bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Genome-wide identification and characterization of HOT regions in the human genome

Hao Li[1], Feng Liu[1], Chao Ren[1], Xiaochen Bo[1*], Wenjie Shu[1*]

[1]Beijing Institute of Radiation Medicine, Department of Biotechnology, Beijing 100850, China

[*]Corresponding author. To whom correspondence should be addressed. Tel & Fax: +86 10 68210077 66932211; Email: shuwj@bmi.ac.cn. Correspondence may also be addressed to: boxc@bmi.ac.cn.

Email addresses:

HL: lihao527_thu@foxmail.com

FL: lf3426@126.com

CR: requiem116@163.com

XB: boxc@bmi.ac.cn

WS: shuwj@bmi.ac.cn

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

HOT (high-occupancy target) regions, which are bound by surprisingly large number of transcription factors, are considered to be among the most intriguing findings of recent years. Improved understanding of the roles that HOT regions play in biology would be afforded by knowing the constellation of factors that constitute these domains and by identifying HOT regions across the spectrum of human cell types. We describe here the population of transcription factors, cofactors, chromatin regulators, and transcription apparatus occupying HOT regions in embryonic stem cells (ESCs) and demonstrate that HOT regions are highly transcribed. We produce a catalogue of HOT regions in a broad range of human cell types and find that HOT regions are associated with genes that control and define the developmental processes of the respective cell and tissue types. We also show evidence of the developmental persistence of HOT regions at primitive enhancers and demonstrate unique signatures of HOT regions that distinguish them from typical enhancers and super-enhancers. Thus, HOT regions play key roles in human cell development and differentiation.

Key words:

HOT regions, cell development and differentiation, bivalent markers, super-enhancers.

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Recent studies have identified a class of mysterious genomic regions that are bound by a suprisingly large number of often functionally unrelated transcription factors (TFs) but lack their consensus binding motifs in *Caenorhabditis elegans* [1, 2], *Drosophila melanogaster* [3-7], and humans [8-10]. These regions are called HOT (high-occupancy target) regions or "hotspots". In *C. elegans*, 22 different TFs were used to identify 304 HOT regions bound to 15 or more TFs [1]. Using the binding profiles of 41 different TFs, nearly 2,000 HOT regions were identified in *D. melanogaster*, each binding an average of 10 different TFs [5]. Many regions that bound to dozens of TFs were also identified in a small number of human cells [9, 10]. The broad presence of these regions in metazoan genomes suggests that they might reflect a general property of regulatory genomes. However, how hundreds of TFs coordinate clustered binding to regulatory DNA to form HOT regions across cell types and tissues is still unclear. Furthermore, the function of HOT regions in gene regulation remains unclear [11, 12], and their proposed roles include functioning as mediators of ubiquitously expressed genes [1], sinks or buffers for sequestering excess TFs [4], insulators [5], DNA origins of replication [5], and patterned developmental enhancers [7]. In addition, it is unknown what effect these regions have on human diseases and cancer. Thus, it is important to systematically analyse HOT regions in a large variety of cell types and tissues and to further understand their functional roles in the control of specific gene expression programs.

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Resolving these challenges requires knowledge of the ensemble of all TF bindings in a cell. However, even predicting where a single TF binds in the genome has proven challenging. Computational motif discovery in regulatory DNA is a commonly used strategy for identifying candidate TF binding sites (TFBSs) for TFs with known binding motifs. Numerous algorithms have been developed for discovering motifs, such as FIMO (Find Individual Motif Occurrences) [13] and HOMER (Hypergeometric Optimization of Motif EnRichment) [14]. It has been reported that TFBSs tend to be DNase I hypersensitive, and only a fraction of the human genome is accessible for TF binding [15]. Remarkably, HOT regions correlate with decreased nucleosome density and increased nucleosome turnover and are primarily associated with open chromatin [1, 5, 6]. DNase I hypersensitive sites (DHSs) in chromatin have been used extensively to mark regulatory DNA and map active *cis*-regulatory elements in diverse organisms [16-18]. Recent advances in Next-Generation Sequencing (NGS) technologies have enabled genome-wide mapping of DHSs in mammalian cells [19-21], laying the foundations for comprehensive catalogues of human regulatory DNA regions. Thus, DHSs, combined with motif discovery algorithms, could be used in a very powerful approach for identifying a large repertoire of TFs in diverse cell and tissue types with high precision. This approach is likely to be widely applicable for investigating cooperativity among TFs that control diverse biological processes.

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Here, we have developed a computational method for the genome-wide mapping of HOT regions in human genome. We have extended our understanding of HOT regions by demonstrating that embryonic stem cell (ESC) HOT regions are highly transcribed; by identifying the population of TFs, cofactors, chromatin regulators, and core transcription apparatus that occupy these regions in ESCs; and by demonstrating the enrichment of known TF motifs in ESC HOT regions. We have created a catalogue of HOT regions for 154 different human cell and tissue types and have shown that these regions are associated with genes encoding cell-type-specific TFs and other components that play important roles in cell-type-specific developmental process. We have shown evidence for the developmental persistence of HOT regions at primitive enhancers and have demonstrated unique signatures of HOT regions that distinguish them from typical enhancers and super-enhancers.

Results

HOT regions in ESCs

Recently, we used Gaussian kernel density estimation across the binding profiles of 542 TFs to identify TFBS-clustered regions and defined a “TFBS complexity” score based on the number and proximity of contributing TFBSs for each TFBS-clustered region (Chen et al., under review). Preliminary inspection of these regions with different TFBS complexity in ESCs revealed an unusual feature: Although the vast majority of TFBS-clustered regions with a median length 5.8 Kb exhibited only low

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

TFBS complexity, a small portion of TFBS-clustered regions spanning as much as 11.9 Kb exhibited TFBS complexity scores greater than 61 (Table S1). Thus, we divided these ESC TFBS-clustered regions into two classes based on TFBS complexity—one class comprised the vast majority of TFBS-clustered regions, which we call LOT (low-occupancy target) regions, and the other encompassed 8533 (10.6%) TFBS-clustered regions with high TFBS complexity, which we call HOT (high-occupancy target) regions (Fig. 1A).

Further characterisation of the ESC HOT regions revealed that they contain many features of LOT regions but at a considerably larger scale (Figs. 1B–D, S1, and Table S1). Previous reports have demonstrated that chromatin modifiers are enriched in enhancer regions. In the present study, we found that the levels of enhancer markers, including histone modifications H3K27ac and H3K4me1 [22, 23] and DNase I hypersensitivity [24], in HOT regions significantly exceed the levels in LOT regions. Similar results were observed for active markers, such as H3K9ac. Interestingly, the permissive histone marker H2AZ was significantly depleted in HOT regions, whereas the repressive marker H4K20me1 was significantly enriched in HOT regions. Strikingly, compared to LOT regions, HOT regions were simultaneously enriched with both permissive histone marker H3K4me3 and repressive marker H3K27me3 signals, which are thought to play an important role in pluripotency by silencing developmental genes in ESCs while keeping them poised for activation upon differentiation [25, 26].

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

RNA polymerase II can transcribe enhancers and produce noncoding RNAs that contribute to enhancer activity [27-31]. We measured the levels of RNA polymerase II in HOT and LOT regions to determine the effect of these regions on transcriptional control. RNA polymerase II was highly enriched in HOT regions relative to LOT regions, which was consistent with RNA signalling levels (Fig. 1C). This result helps to explain why HOT regions drive high-level expression of their associated genes compared to LOT regions (Fig. 1E). Our results suggest that HOT regions could be involved in regulating RNA polymerase II activities and could therefore affect gene expression. Thus, HOT regions may harbour features resembling those of recently identified enhancer RNAs that can contribute to enhancer function [27-29, 32-36].

To further investigate the factors that constitute HOT and LOT regions, we compiled chromatin immunoprecipitation-sequencing (ChIP-seq) data for 13 different chromatin regulators and 30 TFs in ESCs from the ENCODE project [24, 37] (Figs. 1D and S1B, Table S2). It was notable that a broad spectrum of chromatin regulators (12 out of 13, 92%) and transcription regulators (26 out of 30, 87%) that are responsible for cell growth, tissue development, cell cycle progression and developmental events are especially enriched in ESC HOT regions relative to LOT regions, including ATF2, POU5F1, HDAC2, HDAC6, and PHF8. In contrast, four chromatin regulators and TFs, CTCF, RAD21, BCL11A, and MAFK, were significantly enriched in ESC LOT regions relative to HOT regions. Recent studies

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

have revealed that CTCF and RAD21 co-occupy many genomic targets of pluripotency factors in ESCs to play key roles in the control of pluripotency and cellular differentiation [38, 39]. Strikingly, SUZ12 and JARID1 were differentially depleted within HOT and LOT regions. SUZ12, a subunit of PRC2, maintains pluripotency in ESCs by repressing developmental genes that are preferentially activated during ESC differentiation [40]. Recent studies from multiple model organisms, including corn fungus, yeast, *C. elegans*, *Drosophila*, zebrafish, and mice, have demonstrated that JARID1 proteins, as histone H3K4 demethylases, play key roles in development and differentiation [41-43].

Distinct sequence signatures of HOT regions

To gain insight into characteristic sequence features of HOT regions, we studied the enrichment of known TF motifs in HOT and LOT regions using HOMER [14]. Both the genome and the LOT/HOT regions were used as backgrounds in the motif scanning within HOT/LOT regions, respectively. Overall, 226 out of 542 (41.7%) TFs with known motifs exhibited significantly enriched binding in HOT or LOT regions (Fig. 1F, and Table S3). Of these 226 TFs, 59 (26.1%) TFs exhibited specifically enriched binding within HOT regions, relative to the expectations based on the backgrounds of both genome and LOT regions. The majority of these TFs play important roles in development, including MYB, MZF1, TCF7, ZBTB7A/B, HNF4A, POU1F1, PAX2, SRF, XBP1, EGR3 and CREB1, as well as in cell proliferation and differentiation, including RORA, E4F1, MECOM, SP1, RREB1 and FOXM1.

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Thirty-four (15.2%) factors exhibited significantly enriched binding in LOT regions relative to expectations based on the backgrounds of both genome and HOT regions. Strikingly, 12 of these 34 TFs (*p*-value = 0.0012, binomial test) were housekeeping TFs that are associated with the regulation of transcription (NFE2L1, REST, TCF4, NFYC, YY1), protein binding (NFKB1, RBPJ, SMAD4), TF activity (RELA), negative regulation of granulocyte differentiation (RUNX1), multicellular organismal development (TCF12), and the nucleus (SP3). Additionally, we found that a small fraction (8 out of 226, 3.5%) of TFs exhibited specifically enriched binding in both HOT and LOT regions relative to the expectations based on the two backgrounds. These TFs play important roles in development and differentiation, including POU3F2, TCF3, SPY, and MYC, as well as housekeeping roles such as response to oxidative stress, including FOXO1 and NFE2L2.

HOT regions in many cell types

To characterise the HOT regions in as many human cells as possible, we applied a uniform processing pipeline to create a catalogue of HOT regions based on DNase-seq data from 349 samples, including 154 cell and tissue types studied under the ENCODE Project [24, 37] (Fig. 2). We identified an average of 8,036 HOT regions per cell type (range 2,405 to 19,753, Table S4), spanning on average ~1.7% of the genome. In total, we identified 59,986 distinct HOT regions along the genome, collectively spanning 18.8%. To assess the rate of discovery of new HOT regions, we performed saturation analysis as described in a previous study [24] and predicted

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

saturation at approximately 107,184 (standard deviation = 8,608) HOT regions and 774,925,252 bp (standard deviation = 33,534,434) (40.9%) of genome coverage (Fig. 2A). This result indicates that we have discovered more than half of the estimated total HOT regions.

Of these 59,986 HOT regions, 287 localise to UTRs defined by GENCODE, and a collective 9% lie within promoter (*n* = 4,039, 6.7%) and exon (*n* = 1,391, 2.3%) regions. Among the remaining HOT regions, 56.8% (*n* = 34,090) and 33.6% (*n* = 20,179) are positioned in intronic and intergenic regions, respectively (Fig. 2B). We found that HOT regions were more likely localised to genic regions (intron and exon) and less likely localised to intergenic regions compared with LOT regions. HOT regions are typically much more cell-selective than LOT regions (Fig. 2C, 1st column). Promoter proximal HOT regions typically exhibit high accessibility across cell types, with the average proximal HOT region detected in 21 cell types; however, distal HOT regions are largely cell selective, with the average distal HOT region detected in 7 cell types (Fig. 2C, 2nd and 3rd columns).

Gene Ontology (GO) analysis of HOT regions

We next performed GO analysis on HOT region-associated genes (HOT genes). This analysis revealed that HOT genes are linked to developmental processes of the respective cell and tissue types (Fig. 2D). To gain further understanding of the transcriptional regulatory circuitry of development, it would be valuable to identify

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

key developmental TFs that control this process. As the majority of HOT genes are involved in developmental processes, we deduced that candidate key developmental TFs could be identified in human cells by identifying HOT genes that encode TFs. We then performed this analysis in all of the 154 human cells. For cells in which key developmental TFs have already been identified, this analysis captured the vast majority of these factors (Fig. 2E and Table S5). A catalogue of candidate key developmental TFs for other cell types can be found in Table S6. These candidates will be helpful in deducing the transcriptional regulatory circuitry of diverse human cells and in further understanding cell development and cell differentiation.

Associations with validated regulatory elements

To gain understanding of the functional roles of HOT regions, it would be valuable to explore their associations with previously validated regulatory elements. First, we explored the extent to which HOT regions associate with microRNAs, which comprise a major class of regulatory molecules and have been extensively studied, resulting in the consensus annotation of hundreds of conserved microRNA genes [44]. Of 2,633 annotated microRNA transcriptional start sites (TSSs), 1,667 (63%) coincide with a HOT region. The accessibility of HOT regions at microRNA promoters was highly promiscuous compared with GENCODE TSSs (Fig. 2C, 4th column) and showed cell lineage organisation, paralleling the known regulatory roles of well-annotated lineage-specific microRNAs (Fig. 3A). Next, we investigated the association between HOT regions and transposon sequences. A surprising number of

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

these sequences contain highly regulated HOT regions (Fig. 2C, 5th column, and Table S7), which is compatible with the cell type-specific transcription of repetitive elements detected using ENCODE RNA sequencing data [45]. The examples shown in Figure 3B also illustrate the strong cell-selectivity of chromatin accessibility observed for each major repeat class. Finally, we compared HOT regions with an extensive compilation of 373 experimentally validated distal, non-promoter *cis*-regulatory elements, such as insulators, locus control regions (LCRs), transcription initiation platforms (TIPs), and more (Fig. 3C). This analysis revealed that the overwhelming majority (76%) of these elements are encompassed within HOT regions (Table S8), typically with strong cell selectivity (Fig. 2C, 6th column).

Developmental persistence of HOT regions at embryonic enhancers

As HOT regions drive genes that control and define cell development, it is reasonable to surmise that HOT regions could be persistently associated with enhancers active during early development in definitive cells. We compiled 882 early developmental enhancers that were identified through comparative genome analysis and experimental validation of *in vivo* enhancer activity in transgenic mice [46]. Each of these enhancers displays reproducible tissue-staining patterns in one or more embryonic tissues at embryonic day 11.5 (Fig. 4A). Of these 882 non-promoter human enhancers, a surprising proportion (308/882, 35%) occur within HOT regions in at least one definitive human cell type. To quantify the tissue activity spectra of these embryonic enhancers, we systematically examined their lacZ expression

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

patterns in transgenic mice and related these patterns to HOT region patterning at the same elements across different definitive cell types (Fig. 4B). For example, an enhancer that is selectively active in embryonic forebrain tissue (Fig. 4A, 1st column) is selectively found in HOT regions within cells derived from human ESCs (Fig. 4B, 1st column), and an enhancer that is selectively active in embryonic blood vessels (Fig. 4A, 2nd column) is selectively found in HOT regions within endothelial cells (Fig. 4B, 2nd column). In contrast, an enhancer with extremely broad tissue activity (Fig. 4A, 7th column) is found in HOT regions in nearly all definitive cell types (Fig. 4B, 7th column). These findings were further confirmed across the spectrum of enhancers (Figs. 4C–D). A total of 62.5% of enhancers active in embryonic blood vessels are found in HOT regions of endothelial cells, whereas only 26.3% of all other embryonic enhancers are found in endothelial HOT regions. Similarly, 59.3% of enhancers active in embryonic heart tissue are found in HOT regions within cells derived from human heart and great vessel structures, whereas only 30.7% of all other embryonic enhancers are found in HOT regions in these cell types.

Distal HOT regions in super-enhancers

Recently, Richard A. Young and his colleagues identified an unusual class of enhancer domains, called super-enhancers, that drive the high-level expression of genes that control and define cell identity and disease [47-49]. To elucidate the relationship between HOT regions and super-enhancers, we collected the HOT regions and super-enhancers from the same 14 cell types. We found that super-enhancers were

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

highly enriched in HOT regions compared to LOT regions (Fig. 5A, *p*-value < 10^{-4}, binomial test). To determine whether HOT regions might cooperate with super-enhancers to regulate cell type-specific gene regulation, we performed colocalisation analysis of these two types of regions in 14-by-14 cell line combinations, as previously described [50, 51] (Fig. 5B). The diagonal-matched cell line enrichment values (> 1.00 for all comparisons) were much larger than the off-diagonal mismatched cell line values (< 1.00 for all comparisons), indicating that cell type-specific HOT regions tend to strongly colocalise with super-enhancers that function in the corresponding cell types. Furthermore, we compared the densities of chromatin markers, TFs, and RNA polymerase II between HOT regions, enhancers, and super-enhancers. All of these elements exhibited similar DNase I hypersensitivity. As expected, enhancer markers, such as H3K27ac, H3K4me1, and P300, were significantly enriched within enhancers and super-enhancers compared to HOT regions. In addition, RNA Pol II was significantly enriched within enhancers and super-enhancers compared to HOT regions. Notably, HOT regions demonstrated simultaneous significant enrichment of the bivalent markers H3K4me3 and H3K27me3, whereas enhancers and super-enhancers showed both enrichment of H3K4me3 and depletion of H3K27me3 compared to the background genome (Fig. 5C). Finally, we characterised super-enhancer-associated and HOT region-associated genes by GO analysis. Our results revealed that super-enhancer-associated genes are linked to biological processes that largely define the identities of the respective cell and tissue types, which is well consistent with previous study [47]. However, HOT

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

region-associated genes are linked to biological processes that largely define the development and differentiation of the respective cell and tissue types (Fig. 5D).

Discussion and Conclusions

Previous studies have revealed regions in worms [1, 2], flies [3-7], and humans [8-10] with heavily clustered TF binding, termed HOT regions. These reports [1-10] identified HOT regions by the binding peaks of many TFs using ChIP-seq data, whereas we defined HOT regions by a large number of TF motif binding sites on DHSs in DNase-seq data. Although the identifications of HOT regions were based on different data, both definitions demonstrate that HOT regions are a novel class of genomic regions that are bound by a surprisingly large number of TFs and TF motifs. Importantly, our identification of HOT regions using TF motif discovery on DHSs can greatly extend the repertoire of both TFs and cell types in the genome, thus greatly enhancing our understanding of HOT regions.

We have extended our understanding of HOT regions by demonstrating that ESC HOT regions are highly transcribed and by identifying the population of TFs, cofactors, chromatin regulators, and core transcription apparatus that occupy these domains in ESCs. ESCs were chosen for identifying components of HOT regions because the TFs, cofactors, chromatin regulators, and noncoding RNAs that control the ESC state and that contribute to the gene expression program of pluripotency and

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

self-renewal are likely better understood than those for any other cell type [52-54]. HOT regions are occupied by a large portion of enhancer-associated RNA polymerase II and its associated cofactors and chromatin regulators, which may explain how these molecules contribute to the high-level transcription of associated genes. Furthermore, the levels of RNA detected in HOT regions vastly exceed the levels of RNA in LOT regions, and recent evidence suggests that these enhancer RNAs (eRNAs) may contribute to gene activation [27-29, 32-36]. Several additional important insights were gained by studying how more than 40 TFs, cofactors, chromatin regulators, and components of the core transcription apparatus occupy HOT regions and LOT regions in ESCs. All of the enhancer-binding TFs are enriched in HOT regions, with some so highly enriched that they distinguish HOT regions from LOT regions.

By uncovering characteristic sequence signatures of HOT regions, our computational analysis revealed that more than one quarter of enriched TFs exhibited significantly enriched binding within HOT regions, and the majority play essential roles in development and cell differentiation. Strikingly, 12 of 34 TFs (p-value = 0.0012, binomial test) that specifically enriched binding within LOT regions were housekeeping TFs. Our findings, combined with previous observations that HOT regions are depleted in the bound TFs' motifs [1, 3-5] compared with regions bound by single TFs, suggest that HOT regions have distinct sequence features distinguishing them from LOT regions and the genome background, as well as suggesting that information on HOT regions is encoded in the DNA sequence.

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We have generated a catalogue of HOT regions and their associated genes in a broad spectrum of human cell and tissue types. HOT regions tend to be cell type-specific, and the genes associated with these elements are linked to biological processes that largely define the development and differentiation of the respective cell and tissue types. Genes that encode candidate key developmental TFs and noncoding RNAs such as microRNAs are among those associated with HOT regions. Thus, the HOT region catalogue should be a valuable resource for further study of transcriptional control of cell development and differentiation [55-58].

Association analysis between HOT regions and embryonic enhancers presents direct evidence of the systematic developmental persistence of HOT regions at tissue-selective early developmental enhancers and of the persistent imprint of enhancer roles on the formation of cross-cell-type patterning of HOT regions in definitive cells. Additionally, we found that super-enhancers were highly enriched in HOT regions relative to LOT regions, and cell type-specific super-enhancers tend to strongly colocalise with HOT regions that function in the corresponding cell types. Furthermore, all enhancer markers, including DNaseI, H3K27me3, H3K4me1, enhancer-binding TFs and chromatin regulators, are enriched at HOT regions but have lower levels of enrichment that distinguish them from super-enhancers. Strikingly, we observed the paradoxical coexistence of permissive and repressive histone marks, H3K4me3 and H3K27me3, in HOT regions. GO analysis revealed that

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

super-enhancers and HOT regions both drive the expression of genes that define cell identity and cell development of the respective cell and tissue types. Together, our results suggest that HOT regions might therefore represent a novel class of enhancers because they contain many features of enhancers or super-enhancers but at a smaller scale. The activities of HOT regions and super-enhancers are both defined by colocalisation between TFs in these regions but on different genomic scales of colocalisation. A recent study [59] described the relationship between hotspots and super-enhancers in the early phase of adipogenesis, demonstrating that hotspots are highly enriched in large super-enhancers and revealing that hotspots and super-enhancers function as two levels of regulatory hubs that serve to integrate external stimuli through cooperativity between TFs on chromatin. These findings are highly consistent with ours.

Materials and Methods

Data sets

The DNaseI Hypersensitivity by Digital DNaseI data were obtained from the Duke and UW ENCODE groups. Histone modifications according to ChIP-seq data were downloaded from the Broad histone ENCODE group. TFs according to ChIP-seq data were obtained from the HAIB and SYDH TFBS ENCODE groups. Gene annotations were obtained from the GENCODE data (V15). All these data were provided through the ENCODE Project [24, 37], and use of the data strictly adheres to the ENCODE

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Consortium Data Release Policy.

Identifying TFBS-clustered Regions and HOT regions

Position-specific weight matrices of 542 TFs, which corresponded to 796 motif models, were collected from the TRANSFAC [60], JASPAR [61], and UniPROBE [62] databases. The genomic sequence under DHSs from the hg19 genome was used as input for iFORM (Chen et al., in preparation) with a custom library of all 796 motifs scanned for motif instances at a *p*-value threshold of 10^{-18} (corresponding to the FIMO threshold of 10^{-5}). For each TF, motif instances were combined to generate its TFBSs.

An established method [5] was used to perform Gaussian kernel density estimations across the genome (bandwidth 3 kb, centred on each TFBS). Each peak of the density profile was denoted as a TFBS-clustered region. To determine the complexity of each TFBS-clustered region, the Gaussian kernalised distance from a peak to each TFBS that contributed at least 0.1 to the strength was determined. The window around each TFBS-clustered region was derived by finding the maximum distance (in bp) from the TFBS-clustered region to a contributing TF and adding 1.5 kb (one half of the bandwidth). Each window was centred on a TFBS-clustered region.

To identify HOT regions, we first ranked all the TFBS-clustered regions in a cell type by increasing and plotting TFBS complexity (Fig. 1A). This plot revealed a clear

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

point in the distribution of the TFBS-clustered regions at which the complexity signal began increasing rapidly. To geometrically define this point, we first scaled the data such that the *x* and *y* axes were from 0–1. We then found the *x* axis point for which a line with a slope of 1 was tangent to the curve. We defined the TFBS-clustered regions above this point to be HOT regions and the TFBS-clustered regions below that point to be LOT regions. The pipeline for identifying HOT or LOT regions was applied uniformly to datasets from 349 samples, including 154 cell types studied under the ENCODE Project [24, 37]. The classification of the TFBS-clustered regions in each cell type as a HOT or LOT region can be found in Table S4 for diverse human cells and tissues.

Characterisation of HOT Regions

The genome-wide ChIP-seq densities of TF and histone modifications around HOT regions and LOT regions (Figs. 1C, 1D, 3D and S1B) were created by mapping reads to these regions and their corresponding ±5 kb flanking regions. Each HOT/LOT region and its flanking regions were split into 50 equally sized bins. This procedure split all HOT/LOT regions, regardless of their size, into 150 bins. All HOT/LOT regions were then aligned, and the average ChIP-seq density in each bin was calculated to create a genome-wide average in units of reads per kilobase per million (rpkm).

To find sequence motifs enriched in HOT and LOT regions, we analysed the genomic

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

sequences under the DHSs within these regions. HOMER [14] was used to examine whether any of the 542 non-redundant TFs from TRANSFAC [60], JASPAR [61], and UniPROBE [62] were overrepresented with default parameters. Overrepresentation was statistically evaluated using three independent background sets: the entire chromosome 20, all the RefSeq transcription start sites (TSSs) (±2.0 kb), and all the CpG islands annotated in the hg19 genome. A motif was retained only when it was significantly overrepresented ($P \leq 0.01$) compared to all these backgrounds.

Master list and annotation for HOT regions

HOT regions from 154 cell types were consolidated into a master list of 59,986 unique, non-overlapping HOT region positions by first merging these regions across cell types. Then, for each resulting interval of merged regions, the HOT region with the highest TFBS complexity was selected for the master list. Any HOT regions overlapping the regions selected for the master list were then discarded. The remaining HOT regions were merged, and the process was repeated until each original TFBS-clustered region was either incorporated into the master list or discarded.

Genomic annotations from GENCODE annotations (V15) [63], i.e., Basic, Comprehensive, PseudoGenes, 2-way PseudoGenes, and PolyA Transcripts, were used. The promoter class for each GENCODE-annotated TSS was defined as a region from the master list within 1 kb of the TSS. The exon class was defined as any HOT region not in the promoter class that overlapped a GENCODE-annotated "CDS"

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

segment by at least 75 bp. The UTR class was defined as a HOT region not in the promoter or exon class that overlapped a GENCODE-annotated "UTR" segment by at least 1 bp. The intron class was defined as GENCODE segments annotated as "gene" with all "CDS" segments. The intron class also covered any HOT regions not defined by other categories that overlapped introns by at least 1 bp.

The cell-type number was defined for each HOT region by annotating the master list with the number of cell-types with overlapping HOT regions. Plots in Figure 2B were generated using the R function "geom_violin" from the "ggplot2" package, which summarises the distribution of cell-type numbers for distinct categories of HOT regions. The distribution of cell types containing a HOT region was calculated separately for HOT regions observed in 154 cell types.

Gene Ontology Analysis

For gene ontology (GO) analysis, a subset of 19 data sets, representing the diversity of cells in the collection used for this study, were first selected. Each HOT region was assigned to the closest genes annotated in the GENCODE (V15) by determining the distance from the centre of the HOT region to the TSS of each GENCODE gene. For each cell, the genes associated with HOT regions in that cell and no more than six other cells in the subset were analysed using Database for Annotation, Visualization and Integrated Discovery (DAVID) [64]. For each cell, the four top scoring categories (i.e., the categories with the lowest *p*-values) were selected for display. A threshold

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

p-value score of 10^{-6} was incorporated as a minimum requirement filter for scoring as a top category.

Analysis of microRNAs, RepeatMasker, and *cis*-regulatory elements

The microRNA coordinates were downloaded from miRBase (version 20) [44] and used to map microRNAs to their genomic locations. We used the method described in a recent study [15] to assign TSSs for 2633 microRNA loci.

RepeatMasker data were downloaded from the hg19 rmsk table associated with the UCSC Genome Browser. There are 1395 distinctly named repeats in 56 families in 21 repeat classes. The data were analysed by repeat family because this procedure gives a granularity suitable for display. A number of the classes are structural classes rather than classes derived from transposable elements. Bedops utilities [65] were used to count the number of repeat elements that overlapped at least 1 bp with HOT regions. The HOT regions from 154 cell types/tissues were tested for overlap with repeat families. Supplementary Table 7 shows overlap statistics for families of elements with at least 5000 overlapping HOT regions.

Additionally, an extensive compilation of 373 experimentally validated distal, non-promoter *cis*-regulatory elements, including insulators, locus control regions, and so on, were taken from a recent study [15] (Table S8).

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Comparison of HOT regions with known enhancers

Data for tests of human enhancers in a mouse developmental model [46, 66] were downloaded from http://enhancer.lbl.gov/. Embryonic mouse images were downloaded from the VISTA enhancer browser (http://enhancer.lbl.gov/).

To calculate the enrichment of super-enhancers in a HOT region relative to a LOT region in Figure 5A, we first counted the numbers of super-enhancers that overlapped with the HOT and LOT regions and normalised them to the total number of HOT and LOG regions, respectively. We then calculated the log2 ratios of the normalised results to show the enrichment of super-enhancers in HOT and LOT regions.

To perform colocalisation analysis on HOT regions and super-enhancers in *N*-by-*N* (*N* = 14) cell line combinations as similarly described in a previous study [50, 51], we collected a catalogue of super-enhancers for 15 human cells from recent studies. The counts were divided by the corresponding row sum and column sum and multiplied by the matrix sum to obtain enrichment values, which was conducted using the same approach as the χ^2 test. We plotted the enrichment factor for each histone modification in a *N*-by-*N* heat map.

Accession numbers

The identified HOT regions across human cell and tissue types have been deposited

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

with the Gene Expression Omnibus under the accession ID GSE54296.

References

1. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K *et al*: **Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project**. *Science (New York, NY)* 2010, **330**(6012):1775-1787.
2. Araya CL, Kawli T, Kundaje A, Jiang L, Wu B, Vafeados D, Terrell R, Weissdepp P, Gevirtzman L, Mace D *et al*: **Regulatory analysis of the C. elegans genome with spatiotemporal resolution**. *Nature* 2014, **512**(7515):400-405.
3. Moorman C, Sun LV, Wang J, de Wit E, Talhout W, Ward LD, Greil F, Lu XJ, White KP, Bussemaker HJ *et al*: **Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(32):12027-12032.
4. MacArthur S, Li XY, Li J, Brown JB, Chu HC, Zeng L, Grondona BP, Hechmer A, Simirenko L, Keranen SV *et al*: **Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions**. *Genome biology* 2009, **10**(7):R80.
5. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF *et al*: **Identification of functional elements and regulatory circuits by Drosophila modENCODE**. *Science (New York, NY)* 2010, **330**(6012):1787-1797.
6. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R *et al*: **A cis-regulatory map of the Drosophila genome**. *Nature* 2011, **471**(7339):527-531.
7. Kvon EZ, Stampfel G, Yanez-Cuna JO, Dickson BJ, Stark A: **HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature**. *Genes & development* 2012, **26**(9):908-913.
8. Yan J, Enge M, Whitington T, Dave K, Liu J, Sur I, Schmierer B, Jolma A, Kivioja T, Taipale M *et al*: **Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites**. *Cell* 2013, **154**(4):801-813.
9. Chen RA, Stempor P, Down TA, Zeiser E, Feuer SK, Ahringer J: **Extreme HOT regions are CpG-dense promoters in C. elegans and humans**. *Genome research* 2014, **24**(7):1138-1146.
10. Foley JW, Sidow A: **Transcription-factor occupancy at HOT regions quantitatively predicts RNA polymerase recruitment in five human cell lines**. *BMC genomics* 2013, **14**:720.
11. Furlong EE: **Molecular biology: A fly in the face of genomics**. *Nature* 2011, **471**(7339):458-459.
12. Blaxter M: **Genetics. Revealing the dark matter of the genome**. *Science (New York, NY)* 2010, **330**(6012):1758-1759.
13. Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif**. *Bioinformatics (Oxford, England)* 2011, **27**(7):1017-1018.
14. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK:

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* 2010, **38**(4):576-589.

15. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B *et al*: **The accessible chromatin landscape of the human genome**. *Nature* 2012, **489**(7414):75-82.
16. Gaszner M, Felsenfeld G: **Insulators: exploiting transcriptional and epigenetic mechanisms**. *Nature reviews Genetics* 2006, **7**(9):703-713.
17. Gross DS, Garrard WT: **Nuclease hypersensitive sites in chromatin**. *Annual review of biochemistry* 1988, **57**:159-197.
18. Li Q, Harju S, Peterson KR: **Locus control regions: coming of age at a decade plus**. *Trends in genetics* 1999, **15**(10):403-408.
19. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE: **High-resolution mapping and characterization of open chromatin across the genome**. *Cell* 2008, **132**(2):311-322.
20. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS *et al*: **Global mapping of protein-DNA interactions in vivo by digital genomic footprinting**. *Nature methods* 2009, **6**(4):283-289.
21. John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA: **Chromatin accessibility pre-determines glucocorticoid receptor binding patterns**. *Nature genetics* 2011, **43**(3):264-268.
22. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA *et al*: **Histone H3K27ac separates active from poised enhancers and predicts developmental state**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(50):21931-21936.
23. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J: **A unique chromatin signature uncovers early developmental enhancers in humans**. *Nature* 2011, **470**(7333):279-283.
24. Consortium TEP: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414):57-74.
25. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K *et al*: **A bivalent chromatin structure marks key developmental genes in embryonic stem cells**. *Cell* 2006, **125**(2):315-326.
26. Azuara V, Perry P, Sauer S, Spivakov M, Jorgensen HF, John RM, Gouti M, Casanova M, Warnes G, Merkenschlager M *et al*: **Chromatin signatures of pluripotent cell lines**. *Nature cell biology* 2006, **8**(5):532-538.
27. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S *et al*: **Widespread transcription at neuronal activity-regulated enhancers**. *Nature* 2010, **465**(7295):182-187.
28. Lam MT, Cho H, Lesch HP, Gosselin D, Heinz S, Tanaka-Oishi Y, Benner C, Kaikkonen MU, Kim AS, Kosaka M *et al*: **Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription**. *Nature* 2013, **498**(7455):511-515.
29. Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, Merkurjev D, Zhang J, Ohgi K, Song X *et al*: **Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation**. *Nature* 2013, **498**(7455):516-520.

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

30. Natoli G, Andrau JC: **Noncoding transcription at enhancers: general principles and functional models**. *Annual review of genetics* 2012, **46**:1-19.

31. Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, Almada AE, Lin C, Sharp PA, Giallourakis CC *et al*: **Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(8):2876-2881.

32. Lai F, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA, Shiekhattar R: **Activating RNAs associate with Mediator to enhance chromatin architecture and transcription**. *Nature* 2013, **494**(7438):497-501.

33. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q *et al*: **Long noncoding RNAs with enhancer-like function in human cells**. *Cell* 2010, **143**(1):46-58.

34. Ling J, Ainol L, Zhang L, Yu X, Pi W, Tuan D: **HS2 enhancer function is blocked by a transcriptional terminator inserted between the enhancer and the promoter**. *The Journal of biological chemistry* 2004, **279**(49):51704-51713.

35. Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, Chun HB, Tough DF, Prinjha RK, Benner C *et al*: **Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription**. *Molecular cell* 2013, **51**(3):310-325.

36. Mousavi K, Zare H, Dell'orso S, Grontved L, Gutierrez-Cruz G, Derfoul A, Hager GL, Sartorelli V: **eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci**. *Molecular cell* 2013, **51**(5):606-617.

37. Qu H, Fang X: **A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project**. *Genomics, proteomics & bioinformatics* 2013, **11**(3):135-141.

38. Nitzsche A, Paszkowski-Rogacz M, Matarese F, Janssen-Megens EM, Hubner NC, Schulz H, de Vries I, Ding L, Huebner N, Mann M *et al*: **RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity**. *PloS one* 2011, **6**(5):e19470.

39. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS *et al*: **Mediator and cohesin connect gene expression and chromatin architecture**. *Nature* 2010, **467**(7314):430-435.

40. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K *et al*: **Control of developmental regulators by Polycomb in human embryonic stem cells**. *Cell* 2006, **125**(2):301-313.

41. Shi Y: **Histone lysine demethylases: emerging roles in development, physiology and disease**. *Nature reviews Genetics* 2007, **8**(11):829-833.

42. Nottke A, Colaiacovo MP, Shi Y: **Developmental roles of the histone lysine demethylases**. *Development (Cambridge, England)* 2009, **136**(6):879-889.

43. Benevolenskaya EV: **Histone H3K4 demethylases are essential in development and differentiation**. *Biochemistry and cell biology* 2007, **85**(4):435-443.

44. Kozomara A, Griffiths-Jones S: **miRBase: annotating high confidence microRNAs using deep sequencing data**. *Nucleic acids research* 2014, **42**(Database issue):D68-73.

45. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F *et al*: **Landscape of transcription in human cells**. *Nature* 2012, **489**(7414):101-108.

46. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S,

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Dubchak I, Holt A, Lewis KD *et al*: **In vivo enhancer analysis of human conserved non-coding sequences**. *Nature* 2006, **444**(7118):499-502.

47. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA, Young RA: **Super-enhancers in the control of cell identity and disease**. *Cell* 2013, **155**(4):934-947.
48. Loven J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA: **Selective inhibition of tumor oncogenes by disruption of super-enhancers**. *Cell* 2013, **153**(2):320-334.
49. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA: **Master transcription factors and mediator establish super-enhancers at key cell identity genes**. *Cell* 2013, **153**(2):307-319.
50. Xi H, Shulha HP, Lin JM, Vales TR, Fu Y, Bodine DM, McKay RD, Chenoweth JG, Tesar PJ, Furey TS *et al*: **Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome**. *PLoS genetics* 2007, **3**(8):e136.
51. Chen H, Tian Y, Shu W, Bo X, Wang S: **Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome**. *PloS one* 2012, **7**(7):e41374.
52. Ng HH, Surani MA: **The transcriptional and signalling networks of pluripotency**. *Nature cell biology* 2011, **13**(5):490-496.
53. Orkin SH, Hochedlinger K: **Chromatin connections to pluripotency and cellular reprogramming**. *Cell* 2011, **145**(6):835-850.
54. Young RA: **Control of the embryonic stem cell state**. *Cell* 2011, **144**(6):940-954.
55. Zhou Q, Brown J, Kanarek A, Rajagopal J, Melton DA: **In vivo reprogramming of adult pancreatic exocrine cells to beta-cells**. *Nature* 2008, **455**(7213):627-632.
56. Lee TI, Young RA: **Transcriptional regulation and its misregulation in disease**. *Cell* 2013, **152**(6):1237-1251.
57. Graf T, Enver T: **Forcing cells to change lineages**. *Nature* 2009, **462**(7273):587-594.
58. Cherry AB, Daley GQ: **Reprogramming cellular identity for regenerative medicine**. *Cell* 2012, **148**(6):1110-1122.
59. Siersbaek R, Rabiee A, Nielsen R, Sidoli S, Traynor S, Loft A, La Cour Poulsen L, Rogowska-Wrzesinska A, Jensen ON, Mandrup S: **Transcription factor cooperativity in early adipogenic hotspots and super-enhancers**. *Cell reports* 2014, **7**(5):1443-1455.
60. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K *et al*: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes**. *Nucleic acids research* 2006, **34**(Database issue):D108-110.
61. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A: **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles**. *Nucleic acids research* 2010, **38**(Database issue):D105-110.
62. Robasky K, Bulyk ML: **UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions**. *Nucleic acids research* 2011, **39**(Database issue):D124-128.
63. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S *et al*: **GENCODE: the reference human genome annotation for The ENCODE Project**. *Genome research* 2012, **22**(9):1760-1774.
64. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists**

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

using DAVID bioinformatics resources. *Nature protocols* 2009, **4**(1):44-57.

65. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S *et al*: **BEDOPS: high-performance genomic feature operations**. *Bioinformatics (Oxford, England)* 2012, **28**(14):1919-1920.

66. May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C *et al*: **Large-scale discovery of enhancers from human heart tissue**. *Nature genetics* 2012, **44**(1):89-93.

Acknowledgements

We wish to thank the ENCODE Project Consortium for making their data publicly available. This work was supported by grants from the Major Research plan of the National Natural Science Foundation of China (No. U1435222), the Program of International S&T Cooperation (No. 2014DFB30020) and the National High Technology Research and Development Program of China (No. 2015AA020108).

Author Contributions

W.S. conceived the project. W.S., X.B. and S.W. designed all experiments. H.L., H.C., and F.L. performed the experiments. All authors analysed the data and contributed to manuscript preparation. W.S. wrote the manuscript.

Additional information

Supplementary information accompanies this paper at

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure legend

Figure 1. Identification and characterisation of HOT regions in ESCs

(A) Distribution of TFBS complexity signal across the 80,326 ESC TFBS-clustered regions. TFBS-clustered regions are plotted in increasing order based on their TFBS complexity signal. HOT regions are defined as the population of TFBS-clustered regions above the inflection point of the curve. (B) ChIP-seq binding profiles for the indicated DNaseI, histone modifications, transcription factors, chromatin regulators, RNA polymerase II, and MRE at the POLE4 and LINC loci in ESCs. Gene models are depicted below the binding profiles. Region bars and scale bars are depicted above the binding profiles. rpkm, reads per kilobase per million. (C–D) Metagene representations of the mean ChIP-seq signal for the indicated DNaseI, RNA polymerase II, histone modifications, transcription factors, transcriptional cofactors, and chromatin regulators across LOT (blue) and HOT (red) regions. Metagenes are centred on the TFBS-clustered region (5863 bp and 11,890 bp for LOT and HOT regions, respectively) with 5 kb surrounding each TFBS-clustered region. (E) Gene expression level of HOT-specific genes (red) and LOT-specific genes (blue). (F) Motif enrichment in HOT and LOT regions, compared with different backgrounds. Heat map showing the most differentially distributed motifs (multiple testing corrected P-value < 0.01) between HOT regions compared with the genome average values (first column), HOT and LOT regions (second column), LOT regions compared with the genome average values (third column), LOT and HOT regions (fourth column).

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

See also Figure S1 and Tables S1–S3.

Figure 2. General features of HOT regions in many cell types

(A) Saturation analysis of HOT regions. We modelled saturation for element count and length using a Weibull distribution ($r^2 \geq 0.995$) and predicted saturation at approximately 107,184 (sd = 8,608) and 774,925,252 (sd = 33,534,434) for count and length, respectively. The cell line estimation of 95% saturation is 222 and 154 for count and length, respectively. (B) Distribution of 59,986 HOT regions and 301,322 LOT regions with respect to GENCODE gene annotations. Promoter regions are defined as the first region located within 1 kb upstream and downstream of a GENCODE TSS.

(C) Distributions of the number of cell types, from 1 to 154 (*y* axis), in which HOT (red) and LOT (blue) regions in each of six classes (*x* axis) are observed. The width of each shape at a given *y* value shows the relative frequency of regions present in that number of cell types. (D) GO terms for HOT-region-associated genes in 19 human cell and tissue types with corresponding *p*-values. (E) Candidate key developmental transcription factors identified in six cell types. All of these transcription factors were previously demonstrated to play key roles in the development of the respective cell type or facilitate differentiating to the respective cell type.

See also Tables S4–S6.

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 3. Association of HOT regions with validated elements

(A–B) Examples of HOT regions (red line) overlapping microRNA (A) and repetitive elements (B). Peaks are observed in cell types consistent with known functions of the microRNAs and repetitive elements. (C) Examples of known cell-selective experimentally validated distal, non-promoter *cis*-regulatory elements. Shown above each set of DNaseI data are schematics displaying HOT regions relative to the genes they control.

See also Tables S7–S8.

Figure 4. Developmental persistence of HOT regions at embryonic enhancers

(A) Mouse day 11.5 embryonic tissue activity (blue lacZ staining) of seven representative transgenic human enhancer elements from the VISTA database. Shown below each image are the enhancer ID and numbers of individual embryos with enhancer activity (staining) in the indicated anatomical structure. (B) DNaseI hypersensitivity at seven enhancer elements corresponding to (A) across 57 definitive cell types. Note the relationship between the anatomical staining patterns in (A) and the cellular restriction (or lack thereof) of DNaseI hypersensitivity. (C–D) Persistence of HOT regions at embryonic enhancers. (C) Percentage of validated embryonic enhancers from the VISTA database with blood vessel staining (“Blood vessels”) and without blood vessel staining (“NOT Blood vessels”) that overlap a HOT region in any human endothelial cell type. (D) Percentage of validated embryonic enhancers

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

from the VISTA database with heart staining ("Heart") and without heart staining ("NOT Heart") that overlap a HOT region in any human paraxial mesoderm cell type.

Figure 5. Association of distal HOT regions with enhancers

(A) Enrichment of super-enhancers in HOT regions relative to LOT regions in 14 different cell types. (B) Distal HOT regions colocalise with super-enhancers in a cell type-specific manner. Cell type-specific super-enhancers (*y*-axis) are mapped relative to cell-specific distal HOT regions (*x*-axis) in 14 different cell types. (C) ChIP-seq binding profiles of super-enhancer, enhancer, and distal HOT regions for the indicated DNaseI and enhancer-relevant markers including transcription factors, transcriptional cofactors, chromatin regulators, and RNA polymerase II in ESCs. (D) GO analysis of super-enhancer-associated genes and HOT region-associated genes in H1 hESC, CD20, and pancreas cells. The top 10 scoring categories were selected for display. A threshold *p*-value score of 10^{-4} was incorporated as a minimum requirement filter for scoring as a top category.

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplementary figures

Figure S1. Characterisation of HOT regions, related to figure 1

(A) ChIP-seq binding profiles for the indicated DNaseI, histone modifications, transcription factors, chromatin regulators, polymerase II (RNAPII), and MRE at the LINC loci in ESCs. Region bars and scale bars are depicted above the binding profiles. rpkm, reads per kilobase per million. (B) Metagene representations of the mean ChIP-seq signal for the indicated broad histone, chromatin regulator, transcription factor and other markers across LOT (black) and HOT (red) regions. Metagenes are centred on the TFBS-clustered region with 5 kb surrounding each TFBS-clustered region.

Supplementary tables

Table S1. H1hESC TFBS cluster information, related to figure 1

Table showing the TF complexity cutoffs for HOT regions in H1hESCs: the total number of H1hESC TFBS clusters is 80,326, the number of HOT regions is 8,533, and the median length is 11,890 bp and 5,863 bp for HOT regions and LOT regions, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table S2. ChIP-seq density of HOT and LOT regions, related to figure 1

Fold difference values of ChIP-seq signals between LOT and HOT regions for the indicated broad histone, chromatin regulator, transcription factor and other markers. Total signal indicates the mean ChIP-seq signal (total reads) at LOT and HOT regions normalised to the mean value at LOT regions. Density indicates the mean ChIP-seq density at constituent DHSs (rpkm) of LOT regions and HOT regions normalised to the mean value at LOT regions. Read % indicates the percentage of all reads mapped to TFBS-clustered regions that fall in the constituents of LOT or HOT regions.

Table S3. Enrichment of known transcription factor binding motifs, related to figure 1

Known transcription factor binding motifs in HOT and LOT regions using HOMER. Both the backgrounds of the genome and LOT/HOT regions were used in motif scanning within HOT/LOT regions.

Table S4. Information on HOT regions in 154 files, related to figure 2

Table showing the TF complexity cutoffs for HOT regions, HOT region number, total number of TFBS clusters and genome coverage in 154 cell lines.

Table S5. Key TF genes shown in figure 2J, related to figure 2

Functions and references of key developmental transcription factor genes highlighted

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

in figure 2J.

Table S6. Key developmental transcription factor genes, related to figure 4

Development-associated transcription factor genes identified in 154 cell lines.

Table S7. Repetitive elements in HOT regions, related to figure 2

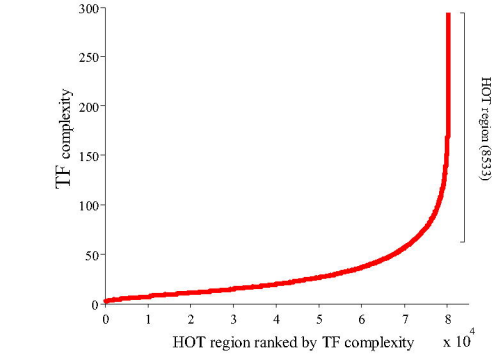
Overlap of repeat-masked elements by repeat family for families with more than 2,000 elements overlapping DHSs. Column 1 shows the repeat family; column 2 shows the repeat class. Column 3 shows the average size of elements in the family; column 4 shows the total number of occurrences of elements of the family in the genome. Column 5 indicates the number of repeat families that overlap a HOT region by at least 95%.

Table S8. The 1046 validated elements in HOT regions, related to figure 2

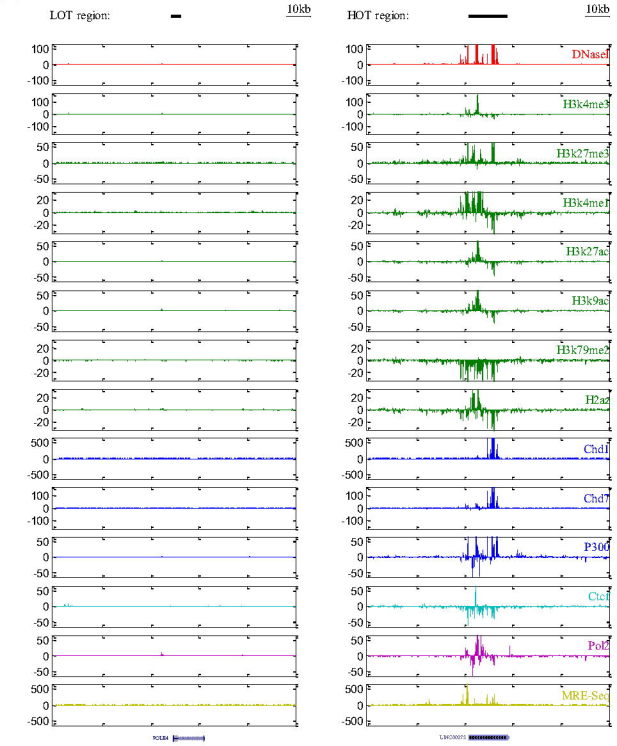
Enrichment of validated elements in HOT and LOT regions. The number of non-VISTA enhancer-associated elements is 373, while the total number of validated elements is 1,046.

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

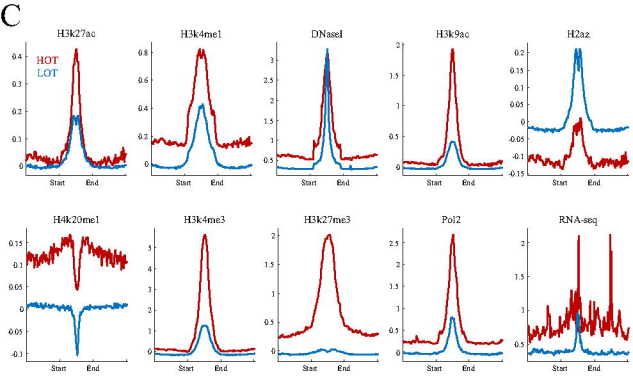
A



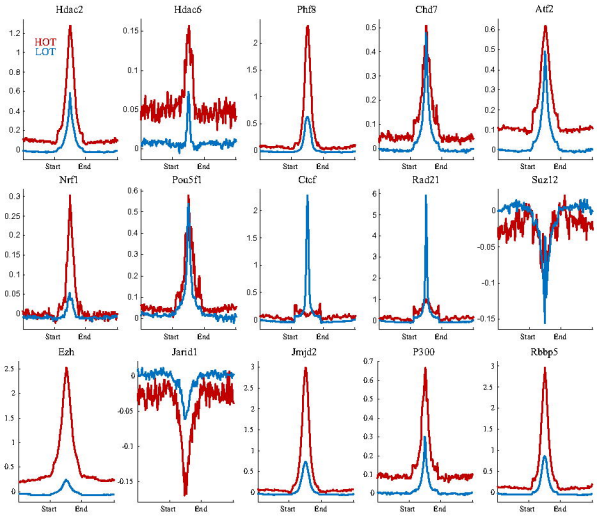
B



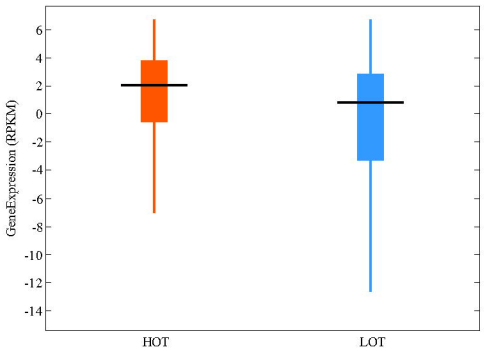
C



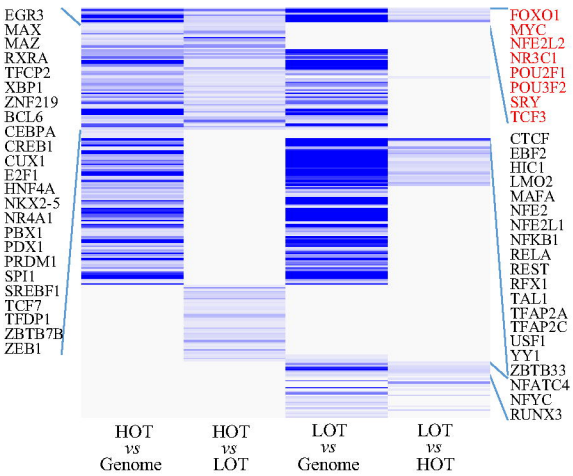
D



E

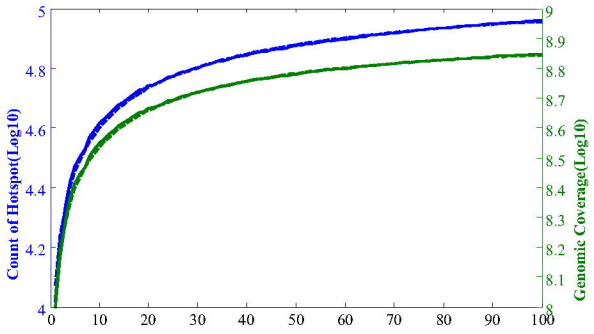


F

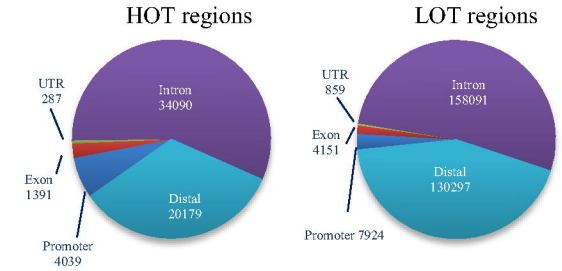


bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

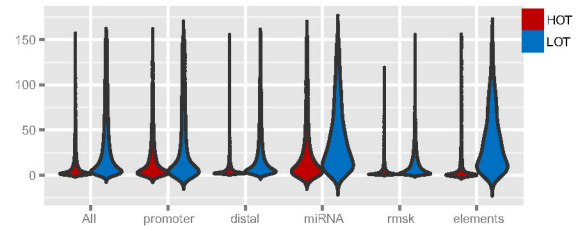
A

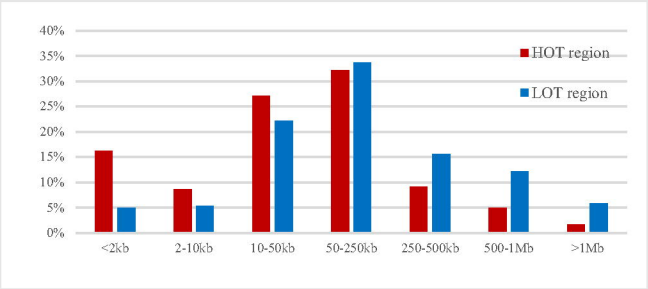


B

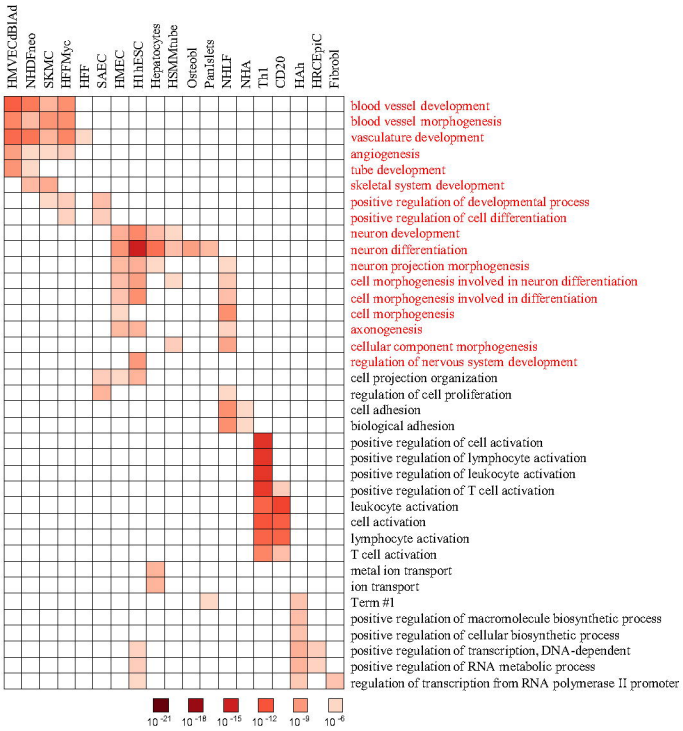


C





D

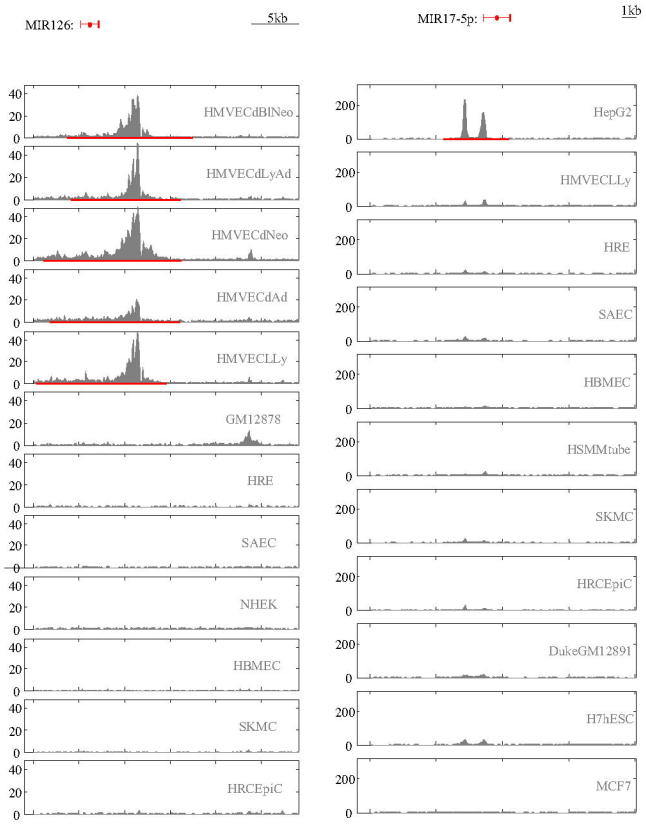


E

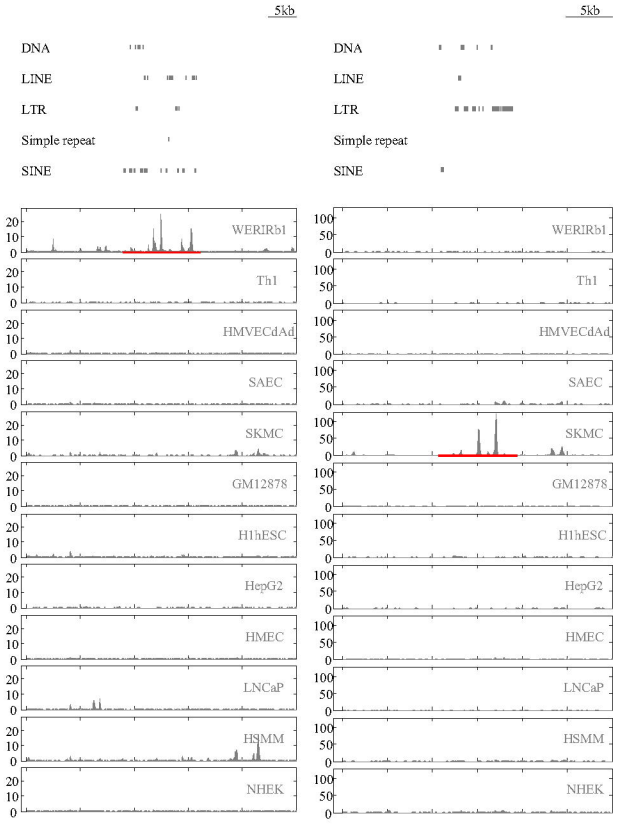
Skin	Treg
P63	*STAT5*
BCL11B	*C-Rel*
ATF2	*FOXP3*
EGR1	
NFE2L2	
Skeletal muscle	**Lung**
SRF	*NFIB*
MYOD1	*NR3C1*
SIX1	*HOXB5*
STAT3	*TBX2/3*
SREBF1	*TBX4/5*
MEF2C	*VEGFA*
Blood vessels	**Breast cancer**
HIF1A	*TBX1*
NFKB1	*ESR1*
NFKB2	*ZNF217*
VEGFA	*GATA3*
TEAD4	*S6K1*

bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

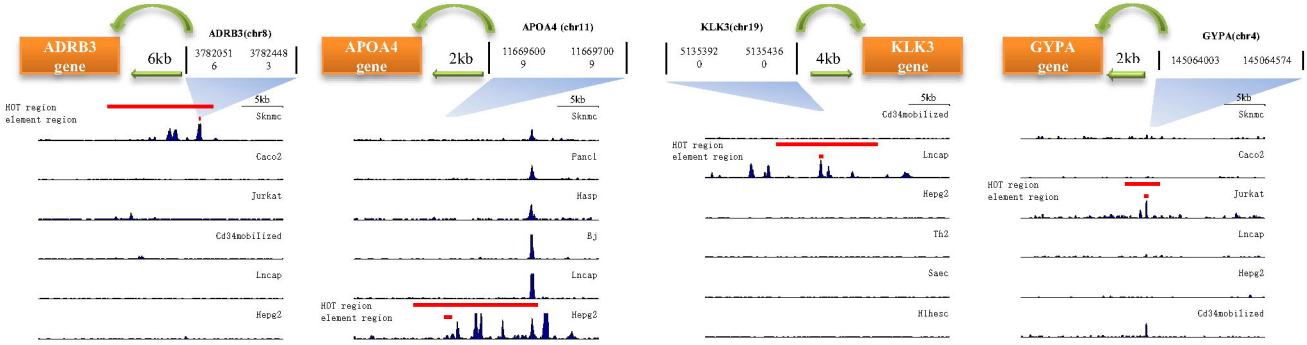
A



B



C

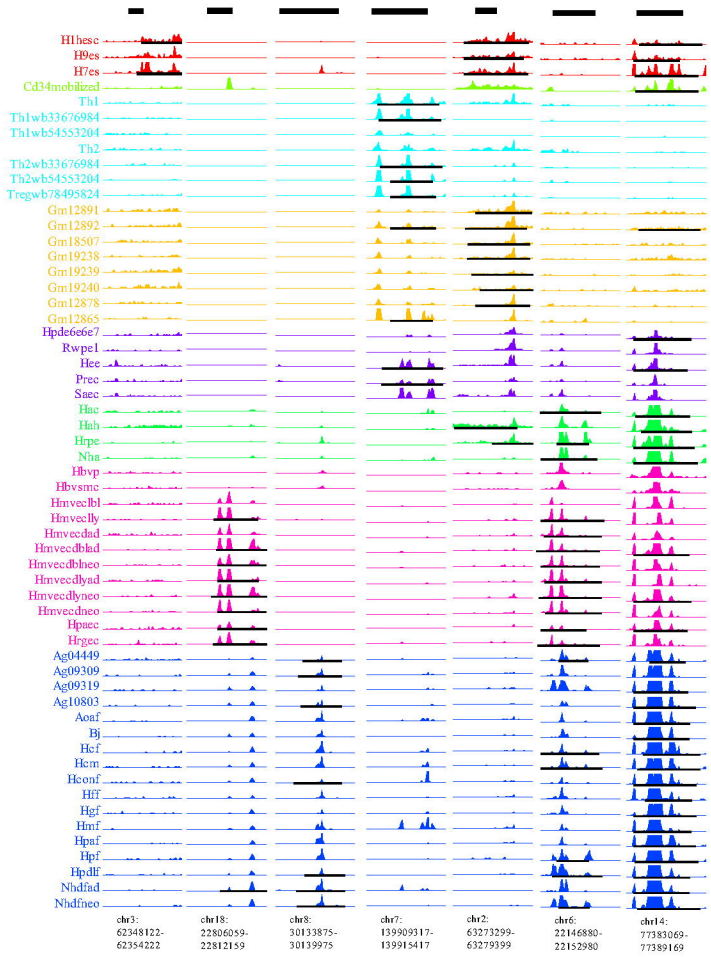


bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

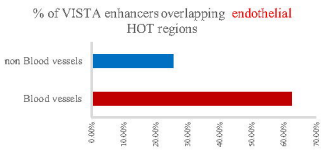
A

Embryonic tissue	*hs434* forebrain[4/4]	*hs1653* blood vessels[5/8]	*hs1962* heart[15/17] melanocytes[7/17]	dorsal root ganglion[3/7] trigeminal V[3/7]	*hs1066* hindbrain[5/5] midbrain[5/5] forebrain[5/5]	*hs1335* neural tube[4/4] hindbrain[3/4] midbrain[4/4] forebrain[3/4]	*hs1466* neural tube[8/8] hindbrain[8/8] midbrain[8/8] dorsal root ganglion[7/8] forebrain[6/8] limb[8/8] branchial arch[8/8] heart[6/8]

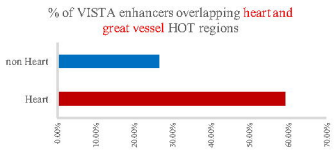
B



C



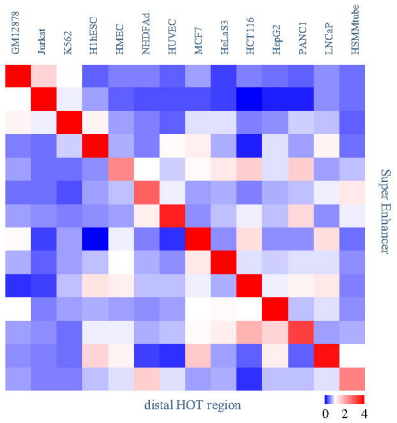
D



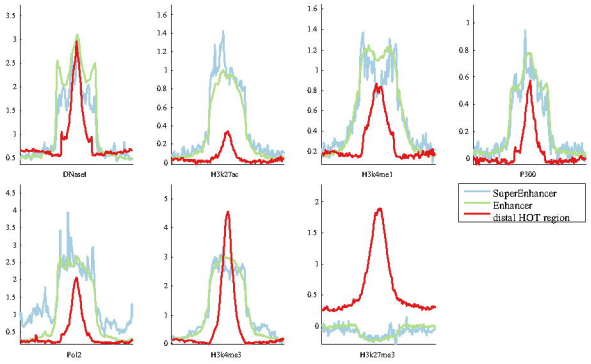
bioRxiv preprint doi: https://doi.org/10.1101/036152; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

A

B



C



D

SuperEnhancer | **HOT region**

H1

GO term	P value	GO term	P value
axon guidance	8.98E-06	neuron differentiation	1.25E-15
axonogenesis	3.17E-05	neuron development	2.42E-10
neuron projection development	5.40E-05	cell morphogenesis involved in differentiation	4.38E-10
cell morphogenesis involved in neuron differentiation	7.57E-05	regulation of nervous system development	1.31E-09
neuron projection morphogenesis	9.27E-05	cell morphogenesis involved in neuron differentiation	1.76E-09
negative regulation of transcription	2.79E-04	regulation of neurogenesis	3.47E-09
regulation of neuron differentiation	3.22E-04	embryonic organ morphogenesis	5.01E-09
negative regulation of RNA metabolic process	3.42E-04	sensory organ development	1.00E-08
cell morphogenesis involved in differentiation	3.84E-04	neuron projection morphogenesis	1.03E-08
cell projection morphogenesis	3.97E-04	axonogenesis	1.30E-08

CD20

GO term	P value	GO term	P value
immune response	3.46E-09	leukocyte activation	1.71E-13
antigen processing	5.05E-07	cell activation	3.31E-12
cell activation	2.12E-06	lymphocyte activation	3.50E-12
regulation of lymphocyte activation	6.12E-06	T cell activation	4.33E-08
positive regulation of cell activation	7.23E-06	positive regulation of T cell activation	2.14E-07
regulation of cell activation	9.59E-06	positive regulation of lymphocyte differentiation	2.26E-07
positive regulation of lymphocyte activation	1.10E-05	regulation of lymphocyte differentiation	6.09E-07
regulation of programmed cell death	1.37E-05	lymphocyte differentiation	7.21E-07
regulation of cell death	1.49E-05	immune response	7.24E-07
regulation of leukocyte activation	2.28E-05	positive regulation of T cell differentiation	8.59E-07

Pancreas

GO term	P value	GO term	P value
regulation of transcription	3.32E-06	neuron differentiation	2.91E-08
endocrine pancreas development	1.12E-05	positive regulation of transcription	9.40E-07
pancreas development	1.12E-05	neuron development	3.93E-06
regulation of cell proliferation	3.76E-05	ion transport	1.04E-05
positive regulation of cell proliferation	5.39E-05	cation transport	1.33E-05
positive regulation of nitrogen compound metabolic process	3.93E-04	regulation of transcription	2.20E-05
positive regulation of macromolecule biosynthetic process	4.62E-04	positive regulation of transcription	2.38E-05
positive regulation of cellular biosynthetic process	7.46E-04	positive regulation of RNA metabolic process	3.03E-05
positive regulation of biosynthetic process	8.64E-04	regulation of cell development	5.74E-05
positive regulation of nucleobase	8.94E-04	cell motion	7.63E-05