Human enhancers associated with immune response harbour specific sequence composition, activity, and genome organization

Charles-Henri Lecellier[1,2,3,*], Wyeth W. Wasserman[3], Remo Rohs[4], and Anthony Mathelier[3,5,6,*]

[1] *Institut de Génétique Moléculaire de Montpellier, Université de Montpellier, 1919 Route de Mende - 34293 Montpellier cedex 5, France*

[2] *Institut de Biologie Computationnelle, 860 rue de St. Priest, 34095 Montpellier cedex 5, France*

[3] *Centre for Molecular Medicine and Therapeutics at the Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, 980 West 28th Avenue, Room 3103, V5Z 4H4, Vancouver, BC, Canada*

[4] *Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA*

[5] *Centre for Molecular Medicine Norway (NCMM), Nordic EMBL partnership, University of Oslo and Oslo University Hospital, Oslo, Norway*

[6] *Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway*

[*] *Co-corresponding authors: charles.lecellier@igmm.cnrs.fr (CHL) and anthony.mathelier@ncmm.uio.no (AM)*

September 30, 2016

Abstract

The FANTOM5 consortium recently characterized 38,554 robust human enhancers from 808 cell and tissue types using the Cap Analysis of Gene Expression technology. We used the distribution of guanine and cytosine nucleotides at enhancer regions to distinguish two classes of enhancers harboring distinct DNA structural properties. 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 TFs involved in immune response. We observed that immune response enhancers were cell type specific, preferentially activated upon bacterial infection, and with long-lasting response activity. Looking at chromatin capture data, we found that the two classes of enhancers were lying in distinct topologically-associated domains and chromatin loops. Our results suggest that specific DNA sequence patterns encode for classes of enhancers that are functionally distinct and specifically organized in the human genome.

Background

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) [1]. These regulatory elements are located within larger regulatory regions, the promoters and enhancers [2]. 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 [3–5]. High-resolution chromatin conformation capture (Hi-C) technology maps genomic regions in spatial proximity within cell nuclei [6]. The Hi-C technology identified specific genomic neighbourhoods of chromatin interactions, the topologically associating domains (TADs), which represent chromatin compartments that are stable between cell types and conserved across species [7, 8].

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 [9,10]. Furthermore, sequence composition is intrinsically linked to the three-dimensional structure of the DNA. Topological studies have used sequence properties to predict four structural features of DNA: helix twist (HelT), minor groove width (MGW), propeller twist (ProT), and Roll [11, 12]. These topological properties have been shown to inform the analysis of protein-DNA interactions obtained from high-throughput experiments [13–16], emphasizing the importance of DNA sequence composition in transcriptional regulation.

DNA sequence composition and other features of promoter regions have been extensively studied, including such key advances as the discovery of CpG is-

lands. The analysis of promoter regions in the human genome was accelerated by the development of the Cap Analysis of Gene Expression (CAGE) technology [17,18], which identifies active TSSs in a high-throughput manner based on 5' capped RNA isolation. Using CAGE data, a large scale identification of the precise location of TSSs in human [19] led to the classification of promoters into four classes based on G+C content (%GC) [20]. The study highlighted that GC-rich promoters are associated with genes involved in various binding and protein transport activities while GC-poor promoters are associated 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 [21].

Large-scale computational analyses of enhancer regions have been hampered by a limited set of bona fide enhancers. An advantage of the CAGE technology is its capacity to identify *in vivo*-transcribed enhancers. Specifically, it identifies active enhancer regions in biological samples by capturing bidirectional RNA transcripts at enhancer boundaries [22]. Using this characteristic of CAGE data, the FANTOM5 project identified 38,554 "robust" human enhancers across 808 samples [22]. Sequence property analysis suggested that the enhancers share properties with CpG-poor promoters. The findings shed light on the structure, organization, and function of human enhancers.

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 1,000 bp, enhancer-promoter links have been shown to be conserved across cell types [23]. As the CAGE technology captures the level of activity for both promoters and enhancers in the same samples, predicting the potential targets of the enhancers was obtained by correlating the activity levels of these regulatory regions over hundreds of human samples from the FANTOM5 consortium [22]. Using the predicted enhancer-gene associations, the authors 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 distribution of G+C nucleotides along the sequences of human CAGE-derived enhancer regions to define two classes of enhancers. The specific sequence features of the two classes encoded for distinct topological DNA shape patterns. The enhancers from the GC-poor class were predicted to be functionally associated with genes involved in the immune response whereas the enhancers from the other class were associated with genes involved in biological processes related to transcription. 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, immune system enhancers were observed with a long-lasting response activity pattern following cell stimulation in time-course data sets. Finally, we observed

that the two classes of enhancers tended to be structurally organized in the human chromosomes within distinct TADs and DNA chromatin loops.

Results

Guanine and cytosine nucleotide patterns identified two classes of human enhancers with distinct DNA structural properties

To analyze the sequence properties of human enhancers, we considered the set of 38,554 CAGE-derived enhancers found to be significantly active in at least one primary cell or tissue sample in the FANTOM5 project [21, 22]. We extracted 500 bp DNA sequences 5' and 3' of the mid-point of the enhancers as defined by Andersson *et al.* [22]. We sought to identify distinct classes of enhancers based on the distribution of guanines (Gs) and cytosines (Cs) along the enhancer regions. Specifically, each enhancer was represented by a 1,001 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 [24] 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 neighbouring clusters [25]. 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 two classes ($k = 2$) of enhancers with distinct distributions of G+C along the enhancer regions (Figure 1a). The two classes were composed of 14,204 and 24,343 enhancers, hereafter referred to as class 1 and class 2, respectively. While enhancers from class 1 were more GC-rich than enhancers from class 2, separating the enhancers solely based on GC content would have resulted in a different classification (i.e. there is an overlap between the classes in terms of G+C content, as shown in Figure 1b).

As DNA sequence and shape are intrinsically linked, we next considered four DNA shape features computed from DNA sequences with the DNAshape tool [12]: helix twist (HelT), minor groove width (MGW), propeller twist (ProT), and Roll. We applied the k-means clustering algorithm with $k = 2$ to vectors combining DNA shape feature values extracted from the GBshape database [11] at 1,001 bp-long enhancer regions centered around enhancers' mid-points. We obtained two sets containing 15,259 (set 1) and 23,288 (set 2) enhancers, respectively. These sets of enhancers derived from DNA shape features were very similar to classes 1 and 2 that were obtained using G+C patterns at enhancer regions. Indeed, class 1 and set 1 have a Jaccard similarity of 0.85; class 2 and set 2 have a Jaccard similarity of 0.90.

We plotted the distribution of the four DNA shape features along the enhancer regions from the two classes obtained with the G+C pattern-based clustering (Figure 1c-f). Similarly, we plotted DNA shape features for the two sets obtained from the DNA shape-based clustering (Figure S2). We consistently

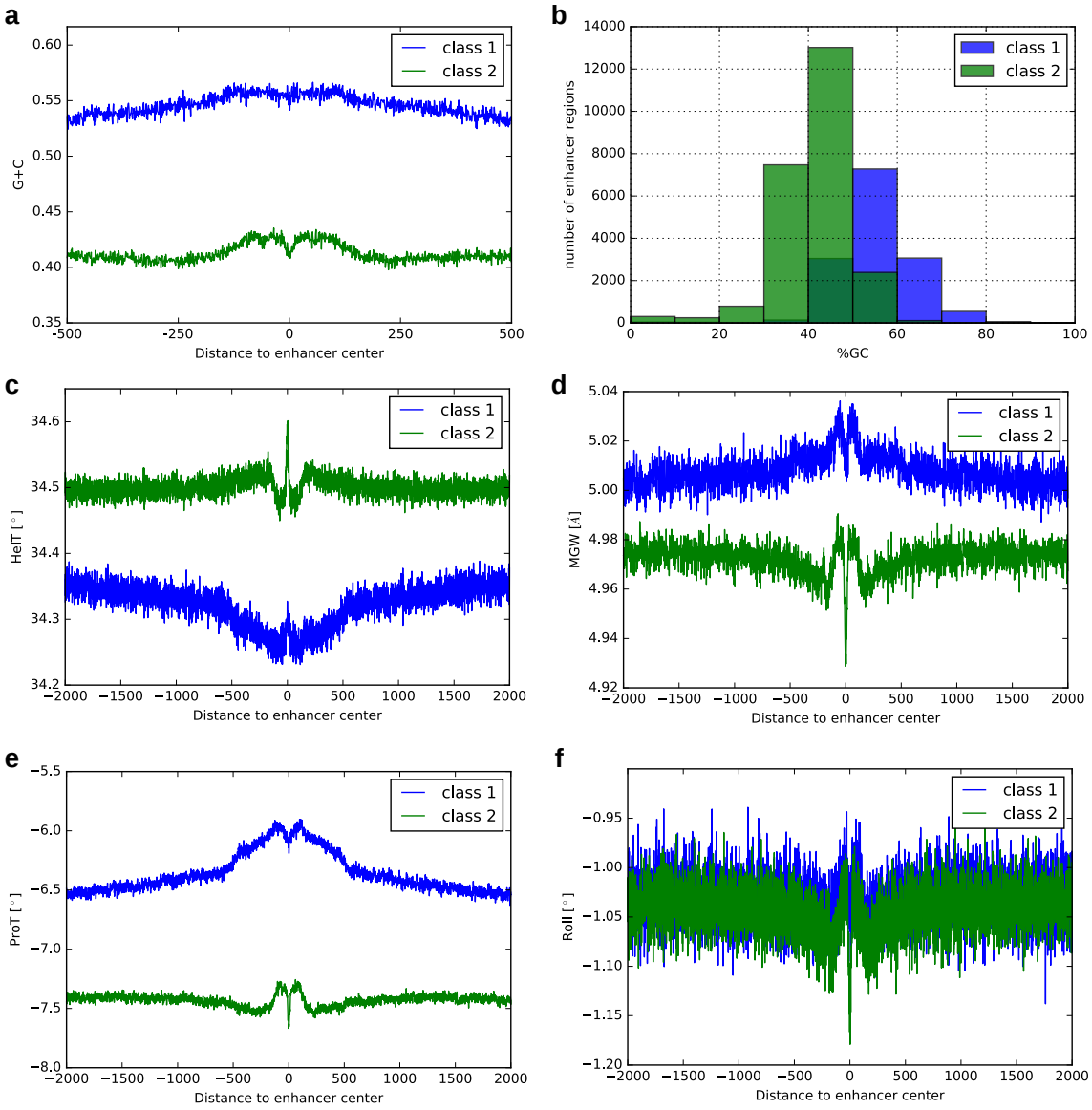

Figure 1: **DNA sequence features at enhancers.** Features associated with human enhancers from class 1 and class 2 are represented in blue and green, respectively. **a.** G+C values (y-axis) of the k-means cluster centers along DNA regions ±500 bp centered at enhancer center points (x-axis). **b.** Histogram of the %GC content of the enhancers. **c-f.** Average DNA shape values (y-axis) along the DNA regions $\pm 2,000$ bp centered at enhancer middle-points (x-axis) for DNA shape features HelT (c), MGW (d), ProT (e), and Roll (f).

observed that class 1 enhancers harboured lower HelT values at the centre of the enhancers as well as about 500 bp away from the enhancers' mid-points (Figure 1c). je observed a symmetrical pattern for MGW with width decrease at the central positions of the enhancers as well as at the edges ($\sim$ 50-150 bp away from the mid-points) of the enhancers (Figure 1d). ProT and Roll signals were also distinct between enhancers from the two classes (Figure 1e-f). The patterns observed for the DNA shape features were in agreement with the two distinct patterns of G+C composition computed along the enhancers from the two classes (Figure 1a).

The similarity between G+C- and DNA shape-based clustering stresses that the G+C pattern is the key discriminant between the two classes of enhancers while the shape represents a secondary effect of the G+C pattern. We therefore focused on the two classes of enhancers derived from their G+C pattern in this report, except otherwise stated. Taken together, these results described two subsets of human enhancers distinguishable by their distribution of G+C along their length and reflected in their DNA structural properties.

The two classes of human enhancers associated with specific biological processes

Different classes of mammalian promoters, derived from their nucleotide composition, were observed to be associated with genes linked to distinct biological functions [20]. Following the same approach, we sought for a functional interpretation of the classification that we obtained. Based on correlations between promoter and enhancer activities derived from CAGE data in human samples, Andersson *et al.* linked enhancers to their potential gene promoter targets [22]. 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. Class 1 enhancers were predicted to target 7,713 genes whereas class 2 enhancers were linked to 7,857 genes (Table S1). In aggregate, the enhancers corresponded to a set of 11,271 genes, of which 4,299 were common to the two classes (representing $\sim 56\%$, $\sim 55\%$ and $\sim 38\%$ of class 1, class 2, and the combined set of genes, respectively). We submitted the two sets of genes associated to class 1 and class 2 enhancers to the GOrilla tool [26] to predict enriched (p-value $< 1 \times 10^{-11}$) gene ontology (GO) biological processes. Note that the aggregated set of 11,271 genes was used as the background set of genes for enrichment analyses.

Biological processes linked to RNA transcription were found to be enriched for genes associated with class 1 enhancers (Figures 2a and S3 and Table S2). Specifically, the directed acyclic graph (Figure S3) of the enriched GO terms highlighted two leafs corresponding to the terms 'transcription, DNA-templated' (FDR q-value $= 8.7 \times 10^{-13}$) and 'regulation of transcription, DNA-templated' (q $= 7.8 \times 10^{-12}$). When considering the genes predicted to be regulated by enhancers from class 2, only two GO biological processes were predicted to be enriched (Figures 2b and S4 and Table S3): 'immune system process' (q $= 6.1 \times 10^{-9}$) and 'regulation of immune response' (q $= 3.2 \times 10^{-8}$).

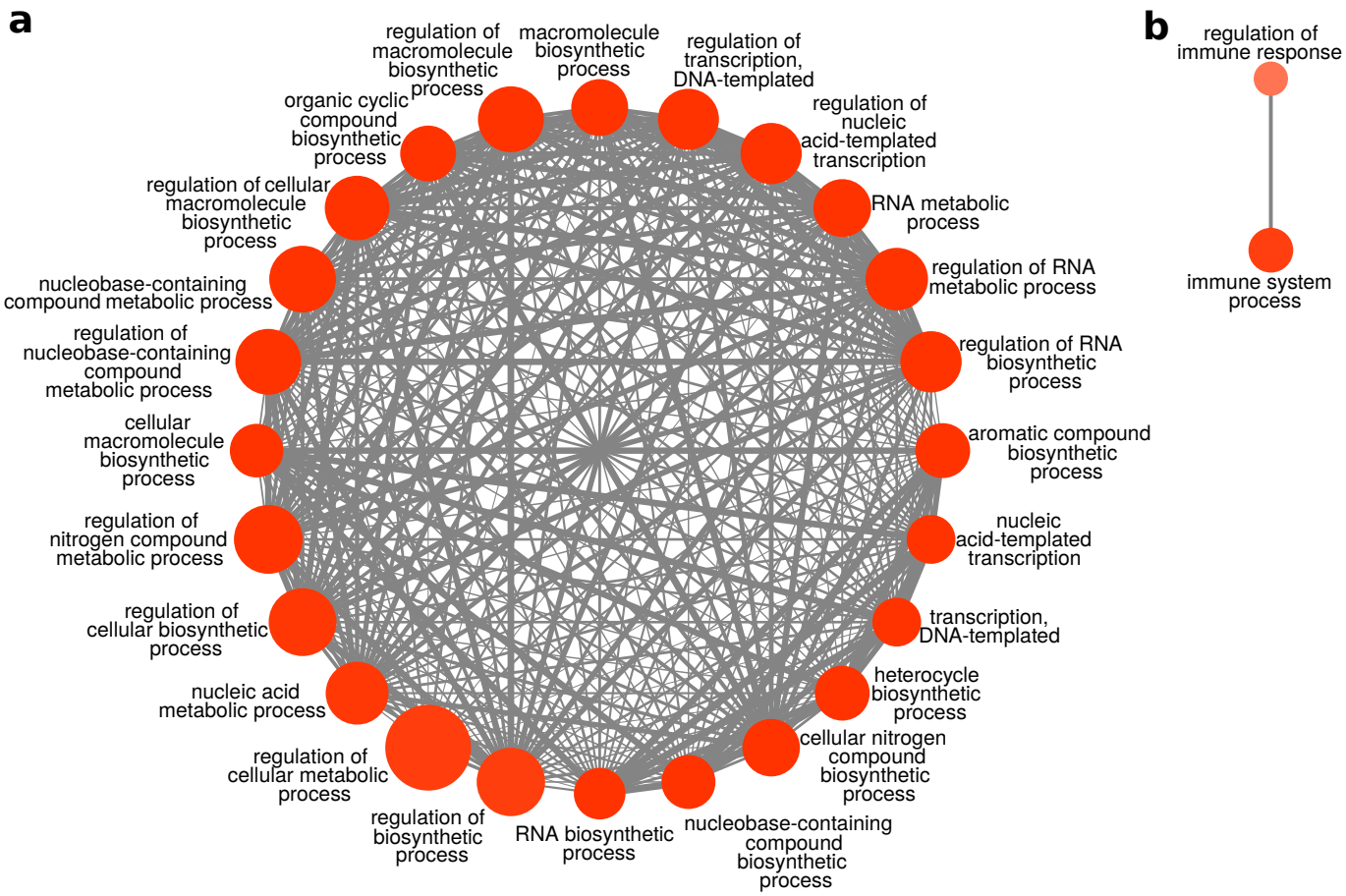

Figure 2: **Functional enrichment analysis.** Enriched GO biological processes associated with genes predicted to be regulated by enhancers from class 1 (**a**) and class 2 (**b**) were obtained using the GOrilla tool [26]. Nodes in the graphs represent enriched GO biological processes. The color of a node represents the FDR q-value of the corresponding enriched GO biological process, the more red, the lower the q-value ($\min = 1.37 \times 10^{-13}$, $\max = 3.8 \times 10^{-8}$). The size of a node represents the number of genes associated with class 1 (a) and class 2 (b) enhancers in the corresponding GO biological process ($\min = 761$, $\max = 3,832$). Edges between two nodes indicate the number of common genes between corresponding processes. The larger the number of overlapping genes ($\min = 264$, $\max = 860$), the larger the edge between the two corresponding nodes. FDR, false discovery rate; GO, gene ontology.

These functional enrichment results were specific to the classification of the enhancers using the pattern of G+C along the enhancer regions. Indeed, we considered a segregation of human enhancer regions solely based on %GC (mean = 46.62, median = 46.15, and standard deviation = 10.56) with GC-poor enhancer regions (%GC $<$ 46.15) assigned to a first set and GC-rich ones (%GC $>$ 46.15) to a second set. We submitted the sets of genes linked to the enhancers from the %GC-based classification to GOrilla and observed that the immune system-related GO terms were found enriched for both sets (Figures S5 and S6).

Taken together, the functional enrichment results revealed that a classification based on the distribution of Gs and Cs along human enhancer regions featured two sets of enhancers predicted to be regulating genes with distinct biological functions. While the first class was linked to genes associated with transcription, the second class highlighted enhancers predicted to regulate immune system related genes.

Distinct transcription factors predicted to act upon the two classes of human enhancers and their predicted promoter targets

We sought to identify TF binding motifs enriched within each class of enhancers, to suggest driving TFs for the distinct biological functions. We considered 1,001 bp-long DNA sequences centered at the enhancers' mid-points. Positional motif enrichment analyses were performed using the Centrimo tool [27] to predict TF binding motifs over-represented around the enhancers' mid-points. Class 1 enhancer regions were compared to class 2 regions and vice-versa to highlight specific motifs (Figure 3a,c and Data S1). Motifs associated with the Specificity Protein/Krüppel-like Factor (SP/KLF) TFs were enriched in class 1 enhancer regions (Figure 3a and Data S1). Members of the SP/KLF family have been associated to a large range of core cellular processes such as cell growth, proliferation, and differentiation [28]. The most enriched motifs in class 2 enhancer regions were associated with nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB)/Rel TFs (Figure 3c and Data S1). As NF-κB is known to have a central role in immune response [29], the enrichment is consistent with an involvement of class 2 enhancers in the immune response biological function (Figure 2b). Other enriched motifs in class 2 enhancers were associated with BACH1/2 TFs, involved in acquired and innate immunity [30], and chromatin remodelling TFs BPTF [31] and SMARCC1 [32].

Linked enhancers and promoters were predicted to be driven by similar sets of TF binding motifs when enhancer-promoter links were derived from a distinct collection of CAGE data from time-course studies [33]. We assessed such associations by extending our positional motif enrichment analyses to the promoters of genes associated with classes 1 and 2 enhancers, respectively. We used Centrimo to predict motifs locally enriched in 1,001 bp regions centered around corresponding TSSs. Motifs associated with SP1 and NF-κB TFs were specifi-

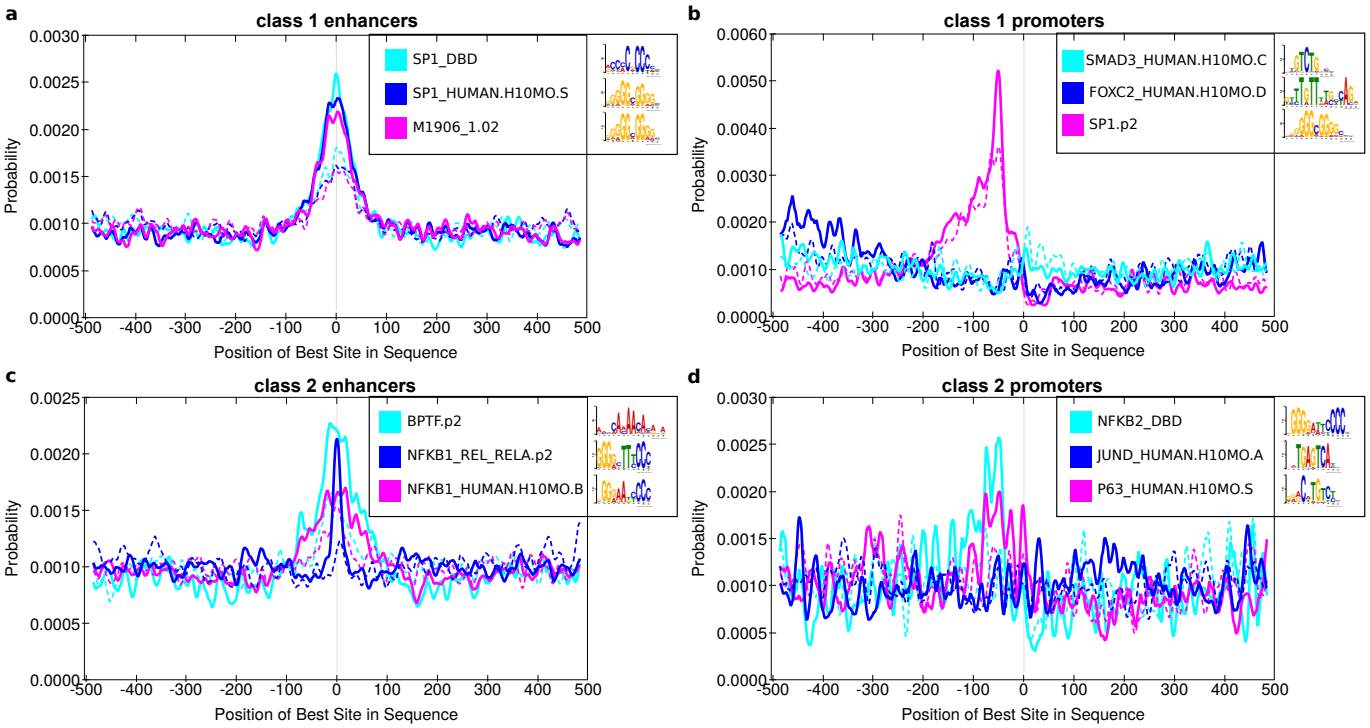

Figure 3: **Motif enrichment analysis at enhancer and promoter regions.** Regions of ±500 bp around enhancer mid-points (**a, c**) and associated genes' TSSs (**b, d**) were subjected to positional motif enrichment analyses using the Centrimo tool [27]. Enhancers and associated gene targets from class 1 (a, b) and class 2 (c, d) were analyzed separately. The x-axis represents the distance to the enhancer mid-point (a, c) and associated gene TSSs (b, d), respectively. 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 in panels a-b and class 2 in panels c-d). Similarly, dashed lines represent the distribution of predicted TFBSs in the background sequences (from class 2 in panels a-b and class 1 in panels c-d). TSSs, transcription start sites; TFBSs, transcription factor binding sites.

cally enriched about 100 bp and 80 bp upstream of TSSs associated with class 1 and class 2 enhancers, respectively (Figure 3b,d and Data S1). It confirmed that promoters predicted to be targets of enhancers shared the same motifs. Centrimo also predicted SMAD3 and FOXC2 motifs in class 1 promoters and JUN and P63 motifs in class 2 promoters, upstream of TSSs.

We confirmed the motif-based enrichment of NF-κB binding in class 2 regions by using ChIP-seq data obtained in GM12878 cells for the RELA TF, which is 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 [34] and Segway [35] tools segment the genome into regions associated to specific chromatin states. Focusing on predictions from ChromHMM and Segway combined, we found 1,802 ($\sim$ 12%) and 2,813 ($\sim$ 11%) active enhancer regions from classes 1 and 2, respectively. We observed that class 2 enhancers were preferentially bound by RELA. Specifically, 591 active class 1 enhancers and 1,226 active class 2 enhancers overlapped RELA ChIP-seq peaks (p-value $= 2.3 \times 10^{-13}$). A similar analysis focusing on predicted promoters identified an enrichment for active promoters in class 2 (8,962, $\sim$ 59%, class 1 and 10,752, $\sim$ 63%, class 2 active promoters, p-value $= 4.5 \times 10^{-9}$). Furthermore, class 2 promoters were preferentially bound by RELA with 1,966 class 1 and 3,179 class 2 active promoters overlapping RELA ChIP-seq peaks (p-value $< 2.2 \times 10^{-16}$).

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 in both enhancers and predicted target promoters.

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 [5, 34–36]. We considered these two approaches. Namely, we considered histone modification data from 6 cell lines and CAGE data from 71 cell types produced by the ENCODE [37] and FANTOM5 [22] projects, respectively.

We retrieved the segmentation of the human genome obtained using a combination of ChromHMM and Segway in the tiers 1 and 2 cell types from ENCODE [37]. For each cell type, we overlapped enhancers with predicted genome segments to assign an activity state to the enhancer. As an example, Figure 4a presents the proportion of enhancers from classes 1 and 2 that were overlapping with segments associated with active, CTCF, and repressed chromatin states in embryonic stem cells (H1-hESC). We consistently observed that enhancers from class 1 were significantly more active than those from class 2, which were found to be enriched in repressed genomic segments (Figures 4a and S7). Class 1 enhancers were also associated with segments characterized by CTCF binding.

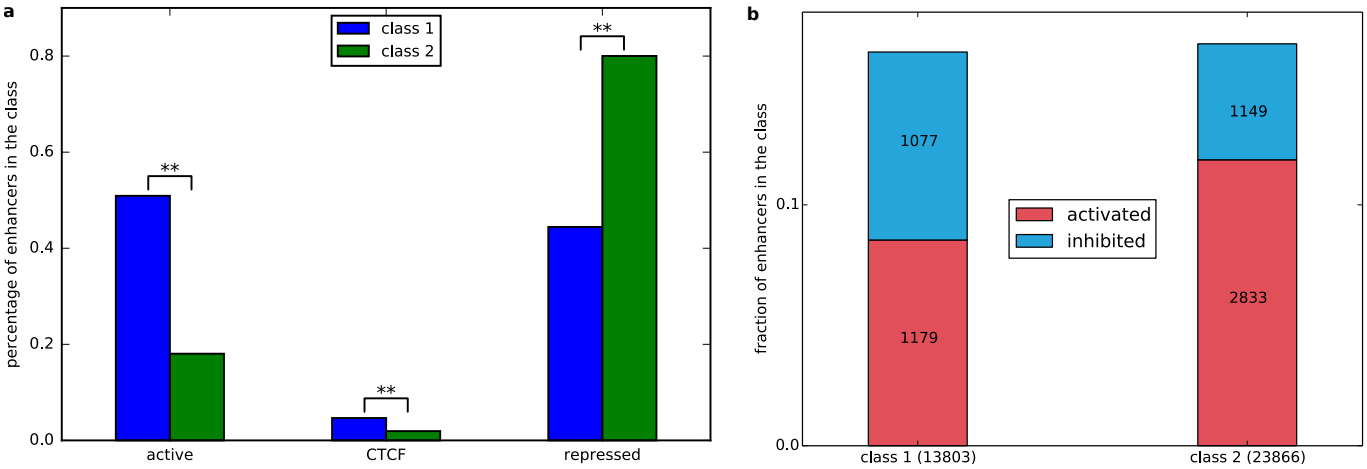

Figure 4: **Human enhancers and genome segmentation.** **a**. Histogram of the proportion of human enhancers (y-axis) in class 1 (blue) and class 2 (green) lying within genome segments (x-axis) as annotated by combined predictions from ChromHMM [34] and Segway [35] on human embryonic stem cells (H1-hESC from the ENCODE project [37]). Statistical significance (Bonferroni-corrected p-value < 0.01) of enrichment for enhancers from a specific class is indicated by '**'. **b**. Stacked histogram of the fraction of human enhancers (y-axis) from class 1 and class 2 predicted to be activated (red) or inhibited (blue). Predictions were obtained using genomic segments predicted by ChromHMM [34] on human dendritic cells before and after infection with *Myobacterium tuberculosis* [38]. Stacked histogram including unchanged activity is provided in Figure S8.

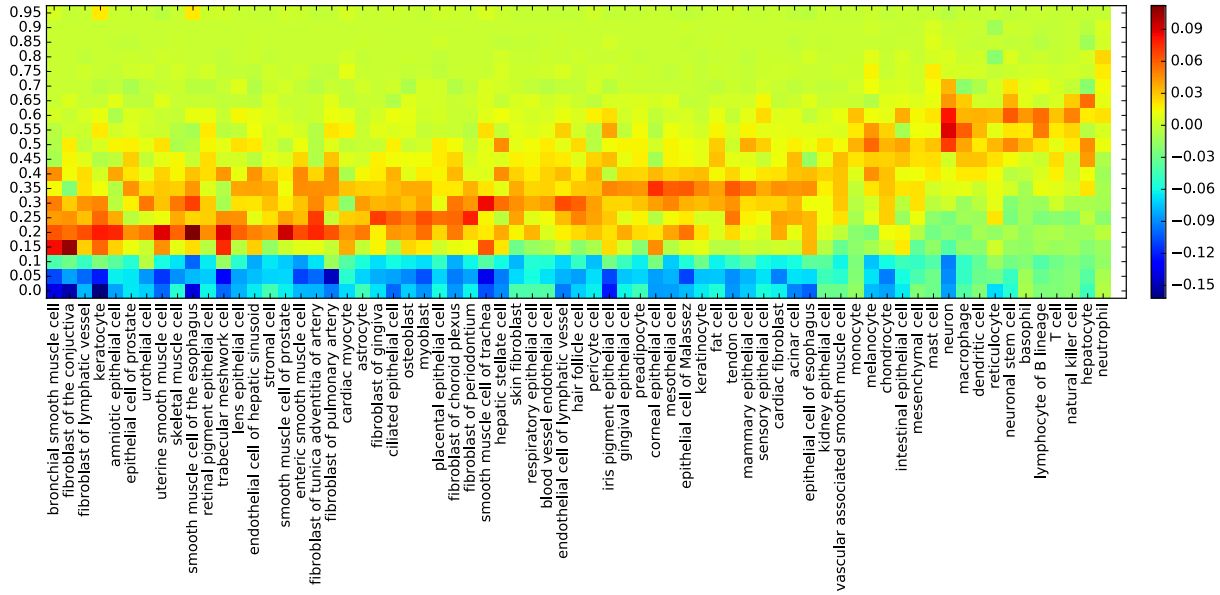

Figure 5: **Cell type expression specificities of human enhancers.** The difference in cell type expression specificities derived from FANTOM5 CAGE datasets [22] for enhancers in class 1 and class 2 is provided as a heat map. The color (see scale) represents the difference in fraction of expressed enhancers in each cell type (columns) found in each expression specificity range (rows). Positive (respectively negative) values indicate a higher fraction of class 2 (respectively class 1) enhancers. The heat maps corresponding to the enhancers in each class are provided in Figure S9. CAGE, Cap Analysis of Gene Expression.

Through CAGE expression analysis in 808 human samples, Andersson *et al.* [22] assigned a cell type-specificity score to human enhancers. Seventy one cell types were defined by grouping cell and tissue samples [22]. Following the enhancer expression specificity analysis performed by Andersson *et al.*, we considered enhancers from class 1 and class 2 separately to highlight potential activity differences in the 71 cell types (Figure S9). Comparing enhancer activity specificity over all the cell types between class 1 and class 2, enhancers from class 2 appeared to be more cell type specific (Figure 5). While immune cells, neurons, neuronal stem cells, and hepatocytes were previously described to use a higher fraction of human enhancers [22], the elevated utilization was even more pronounced for class 2 enhancers (Figures 5 and S9b).

Taken together, these results derived from histone marks and transcriptional data highlighted that enhancers from class 1 were more ubiquitously active over human cell types than enhancers from class 2, 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. Class 1 enhancers were found to be enriched in transcriptional biological processes, which are required for transcription in all cell types. Enhancers from class 2 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 2 enhancers with transcriptional control of immune responses. Pacis *et al.* [38] generated genome-wide DNA methylation, histone marks, and chromatin accessibility data in normal dendritic cells (DCs) and DCs after infection with *Mycobacterium tuberculosis* (MTB). The data provided the opportunity to study the chromatin state changes after infection obtained using the ChromHMM tool [34]. As for the above analysis, 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 (Figures 4b and S8). We observed that the enhancers from class 2 were significantly more activated (p-value $< 2.2 \times 10^{-16}$) and less inhibited (p-value $< 2.2 \times 10^{-16}$) when compared to class 1 enhancers upon MTB infection (Figure 4b). These results reinforced the potential role of class 2 enhancers in immune response.

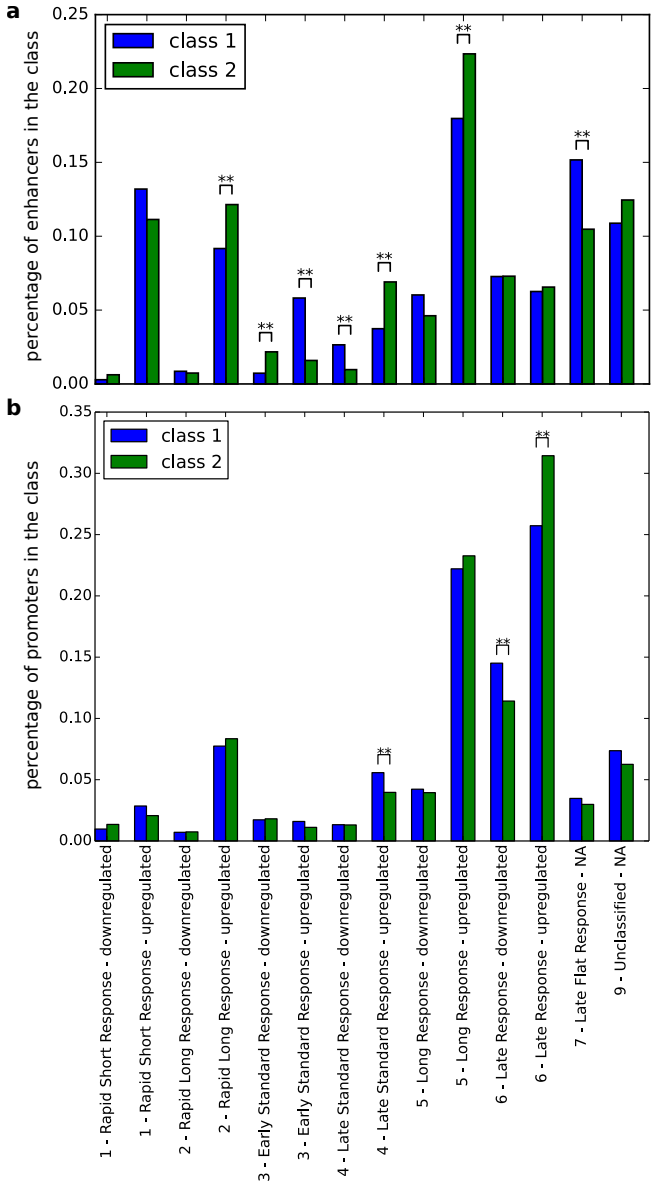

Figure 6: **Expression dynamics of human enhancers and associated promoters.** Response patterns (x-axis) of human enhancers (**a**) and promoters (**b**) in time courses were classified by Arner *et al.* [33]. The percentage (y-axis) of enhancers and promoters from class 1 (blue) and class 2 (green) in each response pattern category are provided as histograms in the two panels. A significant difference (Bonferroni-corrected p-value < 0.01) between class 1 and class 2 enhancers or promoters in a specific category is highlighted by '**'.

Predicted immune system enhancers showed long lasting response activity

Based on time-courses of differentiation and activation, Arner *et al.* analyzed the transcriptional dynamics of enhancers and promoters [33]. They profiled time-courses with CAGE at a high temporal resolution within a 6 hour time-frame to classify enhancers and promoters into distinct dynamic response patterns of early response activity. We overlaid our classification of human enhancers and their predicted target promoters with the dynamic response pattern data (Figure 6). Within the enhancers associated to any dynamic response pattern ($n = 2,694$; 1,533 and 1,161 from class 1 and class 2, respectively), class 1 enhancers were enriched (p-value $< 2.2 \times 10^{-16}$).

We focused on the set of 2,694 enhancers classified in the dynamic response patterns. Looking at the peaks of activity specific to early time points ('rapid short response' and 'early standard response'), class 1 enhancers were found to be upregulated while class 2 enhancers were downregulated (Figure 6a). Enhancers from class 2 showed significant activity dynamics corresponding to long lasting and later responses (up-regulated in 'rapid long response', 'late standard response', and 'long response') when compared to class 1 enhancers (down-regulated in 'late standard response') (Figure 6a). The promoters associated with class 1 were upregulated in the 'late standard response' dynamic. Class 2 promoters exhibited significant up-regulation in the 'late response' dynamic while class 1 promoters were downregulated in the same dynamic.

Taken together, these results identified different dynamics between class 1 and class 2 enhancers. Class 1 enhancers were more dynamic than class 2 enhancers in the FANTOM5 time-course, activated early and for a short period of time while class 2 enhancers harboured long-lasting rapid and late activities. As previously observed [33], the activity of the enhancers were followed by peaks of activity for the associated promoters at later stages (enrichment in late response categories).

Enhancers from the same class co-localized within chromatin domains

The organization of the chromatin in cell nuclei is a key feature in gene expression regulation by forming regulatory region interactions within TADs [8]. Genes within the same TAD tend to be coordinately expressed across cell types and tissues, and clusters of functionally related genes requiring co-regulation tend to lie within the same TADs [8,40]. 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 a set of TADs [7]. 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 classes 1 and 2 labels to the enhancers. The results high-

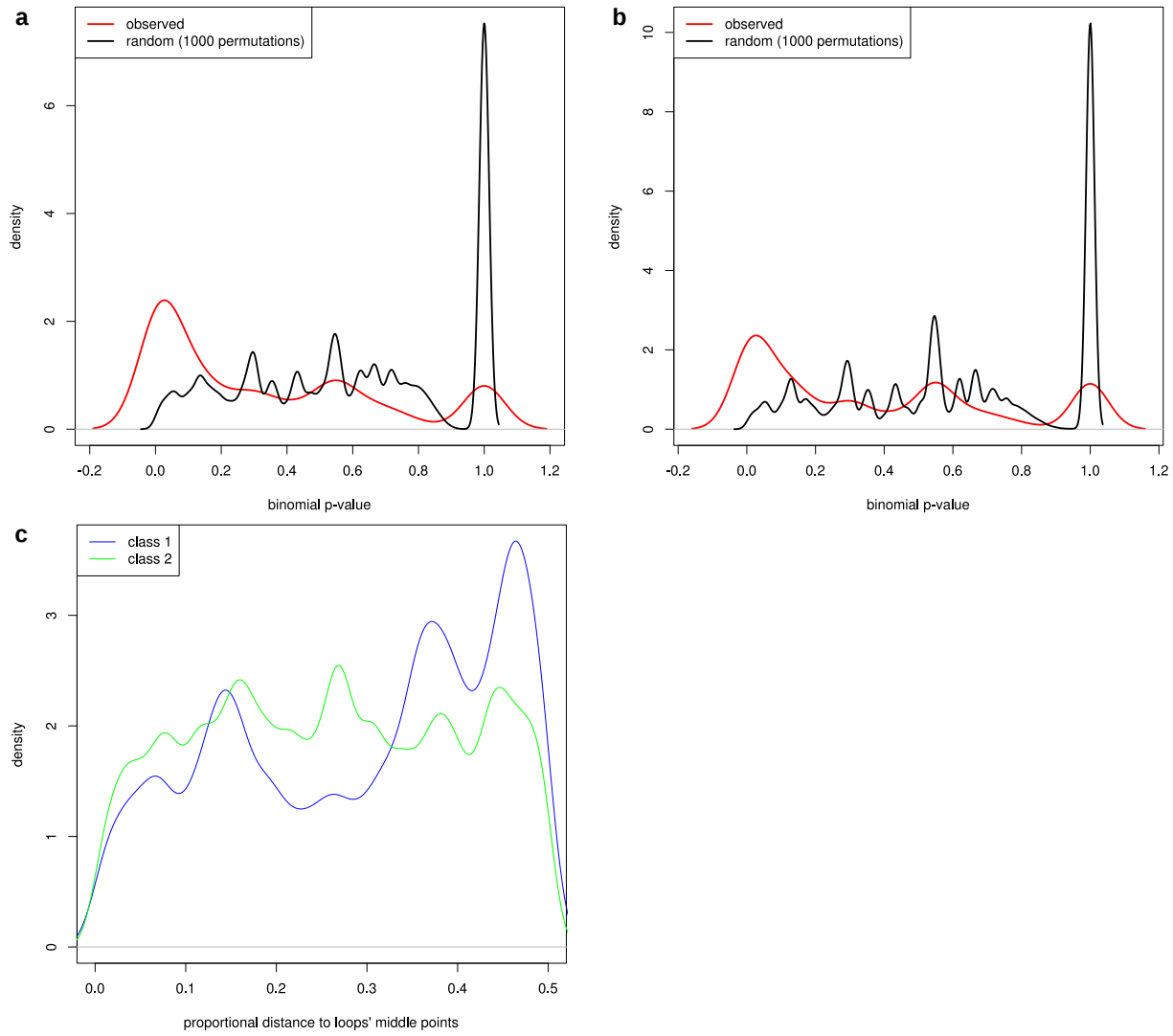

Figure 7: **Chromosomal organization of class 1 and class 2 enhancers. a**. For each TAD [7], we computed the p-value of the Binomial test to assess the enrichment for enhancers from a specific class. The plot compares the density (y-axis) of p-values for Binomial tests (x-axis) applied to classes 1 and 2 enhancers (red) and 1,000 random assignments of class labels to the enhancers (black). **b**. The same analysis as in panel a. was performed using chromatin loops predicted in lymphoblastoid GM12878 cells [39]. **c**. Density (y-axis) of distances (x-axis) between enhancers and chromatin loop centers defined using Hi-C data in GM12878 cells [39]. 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 blue and green, respectively.

lighted that TADs were enriched for enhancers from a specific class (Figure 7a), 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 [39]. We refined our analyses of class-based enhancer co-localization by focusing on chromosomal loops derived from 8 cell lines [39]. Similar to what we observed at the TAD level, we found that chromatin loops tended to contain enhancers from a specific class (Figures 7b and S10). Furthermore, class 2 enhancers were evenly distributed within the chromatin loops whereas enhancers from class 1 were consistently observed to be situated close to the loop boundaries (Figure 7c). This observation is in agreement with the enrichment for class 1 enhancers in CTCF chromatin segments (Figure 4a) as chromatin loop boundaries are known to be enriched for CTCF binding [39].

Discussion

We have analyzed the sequence properties of FANTOM5 human enhancers derived from CAGE experiments to reveal that a subset with low G+C content is associated with immune response genes. This set of enhancers harbours a G+C pattern that corresponds to characteristic HelT, MGW, ProT and Roll conformation of the DNA. The predicted immune system enhancers tend to co-localize within chromatin domains, exhibit cell type specificity, are activated upon infection, and are observed with long lasting response activity. In summary, our study of sequence composition patterns along enhancer regions culminates with the identification of human enhancers associated with immune response that harbour specific sequence composition, activity, and genome organization.

The analyses of sequence properties in regulatory regions, most prominently CpG islands at promoters, have been key to understanding gene expression regulation [9, 10, 20]. The predicted immune response enhancers exhibit a cell-type specific expression pattern and have low %GC. Nevertheless, it remains unclear how and why immune response enhancers have emerged with these sequence properties.

While enhancers predicted to be associated with immune response are GC-poor, a dichotomy of enhancers solely based on GC-content did not highlight a specific set of enhancers associated with immune response genes. This observation might reflect the importance of the DNA sequence, with the DNA shape conformation at enhancer regions as a secondary effect. DNA structural properties were shown to be linked with DNA flexibility, nucleosome positioning, and gene expression regulation [41–45]. The more negative ProT in class 2 (Figure 1e) corresponds to lower GC content [46]. This could relate to 3 hydrogen bonds in G/C pairs versus 2 in A/T pairs, which determines the ability of a base pair to form a ProT angle. Moreover, the differences in DNA shape features between the two classes of enhancers might relate to differences in conformational flexibility. Indeed, we observed class 1 enhancers with less negative ProT, lower

HelT, wider MGW, larger Roll compared to class 2 enhancers (Figure 1c-f). These characteristics all relate to increased flexibility of the DNA [47], which could provide a topological explanation for the differences observed between the two classes.

The classification of human enhancers was performed from basic feature vectors summarizing G+C patterns along enhancer regions. Our observations are reminiscent of the "enhancer–core-promoter specificity" observed in Drosophila [48]. Zabidi *et al.* uncovered two classes of enhancers regulating "housekeeping" versus "developmental" genes, which differ in genomic distribution and in the presence of distinct regulatory elements [48]. Each enhancer appears to have acquired specific DNA features to most effectively regulate the particular promoters it has to regulate [49]. This hypothesis is in agreement with our identification of two classes of enhancers that were defined based on DNA sequence composition and predicted to target promoters associated with genes of distinct biological functions. Similar to Zabidi *et al.* [48], we found distinct genomic organizations between two classes of enhancers, one more ubiquitously active while the other was more cell type-specific.

A recent perspective on immunological memory suggested that transient modifications to chromatin along with inducible noncoding RNAs could mediate a "short-term memory" in immune cells [50]. The principle of this short-term memory is that some histone modifications and regulatory molecules like microRNAs and TFs would be persistent after stimulation, even though limited in time. Mediators of response to stimuli were categorized into (i) labile mediators of activation and (ii) long-lasting mediators of short term memory [50]. It represents a way for both the adaptive and innate immune cells to be more effective in responding to secondary stimulations. Our results suggest that enhancers could also be mediators of the immune response to stimuli lying in the two categories presented by Monticelli *et al.* [50]. Indeed, we observed that class 2 enhancers, associated with immune response genes, were showing a long-term response activity to stimuli, as opposed to short-term patterns of activity for class 1 enhancers (Figure 6). We hypothesize that (i) class 1 enhancers are used for a rapid but short response to stimuli, representing labile mediators of activation and (ii) class 2 enhancers correspond to robust long-lasting mediator of short-term memory to memorize which genes need to be activated after stimulation. Dedicated experiments will be necessary to assess this hypothesis.

Materials and Methods

Human enhancers clusterization

We retrieved the hg19 positions of the 38,554 FANTOM5 robust human enhancers in BED12 format from `http://enhancer.binf.ku.dk/presets/robust_enhancers.bed` [22]. We extracted DNA sequences for regions of 1,001 bp centered at the enhancer mid-points (columns 7-8 of the BED12 file) using the BEDTools [51]. We created binary vectors representing the enhancer sequences

with 1s and 0s corresponding to G or C and A or T, respectively. Note that 7 enhancers were not considered as the 1,001 bp regions contained undefined nucleotides (Ns). The vectors were clustered into $k = 2$ classes using the k-means algorithm implemented in the *KMeans* function of the *scikit* Python module [52]. The silhouette plots (Figure S1) were constructed for $k \in [2, 5]$ using the *silhouette_samples* function of the *scikit* Python module. Formally, the silhouette plots display the silhouette coefficient for each enhancer as $\frac{(b-a)}{max(a,b)}$ where a is the mean intra-cluster euclidian distance and b the mean nearest-cluster euclidian distance.

Similarly, we created vectors representing the enhancers by combining the values of HelT, MGW, ProT, and Roll at the enhancer sequences. The DNA shape values for the hg19 version of the human genome were retrieved as bigWig files from the GBshape database [11]. DNA shape feature values at enhancer regions were obtained by using *bwtool* [53]. It resulted in vectors of 4,004 values each that were submitted to the *Kmeans* function of the *scikit* Python module.

Enhancer gene targets

The enhancer-RefSeq promoter associations were retrieved from `http://enhancer.binf.ku.dk/presets/enhancer_tss_associations.bed` [22]. The corresponding official gene symbols were considered for the functional enrichment analyses.

DNA shape feature plots

The values of DNA structural features HelT, MGW, ProT, and Roll computed using the DNAshape tool [12] were obtained from the GBshape browser [11] 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 [53].

Gene ontology functional enrichment

Official symbols corresponding to the RefSeq promoters associated with enhancers from class 1 and class 2 were submitted to GOrilla [26] at `http://cbl-gorilla.cs.technion.ac.il/` using the March 5th 2016 update. We used the two unranked list option with genes associated with class 1 or class 2 enhancers as targets and the aggregated set of 11,271 genes associated with the full set of enhancers as background. We searched for enriched GO biological processes with the most stringent p-value threshold ($< 10^{-11}$). The DAG representation of the results in Figures S3-S4 were downloaded from the output page of GOrilla. The visual representation of the results in Figure 2 was constructed manually using Cytoscape 3.4.0 [54]. The same procedure has been applied to genes associated with enhancers classified with respect to their GC-content (Figures S2-S3).

Motif enrichment

We applied Centrimo [27] from the MEME suite version 4.11.1 with default parameters to DNA sequences of regions ±500 bp around the mid-points of enhancers from class 1 and class 2. 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 [55] (jolma2013.meme), JASPAR [56] (JASPAR_CORE_2016_vertebrates.meme), Cis-BP [57] (Homo_sapiens.meme), Swiss Regulon [58] (Swiss_Regulon_human_and_mouse.meme), and HOCOMOCO [59] (HOCOMOCOv10_HUMAN_mono_meme_format.meme). The same procedure was applied to promoter regions (±500 bp around TSSs) associated with class 1 and class 2 enhancers.

Figure 3 has been obtained from the html output of Centrimo by selecting the 3 most enriched motifs (ranked using the Fisher E-value). We did not consider inferred motifs in Cis-BP [57].

Genome segmentation

ENCODE genome segmentation

The genome segmentation using the combination of results from ChromHMM [34] and Segway [35] for ENCODE tier 1 and tier 2 cell types GM12878, H1hesc, HelaS3, HepG2, HUVEC, and K562 were retrieved at `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/`.

Genome segmentation in dendritic cells

The genome segmentation of DCs before and after MTB infection [38] was computed using ChromHMM [34] 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 overlap 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 1,001 bp-long regions around enhancer's mid-points and genome segments

predicted by ChromHMM and Segway combined as enhancer or weak enhancer. Similarly, active FANTOM5 promoters were obtained by overlapping 1,001 bp-long regions around TSSs and genome segments predicted as TSS or promoter flanks by ChromHMM and Segway combined. The identified 1,001 bp-long active enhancer and promoter regions were further overlapped with RELA ChIP-seq peaks. All overlaps were computed with the *intersect* subcommand of the BEDTools.

Enhancer expression specificity

The cell-type expression specificity of enhancers was computed as

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

in [22]. 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-S11 in [22]. Heat maps in Figure 5 were computed using the *colormesh* function of the *matplotlib.pyplot* Python module [60].

Enhancer dynamics

FANTOM5 classification in the 14 dynamics displayed in Figure 6 was obtained from Auxiliary data table S3 in [33]. The classification provided response class assignments to 1,533 and 1,161 class 1 and class 2 enhancers, respectively. Response classes were assigned to 2,311 and 2,407 promoters associated with class 1 and class 2 enhancers, respectively. Note that enhancers and promoters can be assigned to multiple response classes.

Corresponding plots (Figures 6) and enrichment analyses were performed using *pandas* Python data structure [61] and the *scipy* Python library [62] in the *IPython* environment [63].

Chromatin conformation data

The enrichment for enhancers associated to 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 [64]. 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 1,000 random trials.

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 [7]. 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 [39] 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`.

Enrichment p-values

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

Acknowledgements

As research parasites [65], 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 for his help with DNA shape data and tools and for insightful discussions and Robin Andersson for his help with FANTOM5 enhancer data. CHL was supported by funding from CNRS, *Plan d'Investissement d'Avenir* and *Institut de Biologie Computationnelle* (Young Investigator grant). AM and WWW were supported by the Genome Canada 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 AM and WWW. AM was also supported by funding from the Norwegian Research Council, Helse Sør-Øst, and the University of Oslo through the Centre for Molecular Medicine Norway (NCMM) and the Oslo University Hospital Radiumhospitalet.

Author contributions

CHL and AM conceived and designed the project. CHL and AM implemented and performed experiments. CHL, WWW, and AM analyzed and interpreted the results. RR was involved in the interpretation of the results relative to the DNA shape features. CHL, WWW, and AM wrote the manuscript.

References

[1] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nature Reviews Genetics*, vol. 5, pp. 276–287, Apr. 2004.

[2] A. Mathelier, W. Shi, and W. W. Wasserman, "Identification of altered cis-regulatory elements in human disease," *Trends in Genetics*, vol. 31, no. 2, pp. 67–76, 2015.

[3] A. Visel, M. J. Blow, Z. Li, *et al.*, "ChIP-seq accurately predicts tissue-specific activity of enhancers," *Nature*, vol. 457, pp. 854–858, Feb. 2009. 00964.

[4] D. Babu and M. J. Fullwood, "3D genome organization in health and disease: emerging opportunities in cancer translational medicine," *Nucleus (Austin, Tex.)*, vol. 6, pp. 382–393, Sept. 2015. 00002.

[5] R. Andersson, A. Sandelin, and C. G. Danko, "A unified architecture of transcriptional regulatory elements," *Trends in genetics: TIG*, vol. 31, pp. 426–433, Aug. 2015. 00013.

[6] E. Lieberman-Aiden, N. L. v. Berkum, L. Williams, *et al.*, "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science*, vol. 326, pp. 289–293, Oct. 2009.

[7] J. R. Dixon, S. Selvaraj, F. Yue, *et al.*, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, pp. 376–380, May 2012. 01078.

[8] J. R. Dixon, D. U. Gorkin, and B. Ren, "Chromatin Domains: The Unit of Chromosome Organization," *Molecular Cell*, vol. 62, pp. 668–680, June 2016.

[9] D. Tillo and T. R. Hughes, "G+C content dominates intrinsic nucleosome occupancy," *BMC Bioinformatics*, vol. 10, p. 442, Dec. 2009.

[10] A. Hughes and O. J. Rando, "Chromatin 'programming' by sequence - is there more to the nucleosome code than %GC?," *Journal of Biology*, vol. 8, p. 96, 2009.

[11] T.-P. Chiu, L. Yang, T. Zhou, *et al.*, "GBshape: a genome browser database for DNA shape annotations," *Nucleic Acids Research*, vol. 43, no. D1, pp. D103–D109, 2015.

[12] T. Zhou, L. Yang, Y. Lu, *et al.*, "DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale," *Nucleic Acids Research*, vol. 41, pp. W56–W62, May 2013.

[13] R. Rohs, S. M. West, A. Sosinsky, *et al.*, "The role of DNA shape in proteinDNA recognition," *Nature*, vol. 461, pp. 1248–1253, Oct. 2009.

[14] T. Zhou, N. Shen, L. Yang, *et al.*, "Quantitative modeling of transcription factor binding specificities using DNA shape," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 4654–4659, Mar. 2015.

[15] M. Slattery, T. Zhou, L. Yang, *et al.*, "Absence of a simple code: how transcription factors read the genome," *Trends in Biochemical Sciences*, vol. 39, pp. 381–399, Sept. 2014.

[16] A. Mathelier, B. Xin, T.-P. Chiu, *et al.*, "DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo," *Cell Systems*, no. 3, pp. 1–9, 2016.

[17] R. Kodzius, M. Kojima, H. Nishiyori, *et al.*, "CAGE: cap analysis of gene expression," *Nature methods*, vol. 3, pp. 211–222, Mar. 2006.

[18] T. Shiraki, S. Kondo, S. Katayama, *et al.*, "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage," *Proceedings of the National Academy of Sciences*, vol. 100, no. 26, pp. 15776–15781, 2003.

[19] P. Carninci, T. Kasukawa, S. Katayama, *et al.*, "The Transcriptional Landscape of the Mammalian Genome," *Science*, vol. 309, pp. 1559–1563, Sept. 2005.

[20] V. B. Bajic, S. L. Tan, A. Christoffels, *et al.*, "Mice and Men: Their Promoter Properties," *PLOS Genet*, vol. 2, p. e54, Apr. 2006.

[21] The FANTOM Consortium and the RIKEN PMI and CLST (dgt), "A promoter-level mammalian expression atlas," *Nature*, vol. 507, pp. 462–470, Mar. 2014.

[22] R. Andersson, C. Gebhard, I. Miguel-Escalada, *et al.*, "An atlas of active enhancers across human cell types and tissues," *Nature*, vol. 507, pp. 455–461, Mar. 2014.

[23] T. R. O'Connor and T. L. Bailey, "Creating and validating cis-regulatory maps of tissue-specific gene expression regulation," *Nucleic Acids Research*, vol. 42, pp. 11000–11010, Jan. 2015.

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

[25] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987.

[26] E. Eden, R. Navon, I. Steinfeld, *et al.*, "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists," *BMC Bioinformatics*, vol. 10, p. 48, Feb. 2009.

[27] T. L. Bailey and P. Machanick, "Inferring direct DNA binding from ChIP-seq," *Nucleic Acids Research*, vol. 40, pp. e128–e128, May 2012.

[28] J. S. Presnell, C. E. Schnitzler, and W. E. Browne, "KLF/SP Transcription Factor Family Evolution: Expansion, Diversification, and Innovation in Eukaryotes," *Genome Biology and Evolution*, vol. 7, pp. 2289–2309, Aug. 2015.

[29] H.-C. Liou, ed., *NF-[kappa]B/Rel transcription factor family.* Molecular biology intelligence unit, Georgetown, Tex. : New York, N.Y: Landes Bioscience/Eurekah.com ; Springer Science+Business Media, 2006. OCLC: ocm68133074.

[30] A. Itoh-Nakadai, R. Hikota, A. Muto, *et al.*, "The transcription repressors Bach2 and Bach1 promote B cell development by repressing the myeloid program," *Nature Immunology*, vol. 15, pp. 1171–1180, Dec. 2014.

[31] M. H. Jones, N. Hamana, and M. Shimane, "Identification and Characterization of BPTF, a Novel Bromodomain Transcription Factor," *Genomics*, vol. 63, pp. 35–39, Jan. 2000.

[32] W. Wang, Y. Xue, S. Zhou, *et al.*, "Diversity and specialization of mammalian SWI/SNF complexes.," *Genes & Development*, vol. 10, pp. 2117–2130, Sept. 1996.

[33] E. Arner, C. O. Daub, K. Vitting-Seerup, *et al.*, "Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells," *Science*, vol. 347, pp. 1010–1014, Feb. 2015.

[34] J. Ernst and M. Kellis, "ChromHMM: automating chromatin-state discovery and characterization," *Nature Methods*, vol. 9, pp. 215–216, Mar. 2012.

[35] M. M. Hoffman, O. J. Buske, J. Wang, *et al.*, "Unsupervised pattern discovery in human chromatin structure through genomic segmentation," *Nature Methods*, vol. 9, pp. 473–476, May 2012.

[36] G. Natoli and J.-C. Andrau, "Noncoding Transcription at Enhancers: General Principles and Functional Models," *Annual Review of Genetics*, vol. 46, no. 1, pp. 1–19, 2012.

[37] I. Dunham, A. Kundaje, S. F. Aldred, *et al.*, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57–74, Sept. 2012.

[38] A. Pacis, L. Tailleux, A. M. Morin, *et al.*, "Bacterial infection remodels the DNA methylation landscape of human dendritic cells," *Genome Research*, vol. 25, pp. 1801–1811, Dec. 2015.

[39] S. S. P. Rao, M. H. Huntley, N. C. Durand, *et al.*, "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping," *Cell*, vol. 159, pp. 1665–1680, Dec. 2014.

[40] J. H. Gibcus and J. Dekker, "The Hierarchy of the 3D Genome," *Molecular Cell*, vol. 49, pp. 773–782, Mar. 2013.

[41] S. C. J. Parker, L. Hansen, H. O. Abaan, *et al.*, "Local DNA Topography Correlates with Functional Noncoding Regions of the Human Genome," *Science*, vol. 324, pp. 389–392, Apr. 2009.

[42] M. Bansal, A. Kumar, and V. R. Yella, "Role of DNA sequence based structural features of promoters in transcription initiation and gene expression," *Current Opinion in Structural Biology*, vol. 25, pp. 77–85, Apr. 2014.

[43] T. Raveh-Sadka, M. Levo, U. Shabi, *et al.*, "Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast," *Nature Genetics*, vol. 44, pp. 743–750, July 2012.

[44] I. Tirosh, J. Berman, and N. Barkai, "The pattern and evolution of yeast promoter bendability," *Trends in Genetics*, vol. 23, pp. 318–321, July 2007.

[45] K. Struhl and E. Segal, "Determinants of nucleosome positioning," *Nature Structural & Molecular Biology*, vol. 20, pp. 267–273, Mar. 2013.

[46] I. Dror, T. Golan, C. Levy, *et al.*, "A widespread role of the motif environment on transcription factor binding across diverse protein families," *Genome Research*, p. gr.184671.114, July 2015.

[47] F. Comoglio, T. Schlumpf, V. Schmid, *et al.*, "High-Resolution Profiling of Drosophila Replication Start Sites Reveals a DNA Shape and Chromatin Signature of Metazoan Origins," *Cell Reports*, vol. 11, no. 5, pp. 821–834, 2015.

[48] M. A. Zabidi, C. D. Arnold, K. Schernhuber, *et al.*, "Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation," *Nature*, vol. 518, pp. 556–559, Feb. 2015. 00049.

[49] D. S. Lorberbaum and S. Barolo, "Enhancers: holding out for the right promoter," *Current biology: CB*, vol. 25, pp. R290–293, Mar. 2015. 00001.

[50] S. Monticelli and G. Natoli, "Short-term memory of danger signals and environmental stimuli in immune cells," *Nature Immunology*, vol. 14, pp. 777–784, Aug. 2013.

[51] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics (Oxford, England)*, vol. 26, pp. 841–842, Mar. 2010.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[53] A. Pohl and M. Beato, "bwtool: a tool for bigWig files," *Bioinformatics*, vol. 30, pp. 1618–1619, June 2014.

[54] P. Shannon, A. Markiel, O. Ozier, *et al.*, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, pp. 2498–2504, Nov. 2003.

[55] A. Jolma, J. Yan, T. Whitington, *et al.*, "DNA-Binding Specificities of Human Transcription Factors," *Cell*, vol. 152, pp. 327–339, Jan. 2013.

[56] A. Mathelier, O. Fornes, D. J. Arenillas, *et al.*, "JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles," *Nucleic Acids Research*, vol. 44, pp. D110–D115, Nov. 2015.

[57] M. T. Weirauch, A. Yang, M. Albu, *et al.*, "Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity," *Cell*, vol. 158, pp. 1431–1443, Sept. 2014.

[58] M. Pachkov, I. Erb, N. Molina, and E. v. Nimwegen, "SwissRegulon: a database of genome-wide annotations of regulatory sites," *Nucleic Acids Research*, vol. 35, pp. D127–D131, Jan. 2007.

[59] I. V. Kulakovskiy, Y. A. Medvedeva, U. Schaefer, *et al.*, "HOCOMOCO: a comprehensive collection of human transcription factor binding sites models," *Nucleic Acids Research*, vol. 41, pp. D195–D202, Nov. 2012.

[60] J. D. Hunter, "Matplotlib: A 2d Graphics Environment," *Computing in Science Engineering*, vol. 9, pp. 90–95, May 2007.

[61] W. McKinney, "Data Structures for Statistical Computing in Python," pp. 51–56, 2010.

[62] E. Jones, T. Oliphant, P. Peterson, *et al.*, "SciPy: Open source scientific tools for Python," 2001–.

[63] F. Perez and B. E. Granger, "IPython: A System for Interactive Scientific Computing," *Computing in Science and Engineering*, vol. 9, no. 3, pp. 21–29, 2007.

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

[65] D. L. Longo and J. M. Drazen, "Data Sharing," *New England Journal of Medicine*, vol. 374, pp. 276–277, Jan. 2016.

Supplementary Figure and Table legends

Figure S1: **Silhouette plots of k-means clusters.** Silhouette plots for clusters obtained using the k-means clusterization algorithm for $k = 2$ (**a**), $k = 3$ (**b**), $k = 4$ (**c**), and $k = 5$ (**d**). After clusterization of the enhancers using the k-means algorithm, the silhouette score was computed for each enhancer vector. For each k-means clusterization corresponding to each panel, clusters are represented with different colors. The scores range from -1 to 1 with -1 indicating a possible assignment of the enhancer vector to the wrong cluster, 0 indicating that the vector is close to the boundary between two clusters, and 1 indicating that the vector is far away from the boundary between two clusters. The silhouette score is calculated as $(b-a)/\max(a,b)$ where a is the mean intra-cluster distance and b the mean nearest-cluster distance for each sample as implemented in the scikit learn *silhouette_score* function. The red dashed lines represent the average silhouette score over all the enhancer vectors (0.019 for $k = 2$, 0.009 for $k = 3$, 0.005 for $k = 4$, and 0.002 for $k = 5$).

Figure S2: **DNA shape features at enhancers from DNA shape-based clusterization.** DNA shape feature values are provided for human enhancers from set 1 and set 2 in blue and green, respectively. Average DNA shape values (y-axis) along the DNA regions $\pm 2,000$ bp centered at enhancer mid-points (x-axis) for DNA shape features HelT (a), MGW (b), ProT (c), and Roll (d).

Figure S3: **Functional enrichment for genes associated to class 1 enhancers.** Directed acyclic graph of the enriched GO biological processed obtained using GOrilla on the set of genes predicted to be regulated by enhancers from class 1.

Figure S4: **Functional enrichment for genes associated to class 2 enhancers.** Directed acyclic graph of the enriched GO biological processed obtained using GOrilla on the set of genes predicted to be regulated by enhancers from class 2.

Figure S5: **Functional enrichment for genes associated to GC-poor enhancers.** Genes predicted to be regulated by %GC poor (%GC < 46.15) were submitted to GOrilla. The figure represents the directed acyclic graph of the enriched GO biological processed obtained.

Figure S6: **Functional enrichment for genes associated to GC-rich enhancers.** Genes predicted to be regulated by %GC poor (%GC > 46.15) were submitted to GOrilla. The figure represents the directed acyclic graph of the enriched GO biological processed obtained.

Figure S7: **Human enhancers and genome segmentation.** Histograms of the proportion of human enhancers (y-axis) in class 1 (blue) and class 2 (green) lying within genome segments (x-axis) as annotated by combined results of ChromHMM [34] and Segway [35] on human lymphoblastoid (GM12878; **a**), cervical cancer (HeLa-S3; **b**), liver carcinoma (HepG2; **c**), umbilical vein endothelial (HUVEC; **d**), and chronic myelogenous leukemia (K562; **e**) cell lines from the ENCODE project [37]. Statistical significance (Bonferroni-corrected p-value < 0.01) of enrichment for enhancers from a specific class is indicated by '**'.

Figure S8: **Enhancer activation upon MTB infection.** Stacked histogram of the fraction of human enhancers (y-axis) from class 1 and class 2 predicted to be activated (red), inhibited (blue), or with unchanged activity (grey). Predictions were obtained using genomic segments predicted by ChromHMM [34] on human dendritic cells before and after infection with Myobacterium tuberculosis [38].

Figure S9: **Cell type expression specificities of human enhancers.** The cell type expression specificities derived from FANTOM5 CAGE datasets [22] is provided as a heat map for human enhancers in class 1 (**a**) and class 2 (**b**). The color (see scale) represents the fraction of expressed enhancers in each cell type (columns) found in each expression specificity range (rows). CAGE, Cap Analaysis of Gene Expression.

Figure S10: **Enrichment of enhancers from a single class within chromatin loops.** For each chromatin loop predicted in HeLa (**a**), HMEC (**b**), HUVEC (**c**), IMR90 (**d**), K562 (**e**), KBM7 (**f**), and NHEK (**g**) cell lines, we computed the p-value of the Binomial test to assess enrichment for enhancers from a single class. The plots compare the density (y-axis) of p-values for Binomial tests (x-axis) applied to classes 1 and 2 enhancers (red) and 1,000 random assignments of class labels to the enhancers (black).

Figure S11: **Organization of enhancers within chromatin loops.** Density (y-axis) of distances (x-axis) between enhancers and chromatin loop centers/anchors. The distances were normalized by the lengths of the loops. Enhancers at the center of the loops were found at distance 0.0 while enhancers at chromatin loops boundaries/anchors were found at distance 0.5. Results associated with class 1 and class 2 enhancers are depicted in blue and green, respectively. HeLa (**a**), HMEC (**b**), HUVEC (**c**), IMR90 (**d**), K562 (**e**), KBM7 (**f**), and NHEK (**g**) cell lines were considered.

Table S1: List of genes associated with class 1 (first column) and class 2 (second column) enhancers derived from Andersson et al. [22].

Table S2: Enriched GO biological processes associated with genes predicted to be regulated by class 1 enhancers.

Table S3: Enriched GO biological processes associated with genes predicted to be regulated by class 2 enhancers.

Data S1: **Centrimo motif enrichment analyses.** Centrimo was applied to regions of 1,001 bp centered around enhancers' mid-points from class 1 (centrimo_class1vs2_enhancers.html) and class 2 (centrimo_class2vs1_enhancers.html). Promoter regions of 1,001 bp centered around TSSs associated with class 1 (centrimo_class1vs2_promoters.html) and class 2 (centrimo_class2vs1_promoters.html) were also subjected to Centrimo. Position enrichment of motifs were focussed around enhancer mid-points for enhancers and all regions for promoters.