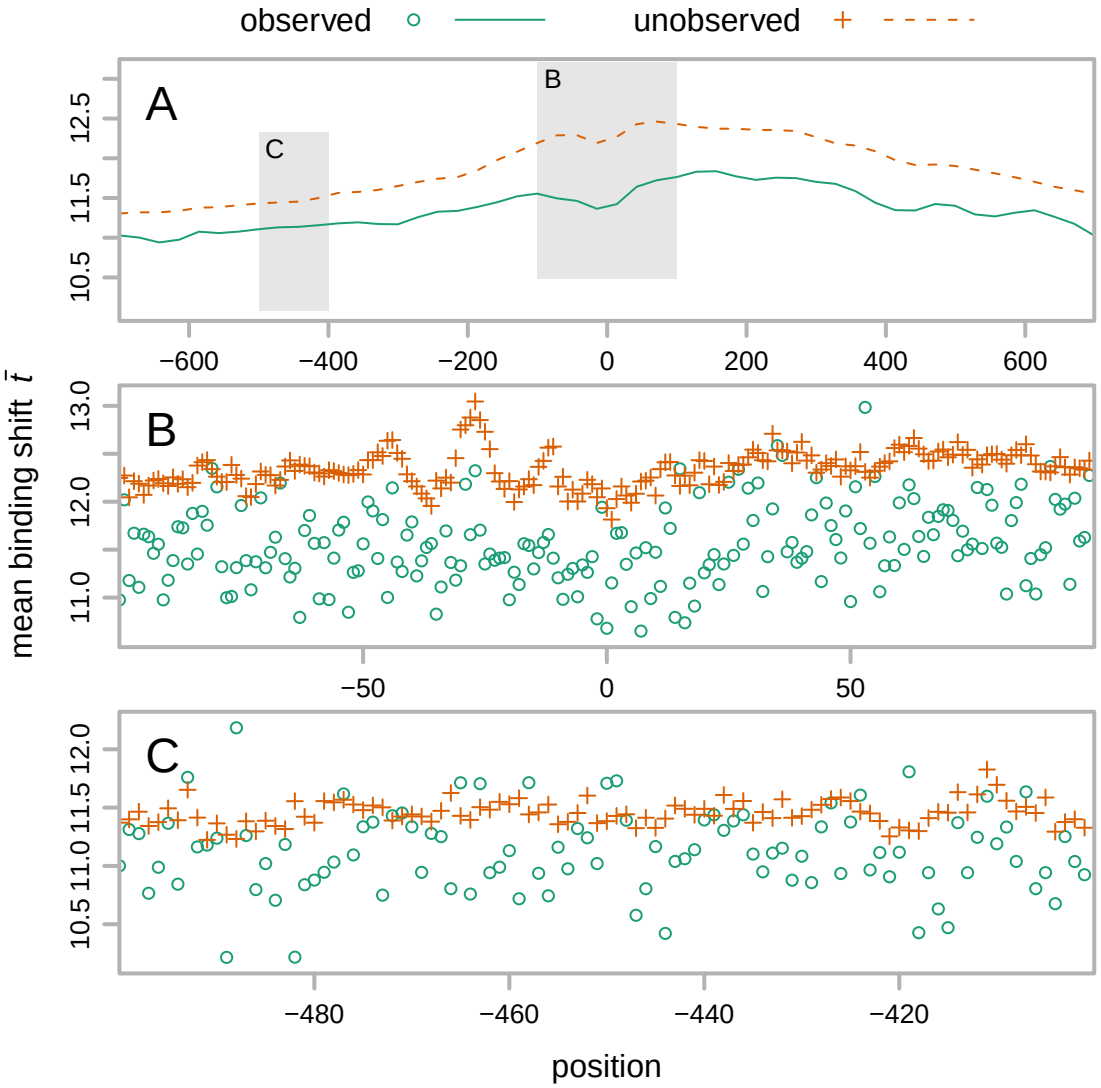
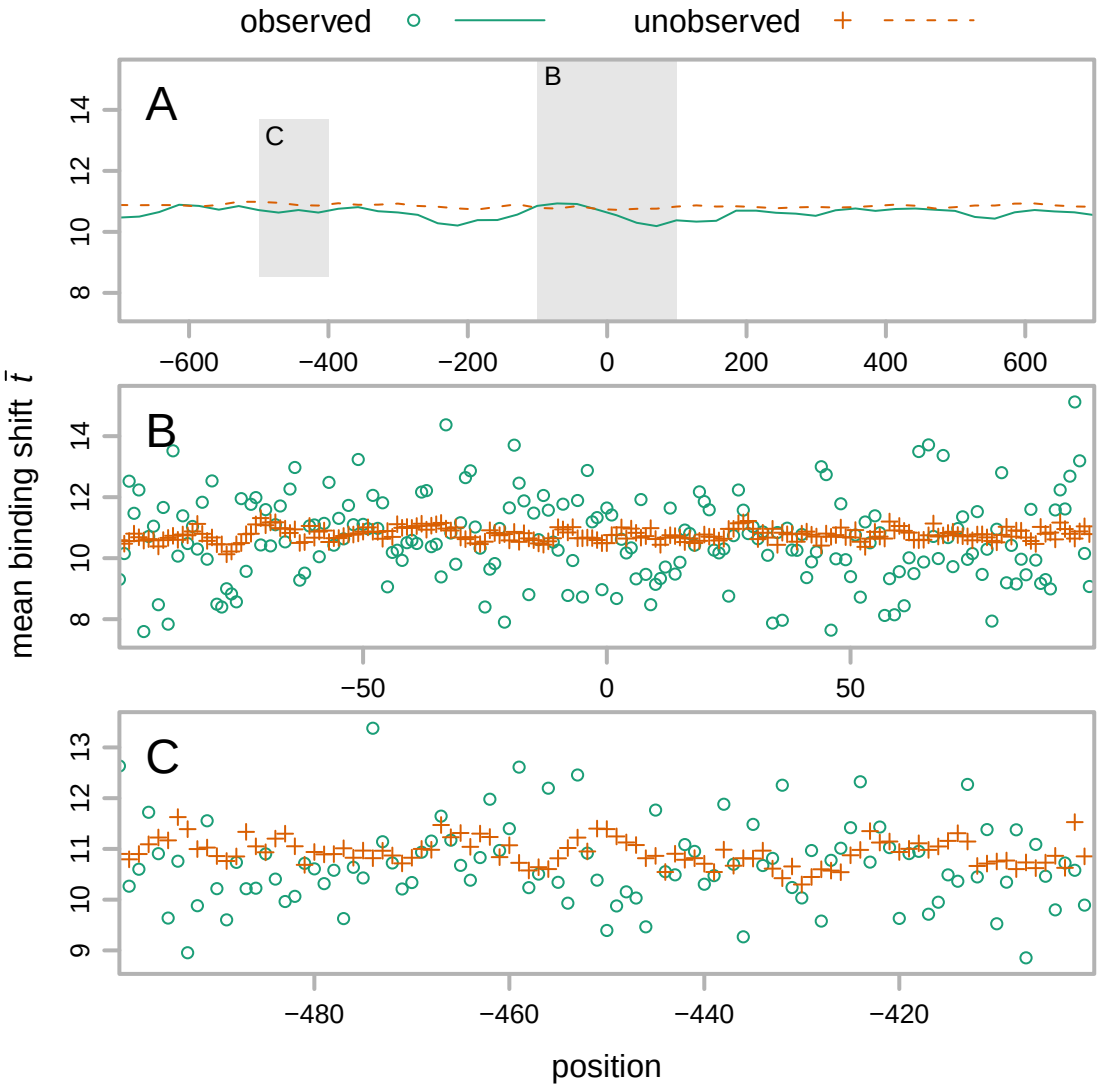
Supplementary Information

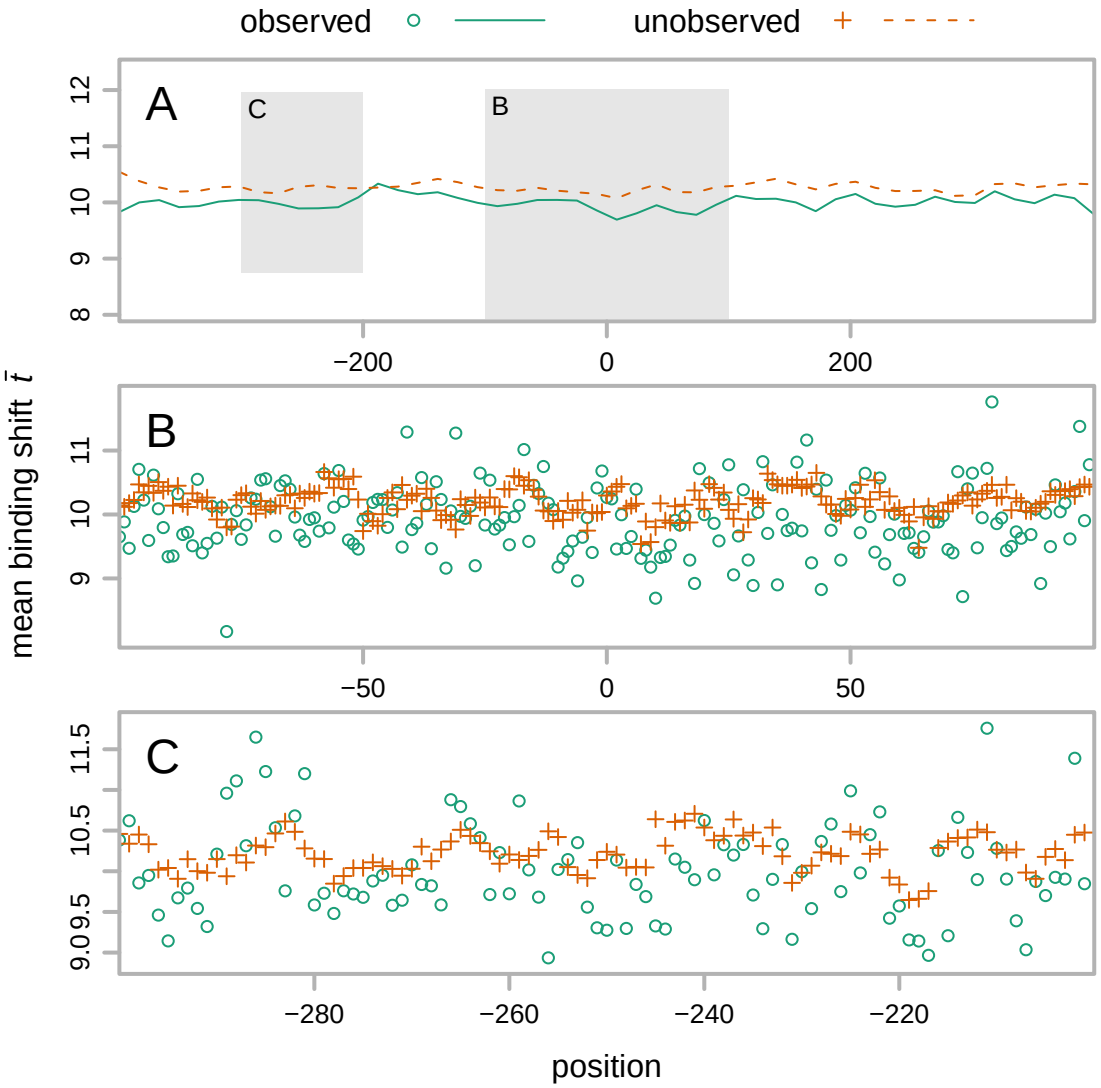
Supplementary Figures and Legends



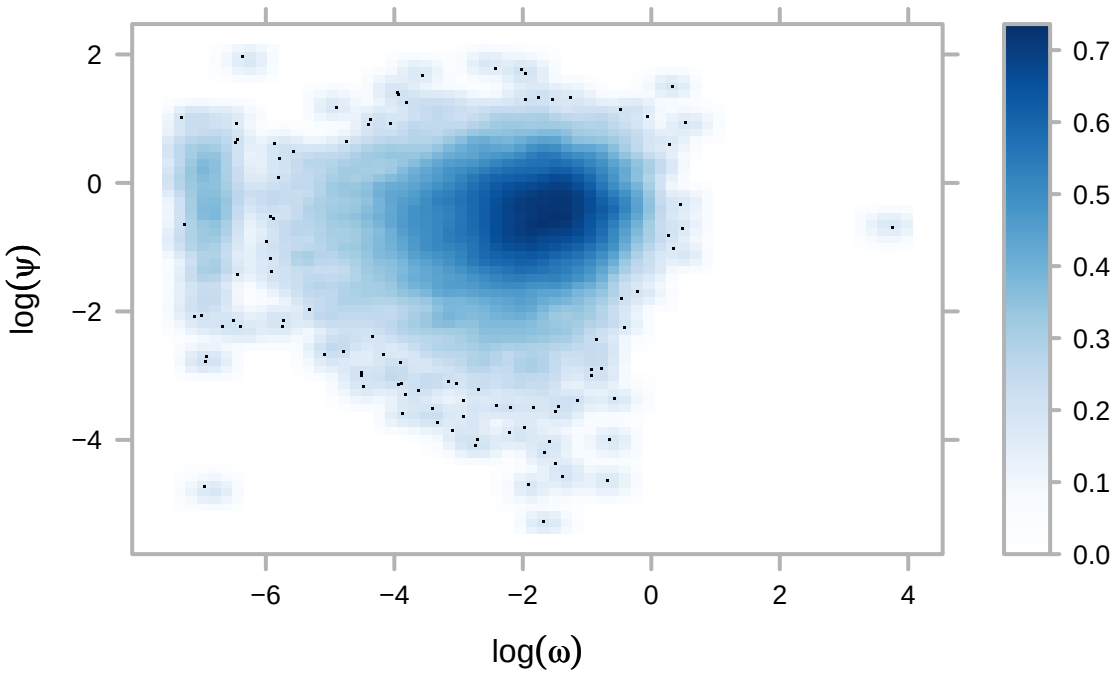
Supplementary Figure 1: Aggregation plot of the binding shifts of 17,194 mouse genes, averaged within two groups: one where the simulated mutation was observed in rat (green circles and solid line), and one where it was unobserved (orange crosses and dashed line). (A) Local regressions for ± 700 bp around the TSS, estimated with the loess (Cleveland and Devlin 1988) function in R (R Development Core Team 2007), with second-degree polynomials and $\alpha = 0.1$. Shaded regions in this plot are magnified as separate panels beneath to show mean binding shifts at individual positions (B) proximal to and (C) more distal from the TSS.



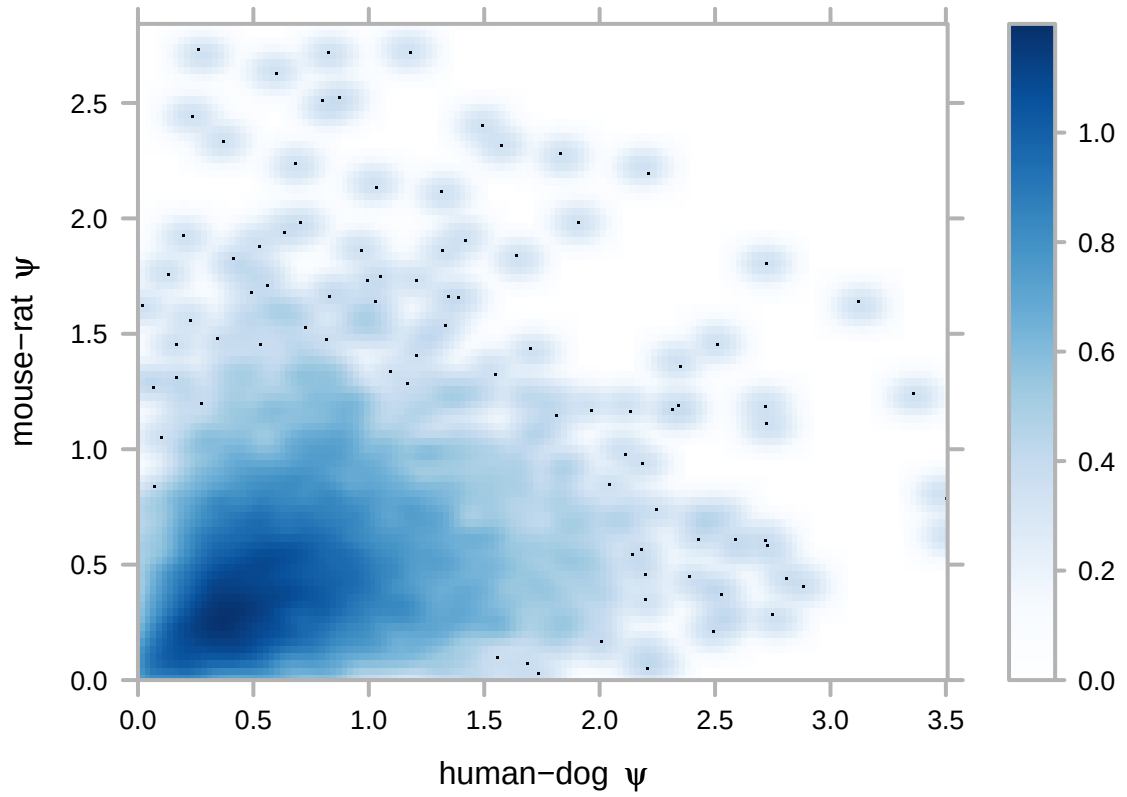
Supplementary Figure 2: Aggregation plot of the binding shifts of 496 human regions with enhancer activity validated in transgenic mice (Pennacchio et al. 2006), averaged within two groups: one where the simulated mutation was observed in mouse (green circles and solid line), and one where it was unobserved (orange crosses and dashed line). (A) Local regressions for ± 700 bp around the midpoint of the enhancer region, estimated with the loess (Cleveland and Devlin 1988) function in R (R Development Core Team 2007), with second-degree polynomials and $\alpha = 0.1$. Shaded regions in this plot are magnified as separate panels beneath to show mean binding shifts at individual positions (B) proximal to and (C) more distal from the midpoint of the enhancer region.



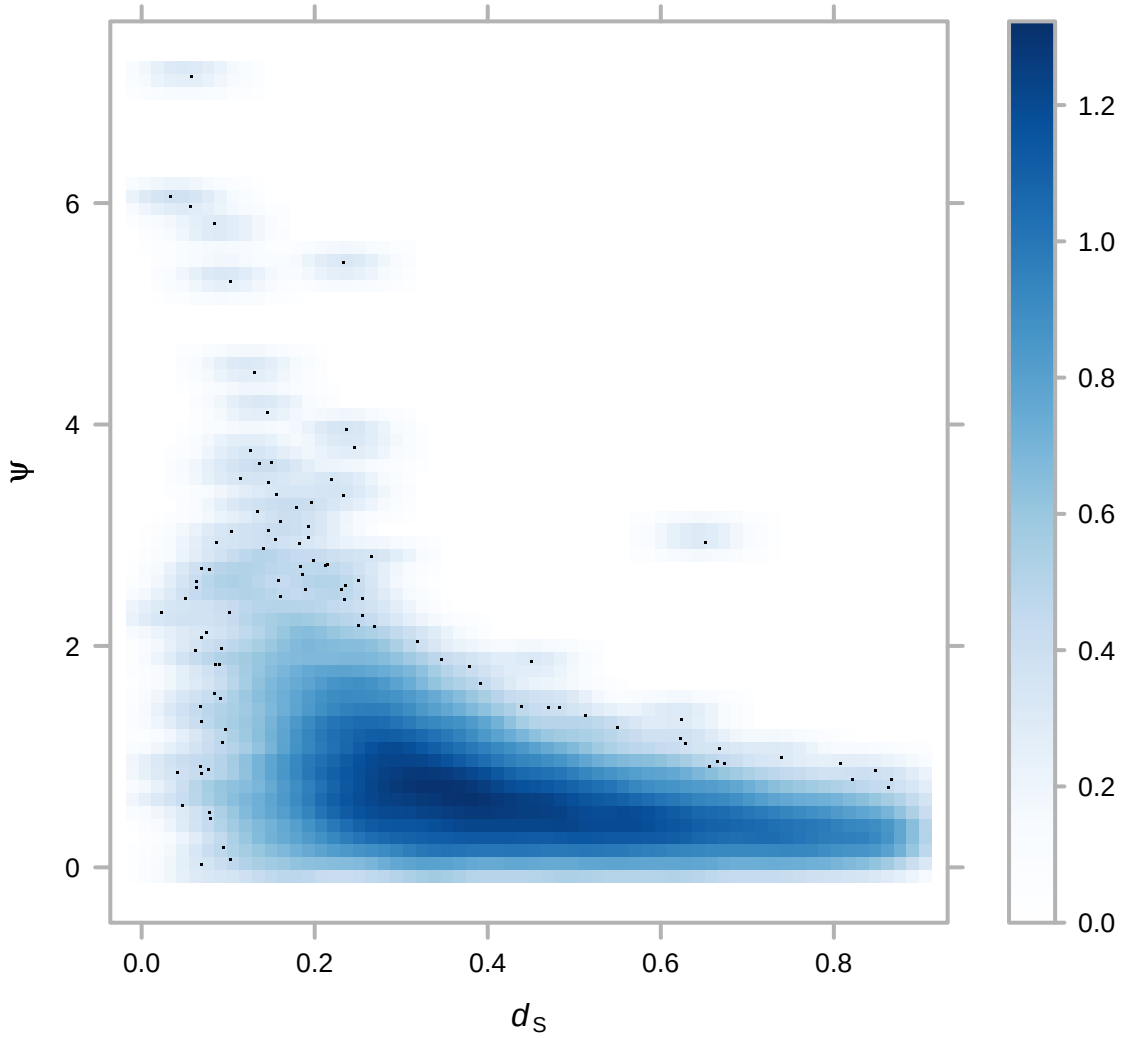
Supplementary Figure 3: Aggregation plot of the binding shifts of 790 human ancestral repeat regions (Paten et al. 2008) with length of at least 1000, averaged within two groups: one where the simulated mutation was observed in mouse (green circles and solid line), and one where it was unobserved (orange crosses and dashed line). (A) Local regressions for ± 400 bp around the midpoint of the ancestral repeats, estimated with the loess (Cleveland and Devlin 1988) function in R (R Development Core Team 2007), with second-degree polynomials and $\alpha = 0.1$. Shaded regions in this plot are magnified as separate panels beneath to show mean binding shifts at individual positions (B) proximal to and (C) more distal from the midpoint of the ancestral repeats.



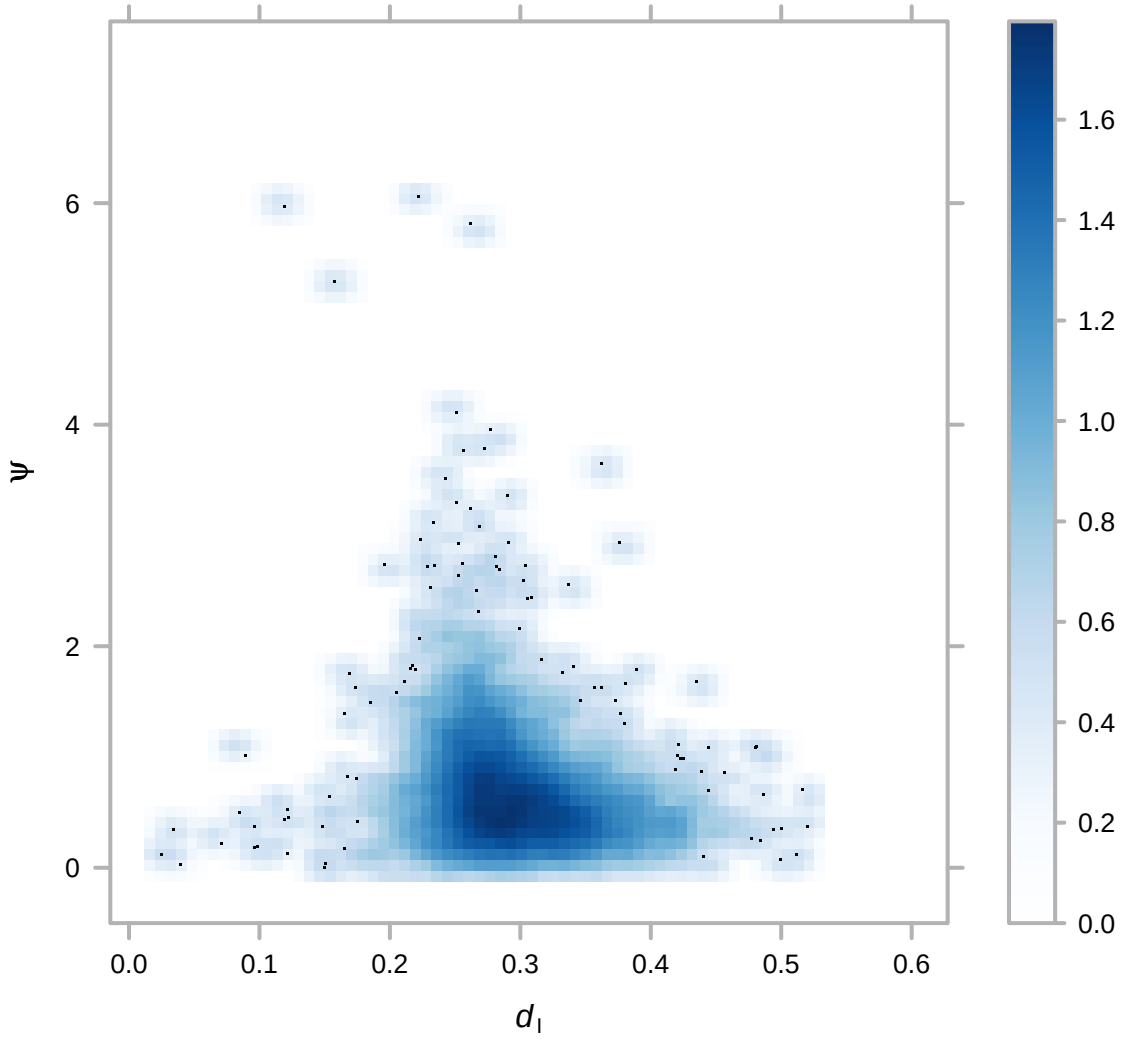
Supplementary Figure 4: Smoothed log-log scatter plot of ψ against ω for 9060 human–dog transcripts, excluding transcripts where d_T, d_N, or d_S is 0. The plot space is divided into a number of cells, and the color of each cell indicates the fourth root of a two-dimensional kernel density estimate of the number of data points in that cell, as performed by the `geneplotter` package of Bioconductor (Gentleman et al. 2004). The color key to the right of the plot indicates how various colors correspond to the transformed density estimate values.



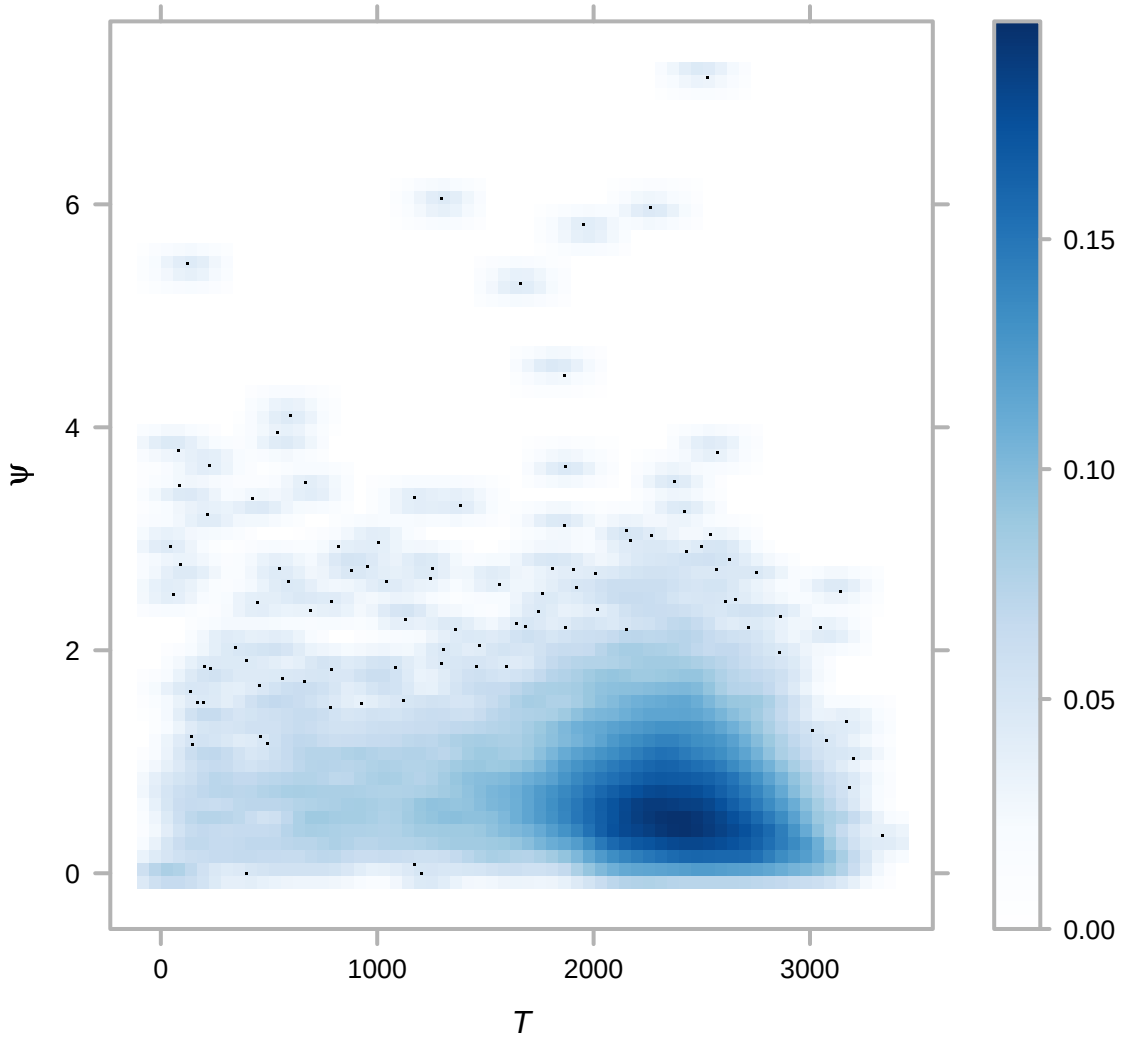
Supplementary Figure 5: Scatter plot of mouse–rat ψ against human–dog ψ for 4547 1:1:1:1 orthologous transcripts. The plot space is divided into a number of cells, and the color of each cell indicates the fourth root of a two-dimensional kernel density estimate of the number of data points in that cell, as performed by the `geneplotter` package of Bioconductor (Gentleman et al. 2004). Only the bottom 99.9 percent of mouse–rat and human–dog ψ values are displayed, to eliminate the distorting effect of outliers. The color key to the right of the plot indicates how various colors correspond to the transformed density estimate values.



Supplementary Figure 6: Smoothed scatter plot of ψ against d_S for 17,600 human–dog transcripts. The plot space is divided into a number of cells, and the color of each cell indicates the fourth root of a two-dimensional kernel density estimate of the number of data points in that cell, as performed by the `geneplotter` package of Bioconductor (Gentleman et al. 2004). The color key to the right of the plot indicates how various colors correspond to the transformed density estimate values.



Supplementary Figure 7: Smoothed scatter plot of ψ against d_I for 17,600 human–dog transcripts. The plot space is divided into a number of cells, and the color of each cell indicates the fourth root of a two-dimensional kernel density estimate of the number of data points in that cell, as performed by the `geneplotter` package of Bioconductor (Gentleman et al. 2004). The color key to the right of the plot indicates how various colors correspond to the transformed density estimate values.



Supplementary Figure 8: Smoothed scatter plot of ψ against T for 17,600 human–dog transcripts. The plot space is divided into a number of cells, and the color of each cell indicates the fourth root of a two-dimensional kernel density estimate of the number of data points in that cell, as performed by the `geneplotter` package of Bioconductor (Gentleman et al. 2004). The color key to the right of the plot indicates how various colors correspond to the transformed density estimate values.

Supplementary Tables

Supplementary Table 1: Identifiers used by JASPAR CORE for the 89 TFs used in the model, along with the HGNC symbol for the TF's gene or its human ortholog.

JASPAR identifier	HGNC symbol
Ar	*AR*
Arnt	*ARNT*
Arnt-Ahr	*AHR*
Bapx1	*NKX3-2*
c-ETS	*ETS1*
cEBP	*CPEB1*
Chop-cEBP	*DDIT3*
CREB1	*CREB1*
deltaEF1	*ZEB1*
E2F1	*E2F1*
ELK1	*ELK1*
ELK4	*ELK4*
En1	*EN1*
ESR1	*ESR1*
Evi1	*EVI1*
Fos	*FOS*
Foxa2	*FOXA2*
FOXC1	*FOXC1*
FOXD1	*FOXD1*
Foxd3	*FOXD3*
FOXF2	*FOXF2*
FOXI1	*FOXI1*
FOXL1	*FOXL1*
Foxq1	*FOXQ1*
GABPA	*GABPA*
Gata1	*GATA1*
GATA2	*GATA2*
GATA3	*GATA3*
Gfi	*GFI1*
HAND1-TCF3	*TCF3*
HLF	*HLF*
HNF4	*HNF4A*
IRF1	*IRF1*
IRF2	*IRF2*
Klf4	*KLF4*
MafB	*MAFB*
MAX	*MAX*
MEF2A	*MEF2A*
Myb	*MYB*

Supplementary Table 1: (continued)

JASPAR identifier	HGNC symbol
MYC-MAX	*MYC*
Mycn	*MYCN*
Myf	*MYOD1*
NF-kappaB	*NFKB1*
NFIL3	*NFIL3*
NFKB1	*NFKB1*
NHLH1	*NHLH1*
Nkx2-5	*NKX2-5*
NR1H2-RXR	*NR1H2*
NR2F1	*NR2F1*
NR3C1	*NR3C1*
Pax2	*PAX2*
Pax4	*PAX4*
Pax5	*PAX5*
Pax6	*PAX6*
Pbx	*PBX1*
PPARG	*PPARG*
PPARG-RXRA	*RXRA*
Prrx2	*PRRX2*
REL	*REL*
RELA	*RELA*
Roaz	*ZNF423*
RORA	*RORA*
RORA1	*RORA*
RREB1	*RREB1*
RUNX1	*RUNX1*
RUSH1-alfa	*GATA4*
RXR-VDR	*VDR*
Sox17	*SOX17*
Sox5	*SOX5*
SOX9	*SOX9*
SP1	*SP1*
SPI1	*SPI1*
SPIB	*SPIB*
Spz1	*SPZ1*
SRF	*SRF*
SRY	*SRY*
Staf	*ZNF143*
T	*T*
TAL1-TCF3	*TCF3*
TBP	*TBP*
TCF1	*HNF1A*

Supplementary Table 1: (continued)

JASPAR identifier	HGNC symbol
TCF11-MafG	*NFE2L1*
TEAD	*TEAD1*
TFAP2A	*TFAP2A*
TP53	*TP53*
USF1	*USF1*
YY1	*YY1*
ZNF42_1-4	*MZF1*
ZNF42_5-13	*MZF1*

	low T		not in JASPAR PWMs	
GO term	p	q	p	q
immune response	4×10^{-22}	$< 1 \times 10^{-4}$	0.5	1
antigen processing and presentation	3×10^{-17}	$< 1 \times 10^{-4}$	0.4	1
defense response	1×10^{-16}	$< 1 \times 10^{-4}$	0.9	1
immune system process	4×10^{-15}	$< 1 \times 10^{-4}$	1	1
sensory perception of smell	6×10^{-15}	$< 1 \times 10^{-4}$	0.3	1
sensory perception of chemical stimulus	4×10^{-14}	$< 1 \times 10^{-4}$	0.3	1
response to stimulus	6×10^{-13}	$< 1 \times 10^{-4}$	1	1
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	7×10^{-10}	$< 1 \times 10^{-4}$	0.8	1
inflammatory response	2×10^{-9}	$< 1 \times 10^{-4}$	1	1
RNA processing	6×10^{-8}	$< 1 \times 10^{-4}$	0.2	1
response to other organism	5×10^{-7}	2×10^{-4}	0.5	1
response to wounding	7×10^{-7}	1×10^{-4}	0.8	1
multi-organism process	9×10^{-7}	1×10^{-4}	0.3	1

Supplementary Table 2: GO biological process terms enriched in genes with low-T TSS-flanking regions (Wilcoxon rank sum test; $p < 10^{-6}$). The global analysis is significant (Kolmogorov-Smirnov–like test; $p = 6 \times 10^{-3}$). For each term the significance of its enrichment in genes not in the JASPAR PWM set is also shown.

	low T		not in JASPAR PWMs	
GO term	p	q	p	q
olfactory receptor activity	5×10^{-17}	$< 1 \times 10^{-4}$	0.2	1
chemokine receptor binding	7×10^{-8}	$< 1 \times 10^{-4}$	0.8	1
MHC class II receptor activity	1×10^{-7}	$< 1 \times 10^{-4}$	1	1
chemokine activity	3×10^{-7}	$< 1 \times 10^{-4}$	0.8	1
G-protein-coupled receptor binding	7×10^{-7}	$< 1 \times 10^{-4}$	0.7	1

Supplementary Table 3: GO molecular function terms enriched in genes with low-T TSS-flanking regions (Wilcoxon rank sum test; $p < 10^{-6}$). The global analysis is significant (Kolmogorov-Smirnov–like test; $p < 10^{-4}$). For each term the significance of its enrichment in genes not in the JASPAR PWM set is also shown.

	low T		not in JASPAR PWMs	
GO term	p	q	p	q
MHC protein complex	1×10^{-17}	$< 1 \times 10^{-4}$	0.5	1
MHC class II protein complex	3×10^{-13}	$< 1 \times 10^{-4}$	0.7	1
ribonucleoprotein complex	3×10^{-9}	$< 1 \times 10^{-4}$	0.07	0.6

Supplementary Table 4: GO cellular component terms enriched in genes with low-T TSS-flanking regions (Wilcoxon rank sum test; $p < 10^{-6}$). The global analysis is significant (Kolmogorov-Smirnov–like test; $p = 3 \times 10^{-3}$). For each term the significance of its enrichment in genes not in the JASPAR PWM set is also shown.

	high T		in JASPAR PWMs	
GO term	p	q	p	q
post-translational protein modification	5×10^{-15}	$< 1 \times 10^{-4}$	1	1
protein modification process	9×10^{-14}	$< 1 \times 10^{-4}$	1	1
biopolymer modification	4×10^{-13}	$< 1 \times 10^{-4}$	1	1
protein amino acid phosphorylation	5×10^{-12}	$< 1 \times 10^{-4}$	1	1
phosphorus metabolic process	4×10^{-10}	$< 1 \times 10^{-4}$	1	1
phosphate metabolic process	4×10^{-10}	$< 1 \times 10^{-4}$	1	1
intracellular signaling cascade	3×10^{-9}	$< 1 \times 10^{-4}$	0.8	1
phosphorylation	8×10^{-9}	$< 1 \times 10^{-4}$	1	1
cell communication	2×10^{-8}	$< 1 \times 10^{-4}$	0.6	1
small GTPase mediated signal transduction	8×10^{-8}	$< 1 \times 10^{-4}$	1	1
organ morphogenesis	1×10^{-7}	$< 1 \times 10^{-4}$	6×10^{-14}	$< 1 \times 10^{-4}$
signal transduction	1×10^{-7}	$< 1 \times 10^{-4}$	0.8	1
localization	2×10^{-7}	$< 1 \times 10^{-4}$	0.9	1
anatomical structure morphogenesis	2×10^{-7}	$< 1 \times 10^{-4}$	3×10^{-16}	$< 1 \times 10^{-4}$
developmental process	2×10^{-7}	$< 1 \times 10^{-4}$	3×10^{-22}	$< 1 \times 10^{-4}$
biological regulation	3×10^{-7}	$< 1 \times 10^{-4}$	1×10^{-26}	$< 1 \times 10^{-4}$
endocytosis	4×10^{-7}	$< 1 \times 10^{-4}$	0.6	1
membrane invagination	4×10^{-7}	$< 1 \times 10^{-4}$	0.6	1
regulation of biological process	4×10^{-7}	$< 1 \times 10^{-4}$	6×10^{-29}	$< 1 \times 10^{-4}$
regulation of cellular process	7×10^{-7}	$< 1 \times 10^{-4}$	1×10^{-29}	$< 1 \times 10^{-4}$
anatomical structure development	7×10^{-7}	$< 1 \times 10^{-4}$	4×10^{-19}	$< 1 \times 10^{-4}$

Supplementary Table 5: GO biological process terms enriched in genes with high-T TSS-flanking regions (Wilcoxon rank sum test; $p < 10^{-6}$). The global analysis is significant (Kolmogorov-Smirnov–like test; $p < 10^{-4}$). For each term the significance of its enrichment in genes in the JASPAR PWM set is also shown.

	high T		in JASPAR PWMs	
GO term	p	q	p	q
phosphotransferase activity, alcohol group as acceptor	1×10^{-14}	$< 1 \times 10^{-4}$	1	1
kinase activity	3×10^{-14}	$< 1 \times 10^{-4}$	1	1
transferase activity, transferring phosphorus-containing groups	8×10^{-13}	$< 1 \times 10^{-4}$	1	1
protein kinase activity	2×10^{-12}	$< 1 \times 10^{-4}$	1	1
protein serine/threonine kinase activity	2×10^{-12}	$< 1 \times 10^{-4}$	1	1
protein tyrosine kinase activity	3×10^{-11}	$< 1 \times 10^{-4}$	1	1
transferase activity	2×10^{-10}	$< 1 \times 10^{-4}$	1	1
protein binding	3×10^{-8}	$< 1 \times 10^{-4}$	8×10^{-4}	0.01
magnesium ion binding	3×10^{-7}	$< 1 \times 10^{-4}$	1	1

Supplementary Table 6: GO molecular function terms enriched in genes with high-T TSS-flanking regions (Wilcoxon rank sum test; $p < 10^{-6}$). The global analysis is significant (Kolmogorov-Smirnov–like test; $p < 10^{-4}$). For each term the significance of its enrichment in genes in the JASPAR PWM set is also shown.

	high T		in JASPAR PWMs	
GO term	p	q	p	q
cytoplasm	4×10^{-12}	$< 1 \times 10^{-4}$	1	1

Supplementary Table 7: GO cellular component terms enriched in genes with high-T TSS-flanking regions (Wilcoxon rank sum test; $p < 10^{-6}$). The global analysis is significant (Kolmogorov-Smirnov–like test; $p < 10^{-4}$). For each term the significance of its enrichment in genes in the JASPAR PWM set is also shown.

	low ψ		high number of gaps	
GO term	p	q	p	q
extracellular region	2×10^{-17}	$< 1 \times 10^{-4}$	0.4	1
extracellular region part	5×10^{-12}	$< 1 \times 10^{-4}$	0.3	1
extracellular space	6×10^{-8}	$< 1 \times 10^{-4}$	0.2	1
plasma membrane	3×10^{-7}	$< 1 \times 10^{-4}$	1	1
plasma membrane part	4×10^{-7}	$< 1 \times 10^{-4}$	1	1
extracellular matrix	5×10^{-7}	$< 1 \times 10^{-4}$	0.6	1
proteinaceous extracellular matrix	7×10^{-7}	$< 1 \times 10^{-4}$	0.6	1

Supplementary Table 8: GO cellular component terms enriched in genes with low-ψ TSS-flanking regions (Wilcoxon rank sum test; $p < 10^{-6}$). The global analysis is significant (Kolmogorov-Smirnov–like test; $p = 2 \times 10^{-3}$). For each term the significance of its enrichment in genes with a high number of gaps is also shown.

	high ψ		high number of gaps	
GO term	p	q	p	q
cytoplasm	$< 2 \times 10^{-308}$	$< 1 \times 10^{-4}$	0.3	1
cytoplasmic part	8×10^{-16}	$< 1 \times 10^{-4}$	0.6	1
mitochondrion	5×10^{-15}	$< 1 \times 10^{-4}$	0.3	1
intracellular	2×10^{-12}	$< 1 \times 10^{-4}$	1	1
intracellular part	2×10^{-11}	$< 1 \times 10^{-4}$	1	1
organelle membrane	1×10^{-8}	$< 1 \times 10^{-4}$	1	1
cell part	6×10^{-8}	$< 1 \times 10^{-4}$	0.8	1
cell	8×10^{-8}	$< 1 \times 10^{-4}$	0.8	1
ribonucleoprotein complex	2×10^{-7}	$< 1 \times 10^{-4}$	1	1
organelle part	9×10^{-7}	$< 1 \times 10^{-4}$	1	1

Supplementary Table 9: GO cellular component terms enriched in genes with high-ψ TSS-flanking regions (Wilcoxon rank sum test; $p < 10^{-6}$). The global analysis is significant (Kolmogorov-Smirnov–like test; $p < 10^{-4}$). For each term the significance of its enrichment in genes with a high number of gaps is also shown.