Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes

Petra C Schwalie1,5†, Michelle C Ward2,3†, Carolyn E Cain3, Andre J Faure1, Yoav Gilad3, Duncan T Odom2,4* and Paul Flicek1,4*

Background: The genomic binding of CTCF is highly conserved across mammals, but the mechanisms that underlie its stability are poorly understood. One transcription factor known to functionally interact with CTCF in the context of X-chromosome inactivation is the ubiquitously expressed YY1. Because combinatorial transcription factor binding can contribute to the evolutionary stabilization of regulatory regions, we tested whether YY1 and CTCF co-binding could in part account for conservation of CTCF binding.

Results: Combined analysis of CTCF and YY1 binding in lymphoblastoid cell lines from seven primates, as well as in mouse and human livers, reveals extensive genome-wide co-localization specifically at evolutionarily stable CTCF-bound regions. CTCF-YY1 co-bound regions resemble regions bound by YY1 alone, as they enrich for active histone marks, RNA polymerase II and transcription factor binding. Although these highly conserved, transcriptionally active CTCF-YY1 co-bound regions are often promoter-proximal, gene-distal regions show similar molecular features.

Conclusions: Our results reveal that these two ubiquitously expressed, multi-functional zinc-finger proteins collaborate in functionally active regions to stabilize one another's genome-wide binding across primate evolution.

Background
CTCF is a highly conserved, 11-zinc finger multi-functional protein [1,2] important in regulating gene expression [3-5], insulating against enhancer-promoter interactions [6,7], regulating splicing [8], as well as ensuring allele-specific expression at imprinted genes [7] and on the inactive X chromosome [9]. Genome-wide studies have suggested that CTCF binding demarcates active and repressive domains [10-12] and contributes to nucleosome positioning [13], as well as nuclear organization and higher order chromatin structure [14].

CTCF's binding profile is largely (but not entirely [15]) invariant across mouse tissues [16], human cell lines [10] and divergent species compared to those of tissue-specific transcription factors (TFs) [17-22]. Comparisons of CTCF binding have revealed a high level of conservation in liver tissue of species separated by up to 180 million years [21], as well as in cell lines from human, mouse, and chicken [19]. Additionally, CTCF has been shown to bind transposable elements in both embryonic stem cells [18] and differentiated tissue [21]. While certain repeat elements have expanded CTCF target sites in several mammalian lineages, thus far there is no evidence of this process being prevalent in primates based on experiments in human and rhesus macaque [21].

The availability of sequenced primate genomes [23-26] and the ability to transform blood B cells into immortal lymphoblastoid cell lines (LCLs) with the Epstein-Barr virus (EBV) [27] facilitates functional genomics comparisons across different primate species. To date, such inter-primate studies have been carried out primarily at the level of gene expression [28-32]. However, it had already been proposed in the 1970s that phenotypic differences between primates are largely due to regulatory differences [33]. While comparative evolutionary studies in mammals...
have provided insight into regulatory mechanisms, limited information is available within the primate order.

Inter-primate comparisons of regulatory evolution have been performed for histone modifications, which can explain 7% of gene expression differences among human, chimpanzee, and rhesus macaque cell lines [34]. Further, DNA methylation studies revealed that promoter methylation differences underlie 12 to 18% of gene expression differences between humans and chimpanzees and that approximately 10% of CpG islands are significantly differentially methylated between the two species [35,36]. Differences in the binding of transcriptional regulators have been inferred from the presence of several hundred species-specific DNase 1 hypersensitive sites near genes differentially expressed between humans and chimpanzees [37]. Regulatory DNA element comparisons among primates are emerging [38,39]; however, a comprehensive analysis of the binding of a sequence-specific factor such as CTCF across primate species has yet to be performed.

CTCF can exert its different functions through interactions with diverse protein factors [40,41]. One such factor is Yin Yang 1 (YY1), which was originally shown to trans-activate the Tsix ncRNA during X-chromosome inactivation through its interaction with CTCF [9]. There is a strong pattern of co-localization between these two factors at predicted boundary elements, suggesting that they could act synergistically in delimiting chromatin domains [42]. Genome-wide chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) data have recently indicated global co-localization of CTCF and YY1 in human cells [43] with a specific distance constraint [44].

YY1 was first identified as both a repressor and an activator of the adeno-associated virus under different conditions [45], but, similar to CTCF, it has been attributed a broad range of distinct functions, including roles in imprinting [46-48], X-chromosome inactivation [49], and chromatin structure maintenance [50]. YY1 is essential in mouse development, as its deletion results in peri-chromatin structure maintenance [50]. YY1 is essential in mouse development, as its deletion results in peri-chromatin structure maintenance [50]. YY1 binding motifs are overrepresented in core promoters [54], with approximately 10% of human promoters containing it [55]. Additionally, YY1 is important for initiating transcription of various transposable elements such as LINE-1s [56,57], Alu SINEs [58], Herv-Ks [59] and LTRs [60].

Here, we map genome-wide CTCF binding at high resolution in seven primate species and propose that the evolutionary stability of CTCF genomic occupancy is, at least in part, linked to its co-binding with the ubiquitous TF YY1.

Results

Evolution of CTCF binding in seven primates

CTCF binding in distantly related mammalian species is highly conserved compared to that of tissue-specific TFs [17-22]. Here, we analyzed the evolution of CTCF binding at high resolution in LCLs from seven primates, spanning 40 million years of evolution, to determine what mechanism(s) contribute to binding conservation. We experimentally profiled CTCF in most of the great apes (Homo sapiens - H.sap, Pan troglodytes - P.tro, Gorilla gorilla - G.gor, and Pongo pygmaeus - P.pyg), two species of Old World monkey (Macaca mulatta - M.mul and Papio hamadryas - P.ham), and one species of New World monkey (Sanguinus oedipus - S.oed) (Table S1 in Additional file 1). In each species, we performed ChIP-seq in at least two replicates and used naked DNA (input) as control (Figure 1A; Additional file 1: Table S2). Species were aligned to their respective genome except for P.ham, which was aligned to the M.mul genome (84% reads aligned), and S.oed, which was aligned to the Callithrix jacchus (C.jac) genome (65% reads aligned), as there are currently no published genomes available for these species, and M.mul and C.jac represent the closest sequenced relatives. We determined regions of significant ChIP enrichment (referred to as ‘bound regions’ or ‘binding events’) compared to input samples with CCAT 3.0 (Materials and methods) using a fixed false discovery rate (FDR) of 0.05 across species (Materials and methods). Inter-species comparisons were made using Ensembl release 60 genome-wide 6-way EPO primate multiple alignments [61], which include H.sap, P.tro, G.gor, P.pyg, M.mul and C.jac (Table S1 in Additional file 1). Our universal cutoff approach is unbiased in that it does not assume that binding events are conserved across species; however, it favors a model where differences are more likely than shared binding and as such is likely to marginally underestimate the fraction of conserved binding events [21,22] (Materials and methods).

In each of the analyzed primates, we detected thousands of shared and species-specific CTCF-bound locations (Figures 1B and 2A). To determine the extent of binding conservation, we split CTCF-bound regions into six different classes: (i) species-specific and not included in the genome-wide multiple alignments, (ii) species-specific and included in the alignments, (iii) shared with exactly one other species, (iv) shared with two, three, or four other species, (v) shared with exactly five other species (that is, missing in exactly one species), and (vi) shared across all primates (Figure 2A; Figure S2A in Additional file 1). On average, approximately 40% of CTCF-bound regions are shared across six or seven species (highly shared). Conversely, thousands of regions (on average 20% of all CTCF binding events) are species-specific or only shared with a single other species. The vast majority (>80%) of CTCF-
Figure 1 CTCF ChIP-seq in seven primate species. (A) Lymphoblastoid cell lines (LCLs) from seven primate species: Homo sapiens (Hsap), Pan troglodytes (Ptro), Gorilla gorilla (Ggor), Pongo pygmaeus (Ppyg), Macaca mulatta (Mmul), Papio hamadryas (Pham) and Sanguinus Oedipus (Soed) were formaldehyde cross-linked prior to CTCF ChIP-seq experiments. YY1 ChIP-seq was performed in a subset of primates - Hsap, Ptro, Ppyg and Soed. Sequencing reads were aligned to each respective species’ genome and peaks called at a fixed false discovery rate (FDR). Inter-species comparisons were based on the EPO multiple sequence alignments. (B) A primate-shared CTCF binding event at the MRPL39 gene, and a human-specific binding event within the COQ7 gene in LCLs from each primate species [representing the great ape, Old World monkey (OWM) and New World monkey (NWM) clades] as well as Mus musculus (Mmus) liver.
bound regions are shared between at least two species and 11,446 binding events are shared across all analyzed primates (Figure 2A; Figure S2A,B in Additional file 1). Of the binding events common to seven primates, 98% are also bound in human liver (11,256 regions), the majority of which are also bound in mouse liver (6,776 regions; Figure 2B; Figure S2C in Additional file 1). In other words, relatively few CTCF binding events are exclusively found in primates, and not in other mammalian lineages.

The differences in CTCF binding accumulate in line with the evolutionary distance between compared species, as has been observed for more distant mammals [21], and pairwise binding overlap fractions between each primate species and human correlate negatively with evolutionary distance (Pearson’s r = −0.92, P = 0.004; Figure 2C; Figure S2B in Additional file 1). Pairwise conservation estimates are consistent with previously published comparisons of CTCF binding in rhesus macaque and mouse liver [21] as well as human and gorilla LCLs [26]. Similarly, as expected based on prior reports [21,26], highly shared binding events show stronger ChIP enrichment, a better match to the consensus motif and a higher overlap with other cell types/tissues (in this case, liver) than species-specific binding events (Figure S2D-G in Additional file 1).

In order to determine the relationship between inter-individual and inter-species variation, we analyzed
independently derived LCLs from *H.sap* (two LCLs), *P.tro* (three LCLs) and *M.mul* (four LCLs). While over three-quarters of the regions bound by CTCF in all four probed *M.mul* cell lines are also shared with five or six other primates, less than one-quarter of cell line-specific bound regions show such high overlap (Figure 2D). Conversely, binding events shared across seven species are present in the vast majority (over 80%) of the individual LCLs, in contrast to roughly half of CTCF-bound regions unique to one or two species. (Figure S2H in Additional file 1). These results indicate that regions bound by CTCF across individuals are more likely to be evolutionarily conserved than individual-specific bound regions. In sum, CTCF binding that is unstable between

Figure 3 Evolutionarily conserved CTCF-bound regions are highly associated with YY1. (A) Fraction of CTCF binding events in each conservation category associated with YY1 binding in human LCLs [43]. Error bars represent the standard error between at least three cell lines. *P*< 0.05, Wilcoxon rank-sum test. (B) Fraction of CTCF binding events in each conservation category associated with repetitive elements, CpG islands, and transcript annotations across all primate species. Error bars represent the standard error across seven species. *P*< 0.05, Wilcoxon rank-sum test. (C) Conserved CTCF and YY1 binding events at the CD81 and TSSC4 genes in *H.sap*, *P.tro*, *P.pyg*, and *P.ham* LCLs. The CTCF binding event is also present in *M.mus* liver, while the YY1 binding event is not. (D) Fraction of CTCF-only (dark blue) and CTCF-YY1 (light blue) binding events in each species overlapping the corresponding regions in *H.sap*. (E) Fraction of YY1-only (dark brown) and YY1-CTCF (light brown) binding events in each species overlapping the corresponding regions in *H.sap*. 

Schwalie et al. Genome Biology 2013, 14:R148 http://genomebiology.com/2013/14/12/R148
species also tends to be unstable between individuals within a species, and vice versa.

**Evolutionarily conserved CTCF-bound regions are co-bound by YY1**

As combinatorial binding of TFs can stabilize regulatory regions [62], and as CTCF has previously been shown to co-localize and functionally interact with the TF YY1 [42-44,51], we asked whether co-binding with YY1 could help explain high CTCF binding conservation. Indeed, we found that almost half (41%) of the primate-shared regions are also bound by YY1 in human LCLs compared to less than 20% of species-specific CTCF binding events (Figure 3A). CTCF-YY1 co-bound regions enrich for all-primate shared CTCF binding regardless of the proximity to transcription start sites (Figure S3A in Additional file 1). In contrast, less than 15% of the regions bound by tissue-specific TFs such as NFKB and PAX5 are also bound by CTCF, irrespective of evolutionary class (Figure S3B in Additional file 1). Evolutionarily conserved CTCF-bound regions are not specifically enriched for repetitive elements, CpG islands or transcripts (Figure S3B).

In order to establish whether YY1 co-binding stabilizes CTCF binding in evolution, we performed YY1 ChIP-seq experiments in *H. sapiens, P. troglodytes, P. pygmaeus* and *P. hamadryas* LCLs, as well as in primary liver tissue from human and mouse. YY1 binds tens of thousands of locations in all interrogated primates, as well as in mouse liver (Figure S3C in Additional file 1). Almost 10,000 regions bound by YY1 are shared across the four primates included in this analysis (4-way shared), 61% of which are also bound in human liver and 40% of which are shared with mouse liver (Figure S3C in Additional file 1). In comparison, virtually all of the 18,000 4-way shared CTCF LCL binding events are present in human liver, and approximately 50% of these are also bound in mouse liver. In other words, for both CTCF and YY1, about half of the binding events found in multiple primate species are also bound in mouse liver. Overall, pairwise YY1 binding conservation is typically lower than observed for CTCF (Figures S2B and S3D in Additional file 1). For instance, *H. sapiens* and *P. troglodytes* share 48% of YY1 binding events compared to 69% of CTCF binding events. *P. troglodytes* and *P. pygmaeus* share 62% YY1 versus 66% CTCF and *P. pygmaeus* and *P. hamadryas* share 59% versus 71% bound regions. Nevertheless, like CTCF, YY1 binding is more highly conserved than that observed for tissue-specific TFs such as CEBPA and HNF4A [22].

After assessing CTCF and YY1 binding independently, we combined the two datasets to analyze the stability of regions co-bound by CTCF and YY1 (Figure 3C-E). CTCF binding events that co-localize with YY1 (CTCF-YY1) in one species are more likely to be shared with a second species (in this case human), and a similar effect is observed for YY1-bound regions that co-localize with CTCF (Figure 3D,E); reflecting this, we observed stronger sequence conservation of CTCF-YY1 co-bound locations (Figure S3E in Additional file 1). For each pair of species, the regions co-bound by CTCF and YY1 are consistently more evolutionarily stable than those bound by either one of the factors in isolation.

In summary, regions co-bound by CTCF and YY1 show enhanced sequence conservation and are more likely to exist and be bound in a second mammalian species, at both shorter and wider evolutionary distances.

**Binding to repeat elements by either CTCF or YY1 does not explain most species-specific regions**

As both CTCF and YY1 have previously been shown to bind to and expand species-specifically via repetitive elements [18,21,56-60] we analyzed the association between binding events and the repetitive genome. We found that CTCF-YY1 bound regions are less likely to overlap annotated repeats than either CTCF-only or YY1-only regions (P < 0.05; Figure S3F in Additional file 1), which is not unexpected given their high conservation. In order to determine the extent to which repetitive elements contribute to species-specific binding, we further analyzed CTCF and YY1 binding events independently.

First, we identified the annotated repetitive elements in each species bound by CTCF (Figure 4A; Figure S4A and Table S3A in Additional file 1). We detected few species-specific repeat associations, but observed consistent enrichments of older mammalian repeats, such as LTR41 and LTR50, as well as overrepresentation of CTCF binding to primate-specific repeats, such as LTR13 [63,64] (Figure 4B). Most were members of the LTR repeat family, consistent with previous evidence of functional exaptation of LTRs in primates (Figure S4A in Additional file 1) [65].

We also searched for repeat-specific CTCF motif words, which previously revealed CTCF-bound repeat expansions in more diverse mammalian species [21]. We detected no species-specific motif words and only a limited number of words bound at a higher frequency in primates than in other mammalian species (Figure S4B in Additional file 1). LTR13 was again identified as a CTCF-bound primate-specific repeat; however, less than 300 binding events account for this enrichment across primates, a comparatively low number considering the tens of thousands of B3-specific motif words that have shaped the CTCF binding landscape in rodent genomes [21].

We similarly identified which annotated repetitive elements are associated with YY1 binding. MLT1-type repeats of the ERVL-MaLR family, including MLT1F and MLT1J, are found to be significantly associated with YY1 binding (Figure 4C,D; Figure S4D and Table S3B in Additional file 1). MLT1-type repeats are not enriched in YY1 binding events in human and mouse livers, suggesting that the genomic interaction of YY1 and this repeat
class might be tissue-specific, unlike the observed LTR repeat association with CTCF binding events (Figure 4E,F). Finally, we did not find an enrichment of repeat-embedded YY1 binding at or in close proximity to repeat-associated CTCF binding locations, indicating that these factors bind distinct repeats. Thus, repeats do not appear to be involved in CTCF-YY1 co-binding in the genome.

In sum, active repeat expansions have not substantially contributed to the CTCF binding repertoire in seven major primate lineages, and most species-specific CTCF and YY1 binding events do not appear to be mediated by repeat elements.

YY1 couples CTCF binding to transcriptional activity
In primate LCLs, on average approximately one-third of CTCF-bound regions are co-bound by YY1, and nearly half of YY1-bound regions are co-bound by CTCF (Figure 5A; Figure S5A in Additional file 1). We asked whether any molecular or sequence features (aside from repeat elements) could differentiate isolated CTCF binding events from those co-bound by YY1. Overall, the binding intensities of CTCF at regions co-bound by YY1 are no greater than those of isolated CTCF binding events, suggesting that the observed pattern is not simply driven by ChIP enrichment class (\(P > 0.05\); Figure S5B in Additional file 1). De novo motif discovery identified the canonical motifs for both CTCF and YY1 at shared CTCF-YY1 bound locations, indicating that both factors directly bind to DNA in general (Figure S5C in Additional file 1). However, we did not observe a consistent spacing constraint between the two motifs at co-bound regions (data not shown). Importantly, we found CTCF-YY1 co-bound...
regions to be significantly more associated with CpG islands ($P < 0.05$) and CpG island promoters ($P < 0.05$) than isolated CTCF binding events, indicating that co-bound regions may be more transcriptionally active (Figure 5B,C; Figure S5D in Additional file 1).

To further explore the relationship between transcription and conservation of CTCF-YY1 co-bound regions, we analyzed the ChIP-seq data from mouse and human liver with corresponding functional data for basal transcriptional machinery, tissue-specific TFs, and histone marks [22,66]. We found that CTCF-YY1 co-bound regions overlap marks of transcriptional activity, including RNA polymerase II (RNA Pol II), the active H3K4me3 histone modification, as well as liver-specific transcriptional regulators such as HNF4A and CEBPA (Figure 6A,B; Figure S6A,B in Additional file 1). In contrast, CTCF-bound regions lacking YY1 (CTCF-only) rarely co-localize with marks of active transcription and tissue-specific TF binding. CTCF-YY1 co-bound regions in liver tend to be associated with core liver functions such as lipid metabolism and transport in both human and mouse (Figure S6C in Additional file 1).

In order to determine whether the presence of YY1 at CTCF-bound regions has an effect on transcriptional output, the expression of genes bound by only CTCF versus those bound by CTCF and YY1 were compared. Genes overlapping YY1 binding events (including YY1-only and YY1-CTCF) are significantly more highly expressed than are genes overlapping CTCF-only binding events ($P < 10^{-16}$) in both human and mouse liver (Figure 6C; Figure S6D in Additional file 1).

In sum, CTCF-YY1 co-bound regions are functionally similar to YY1-only regions based on their association with increased gene expression and enrichment of Pol II, H3K4me3 and liver-specific transcriptional regulators. This means that the CTCF occupied regions co-bound by YY1 show not only stronger evolutionary stability, but also increased transcriptional activity, in contrast to the bulk of CTCF-bound regions.

**Discussion**

Our mapping and inter-species comparison of CTCF binding in cell lines from *H.sap*, *P.tro*, *P.pyg*, and *P.ham* LCLs has revealed over 11,000 genomic locations bound by CTCF across primates, consistent with the high conservation of CTCF binding observed in more distant mammalian species [19,21,67]. This estimate was obtained despite the fact that our analytical approach did not assume that CTCF binding is conserved, and as such is likely to underestimate true inter-species overlaps [21,22]. In contrast, other related studies have assumed conservation and minimized species-specific differences by using a dual cutoff, anchored in a single reference species [34,68].

Despite our conservative approach, we found that 60% of CTCF-bound regions are shared between *H.sap* and *M.mul*, whereas a recent comparison across 25 million years of *Drosophila* sp. evolution, a similar divergence time [69], has revealed approximately 30% binding conservation [70]. This discrepancy could be caused by differences in CTCF function between the chordate and arthropod phyla: CTCF is the only well-characterized insulator-binding protein in vertebrates, where it seems to be largely responsible for three-dimensional genome
organization, as well as regulating various transcriptional and gene regulatory processes in collaboration with cohesin [71]. In contrast, Drosophila sp. have multiple insulator proteins and thus may place lower constraint on CTCF binding sites. Previous studies have shown that the expansion of repetitive elements appears to be a major mechanism by which CTCF increases its target landscape in individual mammalian lineages [18,21,72]. Here, we identified particular repeat types associated with CTCF binding across primates, some of which have previously been shown to associate with CTCF in human, including LTR41 in human embryonic stem cells [18] and LTR13 across multiple cell types [18,63]. However, we did not find systematic evidence for a repeat-mediated expansion of CTCF binding in primate clades. This comparative quiescence of repeats that carry CTCF binding sequences in primates relative to rodents might be in part due to differences in genome transposon content and activity. For instance, mice have more lineage-specific repeat elements than humans, as well as greater transposon activity and fewer ancestral repeats [73,74].

Conclusions
In searching for factors contributing to the conservation of CTCF binding we discovered that co-binding by YY1 is an ancient regulatory mechanism that appears to increase the evolutionary stability of CTCF binding in multiple mammalian species. CTCF and YY1 have previously been shown to co-localize and physically interact [9,42,43,46,75] but their combined, genome-wide interaction has not been investigated across multiple species. Our analysis revealed that one mechanism stabilizing the protein-DNA contacts in regions co-bound by CTCF and YY1 may be their association with active chromatin and gene expression (previously reported for CTCF in [10,76-78]). It has been shown that CTCF can interact directly with Pol II and target it to a subset of CTCF sites genome-wide [79]. Our data now reveal that this likely occurs in the presence of YY1, because in its absence, CTCF-bound regions almost never co-localize with Pol II, H3K4me3, or tissue-specific TFs, whether proximal or distal to genes. These observations suggest that with respect to transcriptional activity, YY1 is the functionally dominant factor at co-bound locations.
Experiments performed at a single genetic locus have shown that CTCF can form a complex with YY1 and the tissue-specific factor Oct4 that binds Tsix and Xite to control X-chromosome pairing and counting in embryonic stem cells [80]. Co-binding of CTCF and YY1 thus appears to indicate globally to the chromatin remodeling machinery which euchromatic regions are to be activated [81]. The integration of these discoveries suggests a model wherein co-transcriptional activity of YY1-bound regions may help conserve CTCF and YY1 binding via functional deployment.

In summary, CTCF-YY1 co-bound regions are not only preferentially and highly conserved but also show hallmarks of transcriptional function that could provide selective pressure to preserve specific protein-DNA contacts across millions of years of mammalian evolution.

Materials and methods

Cell line material

Lymphoblast cell lines were obtained for seven primate species. The species, cell line and source are shown in Table S1 in Additional file 1. All cell lines were transformed by the Epstein-Barr virus except for the M.mul cell lines, which were transformed by Herpesvirus papio.

Cells were grown in suspension at a confluency of 200,000 cells/ml to 1 × 10⁸ cells/ml in RPMI1640 media supplemented with 15% fetal bovine serum and 2 mM L-glutamine, 100 I.U/ml penicillin and 100 μg/ml streptomycin. We cross-linked 1 × 10⁸ cells with 1% formaldehyde as previously described [82].

Tissue material

Mouse

C57BL/6 J mice were housed in the Biological Resources Unit under UK Home Office licensing. Tissue was obtained from at least two independent males and formaldehyde cross-linked as described in [82].

Human

Male and female human tissue samples were obtained from biopsied tissue collected at Addenbrooke’s Hospital, Cambridge, and provided by the Biobank under human tissue license 08/H0308/117. Liver tissue was also obtained from the Liver Tissue Distribution Program (NIDDK contract number N01-DK-9-2310) at the University of Pittsburgh.

ChIP-seq

ChIP-seq assays were performed as previously described [82]. Protein-bound DNA was immunoprecipitated with antibodies against CTCF (Millipore, Billerica, MA, USA, 07–729), or YY1 (Santa Cruz Biotechnology, Dallas, TX, USA, sc-281). Immunoprecipitated DNA was end-repaired, A-tailed and ligated to single-end Illumina sequencing adapters before 18 cycles of PCR amplification. DNA fragments (200 to 300 bp) were selected and 36 bp reads sequenced on an Illumina Genome Analyser II according to the manufacturer’s instructions.

Published ChIP-seq experiments

The following published ChIP-seq data were used: mouse liver CTCF [21], mouse liver H3K4me3 and Pol II [83], human H3K4me3 and Pol II [66], human and mouse liver HNF4A and CEBPA [22]. ENCODE data used were from the following cell lines: for CTCF, GM12878, GM12891, GM12892, GM19239, GM19240, HepG2, H1-hESC; for YY1, GM12878, GM12891, GM12892, HepG2, H1-hESC; for NFKB, GM12878 (no treatment), GM12891 (tumor necrosis factor alpha treatment); for Pax5, GM12878, GM12891, GM12892; and for Pol II, GM12878 [43]. Published mouse embryonic stem cell ChIP-seq data for CTCF and YY1 were also used [19,84]).

Computational methods

All computational analyses were performed with scripts written in Perl, Bioperl 1.2.3, and R version 2.11.1, using packages available in Bioconductor 2.6 (Additional file 2). Displayed error bars represent the standard error of the mean and significance levels were estimated using one-sided Wilcoxon rank-sum tests if not otherwise stated.

Read alignment and peak-calling

ChIP and input sequencing reads from all LCL datasets were aligned using Bowtie [85] 0.12.7 with the parameters ‘-n 2 -m 3 -k 1 -best’ to the following genome assemblies: human GRCh37, chimpanzee CHIMP2.1, gorilla gorGor3, orangutan PPYG2, macaque Mmul 1, and marmoset C. jacchus 3.2.1. All sequence, genome annotations (genes, transcripts, CpG islands) and comparative genomics data were taken from Ensembl release 60. Repeat element annotation was downloaded from the UCSC Table Browser for all species. The baboon data were aligned to the macaque genome, as it was the closest fully assembled genome. When available (all species except for marmoset), only chromosomes and not unmapped contigs were used. Aligned reads were filtered for duplicates, uncalled bases (a maximum of three Ns were allowed) and low complexity reads. Regions of high ChIP enrichment (peaks/bound regions/binding events) were detected with CCAT 3.0 [86] on individual replicates using the parameters ‘fragmentSize 100, slidingWinSize 150, movingStep 10, isStrandSensitiveMode 1, minCount 10, minScore 4.0, bootstrapPass 50’ for the two TFs and ‘fragmentSize 200, slidingWinSize 100, movingStep 20, isStrandSensitiveMode 0, minCount 10, minScore 4.0, bootstrapPass 50’ for Pol II. Naked DNA (input) was used as control and the FDR cutoff was set.
three different LCLs for chimpanzee and four different LCLs for rhesus macaque. The median value is displayed in Figure 2C based on primate (blue square) and mammalian (grey circle) alignments. Evolutionary time between the species was obtained from [88] (median).

For the comparative analyses displayed in Figures 2 and 3 and Figures S2 and S3 in Additional file 1 we divided the bound regions into six different categories in each of the seven primate species (we refer to these categories as ‘conservation classes’): (i) species-specific and not included in the genome-wide multiple alignments, (ii) species-specific and included in the alignments, (iii) shared between two species only, (iv) shared among three to five species, (v) shared among six species, and (vi) shared among all seven analyzed primates. We calculated the relative fraction of bound regions belonging to these six categories and displayed them as barplots in Figure S2A in Additional file 1. The median values across all seven species are shown in Figure 2A as well as the number of seven-way shared peaks (based on the human genome).

For the sequence conservation analysis of CTCF-only and CTCF-YY1 regions in Figure S3E in Additional file 1, the Phastcons tool [89] in Galaxy Cistrome [90] was used.

Properties of different peak categories and CTCF-YY1 binding event classes

Four peak categories (species-specific (1), shared between exactly two species (2), shared among exactly six species (6) and shared among all seven species (7)) were further analyzed for diverse properties in each single species: CCAT score (proportional to ChIP enrichment), the top NestedMica motif match score distribution (with 0 corresponding to the consensus motif), the numbers of peaks with at least one motif, overlaps with peaks called in distinct LCL cell lines when available (human, two; chimpanzee, three; macaque, four), overlaps with transcripts, CpG islands, repetitive elements, Pol II, and publicly available TF binding data from ENCODE. Barplot widths are proportional to the number of regions belonging to each category. We also performed a detailed conservation-inter-individual overlap analysis using four distinct rhesus macaque LCL lines. We selected two types of CTCF-bound regions: (1) bound in only one of the four cell lines and (2) bound in all four cell lines. We then asked how often these regions were shared with the other species and displayed the relative fractions as pie charts in Figure 2D.

Three distinct classes of CTCF/YY1 bound regions (CTCF-only, CTCF-YY1 and YY1-only regions) were analyzed for their properties, using data from four primate LCLs (human, chimpanzee, orangutan, and baboon), as well as human and mouse liver data. Additionally, previously published Pol II, H3K4me3, CEBPA, and HNF4A
data in human and mouse liver were intersected with the three classes of bound regions.

**Repeat element association**

We tested genome-wide association of annotated repeat elements with LCL CTCF/YY1 binding events in each single species by using a binomial test. We estimated background probabilities from median overlaps of repeat elements with randomized CTCF/YY1 binding events, and corrected for multiple testing by the Benjamini-Hochberg method. Repeats that obtained a \( P \)-value \( \leq 0.01 \) are included in Figure S4A,D in Additional file 1; repeats with a \(-\log_{10} P\)-value >10 in at least one species are displayed in Figure 4 and Table S3 in Additional file 1.

We estimated the repeat divergence from the consensus sequence based on the number of substitutions from the consensus ('milliDiv' column in the UCSC-obtained RepeatMasker tracks) and the age of individual bound repeat elements by dividing the substitution number by the mutation rate estimated for mammalian species (2.2 \( \times \) 10\(^9\) per base pair per year) [91] and rodents (4.5 \( \times \) 10\(^9\) per base pair per year) [74]. Repeat ages were used to order the heatmaps shown in Figure 4. Repeats were sorted by class in Figure S4 in Additional file 1 [92].

Functional association analysis

CTCF regions co-bound with YY1 were analyzed relative to all CTCF-bound regions in Figure S6C in Additional file 1 to determine whether these regions were associated with common biological pathways using cPath within the GREAT bioinformatic tool [96,97].

**Motif analysis**

Motif discovery was conducted with NestedMica [94] using the parameters `-minLength 5 -maxLength 30 -numMotifs 6` and a fourth order background model trained on mammalian regulatory regions (DHS) data. Discovered motifs were confirmed using MEME [95], with the options `-ntmotifs 5 -minsites 100 -minw 6 -maxw 25 -revcomp -maxsize 500000 -dna`. We selected the top 1,000 peaks ordered by CCAT score and used 25 bp upstream and downstream of the peak summit as input for motif discovery. As the obtained top motifs were virtually identical in all studied species, we merged them into a single PWM that we used in further motif analysis steps. NestedMica's nmscan with a cutoff of \(-15\) was used for motif matching (a score of 0 corresponds to a perfect match to the motif consensus) displayed in Figure S5C in Additional file 1 are obtained from all sequences that match the CTCF and YY1 PWMs inside regions positive for both CTCF and YY1 ChIP signal.
display the data as boxplots in Figure 6C and Figure S6D in Additional file 1.

Data access
CTCF and YY1 ChIP-seq data have been deposited under Arrayexpress, accession number E-MTAB-1511.

Additional files

Additional file 1: Supplementary Figures S1 to S6 and Tables S1 to S3. Figure S1. CTCF ChIP-seq read correlations. Figure S2. properties of conserved and species-specific CTCF binding events. Figure S3. properties of CTCF and YY1 binding events; Figure S4. association of CTCF and YY1 binding events with repeats; Figure S5. characterization of CTCF-YY1 binding events. Figure S6. association of CTCF-YY1 binding events with marks of active transcription. Table S1. cell line sources. Table S2. ChIP-seq library summary. Table S3. CTCF and YY1 binding event repeat associations.

Additional file 2: Scripts in Perl and R used for computational analyses.

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

Authors’ contributions
PCS, MCW, DTO, and PF wrote the manuscript with input from all authors. YG, DTO, and PF oversaw the work. All authors read and approved the final manuscript.

Acknowledgements
We are grateful to the Cambridge Institute Genomics and Bioinformatics. Core facilities and Margus Lukk and Tim F. Rayner for computational support. Thank you to the Health Protection Agency Cell culture collections and the ECACC for culturing and cross-linking the macaque cell line. PCS, MCW, DTO, and PF wrote the manuscript with input from all authors. YG, DTO, and PF oversaw the work. All authors read and approved the final manuscript.

Author details
1European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
2University of Cambridge, Cancer Research UK-Cambridge Institute, Robinson Way, Cambridge CB2 0RE, UK.
3Current address: Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA.
4Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK.
5Current address: Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne CH-1015, Switzerland.

Received: 13 November 2013 Accepted: 31 December 2013 Published: 31 December 2013

References
1. Kenova EM, Nicolas RH, Paterson HF, Carne AF, Heath CM, Goodwin GH, Neiman PE, Lobanenkov VV: CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. Mol Cell Biol 1993, 13:7612–7624.
2. Moon H, Filippova G, Loukinov D, Pugacheva E, Chen Q, Smith ST, Munhall A, Greve B, Barkhurn M, Arnold R, Burke LJ, Renkawitz-Pohl R, Ohlsson R, Zhou J, Renkawitz R, Lobanenkov V. CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator. EMBO Rep 2005, 6:165–170.
3. Banasimad A, Steiner C, Kohne AC, Renkawitz R. Modular structure of a chicken lysozyme silencer: involvement of an unusual thyroid hormone receptor binding site. Cell 1990, 61:505–514.
4. Lobanenkov W, Nicolas RH, Adler W, Paterson H, Kenova EM, Polotskaya AV, Goodwin GH: A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene 1990, 5:1743–1753.
5. Vostrov AA, Quitschke WW. The zinc finger protein CTCF binds to the AP2beta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. J Biol Chem 1997, 272:33333–33339.
6. Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell 1999, 98:387–396.
7. Hark AT, Schoeninger CJ, Katz DI, Ingram RS, Levone JM, Tilghman SM. CTCF mediates methylation-sensitive-enhancer-blocking activity at the H19/igf2 locus. Nature 2000, 405:486–489.
8. Shukla S, Kavak E, Gregory M, Muthusami M, Shutinoski B, Kashlev M, Oberdoeffler P, Sandberg R, Oberdoeffler S. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature 2011, 479:74–79.
9. Donohoe ME, Zhang LF, Xu N, Shi Y, Lee JT. Identification of a CTCF cofactor, Yy1, for the X chromosome binary switch. Mol Cell 2007, 25:43–56.
10. Chen H, Tian Y, Shu W, Bo X, Wang S. Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. PLoS One 2012, 7:e41374.
11. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res 2009, 19:24–32.
12. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov W, Ren B. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. PLoS Genet 2008, 4:e1000138.
13. Handoko L, Xu H, Li G, Nangan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F, Wong E, Sheng J, Zhang Y, Poh T, Chan CS, Kunaro G, Shahab A, Bourque G, Cacheux-Rataboul V, Sung WK, Ruan Y, Wei CL. CTCF-mediated functional chromatin interaction in pluripotent cells. Nat Genet 2011, 43:630–638.
14. Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, Thurman RE, Kaul R, Myers RM, Stamatoyannopoulos JA. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res 2012, 22:1680–1688.
15. Shen Y, Yue F, McClleary DF, Ye Z, Edwall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov W, Ren B. A map of the cis-regulatory sequences in the mouse genome. Nature 2012, 488:116–120.
16. Borneman AR, Ganoulis TA, Zhang ZD, Yu H, Rozovskiy J, Seringhaus MR, Wang LY, Gerstein M, Snyder M. Divergence of transcription factor binding sites across related yeast species. Science 2007, 317:815–819.
17. Kunaro G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan TS, Ng HH, Bourque G. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet 2010, 42:631–634.
18. Martin D, Pantoja C, Fernandez Minan A, Valdes-Quesada C, Molto E, Mateusar F, Bogdanovic O, de la Calle-Mustienes E, Dominguez O, Taher L, Furlan-Magaril M, Alcina A, Canin S, Fedele M, Blasco MA, Pereira PS, Ovcharenko I, Recillas-Targa F, Montoliu L, Manzanares M, Guigo R, Serano M, Casares F, Gomez-Skarmeta JL. Genome-wide CTCF distribution in
vertebrates defines equivalent sites that aid the identification of disease-associated genes. Nat Struct Mol Biol 2011, 18:708–714.

20. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danфорд TW, Маскадс КД, Rolfe PA, Conboy CM, Gifford DK, Frankei Е. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet 2007, 39:70–75.

21. Schadt EE, Schwaie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 2012, 148:335–348.

22. Schmidt D, Wilson MD, Ballester B, Schwaie PC, Brown GD, Marshall A, Kutter C, Watts S, Martinez-Jimeines CM, Mackay S, Tallianids I, Flicek P, Odom DT. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 2010, 328:1036–1040.

23. CSAC: Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 2005, 437:69–87.

24. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, CSAC: The chimpanzee genome sequence reveals high levels of gene and sequence evolution. Nature 2004, 430:691–713.

25. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Elements of the human genome. Nature 2005, 437:691–713.

26. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, et al: The Drosophila euchromatic state. Nat Rev Genet 2012, 13:662–676.

27. Xi H, Yu Y, Fu Y, Foley J, Halees A, Weng Z, et al: Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. Genome Res 2007, 16:78–806.

28. Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ Jr: Functional characterization of innate immune signaling pathways in primates. PLoS Genet 2010, 6:e1001249.

29. Becker KG, Swergold GD, Ozato K, Thayer RE: Analysis of a pol gene in human LINE-1 transcription factor results in peri-implantation lethality. Mol Cell Biol 2005, 26:245–252.

30. Loo YY, Yu Y, Fu Y, Foley J, Hakes A, Weng Z: Using reporter gene assays to identify cis regulatory differences between human and chimpanzees. Genetics 2007, 176:2069–2076.

31. Cuenca-Rodriguez O, Abyazov A, Malakulvich I, Zhu Q, Sukumar T, Pai AA, Mills RE, Habegger L, Lusinovich DA, Rube MA, Perry GH, Gerstein M, Stone AC, Gilad Y, Lee C: Regulatory element copy number differences shape primate expression profiles. PLoS Nat Acad Sci U S A 2012, 10:12656–12661.

32. Weth O, Renkawitz R: CTCF function is modulated by neighboring DNA binding factors. Biochem Cell Biol 2011, 89:459–468.

33. Zlatanov J, Jiafa P: CTCF and its protein partners: divide and rule? J Cell Sci 2009, 122:1275–1284.

34. Wang J, Lunya W, Jordan IK: Genome-wide prediction and analysis of human chromatin boundary elements. Nucleic Acids Res 2011, 40:511–529.

35. Dunham I, Kundaje A, Aldred SF, Collins PJ, Epstein CB, Dunham I, et al: An integrated encyclopedia of DNA elements in the human genome. Nature 2012, 489:57–74.

36. Guo Y, Mahony S, Gifford DK: High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Comput Biol 2012, 8:e1002638.

37. Shi Y, Seto E, Chang LS, Shenk T: Transcriptional repression by YY1, a human GLI-Kruppel-related protein, and relief of repression by adenosine EI4A protein. Cell 1991, 67:377–388.

38. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

39. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

40. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

41. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

42. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

43. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

44. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

45. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

46. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

47. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

48. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

49. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

50. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

51. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

52. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

53. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

54. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

55. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

56. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.

57. Schadt EE, Schwaie PC, Tang YA, Ward MC, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, et al: Gene expression differences between chimpanzees and humans. Nature 2007, 449:228–232.
