Cell Reports
Article

Cell PRESS
Open ACCESS

A DNA-Centric Protein Interaction Map of Ultraconserved Elements Reveals Contribution of Transcription Factor Binding Hubs to Conservation

Tar Viturawong,[1] Felix Meissner,[1] Falk Butter,[1,*] and Matthias Mann[1,*]
[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany
*Correspondence: butter@biochem.mpg.de (F.B.), mmann@biochem.mpg.de (M.M.)

http://dx.doi.org/10.1016/j.celrep.2013.09.022

SUMMARY

Ultraconserved elements (UCEs) have been the subject of great interest because of their extreme sequence identity and their seemingly cryptic and largely uncharacterized functions. Although in vivo studies of UCE sequences have demonstrated regulatory activity, protein interactors at UCEs have not been systematically identified. Here, we combined high-throughput affinity purification, high-resolution mass spectrometry, and SILAC quantification to map intrinsic protein interactions for 193 UCE sequences. The interactome contains over 400 proteins, including transcription factors with known developmental roles. We demonstrate based on our data that UCEs consist of strongly conserved overlapping binding sites. We also generated a fine-resolution interactome of a UCE, confirming the hub-like nature of the element. The intrinsic interactions mapped here are reflected in open chromatin, as indicated by comparison with existing ChIP data. Our study argues for a strong contribution of protein-DNA interactions to UCE conservation and provides a basis for further functional characterization of UCEs.

INTRODUCTION

Transcriptional regulation is determined by complex interactions of DNA, transcription factors (TFs), and chromatin states. Transcriptional regulatory elements capable of modulating gene expression have been of much interest due to their role in development and disease (Spitz and Furlong, 2012; Williamson et al., 2011). Conservation analysis, chromatin-modification state analysis, and in vivo reporter assays have been used to identify several hundreds of such transcriptional enhancers (Berman et al., 2004; Pennacchio et al., 2006; Visel et al., 2007). Among these, ultraconserved elements (UCEs)—DNA elements defined by their 100% sequence identity over 200 bp between human and mouse genomes—have been identified as tissue- and stage-specific enhancers (Bejerano et al., 2004; Pennacchio et al., 2006; Visel et al., 2007). UCE sequences were predicted to be enriched in binding sites for development-associated TFs, suggesting important developmental regulatory roles. However, relatively few phenotypic alterations have been associated with the loss or mutation of UCEs (Martínez et al., 2010; Poitras et al., 2010; Yang et al., 2008), and whereas several hypotheses have been proposed (Siepel et al., 2005), little has been attempted experimentally to account for the ultraconservation of these loci. Similarly, although the regulatory potential of UCEs has been demonstrated through embryonic reporter assays, the function and mechanism of these regulatory elements largely remain to be explored.

One starting point to enhancer characterization is through interactor mapping. Recently, chromatin immunoprecipitation (ChIP) has mapped out interaction of the genome to several TFs in great detail (Bernstein et al., 2012). ChIP is protein centric, i.e., it maps out target DNA sequences bound to prechosen TFs, limiting the diversity of interaction profiles to a priori knowledge. Furthermore, ChIP data reflect an endpoint of gene regulation, incorporating aspects such as chromatin homeostasis and long-range interactions, rendering the contribution of the underlying DNA sequence difficult to determine. Evidence from a small number of genomic loci as well as whole-chromosome analysis has demonstrated the genetic contribution to the establishment of epigenetic states (Arbab et al., 2013; Wilson et al., 2008). Thus, DNA-centric study of intrinsic interactions between DNA sequences and DNA binding nuclear proteins in the absence of initial epigenetic priming is valuable to understanding the genetic contribution to transcriptional regulation, which is especially important for dissecting per-nucleotide conservation of UCEs.

Past studies have employed a DNA-centric approach to identify potential binders of small numbers of DNA sequences (Butter et al., 2012; Déjardin and Kingston, 2009; Mirzaei et al., 2013; Mittler et al., 2009; Tacheny et al., 2012). Here, we have developed a high-throughput platform to screen unbiased interaction profiles for hundreds of DNA sequences, based on our previously described pull-down method using high-resolution mass spectrometry (MS) and SILAC quantitation (Mittler et al., 2009). We applied this technology to obtain an interaction map for 193 UCEs, including over half of all nonexonic (nx) UCEs in the

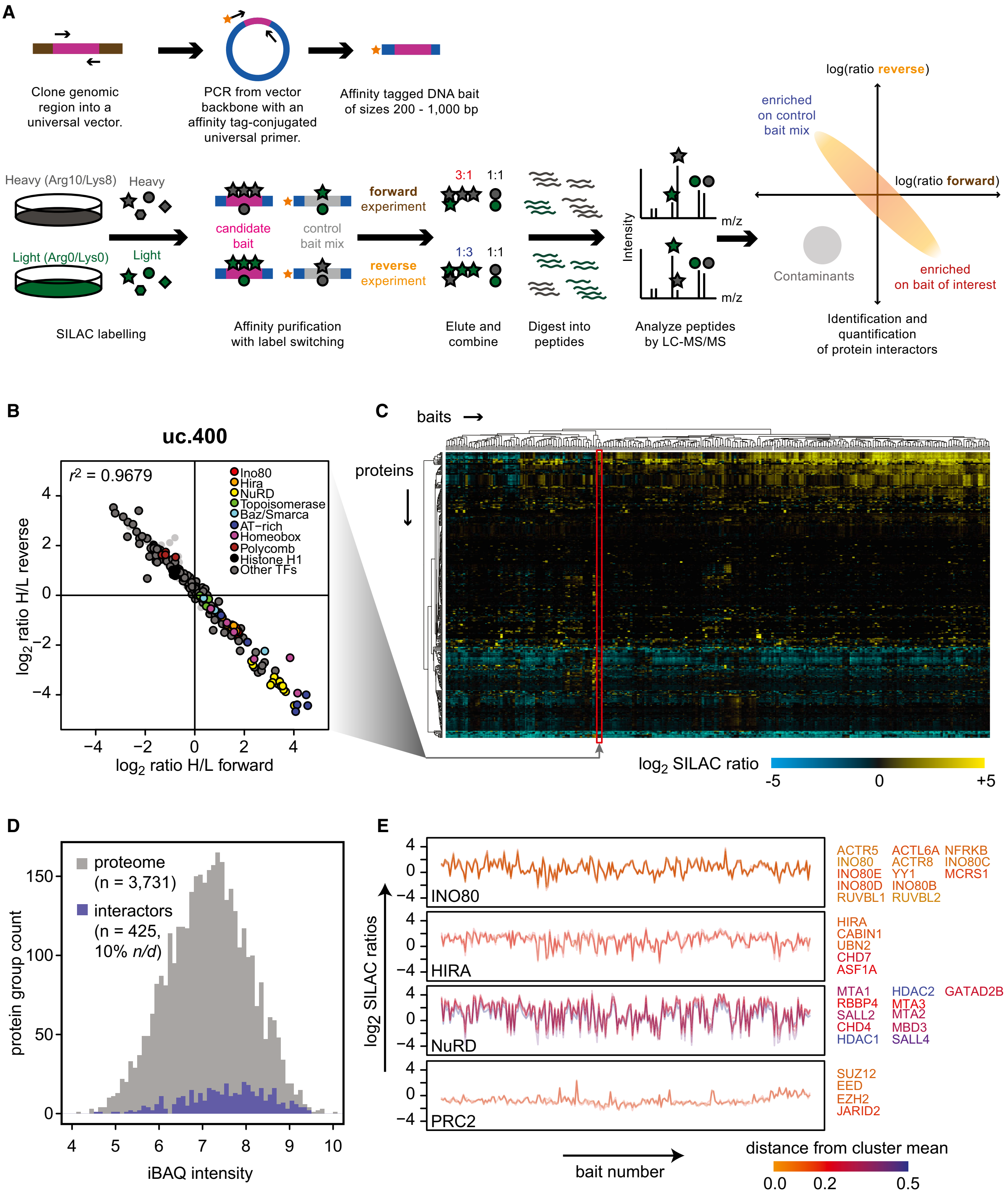

Figure 1. Overview of SILAC Affinity Purification for Protein-DNA Interaction Screen for UCE Sequences

(A) Scheme of bait generation and pull-down pipeline. Genomic loci of interest spanning between 200 and 1,000 bp were amplified using specific PCR primers and cloned into a universal vector. DNA baits were then generated by PCR amplification using affinity-tagged primers binding to the flanking sequence on the universal vector backbone. For each locus, we purified SILAC-labeled nuclear extracts on the DNA baits in two experiments: In the forward experiment, the heavy

(legend continued on next page)

genome. We found nx UCE sequences to bind TFs and chromatin remodelers with known roles in developmental regulation, whereas proteins that promote chromatin compaction were relatively depleted. We inferred that the protein interactors bind to UCE sequences through densely distributed and often overlapping canonical transcription factor binding sites (TFBSs). Individual DNA bases that are part of overlapping TFBSs were on average more stringently conserved among vertebrates. We also obtained mapped intrinsic interactions of one UCE to five-nucleotide resolution and found a high frequency of both gain and loss of binding to occur upon mutation. Finally, comparison of our intrinsic interaction map with existing ChIP-seq data as well as reporter assays linking previous independent observations (Palmer et al., 2007; Thompson et al., 2007) highlight the functional relevance of these interactions. Overall, our interaction map points toward extremely high information content and complex transcription regulation logic behind many UCEs.

RESULTS

The UCE Interactome

We obtained the interaction map for 129 of 256 nx, 36 of 114 putative-exonic (px), 28 of 111 exonic (ex) UCEs, as well as 21 human and 3 mouse random genomic loci by affinity purification, high-resolution MS, and SILAC quantitation in high throughput (Eberl et al., 2013). We used topoisomerase-assisted cloning to insert bait sequences amplified from human or mouse genomic DNA into a universal vector backbone. This backbone enabled us to amplify the baits by parallel PCR, where one primer was labeled with desthiobiotin to allow streptavidin capture and specific elution of protein-DNA complexes (Figure 1A). Our interaction map was generated in the context of the R1/E mouse embryonic stem cell line, in keeping with the proposed relevance of UCEs in gene regulation during development, and exploiting the sequence identity of UCEs between mouse and human genomes.

We performed two experiments for each DNA bait of interest. In one set of pull-downs (called "forward"), we incubated heavy-labeled nuclear extracts with the UCE bait, and unlabeled extracts with the mix of 24 random genomic sequences, to dilute out any binding sites arising by chance. SILAC enabled us to accurately quantify the enrichment of interactors of DNA bait over control (Ong et al., 2002). In the "reverse" pull-downs, we switched the SILAC labels with respect to the baits, enabling two-dimensional separation of true interactors from false positives (Butter et al., 2012) (Figure 1A).

Our screen identified a total of 1,709 proteins across the entire interactome, with an average of 870 proteins per MS run. Of these, 223 (13%) were quantified on all UCE baits, and 660 (39%) were quantified in at least half of the baits. We found 425 proteins with enrichment ratio greater than 1.4 for at least three baits (Figure 1C; Table S1). These proteins represented 10.3% of the R1/E nuclear proteome, which we measured for comparison, and showed a slight bias of 2.8-fold toward high-abundance proteins over the 10,000-fold abundance range ($p < 10^{-16}$; Figure 1D)—arguing that endogenous proteins of most expression levels were accessible from our screen. There was excellent reproducibility of SILAC ratios between the forward and reverse pull-downs (Figures 1B and S1; median SILAC ratio $r^2 = 0.91$). Binding profiles of members belonging to the same complex were extremely tightly correlated (Figure 1E), indicating that the proteins bound to the baits as complexes and providing further positive control. In sum, we have generated an unbiased intrinsic protein interactome for UCE sequences that preserves cell-specific protein-protein interactions and takes into account the cell's nuclear context.

Interactors of Nx UCEs Are Biased for Development and Chromatin-Access Functions

Previous in silico sequence analysis of UCEs proposed a role of transcriptional regulatory "hubs" that recruit developmentally functional TFs (Siepel et al., 2005). Our UCE interactome showed that nx UCE sequences (nxUCE) were more enriched in interactors regardless of SILAC ratio threshold used for interactor calling, followed by possibly exonic (pxUCE), exonic (exUCE), and random genomic sequences (Figure 2A). Annotation enrichment analysis based on SILAC ratios identified Gene Ontology (GO) terms containing the annotations neural, nerve, forebrain, hindbrain, limb, and axis as significant classifications for UCE interactors (Figure 2B). Domain enrichment analysis based on Pfam showed that homeobox TFs were most significantly enriched at nxUCEs ($p < 10^{-31}$) and, to a lesser extent, at pxUCEs ($p < 10^{-12}$) and exUCEs ($p < 0.01$) (Figure 2C). Interestingly, we also found enrichment of leucine zipper family TFs at nxUCEs ($p < 10^{-4}$), a finding not previously predicted from motif analysis based on the JASPAR TF binding motif database (Figure S2).

nuclear proteins were purified with bait of interest and the light nuclear proteins with a mix of random genomic control baits. In the reverse experiments, labels were switched. Eluted proteins from each experiment were combined and digested into peptides. Peptides were then separated and analyzed by liquid chromatography coupled to high-resolution MS. Specific interactors of the bait of interest are expected to have high (>1) forward and low (<1) reverse ratios, whereas background binders are expected to have ratios around 1:1. Interactors relatively de-enriched at the bait of interested have low (<1) forward ratio and high (>1) reverse ratios. Only false positives appear in the upper-right quadrant.

(B) Forward-reverse scatterplot of SILAC ratios for uc400 interactors. See (A) for interpretation of the plot. Colored points indicate proteins belonging to annotated complexes or protein families. H/L, high/low. See also Figure S1.

(C) Summary interaction map of all 216 DNA baits and 425 interactors. Color bar indicates SILAC ratios of proteins bound to each bait over the mix of 24 random genomic loci. Missing values were filled with zeros for visualization only, and not for any analysis.

(D) Distribution of iBAQ intensity (a measure of protein abundance; Schwanhäusser et al., 2011) for R1/E nuclear proteins and for the proteins identified as interactors in our screen. *n/d* denotes the proportion of the 425 identified interactors that were not identified in the nuclear proteome.

(E) SILAC ratios of members of complexes inferred ab initio from highly correlated interaction profiles. Each color trace represents one protein. Trace colors indicate mean absolute SILAC ratio difference from the average profile of the complex. The protein names are given to the right of each complex profile, colored as in the traces.

See also Figure S1 and Table S1.

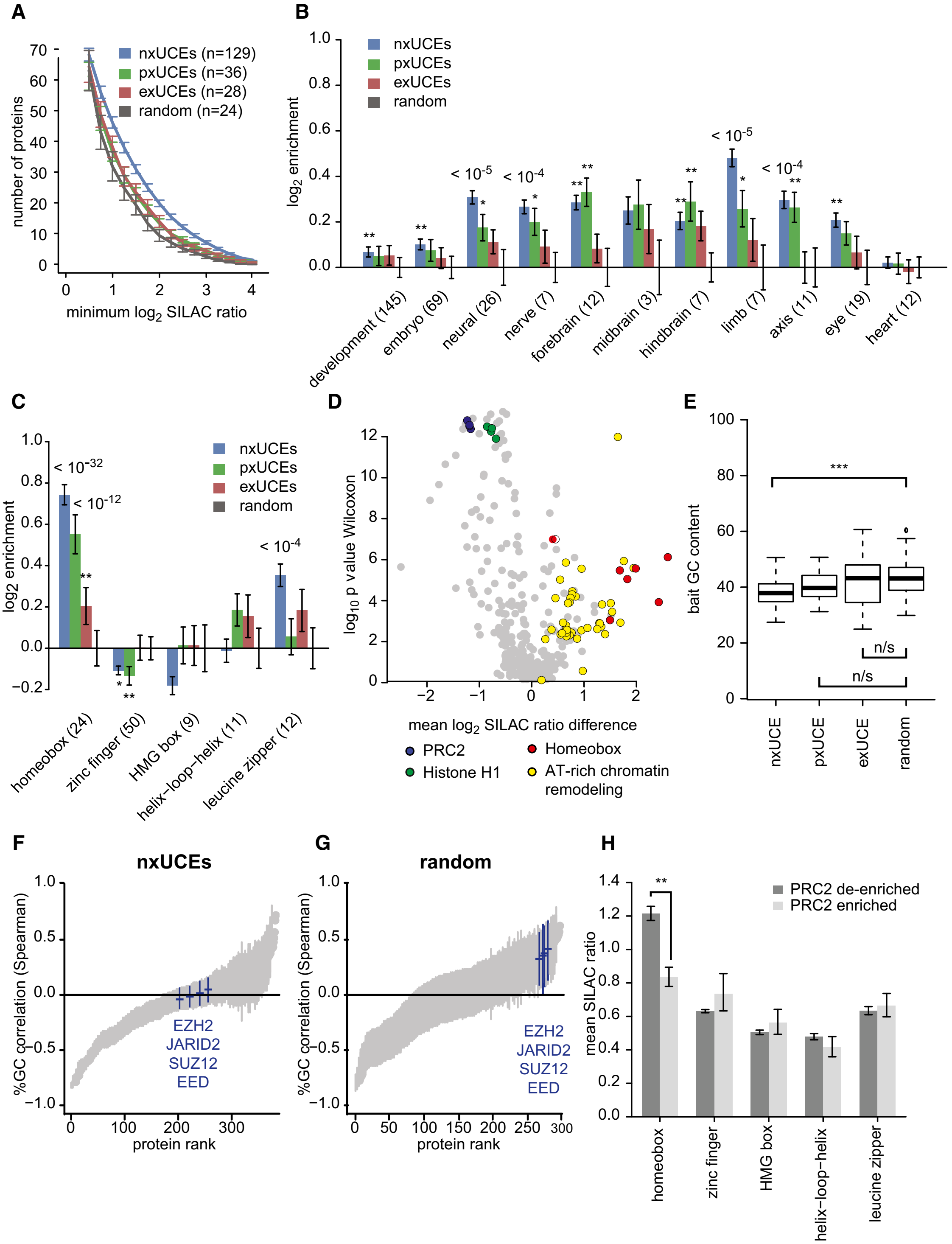

Figure 2. UCE Interactome Shows TF Hub Characteristics

(A) Number of proteins quantified with SILAC ratios greater than those indicated on the x axis, summarized for nxUCEs, pxUCEs, exUCEs, and random genomic baits (mean ± SEM).

(B) Enrichment of proteins containing GO terms indicated on the x axis for nxUCEs, pxUCEs, and exUCEs compared to random genomic baits (mean ± SEM). The numbers of proteins containing indicated GO words are given in parentheses. Significance is indicated: $^*p < 0.05$ and $^{**}p < 0.01$. More significant p values are displayed explicitly.

(C) Enrichment of proteins belonging to the indicated TF classes based on our SILAC pull-down data set (mean ± SEM). Significance is indicated: $^*p < 0.05$ and $^{**}p < 0.01$. More significant p values are displayed explicitly.

(legend continued on next page)

The TF binding hub proposal demands that the chromatin be accessible for function. Intrinsic open chromatin propensity for UCE sequences could be expected owing to AT richness predicted to result in poor nucleosome occupancy (Tillo and Hughes, 2009). Indeed, in addition to homeobox TFs, nxUCEs also favored binding of several chromatin remodelers and other AT-rich factors including the INO80, NuRD, HIRA, and SMARCA/BAZ complexes as well as DNA topoisomerases (Figure 2D). Many of the chromatin remodelers observed in our interactome possess nucleosome shifting or destabilization activity (Aalfs et al., 2001; Jin and Felsenfeld, 2007; Rai et al., 2011; Udugama et al., 2011; Xie et al., 2012). Importantly, although nxUCEs are slightly more AT rich than random genomic loci (median GC content 37.9% and 43.1%, respectively, Figure 2E), preferential enrichment of nxUCEs for AT-rich binders including homeoboxes generally held significant even when we binned our baits by comparable GC content (Figure S2), indicating that the observed enrichment cannot be explained solely by sequence nucleotide composition.

To further explore possible manifestation of intrinsic open chromatin propensity, we investigated the binding of histone H1 and the PRC2, proteins known to promote heterochromatin formation (Cao et al., 2002; Lu et al., 2009; Thoma and Koller, 1977). Indeed, nxUCEs were relatively depleted in histone H1 and PRC2 ($p < 10^{-12}$, Figure 2D), and this effect was equally strong in pxUCEs and exUCEs ($p < 10^{-14}$ and $p < 10^{-6}$, respectively; see Figure S2). PRC2 binding is known to depend partially on TFBS density, with absence of TFBS allowing PRC2 to bind to GC-rich regions (Mendenhall et al., 2010). Strikingly, we found that PRC2 members were among the interactors with strongest GC preference, but only if random genomic sequences were considered on their own. At nxUCE sequences where interactions were more prevailing, the binding of PRC2 showed no GC preference at all ($p < 0.05$ for SUZ12, $p < 0.01$ for EZH2, EED and JARID2; see also Figures 2F, 2G, and S2), indicating that a different rule than GC content governs binding of PRC2 to nxUCE sequences. Furthermore, we found that the homeobox class of interactors—the class most enriched for nxUCEs—is significantly depleted at PRC2-enriched nxUCE baits over PRC2-de-enriched nxUCE baits ($p < 0.01$; Figure 2H). The differential enrichment became even more significant when the comparison was extended to all the baits ($p < 10^{-5}$). These results demonstrate the inverse relationship between binding of TFs and binding of PRC2 in the context of UCE sequences and suggest that nxUCE sequences may avoid heterochromatinization in part by exclusion of PRC2 owing to a large population of interactors.

In conclusion, we have shown that nxUCEs are not only enriched in developmentally relevant TFs but are also enriched in chromatin-destabilization proteins as well as relatively devoid of heterochromatin-promoting proteins. These observations illustrate the inherent biochemical properties of nxUCE sequences appropriate to serve as TF binding hubs.

UCEs Are Strongly Enriched in Overlapping TFBSs with Conservation Bias in Overlapped Sites

One proposed explanation for ultraconservation of UCEs is that of high density of functional TFBSs providing multiple constraints accounting for higher evolutionary pressure. High density of TFBSs could result in information compression in the form of overlapping TFBSs, a concept that has been postulated for UCEs and indeed observed in several other instances (Hermsen et al., 2006; Ngondo-Mbongo et al., 2013; Siepel et al., 2005). Our data set provided an opportunity to address the multiple-constraint hypothesis directly.

We first used our quantitative UCE interactome to derive binding motifs that are directly relevant to UCEs. We tested for association between differential interactor enrichment and all possible motifs up to eight nucleotides in length and found that 439 motifs associated with enrichment of 161 interactors at 5% false discovery rate (FDR). These included a large number of homeobox, E box, and leucine zipper, and several other motifs, as well as a number of putative motifs for several factors (see Experimental Procedures). We also correctly found very short motifs for a number of factors. For instance, we identified the CpG dinucleotide as a binding motif for KDM2B ($p < 10^{-17}$), a H3K36 demethylase known to bind to unmethylated CpG at c-*jun* promoter through its CxxC zinc finger (Koyama-Nasu et al., 2007). Binding of TFAP2 can be described by the presence of a single-nucleotide motif "G," reflecting the GC content as the major influence on the interaction. As a measurement of validity of our motif enrichment, Table 1 compares some of the most significant motifs rediscovered ab initio from our data set to the corresponding known motifs. See also Table S2 for the full enumeration of motifs.

To test the overlapping TFBS hypothesis and its relevance for ultraconservation, we mapped the derived motifs to UCE sequences and other sequences and then compared motif distribution as well as conservation of unmapped bases, singly mapped bases, and repeatedly mapped (superimposed) bases (Figure 3A; see also Experimental Procedures). To allow an exhaustive analysis, we included all 481 UCEs, 720 additional enhancers available from the VISTA database of in vivo enhancer activity of conserved genomic loci (Visel et al., 2007) classed by whether they contain UCEs (ucVISTA) or not (ncVISTA), and 791 randomly picked genomic regions.

We found nxUCEs to be most highly enriched for motif superimposition over random genomic loci ($p < 10^{-48}$), followed by pxUCEs ($p < 10^{-11}$; Figure 3B), but not exUCEs. Similarly, ucVISTA sequences were more enriched for superimposition over

(D) Volcano plots of interactors enriched in nxUCE compared against random genomic loci. Enrichment significance (Wilcoxon rank sum test) is plotted against the mean enrichment. Colored points indicate proteins belonging to the annotated complexes or groups of proteins.
(E) GC content of nxUCE, pxUCE, exUCE, and random genomic baits.
(F and G) Spearman rank correlation coefficients of SILAC ratio with bait GC content for each interactor given the context of nxUCEs or random genomic loci (estimate ± 95% CI). Only interactors with at least 20 quantifications were considered. Colored bars indicate proteins belonging to the PRC2. See also Figure S2.
(H) SILAC ratio of TF classes for PRC2-enriched and PRC2-de-enriched nxUCE baits (mean ± SEM).
See also Figure S2.

Table 1. Comparison of Best Motifs Rediscovered Ab Initio from the UCE Interactome with Known Motifs

	Motif		p Value	Benjamini-Hochberg	Reference
Interactor	This Study	Literature	Wilcoxon	q Value	(PMID or JASPAR)
Zfp384	TTTTTT	SAAAAA(A)	3.262×10^{-23}	3.624×10^{-17}	10669742
Kdm2b	CG	CG	3.500×10^{-18}	6.667×10^{-13}	20417597
Zfp281	GGGGG	TCCCCCCCCCCCCCC/AGGAGACCCCCAATTTG	4.578×10^{-18}	8.571×10^{-13}	JASPAR
Vax2	TAATTA	GTGCACTAATTAAGAC	5.190×10^{-18}	9.556×10^{-13}	JASPAR
Arid3a	TTAAT	GGGTTTAATTAAAATTC	7.330×10^{-18}	1.285×10^{-12}	JASPAR
Arid5b	AT	CTAATATTGCTAAA	4.811×10^{-17}	5.479×10^{-12}	JASPAR
Nanog	TAAT	TAATKK	7.385×10^{-14}	3.343×10^{-9}	12787504
Pou2f3	ATTTGCAT	TTGTATGCAAATTAGA	1.822×10^{-11}	4.502×10^{-7}	JASPAR
Klf2	GGGCG	GGGCG	1.972×10^{-10}	3.612×10^{-6}	19843526; 15774581
Klf4	GGGCG	GGGCG	1.972×10^{-10}	3.612×10^{-6}	19843526; 15774581
Tfeb	CATGTG	CANNTG/GTCACGTGAC	7.551×10^{-10}	1.151×10^{-5}	9806910; 16936731
Atf1	GTCAT	ACGATGACGTCATCGA	8.588×10^{-10}	1.284×10^{-5}	JASPAR
Otx2	TAATCC	TGTAGGGATTAATTGTC	1.837×10^{-9}	2.442×10^{-5}	JASPAR
Creb1	GTCAT	GTCAT	2.086×10^{-9}	2.703×10^{-5}	8458331
Hoxc12	ATT	TTAGGTCGTAAAATTTC	3.022×10^{-9}	3.794×10^{-5}	JASPAR
Sox2	ATTGTT	CCTTTGTTATGCAAA	3.197×10^{-8}	3.066×10^{-4}	JASPAR
Zfx	GGCCT	GGGGCCGAGGCCTG	2.932×10^{-7}	2.015×10^{-3}	JASPAR
Sp3	TCCTCCC	GGTCCCGCCCCCTTCTC	4.663×10^{-7}	3.005×10^{-3}	JASPAR
Tcf7l1	CTTTGAT	ATTTCCTTTGATCTATA/GAAGATCAATCACTAA	5.921×10^{-7}	3.695×10^{-3}	JASPAR
Jund	GTCAT	CCGATGACGTCATCGT	8.441×10^{-7}	4.942×10^{-3}	JASPAR
Eed	TCGG	TCG	2.110×10^{-6}	1.031×10^{-2}	14602076
Ezh2	TCGG	TCG	2.110×10^{-6}	1.031×10^{-2}	14602076
Gtf2ird1	TTAATCT	GATTA	2.210×10^{-6}	1.069×10^{-2}	17346708
Tfap2a	G	GCCCGGGGG	2.315×10^{-6}	1.113×10^{-2}	JASPAR
Cdx1	TTAATT	TAAGGTAATAAAATTA	2.602×10^{-6}	1.223×10^{-2}	JASPAR
Tfap2c	G	ATTGCCTGAGGCGAA/CCGCCCAAGGGCAG	4.085×10^{-6}	1.700×10^{-2}	JASPAR
Lhx2	TAATTAGT	TAAACTAATTAGTGAAC	1.051×10^{-5}	3.499×10^{-2}	JASPAR
Tcf7	CTTTGAT	TATAGATCAAAGGAAAA/CCGTATTATAAACAA	3.079×10^{-5}	7.932×10^{-2}	JASPAR
Six4	TGAGATC	TGATAC	3.381×10^{-5}	8.495×10^{-2}	JASPAR
Nfya	GGCCAAT	CTCAGCCAATCAGCGC	3.553×10^{-5}	8.827×10^{-2}	JASPAR
Usf1	TCACATG	CACGTGG	7.729×10^{-5}	1.549×10^{-1}	JASPAR

See the Results for an interpretation of very short (less than four nucleotide) motifs.
See also Table S2 for full enumeration of motifs.

random genomic loci than ncVISTA ($p < 10^{-25}$ and $p < 10^{-16}$, respectively) but less enriched than nxUCEs, consistent with UCEs being the most conserved core of ucVISTA enhancers. No superimposition enrichment was observed when we instead used nonenriched motifs taken randomly from the UCEs. Our finding that superimposition degree increases from ncVISTA to ucVISTA and finally nxUCE and that exUCEs did not show such enrichment, indicate that nxUCEs represent the extreme case of overlapping TFBSs.

To exclude the possibility that AT richness is solely responsible for the increased motif superimposition at nxUCEs, we shuffled the nucleotides in all sequences used for superimposition analysis to generate synthetic sequences of equivalent GC content. Superimposition enrichment on these sequences was severely abrogated (Figure 3C), indicating that AT richness contributes to but is in itself insufficient to achieve the extent of superimposition observed with nxUCEs by chance. To support this in silico finding, we performed pull-downs on random, highly heterogeneous DNA sequences with average GC content of 20% or 40%. Our experiment showed that only some of the proteins that bound preferentially to UCEs also bound preferentially to the synthetic AT-rich bait population (Figure S3). Generally, there was insignificant correlation between factor preference for AT-rich sequences and enrichment at nxUCEs (Spearman's $\rho = 0.05$; $p > 0.1$). Notably, factors bound to synthetic GC-rich bait populations were also enriched at nxUCEs, ruling out AT richness as the sole explanation for motif occurrence and thus superimposition at nxUCEs. Together with the inherent conservation bias for GC nucleotides over AT nucleotides in UCEs (Figure S3), we speculate that GC-rich TFBSs may be under greater selective pressure in AT-rich UCEs in order to preserve certain regulatory function.

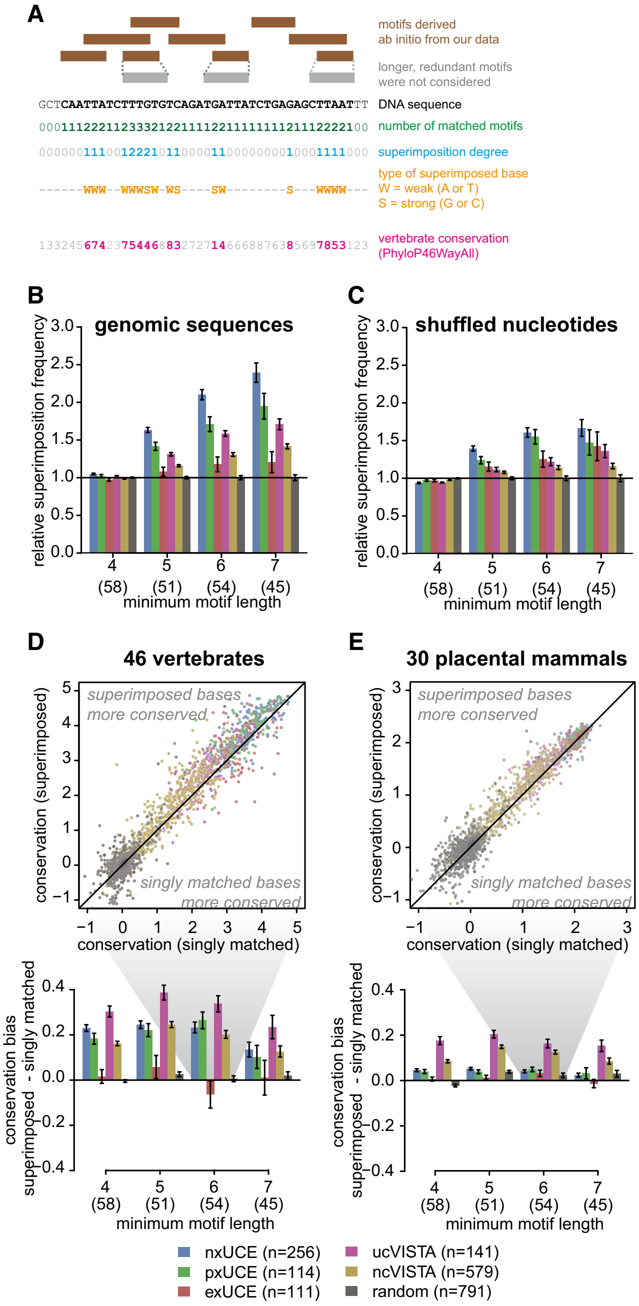

Figure 3. nxUCEs Are Extremely Concentrated in Overlapping, Ultraconserved TFBSs

(A) Outline of superimposition analysis. Motifs derived ab initio from our analysis were mapped back onto DNA sequences. First, a minimum motif length was decided, and longer motifs containing an existing shorter motif were discarded from mapping to avoid counting redundant motifs. The number of motifs mapped onto each base was counted, and bases were then classed as unmatched, singly matched, or superimposed. Frequencies of each base class and their conservation were then compared. See Experimental Procedures for more information.

(B) Relative fraction of superimposed bases given indicated minimum motif length for UCEs and VISTA enhancers, normalized to random genomic loci (mean ± SEM). Numbers in parentheses indicate the number of motifs considered given each minimum motif length. See text for p values. See also Figure S3.

(C) Relative fraction of superimposed bases as with (B) nucleotide-shuffled versions of UCEs and VISTA enhancers (mean ± SEM). Numbers in parentheses indicate the number of motifs considered given each minimum motif length. See text for p values.

(D and E) Conservation bias of superimposed bases over singly matched bases for [AT] bases calculated over 46 vertebrates or 30 placental mammals (mean ± SEM). Outset: Absolute mean conservation scores of singly matched and superimposed bases given a minimum motif length of six nucleotides. See text for p values.

See also Figure S3.

If superimposition of TFBSs also played important biological roles, we would expect DNA bases involved in superimposition to be more deeply conserved. We therefore investigated the extent of DNA base conservation in 46 vertebrates, using an established conservation-scoring scheme (Meyer et al., 2013). For sequences that were putative enhancers, the bases matched by multiple motifs were on average slightly but significantly more conserved than bases mapped only to a single motif ($p < 0.001$). Strikingly, this conservation bias became massively amplified when only AT bases were considered ($p < 10^{-10}$; Figure 3D), consistent with the presence of many AT-rich motifs derived from our data. Conservation bias was also observed in ucVISTA and ncVISTA sequences, concordant with functional overlapping TFBSs reported for loci other than UCEs. The larger difference in VISTA enhancers compared to UCEs can be attributed to the lower conservation baseline for ncVISTA enhancers (Figure 3D). We also found the conservation bias to be reduced when the scoring was restricted to placental mammals (Figure 3E), suggesting early origins of these overlapped sites. In conclusion, we have shown that nxUCEs represent the extreme case of overlapping, deeply conserved, biochemically functional TFBSs among enhancers.

UCE Scanning Mutagenesis Defines Protein Binding Characteristics and Correlates Gain of Interaction with Nucleotide Conservation

Although an implication of the multiple-constraint hypothesis is that mutation of nxUCEs causes deleterious consequences, it has been difficult to identify the exact systems that are affected. However, the conservation bias implies that the multiple-constraint hypothesis would at least manifest itself in terms of change in protein binding capacity, which in turn could result in regulatory logic alteration at UCEs.

In order to test this hypothesis, we performed a scanning mutagenesis of uc325, a nx UCE that is part of a midbrain/eye development enhancer (Visel et al., 2007). Each nonoverlapping five-nucleotide window of uc325 was mutated transitionally—the most frequent mode of nucleotide substitution in vivo (Collins and Jukes, 1994). Pull-down was performed on the resultant series of baits against the wild-type bait (Figure 4A), and interactors were defined as proteins whose SILAC ratios were in the most extreme 5% of all quantified ratios. We discovered 55 interactors for the uc325 set but only 10 for the control set based on a random genomic sequence with comparable GC content. Both gain and loss of interactions were found for uc325, covering the entire span of the bait (Figure 4C). Most of the prominent interaction losses were found in contiguous variants—reflecting binding sites that span more than five nucleotides—whereas

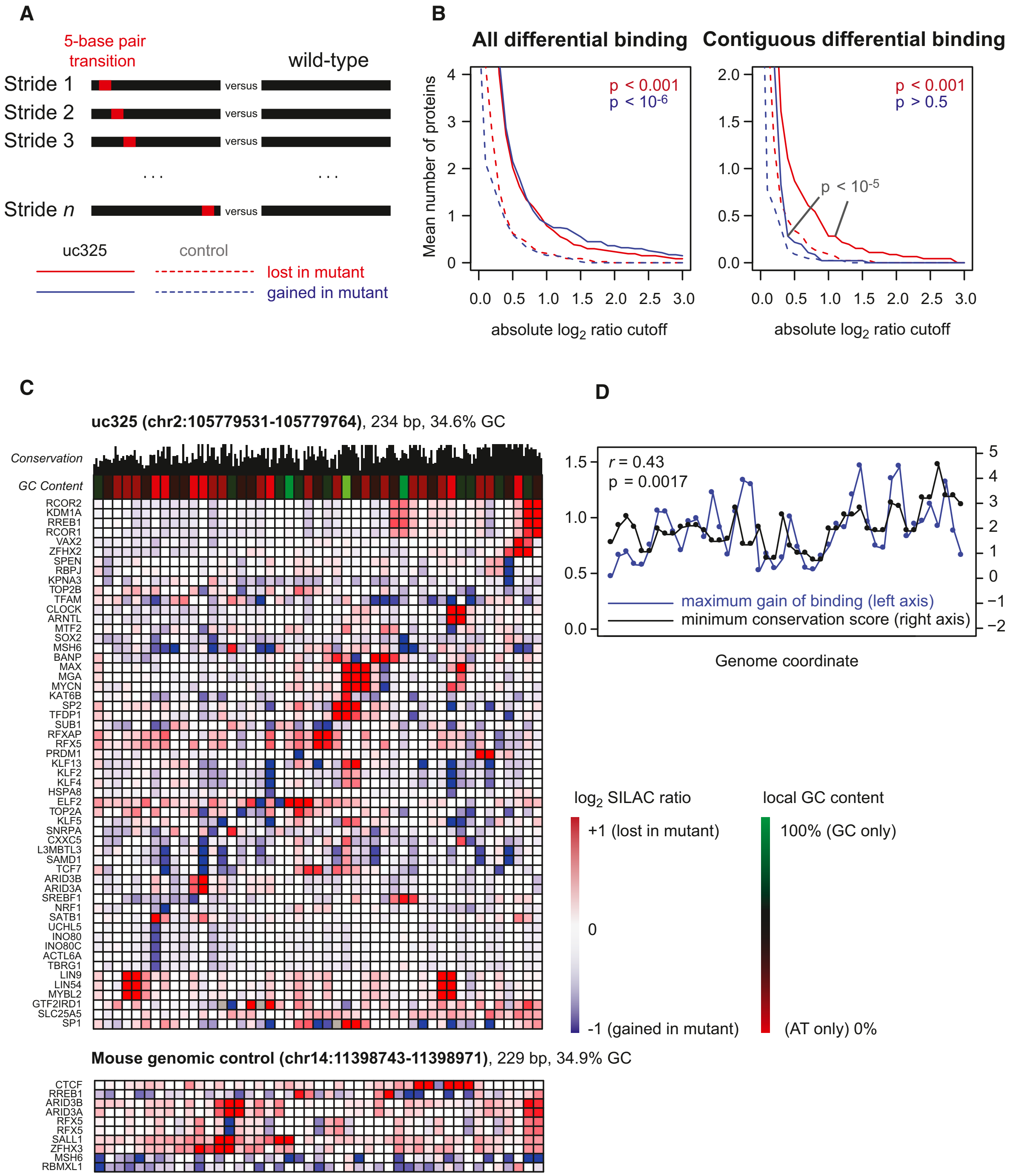

Figure 4. A Fine Intrinsic Interaction Map of uc325

(A) Nonoverlapping five nucleotide windows spanning the 234 bases of uc325 were mutated transitionally (A ↔ G, C ↔ T) and interactors compared to the wild-type in a series of SILAC pull-downs. The same was done for a 229-base random genomic sequence with comparable GC content.

(B) Proportion of variant strides that show either loss or gain of binding of at least one protein owing to the mutation for uc325 and control, given as a function of cutoff ratio used for calling gain/loss. Contiguous differential binding refers to differential binding that spans at least two strides.

(C) Enrichment of proteins where complete pairwise quantification was achieved for at least 80% of all strides and where at least one stride showed localized differential enrichment with magnitude exceeding 95% of all quantified ratios. This 95% cutoff corresponded to log2 ratio of 0.59 for uc325 and 0.83 for the control. Interactors were ordered by the location of a prominent differential binding (magnitude exceeding 99% of ratios across all strides). Genome coordinates are based on the mm10 build. Conservation is based on 60-way vertebrate comparison (Meyer et al., 2013).

(D) Comparison of maximum magnitude of binding gain/loss of each five-nucleotide blocks to the minimum nucleotide conservation smoothened over two successive blocks, giving an effective resolution of ten nucleotides.

interaction gains tend to appear stochastically ($p < 10^{-5}$, Kolmogorov-Smirnov test, Figure 4B). In contrast, only a small region in the control bait appeared to contain prominent interactors (Figures 4B and 4C). These data indicate that uc325 indeed possesses a hub-like characteristic with numerous and diverse TFBSs as well as latent sites that could be reached within a few transition mutations.

We next investigated whether any relationship exists between uc325 conservation and its scanning mutant interactome. Initially, we had expected the conservation to be correlated to the loss of binding owing to transition mutation, but this turned out not to be the case ($p > 0.5$). Surprisingly, we found that conservation of uc325 strides to be significantly correlated with the maximum binding gain owing to mutation ($p = 0.0017$; Figure 4D), whereas such correlation was weaker for the control ($p = 0.027$). When at least two noncorrelating proteins were required to be enriched in the mutant, the correlation with conservation remained significant for uc325 ($p = 0.0020$), but not for the control ($p = 0.12$). Interestingly, AT-rich strides tended to give more drastic binding gain upon mutation (correlation with GC content, -0.36; $p = 0.0062$; Figure 4C). We speculate that these AT-rich strides are under selective pressure against developing such TFBSs, which could alter the regulatory logic of the UCE. Alternatively, an apparent strong gain of binding could be observed if the mutation turned a promiscuous binding site capable of binding several factors weakly into a well-defined, specialized binding site, thereby destroying the "hub" characteristics that may be required for fine-tuned regulatory function.

Regulatory Consequence of the UCE Interactome

Evidence for the regulatory consequence of UCE interactors could be obtained from perturbation experiments and reporter assays. Although it may be difficult to discern the regulatory logic of such complex enhancers without performing very deep perturbation, it should still be possible to address the functionality of certain interactions given existing biological knowledge. To demonstrate such a case, we investigated the functionality of the interaction between uc400 and the protein GTF2IRD1.

The 860 bp genomic region containing uc400 possesses forebrain-specific enhancer activity during embryonic day 11.5 (E11.5) (Pennacchio et al., 2006). We found that uc400 interacts specifically with the Williams Beuren syndrome protein GTF2IRD1 with a SILAC ratio of around 6:1 in R1/E cells and also with hGTF2IRD1 in HeLa cells (Figures 5A and S4). GTF2IRD1 is known to act as a repressor via its interaction with the conserved DNA motif containing the core sequence GATTA (Thompson et al., 2007). Consistently, our motif analysis rediscovered GATTA as a binding motif for GTF2IRD1 (Table 1), which is present in three copies in uc400. GTF2IRD1 is expressed ubiquitously with the exclusion of the forebrain during E10.5 (Palmer et al., 2007), a finding in agreement with the forebrain-specific activity of uc400, the role of Gtf2ird1 as a repressor, and our interaction data. Given the degree of corroboration between existing literature and our data, we decided to investigate possible regulatory modulation of uc400 by hGTF2IRD1.

We first confirmed that hGTF2IRD1 bound to uc400 via the GATTA motif, by mutating all occurrences of such motifs to GAGGA. MS analysis showed hGTF2IRD1 to be the only DNA binding protein bound preferentially to the wild-type uc400 bait compared to the mutant bait (Figure 5B). Interestingly, the data immediately revealed that the mutant uc400 had also gained specific binding of another TF, namely hTEAD1. We then performed reporter assays using wild-type or mutant uc400 as an enhancer driving luciferase reporter, under nontargeting condition or GTF2IRD1 knockdown. Owing to autoregulation of Gtf2ird1 (Palmer et al., 2010), we also monitored mRNA expression levels together with luciferase reporter activity over a time course (Figure S4). We found that hGTF2IRD1 knockdown resulted in differential reporter activity modulation of wild-type uc400 relative to the mutant uc400 (Figure 5D). Because our mutagenesis of uc400 reporter resulted in the gain of hTEAD1 binding site (Figure 5B), we also excluded the indirect effects of hGTF2IRD1 knockdown on reporter activity through hTEAD1 by showing that its mRNA expression level was only modestly affected throughout the course of the experiment (Figures 5C and S4). In conclusion, we have demonstrated a regulatory consequence of the interaction between uc400 and the hGTF2IRD1 protein.

To further explore the regulatory relevance of UCE interactors more globally in cellular contexts, we compared our interaction data with existing ChIP-seq data from the ENCODE consortium (Bernstein et al., 2012). We found 12 TFs from our screen with corresponding ChIP-seq data obtained from the H1 human embryonic cell line, giving rise to 31 *cis-trans* interaction pairs relevant to our loci of interest. ChIP-seq measures if a TF is present at a genomic locus; therefore, if there is a signal in ChIP-seq and a pull-down experiment has been performed on the sequence, then we should have also identified the factor by MS. This was indeed true in 90% of the cases. Although we do not expect the strength of a ChIP-seq signal to directly correlate with the MS measurements—because of the different nature of the experiments—in 65% of the cases (20 interactions), the SILAC ratios indicated clear enrichment over random genomic sequences. In a few cases, the SILAC ratios loosely correlated with the ChIP-seq scores (Figure S4). We also found a highly significant tendency for loci with congruent interactions to have more accessible chromatin than the remaining loci, as deduced by DNase I-hypersensitivity signal (Figure 5E). This suggests that open chromatin has an influence on observing intrinsic interactions in the cell. Overall, the available ChIP-seq data validate the relevance of our UCE interactome in a native genomic context.

Regulatory relevance of our interactome in cellular context should also be reflected in cellular chromatin states associated with enhancer and repressor activity. We therefore correlated our SILAC profiles with several histone methylation and acetylation ChIP-seq tracks as well as with the DNase I-hypersensitivity track. Initial analysis of H1-hESC ChIP-/DNase-seq data obtained from ENCODE revealed that, regardless of the track under consideration, proteins whose SILAC ratios most strongly correlated with the ChIP-/DNase-seq signal were those with strong GC-content preference. To correct for this known bias of ChIP-seq data sets (Dohm et al., 2008), we report association between SILAC profiles and ChIP-/DNA-seq profiles in terms of deviation from correlation expected of the interactor's GC

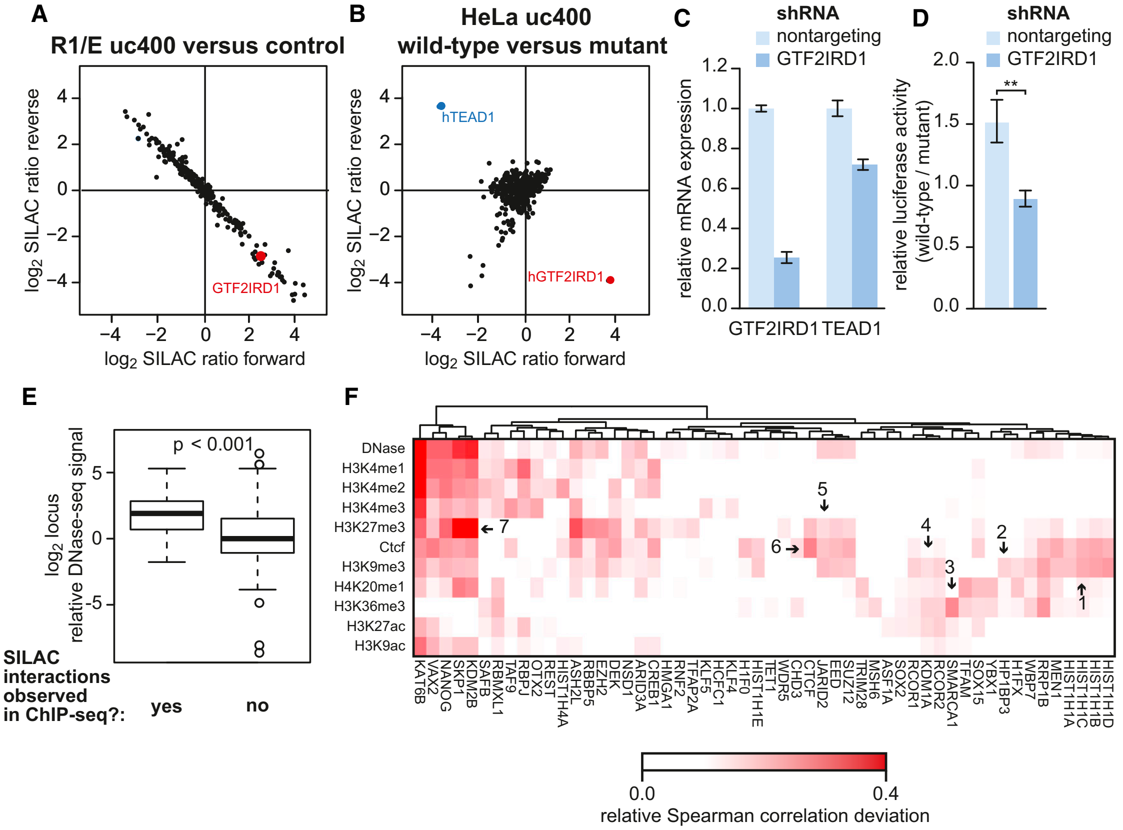

Figure 5. Regulatory Consequence of UCE Interactions

(A) Interaction of uc400 with TF GTF2IRD1 in R1/E compared to random genomic loci.

(B) Disruption of uc400 interaction with GTF2IRD1 in a GATTA → GAGGA mutant form uc400 compared to wild-type.

(C) Relative mRNA expression levels of GTF2IRD1 and TEAD1 at 12 hr posttransfection for nontargeting and GTF2IRD1-knockdown conditions (mean ± SEM). See also Figure S4.

(D) Luciferase reporter activity of wild-type uc400 normalized to that of the GATTA → GAGGA variant at 21 hr (mean ± SEM). Significance levels are **p < 0.01.

(E) Distribution of relative DNase-seq signal for the baits containing ChIP-seq interaction congruent to the SILAC interactome, compared to the baits where no ChIP-seq signal was detected. See also Figure S4.

(F) Heatmap of correlation deviation for all identified chromatin proteins with in vivo chromatin states associated with enhancer activity, repression, and active transcription. Arrows indicate associations consistent with the existing literature (see also Table S3).

See also Figure S4 and Table S3.

preference. We validated our analysis by comparing SILAC profiles to the CTCF ChIP-seq track and, indeed, found the SILAC profile of CTCF to be most strongly associated with its own binding in H1-hESCs (Figure 5F, arrow 6).

The analysis recovered several known relationships between intrinsic interactors and cellular chromatin states at corresponding loci. For example, the PRC1 was most strongly correlated with the classical Polycomb mark H3K27me3, but also to a lesser extent with the enhancer marks (Figure 5F, arrow 6), a finding in line with the bivalent nature of H3K27 methylation and H3K4 methylation (Bernstein et al., 2006; Ku et al., 2008). In contrast, no correlation was observed for the PRC1 with H3K27ac, a mark that counteracts Polycomb silencing (Pasini et al., 2010; Tie et al., 2009). Table S3 summarizes the full set of associations between our interaction data and chromatin data along with functional interpretation. These associations indicate that proteins involved in chromatin-modification pathways already bind even in the initial absence of epigenetic priming. Taken together, our analyses demonstrate the regulatory relevance of our interactome by illustrating congruence between cell-type-specific intrinsic interaction at UCEs and in cellulo chromatin-modification states.

The UCE Interactome Is Determined by the Cellular Context

It is conceivable for DNA sequences of high regulatory information density such as UCEs that regulation is cell-type specific. Such variation in regulatory logic should reflect itself in change in interactions. To explore this, we also obtained interaction data for a subsample of UCEs in the context of HeLa cells. Comparison between the two data sets revealed that homologous interactors with high sequence identity between mouse and human are more likely to have highly correlated binding. Examples of such homolog pairs include CHD7, TFAP4, and RCOR1

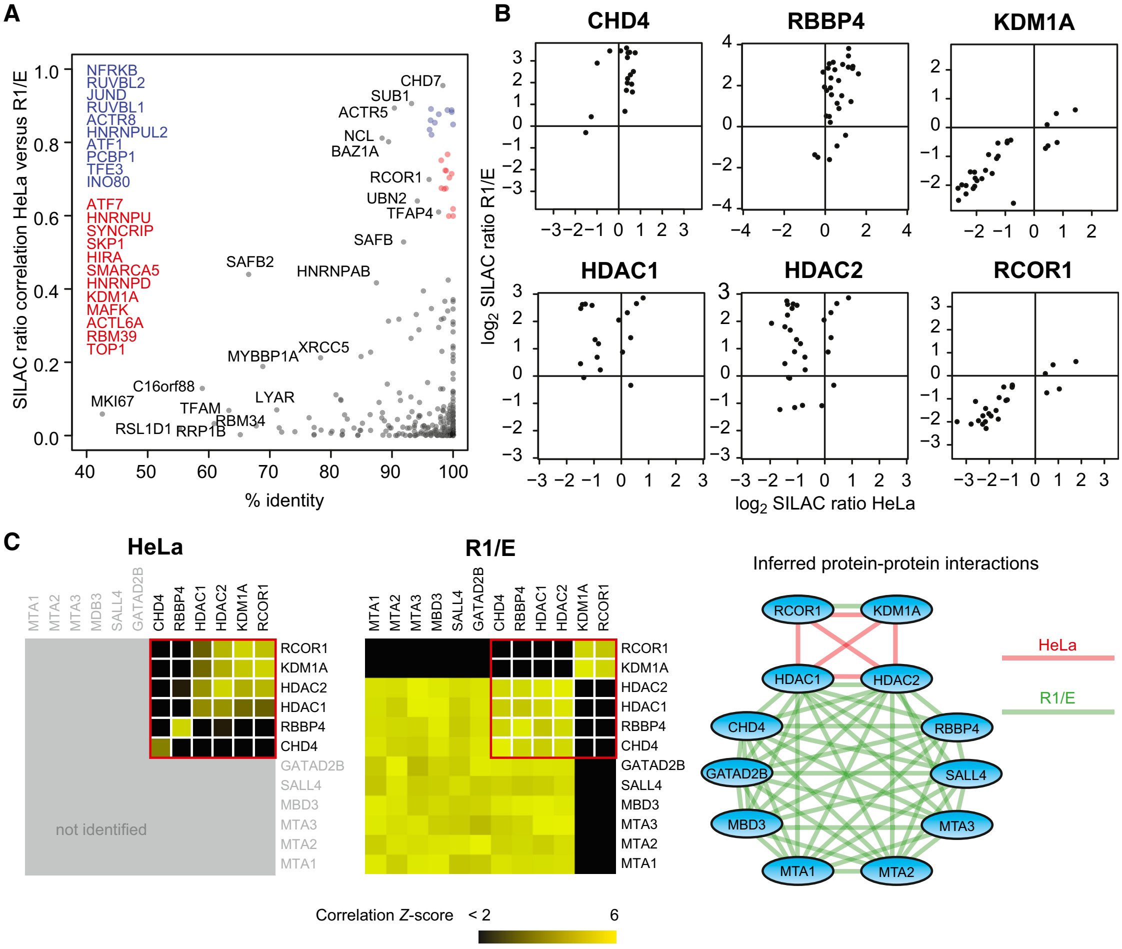

Figure 6. Comparison of UCE Interactomes Obtained in HeLa and R1/E Backgrounds

(A) Scatterplot showing SILAC ratio correlation between R1/E and HeLa data sets against human-mouse protein sequence identity. Names of the proteins represented by colored points are given on the left.

(B) Example profiles of proteins with high human-mouse sequence identity.

(C) Comparison of protein-protein interactions deduced from profile correlation (see also Figure 1), showing members of the REST corepressor complex and the NuRD complex, and the switch of complex membership of HDAC1/Hdac1 and HDAC2/Hdac2.

(Figure 6A). However, many highly identical homolog pairs also behave differently between cell lines, indicating effects of cellular context upon intrinsic interaction with our baits (Figures 6A and 6B). For example, by using profile correlation across baits as a measure for complex organization (Figure 1E), we found that the proteins HDAC2 and HDAC1 bound to our baits in differing contexts: as part of the REST corepressor complex in the HeLa background, and as part of the NuRD complex in the R1/E background (Figure 6C). Thus, UCE sequences are capable of recruiting different interactors based on the nuclear proteome and protein-protein interactome of the cell.

DISCUSSION

Despite the comprehensive tabulation of enhancer activities of UCEs, the candidate interactors responsible for regulation have not been systematically characterized. Although protein-centric approaches such as ChIP-seq have long allowed for global analysis of interactions of candidate proteins with the genome, a DNA-centric approach is particularly suited to answering this question. We have applied DNA-centric interaction screening to map intrinsic interactions of the sequences of hundreds of UCEs to obtain two highly information-rich data sets: the UCE interactome, and the uc325 differential interactome. The exquisite quantitative accuracy of SILAC, combined with the large scale of the interactome study, allowed us to provide candidate interactors that can be used for follow-up studies of UCE regulatory logic, as well as to quantitatively address interaction tendencies of UCEs as a family of sequences—a question not previously addressable in smaller-scale applications of the DNA-centric paradigm.

The analyses demonstrated that the sequences of nxUCEs represent the extreme case when compared to pxUCEs, exUCEs, and random genomic sequences in many aspects of protein-DNA interactions. They were most enriched in intrinsic interactors, especially those annotated to be important in tissue-specific development, they were most refractory to intrinsically GC-rich binding of the heterochromatin-promoting PRC2 (Figure 2), and they were most enriched in deeply conserved, overlapping TFBSs (Figure 3). The latter phenomenon is in the extreme even compared to other nonultraconserved enhancers in the genome. Although the extent to which individual interactions contribute to the regulatory output remains to be determined, we have shown that interactions are recapitulated in cells by ChIP-seq and, as a whole, corroborate with observed chromatin states that reflect regulatory consequences (Figure 5). Furthermore, UCEs appear to bind different factors in different cellular background, which can be explained in part by rewired protein-protein interaction (Figure 6). All these findings provide strong experimental support to the hypothesis of nxUCEs as highly constrained transcriptional regulatory modules (Bejerano et al., 2004; Siepel et al., 2005).

If nxUCEs are highly information-dense regulatory circuits, it is conceivable that any mutation would result in regulatory alterations with adverse effects to the organism. This is supported by the conservation bias of overlapping TFBSs inferred from the UCE interactome and the sensitivity of uc325 to mutation with respect to the gain and loss of binders (Figures 3 and 4). Our observation that mutating hGTF2IRD1 binding sites in uc400 results in the gain of Tead1 binding further exemplifies the idea that functional binding sites can be gained spontaneously through mutation of an existing motif (Figure 5). Our finding that fine-resolution conservation of uc325 correlated with the tendency to gain interactors also lends possibility to the concept that UCEs are under selective pressure that not only prevents loss of regulatory function but also its logical alteration (Figure 4). This is supported by the discovery that whereas many TFBSs can be functional regardless of their context with neighboring TFBSs, some TFs do indeed have a strict contextual prerequisite (Smith et al., 2013). Context-dependent binding might provide cell-type-specific logic that provides further conservational constraints not yet explored in this study. Still, further contribution may come from functional constraints beyond enhancer function (Licastro et al., 2010; Ni et al., 2007; Scaruffi, 2011).

We found that pxUCEs and exUCEs were less extreme in their transcriptional regulatory characteristics as indicated by their intrinsic interactions, in line with their possible functional roles beyond transcriptional regulation. We found pxUCEs to behave similarly to nxUCEs in some aspects (Figures 2B and 3D), to exUCEs in others (Figure 2A), and often as an average between nxUCEs and exUCEs (Figures 2B, 2C, and 3B). This raises the possibility that some of the putative exons coinciding with pxUCEs may in fact be functional exons, and others may be enhancers.

There remains the general challenge that certain deletions or mutations of UCEs have failed to produce observable deleterious phenotypes (Ahituv et al., 2007), which can be interpreted against the high-constraint hypothesis. However, this absence of evidence is not surprising, given that almost all ultraconserved enhancers remain to be systematically characterized at the regulatory level, where the context and environment under which they become indispensable need to be determined. Indeed, it is now known that some enhancers contribute to robust regulation and are indispensable only under certain extreme conditions (Perry et al., 2010). Full, systematic ab initio functional characterization of regulatory elements, including upstream events, context-dependent regulatory logic, and downstream consequences, remains a daunting task. Here, we have demonstrated the utility of our approach as a crucial initial step in the process, and complementary to the VISTA enhancer data that tabulated enhancer activity of UCEs, we provide their potential interactors. The use of insertional ChIP where the interaction was queried in vivo would be a very attractive follow-up in order to ascertain the exact cell specificity of interactions (Hoshino and Fujii, 2009). Further integration with data obtained for in vivo protein-DNA interactions, protein-protein interactions, long-range DNA interactions, as well as gene expression data, reporter assays, and perturbation experiments will allow deep functional characterization of UCEs with the aim to discover their target genes and functional contexts as well as to decode their exact regulatory logic.

EXPERIMENTAL PROCEDURES

Stem Cell Culture and Nuclear Extract Preparation

R1/E cells were SILAC labeled in SILAC DMEM (PAA Laboratories) containing either 73 mg/l Lys-8 HCl and 42 mg/l Arg-10 HCl, or the same concentration of Lsy-0 HCl and Arg-0 HCl. Medium was supplemented with 10% dialyzed FBS (PAA Laboratories), 1× nonessential amino acids (GIBCO Life Technologies), 1 mM sodium pyruvate (GIBCO Life Technologies), 3 μM CT-99021 (Biomol), 1 μM PD-0325901 (Biomol), 50 μM 2-mercaptoethanol (GIBCO Life Technologies), 100 u/ml LIF (Millipore), and penicillin-streptomycin-glutamate. Nuclear extracts were prepared as previously described by Dignam et al. (1983) except for a reduced NP40 concentration of 0.5% to preserve nuclear integrity during cell lysis. Extracts were controlled for the presence of Oct4 by western blot.

Cloning and DNA Bait Generation

UCEs and 24 random mouse and human genomic loci were cloned into pCR8/TOPO/TA (Life Technologies). See Table S4 for genome coordinates of the inserts. Desthiobiotin-conjugated DNA baits of 200–1,000 bp were generated by PCR using the following primers: forward, 5′-desthiobiotin-CAGGCTCCGAATTCGCCCTT-3′; and reverse, 5′-GAAAGCTGGGTCGAATTCGCC-3′. PCR products were concentrated by ethanol precipitation and purified from unincorporated primers on G-50 Sephadex columns (GE Healthcare). Baits for uc325 scanning pull-downs were produced by site-directed mutagenesis PCR.

DNA Pull-Downs and Mass Spectrometric Analysis

DNA pull-downs and sample preparation for mass spectrometric analysis were performed as previously described (Butter et al., 2012). Peptides derived from the bound proteins were separated by HPLC over a 140 min gradient from 2% to 60% acetonitrile and analyzed in an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific). Full-scan MS was acquired with 120,000 resolution in the Orbitrap analyzer, and up to the ten-most intense ions from each full scan were fragmented with collision-induced dissociation and analyzed in the linear ion trap. Mass spectrometric data were processed with the MaxQuant software version 1.2.6.20 (Cox and Mann, 2008). The complete pull-down data set from R1/E and the nuclear proteome data set were searched against the mouse UniProt database. We mapped GO (Ashburner et al., 2000) and Pfam (Bateman et al., 2004) annotations to protein groups using the Perseus module in the MaxQuant software suite.

Nuclear Proteome of R1/E Cells

R1/E nuclear extracts were precipitated in four-volume acetone. The pellet of nuclear proteins was resuspended in 8 M urea, and proteins were digested in solution. Peptides were separated by HPLC over a 240 min gradient from 2% to 60% acetonitrile and analyzed in a Q-Exactive mass spectrometer (Thermo Fisher Scientific) (Michalski et al., 2011). Five replicates were measured to extend proteome coverage. Mass spectrometric data were processed with MaxQuant version 1.2.6.20.

Reporter Assays

We cloned uc.400 into a modified pGL3/Basic firefly luciferase reporter vector containing a minimum mouse heat shock promoter via the Gateway system as previously described (Butter et al., 2012). Primers for amplifying uc.400 were as follows: forward, 5′-GCCTCTCTGAAGCGTTCATC-3′; and reverse, 5′-TGGTGTTACGGATCACAACG-3′. The mutant variant of uc.400 was generated by PCR using mutagenizing primers and subcloned into pCR8/TOPO vector.

Transfection and reporter assays were performed as previously described (Butter et al., 2012). Knockdown of hGTF2IRD1 was achieved using shRNA vector generated using pSUPERIOR vector system, following the manufacturer's protocol. The shRNA core half-sequences for GTF2IRD1 and nontargeting construct were CAGAAAGACTAAAGGAAAT and GACTAGAAGGCA CAGAGGGAG, respectively. Knockdown was quantified using quantitative real-time PCR and SYBR green system, using the standard $\Delta\Delta C_t$ method and normalizing over GAPDH. Primers used for quantitative real-time PCR were as follows: GAPDH, 5′-CAAGGTCATCCATGACAACTTTG-3′ and 5′-GTCCACCACCCTGTTGCTGTAG-3′; GTF2IRD1, 5′-ATCATCACCAGCCTC GTGTC-3′ and 5′-CACCTTCTTGGGGTGCTCT-3′; and TEAD1, 5′-CATGTC CTCAGCCCAGATCG-3′ and 5′-AGGCTCAAACCCTGGAATGG-3′.

Data Analysis

Preprocessing

SILAC ratios were corrected to account for residual proteome differences between heavy and light nuclear extracts (see Supplemental Experimental Procedures for details). Protein groups were then filtered for having a coefficient of determination of SILAC ratios greater than 0.2 across all baits and for having log2 SILAC ratios exceeding 0.5 in at least three baits. For subsequent analyses, we applied a GO annotation filter, requiring the protein groups to contain at least one of these words or their variants as a substring of the GO terms: chromatin, DNA, enhancer, genome, helicase, histone, nuclear, promoter, RNA, splicing, transcription, and translation.

Imputation

Where imputation was required, we filled missing logarithmized quantifications with a normal distribution with the mean equal to the minimum SILAC ratio for each protein and the SD of 0.5. This number was empirically determined to best simulate the errors of SILAC ratios in the data set.

Annotation Enrichment Analysis

We used Pfam annotation to class interactors by domain, and imputed SILAC ratios were used to calculate enrichment. For JASPAR prediction (Bryne et al., 2008), we used the standard position weight matrix-scoring procedure, normalizing the scores to the maximum value attainable for each motif.

Ab Initio Motif Enrichment

For each k-mer motif where $1 \leq k \leq 8$ (excluding reverse complement redundancies), the median motif occurrence in both orientations was determined. DNA baits were then divided into those having less than or equal to the median occurrence of the motif (“low occurrence”), and those having greater than the median occurrence (“high occurrence”). Wilcoxon rank sum test was then used to calculate significance in difference in imputed SILAC ratios between the “high motif occurrence” and “low motif occurrence” bait sets. We used Benjamini-Hochberg FDR to adjust the p value for multiple comparisons (Benjamini and Hochberg, 1995).

Superimposition Analysis

We chose a minimum motif length λ, where $4 \leq \lambda \leq 7$. To exclude counting the overlapping of different length but otherwise redundant motifs, we applied two criteria for keeping a motif: (1) that the motif length was at least λ, and (2) that there existed no shorter motif that was a substring of the motif being considered or its reverse complement. Motifs only significantly associated with de-enrichment of interactors but not enrichment were not considered. Conservation data were obtained from the UCSC Genome Browser (Build hg19). Nonultraconserved VISTA enhancer coordinates were obtained from the VISTA database (Visel et al., 2007). Conservation data were obtained from the phylop46wayAll and phylop46wayPlecantal tracks of hg19, respectively (Meyer et al., 2013).

ENCODE Data Set Integration

Broad histone ChIP-seq signal for histone modifications and peaks for TFBSs were obtained from the ENCODE histone ChIP-seq or DNase-seq tracks mapped to the hg19 build using the UCSC Genome table browser. See Table S3 for the track listing. Only loci corresponding to bait sequences with nonzero signal in both the DNase-/ChIP-seq track and in the control track were considered. For each protein, the Spearman correlation coefficient was determined between SILAC ratios logarithmized DNase-/ChIP-seq signal density normalized to control signal density. Correlation coefficient deviation was calculated by subtracting the expected DNase-/ChIP-seq-to-SILAC ratio correlation given the bait GC content-to-SILAC ratio correlation and then normalized to the minimum value.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2013.09.022.

ACKNOWLEDGMENTS

The authors thank Dennis Kappei for the establishment of SILAC 2i mouse embryonic stem cell culture; Christian Eberl, Charo Robles, and Jürgen Cox for discussions; and Marta Murgia, Paul Boersema, Michael Wierer, and Franziska Paul for suggestions and comments on the manuscript. This project was supported by the EU grant HEROIC (18883; FP6) and the Max-Planck Society for the Advancement of Science.

Received: May 24, 2013
Revised: August 6, 2013
Accepted: September 11, 2013
Published: October 17, 2013

REFERENCES

Aalfs, J.D., Narlikar, G.J., and Kingston, R.E. (2001). Functional differences between the human ATP-dependent nucleosome remodeling proteins BRG1 and SNF2H. J. Biol. Chem. *276*, 34270–34278.

Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L.A., and Rubin, E.M. (2007). Deletion of ultraconserved elements yields viable mice. PLoS Biol. *5*, e234.

Arbab, M., Mahony, S., Cho, H., Chick, J.M., Rolfe, P.A., van Hoff, J.P., Morris, V.W., Gygi, S.P., Maas, R.L., Gifford, D.K., and Sherwood, R.I. (2013). A multiparametric flow cytometric assay to analyze DNA-protein interactions. Nucleic Acids Res. *41*, e38.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al. (2004). The Pfam protein families database. Nucleic Acids Res. *32*(Database issue), D138–D141.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321–1325.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B Met. *57*, 289–300.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., and Celniker, S.E. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. Genome Biol. *5*, R61.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M.; ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Bryne, J.C., Valen, E., Tang, M.H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. *36*(Database issue), D102–D106.

Butter, F., Davison, L., Viturawong, T., Scheibe, M., Vermeulen, M., Todd, J.A., and Mann, M. (2012). Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. PLoS Genet. *8*, e1002982.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science *298*, 1039–1043.

Collins, D.W., and Jukes, T.H. (1994). Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics *20*, 386–396.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

Déjardin, J., and Kingston, R.E. (2009). Purification of proteins associated with specific genomic Loci. Cell *136*, 175–186.

Dignam, J.D., Lebovitz, R.M., and Roeder, R.G. (1983). Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. Nucleic Acids Res. *11*, 1475–1489.

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. *36*, e105.

Eberl, H.C., Spruijt, C.G., Kelstrup, C.D., Vermeulen, M., and Mann, M. (2013). A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics. Mol. Cell *49*, 368–378.

Hermsen, R., Tans, S., and ten Wolde, P.R. (2006). Transcriptional regulation by competing transcription factor modules. PLoS Comput. Biol. *2*, e164.

Hoshino, A., and Fujii, H. (2009). Insertional chromatin immunoprecipitation: a method for isolating specific genomic regions. J. Biosci. Bioeng. *108*, 446–449.

Jin, C., and Felsenfeld, G. (2007). Nucleosome stability mediated by histone variants H3.3 and H2A.Z. Genes Dev. *21*, 1519–1529.

Koyama-Nasu, R., David, G., and Tanese, N. (2007). The F-box protein Fbl10 is a novel transcriptional repressor of c-Jun. Nat. Cell Biol. *9*, 1074–1080.

Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X., Chi, A.S., et al. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet. *4*, e1000242.

Licastro, D., Gennarino, V.A., Petrera, F., Sanges, R., Banfi, S., and Stupka, E. (2010). Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements. BMC Genomics *11*, 151.

Lu, X., Wontakal, S.N., Emelyanov, A.V., Morcillo, P., Konev, A.Y., Fyodorov, D.V., and Skoultchi, A.I. (2009). Linker histone H1 is essential for *Drosophila* development, the establishment of pericentric heterochromatin, and a normal polytene chromosome structure. Genes Dev. *23*, 452–465.

Martínez, F., Monfort, S., Roselló, M., Oltra, S., Blesa, D., Quiroga, R., Mayo, S., and Orellana, C. (2010). Enrichment of ultraconserved elements among genomic imbalances causing mental delay and congenital anomalies. BMC Med. Genomics *3*, 54.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet. *6*, e1001244.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., et al. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. *41*(Database issue), D64–D69.

Michalski, A., Damoc, E., Hauschild, J.P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011). Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol. Cell. Proteomics *10*, M111.011015.

Mirzaei, H., Knijnenburg, T.A., Kim, B., Robinson, M., Picotti, P., Carter, G.W., Li, S., Dilworth, D.J., Eng, J.K., Aitchison, J.D., et al. (2013). Systematic measurement of transcription factor-DNA interactions by targeted mass spectrometry identifies candidate gene regulatory proteins. Proc. Natl. Acad. Sci. USA *110*, 3645–3650.

Mittler, G., Butter, F., and Mann, M. (2009). A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. Genome Res. *19*, 284–293.

Ngondo-Mbongo, R.P., Myslinski, E., Aster, J.C., and Carbon, P. (2013). Modulation of gene expression via overlapping binding sites exerted by ZNF143, Notch1 and THAP11. Nucleic Acids Res. *41*, 4000–4014.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., and Ares, M., Jr. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. *21*, 708–718.

Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics *1*, 376–386.

Palmer, S.J., Tay, E.S., Santucci, N., Cuc Bach, T.T., Hook, J., Lemckert, F.A., Jamieson, R.V., Gunnning, P.W., and Hardeman, E.C. (2007). Expression of Gtf2ird1, the Williams syndrome-associated gene, during mouse development. Gene Expr. Patterns *7*, 396–404.

Palmer, S.J., Santucci, N., Widagdo, J., Bontempo, S.J., Taylor, K.M., Tay, E.S., Hook, J., Lemckert, F., Gunning, P.W., and Hardeman, E.C. (2010). Negative autoregulation of GTF2IRD1 in Williams-Beuren syndrome via a novel DNA binding mechanism. J. Biol. Chem. *285*, 4715–4724.

Pasini, D., Malatesta, M., Jung, H.R., Walfridsson, J., Willer, A., Olsson, L., Skotte, J., Wutz, A., Porse, B., Jensen, O.N., and Helin, K. (2010). Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes. Nucleic Acids Res. *38*, 4958–4969.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499–502.

Perry, M.W., Boettiger, A.N., Bothma, J.P., and Levine, M. (2010). Shadow enhancers foster robustness of *Drosophila* gastrulation. Curr. Biol. *20*, 1562–1567.

Poitras, L., Yu, M., Lesage-Pelletier, C., Macdonald, R.B., Gagné, J.P., Hatch, G., Kelly, I., Hamilton, S.P., Rubenstein, J.L., Poirier, G.G., and Ekker, M. (2010). An SNP in an ultraconserved regulatory element affects Dlx5/Dlx6 regulation in the forebrain. Development *137*, 3089–3097.

Rai, T.S., Puri, A., McBryan, T., Hoffman, J., Tang, Y., Pchelintsev, N.A., van Tuyn, J., Marmorstein, R., Schultz, D.C., and Adams, P.D. (2011). Human CABIN1 is a functional member of the human HIRA/UBN1/ASF1a histone H3.3 chaperone complex. Mol. Cell. Biol. *31*, 4107–4118.

Scaruffi, P. (2011). The transcribed-ultraconserved regions: a novel class of long noncoding RNAs involved in cancer susceptibility. ScientificWorldJournal *11*, 340–352.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005).

Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet. *45*, 1021–1028.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Tacheny, A., Michel, S., Dieu, M., Payen, L., Arnould, T., and Renard, P. (2012). Unbiased proteomic analysis of proteins interacting with the HIV-1 5'LTR sequence: role of the transcription factor Meis. Nucleic Acids Res. *40*, e168.

Thoma, F., and Koller, T. (1977). Influence of histone H1 on chromatin structure. Cell *12*, 101–107.

Thompson, P.D., Webb, M., Beckett, W., Hinsley, T., Jowitt, T., Sharrocks, A.D., and Tassabehji, M. (2007). GTF2IRD1 regulates transcription by binding an evolutionarily conserved DNA motif 'GUCE'. FEBS Lett. *581*, 1233–1242.

Tie, F., Banerjee, R., Stratton, C.A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., Diaz, M.O., Scacheri, P.C., and Harte, P.J. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes *Drosophila* Polycomb silencing. Development *136*, 3131–3141.

Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics *10*, 442.

Udugama, M., Sabri, A., and Bartholomew, B. (2011). The INO80 ATP-dependent chromatin remodeling complex is a nucleosome spacing factor. Mol. Cell. Biol. *31*, 662–673.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L.A. (2007). VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Res. *35*(Database issue), D88–D92.

Williamson, I., Hill, R.E., and Bickmore, W.A. (2011). Enhancers: from developmental genetics to the genetics of common human disease. Dev. Cell *21*, 17–19.

Wilson, M.D., Barbosa-Morais, N.L., Schmidt, D., Conboy, C.M., Vanes, L., Tybulewicz, V.L., Fisher, E.M., Tavaré, S., and Odom, D.T. (2008). Species-specific transcription in mice carrying human chromosome 21. Science *322*, 434–438.

Xie, W., Ling, T., Zhou, Y., Feng, W., Zhu, Q., Stunnenberg, H.G., Grummt, I., and Tao, W. (2012). The chromatin remodeling complex NuRD establishes the poised state of rRNA genes characterized by bivalent histone modifications and altered nucleosome positions. Proc. Natl. Acad. Sci. USA *109*, 8161–8166.

Yang, R., Frank, B., Hemminki, K., Bartram, C.R., Wappenschmidt, B., Sutter, C., Kiechle, M., Bugert, P., Schmutzler, R.K., Arnold, N., et al. (2008). SNPs in ultraconserved elements and familial breast cancer risk. Carcinogenesis *29*, 351–355.