A method to predict the impact of regulatory variants from DNA sequence

Dongwon Lee1,4, David U Gorkin1,3,4, Maggie Baker1, Benjamin J Strober2, Alessandro L Asoni2, Andrew S McCallion1 & Michael A Beer1,2

Most variants implicated in common human disease by genome-wide association studies (GWAS) lie in noncoding sequence intervals. Despite the suggestion that regulatory element disruption represents a common theme, identifying causal risk variants within implicated genomic regions remains a major challenge. Here we present a new sequence-based computational method to predict the effect of regulatory variation, using a classifier (gkm-SVM) that encodes cell type–specific regulatory sequence vocabularies. The induced change in the gkm-SVM score, deltaSVM, quantifies the effect of variants. We show that deltaSVM accurately predicts the impact of SNPs on DNase I sensitivity in their native genomic contexts and accurately predicts the results of dense mutagenesis of several enhancers in reporter assays. Previously validated GWAS SNPs yield large deltaSVM scores, and we predict new risk-conferring SNPs for several autoimmune diseases. Thus, deltaSVM provides a powerful computational approach to systematically identify functional regulatory variants.

Sequence variation in DNA regulatory elements is hypothesized to contribute substantially to risk for common diseases. Variants associated with human disease by GWAS predominantly lie in noncoding genomic regions1 and occur within putative regulatory elements far more often than expected by chance2,3, suggesting that disruption of regulatory function is a common mechanism by which noncoding sequence variants contribute to human disease. Linkage disequilibrium (LD) and the absence of regulatory vocabularies complicate the discrimination of regulatory risk variants from other variation within disease-associated intervals. Therefore, there is a pressing need for methods to predict the impact of regulatory sequence variation, expediting targeted functional validation and the exploration of disease-implicated pathways. However, few formal computational methods have been developed to predict the impact of SNPs on regulatory element activity4,5.

Regulatory elements modulate the expression of their target genes through the direct binding of sequence-specific transcription factors6. Although consensus on the mechanisms of regulatory element activity is emerging, a predictive model is lacking that is capable of (i) specifying the cell types and environmental conditions under which an element would modulate the expression of its target gene(s) and (ii) describing how specific mutations to that sequence would influence its activity. Here we develop a computational model that addresses the latter: given a regulatory element active in a specific cell type, it computes the effect of a given DNA sequence variation within the element. When trained on a set of putative regulatory sequences, our established gapped k-mer support vector machine (gkm-SVM)7 identifies sequence features within these regulatory regions that determine their cell type–dependent activity. We then use this gkm-SVM to quantify the effect of sequence changes within regulatory elements via a metric we term deltaSVM (overview in Fig. 1). In this systematic, quantitative approach, we leverage high-quality catalogs of human regulatory elements, generated for DNase I hypersensitivity, distinctive histone modifications and transcription factor binding8,9.

For example, if a gkm-SVM is trained on DNase I–hypersensitive sites (DHSs), it identifies sequence features that determine chromatin accessibility in the corresponding cellular context. Our method is, however, blind to extant databases or binding motif data and consequently can uncover new motifs, combinatorial constraints and key accessory factors and can quantify the significance of their individual contributions to regulatory element activity.

RESULTS

Model training and validation

We previously demonstrated that a properly trained SVM can predict cell type–specific regulatory elements from primary genome sequence alone10–12. To test whether this SVM-based approach could be adapted to predict the functional consequences of sequence variation within regulatory elements, we first took advantage of a large set of dsQTLs (DNase I–sensitivity quantitative trait loci) identified in a collection of human lymphoblastoid cell lines (LCLs)13–15. These are SNPs within putative regulatory regions (marked by DNase I hypersensitivity) and are associated with altered DNase I sensitivity therein. We first trained a gkm-SVM on the top DHSs in the LCL GM12878 (ref. 8). The gkm-SVM produced a scoring function characterized by a set of weights quantifying the contribution of each possible 10-mer to a region’s DNase I sensitivity in GM12878 cells. We then can

1McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland, USA. 2Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA. 3Present address: Ludwig Institute for Cancer Research, University of California, San Diego, La Jolla, California, USA.
4These authors contributed equally to this work. Correspondence should be addressed to M.A.B. (mbeer@jh.edu) or A.S.M. (andy@jhmi.edu).

Received 12 February; accepted 8 May; published online 15 June 2015; doi:10.1038/ng.3331

NATURE GENETICS  VOLUME 47 | NUMBER 8 | AUGUST 2015  955
calculated deltaSVM, the predicted impact of any single-nucleotide variant (SNV) on chromatin accessibility in LCLs, by summing the change in weight between alleles for each of the ten 10-mers encompassing the SNV, as shown in Figure 2a for the dsQTL rs4953223 (ref. 13). In this case, the implicated SNP allele disrupts a nuclear factor (NF)-κB-binding site, which in our model reduced the strong positive contribution of several 10-mers to SVM score. Two neighboring SNPs contributed much smaller changes to the weights (as shown graphically in Fig. 2b), and the score of each allele was the sum of the weights across this region. Similarly, we can extend this method to indels and multiple substitutions by summing weights across all affected bases.

To systematically assess the ability of deltaSVM to predict the impact of SNPs on DNase I sensitivity, we examined the relationship between deltaSVM and the effect sizes β for the set of dsQTLs13. The correlation between deltaSVM and effect size for the 579 SNPs within 100 bp of a DHS was highly significant, with Pearson correlation coefficient \( C = 0.721 \) (t-distribution \( P = 7.68 \times 10^{-94} \)) (Fig. 2c). This correlation fell off rapidly with increasing distance from a DHS (Supplementary Fig. 1); thus, our analysis is consistent with local action of dsQTLs. However, if our predictions are accurate, deltaSVM analyses on non-dsQTL SNPs should yield low scores, thereby limiting false positive predictions. We chose a negative set of non-dsQTL SNPs 50 times larger than the dsQTL set with comparable levels of DNase I sensitivity, as there are typically 50–100 SNPs within a single LD block15. We generated the receiver operating characteristic (ROC) curve (Fig. 2d), plotting the true positive rate (true positives over positives, TP/P) against the false positive rate (false positives over negatives, FP/N), and the precision recall curve (Fig. 2e), plotting precision (true positives over predicted positives, PP/TP) against recall (TP/P), for our method (gkm-SVM deltaSVM) in comparison to four other methods4,5,10,16.

Here, as is typically the case for genomic predictions where the search space is large, the lower-left quadrant of the ROC curve, where the false positive rate is low, had the most dramatic effect on the accuracy (precision) of the predictions17. At a recall rate of 10%, 55.9% of the gkm-SVM predictions were accurate, ~5 times more than when using deltaSVM based on smaller 6-mers (kmer-SVM)10 (Fig. 2e); this difference arises because, although kmer-SVM can predict the effects of full regions very accurately by averaging many weights, the k-mer weights needed to evaluate individual SNPs are determined from a small set of support vectors and are noisy. By contrast, gkm-SVM greatly reduces the false positive rate by using much more statistically robust gapped k-mer weights18. Additionally, in comparison to conservation prediction (GERP score16) and two recently published methods integrating functional genomic data

**Figure 1** Overview of our deltaSVM method. Left, the first step in calculating deltaSVM is to train a gkm-SVM classifier using a positive training set of putative regulatory sequences (identified by DNase I hypersensitivity, for example) and a negative training set of matched negative-control sequences. The gkm-SVM generates a regulatory sequence vocabulary—a weighted list of all possible 10-mers, where each 10-mer receives an SVM weight that quantifies its contribution to the prediction of regulatory function. Right, after training, this regulatory sequence vocabulary can be used to score the predicted impact of any sequence variant on regulatory activity, as shown here for a single-nucleotide substitution in a melanocyte enhancer of Tyrp1. WT, wild type.

**Figure 2** deltaSVM can accurately predict SNPs associated with DNase I hypersensitivity. (a) An example of a deltaSVM calculation using a known dsQTL SNP (rs4953223). Ref, reference; alt, alternate. (b) The 10-mer gkm-SVM scores across the dsQTL locus containing rs4953223 are shown. Only the functional SNP produces dramatic changes in gkm-SVM score. (c) The effect sizes of the dsQTL SNPs from ref. 13 are well correlated with the deltaSVM scores for these variants. (d,e) deltaSVM predicts dsQTLs with far greater accuracy than existing methods. Discriminative powers are compared among various methods using a 50 times larger control SNP set. (d) ROC curve. (e) Precision recall curve.
Three key features contribute to the dramatically improved accuracy of our method. First, we train gkm-SVM on a set of regulatory elements whose activity is specific to the relevant cell type. Second, this large training set (with thousands of elements) includes both positive and negative elements to statistically determine the DNA sequence elements required for activity, rather than relying on the precise state of any specific regulatory element in a specific assay. Third, we identify a complete catalog of both positive and negative sequence features, as many SNPs contribute more to deltaSVM on the basis of what the variant changes to, rather than what the sequence was in the reference or assayed genome. In our discriminative approach, gkm-SVM identifies these negative sequence elements by their presence in the negative set and their absence in the positive set. This is critical to accurately assess the effect of variants, as many of our predicted functional SNPs modulate intermediate-strength binding sites.

Ultimately, we are interested in how a variant modulates the expression of its target gene(s). One hundred twenty-five of the 579 dsQTLs are also expression quantitative trait loci (eQTLs)\(^19\) (variants associated with differential gene expression), but some dsQTLs are anticorrelated with eQTLs\(^17\). Both classes of dsQTLs showed strong positive

---

### Figure 3

**deltaSVM correlations with dsQTL and eQTL effect sizes.** deltaSVM is strongly positively correlated with dsQTL effect size and is positively or negatively correlated with eQTL effect size, depending on the sign of the correlation for the dsQTL and eQTL. Degner *et al.*\(^13\) reported that 16% of dsQTLs were also eQTLs but that 30% of eQTL dsQTLs were anticorrelated with the expression change. Our predictions are consistent with this observation.

(a-c) deltaSVM is always positively correlated with dsQTL effect size (\(\beta\)) (a); however, because eQTL and dsQTL \(\beta\) values are anticorrelated 30% of the time (b), deltaSVM and eQTL \(\beta\) values are only correlated (positively or negatively) if we consider the activating dsQTLs (red) and repressive dsQTLs (blue) separately (c).

(d) Bases predicted to reduce the activity of functional regions (the 1% of bases with the most negative deltaSVM; red) are evolutionarily conserved relative to neutral bases (the 1% of bases with deltaSVM near zero; gray) or positive bases (the 1% of bases with the most positive deltaSVM; blue).

### Figure 4

**deltaSVM accurately predicts change in luciferase expression in targeted mutagenesis of Tyr and Tyrp1 mouse melanocyte enhancers.**

(a, b) Base-by-base evaluation of all possible substitutions as scored by deltaSVM for the Tyr (a) and Tyrp1 (b) enhancers. Black circles mark substitutions that were tested in luciferase assays. Orange bars show the positions of previously characterized binding sites.

(c, d) Correlation of deltaSVM prediction and observed normalized luciferase expression for the Tyr (c) and Tyrp1 (d) enhancers. Blue circles indicate previously tested binding site\(^20,21\). Error bars are 1 s.d. of the changes in luciferase expression (four biological replicates per variant).
correlation with deltaSVM (Fig. 3a–c). Thus, surprisingly but consistent with earlier analysis, we found that, as 22% of the dsQTLs became more accessible, they repressed target gene expression. We also analyzed the relationship between deltaSVM and evolutionary sequence conservation. Interestingly, although bases predicted to either reduce or increase DNase I sensitivity when mutated are more conserved than bases predicted to be neutral, we found that bases with negative deltaSVM were much more conserved than bases with positive deltaSVM (Fig. 3d and Supplementary Fig. 2).

gkm-SVM predicts functional impact of enhancer variants

To directly test the ability of our SVM-based approach to predict the functional consequence of sequence variation on enhancer activity, we first turned to the well-characterized enhancers of the mouse pigment-gene genes Tyr and Tyrp1 (refs. 20,21) (Fig. 4a,b). We trained a melanocyte-specific gkm-SVM on a large set of putative melanocyte enhancers marked by EP300 binding and monomethylation of histone H3 at lysine 4 (H3K4me1)12. We then scored all possible SNVs in the Tyr and Tyrp1 enhancers, selecting and synthesizing more than 40 SNVs across a range of deltaSVM scores, and tested each variant independently in luciferase reporter assays. For both enhancers, deltaSVM was strongly correlated with the observed difference in luciferase reporter activity between mutant and wild-type enhancer constructs (Pearson $C = 0.778$, $P < 2 \times 10^{-5}$ for Tyr and $C = 0.529$, $P < 0.0095$ for Tyrp1; Fig. 4c,d).

Despite their depth, our analyses of the Tyr and Tyrp1 enhancers tested only a subset of all possible variants therein and relied on in vitro reporter assays. Therefore, we turned to a data set in which all possible variants within a 259-bp liver-specific enhancer of the ALDOB gene were tested using a massively parallel reporter assay in vivo in mouse liver22. We trained a gkm-SVM on a large set of putative liver enhancers marked by DNase I hypersensitivity and H3K4me1 signal in adult mouse liver22. We then compared the deltaSVM for each of the tested mutant regions to the observed functional output. Again, we saw a very high correlation ($C = 0.630$, $P < 3.24 \times 10^{-81}$) between the predicted impact of mutation using our sequence-based model and the observed change in enhancer activity relative to wild-type sequence (Fig. 5a). If we further used the ‘aggregate score’ model22, averaging the deltaSVM values for each of the three possible base substitutions, the correlation reached $C = 0.691$ (Supplementary Fig. 3).

We next asked how deltaSVM performed in predicting functional variants in diverse sets of enhancers. We analyzed data from another massively parallel reporter assay using targeted mutation of enhancers predicted to be active in K562 and HepG2 cells24. For each wild-type construct that was significantly expressed in either cell line, we separately scored all 1-bp and motif-scrubbling mutations using a gkm-SVM trained on K562 and HepG2 DHS regions8, and we compared the measured expression change to the predicted deltaSVM score in each cell line (Fig. 5b,c). For both data sets, we again found high correlation ($C = 0.626$, $P < 1.34 \times 10^{-31}$ for K562 cells and $C = 0.646$, $P < 3.84 \times 10^{-34}$ for HepG2 cells). Because all elements were tested and scored in both cell types, this high correlation underscores the accuracy of deltaSVM’s cell-type-specific predictions and is further supported by the low correlation of deltaSVM scores from gkm-SVMs trained on non-relevant cell types (Table 1).

Table 1. Cell-type specificity of deltaSVM predictions

| gkm-SVM       | LCL dsQTL | Tyr | Tyrp1 | ALDOB | K562 enhancers | HepG2 enhancers |
|---------------|-----------|-----|-------|-------|----------------|----------------|
| GM12878 DHSs  | 0.721 (7.68 $\times 10^{-5}$) | 0.302 (0.172) | 0.117 (0.595) | 0.112 (0.00256) | 0.204 (0.00062) | 0.201 (0.00076) |
| Mouse melanocyte EP300 sites | 0.245 (2.19 $\times 10^{-5}$) | 0.78 (2.0 $\times 10^{-5}$) | 0.53 (0.0095) | 0.147 (7.42 $\times 10^{-5}$) | 0.204 (0.00062) | 0.194 (0.00116) |
| Mouse liver DHSs | 0.131 (0.00157) | 0.282 (0.203) | 0.056 (0.798) | 0.630 (3.24 $\times 10^{-4}$) | −0.329 (2.04 $\times 10^{-6}$) | 0.551 (2.07 $\times 10^{-23}$) |
| K562 DHSs | 0.581 (1.45 $\times 10^{-53}$) | 0.390 (0.0726) | 0.104 (0.638) | 0.092 (0.0137) | 0.626 (1.34 $\times 10^{-31}$) | −0.042 (0.483) |
| HepG2 DHSs | 0.518 (3.84 $\times 10^{-4}$) | 0.551 (0.00791) | 0.166 (0.450) | 0.547 (1.01 $\times 10^{-57}$) | −0.184 (0.0021) | 0.646 (3.84 $\times 10^{-24}$) |

We compare the Pearson correlation coefficients between deltaSVM and expression change using weights trained on all five cell types and expression measurements in all five cell lines. Values are in parentheses, and the most significant row in each column is highlighted in bold. Although some transcription factors are active in more than one cell type, we generally observe that deltaSVM predictions are cell-type specific.
lead SNPs and an additional 2,700 SNPs in tight LD (as defined in the Online Methods) as well as random SNPs, including equivalently sized flanking sets as a control. An example locus in BACH2, associated with several autoimmune diseases, is shown in Figure 6b. We identified SNPs with high deltaSVM scores for 17 independent disease associations (Table 2), which we predict to be expression-perturbing SNPs with high confidence ($P < 0.02$), whereas at this threshold random sampling produced 8 SNPs (Fig. 6c and Supplementary Fig. 4). Most of these high-scoring SNPs were not the lead SNP and thus represent new predictions for the causal SNP.

**DISCUSSION**

One of the greatest challenges facing contemporary genetics and precision medicine is the interpretation of noncoding sequence variation. This challenge remains in the face of substantial advances in genome sequence technologies that now make possible the generation of huge quantities of noncoding sequence data. Despite the wealth of evidence linking noncoding sequence variation to disease risk, it is not yet possible to directly interpret primary noncoding sequence data and, hence, to infer the biological consequences of disease-associated variation. Our efforts in this study take major strides in creating a framework within which regulatory lexicons can be computationally

**Figure 6** deltaSVM only identifies validated causal SNPs when trained on the appropriate cell type. (a) Three validated GWAS SNPs from RFX6 (first column), BCL11A (second column) and SORT1 (third column) and flanking negative SNPs were each scored with deltaSVM trained on all three relevant cell types individually. The validated SNPs are properly identified from among flanking SNPs when trained on the appropriate cell type (red) but not when trained on other cell types (blue). (b,c) Scoring autoimmune GWAS loci with deltaSVM trained on T41 sites yielded the high-confidence causal SNPs listed in Table 2. (b) The BACH2 locus is shown as an example. (c) deltaSVM identified 17 significant independent associations, whereas 8 were expected by chance (binomial test, $P < 0.004$). Error bars are 1 s.d. of the expected binomial distribution.
Table 2 deltaSVM-predicted causal autoimmune SNPs

| Disease                          | Lead SNP    | Predicted SNP  | Chr. | Lead SNP position | Predicted SNP position | Distance (bp) | Nearest gene | deltaSVM | Ref. |
|---------------------------------|-------------|----------------|------|-------------------|------------------------|---------------|--------------|----------|------|
| Vitiligo                        | rs853308    | rs860475       | 8    | 133,929,917       | 133,929,799            | −118          | TG          | 16.797   | 29   |
| Type 1 diabetes                 | rs16872571  | rs4697651      | 4    | 10,726,853        | 10,721,433             | −5,420        | GLIS3       | 13.591   | 30   |
| Type 1 diabetes                 | rs7020673   | rs4380994      | 9    | 4,291,747         | 4,282,536              | −9,211        | CDKAL1      | 13.177   | 31,32|
| Crohn’s disease                 | rs6908425   | rs7748720      | 6    | 20,728,731        | 20,689,945             | −38,786       | ZMIZ1       | 11.039   | 33   |
| Multiple sclerosis              | rs1782645   | rs1250568      | 10   | 81,048,611        | 81,045,280             | −3,331        | ZMIZ1       | 11.039   | 34   |
| Celiac disease                  | rs1250552   | rs1250568      | 10   | 81,058,027        | 81,045,280             | −12,747       | ZMIZ1       | 11.039   | 33   |
| Crohn’s disease                 | rs10801047  | rs6665749      | 1    | 191,559,356       | 191,577,594            | 18,238        | NAB1        | 10.646   | 35   |
| Crohn’s disease                 | rs2797685   | rs2797685      | 1    | 7,879,063         | 7,879,063              | 0             | PER3        | 10.411   | 31   |
| Allergy                         | rs962993-T  | rs962993       | 10   | 9,053,132         | 9,053,132              | 0             | NAB1        | 10.008   | 36   |
| Primary biliary cirrhosis       | rs10931468  | rs3771317      | 2    | 191,538,562       | 191,543,962            | 5,400         | BACH2       | 9.787    | 38   |
| Celiac disease                  | rs7753008   | rs72928038     | 6    | 90,809,639        | 90,976,768             | 167,129       | BACH2       | 9.787    | 37   |
| Rheumatoid arthritis            | rs72928038  | rs72928038     | 6    | 90,976,768        | 90,976,768             | 0             | BACH2       | 9.787    | 33,39,40|
| Autoimmune thyroid disease      | rs564976    | rs485789       | 3    | 159,729,059       | 159,730,148            | 1,089         | AK097161    | 8.782    |      |
| Multiple sclerosis              | rs733724    | rs733724       | 6    | 105,223,864       | 105,223,864            | 0             | HACE1       | 7.360    | 41   |
| Primary biliary cirrhosis       | rs3024921   | rs3024921      | 2    | 191,943,272       | 191,943,272            | 0             | STAT4       | 6.931    | 42   |

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

To ensure very high-confidence predictions, we have limited our initial analyses to the highest deltaSVM scores in Table 2. However, comparison with validated SNPs (Fig. 6a) shows that many SNPs with more moderate deltaSVM scores will also perturb regulatory activity but will likely do so with relatively diminished effect. In this sense, our random control sampling is highly conservative, as the positive loci are all known to be associated with disease. The high accuracy and low false positive rate of deltaSVM (Fig. 2e) is crucial to identifying these causal SNPs with high accuracy. Together with our observations at lymphoblastoid dsQTLs and our analyses of the Tfr and Tfrp1 enhancers, the ALDOB enhancer, and those in K562 and HepG2 cells, these results clearly demonstrate that deltaSVM can broadly predict the empirically measured, cell type–specific functional consequences of enhancer sequence variants, given an appropriate training substrate.

Our results strongly support the hypothesis that noncoding disease-associated SNPs that disrupt DHS or enhancer function do so directly through modulation of local transcription factor–DNA interactions, leading to concomitant changes in chromatin state and gene expression. Many of these sequence determinants are recognizable as transcription factor binding sites. Additionally, gkm-SVM is also in principle capable of capturing the sequence determinants of structural properties of transcription factor–DNA interactions, for example, constrained sequence flanking transcription factor binding sites that may contribute to the activity or stability of the enhancer-promoter regulatory complex. Precise variant evaluation requires an accurate assessment of the relative contributions of moderate-strength and weakly binding sites or other variants that affect chromatin accessibility, which we estimate requires at least ~1,000 training elements and a robust classifier. We have shown that these sequence features are robust predictors of chromatin accessibility and are also predictive of gene expression changes when enhancer-promoter connections are established. But chromatin accessibility is sometimes negatively correlated with gene expression changes, and our results suggest that enhancer-enhancer and enhancer-promoter interactions on a larger scale will ultimately determine a sequence variant’s impact on gene expression.

In their application, both gkm-SVM and deltaSVM reinforce the recognized importance of cell type, developmental timing and biological state when connecting disease mechanisms to molecular events. Here the biological state defines a set of active nuclear transcription factors in the cell, which in turn map sequence features to their target regulatory elements through combinatorial binding. It is understandable, therefore, that the predictive power of deltaSVM for a given variant implicated in any given disease process is heavily dependent on the availability of a biologically appropriate substrate on which to train the gkm-SVM. Consistent with this expectation, we have demonstrated that deltaSVM predictions are highly cell-type dependent: that is, deltaSVM from weights trained on one cell type are weak predictors of expression changes in other cell types (Table 1). Further, deltaSVM only identified the validated disease-associated SNPs shown in Figure 6a if trained on an appropriate cell type. Although the Encyclopedia of DNA Elements (ENCODE)8,23 and Epigenomics Roadmap9 projects have provided a wealth of such training data, our results indicate that future progress in common disease etiology will be greatly facilitated by coupling sequence-based computational analysis with the generation of functional genomics data targeting disease-relevant developmental stages and cell types. What we provide here is evidence that such integration of computational and disease-informed biological strategies can be used to illuminate the roles of noncoding regulatory variation in disease.

URLs. All gkm-SVM weights and source code used in this study are available at http://www.beerlab.org/deltasvm/. Combined Annotation-Dependent Depletion (CADD), http://cadd.gs.washington.edu/; Genome-Wide Annotation of Variants (GWAVA), ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/; eQTL data, http://eqtl.uchicago.edu/RNA_Seq_data/results/.

METHODS

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

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.
ACKNOWLEDGMENTS

This research was supported in part by US National Institutes of Health grant R01 NS62972 to A.S.M. and by grant R01 HG007348 to M.A.B.

AUTHOR CONTRIBUTIONS

M.A.B., A.S.M., D.L. and D.U.G. designed the study and wrote the manuscript. D.U.G. and M.B. performed the experiments and analyzed the data. D.L. and M.A.B. developed the computational algorithms and analyzed the data. B.J.S. and A.L.A. contributed computational analysis. D.L. and D.U.G. contributed equally to this work.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. 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).
2. Maurore, M.T. et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–1195 (2012).
3. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M.A. Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput. Biol. 10, e1003711 (2014).
4. ENCODE Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).

18. Ghandi, M., Mohammad-Noori, M. & Beer, M.A. Robust k-mer frequency estimation using gapped k-mers. J. Math. Biol. 69, 469–500 (2014).
19. Pickrell, J.K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464, 768–772 (2010).
20. Muriisier, F., Guichard, S. & Beermann, F. A conserved transcriptional enhancer that specifies Tip61 expression to melanocytes. Dev. Biol. 298, 644–655 (2006).
21. Muriisier, F., Guichard, S. & Beermann, F. The Tip61 enhancer is activated by Sox10 and Miff in mouse melanocytes. Pigment Cell Res. 20, 173–184 (2007).
22. Patwardhan, R.P. et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. 30, 265–270 (2012).
23. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature 515, 355–364 (2014).
24. Kheradpour, P. et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. 23, 800–811 (2013).
25. Huang, Q. et al. A prostate cancer susceptibility allele at 6q22 increases RXF6 expression by modulating HOX813 chromatin binding. Nat. Genet. 46, 126–135 (2014).
26. Bauer, D.E. et al. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science 342, 253–257 (2013).
27. Musunuru, K. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466, 714–719 (2010).
28. Farh, K.K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343 (2015).
29. Jin, Y. et al. Genome-wide association analyses identify 13 new susceptibility loci for Crohn’s disease susceptibility. Proc. Natl. Acad. Sci. USA 109, 1045–1048 (2012).
30. Barrett, J.C. et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. 41, 703–707 (2009).
31. Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn’s disease susceptibility loci. Nat. Genet. 42, 1118–1125 (2010).
32. Bartlett, J.C. et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn’s disease. Nat. Genet. 40, 955–962 (2008).
33. International Multiple Sclerosis Genetics Consortium. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat. Genet. 45, 1353–1360 (2013).
34. Dubois, P.C.A. et al. Multiple common variants for celiac disease influencing immune gene expression. Nat. Genet. 42, 295–302 (2010).
35. Parkes, M. et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn’s disease susceptibility. Nat. Genet. 39, 830–832 (2007).
36. Hinds, D.A. et al. A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. Nat. Genet. 45, 907–911 (2013).
37. Mells, G.F. et al. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. Nat. Genet. 43, 329–332 (2011).
38. Trynka, G. et al. Denise genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat. Genet. 43, 1193–1201 (2011).
39. Eye, S. et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat. Genet. 44, 1336–1340 (2012).
40. Cooper, J.D. et al. Seven newly identified loci for autoimmune thyroid disease. Hum. Mol. Genet. 21, 5202–5208 (2012).
41. Gounaud, P.-A. et al. A genome-wide association study of brain lesion distribution in multiple sclerosis. Brain 136, 1012–1024 (2013).
42. Liu, L.Z. et al. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. Nat. Genet. 44, 1137–1141 (2012).
ONLINE METHODS
gkm-SVM and deltaSVM. We trained a gkm-SVM by following previously reported methods with minor modifications10–12. Briefly, we first defined a positive training set using publically available DNase I sensitivity (DNase I–seq) and chromatin immunoprecipitation and sequencing (ChIP-seq) data sets. We then generated a negative training set by randomly sampling from the genome an equal number of regions that matched the length, GC content and repeat fraction of the positive training set. To remove false negative regions as much as possible, we excluded any regions with \( P < 1 \times 10^{-5} \) (MACS13) from sampling. We then trained a gkm-SVM with default parameters (word length \( l = 10 \), informative columns \( k = 6 \) and truncated filter \( d = 3 \)) and measured the classification performance using ROC curves with fivefold cross-validation. Scaling of performance with gkm-SVM feature length is shown in Supplementary Figure 5. To calculate deltaSVM, 10-mer SVM scores were used as a proxy for weights. We generated final weights by averaging gkm-SVMs trained on five independently generated equally sized negative sets. When we compared deltaSVM between different training sets, we normalized weights by the standard deviation of the weight distribution, but we have reported raw weights here for simplicity. This correction typically had a small effect (<30%).

Training set for DNase I–hypersensitive regions in lymphoblastoid cell lines. GM12878 DNase I–seq peaks were first defined by MACS13 (\( P < 1 \times 10^{-9} \)) for each replicate independently. We then chose peaks that were consistently found in both replicates. These peaks were further trimmed, and 300-bp central DHSs that maximized the DNase I–hypersensitive signals were determined. We also excluded any regions with >70% repeats and regions overlapping with dsQTLs, to avoid possible overfitting when scoring dsQTLs. We ultimately obtained 22,384 300 bp of DHSs as the positive training set.

Training set for mouse melanocyte enhancers. To train a gkm-SVM appropriate for the Tyr and Tyrp1 enhancers in mouse melanocytes, we defined 4,337 EP300-bound regions in the mouse melanocyte cell line melan-Ink4a-Arf12 as the positive training set by following the above protocol with some adjustments (MACS, \( P < 0.002 \)). Promoter-proximal regions and repeats were excluded from the training set. Because this positive set was much smaller than the others, we generated negative training sets that were ten times larger to obtain more robust weights for deltaSVM analysis.

Training set for mouse liver enhancers. Similar to the training set for DHSs in LCLs, we defined a positive training set (\( n = 19,590 \)) relevant to the ALDOB enhancer by integrating DNase I–seq and H3K4me1 ChIP-seq data from adult mouse liver tissue13. To specify liver enhancers, we additionally excluded all promoter-proximal DHSs (defined as regions with a distance to the nearest known transcription start site (TSS) of <2 kb) from the training set, after determining the 300-bp core DHSs as described above. We further selected DHSs that overlapped with H3K4me1 ChIP-seq peaks, which are well-known markers for enhancer activity44,45, and defined these as the positive training set.

DeltaSVM analysis of dsQTL SNPs. We used dsQTL tables and raw data files downloaded from the Gene Expression Omnibus (GEO) database (accession GSE31388) to define the positive and control sets of dsQTL SNPs. Because association alone does not necessarily imply causation, owing in part to the LD problem, we further applied more stringent rules to determine the most likely causal dsQTL SNPs. We first restricted to 1,296 SNPs within their associated 100-bp DHSs to ensure that the changes in DNase I sensitivity were physically linked to changes in their DNA sequences. We also applied a more strict association \( P \)-value threshold (\( P < 1 \times 10^{-5} \)) to reduce false positive associations, finally resulting in 579 SNPs. As a control SNP set, we generated a 50 times larger set of common, randomly selected SNPs (\( n = 28,930 \); minor allele frequency > 5%) sampled only from the top 5% of DHS regions that had been used to identify dsQTLs in the previous study13. To reduce the incidence of false negative SNPs, we excluded from sampling any DHSs that had been found to be significantly associated with any of the dsQTL SNPs. Weights from a gkm-SVM and kmer-SVM trained on the GM12878 DHSs were then used to calculate deltaSVM scores. We confirmed that training a gkm-SVM on negative sequences constrained to match the distribution of distances to a TSS for the positive sequences did not affect overall performance (Supplementary Fig. 6).

We further confirmed that using negative dsQTL control SNPs constrained to match the dsQTL distance to a TSS and LD distribution of the positive set did not affect overall performance (Supplementary Fig. 7). For comparison, we considered three different scoring metrics: CADD4, GWAVA5 and conservation (GERP6) scores. We downloaded precomputed CADD scores for all 1000 Genomes Project variants, from which the scores for the dsQTLs and control SNPs were extracted. We also extracted the corresponding GWAVA scores from the precalculated table downloaded from the GWAVA website. We analyzed all three different GWAVA models (region, TSS and unmatched) and chose the best one (region), as determined by the area under the ROC curve (AUC), for the main analysis. The GERP scores were extracted from the same GWAVA result files. To carry out a fair comparison, we only considered SNPs for which all 5 scores were available, resulting in 574 positive SNPs and 27,735 control SNPs. The entire prediction results are available in Supplementary Table 1. eQTL \( \beta \) values were calculated using quantile-normalized gene expression from the eQTL website.

Melanocyte luciferase assays and deltaSVM analysis. We selected 22 and 23 SNVs for functional testing in the Tyr (mm10 coordinates, chr. 7: 87,508,164–87,508,388; 226 bp) and Tyrp1 (mm10 coordinates, chr. 4: 80,819,561–80,819,851; 291 bp) enhancers, respectively. These SNVs were randomly selected as follows: ten SNVs in each enhancer predicted to reduce the enhancer’s activity (negative dsSVM), four SNVs in each enhancer predicted to increase the enhancer’s activity (positive deltaSVM), four SNVs in each enhancer predicted to have a neutral impact on the enhancer’s activity (deltaSVM near 0), and four (Tyr) and five (Tyrp1) additional SNVs that overlapped key motifs identified in previous reports20,21. Reference and SNV enhancer sequences were synthesized (Genewiz), verified by Sanger sequencing and cloned into a luciferase reporter plasmid containing a minimal promoter and a luciferase reporter gene. For each SNP, we performed four biological replicate assays (each with an independent plasmid DNA clone) to control for differences that might arise from random mutations in the plasmid backbone or from variation in the quality of the plasmid preps. We transfected each reporter plasmid into the mouse melanocyte cell line melan-Ink4a-Arf (not tested for mycoplasma) and measured luciferase activity 24 h later using the Dual-Luciferase Reporter Assay System (Promega). We compared the activity of each variant enhancer sequence to the activity of the reference sequence (normalized to 1) and were thus able to quantify the impact of each SNP on the enhancer’s activity. The complete deltaSVM predictions and luciferase assay results are provided in Supplementary Tables 2 and 3, respectively.

deltaSVM analysis of massively parallel reporter assays. To compare deltaSVM with exhaustive single-nucleotide mutagenesis of the ALDOB enhancer22, we trained a gkm-SVM on adult mouse liver DHSs, scored each SNP with deltaSVM and compared the resulting deltaSVM values with measured changes in expression in vivo (Supplementary Table 4)22. To compare with the directed mutagenesis of putative K562 and HepG2 enhancers23, we trained K562- and HepG2-specific gkm-SVMs on the top 10,000 500-bp DHS regions in K562 and HepG2 cells23, after excluding regions that were DHSs in more than 30% of human ENCODE cell lines or were near promoters (<2 kb from a TSS), and against an equally sized GC content– and repeat fraction–matched training set. We compared the deltaSVM and expression change for each pair of mutant and wild-type constructs for all wild-type constructs significantly expressed in either cell line (mean normalized expression > 3.5), which yielded 175 wild-type constructs and 277 mutant constructs: 102 of the variants were single-base mutations and 175 were motif scrambling (8–17 bp changed) (Supplementary Table 5). For the motif-scrambling mutations, we summed all 10-mers scores spanning the mutated motif.

Training set for validated enhancers. For each appropriate cell line, we trained on the top 10,000 500-bp DHS regions, after excluding regions that were DHSs in more than 30% of human or mouse ENCODE cell lines and tissues or were near promoters (<2 kb from a TSS), and against an equally sized GC content– and repeat fraction–matched training set. The cell lines chosen were human LNCaP cells8 for RFX6, mouse erythroleukemia (MEL)23 cells for BCL11A and human HepG2 cells8 for SORT1.
Scoring of autoimmune-related variants. We selected 11 autoimmune traits enriched in T_{H1} H3K27ac marks as shown in Figure 3 of ref. 28. We made predictions for 413 lead SNPs associated with 11 autoimmune diseases enriched in T_{H1} H3K27ac regions (type 1 diabetes, Crohn’s disease, multiple sclerosis, celiac disease, primary biliary cirrhosis, rheumatoid arthritis, autoimmune thyroid disease, ulcerative colitis, vitiligo and systemic lupus erythematosus)\textsuperscript{28}. We trained a gkm-SVM on the top 10,000 500-bp T_{H1} DHS regions, after excluding regions that were DHSs in more than 30% of human ENCODE cell lines or were near promoters (<2 kb from a TSS), and against an equally sized GC content– and repeat fraction–matched training set. We scored the lead SNP and all flanking off-lead candidates in LD (defined by $r^2 > 0.5$ and PICS\textsuperscript{28} probability > 0.0275), yielding 3,113 total SNPs. Because the significance of the maximum deltaSVM score in a locus will depend on the number of SNPs in that locus, as a control, we scored randomly selected SNPs and equally sized flanking sets. To determine the cutoff, we first determined the second-percentile deltaSVM score from 10,000 random permutations for each number of flanking SNPs (1–30), and then calculated the mean and standard deviation of the 100 repeated experiments as the final cutoff. We identified 17 high-scoring deltaSVM SNPs that we predict to be expression-perturbing SNPs with high confidence ($P < 0.02$), whereas at this threshold random sampling produced 8 SNPs (binomial test $P < 0.004$; Supplementary Fig. 4). deltaSVM scores for all 3,113 SNPs are provided in Supplementary Table 6.

43. Zhang, Y. et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 9, R137 (2008).
44. Heintzman, N.D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. 39, 311–318 (2007).
45. Heintzman, N.D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459, 108–112 (2009).