Edinburgh Research Explorer

Mammalian transcriptional hotspots are enriched for tissue specific enhancers near cell type specific highly expressed genes and are predicted to act as transcriptional activator hubs

Citation for published version:
Joshi, A 2014, 'Mammalian transcriptional hotspots are enriched for tissue specific enhancers near cell type specific highly expressed genes and are predicted to act as transcriptional activator hubs' BMC Bioinformatics, vol 15, no. 1, 412. DOI: 10.1186/s12859-014-0412-0

Digital Object Identifier (DOI):
10.1186/s12859-014-0412-0

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
BMC Bioinformatics

Publisher Rights Statement:
© 2014 Joshi; licensee BioMed Central. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Mammalian transcriptional hotspots are enriched for tissue specific enhancers near cell type specific highly expressed genes and are predicted to act as transcriptional activator hubs

Anagha Joshi

Abstract

Background: Transcriptional hotspots are defined as genomic regions bound by multiple factors. They have been identified recently as cell type specific enhancers regulating developmentally essential genes in many species such as worm, fly and humans. The in-depth analysis of hotspots across multiple cell types in same species still remains to be explored and can bring new biological insights.

Results: We therefore collected 108 transcription-related factor (TF) ChIP sequencing data sets in ten murine cell types and classified the peaks in each cell type in three groups according to binding occupancy as singletons (low-occupancy), combinatorials (mid-occupancy) and hotspots (high-occupancy). The peaks in the three groups clustered largely according to the occupancy, suggesting priming of genomic loci for mid occupancy irrespective of cell type. We then characterized hotspots for diverse structural functional properties. The genes neighbouring hotspots had a small overlap with hotspot genes in other cell types and were highly enriched for cell type specific function. Hotspots were enriched for sequence motifs of key TFs in that cell type and more than 90% of hotspots were occupied by pioneering factors. Though we did not find any sequence signature in the three groups, the H3K4me1 binding profile had bimodal peaks at hotspots, distinguishing hotspots from mono-modal H3K4me1 singletons. In ES cells, differentially expressed genes after perturbation of activators were enriched for hotspot genes suggesting hotspots primarily act as transcriptional activator hubs. Finally, we proposed that ES hotspots might be under control of SetDB1 and not DNMT for silencing.

Conclusion: Transcriptional hotspots are enriched for tissue specific enhancers near cell type specific highly expressed genes. In ES cells, they are predicted to act as transcriptional activator hubs and might be under SetDB1 control for silencing.

Keywords: ChIP sequencing, Transcription hotspots, Transcription regulation, Data analysis
Results and discussion

Combinatorial binding events overlap across multiple cell types

A Chip sequencing experiment typically identifies thousands to hundreds of thousands of genome-wide binding sites of a TF in a cell type. In order to investigate if distinct characteristics of peaks grouped based on the number of binding factors at a binding location, we collected genome-wide binding events of transcription-related factors (TFs) across 10 different normal and cancer murine cell types. In each cell type, ChIP sequencing data sets (see Methods for details) for 6 to 21 TFs (108 samples in total) were collected. Peaks were called in each sample (see Methods for details) with the number of peaks ranging from 1,000 to 100,000 in 108 samples (Figure 1B). The number of TFs at binding sites followed an exponential curve such that about half of the regions were bound by only one TF in any given cell type and less than 0.5% occupied by all TFs studied. We divided all peaks in a cell type into three groups: singletons or low-occupancy, peaks bound by only one transcription-related factor (Figure 1A), combinatorials or mid-occupancy, peaks bound by a combination of transcription-related factors (Figure 1A) and hotspots or high-occupancy, peaks bound by more than five TFs studied in a given cell type (Figure 1A, Additional file 1: Table S31). On average about 50% of all the peaks in a cell type were singletons and combinatorials while only 0.1-2% was classified as hotspots (Figure 1C). We assigned peaks in these three groups to genomic locations: promoter, 5' UTR, 3' UTR, exon, intron or inter-genic using HOMER [7]. The location of peaks followed a uni-modal distribution with a peak at the TSS (Additional file 2: Figure S1). Singleton peaks were specifically under-represented in promoter regions (P value: 5.2e-3) and 5' UTR regions (P value: 8.9e-3) (Additional file 2: Figure S2).

In order to investigate the overlap of the three groups across cell types, we constructed a hierarchical tree of 30 peak sets (three groups each in 10 cell types). To construct hierarchical tree, we first integrated of all 108 peak sets across 10 cell types resulted in 408,003 unique peaks. We then constructed a binary matrix of 408,003 rows and 30 columns recording the presence or absence of peaks in each group for each cell type and built hierarchical tree using Pearson's correlation coefficient as a distance measure. Many samples clustered according to the group across multiple cell types (Figure 1D). The peaks in three groups in a given cell type are mutually exclusive by definition but the fact that combinatorials across many cell types cluster according to the groups is a notable observation. This suggests that each genomic location is marked for occupancy i.e. mid-occupied loci in one cell type were more likely to be bound by multiple factors rather than one factor in another cell type. Hotspots of related cell types cluster into two tight clusters. The two embryonic stem (ES) cell samples were the most similar to each other as expected. The four related blood cell types – HPC, Erythroid, MK progenitors and B cells formed second separate dense cluster of hotspot peaks.

important genes [3]. The presence of hotspot regions was further confirmed in other species such as worm [4] and humans [5]. Hotspots in C. elegans and D. melanogaster were enriched for sequence binding motifs of many TFs including signature motifs such as GAGA and Zelda [3]. However, analysis of human ENCODE data did not support these observations. In humans, hotspots were specifically deprived of cis-regulatory motifs and no signature motif similar to the ‘GAGA’ motif was found [5]. At the other end of the spectrum, most TF binding events were ‘singletons’ (bound by only one TF) accounting for nearly one third of the binding events in a cell type. These regions when studied in D. melanogaster transgenic assays did not drive patterned reporter gene expression leading to the conclusion that they do not act as strong developmental enhancers [6]. This leads to a series of questions about properties of both hotspots and singletons. For example, are they truly distinct genomic regions, do they have a characteristic sequence or chromatin signature?

In order to answer the above questions, we collected genome-wide binding patterns of multiple transcription-related factors in ten murine cell types. For each cell type, the peaks were classified into three groups: singleton genomic regions occupied by only one TF (low-occupancy), hotspot genomic regions occupied by most TFs under study (high-occupancy) and combinatorial genomic regions occupied by a combination of TFs (mid-occupancy). The genomic regions largely clustered according to the group suggesting distinct genomic regions marked for occupancy independent of cell type. The singletons and combinatorials were bound neighbouring similar genes in all cell types while hotspot peaks occurred near a distinct set of genes in each cell type and showed functional enrichment for cell type specific genes. Though hotspots were enriched for many TF sequence motifs, no signature motif such as GAGA motif was found in murine cell types. We identified H3K4me1 chromatin modification distinguishing hotspots from singletons where hotspots showed a bimodal H3K4me1 peak whereas singletons were mono-modal. Finally, we collected the differentially regulated genes after perturbations of multiple transcription-related factors in ES cells, to show that hotspots were preferentially bound by activators and not repressors. As genes differentially expressed after Setdb1 knockout but not Dnmt knockout were enriched for ES hotspot genes we suggest that Setdb1 might be involved in silencing hotspots.

Results and discussion

Combinatorial binding events overlap across multiple cell types

A Chip sequencing experiment typically identifies thousands to hundreds of thousands of genome-wide binding
**Figure 1** (See legend on next page.)
In *D. melanogaster*, peaks occupied at high levels by transcription factors in vivo drive patterned gene expression, whereas those occupied only at lower levels mostly do not [6]. We calculated average peak height for the three groups where peak height was defined as the total number of tags in a 400 bp window around the peak summit (see Methods). Singletons consistently showed significantly lower peak height compared to combinatorials and hotspots (Additional file 1: Table S32). For example, the average peak height of *Fli1* bound singleton regions in T cells is significantly lower than combinatorials (P value < 1e-323) as were the average peak heights of *Cebpβ* bound regions in macrophages where singletons were significantly lower (P value < 1e-323) than combinatorials (Figure 1E).

Finally, we collected ChIP sequencing data for Oct4 in ES cells generated by three independent labs [8-10] and found about 1000 Oct4 peaks overlapping in all three experiments. The 1000 peaks were enriched for combinatorials and hotspots but were depleted for singleton regions (Additional file 2: Figure S3). Combinatorials and hotspot binding events were found consistently across different experiments.

Taken together, singletons were depleted in promoter and 5’ UTR regions, had lower peak height and were not consistently found across multiple experiments.

**Hotspots preferentially bind in cell type-specific gene neighbourhoods**

The hierarchical clustering of 30 peak sets clustered them largely according to occupancy. To investigate if the group-wise clustering of peaks is also reflected at a gene level, we mapped all peaks to the nearest gene. Peaks occupied by combinatorial factors such as Oct, Sox, and Nanog form super-enhancers and noted that about 30% of super-enhancers overlapped with hotspots in ES cells (Figure 2B, P value:2.7e-8). Karwacki-Neisius et al. [12] showed that master transcription factors such as Oct, Sox, and Nanog form ‘super-enhancers’ at key cell identity genes that span large domains and drive cell-type-specific gene expression program. We calculated the overlap of ES cell peaks in all three groups with super-enhancers and noted that about 30% of super-enhancers overlapped with hotspots in ES cells (Figure 2B, P value:2.7e-8). Karwacki-Neisius et al. [12] showed that Oct4+/− (reduced Oct4 expression) ES cells have increased Oct4 binding at pluripotency enhancers. Only the enhancer regions strongly bound by Oct4 in Oct4+/−, compared to Oct4+/+ had a preferential bias towards hotspot regions (Figure 2C, P value: 4.61e-29). As Oct4 binding in Oct4+/− was shown to be specifically enriched near pluripotent genes, this confirmed that hotspots occupied cell type specific gene loci. Stadler et al. [13] generated a methylome map in murine ES cells and noted that low methylated regions mark multi-TF bound enhancer regions in the vicinity of cell type specific genes. The overlap of high- and low-methylated regions with three groups in ES cells confirmed that hotspots were specifically depleted for high-methylated and enriched for low-methylated regions in ES cells (Additional file 2: Figure S6). Taken together, by comparing the three groups against three complementary experimental datasets we concluded that hotspots lie preferentially in the vicinity of cell type specific genes.

**Hotspots are enriched for sequence motifs of multiple transcription factors and are preferentially bound by pioneering factors**

In order to search for properties distinguishing the three groups, we calculated the average mammalian conservation
### Functional Categories enriched in hotspot genes (P values)

| Cell type       | Functional Categories                                                                 |
|-----------------|----------------------------------------------------------------------------------------|
| B cells         | B cell receptor signalling pathway (1.25e-8), B cell activation (1.09e-7), B cell    |
|                 | differentiation (3.17e-4)                                                              |
| Dendritic cells | Antigen processing and presentation of peptide antigen (2.61e-9), Regulation of     |
|                 | inflammatory response to antigenic stimulus (1.03e-7)                                  |
| Erythroid       | Regulation of erythrocyte differentiation (4.49e-9), erythrocyte homeostasis (3.99e-8), |
|                 | erythrocyte development (2.98e-4)                                                     |
| ES (Ng data set)| Stem cell differentiation (3.26e-11), stem cell development (1.52e-9), stem cell    |
|                 | maintenance (1.14e-8), cell differentiation (3.37e-29)                                 |
| ES (Young data  | set) Stem cell differentiation (1.49e-11), stem cell development (9.63e-11), stem cell |
|                 | maintenance (9.14e-10), chordate embryonic development (8.96e-25)                     |
| HPC             | Hemopoiesis (3.93e-21), hemopoietic or lymphoid organ development (1.48e-20), blood     |
|                 | vessel development (1.83e-10)                                                          |
| Macrophage      | Regulation of immune system process (4.99e-16), response to bacterium (9.46e-9)        |
| MEL             | Hemopoiesis (1.54e-16), myeloid cell differentiation (1.17e-13), regulation of cell    |
|                 | death (1.4e-12), erythrocyte differentiation (9.79e-11)                                |
| MK progenitors  | Thrombin receptor signalling pathway (8.77e-5), regulation of cell communication       |
|                 | (3.90e-5), regulation of inflammatory response (3.43e-5), negative regulation of mast   |
|                 | cell activation (7.98e-5)                                                             |
| T cells         | T cell activation (1.32e-29), T cell differentiation (4.80e-20), positive regulation  |
|                 | of T cell activation (9.27e-13), regulation of T cell receptor signalling pathway       |

**Figure 2** (See legend on next page.)
score for the genome sequences of peaks in all three groups. There was no statistically significant difference in the fraction of conserved peaks between the three groups (Additional file 1: Table S33). In C. elegans, hotspot-associated genes were more likely to be essential and were enriched for a variety of functions such as growth, reproduction, and larval embryonic development [4]. To check if this observation holds true in the mammalian system, we downloaded essential genes in mouse [14] and investigated whether any of the three groups were preferentially bound near essential genes. Though there was a relatively smaller fraction of essential genes in singletons compared to hotspots, the difference was not statistically significant (Additional file 1: Table S34). Thus, genomic conservation and gene essentiality did not distinguish the three groups.

In D. melanogaster, hotspots were enriched for sequence motifs of many but not all TFs under study and those without sequence motif enrichment were shown to bind via protein–protein interaction [2]. They were enriched for sequence motifs of two global activators Zelda and GAGA and this sequence signature alone is sufficient to distinguish the hotspot regions [3]. They were also enriched for BEAF-32 and Trl/GAF insulator motifs [15]. In C. elegans, cis-regulatory sequences of TFs are not always present in hotspots, the binding was associated with open chromatin [4]. In line with this, hotspots were enriched for ‘GAGA’ sequence motifs [16]. On the other hand, the depletion of cis-regulatory motifs was observed in human hotspots where the sequence motif for TFs in ‘cold’ regions were more than two-fold more enriched than in hotspot regions and no ‘GAGA’ or similar motif was found [5], but they mentioned the possible presence of other sequence-specific motifs.

Indeed Foley et al. found that on average, any given HOT promoter contained about four different types of motifs [17]. In summary there is no evidence so far regarding a specific cis-regulatory signature in hotspot peaks in mammals, nor is there a consensus whether hotspots are enriched or depleted for TF sequence motifs.

To check if any sequence motif signature discriminates the three groups, we calculated the number of occurrences of mammalian cis-regulatory sequence motifs and compared them between the three groups. In most cell types, the motifs associated with the transcription factors known to be key players for that cell type were significantly enriched in hotspot regions compared to singletons. For example, Gata (7.0e-2) and Runx (1.94e-8) motifs were enriched in B cell hotspots, E-box (6.26e-8) and Gfi1 (3.00e-3) motifs were enriched in Erythroid hotspots while Meis (1.58e-17) and Scl-Gata (2.44e-13) motifs were enriched in HPC hotspots (Figure 3A, Additional file 1: Table S35). As GC content can bias the motif enrichment, we confirmed that the GC content across three groups was not significantly different across ten cell types (Additional file 1: Table S36). Moreover, we validated that the motif enrichment was consistent only when the promoter peaks were considered in three groups (Additional file 1: Figure S5).

Siersbaek et al. identified ~1,000 TF hotspots during adipogenesis and demonstrated that Cebpβ marks a large number of these hotspots and is required for chromatin remodeling during differentiation [18]. This hypothesizes that pioneering TFs in a cell type mark hotspots. We identified TFs preferentially binding in hotspots (more than 90% regions bound) and found Pu.1 marking B cell hotspots and Cebps marking macrophage hotspots (Figure 3B and Additional file 1: Table S37). Med12 was bound at more than 99% of ES hotspots which proposes hotspots to be under control of mediator complex [11]. Thus hotspots regions were indeed bound by pioneering factors across multiple cell types. On the other hand, only around 1% of peaks of Suz12 overlapped with hotspots (under-representation p value < 0.05). As Suz12 predominantly acts as a transcriptional repressor this suggests that hotspots might act as transcriptional hubs mainly for activation.

Altogether, murine hotspots did not contain a signature sequence motif but were enriched for motifs of many transcription factors including pioneering factors which mainly act as transcriptional activators.
Hotspots lie near highly expressed genes and are enriched for enhancers in nucleosome flanked regions

In order to investigated whether each group shows a distinct chromatin signature, we collected data for three activating chromatin modifications in enhancers and promoters (H3K27ac, H3K4me1, H3K4me3), CCCTC-binding factor (CTCF), RNA polymerase II (PolII) binding as well as RNA sequencing data in murine erythroleukemia (MEL) cells from the ENCODE resource. Hotspot peaks showed significantly higher binding in all six datasets compared to the other two groups. In D. melanogaster, hotspot genes were expressed at higher levels during early embryogenesis but the difference is reduced later in development [2]. In humans, using ENCODE data, it was shown that genes associated with ‘hot’ regions were expressed at higher levels [4]. In line with this, H3K4me3, PolII binding as well as gene expression measured in RPKM values using RNA sequencing data (Figure 4A) in MEL cells were significantly higher in hotspots (P values: 1.49e-116, 2.40e-162 and 1.9e-20 respectively).

**Figure 3** Hotspots are enriched for sequence motifs of multiple transcription factors. A. Bar graphs of fraction of peaks with a given sequence motif in the three groups (singletons – red, combinatorials - green, hotspots - blue) with the name of the sequence motif along with the sequence in IUPAAC format for three representative motifs for each of B cells, erythroid and HPCs (see Additional file 1: Figure S1 for the complete list of motifs). B. A table of cell type, transcription-related factors analysed by ChIP sequencing, transcription-related factors binding to more than 90% of hotspot peaks and statistically significant known sequence motifs found in hotspots for B cells, erythroid and HPCs (see Additional file 1: Figure S2 for the complete table with 10 cell types).
Pennachio et al. [20] predicted about 5,500 high-confidence murine tissue-specific enhancers for 61 tissue types by integrating tissue-specific expression data, conservation information and cis-regulatory motifs and further experimentally validated them using transgenic assays in mouse (VISTA enhancers). MEL hotspots were enriched in VISTA enhancers compared to the singletons and combinatorials (Figure 4B, P value: 4.64e-6). Hotspots in other cell types were also enriched for VISTA enhancers (Additional file 1: Table S38). MEL hotspots had significantly higher H3K4me1 as well as H3K27ac (P values 5.81e-120 ad 2.63e-275 respectively). Histone H3K4me1 marks for enhancer regions whereas histone H3K27ac distinguishes active enhancers from inactive/poised enhancer elements containing H3K4me1 alone [21]. Hotspots therefore mark active enhancer regions in a given cell type.

Hoffman et al. [22] showed high nucleosome positioning at the center of bimodal H3K4me1 peaks compared to mono-modal in mouse islets and liver. The average density plots of H3K4me1 showed a very strong bimodal distribution at hotspots while singletons showed a mono-modal distribution (Figure 4C). This suggests that multiple transcription-related factors can indeed displace nucleosomes for strong enhancer activity at hotspot regions. Hotspots were therefore enriched for nucleosome flanked multiple TF-bound loci. As expected, genes associated with flanked TF-bound loci were more abundantly expressed than those associated with nucleosomal loci, consistent with flanked sites being active enhancer elements [23].

In order to validate these observations in another cell type, we collected genome-wide chromatin modification
data sets in ES cells. Similar to MEL cells [24], all activating chromatin modifications as well as expression values in RPKM using RNA sequencing data showed significant enrichment in hotspots (Additional file 2: Figure S6). The repressive chromatin modifications such as H3K27me3 were depleted in hotspots (Additional file 2: Figure S6). Taken together, we integrated genome wide chromatin binding patterns with TF ChIP sequencing data to show that hotspot regions were enriched for enhancer regions which are located near highly expressed cell type specific genes. Furthermore, the H3K4me1 profile showed a strong bimodal peak at hotspots suggesting nucleosome displacement for transcription activation.

ES hotspots are preferentially occupied by activators and hotspot regions and might be under Setdb1 control for silencing

In order to establish the functional relevance of the binding in three groups, we collected a range of differentially expressed gene sets after TF perturbations in ES cells. Firstly we collected differentially expressed genes after Oct4 knockdown from two independent experiments [25,26]. ES hotspot genes were specifically enriched for down-regulated genes (P value 3.23e-17) after Oct4 knockdown but not up-regulated genes (Figure 5A). This suggests that Oct4 binds to hotspot regions to act as a transcriptional activator.

Nishiyama et al. [27] identified potential functional targets by identifying differentially expressed genes after overexpressing 50 TFs in ES cells. ES hotspot genes were over-represented for differentially expressed regulators such as Etv3 (8.28e-3). However, ES hotspot genes were under-represented (Additional file 3: Figure S7) in differentially expressed genes after overexpression of repressors such as Atf3 (6.62e-3), Gaddd4sa (6.62e-3), Mybl2 (1.65e-3), Rhox6 (1.65e-3) and Tcf4 (4.97e-3). This corroborates the previous observation that hotspot regions were predominantly occupied by activators of transcription (Additional file 1: Table S39).

Not only transcription factors but also other chromatin modifiers such as Baf250 [28], a subunit of the BAF chromatin remodelling complex, as well as Hdac1 [29], a histone deacetylase involved in the removal of acetyl groups from core histones, were preferentially bound at ES hotspots (Additional file 1: Table S40). Enrichment in hotspot loci suggests Hdac1 might act as a transcriptional activator. Though HDACs are generally considered repressors of transcription, it has been noted recently that Hdac1 can indeed act as an activator [29].

As hotspots mark highly active enhancers driving cell type specific gene expression, these regions need to be silenced in other cell types. Karimi et al. [30] showed DNA methylation and Setdb1 methylation work on different sets of genes, where only 7% of the genes up-regulated in a Setdb1 knockout also were up-regulated in Dnmt knock out in ES cells. No significant difference was observed between singletons, combinatorial or hotspot genes for differentially expressed genes after Dnmt knock out (Figure 5B). Challen et al. [31] found ~1,500 loci under Dnmt3a control in HPCs. There was no bias in Dnmt3 occupied loci in HPC, neither towards CpG islands in the three groups. However, genes differentially expressed after Setdb1 knock out were enriched for hotspots (Figure 5B) suggesting that hotspots might be regulated by Setdb1.

Altogether, hotspots might be used by activators to control expression of neighbouring genes as they were enriched for differentially expressed genes after perturbation of activators in ES cells. Furthermore, we suggested that ES hotspots might be under Setdb1 control for silencing.

Conclusion

The combinatorial action of multiple TFs controls regulation of gene expression where TFs activate or repress gene expression via binding to cis-regulatory sequence elements in gene loci. Identification of functional target genes regulated by specific transcription factors is critical for understanding the molecular mechanisms behind transcriptional control. Multi-TF ChIP sequencing studies have led to a remarkable observation that there were regions in the genome bound by almost all of the factors, the so called ‘Transcriptional Hotspots’ [1-4]. These regions acted as active enhancers of developmentally essential genes [3]. On the other hand, low occupancy regions did not drive reporter gene expression [6] questioning the functional significance of these peaks.

In order to study whether there are indeed classes of peaks with distinct characteristics based on their occupancy, we collected 108 TF binding ChIP sequencing datasets in 10 murine cell types. All peaks in a cell type were classified into three groups based on the number of factors bound: singletons (low occupancy), combinatorial (mid occupancy) and hotspots (high occupancy). Singletons were depleted in promoters and 3’ UTR, had a low peak height and were less conserved across multiple biological replicates putting in doubt their functional relevance. On the other hand, conserved cis-regulatory sequences were thought be more functionally relevant but the three groups could not be distinguished with respect to sequence conservation. Indeed, it has been shown in yeast that the low affinity protein-DNA interactions were conserved as well as being functionally relevant [32].

The hierarchical clustering of 30 peak sets (three groups each in 10 cell types) clustered combinatorials according to occupancy. Putative target gene sets of singletons and combinatorial peaks clustered into one tight cluster whereas each hotspot gene set consisted of unique set
of genes with cell type specific functional gene enrichment. The presence of hotspot regions near cell type specific genes has previously been observed in many cell types across species (e.g. Siersbaek et al. [18]). We have shown previously that most TF peaks cluster according to the cell type [33].

GAGA and Zelda signature sequence motifs were associated with hotspots in C. elegans and D. melanogaster [16]. We did not find a signature sequence motif for any of the groups but hotspots were enriched for sequence motifs of many key transcription factors as well as for pioneering factors in a cell type. Hotspots were specifically depleted for CTCF sequence motifs across all cell types. Though depleted in CTCF sequence motifs compared to singletons, the fraction of CTCF peaks and their mean peak height at hotspots is higher (data not

---

**Figure 5** ES hotspots are preferentially occupied by activators. A. Overlap of genes in three groups in ES Ng dataset with up- and down-regulated genes upon Oct4 knockdown in ES cells showing hotspots were enriched for only down-regulated genes. B. Fraction of up- and down-regulated genes upon Dnmt and Set1db knock out in three groups in ES Ng dataset showing hotspot enrichment only in Set1db KO down-regulated genes.
shown). We further intersected three groups with diverse chromatin marks and concluded that hotspots were enriched for enhancers present in the vicinity of highly expressed genes. Moreover, we identified a chromatin modification that can distinguish hotspots from singletons. Bimodal H3K3me1 peaks were specifically associated with hotspots, and mono-modal peaks with singletons. This suggests that three events, binding by multiple factors, creation of a nucleosome-depleted region at the binding site and establishment of bivalent H3K4me1 marks go hand in hand.

We so far confirmed that hotspots had strong occupancy (high peak height) and were enriched for cell type specific enhancers but their functional relevance is still a puzzle. There are various theories such as it has been suggested that they might serve as sinks for TFs [34]. In humans, as hotspots were depleted for TF motifs, hotspots were thought to be highly open chromatin marks where TFs bind non-specifically. Contrary to this, we find hotspots enriched for TF motifs especially for pioneering cell type specific factors such as Pu.1 in B cells. Pu.1 binding has been shown to initiate nucleosome modeling followed by H3K4 mono-methylation [7]. Our hypothesis therefore is that pioneering factors establish stable open chromatin regions at hotspots by recruiting specific chromatin marks which in turn facilitates binding of additional TFs at their motifs or non-specifically.

Finally, as hotspots are cell type-specific active enhancer regions, they need to be silenced in other cell types. Hotspot genes in ES cells significantly overlapped with genes perturbed after Setdb1 knockout and not Dnmt3 knockout suggesting that Setdb1 might play a role in ES hotspot silencing.

**Methods**

**Data collection and classification**

Genome wide binding patterns of transcription-related factors were obtained for five blood lineages (B cells, erythroid, hematopoietic progenitor cells (HPC), macrophage and T cells) [35], two embryonic stem cells [36], dendritic cells [37], MEL cells [38]. Peaks were called, with default parameters, using the peak calling program MACS [39] using an input control if provided in the original study. In total 108 peak sets were collected across 10 cell types (Additional file 1: Table S31). The binding events (peaks) of all transcription-related factors in a given cell type were merged to produce the total number of peaks in each cell type. These peaks were then divided into three groups according to the occupancy. To divide the regulatory peaks into three groups, we defined singletons as the genomic regions bound by one factor only in a cell type; hotspots were defined as genomic regions bound by more than five transcription-related factors. The choice of criterion for defining hotspots is discussed in detail in Additional file 3. Combinatorials were then defined as peaks which did not belong to either of the two groups. Additional file 3: Tables S1–30 are the peaks in BED format using mm9 genome assembly for all 30 peak sets (three peak lists corresponding to the three groups in each cell type).

**Clustering, overlap and enrichment calculation**

Genomic location bias such as promoters, 5’ UTR, 3’ UTR, exon and intron as well as known and de-novo cis-regulatory sequence motif enrichments and functional enrichments were calculated using HOMER [7]. The promoter was defined as -1 kb to +100 bp region around transcription start site. Mammalian conservation tracks were downloaded from the UCSC genome browser. To calculate peak height, each peak was made of uniform width 400 bp and the total number of tags in the 400 bp window was used as a proxy for peak height. UCSC gene annotation was used to map peaks to genes. Peaks present in the promoter or gene body were assigned to the corresponding gene and inter-genic peaks were assigned to the nearest gene within 50 kb (both upstream and downstream). The total of unique genome-wide peak locations was 408,002 which mapped to 21,607 unique genes. A binary peak matrix of 408,002 rows and 30 columns (three groups in 10 cell types) was created where ‘1’ represents presence of a peak and ‘0’ the absence. Similarly, a gene matrix of 21,607 rows and 30 columns was created. Hierarchical clustering of peak and gene matrices was performed using Pearson’s correlation coefficient as a similarity measure and complete linkage. All boxplots and bar charts were generated using R. Genomic sequences of the peaks were downloaded using UCSC galaxy tool and the occurrences of cis-regulatory sequence motifs were obtained using TBFSeach [40]. The percentage of total hotspot peaks occupied by each transcription-related factor in a given cell type was calculated and TFs with more than 90% hotspot binding were tabulated in Figure 4 (Additional file 2: Table S37 for full list). Statistical significance of the differences between groups was calculated using the Mann Whitney U test.

**Additional datasets**

CpG islands in mouse were downloaded from CpG_MI [41]. Methylation density plots for the three groups using H3K4me1 data in MEL cells were generated using SeqMiner [42]. Experimentally validated tissue specific enhancers (VISTA enhancers) in mouse were obtained from VISTA [19]. The lists of differentially expressed genes upon overexpression of transcription factors in embryonic stem (ES) cells were obtained from the data of Nishiyama et al. [27]. ChIP sequencing data for chromatin modifications in ES cells was obtained from Teif et al. [24] whereas methylation profiling in ES cells was
obtained from Stadler et al. [13]. The list of super-enhancers in ES cells were downloaded from Whyte et al. [11] and the genomic regions differentially occupied in Oct4+/− compared to Oct4+/+ were downloaded from Karwacki-Neisius et al. [12]. Differentially expressed genes at multiple time points after Oct4 knockdown were obtained from Hall et al. [25] and Sharov et al. [26]. Differentially expressed genes after Baf250 and Hdacl were obtained from Gao et al. [28] and Zupkovitz et al. [29]. For consistency the differentially expressed gene lists after perturbation of transcription factors in ES cells were compared against both the ES Ng dataset (Figure 5) and the ES Young dataset (Additional file 1: Figure S6). The RNA sequencing data for ES cells is available at GEO (GSE42152).

Availability of supporting data
This study was purely analysis of existing public data sets. The raw data used in this study is available at GEO with accession numbers GSE11431, GSE31233, GSE11724, GSE36104, GSE36030, GSE22178, GSE21314, GSE33913, GSE12346, GSE21512, GSE35024, GSE21978, GSE21614, GSE30142 and GSE29193.

Additional files

Additional file 1: Supplementary data and supplementary tables 31-40.
Additional file 2: Supplementary figures.
Additional file 3: Supplementary tables 1:30.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
AJ conceived the project, carried out analysis and wrote the paper.

Authors’ information
AJ is leads a bioinformatics research group within the developmental biology division at the Roslin Institute, Edinburgh, UK.

Acknowledgement
I thank Professor B. Gottgens and his team members at the University of Cambridge, UK for the data collection and discussions. I thank Anna Mantoski for RNA sequencing data analysis. I also thank Patricia Hart and Tom Michoel at the University of Cambridge, UK for the data collection and discussions. I thank Anna Mantsoki and Professor B. Gottgens and his team members at the University of Edinburgh Chancellors Fellowship awarded to AJ and strategic funding from the Biotechnology and Biological Sciences Research Council.

Received: 2 June 2014 Accepted: 8 December 2014
Published online: 30 December 2014

References
1. Wilson NK, Foster SD, Wang X, Knozevic K, Schütte J, Kaimakis P, Chlarska PA, Kinston S, Ouwendijk WH, Dzierzak E, Pimanda JE, de Bruin MFFR, Gottgens B: Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell 2010, 7:32–44.
2. Moorman C, Sun LV, Wang J, de Wit E, Talhout W, Ward LD, Greil F, Lu X-J, White KP, Bussiemaker HJ, van Steensel B: Hotspots of transcription factor colonization in the genome of Drosophila melanogaster. Proceedings of the National Academy of Sciences 2006, 103(1):12027–12032.
3. 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, 26:908–913.
4. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtschaffen A, Ieeng A, Alves P, Chateigner A, Perry M, Morris M, Auerbach RB, Feng S, Xiong L, Veli E, Aiu N, Wu J, Rhiannonika K, Agarwal A, Alexander RP, Barber G, Bickel CW, Brennan J, Brouillet JG, Carr A, Cheng-M CM, Clavson H, Contrino S, et al: Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science 2010, 330:1775–1787.
5. Yip KY, Cheng C, Bhardwaj N, Brown JB, Ling J, Kundaje A, Rozovsky J, Birney E, Bickel P, Snyder M, Gerstein M: Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biology 2012, 13:948.
6. Fisher WW, Li JJ, Hammonds AS, Brown JB, Pfeiffer BD, Weissmann R, MacArthur S, Thomas S, Stamatoyanopoulos JA, Eisen MB, Bickel PJ, Biggin MD, Celniker SE: DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. Proc. Natl. Acad. Sci. U.S.A. 2012, 109:21330–21335.
7. Heira S, Bennet C, Spann N, Bertolino E, Lin YC, Lasko P, Cheng JX, Murre C, Singh H, Glass CK: Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 2010, 38:576–589.
8. Ang Y-S, Tsai S-Y, Lee D-F, Monk J, Su J, Ratnakumar K, Ding J, Ge Y, Harr D, Chang B, Wang J, Rendi M, Bernstein E, Schaniel C, Lembicka R: Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. Cell 2011, 145:185–197.
9. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Ofos OY, Zhang W, Jiang J, Loh Y-H, Yeo H-C, Yeo ZX, Naranang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bouqueir G, Sung W-K, Clarke ND, Wei C-L, Ng H-H: Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 2008, 133:1106–1117.
10. Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM: Drosophila modENCODE. Nature 2012, 490:459–469.
11. Liao B-Y, Zhang J, Davis S, Minovitsky S, Dubchak I, Pennacchio LA: Rapid identification of regulatory elements in human enhancer databases. Genome Biology 2013, 14:R60.
12. Liao B-Y, Zhang J: Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 2013, 153:307–319.
13. Karwacki-Neisius V, Göke J, Ociro R, Halbritter F, Ng JH, Weiße AY, Wong FCO, Gagliardi A, Mullin NP, Festuccia N, Colby D, Tomlinson SR, Ng H-H, Chambers I: Reduced Oct4 expression directs a robust pluripotent state with distinct signaling activity and increased enhancer occupancy by Oct4 and Nanog. Cell Stem Cell 2013, 12:531–545.
14. Young RA: Connecting microRNA genes to the core regulatory circuitry of embryonic stem cells. Cell 2008, 134:521–533.
15. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin C, Kaye MH, Rahl PB, Lee TI, Young RA: Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 2013, 153:307–319.
16. Davies AM, Mathis S, Skepper JN, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM: In vivo enhancer
analysis of human conserved non-coding sequences. Nature 2006, 444:499–502.
21. Creighton MP, Cheng AW, Welstead GD, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R: From the Cover: Histone H3K27ac separates active from poised enhancers and predicts developmental state. P Natl Acad Sci 2010, 107:21931–21936.
22. Hoffmann BG, Robertson G, Cawagna B, Beach M, Cullum R, Lee S, Soukhatcheva G, Li L, Wederell ED, Thiesen N, Bilenky M, Cezar D, Tam A, Kanoh B, Birol I, Dai D, Zhao Y, Hirst M, Vercherie CB, Heltgason CD, Mara MA, Jones SJM, Hoodless PA: Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate the functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver. Genome Research 2010, 20:1037–1051.
23. Zhang X, Robertson G, Wao S, Moffit BS, Gattardo R: Probabilistic inference for nucleosome positioning with MNase-Based or Sonicated Short-Read Data. PLoS ONE 2012, 7:e32095.
24. Teif VB, Vainshtein Y, Caudron-Herger M, Mallm J-P, Marth C, Höfer T, Rippe K: Genome-wide nucleosome positioning during embryonic stem cell development. Nat. Struct. Mol. Biol. 2012, 19:1185–1192.
25. Hall J, Guo G, Wray J, Eyres I, Nichols J, Grotewold L, Morfopoulou S, Humphreys P, Mansfield W, Walker R, Tomlinson S, Smith A: Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. BMC Genomics 2008, 9:269.
26. Nishiyama A, Xin L, Sharov AA, Thomas M, Mower G, Meyers E, Piao Y, Mehta S, Yee S, Nakatake Y, Correa-Cerro LS, Li L, Wederell ED, Li W, Goodell MA: Identifying two novel approaches to identify functional CpG islands in mammalian genomes. Nucleic Acids Research 2010, 38:R80.
27. Zupkovitz G, Tischler J, Posch M, Sadzak I, Ramsauer K, Egger G, Grausenburger R, Schweifer N, Chiocca S, Deckter T, Seiser C: Negative and positive regulation of gene expression by mouse histone deacetylase 1. Mol. Cell. Biol. 2006, 26:7913–7928.
28. Kamoh B, Birol I, Dai D, Zhao Y, Hirst M, Vercherie CB, Heltgason CD, Mara MA, Jones SJM, Hoodless PA: Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs by systematic integration of factor binding data. Cell Stem Cell 2009, 5:420–433.
29. Challen GA, Sun D, Jeong M, Luo M, Jelinek J, Berg JS, Bock C, Humphreys P, Mansfield W, Walker R, Tomlinson S, Smith A: Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. BMC Genomics 2008, 9:269.
30. Zupkovitz G, Tischler J, Posch M, Sadzak I, Ramsauer K, Egger G, Grausenburger R, Schweifer N, Chiocca S, Deckter T, Seiser C: Negative and positive regulation of gene expression by mouse histone deacetylase 1. Mol. Cell. Biol. 2006, 26:7913–7928.
31. Karimi MW, Goyal P, Makakowa IA, Bilenky M, Leung D, Tang JX, Shinkay K, Mage DL, Jones S, Hirst M, Lorincz MC: DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. Cell Stem Cell 2011, 8:576–587.
32. Challen GA, Sun D, Jeong M, Luo M, Jelinek J, Berg JS, Bock C, Humphreys P, Mansfield W, Walker R, Tomlinson S, Smith A: Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. BMC Genomics 2008, 9:269.
33. Zupkovitz G, Tischler J, Posch M, Sadzak I, Ramsauer K, Egger G, Grausenburger R, Schweifer N, Chiocca S, Deckter T, Seiser C: Negative and positive regulation of gene expression by mouse histone deacetylase 1. Mol. Cell. Biol. 2006, 26:7913–7928.
34. Karimi MW, Goyal P, Makakowa IA, Bilenky M, Leung D, Tang JX, Shinkay K, Mage DL, Jones S, Hirst M, Lorincz MC: DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. Cell Stem Cell 2011, 8:576–587.
35. Challen GA, Sun D, Jeong M, Luo M, Jelinek J, Berg JS, Bock C, Humphreys P, Mansfield W, Walker R, Tomlinson S, Smith A: Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. BMC Genomics 2008, 9:269.
36. Zupkovitz G, Tischler J, Posch M, Sadzak I, Ramsauer K, Egger G, Grausenburger R, Schweifer N, Chiocca S, Deckter T, Seiser C: Negative and positive regulation of gene expression by mouse histone deacetylase 1. Mol. Cell. Biol. 2006, 26:7913–7928.
37. Challen GA, Sun D, Jeong M, Luo M, Jelinek J, Berg JS, Bock C, Humphreys P, Mansfield W, Walker R, Tomlinson S, Smith A: Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. BMC Genomics 2008, 9:269.
38. Zupkovitz G, Tischler J, Posch M, Sadzak I, Ramsauer K, Egger G, Grausenburger R, Schweifer N, Chiocca S, Deckter T, Seiser C: Negative and positive regulation of gene expression by mouse histone deacetylase 1. Mol. Cell. Biol. 2006, 26:7913–7928.
39. Zupkovitz G, Tischler J, Posch M, Sadzak I, Ramsauer K, Egger G, Grausenburger R, Schweifer N, Chiocca S, Deckter T, Seiser C: Negative and positive regulation of gene expression by mouse histone deacetylase 1. Mol. Cell. Biol. 2006, 26:7913–7928.