Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

A ChIP-on-chip tiling array approach detects functional histone-free regions associated with boundaries at vertebrate HOX genes

Surabhi Srivastava, Divya Tej Sowpati, Hita Sony Garapati [1], Deepika Puri, Jyotsna Dhawan, Rakesh K. Mishra *

Centre for Cellular and Molecular Biology, Council for Scientific and Industrial Research, Uppal Road, Hyderabad, Andhra Pradesh 500007, India

ARTICLE INFO

Article history:
Received 24 April 2014
Available online 10 May 2014

Keywords:
ChIP-on-chip
Chromatin domain boundary
Enhancer blocking
GAGA factor
Hox
Histone H3

ABSTRACT

Hox genes impart segment identity to body structures along the anterior–posterior axis and are crucial for proper development. A unique feature of the Hox loci is the collinearity between the gene position within the cluster and its spatial expression pattern along the body axis. However, the mechanisms that regulate collinear patterns of Hox gene expression remain unclear, especially in higher vertebrates. We recently identified novel histone-free regions (HFRs) that can act as chromatin boundary elements demarcating successive murine Hox genes and help regulate their precise expression domains (Srivastava *et al.*, 2013). In this report, we describe in detail the ChIP-chip analysis strategy associated with the identification of these HFRs. We also provide the Perl scripts for HFR extraction and quality control analysis for this custom designed tiling array dataset.

Specifications	
Organism/cell line/tissue	*Mus musculus/C2C12 myoblasts*
Sex	*Female*
Sequencer or array type	*Agilent ChIP-chip custom tiling array (AMADID-0245671)*
Data format	*Raw data: gunzipped txt and analyzed data: bed format*
Experimental features	*Suspension culture to induce synchronized reversible G0 arrest; ChIP for histone H3*

Direct link to deposited data

Deposited data can be found here:http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42941

Experimental design, materials and methods

Array design and data normalization

A custom 1 million probe tiling array chip was designed (manufactured by Agilent Technologies) representing 52 Mb of the mouse genome (NCBI Build 37), of which 1.1 Mb was tiled from the four Hox clusters. The entire genomic sequence of the Hox clusters was tiled including all the flanking regions till the neighbouring genes at either side in each cluster. Approximately 887,000 repeat masked probes were designed for the array, of which 16,200 probes were specific for the Hox clusters. The size of each probe was 60 bp and these were designed with a 10 bp overlap between successive probes to achieve high resolution.

The tiling arrays were subjected to dual color hybridization with Input DNA and ChIP samples obtained by chromatin immunoprecipitation using standard protocols as described in [1] with an antibody designed against the invariant portion of the core histone H3 (Abcam, #ab1791). This antibody recognizes all forms and modifications of histone H3 and hence is useful to report for nucleosomal presence. Following background subtraction and data extraction by Agilent's Feature Extraction, normalized enrichment values for the probes were identified using DNA Analytics (Agilent) software. Probe data were normalized using default blank subtraction and intra-array dye-bias median normalization against the Input to obtain normalized log ratio (NLR) IP/Input values. Fig. 1 shows the distribution of the log ratio values for all the probes in the array. The data were deposited in the Gene Expression Omnibus (GEO; [2]) database. Probe enrichment was visualized on the Mouse NCBI37/mm9 Assembly in the UCSC genome browser [3].

Identification of histone-free regions

To detect histone-free regions (HFRs), the dataset was mined to extract contiguous probes from the Hox clusters that showed no positive enrichment with the histone H3 antibody. A custom Perl script

Abbreviations: ChIP, chromatin immunoprecipitation; DTT, dithiothreitol; GAF, GAGA binding factor; HFR, histone-free region; HS, hypersensitive; NLR, normalized log ratio; PMSF, phenylmethanesulfonylfluoride.

* Corresponding author. Tel.: + 91 40 2719 2658; fax: +91 40 2716 0591.
E-mail address: mishra@ccmb.res.in (R.K. Mishra).

[1] Present address: School of Life Sciences, University of Hyderabad, Hyderabad, Andhra Pradesh 500007, India.

http://dx.doi.org/10.1016/j.gdata.2014.05.001
2213-5960/© 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

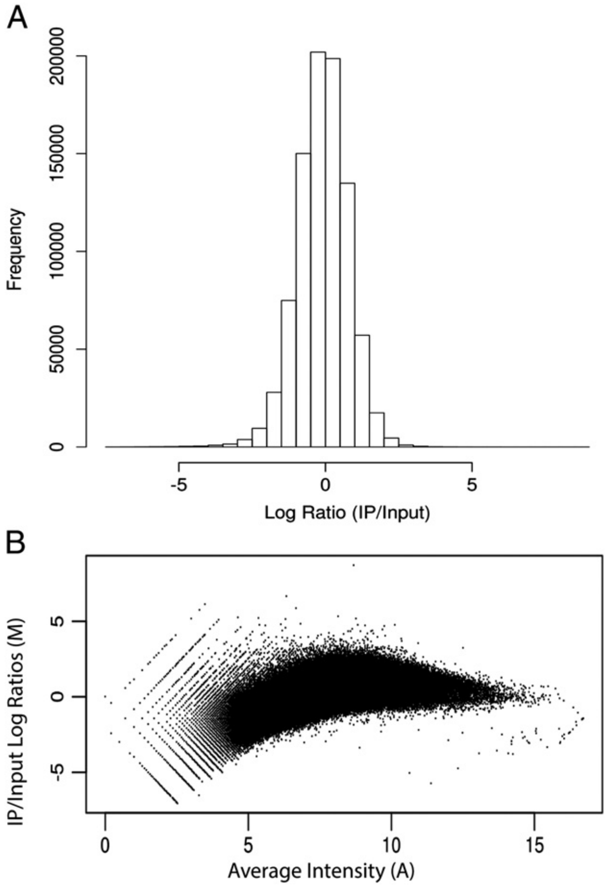

Fig. 1. Probe intensity and distribution of log ratio values A) Histogram depicting frequency of log ratios in the Histone H3 ChIP-chip dataset. B) MA plot showing correlation between log ratios (*y* axis) and their respective average intensities (*x* axis). Log ratios are calculated as log2 (IP)/log2 (Input). Average intensity is given by the formula (log2 (IP) + log2 (Input)) / 2.

(Supplementary file 1) was used to identify blocks of regions that had probes with very low or negative NLR values, corresponding to HFRs.

The HFR script accepts three parameters for running, i) an input file containing information of all probes on the array, ii) an output file into which the coordinates of HFRs are printed out, and iii) the Hox cluster that is to be analyzed. The script works by first extracting from the array all the probe information of the Hox cluster specified by the user. Using the extracted probe information, the probe coordinates and their corresponding NLR values are stored into arrays. The next part of the script takes this array of normalized log ratios and returns blocks of H3-unenriched regions. A block is considered unenriched if 3 out of 5 probes contributing to the block have an NLR <0, and is classified as an ‘n’ block. If more than 2 probes have an NLR >0, it is classified as a ‘p’ block. The enrichment information for all consecutive blocks (n and p) is stored in a separate array. Using the three arrays of coordinates, normalized log ratios and enrichment information, the final part of the script returns a list of coordinates corresponding to regions unenriched for histone H3 or histone-free regions (HFRs). For this, starting with each negative (‘n’) block the negative region is extended till two consecutive ‘p’ blocks are encountered. The script also terminates negative regions where the genomic distance between two consecutive probes is greater than 200 bp. The minimum size cut-off for the negative regions is set to 500 bp. The final list of HFRs is output to a user-specified file in bed format. The total number of HFRs identified in each of the four Hox clusters is provided in Table 1.

Table 1
Total number of histone-free regions.

Cluster	No. of HFRs	Unenriched in H3K4me3 array	Unenriched in H3K27me3 array
HOXA	74	74	68
HOXB	22	22	18
HOXC	67	65	64
HOXD	52	52	47

The total number of HFRs identified in each Hox cluster. The number of HFRs that were also found unenriched in H3K4me3 and H3K27me3 arrays is indicated.

Example usage: perl hox_final.pl -in C2C12_G0_H3.txt -out hoxa_HFRs.bed -cluster hoxa

Quality control analysis

To confirm that the HFRs returned by the custom script were consistent, we analyzed probe binding data at these regions from two additional ChIP-chip custom arrays using modified histone H3 antibodies specific for H3K4me3 and H3K27me3. In most cases, more than 50% of the probes contributing to each of these histone-free regions were found to have NLR <0 in both the H3K4me3 and H3K27me3 arrays as well. The number of HFRs from each cluster found unenriched in both these arrays is listed in Table 1.

To further rule out the possibility of false detection of positively enriched regions as HFRs, we analyzed a test subset of probes from these tiling array datasets using our custom script. The test dataset was made of probes corresponding to genomic regions that were found positively enriched for either H3K4me3 or H3K27me3 in ChIP-qPCR experiments, and hence should not be negative for histone H3. As expected, no HFR was returned when this test dataset was submitted to the HFR script.

Real time qPCR validations

Real time primers were designed for sequences specific for ten intergenic histone-free regions from the four Hox clusters as well as ten control H3-positive regions from within the Hox gene bodies (Supplementary file 2: Table S1). Chromatin immunoprecipitation was performed with the histone H3 antibody and a non-specific IgG control antibody (Diagenode #kch-803-015). Real time quantitative PCR assays were performed to calculate the relative abundance of histone H3 at the target HFRs and at the control regions using Power SYBR Green qPCR Master mix (Applied Biosystems) on an ABI7900HT Fast Real-Time PCR System (2 min at 50 °C; 10 min at 95 °C; 40 cycles of 15 s at 94 °C, 30 s at 60 °C and 30 s at 68 °C; followed by dissociation curve analysis). Enrichment in the ChIP DNA was determined as percentage Input, where Input DNA represents an aliquot of the same crosslinked and sonicated chromatin used for ChIP and processed in parallel. The enrichment for H3 was then normalized to that observed for IgG at each locus (Fig. 2) and was found to be much lower at HFRs compared to control regions. The difference between the two groups was extremely statistically significant ($P < 0.0001$) using the paired *t*-test.

Sequence analysis

The sequences corresponding to the genomic coordinates of the predicted HFRs in each Hox cluster (including 10 kb upstream and downstream genomic regions of each cluster) were extracted from the mouse reference assembly (MGSCv37-C57BL/6 J; NCBI build 37.2) and analyzed for common motifs using a MEME online tool [4]. The number of expected motifs was set as 5 and the minimum and maximum motif widths were set as 4 and 8 respectively, while allowing for any number of repetitions of the motif on a sequence. The GAGA binding factor (GAF) motif was identified as the top hit with high significance ($p < 0.0005$) in HFRs at all the clusters except HoxB. The binding of *Th-POK* (vertebrate homolog of *Drosophila* GAF, [5]) at these regions was subsequently tested by ChIP-qPCR as described in [1].

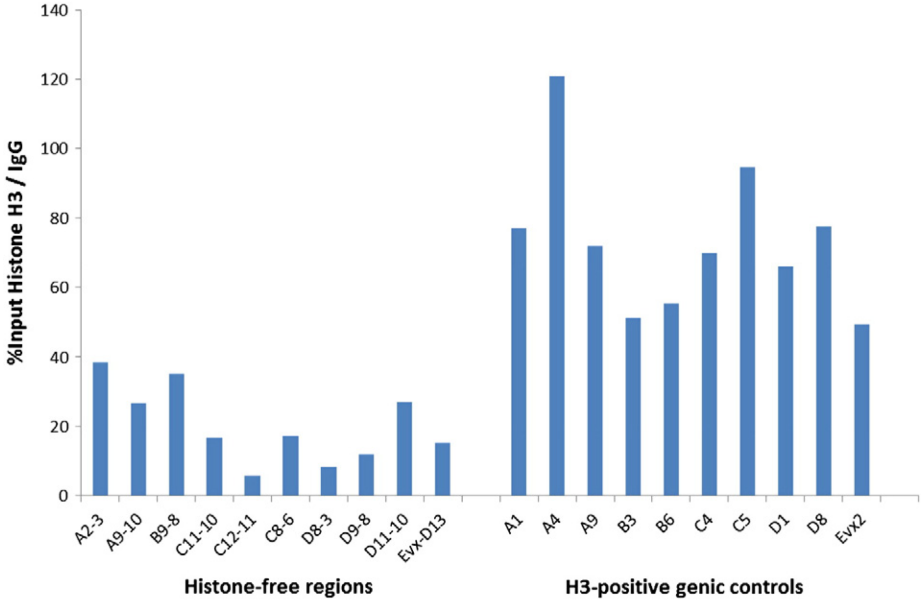

Fig. 2. HFRs show poor enrichment for histone H3. Presence of histone H3 at ten intergenic HFRs was assessed by ChIP with histone H3 antibody followed by real time quantitative PCR. Enrichment for histone H3 was found to be significantly lower at HFRs than at control regions using primers designed within Hox gene bodies.

To check if the predicted HFRs mapped to sites of DNaseI hypersensitivity and could therefore be considered as reliable markers for sites of histone disruption, the coordinates of DNaseI HS sites in skeletal muscle (tissue of origin), mesoderm and embryonic stem cells from the mouse ENCODE project (DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington) were obtained from UCSC browser. These were compared with the start and stop coordinates of each of the HFRs associated with the Hox clusters, including 10 kb upstream and downstream of each cluster (93 HFRs in total). DNaseI HS peaks called in any of the three cell types that either overlapped with histone free regions or lay within close proximity (500 bp) to an HFR were considered. The HFRs and DNaseI HS sites from the different cell types were mapped in the context of the Hox genes using R script. Table 2 lists all the HFRs that were found to be associated with a DNaseI HS peak in at least one cell type, supporting the hypothesis that these HFRs indeed represent histone-free regions likely to harbor sites for regulatory activity.

The coordinates of the non-coding transcripts within the Hox clusters were obtained from the TROMER transcriptome database. In case of overlapping transcripts in the same orientation, only the longer one was mapped for clarity. The co-ordinates of the 93 HFRs located within the Hox clusters (including 10 kb upstream and downstream of the clusters) were similarly compared with those for all the non-coding transcripts, CpG islands, gene bodies, and 5′ and 3′ ends of the genic regions (co-ordinates obtained from the UCSC mm9 database) and the majority of the HFRs showed no overlap with these known features of the Hox clusters.

Conclusions

We describe here the detailed methods and Perl script used to analyze the dataset obtained from a custom designed histone H3 ChIP-chip tiling array. The probe data offers very high resolution due to the array design thus enabling the identification of novel histone-free regions at the Hox clusters. This dataset has helped delineate novel *cis* elements that are likely involved in organizing higher order chromatin and governing the tightly regulated expression domains of vertebrate homeotic genes.

The HFR Perl script and real time qPCR primer sequences are provided as supplementary material. Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gdata.2014.05.001.

Acknowledgments

This work was supported by grants from the Indo-Australian Biotechnology Fund-Department of Biotechnology (DBT-IABF), Government of India (BT/Indo-Aus/05/36/2010). DP was supported by a Senior Research Fellowship from the Council for Scientific and Industrial Research (CSIR), India.

References

[1] S. Srivastava, D. Puri, H.S. Garapati, J. Dhawan, R.K. Mishra, Vertebrate GAGA factor associated insulator elements demarcate homeotic genes in the HOX clusters. Epigenetics Chromatin 6 (1) (Apr 22 2013) 8, http://dx.doi.org/10.1186/1756-8935-6-8.

[2] R. Edgar, M. Domrachev, A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30 (1) (Jan 1 2002) 207–210.

[3] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC. Genome Res. 12 (6) (Jun 2002) 996–1006 (http://genome.ucsc.edu/).

[4] T.L. Bailey, M. Boden, F.A. Buske, M. Frith, C.E. Grant, L. Clementi, J. Ren, W.W. Li, W.S. Noble, MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 37 (2009) W202–W208.

[5] N.K. Matharu, T. Hussain, R. Sankaranarayanan, R.K. Mishra, Vertebrate homologue of Drosophila GAGA factor. J. Mol. Biol. 400 (2010) 434–447.

Table 2
HFRs associated with DNase hypersensitive sites.

S. no	HFR	SM	MD	ESC
HoxA				
1	A_DOWN-1.1		✓	✓
2	**A_1-2.1**	✓	✓	✓
3	**A_1-2.2**	✓	✓	✓
4	A_1-2.3	✓	✓	
5	**A_2-3.1**	✓	✓	✓
6	**A_3-4.1**	✓	✓	✓
7	A_3-4.2		✓	
8	A_4-5.1		✓	
9	A_4-5.2		✓	✓
10	A_5.1		✓	✓
11	A_6-7.1			✓
12	A_6-7.2	✓	✓	
13	A_7-9.1	✓	✓	
14	**A_7-9.2**	✓	✓	✓
15	A_7-9.3	✓		
16	**A_9-10.1**	✓	✓	✓
17	A_10-11.1	✓		
18	A_10-11.2		✓	
19	A_11.1		✓	✓
20	A_11-13.1		✓	✓
21	**A_11-13.2**	✓	✓	✓
22	A_UP.1		✓	✓
23	A_UP.2		✓	✓
24	A_UP.3		✓	
HoxB				
1	B_13-9.1		✓	
2	B_13-9.2		✓	
3	**B_9-8.1**	✓	✓	✓
4	B_7-6.1		✓	
5	B_5-4.1		✓	
6	B_4-3.1		✓	
7	B_3.1		✓	
8	**B_3.2**	✓	✓	✓
9	B_2-1.1		✓	✓
10	B_2-1.4		✓	✓
HoxC				
1	C_UP.26		✓	✓
2	C_UP.27		✓	✓
3	**C_UP.29**	✓	✓	✓
4	C_13.2		✓	✓
5	C_12-11.1			✓
6	**C_12-11.2**	✓	✓	✓
7	C_12-11.3		✓	✓
8	C_11-10.1		✓	
9	**C_11-10.2**	✓	✓	✓
10	**C_11-10.3**	✓	✓	✓
11	**C_11-10.4**	✓	✓	✓
12	**C_11-10.5**	✓	✓	✓
13	C_10.1	✓		
14	**C_9.1**	✓	✓	✓
15	C_8-6.2	✓		
16	**C_8-6.3**	✓	✓	✓
17	C_8-6.4	✓	✓	
18	**C_6-5.1**	✓	✓	✓
19	C_5-4.1		✓	
20	C_5-4.2		✓	
21	**C_5-4.3**	✓	✓	✓
22	C_5-4.4		✓	
23	C_4-DOWN.1		✓	
24	C_DOWN.1	✓	✓	
HoxD				
1	**D_UP.13**	✓	✓	✓
2	**D_13.1**	✓	✓	✓
3	**D_11-10.1**	✓	✓	✓
4	**D_11-10.2**	✓	✓	✓
5	**D_10-9.1**	✓	✓	✓
6	**D_9-8.1**	✓	✓	✓
7	**D_8-4.1**	✓	✓	✓
8	D_8-4.2		✓	✓
9	**D_8-4.3**	✓	✓	✓
10	D_8-4.4		✓	
11	D_4-3.1		✓	✓
12	D_4-3.2		✓	
13	**D_3-1.1**	✓	✓	✓
14	D_3-1.2		✓	✓
15	D_3-1.5		✓	
16	D_1-DOWN.1			✓

The HFRs overlapping with DNaseI hypersensitive peaks in different tissues as obtained from ENCODE data are tabulated. HFRs that are consistently associated with DNaseI HS sites in all three tissues are highlighted in bold. SM = skeletal muscle, MD = mesoderm, ESC = embryonic stem cells.