nature structural & molecular biology

# Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes

David Martin[1], Cristina Pantoja[2,11], Ana Fernández-Miñán[3,11], Christian Valdes-Quezada[4,11], Eduardo Moltó[5,6,11], Fuencisla Matesanz[7,11], Ozren Bogdanović[3,11], Elisa de la Calle-Mustienes[3], Orlando Domínguez[2], Leila Taher[8], Mayra Furlan-Magaril[4], Antonio Alcina[7], Susana Cañón[9], María Fedetz[7], María A Blasco[2], Paulo S Pereira[10], Ivan Ovcharenko[8], Félix Recillas-Targa[4], Lluís Montoliu[5,6], Miguel Manzanares[9], Roderic Guigó[1], Manuel Serrano[2], Fernando Casares[3] & José Luis Gómez-Skarmeta[3]



**Many genomic alterations associated with human diseases localize in noncoding regulatory elements located far from the promoters they regulate, making it challenging to link noncoding mutations or risk-associated variants with target genes. The range of action of a given set of enhancers is thought to be defined by insulator elements bound by the 11 zinc-finger nuclear factor CCCTC-binding protein (CTCF). Here we analyzed the genomic distribution of CTCF in various human, mouse and chicken cell types, demonstrating the existence of evolutionarily conserved CTCF-bound sites beyond mammals. These sites preferentially flank transcription factor–encoding genes, often associated with human diseases, and function as enhancer blockers *in vivo*, suggesting that they act as evolutionarily invariant gene boundaries. We then applied this concept to predict and functionally demonstrate that the polymorphic variants associated with multiple sclerosis located within the *EVI5* gene impinge on the adjacent gene *GFI1*.**

Although only a small proportion of the genome codes for proteins and regulatory RNAs, *cis*-regulatory elements (CREs), the DNA sequences controlling the expression of the coding segments, are located in the vast noncoding portion of the genome[1]. Genome-wide association studies (GWASs) are increasingly linking human diseases to noncoding DNA that probably has regulatory function (reviewed in refs. 2,3). However, in these cases, the assignment of the candidate disease gene may not be straightforward: a CRE can act over long distances, and its target gene may not be the one closest to the CRE (see, for example, ref. 4). Therefore, methods to predict which gene (or genes) is regulated by particular noncoding genome segments should help identify the candidate disease gene when the lesion lies in noncoding regions.

CTCF contributes to the regulation of gene expression and higher-order organization of the genome[5]. It is evolutionarily conserved and widely distributed along the vertebrate and *Drosophila melanogaster* genomes[6–9]. Although the primary function(s) of CTCF cannot now be directly derived from its genomic distribution, some of the CTCF-bound sites function as regulatory boundaries, confining the range of actions of CREs to genes within those boundaries (reviewed in refs. 5,10). Different cofactors can interact with CTCF, including the SNF2-like chromodomain helicase CHD8 and the DEAD-box RNA helicase p68 (refs. 11,12). CTCF also binds to the cohesin complex at many genomic sites[13–15]. Indeed, at several loci, the cohesin complex seems to regulate this insulator activity[13–15]. Constitutive CTCF-bound sites probably serve this function, whereas more labile sites may be involved in regulating tissue-specific gene expression. In fact, a proportion of CTCF sites are constitutively occupied in several human cell types and are conserved between human and mouse at the sequence level[7,16]. This conservation might extend even further evolutionarily, because the development of the shared body plan of vertebrates is controlled by a shared set of transcription factors and signaling molecules deployed in similar patterns[17]. However, genome-wide CTCF distribution has not yet been examined beyond mammals. If CTCF-bound sites are found at syntenic positions in different vertebrates, these evolutionarily conserved boundaries could be used to

[1]Center for Genomic Regulation, Universitat Pompeu Fabra, Barcelona, Spain. [2]Centro Nacional de Investigaciones Oncológicas, Madrid, Spain. [3]Centro Andaluz de Biología del Desarrollo, Consejo Superior de Investigaciones Científicas (CSIC)–Universidad Pablo de Olavide (UPO), Seville, Spain. [4]Instituto de Fisiología Celular, Departamento de Genética Molecular, Universidad Nacional Autónoma de México, México DF, México. [5]Department of Molecular and Cellular Biology, Centro Nacional de Biotecnología, CSIC, Madrid, Spain. [6]Centro de Investigación Biomédica en Red de Enfermedades Raras, Instituto de Salud Carlos III (ISCIII), Madrid, Spain. [7]Instituto de Parasitología y Biomedicina Lopez-Neyra, CSIC, Granada, Spain. [8]National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health (NIH), Bethesda, Maryland, USA. [9]Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain. [10]Instituto de Biologia Molecular e Celular (IBMC), Universidade do Porto, Oporto 4150-180, Portugal. [11]These authors contributed equally to this work. Correspondence should be addressed to F.C. (fcasfer@upo.es) or J.L.G.-S. (jlgomska@upo.es).

Received 7 October 2010; accepted 15 March 2011; published online 22 May 2011; corrected after print 3 June 2011; doi:10.1038/nsmb.2059

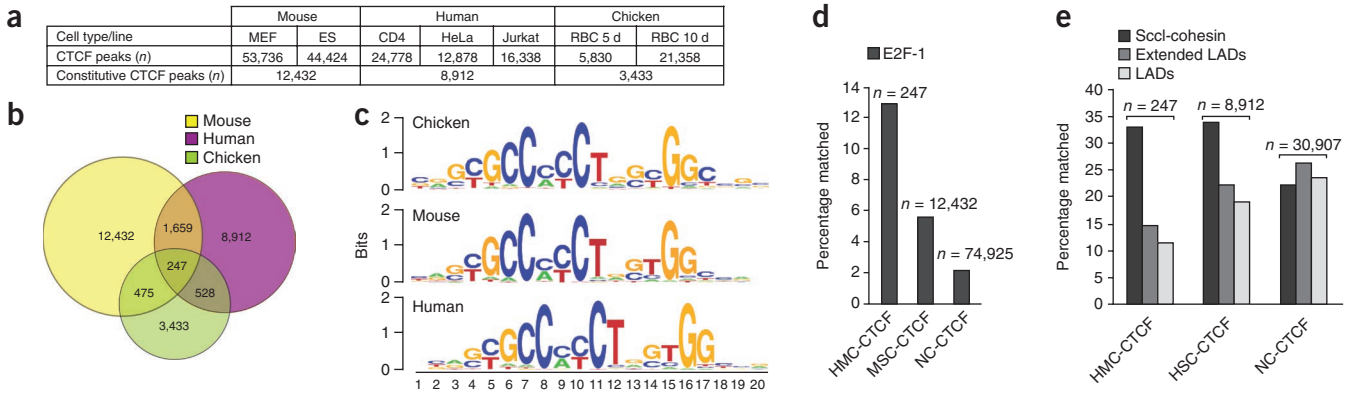


**Figure 1** Detection and conservation of CTCF-binding sites. (**a**) Number of CTCF peaks detected in each cell type or line investigated in each species. Bottom row, number of constitutive peaks within each species. (**b**) Venn diagram of interspecies conservation of CTCF sites. (**c**) Canonical CTCF motifs obtained by *de novo* motif discovery. (**d**) Genomic intersections (overlap ≥ 50%) of mouse CTCF sites with E2F-1-binding sites. CTCF sites were grouped according to their conservation status into HMC, human-mouse-chicken conserved; MSC, mouse genome–specific; NC, nonconserved. (**e**) Genomic intersections of HMC, NC and human-specific (HSC) CTCF sites with LADs and Sccl-cohesin. Extended LADs, LADs extended for 10 kb to each side.



resolve ambiguous associations of target genes affected by mutation in noncoding regions in human diseases, as is the case of multiple sclerosis and the *GFI1* and *EVI5* genes (encoding growth factor–independent 1 and ecotropic viral integration site 5, respectively).

Multiple sclerosis (MIM 126200) is the most common progressive and disabling neurological condition affecting young adults in the world today. The overall prevalence of multiple sclerosis is 2–150 per 100,000 individuals. Pathogenetically, multiple sclerosis is an autoimmune disease leading to the demyelination of central nervous system (CNS) axons[18]. Genetically, it is a complex disorder resulting from a combination of genetic and nongenetic factors[19]. In addition to the human leukocyte antigen (HLA), which is the strongest locus for multiple sclerosis in most populations, other genetic factors involved in multiple sclerosis remained elusive until the arrival of GWASs (MSGene Database, http://www.msgene.org/). So far, seven GWASs have been carried out for multiple sclerosis; although study design and results vary substantially among experiments, some new susceptibility genes have been identified and replicated using this approach[20]. However, even after convincing replications, the localization of the causal variant(s) of most of these loci remains to be determined. Several GWASs have found a set of multiple sclerosis–associated polymorphisms belonging to the same linkage disequilibrium block, which is in a region containing *GFI1*, *EVI5*, *RPL5* and *FAM69A* (encoding ribosomal protein L5 and family with sequence similarity 69, respectively)[21–23]. Fine mapping of this genomic region has pointed to polymorphisms within the 17th intron of the *EVI5* gene as the most probable causal variants of the association[24]. However, these findings do not clarify the functional role of this *EVI5* risk region. Our analysis of the CTCF sites within this genetic block indicates that the 17th intron of the *EVI5* gene probably belongs to the *GFI1*, and not the *EVI5*, regulatory domain. We further demonstrate that this intron contains CREs that contact *GFI1* but not *EVI5*. We finally show that increased expression of *GFI1*, but not of *EVI5*, is associated with the multiple sclerosis risk haplotype. We therefore conclude that *GFI1*, and not *EVI5*, is the causal gene associated with multiple sclerosis.

## RESULTS

### CTCF occupies syntenic positions in vertebrate genomes

An important fraction of CTCF sites in human cells are constitutive: that is, they are occupied by CTCF regardless of the cell type analyzed[7]. We wondered to what extent these constitutive sites are also bound *in vivo* by CTCF in equivalent syntenic positions (that is, surrounded by the same orthologous genes) across vertebrate genomes. To investigate this, we collected available genome-wide CTCF chromatin immunoprecipitation sequencing (ChIP-Seq) data from human cells ($CD4^+$, HeLa and Jurkat cells[6,7]) and produced CTCF ChIP-Seq data for mouse (*Mus musculus*) embryonic stem (ES) cells and embryonic fibroblasts (MEFs), as well as for chicken (*Gallus gallus*) red blood cells isolated from embryos at 5 and 10 d of development. We derived potential CTCF-bound sites from ChIP-Seq using described protocols[6,7,25] (see **Supplementary Fig. 1** for details). Within each species, we identified the set of sites common to all cell lines (species-specific constitutive CTCF sites), considering two sites as common between two cell types if their sequences overlapped by ≥50% of their length (see Online Methods for details). A large fraction of CTCF sites seem to be constitutive for the different cell types within each species (**Fig. 1a**).

We next wanted to identify the set of constitutive CTCF sites that are, in addition, evolutionarily conserved in all investigated species. 74% and 5% of human constitutive CTCF sites lie within sequences conserved in mouse and chicken, respectively, showing at least 50% identity in pairwise alignments. The global nucleotide coverage is 61.5% and 3.9% in mouse and chicken, respectively, as inferred from multiple sequence alignments of 44 mammalian genomes[26]. However, we reasoned that CTCF-bound sites located at equivalent positions in two species (for example, between two paralogous genes) could have equivalent functions (that is, be evolutionarily conserved), even if the sites did not reside in conserved sequences. Therefore, relying exclusively on sequence conservation was not sufficient to identify these evolutionarily equivalent sites occupied by CTCF. We therefore considered two CTCF sites as equivalent evolutionarily if they were syntenic in two species—that is, if they separated the same evolutionarily conserved regions (ECRs[27]), either coding or noncoding, at orthologous genome loci (see Online Methods and **Supplementary Fig. 2**). We found 247 constitutive and syntenic (CONSYN)-CTCF sites among the three investigated genomes (**Fig. 1b**). This is probably an underestimate of the number of CONSYN sites because of the stringency in the definition of vicinity to conserved ECRs we used, as well as the relatively low overall degree of sequence conservation between chicken and mammals. Still, these 247 CONSYN-CTCF sites correspond to 7% of the total constitutive sites in chicken, the species with the lowest number of identified CTCF-bound regions.

**Table 1 Transcription factor–binding motifs overrepresented in the CONSYN-CTCF set as compared to species-specific constitutive sites**

| Rank | TRANSFAC ID | Factor | Average rank | *P* (chicken) | *P* (human) | *P* (mouse) | Domain |
|---|---|---|---|---|---|---|---|
| 1 | M01167 | SAP-1a | 9.3 | $9.62 \times 10^{-5}$ | $4.72 \times 10^{-7}$ | $2.77 \times 10^{-5}$ | ETS |
| 2 | M00938 | E2F-1 | 22.0 | 0.00036 | $1.05 \times 10^{-5}$ | 0.0028 | Forkhead |
| 3 | M01073 | HIC1 | 23.3 | $4.01 \times 10^{-5}$ | 0.00035 | 0.0015 | CH+BTB/POZ |
| 4 | M00800 | AP-2 | 27.0 | 0.00012 | 0.00015 | $3.38 \times 10^{-6}$ | bHSH |
| 5 | M00803 | E2F | 30.0 | 0.00026 | $6.38 \times 10^{-5}$ | 0.0061 | Forkhead |
| 6 | M00469 | AP-2α | 36.3 | 0.0002 | 0.00055 | 0.0045 | bHSH |
| 7 | M00470 | AP-2γ | 36.7 | 0.0003 | 0.00029 | 0.0085 | bHSH |
| 8 | M00341 | GA BP | 43.7 | 0.0028 | $2.42 \times 10^{-5}$ | 0.0075 | ETS |
| 9 | M01165 | Elk-1 | 44.7 | 0.00223 | 0.00147 | 0.0015 | ETS |
| 10 | M00333 | ZF5 | 46.3 | 0.00315 | 0.0013 | 0.00012 | CH+BTB/POZ |
| 11 | M01072 | HIC1 | 50.7 | 0.00253 | 0.0005 | 0.007 | CH+BTB/POZ |

## Genomic features of CONSYN sites

Next, we analyzed sequence features associated with this set of CTCF-binding sites. We first found that the most overrepresented motifs were three highly similar position weight matrixes that matched the previously established CTCF consensus motif (**Fig. 1c**), consistent with the high conservation of the CTCF protein from mammals to birds[28]. Using motif discovery on CONSYN sites (see Online Methods), we identified several additional motifs that are overrepresented in the CONSYN-CTCF set as compared with the species-specific constitutive sites (**Table 1**). This suggests that other nuclear factors may cooperate with CTCF at these sites. Among the top-ranking factors, we found SAP-1a, E2F-1, HIC1 and AP-2. ChIP-Seq data for E2F-1 in mouse[29] confirms that E2F-1 is more frequently found in the vicinity of CONSYN-CTCF sites than near nonconstitutive sites or species-specific constitutive sites (**Fig. 1d**). Using a very stringent set of simulated CONSYN sites as a random control (see **Supplementary Methods**), we found that the association of CONSYN-CTCF sites with E2F-1 sites was statistically significant ($P < 0.001$).

CTCF sites have been proposed to serve four types of functions: (i) enhancer blockage, (ii) barriers that prevent repressive heterochromatin spreading , (iii) three-dimensional genome organization and (iv) transcriptional enhancement[5,10,30–32]. In certain contexts cohesins act as mediators for the enhancer-blocking and/or three-dimensional genome organizing activities of CTCF[15]. Conversely, CTCF flanks lamina-associated domains (LADs), where it has been proposed to have a barrier function, preventing the spread of heterochromatin into transcriptionally active chromosomal domains[33]. To determine whether CONSYN-CTCF sites could be linked preferentially to either of these functions, we examined the correlation of these sites with SccI-cohesin and LAD peaks[33]. We found that CONSYN sites overlapped with cohesin-associated loci, but tended not to overlap with LADs, even when we extended the search to up to 10 kb around each LAD site (**Fig. 1e**). Both the overlap with cohesin-associated loci and the avoidance of LADs are statistically significant when compared with the control set of simulated CONSYN sites ($P < 0.001$ in both cases). These data are consistent with CONSYN-CTCF sites having an enhancer-blocking activity. To test this hypothesis, we assayed the insulator activity on six human CTCF sites, three conserved between human and mouse and three CONSYN sites, in two ways, through luciferase enhancer-blocking assays in human HEK293 cells[34,35] and *in vivo*, using a recently described insulator assay in zebrafish[36] (**Fig. 2**).

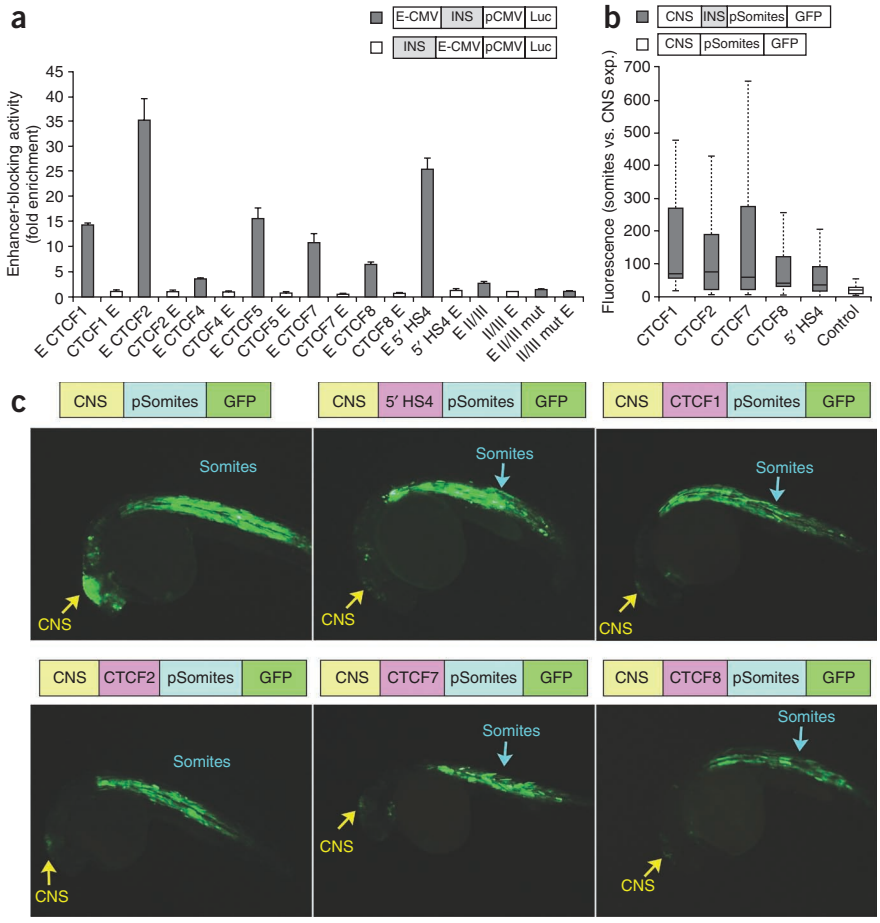


**Figure 2** Functional validation of CTCF sites as insulators. (**a**) Three CTCF sites conserved between human and mouse (CTCF1, CTCF4 and CTCF5) and three CONSYN sites (CTCF2, CTCF7 and CTCF8) were assessed for enhancer-blocking activity through an *in vitro* assay[34]. Enhancer-blocking activity (fold enrichment) normalized to the reference pELuc vector ± s.d. The 5′ HS4 chicken β-globin insulator (5′ HS4) and the internal II/III element were used as positive controls. A mutated II/III element with an altered CTCF site was used as a negative control[35]. (**b**) Enhancer-blocking activity *in vivo* using a transgenic zebrafish assay in which the insulator is placed between a CNS enhancer and a promoter driving GFP expression in somites. Bars, ratios between fluorescence in somites versus CNS. Only the four CTCF sites with significant insulator activity are shown. (**c**) Images from zebrafish embryos after microinjection of each of the CTCF sites in **b**, along with positive (5′ HS4) and negative (empty) controls. Top, construct used.

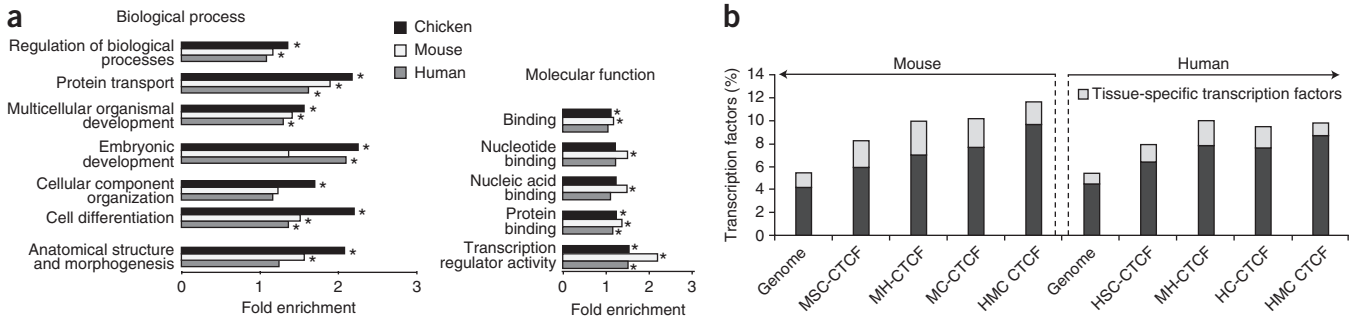


**Figure 3** CONSYN-CTCF sites preferentially flank transcription factors involved in developmental processes. (**a**) Significantly enriched ($P \leq 0.05$) gene ontology terms in genes associated with CONSYN-CTCF sites in chicken, mouse and human. *$P \leq 0.01$. (**b**) Proportion of transcription factor–encoding genes associated with CONSYN CTCF sites in mouse and human genomes. MSC ($n = 12,432$) and HSC ($n = 8912$) are constitutive sites in mouse and human cells, respectively. MH ($n = 1,659$), MC ($n = 475$), HC ($n = 528$) and HMC ($n = 247$) are mouse-human, mouse-chicken, human-chicken and human-mouse-chicken conserved sites, respectively.

All six sites showed consistent enhancer-blocking activity in the *in vitro* assays (more than two-fold; **Fig. 2a**) and four of them, including the three CONSYN sites, reproducibly showed robust enhancer-blocking activity *in vivo* (**Fig. 2b**,**c**).

## CONSYN sites flank developmental and disease genes

All these facts led us to hypothesize that CONSYN-CTCF sites separate genes whose expression should be tightly regulated and whose chromatin should be organized at the genomic level, at least from chicken to humans. To identify these genes, we assigned two neighboring genes in each direction for each CONSYN site (lists are in **Supplementary Table 1**). Using gene ontology term analysis, we identified enrichment in transcription factors involved in cell differentiation and embryonic development (**Fig. 3a** and **Supplementary Tables 2** and **3**). Indeed, although transcription factors constitute ~5% of all genes in mouse and human genomes, ~12% and 10% of murine and human genes adjacent to CONSYN-CTCF sites are transcription factors. The difference is statistically significant compared with a set of random genomic sites ($P < 0.01$, **Fig. 3b**). We analyzed recent expression data for human and mouse[37] and found that transcription factors flanking CONSYN sites are not preferentially associated with tissue-specific expression patterns (**Fig. 3b**). However, we observed that adjacent genes separated by CTCF-binding sites have different patterns of expression as compared with all genes in the genome (**Fig. 4a** and **Supplementary Fig. 3**), suggesting that CTCF-bound sites function as regulatory domain boundaries.

Altered regulation of genes is often associated with human diseases[3,38]. We therefore examined whether the set of genes flanking constitutive CTCF sites is enriched in disease-associated genes. When subjected to MeSH analysis (http://www.nlm.nih.gov/mesh/), the human genes linked to human-mouse conserved CTCF sites and to human-mouse-chicken (CONSYN) sites were significantly associated with disease, including cardiovascular disease, neuroectodermal tumors and lymphomas (**Fig. 4b**).

## CONSYN sites predict *GFI1* association to multiple sclerosis

A considerable number of GWASs indicate that many human diseases are caused by mutations in CREs. However, identifying the target gene of each of these CREs is not trivial, because regulatory elements may be located hundreds of kilobases away from its target promoter, and even inside neighboring genes. Thus, the gene implicated in the development of a particular disease often cannot be directly identified. Because CONSYN-CTCF sites seem to define evolutionarily conserved gene-regulatory boundaries and these boundaries are preferentially linked to genes encoding transcription factors whose malfunction is frequently associated with human diseases, we reasoned that these sites could aid in linking mutations or polymorphisms in CREs associated with human diseases to their target 'disease' genes.

As a proof of principle, we used the *GFI1*–*EVI5* genomic region, which is associated with multiple sclerosis[21]. The most probable causal variants of the association with

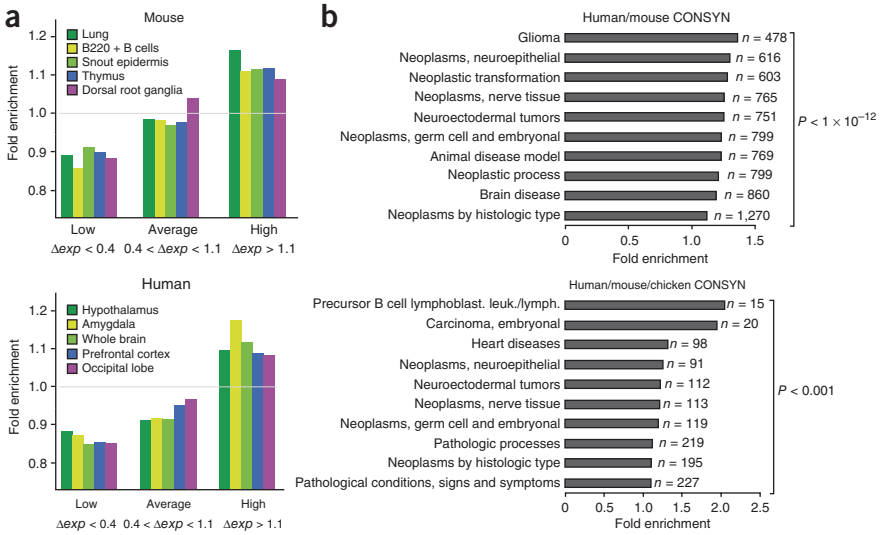


**Figure 4** Genes separated by CTCF sites have differential expression patterns and are associated with human diseases. (**a**) Five tissues showing most significant differences in expression between genes separated by CTCF (Δ*exp*) sites in mouse and human CTCF. (**b**) MeSH analysis of genes flanking CONSYN-CTCF sites show high enrichment in disease-associated genes. The top ten overrepresented MeSH-disease terms in the input gene set (ranked by fold enrichment) are listed with respective number of genes observed for each term and *P* values for each set.

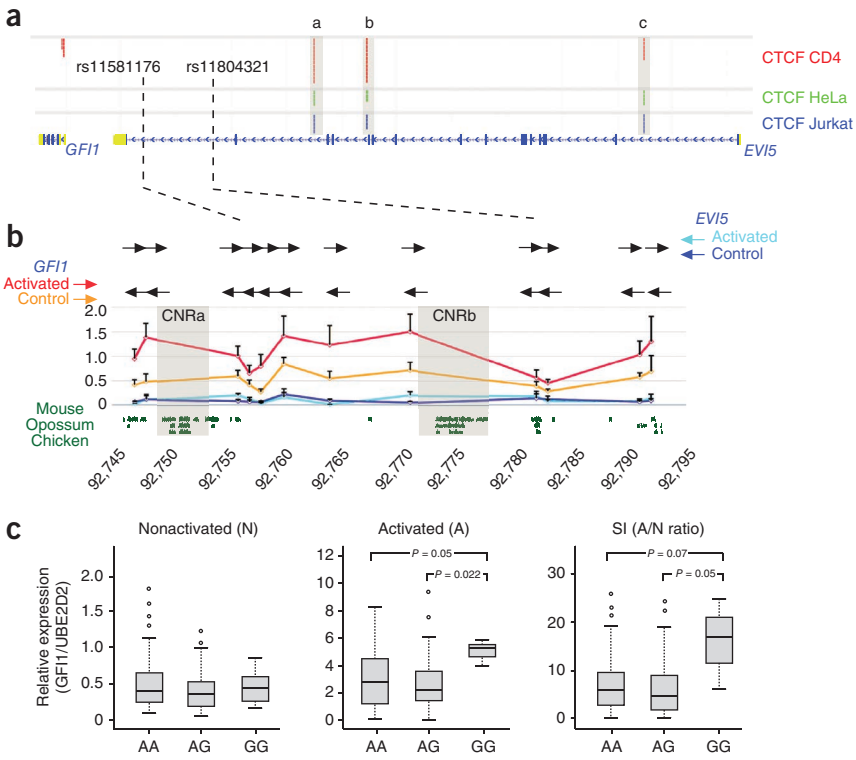


**Figure 5** Constitutive CTCF sites help assign target genes for noncoding mutations. (**a**) Distribution of CTCF-bound sites in human cell types along the *GFI1–EVI5* genomic regions. Constitutive CTCF sites (gray boxes) separate the last intron of *EVI5* (blue rectangle) from its promoter. Two SNPs associated with multiple sclerosis, black dashed vertical lines. (**b**) Close view of 3C results obtained from this intron and the two flanking promoters (coordinates below in kilobases). Primers used along the intron, horizontal black arrows; fixed primers at the promoters of *EVI5* (blue and cyan) and *GFI1* (orange and red), colored horizontal arrows for control and activated PBMCs, respectively. Shadowed boxes, conserved contacting regions tested in functional assays (**Fig. 6**). (**c**) Association of *GFI1* transcription with genotypes of the *EVI5* rs11804321 polymorphism in PBMCs. Gene expression of *GFI1* was assayed by rtPCR normalized to the *UBE2D2* gene. Expression from nonactivated (N) and activated PBMCs (A) and ratio A/N (stimulation index, SI); significant ($P \leq 0.05$) or trend ($P \leq 0.08$) differences are indicated (Mann-Whitney test for 108 independent PBMC samples, genotyped for rs11804321 with 59 AA, 46 AG and 3 GG).

multiple sclerosis are located in the last intron of the *EVI5* gene[24]. Thus, one or several CREs within this intron may be affected in the risk haplotypes. On the basis of this evidence, *EVI5* has been suggested to be the potential target of these CREs[22,23]. However, we examined the human constitutive CTCF-binding sites in the *GFI1–EVI5* genomic locus and found three sites separating the risk genomic area from the *EVI5* promoter (**Fig. 5a**). Strong CTCF-binding sites also separate this last *EVI5* intron from its promoter in mouse and chicken genomes, and in similar positions (**Supplementary Fig. 4**). Although we could not identify these CTCF sites as syntenic sites in our pipeline owing to the stringent criteria imposed, they probably constitute functionally equivalent CTCF sites. The evolutionarily conserved architecture of the *GFI1–EVI5* genomic locus, with CTCF-bound sites separating the last intron of *EVI5* from its promoter in all vertebrates strongly suggests that potential CREs within this intron preferentially act on the neighboring *GFI1* gene, and not on *EVI5*. Multiple sclerosis is a heterogeneous immunopathy probably caused by the joint participation of different peripheral blood cells in the CNS[39]. Notably, malfunction of *GFI1*, which encodes a zinc-finger transcription factor, causes abnormal or malignant hematopoiesis (reviewed in ref. 40), and therefore could have a role in an autoimmune diseases such as multiple sclerosis. To evaluate whether CREs in the last intron of *EVI5* act on either the *EVI5* or *GFI1* promoter, we carried out chromatin conformation capture (3C) assays on control and phorbol myristate acetate plus ionomycin (PMA + Io)-activated human peripheral blood mononuclear cells (PBMCs). In these 3C assays we used several different anchor primers, one on the promoter region of each gene, and multiple primers spanning the whole genomic region of the last *EVI5* intron (**Fig. 5b**). These primers allow the detection of DNA interactions between different regions across the whole intron and either of the two promoters. We normalized PCR product values for each primer pair against those obtained in control samples containing bacterial artificial chromosome clones spanning the tested genomic region (see **Supplementary Methods**). In nonactivated PBMCs, we found no significant interaction between any of the promoters and different regions of the intron (**Fig. 5b**, blue trace). The same was true in activated cells when we surveyed the *EVI5* promoter (**Fig. 5b**, cyan trace). In contrast, the *GFI1* promoter interacted with several regions of the intron, more strongly in the activated than in the control PBMCs (**Fig. 5b**, red and orange traces, respectively).

These results suggest that the *EVI5* intron contains CREs that act on the promoter of *GFI1*, and not on that of *EVI5*. Accordingly, *GFI1* is robustly upregulated in activated PBMCs, whereas *EVI5* is undetectable in both activated and nonactivated blood cells (**Supplementary Fig. 5**). In support of this hypothesis, a probable *GFI1* hematopoietic stem cell–specific enhancer has recently been identified in this genomic area[41].

Notably, we predicted from these data that the expression of *GFI1*, and not that of *EVI5*, would be altered in individuals carrying a risk haplotype. Indeed, we confirmed this: in PBMCs of the risk (G) allele within single nucleotide polymorphism (SNP) rs11804321, *GFI1* expression was greater than in samples carrying the protective (A) allele either heterozygously or homozygously (**Fig. 5c**). In contrast, we found no differences in *EVI5* expression levels (data not shown). This is consistent with a recent finding that *GFI1* expression is also increased in peripheral blood cells of individuals who later develop multiple sclerosis[42], indicating that increased *GFI1* is linked to higher risk of developing the disease. The regions from the *EVI5* intron that interact with the *GFI1* promoter in our 3C studies contain two evolutionarily conserved noncoding sequence blocks (CNR-A and CNR-B; **Fig. 5b**), suggesting a possible regulatory function for these regions. To examine this possibility, we PCR-amplified these two regions and tested their potential enhancer or repressor activities in luciferase assays in THP-1 human acute monocytic leukemia cells. Both regions showed clear repression activity in these assays (**Fig. 6a**). Therefore, our results are compatible with a scenario in which an increased risk for multiple sclerosis is caused by a mutation in any of these two, or even other, repressors located in the last *EVI5* intron, which would then lead to abnormally high expression of *GFI1*.

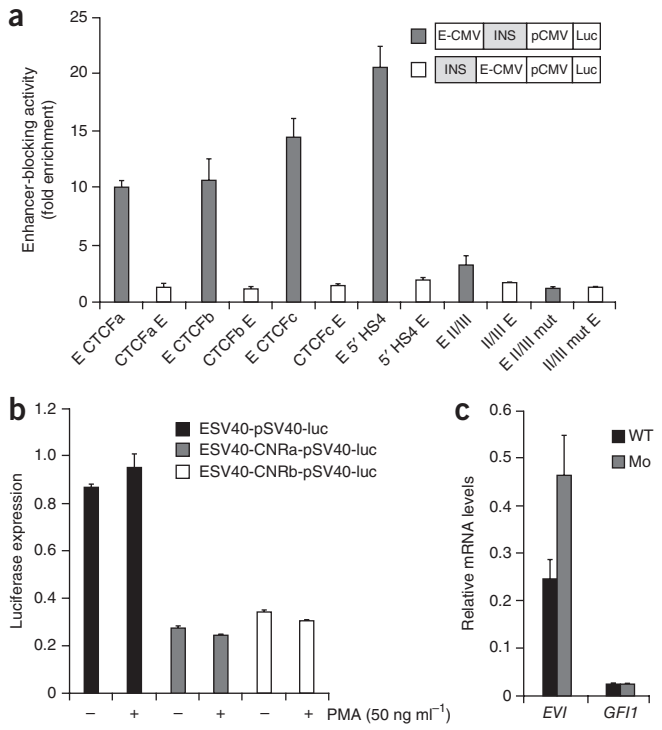


**Figure 6** CTCF sites in the *EVI5* gene act as insulators that prevent the interaction of *GFI1*-associated CREs with the *EVI5* promoter. (**a**) Enhancer-blocking activity assays carried out on three human CTCF-bound sites (a, b and c). (**b**) Luciferase repressor assays performed with CNRA and CNRB in control and activated THP-1 human acute monocytic leukemia cells. (**c**) *evi5* and *gfi1* mRNA levels in wild-type (WT) and CTCF-depleted (Mo) zebrafish embryos 48 h post fertilization, measured by qrtPCR. Values are mean ± s.e.m. of three experiments.

Our earlier prediction of a functional linkage between the risk haplotype and *GFI1* was based on the location of a potential enhancer barrier separating the risk region and the *EVI5* promoter. To test whether any of these three human CTCF-bound sites can function as insulators, we carried out functional enhancer-barrier cell culture assays with all three of them. Much like other CTCF-bound sites we tested, all three regions acted as insulators in these experiments (**Fig. 6b**). These results suggest that these CTCF sites are insulators separating *GFI1* and *EVI5* regulatory landscapes. If so, we expect that reducing CTCF function would affect this boundary, leading to misregulation of either of these two genes. Because organization of *GFI1* and *EVI5* is syntenic in zebrafish, we tested this possibility by knocking down CTCF function with a splicing-specific morpholino (*MOsp1CTCF*, see Online Methods and **Supplementary Fig. 6** for details) in this organism. The *MOsp1CTCF* morpholino partially inhibits the correct removal of intron 2. The inclusion of intron 2 in the mRNA introduces several premature stop codons that eliminate key domains of the CTCF protein (**Supplementary Fig. 6**). We then determined by quantitative real-time PCR (qrtPCR) the expression levels of both *gfi1* and *evi5* genes in control and morphant embryos. In the CTCF morphant embryos the expression of *evi5* is higher than in control individuals, whereas that of *gfi1* does not vary (**Fig. 6c**). These results indicate that reducing CTCF causes *evi5* misregulation, which could be due to inappropriate contact with neighboring regulatory regions. Because the genomic organization of these two genes is conserved throughout vertebrate evolution, we predict that a similar situation may also occur in humans.

## DISCUSSION

A large fraction of CTCF-bound sites in different human and mouse cell types are conserved within species[7,16], defining what we call constitutive CTCF-bound sites. Moreover, several of these CTCF-bound sites lie within sequence stretches conserved between human and mouse genomes, and therefore are evolutionarily conserved in mammals[16]. However, we think that this criterion is too restrictive. CTCF sites may serve similar insulator or enhancer-blocking functions in two species if they are located at equivalent genomic positions (that is, are syntenic), regardless of whether they are located in conserved sequences. Therefore, in this study we extended the set of conserved CTCF-bound sites to include those that are syntenic. Using this approach, we identified at least twice as many likely equivalent CTCF-bound sites in mammalian genomes than by using only sequence conservation, corresponding to 18% of the human constitutive sites. To search for deeply conserved CTCF syntenic sites in a nonmammalian genome, we determined by ChIP-Seq the genome-wide CTCF distribution in two chicken cell types. As in other species, a large fraction of CTCF sites were constitutively occupied in the two different chicken cell types we analyzed (59% of the sites from the cell type with fewer reads). Moreover, 7% of these chicken constitutive sites are located at syntenic positions in mice and humans, but most of them are not conserved at the sequence level. We call these sites CONSYN-CTCF sites. We therefore conclude that using synteny is a more useful way to identify equivalent positions occupied by a transcription factor in different species than using sequence conservation alone.

Notably, our work demonstrates that these deeply evolutionarily conserved CTCF-bound sites show enhancer-blocking activity and flank developmental genes associated with human diseases. Therefore, our work identifies a set of gene boundaries that have remained constant through evolution, at least from chicken to humans. This conservation may stem from the need to avoid regulatory interference within and between these essential genes. Disruption of these genes' boundaries would probably impair development or cause disease. Therefore, we propose that evolutionarily conserved CTCF sites can serve as a useful guide in assigning noncoding mutations to target genes, including those associated with human diseases. Indeed, as a proof of principle, in this study we used this knowledge of conserved gene boundaries to identify a probable target gene affected by haplotypes associated with an increased risk of multiple sclerosis located in the *GFI1–EVI5* genomic region. Although *EVI5* has been considered the target gene involved in this disease[22,23], we demonstrate that the last intron of this gene, which contains the multiple sclerosis risk haplotypes, is separated from its promoter by several syntenic CTCF-bound sites that can function as insulators. Indeed, the presence of these syntenic CTCF-bound sites suggests that the last *EVI5* intron is within the *GFI1* gene regulatory landscape. Therefore, CTCF could potentially prevents the interaction of several *GFI1* regulatory elements present in this *EVI5* intron with *EV15*'s own promoter. Accordingly, *evi5* expression is misregulated in zebrafish embryos with reduced CTCF function. We also found two repressor elements within this intron that are candidate regions for mutation in multiple sclerosis risk haplotypes. Accordingly, as we expected on the basis of malfunction of these repressors, we found that individuals carrying homozygosity in one of the multiple sclerosis risk SNPs have higher expression of *GFI1*, but not *EVI5*, in peripheral blood cells. Finally, in these cell types, and using 3C experiments, we further demonstrated that these repressors physically contact with the *GFI1* promoter, but not that of *EVI5*. Taken together, our results indicate that *GFI1*, but not *EVI5*, is potentially associated with higher risk of developing multiple sclerosis; this prediction was originally based

on the distribution of syntenic CTCF sites in multiple vertebrates. Therefore, the location of these sites might clarify the associations between disease-linked SNPs at noncoding DNA and target genes by defining regulatory domains throughout the genome.

## METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/nsmb/.

**Accession codes.** GEO: chicken CTCF binding maps, GSE27945; mouse CTCF binding maps, GSE27944.

*Note: Supplementary information is available on the Nature Structural & Molecular Biology website.*

### ACKNOWLEDGMENTS

This research was supported by the following grants: BFU2007-60042/BMC, BFU2010-14839, Petri PET2007_0158, CONSOLIDER CSD2007-00008 (Spanish Ministerio de Ciencia e Innovación (MICINN)) and Proyecto de Excelencia CVI-3488 (Junta de Andalucía) to J.L.G.-S.; BFU2009-07044 (MICINN) and Proyecto de Excelencia CVI 2658 (Junta de Andalucía) to F.C.; FIS PI081636 (ISCIII) to F.M.; PN-SAF2009-11491 (MICINN) and Proyecto de Excelencia P07-CVI-02551 (Junta de Andalucía) to A.A.; BFU2008-00838, CONSOLIDER CSD2007-00008 (MICINN), Regional Government of Madrid (CAM S-SAL-0190-2006) and the Pro-CNIC Foundation to M.M.; BFU2006-12185 and BIO2009-12697 (MICINN) to L.M.; Dirección General de Asuntos del Personal Académico, Universidad Nacional Autónoma de México (IN209403, IN214407 and IN203811) and Consejo Nacional de Ciencia y Tecnología, México (CONACyT: 42653-Q, 58767 and 128464) to F.R.-T.; Intramural Research Program of the US NCBI (NIH) to I.O. and BIO2006-03380, CONSOLIDER CSD2007-00050 (MICINN) and RETICS RD07/0067/0012 (Spanish MICINN) to R.G. L.M. thanks A. Fernández for technical assistance and L. Barrios for statistical analysis. F.R.-T. thanks G.G. Avendaño for technical assistance.

### AUTHOR CONTRIBUTIONS

J.L.G.-S. and F.C. conceived the study, designed the experiments, interpreted results and wrote the manuscript. D.M. devised bioinformatics methods, carried out data analysis and wrote the paper. C.P., M.S. and M.A.B. conducted mouse ChIP experiments. C.V.-Q., M.F.-M. and F.R.-T. carried out chicken ChIP experiments. E.C.-M., E.M. and L.M. carried out insulator assays. A.F.M. conducted the 3C experiments. F.M., A.A. and M.F. provided PBMCs from blood cells and carried out qrtPCR, CNRA/CNRB activity assay in a luciferase reporter assay, quantification of *GFI1* relative expression of 108 PBMC samples, genotyping of the *EVI5* rs11804321 and statistical analysis. O.D. carried out the high-throughput sequencing. O.B., L.T., I.O., S.C. and P.S.P. did data analysis. M.M. and R.G. collaborated in the experimental design, discussion of results and in writing the manuscript.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nsmb/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Elgar, G. & Vavouri, T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet.* **24**, 344–352 (2008).
2. Manolio, T.A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363**, 166–176 (2010).
3. Epstein, D.J. *Cis*-regulatory mutations in human disease. *Brief. Funct. Genomics Proteomics* **8**, 310–316 (2009).
4. Ragvin, A. *et al.* Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. *Proc. Natl. Acad. Sci. USA* **107**, 775–780 (2010).
5. Phillips, J.E. & Corces, V.G. CTCF: master weaver of the genome. *Cell* **137**, 1194–1211 (2009).
6. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
7. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* **19**, 24–32 (2009).
8. Bushey, A.M., Ramos, E. & Corces, V.G. Three subclasses of a *Drosophila* insulator show distinct and cell type-specific genomic distributions. *Genes Dev.* **23**, 1338–1350 (2009).
9. Nègre, N. *et al.* A comprehensive map of insulator elements for the *Drosophila* genome. *PLoS Genet.* **6**, e1000814 (2010).
10. Ohlsson, R., Bartkuhn, M. & Renkawitz, R. CTCF shapes chromatin by multiple mechanisms: the impact of 20 years of CTCF research on understanding the workings of chromatin. *Chromosoma* **119**, 351–360 (2010).
11. Ishihara, K., Oshimura, M. & Nakao, M. CTCF-dependent chromatin insulator is linked to epigenetic remodeling. *Mol. Cell* **23**, 733–742 (2006).
12. Yao, H. *et al.* Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. *Genes Dev.* **24**, 2543–2555 (2010).
13. Parelho, V. *et al.* Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422–433 (2008).
14. Rubio, E.D. *et al.* CTCF physically links cohesin to chromatin. *Proc. Natl. Acad. Sci. USA* **105**, 8309–8314 (2008).
15. Wendt, K.S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796–801 (2008).
16. Mikkelsen, T.S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156–169 (2010).
17. Shubin, N., Tabin, C. & Carroll, S. Deep homology and the origins of evolutionary novelty. *Nature* **457**, 818–823 (2009).
18. Oksenberg, J.R., Baranzini, S.E., Sawcer, S. & Hauser, S.L. The genetics of multiple sclerosis: SNPs to pathways to pathogenesis. *Nat. Rev. Genet.* **9**, 516–526 (2008).
19. Handel, A.E., Handunnetthi, L., Giovannoni, G., Ebers, G.C. & Ramagopalan, S.V. Genetic and environmental factors and the distribution of multiple sclerosis in Europe. *Eur. J. Neurol.* **17**, 1210–1214 (2010).
20. Hoffjan, S. & Akkad, D.A. The genetics of multiple sclerosis: an update 2010. *Mol. Cell. Probes* **24**, 237–243 (2010).
21. Hafler, D.A. *et al.* Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* **357**, 851–862 (2007).
22. Hoppenbrouwers, I.A. *et al.* EVI5 is a risk gene for multiple sclerosis. *Genes Immun.* **9**, 334–337 (2008).
23. Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene). Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nat. Genet.* **41**, 824–828 (2009).
24. Alcina, A. *et al.* Tag-SNP analysis of the GFI1–EVI5-RPL5–FAM69 risk locus for multiple sclerosis. *Eur. J. Hum. Genet.* **18**, 827–831 (2010).
25. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.* **36**, 5221–5231 (2008).
26. Rhead, B. *et al.* The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.* **38**, D613–D619 (2010).
27. 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–W286 (2004).
28. Filippova, G.N. *et al.* An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol. Cell. Biol.* **16**, 2802–2813 (1996).
29. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
30. Pikaart, M.J., Recillas-Targa, F. & Felsenfeld, G. Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators. *Genes Dev.* **12**, 2852–2862 (1998).
31. Recillas-Targa, F. *et al.* Position-effect protection and enhancer blocking by the chicken β-globin insulator are separable activities. *Proc. Natl. Acad. Sci. USA* **14**, 6883–6888 (2002).
32. Wallace, J.A. & Felsenfeld, G. We gather together: insulators and genome organization. *Curr. Opin. Genet. Dev.* **17**, 400–407 (2007).
33. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
34. Lunyak, V.V. *et al.* Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* **317**, 248–251 (2007).
35. Recillas-Targa, F., Bell, A.C. & Felsenfeld, G. Positional enhancer-blocking activity of the chicken β-globin insulator in transiently transfected cells. *Proc. Natl. Acad. Sci. USA* **96**, 14354–14359 (1999).
36. Bessa, J. *et al.* Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of *cis*-regulatory regions in zebrafish. *Dev. Dyn.* **238**, 2409–2417 (2009).
37. Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744–752 (2010).
38. Kleinjan, D.A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32 (2005).
39. Hemmer, B., Cepok, S., Zhou, D. & Sommer, N. Multiple sclerosis–a coordinated immune attack across the blood brain barrier. *Curr. Neurovasc. Res.* **1**, 141–150 (2004).
40. Phelan, J.D., Shroyer, N.F., Cook, T., Gebelein, B. & Grimes, H.L. Gfi1-cells and circuits: unraveling transcriptional networks of development and disease. *Curr. Opin. Hematol.* **17**, 300–307 (2010).
41. Wilson, N.K. *et al.* Gfi1 expression is controlled by five distinct regulatory regions spread over 100 kilobases, with Scl/Tal1, Gata2, PU.1, Erg, Meis1, and Runx1 acting as upstream regulators in early hematopoietic cells. *Mol. Cell. Biol.* **30**, 3853–3863 (2010).
42. Achiron, A. *et al.* Microarray analysis identifies altered regulation of nuclear receptor family members in the pre-disease state of multiple sclerosis. *Neurobiol. Dis.* **38**, 201–209 (2010).

## ONLINE METHODS

**Chromatin immunoprecipitation.** Mouse CTCF-ChIPs in ES cells and fibroblasts were carried out as described[43]. For immunoprecipitation we used a rabbit polyclonal antibody to CTCF (07-729, Millipore). To evaluate the CTCF-ChIP quality, positive PCR controls were carried out for the H19 imprinting control region (**Supplementary Fig. 7a**). Chicken red blood cell CTCF-ChIP experiments were carried out as described[44,45]. We evaluated the CTCF-ChIP quality with several positive and negative PCR controls (**Supplementary Fig. 7b**).

**Sequencing.** Sequencing libraries were produced using the Illumina ChIP-Seq sample preparation kit, according to the manufacturer's instructions. Single read sequencing was carried out on the Illumina Genome Analyzer platforms I and II, and images were analyzed using Illumina pipeline versions 1.3.2 and 1.4.

**Short-read mapping and peak calling.** Genomic coordinates of chicken and mouse ChIP-Seq sequence reads were obtained using the GEM mapping software suite (http://gemlibrary.sourceforge.net/). We used the quality files provided by the Illumina Genome Analyzer, and mapped them against the corresponding genome sequence (Galgal3 and mm9 for chicken and mouse, respectively). Using the same program, we determined the corresponding genome fractions to be used with the peak-calling program. For human data, because the quality files were not available, we used the provided mapping (Eland) on the hg18 genome assembly. We filtered out reads that did not map uniquely to the reference genome sequence. Details are provided in **Supplementary Figure 1**.

Peak calling was carried out using SISSRs[25], with the same parameters as described for detection of the human CTCF-binding sites[6,7]. The selected regions were then extended 200 base pairs to each side, centered on the middle coordinate of the peak.

**Evolutionary conservation.** For sequence conservation analysis among species, we retrieved multiple alignment format (MAF) blocks from the UCSC genome browser repository using one of the species as reference coordinate, and examined whether these blocks overlapped a peak in the query species. The retrieval of the blocks was carried out using the "extract MAF blocks" module from the Galaxy suite[46,47].

We developed three methods (**Supplementary Fig. 2**) to assess conservation on the basis of anchoring peaks from one species (reference species) to peaks in another (query) species through sequence features. This allowed us to detect conserved peaks having no sequence conservation. Further details can be found in **Supplementary Methods**.

**Motif analysis.** We carried out two types of analysis, *de novo* motif discovery and known motif overrepresentation. To assess motif overrepresentation, we used Pscan[48] with the Transfac[49] Pro 2009.2 motif collection (892 matrices). Nonvertebrate motifs and low-quality matrices (Q5 and Q6) were removed from the collection. *De novo* motif discovery analysis was carried out with MEME[50] by trying all possible motif widths from 6 to 20 base pairs, asking for five motifs to be found per run and using two different distribution models, one occurrence of the motif per sequence (oops) and zero or one occurrence per sequence (zoops).

**Gene ontology analysis.** Gene ontology term enrichment analyses were computed using the GOToolBox suite[51]. Term overrepresentation was calculated using a hypergeometric-based test. *P* values were corrected for multiple testing using a Benjamini and Hochberg correction. We consider $P \leq 0.01$ highly significant and $P \leq 0.05$ significant. We also computed the enrichment ratio for each over- or underrepresented term, dividing the frequency of this term in our gene set by its frequency in the whole genome. For clarity, we also mapped the over- and underrepresented terms to the generic slim ontology provided by the GO consortium.

**Identification of tissue differential expression of genes separated by CTCF peaks.** First, we used the set of UCSC known genes[52] to define nonoverlapping gene clusters. Second, we associated each microarray probe in the Gene Expression Atlas 2 data provided by the Genomics Institute of the Novartis Research Foundation (GNF[53]) with a given gene cluster. The GNF database (gnfAtlas2) contains two replicates each of 61 mouse tissues and 79 human tissues run over Affymetrix microarrays. The $\log_2$ ratios of the signals (expression score) of all probes associated with a given cluster were averaged. For each tissue, we then computed the absolute difference in the expression scores between pairs of adjacent gene clusters not isolated by CTCF-binding events. This background distribution was subsequently compared to that corresponding to gene clusters isolated by CTCF-binding events using the Wilcoxon-Mann-Whitney test. We corrected the resulting *P* values for multiple testing with Bonferroni's method using the number of tissues.

**Enhancer-blocking assay.** Enhancer-blocking assays were used to address the insulator activity of selected CTCF elements amplified with primers listed in **Supplementary Table 4**, using the pELuc backbone plasmid and human HEK 293 cells as reported[34]. For details, see **Supplementary Methods**.

**Disease association of CONSYN-CTCF sites.** MeSH analysis of genes flanking CONSYN-CTCF sites was performed using the Genomatix Software Suite. Of the set of 2,562 CTCF-bound genes conserved between human and mouse, Genomatix identified 2,273 genes, of which 1,714 had a MeSH annotation. Of the set of 412 CTCF-bound genes conserved between human, mouse and chick, Genomatix identified 360 genes, of which 262 had a MeSH annotation.

**Repressor luciferase assays.** CNRA and CNRB were PCR-amplified, cloned in TOPO T/A vector and transferred using the Gateway system to the destiny vector pGL3 control, which contains the SV40 enhancer and the SV40 minimal promoter. The constructs were transfected to exponentially growing THP1 cells. Triplicate samples of transfected cell cultures were treated or untreated with PMA + Io for 4 h and then harvested. Luciferase activity was evaluated using a Dual-Luciferase assay (Promega). For more details see **Supplementary Methods**.

***In vivo* insulator activity in zebrafish and morpholino injections.** The insulator activity of selected CTCF elements was analyzed *in vivo* by microinjection in one-cell zebrafish embryos as reported[36]. About 10–40 individual zebrafish were analyzed and quantified for each condition. Each set of experimental constructs was injected and analyzed with its corresponding set of controls. LaserPix (Bio-Rad) image analysis software was used for quantification.

*MOsp1* was designed to bind to the acceptor-splicing site between intron 2 and exon 3 (5′-AGCAAATATCACACACTCACCTTTC-3′). A total of 15 ng of *MOsp1* morpholino was injected into one cell-stage embryos. More details can be found in **Supplementary Methods**.

**Chromosome conformation capture assay.** Chromosome conformation capture assay was carried out in control and PMA + Io–activated human PBMCs cells as described[54]. See **Supplementary Methods** for details.


43. Gonzalez, S. *et al.* Oncogenic activity of Cdc6 through repression of the INK4/ARF locus. *Nature* **440**, 702–706 (2006).
44. Escamilla-Del-Arenal, M. & Recillas-Targa, F. GATA-1 modulates the chromatin structure and activity of the chicken α-globin 3′ enhancer. *Mol. Cell. Biol.* **28**, 575–586 (2008).
45. Rincón-Arano, H., Guerrero, G., Valdes-Quezada, C. & Recillas-Targa, F. Chicken α-globin switching depends on autonomous silencing of the embryonic π globin gene by epigenetics mechanisms. *J. Cell. Biochem.* **108**, 675–687 (2009).
46. Blankenberg, D. *et al.* Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* **19**, 19.10.1–19.1.21 (2010).
47. Blankenberg, D *et al.* A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res.* **17**, 960–964 (2007).
48. Zambelli, F., Pesole, G. & Pavesi, G. Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res.* **37**, W247–W252 (2009).
49. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
50. Bailey, T.L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 21–29 (1995).
51. Martin, D. *et al.* GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol.* **5**, R101 (2004).
52. Hsu, F. *et al.* The UCSC known genes. *Bioinformatics* **22**, 1036–1046 (2006).
53. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
54. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).

# Erratum: Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes

**David Martin, Cristina Pantoja, Ana Fernández Miñán, Christian Valdes-Quezada, Eduardo Moltó, Fuencisla Matesanz, Ozren Bogdanović, Elisa de la Calle-Mustienes, Orlando Domínguez, Leila Taher, Mayra Furlan-Magaril, Antonio Alcina, Susana Cañón, María Fedetz, María A Blasco, Paulo S Pereira, Ivan Ovcharenko, Félix Recillas-Targa, Lluís Montoliu, Miguel Manzanares, Roderic Guigó, Manuel Serrano, Fernando Casares, & José Luis Gómez-Skarmeta**

*Nat. Struct. Mol. Biol.* **18, 708–714 (2011); published online 22 May 2011; corrected after print 3 June 2011.**

In the version of this article initially published, the affiliation for authors at the Department of Molecular and Cellular Biology, Centro Nacional de Biotecnología, Madrid, Spain, was incomplete. The full affiliation is "Department of Molecular and Cellular Biology, Centro Nacional de Biotecnología, CSIC, Madrid, Spain." The error has been corrected in the HTML and PDF versions of the article.