nature genetics

Chromatin marks identify critical cell types for fine mapping complex trait variants

Gosia Trynka[1–4,8], Cynthia Sandor[1–4,8], Buhm Han[1–4], Han Xu[5], Barbara E Stranger[1,4,7], X Shirley Liu[5] & Soumya Raychaudhuri[1–4,6]

If trait-associated variants alter regulatory regions, then they should fall within chromatin marks in relevant cell types. However, it is unclear which of the many marks are most useful in defining cell types associated with disease and fine mapping variants. We hypothesized that informative marks are phenotypically cell type specific; that is, SNPs associated with the same trait likely overlap marks in the same cell type. We examined 15 chromatin marks and found that those highlighting active gene regulation were phenotypically cell type specific. Trimethylation of histone H3 at lysine 4 (H3K4me3) was the most phenotypically cell type specific ($P < 1 \times 10^{-6}$), driven by colocalization of variants and marks rather than gene proximity ($P < 0.001$). H3K4me3 peaks overlapped with 37 SNPs for plasma low-density lipoprotein concentration in the liver ($P < 7 \times 10^{-5}$), 31 SNPs for rheumatoid arthritis within CD4$^+$ regulatory T cells ($P = 1 \times 10^{-4}$), 67 SNPs for type 2 diabetes in pancreatic islet cells ($P = 0.003$) and the liver ($P = 0.003$), and 14 SNPs for neuropsychiatric disease in neuronal tissues ($P = 0.007$). We show how cell type–specific H3K4me3 peaks can inform the fine mapping of associated SNPs to identify causal variation.

Recent work showing that common phenotypically associated SNPs are enriched for expression quantitative trait loci (eQTLs)[1–6] suggests that they might act by altering gene regulatory regions. One example is a common non-coding variant associated with plasma low-density lipoprotein (LDL) concentration. This variant modifies a CEBPB transcription factor–binding site in an enhancer and, in doing so, alters the expression of *SORT1*, a gene that affects plasma LDL concentration[7]. Another similar example is an intergenic risk allele for systemic lupus erythematosus (SLE) that decreases *TNFAIP3* transcription by modifying the nuclear factor (NF)-κb–binding site within a promoter[8]. Whereas many eQTLs and regulatory variants act universally, the ones most relevant to disease might have tissue specific activity[6]. The cell type specificity of regulatory elements is one of the major limitations in pursuing functional studies to investigate the regulatory potential of common alleles[9–13].

One approach to identify regulatory elements influenced by common variants involves assaying epigenetic chromatin marks[14–16]. For example, H3K4me3 and monomethylation at H3K4 (H3K4me1) highlight active promoters and enhancers. But, a practical challenge of this approach is that dozens of chromatin marks might potentially be assayed[17], and it is prohibitive to conduct studies on all of them in large numbers of different tissues or in samples collected from many individuals. However, because chromatin marks colocalize[18], the status of a small subset of the most informative marks might be characterized, allowing for more focused assays in tissue libraries and populations to link variants to regulatory mechanisms. Additionally, it is challenging for a given phenotype to know which cell type(s) are most useful to assay chromatin marks in order to fine map risk alleles. If the critical cell types were known, then it might be possible to identify the biologically important cell type–specific eQTLs.

Here, we hypothesize that a proportion of alleles for a given phenotype influence gene regulation by altering regulatory elements that control expression within the cell types most relevant to the phenotype. If this is the case, then variants associated with the same phenotype should overlap marks preferentially occurring within the same cell type. Therefore, to identify the most informative chromatin marks, we quantify the degree to which their activity in specific cell types near phenotypically associated variants tracks with phenotype. We then show how those chromatin marks that are most phenotypically cell type specific can identify causal cell types, asserting that cell type–specific marks might be used to fine map and identify the plausible causal variant at a particular locus.

RESULTS

Summary of statistical methods

We first sought to define a score that corresponds to the possibility that a phenotypically associated SNP or a variant in tight linkage disequilibrium (LD) with it can alter cell type–specific gene regulation, as highlighted by a specific chromatin mark. We define chromatin marks as precise positions in the genome where there is a significant

[1]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [2]Division of Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [3]Partners Center for Personalized Genetic Medicine, Boston, Massachusetts, USA. [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [5]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts, USA. [6]Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK. [7]Present addresses: Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, Illinois, USA and Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to S.R. (soumya@broadinstitute.org).

Received 6 July; accepted 28 November; published online 23 December 2012; doi:10.1038/ng.2504

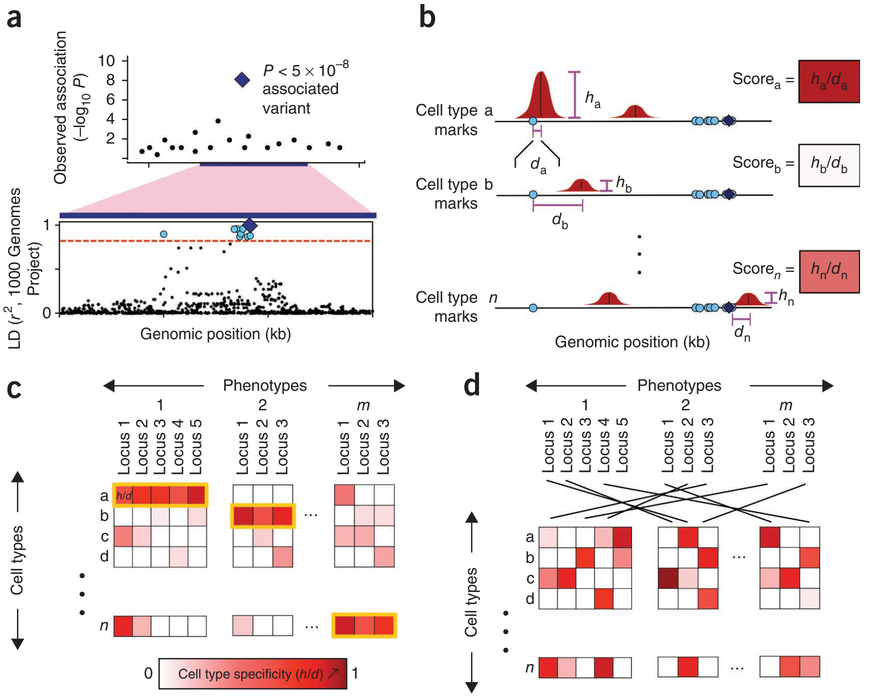

Figure 1 Overview of the statistical approach. (**a**) For phenotypically associated variants, other variants in tight LD are found. For each SNP associated with a phenotype from genetic studies (lead SNP, blue diamond; top), we define a locus by identifying SNPs in tight LD ($r^2 > 0.8$, dashed red line; bottom) using data from the 1000 Genomes Project (blue dots; bottom). (**b**) Each locus is scored on the height and distance of the nearest peak to a variant in LD. For a selected chromatin mark, we define peaks (red) in *n* cell types across the genome. For each SNP in the locus (blue diamond and light-blue circles), we compute a score equal to the height of the closest peak (vertical purple line) divided by the distance to the summit in each of the *n* cell types (horizontal purple line). In each locus within each cell type, we note the value of the SNP with the highest score: this measure reflects the overlap between a locus and a cell type–specific regulatory element. (**c**) Across many phenotypes, we assess whether marks overlap alleles in specific cell types. Here, the measure of cell type specificity of each risk locus is represented by the intensity of red color. A phenotypically cell type–specific mark should consistently give signal in one or a small number of cell types for a given phenotype (yellow outline). We quantify the phenotypic cell type specificity of each mark. (**d**) Permutations are performed to assess the significance of phenotypic cell type specificity. To compute the significance of the phenotypic cell type specificity for a chromatin mark, we permutate SNPs from different loci across phenotypes; this preserves tissue-specific signals without altering the correlation and prevalence of tissue-specific signals.

excess of reads from chromatin immunoprecipitation and sequencing (ChIP-seq) data over control sequencing data. We assume that variants close to or directly under tall chromatin mark peaks in specific cell types might be involved in cell type–specific gene regulation; on the other hand, variants that are far from chromatin mark peaks are much less likely to have a direct role in gene regulation. First, for each phenotypically associated SNP, we identified each SNP or insertion and/or deletion (indel) in tight LD ($r^2 > 0.8$ in 1000 Genomes Project data[19]; **Fig. 1a**). Next, for each cell type, we assigned each variant in LD a score proportional to the height of the nearest chromatin mark peak (referred to as *h*; Online Methods) divided by the physical distance to the summit (*h*/*d* in **Fig. 1b**; referred to as *s*; Online Methods). If the physical distance to the nearest peak is more than 2.5 kb, then the score is set to 0 to obviate any confounding distal effects. Thus, a variant in LD directly under a strong peak will receive a very high score. For each cell type, we assigned the phenotypically associated SNP the maximum score achieved by any of its variants in LD. To quantify the specificity of signals across cell types (as opposed to the absolute magnitude), we normalized the *h*/*d* scores so that the Euclidean metric across cell types was one (normalized *h*/*d* scores (*sn*; Online Methods). Thus, a SNP within a chromatin mark that is active in only one cell type will have a high score of 1 in that cell type and 0 in others. In contrast, a SNP close to chromatin marks that are not cell type specific will have similarly modest scores across cell types.

Then, we wanted to quantify the phenotypic cell type specificity of the overlap between SNPs and chromatin marks. To do this, we identified sets of SNPs associated with different phenotypes and then assessed the phenotypic cell type specificity of different marks (**Fig. 1c**). For informative marks, one or few cell types should consistently score highly across many of the SNPs for a given phenotype. For an uninformative chromatin mark, the cell types with the greatest scores vary from SNP to SNP within the same phenotype. Therefore, for informative marks, there should be minimal deviation of scores within a phenotype across multiple cell types. To quantify the phenotypic cell type specificity of a chromatin mark, we defined a metric representing the variation of signal seen within a cell type within a specific phenotype (referred to as *d*; Online Methods). We evaluated the statistical significance of this metric with permutations with which we randomly reassigned SNPs to phenotypes (**Fig. 1d**). This permutation strategy restricts analysis to only phenotypically associated SNPs and, in doing so, avoids biases that might result from known differences between phenotypically associated SNPs and non–phenotypically associated SNPs in local LD structure, gene density and epigenetic activity. We note that this approach accurately estimates type I error (**Supplementary Fig. 1a**).

Active gene regulation is phenotypically cell type specific

To test the phenotypic cell type specificity of individual marks, we identified a set of SNPs associated with any one of many complex traits[20]. We selected only SNPs associated in European populations to facilitate LD calculations. To ensure adequate power, we selected only those traits that had at least 15 reported associations in European populations. Then, we pruned SNPs by LD so that they were all independent ($r^2 < 0.1$ and >100 kb away from other associated SNPs in the genome; Online Methods). This resulted in a set of 510 independent SNPs associated with 31 complex traits. After defining the genomic locations and heights of peaks for 15 chromatin marks assayed in 14 Encyclopedia of DNA Elements (ENCODE) cell types[15] (**Supplementary Table 1**), we observed statistically significant phenotypic cell type specificity for 4 marks ($P < 0.0033 = 0.05/15$; **Fig. 2**). The most strongly associated chromatin marks were H3K4me3 and acetylation of histone H3 at lysine 9 (H3K9ac) ($P < 1 \times 10^{-6}$), which are known to highlight active gene promoters[16,21]. In fact, all four most significant modifications are known to occur at regions of the

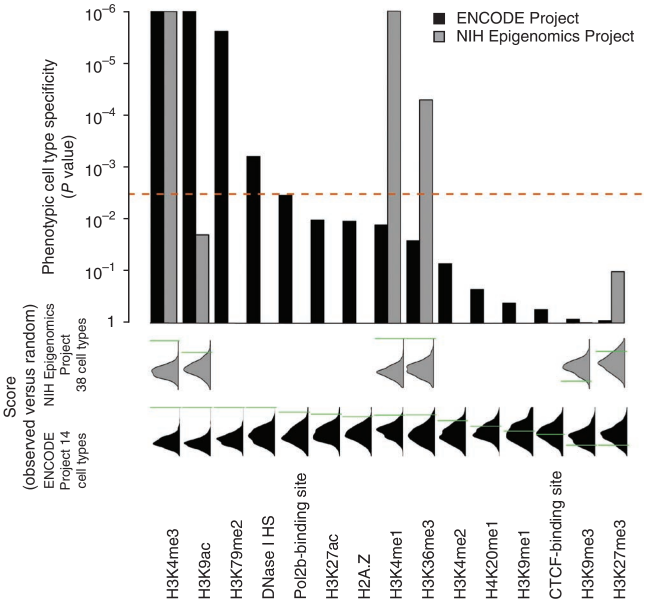

Figure 2 Evaluating the significance of phenotypic cell type specificity for different marks. We used two data sets of marks assayed in different cell types: the ENCODE Project and NIH Epigenomics Project. For each mark, we performed up to 1 million permutations of SNPs and phenotypes to calculate the null distribution of phenotypic cell type specificity for comparison to observed phenotypic cell type specificity. Below, we show the observed phenotypic cell type specificity (green lines) against the null distribution (black and gray density plots). Above, we plot the corresponding *P* values. The red dashed line indicates the significance threshold after correcting for the testing of multiple independent hypotheses.

genome involved in active gene transcription; DNase I hypersensitivity sites (DHSs; $P < 1 \times 10^{-3}$) and dimethylation of histone H3 at lysine 79 (H3K79me2; $P < 1 \times 10^{-5}$) identify active promoter, enhancer or transcribed regions. Because some chromatin marks colocalize (**Supplementary Fig. 2**), we performed conditional analyses to assess whether chromatin marks contributed to phenotypic cell type specificity independently (**Supplementary Fig. 3**). We observed that the highly significant associations of H3K4me3, DHSs and H3K9ac were generally not independent. In contrast, we found that chromatin marks that did not correspond to active gene regulation were not phenotypically cell type specific. In particular, H3K9me1, H3K9me3, CTCF-binding sites and trimethylation at histone H3 lysine 27 (H3K27me3), highlighting transcriptionally repressed heterochromatic insulator and polycomb-repressed regions, respectively, showed no evidence of being phenotypically cell type specific ($P > 0.40$).

To assess the reproducibility of these results, we conducted a similar analysis of data from the US National Institutes of Health (NIH) Epigenomics Project, consisting of assays for 6 different chromatin marks in 38 different cell types[22] (**Supplementary Table 2**). We again observed that the most informative mark was H3K4me3 ($P < 1 \times 10^{-6}$), along with H3K4me1 (**Fig. 2**). H3K9ac was more nominally significant ($P = 0.03$), perhaps owing to the fewer cell types assayed in this experiment. The concordance of the results from these two data sets was reassuring when considering that the data from the ENCODE Project were obtained on cell lines, whereas most of the NIH Epigenomics Project data were obtained using primary cell types.

Our approach benefits from taking advantage of 1000 Genomes Project data to identify variants in LD (**Fig. 1a**). Repeating our analysis using only the reported lead SNPs and not examining SNPs in LD resulted in considerably less significant results (**Supplementary Fig. 1b**). We note that some of the variation in phenotypic cell type specificity could be related to the variable number of assayed cell types for different chromatin marks; power to detect phenotypic cell type specificity correlates with the number of assayed cell types (**Supplementary Fig. 4**).

Variants colocalize with cell type–specific H3K4me3 peaks

Because chromatin marks tend to concentrate in and around genes, we considered the possibility that the observed overlap between H3K4me3 peaks and variants might be an artifact of proximity to gene transcript sequences with phenotypically cell type specific expression. To assess the role of the specific peak locations versus proximity to specifically expressed genes, we repeated our analyses after randomly shifting the specific location of peaks locally (± 10 kb, s.d. of 2.5 kb) within phenotypically associated loci. While these small shifts would maintain the proximity of peaks to genes, they would disrupt the specific colocalization of variants and H3K4me3 peaks. Indeed, in 1,000 such experiments, we found that shifting peak locations lowered the significance of phenotypic cell type specificity (median $P = 0.03$), and we did not observe any instance where the phenotypic cell type specificity was more significant than it was in the actual data (**Supplementary Fig. 5**). This result strongly suggests that the specific colocalization of variants in LD with phenotypically associated SNPs and H3K4me3 peaks rather than proximity to gene structures is driving the phenotypic cell type specificity signal ($P < 0.001$ by permutation).

Enhancers and promoters underlie phenotypic cell type specificity

To understand whether the phenotypic cell type specificity that we observed was driven by the activity of promoters or enhancers, we divided chromatin peaks into those falling within proximal promoter regions (including the transcriptional start site (TSS) ± 2 kb) and those falling outside of promoter regions and repeated our analysis. Whereas phenotypic cell type specificity was seen both within and outside of the immediate promoter regions, H3K4me3, H3K79me2 and DHSs were more significantly phenotypically cell type specific outside of promoter regions than within (**Supplementary Fig. 6**). We note that, although H3K4me3 marks are not generally thought of as being enriched in enhancers, there was evidence that they can be enriched in strong and disease-associated enhancers[9,23,24]. Alternatively, H3K4me3 enrichment outside of promoter sites might also represent unannotated sites.

We further assessed the phenotypic cell type specificity of previously published functional annotations on the basis of hidden Markov model states capturing information on nine separate chromatin marks[9]. We observed that hidden states 4 and 5, corresponding to active proximal enhancers and active distal enhancers, respectively, were most significantly phenotypically cell type specific (**Supplementary Fig. 7**). State 4 is highly enriched for H3K4me3 peaks, the mark that we observed to be the most phenotypically cell type specific.

Identification of key cell types for four phenotypes

We identified the cell types within which common variants likely influence gene regulation using published SNPs for 4 distinct phenotypes (**Fig. 3** and **Supplementary Table 3**) and H3K4me3 data from the Epigenomics Project for a panel of 34 cell-types[22]. We selected these phenotypes because there is a reasonable sense of what the critical cell types might be and because a sufficient number of associated SNPs had been identified. For each phenotype, we assigned a cell type specificity score to each of its associated variants (**Fig. 1a**,**b** and

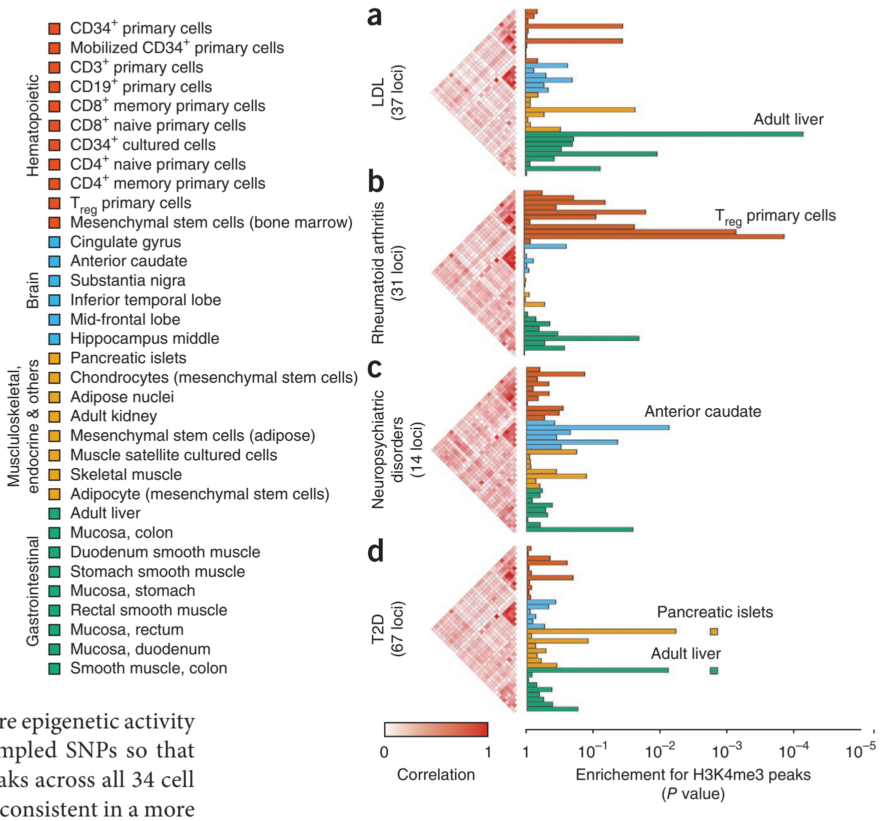

Figure 3 SNPs for four complex traits overlap H3K4me3 marks in specific cell types. (**a**–**d**) We considered four phenotypes: LDL cholesterol plasma concentration (**a**), rheumatoid arthritis (**b**), neuropsychiatric disorders (schizophrenia and bipolar disease) (**c**) and T2D (**d**). For each phenotype, we calculated the cell type–specific overlap with H3K4me3 histone modification peaks in 34 tissues (listed on the left). The histograms on the right show the significance of the overlap for each tissue with variants from each of the phenotypes, estimated by sampling sets of SNPs matched so that the total number of peaks overlapping SNPs in LD was the same as in the test set. Adjacent to each histogram, we present correlation coefficients between two tissues based on scores computed from randomly sampled sets of independent loci. Colored boxes in **d** show independent *P* values for pancreatic islets and liver computed by removing the SNPs scoring highly in one tissue but not the other.

Online Methods) and compared to scores from equal-sized sets of matched SNP sets sampled from 45,950 LD-pruned SNPs[3]. Because phenotypically associated SNPs have more epigenetic activity than other SNPs, we were careful to match sampled SNPs so that they had similar total numbers of H3K4me3 peaks across all 34 cell types as associated SNPs. Results were generally consistent in a more stringent analysis when we sampled instead from only phenotypically associated SNPs from the National Human Genome Research Institute (NHGRI) genome-wide association study (GWAS) catalog[20] (**Supplementary Fig. 8**). In addition to these phenotypes, we present separately the results for four additional phenotypes, B-cell–specific *cis* eQTL associations, SLE, type 1 diabetes (T1D) and body mass index (BMI) (**Supplementary Fig. 9**); in all of those instances, except BMI, we were able to identify highly significant cell types.

Application to plasma LDL concentration implicates liver

As a positive control, we tested 37 SNPs associated with LDL concentration[25] for overlap with H3K4me3 marks in different tissues. These variants should implicate regulatory activity within the liver, according to previous work[7,26,27]. In aggregate, we observed that the 37 SNPs implicated a total of 1,501 H3K4me3 peaks in 34 different cell types. The most significant cell type was adult liver tissue ($P = 7.2 \times 10^{-5}$; **Fig. 3a**). We observed overlap with liver-specific peaks using other phenotypically cell type–specific marks, including H3K9ac ($P = 0.003$) and H3K4me1 ($P = 0.002$). In contrast, we observed little association with liver for the H3K27me3 or H3K9me3 marks (**Fig. 2** and **Supplementary Table 4**). Examining the relative proximity and specificity of the SNPs within 10,000 sets of matched SNP sets used to calculate statistical significance, we identified the 95th-percentile threshold at a score of 0.58 (**Fig. 4a**). Of the 37 SNPs associated with LDL concentration, 7 (19%) were near to a highly liver-specific chromatin mark at this threshold. These seven SNPs are generally in tight LD with a variant that is very close to cell type–specific H3K4me3 peaks (median of 132 bp away; see **Supplementary Table 3** for details on the specific SNPs).

Application to rheumatoid arthritis implicates CD4+ T_{reg} cells

For rheumatoid arthritis and other autoimmune diseases, the critical immune cell types are often not clearly defined in the literature and can be controversial[28–30]. When we tested the 31 SNPs associated with rheumatoid arthritis[31], we observed that they implicated 1,328 H3K4me3 peaks in 34 tissues, with the most significant association to CD4+ T cells and, in particular, CD4+ regulatory T (T_{reg}) cells ($P = 1.3 \times 10^{-4}$; **Fig. 3b**). The phenotypically similar CD4+ memory T cells were also highly significantly associated ($P = 7.0 \times 10^{-4}$)[32]. Of the 31 SNPs associated with rheumatoid arthritis, we found that 6 (19.3%) were close to chromatin marks that were highly specific to CD4+ T_{reg} cells, with relative specificity of 0.53 or greater (permuted 95th-percentile threshold; **Fig. 4b**). These 6 SNPs are generally in tight LD with a variant that is very close to cell type–specific H3K4me3 peaks (median of 37 bp away; see **Supplementary Table 3** for details on the specific SNPs).

In instances where dense genotyping has been applied to localize the association signal, we speculate that cell type–specific overlap might become more apparent. Indeed, for the 31 loci associated with rheumatoid arthritis, we examined recent results from a fine-mapping study using the dense genotyping platform the Immunochip[33]. Indeed, when repeating the analysis with the newly defined index SNPs from each locus using dense genotyping data, we found that the significance of the enrichment for CD4+ T_{reg} cells increased (5.1×10^{-5}; **Supplementary Fig. 10**) and that the median specificity score for each locus increased from 0.13 to 0.16.

Application to psychiatric disorders implicates neuronal tissues

The 14 independent SNPs from neuropsychiatric disorders[34,35] mapped within 874 H3K4me3 peaks. Despite the limited power of this analysis, we were encouraged to see that these SNP associations implicated multiple neuronal tissues, including the anterior caudate nucleus ($P = 0.0076$) and the mid-frontal lobe of the brain ($P = 0.044$) (**Fig. 3c**); we also observed a likely spurious association with colonic smooth muscle ($P = 0.026$). The role of the frontal lobe in neuropsychiatric disease in particular has long been appreciated[36–38]. Although none

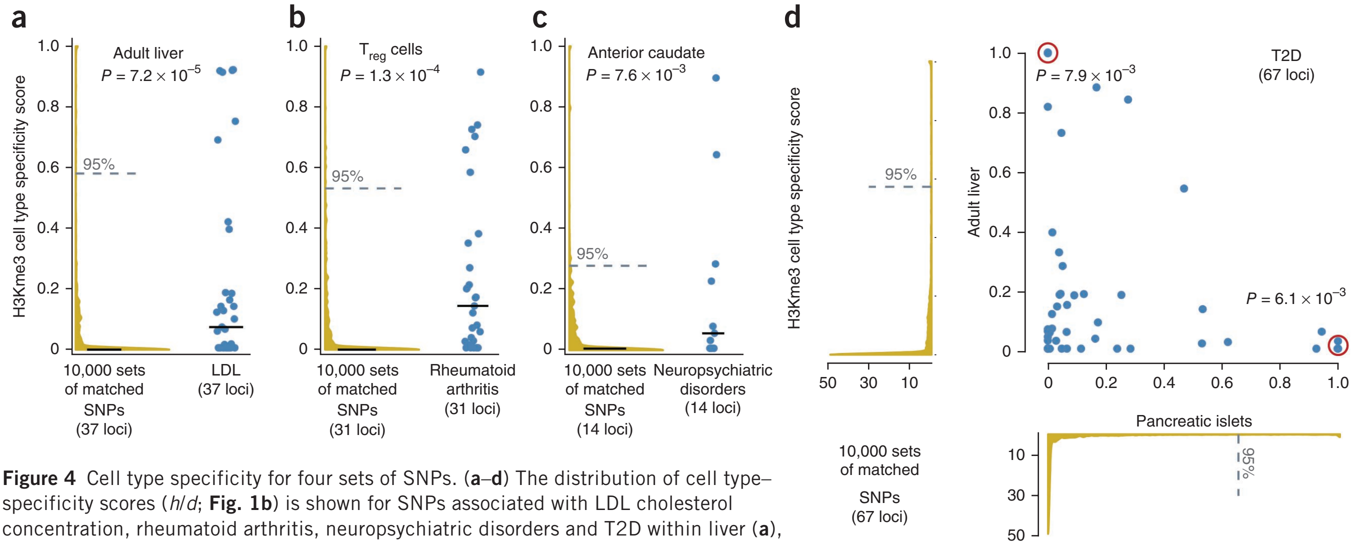

Figure 4 Cell type specificity for four sets of SNPs. (**a**–**d**) The distribution of cell type–specificity scores (*h*/*d*; **Fig. 1b**) is shown for SNPs associated with LDL cholesterol concentration, rheumatoid arthritis, neuropsychiatric disorders and T2D within liver (**a**), CD4$^+$ T$_{reg}$ cells (**b**), anterior caudate nucleus (**c**) and jointly in pancreatic islets (*x* axis) and liver (*y* axis) (**d**). Blue points represent cell type specificity scores. Red circles indicate overlapping points, representing SNPs with very similar scores. We compare these scores to specificity scores in the same tissue of 10,000 sampled sets of matched SNPs from HapMap (yellow density plots). We plot the median specificity for both the distribution of observed SNPs and the sampled sets of matched SNPs (solid lines). Also, we present the 95th-percentile threshold for the sampled sets of matched SNPs (dashed line), which we use as a specificity cutoff. For each phenotype, about one-fourth of variants overlap cell type–specific H3K4me3 peaks.

of these results reached a conservative level of significance after correcting for multiple-hypothesis testing, we are hopeful that additional SNP discoveries will help clarify this result further. Of the 14 SNPs associated with neuropsychiatric disorders, 3 (21%) had a tissue-specific chromatin mark within the anterior caudate, with a relative specificity of 0.28 or greater (permuted 95th percentile; **Fig. 4c**).

Application to T2D implicates pancreatic islets and liver

In certain instances, it might be plausible that multiple tissues could be implicated in a disease. When we examined 67 SNPs for type 2 diabetes (T2D)[39–50], implicating a total of 2,776 H3K4me3 peaks within 34 different cell types, we observed the most significant enrichment in pancreatic islets ($P = 0.0061$) and the liver ($P = 0.0079$) (**Fig. 3d**). In particular, of the 67 SNPs associated with type 2 diabetes, 14 (20.1%) were either highly specific for chromatin marks within the liver (at a 0.57 permuted 95th-percentile threshold) or pancreatic islets (at a 0.65 permuted 95th-percentile threshold); these SNPs are in tight LD with a marker that has a median distance of 46 bp from the summit of a cell type–specific peak. When we tested the pancreatic islet and liver tissues together, we found that the combination of liver and pancreatic islets was even more significant than the tissues individually ($P = 2.0 \times 10^{-4}$; Online Methods) and was more significant than all other possible tissue pairs. We found that the SNPs driving the overlap in the two tissues were distinct (**Fig. 4d**). When we removed the SNPs most specific for pancreatic islet marks (score > 0.3), we observed that enrichment in liver was even more apparent ($P = 0.0032$); similarly, when we removed the SNPs most specific for overlap with liver marks (score > 0.3), we observed that the enrichment in pancreatic islets was also more apparent ($P = 0.0026$). Both islet cells and the liver have long been known to have a key role in mediating glucose synthesis, insulin secretion and diabetes[51].

Fine mapping with cell type–specific H3K4me3 peaks

One of the major challenges in understanding complex trait associations is to identify the causal variants and the mechanisms through which they affect genes. Associated variants can be fine mapped to variants in tight LD within cell type–specific chromatin marks in the appropriate cell type. Here, we present examples where cell type–specific H3K4me3 peaks can potentially be used to localize associated variants to causal variants.

First, we considered the rs629301 SNP that is associated with plasma LDL concentration in the region including the *SORT1* gene (**Fig. 5a**). A liver-specific H3K4me3 peak, not seen as prominently in other cell types, overlapped with this SNP and three other variants in tight LD with it. This H3K4me3 peak is located far from the TSS region and corresponds to a hepatocyte enhancer region[7]. The closest SNP to the summit of the peak (87 bp away) is the rs12740374 functional variant. This variant is known to alter a CEBPB-binding site within the enhancer region controlling *SORT1*.

Another example is provided by the locus for T2D represented by the rs704184 reported SNP association. rs10814915, tightly in LD with the reported GWAS SNP ($r^2 = 0.93$), scored highly for pancreatic islets but showed no tissue specificity for the liver (**Fig. 5b**). This SNP located only 84 bp away from the summit of a highly pancreatic islet–specific peak. rs10814915 is predicted to be present within a sequence bound by the glucocorticoid receptor (GR)[52], which is known to have a role in pancreatic islets and glucose regulation. The SNP resides within an intron of the *GLIS3* gene, which is involved in the development of pancreatic islets.

Finally, we examined the locus for rheumatoid arthritis defined by a reported association with the rs13119723 SNP in the intron of a gene with unknown function, *KIAA1109*. This SNP is in LD with other variants spanning over 500 kb within this locus, rendering fine-mapping efforts particularly challenging. We identified a SNP, rs13140464, in tight LD with rs13119723 ($r^2 = 0.9$) (**Fig. 5c**), which maps only 116 bp from the summit of the H3K4me3 peak, which is highly specific to CD4$^+$ T$_{reg}$ cells with a score of 0.63. This SNP is located between the *IL2* and *IL21* genes, 122 kb downstream of *IL2* and 34 kb upstream of *IL21*, and is 280 kb away from the published SNP. It is tempting to speculate that rs13140464 might act by altering a highly cell type–specific regulatory sequence controlling *IL2* expression, which has a key role in CD4$^+$ T$_{reg}$ maturation[53].

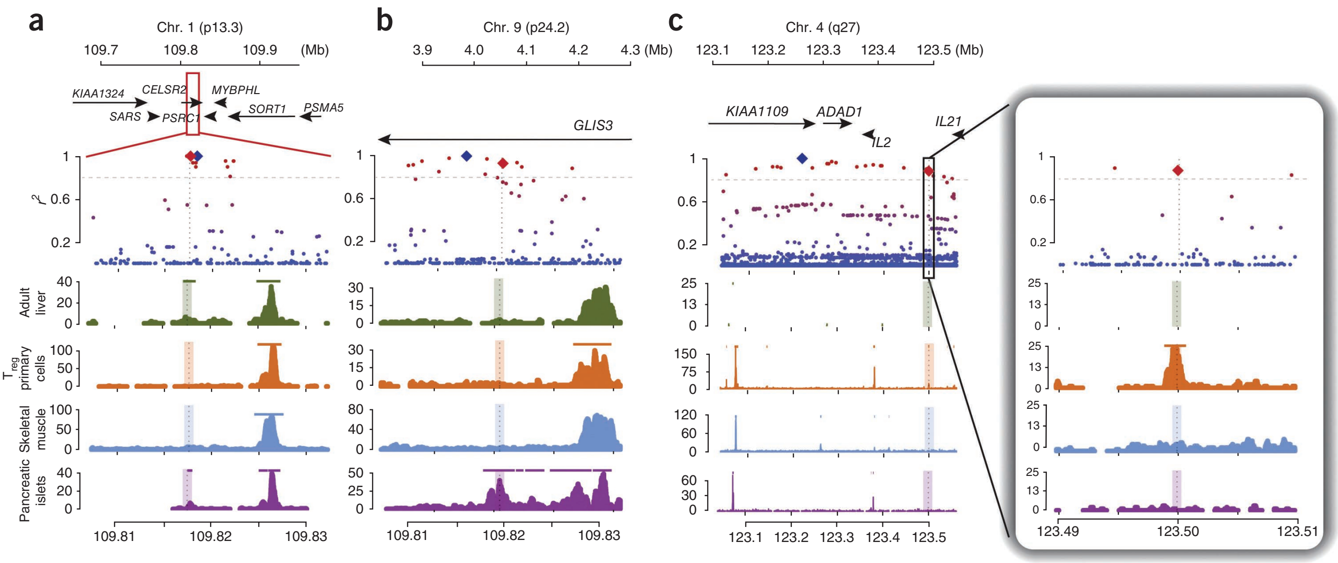

Figure 5 Selected phenotypically associated loci with high cell type specificity. We present three examples of loci with cell type–specific overlap with H3K4me3 peaks. Top, genomic coordinates and genes near the associated SNP. Middle, lead SNP (blue diamond) and other nearby SNPs from the 1000 Genomes Project (red dots correspond to those with high r^2, blue dots correspond to those with low r^2). We also show the SNP that is closest to the cell type–specific peak (red diamond). Bottom, H3K4me3 sequence tag counts for selected cell types. Colored horizontal lines in the tissue panels correspond to peak calls. Dashed vertical lines mark the summits of phenotypically cell type–specific peaks. (**a**–**c**) Shown are the *SORT1* locus for LDL (**a**), the *GLIS3* locus for T2D (**b**) and the *IL2*-*IL21* locus for rheumatoid arthritis (**c**).

DISCUSSION

In this study, we demonstrated that chromatin marks highlighting active regulatory regions, such as H3K4me3, H3K9ac and DHSs, overlap phenotypically associated variants; furthermore, this overlap is phenotypically cell type specific. These results strongly support the hypothesis that many complex disease and trait alleles might act by influencing gene regulation in a cell type–specific manner. In addition, we quantified the degree to which different marks are cell type specific in their overlap with phenotypically associated SNPs. These cell type–specific marks might not only be used to connect phenotypes to specific cell types, but they might also be useful in mapping phenotype-associated SNPs to potential regulatory variants. In particular, we consistently observed that H3K4me3 marks could be used to effectively identify specific cell types that are enriched among specific phenotypes. We note that this statistical approach can be applied to assess the significance of other chromatin marks or other cell type–specific gene annotations as they become available.

In the phenotypes that we examined, we found that about one-fourth of associated variants could be connected to a highly cell type–specific mark within a critical cell type (**Fig. 5**). In instances where we do not observe a SNP in tight LD within a highly cell type–specific H3K4me3 peak, it is possible that a regulatory region that is not cell type specific might be altered. Alternatively, in some instances the reported SNP association will need to be further refined with dense genotyping, or undiscovered variants in tight LD will need to be ascertained through sequencing, before the effect of a cell type–specific peak can be identified. Finally, for many phenotypes, multiple cell types could be involved, in which case this approach might have limited efficacy. We demonstrated one example of this type of scenario in T2D, where we detected effects both in liver and pancreatic islet cell types.

We acknowledge that our approach is potentially sensitive to the diversity and number of cell types assayed. For instance, a limited application to a set of hematopoietic cell types might not be particularly informative if a set of purely neurological phenotypes is assayed. We note that our approach depends critically on technical factors—for instance, the quality of antibody reagents, experimental protocols or other technical factors that might introduce noise into specific chromatin mark assays could mitigate true signals. Our approach may perform better on the chromatin marks with higher quality assays.

Once variants and cell types are identified, they will likely be excellent candidates for cell type–specific functional investigations, including allelic imbalance assays to define *cis*-eQTL activity[54], cell type–specific DHS quantitative trait locus (dsQTL) analyses[55] and identification of active transcription factor–binding sites. These cell type–specific investigations in appropriately chosen cell types might ultimately help to lead investigators from common disease variation to causal variants and molecular mechanisms.

URLs. All software is available online at http://www.broadinstitute.org/mpg/epigwas/. ENCODE, http://genome.ucsc.edu/ENCODE/downloads.html; NIH Roadmap Epigenomics Mapping Consortium, http://www.roadmapepigenomics.org/; NHGRI GWAS catalog, http://www.genome.gov/gwastudies/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Daly, M. Kellis, D. Diogo, X. Hu, Y. Okada, R. Plenge, S. Ripke, G. Srivastava, E. Stahl and S. Sunyaev for critical feedback and discussion. G.T. is supported by the Rubicon grant from The Netherlands Organization for Scientific Research (NWO). B.E.S. and S.R. are supported by the Harvard University Milton Fund, and Brigham and Women's Hospital. S.R. is also supported by funds from the US NIH (K08AR055688 and U01HG0070033) and the Arthritis Foundation. X.S.L. is also supported by funds from the US NIH (R01 HG004069). We thank the

ENCODE Project, supported by the NHGRI, and the NIH Roadmap Epigenomics Mapping Consortium for making data available.

AUTHOR CONTRIBUTIONS

S.R. led the study. G.T., C.S., S.R., B.H. and H.X. performed the analysis. G.T., C.S., S.R., B.E.S. and X.S.L. wrote the manuscript. All authors reviewed the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2504.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
2. Fraser, H.B. & Xie, X. Common polymorphic transcript variation in human disease. *Genome Res.* **19**, 567–575 (2009).
3. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
4. Nica, A.C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
5. Fehrmann, R.S. *et al.* eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7**, e1002197 (2011).
6. Fairfax, B.P. *et al.* Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
7. Musunuru, K. *et al.* From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
8. Adrianto, I. *et al.* Association of a functional variant downstream of *TNFAIP3* with systemic lupus erythematosus. *Nat. Genet.* **43**, 253–258 (2011).
9. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
10. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **107**, 21931–21936 (2010).
11. Waki, H. *et al.* Global mapping of cell type–specific open chromatin by FAIRE-seq reveals the regulatory role of the NFI family in adipocyte differentiation. *PLoS Genet.* **7**, e1002311 (2011).
12. Atchison, M.L. Enhancers: mechanisms of action and cell specificity. *Annu. Rev. Cell Biol.* **4**, 127–153 (1988).
13. Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* **21**, 1757–1767 (2011).
14. Boyle, A.P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
15. Encode Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
16. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
17. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
18. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* **40**, 897–903 (2008).
19. Thousand Genomes Project. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
20. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
21. Bernstein, B.E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169–181 (2005).
22. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
23. Pekowska, A. *et al.* H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.* **30**, 4198–4210 (2011).
24. Jia, L. *et al.* Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet.* **5**, e1000597 (2009).
25. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
26. Smith, L.C., Pownall, H.J. & Gotto, A.M. Jr. The plasma lipoproteins: structure and metabolism. *Annu. Rev. Biochem.* **47**, 751–757 (1978).
27. Hobbs, H.H., Brown, M.S. & Goldstein, J.L. Molecular genetics of the LDL receptor gene in familial hypercholesterolemia. *Hum. Mutat.* **1**, 445–466 (1992).
28. Firestein, G.S. Evolving concepts of rheumatoid arthritis. *Nature* **423**, 356–361 (2003).
29. Lee, D.M. *et al.* Mast cells: a cellular link between autoantibodies and inflammatory arthritis. *Science* **297**, 1689–1692 (2002).
30. Boilard, E. *et al.* Platelets amplify inflammation in arthritis via collagen-dependent microparticle production. *Science* **327**, 580–583 (2010).
31. Stahl, E.A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).
32. Akbar, A.N., Vukmanovic-Stejic, M., Taams, L.S. & Macallan, D.C. The dynamic co-evolution of memory and regulatory CD4$^+$ T cells in the periphery. *Nat. Rev. Immunol.* **7**, 231–237 (2007).
33. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44**, 1336–1340 (2012).
34. Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near *ODZ4*. *Nat. Genet.* **43**, 977–983 (2011).
35. Schizophrenia Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–976 (2011).
36. Goldman-Rakic, P.S. & Selemon, L.D. Functional and anatomical aspects of prefrontal pathology in schizophrenia. *Schizophr. Bull.* **23**, 437–458 (1997).
37. Goldstein, J.M. *et al.* Cortical abnormalities in schizophrenia identified by structural magnetic resonance imaging. *Arch. Gen. Psychiatry* **56**, 537–547 (1999).
38. Strakowski, S.M., Delbello, M.P. & Adler, C.M. The functional neuroanatomy of bipolar disorder: a review of neuroimaging findings. *Mol. Psychiatry* **10**, 105–116 (2005).
39. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
40. Cho, Y.S. *et al.* Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat. Genet.* **44**, 67–72 (2012).
41. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105–116 (2010).
42. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).
43. Kooner, J.S. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.* **43**, 984–989 (2011).
44. Perry, J.R. *et al.* Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in *LAMA1* and enrichment for risk variants in lean compared to obese cases. *PLoS Genet.* **8**, e1002741 (2012).
45. Qi, L. *et al.* Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Hum. Mol. Genet.* **19**, 2706–2715 (2010).
46. Saxena, R. *et al.* Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci. *Am. J. Hum. Genet.* **90**, 410–425 (2012).
47. Shu, X.O. *et al.* Identification of new genetic risk variants for type 2 diabetes. *PLoS Genet.* **6**, pii: e1001127 (2010).
48. Tsai, F.J. *et al.* A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLoS Genet.* **6**, e1000847 (2010).
49. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
50. Yamauchi, T. *et al.* A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at *UBE2E2* and *C2CD4A-C2CD4B*. *Nat. Genet.* **42**, 864–868 (2010).
51. Seino, S., Shibasaki, T. & Minami, K. Dynamics of insulin secretion and the clinical implications for obesity and diabetes. *J. Clin. Invest.* **121**, 2118–2125 (2011).
52. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
53. Setoguchi, R., Hori, S., Takahashi, T. & Sakaguchi, S. Homeostatic maintenance of natural Foxp3$^+$ CD25$^+$ CD4$^+$ regulatory T cells by interleukin (IL)-2 and induction of autoimmune disease by IL-2 neutralization. *J. Exp. Med.* **201**, 723–735 (2005).
54. McCarroll, S.A. *et al.* Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nat. Genet.* **40**, 1107–1112 (2008).
55. Degner, J.F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).

ONLINE METHODS

Chromatin mark data. We obtained two publicly available data sets for chromatin mark assays on different sets of tissues. We use the term chromatin mark broadly to include histone modifications and DHSs, as well as common epigenetic features, such as CTCF-binding sites.

First, we used data from the ENCODE Project, which included sequence reads from ChIP-seq assays and controls in up to 14 different cell types from a diverse set of 15 chromatin marks: CTCF-binding sites, the variant H2A histone (H2A.Z), H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me1, H3K9me3, H4K20me1, Pol2b-binding sites and DHSs[15] (**Supplementary Tables 1** and **2**). We separately obtained hidden chromatin state annotations for 8 of the 14 cell types defined by clustering chromatin marks[9].

Second, we used data from the NIH Roadmap Epigenomics Mapping Consortium that assayed only six chromatin marks on a large number of cell types[22]. This data set included sequence reads from ChIP-seq assays and controls for 6 histone modifications—H3K27me3, H3K4me3, H3K36me3, H3K9ac, H3K4me1 and H3K9me3—assayed in 38 adult and fetal tissues (**Supplementary Table 2**).

For both of these data sets, we downloaded data comprising hg19-mapped sequence reads. In instances where there were multiple replicates of a given ChIP-seq assay for the same tissue, we aggregated sequence reads for the individual assays. We also obtained mapped reads from control data comprising sequenced genomic DNA. We ran MACS (v1.4) software to identify significant peaks ($P < 1 \times 10^{-5}$), specific locations within the genome with enrichment of tag sequences, setting the bandwidth parameter to 300 bp[56]. For each chromatin mark, we located its summit, which represents the position with the highest pileup of sequence tags.

Processing chromatin mark data. Once we identified peaks, we used MACS to determine the fold enrichment of tags compared to controls, using the equation

$$f = \frac{\lambda_{\text{mean}}}{\lambda_{\text{local}}} \tag{1}$$

where λ_{peak} and λ_{local} are parameters for a Poisson distribution determined by fitting the local sequence tag distributions in the peak region from ChIP-seq data and control data, respectively. We considered f as the height of peak instead of the raw number tags, as this approach leverages control data to account for local biases in the genome (due to sequencing bias, mapping bias, chromatin structure and genome copy-number variations) and improves the robustness and specificity of the estimation.

We then corrected for global variation in multiple experiments for the same chromatin mark in different cell types, using the equation

$$h_{i,j,\text{norm}} = f_{i,j} \frac{\max\limits_{i \in \text{cell type}} \left\{ \sum\limits_{j} f_{i,j} \right\}}{\sum\limits_{j} f_{i,j}} \tag{2}$$

where $f_{i,j}$ corresponds to fold enrichment for the peak j in the cell type i before normalization, and $h_{i,j,\text{norm}}$ is the fold enrichment after normalization (or the height of the peak).

Phenotypes and associated SNPs. To estimate the phenotypic cell type specificity of each chromatin mark, we identified a comprehensive set of independent SNPs associated with unique phenotypes. We used data from a catalog summarizing results from recent GWAS[20] (downloaded January 2012). We selected only the phenotype-associated SNPs with highly statistically significant associations ($P < 5 \times 10^{-8}$). To ensure the applicability of the 1000 Genomes Project resource, we used only those SNPs associated in populations of European descent. To limit the analysis to phenotypes that have an adequate number of SNP associations, we selected only phenotypes with at least 15 such SNP associations. To ensure the independence of the associated SNPs, we removed SNPs with $r^2 > 0.1$ and those that were <100 kb from a more strongly associated variant in the genome. To preserve a priori specific phenotypes for independent testing, we removed SNPs associated with rheumatoid arthritis, BMI and LDL plasma cholesterol concentration as well as height. For variants associated with multiple phenotypes, we selected a single phenotype association and discarded others; we selected the SNP associated with the phenotype with the fewest SNPs. Our final data set consisted of 510 risk variants associated with 31 diseases or traits.

To test our approach, we also separately identified in the literature 37 SNPs associated with LDL plasma concentration[25], 31 SNPs associated with rheumatoid arthritis risk[31], 67 SNPs associated with T2D risk[39–50] and 14 risk loci for neuropsychiatric disorders[34,35].

Evaluating marks for their phenotypically cell–type specific overlap with variants. *Step 1. Identifying variants in LD with associated SNPs.* We recognized that the observed phenotype association of a given variant might be the consequence of other variants tightly linked to the associated variant (**Fig. 1a**). We therefore comprehensively ascertained variants from the 1000 Genomes Project to identify all variants (SNPs and indels) in LD[19] ($r^2 > 0.8$) on the basis of haplotypes reconstructed with Beagle from the subset of 379 individuals of European descent.

Step 2. Scoring regulatory activity near a risk SNP. Next, we examined chromatin marks in the different cell types located near associated SNPs (**Fig. 1b**). We assumed that the closer an associated SNP (or variant in LD) was to a tall peak, the greater the chance that it might influence a regulatory element highlighted by that peak. We scored each associated SNP k within each cell type by identifying a SNP k' (or indel) in tight LD that was closest to a chromatin mark peak j in tissue i. We then assigned a score $s_{j,k}$ equal to the height of peak j in the tissue i, $h_{i,j,\text{norm}}$ (referred to as h in the main text) divided by the distance d between the SNP k' and the summit of the peak j. If there was no peak within 2.5 kb of each SNP in LD with SNP k, then $s_{i,k}$ was set to zero.

Step 3. Normalization to obtain a cell type specificity score. For each associated SNP k and chromatin mark, we obtained a vector of scores for multiple cell types i. To compare the cell type specificity score across risk variants and phenotypes, we applied Euclidean normalization in the following equation:

$$sn_{i,k} = \frac{s_{i,k}}{\sqrt{\sum\limits_{i} s_{i,k}^2}} \tag{3}$$

This ensured that $sn_{i,k}$ emphasized cell type specificity instead of the magnitude of the signal. For associated risk variants not near any peak, where $s_{i,k}$ is zero for all i, we replaced values with the average of values of other associated SNPs with at least one nonzero $s_{i,k}$ value over all cell types.

Step 4. Estimating the phenotypic cell type specificity of a chromatin mark. If a chromatin mark is informative for phenotypic cell type specificity, then the deviance of chromatin mark overlap for associated SNPs ($sn_{i,k}$) should be minimal for a given phenotype and tissue. If a chromatin mark is not informative, then the deviance of chromatin mark overlap for associated SNPs will be high for a phenotype and tissue.

Therefore, we defined a deviance-based metric of phenotypic cell type specificity for a mark, which was the aggregate sum of the squared differences between $sn_{i,k}$ values and mean values for fixed phenotypes p and cell types i,

$$d = \sum_{i \in \text{cell types}} \sum_{p \in \text{phenotypes}} \left[\sum_{k \in p} \left(\text{mean}_{i,p}(\mathbf{sn}) - sn_{i,k} \right)^2 \right] \tag{4}$$

where $\text{mean}_{i,p}(\mathbf{sn})$ is the mean of the normalized cell specificity scores in the cell type i for SNPs associated with phenotype p. If a mark is informative, then sn scores are dependent on the phenotype and cell type, and this sum of squares should be relatively small.

Step 5. Evaluating the statistical significance of phenotypic cell type specificity. To evaluate the statistical significance of phenotypic cell type specificity for particular marks, we conducted up to 1 million permutations reassigning SNPs to phenotypes randomly. This ensures that the properties of associated SNPs in the analysis are maintained, only disrupting their phenotypic associations.

We recalculated *d* after each permutation. To compute *P* values, we calculated the proportion of *d* scores from permutations (these correspond to the null hypothesis) that were greater than the observed *d* score.

Using overlap with chromatin marks to identify the critical cell type(s) for a specific phenotype. After identifying SNPs associated with a selected phenotype, we compute a cell type specificity score $c_{i,p}$ for a phenotype *p* by summing the normalized $sn_{i,k}$ scores for a cell type *i* and associated SNPs *k* in the following equation:

$$c_{i,p} = \sum_{k \in p} sn_{i,k} \qquad (5)$$

To evaluate the statistical significance of cell type specificity scores $c_{i,p}$, we defined matched sets of SNPs not associated with phenotype *p* and used them to calculate cell type specificity scores. Statistical significance was calculated as the proportion of SNP sets with cell type specificity scores exceeding the observed scores for actual phenotypic SNPs.

To define the matched SNP sets, we required that the sampled SNPs had the same total number of chromatin mark peaks in the region in LD across all cell types as associated SNPs. This ensures that randomly selected SNPs have similar nearby regulatory activity. For the primary analysis, we drew random SNPs from 45,950 independent HapMap SNPs that were clustered to ensure minimal independence[3]. In a secondary analysis, we drew SNPs from phenotypically associated SNPs from the NIH GWAS catalog[20].

Using overlap with marks to identify pairs of critical cell types for a specific phenotype. To test possible pairs of *n* cell types for association, we constructed $(n - 1) \times n/2$ artificial ChIP-seq profiles for each tissue pair. Each artificial profile consisted of all of the peaks defined in both tissues, where the peak heights were reduced to half of their original heights. We then tested for association with cell type pairs in the same way as for single cell types, except that we replaced individual cell type scores with scores for cell type pairs.

56. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).