Human Enhancers Harboring Specific Sequence Composition, Activity, and Genome Organization Are Linked to the Immune Response

Charles-Henri Lecellier,*,†,‡,1 Wyeth W. Wasserman,‡ and Anthony Mathelier‡,§,,1**

*Institut de Génétique Moléculaire de Montpellier, University of Montpellier, Centre National de la Recherche Scientifique (CNRS), 34293 Montpellier cedex5, France, †Institut de Biologie Computationnelle, 34095 Montpellier, France, ‡Centre for Molecular Medicine and Therapeutics at the Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, Vancouver, V5Z 4H4, Canada, §Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, Faculty of Medicine, University of Oslo, 0349 Oslo, Norway, and **Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0372 Oslo, Norway

ORCID IDs: 0000-0002-0229-5434 (C.-H.L.); 0000-0001-5127-5459 (A.M.)

ABSTRACT The FANTOM5 consortium recently characterized 65,423 human enhancers from 1829 cell and tissue samples using the Cap Analysis of Gene Expression technology. We showed that the guanine and cytosine content at enhancer regions distinguishes two classes of enhancers harboring distinct DNA structural properties at flanking regions. A functional analysis of their predicted gene targets highlighted one class of enhancers as significantly enriched for associations with immune response genes. Moreover, these enhancers were specifically enriched for regulatory motifs recognized by transcription factors involved in immune response. We observed that enhancers enriched for links to immune response genes were more cell-type specific, preferentially activated upon bacterial infection, and with specific response activity. Looking at chromatin capture data, we found that the two classes of enhancers were lying in distinct topologically associating domains and chromatin loops. Our results suggest that specific nucleotide compositions encode for classes of enhancers that are functionally distinct and specifically organized in the human genome.

KEYWORDS transcription; enhancer; sequence-level instructions; immune response

GENE expression is regulated through many layers, one of which being the regulation of the transcription of DNA segments into RNA. Transcription factors (TFs) are key proteins regulating this process through their specific binding to the DNA at regulatory elements, the TF binding sites (TFBSs) (Wasserman and Sandelin 2004). These regulatory elements are located within larger regulatory regions such as promoters and enhancers (Mathelier *et al.* 2015b). While promoters are situated around transcription start sites (TSSs), enhancers are distal to the genes they regulate. The canonical view is that chromatin conformation places enhancers in close 3D proximity to their target gene promoters through DNA looping (Visel *et al.* 2009; Andersson *et al.* 2015; Babu and Fullwood 2015). The high-resolution chromatin conformation capture (Hi-C) technology maps genomic regions in spatial proximity within cell nuclei (Lieberman-Aiden *et al.* 2009). It identifies specific genomic neighborhoods of chromatin interactions, the topologically associating domains (TADs), which represent stable chromatin compartments between cell types and conserved across species (Dixon *et al.* 2012, 2016).

Studies have shown relationships between the composition of a DNA sequence in guanine (G) and cytosine (C) and chromatin organization, for instance in relation to nucleosome positioning (Hughes and Rando 2009; Tillo and Hughes 2009) and chromatin architecture (Jabbari and Bernardi 2017). DNA sequence composition and other features of promoter regions, such as CpG islands, have been extensively studied. The Cap Analysis of Gene Expression (CAGE)

doi: https://doi.org/10.1534/genetics.118.301116
Manuscript received December 18, 2018; accepted for publication June 1, 2018; published Early Online May 6, 2018.
Supplemental material available at Zenodo: https://doi.org/10.5281/zenodo.1283306.
[1]Corresponding authors: AM: Centre for Molecular Medicine Norway (NCMM), University of, Gaustadalléen 21, Oslo 0349, Norway; CHL: University of Montpellier, Centre National de la Recherche Scientifique (CNRS), 34294 Montpellier cedex5, France. E-mail: anthony.mathelier@ncmm.uio.no; and charles.lecellier@igmm.cnrs.fr

technology (Shiraki *et al.* 2003; Kodzius *et al.* 2006), which identifies active TSSs in a high-throughput manner based on 5′ capped RNA isolation, accelerated our capacity to analyze human promoters. Using CAGE data, a large-scale identification of the precise location of TSSs in human (Carninci *et al.* 2005) led to the classification of promoters into four classes, based on G+C content (%GC) (Bajic *et al.* 2006). GC-rich promoters are associated with genes involved in various binding and protein transport activities and GC-poor promoters with genes responsible for environmental defense responses. While promoters overlapping CpG islands are commonly assumed to be ubiquitous drivers of housekeeping genes, comprehensive analysis of CAGE data from >900 human samples showed that a subset deliver cell-type-specific expression [FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* 2014].

Large-scale computational analyses of enhancer regions have been hampered by a limited set of *bona fide* enhancers. The CAGE technology can identify *in vivo*-transcribed enhancers. Specifically, it identifies active enhancer regions in biological samples by capturing bidirectional RNA transcripts at enhancer boundaries (Andersson *et al.* 2014). Using this characteristic of CAGE data, the FANTOM5 project identified 65,423 human enhancers across 1829 CAGE libraries [Andersson *et al.* 2014; FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* 2014; Arner *et al.* 2015]. Sequence property analysis suggested that the enhancers share properties with CpG-poor promoters (Andersson *et al.* 2014).

As enhancers are distal to the genes they regulate, it is challenging to predict these relationships. Based on cross-tissue correlations between histone modifications at enhancers and CAGE-derived expression at promoters within 1000 bp, enhancer–promoter links have been shown to be conserved across cell types (O'Connor and Bailey 2015). As the CAGE technology captures the level of activity for both promoters and enhancers in the same samples, recent studies predicted enhancer targets by correlating the activity levels of these regulatory regions over hundreds of human samples (Andersson *et al.* 2014; Cao *et al.* 2017). Enhancer–gene associations were supported by experimental data from ChIA-PET and Hi-C, and eQTL data (Andersson *et al.* 2014; Cao *et al.* 2017). Further, Andersson *et al.* (2014) unveiled that closely spaced enhancers were linked to genes involved in immune and defense responses. These results stress that predictions of enhancer–promoter associations are critical to decipher the functional roles of enhancers.

Here, we used the G+C content at human, CAGE-derived, enhancer regions to define two classes of enhancers. Based on the enhancer–gene target pairs characterized by both Andersson *et al.* (2014) and Cao *et al.* (2017), we showed that the class of enhancers with a lower G+C content was predicted to be functionally associated with genes involved in the immune response. Accordingly, regulatory motifs associated with immune response TFs like NF-κB are enriched in the DNA sequence of the immune response-related set of enhancers. Independent functional analysis of histone modification and CAGE data highlighted a cell-type specificity of these enhancers along with their activation upon bacterial infection. Moreover, the class of enhancers enriched for associations with immune system genes was observed with a distinct response activity pattern following cell stimulation in time-course data sets. Finally, the two classes of enhancers tended to be structurally organized within distinct TADs and DNA chromatin loops.

Materials and Methods

Human enhancers

We retrieved the hg19 positions of the 65,423 human enhancers from phases 1 and 2 of the FANTOM5 project in BED12 format from http://fantom.gsc.riken.jp/5/datafiles/phase2.2/extra/Enhancers/human_permissive_enhancers_phase_1_and_2.bed.gz along with the 1829 CAGE library ids from http://fantom.gsc.riken.jp/5/datafiles/phase2.2/extra/Enhancers/Human.sample_name2library_id.txt [Andersson *et al.* 2014; FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* 2014; Arner *et al.* 2015]. We extracted DNA sequences for regions of 1001 bp centered at the enhancer midpoints (columns 7–8 of the BED12 file) using the BEDTools (Quinlan and Hall 2010) to compute G+C contents. We considered the distribution of the G+C content of all the enhancers (mean $\sim$ 45%, median $\sim$ 44%, SD $\sim$ 8) to distinguish two classes of enhancers (%GC below the median, class 1; and %GC above the median, class 2; Figure 1a).

Clusterization of human enhancers based on positional distribution of G+C content

DNA sequences of 1001 bp centered at the enhancer midpoints were converted to binary vectors with 1 encoding G or C and 0 encoding A or T. Eleven enhancers were not considered as the considered sequences contained undefined nucleotides (N in the International Union of Pure and Applied Chemistry (IUPAC) notation). The vectors were clustered using the k-means algorithm implemented in the KMeans function of the *scikit* Python module (Pedregosa *et al.* 2011). The silhouette plots (Supplemental Material, Figure S1) were constructed for $k \in [2, 5]$ (*silhouette_samples* function of the *scikit* Python module). Formally, the silhouette plots display the silhouette coefficient for each enhancer as $(b - a)/\max(a, b)$ where a is the mean intra-cluster euclidean distance and b the mean nearest-cluster euclidean distance.

Distribution of enhancers in the human genome

The distribution of enhancers from the two classes in 3′ UTR, 5′ UTR, intergenic regions, transcription termination sites, intronic regions, noncoding and coding exons, and promoter regions in Figure S3a were obtained using the HOMER (v.4.7.2) *annotatePeaks.pl* script with annotations from the human genome hg19 v.5.4 (http://homer.ucsd.edu/homer/). Distances to TSSs for Figure S3b were obtained using the same script.

Repetitive elements

Repetitive elements coordinates (hg19) were retrieved from the RepeatMasker track of the University of California Santa Cruz (UCSC) Table browser tool (https://genome.ucsc.edu/cgi-bin/hgTables). The overlaps between enhancers and repetitive elements were obtained using the *intersect* subcommand of the BEDTools requiring a minimum overlap of 50% of the enhancer lengths.

Expression quantitative trait loci

The v6p GTEx *cis*-eQTL (expression quantitative trait loci) corresponding to eGene with significant SNP–gene associations based on permutations (GTEx_Analysis_v6p_eQTL.tar) were downloaded from the GTEx Portal at https://www.gtexportal.org/home/datasets. The *cis*-eQTL (all tissues combined) located in enhancers were retrieved with the *intersect* subcommand of the BEDTools. Enhancers were linked to genes following *cis*-eQTL variant–gene associations (2459 and 5857 genes for class 1 and class 2, respectively, Table S2). Four hundred thirty-seven genes were common to the two classes. Significance of the intersection was assessed with 1000 random assignments of the two classes to enhancers (with 32,487 and 32,936 enhancers in each class). No intersection of ≤ 437 potential target genes was obtained (empirical *P*-value $< 10^{-3}$).

Enhancer–gene targets

We considered two sets of enhancer–gene pairs derived from the correlation of CAGE signal at enhancers and (i) CAGE-derived or (ii) RefSeq-annotated TSSs [Andersson *et al.* (2014)] (http://enhancer.binf.ku.dk/presets/human.associations.hdr.txt.gz and http://enhancer.binf.ku.dk/presets/enhancer_tss_associations.bed). A third set of enhancer–gene pairs was obtained from Cao *et al.* (2017) where associations were obtained by first computing a lasso-based multiple regression between CAGE signals at each TSS and proximal enhancers from all FANTOM5 samples and then using sample-specific information to obtain sample-specific pairs (http://yiplab.cse.cuhk.edu.hk/jeme/fantom5_lasso.zip). To rely on high-quality enhancer–gene pairs, we considered the intersection of the pairs of enhancer–gene targets predicted by the three data sets.

We used these pairs for promoter sequence analyses. We extracted DNA sequences of ±500 bp around gene starts (ENSEMBL hg19 coordinates) using the BEDTools *getfasta* subcommand (Quinlan and Hall 2010) and computed average G+C content using the *GC* function of Biopython (Cock *et al.* 2009).

MNase profiles

MNase-seq signal from ENCODE for cell lines GM12878 and K562 were obtained as bigWig files from the UCSC genome browser website at http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeSydhNsome/wgEncodeSydhNsomeGm12878Sig.bigWig and http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeSydhNsome/wgEncodeSydhNsomeK562Sig.bigWig. Average MNase-seq signal values at enhancer regions were computed using the *agg* subcommand of the *bwtool* tool (Pohl and Beato 2014) with regions spanning ±2000 bp around the enhancers' midpoints.

DNA shape feature plots

The values of 13 DNA structural features were retrieved from the GBshape browser (Chiu *et al.* 2015) as bigwig files at ftp://rohslab.usc.edu/hg19/. We retrieved the averaged DNA shape values at the enhancer regions from class 1 and class 2 using the *agg* subcommand of the *bwtool* tool (Pohl and Beato 2014). The normalized averaged DNA shape values were computed independently for each enhancer class using the equation:

$$norm_{value} = (value - min_{value})/(max_{value} - min_{value})$$

where $norm_{value}$ is the normalized value to be computed for a DNA shape at a specific position in the DNA sequence, *value* is the averaged DNA shape value at this position for the enhancers in the class, and min_{value} (max_{value}) is the minimum (maximum) averaged DNA shape value for the enhancers in the class. The 90% confidence intervals for each DNA shape feature at each position was computed using a bootstrap approach. Specifically, random subsampling of enhancers was used to construct 100 sets of 10,000 randomly selected enhancers from classes 1 and 2. Average DNA shape values were computed for each random set, and values from the 5th and 95th percentages were used to define the 90% confidence intervals.

DNA sequence shuffling

The ±2000 bp DNA sequences around enhancer midpoints were shuffled with the *m* subcommand of the *BiasAway* tool to conserve mononucleotide composition (Worsley Hunt *et al.* 2014).

Gene ontology functional enrichment

To construct Figure 2a, official symbols corresponding to the promoters associated with enhancers from class 1 and class 2 (Table S3) were submitted to GOrilla (Eden *et al.* 2009) at http://cbl-gorilla.cs.technion.ac.il/ (January 7, 2017 update). We used the two unranked lists option with genes associated with enhancers from either class 1, class 2, specific to class 1, specific to class 2, or common to class 1 and class 2 as targets and the aggregated set of genes associated with the full set of enhancers as background. We submitted the enriched gene ontology (GO) biological processes with FDR < 0.01 to REViGO (Supek *et al.* 2011) at http://revigo.irb.hr/ asking for a "small" output list. GOrilla outputs are provided in Tables S4 and S5.

To construct Figure 2b, genes were ranked based on the number of enhancers predicted to target them in decreasing order (Tables S6 and S7). This ranked list was submitted to

GOrilla using the ranked list option. Enriched GO terms were retrieved as described above (Table S8).

Motif enrichment

We applied Centrimo (Bailey and Machanick 2012) (MEME suite version 4.11.1) with default parameters to 1001-bp-long DNA sequences around the midpoints of enhancers. Class 1 enhancer regions were used as foreground and class 2 enhancer regions as background and vice versa. The MEME databases of motifs considered for enrichment were derived from Jolma *et al.* (2013) (jolma2013.meme), JASPAR (Mathelier *et al.* 2015a) (JASPAR_CORE_2016_vertebrates.meme), Cis-BP (Weirauch *et al.* 2014) (Homo_sapiens.meme), Swiss Regulon (Pachkov *et al.* 2013) (Swiss_Regulon_human_and_mouse.meme), and HOCOMOCO (Kulakovskiy *et al.* 2016) (HOCOMOCOv10_HUMAN_mono_meme_format.meme). The same procedure was applied to promoter regions (±1001 bp around gene starts) associated with class 1 and class 2 enhancers.

Figure 3, a and b has been obtained from the html output of Centrimo by selecting the three most enriched motifs (ranked using the Fisher *E*-value; Data S1 and S2, https://doi.org/10.5281/zenodo.1283306).

Genome segmentation

ENCODE genome segmentation: The genome segmentation from combined ChromHMM (Ernst and Kellis 2012) and Segway (Hoffman *et al.* 2012) for ENCODE tier 1 and tier 2 cell types GM12878, H1hesc, HelaS3, Hep G2, HUVEC, and K562 were retrieved from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/.

Genome segmentation in dendritic cells: The genome segmentation of dentritic cells before and after *Mycobacterium tuberculosis* infection (Pacis *et al.* 2015) was computed using ChromHMM (Ernst and Kellis 2012) and retrieved at http://132.219.138.157:8080/DC_NI_7_segments_modID.bed.gz and http://132.219.138.157:8080/DC_MTB_7_segments_modID.bed.gz.

Genome segmentation overlap with enhancers: The overlaps between enhancers and genome segments were obtained using the *intersect* subcommand of the BEDTools requiring a minimum overlap of 50% of the enhancer lengths. We considered enhancers as in active states if they overlapped the TSS, promoter flank, enhancer, weak enhancer, and transcribed segments.

RELA ChIP-seq data analyses

The ENCODE RELA ChIP-seq data in GM12878 cells was retrieved at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/wgEncodeAwgTfbsSydhGm12878NfkbTnfaIggrabUniPk.narrowPeak.gz. To identify active FANTOM5 enhancers in GM12878, we considered the overlap between 1001-bp-long regions around enhancers' midpoints and genome segments predicted by ChromHMM and Segway combined as enhancer or weak enhancer. The identified 1001-bp-long active enhancer regions were overlapped with RELA ChIP-seq peaks with the *intersect* subcommand of the BEDTools.

Enhancer expression specificity

The cell-type expression specificity of enhancers was computed as

$$1-\left(\frac{\text{entropy}(\text{enhancer expression})}{\log_2(\text{number of cell types})}\right)$$

in Andersson *et al.* (2014). Each enhancer expression was represented by a vector of expression values in each cell type, which corresponded to the mean of the enhancer expression in the samples associated with the cell types. The binary matrix of enhancer usage across FANTOM5 samples was obtained at http://enhancer.binf.ku.dk/presets/hg19_permissive_enhancer_usage.csv.gz. The association between FANTOM5 samples and cell types was obtained from Tables S10 and S11 in Andersson *et al.* (2014). Heat maps in Figure 6 were computed using the *colormesh* function of the *matplotlib.pyplot* Python module (Hunter 2007).

Enhancer dynamics

FANTOM5 classification in the 14 dynamics displayed in Figure 8 was obtained from Auxiliary data table S3 in Arner *et al.* (2015). The classification provided response class assignments to 1294 and 2800 class 1 and class 2 enhancers, respectively. Response classes were assigned to 2827 and 4406 genes associated with class 1 and class 2 enhancers, respectively. Note that enhancers and promoters can be assigned to multiple response classes.

Corresponding plots (Figure 8) and enrichment analyses were performed using *pandas* Python data structure (McKinney 2010) and the *scipy* Python library (http://www.scipy.org/) in the *IPython* environment (Perez and Granger 2007).

Chromatin conformation data

The enrichment for enhancers associated with a specific class in each TAD or chromatin domain (see below) was computed using Binomial test *P*-values as implemented by the *binom.test* function in the *R* environment (R Core Team 2016). As a control, we randomly assigned the labels class 1 and class 2 to the enhancers and computed the corresponding Binomial test *P*-values; this procedure was applied to 1000 random trials. Density plots were obtained using the *density* function in the *R* environment with the *adjust* parameter set to 0.5.

Topologically associating domains: As TADs have been shown to be conserved between cell types and species, we retrieved the TADs defined in the first study describing them (Dixon *et al.* 2012). The TADs were predicted in mouse embryonic stem cells and we used the *liftOver* tool from the

UCSC genome browser at https://genome.ucsc.edu/cgi-bin/hgLiftOver to map them to hg19 coordinates.

Chromatin loops: The positions of the chromatin loops computed with the HICCUPS tools (Rao *et al.* 2014) from Hi-C data on the GM12878, HMEC, HUVEC, HeLa, IMR90, K562, KBM7, and NHEK human cell lines were retrieved from GEO at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525. Density plots were obtained using the *density* function in the *R* environment with the *adjust* parameter set to 0.5.

Enrichment P-values

P-values throughout the manuscript were computed using the Fisher's exact test except otherwise stated.

Data availability

All of the supplemental materials are provided on zenodo at https://doi.org/10.5281/zenodo.1283306.

Results

Guanine and cytosine nucleotide content identified two classes of human enhancers

To analyze the sequence properties of human enhancers, we considered regions of 1001 bp around the midpoints of the 65,423 CAGE-derived enhancers from phases 1 and 2 of the FANTOM5 project [Andersson *et al.* 2014; FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* 2014; Arner *et al.* 2015].

We sought to identify distinct classes of enhancers based on the positional distribution of guanines (Gs) and cytosines (Cs) along the enhancer regions. Specifically, each enhancer was represented by a 1001-bp-long binary vector with 1s representing G+C and 0s representing adenines (As) and thymines (Ts). We clustered the enhancers by applying the k-means clustering algorithm (MacQueen 1967) on the vectors. To select the number of clusters k, we considered silhouette plots, which provide a visual representation of how close each enhancer in one cluster is to enhancers in neighboring clusters (Rousseeuw 1987). A visual inspection of cluster silhouettes with $k \in [2, 5]$ revealed that the best clustering was obtained with $k = 2$ (Figure S1). We extracted the two clusters ($k = 2$) of enhancers (containing 42,248 and 23,164 enhancers, respectively) and observed distinct average G+C compositions (Figure S2).

As the clusterization highlighted distinct G+C content between the two classes of enhancers (mean $\sim 45\%$, median $\sim 44\%$, SD ~ 8), we distinguished enhancers with lower (%GC below the median; 32,487 enhancers) and higher (%GC above the median; 32,936 enhancers) %GC content (Figure 1a). The two classes are hereafter referred to as class 1 (with lower %GC content) and class 2 (with higher %GC content). As expected, we observed a large overlap between the two clusters obtained from the k-means algorithm applied to the positional patterns of G+C and the two classes defined from %GC (Jaccard similarity coefficients of 0.77 and 0.7, respectively).

As the mere %GC was sufficient to distinguish classes of enhancers and given the simplicity of this criterion, we used this classification in the following analyses. We sought to further explore the positional distribution of the %GC along the enhancer and their flanking regions in classes 1 and 2. We considered ±2000-bp DNA sequences 5′ and 3′ of the midpoints of the enhancers and computed the %GC at each position (Figure 1b). We observed distinct positional patterns of G+C content at DNA sequences flanking the enhancers from the two classes. The distinct patterns were emphasized when focusing on the differences in positional patterns of %GC from class 1 and class 2 enhancers represented as the normalized average G+C content separately for the two classes (Figure S5a). This result was expected as the k-means clusterization of the G+C positional patterns provided a classification of enhancers similar to classes 1 and 2. Class 1 enhancers harbored a stronger decrease in %GC at their midpoints when compared to class 2 enhancers. Moreover, the regions surrounding the class 1 enhancers harbored a symmetric decrease in %GC going away from the midpoints with a minimum at ~300–400 bp from the midpoints; it was followed by an increase in %GC. On the contrary, we observed a continuous symmetric decrease in %GC composition going away from class 2 enhancer midpoints. Nevertheless, both class 1 and class 2 enhancers harbored a symmetrical decrease of %GC in regions of ~300–400 bp around midpoints.

The two classes of enhancers are associated with distinct interspersed nuclear elements

The two classes of enhancers harbored similar proportions of enhancers located in intronic (~ 55 and ~ 49% of class 1 and class 2 enhancers, respectively) and intergenic (~ 44 and ~ 48% of class 1 and class 2 enhancers, respectively) regions (Figure S3a) but class 2 enhancers were found closer to TSSs than class 1 enhancers (Figure S3b). A third of class 1 enhancers (10,791) and 22% of class 2 enhancers (7165) overlapped repetitive elements. In agreement with their nucleotide composition, class 1 enhancers were enriched in (A)n and (T)n simple repeats and in AT-rich low complexity sequences while class 2 enhancers harbored G-rich and C-rich low complexity sequences (Table S1). Further, long interspersed nuclear elements were enriched in class 1 enhancers while no difference was observed for short interspersed nuclear elements (Table S1).

DNA regions flanking the two classes of human enhancers harbored distinct DNA structural properties

As DNA sequence and shape are intrinsically linked, we next considered 13 DNA shape features computed from DNA sequences with the DNAshape tool (Zhou *et al.* 2013; Chiu *et al.* 2016): buckle, helix twist (HelT), minor groove width (MGW), opening, propeller twist (ProT), rise, roll, shear, shift, slide, stagger, stretch, and tilt (Li *et al.* 2017). We plotted the distribution of these DNA shape features along the enhancers and their flanking regions for the two classes following the same procedure used for analyzing the G+C content (Figure 1, c–e, Figure S4, and Figure S5). We assessed the pattern differences between class 1 and class 2 enhancers by computing Kolmogorov–Smirnov (K-S) statistics. The three largest K-S statistics were obtained when considering MGW, stretch, and HelT (Figure 1, c–e). The main differences between class 1 and class 2 enhancers were observed for regions flanking the enhancers while the regions $< \sim 200$ bp away from the midpoints harbored very similar patterns; this observation was consistent between all 13 DNA shape features (Figure 1, c–e and Figure S5, b–n). These observations were in agreement with the %GC-patterns observed close to the enhancer midpoints with a symmetric decrease in G+C content and differences when considering flanking regions (Figure 1b). The DNA shape patterns are lost when shuffling the DNA sequences (Figure S6).

Taken together, these results described two subsets of human enhancers distinguishable by their G+C content with distinct positional distribution of %GC along the regions flanking the enhancers, which were reflected in their DNA structural properties. Importantly, we observed that the enhancer classification based on %GC highlighted distinct patterns of DNA shape features along the regions immediately flanking the enhancers but not at the enhancer central regions, indicating that the two classes of enhancers are located in distinct genomic environments.

The two classes of human enhancers associated with distinct biological processes

Different classes of mammalian promoters, derived from their nucleotide composition, were observed to be associated with genes linked to distinct biological functions (Bajic *et al.* 2006). Following the same approach, we sought for a functional interpretation of the %GC-based classification of human enhancers. We first aimed at characterizing whether the enhancers from the two classes were associated with distinct sets of target genes based on *cis*-eQTL associations from the GTEx project. For each enhancer class, a list of potential target genes was obtained for enhancers overlapping with *cis*-eQTL single nucleotide polymorphisms (see *Materials and Methods*). We found 2459 and 5857 genes linked to class 1 and class 2 enhancers, respectively (Table S2), with only 437 genes in common (*P*-value $< 10^{-3}$, Fisher's exact test, see *Materials and Methods*). It suggests that the %GC-based classification of human enhancers distinguished enhancers regulating different sets of genes.

As the number of enhancer–gene associations is limited from the *cis*-eQTL analysis (since it requires SNPs in QTL to be located within the enhancers), we drew CAGE-derived enhancer–gene pairs from two previous studies (Andersson *et al.* 2014; Cao *et al.* 2017). Based on correlations between promoter and enhancer activities derived from CAGE data in human samples, Andersson *et al.* (2014) linked enhancers to their potential gene promoter targets. Two sets of associations were computed, based on either CAGE-derived TSSs from FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* (2014) or RefSeq-annotated TSSs. More recently, the JEME method (Cao *et al.* 2017) used CAGE signal to predict enhancer–gene pairs based on two steps: (1) multiple regression between a TSS and all enhancers in the neighborhood of this TSS across samples; and (2) extraction of sample-specific enhancer–gene pairs. Application of JEME to FANTOM5 samples resulted in a third set of enhancer–gene pairs (Cao *et al.* 2017). Importantly, both FANTOM5 and JEME-based associations were supported by ChIA-PET, Hi-C, and eQTL data (Andersson *et al.* 2014; Cao *et al.* 2017). We intersected the predictions of enhancer–gene pairs from these three data sets to rely on high-confidence pairs. We noticed that G+C content of promoters targeted by class 1 enhancers is lower than that of promoters targeted by class 2 enhancers (Figure S7, Wilcoxon test *P*-value $<2.2e-16$).

To infer the biological functions of enhancers, we assumed that each enhancer was associated with the same biological functions as the genes it was predicted to regulate. We submitted the two sets of genes associated with class 1 and class 2 enhancers to the GOrilla and REViGO tools (Eden *et al.* 2009; Supek *et al.* 2011) to predict enriched (FDR *q*-value < 0.01) GO biological processes. Class 1 enhancers were predicted to target 1413 genes whereas class 2 enhancers were linked to 2838 genes (Table S3). In aggregate, the enhancers corresponded to a set of 3575 genes, of which 676 were common to the two classes (representing ~ 48, ~ 24, and ~ 19% of class 1, class 2, and the combined set of genes, respectively). The aggregated set of 3575 genes was used as the background set of genes for enrichment analyses.

Biological processes linked to immune system processes and response to stimulus were found enriched for genes associated with class 1 enhancers: "immune system process" ($q = 6.25 \times 10^{-9}$), "positive regulation of immune system process" ($q = 1.68 \times 10^{-6}$), "defense response" ($q = 5.49 \times 10^{-7}$, and "regulation of response to stimulus" ($q = 3.57 \times 10^{-4}$) (Figure 2a and Table S4). The GO term "immune system process" was found enriched with 647 genes predicted to be targets of class 1 enhancers (Table S4). When considering the genes predicted to be regulated by enhancers from class 2, no GO biological process term was found enriched.

When focusing on the genes predicted to be exclusively targeted by enhancers from class 1 or class 2, we did not find enriched GO terms. Finally, we considered the set of genes that were predicted to be common targets of enhancers from the two classes. GO terms associated with immune system process, response to stimulus, and cytokine production were found enriched (Table S5). The enrichment of immune system process-related terms was expected as ~48% of the target genes of class 1 enhancers are predicted to be targets of class 2 enhancers as well. While immune system genes are targeted by enhancers from both class 1 and class 2, class

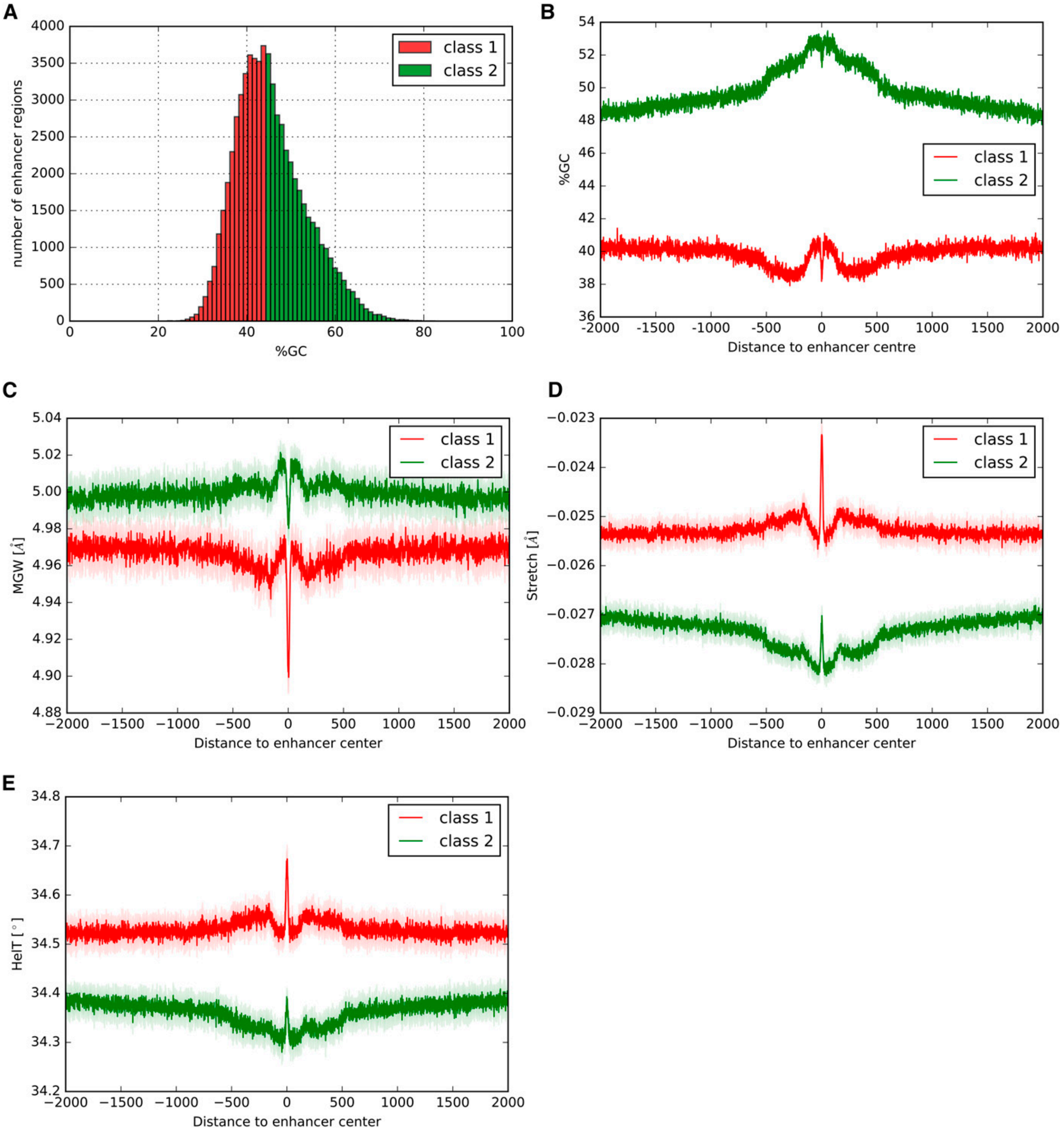

Figure 1 DNA sequence features at enhancers. Features associated with human enhancers from class 1 and class 2 are represented in red and green, respectively. (A) Histogram of the %GC of the enhancers. (B) Distribution of the average %GC (*y*-axis) of the enhancers in classes 1 and 2 along DNA regions of ±2000 bp centered at enhancer center points (*x*-axis). (C–E) Average DNA shape values (*y*-axis) along the DNA regions of ±2000 bp centered at enhancer midpoints (*x*-axis) for DNA shape features MGW (C), Stretch (D), and HelT (E). Shadow regions represent the 90% confidence intervals obtained from bootstrapping (see *Materials and Methods*). The largest Kolmogorov–Smirnov statistics between class 1 and class 2 enhancers were obtained with these three DNA shape features.

1 enhancers are more specifically associated with immune system genes.

Further, we observed that genes linked to the immune response were targeted by a greater number of enhancers than other genes. We ranked the list of genes by the number of enhancers they were associated with (Table S6) and submitted this list to GOrilla for functional enrichment analysis. This list was derived from 8339 pairs (Table S7, 2672 and

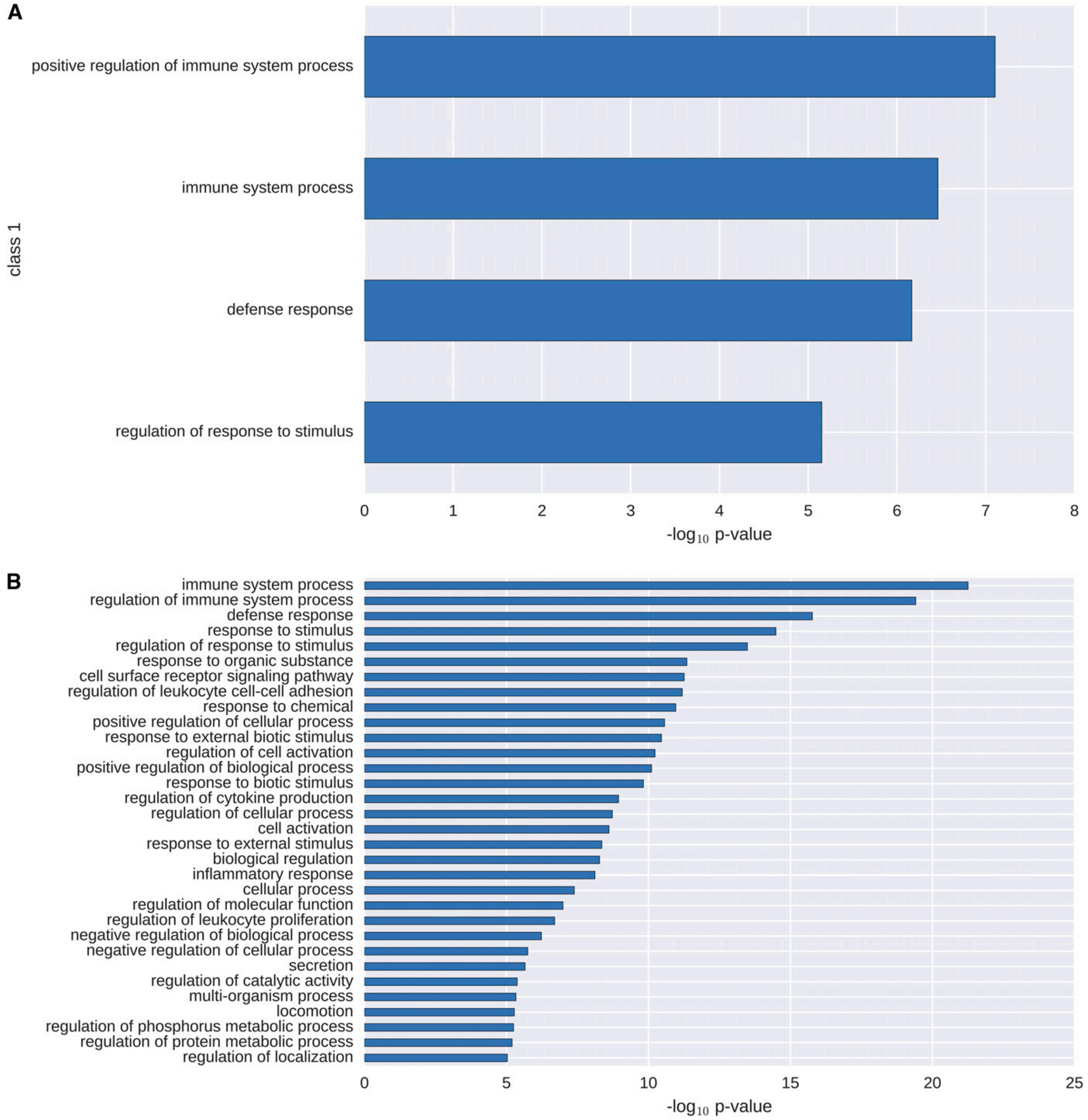

Figure 2 Functional enrichment analysis. Enriched GO biological processes (*y*-axis; $\log_{10}$ *P*-value < -5) obtained using the GOrilla and REViGO tools (Eden *et al.* 2009; Supek *et al.* 2011) with genes predicted to be regulated by enhancers from class 1 (a) and the ranked list of genes (by decreasing number of associated enhancers) predicted to be regulated by class 1 or 2 enhancers (b). The enhancer–gene pairs were recurrently predicted in three sets of enhancer–gene associations (see *Materials and Methods*).

5667 pairs with enhancers from class 1 and class 2, respectively). The GO biological process terms "immune system process," "regulation of immune system process," "defense response," "response to stimulus," and "regulation of response to stimulus" were found at the top of the enriched terms (Figure 2b and Table S8). Furthermore, from the 1661 enhancer–gene pairs where the gene is associated with "immune system process" (Table S9), 621 were derived from class 1 enhancers, showing that class 1 enhancer–gene associations are enriched in the list of enhancer–gene pairs linked to "immune system process" (Fisher's exact test *P*-value = 2.65×10^{-7}).

Taken together, the functional enrichment results revealed that a classification based on the G+C content of human

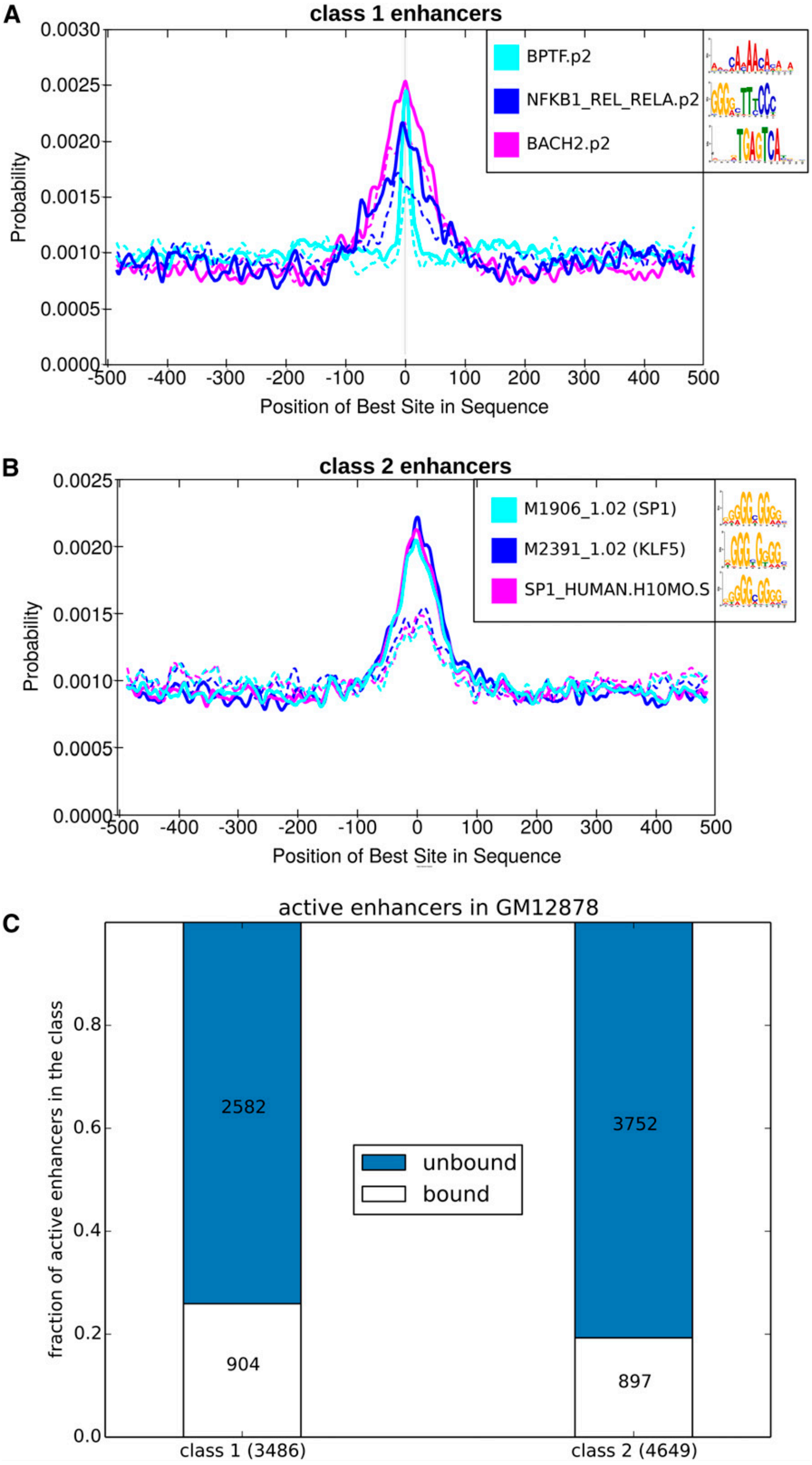

Figure 3 TF binding analysis at enhancer regions. Regions of ±500 bp around enhancer midpoints (a and b) were subjected to positional motif enrichment analyses using the Centrimo tool (Bailey and Machanick 2012) with motifs from JASPAR (Mathelier *et al.* 2015a), Cis-BP (Weirauch *et al.* 2014), Swiss Regulon (Pachkov *et al.* 2013), and HOCOMOCO (Kulakovskiy *et al.* 2016). Enhancers from class 1 (a) and class 2 (b) were analyzed separately. The *x*-axis represents the distance to the enhancer midpoints. The *y*-axis represents the probability of predicting TFBSs associated with the motifs given in the legend boxes. Plain lines represent the distribution of predicted TFBSs in the foreground sequences (from class 1 and class 2). Similarly, dashed lines represent the distribution of predicted TFBSs in the background sequences (from class 1 and class 2). Note that the SP1 PWMs enriched in class 2 enhancers originate from Weirauch *et al.* (2014) (M1906_1.02) and Kulakovskiy *et al.* (2016) (SP1_HUMAN.H10MO.S). (c) Proportion of class 1 (left) and class 2 (right) active enhancers in GM12878 bound or not by the RELA TF (using ChIP-seq data).

enhancer regions featured two sets of enhancers predicted to be regulating genes enriched for distinct biological functions. Specifically, while genes linked to “immune system process” were enriched for being linked to a high number of enhancers, the functional enrichments showed that class 1 enhancers were more specifically targeting genes involved in the immune response.

Distinct TFs predicted to act upon the two classes of human enhancers

We sought to identify TF binding motifs enriched within each class of enhancers, to suggest driving TFs for their distinct biological functions. We considered 1001-bp-long DNA sequences centered at the enhancers’ midpoints. Positional motif enrichment analyses were performed using the Centrimo tool (Bailey and Machanick 2012) to predict TF binding motifs overrepresented at enhancers. Class 1 enhancer regions were compared to class 2 regions and vice versa to highlight specific motifs (Figure 3, a and b and Data S1 and S2). The most enriched motifs in class 1 enhancer regions were related to the nucleosome-remodeling factor subunit BPTF and the nuclear factor κ-light-chain-enhancer of activated B cells (NF-κB)/Rel signaling [NFKB1, REL, and RELA (Liou 2006) and BACH2 (Itoh-Nakadai *et al.* 2014); Figure 3a and Data S1 and S2], in agreement with an involvement of class 1 enhancers in the immune response biological function (Figure 2). Motifs associated with the Specificity Protein/Krüppel-like Factor (SP/KLF) TFs were enriched in class 2 enhancer regions (Figure 3b and Data S1 and S2). Members of the SP/KLF family have been associated with a large range of core cellular processes such as cell growth, proliferation, and differentiation (Presnell *et al.* 2015). A similar analysis based on gene promoters associated with class 1 and class 2 enhancers did not yield enriched motifs.

We confirmed the motif-based enrichment of NF-κB/REL/RELA binding in class 1 enhancers by using ENCODE ChIP-seq data obtained in GM12878 cells for the RELA TF involved in NF-κB heterodimer formation. By combining data capturing histone modification marks, TF binding, and open chromatin regions from a specific cell type, the ChromHMM (Ernst and Kellis 2012) and Segway (Hoffman *et al.* 2012) tools segment the genome into regions associated with specific chromatin states. Focusing on predictions from ChromHMM and Segway combined, we found 3486 ($\sim$ 11%) and 4649 ($\sim$ 14%) active enhancer regions from classes 1 and 2, respectively. We observed that class 1 active enhancers were preferentially bound by RELA. Specifically, 904 active class 1 enhancers and 897 active class 2 enhancers overlapped RELA ChIP-seq peaks (P-value $= 1.2 \times 10^{-12}$, Fisher's exact test; Figure 3c).

Together, these results reinforced the predictions of biological functions specific to class 1 and class 2 enhancers (Figure 2) through the presence of associated TF binding motifs at enhancers.

The two classes of human enhancers exhibited distinct activity patterns

We further investigated the functional differences between the two classes of human enhancers by analyzing their patterns of activity across cell types. In previous studies, enhancer activity has been inferred either from histone modifications or eRNA transcription signatures (Ernst and Kellis 2012; Hoffman *et al.* 2012; Natoli and Andrau 2012; Andersson *et al.* 2015). We considered these two approaches with histone modification data from six cell lines and CAGE data from 71 cell types produced by the ENCODE (ENCODE Project Consortium 2012) and FANTOM5 (Andersson *et al.* 2014) projects, respectively.

We retrieved the human genome segmentation obtained using ChromHMM and Segway in the tiers 1 and 2 cell types from ENCODE (ENCODE Project Consortium 2012). For each cell type, we overlapped enhancers with predicted genome segments to assign activity states to the enhancers. As an example, Figure 4 presents the proportion of enhancers from classes 1 and 2 overlapping with segments associated with active, CTCF, and repressed chromatin states in embryonic stem cells (H1-hESC). We consistently observed that enhancers from class 2 were significantly more active than those from class 1, which were found to be enriched in repressed genomic segments (Figure 4 and Figure S8). Class 2 enhancers were also associated with segments characterized by CTCF binding.

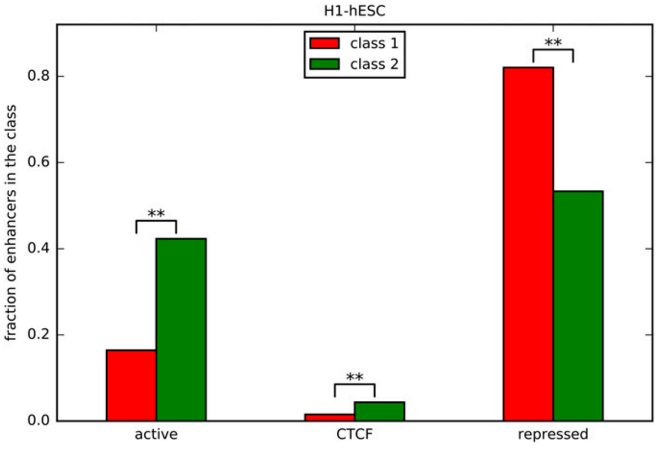

Figure 4 Human enhancers and genome segmentation. Histogram of the proportion of human enhancers (*y*-axis) in class 1 (red) and class 2 (green) lying within genome segments (*x*-axis) as annotated by combined predictions from ChromHMM (Ernst and Kellis 2012) and Segway (Hoffman *et al.* 2012) on human embryonic stem cells [H1-hESC from the ENCODE project (ENCODE Project Consortium 2012)]. Statistical significance (Bonferroni-corrected P-value < 0.01) of enrichment for enhancers from a specific class is indicated by "**."

To further validate these predictions, we specifically investigated nucleosome occupancy at classes 1 and 2 enhancers and extracted available MNase data at these regions from two cell lines, GM12878 and K562 (Figure 5). Nucleosomes occupancy at class 1 enhancers was found higher than that at class 2 enhancers (Figure 5) in both cell lines, indicating that class 2 enhancers were more associated with nucleosome-depleted regions, in agreement with their higher transcriptional activity than class 1 enhancers in GM12878 and K562 cells (Figure S8, a and e).

From the human samples with CAGE expression from the FANTOM5 project (Andersson *et al.* 2014), 71 cell types were defined by grouping cell and tissue samples. For each enhancer, Andersson *et al.* (2014) computed the entropy of expression of the enhancer across all the cell types. The entropy was used to compute a cell-type-specificity score for each enhancer [see *Materials and Methods* and Andersson *et al.* (2014)]. The cell-type-specificity score ranges from 0 to 1 with 0 indicating ubiquitous expression and 1 exclusive expression in one cell type. Using this enhancer expression specificity computation, we considered enhancers from class 1 and class 2 separately to highlight potential activity

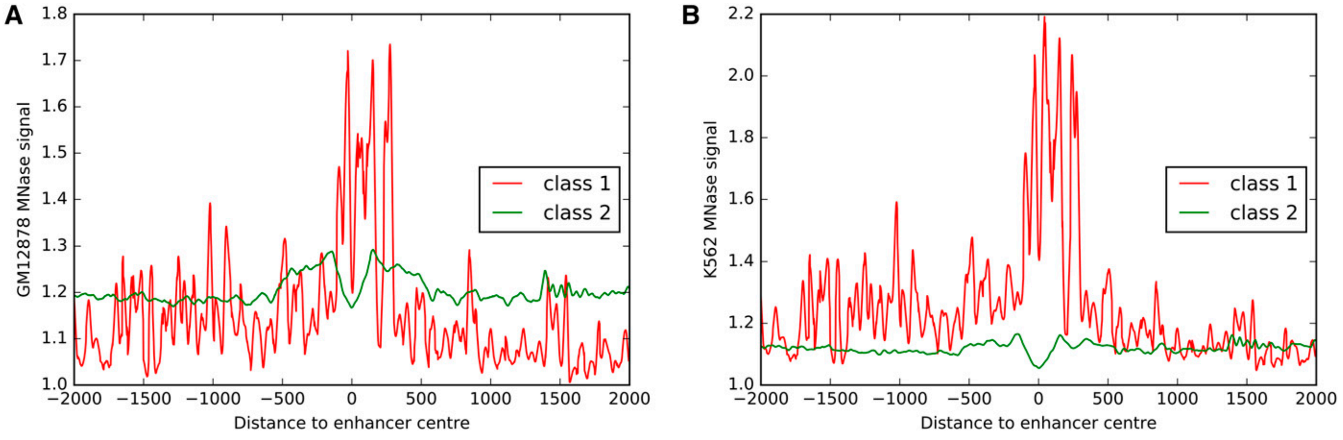

Figure 5 Nucleosome positioning at enhancers. MNase-seq signal from ENCODE for GM12878 (a) and K562 (b) cell lines at regions of ±2000 bp around enhancer centers from classes 1 (red) and 2 (green).

differences in the 71 cell types (Figure 6, a and b). Comparing enhancer activity specificity over all the cell types, enhancers from class 1 appeared to be more cell-type specific (Figure 6c). While immune cells, neurons, neuronal stem cells, and hepatocytes were previously described to use a higher fraction of human enhancers (Andersson *et al.* 2014), the elevated utilization was even more pronounced for class 1 enhancers (Figure 6c).

Taken together, these results derived from histone marks and transcriptional data highlighted that enhancers from class 2 were more ubiquitously active over human cell types than enhancers from class 1, which were more cell-type specific. In our previous functional analyses, we inferred the biological functions of the two classes of enhancers from the genes they were predicted to regulate. Here, we further confirmed specific functionalities for the two classes based on enhancer activity analyses, which corroborated with our functional analysis described above. Enhancers from class 1 were more cell-type specific, with an emphasis in cell types associated with the immune system, in agreement with the functional enrichment analysis.

Predicted immune system enhancers were activated upon cell infection

We sought to further confirm the association of class 1 enhancers with transcriptional control of immune responses. Pacis *et al.* (2015) generated genome-wide DNA methylation, histone marks, and chromatin accessibility data in normal dendritic cells (DCs) and DCs after infection with *Mycobacterium tuberculosis* (MTB). It provided the opportunity to study the chromatin state changes after infection obtained using the ChromHMM tool (Ernst and Kellis 2012). We overlapped chromatin state information with the enhancers from classes 1 and 2. To highlight the key epigenetic changes at enhancers, we classified the transition of activities before and after MTB infection into three groups: activated (from inactive before MTB infection to active after infection), inhibited (active to inactive), or unchanged (Figure 7 and Figure S9). We observed that the enhancers from class 1 were significantly more activated (P-value $= 4.5 \times 10^{-8}$, Fisher's exact test) and less inhibited (P-value $< 2.2 \times 10^{-16}$, Fisher's exact test) when compared to class 2 enhancers upon MTB infection (Figure 7). These results reinforced the potential role of class 1 enhancers in immune response.

Predicted immune system enhancers showed specific response activity

Arner *et al.* (2015) profiled time-courses with CAGE at a high temporal resolution within a 6-hr time-frame to analyze the transcriptional dynamics of enhancers and promoters on the terminal differentiation of stem cells and committed progenitors as well as on the response to stimuli for differentiated primary cells and cell lines. It highlighted distinct dynamic response patterns of early response activity. We overlaid our classification of human enhancers and their predicted target promoters with the response pattern data (Figure 8). Within the enhancers associated with any response pattern ($n = 4094$; 1294 and 2800 from class 1 and class 2, respectively), class 2 enhancers were enriched (P-value $= 3.9 \times 10^{-129}$, hyper-geometric test).

We focused on the set of 4094 enhancers classified in the dynamic response patterns. Class 1 enhancers were found enriched for down-regulation in "early standard response" and up-regulation in "late standard response." Conversely, class 2 enhancers appeared enriched for up-regulation in "early standard response" and for down-regulation in "late standard response." These results highlighted opposite activity dynamics between class 1 and class 2 enhancers. Note that class 2 was also enriched for up-regulation in "late response." On the other hand, promoters associated to class 1 were enriched in up-regulation of "rapid long response," "late response," and "long response," while promoters associated to class 2 enhancers were enriched in up-regulation of "early response," in down-regulation of "late response" and "long response," and in "late flat response" (Figure 8b), further reinforcing the existence of different transcriptional responses associated with class 1 and class 2 enhancers.

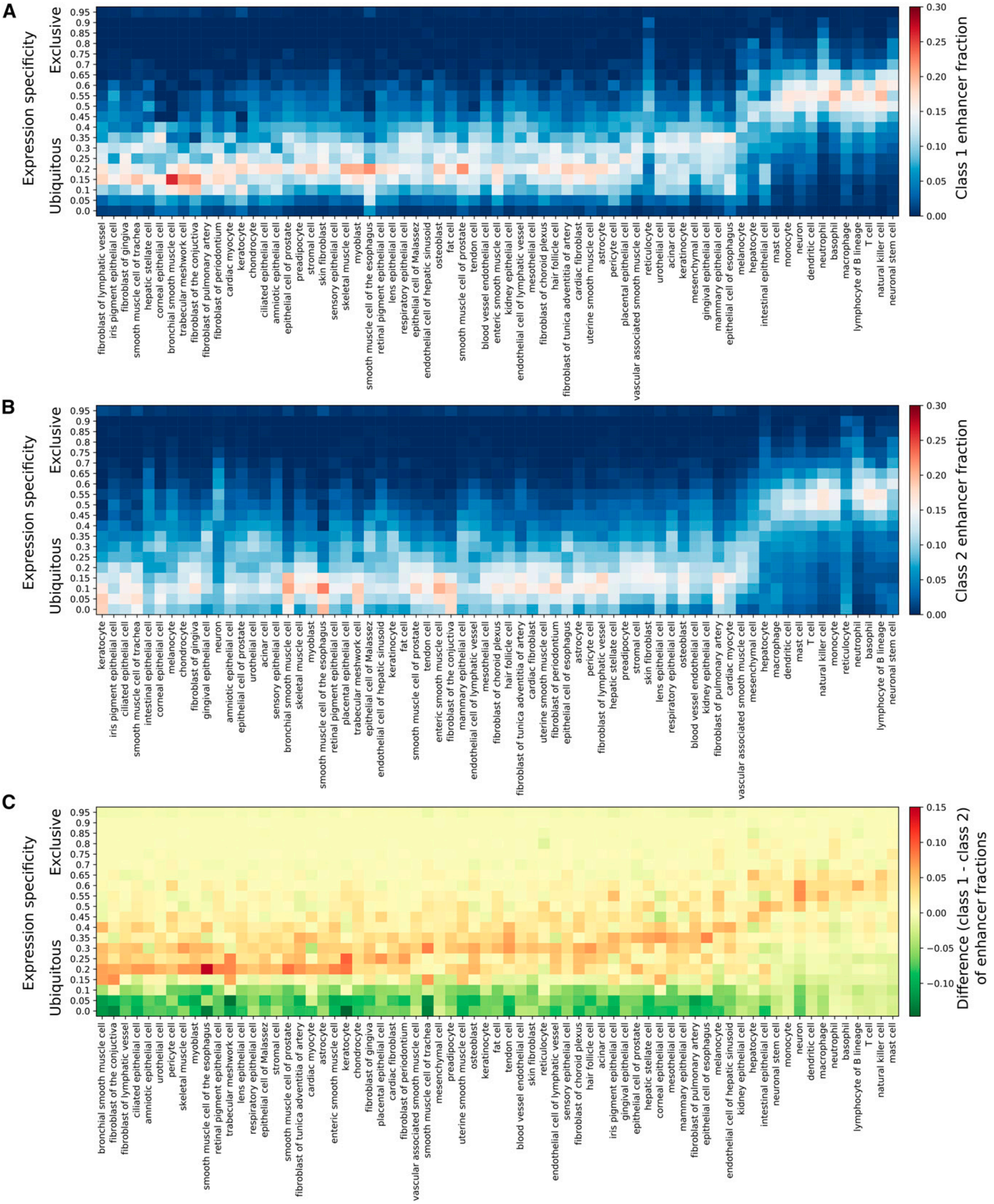

Figure 6 Cell-type expression specificities of human enhancers. The cell-type expression specificities (see *Materials and Methods* for details on cell-type specificity computation) derived from FANTOM5 CAGE data sets (Andersson *et al.* 2014) are provided as a heat map for human enhancers in class 1 (a) and class 2 (b). The colors (see scales) in a and b represent the fraction of expressed enhancers in each cell type (columns) found in each expression specificity range (rows). The differences in cell-type expression specificities between class 1 and class 2 enhancers are provided as a heat map (c). Positive (respectively negative) values are represented in red (respectively green) and indicate a higher fraction of class 1 (respectively class 2) enhancers. CAGE, Cap Analysis of Gene Expression.

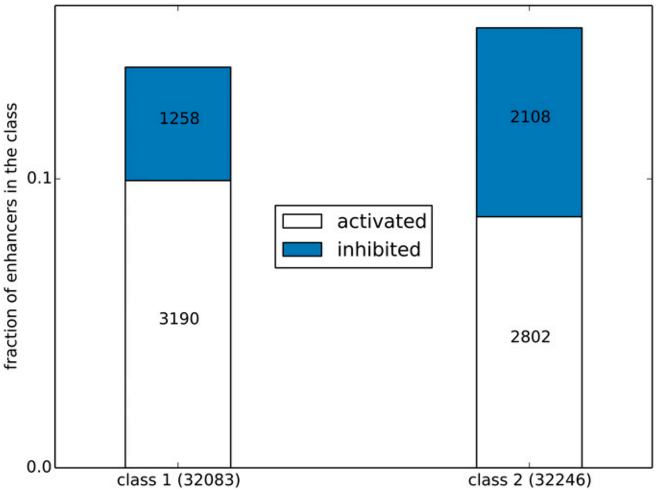

Figure 7 Enhancer activation upon cell infection. Stacked histogram of the fraction of human enhancers (*y*-axis) from class 1 and class 2 predicted to be activated (white) or inhibited (blue). Predictions were obtained using genomic segments predicted by ChromHMM (Ernst and Kellis 2012) on human dendritic cells before and after infection with *Myobacterium tuberculosis* (Pacis *et al.* 2015). Stacked histogram including unchanged activity is provided in Figure S9.

Enhancers from the same class colocalized within chromatin domains

The organization of the chromatin in cell nuclei is a key feature in gene expression regulation by forming regulatory interactions within TADs (Dixon *et al.* 2016). Genes within the same TAD tend to be coordinately expressed across cell types and tissues, and clusters of functionally related genes requiring coregulation tend to lie within the same TADs (Gibcus and Dekker 2013; Dixon *et al.* 2016). Similar to these studies analyzing gene organization observed in chromatin domains, we focused on how the two classes of enhancers were organized with respect to TADs. We compared the distribution of enhancers from the two classes within TADs (Dixon *et al.* 2012). Specifically, we assessed whether individual TADs were biased for containing more enhancers associated with a specific class than expected by chance using the Binomial test. The distribution of the corresponding *P*-values was compared to those obtained by randomly assigning class 1 and 2 labels to the enhancers. TADs were enriched for enhancers from a specific class (Figure 9a), showing a genomic organization of human enhancers with respect to chromatin domains.

TADs represent interactions within megabase-sized domains of chromatin, which can be subdivided into kilobase-sized chromatin loops of chromatin interactions (Rao *et al.* 2014). We refined our analyses of class-based enhancer colocalization by focusing on chromosomal loops derived from eight cell lines (Rao *et al.* 2014). We found that chromatin loops tended to contain enhancers from a specific class (Figure 9b and Figure S10), similar to TADs. Furthermore, class 1 enhancers were evenly distributed within the chromatin loops whereas enhancers from class 2 were observed to be situated close to the loop boundaries (Figure 9c and Figure S11). This observation is in agreement with the enrichment for class 2 enhancers in CTCF chromatin segments (Figure 4) as chromatin loop boundaries are known to be enriched for CTCF binding (Rao *et al.* 2014).

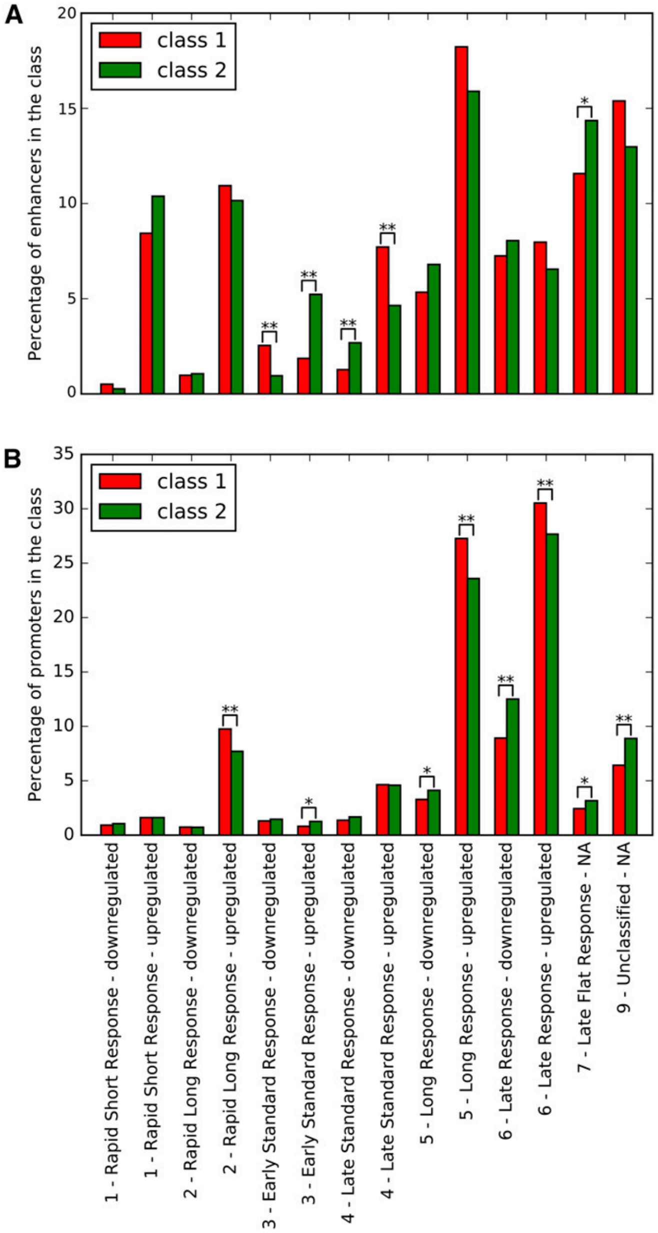

Figure 8 Expression dynamics of human enhancers and associated promoters. Response patterns (*x*-axis) of human enhancers (a) and promoters (b) in time courses were classified by Arner *et al.* (2015). The percentage (*y*-axis) of enhancers (top) and promoters (bottom) from class 1 (red) and class 2 (green) in each response pattern category are provided as histograms in the two panels. A significant difference between class 1 and class 2 enhancers or promoters in a specific category is highlighted by "*" or "**" for Bonferroni-corrected *P*-value < 0.05 or < 0.01, respectively (Fisher's exact tests).

Discussion

We have analyzed the sequence properties of FANTOM5 human enhancers derived from CAGE experiments to reveal

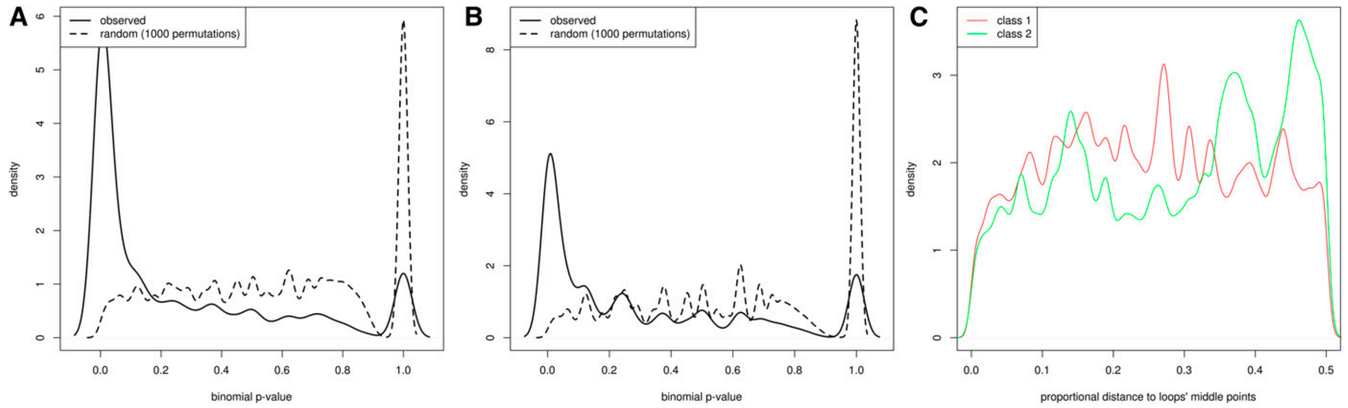

Figure 9 Chromosomal organization of class 1 and class 2 enhancers. (a) For each TAD (Dixon *et al.* 2012), we computed the *P*-value of the Binomial test to assess the enrichment for enhancers from a specific class. The plot compares the density (*y*-axis) of *P*-values for Binomial tests (*x*-axis) applied to classes 1 and 2 enhancers (plain line) and 1000 random assignments of class labels to the enhancers (dashed line). (b) The same analysis as in a was performed using chromatin loops predicted in lymphoblastoid GM12878 cells (Rao *et al.* 2014). (c) Density (*y*-axis) of distances (*x*-axis) between enhancers and chromatin loop centers defined using Hi-C data in GM12878 cells (Rao *et al.* 2014). The distances were normalized by the length of the loops. Enhancers at the center of the loops were found at distance 0.0 while enhancers at chromatin loops boundaries were found at distance 0.5. Results associated with class 1 and class 2 enhancers are depicted in red and green, respectively.

that a subset with lower G+C content is more specifically associated with immune response genes. This set of enhancers tends to colocalize within chromatin domains, exhibits cell-type specificity, is activated upon infection, and is observed with specific response activity. In summary, our study of enhancer DNA sequence composition culminated with the identification of human enhancers associated with genes enriched for immune response functions that harbor specific sequence composition, activity, and genome organization.

While immune response genes were found to be targeted by enhancers from both classes, the enhancers with lower %GC were more specifically targeting these genes (Figure 2). As these enhancers are active in a more cell-type-specific manner, it suggests that activation of immune response genes in specific cell types is driven by such enhancers bound by NFκB and related TFs (Figure 3).

The analyses of sequence properties in regulatory regions, most prominently CpG islands at promoters, have been key to understanding gene expression regulation (Bajic *et al.* 2006; Hughes and Rando 2009; Tillo and Hughes 2009). We observed that enhancers with a higher G+C content were more broadly activated than the enhancers with a lower G+C content. A recent study highlighted that human enhancers with broad regulatory activity across cellular contexts were enriched for GC-rich sequence motifs, in line with the fact that broadly active human TFs bind to GC-rich motifs (Colbran *et al.* 2017). The enhancers more specifically associated with immune response genes predicted here exhibit a cell-type-specific expression pattern and have lower %GC. It remains unclear how and why this set of enhancers has emerged with such sequence properties. In line with their lower G+C content, they were associated with (A)n and (T)n simple repeats and AT-rich low complexity sequences. They were also strongly enriched in long interspersed nuclear elements (LINEs) when compared to other enhancers. Provided that repetitive elements represent both molecular parasites and evolutionary drivers (Elbarbary *et al.* 2016), these observations may explain, at the sequence level, the differences observed between the two classes of enhancers. Besides, they suggest that, similar to Alu elements and endogenous retroviruses (Sasaki *et al.* 2008; Chuong *et al.* 2013; Su *et al.* 2014), LINEs can exert enhancer activities that are specifically related to the immune response process. In line with this proposal, expression of LINE-1 has been shown to trigger inflammatory pathways in systemic autoimmune disease (Crow 2010; Mavragani *et al.* 2016).

When comparing the genomic distribution of the two sets of enhancers, we found that they lie in specific genomic environments associated with a distinct local DNA shape pattern. DNA structural properties were shown to be linked with DNA flexibility, nucleosome positioning, and gene expression regulation (Tirosh *et al.* 2007; Parker *et al.* 2009; Raveh-Sadka *et al.* 2012; Struhl and Segal 2013; Bansal *et al.* 2014). We noticed that the DNA shape conformation at enhancers was similar between the two classes but distinct at their flanking regions (Figure 1, c–e and Figure S4). The differences in DNA shape features between the two classes of enhancers might relate to differences in conformational flexibility. Indeed, we observed that class 1 enhancer flanking regions harbored increasing MGW, stagger, and opening combined with decreasing HelT close to enhancers compared to class 2 enhancers (Figure 1, c–e and Figure S4). These characteristics all relate to distinct flexibility of the DNA, which could provide a topological explanation for the differences observed between the two classes.

One simple explanation for the observation that enhancers from the same class are colocalized would be that enhancers from the same TADs can be found within the same isochore, simply because isochores can help define TADs (Jabbari and Bernardi 2017). For enhancer function, this implies that enhancer–promoter associations can be governed by sequence-level instructions, like G+C content. This idea is in line with the existence of a sequence-encoded enhancer–promoter specificity as unveiled by Zabidi *et al.* (2015) and Singh *et al.* (2018), which is also supported by our findings related to the association of enhancers with gene promoters of similar %GC composition (Figure S7).

Acknowledgments

As research parasites (Longo and Drazen 2016), we are indebted to the researchers around the globe who generated experimental data and made them freely available. We thank Miroslav Hatas and Georgios Magklaras for systems support, Dora Pak for management support, Chih-Yu Chen for providing the code for enrichment analyses in chromatin conformation data, Tsu-Pei Chiu, Jinsen Li, and Remo Rohs for providing DNA shape feature computation before their publication and their help in interpreting the DNA shape results, Robin Andersson for his help with FANTOM5 enhancer data, and Vladimir Teif for insightful discussion on nucleosome positioning. C.-H.L. was supported by funding from Le Centre National de la Recherche Scientifique (CNRS), *Plan d'Investissement d'Avenir* #ANR-11-BINF-0002 *Institut de Biologie Computationnelle* (Young Investigator grant). A.M. and W.W.W. were supported by the Genome Canada/Genome BC Large Scale Applied Research Grant 174CDE. Funding was provided by the Child and Family Research Institute and the British Columbia Children's Hospital Foundation, Vancouver, to A.M. and W.W.W. A.M. was also supported by funding from the Norwegian Research Council (Project Number 187615), Helse Sør-Øst, and the University of Oslo through the Centre for Molecular Medicine Norway (NCMM) and the Oslo University Hospital Radiumhospitalet.

Author contributions: C-HL and AM conceived and designed the project. C-HL and AM implemented and performed experiments. C-HL, WWW, and AM analyzed and interpreted the results. C-HL and AM wrote the manuscript with revisions from WWW.

Literature Cited

Andersson, R., C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt *et al.*, 2014 An atlas of active enhancers across human cell types and tissues. Nature 507: 455–461. https://doi.org/10.1038/nature12787

Andersson, R., A. Sandelin, and C. G. Danko, 2015 A unified architecture of transcriptional regulatory elements. Trends Genet. 31: 426–433. https://doi.org/10.1016/j.tig.2015.05.007

Arner, E., C. O. Daub, K. Vitting-Seerup, R. Andersson, B. Lilje *et al.*, 2015 Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science 347: 1010–1014. https://doi.org/10.1126/science.1259418

Babu, D., and M. J. Fullwood, 2015 3D genome organization in health and disease: emerging opportunities in cancer translational medicine. Nucleus 6: 382–393. https://doi.org/10.1080/19491034.2015.1106676

Bailey, T. L., and P. Machanick, 2012 Inferring direct DNA binding from ChIP-seq. Nucleic Acids Res. 40: e128. https://doi.org/10.1093/nar/gks433

Bajic, V. B., S. L. Tan, A. Christoffels, C. Schönbach, L. Lipovich *et al.*, 2006 Mice and men: their promoter properties. PLoS Genet. 2: e54. https://doi.org/10.1371/journal.pgen.0020054

Bansal, M., A. Kumar, and V. R. Yella, 2014 Role of DNA sequence based structural features of promoters in transcription initiation and gene expression. Curr. Opin. Struct. Biol. 25: 77–85. https://doi.org/10.1016/j.sbi.2014.01.007

Cao, Q., C. Anyansi, X. Hu, L. Xu, L. Xiong *et al.*, 2017 Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. Nat. Genet. 49: 1428–1436. https://doi.org/10.1038/ng.3950

Carninci, P., T. Kasukawa, S. Katayama, J. Gough, M. C. Frith *et al.*, 2005 The transcriptional landscape of the mammalian genome. Science 309: 1559–1563. https://doi.org/10.1126/science.1112014

Chiu, T.-P., L. Yang, T. Zhou, B. J. Main, S. C. J. Parker *et al.*, 2015 GBshape: a genome browser database for DNA shape annotations. Nucleic Acids Res. 43: D103–D109. https://doi.org/10.1093/nar/gku977

Chiu, T.-P., F. Comoglio, T. Zhou, L. Yang, R. Paro *et al.*, 2016 DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics 32: 1211–1213. https://doi.org/10.1093/bioinformatics/btv735

Chuong, E. B., M. A. Rumi, M. J. Soares, and J. C. Baker, 2013 Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat. Genet. 45: 325–329. https://doi.org/10.1038/ng.2553

Cock, P. J., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox *et al.*, 2009 Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Colbran, L. L., L. Chen, and J. A. Capra, 2017 Short DNA sequence patterns accurately identify broadly active human enhancers. BMC Genomics 18: 536. https://doi.org/10.1186/s12864-017-3934-9

Crow, M. K., 2010 Long interspersed nuclear elements (LINE-1): potential triggers of systemic autoimmune disease. Autoimmunity 43: 7–16. https://doi.org/10.3109/08916930903374865

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li *et al.*, 2012 Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485: 376–380. https://doi.org/10.1038/nature11082

Dixon, J. R., D. U. Gorkin, and B. Ren, 2016 Chromatin domains: the unit of chromosome organization. Mol. Cell 62: 668–680. https://doi.org/10.1016/j.molcel.2016.05.018

Eden, E., R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, 2009 GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10: 48. https://doi.org/10.1186/1471-2105-10-48

Elbarbary, R. A., B. A. Lucas, and L. E. Maquat, 2016 Retrotransposons as regulators of gene expression. Science 351: aac7247. https://doi.org/10.1126/science.aac7247

ENCODE Project Consortium, 2012 An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74. https://doi.org/10.1038/nature11247

Ernst, J., and M. Kellis, 2012 ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods 9: 215–216. https://doi.org/10.1038/nmeth.1906

FANTOM Consortium and the RIKEN PMI and CLST (DGT)Forrest, A. R., H. Kawaji, M. Rehli, J. K. Baillie *et al.*, 2014 A promoter-level mammalian expression atlas. Nature 507: 462–470. https://doi.org/10.1038/nature13182

Gibcus, J. H., and J. Dekker, 2013 The hierarchy of the 3D genome. Mol. Cell 49: 773–782. https://doi.org/10.1016/j.molcel.2013.02.011

Hoffman, M. M., O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes *et al.*, 2012 Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods 9: 473–476. https://doi.org/10.1038/nmeth.1937

Hughes, A., and O. J. Rando, 2009 Chromatin 'programming' by sequence – is there more to the nucleosome code than %GC? J. Biol. 8: 96. https://doi.org/10.1186/jbiol207

Hunter, J. D., 2007 Matplotlib: a 2d graphics environment. Comput. Sci. Eng. 9: 90–95. https://doi.org/10.1109/MCSE.2007.55

Itoh-Nakadai, A., R. Hikota, A. Muto, K. Kometani, M. Watanabe-Matsui *et al.*, 2014 The transcription repressors Bach2 and Bach1 promote B cell development by repressing the myeloid program. Nat. Immunol. 15: 1171–1180. https://doi.org/10.1038/ni.3024

Jabbari, K., and G. Bernardi, 2017 An isochore framework underlies chromatin architecture. PLoS One 12: e0168023. https://doi.org/10.1371/journal.pone.0168023

Jolma, A., J. Yan, T. Whitington, J. Toivonen, K. R. Nitta *et al.*, 2013 DNA-binding specificities of human transcription factors. Cell 152: 327–339. https://doi.org/10.1016/j.cell.2012.12.009

Kodzius, R., M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda *et al.*, 2006 CAGE: cap analysis of gene expression. Nat. Methods 3: 211–222. https://doi.org/10.1038/nmeth0306-211

Kulakovskiy, I. V., I. E. Vorontsov, I. S. Yevshin, A. V. Soboleva, A. S. Kasianov *et al.*, 2016 HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Res. 44: D116–D125. https://doi.org/10.1093/nar/gkv1249

Li, J., J. M. Sagendorf, T.-P. Chiu, M. Pasi, A. Perez *et al.*, 2017 Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. Nucleic Acids Res. 45: 12877–12887. https://doi.org/10.1093/nar/gkx1145

Lieberman-Aiden, E., N. L. Berkum, L. Williams, M. Imakaev, T. Ragoczy *et al.*, 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289–293. https://doi.org/10.1126/science.1181369

Liou, H.-C., (Editor), 2006 *NF-[kappa]B/Rel Transcription Factor Family*. Molecular Biology Intelligence Unit. Landes Bioscience/Eurekah.com. Springer Science+Business Media, Georgetown, TX,/New York. OCLC: ocm68133074.

Longo, D. L., and J. M. Drazen, 2016 Data sharing. N. Engl. J. Med. 374: 276–277. https://doi.org/10.1056/NEJMe1516564

MacQueen, J., 1967 Some methods for classification and analysis of multivariate observations, pp. 281–297 in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. University of California Press, Berkeley, CA.

Mathelier, A., O. Fornes, D. J. Arenillas, C.-y. Chen, G. Denay *et al.*, 2015a JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 44: D110–D115. https://doi.org/10.1093/nar/gkv1176

Mathelier, A., W. Shi, and W. W. Wasserman, 2015b Identification of altered cis-regulatory elements in human disease. Trends Genet. 31: 67–76. https://doi.org/10.1016/j.tig.2014.12.003

Mavragani, C. P., I. Sagalovskiy, Q. Guo, A. Nezos, E. K. Kapsogeorgou *et al.*, 2016 Expression of long interspersed nuclear element 1 retroelements and induction of type I interferon in patients with systemic autoimmune disease. Arthritis Rheumatol. 68: 2686–2696. https://doi.org/10.1002/art.39795

McKinney, W., 2010 Data structures for statistical computing in Python, pp. 51–56 in *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman editors.

Natoli, G., and J.-C. Andrau, 2012 Noncoding transcription at enhancers: general principles and functional models. Annu. Rev. Genet. 46: 1–19. https://doi.org/10.1146/annurev-genet-110711-155459

O'Connor, T. R., and T. L. Bailey, 2015 Creating and validating cis-regulatory maps of tissue-specific gene expression regulation. Nucleic Acids Res. 42: 11000–11010. https://doi.org/10.1093/nar/gku801

Pachkov, M., P. J. Balwierz, P. Arnold, E. Ozonov, and E. Nimwegen, 2013 SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic Acids Res. 41: D214–D220. https://doi.org/10.1093/nar/gks1145

Pacis, A., L. Tailleux, A. M. Morin, J. Lambourne, J. L. MacIsaac *et al.*, 2015 Bacterial infection remodels the DNA methylation landscape of human dendritic cells. Genome Res. 25: 1801–1811. https://doi.org/10.1101/gr.192005.115

Parker, S. C. J., L. Hansen, H. O. Abaan, T. D. Tullius, and E. H. Margulies, 2009 Local DNA topography correlates with functional noncoding regions of the human genome. Science 324: 389–392. https://doi.org/10.1126/science.1169050

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion *et al.*, 2011 Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12: 2825–2830.

Perez, F., and B. E. Granger, 2007 IPython: a system for interactive scientific computing. Comput. Sci. Eng. 9: 21–29. https://doi.org/10.1109/MCSE.2007.53

Pohl, A., and M. Beato, 2014 bwtool: a tool for bigWig files. Bioinformatics 30: 1618–1619. https://doi.org/10.1093/bioinformatics/btu056

Presnell, J. S., C. E. Schnitzler, and W. E. Browne, 2015 KLF/SP transcription factor family evolution: expansion, diversification, and innovation in eukaryotes. Genome Biol. Evol. 7: 2289–2309. https://doi.org/10.1093/gbe/evv141

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rao, S. S. P., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov *et al.*, 2014 A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159: 1665–1680 (erratum: Cell 162: 687–688). https://doi.org/10.1016/j.cell.2014.11.021

Raveh-Sadka, T., M. Levo, U. Shabi, B. Shany, L. Keren *et al.*, 2012 Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat. Genet. 44: 743–750. https://doi.org/10.1038/ng.2305

R Core Team, 2016 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rousseeuw, P. J., 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20: 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sasaki, T., H. Nishihara, M. Hirakawa, K. Fujimura, M. Tanaka *et al.*, 2008 Possible involvement of SINEs in mammalian-specific brain formation. Proc. Natl. Acad. Sci. USA 105: 4220–4225. https://doi.org/10.1073/pnas.0709398105

Shiraki, T., S. Kondo, S. Katayama, K. Waki, T. Kasukawa *et al.*, 2003 Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. USA 100: 15776–15781. https://doi.org/10.1073/pnas.2136655100

Singh, S., Y. Yang, B. Poczos, and J. Ma, 2018 Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. bioRxiv. https://doi.org/10.1101/085241.

Struhl, K., and E. Segal, 2013 Determinants of nucleosome positioning. Nat. Struct. Mol. Biol. 20: 267–273. https://doi.org/10.1038/nsmb.2506

Su, M., D. Han, J. Boyd-Kirkup, X. Yu, and J. D. Han, 2014 Evolution of Alu elements toward enhancers. Cell Rep. 7: 376–385. https://doi.org/10.1016/j.celrep.2014.03.011

Supek, F., M. Bošnjak, N. Škunca, and T. Šmuc, 2011 REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6: e21800. https://doi.org/10.1371/journal.pone.0021800

Tillo, D., and T. R. Hughes, 2009 G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics 10: 442. https://doi.org/10.1186/1471-2105-10-442

Tirosh, I., J. Berman, and N. Barkai, 2007 The pattern and evolution of yeast promoter bendability. Trends Genet. 23: 318–321. https://doi.org/10.1016/j.tig.2007.03.015

Visel, A., M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama *et al.*, 2009 ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457: 854–858. https://doi.org/10.1038/nature07730

Wasserman, W. W., and A. Sandelin, 2004 Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. 5: 276–287. https://doi.org/10.1038/nrg1315

Weirauch, M. T., A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero *et al.*, 2014 Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158: 1431–1443. https://doi.org/10.1016/j.cell.2014.08.009

Worsley Hunt, R., A. Mathelier, L. Peso, and W. W. Wasserman, 2014 Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. BMC Genomics 15: 472. https://doi.org/10.1186/1471-2164-15-472

Zabidi, M. A., C. D. Arnold, K. Schernhuber, M. Pagani, M. Rath *et al.*, 2015 Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature 518: 556–559. https://doi.org/10.1038/nature13994

Zhou, T., L. Yang, Y. Lu, I. Dror, A. C. Dantas Machado *et al.*, 2013 DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. 41: W56–W62. https://doi.org/10.1093/nar/gkt437

Communicating editor: C. Kaplan