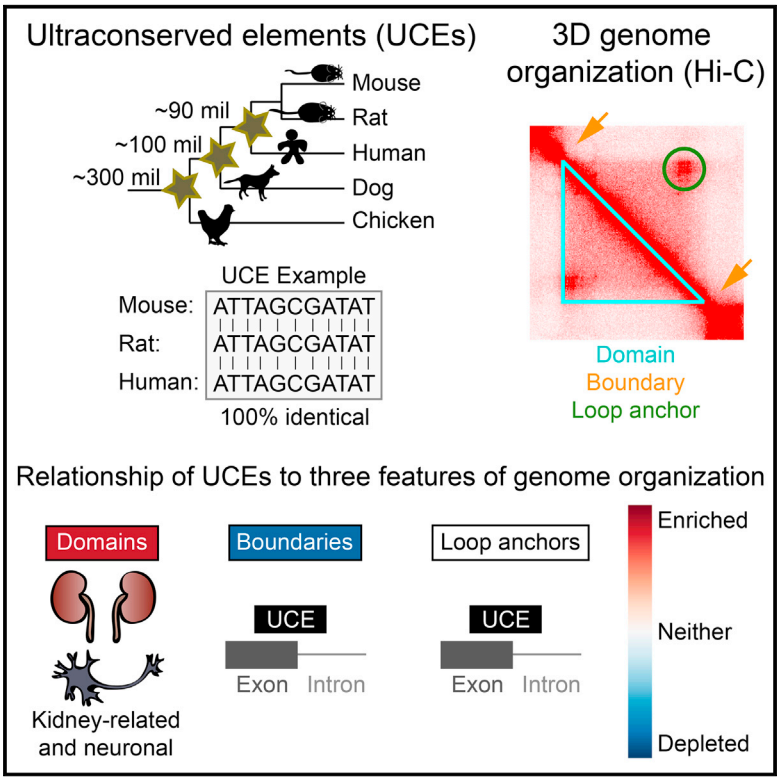
Article

Cell Reports

Ultraconserved Elements Occupy Specific Arenas of Three-Dimensional Mammalian Genome Organization

Graphical Abstract



Highlights

- UCEs are enriched in domains, depleted from boundaries, and neither at loop anchors
- UCEs in domains shared across cell types are linked to kidney and neuronal processes
- UCEs that do occur in boundaries and loop anchors are predominantly exonic
- UCEs that are present in loop anchors are enriched in splice sites

Authors

Ruth B. McCole, Jelena Erceg, Wren Saylor, Chao-ting Wu

Correspondence

twu@genetics.med.harvard.edu

In Brief

McCole et al. demonstrate the non-random relationship between the positions of perfectly conserved genomic regions, termed the ultraconserved elements (UCEs), and three-dimensional genome organization within mammalian nucleus as defined by Hi-C studies. They postulate that these connections aid in orchestrating genome packaging and preserving genome function and integrity.

McCole et al., 2018, Cell Reports *24*, 479–488
July 10, 2018
https://doi.org/10.1016/j.celrep.2018.06.031

CellPress

Cell Reports
Article

Ultraconserved Elements Occupy Specific Arenas of Three-Dimensional Mammalian Genome Organization

Ruth B. McCole,[1,2] Jelena Erceg,[1,2] Wren Saylor,[1] and Chao-ting Wu[1,3,*]
[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[2]These authors contributed equally
[3]Lead Contact
*Correspondence: twu@genetics.med.harvard.edu
https://doi.org/10.1016/j.celrep.2018.06.031

SUMMARY

This study explores the relationship between three-dimensional genome organization and ultraconserved elements (UCEs), an enigmatic set of DNA elements that are perfectly conserved between the reference genomes of distantly related species. Examining both human and mouse genomes, we interrogate the relationship of UCEs to three features of chromosome organization derived from Hi-C studies. We find that UCEs are enriched within contact domains and, further, that the subset of UCEs within domains shared across diverse cell types are linked to kidney-related and neuronal processes. In boundaries, UCEs are generally depleted, with those that do overlap boundaries being overrepresented in exonic UCEs. Regarding loop anchors, UCEs are neither overrepresented nor underrepresented, but those present in loop anchors are enriched for splice sites. Finally, as the relationships between UCEs and human Hi-C features are conserved in mouse, our findings suggest that UCEs contribute to interspecies conservation of genome organization and, thus, genome stability.

INTRODUCTION

Chromosome organization in the mammalian nucleus is strikingly orchestrated, like a symphony played throughout the organism's life span, composed by evolutionary forces. To explore this process of evolutionary "composition," we are investigating the relationships between chromosome organization and sequence evolution in the mammalian genome, focusing on some of the most highly conserved regions—the ultraconserved elements (UCEs) (Bejerano et al., 2004; Sandelin et al., 2004; Woolfe et al., 2005). UCEs show staggering levels of interspecies sequence conservation, demonstrating perfect sequence identity extending ≥200 bp between species that diverged 90–300 million years ago and comprising one of the most puzzling findings in comparative genomics (Harmston et al., 2013; Polychronopoulos et al., 2017). While UCEs have been found to encompass a variety of functions, including enhancer, promoter, splicing, and repressive activities (Bejerano et al., 2004; Dickel et al., 2018; Kushawah and Mishra, 2017; Pennacchio et al., 2006; Poitras et al., 2010; Sandelin et al., 2004; Warnefors et al., 2016), these functions arguably fall short of explaining ultraconservation, per se. We have suggested that UCEs may maintain their sequence conservation through a mechanism involving the pairing and comparison of allelic UCEs, followed by loss of fitness should mutations or rearrangements that disrupt UCE pairing be detected (Chiang et al., 2008; Derti et al., 2006; McCole et al., 2014) (see also Elgar and Vavouri, 2008; Kritsas et al., 2012). Such a mechanism would protect genome integrity in the body overall and, at the organismal level, promote ultraconservation over evolutionary timescales. Consistent with this model, UCEs are associated with regions of elevated synteny (Dimitrieva and Bucher, 2012; Dong et al., 2009; Irimia et al., 2012; Kikuta et al., 2007; Polychronopoulos et al., 2014, 2016; Sandelin et al., 2004; Sun et al., 2006, 2009). Furthermore, and in line with our proposal that disruptions of UCEs or UCE pairing lead to loss of fitness, the genomes of healthy individuals are generally not disrupted in the vicinity of UCEs (Chiang et al., 2008; Derti et al., 2006; McCole et al., 2014), while this pattern does not hold for genomes representing the cancerous state, or individuals with neurodevelopmental disorders or mental delay and congenital anomalies (Martínez et al., 2010; McCole et al., 2014). Highly conserved noncoding sequences can also interact in three dimensions (Robyr et al., 2011), adding weight to our proposal that interactions between UCEs in the nucleus may be important to their function. Finally, and of direct relevance to the proposal that allelic UCEs may pair, is the capacity of somatic genomes to support localized or whole chromosome pairing in a wide range of species (as reviewed by Joyce et al., 2016), with the most dramatic example in mammals being observed in renal oncocytoma (Koeman et al., 2008).

Here, we examine UCEs in the context of the three-dimensional organization of the genome, considering three features revealed by chromosome conformation capture (Hi-C) studies. We begin with contact "domains" (also called topologically associated domains [TADs]) and "boundaries"; contact domains are regions displaying frequent intra-regional interactions, while boundaries, which flank contact domains, are characterized by a paucity of interactions that traverse them (Bonev and Cavalli, 2016; Dekker et al., 2002; Denker and de Laat, 2016; Dixon et al., 2012, 2016; Liu and Weigel, 2015; Nora et al., 2012; Rao et al., 2014; Sexton et al., 2012). A third type of interaction involves the association of *cis*-linked regions known as "loop anchors," wherein the intervening genomic segment forms a loop (Rao et al., 2014). In concordance with the functional

importance of these three features, the positions of approximately half of domains, boundaries, and loops are conserved (Dixon et al., 2012; Rao et al., 2014), with domains preserved as units when positions are not conserved (Vietri Rudan et al., 2015). Thus, disrupting three-dimensional contacts inside domains may be disadvantageous, perhaps even oncogenic (Corces and Corces, 2016; Hnisz et al., 2016; Lupiáñez et al., 2016; Valton and Dekker, 2016; Weischenfeldt et al., 2017).

This study considers our proposal that ultraconservation protects genome integrity (Chiang et al., 2008; Derti et al., 2006; McCole et al., 2014) and hypothesizes that UCEs contribute to the preservation of domains over evolutionary time. In particular, we predicted that UCEs would be enriched within domains. In line with this, a recent publication reported that clusters of highly conserved noncoding elements (CNEs) correlate with the spans of domains encompassing genes involved in development (Harmston et al., 2017); although the thresholds for the length and identity used in this publication to define CNEs (>50 bp of 70%–90% conservation between human and chicken genomes) are much less stringent than those used to define UCEs, the findings are intriguing in light of our proposal. To test our hypothesis, we examined ten human and six mouse Hi-C datasets (Dixon et al., 2012; Fraser et al., 2015; Rao et al., 2014) and asked whether UCEs are enriched in or depleted from domains, boundaries, or loop anchors. Excitingly, UCEs proved to be significantly enriched in domains, and domains containing UCEs tend to be larger and relatively gene sparse, possibly suggesting a more structural role for these domains. In contrast, UCEs are generally depleted from boundaries and neither enriched nor depleted from loop anchors. The UCEs that do, nevertheless, occur in boundaries and loop anchors are predominantly exonic, with those in loop anchors enriched in splice sites. Our findings demonstrate that UCEs show specific, conserved relationships to domains, boundaries, and loops, hinting that UCEs may play a role in establishing and maintaining genomic organization.

RESULTS

UCEs Are Enriched within Domains, Depleted from Boundaries, and Indifferent to Loop Anchors

We began our studies by delineating how the Hi-C annotated genomic features of domains, boundaries, and loop anchors are related to the positioning of UCEs. To do this, we first collected published Hi-C datasets derived from nine human and five mouse tissues (Table S1), representing a variety of cell types (Dixon et al., 2012; Fraser et al., 2015; Rao et al., 2014). As Hi-C annotated regions vary between studies due to differences in cell type, species examined, amount of starting material, Hi-C protocol (in-solution [Dixon et al., 2012; Fraser et al., 2015] or in-nucleus [Rao et al., 2014]), and sequencing depth, we examined each dataset individually in addition to querying datasets combined according to species and genomic feature (Table S1) (Dixon et al., 2012; Fraser et al., 2015; Rao et al., 2014). Regarding UCEs, our analyses used our previously defined dataset (Table S2C), which comprises 896 elements that are ≥200 bp in length and identical in sequence within at least one of three groups of reference genomes (Derti et al., 2006; McCole et al., 2014). The three groups consist of the reference genomes of human, mouse, and rat (HMR), of human, dog, and mouse (HDM), and of human and chicken (HC), with the combined dataset of 896 UCEs designated as HMR-HDM-HC (Table S2C). To obtain UCE positions in the mouse genome, we aligned human UCE sequences to the mouse genome and recovered 893 orthologs (Supplemental Experimental Procedures; Table S2D). UCEs were also subdivided into exonic, intronic, and intergenic categories, which were then examined jointly and separately for enrichment or depletion within the Hi-C annotations (Experimental Procedures). Of note, a UCE is considered exonic if any part overlaps an exon; hence, exonic UCEs may overlap splice sites and contain intronic sequence.

To assess whether UCEs are significantly enriched in or depleted from domains, boundaries, and loop anchors, we used our previously established method (Chiang et al., 2008; Derti et al., 2006; McCole et al., 2014) (Figure 1), which compares "observed overlaps," in base pairs, between UCEs and Hi-C annotated regions to "expected overlaps" between a set of regions matched to UCEs in terms of number and length, but randomly positioned in the genome. Expected overlaps are generated 1,000 times to produce a distribution of expected overlaps, which, when normally distributed, is subjected to a Z-test to compare the observed overlap with the distribution of expected overlaps. In cases where normality is not observed, the proportion of expected overlaps equal to, or more extreme than, the observed overlap is reported. In all cases, we report the ratio of observed to mean expected overlap (obs/exp). This tailored approach for each Hi-C dataset enables comparison of datasets that differ in number of identified regions, median region size, and percentage of genome covered.

We first analyzed ten datasets of domains, drawn from Dixon et al. (2012) and Rao et al. (2014), that examined nine human cell lines, whose origins spanned embryonic (human embryonic stem cell [hESC]) and fetal (IMR90 lung fibroblast) development, cancer (HeLa, K562, and KBM7), and differentiated tissues (GM12878, human mammary epithelial cell [HMEC], human umbilical vein endothelial cell [HUVEC], and normal human epidermal keratinocyte [NHEK]), with IMR90 studied by both Dixon et al. and Rao et al. and thus contributing two datasets (Table S1). The domains described by these datasets range in coverage from 83.2% of the genome for hESC domains from Dixon et al. (2012) to 40.1% for HMEC domains from Rao et al. (2014). Excitingly, we observed significant enrichment for UCEs within domains in eight out of ten datasets ($4.22 \times 10^{-15} \leq p \leq 0.020$, $1.061 \leq obs/exp \leq 1.167$; Table S2A); the two in which enrichment was not seen represented HMEC and NHEK cells from Rao et al. (2014) (Table S2A). Combining all ten datasets, which included merging overlapping regions, produced a dataset, called "pooled domains," containing 293 regions covering 89.1% of the genome (Table S1) that is also significantly enriched for UCEs ($p = 2.77 \times 10^{-6}$, obs/exp = 1.025; Figure 2A; Table S2A). These results show that UCEs are overrepresented within Hi-C domains across many cell types, supporting the idea that there is an interrelationship between UCEs and three-dimensional chromosome conformation.

We then examined datasets of boundaries from Dixon et al. (2012). These datasets, which represent hESC and IMR90 cells

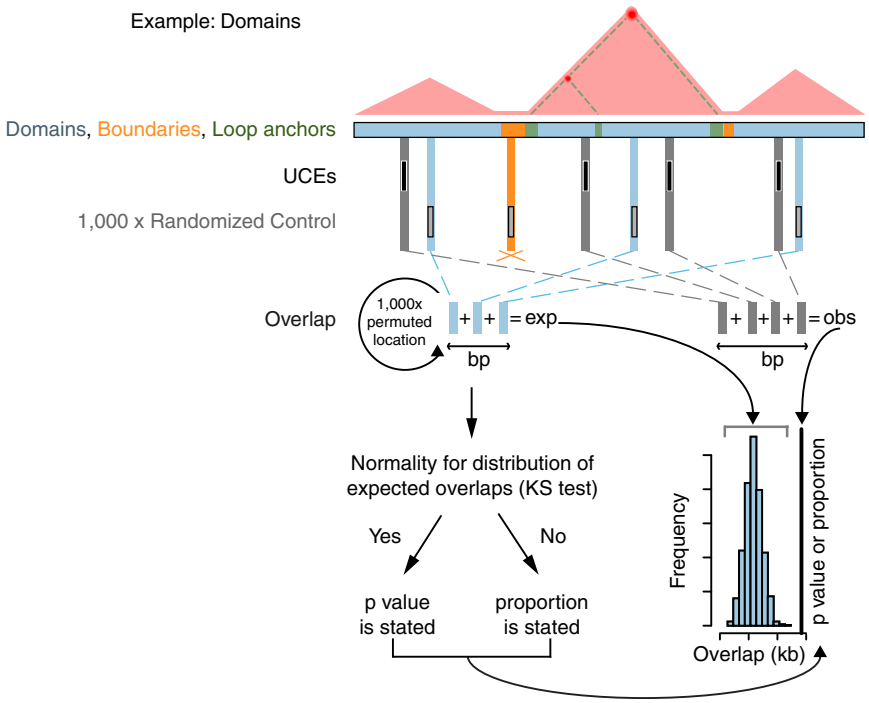


Figure 1. Strategy for Assessing the Relationship between UCEs and Domains, Boundaries, and Loop Anchors

We assess the relationship of UCEs (black) to domains (blue), boundaries (orange), and loop anchors (green) via a multi-step process, illustrated here with respect to domains. Throughout this and other figures, blue, orange, and green refer to analyses related to domains, boundaries, and loop anchors, respectively. First, overlaps between UCEs and all domains in a dataset are summed to produce the observed overlap; as this example concerns domains, overlap between UCEs and boundaries are not tallied (orange cross). The observed overlap is then compared to a distribution of expected overlaps generated from the overlap of domains with each of 1,000 sets of control genomic sequences, matched to UCEs in number and length and randomly positioned in the genome. Finally, the distribution of the resulting 1,000 control overlaps is tested for normality using the Kolmogorov-Smirnov (KS) test, and when normality is observed, a Z-test p value is reported to describe the significance of the deviation of the observed overlap from the distribution of expected overlaps. If normality is not observed, the proportion of expected overlaps equal to, or more extreme than, the observed overlap is stated. See also Tables S1 and S2.

and cover 4.0% and 3.6% of the genome, respectively (Table S1), are significantly depleted of UCEs ($p = 0.002$, obs/exp = 0.516, and $p = 0.025$, obs/exp = 0.669, respectively; Table S2A; although for IMR90, the p value hovers at our significance cutoff). Merging the two datasets created a "pooled boundary" dataset, containing 3,715 regions and covering 6.6% of the genome (Table S1), that is also depleted for UCEs ($p = 7.51 \times 10^{-4}$, obs/exp = 0.609; Figure 2A; Table S2A). These findings reinforce our observation that UCEs do not commonly occur within Hi-C boundaries and complement our previous observation that UCEs preferentially occur within domains.

Our next analysis concerned eight datasets of loop anchors provided by Rao et al. and representing GM12878, HeLa, HMEC, HUVEC, IMR90, K562, KBM7, and NHEK cells, with genome coverage ranging from 2.3% to 5.9% (Table S2A). For all but two datasets, UCEs are neither enriched nor depleted ($0.006 \leq p \leq 0.480$, $0.710 \leq$ obs/exp ≤ 1.334; Table S2A). Merging all eight datasets produced a dataset of "pooled loop anchors," comprising 18,331 regions and covering 13.6% of the genome (Table S1), that is also neither enriched nor depleted for UCEs ($p = 0.073$, obs/exp = 1.124; Figure 2A; Table S2A). The overall lack of UCE enrichment in loop anchors is surprising, since many UCEs show enhancer-like properties (Bhatia et al., 2013; Lampe et al., 2008; McBride et al., 2011; Pauls et al., 2012; Pennacchio et al., 2006; Poitras et al., 2010; Poulin et al., 2005; Visel et al., 2008; Woolfe et al., 2005), and enhancer-promoter interactions have been proposed to generate loops (Rao et al., 2014). Indeed, we did observe enrichment of UCEs in two of the eight datasets, HUVEC ($p = 0.020$, obs/exp = 1.322; Table S2A) and NHEK ($p = 0.006$, obs/exp = 1.334; Table S2A), suggesting that UCEs might be particularly involved in loop anchors in endothelial and epidermal cell types (HUVEC and NHEK cells, respectively).

Relationships of UCEs to Hi-C Annotations Are Robust

Having revealed positional relationships between UCEs and domains, boundaries, and loop anchors, we examined whether these relationships are robust to co-correlation with nine other genomic features. These features, which can be considered controls, included six that were previously shown to be non-randomly associated with UCE positions: copy number variants (CNVs), cancer-specific copy number alterations (CNAs), genes, exons, introns, and segmental duplications (SDs) (Chiang et al., 2008; Derti et al., 2006; McCole et al., 2014). They also included open chromatin, since UCEs have been linked to transcriptional activity (reviewed in Baira et al., 2008; Fabris and Calin, 2017; Harmston et al., 2013), repetitive elements, which UCEs avoid (Bejerano et al., 2004; Chiang et al., 2008; Derti et al., 2006; McCole et al., 2014), and GC content, which is associated with the positions of CNVs (Koren et al., 2012). We divided the genome into equally sized bins and, because domains and the nine control features span a vast range of sizes, our analyses involved multiple iterations using a range of bin sizes (20, 50, and 100 kb). Within each bin, the fraction of sequence occupied by each control feature was calculated, as was that of UCEs, except in the case of GC content, where it was calculated as the fraction of G + C (Experimental Procedures). Genome-wide correlations were then determined with respect to each control within each bin.

Using pairwise Spearman correlation coefficients and associated p values for the strength of correlation, we first determined that UCEs are significantly and positively associated with pooled

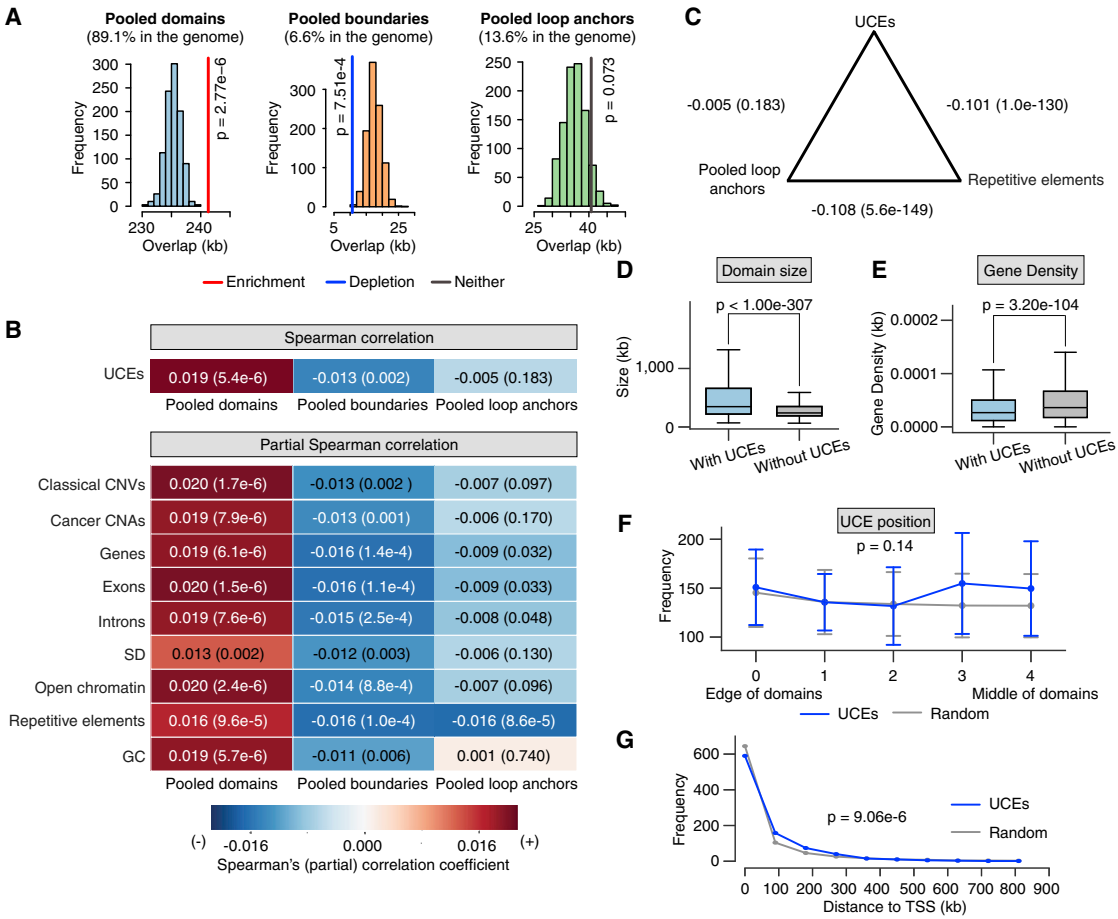


Figure 2. UCEs Are Enriched in Pooled Domains, Depleted from Pooled Boundaries, and Indifferent to Pooled Loop Anchors

(A) In the case of pooled domains, the observed overlap (colored vertical line) of UCEs is significantly greater than the expected overlaps (red line; $p = 2.76 \times 10^{-6}$, obs/exp = 1.025). For pooled boundaries, the observed overlap is significantly below expectation (blue line; $p = 7.51 \times 10^{-4}$, obs/exp = 0.609). Observed overlap between UCEs and pooled loop anchors does not deviate significantly from expectation (gray line; p = 0.073, obs/exp = 1.124). Note that pooled domains may include pooled boundaries, because the boundaries of some cell types may be organized as domains in other cell types.

(B) Correlation analyses. Spearman correlation: using pairwise Spearman correlation and splitting the genome into 50-kb bins, the representation of UCEs is positively correlated with that of pooled domains ($p = 5.4 \times 10^{-6}$), negatively correlated with that of pooled boundaries (p = 0.002), and not significantly correlated with that of pooled loop anchors (p = 0.183). Partial Spearman correlation: the positive and negative correlations between the positions of UCEs and pooled domains (first column), and negative correlation between the positions of UCEs and pooled boundaries (second column) remain significant even after accounting for the correlation between the positions of UCEs and nine control genomic features. The representation of UCEs and pooled loop anchors (third column) is not significantly positively nor negatively correlated except when controlling for repetitive elements, explored in (C).

(C) Although UCEs and pooled loop anchors are not significantly correlated with each other (p = 0.183), pairwise correlation analyses of both UCEs and pooled loop anchors show a highly significant negative correlation with repetitive elements ($p = 1.0 \times 10^{-130}$ and $p = 5.6 \times 10^{-149}$, respectively). In (B) and (C), Spearman (partial) correlation coefficients are reported in each box and by a heatmap; p values are reported in parentheses.

(D and E) Domains containing UCEs are significantly larger (D) ($p < 1.00 \times 10^{-307}$) and relatively gene sparse (E) ($p = 3.20 \times 10^{-104}$) as compared to domains without UCEs. p values were calculated by Mann-Whitney U test; box: interquartile range; whisker: 1.5 × interquartile range.

(F) UCEs are positioned roughly evenly across domains, with the distribution differing insignificantly from expectation (p = 0.14; K-S test; error bar: SD).

(G) UCEs are positioned further than expected from the nearest transcription start site (TSS) ($p = 9.06 \times 10^{-6}$; Anderson-Darling test).

See also Figure S1 and Tables S2 and S3.

domains ($p = 5.4 \times 10^{-6}$; Figure 2B), significantly negatively correlated with pooled boundaries (p = 0.002; Figure 2B), and not correlated with pooled loop anchors (p = 0.183; Figure 2B). These results correspond well to UCE enrichment, depletion, and neither enrichment in nor depletion from pooled domains, boundaries, and loops, respectively (Figure 2A). Then, using a partial correlation approach, we asked whether these correlations, or lack thereof, are influenced by co-correlation with any of the nine control genomic features. With a bin size of 50 kb, the correlation between UCEs and pooled domains remains significantly positive in all cases, indicating that it is robust to contributions from the control features (Figure 2B). Similarly, the negative correlation between UCEs and pooled boundaries remains robust to all control features (Figure 2B). As for pooled

loop anchors, the correlation with UCEs is insignificant in all cases but one, consistent with UCEs being neither enriched nor depleted in pooled loop anchors (Figure 2B). The one exception pertains to repetitive elements, where the correlation is significantly negative. Investigating this further, we discovered a negative correlation between UCEs and repetitive elements ($p = 1.0 \times 10^{-130}$; Figure 2C), which is unsurprising, as UCEs are non-repetitive (Bejerano et al., 2004; Chiang et al., 2008; Derti et al., 2006) and avoid insertions of repetitive elements (Zhang et al., 2017). We also uncovered a strong negative correlation between pooled loop anchors and repetitive elements ($p = 5.6 \times 10^{-149}$; Figure 2C), which may again be expected as loop anchors are derived from Hi-C analyses that exclude reads from repetitive regions (Rao et al., 2014). Thus, while a significant negative correlation exists between UCEs and pooled loop anchors, it may be secondary to the strong negative correlation between repetitive elements and both UCEs and pooled loop anchors. Altering the sizes of the genomic bins to 20 kb (Figure S1A) and 100 kb (Figure S1B) produced very similar results. Taken together, the positioning of UCEs relative to domains, boundaries, and loop anchors is robust to co-correlation with nine other genomic features.

UCEs Occur Evenly across Large, Gene-Sparse Domains and Are Somewhat Distant from Transcription Start Sites

We next investigated the properties of domains containing UCEs. Considering all cell types together, we found that domains containing UCEs are larger than those without UCEs ($p < 1.00 \times 10^{-307}$; Figure 2D; Table S3A) and have a lower density of genes ($p = 3.20 \times 10^{-104}$; Figure 2E; Table S3C), with a distribution of UCEs being relatively even across domains and not significantly different to that of random control regions (Experimental Procedures) ($p = 0.14$; Figure 2F; Table S3E). Nevertheless, we found slightly fewer UCEs within 100 kb of the nearest transcription start site (TSS), but slightly more 100–300 kb from the nearest TSS, compared to within random control regions ($p = 9.06 \times 10^{-6}$; Figure 2G; Table S3G). With regard to domain size, gene density, UCE position, and distance from UCE to TSS (Table S3), the domains of individual cell types followed the trends observed for all domains combined, except for the domains of HUVEC and IMR90 cells as described by Rao et al. (2014), where UCEs tended to occupy the center of domains (Table S3E). In summary, UCEs are arranged roughly evenly across large, gene-sparse domains and are slightly distanced from TSSs, perhaps highlighting a potential role for UCEs in the maintenance of genome structure.

Positioning of UCEs within Hi-C Annotations Is Conserved between Human and Mouse

Since UCEs are defined by their extreme evolutionary conservation between species, we next asked whether the relationships observed between UCEs and domains, boundaries, and loop anchors in the human genome are conserved in the mouse genome. Accordingly, we turned to the 893 mouse orthologs (Table S2D) of our human UCEs and three Hi-C studies (Dixon et al., 2012; Fraser et al., 2015; Rao et al., 2014), addressing mouse embryonic stem cells (mESCs), blood (B-lymphoblasts), neuronal precursor cells (NPCs), post-mitotic neurons, and cortical tissue (Table S1). We found that the relationships of UCEs to domains, boundaries, and loop anchors are evolutionarily conserved. For domains, we examined six datasets covering between 29.0% (lymphoblasts from Rao et al.) and 92.5% of the genome (neurons from Fraser et al. [Table S1]). All six datasets are significantly enriched for UCEs ($5.52 \times 10^{-10} \leq p \leq 0.002$; $1.020 \leq$ obs/exp ≤ 1.260; Figure S1C; Table S2B), with domains containing UCEs being larger ($p = 1.24 \times 10^{-240}$; Table S3B), and more gene sparse ($p = 1.37 \times 10^{-13}$; Table S3D) as compared to domains without UCEs, recapitulating our findings for human domains. For boundaries, we examined a dataset described by Dixon et al. (2012) to be common to both mESC and cortex tissue and covering 8.1% of the genome, calling this dataset "mouse common boundaries" (Table S1). This dataset shows significant depletion for UCEs ($p = 4.68 \times 10^{-7}$, obs/exp = 0.452; Figure S1C; Table S2B). Finally, we examined one dataset representing loop anchors in lymphocytes from Rao et al. This dataset covers 2.2% of the genome (Table S1) and is neither enriched in nor depleted of UCEs ($p = 0.090$, obs/exp = 0.701; Figure S1C; Table S2B).

When UCEs Are Found in Boundaries and Loop Anchors, They Show an Excess of Exonic UCEs Associated with RNA Processing

Having established that UCEs are differentially associated with domains, boundaries, and loop anchors, we queried whether specific subsets of UCEs might be driving the associations. We examined intergenic, intronic, and exonic UCEs separately, since these subdivisions have behaved distinctly in our previous studies; for example, CNVs are more depleted for intergenic and intronic UCEs than for exonic UCEs (Chiang et al., 2008; Derti et al., 2006; McCole et al., 2014). First, we examined all of the individual datasets for domains as well as pooled domains and found that, in all cases, there is no significant deviation from expected in the observed proportions of intergenic, intronic, and exonic UCEs ($0.131 \leq p \leq 0.892$; Figure 3A; Tables S3A and S3B). That the proportions of UCEs in pooled domains are the same as those within the entire UCE dataset is not surprising, as pooled domains contain all 896 UCEs, including boundary UCEs, since boundaries in some cell types are organized as domains in other cell types.

For pooled boundaries, the distribution of intergenic, intronic, and exonic UCEs deviates significantly from that of the full set of UCEs ($p = 3.46 \times 10^{-7}$; Figure 3B; Table S4A). We found a depletion of intergenic and intronic UCEs, with 21.6% (8 out of 37) and 21.6% (8 out of 37), respectively, in boundaries, as compared to the expected 32.4% and 47.0%, respectively. In contrast, exonic UCEs are overrepresented, with 56.8% (21 out of 37) in pooled boundaries, while making up only 20.6% of all UCEs. The overrepresentation of exonic UCEs is especially striking since the majority (52.7%) of pooled boundary DNA is intronic (109 Mb), with only a small fraction (5.3%) being exonic (11 Mb) ($p = 2.89 \times 10^{-44}$; Experimental Procedures; Table S4C).

We also found significant deviation of the proportions of intergenic, intronic, and exonic UCEs in pooled loop anchors ($p = 2.79 \times 10^{-4}$; Figure 3C; Table S4A). Intergenic and intronic

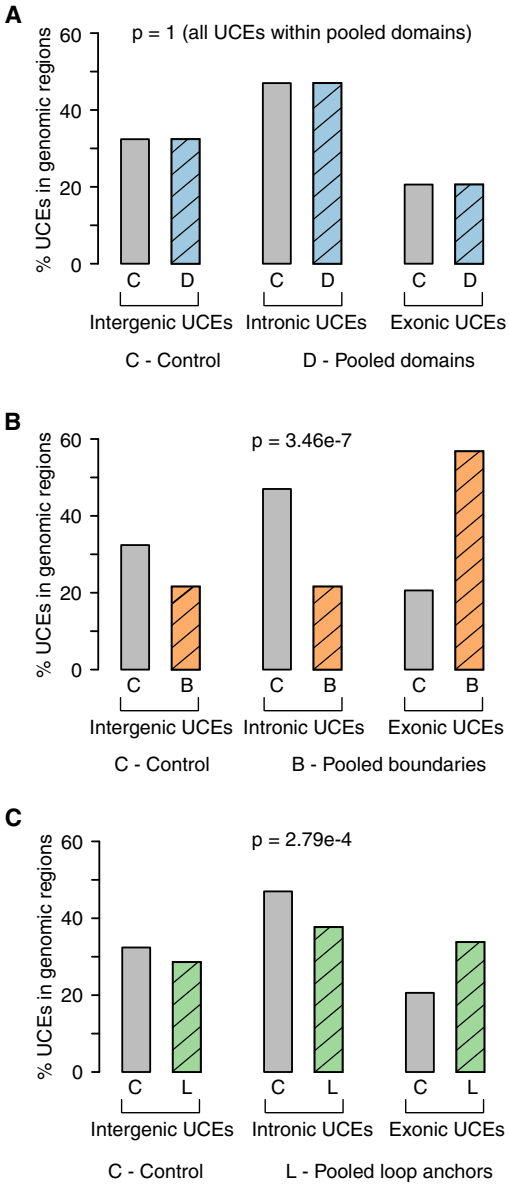


Figure 3. Underrepresentation of Intronic and Intergenic UCEs in Pooled Boundaries and Loop Anchors Are Accompanied by Overrepresentation of Exonic UCEs

Proportions of intergenic, intronic, and exonic UCEs that overlap pooled domains (A, blue), boundaries (B, orange), and loop anchors (C, green) compared to the full set of 896 UCEs as a control (gray).

(A) Pooled domains that are not significantly different compared to the control set since all UCEs fall within pooled domains, so no p value is calculated.

(B and C) Pooled boundaries (B) (chi-square test, $p = 3.46 \times 10^{-7}$) and pooled loop anchors (C) (chi-square test, $p = 2.79 \times 10^{-4}$) both show a significant overrepresentation of exonic UCEs and an underrepresentation of intronic and intergenic UCEs, as compared with the full UCE set.

See also Figures S2 and S3, and Tables S4 and S5.

UCEs represent only 28.6% (44 out of 154) and 37.7% (58 out of 154) of UCEs, respectively, whereas 32.4% and 47.0% of the full UCE set are intergenic and intronic, respectively. As in pooled boundaries, exonic UCEs are overrepresented at 33.8% (52 out of 154) as compared to 20.6% of all UCEs. These proportions deviate significantly from expectation based on the sequence composition of pooled loop anchors, which is 47.0% intronic and only 5.02% exonic ($p = 5.38 \times 10^{-60}$; Table S4C). These results point to intronic, and, to some extent, intergenic, UCEs as drivers of depletion from pooled boundaries and to exonic UCEs as the dominant type of UCE within both pooled boundaries and loop anchors.

We next used the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010) and discovered that exonic UCEs in pooled boundaries and pooled loop anchors are enriched for gene ontology (GO) terms associated with RNA processing (Figures S2A and S2B), and this is in line with previous reports that exonic UCEs are associated with RNA processing, including splicing (Baira et al., 2008; Bejerano et al., 2004; Lareau and Brenner, 2015; Lareau et al., 2007; Lupiáñez et al., 2016; Ni et al., 2007; Pirnie et al., 2017; Rödelsperger et al., 2009). Considering further the structure of exonic UCEs themselves, we found that 76% (16 out of 21; Table S5A) and 82% (43 out of 52; Table S5B) of exonic UCEs in pooled boundaries and loop anchors, respectively, partially overlap introns and hence cover splice sites, as compared to 57% in the full set of exonic UCEs (107 out of 185; Table S2E). Thus, while exonic UCEs in pooled boundaries are not enriched for splice sites ($p = 0.07$; Table S5A), those in pooled loop anchors are ($p = 1.82 \times 10^{-4}$; Table S5B). These results suggest a two-layered association of UCEs with RNA processing, whereby UCEs are associated with genes involved in RNA processing and UCEs may also help the splicing of these very same genes. This double association is particularly prominent in loop anchors, suggesting that UCEs in loop anchors may assist in particular splicing mechanisms.

UCEs within Domains That Are Shared in Many Cell Types Are Associated with Kidney-Related Processes

While domains vary between cell types, studies suggest that at least 50% are shared across cell types (Dixon et al., 2012; Fraser et al., 2015; Rao et al., 2014). Thus, we next focused on UCEs that occur within domains common across multiple cell types; these might address the functional significance underlying the enrichment of UCEs. We first identified 124 UCEs that overlap domains in all ten individual human datasets across diverse cell types (Table S1), calling these "human invariant domain UCEs" (Table S5C); such UCEs overlap between 30 and 51 domains depending on the individual dataset (Table S3A). For mouse, we identified 310 UCEs that overlap domains identified by all six mouse datasets (Table S1), calling these "mouse invariant domain UCEs" (Table S5D). Using GREAT, these human and mouse invariant domain UCEs were compared to the full UCE sets in humans and mouse, respectively, revealing a surprising association with kidney-related GO terms for human invariant domain UCEs (Figure S3A). Terms related to kidney biology were also obtained in the case of mouse UCEs, although, here, other terms were obtained as well, some with greater significance (Figure S3B). These findings are corroborated by the association with kidney-related processes, as well as neuronal development, of the 74 UCEs shared between the human and

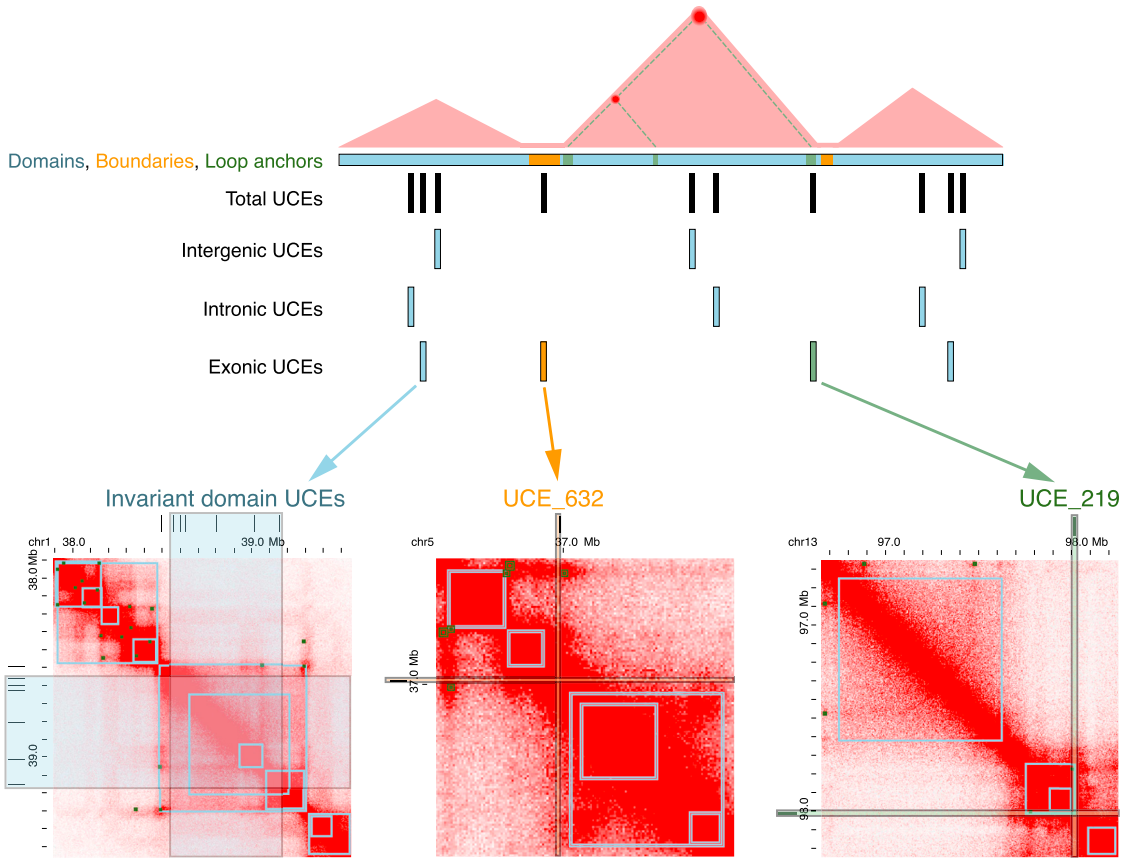


Figure 4. Schematic Representation Summarizing the Relationship between Chromosome Organization and UCEs

Top, domains (blue) are enriched in UCEs, boundaries (orange) are depleted, and loop anchors (green) are neither enriched nor depleted. Bottom, examples of UCEs in each of the three genomic features as defined by Hi-C annotation of human GM12878 cells (Rao et al., 2014) using the Juicebox tool (Durand et al., 2016). Left, invariant domain UCEs; middle, UCE in a boundary; right, UCE in a loop anchor. Domain and loop anchor calls (squares in gray outline) are indicated on the heatmaps as available in Juicebox (Durand et al., 2016). Numbers of UCEs are not representative of their true occupancies.

mouse invariant domain UCE datasets (Figures S3C and S3D; Table S5E). Interestingly, domains containing the 74 shared invariant domain UCEs were smaller and more gene rich than were all UCE-containing domains ($p = 5.45 \times 10^{-4}$ and $p = 5.52 \times 10^{-10}$, respectively; Tables S3A and S3C), suggesting that these domains may be functionally different from domains in general, perhaps with UCEs specifically involved in regulating kidney and neuronal development. Of note, a recent study has demonstrated that UCEs are required for normal brain development (Dickel et al., 2018). In brief, functions related to kidney and neuronal development might be a feature of UCEs within domains shared among diverse cell types.

DISCUSSION

Our findings reveal a non-random UCE distribution among three main arenas of three-dimensional genome organization, with UCEs being enriched in domains, depleted from boundaries, and indifferent to loop anchors (Figure 4). Furthermore, domains containing UCEs are larger and less gene rich than those without UCEs, and while UCEs are distributed relatively evenly across domains, they are slightly further away from TSS than expected, suggesting that UCEs may help maintain the structure of large domains in a role distinct from that of gene regulation. The UCEs that do occupy boundaries and loop anchors display an overrepresentation of exonic UCEs, and in loop anchors, those UCEs are enriched for overlap with splice sites, suggesting a specific involvement of loop anchors containing UCEs in splicing. With respect to UCEs in domains that do not vary between cell types, they are, as a group, significantly associated with kidney-related and neuronal gene ontologies.

These findings tying UCEs to genome organization are especially intriguing in light of the proposal that UCEs may contribute to genome integrity through yet another potent organizational feature of genomes—allelic and homolog pairing (Chiang et al., 2008; Derti et al., 2006; Kritsas et al., 2012; McCole et al., 2014; Vavouri et al., 2007). Indeed, they raise the question of whether UCEs contribute to the establishment of domains, and/or whether the evolution of a domain promotes the fixation of UCEs within the domain. Consistent with this, Harmston et al. (2017) recently reported that clusters of CNEs predict the span of domains, suggesting that CNEs might be involved in

chromatin folding. For example, since some UCEs embody enhancer activity (Bejerano et al., 2004; Bhatia et al., 2013; Lampe et al., 2008; McBride et al., 2011; Pauls et al., 2012; Pennacchio et al., 2006; Poitras et al., 2010; Poulin et al., 2005; Sandelin et al., 2004; Vavouri et al., 2007; Visel et al., 2008; Warnefors et al., 2016; Woolfe et al., 2005) and, thus, are likely to participate in enhancer-promoter interactions, might that activity help define chromosomal contacts? Separately, but not exclusively, might selection against changes that disrupt chromosomal domains promote sequence invariance and, thus, ultraconservation? Specifically, if, as we have proposed (Chiang et al., 2008; Derti et al., 2006; McCole et al., 2014), rearrangements that disrupt the pairing of allelic UCEs are culled, then UCEs will contribute to the structural invariance of genomic regions in which they lie. In this way, UCEs may have enhanced the capacity of certain regions to evolve the intra-regional contacts that, today, define contact domains.

The strong association of invariant domain UCEs with kidney-related and neuronal GO categories was intriguing and merits further exploration. In this light, it may be noteworthy that evolution of the kidney has been argued to be an early defining process in the emergence of vertebrates (Ditrich, 2007). If so, that evolution may have benefitted from the genome stability provided by UCEs.

Our studies have also shown that, while boundaries are generally depleted of UCEs, 21 of the 37 UCEs found in boundaries are exonic, constituting an enrichment of exonic UCEs in boundaries. Of the 21 boundary exonic UCEs, two (UCEs 632 and 633) are in the NIPBL gene, which is a cohesin loading factor that, when mutated, leads to a developmental disorder known as Cornelia de Lange syndrome (Strachan, 2005). Given that cohesin binding is implicated in sister chromatid cohesion and gene expression (Merkenschlager, 2010; Merkenschlager and Odom, 2013), ultraconservation within NIPBL may speak to this gene's importance in genome structure and function. Indeed, a recent study demonstrated that depletion of NIPBL in mouse affects reorganization of chromosome folding (Schwarzer et al., 2017). Furthermore, the evolutionarily conserved position of the NIPBL gene within boundaries may suggest that the lack of three-dimensional associations across a boundary may also be important for its expression.

Turning to loop anchors, their lack of enrichment in UCEs chimes with other findings arguing that loops are evolutionarily dynamic (Vietri Rudan et al., 2015). Their dynamic nature is consistent with the malleability of enhancers over evolutionary time and thus, also, of enhancer-promoter interactions, both of which make the lack of enrichment for UCEs in loop anchors unsurprising. Indeed, unconstrained enhancers may more easily accommodate tissue-specific (Lonfat et al., 2014) or even species-specific regulatory programs (Vietri Rudan et al., 2015).

To conclude, our data describe the pattern of relationships between ultraconservation of DNA sequence and three types of chromosome organization, with domains enriched in UCEs, boundaries being depleted, and loops being neither enriched nor depleted. More generally, they illustrate how different structural arenas of genome organization display distinct degrees of flexibility or stability over evolutionary timescales, as measured by ultraconservation.

EXPERIMENTAL PROCEDURES

Depletion or Enrichment Analysis of UCEs in Specific Genomic Regions

The enrichment or depletion of UCEs in genomic regions of interest such as domains, boundaries, and loop anchors was assessed using established methods previously reported in our publications (Chiang et al., 2008; Derti et al., 2006; McCole et al., 2014). Briefly, observed overlap between UCEs and regions of interest were compared to a distribution of expected overlaps produced using 1,000 randomized sets of elements that match UCEs in number and length. Deviation of the observed overlap from the expected overlaps is indicated by the obs/exp ratio, and statistical significance was determined by a Z-test where appropriate.

Correlation Analyses

The genome was divided into bins of equal sizes. Within each bin, the fraction of sequence occupied by each control feature was calculated, as was that of UCEs, except in the case of GC content, where it was calculated as the fraction of G + C. Then genome-wide correlations within each bin were preformed among feature densities or GC content. The Spearman correlation coefficients and matching p values were provided.

Analyses of Domains Containing UCEs

Custom scripts were used to calculate metrics and p values for domain size, gene density (Mann-Whitney U test), UCE position within domains (K-S test), and distances to the nearest TSS (Anderson Darling test). Expected distributions were defined using 100 sets of regions matched to UCE number and position generated as for Depletion or Enrichment Analysis of UCEs in Specific Genomic Regions.

Distribution of Intergenic, Intronic, and Exonic UCEs That Overlap Domains, Boundaries, and Loop Anchors

The distribution of intergenic, intronic, and exonic UCEs that overlap feature of interest, i.e., either domains, boundaries, or loop anchors (reported in Table S2A), was compared to the full set of 896 UCEs using a χ^2 test.

To determine the proportions of domains, boundaries, and loop anchors that are intergenic, intronic, and exonic, the overlap between two features (i.e., intergenic regions in domains) was calculated using bedtools intersect (Quinlan and Hall, 2010).

Scripts

Custom scripts associated with this study are available at https://github.com/rmccole/UCEs_genome_organization.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, and five tables and can be found with this article online at https://doi.org/10.1016/j.celrep.2018.06.031.

ACKNOWLEDGMENTS

We thank Brian J. Beliveau, Chamith Y. Fonseka, Roxana Tarnita, Kaia Mattioli, Tommy Tullius, and all members of the Wu laboratory for valuable and insightful discussions. We apologize to authors whose work we could not cite due to length restrictions. This work was supported by a William Randolph Hearst Foundation grant to R.B.M., an EMBO Long-Term Fellowship (ALTF 186-2014) to J.E., and awards to C.-t.W. from NIH (DP1GM106412, R01GM123289-01, and R01HD091797) and Harvard Medical School.

AUTHOR CONTRIBUTIONS

R.B.M, J.E., and C.-t.W. designed the research. R.B.M, J.E., and W.S. performed the research. R.B.M, J.E., and W.S. analyzed the data. R.B.M, J.E., and C.-t.W. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 11, 2017
Revised: May 9, 2018
Accepted: June 7, 2018
Published: July 10, 2018

REFERENCES

Baira, E., Greshock, J., Coukos, G., and Zhang, L. (2008). Ultraconserved elements: genomics, function and disease. RNA Biol. *5*, 132–134.

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.

Bhatia, S., Bengani, H., Fish, M., Brown, A., Divizia, M.T., de Marco, R., Damante, G., Grainger, R., van Heyningen, V., and Kleinjan, D.A. (2013). Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. Am. J. Hum. Genet. *93*, 1126–1134.

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. Nat. Rev. Genet. *17*, 661–678.

Chiang, C.W., Derti, A., Schwartz, D., Chou, M.F., Hirschhorn, J.N., and Wu, C.T. (2008). Ultraconserved elements: analyses of dosage sensitivity, motifs and boundaries. Genetics *180*, 2277–2293.

Corces, M.R., and Corces, V.G. (2016). The three-dimensional cancer genome. Curr. Opin. Genet. Dev. *36*, 1–7.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

Denker, A., and de Laat, W. (2016). The second decade of 3C technologies: detailed insights into nuclear organization. Genes Dev. *30*, 1357–1382.

Derti, A., Roth, F.P., Church, G.M., and Wu, C.T. (2006). Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. Nat. Genet. *38*, 1216–1220.

Dickel, D.E., Ypsilanti, A.R., Pla, R., Zhu, Y., Barozzi, I., Mannion, B.J., Khin, Y.S., Fukuda-Yuzawa, Y., Plajzer-Frick, I., Pickle, C.S., et al. (2018). Ultraconserved enhancers are required for normal development. Cell *172*, 491–499.e15.

Dimitrieva, S., and Bucher, P. (2012). Genomic context analysis reveals dense interaction network between vertebrate ultraconserved non-coding elements. Bioinformatics *28*, i395–i401.

Ditrich, H. (2007). The origin of vertebrates: a hypothesis based on kidney development. Zool. J. Linn. Soc. *150*, 435–441.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Gorkin, D.U., and Ren, B. (2016). Chromatin domains: the unit of chromosome organization. Mol. Cell *62*, 668–680.

Dong, X., Fredman, D., and Lenhard, B. (2009). Synorth: exploring the evolution of synteny and long-range regulatory interactions in vertebrate genomes. Genome Biol. *10*, R86.

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. *3*, 99–101.

Elgar, G., and Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet. *24*, 344–352.

Fabris, L., and Calin, G.A. (2017). Understanding the genomic ultraconservations: T-UCRs and cancer. Int. Rev. Cell Mol. Biol. *333*, 159–172.

Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C., Aitken, S., et al.; FANTOM Consortium (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol. Syst. Biol. *11*, 852.

Harmston, N., Baresic, A., and Lenhard, B. (2013). The mystery of extreme non-coding conservation. Philos. Trans. R. Soc. Lond. B Biol. Sci. *368*, 20130021.

Harmston, N., Ing-Simmons, E., Tan, G., Perry, M., Merkenschlager, M., and Lenhard, B. (2017). Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. Nat. Commun. *8*, 441.

Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.

Irimia, M., Tena, J.J., Alexis, M.S., Fernandez-Miñan, A., Maeso, I., Bogdanovic, O., de la Calle-Mustienes, E., Roy, S.W., Gómez-Skarmeta, J.L., and Fraser, H.B. (2012). Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. Genome Res. *22*, 2356–2367.

Joyce, E.F., Erceg, J., and Wu, C.T. (2016). Pairing and anti-pairing: a balancing act in the diploid genome. Curr. Opin. Genet. Dev. *37*, 119–128.

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A.Z., Engström, P.G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., et al. (2007). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. *17*, 545–555.

Koeman, J.M., Russell, R.C., Tan, M.H., Petillo, D., Westphal, M., Koelzer, K., Metcalf, J.L., Zhang, Z., Matsuda, D., Dykema, K.J., et al. (2008). Somatic pairing of chromosome 19 in renal oncocytoma is associated with deregulated EGLN2-mediated [corrected] oxygen-sensing response. PLoS Genet. *4*, e1000176.

Koren, A., Polak, P., Nemesh, J., Michaelson, J.J., Sebat, J., Sunyaev, S.R., and McCarroll, S.A. (2012). Differential relationship of DNA replication timing to different forms of human mutation and variation. Am. J. Hum. Genet. *91*, 1033–1040.

Kritsas, K., Wuest, S.E., Hupalo, D., Kern, A.D., Wicker, T., and Grossniklaus, U. (2012). Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes. Genome Res. *22*, 2455–2466.

Kushawah, G., and Mishra, R.K. (2017). Ultraconserved sequences associated with HoxD cluster have strong repression activity. Genome Biol. Evol. *9*, 2049–2054.

Lampe, X., Samad, O.A., Guiguen, A., Matis, C., Remacle, S., Picard, J.J., Rijli, F.M., and Rezsohazy, R. (2008). An ultraconserved Hox-Pbx responsive element resides in the coding sequence of Hoxa2 and is active in rhombomere 4. Nucleic Acids Res. *36*, 3214–3225.

Lareau, L.F., and Brenner, S.E. (2015). Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. Mol. Biol. Evol. *32*, 1072–1079.

Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., and Brenner, S.E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature *446*, 926–929.

Liu, C., and Weigel, D. (2015). Chromatin in 3D: progress and prospects for plants. Genome Biol. *16*, 170.

Lonfat, N., Montavon, T., Darbellay, F., Gitto, S., and Duboule, D. (2014). Convergent evolution of complex regulatory landscapes and pleiotropy at Hox loci. Science *346*, 1004–1006.

Lupiáñez, D.G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: how alterations of chromatin domains result in disease. Trends Genet. *32*, 225–237.

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.

McBride, D.J., Buckle, A., van Heyningen, V., and Kleinjan, D.A. (2011). DNaseI hypersensitivity and ultraconservation reveal novel, interdependent long-range enhancers at the complex Pax6 cis-regulatory region. PLoS One *6*, e28616.

McCole, R.B., Fonseka, C.Y., Koren, A., and Wu, C.T. (2014). Abnormal dosage of ultraconserved elements is highly disfavored in healthy cells but not cancer cells. PLoS Genet. *10*, e1004646.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.

Merkenschlager, M. (2010). Cohesin: a global player in chromosome biology with local ties to gene regulation. Curr. Opin. Genet. Dev. *20*, 555–561.

Merkenschlager, M., and Odom, D.T. (2013). CTCF and cohesin: linking gene regulatory elements with their targets. Cell *152*, 1285–1297.

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.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Pauls, S., Smith, S.F., and Elgar, G. (2012). Lens development depends on a pair of highly conserved Sox21 regulatory elements. Dev. Biol. *365*, 310–318.

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.

Pirnie, S.P., Osman, A., Zhu, Y., and Carmichael, G.G. (2017). An ultraconserved element (UCE) controls homeostatic splicing of ARGLU1 mRNA. Nucleic Acids Res. *45*, 3473–3486.

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.

Polychronopoulos, D., Sellis, D., and Almirantis, Y. (2014). Conserved noncoding elements follow power-law-like distributions in several genomes as a result of genome dynamics. PLoS One *9*, e95437.

Polychronopoulos, D., Athanasopoulou, L., and Almirantis, Y. (2016). Fractality and entropic scaling in the chromosomal distribution of conserved noncoding elements in the human genome. Gene *584*, 148–160.

Polychronopoulos, D., King, J.W.D., Nash, A.J., Tan, G., and Lenhard, B. (2017). Conserved non-coding elements: developmental gene regulation meets genome organization. Nucleic Acids Res. *45*, 12611–12624.

Poulin, F., Nobrega, M.A., Plajzer-Frick, I., Holt, A., Afzal, V., Rubin, E.M., and Pennacchio, L.A. (2005). In vivo characterization of a vertebrate ultraconserved enhancer. Genomics *85*, 774–781.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Robyr, D., Friedli, M., Gehrig, C., Arcangeli, M., Marin, M., Guipponi, M., Farinelli, L., Barde, I., Verp, S., Trono, D., and Antonarakis, S.E. (2011). Chromosome conformation capture uncovers potential genome-wide interactions between human conserved non-coding sequences. PLoS One *6*, e17634.

Rödelsperger, C., Köhler, S., Schulz, M.H., Manke, T., Bauer, S., and Robinson, P.N. (2009). Short ultraconserved promoter regions delineate a class of preferentially expressed alternatively spliced transcripts. Genomics *94*, 308–316.

Sandelin, A., Bailey, P., Bruce, S., Engström, P.G., Klos, J.M., Wasserman, W.W., Ericson, J., and Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics *5*, 99.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., H Haering, C., Mirny, L., and Spitz, F. (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51–56.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the *Drosophila* genome. Cell *148*, 458–472.

Strachan, T. (2005). Cornelia de Lange syndrome and the link between chromosomal function, DNA repair and developmental gene regulation. Curr. Opin. Genet. Dev. *15*, 258–264.

Sun, H., Skogerbø, G., and Chen, R. (2006). Conserved distances between vertebrate highly conserved elements. Hum. Mol. Genet. *15*, 2911–2922.

Sun, H., Skogerbø, G., Zheng, X., Liu, W., and Li, Y. (2009). Genomic regions with distinct genomic distance conservation in vertebrate genomes. BMC Genomics *10*, 133.

Valton, A.L., and Dekker, J. (2016). TAD disruption as oncogenic driver. Curr. Opin. Genet. Dev. *36*, 34–40.

Vavouri, T., Walter, K., Gilks, W.R., Lehner, B., and Elgar, G. (2007). Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biol. *8*, R15.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D.T., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. *10*, 1297–1309.

Visel, A., Prabhakar, S., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E.M., and Pennacchio, L.A. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat. Genet. *40*, 158–160.

Warnefors, M., Hartmann, B., Thomsen, S., and Alonso, C.R. (2016). Combinatorial gene regulatory functions underlie ultraconserved elements in *Drosophila*. Mol. Biol. Evol. *33*, 2294–2306.

Weischenfeldt, J., Dubash, T., Drainas, A.P., Mardin, B.R., Chen, Y., Stütz, A.M., Waszak, S.M., Bosco, G., Halvorsen, A.R., Raeder, B., et al. (2017). Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat. Genet. *49*, 65–74.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al. (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. *3*, e7.

Zhang, Y., Li, S., Abyzov, A., and Gerstein, M.B. (2017). Landscape and variation of novel retroduplications in 26 human populations. PLoS Comput. Biol. *13*, e1005567.

Cell Reports, Volume *24*

Supplemental Information

Ultraconserved Elements Occupy Specific Arenas of Three-Dimensional Mammalian Genome Organization

Ruth B. McCole, Jelena Erceg, Wren Saylor, and Chao-ting Wu

Supplemental Figures

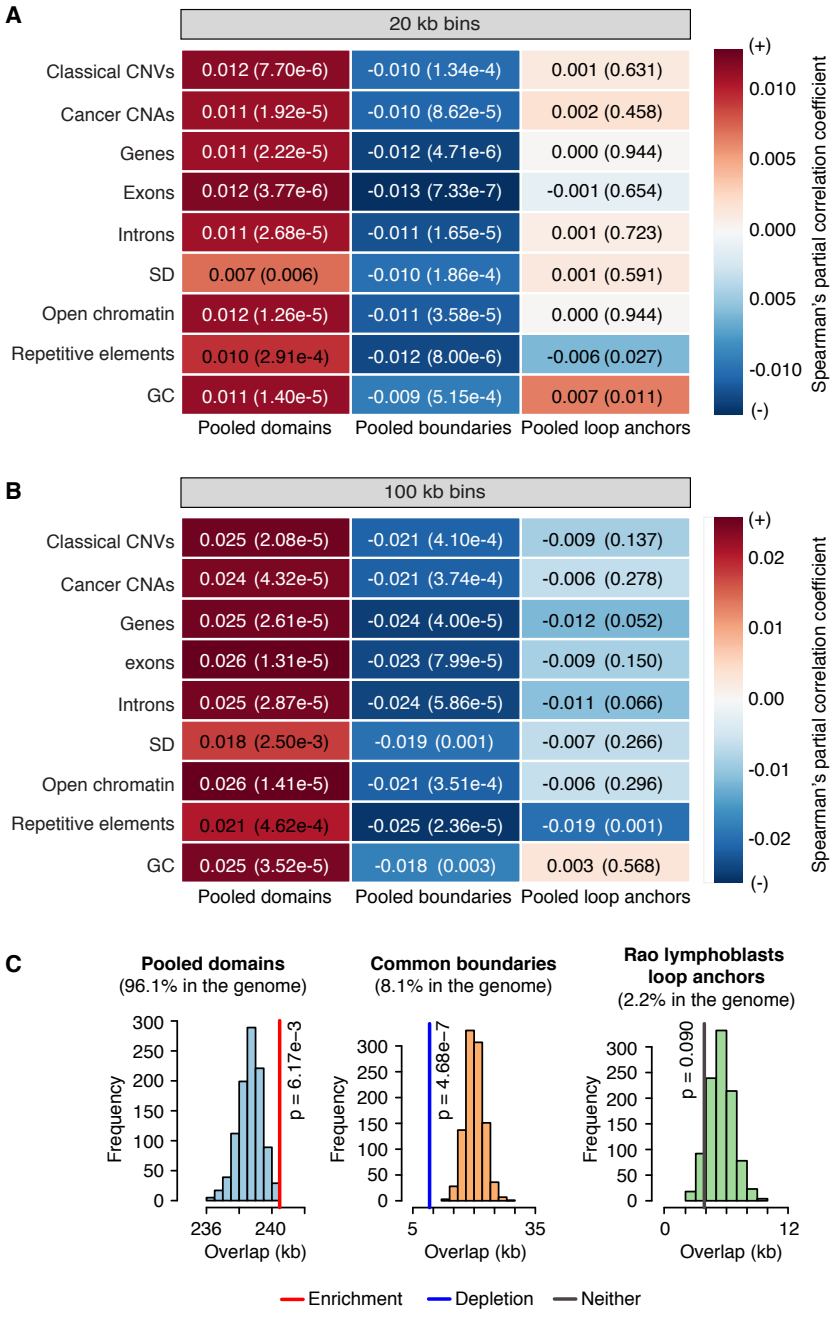


Figure S1. Partial correlation and mouse depletion/enrichment analyses, Related to Figure 2.

(A-B) The positive correlation between representation of UCEs and pooled domains (first column), and negative correlation between UCEs and pooled boundaries (second column) remain even after accounting for co-correlation between the positions of UCEs and nine control genomic features. These findings are robust to the selected size of genomic bin, either (A) 20 kb or (B) 100 kb, in which the number of base pairs encompassed by each genomic feature was assessed; p values are provided in parentheses. (C) Similar to the situation in humans (Figure 2A), UCEs are enriched in mouse pooled domains (red line; $p=6.17\times10^{-3}$, obs/exp=1.007), depleted in (common) boundaries (blue line; $p=4.68\times10^{-7}$, obs/exp=0.452), and neither enriched nor depleted in loop anchors (grey line; p=0.090, obs/exp=0.701). Note that, pooled domains may include common boundaries, because the boundaries of some cell types may be organized as domains in other cell types.

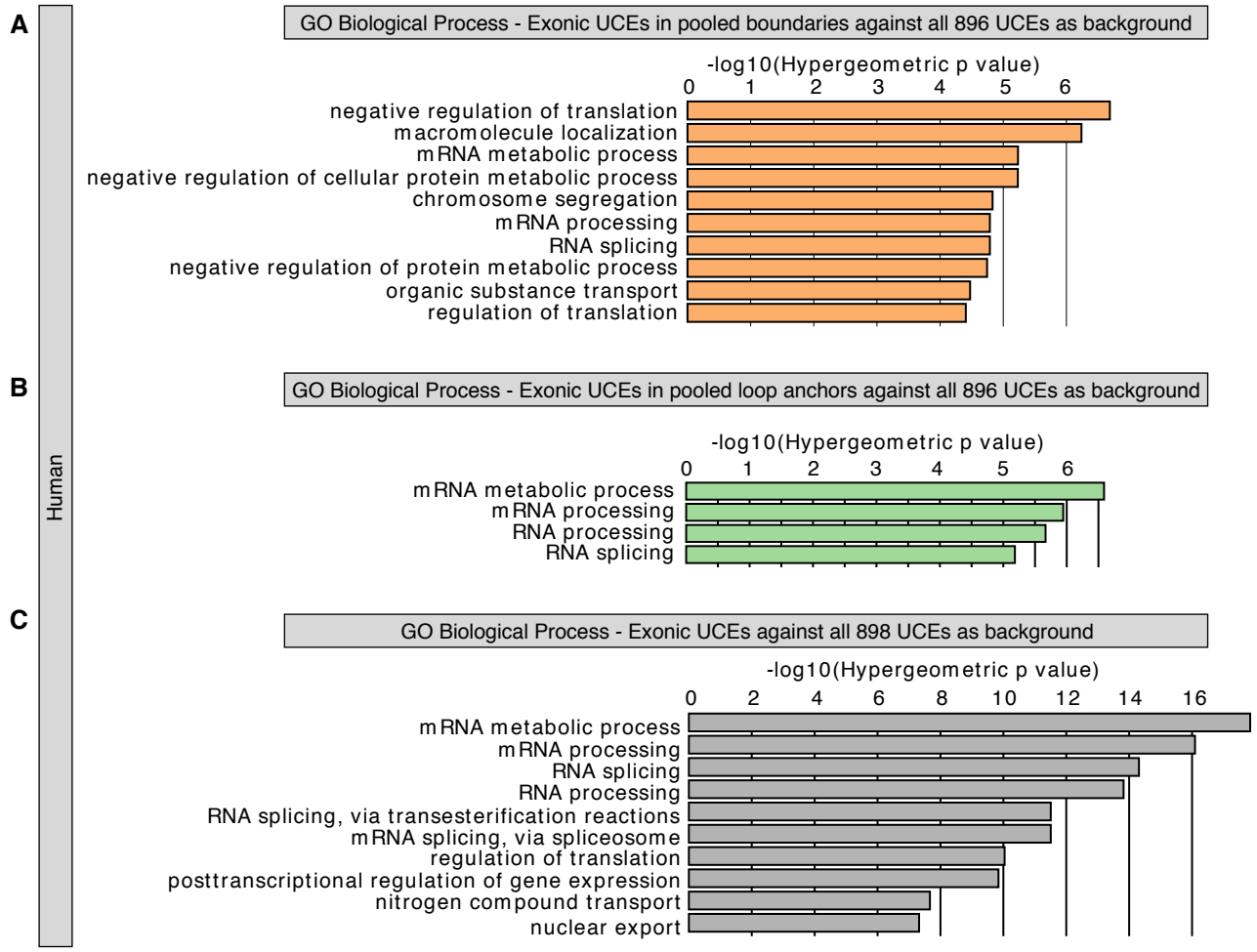


Figure S2. Exonic UCEs that overlap pooled boundaries or pooled loop anchors are functionally associated with RNA processing GO terms, Related to Figure 3.

Gene ontology (GO) terms associated with genes of human exonic UCEs that overlap either pooled boundaries (A) or pooled loop anchors (B) are involved in RNA processing. This association is not unique to exonic UCEs overlapping pooled boundaries and loop anchors since it is also observed with all exonic UCEs (C) when using the full set of 896 UCEs as a background.

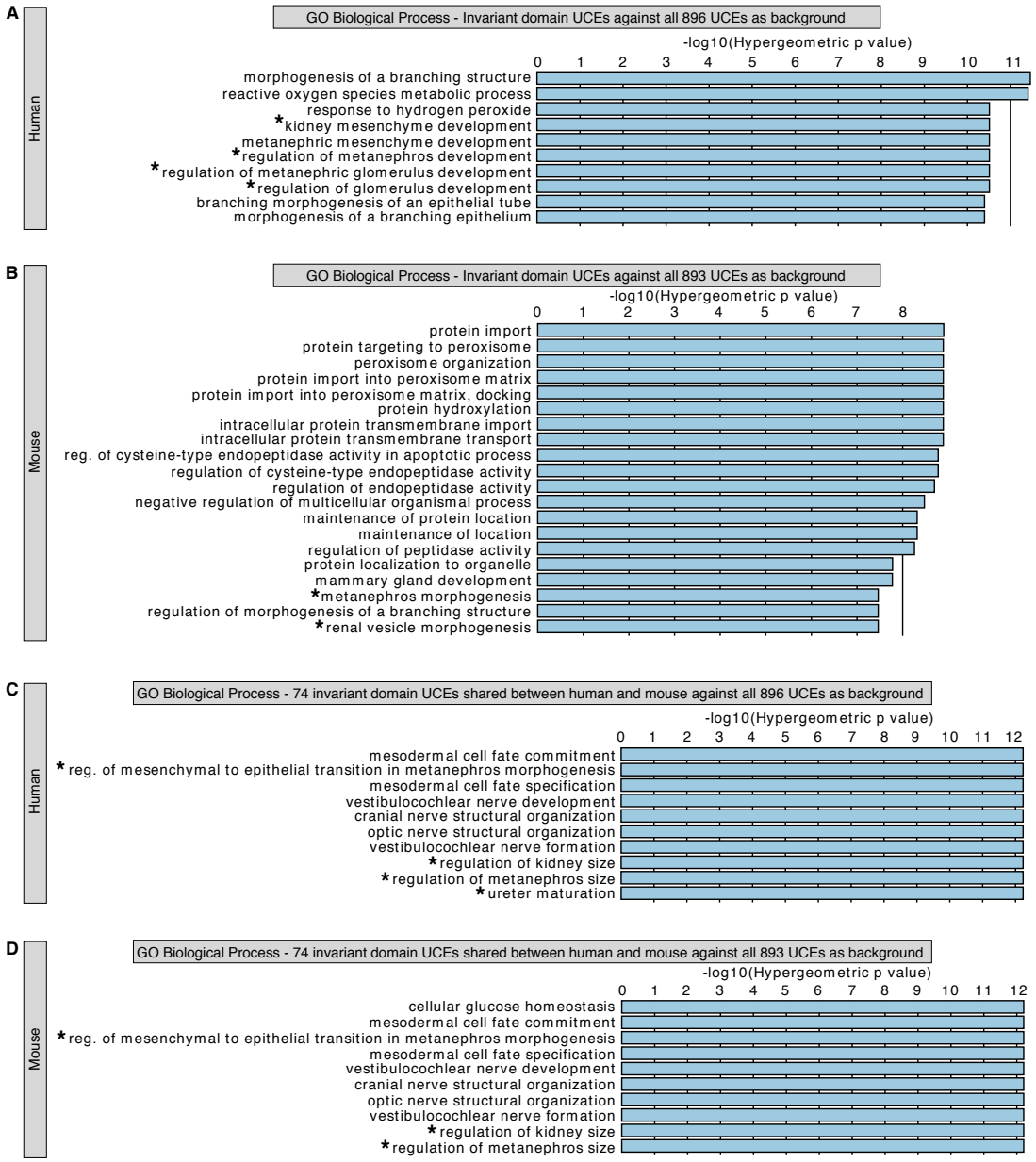


Figure S3. Invariant domain UCEs are associated with kidney-related processes in both human and mouse, Related to Figure 3.

(A-B) Gene ontology (GO) terms associated with genes of invariant domain UCEs were obtained applying GREAT (McLean et al., 2010) against the full set of UCEs. Both human (A) and mouse (B) invariant domain UCEs are linked to kidney-related processes. (C-D) The association with kidney-related processes and neuronal development of 74 invariant domain UCEs shared between human and mouse is observed when assessing either human (A) or mouse (B) UCEs set against all UCEs as a background. The asterisk indicates GO terms that are unambiguously associated with kidney development.

Supplemental Tables

Table S1. Hi-C datasets, Related to Figure 1. *See separate excel sheet.*

Table S2. Depletion/enrichment analysis, Related to Figures 1 and 2. *See separate excel sheet.*

(A) Analysis of UCEs representing the union of Human-Mouse-Rat (HMR), Human-Dog-Mouse (HDM), and Human–Chicken (HC) UCEs, as in Derti *et al.* 2006, from domain, boundary, and loop anchor datasets. (B) Analysis of 893 HMR-HDM-HC UCEs from mouse domain, boundary, and loop anchor datasets. (C) Coordinates in hg19 for UCE sets. (D) Coordinates in mm9 for UCE sets. (E) Human exonic UCEs categorized for whether they contain some intronic DNA (listed as yes).

Table S3. Properties of Hi-C domains, Related to Figure 2. *See separate excel sheet.*

(A) Human domain sizes. (B) Mouse domain sizes. (C) Gene densities in human domains. (D) Gene densities in mouse domains. (E) Human domains: Positions of UCEs within domains. (F) Mouse domains: Positions of UCEs within domains. (G) Human domains: Distances of UCEs to nearest transcription start site (TSS). (H) Mouse domains: Distances of UCEs to nearest transcription start site (TSS). (I) Human domains: UCEs per domain. (J) Mouse domains: UCEs per domain.

Table S4. Investigation of the distribution of UCEs and Hi-C features that overlap intergenic, intronic, and exonic regions, Related to Figure 3. *See separate excel sheet.*

(A) Statistical analysis on the distribution of intergenic, intronic, and exonic UCEs comparing all UCEs (control) to UCEs that overlap pooled domains, pooled boundaries, and pooled loop anchors. (B) Statistical analysis on the distribution of intergenic, intronic, and exonic UCEs comparing all UCEs (control) to UCEs that overlap individual domain sets. (C) Analysis of the distribution of intergenic, intronic, and exonic sequences within pooled domains, pooled boundaries, and pooled loop anchors that fall within intergenic, intronic, and exonic regions.

Table S5. Analysis on UCEs that fall within boundaries, loop anchors, and the individual Hi-C domain datasets, Related to Figure 3. *See separate excel sheet.*

Overlaps of exonic UCEs within human pooled boundaries (A) or (B) pooled loop anchors with each feature. (C) Invariant domain UCEs that are overlapped by all human Hi-C domain datasets (10/10). (D) Invariant domain UCEs that are overlapped by all mouse Hi-C domain datasets (6/6). (E) List of UCEs that overlap between invariant domain human and mouse UCEs. (F) Statistical analysis of the distribution of intergenic, intronic, and exonic UCEs, comparing all UCEs (control) to invariant domain UCEs.

Supplemental Experimental Procedures

Hi-C data sources

The genomic coordinates for domains, boundaries, and loop anchors were obtained from the published Hi-C datasets (Table S1) (Dixon et al., 2012; Fraser et al., 2015; Rao et al., 2014). When required the coordinates were converted using UCSC Genome Browser tool liftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) to hg19 genome assembly. To avoid counting a same region multiple times, overlapping genomic coordinates were merged providing a final list of coordinates that may differ from originally reported ones in the respective publications. For each individual and pooled Hi-C dataset after coordinate merging the information about the number of regions, median size (bp), coverage (bp), and proportion of covered genome (%) was reported (Table S1).

To account for data variability given that the resolution between datasets varied based on the amount of biological material, applied Hi-C protocol (Kalhor et al., 2011; Lieberman-Aiden et al., 2009; Nagano et al., 2015; Rao et al., 2014), and sequencing depth, each chromosomal feature (domain, boundary, loop anchor) was inspected individually within a single dataset, in addition to pooling across multiple datasets.

UCE data sources

The UCEs encompass a dataset representing 896 HMR-HDM-HC elements as previously reported in hg18 genome assembly (Derti et al., 2006; McCole et al., 2014). All UCE genomic coordinates were lifted over to hg19 genome assembly and made available in Table S2C.

To obtain the respective UCE coordinates in mm9 genome assembly, a read aligner tool bowtie2 was used (Langmead and Salzberg, 2012). To start, the UCEs sequences in fasta format were converted to fastq format, and then subsequently mapped using bowtie2 with the following parameters: 'bowtie2 -p 4 -x /bowtie_index/mm9 --very-sensitive -t -S UCEs_mm9.sam -U sequences_hg18_allUCEs.fastq'. The end-to-end alignment was chosen to avoid 'trimming' or 'clipping' of some read characters from ends of the alignment. The output file in SAM format was first converted to a sorted BAM format (samtools view –bS UCEs_mm9.sam | samtools sort – UCEs_mm9_sorted) using SAMtools (Li et al., 2009), and then to a BED file ('bedtools bamtobed –i UCEs_mm9_sorted.bam > UCEs_mm9_sorted.bed') using BEDTools (Quinlan and Hall, 2010). Upon filtering matches in the mouse genome that were <200 bp (UCE_153, UCE_600) or differed more than 15 bp in length to hg19 coordinates (UCE_733), we obtained 893 UCEs in mm9 genome assembly. The output mm9 UCE coordinates adjusted to 1-based system were reported in Table S2D. Three UCEs were not recovered in the mouse genome, namely UCE_153, UCE_600, and UCE_733. The first two omitted UCEs spanned less < 200 bp in the mouse genome, and thus did not satisfy our initial UCE definition that required a UCE to be ≥ 200 bp in length. The third UCE_733 was excluded since it differed more than 15 bp in length to the human counterpart, and it falls within a mouse intergenic region as opposed to its human UCE sequence that lies within the intron of POU6F2 gene.

During analysis UCEs were further subclassified into exonic, intronic or intergenic elements as previously described (Derti et al., 2006; McCole et al., 2014) to allow for finer dissection of UCE relation to chromosome organization (Table S2).

Data sources for correlation analysis

The CNV/CNA data previously assembled (McCole et al., 2014) was inspected in healthy individuals representing classical inherited CNVs (Abecasis et al., 2010; Campbell et al., 2011; Conrad et al., 2010; Drmanac et al., 2010; Jakobsson et al., 2008; Matsuzaki et al., 2009; McCarroll et al., 2008; Shaikh et al., 2009), and CNAs derived from 52 different cancers (Beroukhim et al., 2010; Bullinger et al., 2010; Curtis et al., 2012; Holmfeldt et al., 2013; Kandoth et al., 2013; Network, 2008, 2011, 2012a, b, c; Nik-Zainal et al., 2012; Robinson et al., 2012; Taylor et al., 2010; Walker et al., 2012; Walter et al., 2009; Weischenfeldt et al., 2013; Zhang et al., 2012). All CNV and CNA genomic coordinates were lifted over to hg19 assembly and collapsed.

The genomic coordinates for hg19 assembly were downloaded from the UCSC Table Browser (Karolchik et al., 2004), coordinates for genes, introns, and exons were derived from group - Genes and Gene Predictions, track - UCSC genes; repetitive element were from group - Repeats, track - RepeatMasker, and contain SINE, LINE, and LTR repetitive elements; segmental duplications were from group- Repeats, track- Segmental Dups, CpG islands were from group - Regulation, track - CpG Islands, and open chromatin identified by the ENCODE project (Dunham, 2012) from group - Regulation, track - Open Chrom Synth for cell lines that match Hi-C datasets (GM12878, H1-hESC, K562, HeLa, HeLa-Ifna4h, HUVEC, NHEK).

Depletion or enrichment analysis of UCEs in specific genomic regions

The enrichment or depletion of UCEs in genomic regions of interest such as domains, boundaries, and loop anchors was assessed using established methods previously reported in our publications (Chiang et al., 2008; Derti et al., 2006; McCole et al., 2014). Briefly, observed overlap between for instance UCEs and domains were compared to mean expected overlap, which was produced by placing randomized set of elements that match UCEs in number and length 1,000 times in the genome. The distribution of expected overlaps was assessed for normality using Kolomogorov-Smirnov (KS) test. In a case normality was observed, Z-test comparison between observed and expected overlaps was reported, which when significant indicated depletion if ratio between observed and mean expected overlaps (obs/exp) was below 1.0, or enrichment if the obs/exp ratio was above 1.0. In instances where normality was not observed, the proportion of expected overlaps equal to, or more extreme than the observed overlap together with obs/exp ratio was reported. During further dissection of relation between UCEs, which were classified to intergenic, intronic, and exonic elements, to genomic regions of interest random set of elements used to calculate mean expected overlap was pooled from the matching genomic regions, *i.e.* only intergenic, intronic, or exonic ones.

Distribution of expected overlaps and observed overlap (colored line) were visualized using histograms, where corresponding statistical values were reported based on the results of the KS test.

Correlation analyses

The genome was divided into bins of equal sizes. Within each bin, the fraction of sequence occupied by each control feature was calculated, as was that of UCEs, except in the case of GC content, where it was calculated as the fraction of G + C. Then genome-wide correlations within each bin were preformed among feature densities or GC content. The Spearman correlation coefficients and matching p values were provided in two flavors, either for a pairwise comparison between two features, or as a part of partial correlation approach, which assesses whether the correlation between two features was affected by co-correlation with the third genomic feature. The obtained Spearman correlation coefficients were visualized as a heatmap.

Analysis of domain size, gene density, positions of UCEs within domains and with respect to transcription start sites

Domain properties were analyzed using custom python scripts. Statistical comparison of domain sizes was performed using Mann-Whitney U test. Gene density (kb) refers to the number of unique genes in a region divided by the size of the region in kb. Comparisons of gene density were also carried out using Mann-Whitney U tests. Positions of UCEs across domains were compared to 1,000 sets of random control elements using a K-S test. Domains were 'folded' to create 5 bins of equal size across the domain from edge to middle with, for example, the 'left-hand' and 'right-hand' edge assessed in bin 1. Distances between UCEs and transcription start sites (TSS) were in relation to TSS specified by UCSC known genes track. Distances were compared to 100 sets of random control elements using an Anderson-Darling test.

Distribution of intergenic, intronic, and exonic UCEs that overlap domains, boundaries, and loop anchors

Analysis on the distribution of intergenic, intronic, and exonic UCEs (reported from depletion/enrichment analysis in Table S2A) was performed by comparing a control set (all 896 UCEs from Table S2C) to UCEs that overlap feature of interest, *i.e.* either domains, boundaries, or loop anchors (Table S1). A statistical significance was determined using chi-squared test, and reported for pooled (Table S3A), and other human domain datasets from other cell lines (Table S3B).

To determine the amount of domains, boundaries, and loop anchors that fall within intergenic, intronic, and exonic regions, the overlap between two features (*i.e.* intergenic regions in domains) was calculated using bedtools intersect (Quinlan and Hall, 2010). The information on an overlap such as

number of intervals, median interval size (bp), coverage (bp), percentage of genome, and percentage of a Hi-C feature coverage is reported for pooled domains, boundaries, and loop anchors (Table S3C).

Gene ontology

Functional association of UCEs to the gene ontology (GO) terms of nearby genes was determined against the full set of 896 human (or 893 mouse) UCEs as a background using the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010). The background provided a control that observed functional association for domain invariant UCEs and exonic UCEs that overlap boundaries and loop anchors was inherent to those UCE subsets, and not the full set of UCEs.

Analysis on the number of times each UCE is overlapped by the individual domain dataset

Each assessment of overlap between every UCE and every individual domain dataset is assigned a score of 1 (intersect) or 0 (no intersect). The summation of scores across all individual datasets resulted in a total score for each UCEs. Those UCEs with the highest score, *i.e.* that are confirmed by all individual datasets to overlap domains, were termed domain invariant UCEs, and reported for human (Table S5C), and mouse (Table S5D). Overlap between domain invariant UCEs from human and mouse was performed using UCE ID identifiers. Shared domain invariant UCEs between human and mouse, which are classified as exonic or intronic, were reported together with the assigned gene (UCSC) and gene abbreviation (NCBI) in Table S5E.

Scripts

Custom scripts associated with this study are available at https://github.com/rmccole/UCEs_genome_organization.

Supplemental References

Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899-905.

Bullinger, L., Kronke, J., Schon, C., Radtke, I., Urlbauer, K., Botzenhardt, U., Gaidzik, V., Cario, A., Senger, C., Schlenk, R.F., et al. (2010). Identification of acquired copy number alterations and uniparental disomies in cytogenetically normal acute myeloid leukemia using high-resolution single-nucleotide polymorphism analysis. Leukemia *24*, 438-449.

Campbell, C.D., Sampas, N., Tsalenko, A., Sudmant, P.H., Kidd, J.M., Malig, M., Vu, T.H., Vives, L., Tsang, P., Bruhn, L., et al. (2011). Population-genetic properties of differentiated human copy-number polymorphisms. Am. J. Hum. Genet. *88*, 317-332.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704-712.

Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346-352.

Dale, R.K., Pedersen, B.S., and Quinlan, A.R. (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics *27*, 3423-3424.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science *327*, 78-81.

Dunham (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Holmfeldt, L., Wei, L., Diaz-Flores, E., Walsh, M., Zhang, J., Ding, L., Payne-Turner, D., Churchman, M., Andersson, A., Chen, S.C., et al. (2013). The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat. Genet. *45*, 242-252.

Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature *451*, 998-1003.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2011). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. *30*, 90-98.

Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R., Benz, C.C., et al. (2013). Integrated genomic characterization of endometrial carcinoma. Nature *497*, 67-73.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res *32*, D493-496.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357-359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289-293.

Matsuzaki, H., Wang, P.H., Hu, J., Rava, R., and Fu, G.K. (2009). High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. Genome Biol. *10*, R125.

McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A., et al. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet. *40*, 1166-1174.

Nagano, T., Varnai, C., Schoenfelder, S., Javierre, B.M., Wingett, S.W., and Fraser, P. (2015). Comparison of Hi-C results using in-solution versus in-nucleus ligation. Genome Biol *16*, 175.
Network, T.C.G.A.R. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061-1068.

Network, T.C.G.A.R. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.

Network, T.C.G.A.R. (2012a). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519-525.

Network, T.C.G.A.R. (2012b). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330-337.

Network, T.C.G.A.R. (2012c). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. Cell *149*, 979-993.

Robinson, G., Parker, M., Kranenburg, T.A., Lu, C., Chen, X., Ding, L., Phoenix, T.N., Hedlund, E., Wei, L., Zhu, X., et al. (2012). Novel mutations target distinct subgroups of medulloblastoma. Nature *488*, 43-48.

Shaikh, T.H., Gai, X., Perin, J.C., Glessner, J.T., Xie, H., Murphy, K., O'Hara, R., Casalunovo, T., Conlin, L.K., D'Arcy, M., et al. (2009). High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome Res. *19*, 1682-1690.

Taylor, B.S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B.S., Arora, V.K., Kaushik, P., Cerami, E., Reva, B., et al. (2010). Integrative genomic profiling of human prostate cancer. Cancer Cell *18*, 11-22.

Walker, L.C., Krause, L., Spurdle, A.B., and Waddell, N. (2012). Germline copy number variants are not associated with globally acquired copy number changes in familial breast tumours. Breast Cancer Res. Treat. *134*, 1005-1011.

Walter, M.J., Payton, J.E., Ries, R.E., Shannon, W.D., Deshmukh, H., Zhao, Y., Baty, J., Heath, S., Westervelt, P., Watson, M.A., et al. (2009). Acquired copy number alterations in adult acute myeloid leukemia genomes. Proc. Natl. Acad. Sci. U S A *106*, 12950-12955.

Weischenfeldt, J., Simon, R., Feuerbach, L., Schlangen, K., Weichenhan, D., Minner, S., Wuttig, D., Warnatz, H.J., Stehr, H., Rausch, T., et al. (2013). Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. Cancer Cell *23*, 159-170.

Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S.L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M., et al. (2012). The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature *481*, 157-163.