Supplemental Materials

Supplemental Figures S1-S7 & Legends
Supplemental Tables 1-6
Supplemental Files:
Supplemental File 1: SyntheticLibraryDesign.txt
Supplemental File 2: GenomicLibraryDesign.txt
Supplemental File 3: SYN_ExpressionSummary.txt
Supplemental File 4: GEN_ExpressionSummary.txt
Supplemental File 5: gkmSVM_8merScoredWeights.txt
Supplemental File 6: gkmSVM_8merTop50TOMTOMe27.txt
Supplemental File 7: SYN_FeaturesiRF.txt
Supplemental File 8: gWT_FeaturesiRF.txt
Supplemental File 9: SYN_DemultiplexedBCcounts.zip
Supplemental File 10: GEN_DemultiplexedBCcounts.zip

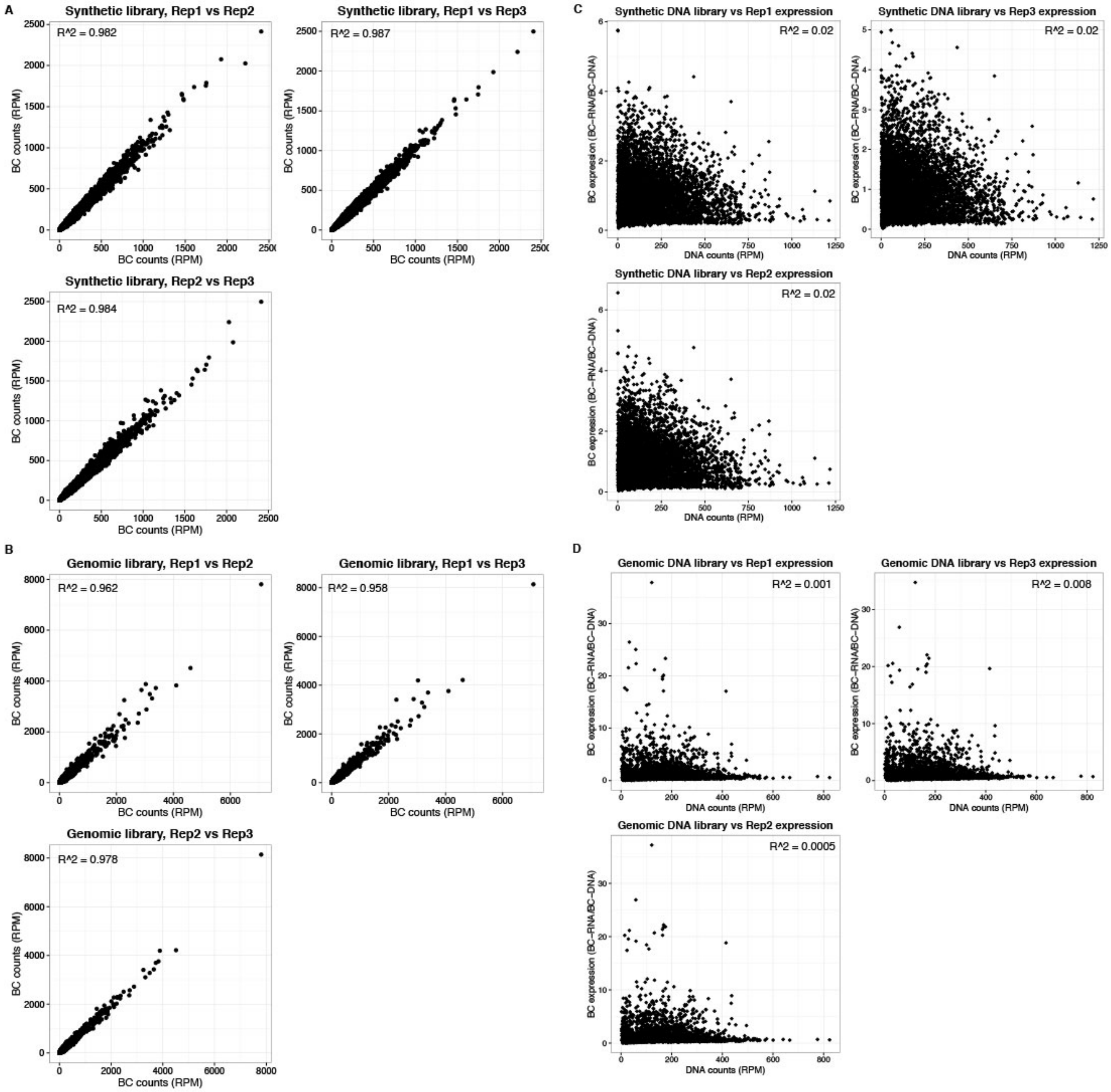

Figure S1. MPRA data quality. Reproducibility of barcode (BC) counts between biological replicates, normalized as reads per million per RNA replicate for (**A**) Synthetic library and (**B**) Genomic, gWT and gMUT, library. Comparison of normalized BC expression (BC_{RNA}/BC_{DNA}) versus DNA counts for (**C**) Synthetic library and (**D**) Genomic, gWT and gMUT, library.

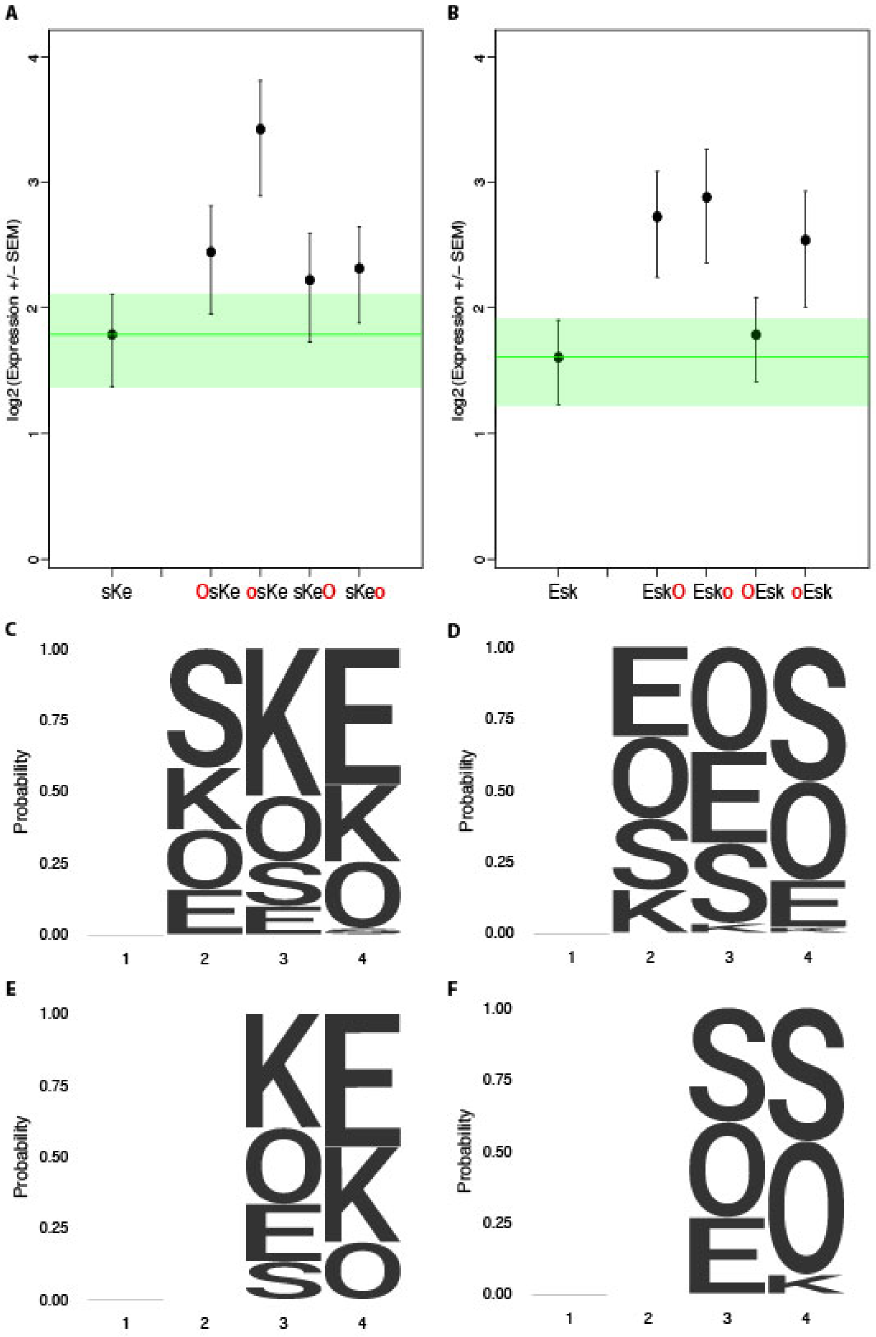

Figure S2. Additional examples of non-additivity in synthetic elements. Comparisons of synthetic 3-mer elements with matched 4-mer elements containing one additional site in the first or fourth position with (**A**) three of four matched 4-mers with overlapping expression despite an additional binding site and (**B**) one of four matched 4-mers with overlapping expression. Activity logos for the top 25% (**C**), bottom 25% (**D**) of 3-mer synthetic elements (n= 48 each), and top 25% (**E**), bottom 25% of 2-mer synthetic elements (n= 12 each). Height of letter is proportional to frequency of site in indicated position. Positions organized as in **Figure 2**.

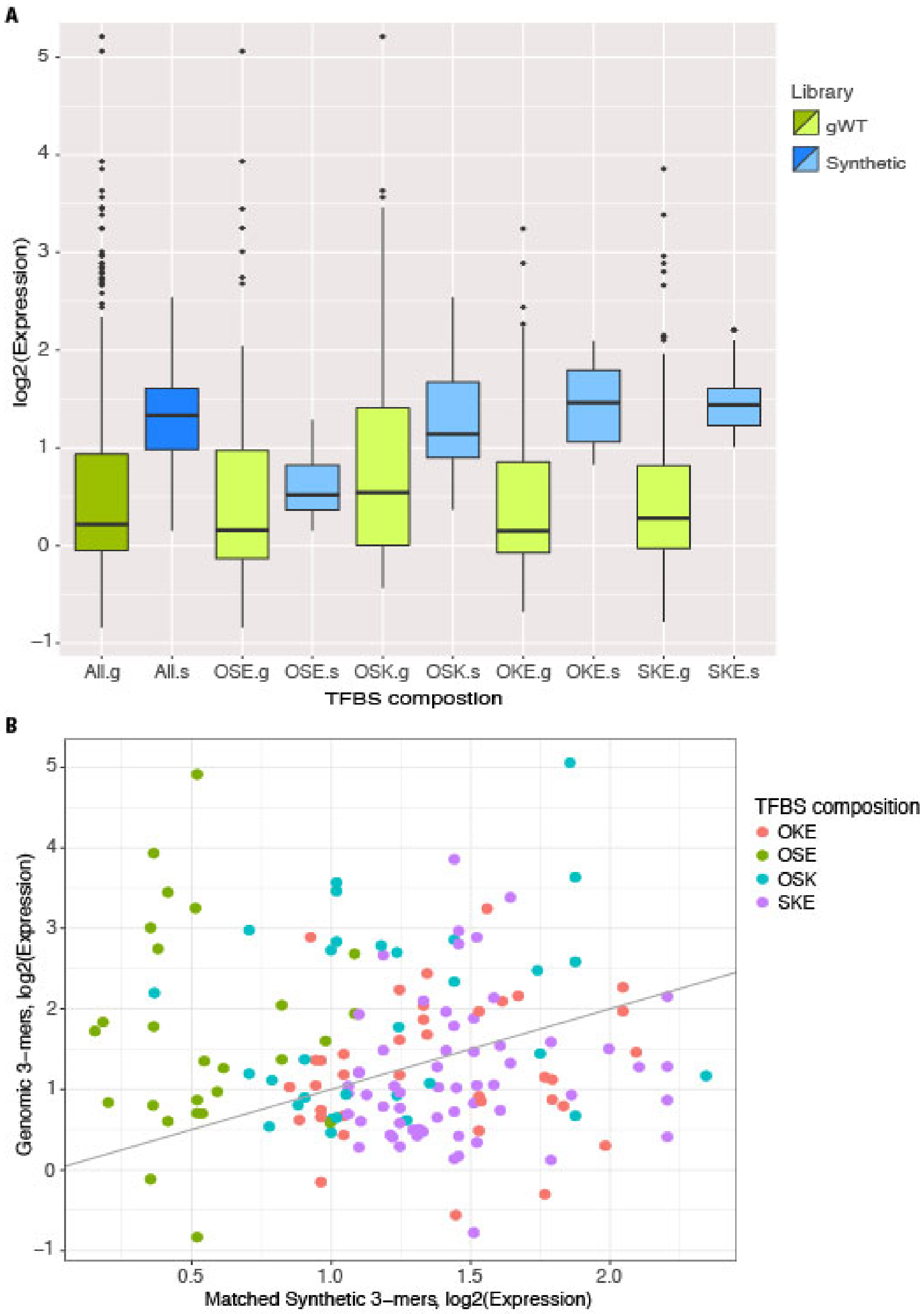

Figure S3. Comparison of synthetic and genomic patterns of transcription factor binding sites (TFBSs). (**A**) Expression (log$_2$) of all synthetic (dark blue) and gWT (dark green) library members subset by TFBS composition (light blue and light green, respectively). (**B**) Expression (log$_2$) of synthetic (x-axis) and gWT (y-axis) library members, matched by composition and order of binding sites for OCT4 (O), SOX2 (S), KLF4 (K), and ESRRB (E). Subsets of TFBS composition indicated by color. Grey line indicates x-y diagonal as axis scales differ.

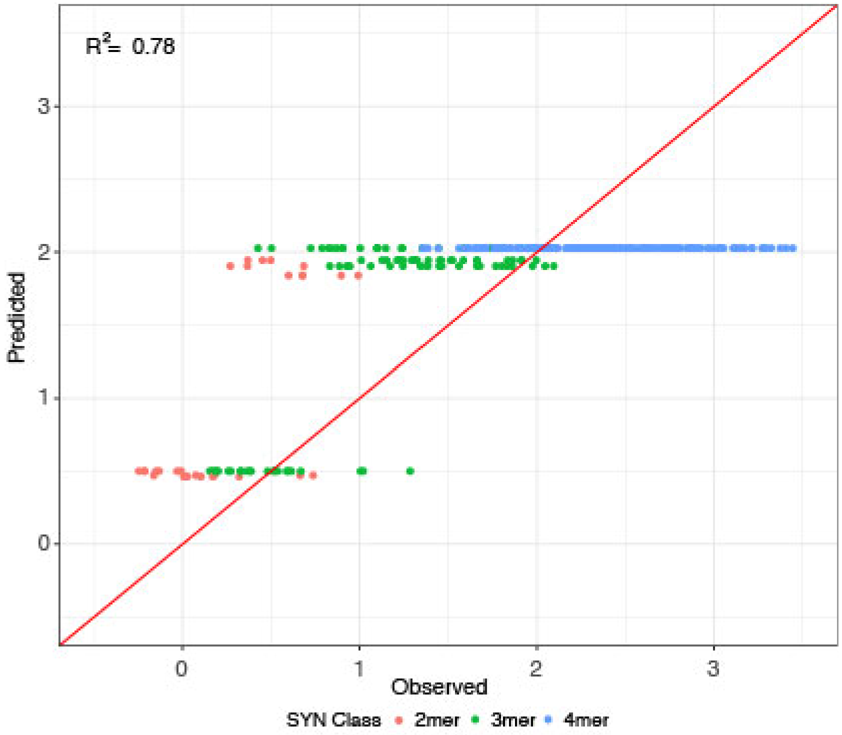

Figure S4. Additive effects in synthetic elements. Iterative random forest (iRF) regression model that includes features for only presence of pluripotency TFBSs to predict the relative expression of synthetic elements. Number of binding site per element indicated as in **Figure 3**. Observed and predicted expression are both plotted in $\log_2$ space.

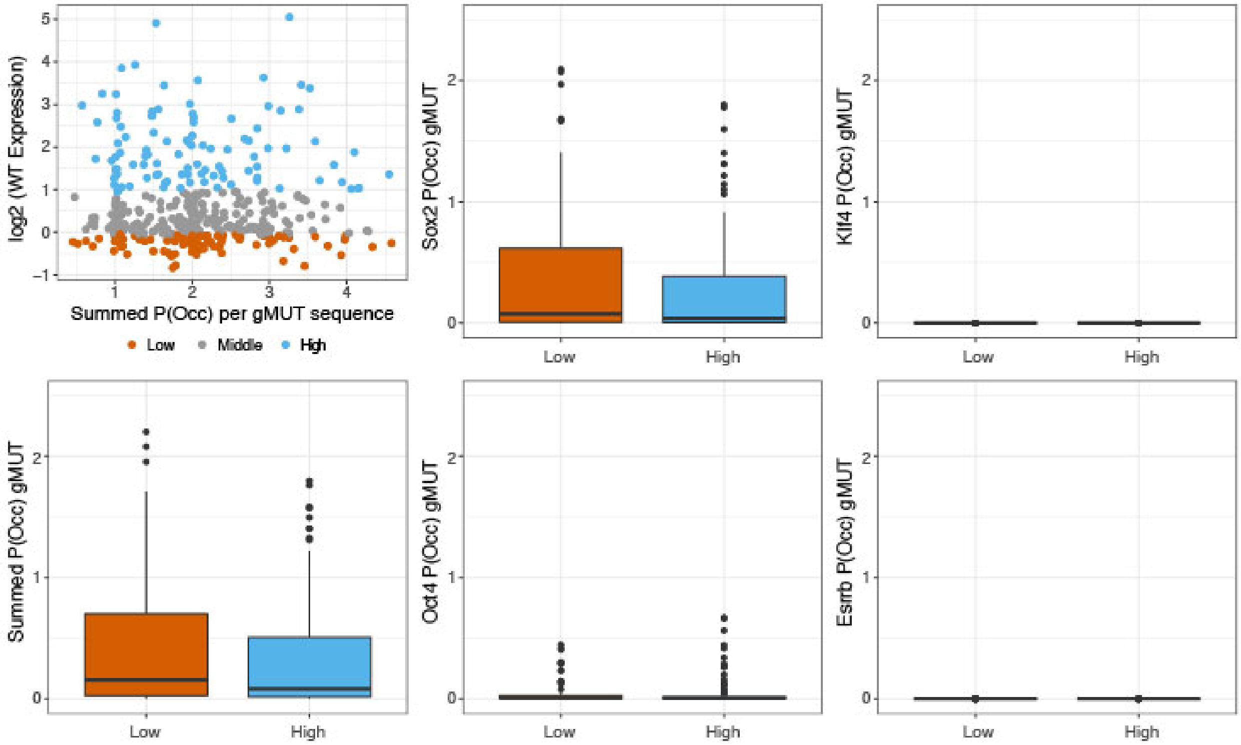

Figure S5. Predicted occupancy of genomic sequences. Predicted occupancy (P(Occ)) for genomic sequences in the absence of the primary pluripotency sites (gMUT sequences) for high assumed protein concentration (mu) for SOX2 (mu = 8), OCT4 (mu =10), KLF4 (mu = 8), and ESRRB (mu = 8) shown in middle and right panels. Summed P(Occ) of all factors per gMUT sequence, compared to expression (top left panel) or binned as low or high library members (bottom 25% and top 25% of sequences, ranked by gWT expression, n = 101).

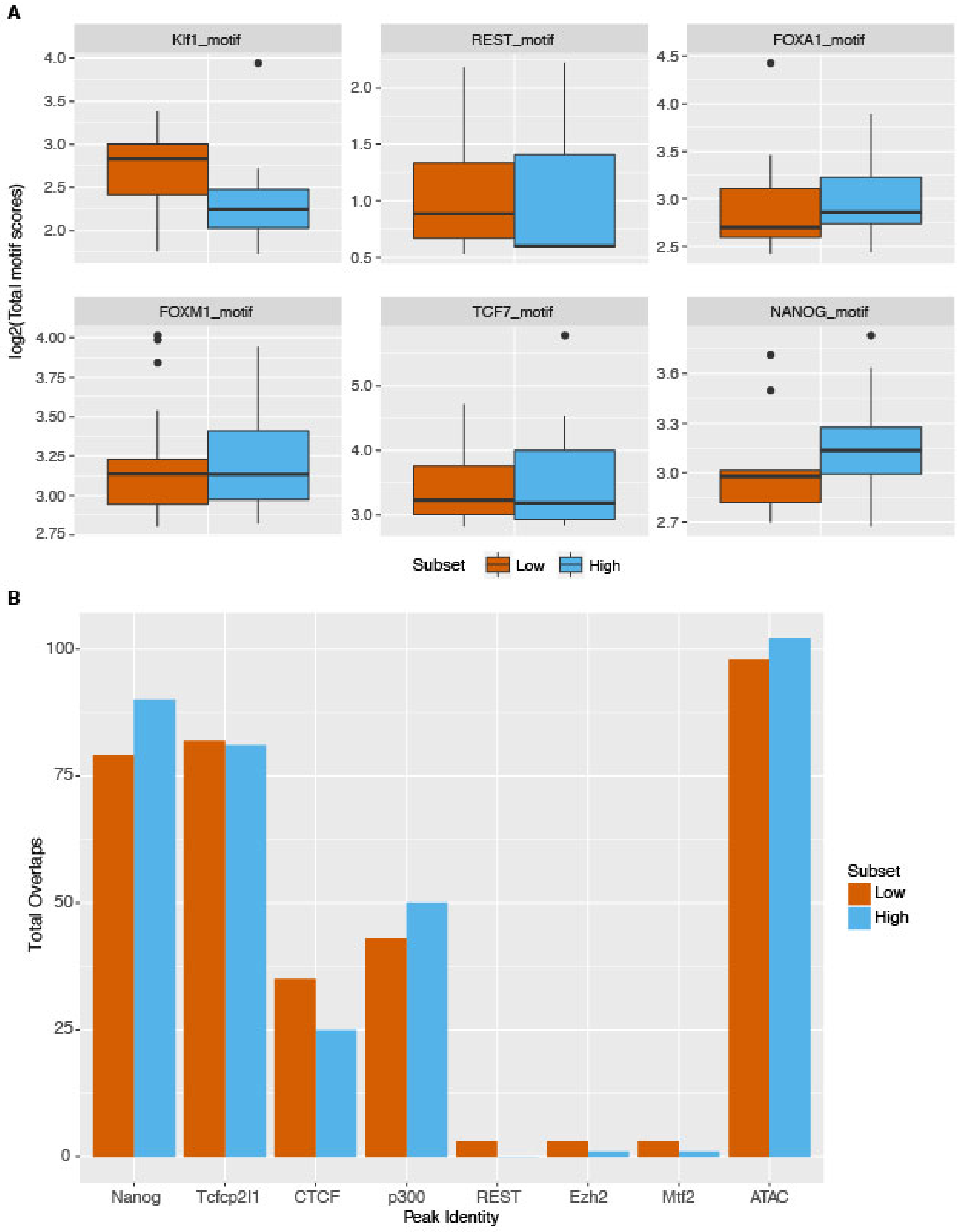

Figure S6. Genomic sequences show signatures for other factors. (**A**) Summed motif scores for indicated motif across genomic sequences, excluding primary pluripotency sites. Site scores output during motif scanning of high (top 25% as ranked by gWT expression, n = 101) and low (bottom 25% as ranked by gWT expression, n = 101) gMUT sequences to prevent scoring of O, S, K, or E TFBS sequences. (**B**) Overlapping TF occupancy, as measured by ChIP-seq, or accessibility, as measured by ATAC-seq, for high (top 25% as ranked by gWT expression, n = 101) and low (bottom 25% as ranked by gWT expression, n = 101) genomic sequence intervals.

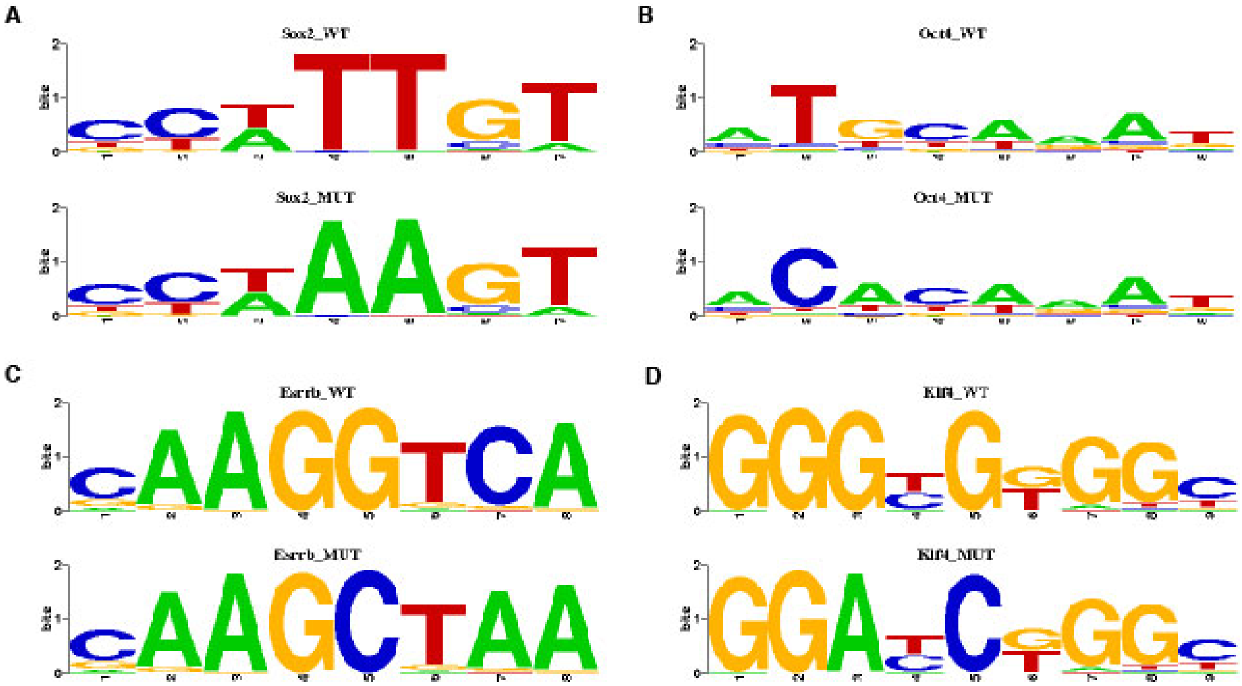

Figure S7. Pluripotency motif substitutions for gMUT sequences. Highest information content positions in each motif were substituted with least frequent nucleotide for that position. (**A**) For mutating Sox2 motifs, the reference nucleotides were substituted for ‘A’ in position 4 and 5. (**B**) For mutating Oct4 motifs, the reference nucleotide was substituted for ‘C’ in position 2 and for ‘A’ in position 3. (**C**) For mutating Esrrb motifs, the reference nucleotide was substituted for ‘C’ in position 5 and ‘A’ for position 7. (**D**) For mutating Klf4 motifs, the reference nucleotide was substituted for ‘A’ in position 3 and ‘C’ in position 5.

Supplemental Table 1: SYN library composition

Element class	Unique Elements	Unique Element-Barcode Pairs
2-mers	48	384
3-mers	192	1,536
4-mers	384	3,072
Basal	1	112
Total library size	625	5,104

Supplemental Table 2: gWT site composition

Sequence composition (Primary sites)	Unique Sequences
OKE	117
OSE	65
OSK	68
SKE	157

Supplemental Table 3: gWT/gMUT library composition

Sequence class	Unique Sequences	Unique Sequence-Barcode Pairs
gWT	407	3,256
gMUT	407	3,256
Basal	1	112
Total library size	815	6,624

Supplemental Table 4: Primer sequences

Name	**Sequence**	**Demultiplexing BC**
Synthetic_FW-1	CTTCTACTACTAGGGCCCA	-
Synthetic_Rev-2	CATGAACTAGCATGTAGAGCTC	-
Genomic_FW-1	GACTTACATTAGGGCCCGT	-
Genomic_Rev-1	CAGTATCGTAGTCCGAGCTC	-
CF121	TAGCGTCGAGGACATCAAGA	-
CF122	TGGTTTGTCCAAACTCATCAA	-
CF150	TACACCGTGGTGGAGCAGTA	-
CF151b	AGCGTACTCGAGTTGTTAACTTGTTTATTGCAGCTT	-
CF52	AATGATACGGCGACCACCGAG	-
CF53	CAAGCAGAAGACGGCATACGA	-
P1_XbaI_1_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAACCTCA	AACCTCA
P1_XbaI_1_R	/5Phos/C*TAGTGAGGTTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT	AACCTCA
P1_XbaI_2_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCTAAGC	TCTAAGC
P1_XbaI_2_R	/5Phos/C*TAGGCTTAGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT	TCTAAGC
P1_XbaI_3_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGTCAT	CTGTCAT
P1_XbaI_3_R	/5Phos/C*TAGATGACAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT	CTGTCAT
P1_XbaI_4_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGGAGGTG	GGAGGTG
P1_XbaI_4_R	/5Phos/C*TAGCACCTCCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT	GGAGGTG
P1_XbaI_5_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTCGAT	GCTCGAT
P1_XbaI_5_R	/5Phos/C*TAGATCGAGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT	GCTCGAT

P1_XbaI_6_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGAGTA	TAGAGTA
P1_XbaI_6_R	/5Phos/C*TAGTACTCTAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT	TAGAGTA
P1_XbaI_7_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCAGTCT	TCAGTCT
P1_XbaI_7_R	/5Phos/C*TAGAGACTGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT	TCAGTCT
P1_XbaI_8_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTTCCAAG	TTCCAAG
P1_XbaI_8_R	/5Phos/C*TAGCTTGGAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT	TTCCAAG
PE2_SIC69_Sal1_F	/5Phos/T*CGAAGATCGGAAGAGCACACGTCTGAACTCCAGTCACAGCGTGCCCATCTCGTATGCCGTCTTCTGCTTG	-
PE2_SIC69_Sal1_R	CAAGCAGAAGACGGCATACGAGATGGGCACGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	-

Supplemental Table 5: **iRF SYN feature matrix**

Model comparison	**Billboard model**	**Positional model**	
Test Set, R^2 (overall)	0.56	0.87	
Test Set, R^2 (4mers)	0.00	0.52	
Features included			**Source**
O_presence	Yes	Yes	Same terms used for SYN iRF Billboard Model. Identity of TFBSs present in the sequence are determined via FIMO.
S_presence	Yes	Yes	
K_presence	Yes	Yes	
E_presence	Yes	Yes	
Position.4_O	(-)	Yes	Same terms used for SYN Billboard + Position iRF Model. Relative position of pluripotency TFBSs as determined via FIMO; as sequences contain only three primary sites Position.1 is FALSE for all factors and omitted from table.
Position.4_S	(-)	Yes	
Position.4_K	(-)	Yes	
Position.4_E	(-)	Yes	
Position.3_O	(-)	Yes	
Position.3_S,	(-)	Yes	
Position.3_K	(-)	Yes	
Position.3_E	(-)	Yes	
Position.2_O	(-)	Yes	
Position.2_S	(-)	Yes	
Position.2_K	(-)	Yes	
Position.2_E	(-)	Yes	
Position.1_O	(-)	Yes	
Position.1_S	(-)	Yes	
Position.1_K	(-)	Yes	
Position.1_E	(-)	Yes	

Supplemental Table 6: iRF gWT Feature Matrix

Model Comparison	**Billboard**	**Positional**	**Spacing**	**Primary sites**	**ChIP**	**All**	
AUROC	0.52	0.47	0.52	0.64	0.56	0.67	
AUPRC	0.22	0.25	0.31	0.34	0.29	0.46	
Features included							**Source**
O_presence	Yes	Yes	(-)	(-)	(-)	(-)	Same terms used for SYN iRF Billboard Model. Identity of

S_presence	Yes	Yes	(-)	(-)	(-)	(-)	TFBSs present in the sequence are determined via FIMO.
K_presence	Yes	Yes	(-)	(-)	(-)	(-)	
E_presence	Yes	Yes	(-)	(-)	(-)	(-)	
Position.4_O	(-)	Yes	(-)	(-)	(-)	(-)	Same terms used for SYN Positional iRF Model. Relative position of pluripotency TFBSs as determined via FIMO; as sequences contain only three primary sites Position.1 is FALSE for all factors and omitted from table.
Position.4_S	(-)	Yes	(-)	(-)	(-)	(-)	
Position.4_K	(-)	Yes	(-)	(-)	(-)	(-)	
Position.4_E	(-)	Yes	(-)	(-)	(-)	(-)	
Position.3_O	(-)	Yes	(-)	(-)	(-)	(-)	
Position.3_S,	(-)	Yes	(-)	(-)	(-)	(-)	
Position.3_K	(-)	Yes	(-)	(-)	(-)	(-)	
Position.3_E	(-)	Yes	(-)	(-)	(-)	(-)	
Position.2_O	(-)	Yes	(-)	(-)	(-)	(-)	
Position.2_S	(-)	Yes	(-)	(-)	(-)	(-)	
Position.2_K	(-)	Yes	(-)	(-)	(-)	(-)	
Position.2_E	(-)	Yes	(-)	(-)	(-)	(-)	
Distance_O.S	(-)	(-)	Yes	(-)	(-)	Yes	Distance between FIMO identified sites (OCT4, SOX2, KLF4, & ESRRB); if a site is absent from the sequence distance between the two factors is sent to total length of sequence (81 or 82 bps).
Distance_K.E	(-)	(-)	Yes	(-)	(-)	Yes	
Distance_K.O	(-)	(-)	Yes	(-)	(-)	Yes	
Distance_K.S	(-)	(-)	Yes	(-)	(-)	Yes	
Distance_E.S	(-)	(-)	Yes	(-)	(-)	Yes	
Distance_E.O	(-)	(-)	Yes	(-)	(-)	Yes	
Distance_1.2	(-)	(-)	Yes	(-)	(-)	Yes	Distance between sites, regardless of identity, present in sequences (1st to 2nd, 2nd to 3rd site).
Distance_2.3	(-)	(-)	Yes	(-)	(-)	Yes	
Total_spacing	(-)	(-)	(-)	(-)	(-)	Yes	Sum of ‘Distance_1.2’ and ‘Distance_2.3’
Oct4.site_affinity	(-)	(-)	(-)	Yes	(-)	Yes	Scores assigned by FIMO, a log-likelihood score ratio based on the PWM provided (Grant et al. 2011).
Sox2.site_affinity	(-)	(-)	(-)	Yes	(-)	Yes	
Klf4.site_affinity	(-)	(-)	(-)	Yes	(-)	Yes	
Esrrb.site_affinity	(-)	(-)	(-)	Yes	(-)	Yes	
OSKE_TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	Sum of OCT4, SOX2, KLF4, & ESRRB site affinities for each sequence (above terms)

Oct4.Occ_10	(-)	(-)	(-)	(-)	(-)	Yes	Total predicted occupancy across the sequence for each pluripotency factor, annotated with custom code (See Methods)
Sox2.Occ_8	(-)	(-)	(-)	(-)	(-)	Yes	
Klf4.Occ_8	(-)	(-)	(-)	(-)	(-)	Yes	
Essrb.Occ_8	(-)	(-)	(-)	(-)	(-)	Yes	
OSKE_P(Occ)	(-)	(-)	(-)	(-)	(-)	Yes	Sum of predicted occupancies for pluripotency factors (above terms)
Klf1_Mut.count	(-)	(-)	(-)	(-)	(-)	Yes	Number of identified sites for gMUT sequences scanned using FIMO with PWMs of SVM supported factors. gMUT sequences were scored to prevent assigning a score for another factor to any of the primary pluripotency sites.
REST_Mut.count	(-)	(-)	(-)	(-)	(-)	Yes	
FOXA1_Mut.count	(-)	(-)	(-)	(-)	(-)	Yes	
FOXM1_Mut.count	(-)	(-)	(-)	(-)	(-)	Yes	
TCF7_Mut.count	(-)	(-)	(-)	(-)	(-)	Yes	
NANOG_Mut.count	(-)	(-)	(-)	(-)	(-)	Yes	
KLF1_Mut.Total Affinity	(-)	(-)	(-)	(-)	(-)	Yes	Sum of scores assigned for FIMO scanning with PWMs of SVM supported factors for gMUT sequences. gMUT sequences were scored to prevent assigning a score for another factor to primary pluripotency sites.
REST_Mut.Total Affinity	(-)	(-)	(-)	(-)	(-)	Yes	
FOXA1_Mut.TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	
FOXM1_Mut.TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	
TCF7_Mut.Total Affinity	(-)	(-)	(-)	(-)	(-)	Yes	
NANOG_Mut.TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	
SVM_TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	Sum of KLF1, REST, FOXA1, FOXM1, TCF7, & Nanog site affinities (above).
Total_site_affinity	(-)	(-)	(-)	(-)	(-)	Yes	Sum of ‘SVM_TotalAffinity’ and ‘OSKE_TotalAffinity’.
Oct4_ChIP	(-)	(-)	(-)	(-)	Yes	Yes	Chen et al. ChIP-seq overlaps. GEO dataset: GSE11431; GEO IDs: GSM288346 (O),
Sox2_ChIP	(-)	(-)	(-)	(-)	Yes	Yes	
Klf4_ChIP	(-)	(-)	(-)	(-)	Yes	Yes	

Essrb_ChIP	(-)	(-)	(-)	(-)	Yes	Yes	GSM288347 (S), GSM288354 (K), GSM288355 (E); E14 mESCs.
Nanog	(-)	(-)	(-)	(-)	Yes	Yes	Additional ChIP-seq peaks from Chen et al. GEO dataset: GSE11431; GEO IDs: GSM288345 (Nanog), GSM288350 (Tcfcp2l1), GSM288351 (CTCF), GSM288359 (p300); E14 mESCs.
Tcfcp2l1	(-)	(-)	(-)	(-)	Yes	Yes	
CTCF	(-)	(-)	(-)	(-)	Yes	Yes	
p300	(-)	(-)	(-)	(-)	Yes	Yes	
Peak_count	(-)	(-)	(-)	(-)	(-)	Yes	Total overlapping pluripotency ChIP seq signals for all peaks from Chen et al. (above), including O,S,K,E, Nanog, Tcfcp2l1, CTCF, & p300.
REST	(-)	(-)	(-)	(-)	Yes	Yes	ChIP-seq from Yu et al. PMID: 21632747; GEO dataset: GSE28233; GEO ID: GSM698696; E14 mESCs.
Mtf2	(-)	(-)	(-)	(-)	Yes	Yes	ChIP-seq from Perino et al. PMID: 29808031; GEO dataset: GSE94300; MAnorm bed files from dataset used for respective factors; E14 mESCs.
Ezh2	(-)	(-)	(-)	(-)	Yes	Yes	
H3K27me3	(-)	(-)	(-)	(-)	Yes	Yes	
H3K4me3	(-)	(-)	(-)	(-)	Yes	Yes	
ATAC	(-)	(-)	(-)	(-)	Yes	Yes	ATAC-seq from Wu et al. PMID: 27309802; GEO ID: GSM2156965; 50k cell stage of mESCs.
H3K27ac	(-)	(-)	(-)	(-)	Yes	Yes	All files downloaded from www.encodeproject.org. Note: all data sets from Yue et al. (PMID: 25409824) are under review due to library complexity and/or read depth issues. E14 mESCs.
H3K36me3	(-)	(-)	(-)	(-)	Yes	Yes	
H3K4me1Ren	(-)	(-)	(-)	(-)	Yes	Yes	
H3K4me1Snyder	(-)	(-)	(-)	(-)	Yes	Yes	
H3K4me3Ren	(-)	(-)	(-)	(-)	Yes	Yes	
H3K4me3Snyder	(-)	(-)	(-)	(-)	Yes	Yes	
H3K9ac	(-)	(-)	(-)	(-)	Yes	Yes	
H3K9me3	(-)	(-)	(-)	(-)	Yes	Yes	