters compared to background expectation.

Comparison to Other Methods for Determining Candidate Causal Variants

We next compared the efficacy of PICS versus prior methods that had used cutoffs of r^2=1.0 and r^2>0.8 to determine likely causal SNPs. Because prior studies had not made use of fine-mapping, we used only the GWAS catalog results for this comparison, and applied PICS, and the two r^2-cutoff criteria. In practice these were much more stringent that prior analyses, because we limited the GWAS catalog studies to those that produced 6 or more genome-wide significant hits, thereby pruning a lot of the underpowered studies, and also required a significance of $p<10^{-6}$ for index SNPs, and merged index SNPs at the same locus to use the strongest and most accurate lead SNP. We found that PICS autoimmunity SNPs were

much more likely to map to immune enhancers than SNPS identified by the other methodologies. In addition, when the PICS SNPs which overlapped the r^2>0.8 and r^2>1.0 sets were removed, the remaining SNPs did not show any enrichment above background. In contrast, the candidate causal SNPS identified by PICS, but missed by both of the other methodologies, were significantly enriched for immune enhancers. Background was calculated based on random SNPs drawn from the same loci (within 50kb, frequency-matched controls) as the candidate causal SNPs.

Tissue-specificity of Diseases

GWAS hits for each disease were compared against the enhancers most specific to each cell type (top 15th percentile of H3K27ac and at least > 1 normalized mean H3K27ac ChIP-seq reads/base). We show a heatmap indicating the p-value for the enrichment of PICS candidate causal SNPs for each disease in each cell type (**Fig. 3E**). In contrast to other analyses, enrichment was measured against a background of all common 1000 Genomes SNPs; this is because for genomic regions, generally all enhancers within the region tend to have similar patterns of activity. Therefore using as background SNPs that are only drawn within 50kb of the candidate causal SNPs results in over-controlling, when the purpose of the analysis is to bring out the tissue-specificity, as the specificity for PICS candidate causal SNPs for enhancers within the locus was already demonstrated. We also mapped the expression patterns of genes with PICS candidate causal coding SNPs associated with Crohn's disease, MS and rheumatoid arthritis (**Fig. S10**).

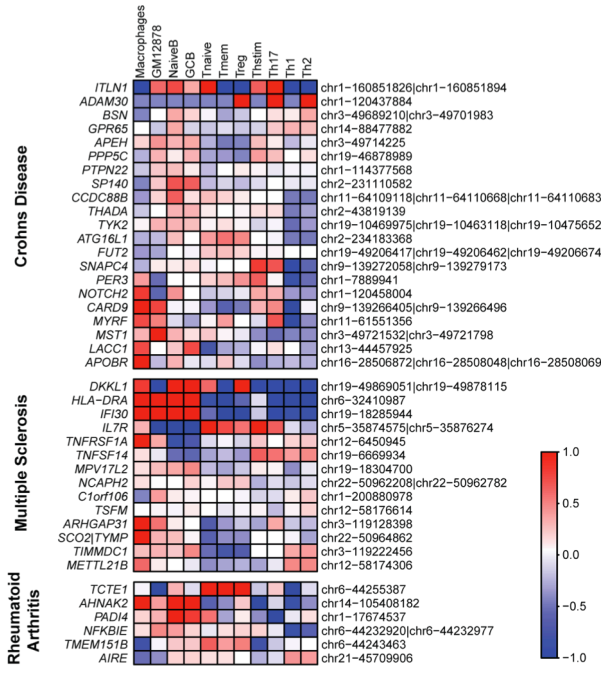

Figure S10. **Expression Pattern of Genes with PICS Autoimmunity Coding SNPs**. Heatmap shows the relative expression level of genes with coding SNPs associated with crohn's disease, MS and rheumatoid arthritis.

Super-enhancer Enrichment

The full set of loci called super-enhancers[37] in $CD4^+$ T-cell subsets (T_{naive}, T_{mem}, Th17, Th_{Stim}) were merged and identified as $CD4^+$ T-cell super-enhancer regions. These regions often contain clusters of discrete enhancers marked with H3K27ac, separated by non-acetylated regions. We assessed if PICS candidate causal SNPs mapping to super-enhancers were more likely to occur in H3K27ac-marked enhancer regions than in intervening regions. Within $CD4^+$ T-cell super-enhancer regions, we compared overlap of PICS candidate causal SNPs with $CD4^+$ T-cell H3K27ac regions, compared to background SNPs drawn from these regions (1000 iterations, frequency-matched). $CD4^+$ T-cell enhancers were called based on being in at least the top 15% percentile in mean H3K27ac among T cells compared to the other 25 cell types. In addition, we compared the overlap of PICS vs background SNPs for stimulated and unstimulated $CD4^+$ T-cell enhancers. Stimulated $CD4^+$ T-cell enhancers were those which had mean increase of at least 25% in H3K27ac in the (average of) Th17, Th_{Stim}, Th0, Th1, Th2 cells, compared the T_{naive}, T_{mem}, T_{reg}; the unstimulated $CD4^+$ T-cell enhancers were the remainder of the $CD4^+$ T-cell enhancer set.

Noncoding RNA Analysis

We next examined the set of disease-associated enhancers, i.e. immune enhancers containing PICS autoimmunity SNPs, and their association with noncoding RNAs. Noncoding RNA transcripts were called based on a RNAseq read density of 0.5 genome-normalized reads per base pair over a window size of at least 2kb, and RNA transcripts overlapping annotated exons or gene bodies of protein coding genes were excluded. We found that enhancers containing autoimmunity candidate causal SNPs were enriched for noncoding transcript production, the majority of which appear to be unspliced enhancer-associated RNAs. Candidate causal SNPs were enriched 1.6-fold within T-cell enhancers that transcribed noncoding RNAs, compared to all T-cell enhancers in general ($p<0.01$). The observation that these disease-associated SNPs are affecting transcribed noncoding sequences suggests another possible mechanism through which they could influence disease risk.

H3K27ac and DNAse Profiles

We measured H3K27ac profiles and DNAse hypersensitivity profiles in a 12kb window centered around candidate causal SNPs, taking the average signal for the 14 immune cell types for which H3K27ac was available, and immune cell types for which DNAse was available. Average normalized reads for H3K27ac and DNAse centered at PICS SNPs are displayed in **Fig. 5A**.

Transcription Factor ChIP-Seq Binding Site Analysis

We assessed the enrichment of PICS autoimmunity SNPs at transcription factor (TF) binding sites. We compared the enrichment of PICS candidate causal SNPs at TF binding sites identified by ENCODE ChIP-Seq[38], relative to random SNPs drawn from the same loci (50kb window around the candidate causal SNPs, frequency matched). We show the results for the 18 TFs whose binding sites are most significantly enriched with PICS autoimmunity SNPs, as well the results for 26 additional immune-relevant TFs. The bar graphs (above) show the overall enrichment for PICS SNPs across all 21 autoimmune diseases in the experimentally obtained ChIP-seq binding sites for each TF; the heatmap (below) shows the enrichment of SNPs associated with each individual disease (**Fig. 5B**).

Motif Creation / Disruption Analysis

We downloaded motifs from Selex[39] and Xie et al.[40]; these motifs were consensus motifs using degenerate nucleotide codes. For this analysis, we used the 853 highest probability non-coding candidate causal SNPs (mean probability = 0.30, cutoff >0.1187), representing 403 different GWAS signals. For each candidate causal SNP, we examined whether it created or disrupted a known motif from Selex or Xie et al.; for comparison we ran 1000 iterations using frequency-matched random SNPs drawn from the same locus (within 50kb of the candidate causal SNPs). We found several known motifs (**Table S1**) to be significantly enriched, including AP1, ETS, NFKB, SOX, PITX, as well as several unknown Xie motifs (**Table S2**). Subtracting the number of motifs found to be disrupted against that expected by background, and dividing by the total number of GWAS signals, we estimate that approximately 11% of noncoding GWAS hits can be attributed to direct disruptions of transcription factor binding motifs.

Neighbouring Motif Analysis

We compared the sequence within 100nt of candidate causal PICS SNPs (cutoff >0.1187) against random flanking sequence (10kb away on either side from the causal SNPs) and looked for enriched motifs using HOMER (http://homer.salk.edu/homer/)[36]. We found strong enrichments for NFKB, RUNX, AP1, ELF1, and PU1. Interestingly, there was a palindromic motif TGGCWNNNWGCCA ($p<10^{-4}$) that was significant both in this method and in the motif disruption analysis, suggesting the role of a yet uncharacterized immune transcription factor motif in autoimmunity.

Determination of phospho-p65 NFκB Activation

PBMCs were prepared from healthy control and age-matched RRMS patients from the Yale MS clinic as described in compliance with institutional review board protocols. Cells were fixed with BD fixation buffer (BD Biosciences) and permeabilized with ice-cold Perm Buffer III (BD Biosciences). Prior to

staining, the cells were washed and Fc receptor blocked (eBioscience). Total PBMCs were stained for CD4 PE (Clone RPA-T4), CD45RA AF700 (clone HI100) , CD45RO PE-Cy7 (clone UCHL1), and pS529 p65 NFκB AF647 (clone K10-895.12.50). Cells were run on a BD Fortessa (BD biosciences).

Expression Quantitative Trait Loci (eQTL) Analysis

We used PICS to determine the candidate causal SNPs from both GWAS studies and the eQTL analyses; for the eQTL analysis, 1000 Genomes summary statistic data was available for all *cis*-eQTLs for all genes. We required a gene to have a *cis*-eQTL with at least a p-value of 10^{-6} for this analysis, giving us 4136 genes. For each gene we applied PICS. We called an overlap between a GWAS hit and an eQTL hit if there was a GWAS candidate causal SNP (again, defined as average probability >10%) in the eQTL candidate causal SNP set with average probability >0.01%. We found that 74/636 (11.6%) of GWAS hits were also eQTL hits. In addition, 15/81 coding GWAS hits (18.5%) showed eQTL effects, suggesting that they may actually operate at the transcriptional level, in addition to any coding effects they may have.

Candidate causal eQTL SNPs were compared against frequency-matched background SNPs drawn from the same loci (within 50kb) in 1000 iterations, and the comparisons are shown in signal-to-background bar graphs for both coding/transcript-related functional elements and for enhancers and promoters (**Fig. S11**). The signal above mean background was calculated for each functional category, and these results were compared against the results for GWAS hits in the two pie charts shown in **Fig. 6A**.

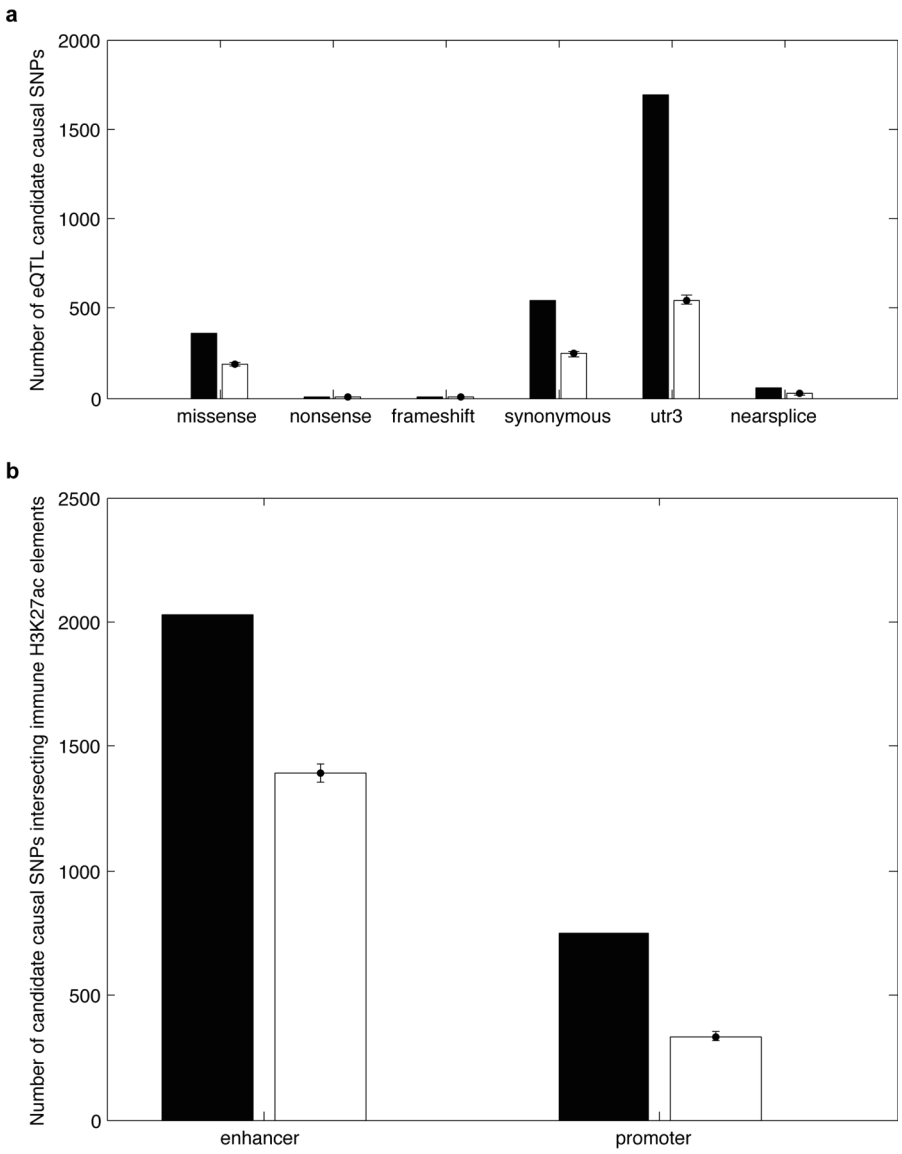

Figure S11**. Enrichment of Candidate Causal eQTL SNPs in Functional Elements: Signal to Background**. (**a**) Candidate causal SNPs for 4136 eQTL genes in peripheral blood were determined by PICS, and their overlap with different types of functional transcript annotations is shown (black bar). The background expectation (white bar) represents the overlap observed for frequency-matched control SNPs drawn from the same loci (within 50kb of the candidate causal SNPs). (**b**) Overlap of candidate causal eQTL SNPs with noncoding immune enhancers and promoters (black bar), versus background expectation (white bar).

Supplementary References

1 Brucklacher-Waldert, V. *et al.* Phenotypical characterization of human Th17 cells unambiguously identified by surface IL-17A expression. *Journal of immunology* **183**, 5494-5501, doi:10.4049/jimmunol.0901000 (2009).

2 Johnston, A., Sigurdardottir, S. L. & Ryon, J. J. Isolation of mononuclear cells from tonsillar tissue. *Current protocols in immunology / edited by John E. Coligan ... [et al.]* **Chapter 7**, Unit 7 8, doi:10.1002/0471142735.im0708s86 (2009).

3 Caron, G., Le Gallou, S., Lamy, T., Tarte, K. & Fest, T. CXCR4 expression functionally discriminates centroblasts versus centrocytes within human germinal center B cells. *Journal of immunology* **182**, 7595-7602, doi:10.4049/jimmunol.0804272 (2009).

4 Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642-654, doi:10.1016/j.cell.2012.12.033 (2013).

5 Gifford, C. A. *et al.* Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149-1163, doi:10.1016/j.cell.2013.04.037 (2013).

6 Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341**, 1237973, doi:10.1126/science.1237973 (2013).

7 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

8 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

9 Beguelin, W. *et al.* EZH2 is required for germinal center formation and somatic EZH2 mutations promote lymphoid transformation. *Cancer cell* **23**, 677-692, doi:10.1016/j.ccr.2013.04.011 (2013).

10 Beyer, M. *et al.* High-resolution transcriptome of human macrophages. *PloS one* **7**, e45466, doi:10.1371/journal.pone.0045466 (2012).

11 Hawkins, R. D. *et al.* Global chromatin state analysis reveals lineage-specific enhancers during the initiation of human T helper 1 and T helper 2 cell polarization. *Immunity* **38**, 1271-1284, doi:10.1016/j.immuni.2013.05.011 (2013).

12 ENCODE Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:nature11247 [pii]
10.1038/nature11247.

13 Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* **42**, D1001-1006, doi:10.1093/nar/gkt1229 (2014).

14 Hindorff LA, M. J. E. B. I., Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. *Available at:* http://www.genome.gov/gwastudies. (Accessed July, 2013).

15 Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature genetics* **43**, 1193-1201, doi:10.1038/ng.998 (2011).

16 Cooper, J. D. *et al.* Seven newly identified loci for autoimmune thyroid disease. *Human molecular genetics* **21**, 5202-5208, doi:10.1093/hmg/dds357 (2012).

17 Liu, J. Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nature genetics* **44**, 1137-1141, doi:10.1038/ng.2395 (2012).

18 Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature genetics* **44**, 1336-1340, doi:10.1038/ng.2462 (2012).

19 Beecham, A. H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature genetics* **45**, 1353-1360, doi:10.1038/ng.2770 (2013).

20 Cortes, A. *et al.* Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nature genetics* **45**, 730-738, doi:10.1038/ng.2667 (2013).

21 Ellinghaus, D. *et al.* High-density genotyping study identifies four new susceptibility loci for atopic dermatitis. *Nature genetics* **45**, 808-812, doi:10.1038/ng.2642 (2013).
22 Liu, J. Z. *et al.* Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nature genetics* **45**, 670-675, doi:10.1038/ng.2616 (2013).
23 Hinks, A. *et al.* Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nature genetics* **45**, 664-669, doi:10.1038/ng.2614 (2013).
24 Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature genetics* **44**, 1341-1348, doi:10.1038/ng.2467 (2012).
25 International HapMap, C. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861, doi:10.1038/nature06258 (2007).
26 International HapMap, C. A haplotype map of the human genome. *Nature* **437**, 1299-1320, doi:10.1038/nature04226 (2005).
27 International Multiple Sclerosis Genetics, C. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature genetics* **45**, 1353-1360, doi:10.1038/ng.2770 (2013).
28 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:nature11632 [pii]
10.1038/nature11632.
29 Pidasheva, S. *et al.* Functional studies on the IBD susceptibility gene IL23R implicate reduced receptor function in the protective genetic variant R381Q. *PloS one* **6**, e25038, doi:10.1371/journal.pone.0025038 (2011).
30 Duerr, R. H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461-1463, doi:10.1126/science.1135245 (2006).
31 Armitage, P. Tests for Linear Trends in Proportions and Frequencies. *Biometrics* **11**, 375-386 (1955).
32 Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics* **42**, 1118-1125, doi:10.1038/ng.717 (2010).
33 Consortium, U. I. G. *et al.* Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nature genetics* **41**, 1330-1334, doi:10.1038/ng.483 (2009).
34 Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature genetics* **40**, 955-962, doi:10.1038/ng.175 (2008).
35 Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature genetics* **43**, 246-252, doi:10.1038/ng.764 (2011).
36 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).
37 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).
38 Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100, doi:10.1038/nature11245 (2012).
39 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327-339, doi:10.1016/j.cell.2012.12.009 (2013).
40 Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338-345, doi:10.1038/nature03441 (2005).

Table S1: Known motifs created or disrupted by candidate causal SNPs

Motif	Observed	Expected	Pvalue <	Annotation
RRACAATG	8	1.6	10^{-3}	SOX
CAGGAARY	5	.82	.01	ETS/ELF1
TGANTCA	8	2.63	.03	AP-1
CCACTTRA	2	.12	.05	NKX2-3
GCTKASTCA	2	.12	.02	MAFK
TTAATCC	2	.24	.05	PITX1
GGGAWWTCC	2	.28	.05	NFKB

Table S2: Unknown motifs created or disrupted by candidate causal SNPs

Motif	Observed	Expected	Pvalue <	Annotation
KMCATNNWGGA	7	.45	10^{-5}	XIE116
TGGNNNNNNKCCAR	4	.65	.01	XIE27
WYAAANNRNNNGCG	2	.12	.02	XIE126
CCCNNNNNNAAGWT	3	.41	.02	XIE158
ATTTCAW	6	1.97	.03	XIE174
CTGRNNNTTGW	3	.61	.04	XIE152