Independent transposon exaptation is a widespread mechanism of redundant enhancer evolution in the mammalian genome

Nicolai K. H. Barth[+], Lifei Li[+] and Leila Taher[+#*]

[+] Division of Bioinformatics, Department of Biology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany.

[#] Current address: Institute of Biomedical Informatics, Graz University of Technology, Graz, Austria

* To whom correspondence should be addressed. Email: leila.taher@tugraz.at

This document contains

Supplementary Figures S1 to S16

Supplementary Tables S1 to S5

Supplementary Materials and Methods

SUPPLEMENTARY FIGURES

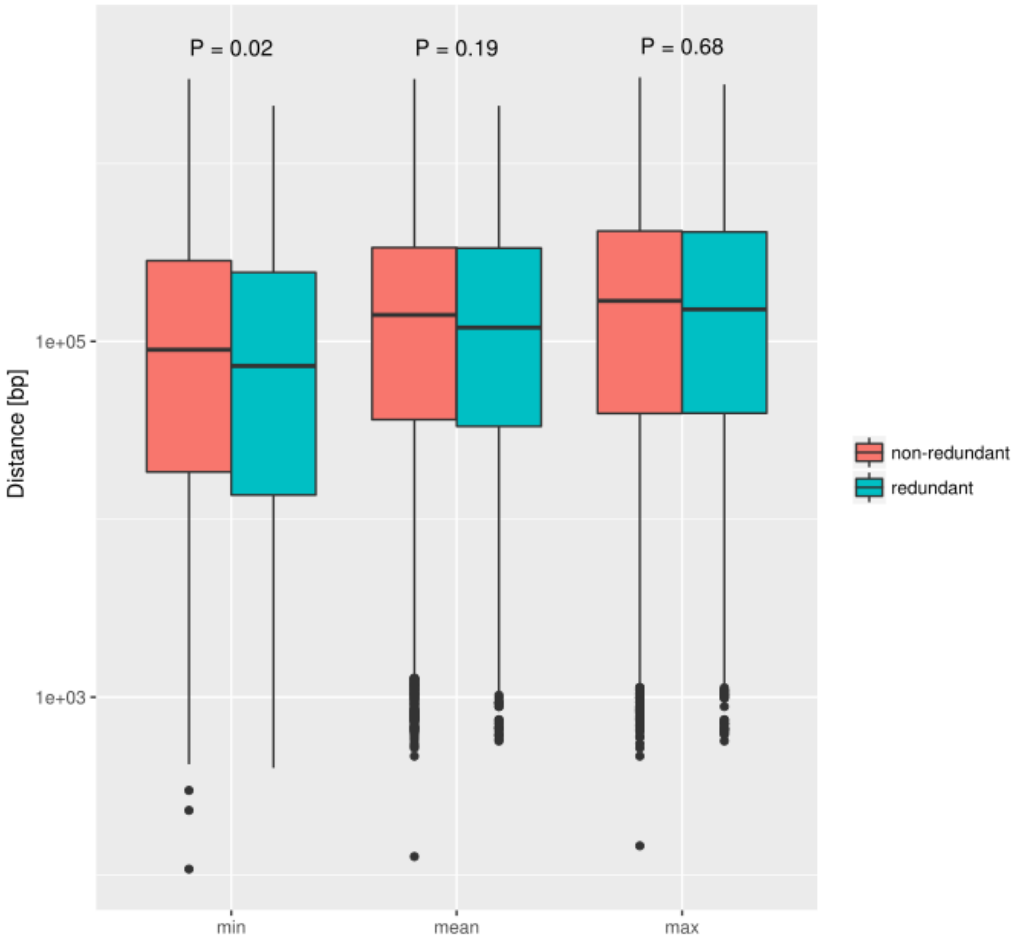

Figure S1. Distance to the (nearest:min, average:mean, farthest:max) TSS of a target gene for 1,280 shadow and 2,243 non-shadow enhancers in the human genome. P-values from Wilcoxon rank-sum tests.

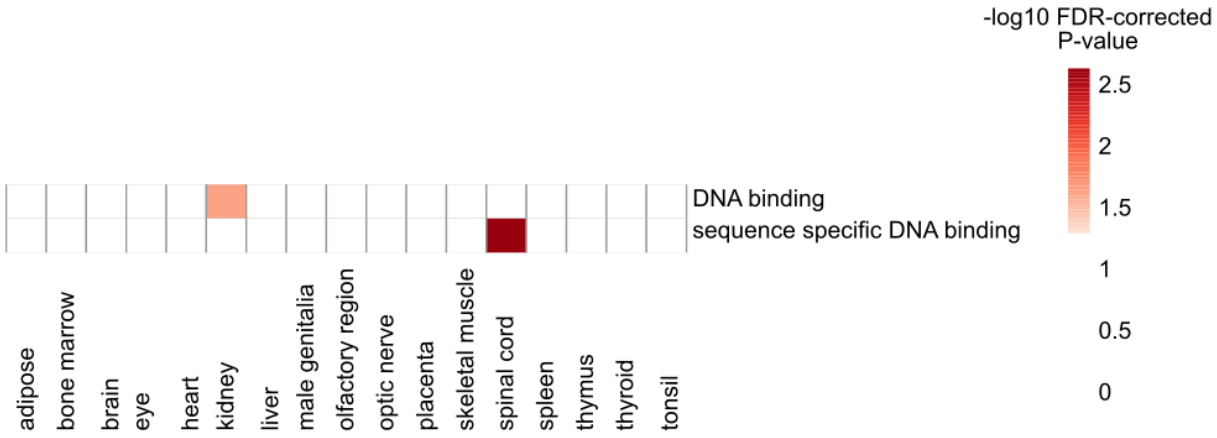

Figure S2. Functional analysis of human redundant enhancer target genes, per facet. The target genes of all correlated enhancers served as background set. Coloring: White cells represent tests with an FDR-corrected P-value >0.05, red represents significant tests, the darker the red tone the lower the P-value. Scale: negative base 10 logarithm of the FDR-corrected P-values.

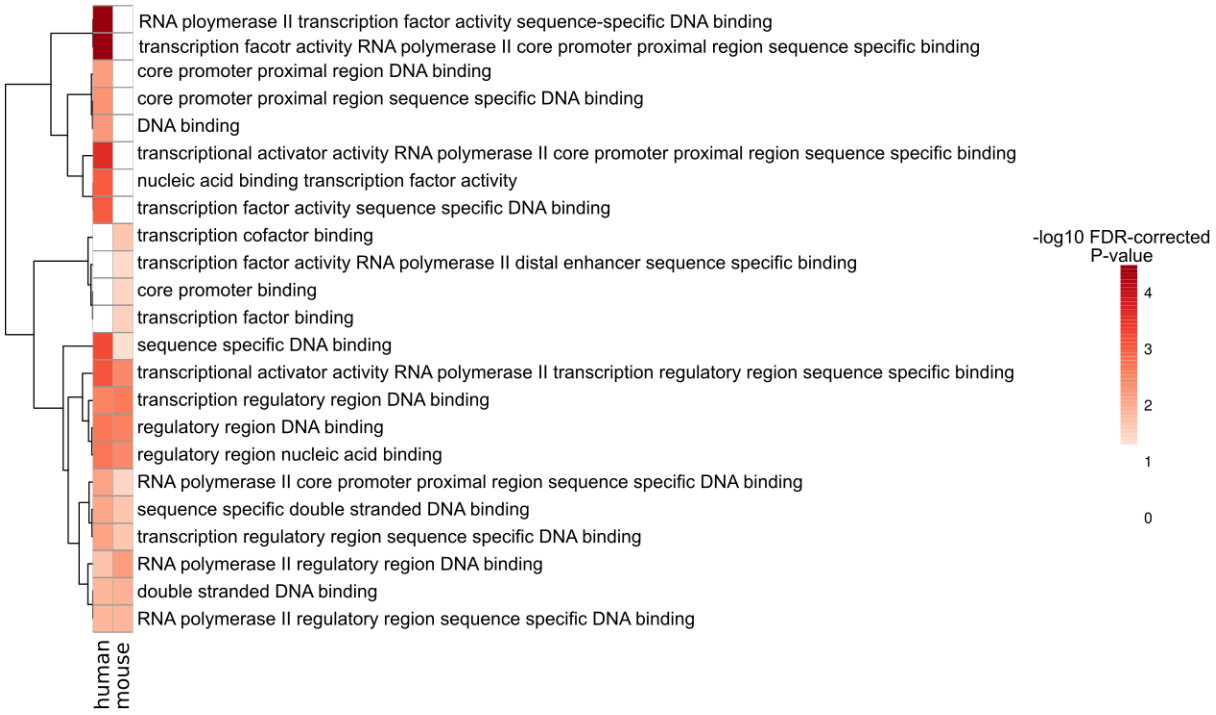

Figure S3. Functional analysis of human and mouse redundant enhancer target genes. The target genes of all correlated enhancers served as background set. See Figure S3 for coloring and scale.

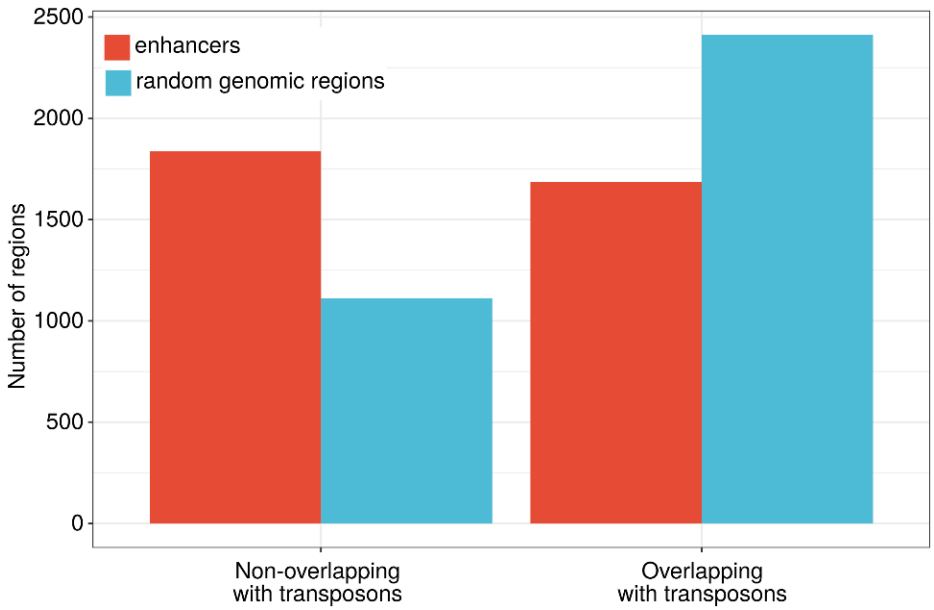

Figure S4. Number of human enhancers annotated as transposons. Random genomic regions show more frequent transposon overlaps than enhancers.

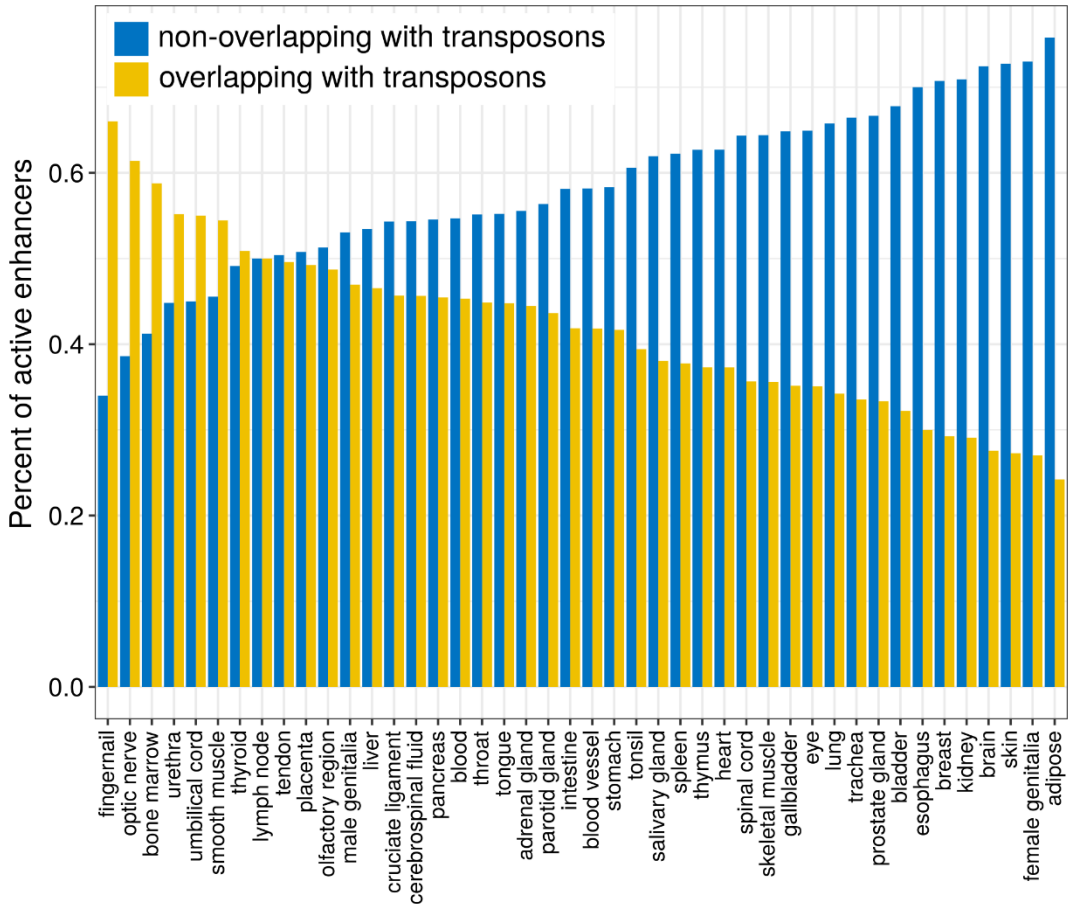

Figure S5. The fraction of transposon overlapping enhancers out of all active enhancers depends on the facet.

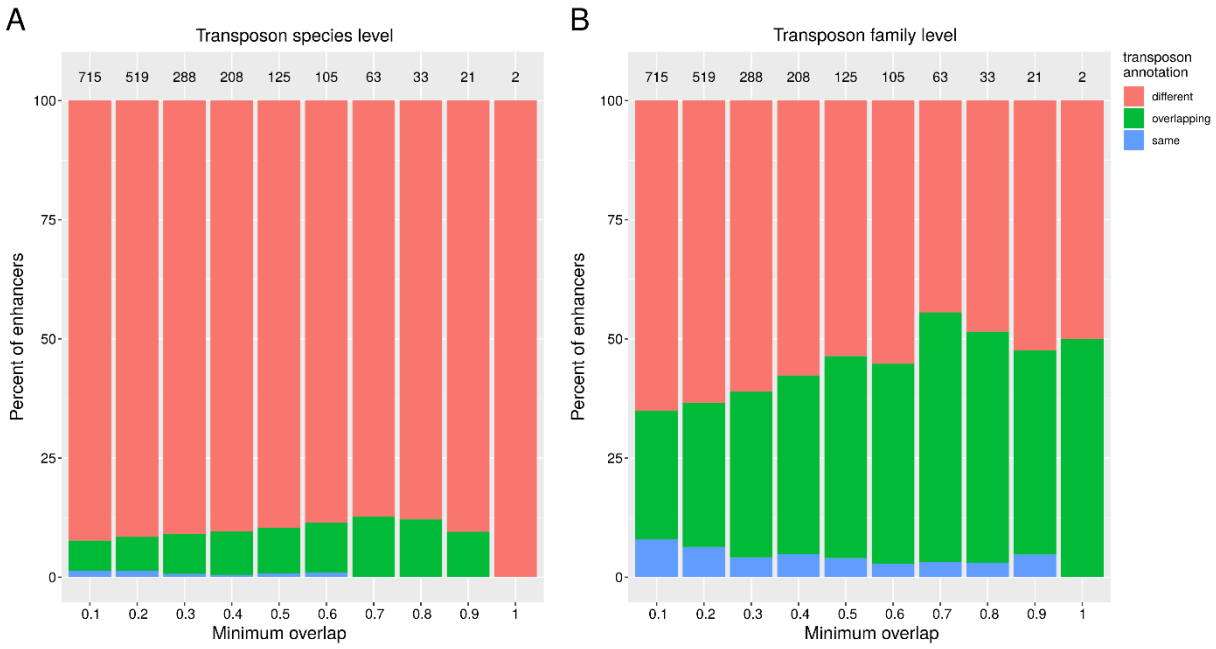

Figure S6. Percent of transposon-shadow enhancer pairs where partners are annotated as transposons from different, shared or identical families, for increasing thresholds of overlap between the enhancer and the transposon sequences for the enhancer to be annotated as a transposon, as fraction of the total enhancer sequence. The numbers above the bars are the total number of pairs where both partners fulfill the overlap criterion. For a 100% overlap there is only a hand full of redundant enhancer pairs left.

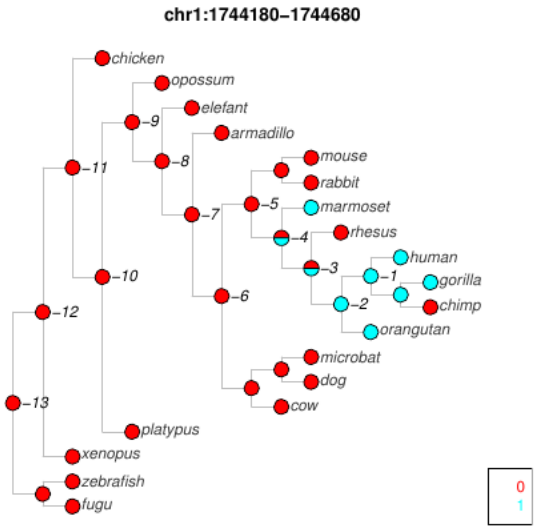

Figure S7. Topology of the phylogenetic tree used for the character state reconstruction of human transposon shadow enhancers. Nodes corresponding to ancestral species are labeled with negative numbers. Ancestral state inferences for the enhancer with coordinates chr1:1,744,180-1,744,680 (hg19). In this example, node -4 is the inferred transposon insertion node.

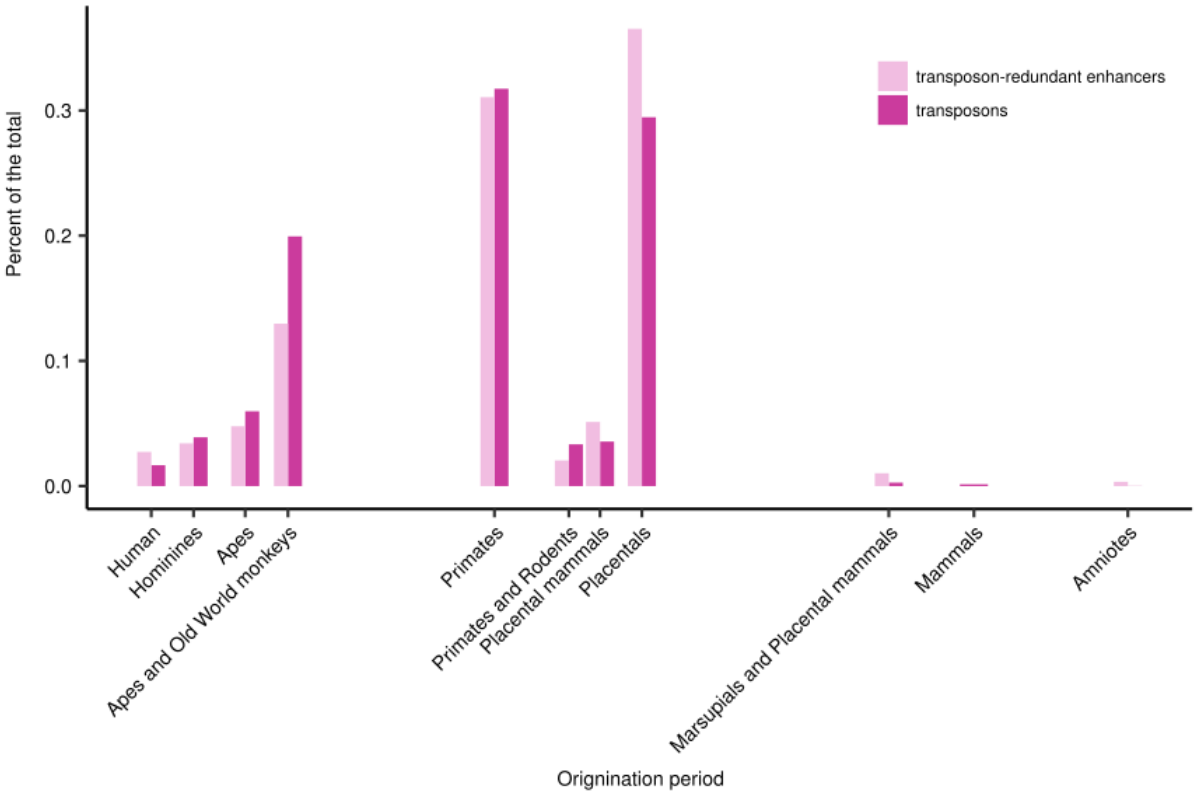

Figure S8. Ages of transposon-shadow enhancers and of random genomic transposon sequences as inferred from the reconstruction of ancestral states in the phylogenetic tree. The nodes of the tree were mapped to clades according to the ENSEMBL COMPARA species tree (www.ensembl.org/info/about/speciestree.html).

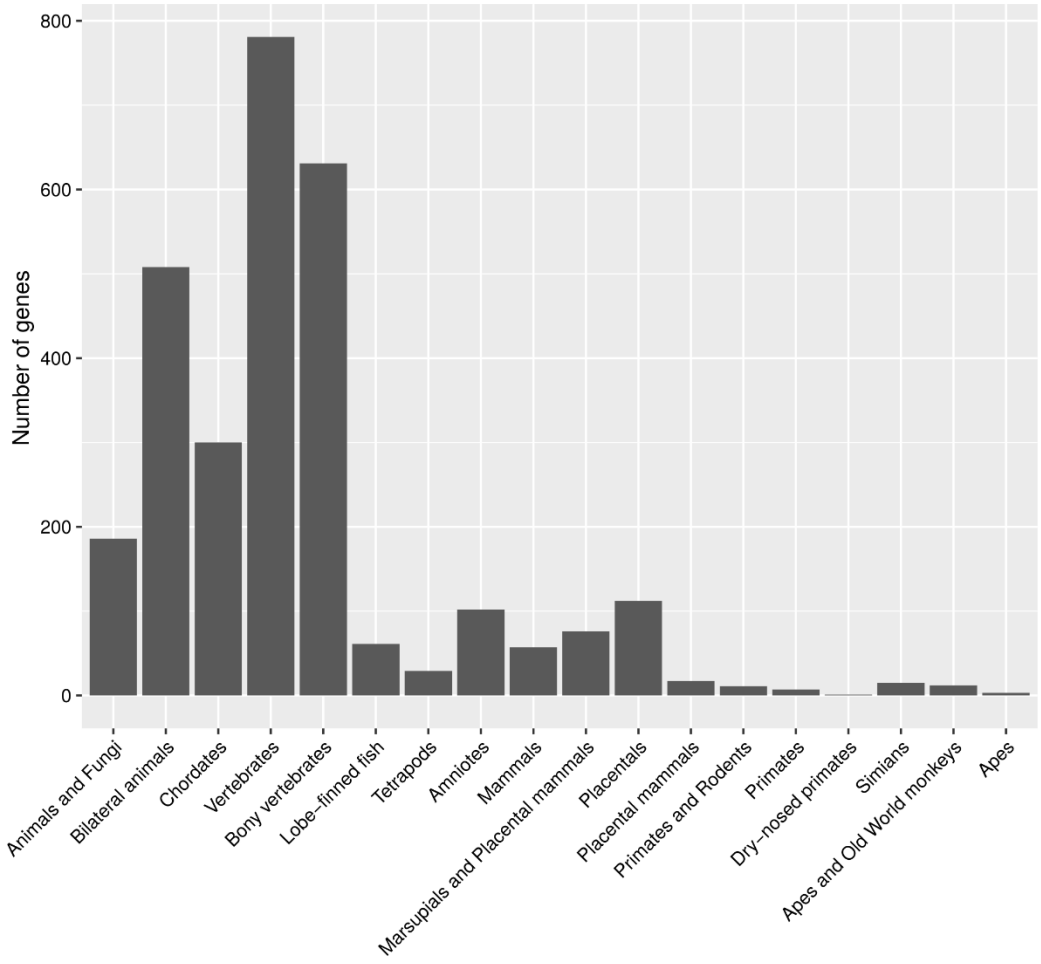

Figure S9. Age of human enhancer target genes. The gene age estimates were extracted from the ENSEMBL COMPARA gene phylogenetic trees. No information was available for 67 genes (excluded from this graph).

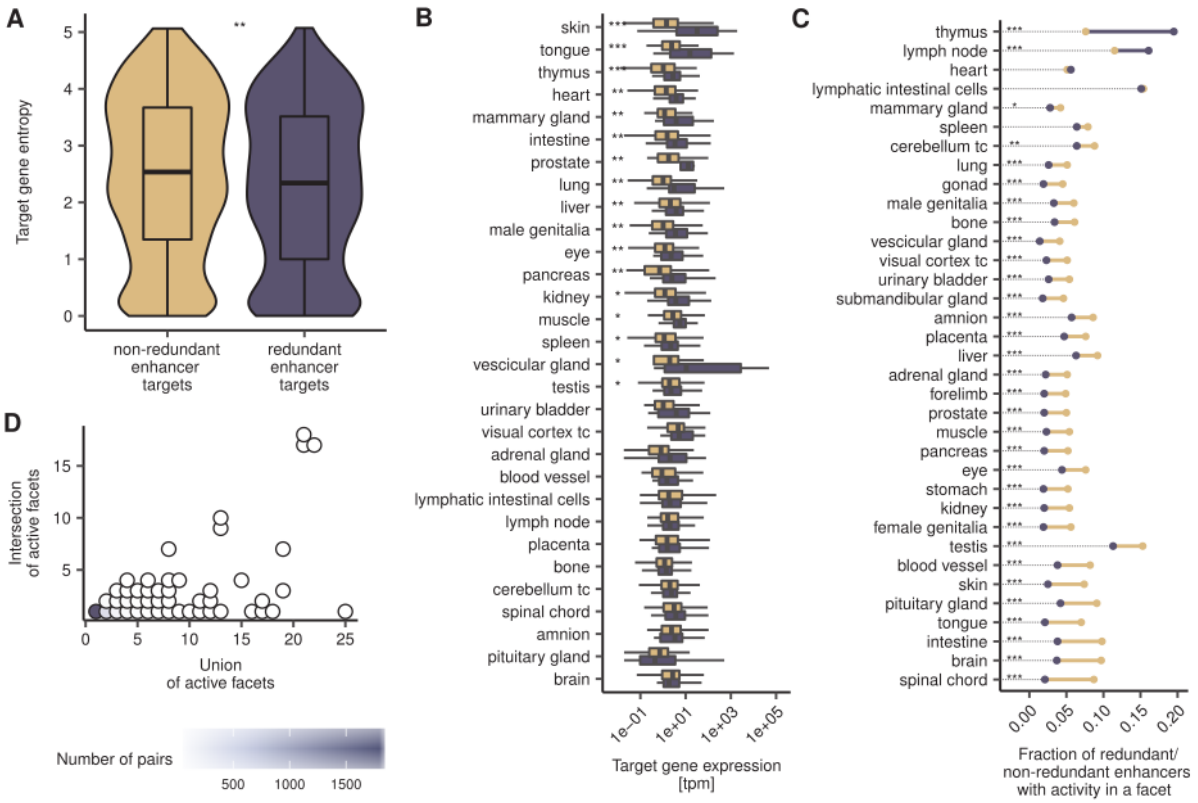

*Figure S10. Comparison of shadow and non-shadow mouse enhancers. A) Target gene entropies of non-shadow (beige) and shadow (purple) enhancer target genes, Wilcoxon rank sum test. '**' P < 0.01. B) Facet-specific expression of non-shadow (beige) and shadow (purple) enhancer target genes, adjusted P-values of Wilcoxon rank sum tests. '*' P < 0.05, '**' P < 0.01, '***' P < 0.001. C) Dumbbell plot showing the fractions of active non-shadow (beige) and shadow (purple) enhancers per facet. If the fraction of active non-shadow enhancers is larger than that of shadow enhancers, the line between the dots is depicted in beige; otherwise, in purple. The only purpose of the dotted lines is to serve as visual aids. Asterisks indicate adjusted P-values of Fisher's exact tests: Significance code see B).) D) Redundancy of shadow enhancer pairs.*

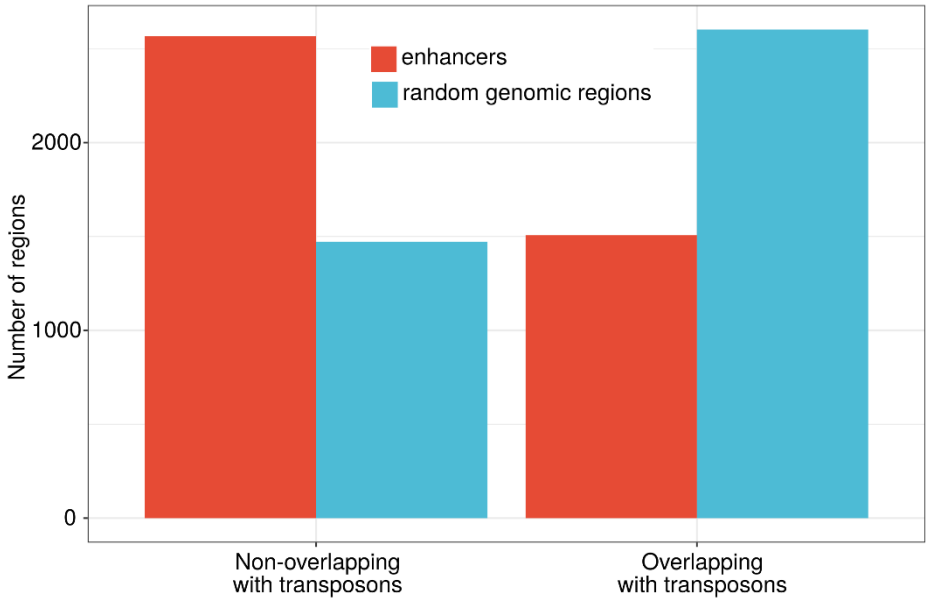

Figure S11. Number of mouse enhancers annotated as transposons. The numbers of elements that overlap with transposons and elements that do not overlap with transposons are reversed in random genomic regions as compared to enhancers.

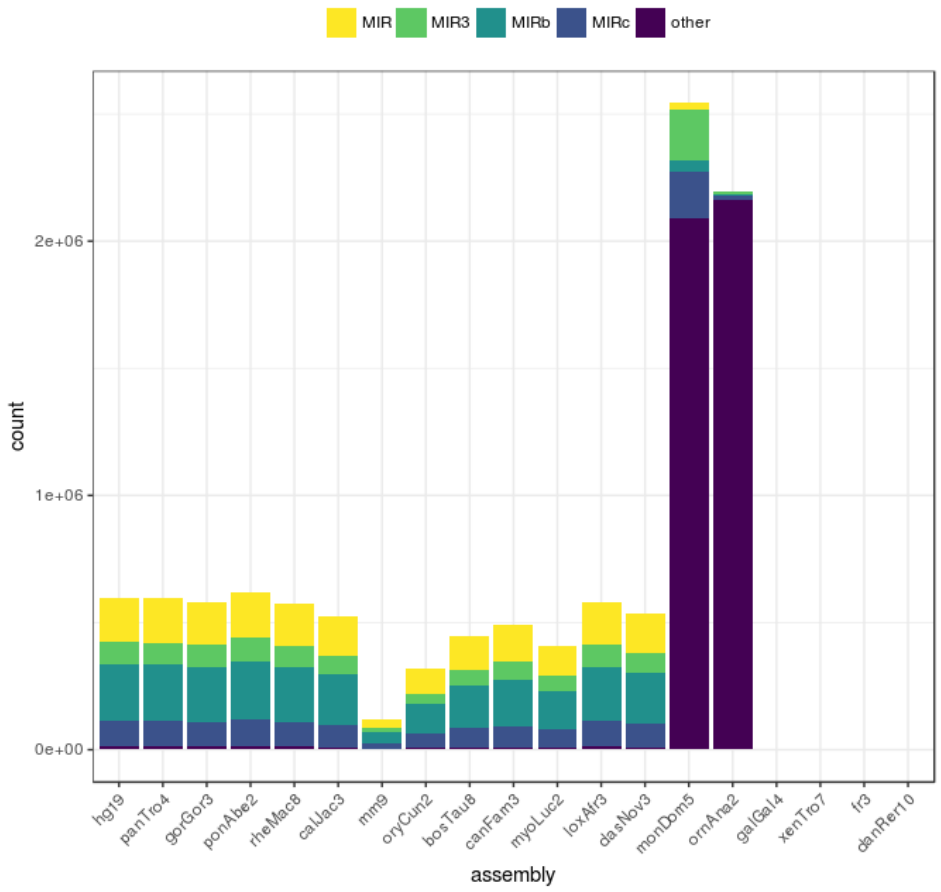

Figure S12: Number of TE instances for different MIR species in each of the assemblies considered in this study. The figure illustrates the number of TE species from the MIR family, which were active before the mammalian radiation. The number of MIRb, MIR3, MIRc and MIR instances in the mouse genome is much smaller than that in the human genome and the genomes of most other mammalian species used in this study. However, the vast majority of these transposons were inserted in the genome of the common ancestor before the mammalian radiation, and thus, are expected to be mostly shared by its descendants. "other" summarizes all other MIR families, which are mostly specific to the monotremata or marsupalia clades. In summary, many such MIRb, MIR3, MIRc and MIR instances appear to have diverged faster in the mouse genome, to the point where they are no longer recognizable. This rationale can be extended to other TE families.

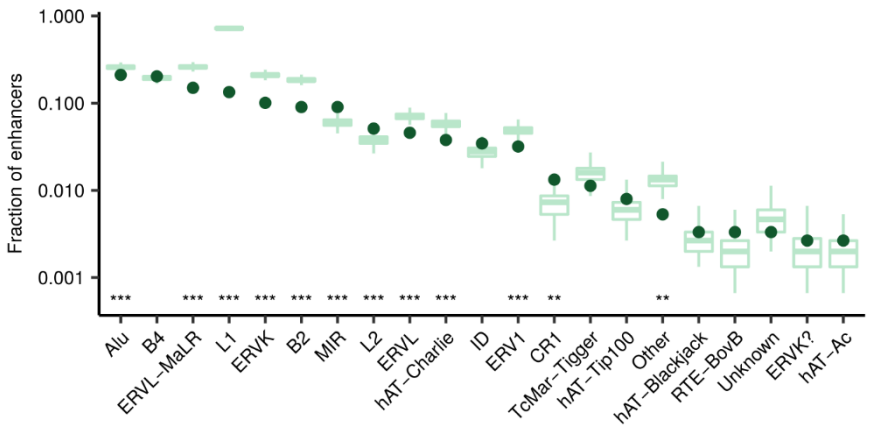

Figure S13. Enrichment of transposon families in transposon enhancers compared to random genomic sequences. Asterisks indicate adjusted empirical P-values. '' P < 0.05, '**' P < 0.01, '***' P < 0.001. For further description refer to Figure 2A.*

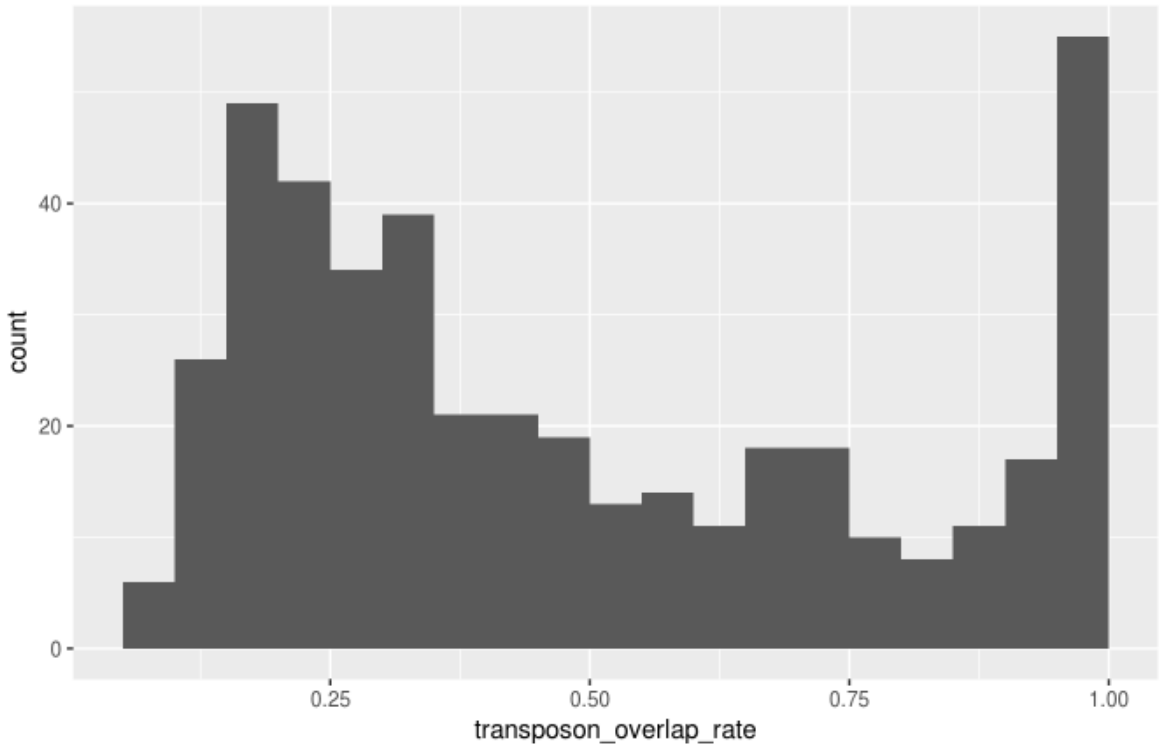

Figure S14. The fraction of transposon-redundant enhancers overlapping with transpsons.

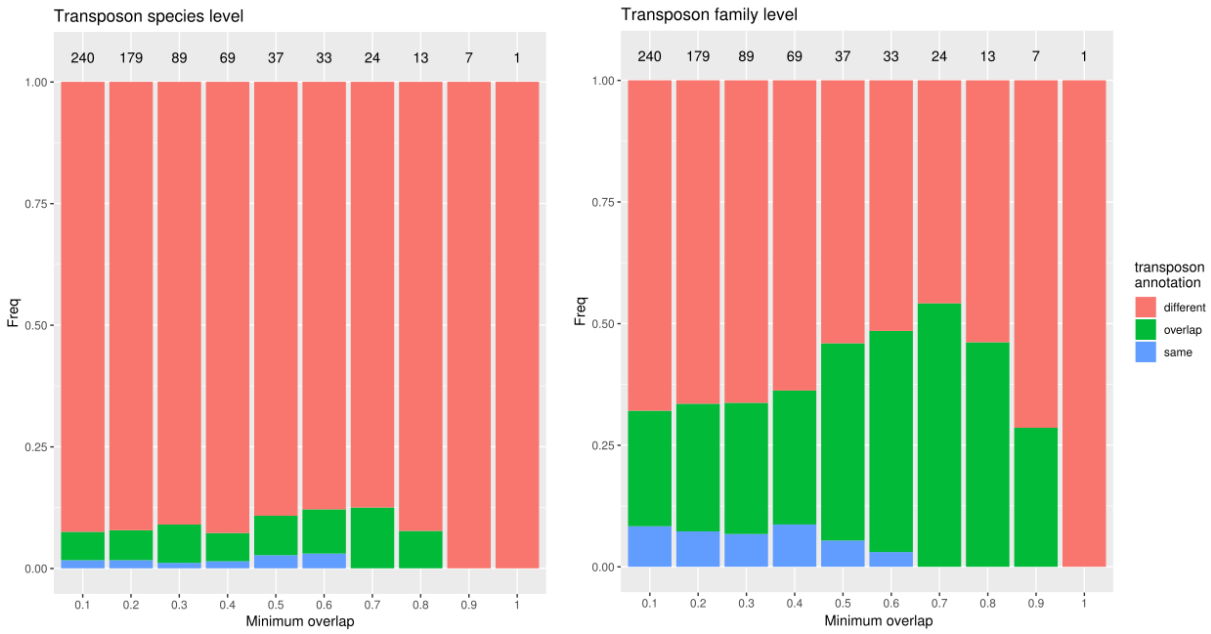

Figure S15. Comparison of transposon redundant enhancer annotation for the redundant enhancer pairs that are based on a correlation coefficient = 1 (see also the caption of Figure S6).

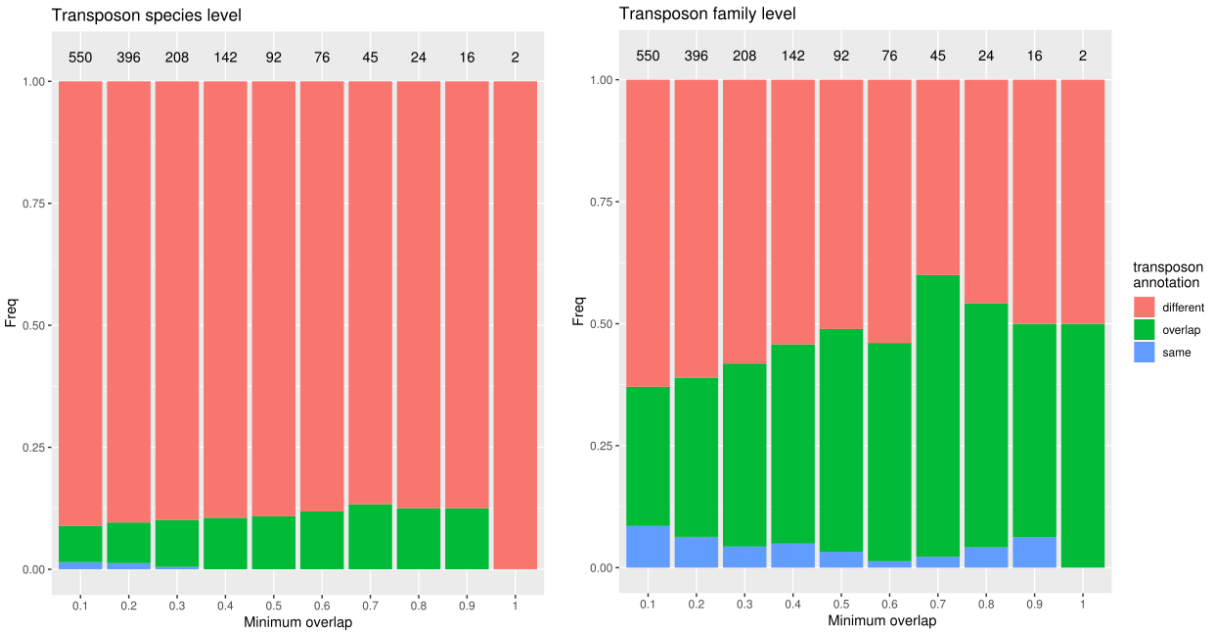

Figure S16. Comparison of transposon redundant enhancer annotation for the redundant enhancer pairs that are based an even stricter target gene assignment via an empirical P-value (see supplementary Materials and Methods and the caption of Figure S6).

SUPPLEMENTARY TABLES

Table S1. Human tissue samples generated by the FANTOM5 consortium (phase 2.5) assigned to each facet. The data were kindly provided by the FANTOM5 consortium.

Sample_ID	Facet
cerebrospinal_fluid	CNhs13437
spleen	CNhs10631, CNhs10651
bone_marrow	CNhs12845
prostate_gland	CNhs10628
cruciate_ligament	CNhs13439
adipose	CNhs10615, CNhs13972, CNhs13973, CNhs13974, CNhs13975
optic_nerve	CNhs13449
lung	CNhs10625, CNhs11680, CNhs11786
tendon	CNhs13435
adrenal_gland	CNhs11793
throat	CNhs12858, CNhs11770
tongue	CNhs13460, CNhs12853, CNhs11768
umbilical_cord	CNhs11765
esophagus	CNhs10620
urethra	CNhs13464
breast	CNhs11792
intestine	CNhs12842, CNhs11794, CNhs10619, CNhs11780, CNhs11781, CNhs12997, CNhs11777, CNhs10630, CNhs11773
skin	CNhs13458, CNhs11785, CNhs11774
placenta	CNhs10627
thymus	CNhs10633, CNhs10650
thyroid	CNhs10634, CNhs11769
blood	CNhs11761
liver	CNhs10624, CNhs11798
smooth_muscle	CNhs11755
lymph_node	CNhs11788
male_genitalia	CNhs12846, CNhs12847, CNhs12850, CNhs12851, CNhs10632, CNhs12998
fingernail	CNhs13445
brain	CNhs13793, CNhs12311, CNhs14078, CNhs11796, CNhs10617, CNhs11797, CNhs13802, CNhs12321, CNhs14232, CNhs14071, CNhs13799, CNhs12323, CNhs11795, CNhs14075, CNhs12840, CNhs10649, CNhs12610, CNhs10648, CNhs10647, CNhs13801, CNhs12319, CNhs14549, CNhs14082, CNhs13795, CNhs12312, CNhs14227, CNhs14081, CNhs10646, CNhs13808, CNhs12322, CNhs14550, CNhs14080, CNhs13796, CNhs12310, CNhs14221, CNhs14069, CNhs13809, CNhs12316, CNhs14229, CNhs14552, CNhs14070, CNhs13800, CNhs12315, CNhs10645, CNhs14079, CNhs10644, CNhs13798, CNhs12320, CNhs14073, CNhs11787, CNhs11784, CNhs10643, CNhs10642, CNhs14226, CNhs13797, CNhs12317, CNhs10641, CNhs11782, CNhs14074, CNhs13804, CNhs12228, CNhs14230, CNhs13805, CNhs12229, CNhs14231, CNhs10640, CNhs10638, CNhs12324, CNhs13912, CNhs14225, CNhs14618, CNhs14083, CNhs13803, CNhs12318, CNhs14224,

	CNhs14076, CNhs10637, CNhs11772, CNhs12996, CNhs13794, CNhs12314, CNhs14223, CNhs14551, CNhs14084
female_genitalia	CNhs10618, CNhs10626, CNhs11676, CNhs11763, CNhs12854
tonsil	CNhs10654
trachea	CNhs10635, CNhs11766
gallbladder	CNhs12848
stomach	CNhs11771
heart	CNhs12855, CNhs12856, CNhs12857, CNhs11757, CNhs11758, CNhs10621, CNhs10653, CNhs11790, CNhs11789
bladder	CNhs10616
salivary_gland	CNhs11677, CNhs12852
parotid_gland	CNhs12849
eye	CNhs13444, CNhs13442, CNhs13443, CNhs13441, CNhs13440, CNhs11762, CNhs10636
blood_vessel	CNhs11760, CNhs12843, CNhs12844
kidney	CNhs10622, CNhs10652
skeletal_muscle	CNhs11779, CNhs13454, CNhs10629, CNhs11776
olfactory_region	CNhs12611
pancreas	CNhs11756
spinal_cord	CNhs13807, CNhs12227, CNhs14222, CNhs11764, CNhs14077

Table S2. Mouse tissue samples generated by the FANTOM5 consortium (phase 2.5) assigned to each facet. The data were kindly provided by the FANTOM5 consortium.

Sample_ID	**Facet**
prostate	CNhs10470
stomach	CNhs10503, CNhs10588, CNhs10603, CNhs11022, CNhs11006, CNhs10999, CNhs11193, CNhs11210, CNhs11104, CNhs11134
mammary_gland	CNhs10480, CNhs10476
adrenal_gland	CNhs10508, CNhs11038, CNhs11004, CNhs11043, CNhs11026, CNhs11191, CNhs11223
spinal_chord	CNhs10505
pituitary_gland	CNhs10493, CNhs11018, CNhs11009, CNhs11037, CNhs10592, CNhs11036, CNhs11039, CNhs11190
amnion	CNhs10488
submandibular_gland	CNhs10469
female_genitalia	CNhs10507, CNhs11040, CNhs11217, CNhs10500, CNhs10509, CNhs10497, CNhs10502
visual_cortex_tc	CNhs13040, CNhs13041, CNhs13821, CNhs13042, CNhs13043, CNhs13044, CNhs13045, CNhs13046, CNhs13048, CNhs13031, CNhs13032, CNhs13033, CNhs13820, CNhs13034, CNhs13035, CNhs13036, CNhs13037, CNhs13038, CNhs13039
urinary_bladder	CNhs10481
pancreas	CNhs10486, CNhs11012, CNhs11042, CNhs11003, CNhs10599, CNhs10580, CNhs11105, CNhs11138, CNhs11139, CNhs11136, CNhs11094, CNhs11182

liver	CNhs10466, CNhs10601, CNhs10524, CNhs10594, CNhs10520, CNhs10523, CNhs10510, CNhs10579, CNhs11117, CNhs11123, CNhs11101, CNhs11103, CNhs11115, CNhs11220, CNhs11198, CNhs11106
forelimb	CNhs10596, CNhs10600, CNhs10589, CNhs10577, CNhs11007, CNhs10598, CNhs11008
blood_vessel	CNhs10498
male_genitalia	CNhs11218, CNhs11199, CNhs10490
testis	CNhs10504, CNhs11031, CNhs11034, CNhs11033, CNhs11029, CNhs11027, CNhs11189, CNhs11222, CNhs11204, CNhs11110, CNhs11130
intestine	CNhs10467, CNhs10468, CNhs13199, CNhs10506, CNhs10496, CNhs11019, CNhs11010, CNhs10602, CNhs10585, CNhs10582, CNhs10526, CNhs11126, CNhs11192, CNhs11102, CNhs11095, CNhs11098, CNhs11187, CNhs11121, CNhs11131, CNhs11114
gonad	CNhs11044
bone	CNhs10483, CNhs11227, CNhs11225, CNhs11195
vescicular_gland	CNhs10491
thymus	CNhs10471, CNhs11041, CNhs11005, CNhs11002, CNhs10581, CNhs10595, CNhs11181, CNhs11137, CNhs11197, CNhs11211, CNhs11194, CNhs11186, CNhs11125, CNhs11132
lymphatic_intestinal_cells	CNhs13211, CNhs13200
lymph_node	CNhs10475
kidney	CNhs10606, CNhs10997, CNhs10584, CNhs11028, CNhs11001, CNhs11214, CNhs11206, CNhs11113, CNhs11122, CNhs11203
heart	CNhs10586, CNhs11015, CNhs11013, CNhs10597, CNhs11017, CNhs11021, CNhs11025, CNhs11030, CNhs11213, CNhs11221, CNhs11118, CNhs11209, CNhs11127, CNhs11196, CNhs11202
placenta	CNhs10472, CNhs10464
brain	CNhs10494, CNhs11135, CNhs10501, CNhs10487, CNhs11226, CNhs10473, CNhs11107, CNhs10482, CNhs11201, CNhs10478, CNhs11228, CNhs10477, CNhs11200, CNhs10489
tongue	CNhs10499
spleen	CNhs10465, CNhs11035, CNhs11011, CNhs11116, CNhs11112, CNhs11099
skin	CNhs10492, CNhs11124, CNhs11215, CNhs11097, CNhs11108
muscle	CNhs11129
lung	CNhs10474, CNhs10522, CNhs10604, CNhs11020, CNhs10998, CNhs10605, CNhs10583, CNhs11224, CNhs11212, CNhs11111, CNhs11219, CNhs11109, CNhs11119, CNhs11133
cerebellum_tc	CNhs12956, CNhs13002, CNhs13014, CNhs12957, CNhs13003, CNhs13015, CNhs12958, CNhs13004, CNhs13016, CNhs12960, CNhs13005, CNhs13017, CNhs12961, CNhs13006, CNhs13018, CNhs13000, CNhs13007, CNhs13019, CNhs12818, CNhs13008, CNhs13020, CNhs12962, CNhs13009, CNhs13021, CNhs12963, CNhs13010, CNhs13022, CNhs13001, CNhs13011, CNhs13024, CNhs12819, CNhs13012, CNhs13025, CNhs12820, CNhs13013, CNhs13026
eye	CNhs10484, CNhs11016, CNhs10521, CNhs10593, CNhs11023, CNhs11207, CNhs11140, CNhs11205, CNhs11188

Table S3. Assignments of human facets to the tissues of ENCODE TF ChIP-seq assays.

Facet	ENCODE tissue
blood	B cell
blood	erythroblast
blood vessel	endothelial cell of umbilical vein
bone marrow	osteoblast
brain	astrocyte
brain	neural cell
liver	liver
lung	fibroblast of lung
skeletal muscle	skeletal muscle myoblast
skeletal muscle	myotube
skin	fibroblast of dermis
smooth muscle	skeletal muscle myoblast
smooth muscle	myotube

Table S4. Construction of the human and mouse enhancer datasets. Number of regions for sets defined based on multiple criteria. Each set of regions is a subset of the set in the previous row.

	Human	Mouse
FANTOM enhancers	65,423	44,459
- **excluding FANTOM promoters and Ensembl coding exons**	54,284	38,662
- **active in our facets**	11,582	9,426
- **in TADs**	10,609	8,805
- **in TADs that comprise at least one active promoter**	10,445	8,762
- **significantly correlated with one or more promoters**	3,523	4,074

Table S5. Construction of the human and mouse promoter datasets. Number of regions for sets defined based on multiple criteria. Each set of regions is a subset of the set in the previous row.

	Human	Mouse
FANTOM promoters	201,799	158,965
- **active in our facets**	165,749	131,995
- **near an Ensembl TSS of a coding gene**	72,272	59,430
- **in TADs**	60,329	52,298
- **in TADs that comprise at least one active enhancer**	55,612	48,323
- **significantly correlated with one or more enhancers**	6,474	8,082

Table S6. Number of enhancers, enhancer pairs and enhancer groups for multiple sets in mouse. For a description of the sets see Table 1 in the main manuscript.

Set	Nr of enhancers	Nr of pairs	Nr of groups	Coloring in main figures
Correlated enhancers	4,074	-	-	Gray
Shadow enhancers	1,939	2,787	670	Purple
Transposon overlapping enhancers	1,507	-	-	Green
Transposon-shadow enhancers	503	493	192	Red/magenta

SUPPLEMENTARY MATERIALS AND METHODS

Target gene assignment with empirical P-values

In order to show the significance of our enhancer promoter correlation coefficients, we tested whether the correlation coefficient was significantly higher than correlations of the enhancer to a set of background genes. Specifically, for each enhancer, we calculated the correlations with each promoter in the 20 neighboring TADs. Since these correlations are unlikely to represent true interactions, we used them as a background distribution. The empirical P-value was defined as the number of background correlations with an equal or more extreme coefficient than the observed correlation divided by the total number of background correlations. The empirical P-values were FDR-corrected for multiple testing. The same thresholds as before were applied to the correlation coefficients and the (FDR-corrected) P-values. The so formed enhancer promoter associations were used to form pairs of redundant enhancers as before and these transposon annotation of these pairs were tested as before (see Figure S16).