Cohesin regulates tissue-specific expression by stabilizing highly occupied *cis*-regulatory modules

Andre J. Faure,[1,7] Dominic Schmidt,[2,3,7] Stephen Watt,[2,3] Petra C. Schwalie,[1] Michael D. Wilson,[2,3,8] Huiling Xu,[4,5] Robert G. Ramsay,[4,5] Duncan T. Odom,[2,3,6] and Paul Flicek[1,6,9]

[1]*European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom;* [2]*Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge CB2 0RE, United Kingdom;* [3]*Department of Oncology, Hutchison/MRC Research Centre, Cambridge CB1 9RN, United Kingdom;* [4]*Differentiation and Transcription Laboratory, Cancer Research Division, Peter MacCallum Cancer Centre, East Melbourne, Victoria 3002, Australia;* [5]*Sir Peter MacCallum Department of Oncology and Department of Pathology, The University of Melbourne, Parkville, Victoria 3000, Australia;* [6]*Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom*

The cohesin protein complex contributes to transcriptional regulation in a CTCF-independent manner by colocalizing with master regulators at tissue-specific loci. The regulation of transcription involves the concerted action of multiple transcription factors (TFs) and cohesin's role in this context of combinatorial TF binding remains unexplored. To investigate cohesin-non-CTCF (CNC) binding events in vivo we mapped cohesin and CTCF, as well as a collection of tissue-specific and ubiquitous transcriptional regulators using ChIP-seq in primary mouse liver. We observe a positive correlation between the number of distinct TFs bound and the presence of CNC sites. In contrast to regions of the genome where cohesin and CTCF colocalize, CNC sites coincide with the binding of master regulators and enhancer-markers and are significantly associated with liver-specific expressed genes. We also show that cohesin presence partially explains the commonly observed discrepancy between TF motif score and ChIP signal. Evidence from these statistical analyses in wild-type cells, and comparisons to maps of TF binding in *Rad21*-cohesin haploinsufficient mouse liver, suggests that cohesin helps to stabilize large protein–DNA complexes. Finally, we observe that the presence of mirrored CTCF binding events at promoters and their nearby cohesin-bound enhancers is associated with elevated expression levels.

[Supplemental material is available for this article.]

The evolutionarily conserved cohesin protein complex plays an essential role in chromosome cohesion during mitosis and meiosis (Peters et al. 2008). The core of the complex is a heterodimer of structural maintenance of chromosome (SMC) subunits (SMC1A and SMC3) connected by a third subunit RAD21 (MCD1/SCC1 in budding yeast), forming an unusual tripartite ring-like structure (Anderson et al. 2002; Haering et al. 2002). RAD21 is bound to a fourth member (either STAG1, STAG2, or STAG3) and it has been proposed that the complex mediates cohesion by embracing sister chromatids (Nasmyth and Haering 2009). Several other proteins are associated with cohesin, including NIPBL (Nipped-B in fly, Scc2 in budding yeast), which is required for loading of cohesin onto chromatin (Rollins et al. 2004).

Although essential for sister chromatid cohesion, Nipped-B was first identified in *Drosophila melanogaster* as a result of its function in gene regulation, where it was suggested to facilitate enhancer–promoter interactions (Rollins et al. 1999). Similarly, mutations in core components of the cohesin complex can affect gene expression, and have been linked to developmental defects in a number of different species (Donze et al. 1999; Bénard et al. 2004; Krantz et al. 2004; Vega et al. 2005; Horsfield et al. 2007; Zhang et al. 2007; Pauli et al. 2008). Beyond its presence on sister chromatids during cell division, cohesin is also expressed in post-mitotic cells and is loaded onto unreplicated chromosomes in telophase (Sumara et al. 2000; Zhang et al. 2007; Wendt et al. 2008). Together these findings point toward an important non-canonical role of cohesin in regulating gene expression.

More recently, genome-wide maps of cohesin binding in mammalian cells reveal that the complex functionally associates with a large proportion of CTCF sites. Both repressor and activator functions have been attributed to CTCF, but its role as an enhancer-blocking insulator is the most extensively studied, and cohesin plays a role in this function (Parelho et al. 2008; Rubio et al. 2008; Stedman et al. 2008; Wendt et al. 2008). Results from chromatin conformation capture (3C) experiments in the well-characterized *H19/Igf2* imprinting control region (ICR) indicate that CTCF regulates allele-specific expression of the *H19* and *Igf2* genes by controlling intrachromosomal looping interactions (Murrell et al. 2004; Kurukuti et al. 2006; Yoon et al. 2007; Engel et al. 2008). Direct evidence from cohesin knockdown experiments implicates the complex in facilitating long-range interactions between CTCF sites at these loci, as well as at others including the *IFNG*, apolipoprotein, and hemoglobin, beta genes (Hadjur et al. 2009; Mishiro et al. 2009; Nativio et al. 2009; Hou et al. 2010).

There is increasing evidence to suggest that changes in higher-order genome structure and subnuclear chromatin localization are crucial for lineage specification and temporal/tissue-specific transcriptional regulation (Misteli 2007). In view of CTCF's

[7]These authors contributed equally to this work.
[8]Present address: SickKids Research Institute and Department of Molecular Genetics, University of Toronto M5G 1X8, Canada.
[9]Corresponding author
E-mail flicek@ebi.ac.uk

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.136507.111. Freely available online through the *Genome Research* Open Access option.

involvement in mediating chromatin loops at specific developmentally regulated genomic loci, it has been suggested that the primary role of CTCF may be the genome-wide organization of chromatin architecture (Phillips and Corces 2009). Globally, this hypothesis is supported by observations that CTCF-binding sites demarcate the borders of regions localized to the nuclear periphery (Guelen et al. 2008) and that they tend to correlate with intra- and interchromosomal interactions in the human genome as measured by Hi-C (Lieberman-Aiden et al. 2009; Botta et al. 2010). However, given the similarities in CTCF occupancy across different cell types (Cuddapah et al. 2009), it is not clear how CTCF alone could configure the three-dimensional structure of the genome in a dynamic way. Considering that cohesin and CTCF colocalize genome wide, it is an attractive hypothesis that cohesin contributes to this global organizational function. Indeed, mutations causing human cohesinopathies such as Cornelia de Lange Syndrome severely disrupt the subnuclear organization of chromatin and cause aberrant nucleolar morphology when induced in budding yeast (Gard et al. 2009). Furthermore, studies at specific loci show that cohesin dynamically controls the spatial conformation of chromatin required for normal development and differentiation, in a cell-division independent way (Hadjur et al. 2009; Seitan et al. 2011).

Using ChIP-seq experiments in MCF-7 and HepG2 human cancer cells, we have recently shown that cohesin binds to thousands of sites in a CTCF-independent manner. In stark contrast to relatively invariant CTCF sites, these CNC-binding events differ dramatically between cell types. CNC sites colocalize with tissue-specific transcription factors (TFs), such as estrogen receptor alpha (ER) in MCF-7 cells, and contribute to global gene expression. Cohesin is also highly enriched at ER-bound regions that participate in interchromosomal looping interactions as assayed by ChIA-PET (Fullwood et al. 2009; Schmidt et al. 2010a). A study in mouse embryonic stem (ES) cells, which highlighted subunits of both the cohesin and mediator complexes as key contributors to ES cell state, found analogous patterns of CNC binding and co-occupancy with pluripotency regulators, such as POU5F1 (also known as OCT4), at interacting promoter and enhancer regions (Kagey et al. 2010). Finally, results from 3C experiments show that cohesin is required for similar promoter–enhancer interactions within the T-cell receptor alpha locus (Seitan et al. 2011), and taken together with previous findings, firmly establish a role for the complex in widespread mediation of long-range transcriptional regulation.

To investigate in vivo patterns of cohesin binding in depth—particularly independent of CTCF—we mapped both factors together with a collection of 10 TFs using chromatin immunoprecipitation experiments followed by high-throughput sequencing in primary mouse liver. We collected additional data from the same tissue for several histone modifications and other functional DNA–protein interactions, providing a comprehensive map of cohesin's role in tissue-specific transcriptional regulation. We show that this role is likely to be functionally similar across multiple tissues by demonstrating that cohesin's presence at binding events of liver-specific TFs parallels its localization with ES cell-specific factors. We observe a positive correlation between the number of distinct TFs bound and cohesin presence, where most multiply bound *cis*-regulatory modules are CNC sites. In contrast to sites with CTCF, CNC sites tend to coincide with the binding of master regulators, including HNF4A, enhancer-markers such as EP300 (also known as p300), and are significantly associated with genes expressed in a liver-specific fashion. We also show that cohesin presence at least partially explains the commonly observed discrepancy between TF motif score and ChIP signal, suggesting a role for cohesin in stabilizing large protein–DNA complexes by enabling TFs to bind sequences less similar to the canonical binding site motif. Indeed, compared with wild-type mouse liver cells, ChIP signals in *Rad21*-cohesin haploinufficient cells are preferentially diminished at binding events without high-scoring motifs. Finally, we identify cases where the presence of a cohesin-bound enhancer/CTCF pair is mirrored by the presence of CTCF near the putative target transcription start site (TSS) and observe differences in gene-expression levels that are associated with these consistent binding patterns.

Results

We performed ChIP-seq experiments in primary mouse liver with antibodies targeted to CTCF, three cohesin subunits (RAD21, STAG1, STAG2), 10 TFs (CEBPA, HNF4A, FOXA1, FOXA2, ONECUT1, HNF1A, PKNOX1, REST, GABPA, E2F4), two coactivators (EP300, CREBBP), five histone-modifications (H3K4me1, H3K4me3, H3K36me3, H3K79me2, H2AK5ac), and RNA polymerase II (RNAP2). See Methods for full experimental details.

The TFs for our analysis were chosen to include both ubiquitously expressed factors and liver-specific regulators, two of which have well-characterized evolutionary dynamics (Schmidt et al. 2010b). We additionally profiled chromatin marks associated with active TSSs, enhancers, and transcribed genes, providing a comprehensive picture of the genome function and the transcriptional regulatory network active in mouse liver cells. Figure 1A displays a number of key regulatory features of the data in the vicinity of the predominantly liver-specific phosphoenolpyruvate carboxykinase 1 (*Pck1*) gene on mouse chromosome 2, including two clusters of TFs: one immediately proximal to the TSS and another ~25 kb upstream of the TSS. Cohesin can be seen colocalizing with CTCF as well as with clusters of TFs. These data are quantitatively and qualitatively comparable to other multifactor experiments in other tissues (Chen et al. 2008).

After short read alignment with Burrows-Wheeler Alignment tool (BWA) (Li and Durbin 2009) and peak-calling with SWEmbl (Wilder et al., in prep.; see Methods), we determined the overlap between sites bound by CTCF and the cohesin subunits. As expected, the three assayed cohesin subunits show highly similar patterns of binding with peaks of the RAD21 subunit coinciding with 99% and 94% of STAG1 and STAG2 peaks, respectively (Schmidt et al. 2010a). By defining cohesin presence as the occurrence of at least one of its subunits, we find that cohesin colocalizes at the majority of CTCF sites (48,487; 87%), but is also present at a similar number of sites independently of CTCF (Fig. 1B). We define this latter set of 46,471 cohesin-binding sites as cohesin-non-CTCF (CNC) sites (Supplemental Table S1).

CTCF-independent cohesin binding is associated with master regulators and enhancers

To determine the binding partners of cohesin at both cohesin-CTCF and CNC sites, we defined a set of putative *cis*-regulatory modules (CRMs) by grouping together CTCF and cohesin binding events with overlapping binding events of the 10 TFs and the coactivators EP300 and CREBBP (see Methods). The resulting CRMs have a median width of 449 bp, but vary in size depending on the number of factors present (SD = 346 bp; see Supplemental Table S3 for all peak width statistics). The comparatively broad regions of

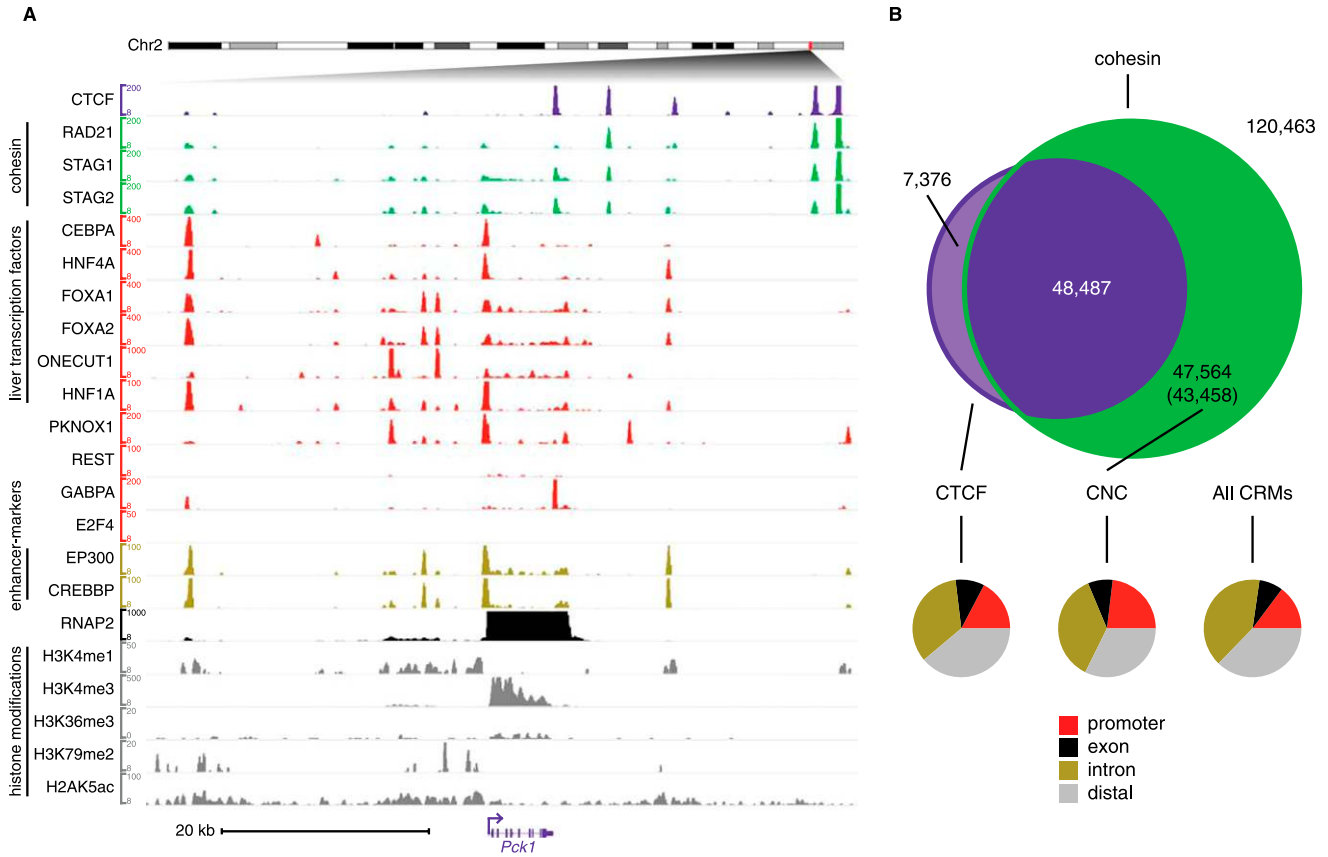


Figure 1. (*A*) Genome-wide occupancy of cohesin, CTCF, tissue-specific and ubiquitous TFs in primary mouse liver as measured by ChIP-seq and shown near the *Pck1* gene. Cohesin colocalizes with CTCF as well as with clusters of transcription factors in the absence of CTCF, one of which can be seen overlapping the TSS of the *Pck1* gene. (*B*) Venn diagram showing CTCF and cohesin (RAD21, STAG1, STAG2) occurrence within CRMs. The pie charts indicate genomic locations of all CRMs (background), as well as those containing CTCF and CNC. The latter occur within promoter regions at a higher relative frequency compared with the other two classes.

the genome associated with the histone modifications (H3K4me1, H3K4me3, H3K36me3, H3K79me2, H2AK5ac) and RNAP2 were not used to define the CRMs themselves. Instead, they were used to annotate the chromatin state of the CRMs post hoc (see Methods).

Of the 223,890 CRMs that were identified, 43,458 (19.4%) are identified as CNC sites. These CRMs mostly occur away from annotated TSSs (77%; Fig. 1B) and tend to coincide with the binding of master regulators, such as HNF4A (69%), and the enhancer markers EP300 and/or CREBBP (66%). The fraction of promoter–proximal CNCs (23%) is nevertheless higher than that of CTCF (17%), and CNC-containing CRMs are significantly enriched for occurrence near TSSs when compared with all CRMs (Fisher's exact test $P < 10^{-15}$; Fig. 1B). CNC sites that occur within promoter regions (≤2.5 kb from the annotated TSS), are highly enriched for RNAP2 binding compared with cohesin-bound promoters in general (Fisher's exact test $P < 10^{-15}$). These results are similar to those in *Drosophila*, where cohesin lacks a functional interaction with CTCF (Bartkuhn et al. 2009), but is preferentially detected at promoters of active genes (Misulovin et al. 2007). Here cohesin selectively binds genes with paused RNAP2 and lacking H3K36me3, a mark associated with transcriptional elongation (Fay et al. 2011). Although we find that cohesin is associated with increased RNAP2 pausing indices in mouse liver cells, cohesin-bound promoters are also associated with elevated expression levels and an enrichment of H3K36me3 within the gene body (Supplemental Fig. S1).

At cohesin sites containing CTCF, we observe a shift in the summit positions of all cohesin subunits with respect to the CTCF summit position when the orientation of the CTCF motif is taken into account (Supplemental Fig. S2). This result is similar to recent reports for RAD21 (Nitzsche et al. 2011) and supports a direct and directional biochemical interaction between cohesin and CTCF. The same directional analysis at CNC sites, however, reveals that the position of cohesin is independent of the peak position and motif orientation of all other sequence-specific factors considered (Supplemental Fig. S2). This demonstrates a specific cohesin–CTCF interaction that is not seen at CNC sites and suggests a different mechanism of cohesin recruitment in the absence of CTCF.

To identify the primary interacting partner proteins within the CRMs, we used the within-CRM ChIP fragment count (i.e., the number of mapped ChIP reads, extended to the estimated fragment length, overlapping the CRM) to measure the binding strength correlation between all ChIP-seq data sets. These correlations highlighted two separate modes of cohesin binding. First, a clear and distinct cluster includes all three cohesin subunits and CTCF (Fig. 2A, purple cluster). Second, cohesin subunits also correlate with tissue-specific factors including FOXA1, HNF4A, and HNF1A (green cluster); TFs are not correlated with CTCF. The cohesin/TF cluster is also marked by the active histone modifications H3K4me3, H3K4me1, and H2AK5ac, as well as with the

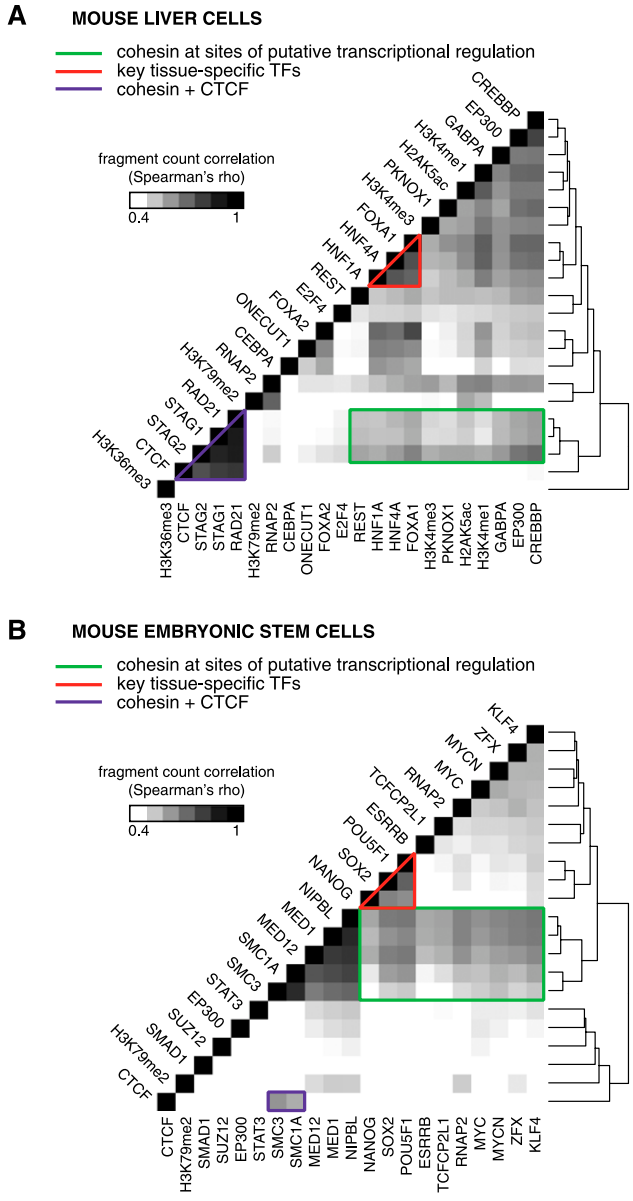


Figure 2. Within-CRM binding correlations reveal distinct modes of cohesin binding in diverse cell types. The number of ChIP fragments (mapped reads extended to the estimated fragment length) overlapping a given CRM was used as a measure of binding strength for each data set. Factors were clustered along both axes based on the similarity in their colocalization profiles. (*A*) Heatmap visualization of all pairwise correlations between all ChIP-seq data sets in mouse liver cells illustrates cohesin subunits (RAD21, STAG1, STAG2) clustered with CTCF. Cohesin also correlates with key tissue-specific TFs (FOXA1, HNF4A, and HNF1A) independently of CTCF as well as with histone modifications associated with transcriptional activity (H3K4me1, H3K4me3, H2AK5ac) and coactivators (EP300 and CREBBP). (*B*) All pairwise correlations between previously published ChIP-seq data sets in mouse embryonic stem cells. Cohesin binding strength (SMC1A, SMC3) correlates with CTCF while also forming a distinct cluster with key regulators of stem cell identity (POU5F1, SOX2, NANOG, MYC), components of the mediator complex, as well as RNAP2. Similar results were obtained by performing the correlation analysis separately on CRMs with CNC and CTCF (see Supplemental Fig. S3).

coactivators EP300 and CREBBP, suggesting that cohesin within CNC sites may play a central role in active transcriptional regulation together with a wide range of TFs.

To investigate whether the correlations between CTCF, cohesin, and tissue-specific TFs are particular to liver or differentiated tissue, we performed the same analysis for a set of previously published ChIP-seq data sets from mouse embryonic stem (ES) cells (Chen et al. 2008; Marson et al. 2008; Seila et a. 2008; Kagey et al. 2010) (see Methods). Although the ES cell ChIP-seq data set contains a different collection of TFs and cohesin subunits, they show patterns highly similar to those observed in primary liver tissue (Fig. 2). Indeed, the SMC1A and SMC3 cohesin subunits correlate with CTCF (Fig. 2B, purple cluster) while also forming a separate, distinct cluster with key regulators of stem cell identity (POU5F1, SOX2, and NANOG), components of the mediator complex, and RNAP2 (Fig. 2B, green cluster). The cohesin loading factor is absent from the SMC1A/SMC3/CTCF cluster, which is consistent with previous observations of NIPBL's preferential association with CNC sites and supports the idea of a different mechanism of cohesin recruitment in the absence of CTCF (Kagey et al. 2010). Overall, cohesin shows two separate modes of binding that have minimal overlap in two transcriptionally divergent and phenotypically distinct mouse tissues: (1) either with CTCF and showing minimal signs of transcriptional activity, or (2) with clusters of tissue-specific TFs showing hallmarks of transcriptional activation.

CNC sites occur preferentially at multiply bound *cis*-regulatory modules (CRMs)

To understand the genomic properties of the identified CRMs, we grouped them into similar clusters based either on the normalized ChIP enrichment or binary presence/absence of the sequence-specific factors using two different clustering methods (K-means and *AutoClass*) (see Methods). A primary difference between these two clustering methods is that K-means requires the number of clusters to be defined a priori, whereas *AutoClass* uses a Bayesian probabilistic approach to automatically optimize the properties of each cluster (as well as the number of clusters) to achieve the best separation. Because the overall clustering results were similar between the two methods, we focused our analysis on the results from K-means (with K = 10) for ease of interpretation (Fig. 3) (see Supplemental Fig. S4 for *AutoClass* results; Supplemental Fig. S5 for a justification of the choice of K).

The 10 clusters, totaling 210,067 CRMs, are visualized in Figure 3, sorted from left to right by the fraction of CNC-containing CRMs in a given cluster. CRMs with CTCF form a large, distinct cluster at the extreme left (cluster 10; 41,368 CRMs). Most CRMs without CTCF fall into three large groups (clusters 7–9; 102,091 CRMs) with an average of less than two sequence-specific factors (singleton CRMs). The remaining six clusters (66,608 CRMs) have increasing numbers of colocalizing TFs, with almost all possessing either HNF4A, FOXA1, or FOXA2 (99%) and nearly half (48%) possessing all three of these factors. Furthermore, these six clusters all show a distinct pattern of chromatin state. For example, compared with singleton CRMs, clusters 1–3 are more strongly enriched for RNAP2, EP300/CREBBP, and H3K4me1 ($P < 10^{-15}$), indicating that these clusters are likely to contain active enhancers.

Ranking the CRM clusters by the proportion of CNC sites in each cluster, we observe a strong positive correlation between the average number of distinct TFs present and CNC presence (Spearman's $\rho = 0.95$, $P < 10^{-15}$). In other words, CNC sites occur preferentially at multiply bound CRMs. The most highly bound cluster of CRMs (cluster 1) is also enriched for well-conserved TF-binding events (Schmidt et al. 2010b) compared with the

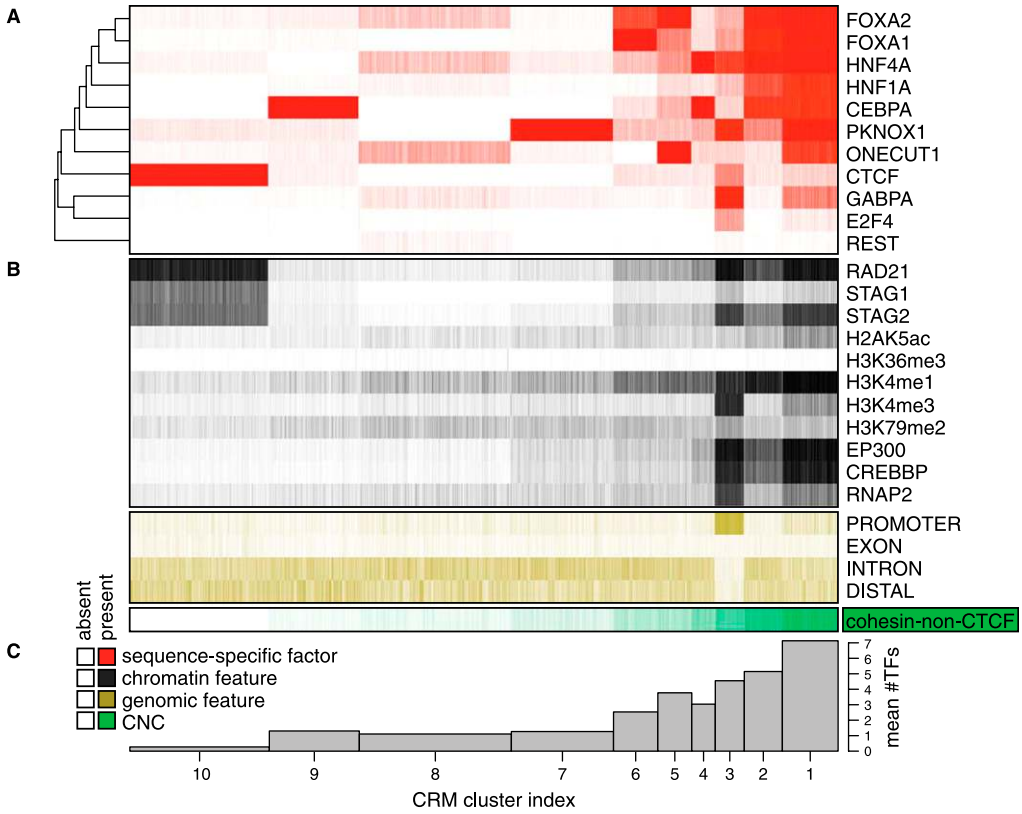


Figure 3. Cohesin-non-CTCF (CNC) binding occurs preferentially at multiply bound CRMs. (*A*) Results from K-means clustering (K = 10) of the binary presence/absence of ChIP-seq peaks corresponding to the 11 sequence-specific factors within 210,067 CRMs containing at least one of these factors. Factors were clustered based on the similarity in their binary occupancy profiles. The clusters were indexed and sorted by the proportion of CRMs with CNC in each cluster (increasing from *left* to *right*). (*B*) The binary presence/absence of ChIP-seq peaks for various chromatin features (non-sequence-specific factors and histone modifications) visualized according to the K-means results in *A*. Genomic location with respect to promoters (≤2.5 kb from an annotated TSS), exons, introns, and gene distal regions, is also indicated. The proportion of CRMs with CNC sites in each cluster is indicated at the *bottom* (increasing from *left* to *right*). (C) Barplot indicating the mean number of distinct TFs within each CRM cluster. Bar widths correspond to the number of CRMs within each cluster.

remaining clusters, including CEBPA-binding events shared in five species from chicken to human (Fisher's exact test $P = 10^{-10}$) and HNF4A-binding events shared in human, mouse, and dog (Fisher's exact test $P < 10^{-15}$). Taken together, these observations suggest a role for cohesin in integrating regulatory information from multiple TFs and stabilizing the binding of large multiprotein complexes to *cis*-regulatory sequences.

CNC presence is associated with liver-specific gene expression

Results from the unsupervised clustering analysis suggested that there might be a direct correlation between the number of TFs bound within a CRM, CNC presence, and the transcriptional activity of the genomic regions. By explicitly grouping CRMs into classes based purely on the number of distinct TFs present, we see that the proportion of CNC-containing CRMs significantly correlates with the number of bound TFs (Spearman's $\rho = 0.89$, $P = 10^{-3}$), whereas CTCF shows no significant correlation (Spearman's $\rho = 0.49$, $P = 0.13$). Indeed, almost two-thirds (62%) of highly occupied CRMs, defined as containing five or more TFs, possess CNC sites. The ratio of CNC- to CTCF-containing CRMs (CNC enrichment) is 0.2 when zero TFs are present, but reaches a maximum of threefold at seven TFs before returning to equivalence at 10 TFs (Fig. 4A). The proportion of promoter proximal CRMs (≤2.5 kb from an annotated TSS) is also correlated with the number of distinct TFs present (Spearman's $\rho = 0.95$, $P < 10^{-15}$), but in contrast to CNC enrichment that peaks at seven TFs, the proportion of both RNAP2 and H3K4me3 increase monotonically from 0 to 10 TFs (Fig. 4B). Other signs of transcriptionally active chromatin, such as the presence of the coactivators EP300/CREBBP, show a similar consistently increasing trend from one to 10 TFs (data not shown).

We next asked how these CRM occupancy patterns may be related to transcriptional output by assigning CRMs to their nearest canonical TSSs and using mouse liver expression data obtained by replicate RNA-seq experiments (Kutter et al. 2011) (see Methods). In addition, we identified 107 genes that are significantly up-regulated in mouse liver (Su et al. 2004). Median gene expression of CRM-associated genes increases when more than six TFs are present (Fig. 4B); however, only CRMs with between six and nine TFs are significantly enriched for the 107 genes significantly up-regulated in mouse liver cells (Fig. 4A) (Fisher's exact test $P < 0.01$) (Su et al. 2004). Strikingly, the peak of enrichment for tissue-specific genes at seven TFs coincides precisely with the peak in CNC enrichment at seven TFs. The three-way correspondence between liver-specific gene expression, CRM occupancy, and CNC sites, provides further evidence of cohesin's CTCF-independent transcriptional regulatory role at regions where multiple TFs assemble to effect tissue-specific expression.

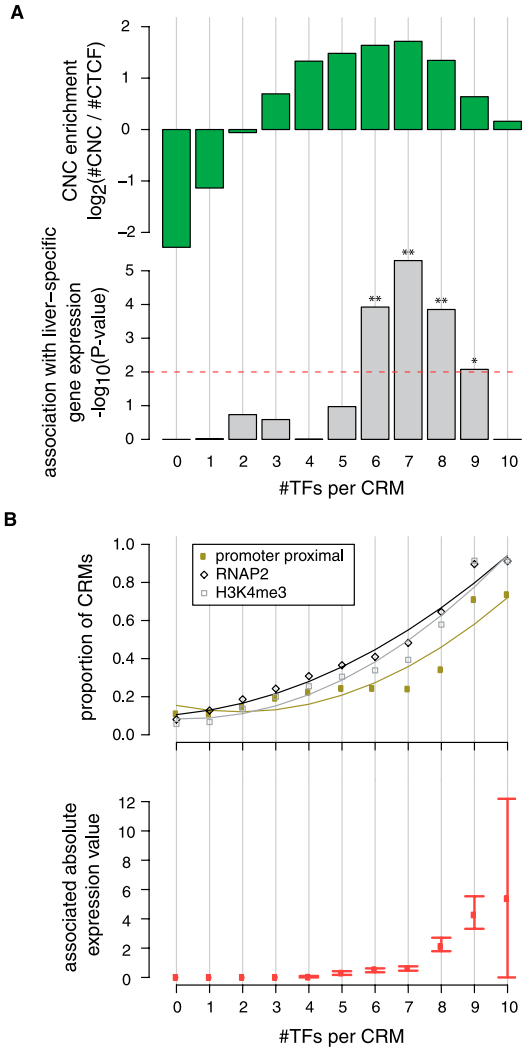


Figure 4. CNC sites are associated with liver-specific gene expression. (*A*) Ratio of CNC-containing CRMs versus those with CTCF (log-fold change) for CRM classes with 0–10 TFs. Each class of CRMs was also tested for association with 107 genes signficantly up-regulated in mouse liver cells (see Methods). The significance of the association (negative-log-transformed Fisher's exact test *P*-values) are indicated. (*) $P < 0.01$; (**) $P < 0.001$. The enrichment of CNC-containing CRMs reaches threefold when seven TFs are present, and coincides with highly significant enrichment for an association with liver-specific gene expression for the same class. (*B*) CRMs with high numbers of colocalizing TFs are associated with increased promoter proximity (≤2.5 kb from an annotated TSS) and characteristics of transcriptional activity (RNAP2 and H3K4me3 ChIP-seq peaks). Likewise, the associated absolute gene expression value increases significantly with the number of bound TFs. Error bars indicate the 95% confidence interval of the median.

Maximally occupied CRMs show similar properties to HOT regions

A total of 34 CRMs contain all 10 assayed TFs. These regions have similar characteristics to recently identified high-occupancy target (HOT) regions (Moorman et al. 2006; Gerstein et al. 2010; Nègre et al. 2011). These CRMs have high ChIP signal for all of the 10 TFs, are highly enriched in promoter–proximal regions (Fisher's exact test $P = 10^{-13}$), and are associated with genes having high absolute expression value—yet none of these are liver-specific genes. However, due to the low number of CRMs with all 10 factors, the confidence intervals for the expression value are large.

The group of genes associated with these HOT regions includes *Polr2a*, which encodes the largest subunit of the RNA polymerase II complex and *Ccnl1*, a gene whose product (cyclin L1) participates in the regulation of the pre-mRNA splicing process (Supplemental Table S2; Dickinson et al. 2002). Another gene with a nearby HOT region, *Grlf1*, encodes a transcription factor that binds to the promoter region of the glucocorticoid receptor (*Nr3c1*), a gene that is expressed in almost all cell types (Adcock and Caramori 2001).

These observations support the idea that HOT regions consist of constitutively open chromatin (Gerstein et al. 2010). Although the number of TFs in this study is limited, and many of those that were included have tissue-specific functions, these are the first HOT regions to be identified in vertebrates with similar properties to those described in the model organisms *D. melanogaster* and *C. elegans*.

Cohesin intensity explains disparities between motif score and ChIP signal

The resolution of ChIP-seq data lends itself to the problem of finding TF-binding site motifs, as the actual binding site is typically within ~50 bp of the peak summit. Nonetheless, the presence of the canonical motif usually explains only a fraction of the original ChIP-seq peaks (Valouev et al. 2008). Although the proportion of peaks with a motif match is dependent on the chosen score threshold, some ChIP-positive sequence regions have no recognizable similarity to the canonical motif (Johnson et al. 2007; Boyle et al. 2011). Furthermore, quantitative TF binding, as measured by either ChIP-chip or ChIP-seq enrichment, is only weakly correlated with motif strength, as measured by the PWM log-odds score (Schmidt et al. 2010b; Wilczyński and Furlong 2010).

In order to investigate these phenomena, we asked whether there was an unexpected correlation between a given factor's motif score and another factor's ChIP signal within our identified CRMs. Briefly, for each sequence-specific factor, we first determined the PWM score of the best motif match within each corresponding peak. We then compared this motif score with the ChIP signal of all other data sets within CRMs containing that peak. Similarly, motif score correlations were calculated for the occupancy count (i.e., the number of distinct TFs present) and the distance to the nearest canonical TSS (Fig. 5A; Supplemental Figs. S6, S7).

As expected due to their roles at the core promoter, high motif scores for both GABPA and E2F4 are most associated with H3K4me3 ChIP signal and tend to occur near to annotated TSSs (Conboy et al. 2007). In addition, CRM occupancy count is anticorrelated with motif scores of all factors except E2F4, indicating that when TFs occur in the absence of other potential binding partners, their binding is more likely to coincide with a high-scoring motif match. However, for only four out of the 11 sequence-specific factors that we tested, the factor's motif score is most strongly correlated with its own ChIP signal. In fact, the strength of motif score for four different factors (ONECUT1, FOXA1, FOXA2, and HNF4A) is most strongly associated with HNF4A ChIP signal.

Interestingly, cohesin ChIP signal is also anticorrelated with motif scores of all assayed factors except CTCF (Spearman's $\rho = 0.11$) and E2F4 (Spearman's $\rho = 0.13$); in other words, stronger cohesin binding is associated with lower-quality motif matches for

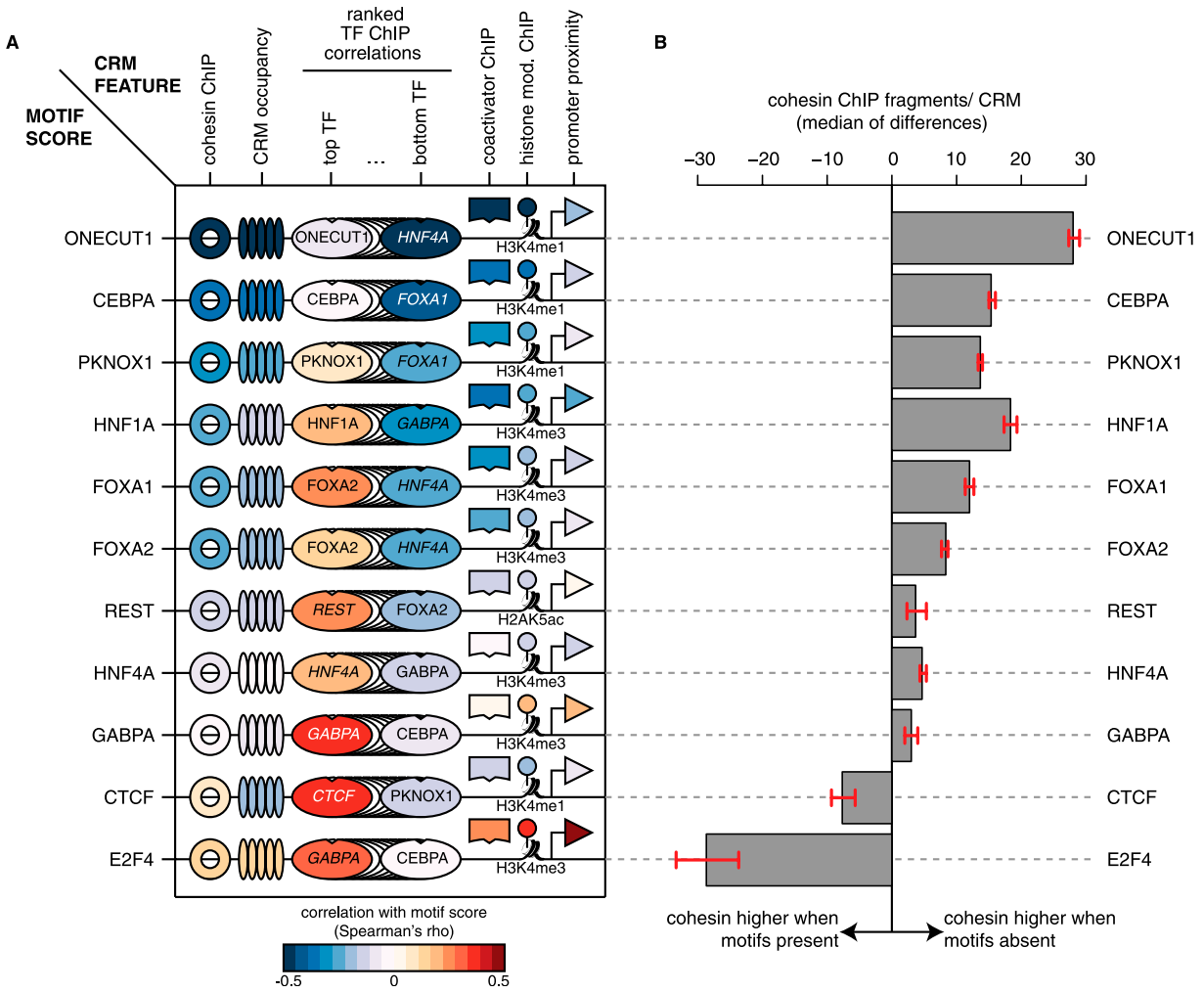


Figure 5. Cohesin ChIP signal is significantly associated with TF motif score. (*A*) Cartoon heatmap representation of correlations between each sequence-specific factor's motif score and the ChIP signal of all available ChIP-seq data sets. Correlations with CRM occupancy (number of distinct TFs present) and promoter proximity (distance to the nearest canonical TSS) are also shown. For each factor, the motif score correlation was calculated on the set of CRMs that contained a ChIP-seq peak for the same factor. Correlations with cohesin and coactivator ChIP signal were averaged over subunits (RAD21, STAG1, STAG2) and family members (EP300, CREBBP), respectively. Heatmap rows were ordered by increasing correlation with cohesin ChIP signal (from *top* to *bottom*). As a visual summary, only the top- and bottom-ranking correlations involving TFs are shown (see Supplemental Figs. S6, S7 for all correlations). (*B*) Increased cohesin ChIP signal at TF binding events without motifs. For each sequence-specific factor, the number of cohesin ChIP fragments within CRMs without high-scoring motifs was compared with that of CRMs with motifs. The 95% confidence intervals shown are based on a normal approximation of the Hodges–Lehmann estimate (median of all possible differences).

co-bound TFs. The two exceptions to this rule, i.e., positive correlations with E2F4 and CTCF, are unsurprising since strong E2F4 motifs and binding are found in highly occupied CRMs and CTCF binding has previously been shown to correlate well with motif quality, and there is evidence that CTCF recruits cohesin to sites where they co-occur (Parelho et al. 2008). For all other factors, stronger cohesin ChIP signals are associated with lower motif scores, particularly for the ONECUT1 motif (Spearman's $\rho = -0.5$) and CEBPA motif (Spearman's $\rho = -0.38$).

We also compared levels of cohesin ChIP signal between explicit groups of CRMs: those with and those without high-scoring motifs according to a minimum PWM score threshold. For all sequence-specific factors except CTCF and E2F4, we observe higher levels of cohesin in the absence of high-scoring motifs (Fig. 5B).

To determine whether cohesin presence could help to explain the discrepancy between TF ChIP signal and motif score, we trained logistic regression classifiers to predict the presence of high-scoring motifs for each sequence-specific factor, with and without cohesin ChIP signal information. For ONECUT1, CEBPA, HNF1A, PKNOX1, FOXA1, FOXA2, REST, and E2F4, cohesin ChIP information markedly improved the performance of the classifier. For GABPA, HNF4A, and CTCF there is minimal improvement in performance with the inclusion of cohesin in the model (Supplemental Fig. S8). These results suggest that cohesin presence is able to partially decouple ChIP signal from motif score for a significant number of TFs, including those that are often found at enhancer elements.

ONECUT1 ChIP signal is reduced at weak motifs in heterozygous *Rad21*$^{+/-}$ mouse liver cells

In order to determine whether cohesin plays an active role in the binding of TFs to their target sequences, particularly in the absence

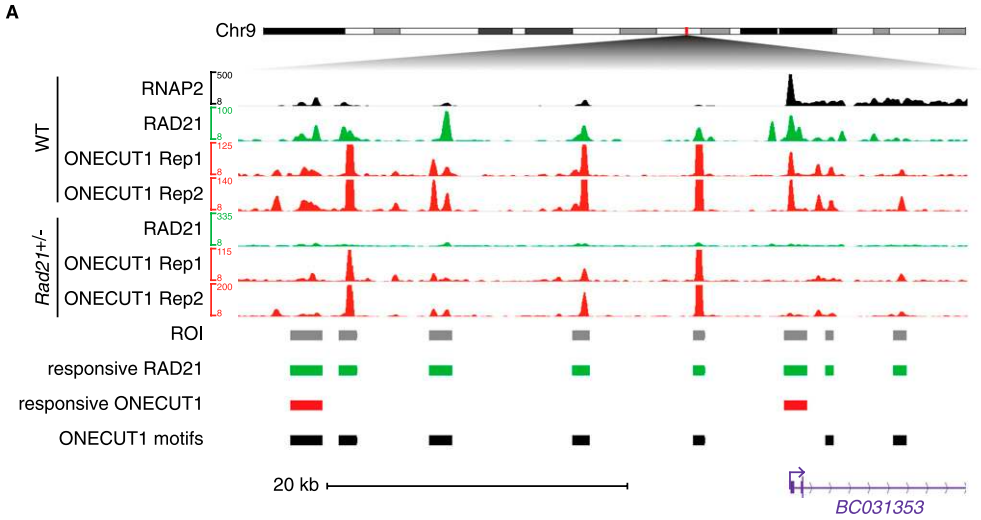
of high-scoring motifs, we used liver tissue from mice with only one functional allele of the *Rad21* gene. Homozygous knockouts of *Rad21* are lethal early in embryogenesis, suggesting that at least one wild-type *Rad21* allele is essential for normal development in mammals. Although heterozygous *Rad21*$^{+/-}$ mice are viable, they possess a number of defects including hypersensitivity to ionizing radiation and impaired DNA repair capacity (Xu et al. 2010). To confirm that the level of cohesin binding is reduced, and to determine whether TF binding is consequently affected, we mapped RAD21, ONECUT1, CEBPA, and HNF4A in heterozygous *Rad21*$^{+/-}$ mouse liver cells using ChIP-seq.

The total number of binding events for all TFs is reduced in heterozygous *Rad21*$^{+/-}$ cells (ONECUT1 45%, CEBPA 63%, HNF4A 18%) and, as expected, the reduction is most severe for RAD21 (14%). A total of 78,625 CRMs lose RAD21 binding according to the absence of an overlapping peak in heterozygous *Rad21*$^{+/-}$ cells. We focus the remainder of our analysis on these sites. In terms of peak loss, CRMs without high-scoring motifs are enriched for binding events lost in heterozygous *Rad21*$^{+/-}$ cells (responsive binding events) for all three assayed TFs (Fisher's exact test $P < 10^{-15}$; Supplemental Fig. S9). We also performed statistically robust differential binding analysis on replicate ONECUT1 and CEBPA ChIP-seq data in order to determine ChIP signal differences between wild-type and heterozygous *Rad21*$^{+/-}$ cells (see Methods). Similar to the peak-level analysis results, CRMs exhibiting significantly reduced ONECUT1 ChIP signal in mutant cells are enriched for ONECUT1 peaks without high-scoring motifs (Fisher's exact test $P = 10^{-4}$; Fig. 6B). However, differential binding analysis for CEBPA revealed no significant differences in ChIP signal.

Interestingly, promoter–proximal CRMs tend to be associated with both reduced ONECUT1 motif scores (Fig. 5A) and *Rad21*$^{+/-}$ responsive ONECUT1 binding events (Fisher's exact test $P < 10^{-15}$). This suggests that cohesin may help to stabilize the binding of ONECUT1 near promoters in particular. One such region is shown in Figure 6A overlapping the *BC031353* promoter, where all but one of the remaining ONECUT1-containing CRMs displayed retain ONECUT1 binding in heterozygous *Rad21*$^{+/-}$ cells (resistant binding events). Note that a high-scoring motif is absent from the *Rad21*$^{+/-}$ responsive ONECUT1 binding event overlapping the TSS, although the effect on *BC031353* expression was not assessed.

Mirrored binding of CTCF near transcription start sites and cohesin-bound enhancers are associated with elevated expression levels

Cohesin has been shown to be crucial for two distinct types of chromatin interactions: (1) looping between individual CTCF binding events (Hadjur et al. 2009; Mishiro et al. 2009; Nativio et al. 2009; Hou et al. 2010), and (2) interactions between promoters and CNC-containing enhancers (Kagey et al. 2010; Schmidt et al. 2010a; Seitan et al. 2011). Reports of long-range chromatin looping mediated by CTCF have suggested that CTCF may influence



B

WT ONECUT1 CRMs	*Rad21*+/- responsive	*Rad21*+/- resistant
no motif	203	2703
motif	774	14693

Fisher's exact test $P_{(enrich)} = 10^{-4}$
odds ratio = 1.4

Figure 6. ONECUT1 ChIP-seq in heterozygous *Rad21*$^{+/-}$ mouse liver cells shows preferential loss of TF binding events where no motif is present. (*A*) Sample region near the *BC031353* gene showing overall reduction in RAD21 ChIP signal in heterozygous *Rad21*$^{+/-}$ cells (responsive RAD21) and associated significant reduction in ONECUT1 ChIP signal within two CRMs (responsive ONECUT1). The ONECUT1 binding event overlapping the TSS contains no ONECUT1 motif. (*B*) WT ONECUT1 CRMs without motifs show a preferential decrease in ChIP signal (FDR < 0.1) in heterozygous *Rad21*$^{+/-}$ mouse liver cells (Fisher's exact test $P = 10^{-4}$). Regions of interest (ROI) are those CRMs where RAD21 binding was ablated in heterozygous *Rad21*$^{+/-}$ mouse liver cells (responsive RAD21).

transcription by facilitating enhancer–promoter interactions (Handoko et al. 2011). In this model, interactions between promoter–proximal and distal CTCF binding events connect enhancers to their target genes by looping out the intervening DNA, thereby reducing the effective distance and increasing the probability of interactions between linearly distant genomic regulatory regions.

We therefore searched for genes where this configuration has the potential to occur, i.e., genes with CTCF/cohesin binding events both nearby the TSS and proximal to their associated enhancers (Fig. 7B). If these consistent binding patterns have biological relevance, we expect their presence to be associated with increased expression levels of the corresponding genes. To test this, we first compiled a list of putative liver-specific enhancers, defined as CRMs >5 kb from their nearest canonical TSS that possess (1) a CNC site, (2) the liver master regulator HNF4A, (3) the EP300 enhancer marker, and (4) the histone signature H3K4me1, but (5) not H3K4me3. We then assigned each identified enhancer to the nearest gene based on distance to the TSS, such that each enhancer is assigned to only one gene.

Of the 5364 genes with nearby enhancers as defined above, 532 genes have CTCF binding events both 2.5 kb from the TSS and nearby their enhancers (≤2.5 kb). Indeed, these genes have significantly greater expression values than genes with CTCF binding near the TSS, but not nearby their enhancers, or vice versa (Fig. 7A) (Mann-Whitney *U*-test $P < 10^{-3}$).

Discussion

Cohesin has multiple vital functions in mammalian cells, including well-established roles in sister chromatid segregation in mitosis and meiosis. Recent results have implicated cohesin in the regulation of gene expression. Because cohesin has no known DNA-binding domain, the mechanism of this transcriptional regulation is assumed to arise from cohesin's ability to stabilize higher-order chromatin structure through interactions with chromatin organization proteins such as CTCF (Hadjur et al. 2009; Nativio et al. 2009). We and others have shown that cohesin plays a role in tissue-specific transcriptional regulation and that this role is at least partially characterized by CTCF-independent cohesin localization with master regulators in several tissues. To better understand cohesin's contribution to gene regulation, we collected genome-wide localization data of 10 TFs, several histone modifications and other functional DNA–protein interactions in primary mouse liver. These data provide a comprehensive map of cohesin's two known roles: one associated with CTCF and another CTCF-independent role in tissue-specific transcriptional regulation. We show that these roles are functionally similar across multiple tissues by demonstrating that cohesin's presence at binding events of liver-specific TFs mirrors its localization with ES cell-specific factors.

To further characterize cohesin's tissue-specific regulatory role, we focused on the properties of cohesin-non-CTCF (CNC) sites. By clustering the binding patterns of sequence-specific factors within CRMs and ranking these clusters by the fraction that overlap CNC sites, we demonstrate that CNC sites occur preferentially at CRMs containing multiple TFs and are less likely to be found at CRMs with singleton binding events that represent the majority of regions bound by any given factor. The class of CRMs with the most TFs is also highly enriched for binding events that are persistent across hundreds of millions of years of evolution (Schmidt et al. 2010b), suggesting that the conservation of these events—and possibly those of other tissue-specific TFs—is attributable to their highly bound state and putative functional context. However, only 5% of the CRMs in the maximally occupied cluster have a deeply shared binding event (i.e., five-species

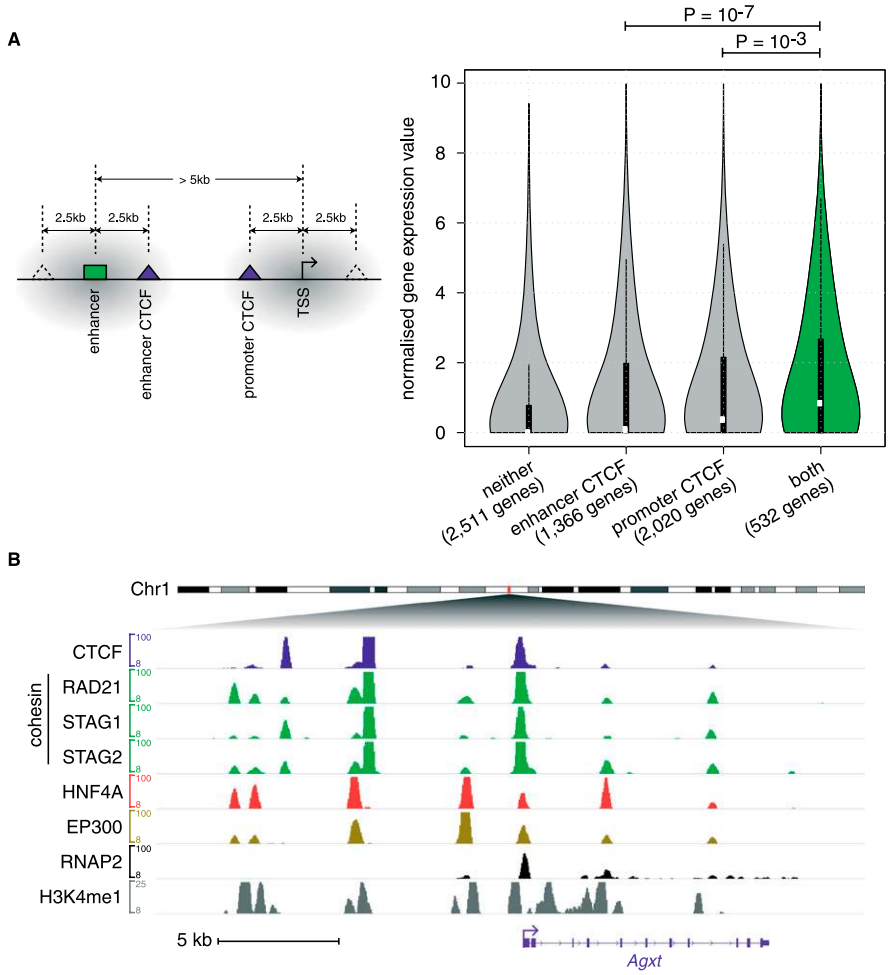


Figure 7. Simultaneous CTCF binding within promoters and nearby enhancers is associated with elevated expression levels. (*A*) Violin plots showing gene expression distributions. Genes with CTCF binding events both within their promoters and nearby their associated enhancers show significantly elevated expression levels over those of the other three indicated classes (Mann-Whitney *U*-test $P < 10^{-3}$). (*B*) Sample region near the liver-expressed *Agxt* gene, where CTCF binds within the core promoter, as well as near putative upstream cohesin-bound enhancers. Note that while CTCF is absent from the enhancers (CNC), it co-binds with HNF4A and EP300 within the *Agxt* promoter.

CEBPA or three-species HNF4A). Thus, the colocalization of a large number of TFs does not mean, a priori, that a binding event will be invariant over evolutionary time.

We observe a striking relationship between CNC enrichment and liver-specific gene expression for CRMs with submaximal numbers of distinct TFs bound. In particular, CRMs with between six and eight of our assayed TFs are more than twice as likely to possess CNC sites than CTCF, and are also the most highly enriched for liver-specific gene expression. Although both CNC enrichment and association with liver-specific expression peaks when CRMs have seven TFs, we find no evidence in the mouse ES cell data that this is a (just right) "goldilocks" number of TFs. However, our results in mouse liver are consistent with previous research in other species, showing that regions with low-to-moderate numbers of transcription factors are most significantly enriched for annotated enhancers and signs of active transcriptional regulation (Nègre et al. 2011). Therefore, although the distribution of tissue-specific and ubiquitous factors is different in the ES cell experiments, this does not rule out the attractive hypothesis that a specific and relatively small number of TFs binding together and stabilized by cohesin is a fundamental characteristic of mammalian tissue-specific gene regulation.

Intriguingly, the most highly occupied CRMs containing all 10 of our assayed TFs are neither associated with liver-specific genes nor CNC enrichment. Instead, these regions seem to be nearby constitutively active genes and have characteristics that are similar to recently described HOT regions (Moorman et al. 2006; Gerstein et al. 2010; Nègre et al. 2011). To our knowledge these are the first HOT regions to be described in vertebrates.

The DNA sequence preferences of TFs are typically described using position weight matrices (PWMs), and referred to as binding-site motifs. These motifs remain a challenge to discover computationally despite the large number of de novo motif discovery algorithms that have been developed to infer these sequence preferences (Nguyen and Androulakis 2009). Previous results have demonstrated that while ChIP-seq data is useful for identifying the specific regions of the genome bound by a given TF, there remains a subset of binding events with either weak or nonexistent motif matches. This lack of a clear relationship between DNA sequence content and TF recruitment has been described as a result of indirect or cooperative binding, and recent approaches tailored specifically to ChIP-seq data have subsequently focused on finding these candidate cofactors (Bailey 2011).

Using both computational and experimental methods, we show that the presence of cohesin likely explains the inverse relationship between ChIP signal and motif score observed for a number of our assayed factors. These TFs bind to stronger motifs in the absence of cohesin. Stated alternatively, we observe higher levels of cohesin in the absence of high-scoring motifs. These results suggest that cohesin enables TFs to bind to suboptimal motif sequences either by stabilizing large protein–DNA complexes at highly occupied CRMs or by inducing binding through specific chromatin contortions. Importantly, we show that computational classifiers trained to predict high-scoring motif occurrence exhibit markedly improved performance when cohesin is incorporated into the model. Furthermore, using ChIP-seq in the livers of a *Rad21*-cohesin haploinsufficient mouse model, we show that heterozygous loss of *Rad21* results in the loss of 86% of RAD21 binding events found in the wild type. This is accompanied by a reduction in ChIP-seq peak numbers for ONECUT1, CEBPA, and HNF4A that disproportionately affects binding events without high-scoring motifs for these TFs. Similarly, we find that sites both without RAD21 peaks and showing a significant loss of ONECUT1 ChIP signal in heterozygous $Rad21^{+/-}$ cells are also significantly depleted for high-scoring ONECUT1 motifs. Taken together with our observations in wild-type cells that cohesin is more abundant at highly occupied CRMs and at those without high-scoring motifs, these results point toward a role for cohesin in stabilizing the binding of TFs to *cis*-regulatory sequences, particularly near promoters. Alternatively, expression level differences of the TFs themselves caused by the loss of cohesin may contribute to the overall reduction in binding events observed in heterozygous $Rad21^{+/-}$ cells.

Promoter regions are important sites of TF binding, where multiple regulatory signals are integrated to coordinate cell-type-specific expression programs. Both CTCF and cohesin have been shown to modulate chromatin structure in order to enable promoter–proximal factors to respond to signals from distant *cis*-regulatory elements, such as enhancers. However, our results indicate that the majority of highly occupied CRMs, which show typical characteristics of enhancers, possess cohesin in the absence of CTCF (CNC sites). An attractive hypothesis is that CTCF may set up indirect chromatin interactions as the primary step toward enabling enhancer–promoter communications (Handoko et al. 2011). We tested whether the dual presence of CTCF-binding events both nearby TSSs and their corresponding enhancers is associated with increased expression levels. Using this simple approach, we observe genome-wide patterns that support the model that concerted CTCF binding to linearly distant regulatory regions is associated with significantly elevated expression levels. Further investigations using 3C-based chromatin conformation assays would be needed to determine whether these patterns are indeed associated with functional chromatin looping interactions between enhancers and promoters.

Methods

ChIP sequencing

ChIP experiments were performed with wild-type primary mouse (C57BL/6 and/or C57BL/6xA/J) liver tissue and antibodies against CTCF (two replicates, two individuals; antibody: Upstate Biotechnology, 07729), STAG1 (three replicates, two individuals; antibody: Abcam, ab4457), STAG2 (singlicate; antibody: Abcam, 4464), RAD21 (singlicate; antibody: Abcam, ab992), CEBPA (six replicates, two individuals; antibody: Santa Cruz Biotechnology, sc9314), HNF4A (two replicates, one individual; antibody: aviva systems biology, ARP31946), FOXA1 (two replicates, two individuals; antibody: Abcam, ab5089), FOXA2 (four replicates, two individuals; antibody: Santa Cruz Biotechnology, sc6554), ONECUT1 (six replicates, two individuals; antibody: Santa Cruz Biotechnology, sc13050), HNF1A (three replicates, one individual; antibody: Santa Cruz Biotechnology, sc6547), PKNOX1 (singlicate; antibody: Santa Cruz Biotechnology, sc6245), REST (singlicate; antibody: Santa Cruz Biotechnology, sc25398), GABPA (two replicates, one individual; antibody: Santa Cruz Biotechnology, sc22810), E2F4 (singlicate; antibody: Santa Cruz Biotechnology, sc1082), EP300 (two replicates, two individuals; antibody: Santa Cruz Biotechnology, sc585), CREBBP (singlicate; antibody: Santa Cruz Biotechnology, sc369), RNAP2 (two replicates, two individuals; antibody: Abcam, ab5408), H3K4me1 (singlicate; antibody: Abcam, ab8895), H3K4me3 (singlicate; antibody: Abcam, ab8580), H3K36me3 (singlicate; antibody: Abcam, ab9050), H3K79me2 (singlicate; antibody: Abcam, ab3594) and H2AK5ac (singlicate; antibody: Abcam, 1764) as recently described (Schmidt et al. 2009). Briefly, the immunoprecipitated DNA was end-repaired, A-tailed, ligated to the sequencing adapters, amplified by 18 cycles

of PCR, and size selected (200–300 bp) followed by single-end sequencing on an Illumina Genome Analyzer according to the manufacturer's recommendations.

ChIP experiments were performed with heterozygous *Rad21*$^{+/-}$ primary mouse liver tissue and antibodies against RAD21 (two replicates, two individuals; antibody: Abcam, ab992), CEBPA (two replicates, two individuals; antibody: Santa Cruz Biotechnology, sc9314), HNF4A (two replicates, two individuals; antibody: aviva systems biology, ARP31946), ONECUT1 (two replicates, two individuals; antibody: Santa Cruz Biotechnology, sc13050) as above.

Read mapping and peak calling

All ChIP sequencing reads from each replicate were aligned to the mouse reference genome assembly (NCBI37/mm9) using BWA (Li and Durbin 2009) with default parameters. After pooling replicate data for each factor/histone-modification, the reads were then filtered to remove low-quality mappings (*phred*-scaled mapping quality <10), multiple reads mapping to the same genomic location and strand, as well as those mapping to the mitochondrial genome. Peaks were then called on all data sets using matched input data and a dynamic programming algorithm (SWEmbl) with -R 0.005 as recently described (Schmidt et al. 2010a). See Supplemental Table S3.

Cohesin-non-CTCF site definition and peak clustering

Firstly, overlapping ChIP-seq peaks for CTCF and the cohesin subunits (STAG1, STAG2, RAD21) were merged to form a set of disjoint genomic regions. Our definition of cohesin-non-CTCF (CNC) sites required the presence of at least one cohesin subunit peak and the absence of CTCF. In order to obtain a high-confidence set of CNC sites, in the absence of significant CTCF ChIP enrichment that may have escaped peak-detection, we required that these sites also satisfied the following criterion: log((norm_CTCF_ChIP)/(norm_Input))<0.68. This cut-off corresponds to the fifth percentile of ChIP enrichment scores within CTCF peaks.

Overlapping peak regions of the sequence-specific factors (CTCF, CEBPA, HNF4A, FOXA1, FOXA2, ONECUT1, HNF1A, PKNOX1, REST, GABPA, E2F4), as well as cohesin (STAG1, STAG2, RAD21), CNC sites, and the coactivators EP300/CREBBP, were merged to define putative *cis*-regulatory modules (CRMs) (Zinzen et al. 2009). A single-linkage clustering approach was used, where a peak overlap of ≥1 bp with at least one other peak within a CRM is sufficient for membership within the CRM. The presence or absence of a particular histone-modification (H3K4me1, H3K4me3, H3K36me3, H3K79me2, H2AK5ac) or RNAP2 binding within a CRM was then determined post hoc by satisfaction of either of the following criteria: (1) the presence of an overlapping peak, or (2) ChIP enrichment within the entire CRM region of at least threefold, where the number of ChIP reads overlapping the CRM≥8.

Motif analysis and selection

We used MEME (Bailey and Elkan 1994) and NestedMica (Down and Hubbard 2005) to perform de novo motif discovery for each sequence-specific factor using peak regions with the top 1000 scores. In each case, 50 bp of DNA sequence surrounding the SWEmbl summit was used to find five frequently occurring sequence motifs up to 25 bp in length (MEME parameters: -nmotifs 5 -maxw 25 -revcomp; NestedMica parameters: -numMotifs 5 -minLength 5 -maxLength 25 -revComp -backgroundOrder 1 -backgroundClasses 4). We scanned all bound (*positive*) regions for each factor with PWMs for all five NestedMica motifs as well as the top-scoring MEME motif to determine the score of the best motif match in each case. We repeated this using equally sized unbound (*negative*) regions, which were randomly sampled from the repeat- and exon-masked genome. The optimal motif for each factor, which was retained for further analysis, was defined as that best able to discriminate between positive and negative regions according to the AUC (area under ROC curve) performance measure.

Mouse embryonic stem cell data analysis

Publicly-available ChIP-seq data sets from mouse embryonic stem cells were downloaded, reprocessed, and analyzed using a similar procedure to that described above: CTCF, MYC, ESRRB, KLF4, MYCN, SMAD1, STAT3, TCFCP2L1, ZFX, EP300, SUZ12 (Chen et al. 2008), NANOG, POU5F1, SOX2, H3K79me2 (Marson et al. 2008), RNAP2 (Seila et al. 2008), NIPBL, SMC1A, SMC3, MED1, MED12 (Kagey et al. 2010). See Supplemental Table S3.

CRM clustering and analysis

To restrict our analysis to sites with possible patterns of combinatorial TF binding, we filtered our data to retain only CRMs containing a binding event of at least one sequence-specific factor. We used two independent methods (K-means and *AutoClass*) to group CRMs into similar clusters.

(1) We used K-means to group CRMs into K similar clusters based on the binary presence/absence of the 11 sequence-specific factors within each CRM. In order to choose an appropriate value for K, we ran the clustering algorithm on a random subset of 20,000 CRMs and determined the median within-cluster sum of squares (WCSS) over 10 replicates of each value of K in the range [2–50]. The WCSS tends to decrease as the number of clusters K increases, but the decrease flattens slightly for values of K near 10 (see Supplemental Fig. S5). We used this "elbow" method to choose a value of K = 10 when running the algorithm on the entire data set.

(2) We used *AutoClass* (Cheeseman and Stutz 1996) to group CRMs into similar classes based on the normalized ChIP enrichment of the 11 sequence-specific factors within each CRM. *AutoClass* uses a Bayesian probabilistic approach to automatically optimize the properties of each class (as well as the number of classes) to achieve the best separation. An advantage of this "fuzzy" clustering approach, not provided by other traditional clustering methods such as K-means, is the availability of a measure (posterior probability) to assess the confidence that each CRM belongs to its assigned class. The *AutoClass* C command-line program was used with the following primary settings: (1) data model: single_normal_cn (factor ChIP enrichments follow conditionally independent normal variables); (2) convergence criterion: converge_3 (most stringent); (3) initial values for the number of class: 2, 3, 5, 7, 10, 15, 25, 35, 45, 55, 65, 75, 85, 95, 105; (4) absolute error of input data: 10% for all factors. We filtered the CRMs in the resulting classification to retain only those with posterior probabilities ≥0.5 and combined classes with high correlation (Spearman's $\rho > 0.9$) between their median ChIP enrichment profiles.

Other CRM attributes such as the ChIP peak presence/absence of other factors/histone-modifications not used in the original clustering were added to aid visualization of the clustering results. Gene annotation information from Ensembl version 60 (Flicek et al. 2010) was used to add genomic localization information for each CRM, where "Promoter" was defined as occurring ≤2.5 kb from an annotated TSS, "Exon" corresponds to overlap with an exon but not a promoter, "Intron" corresponds to overlap with

a gene but neither an exon nor a promoter, and "Distal" for localization elsewhere. Gene annotation information for pseudogenes was ignored throughout the analysis.

Expression analysis

We used previously published RNA-seq data from mouse liver to obtain absolute expression estimates for all genes (Kutter et al. 2011). Briefly, the raw reads were truncated to 35-mers and aligned to mouse transcript sequences (cDNA sequences from Ensembl version 60, NCBI37/mm9) using Bowtie version 0.12.7 (Langmead et al. 2009) with default parameters. Normalized gene expression estimates were obtained using MMSEQ (Turro et al. 2011) and summarized by taking the replicate mean.

We used a previously published data set consisting of expression measurements from 40 diverse mouse tissues to determine sets of genes with liver-specific patterns of expression (Su et al. 2004). The processed data (ArrayExpress accession: E-MTAB-25) was obtained from the Gene Expression Atlas (Kapushesky et al. 2010) where up-regulation in a particular tissue with respect to the remainder was assessed using a *t*-test and $P < 0.05$.

Motif presence prediction

For each sequence-specific factor, we trained logistic regression classifiers to predict the presence of high-scoring motif matches using the ChIP signals (estimated number of ChIP fragments overlapping a given CRM) of various factors. Models were trained using: (1) ChIP signal of the corresponding factor, and (2) both ChIP signal of the corresponding factor and ChIP signals of the cohesin subunits (RAD21, STAG1, STAG2). Motif score cut-offs corresponding to FDR = 0.4 were chosen to determine high-scoring motif match presence/absence (see Supplemental Fig. S10). Ten-fold cross-validation was performed using CRMs containing a peak for the factor of interest, where 50% of these CRMs were randomly selected for the training set and the remaining 50% formed the test set.

Wild-type versus heterozygous *Rad21*$^{+/-}$ differential binding analysis

Read mapping and filtering for CEBPA and ONECUT1 was carried out as described above for both wild-type and heterozygous *Rad21*$^{+/-}$ ChIP-seq data sets, except reads for biological replicates were handled separately (technical replicates were pooled). The DiffBind package (Ross-Innes et al. 2012) was used with default parameters to determine CRMs with significantly lower ChIP signal in heterozygous *Rad21*$^{+/-}$ mouse liver cells versus wild-type liver cells (FDR threshold = 0.1).

Data access

Data deposited under ArrayExpress accession number E-MTAB-941.

Acknowledgments

We thank John Marioni, Benoit Ballester, and Angela Goncalves for helpful discussions, as well as the EBI systems team and the CRI Genomics and Bioinformatics Cores. This research is supported by the European Molecular Biology Laboratory (A.J.F., P.C.S., P.F.), Cancer Research UK (D.S., M.D.W., D.T.O.), the Wellcome Trust (WT079643) and by the European Research Council, EMBO Young Investigator Program, and Hutchinson Whampoa (D.T.O).

Author contributions: A.J.F. and P.C.S. analyzed the data; D.S., P.C.S, P.F., and D.T.O. conceived and designed the experiments; D.S., M.D.W., S.W., H.X., and R.G.R. performed the experiments; A.J.F., D.S., D.T.O., and P.F. wrote the manuscript; and D.T.O and P.F. oversaw the work.

References

Adcock IM, Caramori G. 2001. Cross-talk between pro-inflammatory transcription factors and glucocorticoids. *Immunol Cell Biol* **79:** 376–384.
Anderson DE, Losada A, Erickson HP, Hirano T. 2002. Condensin and cohesin display different arm conformations with characteristic hinge angles. *J Cell Biol* **156:** 419–424.
Bailey TL. 2011. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27:** 1653–1659.
Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2:** 28–36.
Bartkuhn M, Straub T, Herold M, Herrmann M, Rathke C, Saumweber H, Gilfillan GD, Becker PB, Renkawitz R. 2009. Active promoters and insulators are marked by the centrosomal protein 190. *EMBO J* **28:** 877–888.
Bénard CY, Kébir H, Takagi S, Hekimi S. 2004. *mau-2* acts cell-autonomously to guide axonal migrations in *Caenorhabditis elegans*. *Development* **131:** 5947–5958.
Botta M, Haider S, Leung IX, Lio P, Mozziconacci J. 2010. Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol Syst Biol* **6:** 426. doi: 10.1038/msb.2010.79.
Boyle AP, Song L, Lee B-K, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS. 2011. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res* **21:** 456–464.
Cheeseman P, Stutz J. 1996. Bayesian classification (AutoClass): Theory and results. In *Advances in knowledge discovery and data mining*, pp. 153–180. American Association for Artificial Intelligence, Menlo Park, CA.
Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133:** 1106–1117.
Conboy CM, Spyrou C, Thorne NP, Wade EJ, Barbosa-Morais NL, Wilson MD, Bhattacharjee A, Young RA, Tavaré S, Lees JA, et al. 2007. Cell cycle genes are the evolutionarily conserved targets of the E2F4 transcription factor. *PLoS ONE* **2:** e1061. doi: 10.1371/journal.pone.0001061.
Cuddapah S, Jothi R, Schones DE, Roh T-Y, Cui K, Zhao K. 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19:** 24–32.
Dickinson LA, Edgar AJ, Ehley J, Gottesfeld JM. 2002. Cyclin L is an RS domain protein involved in pre-mRNA splicing. *J Biol Chem* **277:** 25465–25473.
Donze D, Adams CR, Rine J, Kamakaka RT. 1999. The boundaries of the silenced HMR domain in *Saccharomyces cerevisiae*. *Genes Dev* **13:** 698–708.
Down TA, Hubbard TJP. 2005. NestedMICA: Sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res* **33:** 1445–1453.
Engel N, Raval AK, Thorvaldsen JL, Bartolomei SM. 2008. Three-dimensional conformation at the *H19/Igf2* locus supports a model of enhancer tracking. *Hum Mol Genet* **17:** 3021–3029.
Fay A, Misulovin Z, Li J, Schaaf CA, Gause M, Gilmour DS, Dorsett D. 2011. Cohesin selectively binds and regulates genes with paused RNA polymerase. *Curr Biol* **21:** 1624–1634.
Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2010. Ensembl 2011. *Nucleic Acids Res* **39:** D800–D806.
Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. 2009. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462:** 58–64.
Gard S, Light W, Xiong B, Bose T, McNairn AJ, Harris B, Fleharty B, Seidel C, Brickner JH, Gerton JL. 2009. Cohesinopathy mutations disrupt the subnuclear organization of chromatin. *J Cell Biol* **187:** 455–462.
Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330:** 1775–1787.
Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, et al. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453:** 948–951.
Hadjur S, Williams LM, Ryan NK, Cobb BS, Sexton T, Fraser P, Fisher AG, Merkenschlager M. 2009. Cohesins form chromosomal *cis*-interactions at the developmentally regulated IFNG locus. *Nature* **460:** 410–413.
Haering CH, Löwe J, Hochwagen A, Nasmyth K. 2002. Molecular architecture of SMC proteins and the yeast cohesin complex. *Mol Cell* **9:** 773–788.

Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CWH, Ye C, Ping JLH, Mulawadi F, et al. 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43:** 630–638.
Horsfield JA, Anagnostou SH, Hu JK, Cho KH, Geisler R, Lieschke G, Crosier KE, Crosier PS. 2007. Cohesin-dependent regulation of Runx genes. *Development* **134:** 2639–2649.
Hou C, Dale R, Dean A. 2010. Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc Natl Acad Sci* **107:** 3651–3656.
Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.
Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467:** 430–435.
Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A. 2010. Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res* **38:** D690–D698.
Krantz ID, McCallum J, DeScipio C, Kaur M, Gillis LA, Yaeger D, Jukofsky L, Wasserman N, Bottani A, Morris CA, et al. 2004. Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of *Drosophila melanogaster Nipped-B*. *Nat Genet* **36:** 631–635.
Kurukuti S, Tiwari VK, Tavoosidana G, Pugacheva E, Murrell A, Zhao Z, Lobanenkov V, Reik W, Ohlsson R. 2006. CTCF binding at the *H19* imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to *Igf2*. *Proc Natl Acad Sci* **103:** 10684–10689.
Kutter C, Brown GD, Gonçalves Â, Wilson MD, Watt S, Brazma A, White RJ, Odom DT. 2011. Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nat Genet* **43:** 948–955.
Langmead B, Trapnell C, Pop M. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** r25. doi: 10.1186/gb-2009-10-3-r25.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.
Lieberman-Aiden E, Van Berkum N, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie B, Sabo P, Dorschner M, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326:** 289–293.
Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J, et al. 2008. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134:** 521–533.
Mishiro T, Ishihara K, Hino S, Tsutsumi S, Aburatani H, Shirahige K, Kinoshita Y, Nakao M. 2009. Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. *EMBO J* **28:** 1234–1245.
Misteli T. 2007. Beyond the sequence: Cellular organization of genome function. *Cell* **128:** 787–800.
Misulovin Z, Schwartz YB, Li XY, Kahn TG, Gause M, MacArthur S, Fay JC, Eisen MB, Pirrotta V, Biggin MD, et al. 2007. Association of cohesin and Nipped-B with transcriptionally active regions of the *Drosophila melanogaster* genome. *Chromosoma* **117:** 89–102.
Moorman C, Sun L, Wang J. 2006. Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster*. *Proc Natl Acad Sci* **8:** 12027–12032.
Murrell A, Heeson S, Reik W. 2004. Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops. *Nat Genet* **36:** 889–893.
Nasmyth K, Haering CH. 2009. Cohesin: Its roles and mechanisms. *Annu Rev Genet* **43:** 525–558.
Nativio R, Wendt KS, Ito Y, Huddleston JE, Uribe-Lewis S, Woodfine K, Krueger C, Reik W, Peters JM, Murrell A. 2009. Cohesin is required for higher-order chromatin conformation at the imprinted *IGF2-H19* locus. *PLoS Genet* **5:** e1000739. doi: 10.171/journal.pgen1000739.
Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, et al. 2011. A *cis*-regulatory map of the *Drosophila* genome. *Nature* **471:** 527–531.
Nguyen TT, Androulakis IP. 2009. Recent advances in the computational discovery of transcription factor binding sites. *Algorithms Mol Biol* **2:** 582–605.
Nitzsche A, Paszkowski-Rogacz M, Matarese F, Janssen-Megens EM, Hubner NC, Schulz H, de Vries I, Ding L, Huebner N, Mann M, et al. 2011. RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. *PLoS ONE* **6:** e19470. doi: 10.1371/journal.pone.0019470.
Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, Jarmuz A, Canzonetta C, Webster Z, Nesterova T, et al. 2008. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132:** 422–433.
Pauli A, Althoff F, Oliveira RA, Heidmann S, Schuldiner O, Lehner CF, Dickson BJ, Nasmyth K. 2008. Cell-type-specific TEV protease cleavage reveals cohesin functions in *Drosophila* neurons. *Dev Cell* **14:** 239–251.
Peters JM, Tedeschi A, Schmitz J. 2008. The cohesin complex and its roles in chromosome biology. *Genes Dev* **22:** 3089–3114.
Phillips JE, Corces VG. 2009. CTCF: Master weaver of the genome. *Cell* **137:** 1194–1211.
Rollins RA, Morcillo P, Dorsett D. 1999. Nipped-B, a *Drosophila* homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes. *Genetics* **152:** 577–593.
Rollins RA, Korom M, Aulner N, Martens A, Dorsett D. 2004. *Drosophila* nipped-B protein supports sister chromatid cohesion and opposes the stromalin/Scc3 cohesion factor to facilitate long-range activation of the *cut* gene. *Mol Cell Biol* **24:** 3100–3111.
Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, Brown GD, Gojis O, Ellis IO, Green AR, et al. 2012. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481:** 389–393.
Rubio ED, Reiss DJ, Welcsh PL, Disteche CM, Filippova GN, Baliga NS, Aebersold R, Ranish JA, Krumm A. 2008. CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci* **105:** 8309–8314.
Schmidt D, Wilson MD, Spyrou C, Brown GD, Hadfield J, Odom DT. 2009. ChIP-seq: Using high-throughput sequencing to discover protein-DNA interactions. *Methods* **48:** 240–248.
Schmidt D, Schwalie PC, Ross-Innes CS, Hurtado A, Brown GD, Carroll JS, Flicek P, Odom DT. 2010a. A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res* **20:** 578–588.
Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010b. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328:** 1036–1040.
Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent transcription from active promoters. *Science* **322:** 1849–1851.
Seitan VC, Hao B, Tachibana-Konwalski K, Lavagnolli T, Mira-Bontenbal H, Brown KE, Teng G, Carroll T, Terry A, Horan K, et al. 2011. A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. *Nature* **476:** 467–471.
Stedman W, Kang H, Lin S, Kissil JL, Bartolomei MS, Lieberman PM. 2008. Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *EMBO J* **27:** 654–666.
Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci* **101:** 6062–6067.
Sumara I, Vorlaufer E, Gieffers C, Peters BH, Peters JM. 2000. Characterization of vertebrate cohesin complexes and their regulation in prophase. *J Cell Biol* **151:** 749–762.
Turro E, Su S-Y, Gonçalves Â, Coin LJ, Richardson S, Lewin A. 2011. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol* **12:** R13. doi: 10.1186/gb-2011-12-2-r13.
Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5:** 829–834.
Vega H, Waisfisz Q, Gordillo M, Sakai N, Yanagihara I, Yamada M, van Gosliga D, Kayserili H, Xu C, Ozono K, et al. 2005. Roberts syndrome is caused by mutations in *ESCO2*, a human homolog of yeast *ECO1* that is essential for the establishment of sister chromatid cohesion. *Nat Genet* **37:** 468–470.
Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, et al. 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451:** 796–801.
Wilczyński B, Furlong EE. 2010. Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol Syst Biol* **6:** 383. doi: 10.1038/msb.2010.35.
Xu H, Balakrishnan K, Malaterre J, Beasley M, Yan Y, Essers J, Appeldoorn E, Tomaszewski JM, Vazquez M, Verschoor S, et al. 2010. *Rad21*-cohesin haploinsufficiency impedes DNA repair and enhances gastrointestinal radiosensitivity in mice. *PLoS ONE* **5:** e12112. doi: 10.1271/journal.pone0012112.
Yoon YS, Jeong S, Rong Q, Park K-Y, Chung JH, Pfeifer K. 2007. Analysis of the *H19ICR* insulator. *Mol Cell Biol* **27:** 3499–3510.
Zhang B, Jain S, Song H, Fu M, Heuckeroth RO, Erlich JM, Jay PY, Milbrandt J. 2007. Mice lacking sister chromatid cohesion protein PDS5B exhibit developmental abnormalities reminiscent of Cornelia de Lange syndrome. *Development* **134:** 3191–3201.
Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE. 2009. Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature* **462:** 65–70.

Received December 16, 2011; accepted in revised form July 9, 2012.