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

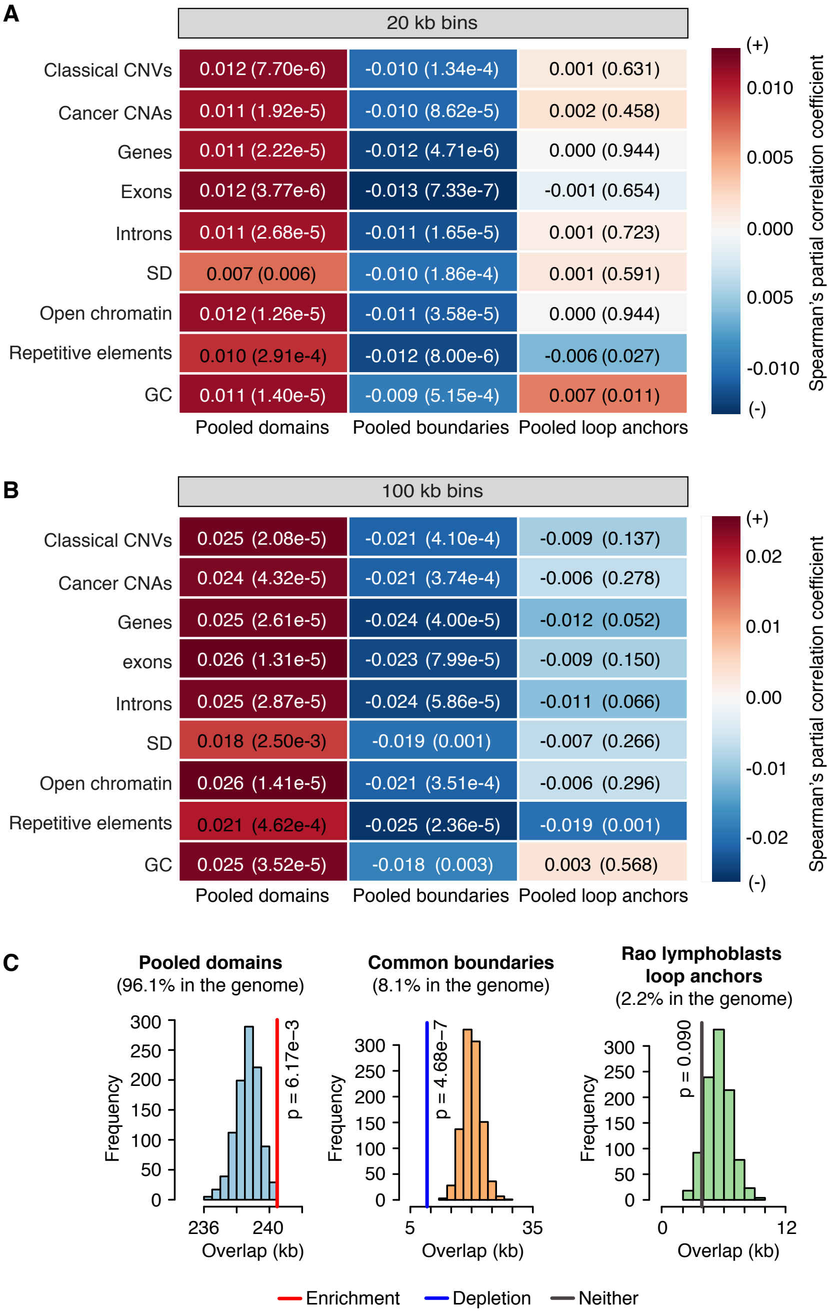

Figure S1. Partial correlation and mouse depletion/enrichment analyses, Related to Figure 2.

(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=6.17×10^{-3}, obs/exp=1.007), depleted in (common) boundaries (blue line; p=4.68×10^{-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.

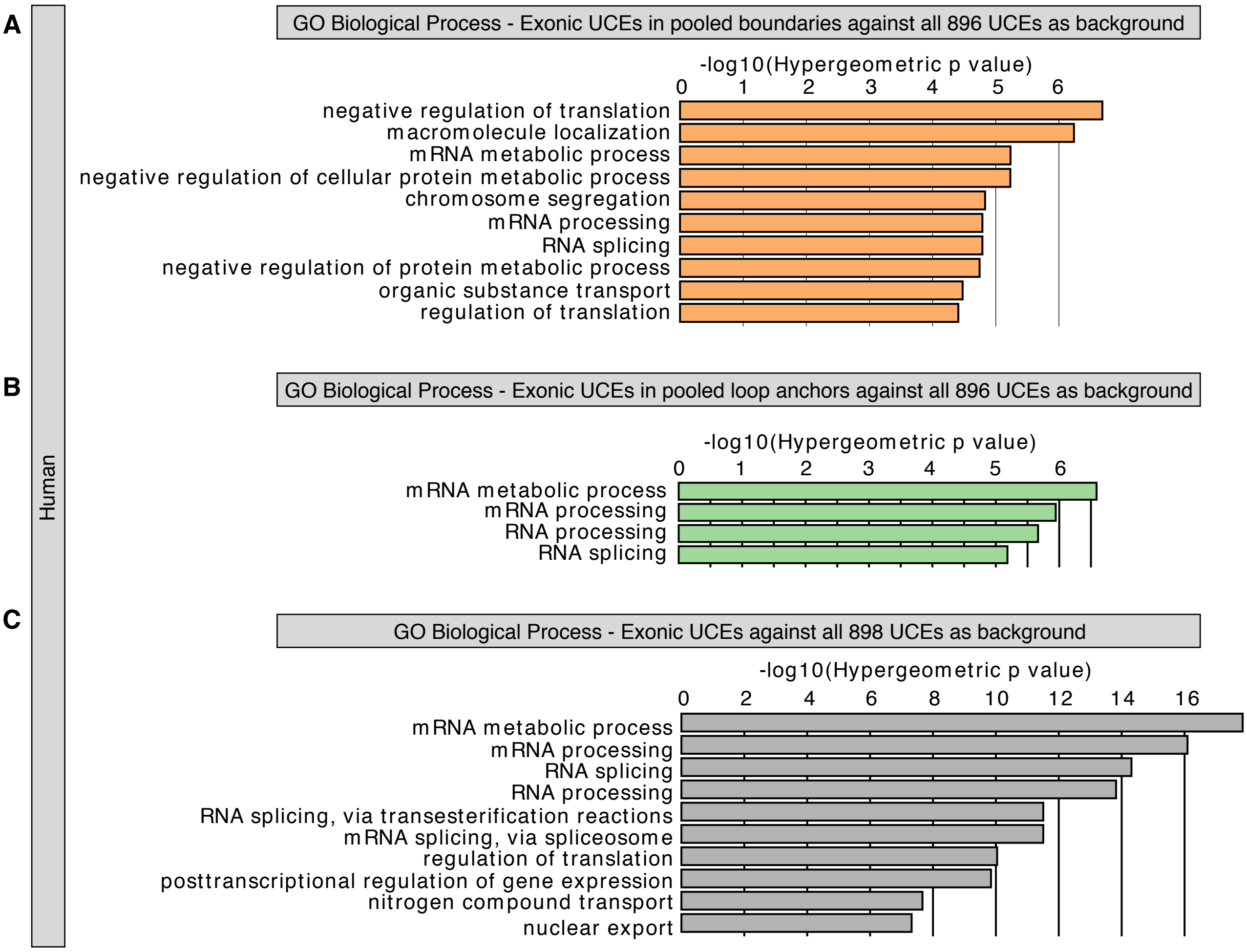

Figure S2. Exonic UCEs that overlap pooled boundaries or pooled loop anchors are functionally associated with RNA processing GO terms, Related to Figure 3.

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.

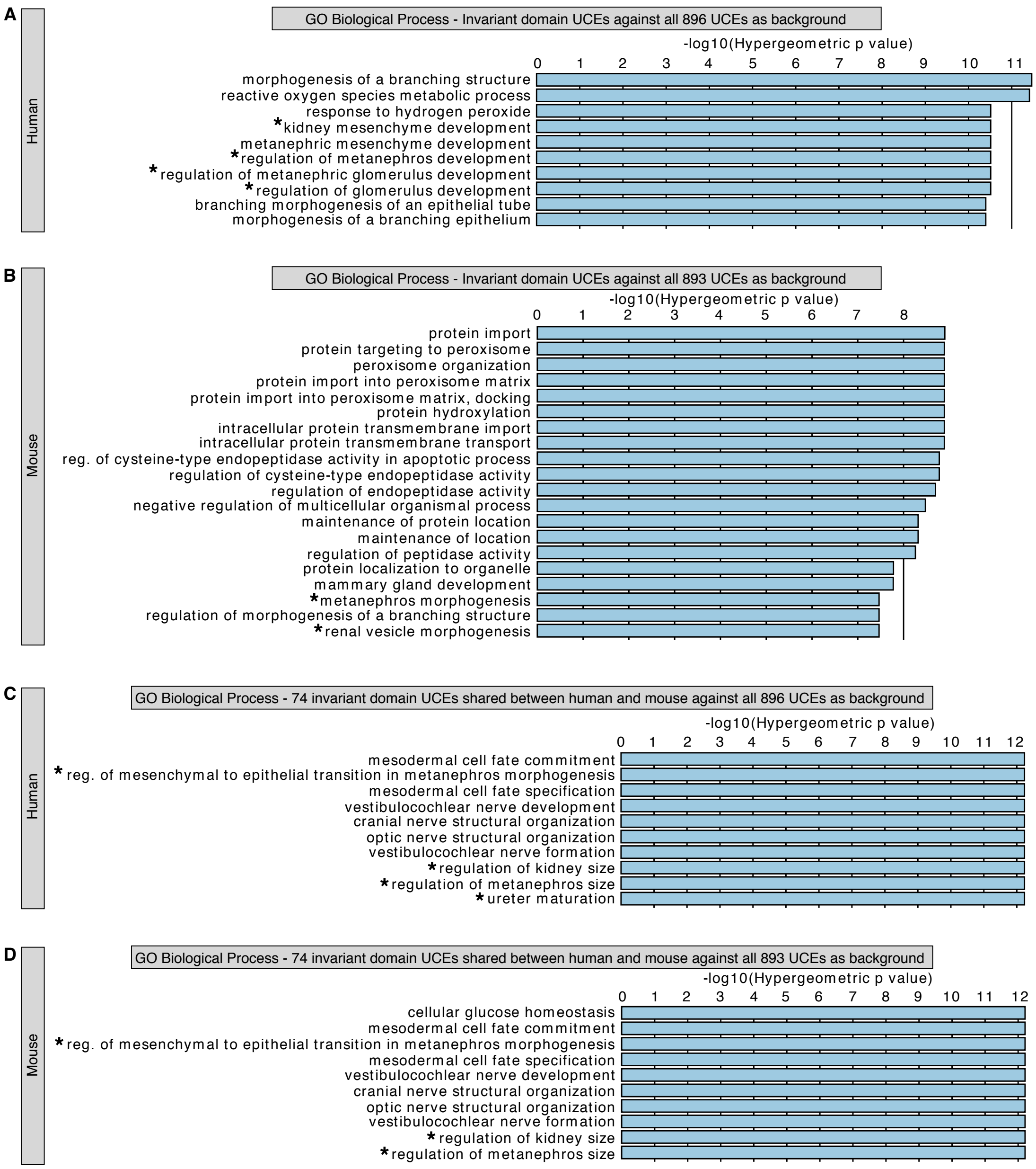

Figure S3. Invariant domain UCEs are associated with kidney-related processes in both human and mouse, Related to Figure 3.

(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.