Supplementary Methods

Evolutionary conservation

To anchor peaks from one reference species to peaks in the query species we used three methods: a) Evolutionary conserved blocks (ECR) from the ECR browser [1]. The closest flanking ECRs were retrieved for each peak, and examined whether this pair of ECRs also flanks a peak in the query species. b) Orthologous neighbouring genes retrieved from EnsEMBL [2]. The closest flanking genes were retrieved for each peak, and then determined whether the orthologs of this pair of genes also flanks a peak in the query species. c) Gene-by-gene approach. For each gene flanking a CTCF site, it was examined whether it has an ortholog in the query species also flanking a CTCF site and oriented the same way with respect to the bound site.

Correlations with genomic features

CTCF peaks were mapped to annotations (EnsEMBL genes) to identify the surrounding genes. Each peak was modelled as its central coordinate, and could fall into one of the following two broad categories: intergenic or intragenic.

The correlations or anti-correlations with various genomic features have been performed by computing the CTCF binding sites sequence overlap with those features. Relevant overlaps correspond to 50% coverage of the smallest element. To generate the control sets for P-value calculations, we first screened the 3 genomes with the CTCF matrix from Transfac (M01200 V$CTCF_01) with a low threshold (70% of the highest score achievable using the 10 most conserved positions from the matrix) to obtain a set of pseudo CTCF sites, then removed overlapping sites by selecting those ones having the best score (7765459, 21079470 and 20486540 sites in chicken, human and mouse respectively), and applied our method to detect CONSYN sites. We ended up with

243284 conserved site triplets, and built 1000 random sets of 247 elements from these pseudo-CONSYN. Then we followed the same process than previously described to match them to LADs and TF-encoding genes in order to assess the relevance of the over- and under-representations we observed for the real set of 247 CONSYN sites. In addition, we also built 1000 purely random sets of 247 elements.

Enhancer-blocking assay

Chosen CTCF elements were amplified from human genome with primers indicated in Supplementary Table 4 and subcloned in TOPO vector. DNA fragments were then cloned between the CMV enhancer and the promoter (XhoI) or upstream of the CMV enhancer (PstI). Data are presented as fold enhancer-blocking activity normalized to the value achieved by the basal pELuc vector. The 1.2 kb chicken 5´HS4 b-globin insulator element was used as positive control. Similarly, the internal II/III boxes, both wild-type and mutated, from the chicken 5'HS4 b-globin insulator element, corresponding to the wild-type and mutated CTCF target binding sites, were used as positive and negative controls, respectively [3].

Repressor luciferase assays

The primers used to amplify the CNRs were: CNRA-F: 5'-TTCCCACTCATGAGGTTTCC-3', CNRA-R: 5'- GCTGTGGCTTTCTTCTAGCC-3', CNRB-F: 5'- TGTAGATGACGGGTTGATGG-3' and CNRB-R: 5'-TTAACAGTGTGCCACCTTGG-3'.

Human THP1 cell line (ATCC TIB-202) was maintained in RPMI 1640 medium, supplemented with 10% fetal calf serum, 2.0 mM glutamine, 10 mM HEPES, 1 mM sodium pyruvate, 4.5 g/l glucose, 100 μg/ml streptomycin, 100 U/ml penicillin, 0.05

mM 2-mercaptoethanol.. Cells were transfected by nucleofection Amaxa® Cell Line Nucleofector® Kit V (Lonza) according to manufacture protocols. Briefly, exponentially growing THP1 cells (2×10^6) were mixed with CNR-pGL3 control plasmid DNA (1 μg) and 0.2 μg of Renilla luciferase reporter gene plasmid pRL-TK (Promega) for normalizing transfection efficiency. The transfected cells were cultured in 6 well-plates for 24 h and treated with phorbol-myristate acetate (PMA, 50 ng/ml) and ionomycin (Io, 10 ng/ml) (Sigma-Aldrich Quimica) for 4 h. Cells were harvested and washed twice in PBS at 4 °C. Luciferase activity was evaluated using Dual-Luciferase® Assay System (Promega) according to manufactures. For analysis, lysates were thawed on ice and the protein concentration was measured by the Bradford method. Then 10 μg of protein from all supernatants were analyzed for both firefly and Renilla luciferase activities in luminometer F12 (Berthold Detection Systems).

Bood cell preparation and qRT-PCR

Blood cells were subjected to density gradient centrifugation by Histopaque-1077 (Sigma-Aldrich, St. Louis, MO 63103, USA) to obtain a homogeneous population of mononuclear cells (PBMCs). We cultured PBMCs in RPMI supplemented with 2mM Glutamine, 50 μM ß-Mercaptoethanol and 10% inactivated fetal bovine serum. Cultured cells were stimulated with 50 ng/ml of PMA and 10 ng/ml Io for 4h at 37ºC and 5% CO_2. qRT-PCR was performed as reported [4] with forward (5'-TCCACACACCTGCTTATCCA-3') and reverse (5'-CCTTCCTCTGGAAACCCTTC-3') primers.

Zebrafish morpholino injections

To evaluate the inhibition of *MOsp1* on CTCF mRNA splicing, we designed primers in exons 2 and 5 (5'- GGAAGAAGAAATGGCTGAACC -3' and 5'- GGCATAACTGCACAGACTGC -3'). A band of 726 bp should be produced only if the morpholino inhibits correct removal of intron 2. The inclusion of this intron in the mRNA introduces several precocious stop codons. For RT-PCR total RNA was extracted at 48 hpf from 25 morphants and control embryos and amplification was carried out for 30 cycles. The following primers were used to detect *gfi1* and *evi5* mRNA levels in control and injected embryos by qRT-PCR assays: evi5F (5'- GCAGAGATATTGCCCGTACC-3'), evi5R (5'- TGACAGTAGCCCACTTCTCG -3'), gfi1F (5'- GACCACCATCACCATCAGC -3'), gfi1% (5'- GATTTCACACCCCGAGTACC -3'). In these experiments we used the *elfa* gene, detected with primers *elfa*F (5'- CTTCTCAGGCTGACTGTGC-3') and *elfa*R (5'- CCGCTAGCATTACCCTCC-3'), as an internal control, as reported [5].

Chromosome conformation capture assay (3C)

Control and PMA+Io-activated human PBMCs were processed as reported [6]. $3x10^7$ cells were fixated with 2% PFA, lysed, and nuclei were digested with HindIII (Roche). DNA was then ligated with T4 DNA ligase (Promega) in low concentration conditions to favour intramolecular ligations. A set of locus-specific primers (Supplementary Table 4) were designed close to the HindIII sites for a 51kb LD block region containing the SNPs rs11804321 and rs11581176. Two primers close to *EVI5* and *GFI1* promoters were acting as fixed primers. Interactions between each of the promoters and the 51kb SNPs-containing region were tested in both control and activated cells. PCR products were run in a 2% agarose gel and quantified using a Typhoon scanner.

Product values were normalised to that obtained in a control sample composed of two BAC clones (RP11-515G5 and RP11-640O24) containing all test fragments in equal amounts [7]. This control sample also allows verifying that all combinations of pair of primers properly amplify the single expected PCR bands. BAC clones were mixed in equimolar amounts, digested with HindIII and ligated with T4 ligase. In mouse, a ubiquitously expressed gene, *ERCC3* (excision repair cross-complementing rodent repair deficiency, complementation group 3) has been previously used to correct for differences in crosslinking and digestion efficiencies between samples [8]. In order to normalize our data, we used the human *ERCC3* locus (human ERCC3 BAC clone: RP11-313N8).

REFERENCES

1. Ovcharenko, I., Nobrega, M.A., Loots, G.G. & Stubbs, L. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res* **32**, W280-6 (2004).
2. Hubbard, T.J. et al. Ensembl 2009. *Nucleic Acids Res* **37**, D690-7 (2009).
3. Recillas-Targa, F., Bell, A.C. & Felsenfeld, G. Positional enhancer-blocking activity of the chicken beta-globin insulator in transiently transfected cells. *Proc Natl Acad Sci U S A* **96**, 14354-9 (1999).
4. Hamalainen, H.K. et al. Identification and validation of endogenous reference genes for expression profiling of T helper cell differentiation by quantitative real-time RT-PCR. *Anal Biochem* **299**, 63-70 (2001).
5. McCurley, A.T. & Callard, G.V. Characterization of housekeeping genes in zebrafish: male-female differences and effects of tissue type, developmental stage and chemical treatment. *BMC Mol Biol* **9**, 102 (2008).

6. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-11 (2002).

7. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Methods* **3**, 17-21 (2006).

8. Splinter, E. et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* **20**, 2349-54 (2006).

Supplementary Figure 1

a Mouse data

Mapping

	mouse ES		mouse MEF	
Total Reads	17.005.213		16.392.325	
Mapped Reads	7.174.521	42.19%	6.113.199	37.29%
0 Mismatch	1.697.681	9.98%	1.412.030	8.61%
>0 Mismatch	5.476.840	32.21%	4.701.169	28.68%
Repeated	1.960.267	11.53%	1.804.971	11.01%
Not Mapped	7.870.425	46.28%	8.474.155	51.70%

Peak calling

	ES	MEF
Nb of Reads	7.174.521	6.113.199
Nb of Sites	**44.424**	**53.736**
Nb Reads in Sites	4.390.916	3.824.819
Avg. Reads per Site	98.84	71.18
% Reads in Sites	61.2	62.57
Avg Site Length	95.40	83.13

b Chicken data

Mapping

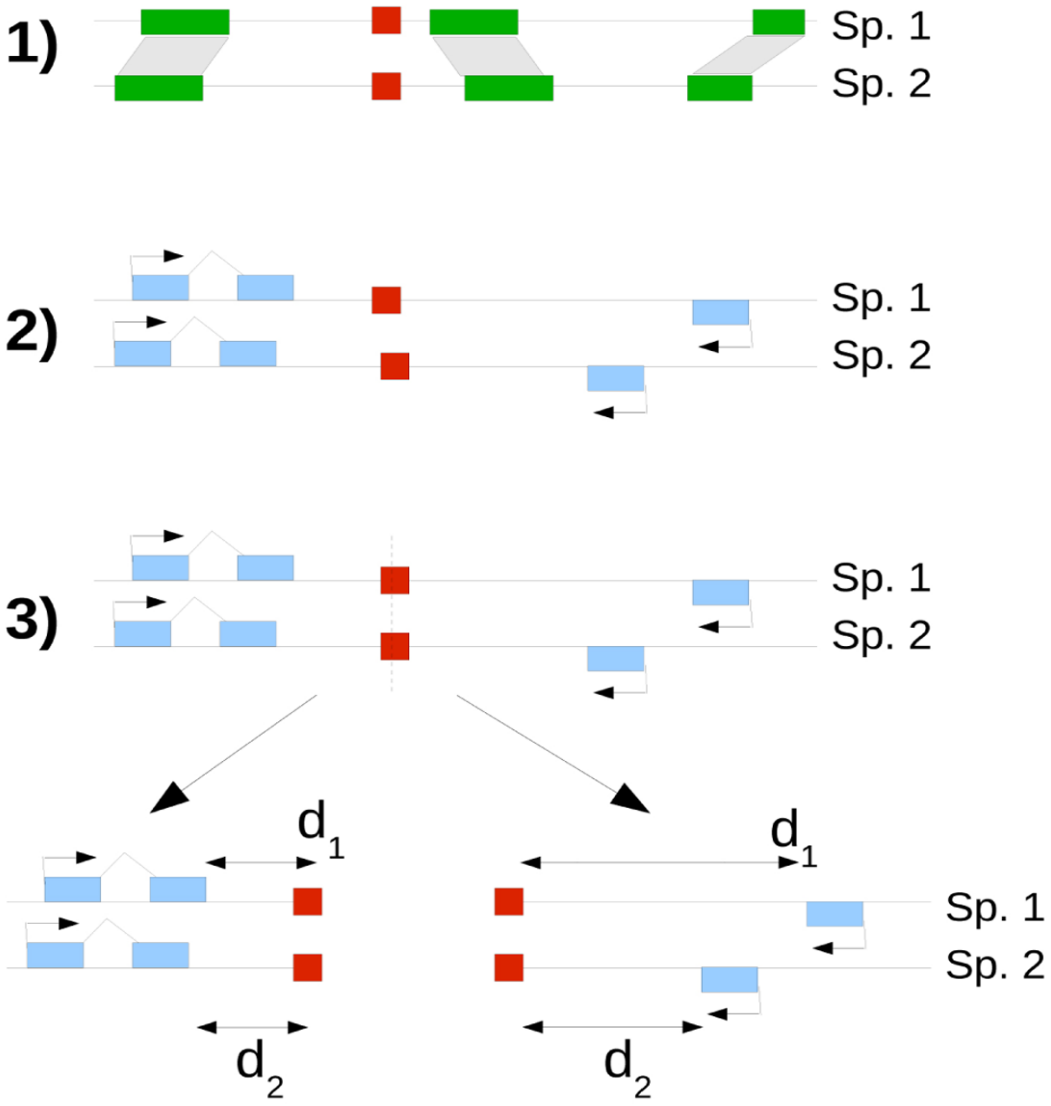
	RBC 5 days		RBC 10 days	
Total Reads	8.793.714		24.531.250	
Mapped Reads	5.278.305	60.02%	15.359.697	62.61%
0 Mismatch	3.831.282	43.57%	10.757.606	43.85%
1 Mismatch	1.014.491	11.54%	3.246.954	13.23%
2 Mismatches	432.532	4.92%	1.355.137	5.52%
Repeated	999.377	11.36%	3.434.599	14.00%
Not Mapped	2.516.032	28.61%	5.736.954	23.38%

Peak calling

	RBC 5 days	RBC 10 days pooled
Nb of Reads	5.278.305	15.359.697
Nb of Sites	**5.830**	**21.358**
Nb Reads in Sites	183.575	696.540
Avg. Reads per Site	31.49	32.61
% Reads in Sites	3.48	4.53
Avg Site Length	114.1	105.6

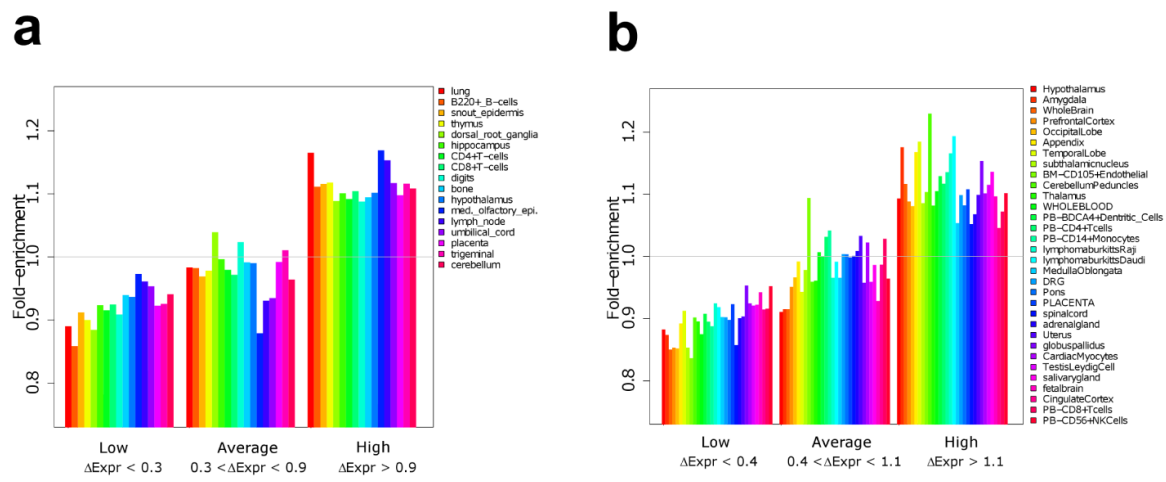
Supplementary Figure 1. Mapping statistics and peak calling. Mapping tables describe read mapping statistics. The mapped reads row refers to the number of reads mapping onto the reference genome uniquely and the respective percentage relative to the total number of reads from sequencing. The peak calling tables provide all numbers relative to the peak calling process, and the number of sites row refers to the number of peaks found in each sample.

Supplementary Figure 2



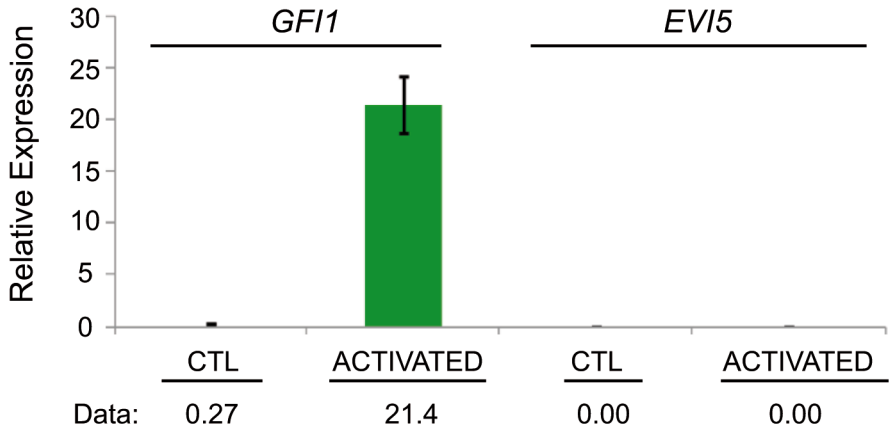
Supplementary Figure 2. Methods for assessing inter-species conservation of CTCF binding sites. Three methods for assessing inter-species conservation of CTCF binding sites: 1- a site in both species (sp.1 and sp.2) is anchored to contiguous flanking ECRs (evolutionary conserved regions). 2- a site is located in the same locus (orthologous genes) in the two species. 3- genes flanking CTCF peaks in two species are orthologous and oriented the same way relative to the peak, at a distance below threshold d1 (species1) and d2 (species2).

Supplementary Figure 3



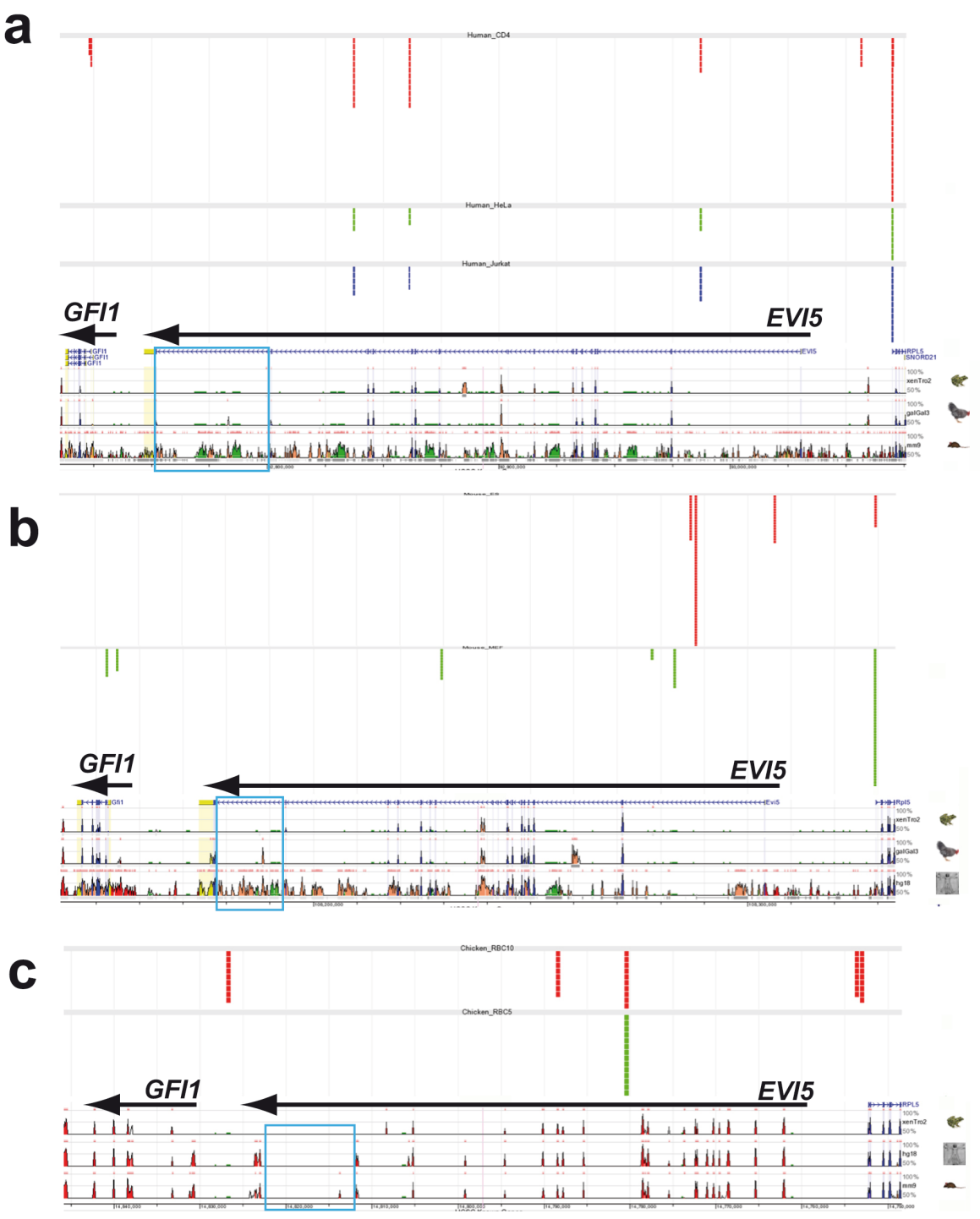
Supplementary Figure 3. Genes separated by CTCF sites have differential expression patterns. Genes separated by CTCF binding events exhibit greater difference in expression levels (ΔExpr) than expected. Genes separated by CTCF (as compared to random expectation) are enriched for pairs with high differences in expression and depleted for genes with low differences in expression, as compared to those with an average difference. This effect is statistically significant for 32 tissues in human (a) and 17 tissues in mouse (b) (corrected p-value < 0.05). Specifically, we computed the distribution of the absolute differences in expression between pairs of genes separated by CTCF binding sites and compared it to the analogous distribution for pairs of genes not separated by CTCF binding sites. We applied the Wilcoxon-Mann-Whytney test to ask whether these two distributions are distinct, and corrected the p-values for multiple testing. For this analysis, we used the 8912 and the 12432 CTCF sites conserved in different cell types in human and mouse, respectively. For ~2,300 pairs of human genes separated by CTCF binding sites for which there is expression data available, we can identify 32 out of 79 tissues displaying differential expression (a); uterus corpus, tonsils and placenta appear among tissues with most significant differential expression. Similarly, ~2,600 pairs of mouse genes separated by CTCF binding sites are more likely to be differentially expressed for 17 out of 61 tissues (b), including epidermis, tongue and bone marrow.

Supplementary Fig. 4



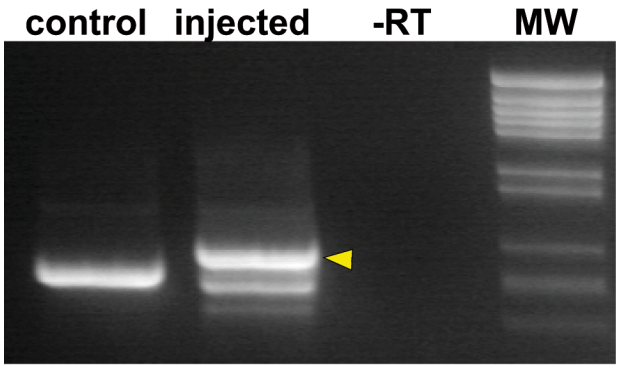
Supplementary Figure 4. Upregulation of *GFI1* expression in the activated PBMCs. Q-PCR showing the strong upregulation of *GFI1* expression in the activated PBMCs.

Supplementary Fig. 5



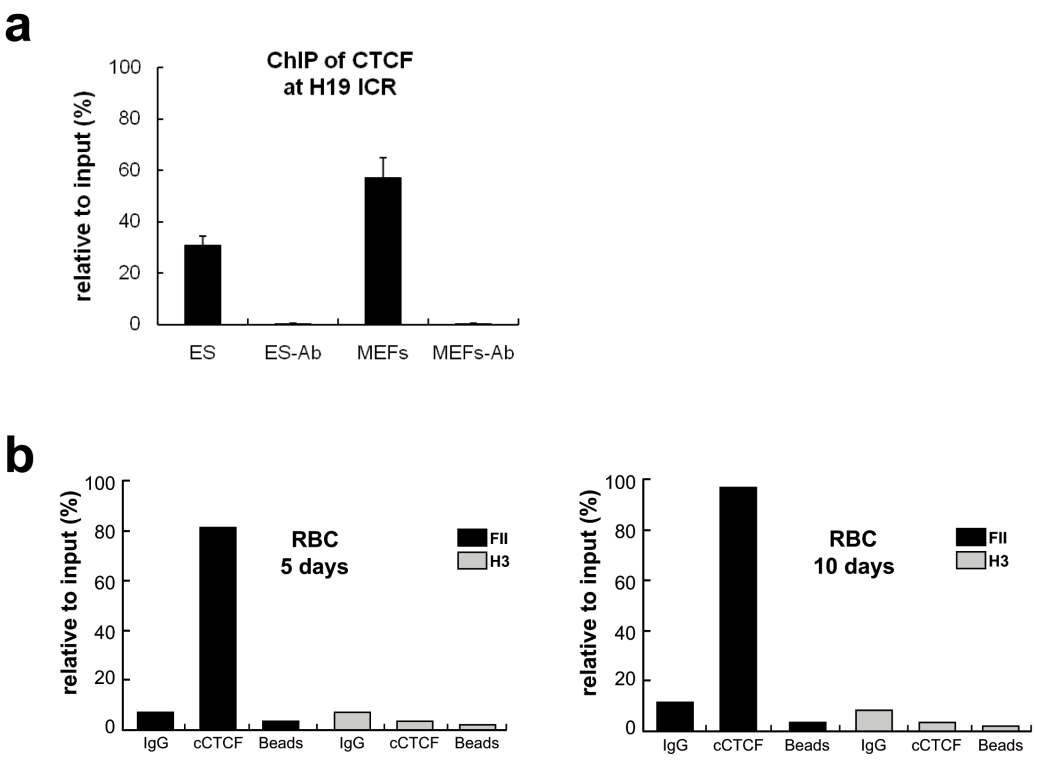
Supplementary Figure 5. CTCF sites separates the last *EVI5* intron from its promoter in all vertebrates. Image from the ECR browser showing the distribution of CTCF bound sites in different cell types in human **(a)**, mouse **(b)** and chicken **(c)**, along the *GFI1-EVI5* genomic regions. Several constitutive CTCF sites (vertical dashed lines) separate the last intron of *EVI5* , marked by blue rectangules, from its promoter in all species.

Supplementary Fig. 6



Supplementary Figure 6. MOsp1CTCF partially inhibits the correct removal of intron 2 of CTCF pre mRNA. The first and second lanes show RT-PCRs with primers from exons 2 and 5 of the zebrafish CTCF gene. Note an extra band of 726 bp in the morphant embryos (yellow arrow), indicating that MOsp1CTCF partially inhibits the correct removal of intron 2. The inclusion of this intron 2 in the mRNA introduces several precocious stops codons that eliminate the key domains of the protein. The third and fourth lanes correspond to a control without RT and the Molecular weigh marker.

Supplementary Fig. 7



Supplementary Figure 7. CTCF-ChIP quality controls in mouse and chicken cell types. (a) In order to evaluate the CTCF-ChIP quality in mouse stem (ES) and fibroblast (MEF) cells, positive PCR controls were performed for the H19 Imprinting control region. **(b)** We verified the CTCF-ChIP from chicken red blood cells (RBC) at 5 and 10 days of development evaluating CTCF enrichment over the Footprint II of the 5´cHS4 insulator of the β-globin gene domain (FII, black bars). We used one region of repetitive sequences structured in heterochromatin and located 16 kb upstream of the 5´cHS4 insulator as a negative control for CTCF binding (H3, grey bars).