Supplemental Information

Table of Contents

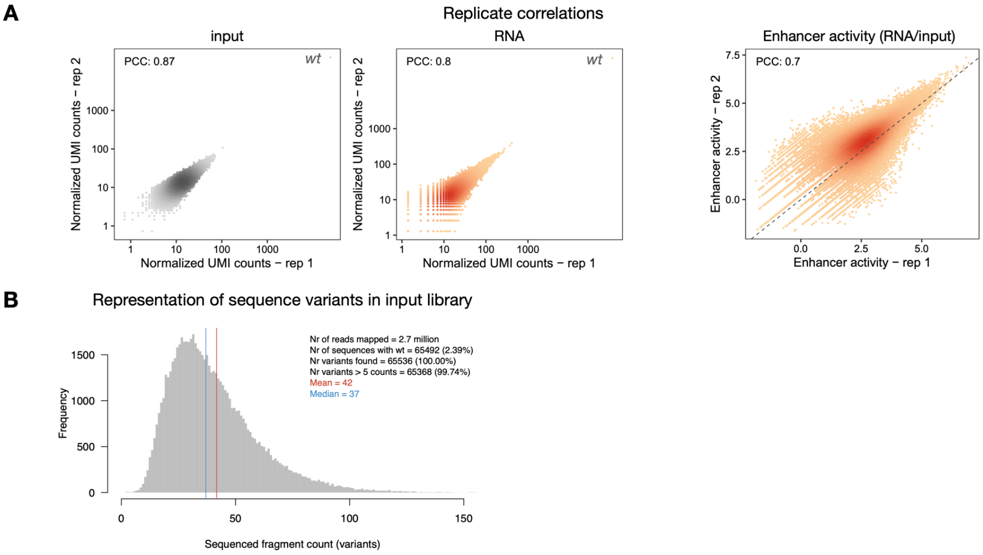
SUPPLEMENTAL TABLES.......27

SUPPLEMENTAL METHODS.......28

REFERENCES.......38

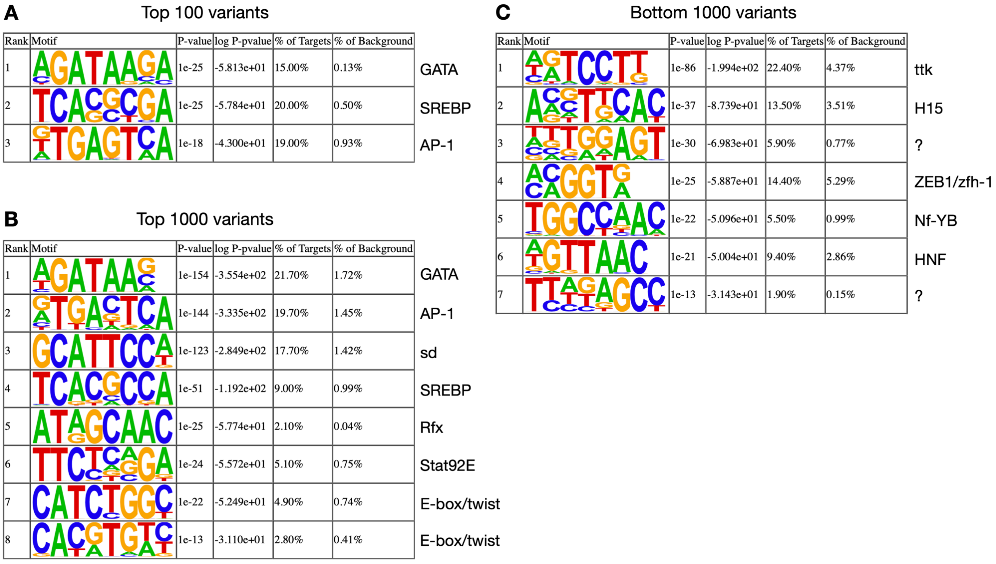
Supplemental Figures

Supplemental Fig S1. STARR-seq comprehensively assesses the activity of random variants in a specific region of the enhancer.



A) Pairwise comparisons of normalized STARR-seq input (left) and RNA (middle) UMI read counts or enhancer activity (RNA/input; right) between two independent biological replicates across all sequence variants tested in the GATA position (pos241) in the *ced-6* enhancer. Color reflects point density. The PCC is denoted for each comparison. Note the overrepresentation of the wild-type sequence both in the input and RNA libraries (top right corner), since it was used as the template for the PCR cloning (see Methods). **B)** Representation of sequence variants in STARR-seq input library. Frequency of variants covered by different number of UMI read counts. Number of sequences matching to wild type and the number of variants recovered are shown, together with the mean and median counts sequenced per variant.

Supplemental Fig S2. *De novo* motif discovery with Homer of top and bottom variants at the GATA position (pos241) in the *ced-6* enhancer.



A Top 100 variants

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1	AGATAAGA	1e-25	-5.813e+01	15.00%	0.13%	GATA
2	TCACGCGA	1e-25	-5.784e+01	20.00%	0.50%	SREBP
3	GTGAGTCA	1e-18	-4.300e+01	19.00%	0.93%	AP-1

B Top 1000 variants

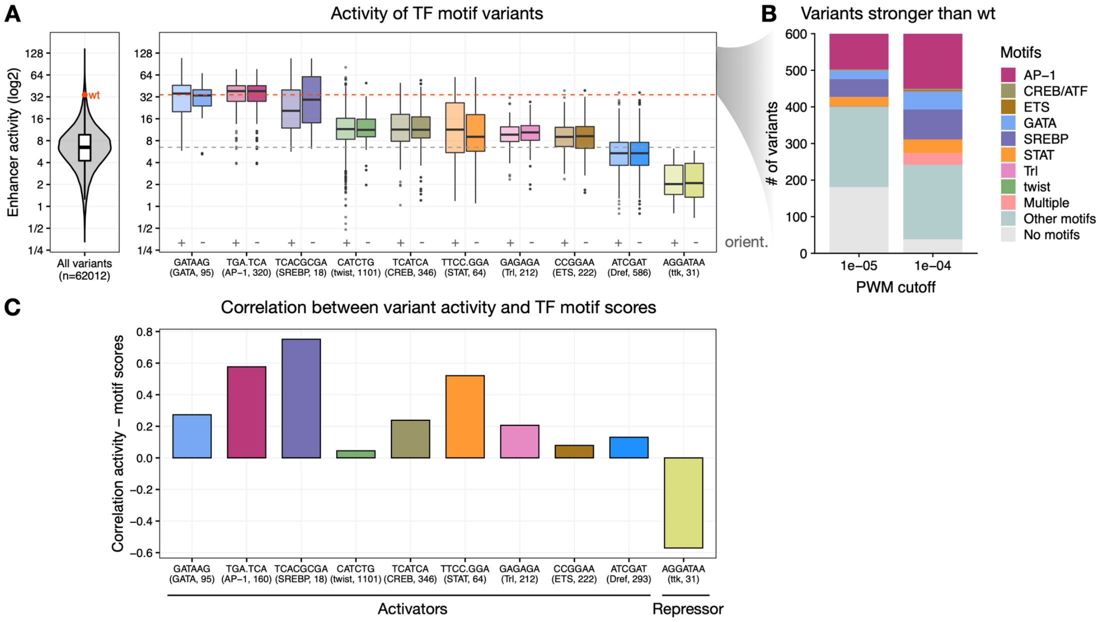
Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1	AGATAAG	1e-154	-3.554e+02	21.70%	1.72%	GATA
2	GTGACTCA	1e-144	-3.335e+02	19.70%	1.45%	AP-1
3	GCATTCCA	1e-123	-2.849e+02	17.70%	1.42%	sd
4	TCACGCCA	1e-51	-1.192e+02	9.00%	0.99%	SREBP
5	ATAGCAAC	1e-25	-5.774e+01	2.10%	0.04%	Rfx
6	TTCTCAGA	1e-24	-5.572e+01	5.10%	0.75%	Stat92E
7	CATCTGGC	1e-22	-5.249e+01	4.90%	0.74%	E-box/twist
8	CACGTGTC	1e-13	-3.110e+01	2.80%	0.41%	E-box/twist

C Bottom 1000 variants

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1	AGTCCTT	1e-86	-1.994e+02	22.40%	4.37%	ttk
2	ACGTTCAC	1e-37	-8.739e+01	13.50%	3.51%	H15
3	TTTGGAGT	1e-30	-6.983e+01	5.90%	0.77%	?
4	ACGGTG	1e-25	-5.887e+01	14.40%	5.29%	ZEB1/zfh-1
5	TGGCCAAC	1e-22	-5.096e+01	5.50%	0.99%	Nf-YB
6	AGTTAAC	1e-21	-5.004e+01	9.40%	2.86%	HNF
7	TTAGAGCC	1e-13	-3.143e+01	1.90%	0.15%	?

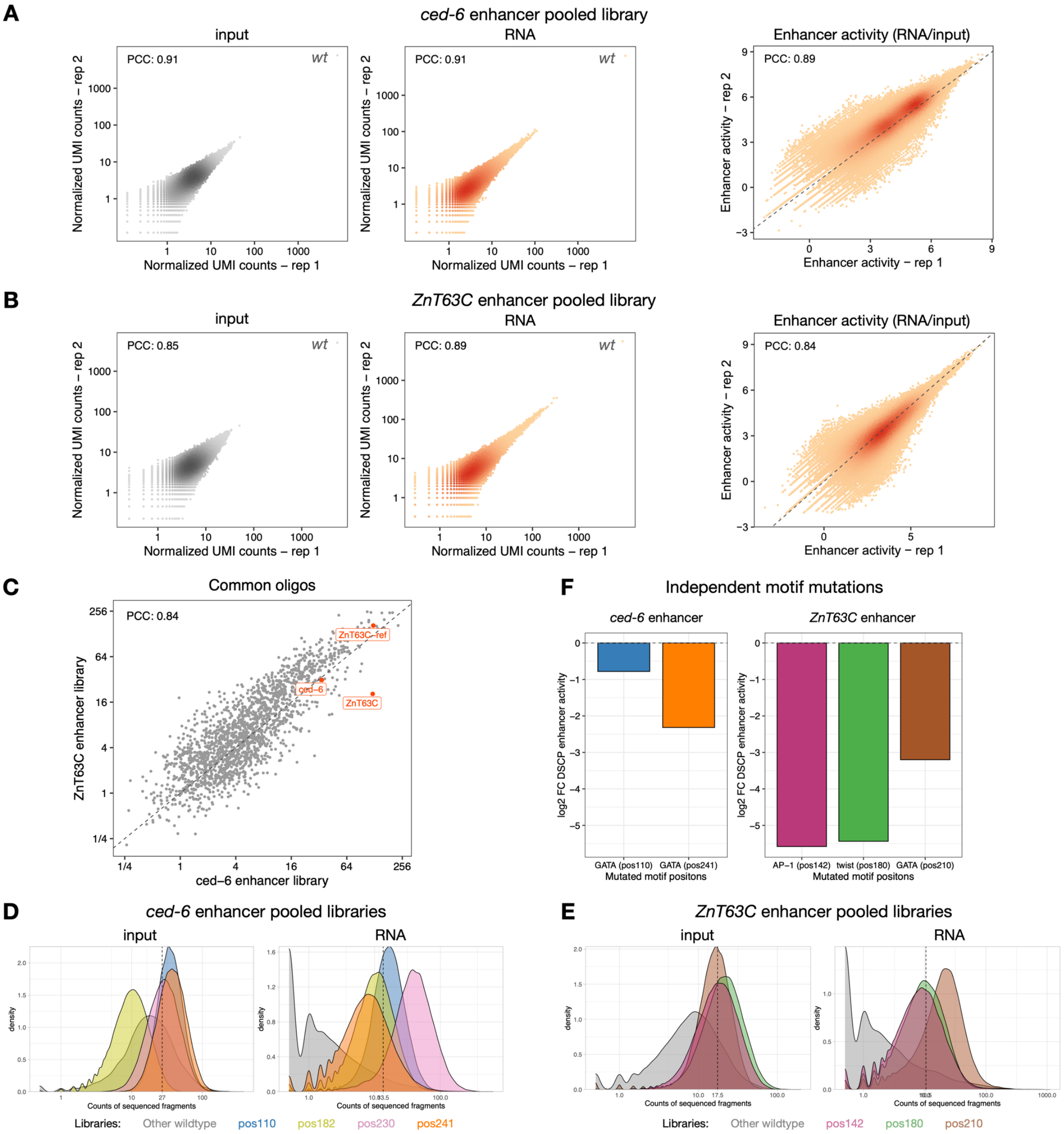
TF motifs found *de novo* (Homer) within the top 100 **(A)**, top 1,000 **(B)** or bottom 1,000 **(C)** variants. Motifs logo, statistics and predicted TF are shown.

Supplemental Fig S3. Activity of variants creating different TF motif types at the GATA position (pos241) in the *ced-6* enhancer.



A) Distribution of enhancer activity for all 62,012 enhancer variants (left) or variants creating each TF motif in either orientation (right; positive and negative orientation are shown in grey). The motif activities are independent of their orientation (Wilcoxon rank sum test p-value > 0.05). The activity of the wild-type sequence (wt, red dot and dashed line) or median of all variants (grey dashed line) are highlighted. The string of each TF motif used for the motif matching and the number of variants matching to each motif are described in the x-axis in the format “motif string (TF motif name, number of variants)”. **B)** Number of variants among the 600 stronger than wild type that match to motifs enriched in S2 developmental enhancers, using two different PWM p-value cutoffs (1e^{-05} and 1e^{-04}). **C)** Pearson correlation coefficient between variant activity and TF motif PWM scores. Note that for repressors, as ttk, the correlation is expected to be negative.

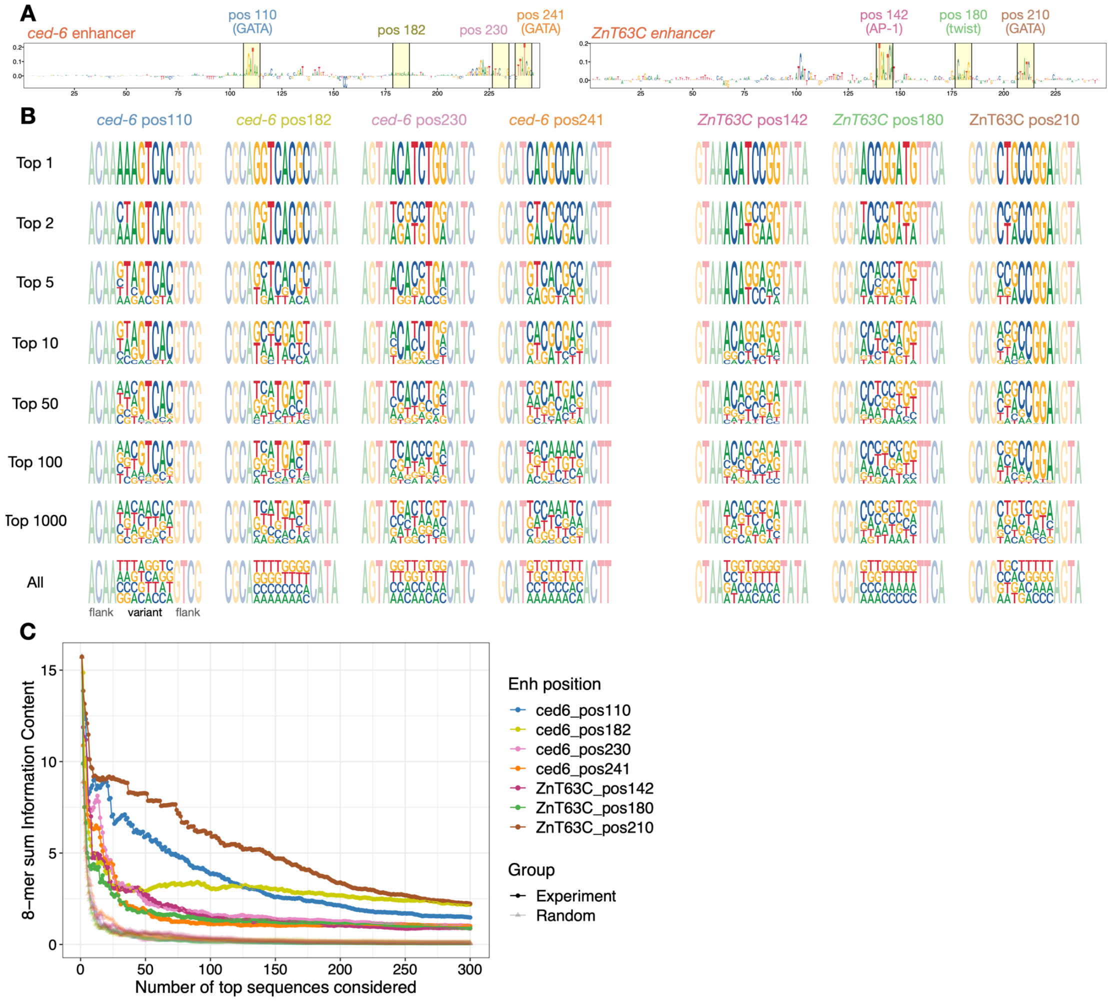
Supplemental Fig S4. STARR-seq screens with random variants in seven positions of two different enhancers.



A, B) Pairwise comparisons of normalized STARR-seq input (left) and RNA (middle) UMI read counts or enhancer activity (RNA/input; right) between two independent biological replicates across all sequence variants tested in positions of the *ced-6* **(A)** or *ZnT63C* **(B)** enhancer. Color reflects point density. The PCC is denoted for each comparison. Note the overrepresentation of the wild-type sequence both in the input and RNA libraries (top right corner), since it was used as the template for the PCR cloning (see Methods). **C)** Comparison of enhancer activity between the two different enhancer pooled libraries for the common oligos (a library of wild-type enhancer or negative sequences; see Methods). The PCC is shown. The respective wild-type enhancers are highlighted. Given the underestimation of the activity of the *ZnT63C* wild-type sequence in its pooled library, we used as reference wildt-ype activity the activity of another enhancer with similar activity that was conserved in both libraries (see Methods). **D, E)** Representation of sequence variants from each individual library (a library of wild-type enhancer and negative sequences, grey, or libraries with random variants in each enhancer position, different colors) in STARR-seq input and RNA pooled libraries of the *ced-6* **(D)** or *ZnT63C* **(E)** enhancer. The mean counts sequenced per variant is shown per pooled library with a dashed line. **F)** Importance of each motif position selected in the *ced-6* (Left) or *ZnT63C*

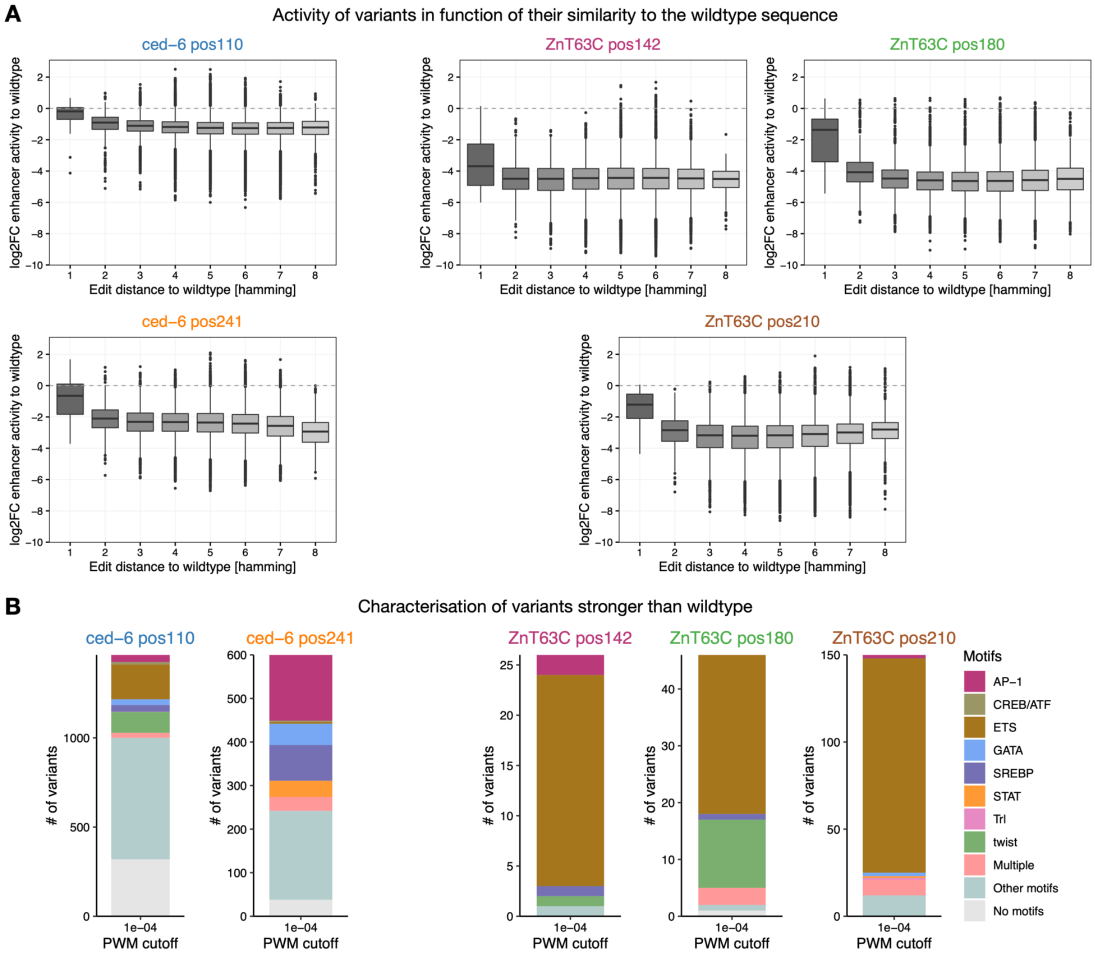
(Right) enhancer as judged by the impact of their individual mutation in enhancer activity ($\log_2$ fold-change). Data retrieved from *de Almeida et al., 2022* (de Almeida et al. 2022).

Supplemental Fig S5. Top active variants at each enhancer position are highly diverse.



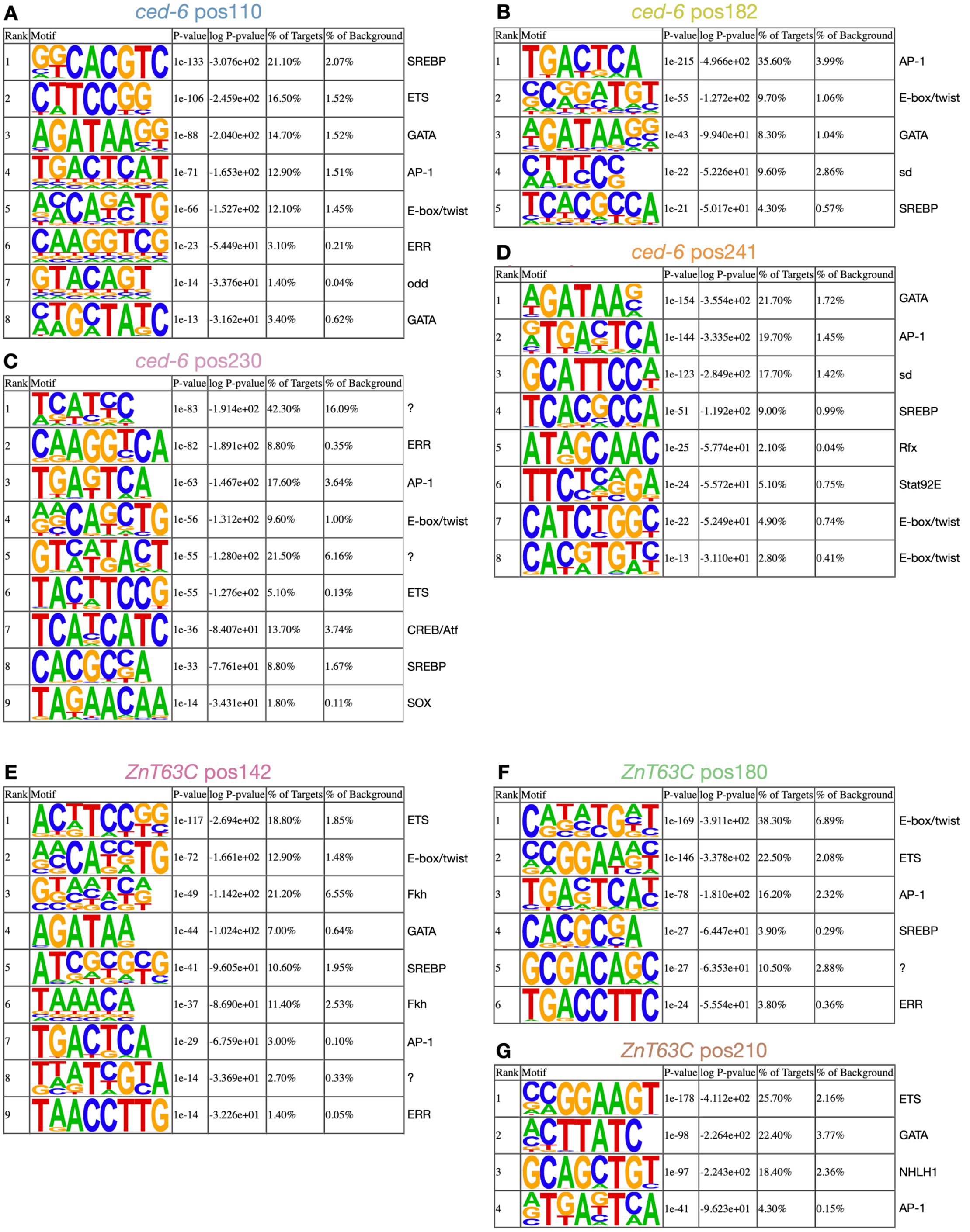
A) DeepSTARR-predicted nucleotide contribution scores for the *ced-6* (left) and *ZnT63C* (right) selected enhancer sequences. Selected 8nt motif positions and non-important control positions are highlighted in yellow with the respective numerical position, TF motif identity and different colors. **B)** Strong sequence variants are highly diverse. Logos with nucleotide frequency of the most-active variants in STARR-seq (1, 2, 5, 10, 50, 100, 1,000 and all) at each enhancer position (colored as in (A)). **C)** Sum of information content within the most-active 8-mers in STARR-seq (colored as in (A)) compared with the same after randomly sorting the variants (grey) for each enhancer position, considering different number of top sequences.

Supplemental Fig S6. Characterization of active variants.



A) Log$_2$ fold-change enhancer activity over the wild-type activity for all enhancer variants grouped by their edit distance (hamming distance) to the wild-type sequence, per enhancer position. **B)** Number of variants stronger than wild type that match to motifs enriched in S2 developmental enhancers is shown (PWM p-value cutoff 1e^{-04}), per enhancer position.

Supplemental Fig S7. *De novo* motif discovery with Homer of the top 1000 variants at the different enhancer positions.



A *ced-6* pos110

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1	GGCACGTC	1e-133	-3.076e+02	21.10%	2.07%	SREBP
2	CTTCCGG	1e-106	-2.459e+02	16.50%	1.52%	ETS
3	AGATAAGG	1e-88	-2.040e+02	14.70%	1.52%	GATA
4	TGACTCAT	1e-71	-1.653e+02	12.90%	1.51%	AP-1
5	ACCAGATG	1e-66	-1.527e+02	12.10%	1.45%	E-box/twist
6	CAAGGTCG	1e-23	-5.449e+01	3.10%	0.21%	ERR
7	GTACAGT	1e-14	-3.376e+01	1.40%	0.04%	odd
8	CTGCTATC	1e-13	-3.162e+01	3.40%	0.62%	GATA

B *ced-6* pos182

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1	TGACTCA	1e-215	-4.966e+02	35.60%	3.99%	AP-1
2	CCGGATGT	1e-55	-1.272e+02	9.70%	1.06%	E-box/twist
3	AGATAAGG	1e-43	-9.940e+01	8.30%	1.04%	GATA
4	CTTTCC	1e-22	-5.226e+01	9.60%	2.86%	sd
5	TCACGCCA	1e-21	-5.017e+01	4.30%	0.57%	SREBP

C *ced-6* pos230

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1	TCATCC	1e-83	-1.914e+02	42.30%	16.09%	?
2	CAAGGTCA	1e-82	-1.891e+02	8.80%	0.35%	ERR
3	TGAGTCA	1e-63	-1.467e+02	17.60%	3.64%	AP-1
4	AACAGCTG	1e-56	-1.312e+02	9.60%	1.00%	E-box/twist
5	GTCATACT	1e-55	-1.280e+02	21.50%	6.16%	?
6	TACTTCCG	1e-55	-1.276e+02	5.10%	0.13%	ETS
7	TCATCATC	1e-36	-8.407e+01	13.70%	3.74%	CREB/Atf
8	CACGCCA	1e-33	-7.761e+01	8.80%	1.67%	SREBP
9	TAGAACAA	1e-14	-3.431e+01	1.80%	0.11%	SOX

D *ced-6* pos241

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1	AGATAAG	1e-154	-3.554e+02	21.70%	1.72%	GATA
2	GTGACTCA	1e-144	-3.335e+02	19.70%	1.45%	AP-1
3	GCATTCCA	1e-123	-2.849e+02	17.70%	1.42%	sd
4	TCACGCCA	1e-51	-1.192e+02	9.00%	0.99%	SREBP
5	ATAGCAAC	1e-25	-5.774e+01	2.10%	0.04%	Rfx
6	TTCTCAGA	1e-24	-5.572e+01	5.10%	0.75%	Stat92E
7	CATCTGGC	1e-22	-5.249e+01	4.90%	0.74%	E-box/twist
8	CACGTGTC	1e-13	-3.110e+01	2.80%	0.41%	E-box/twist

E *ZnT63C* pos142

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1	ACTTCCGG	1e-117	-2.694e+02	18.80%	1.85%	ETS
2	AACACCTG	1e-72	-1.661e+02	12.90%	1.48%	E-box/twist
3	GTAATCA	1e-49	-1.142e+02	21.20%	6.55%	Fkh
4	AGATAA	1e-44	-1.024e+02	7.00%	0.64%	GATA
5	ATCGCGCG	1e-41	-9.605e+01	10.60%	1.95%	SREBP
6	TAAACA	1e-37	-8.690e+01	11.40%	2.53%	Fkh
7	TGACTCA	1e-29	-6.759e+01	3.00%	0.10%	AP-1
8	TTATCGTA	1e-14	-3.369e+01	2.70%	0.33%	?
9	TAACCTTG	1e-14	-3.226e+01	1.40%	0.05%	ERR

F *ZnT63C* pos180

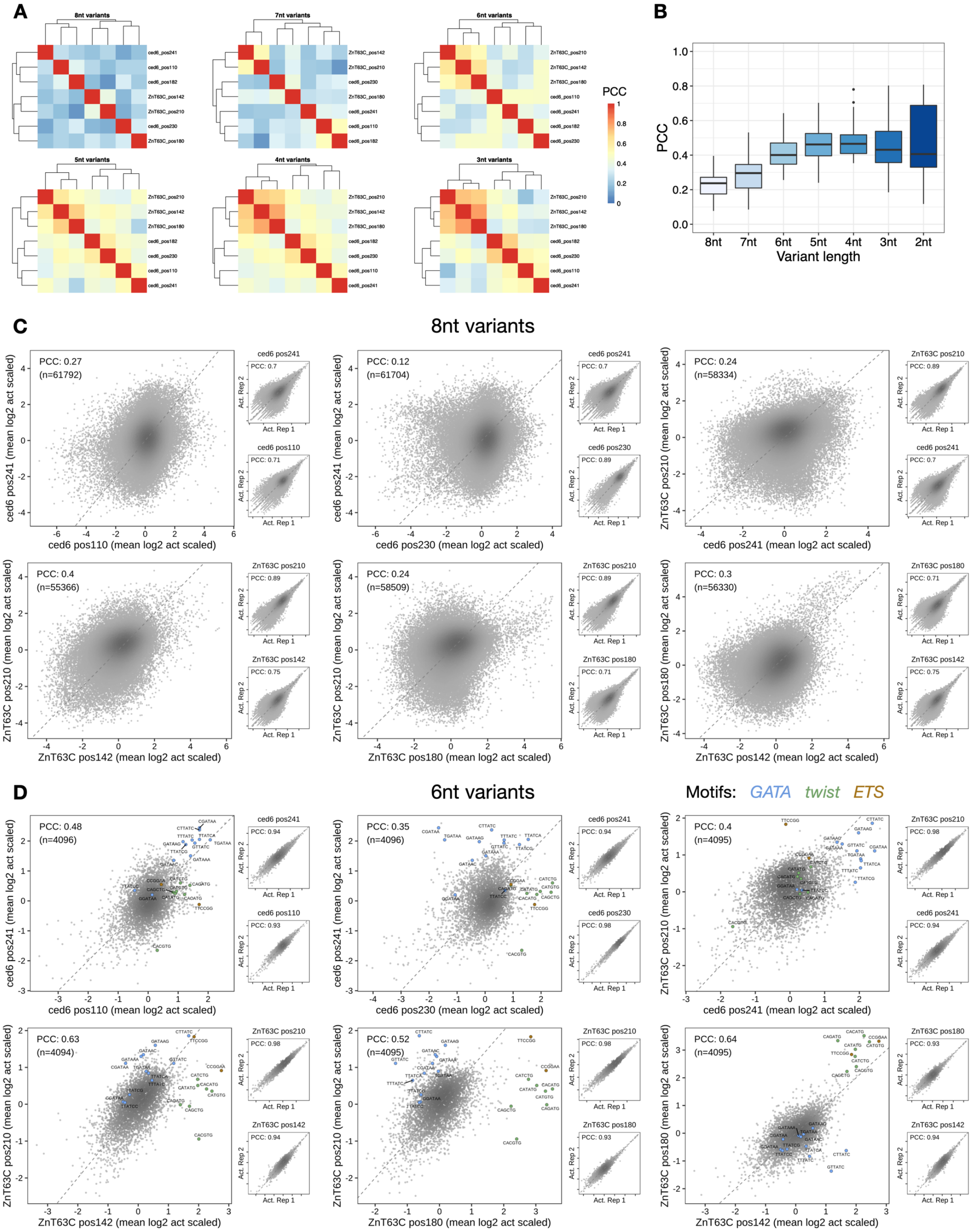
Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1	CATATGAT	1e-169	-3.911e+02	38.30%	6.89%	E-box/twist
2	CCGGAAAC	1e-146	-3.378e+02	22.50%	2.08%	ETS
3	TGACTCAT	1e-78	-1.810e+02	16.20%	2.32%	AP-1
4	CACGCGA	1e-27	-6.447e+01	3.90%	0.29%	SREBP
5	GCGACAGC	1e-27	-6.353e+01	10.50%	2.88%	?
6	TGACCTTC	1e-24	-5.554e+01	3.80%	0.36%	ERR

G *ZnT63C* pos210

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1	CCGGAAGT	1e-178	-4.112e+02	25.70%	2.16%	ETS
2	ACTTATC	1e-98	-2.264e+02	22.40%	3.77%	GATA
3	GCAGCTGT	1e-97	-2.243e+02	18.40%	2.36%	NHLH1
4	ATGAGTCA	1e-41	-9.623e+01	4.30%	0.15%	AP-1

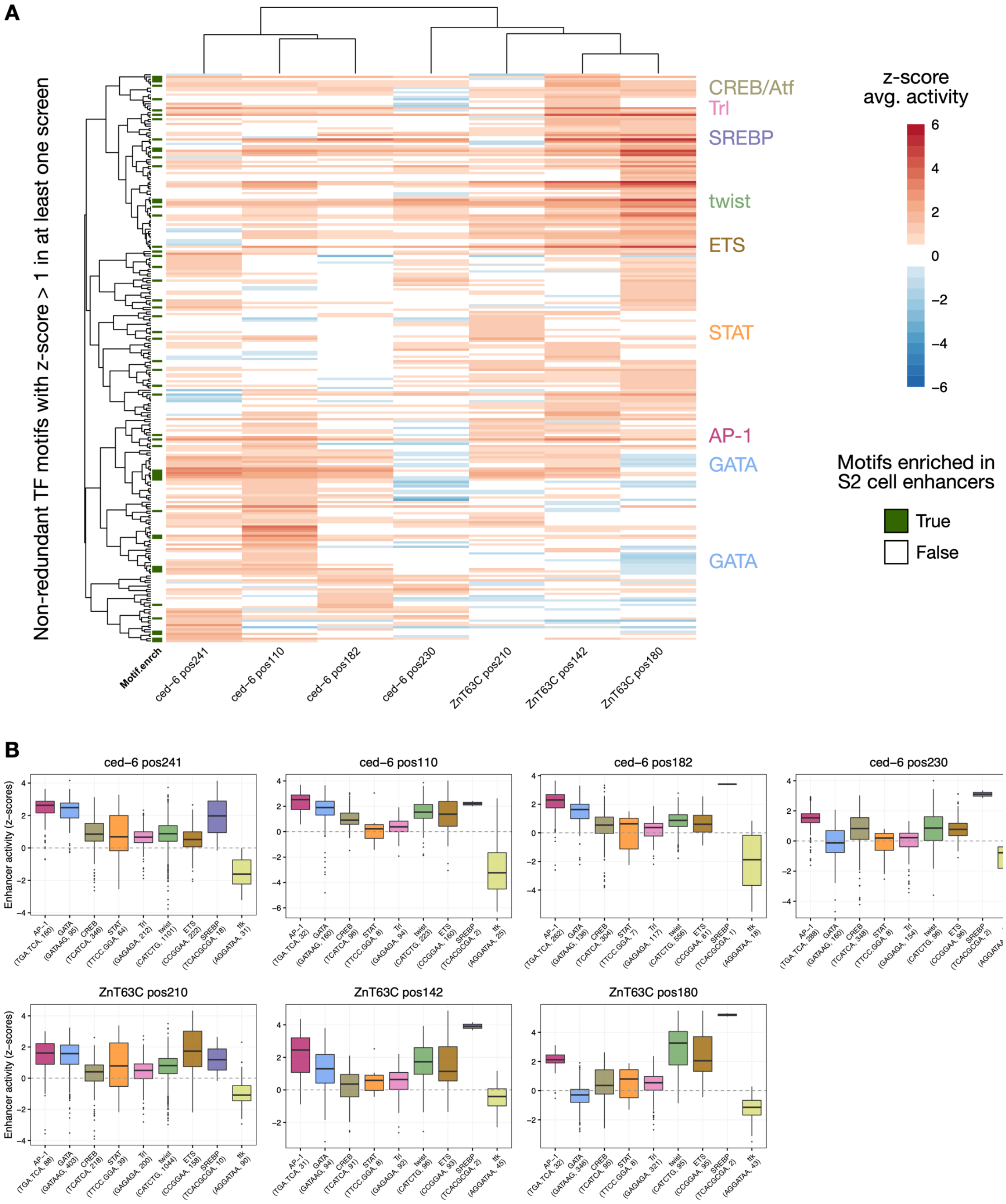
TF motifs found *de novo* (Homer) within the top 1,000 variants at each enhancer position. Motifs logo, statistics and predicted TF are shown.

Supplemental Fig S8. Comparison of all random variants across enhancer positions.



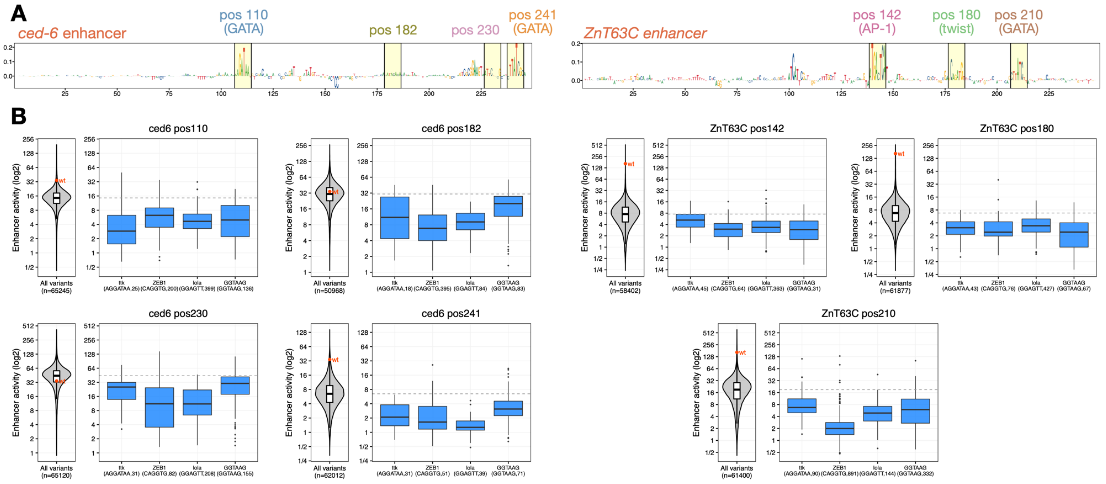
A) Hierarchical clustering of all enhancer positions based on PCC of variant enhancer activities in each position, when considering different lengths of sequence variants (see Methods). **B)** Distribution of PCCs from (A) in function of the length of sequence variants considered. **C,D)** Comparison of *z*-scores of $\log_2$ enhancer activity of all 8nt **(C)** or 6nt **(D**; see Methods**)** variants between enhancer positions (insets show activity for replicates (Act. Rep) 1 versus 2 for each position). Color reflects the enhancer position and point density. PCCs and number of sequence variants are shown. Variants matching to GATA, twist and ETS motifs are highlighted in (D).

Supplemental Fig S9. Activity of TF motif types at different enhancer positions.



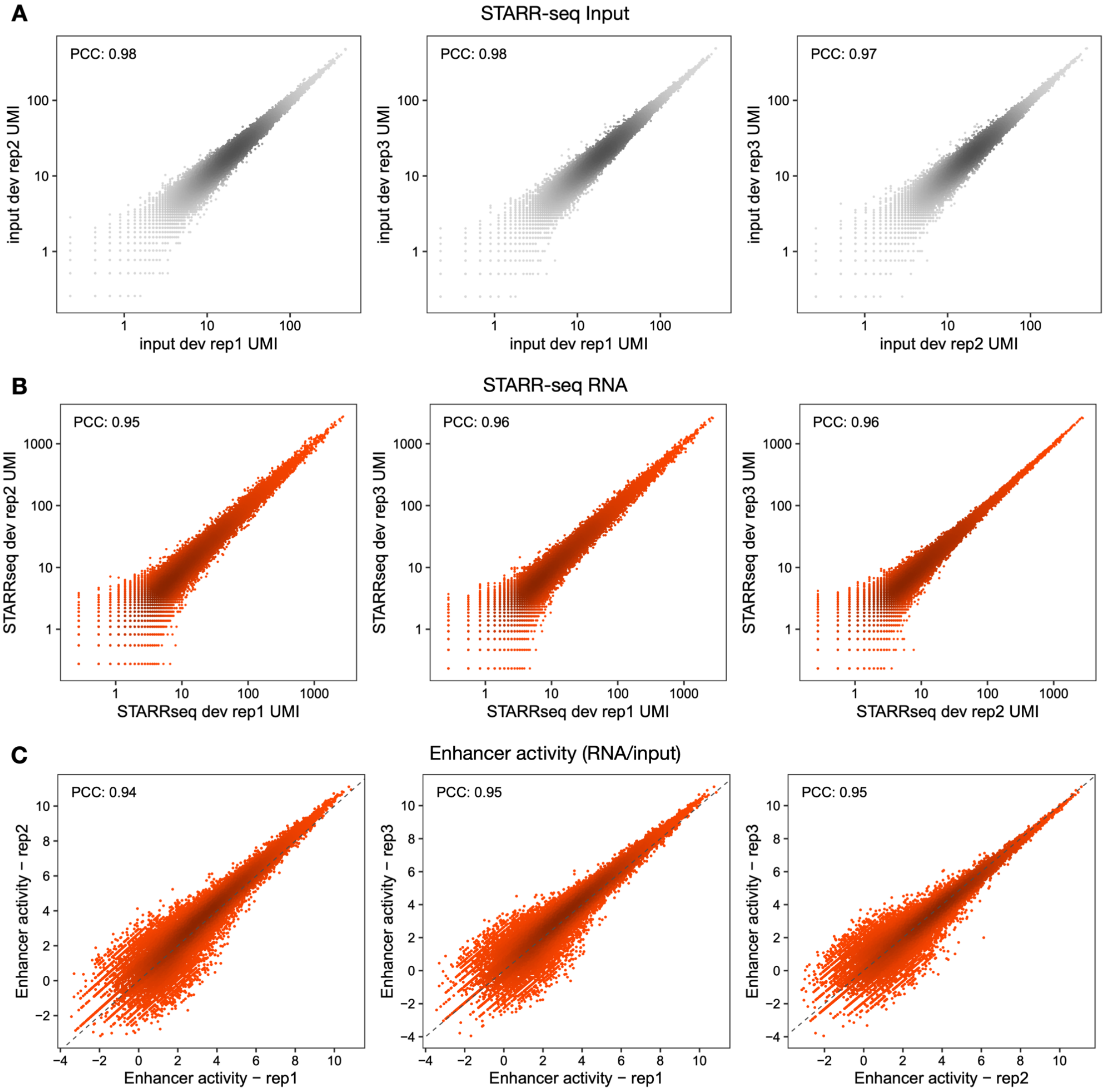
A) Heatmap of average z-scores of $\log_2$ enhancer activity of variants creating each TF motif type across all seven enhancer positions. Only motif types active (average z-score > 1) in at least one position are shown. Motifs and enhancer positions were clustered using hierarchical clustering and their activity is colored in shades of red (activating) and blue (repressing). Motifs enriched in S2 cell enhancers are labelled in green. Motif types used in the motif pasting experiment are highlighted. **B)** Activity of different TF motifs at each enhancer position. Distribution of z-scores of $\log_2$ enhancer activity for variants creating each TF motifs in *ced-6* and *ZnT63C* enhancer positions.

Supplemental Fig S10. STARR-seq identifies known and novel motifs that repress enhancer activity.



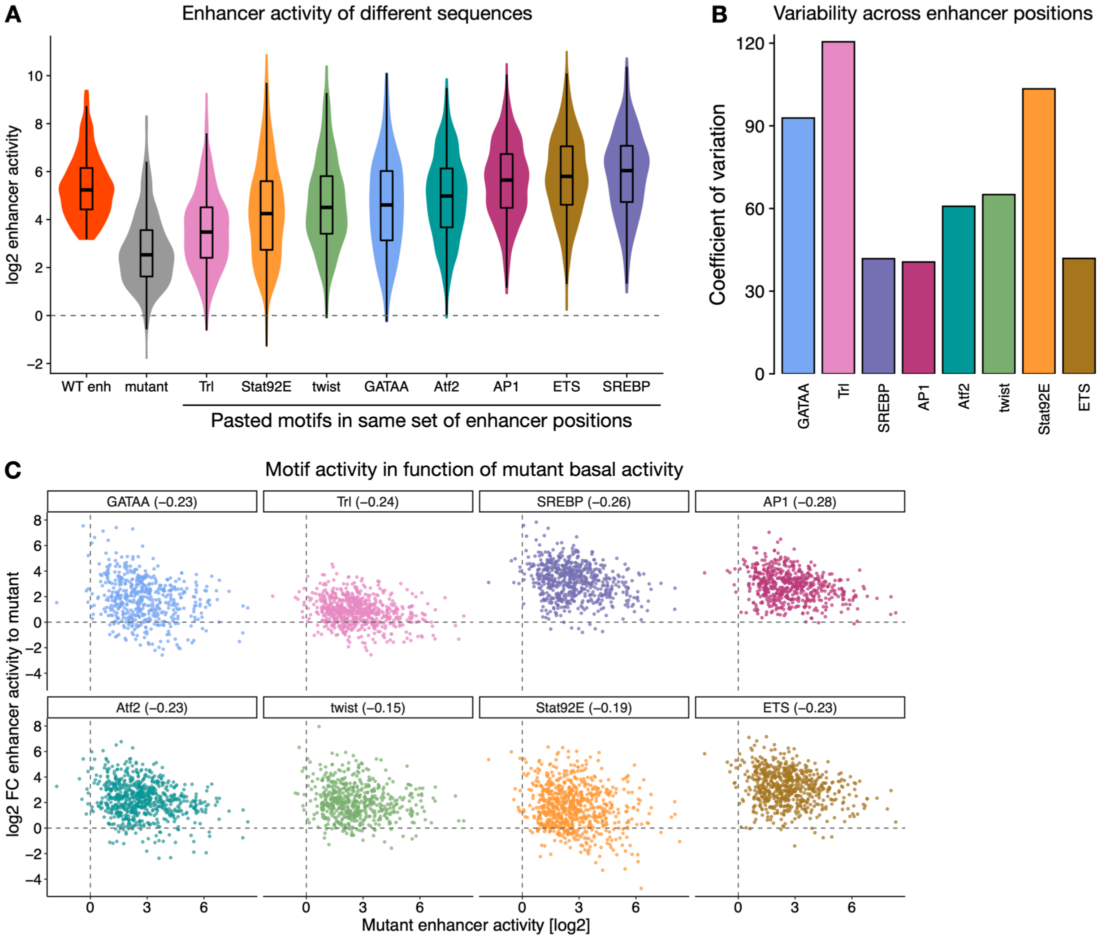
A) DeepSTARR-predicted nucleotide contribution scores for the *ced-6* (left) and *ZnT63C* (right) selected enhancer sequences. Selected 8nt motif positions and non-important control positions are highlighted in yellow with the respective numerical position, TF motif identity and different colors. **B)** Activity of different repressor motifs at each enhancer position. Distribution of enhancer activity for all enhancer variants (left) or variants creating each repressor TF motif (right), per enhancer position. The activity of the wild-type sequence (wt, red) or median of all variants (grey dashed line) are shown. The string of each TF motif used for the motif matching and the number of variants matching to each motif are described in the x-axis: in the format “motif string (TF motif name, number of variants)”.

Supplemental Fig S11. Systematic motif pasting screens in *Drosophila* enhancers.



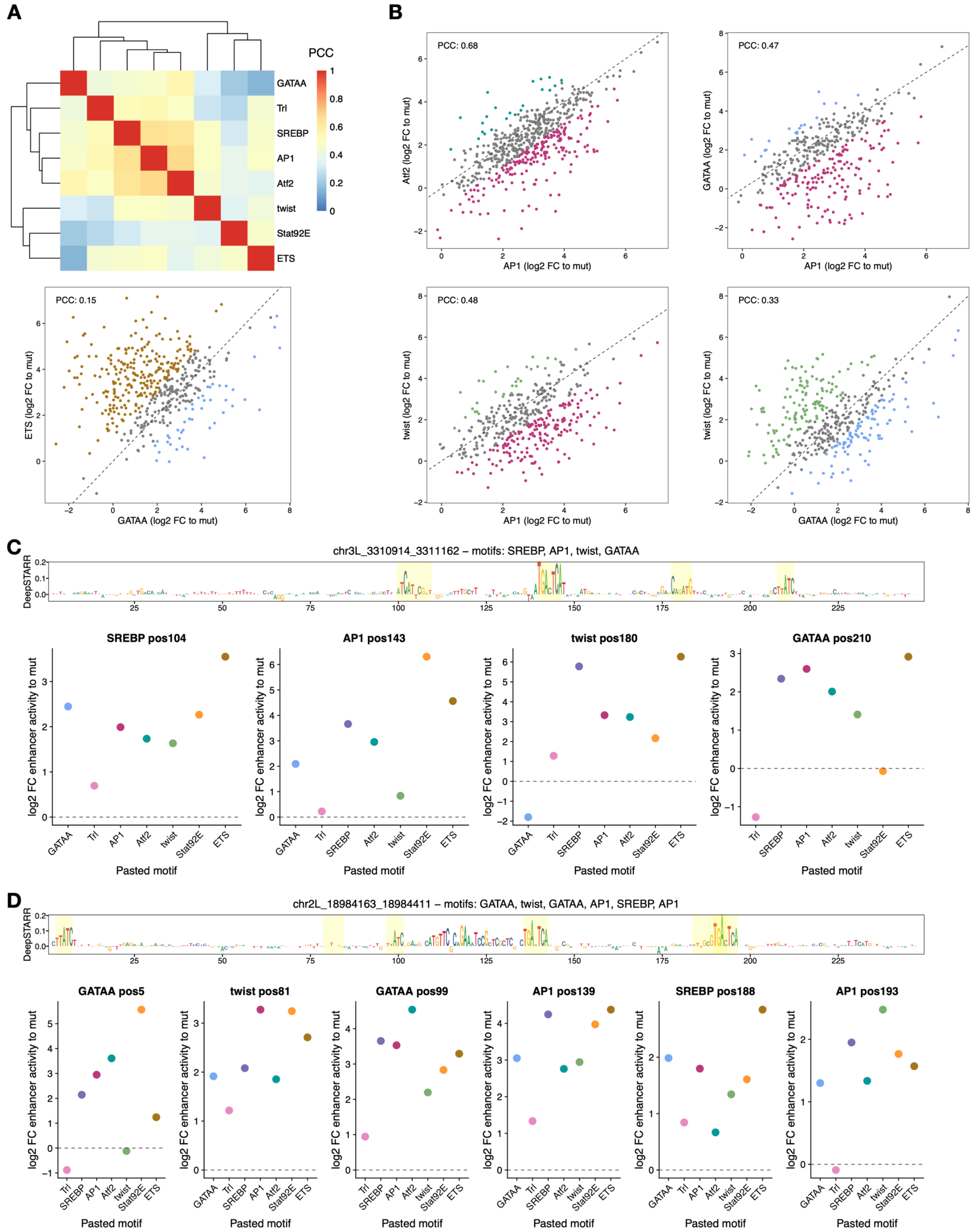
Pairwise comparisons of normalized STARR-seq input **(A)** and RNA **(B)** UMI read counts or enhancer activity (RNA/input) **(C)** between three independent biological replicates across all oligos tested. Color reflects point density. The PCC is denoted for each comparison.

Supplemental Fig S12. Enhancer activity of different sequences in Drosophila enhancers.



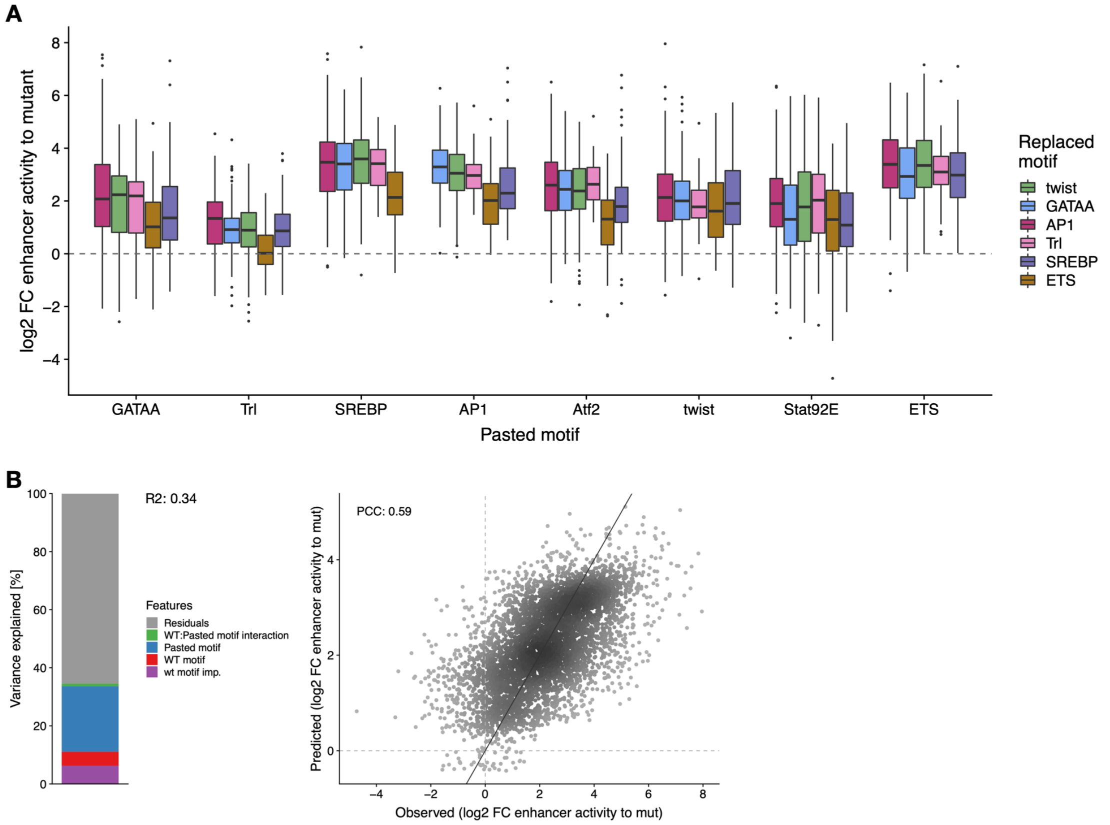
A) Activity of pasted motifs at different enhancer positions. Distribution of enhancer activity changes ($\log_2$) of all wild-type enhancers used and their variants with either mutant sequences or different TF motifs pasted. Few instances show negative values: these are not dependent on the specific mutant sequence but rather correspond to the creation of a repressor motif at the flanks of the pasted motif and the backbone enhancer. **B)** Bar plots showing the coefficient of variation (ratio of the standard deviation to the mean) of the activity of each TF motif across all enhancer positions. **C)** Activity of pasting motifs (y-axis, $\log_2$ fold-change activity over basal motif-mutated enhancer activity) in function of the basal activity (x-axis, activity of motif-mutated enhancer). The PCC is denoted for each motif.

Supplemental Fig S13. Motifs work differently at different enhancer positions.



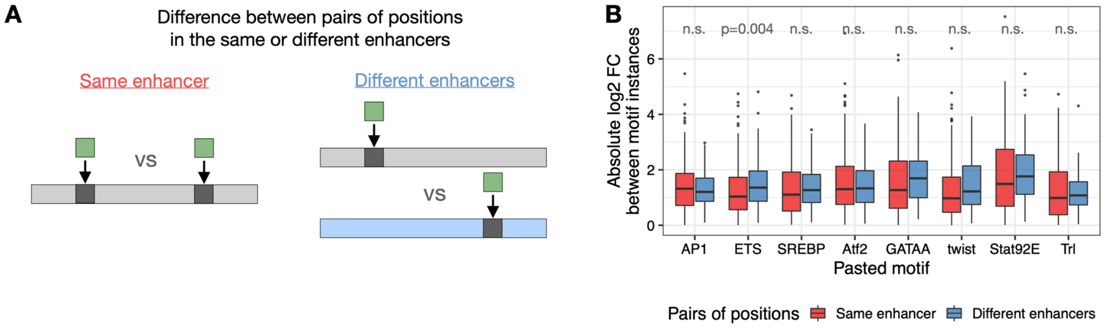
A) Hierarchical clustering of all TF motifs based on PCC of motif activities across all enhancer positions. **B)** Motifs work differently at different enhancer positions. Comparison between enhancer activity changes (log$_2$ FC to mutated sequence) after pasting different TF motifs across all enhancer positions. Positions with stronger activity of each motif (>= 2-fold in respect to the other motif) are colored with the respective colors. PCC: Pearson correlation coefficient. **C,D)** DeepSTARR-predicted nucleotide contribution scores for two enhancers and respective positions (highlighted in yellow, with wild-type motif types described on top) included in the screen. For each position, the enhancer activity changes (log$_2$ FC to mutated sequence) after pasting each TF motif are shown in dot plots (bottom).

Supplemental Fig S14. TF motif activity in function of wild-type motif identity.



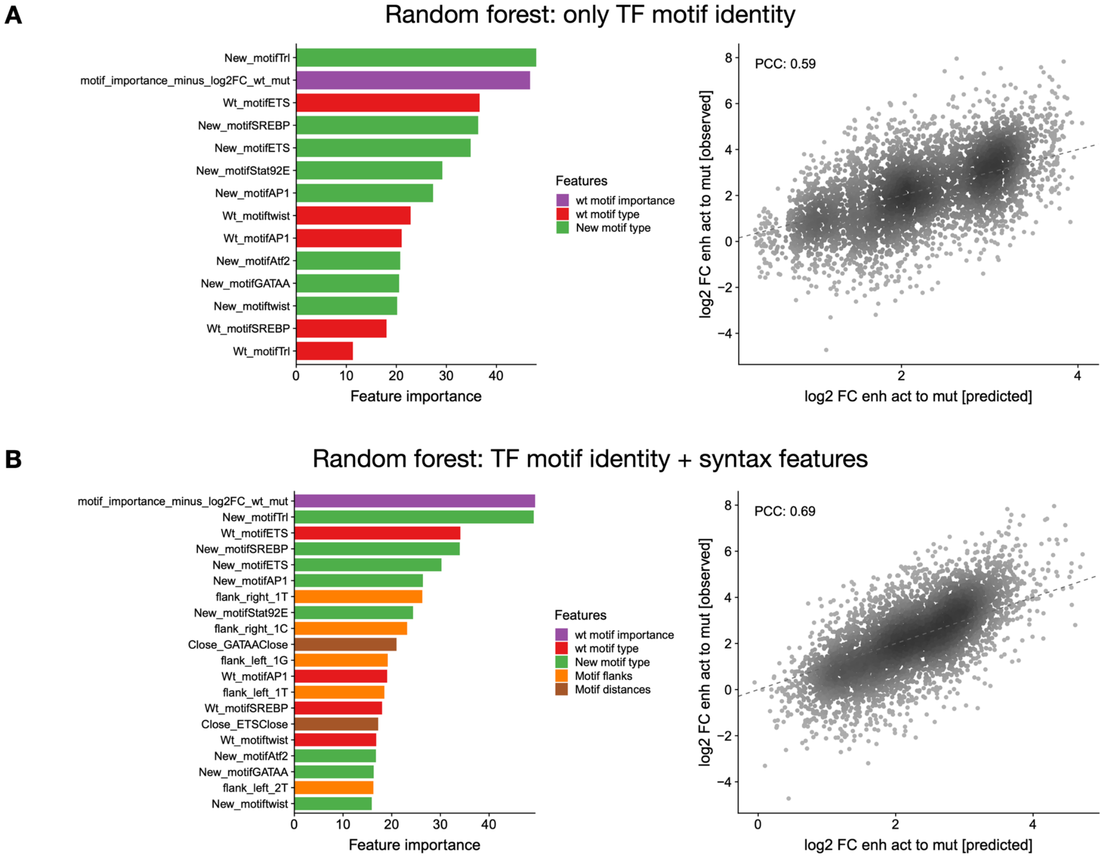
A) Distribution of enhancer activity changes ($\log_2$ FC to mutated sequence) across all enhancer positions for each pasted TF motif, grouped by the identity of the wild-type motif. **B)** Left: Bar plot showing the amount of variance explained by the wild-type motif importance and identity, the pasted motif identity and the interaction between the wild type and pasted motifs, using a linear model fit on all motif pasting results. Right: Scatter plots of predicted (linear model) vs. observed enhancer activity changes ($\log_2$ FC to mutated sequence) across all motif pasting experiments. Color reflects point density. PCC is shown.

Supplemental Fig S15. Motif activity in different positions in the same or different enhancers.



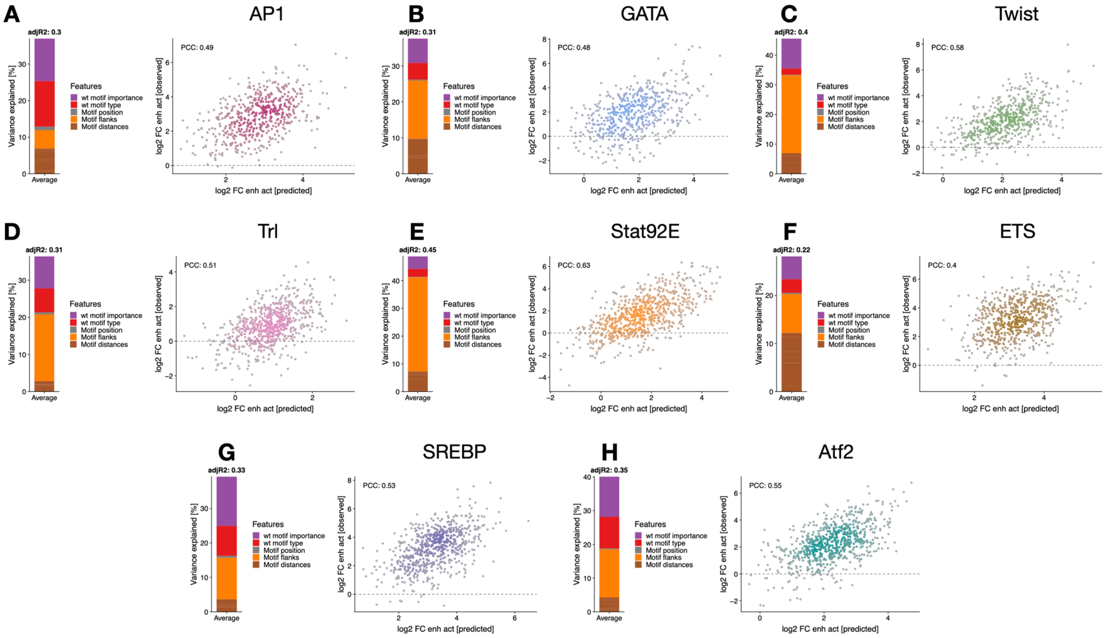
A) Schematics of comparison of motif activity between instances within the same enhancer or in different enhancers. **B)** Absolute $\log_2$ fold-change in enhancer activity between instances within the same enhancer (red) or in different enhancers (blue) for each pasted TF motif type. n.s. non-significant (Wilcoxon signed rank test).

Supplemental Fig S16. Prediction of motif activities using motif syntax features in random forest model.



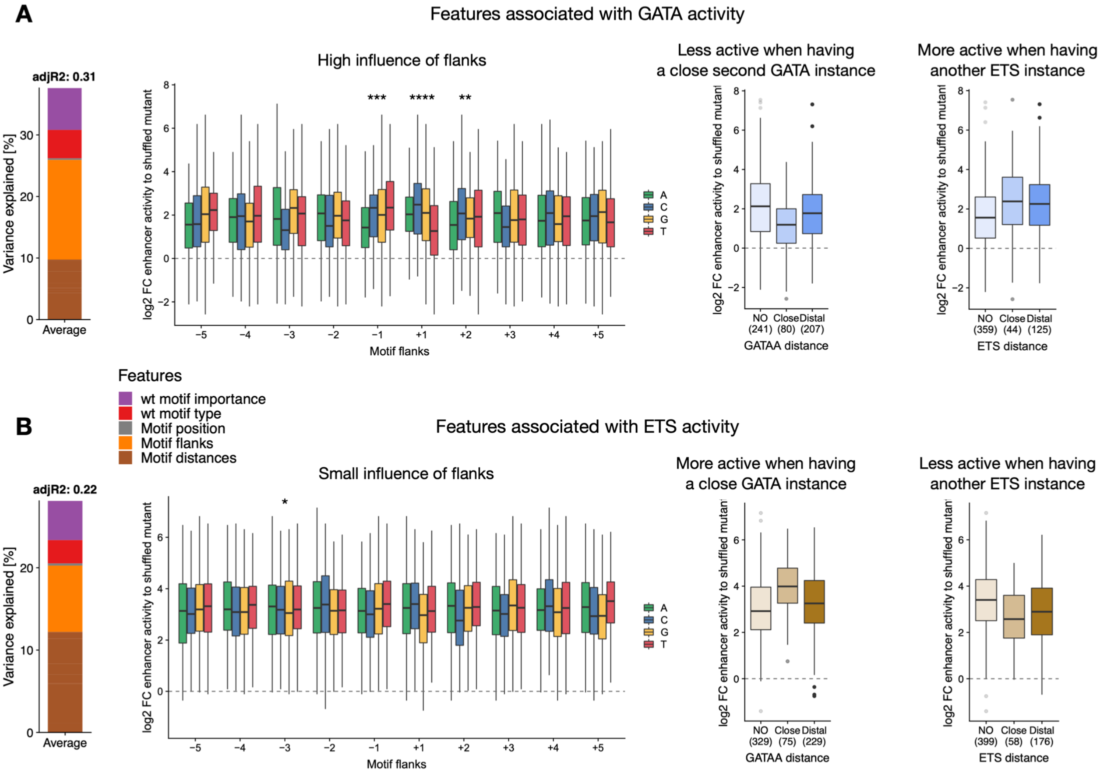
Left: Importance of all features **(A)** or only the top 20 **(B)** included in the random forest models with only TF motif identity **(A)** or also with syntax features **(B)**, sorted by importance and colored by feature type. Right: Scatter plots of predicted vs. observed enhancer activity changes (log$_2$ FC to mutated sequence) across all motif pasting experiments. Color reflects point density. PCC is shown.

Supplemental Fig S17. Linear models with syntax features to predict motif activities.



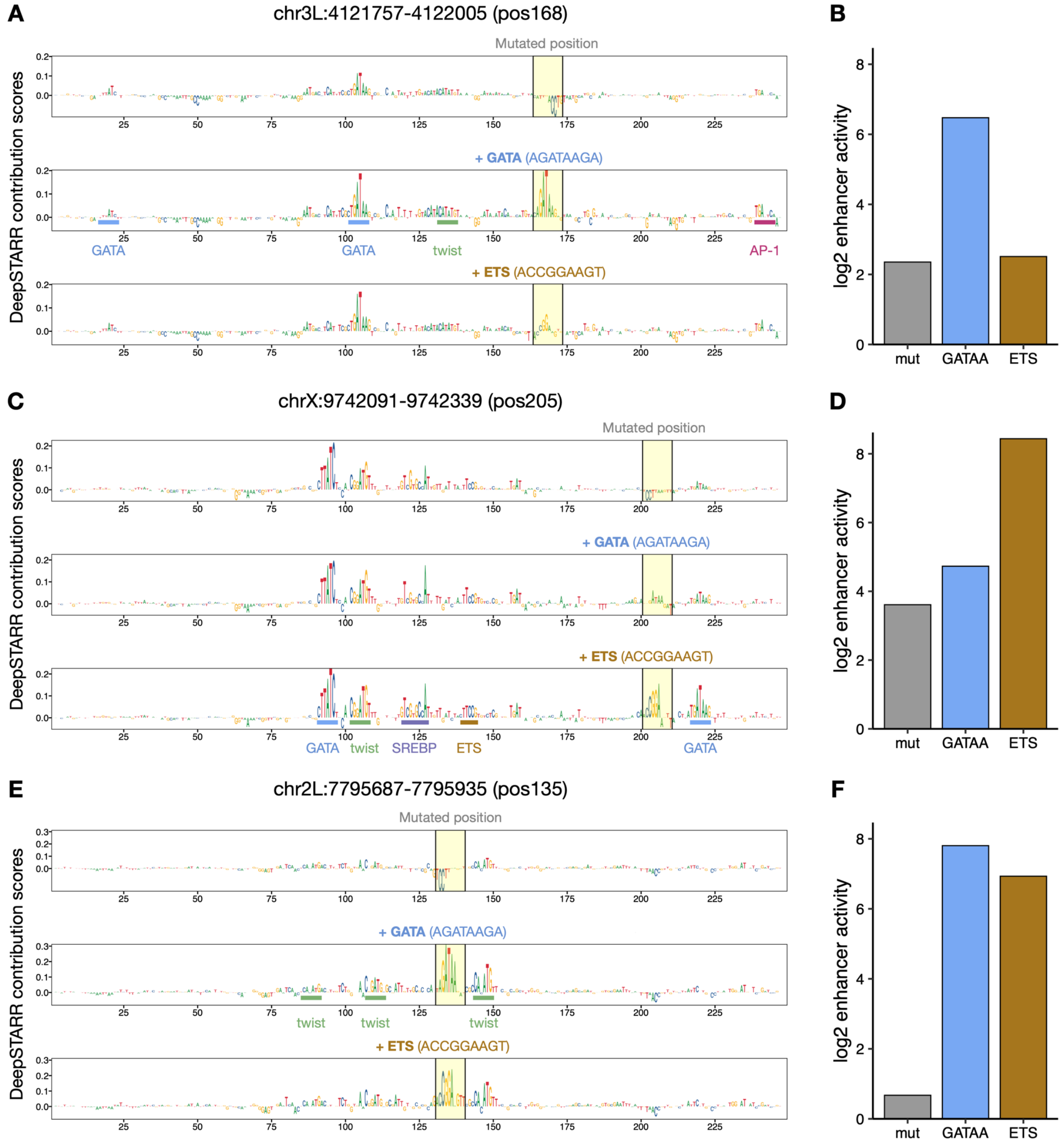
A-H) Left: Bar plot showing the variance explained by the different types of features (color legend) for each of the linear models. Right: Scatter plots of predicted vs. observed enhancer activity changes (log$_2$ FC to mutated sequence) for motif pasting experiments per TF motif type. Color reflects point density. PCC is shown.

Supplemental Fig S18. Characterization of preferred syntax features of GATA and ETS motifs.



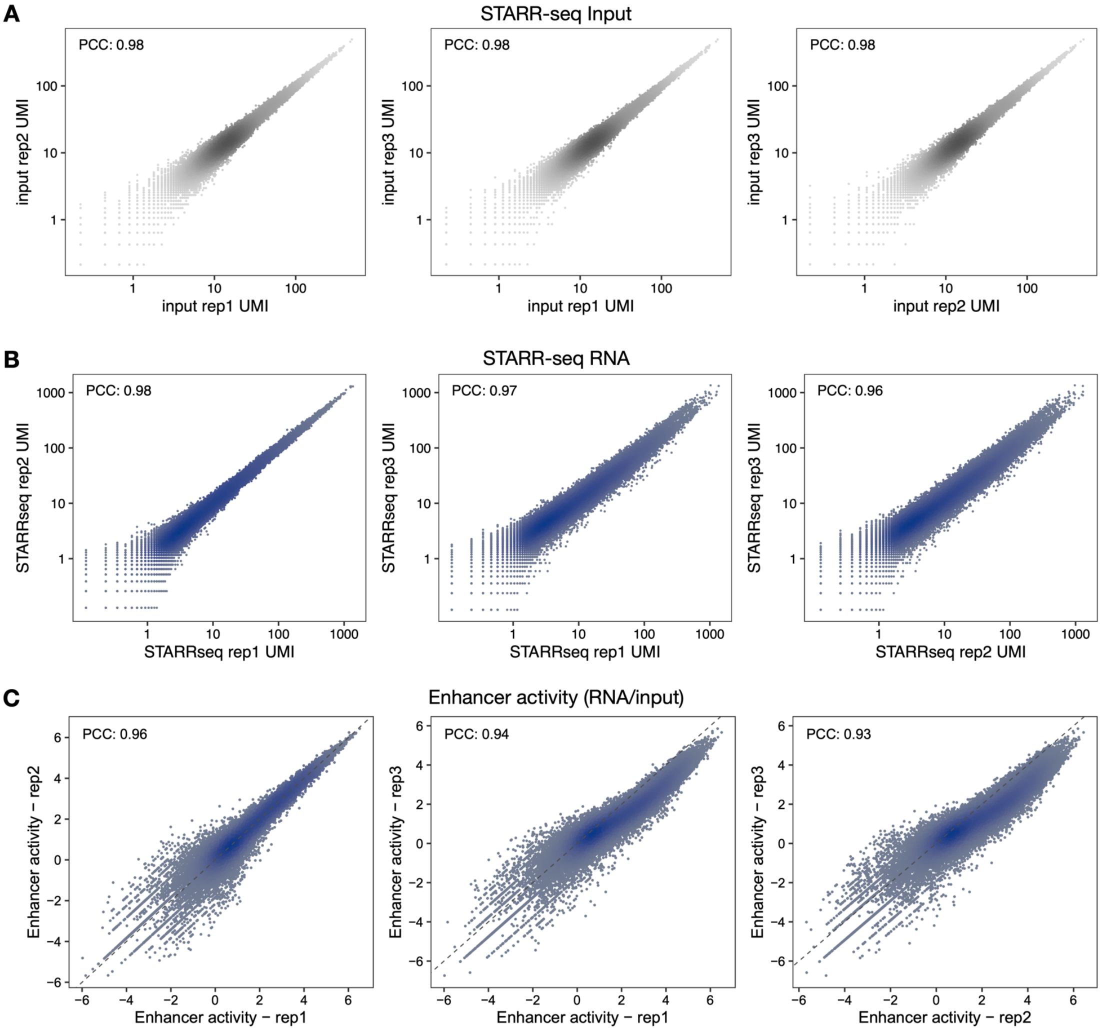
Syntax features associated with GATA **(A)** or ETS **(B)** activity. Left: bar plot showing the variance explained by the different types of features (color legend) for each of the linear models. Middle-left: motif activity according to the different bases at each flanking position, colored by nucleotide identity. Statistics from linear model in Fig 4A: ****P < 0.0001, ***P < 0.001, **P < 0.01, *P < 0.05 (linear regression p-value). Middle-right and right: enhancer activity changes (log$_2$ FC to mutated sequence) after pasting each TF motif in positions with no additional GATA (middle-right) or ETS (right) in the enhancer, or with additional GATA or ETS at close (<= 25 bp) or distal (>25 bp) distances. Number of instances are shown.

Supplemental Fig S19. DeepSTARR-predicted importance scores for pasting GATA or ETS in the same positions.



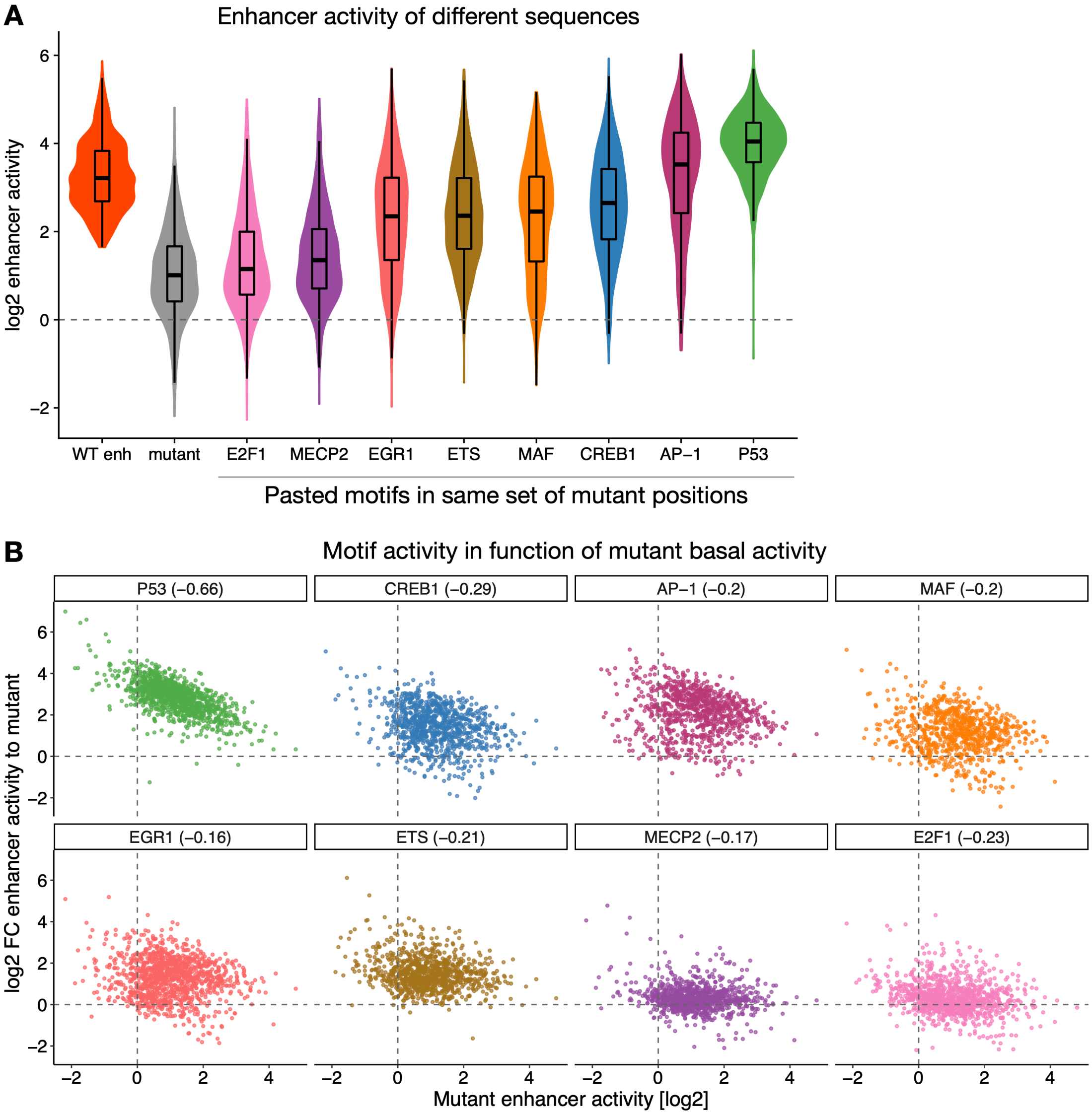
A,C,E) DeepSTARR-predicted nucleotide contribution scores for three different enhancers with a mutant sequence, GATA or ETS pasted at the highlighted positions. Motif sequences pasted are shown. **B,D,F)** Bar plots with enhancer activity ($\log_2$) of variants from (A,C,E).

Supplemental Fig S20. Systematic motif pasting screens in human enhancers.



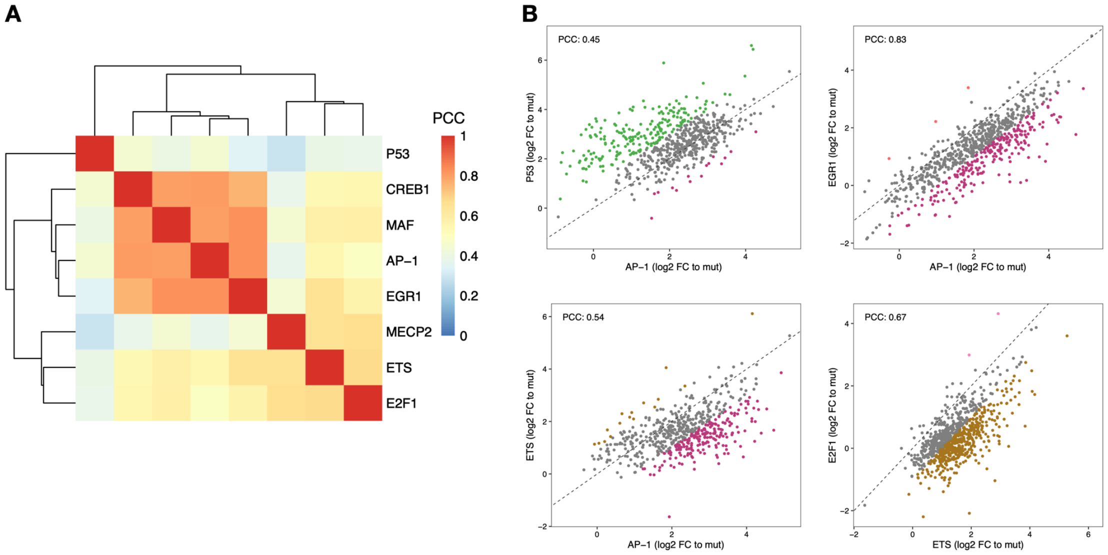
Pairwise comparisons of normalized STARR-seq input **(A)** and RNA **(B)** UMI read counts or enhancer activity (RNA/input) **(C)** between three independent biological replicates across all oligos tested. Color reflects point density. The PCC is denoted for each comparison.

Supplemental Fig S21. Enhancer activity of different sequences in human enhancers.



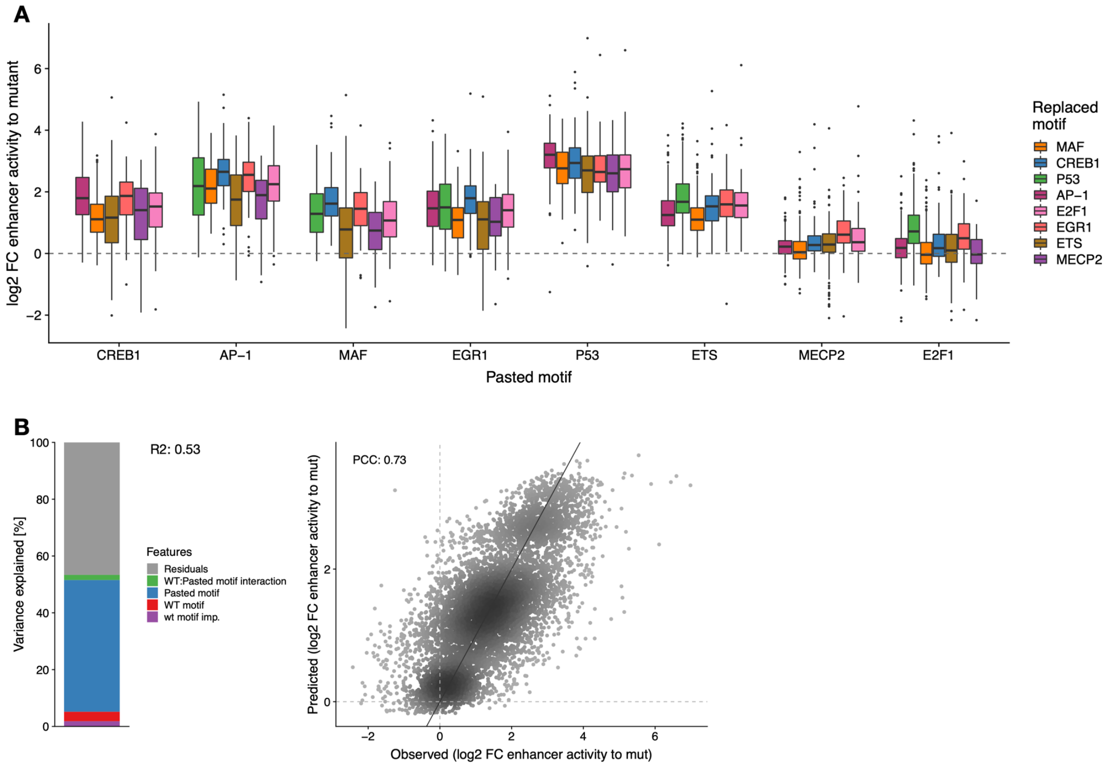
A) Activity of pasted motifs at different enhancer positions. Distribution of enhancer activity changes ($\log_2$) of all wild-type enhancers used and their variants with either mutant sequences or different TF motifs pasted. Few instances show negative values: these are not dependent on the specific mutant sequence but rather correspond to the creation of a repressor motif at the flanks of the pasted motif and the backbone enhancer. **B)** Activity of pasting motifs (y-axis, $\log_2$ fold-change activity over basal motif-mutated enhancer activity) in function of the basal activity (x-axis, activity of motif-mutated enhancer). The PCC is denoted for each motif.

Supplemental Fig S22. Human TF motifs work differently at different enhancer positions.



A) Hierarchical clustering of all TF motifs based on PCC of motif activities across all enhancer positions. **B)** Motifs work differently at different enhancer positions. Comparison between enhancer activity changes ($\log_2$ FC to mutated sequence) after pasting different TF motifs across all enhancer positions. Positions with stronger activity of each motif (>= 2-fold in respect to the other motif) are colored with the respective colors. PCC: Pearson correlation coefficient.

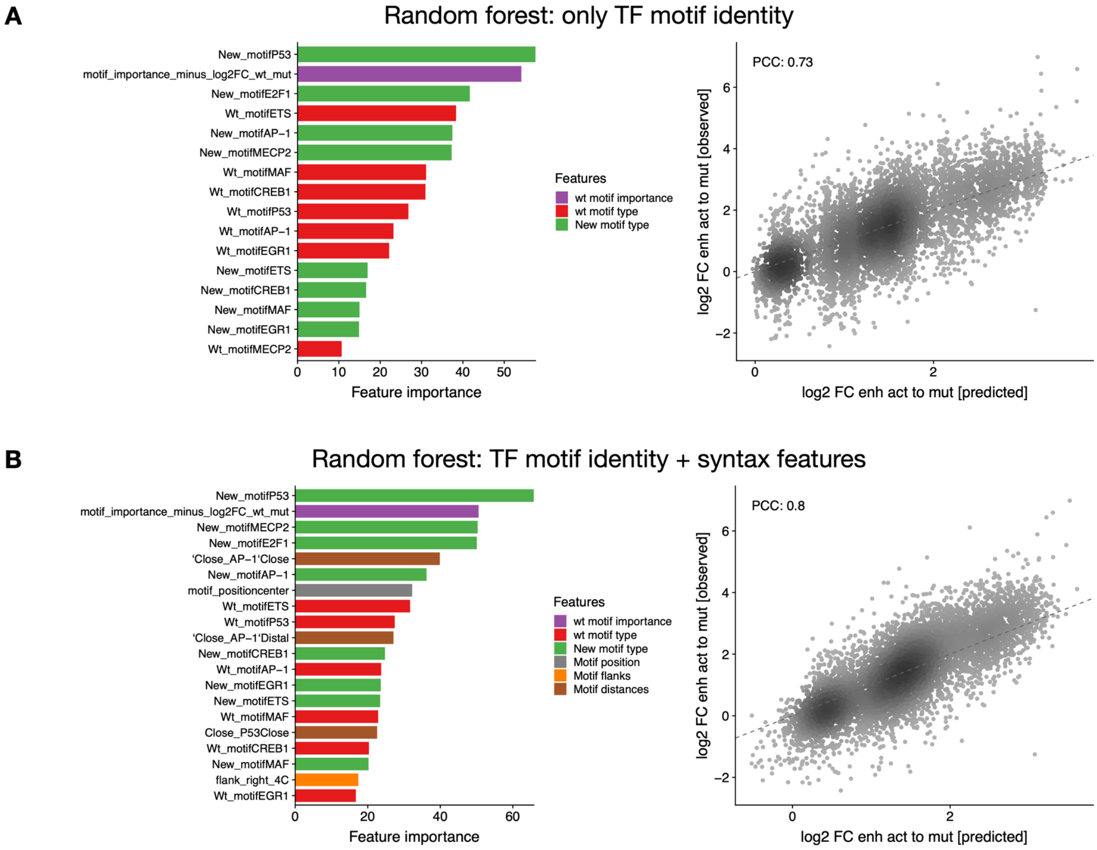
Supplemental Fig S23. TF motif activity in function of wild-type motif identity in human enhancers.



A) Distribution of enhancer activity changes ($\log_2$ FC to mutated sequence) across all enhancer positions for each pasted TF motif, grouped by the identity of the wild-type motif.

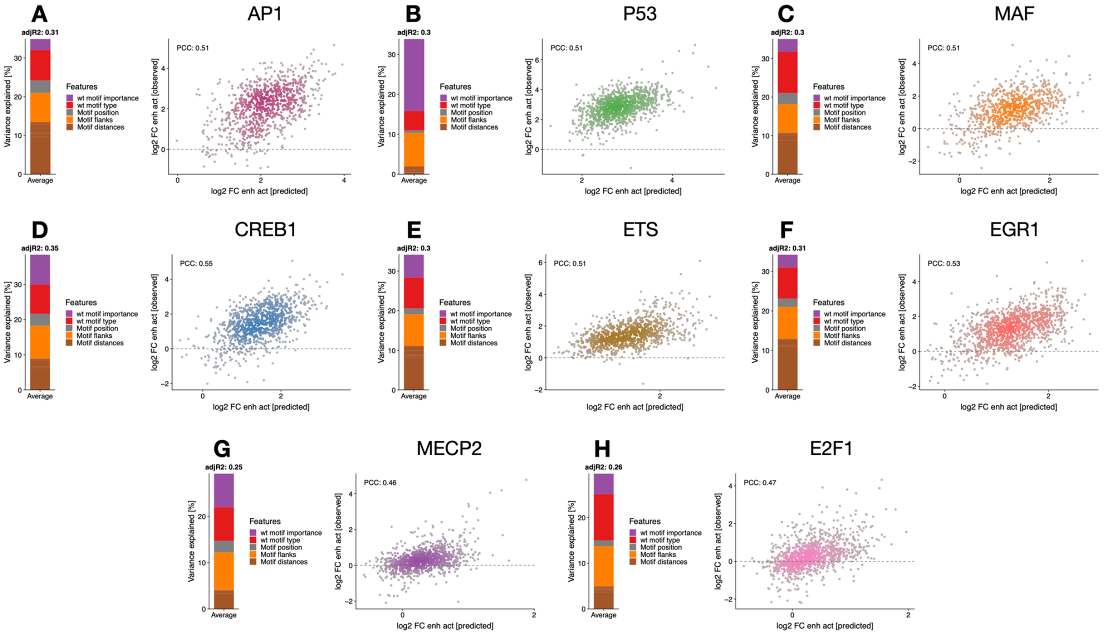
B) Left: Bar plot showing the amount of variance explained by the wild-type motif importance and identity, the pasted motif identity and the interaction between the wild type and pasted motifs, using a linear model fit on all motif pasting results. Right: Scatter plots of predicted (linear model) vs. observed enhancer activity changes ($\log_2$ FC to mutated sequence) across all motif pasting experiments. Color reflects point density. PCC is shown.

Supplemental Fig S24. Prediction of motif activities using motif syntax features in human enhancers.



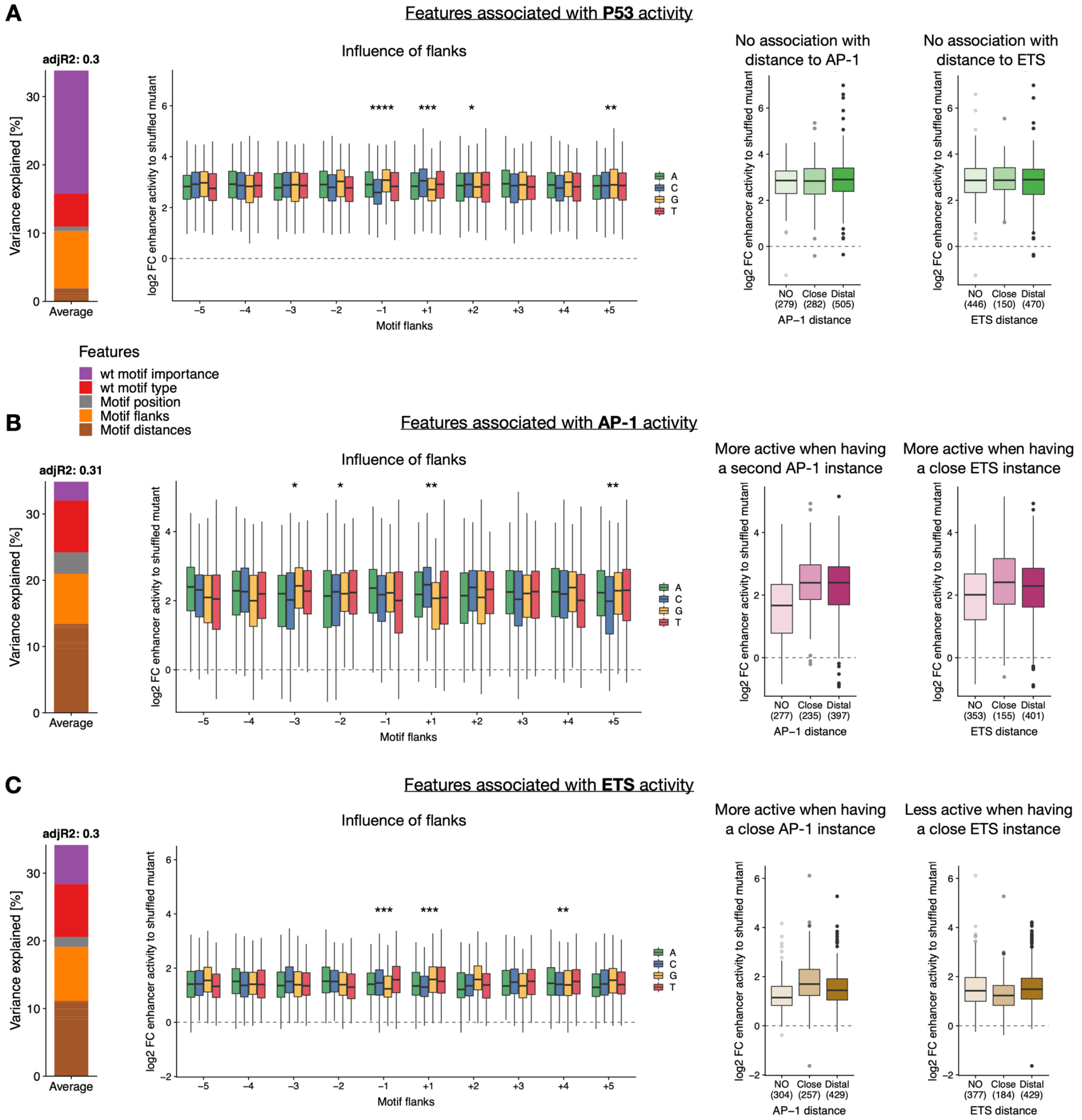
Left: Importance of all features **(A)** or only the top 20 **(B)** included in the random forest models with only TF motif identity **(A)** or also with syntax features **(B)**, sorted by importance and colored by feature type. Right: Scatter plots of predicted vs. observed enhancer activity changes ($\log_2$ FC to mutated sequence) across all motif pasting experiments. Color reflects point density. PCC is shown.

Supplemental Fig S25. Linear models with syntax features to predict motif activities in human enhancers.



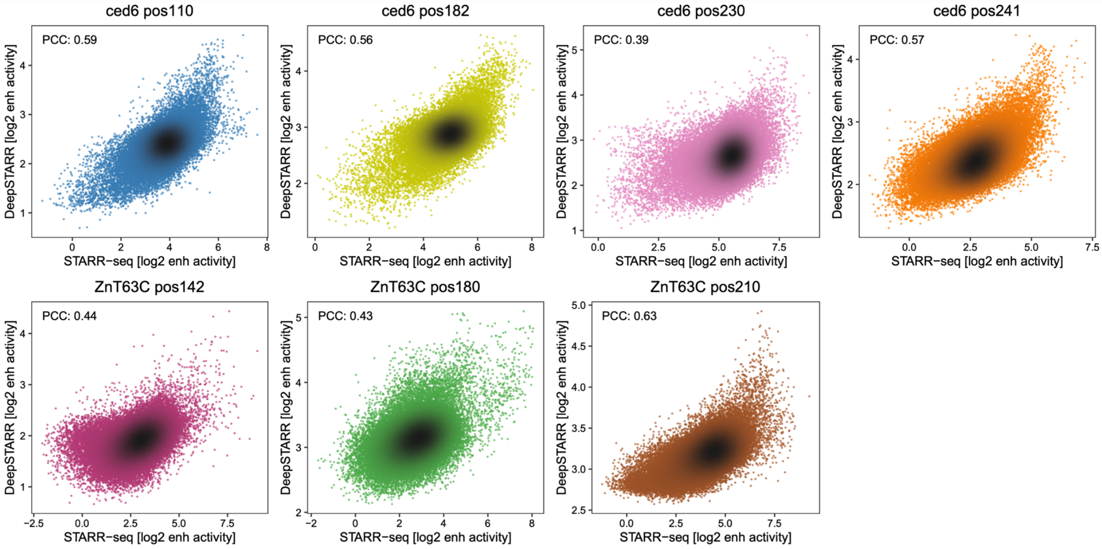
A-H) Left: Bar plot showing the variance explained by the different types of features (color legend) for each of the linear models. Right: Scatter plots of predicted vs. observed enhancer activity changes (log$_2$ FC to mutated sequence) for motif pasting experiments per TF motif type. Color reflects point density. PCC is shown.

Supplemental Fig S26. Sequence features associated with activity of P53, AP-1 and ETS motifs in human enhancers.



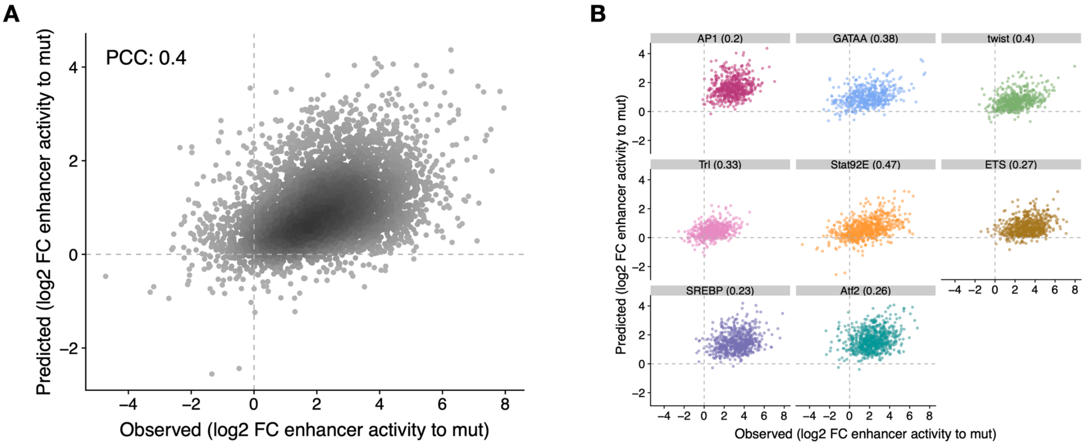
A-C) Left: Bar plot showing the variance explained by the different types of features (color legend) for each of the linear models. Middle-left: Motif activity according to the different bases at each flanking position, colored by nucleotide identity. Statistics from linear model in Fig 5E: ****P < 0.0001, ***P < 0.001, **P < 0.01, *P < 0.05 (linear regression p-value). Middle-right and right: Enhancer activity changes ($\log_2$ FC to mutated sequence) after pasting each TF motif in positions with no additional AP-1 (middle-right) or ETS (right) in the enhancer, or with additional AP-1 or ETS at close (<= 25 bp) or distal (>25 bp) distances. Number of instances are shown.

Supplemental Fig S27. DeepSTARR predicts enhancer sequence changes.



Comparison between DeepSTARR predicted (y-axis) and experimentally measured (x-axis) activity of random sequence variants tested at the different enhancer positions. Color reflects the enhancer position and point density. PCCs are shown.

Supplemental Fig S28. DeepSTARR predicts activity of motifs in different enhancer positions.



A) Comparison between DeepSTARR predicted (y-axis) and experimentally measured (x-axis) enhancer activity changes (log$_2$ FC to mutated sequence) for all motif pasting sequences. Color reflects the enhancer position and point density. PCCs are shown. **B)** Same as in (A) but per pasted TF motif.

Supplemental Tables

Supplemental Table S1. Primers used for UMI-STARR-seq library cloning.
Primers used for UMI-STARR-seq library cloning.

Supplemental Table S2. Random variants and oligo UMI-STARR-seq mapping statistics.
Summary of total sequenced reads, mapped reads and unique fragments (after collapsing by UMIs) for two random variants and three oligo UMI-STARR-seq screens in S2 cells, and three oligo UMI-STARR-seq screens in human HCT-116 cells.

Supplemental Table S3. Activity of random variants in seven enhancer positions.
8nt and 16nt forward and reverse sequences, activities and scaled activities in each of the seven enhancer positions.

Supplemental Table S4. Drosophila and human TF motif sequences used in the motif pasting experiments.
Drosophila and human TF motif sequences used in the motif pasting experiments.

Supplemental Table S5. Results of motif-pasting experiment in Drosophila S2 enhancers.
Table with all oligos used in the analysis of *Drosophila* motif pasting with their DNA sequence, wild-type motif information, pasted motif information, activity of respective enhancer variant, of the original wild type or motif-mutant enhancer, and respective $\log_2$ fold-changes.

Supplemental Table S6. Results of motif-pasting experiment in human HCT-116 enhancers.
Table with all oligos used in the analysis of human motif pasting with their DNA sequence, wild-type motif information, pasted motif information, activity of respective enhancer variant, of the original wild type or motif-mutant enhancer, and respective $\log_2$ fold-changes.

Supplemental Methods

UMI-STARR-seq

Cell culture and transfection

Drosophila Schneider 2 cells were grown in Schneider's *Drosophila* Medium (Gibco; 21720-024) supplemented with 10% heat inactivated FBS (Sigma-Aldrich; F7524) at 27°C. Human HCT116 cells were cultured in DMEM (Gibco; 52100-047) supplemented with 10% heat inactivated FBS (Sigma-Aldrich; F7524) and 2mM L-Glutamine (Sigma-Aldrich; G7513) at 37ºC in a 5% $C0_2$-enriched atmosphere. Both cell types were passaged every 2-3 days.

We used the MaxCyte-STX electroporation system for all library transfections. S2 cells were collected at 300 x g for 5min and washed once in 1:1 Schneider's Drosophila Medium and MaxCyte electroporation buffer (EPB-1). 50 x 10^6 cells were transfected with 5µg of DNA using the "Optimization 1" protocol, recovered for 30min at 27°C and resuspended in 10mL S2 Medium with 10% FBS. HCT116 cells were collected at 200 x g for 5min and washed once in MaxCyte electroporation buffer (EPB-1). Cells were electroporated at a density of 1 x 10^7 cells per 100µL and 20µg of DNA using the preset "HCT116" program, recovered for 20min at 37 °C and resuspended in 10mL DMEM 10% FBS and 2mM L-Glutamine.

Each replicate for a STARR-seq screen was transfected in 2 OC400 cuvettes with a total of 400 x 10^6.

UMI-STARR-seq experiments

Library cloning

Random 8nt variant libraries were generated using a PCR approach with degenerate oligonucleotides. Forward primers (primers see Supplemental Table S1) were designed to anneal directly downstream of the enhancer position of interested followed by 8 degenerate bp (creating 65,536 variants) and another 20 bp complementary stretch. Reverse primers were complementary to the 20 bp 5' of the degenerate stretch. The STARR-seq vector containing the wild-type enhancer of interest (either *ced-6* or *ZnT63C*) was used as a template for the PCR. The PCR was run across the whole STARR-seq plasmid, followed by DpnI digest and a Gibson reaction that re-circularizes the plasmid. Libraries were grown in 2l LB-Amp (final ampicillin concentration 100µg/mL). Variant libraries of the same enhancer i.e. *ced-6* enhancer pos110, pos182, pos230, pos241 and

ZnT63C enhancer pos142, pos180, pos210 were pooled to equimolar ratio, together with another synthetic oligo library containing wt enhancer sequences and negative regions. *Drosophila* and human oligo libraries were synthesized by Twist Bioscience including the 249 bp enhancer sequence and adaptors for library cloning. *Drosophila* library fragments were amplified (primers see Supplemental Table S1) and cloned into *Drosophila* STARR-seq vectors containing the DSCP core-promoters using Gibson cloning (New England BioLabs; E2611S). The oligo library for human STARR-seq screens was amplified (primers see Supplemental Table S1) and cloned into the human STARR-seq plasmid with the ORI in place of the core promoter (Muerdter et al. 2018). Libraries were grown in 2l LB-Amp (final ampicillin concentration 100µg/mL).

All libraries were purified with Qiagen Plasmid *Plus* Giga Kit (cat. no. 12991).

Drosophila S2 cells

UMI-STARR-seq was performed as described previously (Arnold et al. 2013; Neumayr et al. 2019). In brief, we transfected 400 × 10^6 S2 cells total per replicate with 20 µg of the input library (see libraries above). After 24 hr incubation, poly(A) RNA was isolated and processed as described before (Neumayr et al. 2019). Briefly: after reverse transcription and second strand synthesis a unique molecular identifier (UMI) was added to each transcript, allowing the counting of individual RNA molecules. This is followed by two nested PCR steps, each with primers that are specific to the reporter transcripts such that STARR-seq does not detect endogenous cellular RNAs.

Human HCT116 cells

UMI-STARR-seq was performed as described previously (Arnold et al. 2013; Muerdter et al. 2018; Neumayr et al. 2019). Screening libraries were generated from synthesized oligo pools by Twist Bioscience (see above). We transfected 80 × 10^6 HCT116 cells total per replicate with 160 µg of the input library. After 6 hr incubation, poly(A) RNA was isolated and further processed as described before (Neumayr et al. 2019).

Illumina sequencing

High-throughput sequencing was performed at the VBCF NGS facility on an Illumina NextSeq 550 or NovaSeq SP platform, following manufacturer's protocol. Random variants UMI-STARR-seq and Twist-oligo library screens were sequenced as paired-end 150 cycle runs, using standard Illumina i5 indexes as well as unique molecular identifiers (UMIs) at the i7 index. Deep sequencing base-calling was performed with CASAVA (v.1.9.1).

Random variants UMI-STARR-seq data analysis

Dedicated Bowtie indices were created for each enhancer position's N_8 library and combined with an oligo library of thousands of wild-type enhancers and negative sequences (de Almeida et al. 2022) for normalization, all 249 bp-long sequences. UMI-STARR-seq RNA and DNA input reads (paired-end 150 bp) were mapped to these dedicated Bowtie indices using Bowtie v.1.2.2 (Langmead et al. 2009). Since the N_8 variants were all positioned in the last 150 nt of each enhancer, we allowed for flexible mapping in the beginning of the fragments to increase the number of mapped reads while keeping high sensitivity for the different enhancer variants. Specifically, we trimmed the forward reads to 36 bp and mapped them to the indices allowing for 3 mismatches; the full 150 bp-long reverse reads were mapped with no mismatches, to identify all sequence variants; paired-end reads with the correct position, length and strand were kept. This mapping strategy was used for both DNA and RNA reads. For paired-end DNA and RNA reads that mapped to the same variant, we collapsed those that have identical UMIs (10 bp, allowing one mismatch) to ensure the counting of unique molecules (Supplemental Table S2).

We excluded oligos with less than 5 reads in any of the input replicates and less than 1 read in any of the RNA replicates. The enhancer activity of each sequence in each screen was calculated as the $\log_2$ fold-change over input, using all replicates, with DESeq2 (Love et al. 2014). We used the counts of wild-type negative regions in each library as scaling factors between samples.

Oligo library UMI-STARR-seq data analysis

As described previously (de Almeida et al. 2022), oligo library UMI-STARR-seq RNA and DNA input reads (paired-end 150 bp) were mapped to a reference containing the 249 bp-long sequences from the fragments present in the *Drosophila* (dm3) or human (hg19) libraries using Bowtie v.1.2.2 (Langmead et al. 2009). We used these reference genomes to be able to integrate our results with older in-house and published datasets and made sure this choice does not affect the quantifications of enhancer activity. For each library we demultiplexed reads by the i5 and i7 indexes and oligo identity. Mapping reads with the correct length, strand and with no mismatches (to identify all sequence variants) were kept. Both DNA and RNA reads were collapsed by UMIs (10 bp) as above (Supplemental Table S2).

We excluded oligos with less than 10 reads in any of the input replicates and added one read pseudocount to oligos with zero RNA counts. The enhancer activity of each oligo in each screen was calculated as the $\log_2$ fold-change over input, using all replicates, with DESeq2 (Love et al. 2014). We used the counts of wild-type negative regions in each library as scaling factors between samples.

Analyses of random variants at different enhancer positions

Independent motif mutations

Two strong S2 developmental enhancers with different TF motif compositions were selected to test a diversity of random 8 nt variants in different positions: *ced-6* (chr2R:5326628-5326876) and *ZnT63C* (chr3L:3310914-3311162) enhancers. Experimental mutations of GATA, AP-1 and twist motifs in these enhancers were performed in a previous study (Supplemental Fig S4F; (de Almeida et al. 2022)) and used here to select important enhancer positions.

Enhancer random variants libraries and UMI-STARR-seq

We selected five positions important for the activity of the two enhancers (*ced-6* pos110 and pos241; *ZnT63C* pos142, pos180, pos210) and two non-important positions of the *ced-6* enhancer (pos182 and pos230). At each position, we experimentally replaced the respective 8nt stretch of the enhancer with randomized nucleotides (N_8), creating 65,535 enhancer variants in addition to the wild-type sequence per position. For each enhancer, we pooled the libraries of the different positions and combined them with an oligo library of thousands of wild-type enhancers and negative sequences (de Almeida et al. 2022) for normalization. UMI-STARR-seq using the *ced-6* or *ZnT63C* pooled libraries was performed ("UMI-STARR-seq experiments") and analyzed ("Random variants UMI-STARR-seq data analysis") as described above (Supplemental Table S3). We performed two independent replicates per enhancer pooled library screen (Pearson correlation coefficient (PCC)=0.85-0.91; Supplemental Fig S4A-E).

To be able to compare the activity of variants and motifs between enhancer positions, we next scaled the enhancer activity of all variants per position (*z*-scores). This allows to measure the change in activity of a given variant over the average of all variants, correcting for the importance of the different enhancer positions tested.

Comparison between pooled libraries using common oligos

The respective wild-type enhancer sequence was overrepresented in each N_8 library input since it was used as the template for the PCR cloning (Supplemental Fig S4A,B). We compared the activities of the *ced-6* and *ZnT63C* enhancer sequences and all other wild-type enhancers and negative sequences present in both *ced-6* and *ZnT63C* pooled libraries (Supplemental Fig S4C). The activities of the common sequences were similar between both screens, except for the *ZnT63C* enhancer whose activity was underestimated in the *ZnT63C* pooled library, likely due to the technical overrepresentation in the input. We therefore selected another enhancer with the same activity as the *ZnT63C* enhancer (chrX:9273894-9274142) to be used as the reference wild-type activity for the *ZnT63C* enhancer variants (Supplemental Fig S4C, 2B).

Diversity of top active variants and *de novo* motif discovery

The most-active 8nt variants of each screen (1, 2, 5, 10, 50, 100 and 1,000) were retrieved and consolidated into position probability matrices based on the nucleotide frequencies at each position (Fig 1C, S5B). Logos were visualized using the *ggseqlogo* function from R package *ggseqlogo* (v.0.1; (Omar Wagih 2017)). The same was done after randomly sorting the variants of each screen for comparison. The information content of the top sequences at each position was calculated as described in https://bioconductor.org/packages/release/bioc/vignettes/universalmotif/inst/doc/IntroductionToSequenceMotifs.pdf (Schneider and Stephens 1990; Schneider et al. 1986) (Fig 1D, S5C).

The top 100 and 1,000 or bottom 1,000 variants (8nt +/- 4nt flanks) of each screen were used for *de novo* motif discovery analyses using HOMER, taking all detected variants of the respective screen as background (Supplemental Fig S2, S7). HOMER (v4.10.4; (Heinz et al. 2010)) was run with the findMotifs.pl command and the arguments *fly -len 6,7,8*.

Activity of TF motifs created by sequence variants

To robustly assess the activity of a given TF motif, we retrieved the activity of all 16nt variants (8nt +/- 4nt flanks) creating each motif by string-matching. The main motifs used were: GATA – GATAAG, AP-1 – TGA.TCA, SREBP – TCACGCGA, twist – CATCTG, CREB/ATF – TCATCA, STAT – TTCC.GGA, Trl – GAGAGA, ETS – CCGGAA, Dref – ATCGAT, ttk – AGGATAA, ZEB1 – CAGGTG, lola – GGAGTT (format: TF motif – string). For a more systematic comparison across all TF motif types, we matched variants to the optimal string from each TF motif PWM model in a motif database (Supplemental Fig S9A; (de Almeida et al. 2022)). The average activity across variants was defined as the motifs' intrinsic strength. These activities were used in Fig 1E, 2E,D, Supplemental Fig S3A, S6A, S9, S10.

To find how many active variants are explained by the creation of known motifs enriched in S2 developmental enhancers (from (de Almeida et al. 2022)), we performed PWM-based motif scanning of those candidate motifs onto variants (8nt +/- 4 flanks) (Fig 1F, Supplemental Fig S3B, S6B). We used the *matchMotifs* function from R package *motifmatchr* (v.1.4.0; genome = "BSgenome.Dmelanogaster.UCSC.dm3", bg="genome" (Schep 2021)) with p-value cutoffs 1e^{-04} and 1e^{-05}.

Activity of variants in function of their similarity to the wild-type sequence

The similarity of each sequence variant to the wild-type sequence at each enhancer position was measured using the *stringdist* R package and *hamming* distance method (Supplemental Fig S6A).

Comparison of random variants activity across enhancer positions

We compared the activity of all 8nt random variants across enhancer positions using their *z*-score scaled activity (Fig 2C, Supplemental Fig S8; Supplemental Table S3). We calculated pairwise PCCs between the different libraries, performed hierarchical clustering ("complete" method) using the correlation values as similarities, and displayed heatmaps using the *pheatmap* R package (v.1.0.12; (Kolde 2019)). To reduce the impact of the flanking sequence of each position when comparing the activity of variants between them, we repeated the same after consolidating the 8nt into shorter variants by taking the centered sequence and averaging the activity across variants with different flanking nucleotides.

Analyses of motif pasting screens in *Drosophila* and human enhancers

Oligo library design

Drosophila motif pasting library

We selected 1,172 motif positions (among 728 enhancers) that are required for the activity of the respective enhancers, assessed by experimental mutagenesis in a previous study (de Almeida et al. 2022). These wild-type positions cover different contexts and TF motifs: GATA, AP-1, twist, Trl, ETS and SREBP. We next designed sequences of enhancer variants where we pasted a mutant sequence or the optimal sequence of eight TF motifs (GATA, AP-1, twist, Trl, ETS, SREBP, Stat92E and Atf2; one at a time; sequences in Supplemental Table S4) in each of these positions (Fig 3A). To reduce the influence of flanking nucleotides and different motif affinities and focus on differences due to the enhancer context we pasted an extended optimal sequence of each TF motif (as in de Almeida et al. (de Almeida et al. 2022)). This library (Supplemental Table S5) was synthetized and pooled with a previous library containing the wild-type enhancer sequences (de Almeida et al. 2022) to be screened together.

Human motif pasting library

Similar to the *Drosophila* library, we selected 1,456 motif positions important for the activity of 808 enhancers, assessed by experimental mutagenesis in a previous study (de Almeida et al. 2022). These wild-type positions cover different contexts and TF motifs: AP-1, ETS, E2F1, EGR1, MAF, MECP2, CREB1, P53. We next designed sequences of enhancer variants where we pasted a mutant sequence or the optimal sequence of the same eight TF motifs (AP-1, ETS, E2F1, EGR1, MAF, MECP2, CREB1, P53; one at a time; sequences in Supplemental Table S4) in each of these positions. As for the *Drosophila* motifs, we pasted an extended optimal sequence of each TF motif to reduce the influence of flanking nucleotides and different motif affinities and focus on differences due to the enhancer context. This library (Supplemental Table S6) was synthetized and pooled with a previous library containing the wild-type enhancer sequences (de Almeida et al. 2022) to be screened together.

Oligo library synthesis and UMI-STARR-seq

The *Drosophila* and human enhancers' oligo libraries contained each sequences for the wild-type enhancers and enhancers with mutant variants or motifs pasted at the selected positions (Supplemental Table S5 and S6, respectively). All sequences were designed using the dm3 and hg19 genome versions, respectively. The enhancer sequences spanned 249 bp total, flanked by the Illumina i5 (25 bp; 5′ -TCCCTACACGACGCTCTTCCGATCT) and i7 (26 bp; 5′ AGATCGGAAGAGCACACGTCTGAACT) adaptor sequences upstream and downstream, respectively, serving as constant linkers for amplification and cloning. The resulting 300-mer oligonucleotide *Drosophila* and human libraries were synthesized by Twist Bioscience. UMI-STARR-seq using these oligo libraries was performed ("UMI-STARR-seq experiments") and analyzed ("Oligo library UMI-STARR-seq data analysis") as described above (Supplemental Table S5 and S6). We performed three independent replicates for *Drosophila* (correlation PCC=0.95-0.98; Supplemental Fig S11A,B) and human (PCC=0.96-0.98; Supplemental Fig S20A,B) screens.

Quantification of motif activity at different enhancer positions

We used our enhancer activity measures of the wild-type and mutated sequences to stringently select important enhancer positions for further analyses: positions where mutation reduced the activity by at least 2-fold (Supplemental Fig S12A, S21A). These resulted in 763 important positions distributed among 496 *Drosophila* enhancers and 1,354 positions distributed among 753 human enhancers. This was important to select positions where we could reliably measure the increase in enhancer activity after pasting each TF motif – quantified as the $\log_2$ fold-change activity over the mutated enhancer (Fig 3B, 5A). Variability of activity of each motif across enhancer positions was quantified using the coefficient of variation (ratio of the standard deviation to the mean; Supplemental Fig S12B).

We compared the activity of motifs across enhancer positions by pairwise PCCs and performed hierarchical clustering ("complete" method) using the correlation values as similarities. Heatmaps were displayed using the *pheatmap* R package (v.1.0.12; (Kolde 2019)) (Fig 3D, 5B, Supplemental Fig S13A, S22A).

Importance of the wild-type motif

We fitted motif activity values ($\log_2$ fold-change enhancer activity after motif pasting) with linear models using the wild-type TF motif identity and importance ($\log_2$ fold-change activity between wild-type and motif-mutant sequence), the pasted motif identity, and the interaction between the wild-type and pasted motifs as covariates, using the *lm* function (v.3.5.1; (R Core Team 2020)). Variance explained by each covariate was calculated with one-way ANOVAs of the respective models (Fig 5D, Supplemental Fig S14B, S23B).

Difference between pairs of positions in the same or different enhancers

Drosophila enhancers with two positions tested in our assay were selected and the fold-change in motif activity between pairs of positions in the same enhancer was compared with the fold-change between pairs of positions in different enhancers (matched by similar position-mutant baseline activities). For each pasted TF motif, significant differences were assessed through a two-sided Wilcoxon signed rank test followed by FDR multiple testing correction (Supplemental Fig S15).

Prediction of motif activities using motif syntax features

Motif syntax features

To test how motif activities depend on motif syntax features we extracted the following features per tested enhancer position: the position relative to the enhancer center (center: -/+ 25 bp, flanks: -/+25:75 bp, boundaries: -/+75:125 bp), the position flanking nucleotides (5 bp on each side), and the presence and distance to other TF motifs (close: <= 25 bp; distal: >25 bp; between motif centers).

Instances of each TF motif type were mapped across all enhancers using their annotated PWM models (Supplemental Table S3) and the *matchMotifs* function from R package *motifmatchr* (v.1.4.0; (Schep 2021)) with the following parameters: genome = “BSgenome.Dmelanogaster.UCSC.dm3”, p.cutoff = 5e-04, bg="genome". Overlapping instances (minimum 50%) for the same TF motif were collapsed and counted only once.

Random forest models

We used a 10-fold cross-validation scheme to train random forest models to predict *Drosophila* or human motif pasting activities ($\log_2$ fold-change to mutant) using as features the wild-type TF motif identity and importance ($\log_2$ fold-change activity between wild-type and motif-mutant sequence) and the pasted motif identity, together or not with additional syntax features (described above). All models were built using the *Caret* R package (v. 6.0-80; (Kuhn 2018)) and feature importance was calculated using its *varImp* function. Predictions for each held-out test sets were used to compare with the observed motif activities and assess model performance (Supplemental Fig S16, S24).

Linear model with motif syntax rules to predict motif activities

For each TF motif type, we built a multiple linear regression model to predict its activity ($\log_2$ fold-change to mutant) across different enhancer positions using as covariates the wild-type TF motif identity and importance ($\log_2$ fold-change activity between wild-type and motif-mutant sequence) together with additional syntax features (described above). All models were built using the Caret R package (v. 6.0-80; (Kuhn 2018)) and 10-fold cross-validation. Predictions for each held-out test sets were used to compare with the observed $\log_2$ fold- changes and assess model performance (Supplemental Fig S17, S25). The linear model coefficients and respective FDR-corrected p-values were used as metrics of importance for each feature, using the red or blue scale depending on positive or negative associations (Fig 4A, 5E). For flanking positions, we used always red because the direction of the association is not relevant. In addition, we calculated the percentage of variance explained by each covariate in the linear models built for each TF motif with one-way ANOVAs. For each TF motif, we generated 100 different models, randomizing the order of the covariates (since the variance explained depends on the order of covariates entered), quantified the percentage of variance explained of each covariate as its sum of squares divided by the total sum of squares, and used the average value across all 100 models as the final variance explained per covariate (Supplemental Fig S17, S25).

DeepSTARR predictions

Nucleotide contribution scores

Nucleotide contribution scores for wild-type enhancers or enhancer variants (Fig 2A, S5A, S11C,D, S16) were calculated as described previously (de Almeida et al. 2022), using DeepExplainer (the DeepSHAP implementation of DeepLIFT, see refs. (Shrikumar et al. 2017; Lundberg and Lee 2017; Lundberg et al. 2020); update from https://github.com/AvantiShri/shap/blob/master/shap/explainers/deep/deep_tf.py) and visualized using the *ggseqlogo* function from R package *ggseqlogo* (v.0.1; (Omar Wagih 2017)).

DeepSTARR predictions of enhancer sequence changes

DeepSTARR (https://github.com/bernardo-de-almeida/DeepSTARR, (de Almeida et al. 2022)) was used to predict the enhancer activity of N_8 variants in enhancers (Supplemental Fig S27) or the $\log_2$ fold-change enhancer activity of motif pasting sequences (Supplemental Fig S28).

Statistics and data visualization

All statistical calculations and graphical displays have been performed in R statistical computing environment (v.3.5.1; (R Core Team 2020)) and using the R package *ggplot2* (Wickham 2016). In all box plots, the central line denotes the median, the box encompasses 25th to 75th percentile (interquartile range) and the whiskers extend to 1.5× interquartile range.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE211659 or Zenodo at https://doi.org/10.5281/zenodo.7010528. Code used to process the UMI-STARR-seq data as well as to reproduce all analyses, results and figures has been submitted to GitHub (https://github.com/bernardo-de-almeida/Variant_STARRseq) and is available as Supplemental Code.

References

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (1979)* **339**: 1074–1077.

de Almeida BP, Reiter F, Pagani M, Stark A. 2022. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet*.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38**: 576–589.

Kolde R. 2019. pheatmap: Pretty Heatmaps. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap.

Kuhn M. 2018. caret: Classification and Regression Training. R package version 6.0-80. https://CRAN.R-project.org/package=caret.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 1–21.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**: 56–67.

Lundberg SM, Lee S-I. 2017. A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems*.

Muerdter F, Boryn ŁM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, Pagani M, Haberle V, Kazmar T, Catarino RR, et al. 2018. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat Methods* **15**: 141–149.

Neumayr C, Pagani M, Stark A, Arnold CD. 2019. STARR-seq and UMI-STARR-seq: Assessing Enhancer Activities for Genome-Wide-, High-, and Low-Complexity Candidate Libraries. *Curr Protoc Mol Biol* **128**: e105.

Omar Wagih. 2017. ggseqlogo: A “ggplot2” Extension for Drawing Publication-Ready Sequence Logos. R package version 0.1. https://CRAN.R-project.org/package=ggseqlogo.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schep A. 2021. motifmatchr: Fast Motif Matching in R. R package version 1.14.0.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**: 6097.

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. 1986. Information content of binding sites on nucleotide sequences. *J Mol Biol* **188**: 415–431.

Shrikumar A, Greenside P, Kundaje A. 2017. Learning important features through propagating activation differences. *ArXiv* **1704.02685**.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, http://ggplot2.org.*