

Supplementary Tables

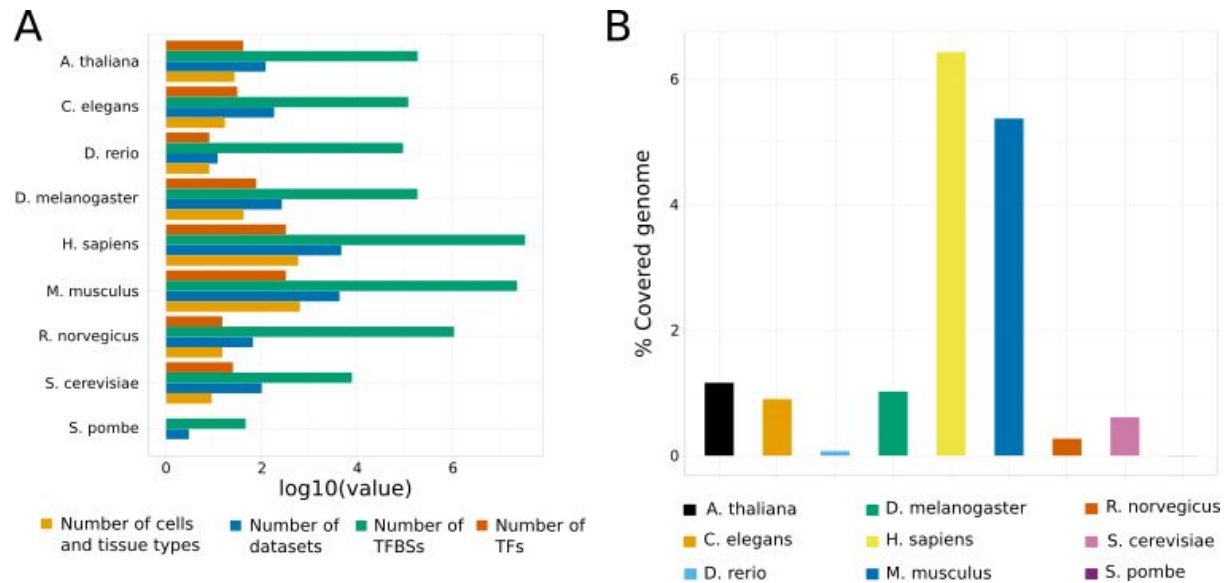
Organism	Number of datasets	Number of TFs	Number of cells / tissue types	Number of TFBSs	Number of CRMs
<i>A. thaliana</i>	121	41	27	182,112	2,217
<i>C. elegans</i>	182	31	17	116,018	871
<i>D. rerio</i>	12	8	8	90,455	1,145
<i>D. melanogaster</i>	264	77	42	181,359	1,990
<i>H. sapiens</i>	4,635	320	581	32,404,467	110,872
<i>M. musculus</i>	4,242	319	629	22,042,987	74,601
<i>R. norvegicus</i>	66	15	15	1,055,551	15,889
<i>S. cerevisiae</i>	100	25	9	7,699	210
<i>S. pombe</i>	3	1	1	46	0
Total	9,625	837	1,329	56,080,694	207,795

Supplementary Table 1. Overview of the permissive collection. Table providing the number of datasets, TFs, cell / tissue types, and TFBSs in the permissive collection of UniBind.

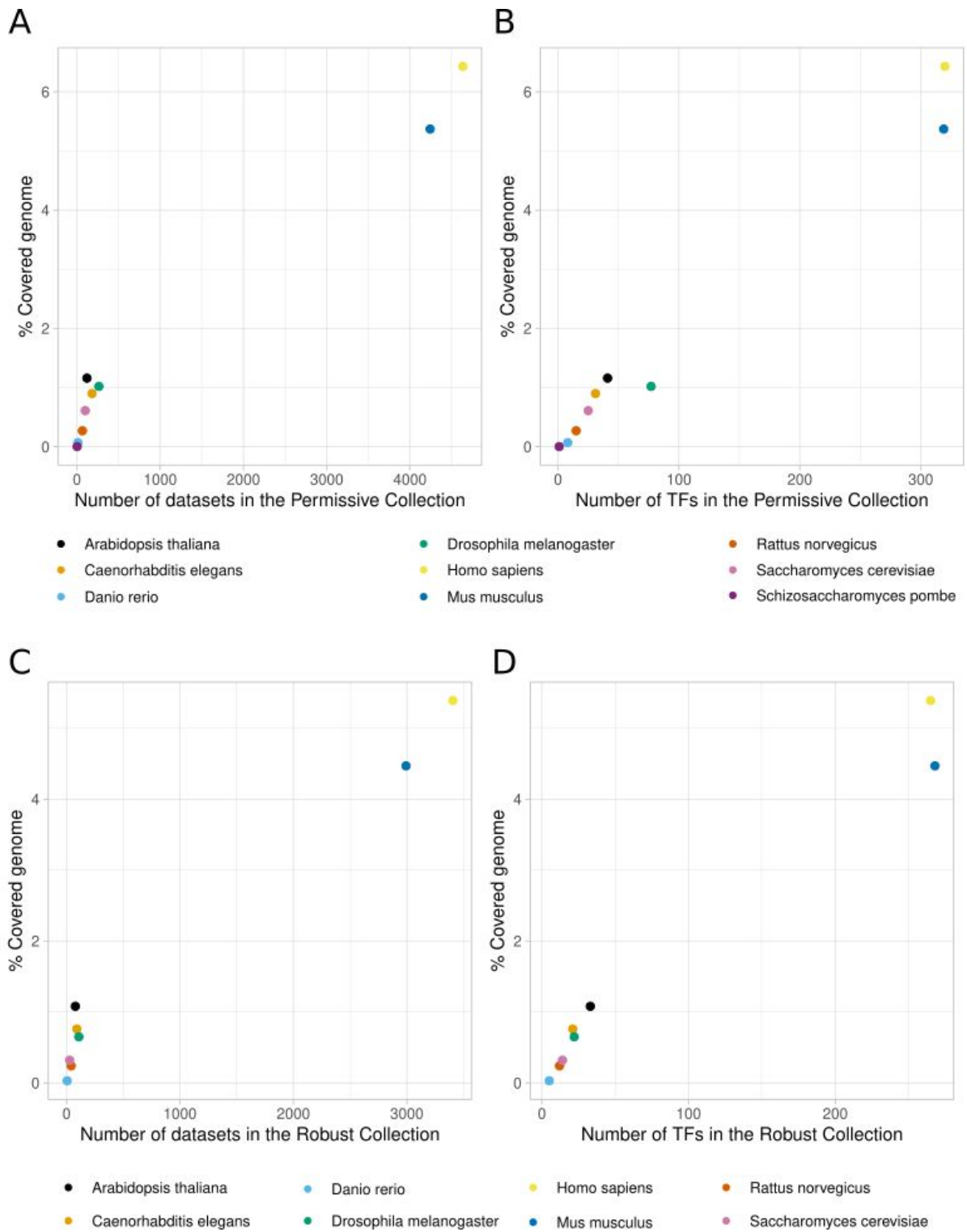
Organism	Number of datasets	Number of TFs	Number of cells / tissue types	Number of TFBSs	Number of CRMs
<i>A. thaliana</i>	78	33	22	169,649	2,211
<i>C. elegans</i>	91	21	12	93,138	691
<i>D. rerio</i>	6	5	4	44,187	619
<i>D. melanogaster</i>	109	22	28	95,856	1,191
<i>H. sapiens</i>	3,406	265	506	25,445,413	101,584
<i>M. musculus</i>	2,993	268	512	17,166,858	70,942
<i>R. norvegicus</i>	41	12	13	939,924	6,190
<i>S. cerevisiae</i>	28	14	4	3,643	119
Total	6,752	640	1,101	43,958,668	183,547

Supplementary Table 2. Overview of the robust collection. Table providing the number of datasets, TFs, cell / tissue types, and TFBSs in the robust collection of UniBind.

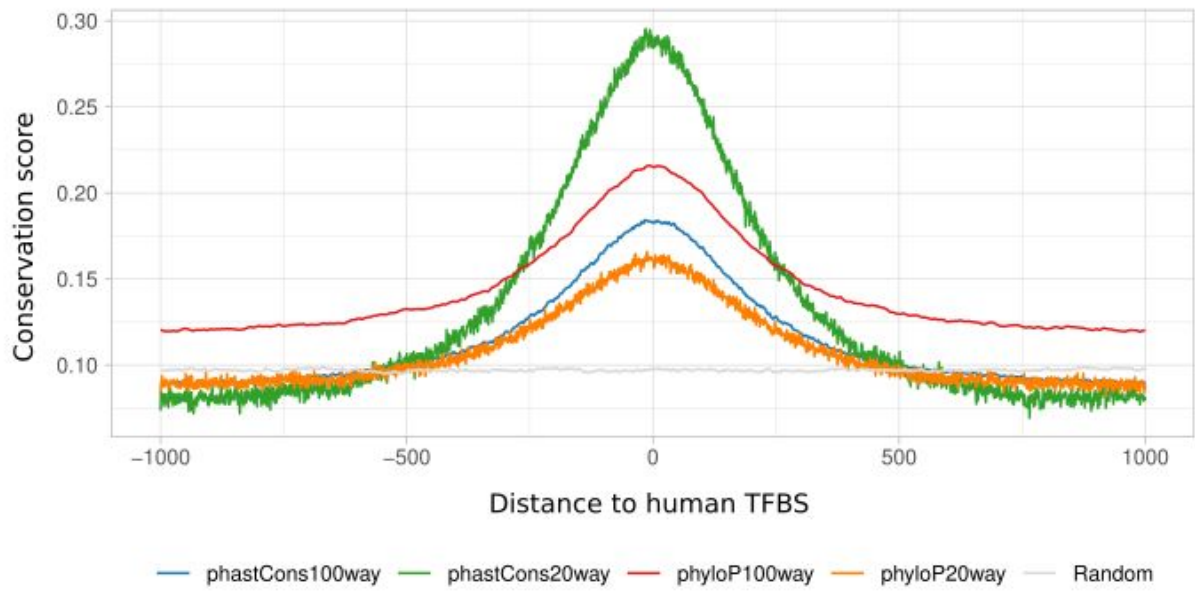
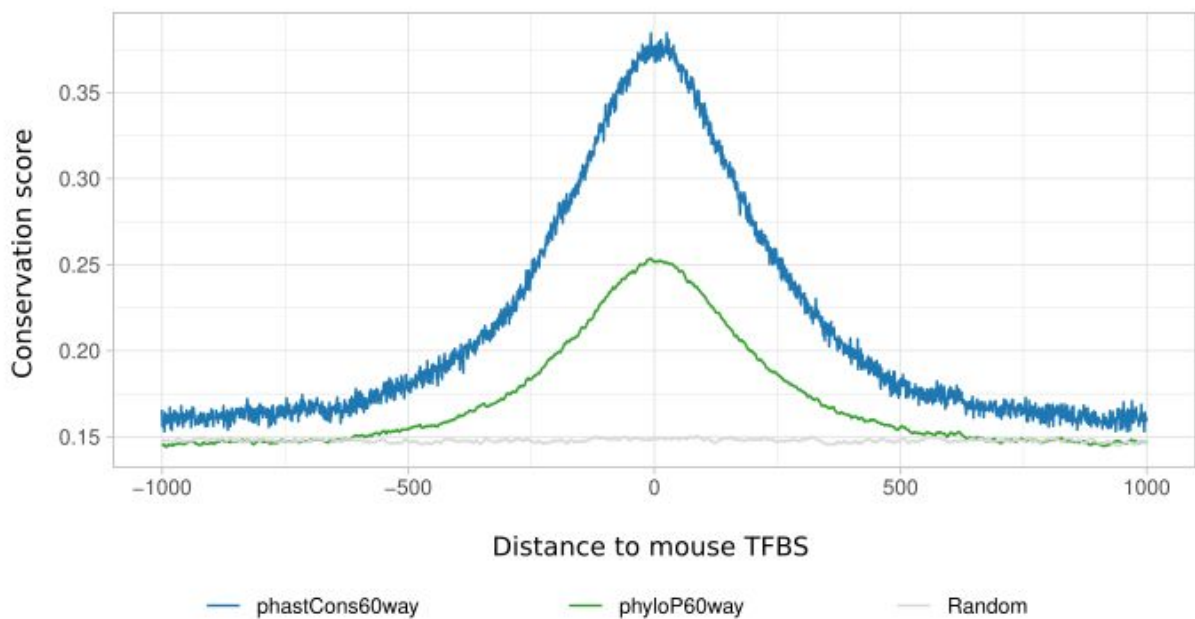
Supplementary Figures



Supplementary Figure 1. Visual overview of the permissive collection. Figure 1. (A) Barplots showing the number of TFs (dark orange), TFBSs (green), datasets (blue), and cell and tissue types (light orange) stored in the permissive collection of UniBind for each analyzed species. All numbers are provided after transformation using the \log_{10} function. **(B)** Distribution of the percentages of the genomes covered by robust TFBSs in each species (one color per species, see legend).



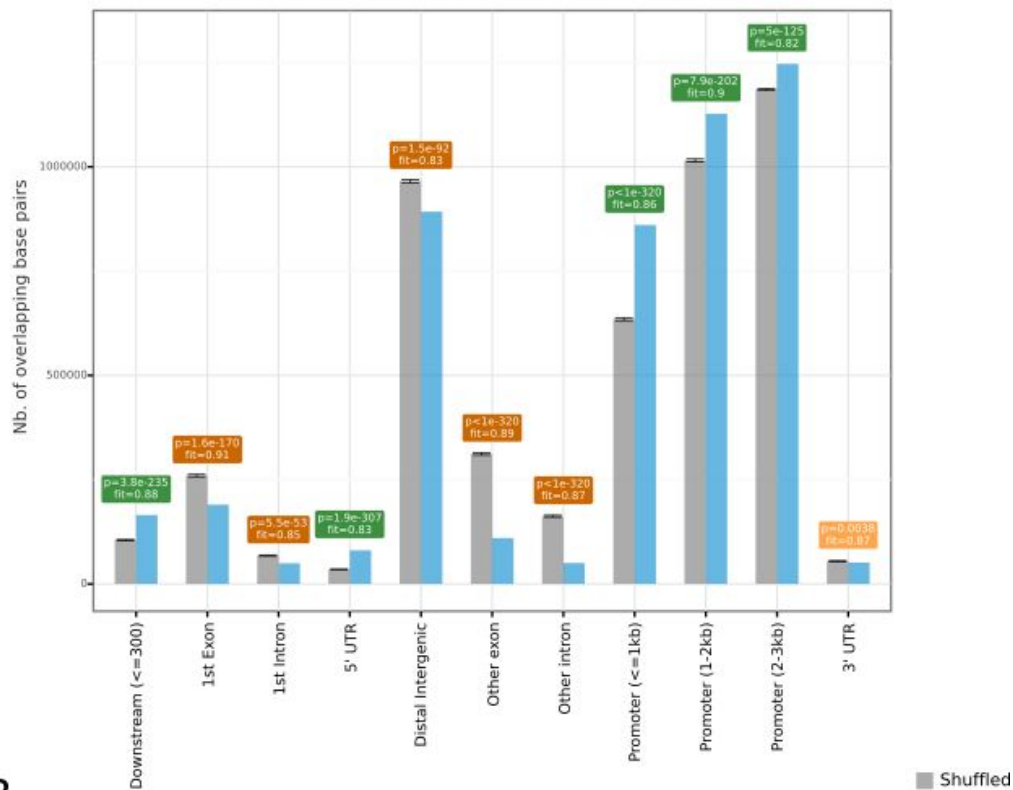
Supplementary Figure 2. Relationship between number of datasets and genome coverage. Scatter plots representing the percentage of genome coverage (y-axes) with respect to the number of datasets in the permissive (**A**) and robust (**C**) collections or the number of TFs in the permissive (**B**) and robust (**D**) collection (x-axes). Each colored point in each panel represents the data associated to one species (see legend for color coding).

A**B**

Supplementary Figure 3. Evolutionary conservation at human and mouse robust CRMs. Distributions of the average base-pair evolutionary conservation scores (phyloP and phastCons scores using multi-species genome alignments, see legend) at regions centered around UniBind human (**A**) and mouse (**B**) CRMs from the robust collection. Conservation of random CRMs was obtained by shuffling the original CRMs and obtaining the conservation score of the new regions.

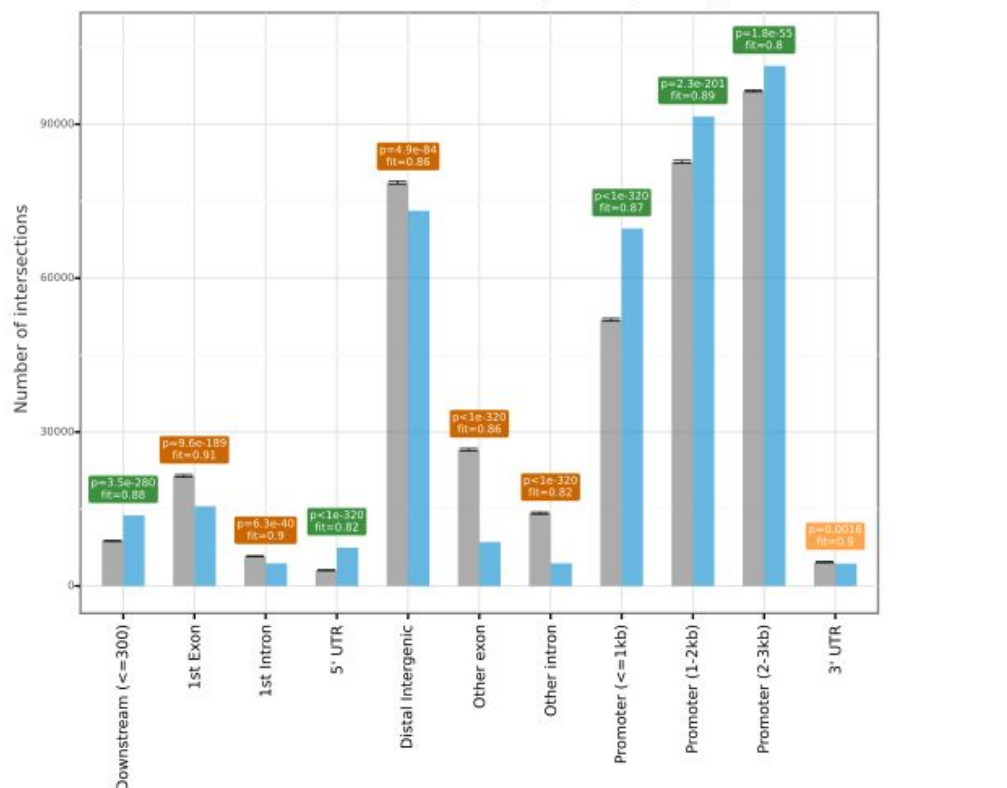
A

Total overlap length per region type

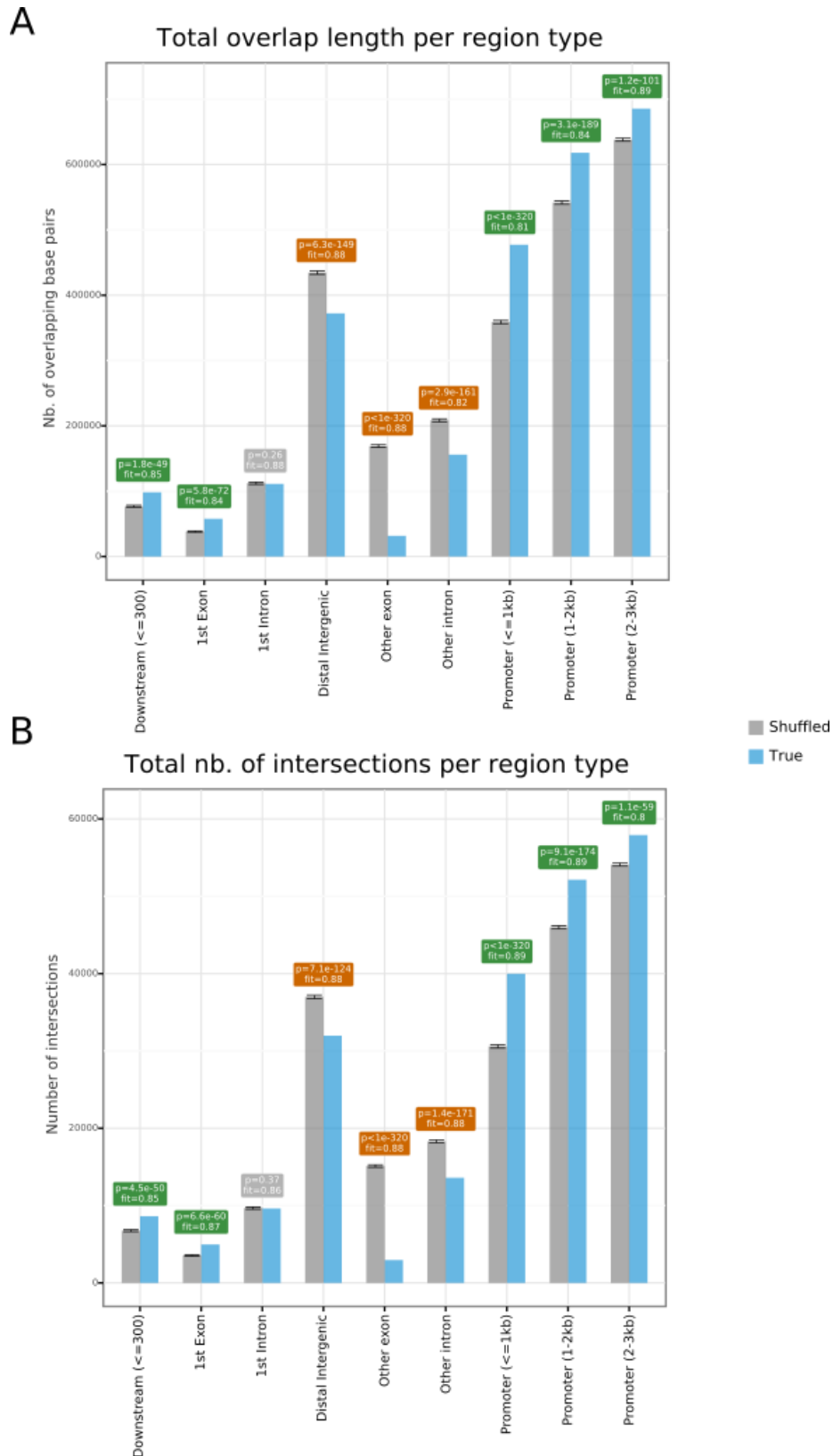


B

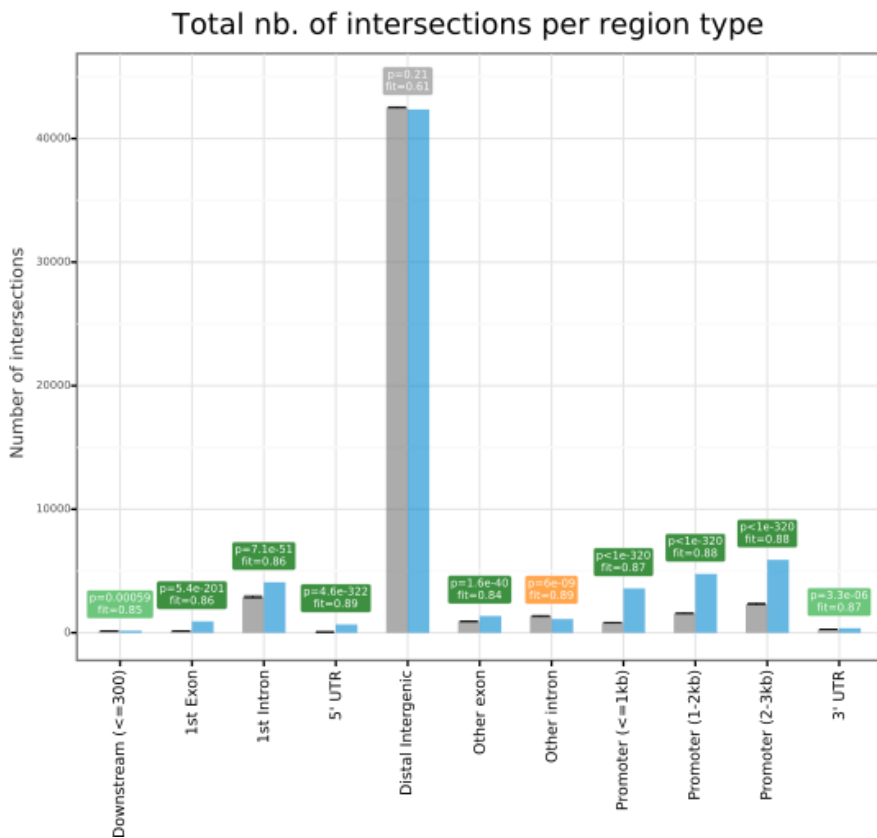
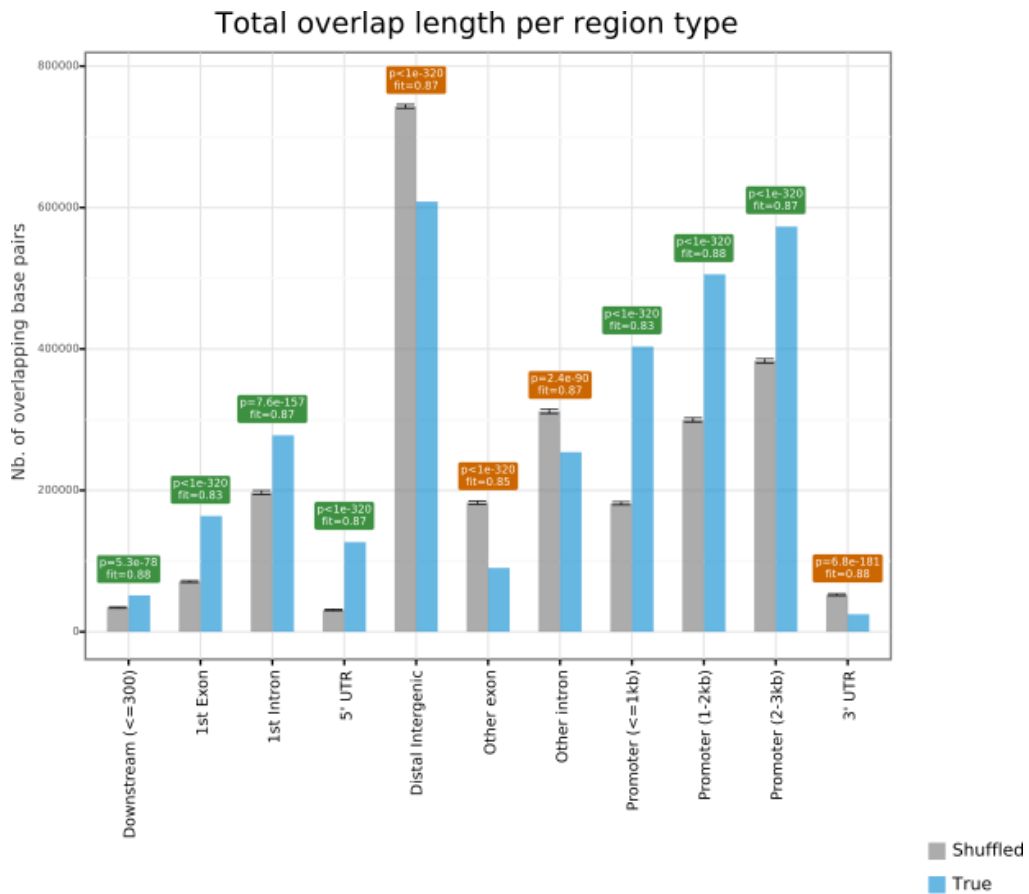
Total nb. of intersections per region type



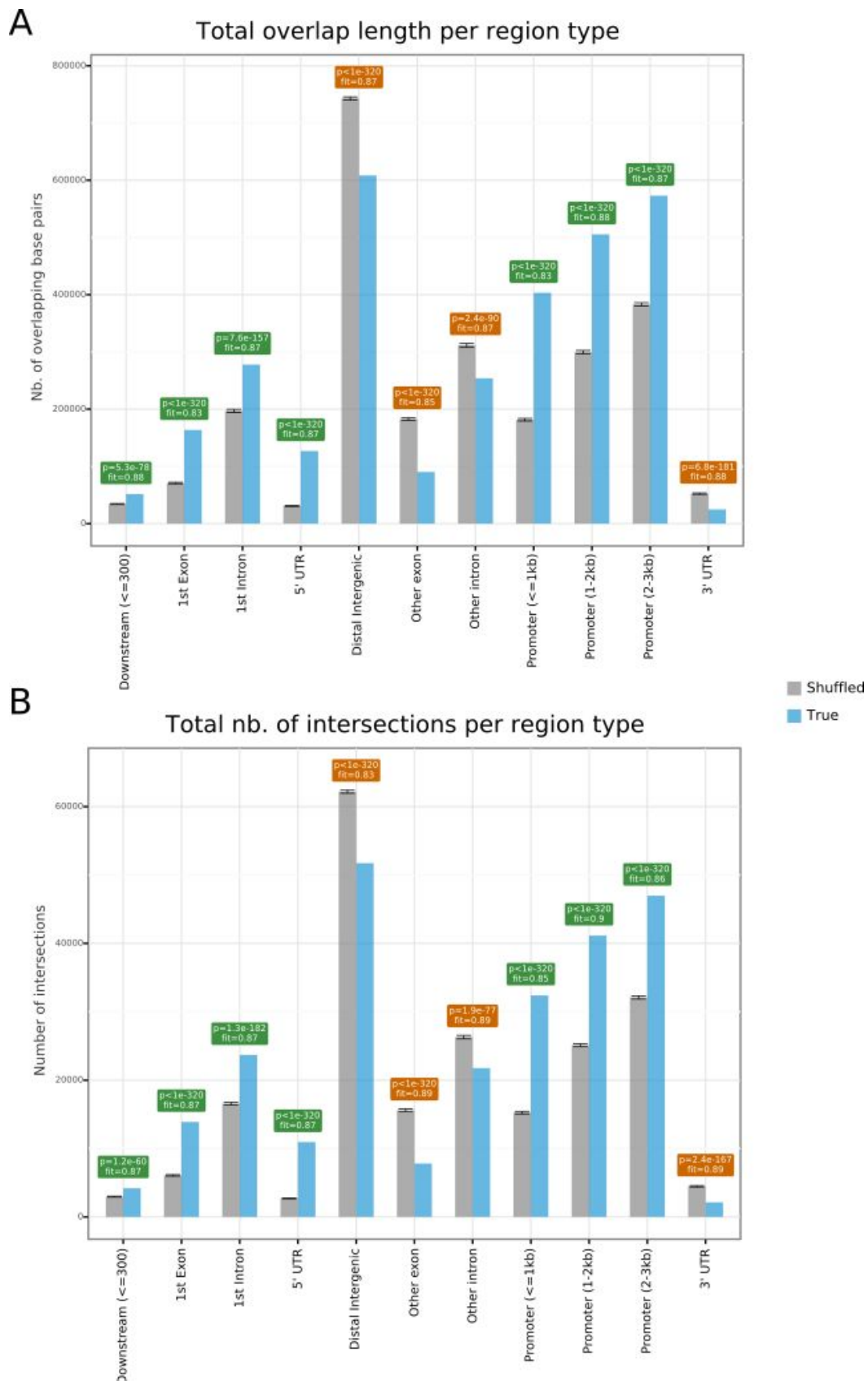
Supplementary Figure 4. Enrichment analysis for *A. thaliana* TFBSs in genomic regions. Barplots representing the expected (grey bars) versus observed (blue bars) overlap lengths (A) or number of intersections (B) between *A. thaliana* TFBSs from the robust collection and genomic annotations (x-axis). The plots and computed p-values (green: enrichment; orange: depletion) were obtained using the OLOGRAM command of the GTF toolkit.



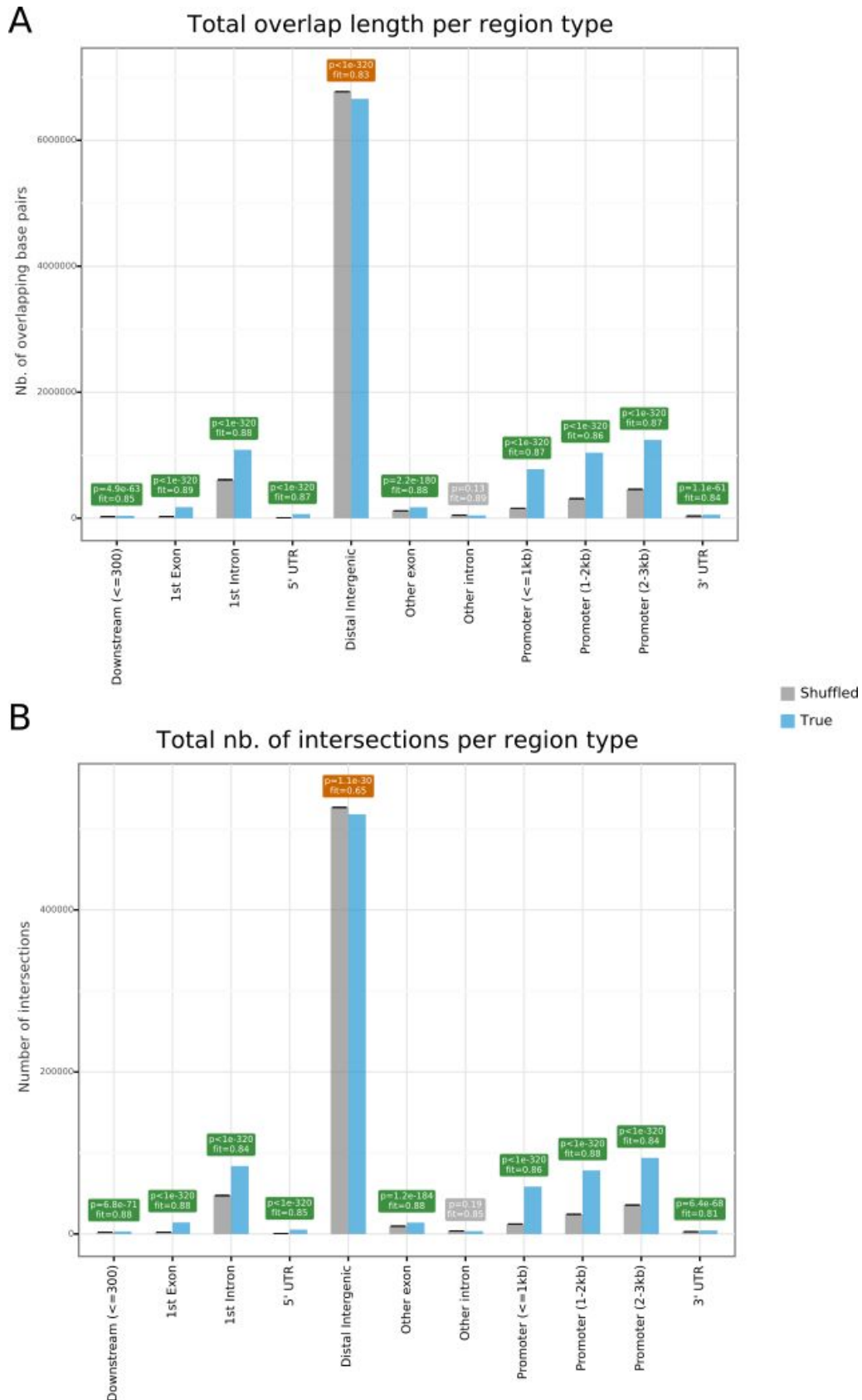
Supplementary Figure 5. Enrichment analysis for *C. elegans* TFBSs in genomic regions. Barplots representing the expected (grey bars) versus observed (blue bars) overlap lengths (A) or number of intersections (B) between *C. elegans* TFBSs from the robust collection and genomic annotations (x-axis). The plots and computed p-values (green: enrichment; orange: depletion) were obtained using the OLOGRAM command of the GTF toolkit.



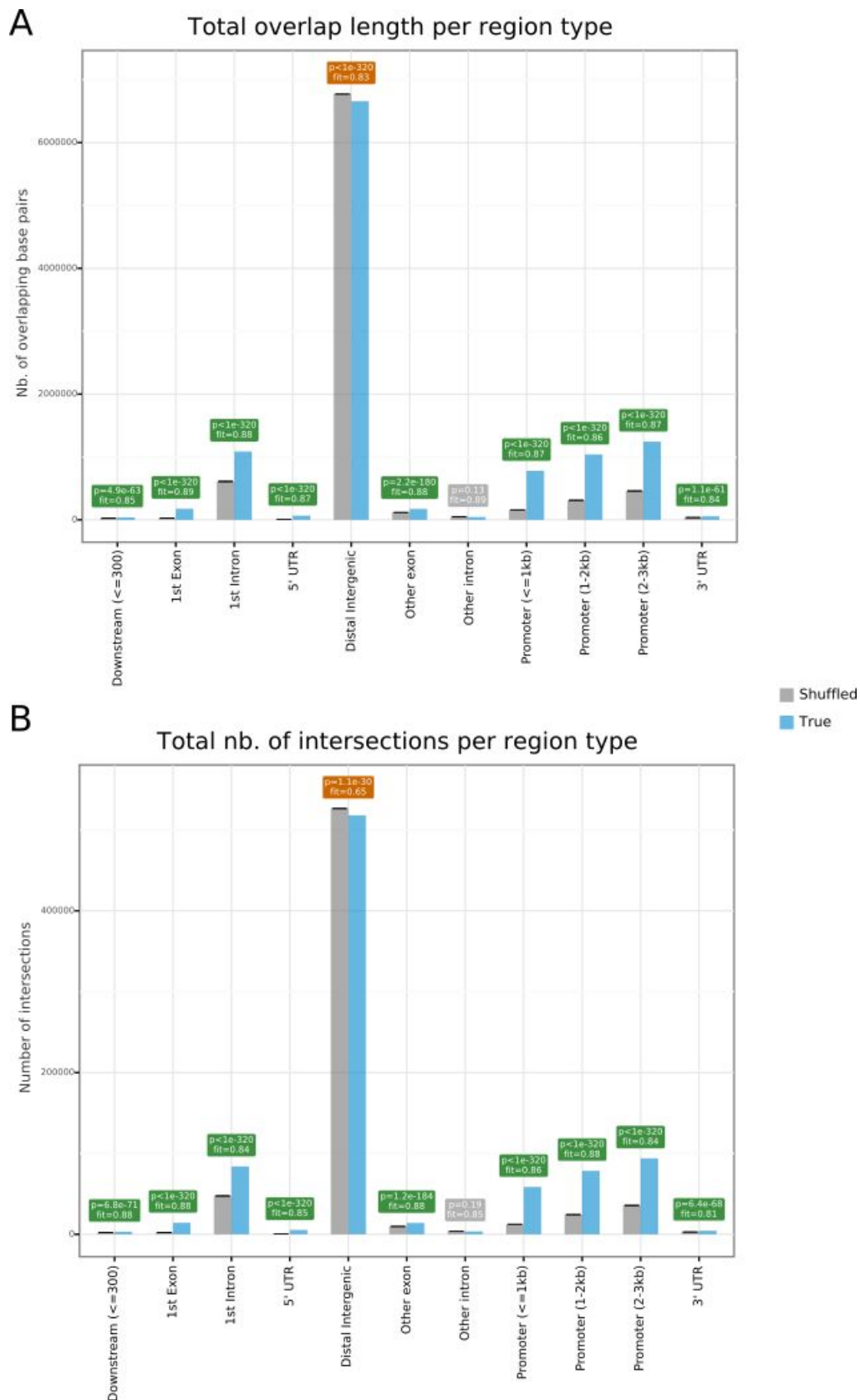
Supplementary Figure 6. Enrichment analysis for *D. rerio* TFBSs in genomic regions. Barplots representing the expected (grey bars) versus observed (blue bars) overlap lengths (A) or number of intersections (B) between *D. rerio* TFBSs from the robust collection and genomic annotations (x-axis). The plots and computed p-values (green: enrichment; orange: depletion) were obtained using the OLOGRAM command of the GTF toolkit.



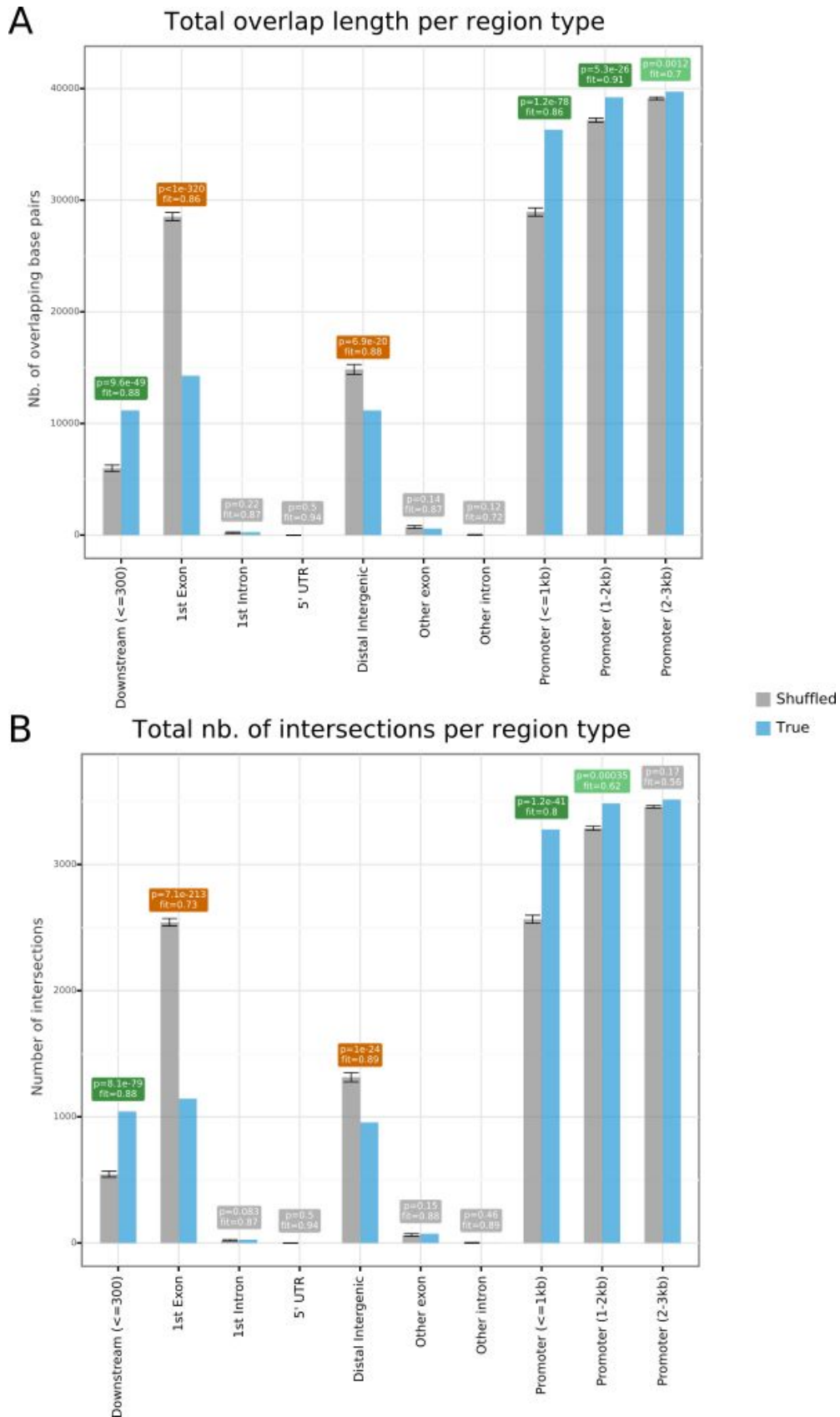
Supplementary Figure 7. Enrichment analysis for *D. melanogaster* TFBSs in genomic regions. Barplots representing the expected (grey bars) versus observed (blue bars) overlap lengths (A) or number of intersections (B) between *D. melanogaster* TFBSs from the robust collection and genomic annotations (x-axis). The plots and computed p-values (green: enrichment; orange: depletion) were obtained using the OLOGRAM command of the GTF toolkit.



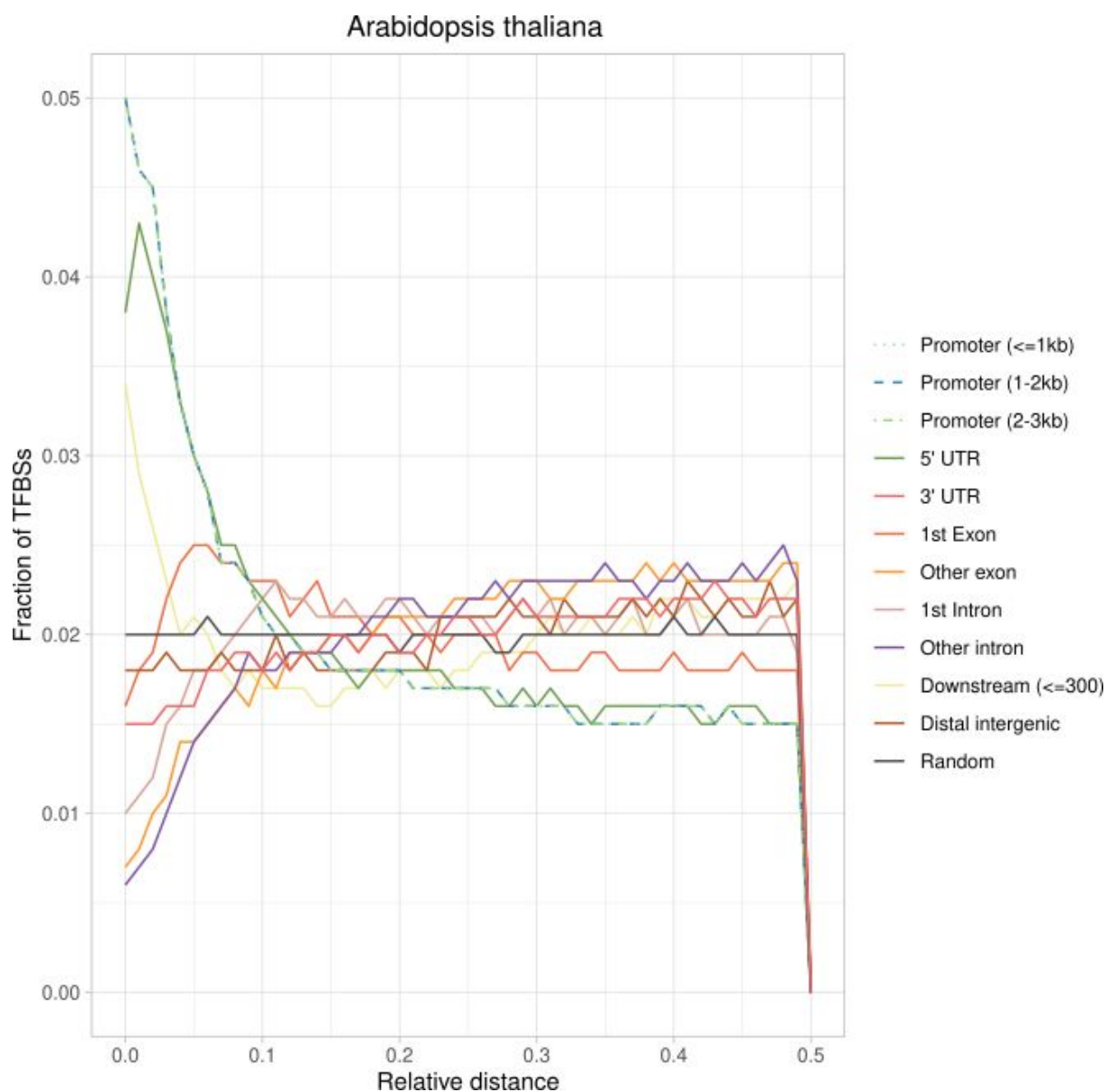
Supplementary Figure 8. Enrichment analysis for *M. musculus* TFBSs in genomic regions. Barplots representing the expected (grey bars) versus observed (blue bars) overlap lengths (A) or number of intersections (B) between *M. musculus* TFBSs from the robust collection and genomic annotations (x-axis). The plots and computed p-values (green: enrichment; orange: depletion) were obtained using the OLOGRAM command of the GTF toolkit.



Supplementary Figure 9. Enrichment analysis for *R. norvegicus* TFBSs in genomic regions. Barplots representing the expected (grey bars) versus observed (blue bars) overlap lengths (A) or number of intersections (B) between *R. norvegicus* TFBSs from the robust collection and genomic annotations (x-axis). The plots and computed p-values (green: enrichment; orange: depletion) were obtained using the OLOGRAM command of the GTF toolkit.

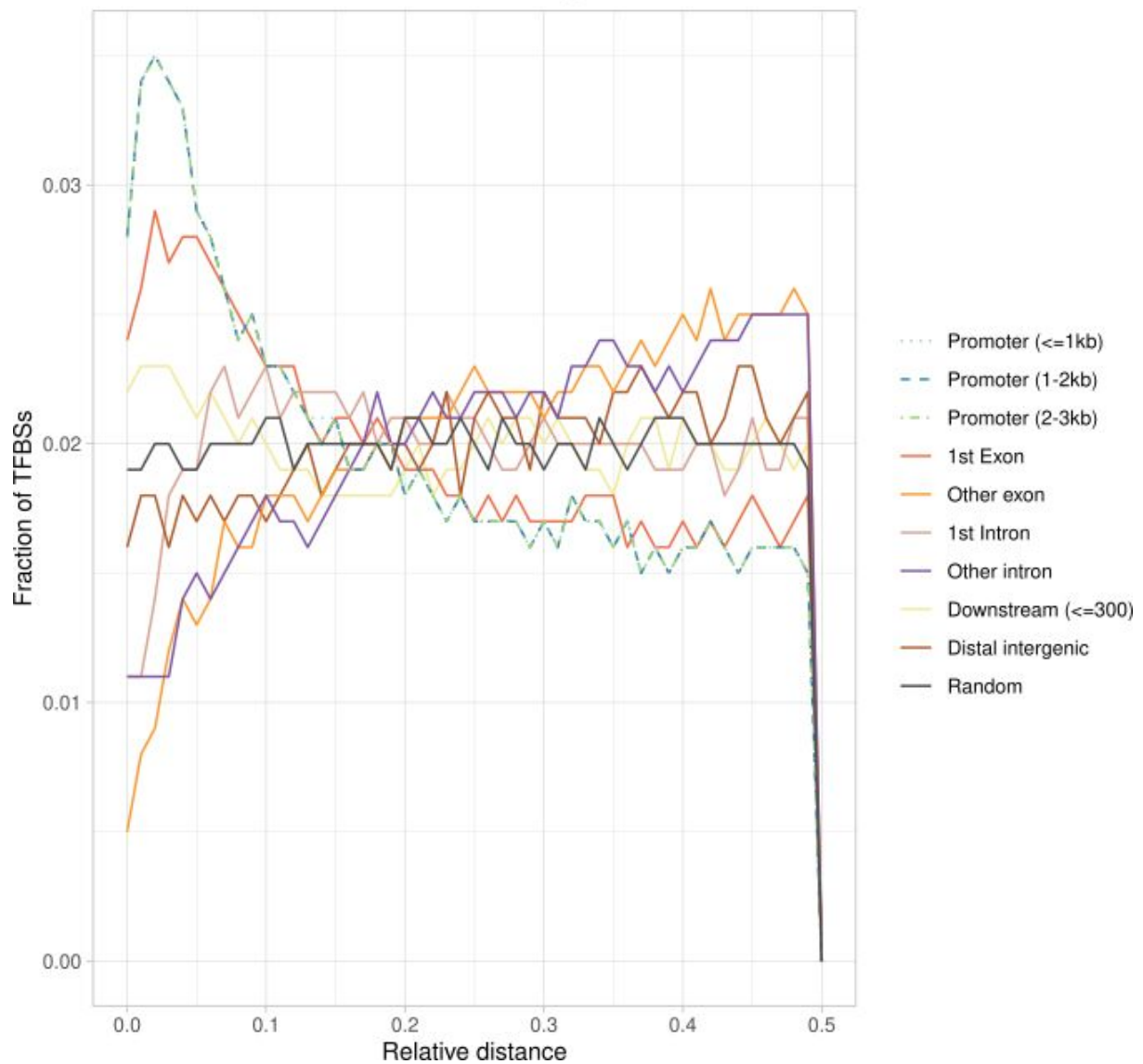


Supplementary Figure 10. Enrichment analysis for *S. cerevisiae* TFBSs in genomic regions. Barplots representing the expected (grey bars) versus observed (blue bars) overlap lengths (A) or number of intersections (B) between *S. cerevisiae* TFBSs from the robust collection and genomic annotations (x-axis). The plots and computed p-values (green: enrichment; orange: depletion) were obtained using the OLOGRAM command of the GTF toolkit.

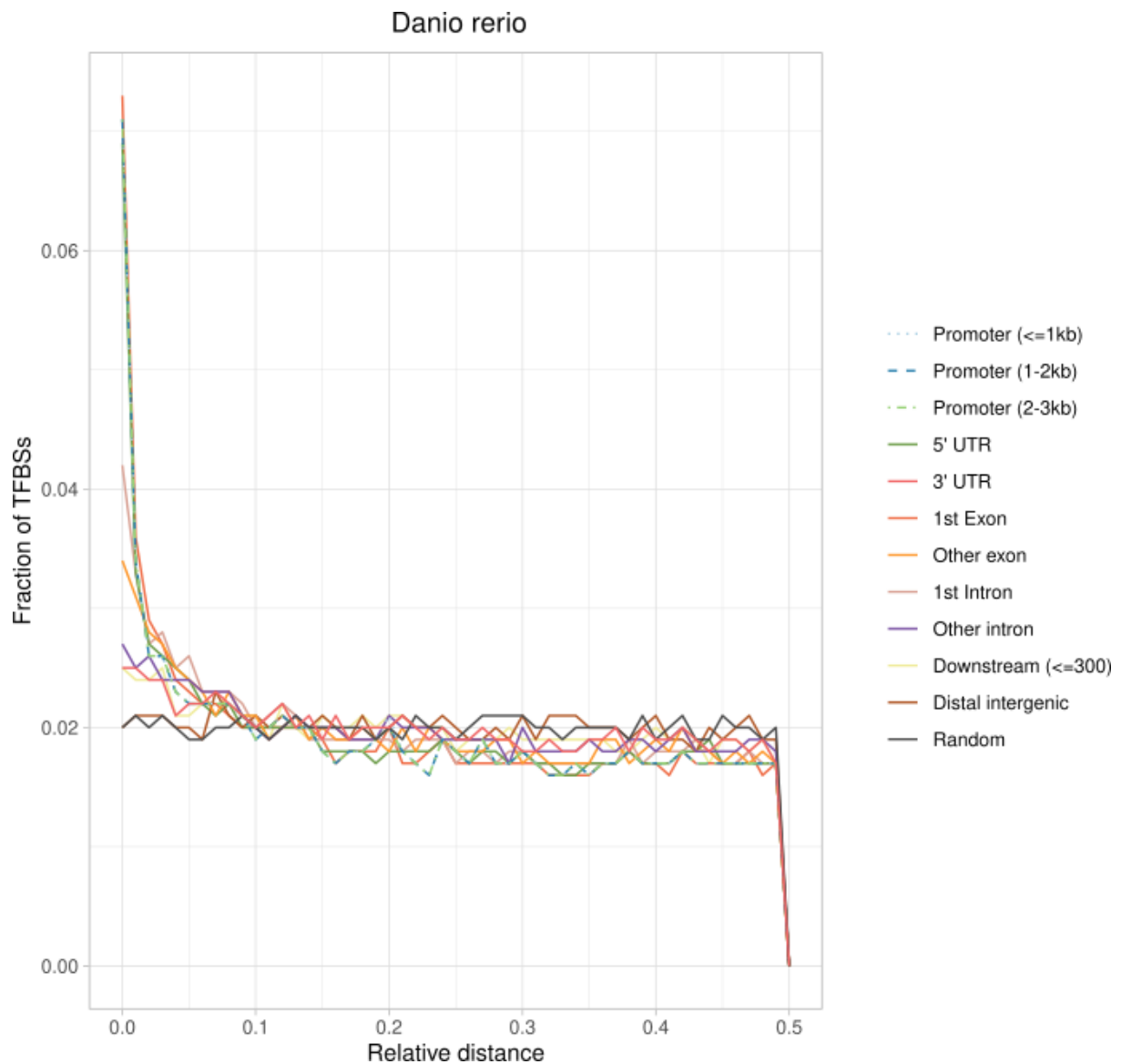


Supplementary Figure 11. Analysis of the overlap of TFBSs with respect to genomic annotations in *Arabidopsis thaliana*. Fraction of TFBSs in the UniBind robust collection (y-axis) with respect to increasing relative distances (x-axis) from different genomic regions computed using the *bedtools reldist* command. When two genomic tracks are not spatially related, one expects the fraction of relative distance distribution to be uniform.

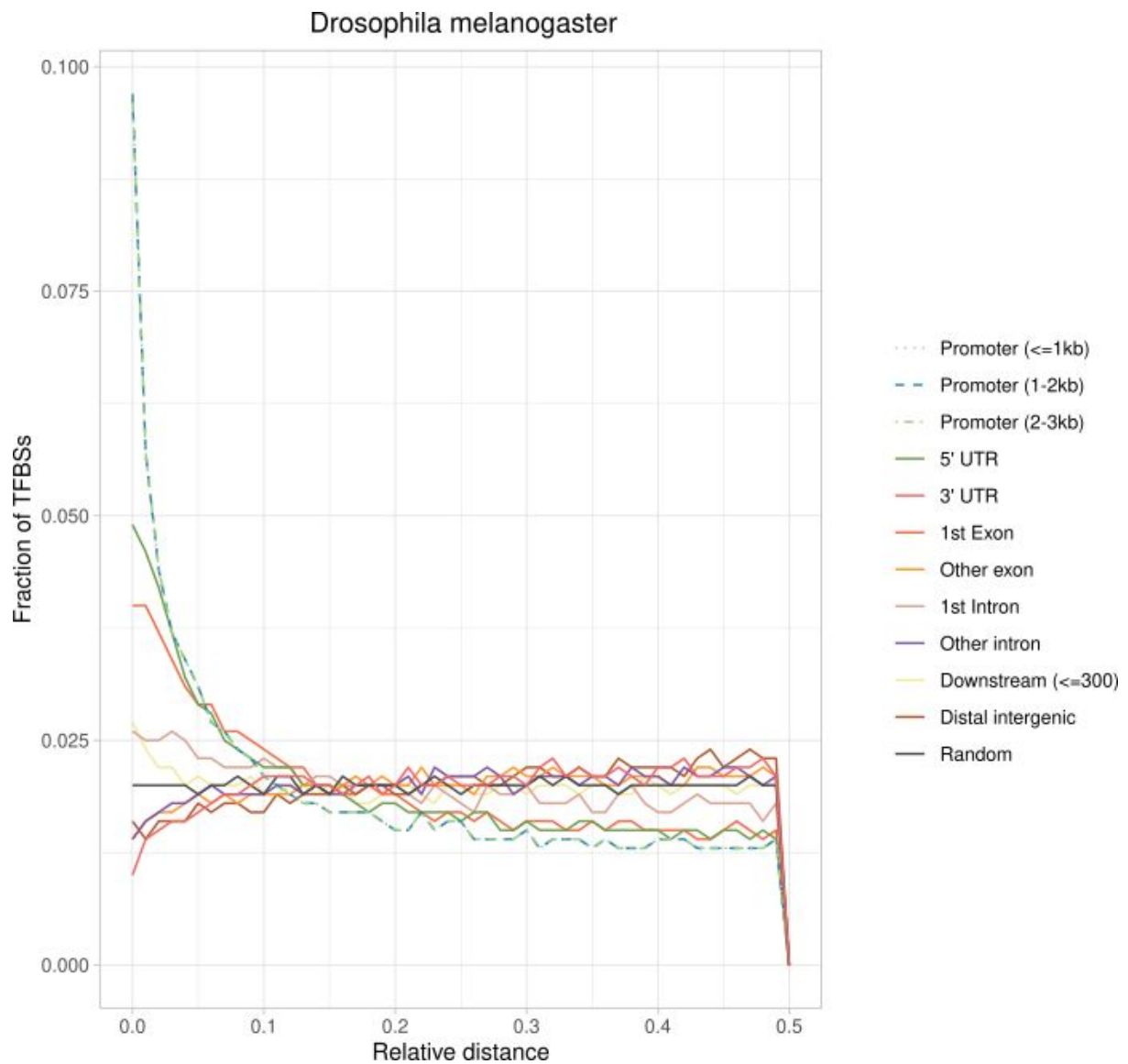
Caenorhabditis elegans



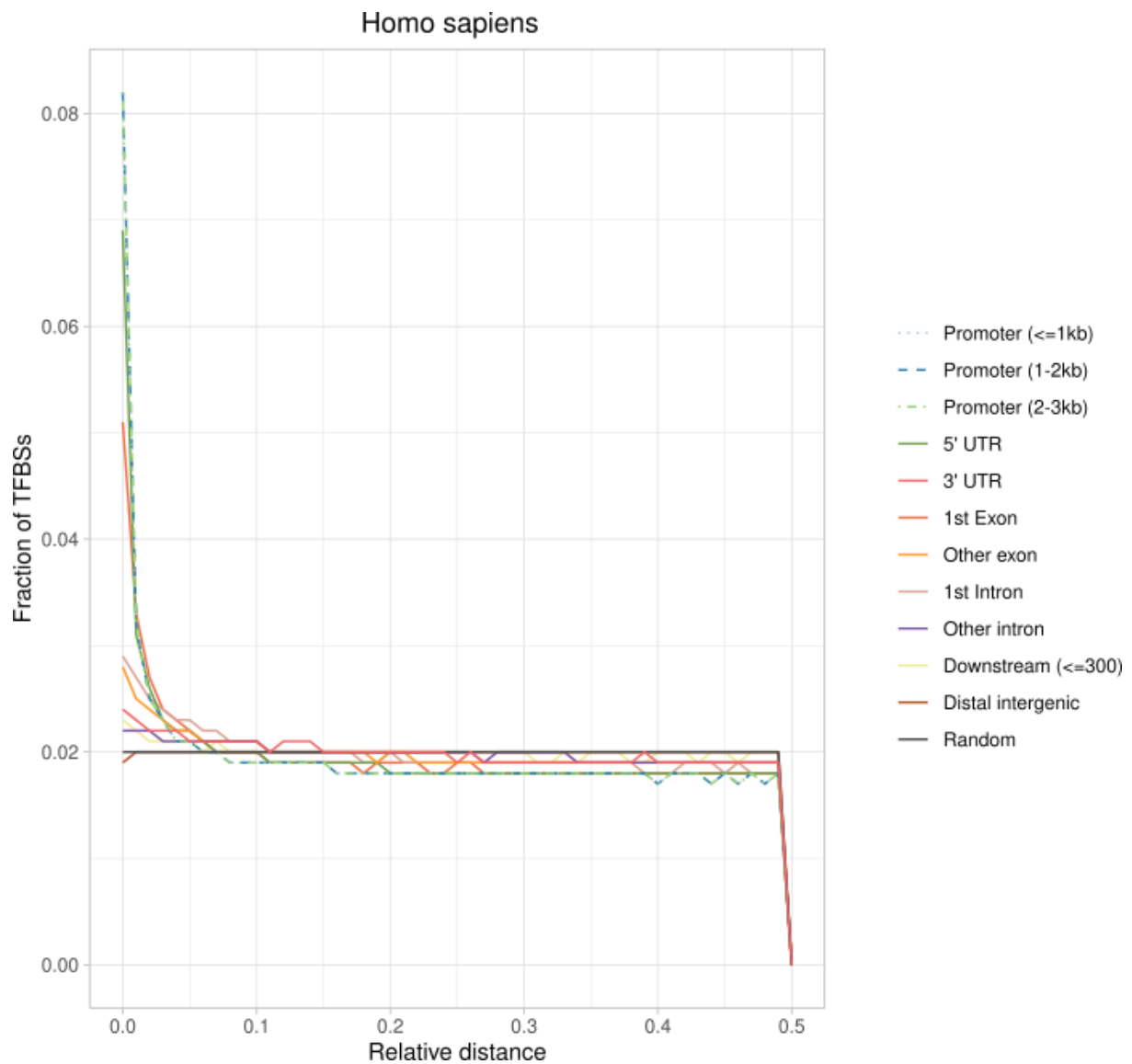
Supplementary Figure 12. Analysis of the overlap of TFBSs with respect to genomic annotations in *Caenorhabditis elegans*. Fraction of TFBSs in the UniBind robust collection (y-axis) with respect to increasing relative distances (x-axis) from different genomic regions computed using the *bedtools reldist* command. When two genomic tracks are not spatially related, one expects the fraction of relative distance distribution to be uniform.



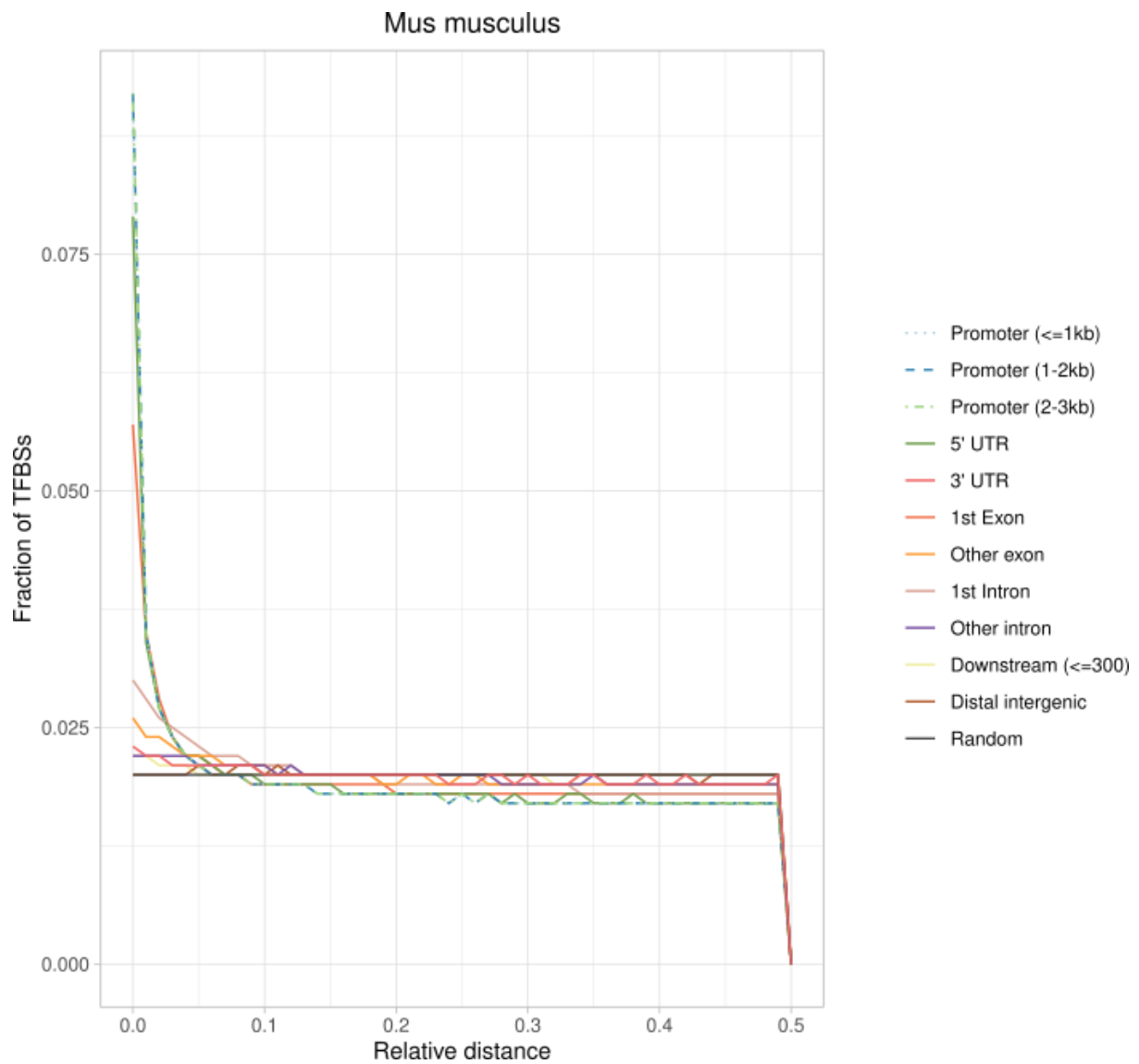
Supplementary Figure 13. Analysis of the overlap of TFBSs with respect to genomic annotations in *Danio rerio*. Fraction of TFBSs in the UniBind robust collection (y-axis) with respect to increasing relative distances (x-axis) from different genomic regions computed using the *bedtools reldist* command. When two genomic tracks are not spatially related, one expects the fraction of relative distance distribution to be uniform.



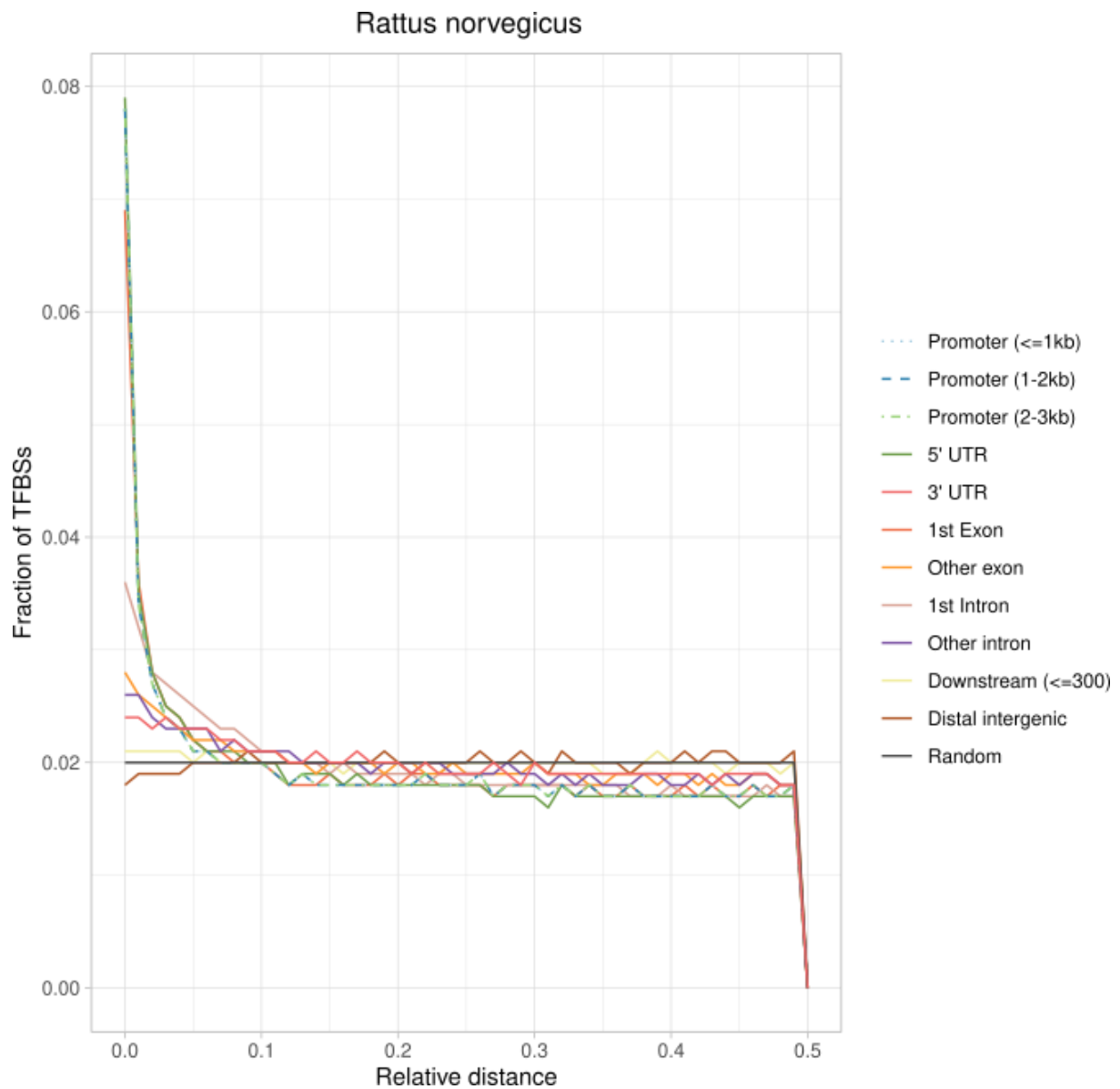
Supplementary Figure 14. Analysis of the overlap of TFBSs with respect to genomic annotations in *Drosophila melanogaster*. Fraction of TFBSs in the UniBind robust collection (y-axis) with respect to increasing relative distances (x-axis) from different genomic regions computed using the *bedtools reldist* command. When two genomic tracks are not spatially related, one expects the fraction of relative distance distribution to be uniform.



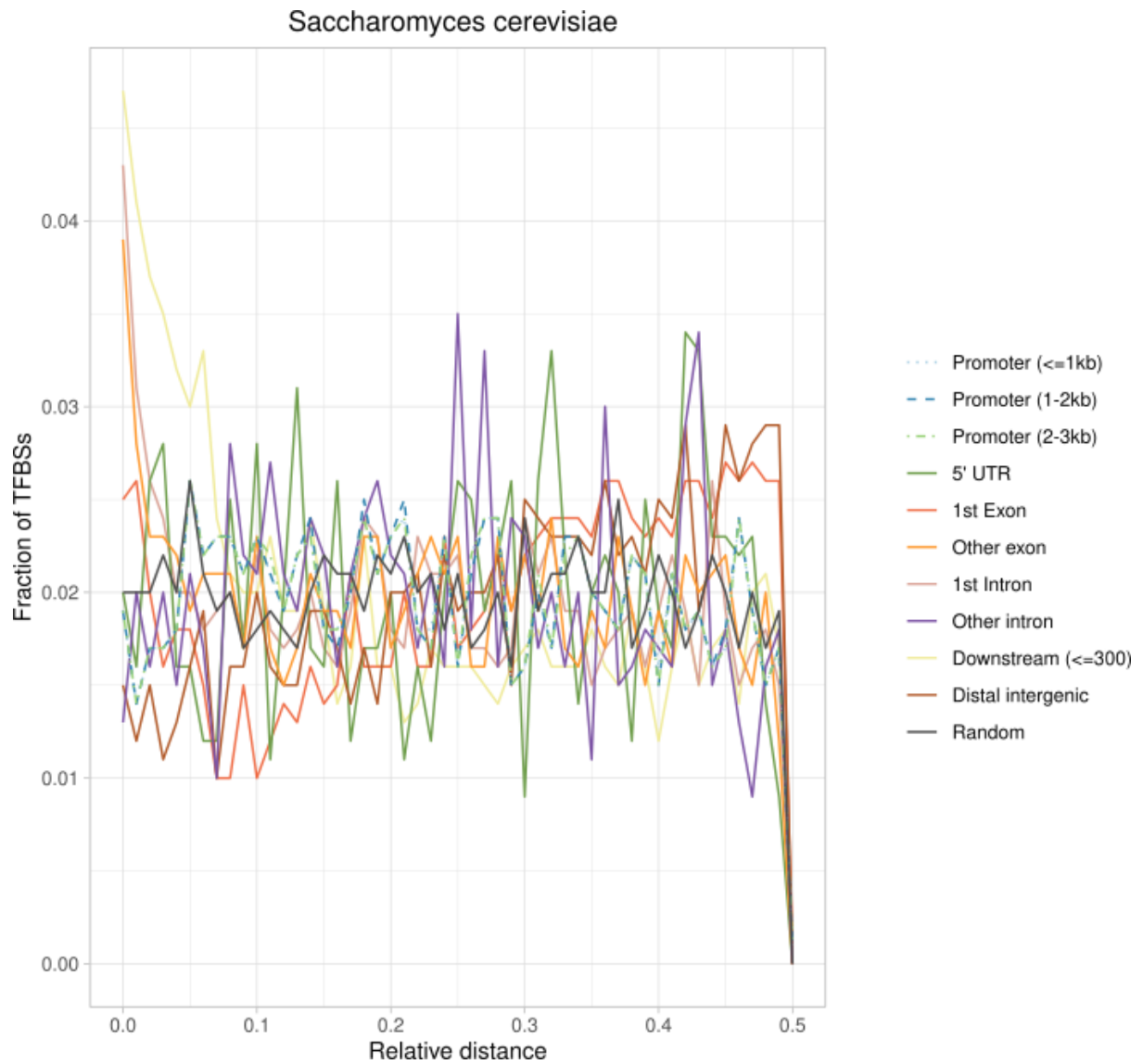
Supplementary Figure 15. Analysis of the overlap of TFBSs with respect to genomic annotations in *Homo sapiens*. Fraction of TFBSs in the UniBind robust collection (y-axis) with respect to increasing relative distances (x-axis) from different genomic regions computed using the *bedtools reldist* command. When two genomic tracks are not spatially related, one expects the fraction of relative distance distribution to be uniform.



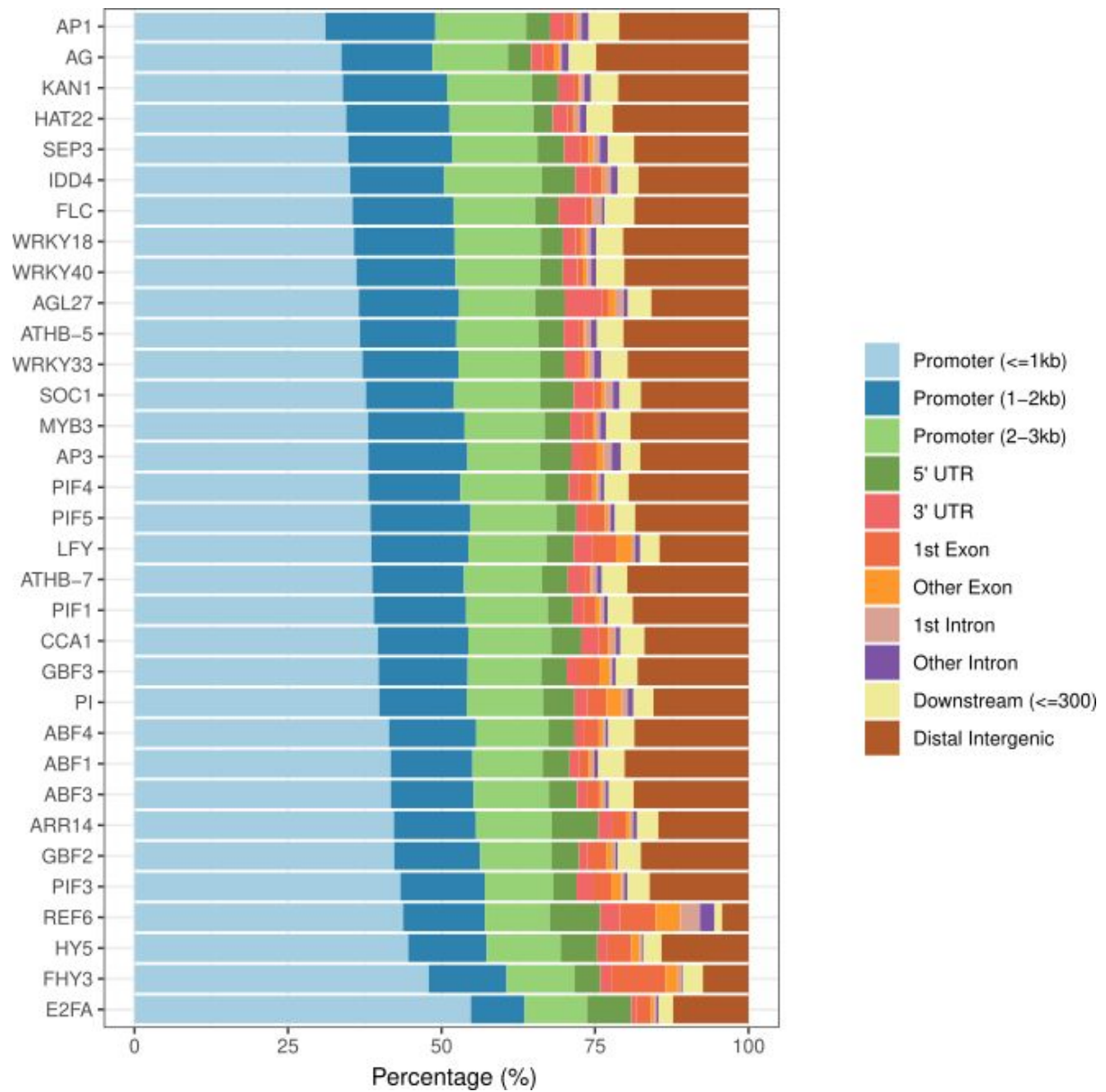
Supplementary Figure 16. Analysis of the overlap of TFBSs with respect to genomic annotations in *Mus musculus*. Fraction of TFBSs in the UniBind robust collection (y-axis) with respect to increasing relative distances (x-axis) from different genomic regions computed using the *bedtools reldist* command. When two genomic tracks are not spatially related, one expects the fraction of relative distance distribution to be uniform.



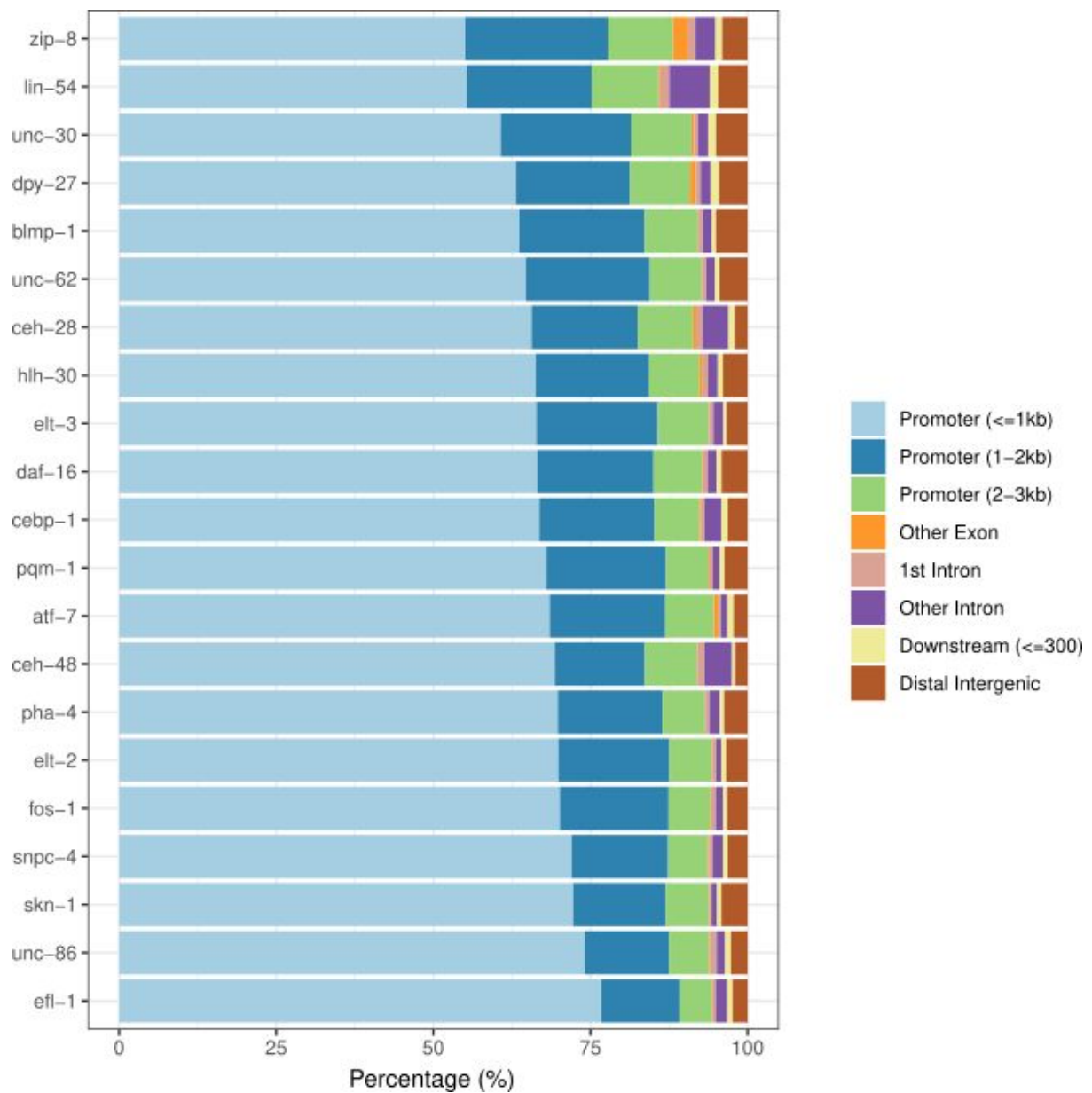
Supplementary Figure 17. Analysis of the overlap of TFBSs with respect to genomic annotations in *Rattus norvegicus*. Fraction of TFBSs in the UniBind robust collection (y-axis) with respect to increasing relative distances (x-axis) from different genomic regions computed using the *bedtools reldist* command. When two genomic tracks are not spatially related, one expects the fraction of relative distance distribution to be uniform.



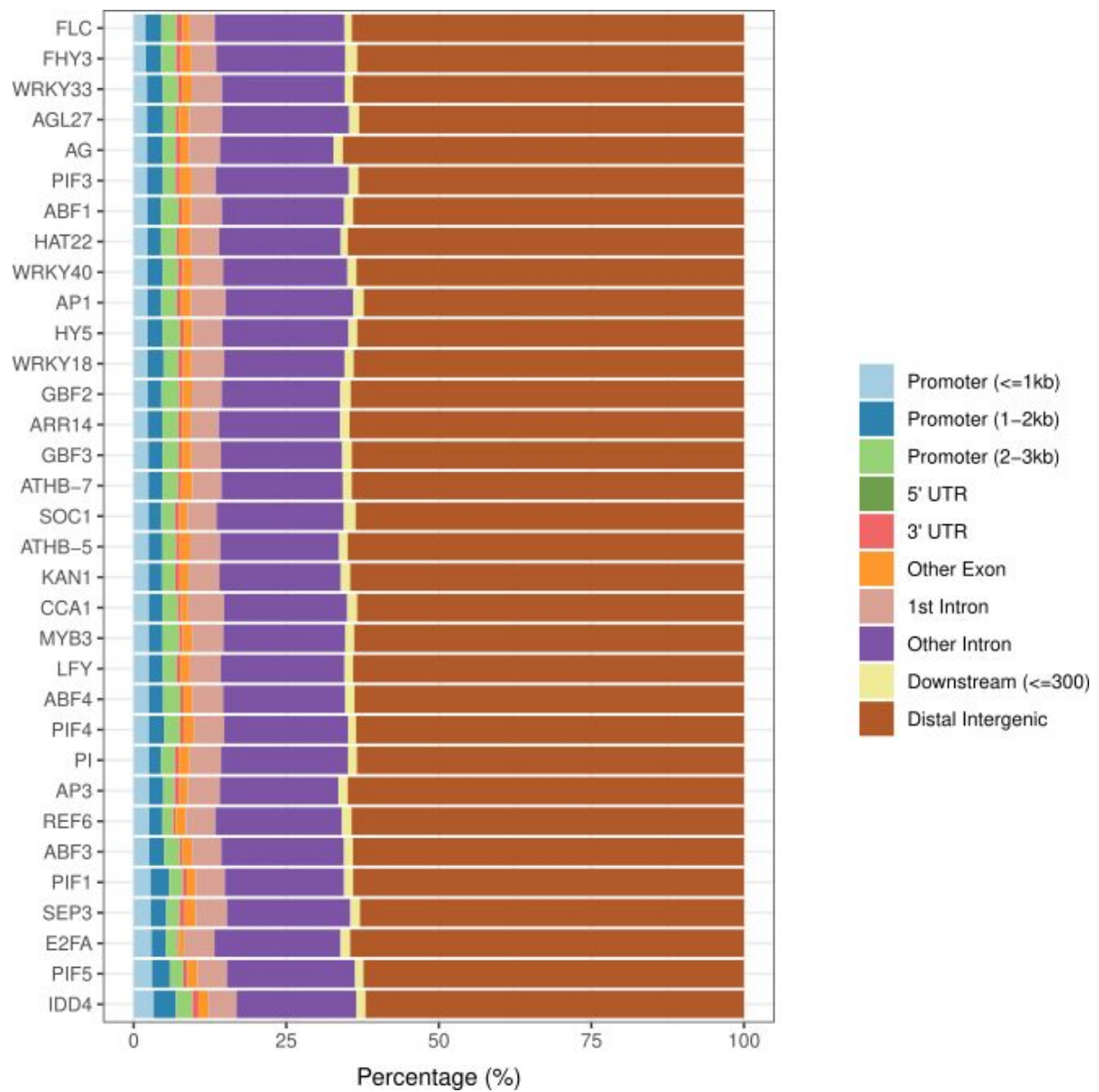
Supplementary Figure 18. Analysis of the overlap of TFBSs with respect to genomic annotations in *Saccharomyces cerevisiae*. Fraction of TFBSs in the UniBind robust collection (y-axis) with respect to increasing relative distances (x-axis) from different genomic regions computed using the *bedtools reldist* command. When two genomic tracks are not spatially related, one expects the fraction of relative distance distribution to be uniform.



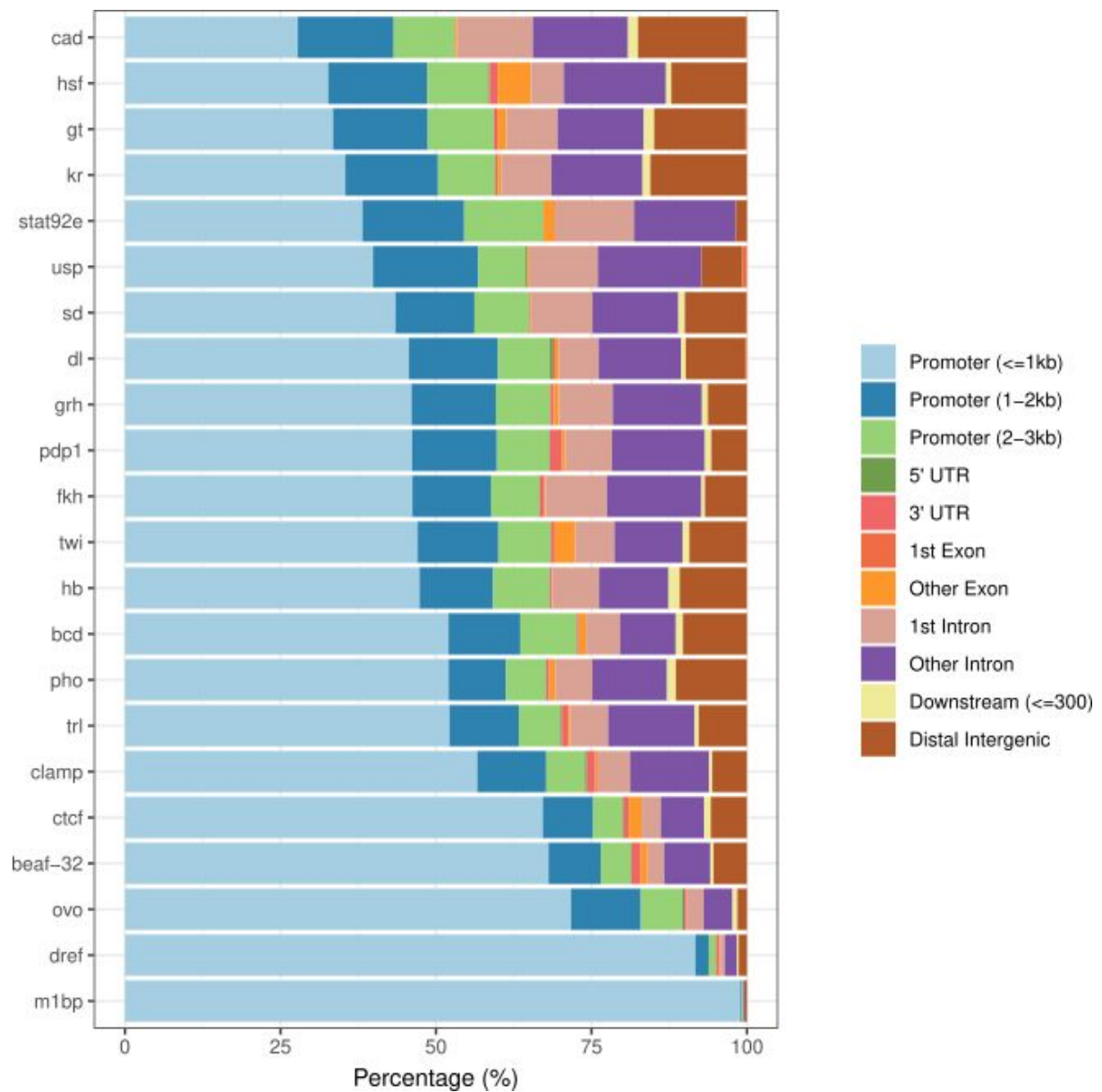
Supplementary Figure 19. Genomic distribution of *A. thaliana* TFBSs. Distribution of the proportion of *A. thaliana* UniBind robust TFBSs overlapping with different types of genomic regions (columns; see legend) across TFs (rows).



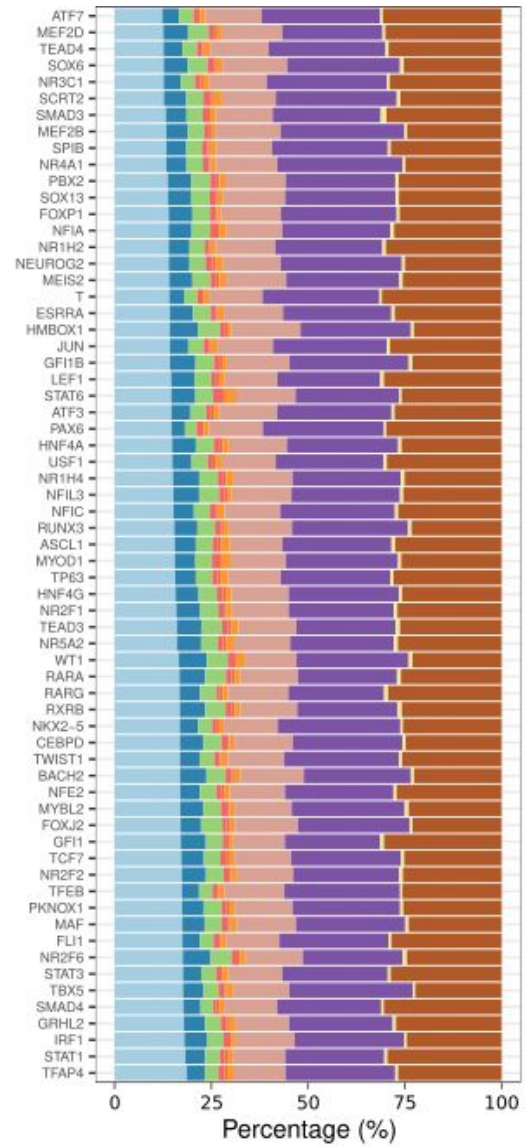
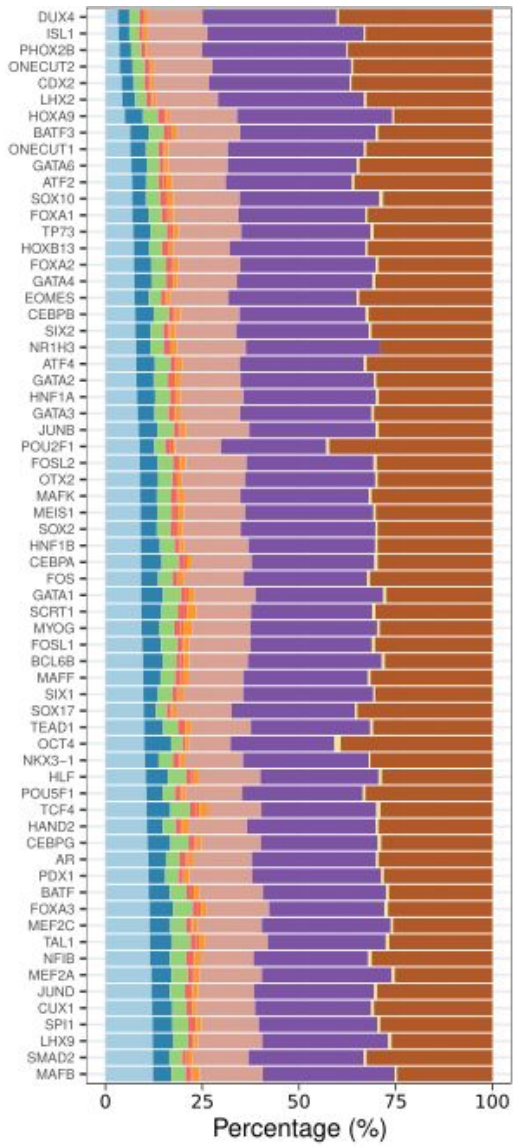
Supplementary Figure 20. Genomic distribution of *C. elegans* TFBSs. Distribution of the proportion of *C. elegans* UniBind robust TFBSs overlapping with different types of genomic regions (columns; see legend) across TFs (rows).



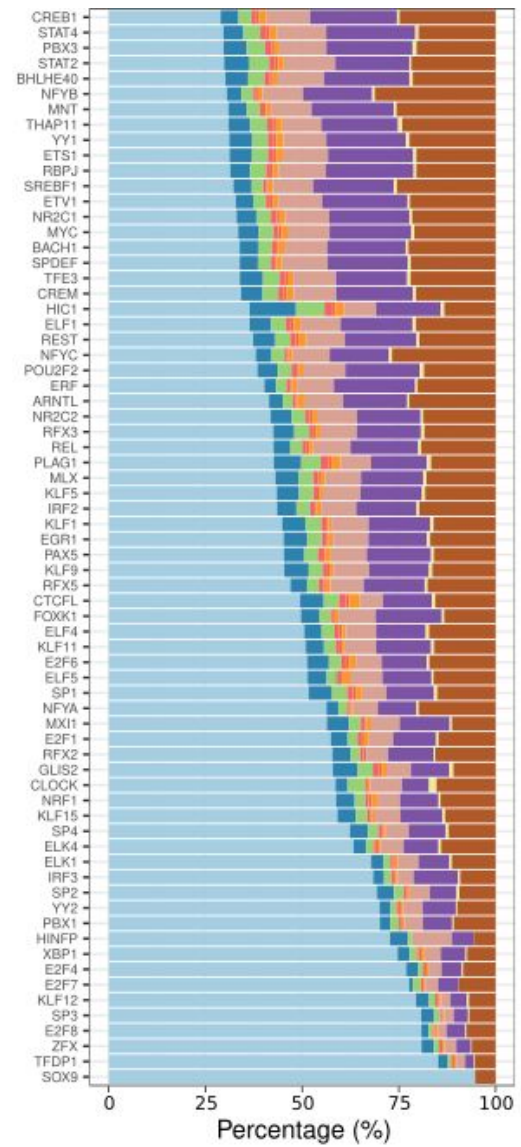
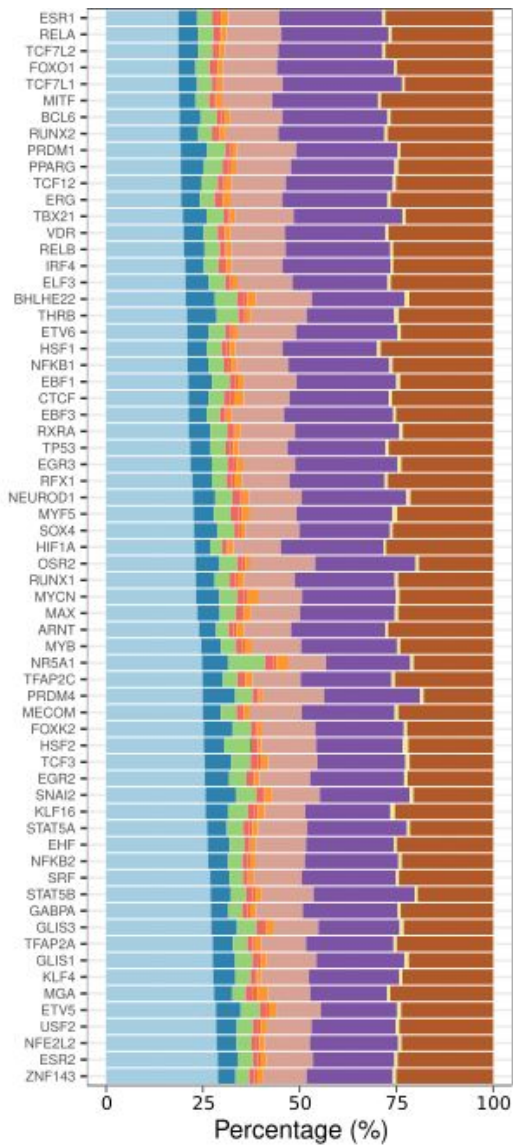
Supplementary Figure 21. Genomic distribution of *D. rerio* TFBSs. Distribution of the proportion of *D. rerio* UniBind robust TFBSs overlapping with different types of genomic regions (columns; see legend) across TFs (rows).



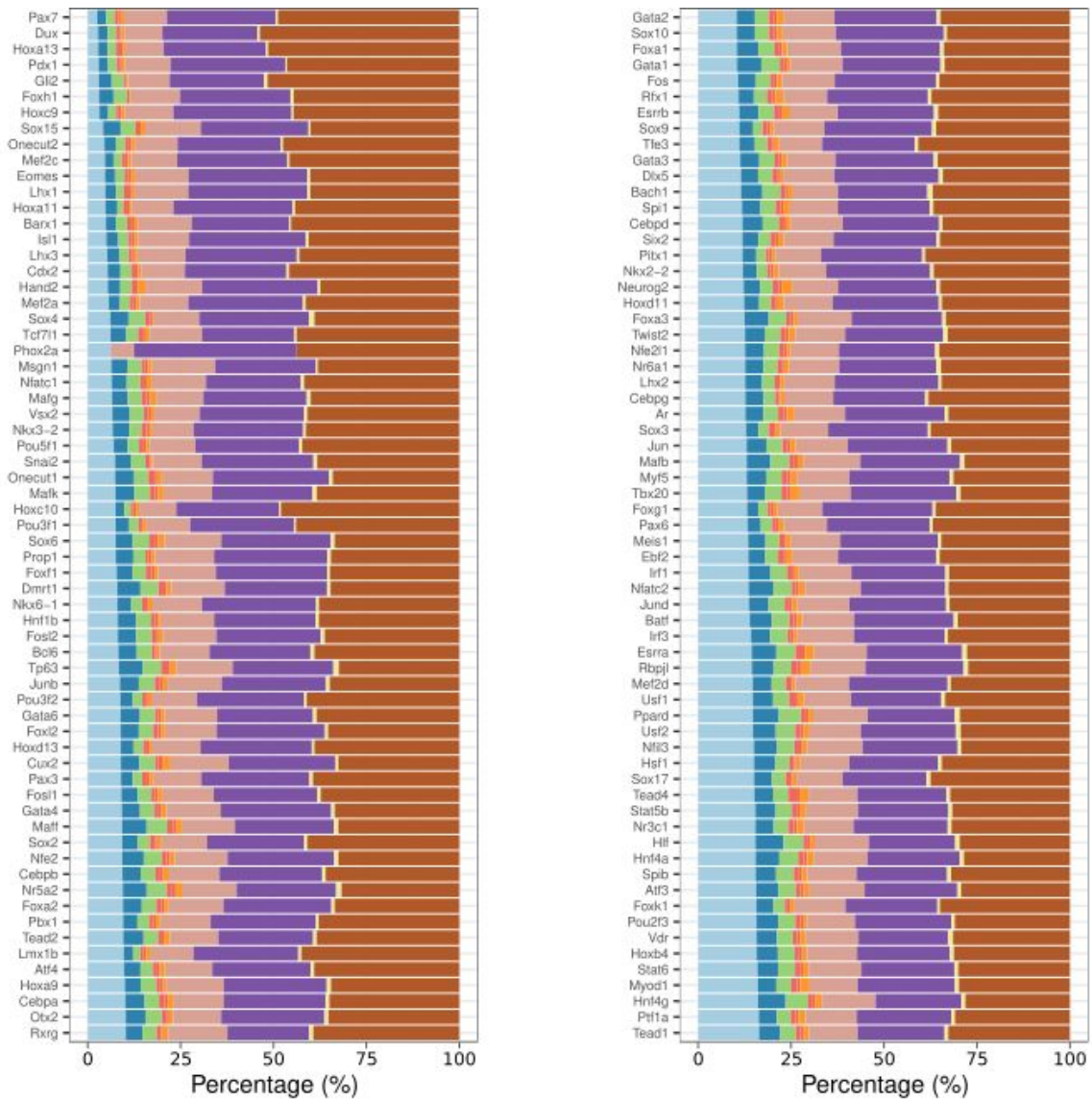
Supplementary Figure 22. Genomic distribution of *D. melanogaster* TFBSs. Distribution of the proportion of *D. melanogaster* UniBind robust TFBSs overlapping with different types of genomic regions (columns; see legend) across TFs (rows).



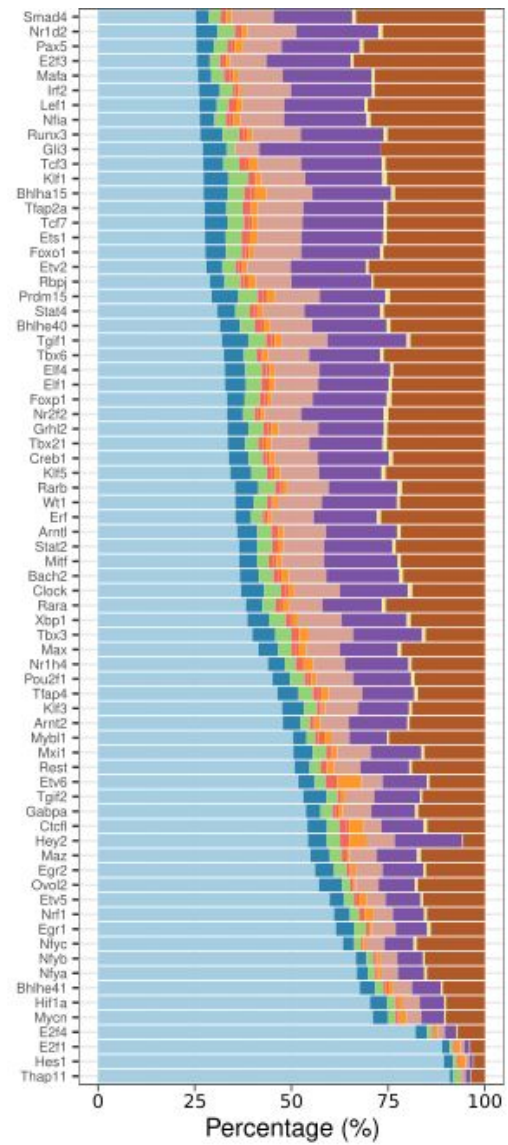
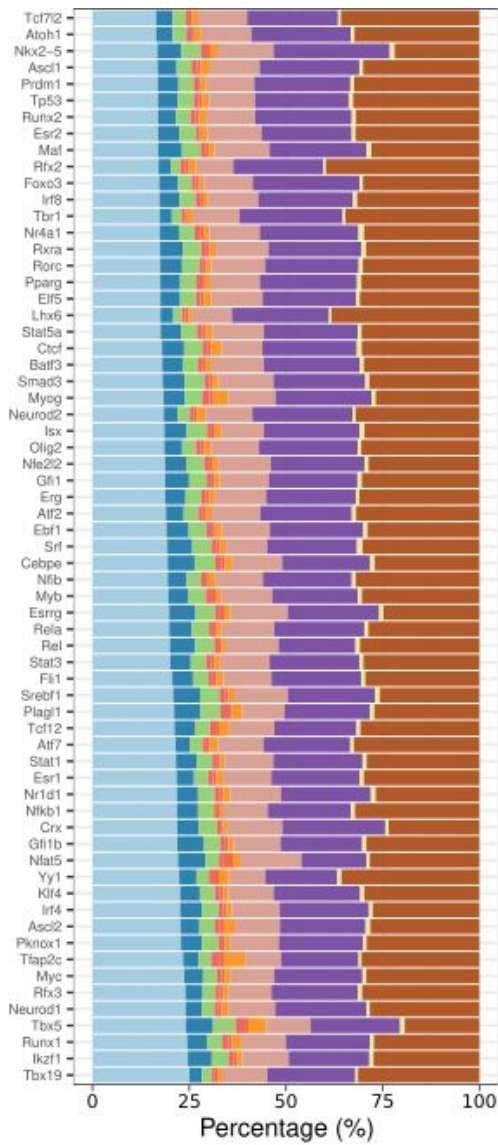
Supplementary Figure 23. Genomic distribution of *H. sapiens* TFBSs. Distribution of the proportion of *H. sapiens* UniBind robust TFBSs overlapping with different types of genomic regions (columns; see legend) across TFs (rows).



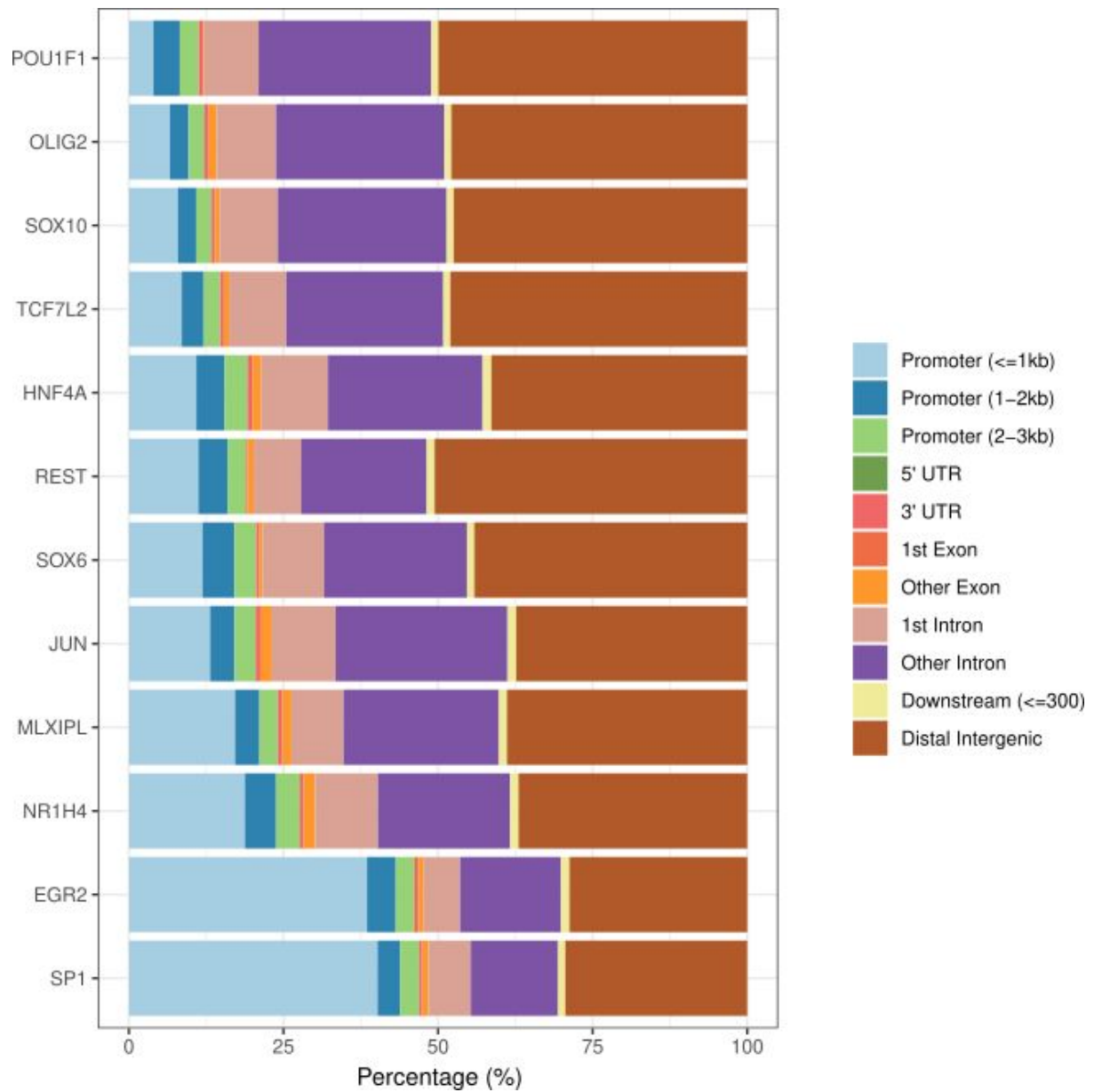
Supplementary Figure 23 (continued). Genomic distribution of *H. sapiens* TFBSs. Distribution of the proportion of *H. sapiens* UniBind robust TFBSs overlapping with different types of genomic regions (columns; see legend) across TFs (rows).



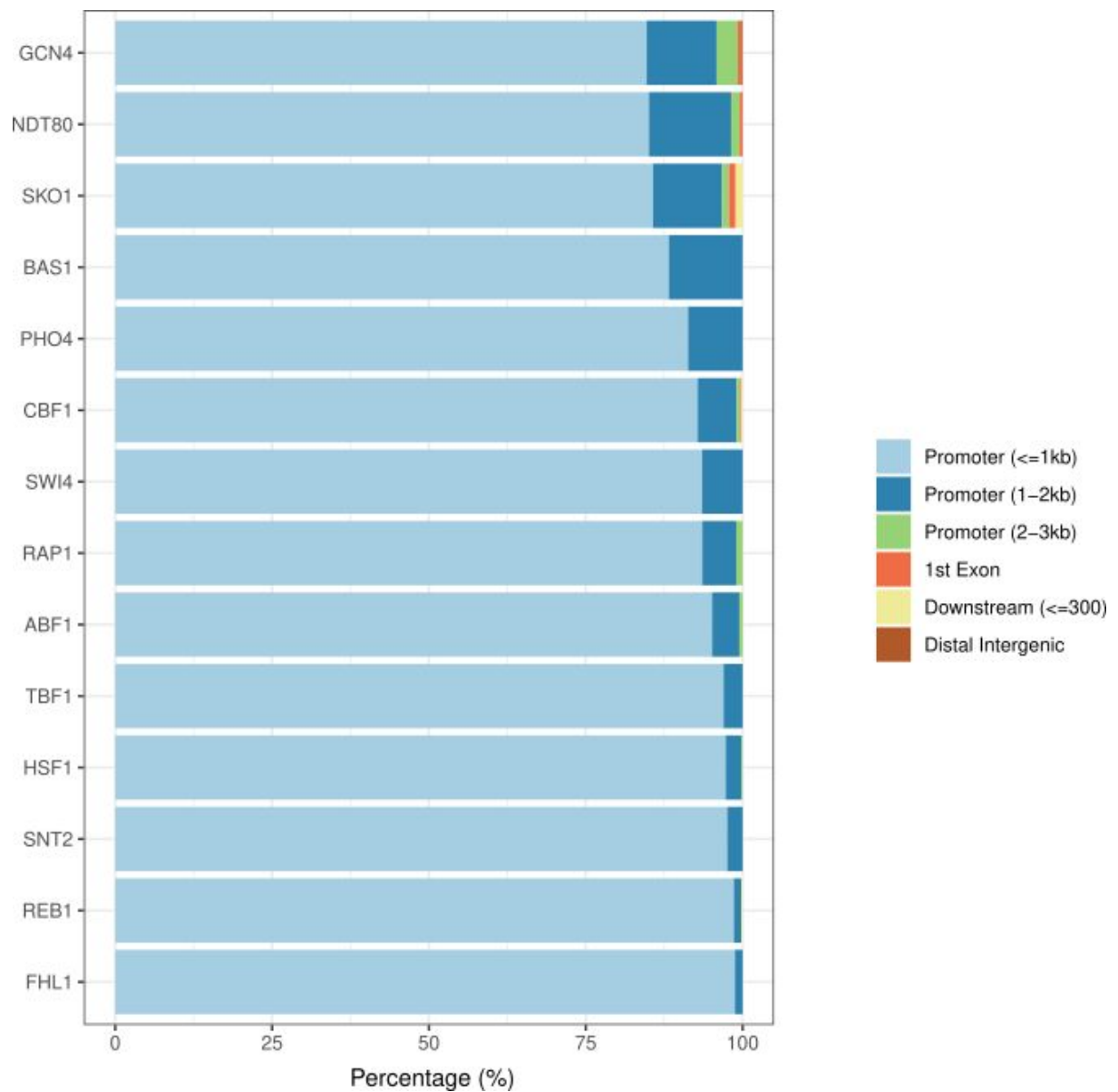
Supplementary Figure 24. Genomic distribution of *M. musculus* TFBSs. Distribution of the proportion of *M. musculus* UniBind robust TFBSs overlapping with different types of genomic regions (columns; see legend) across TFs (rows).



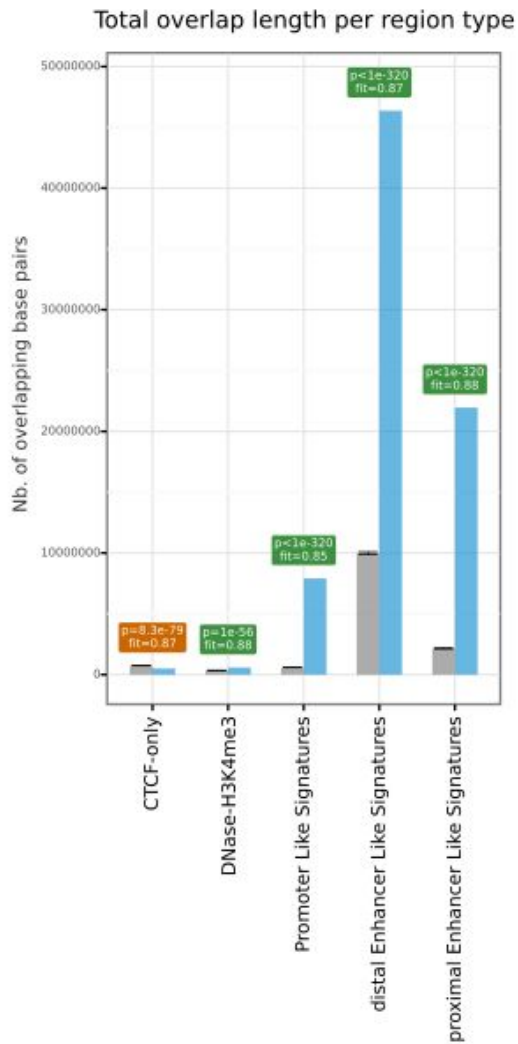
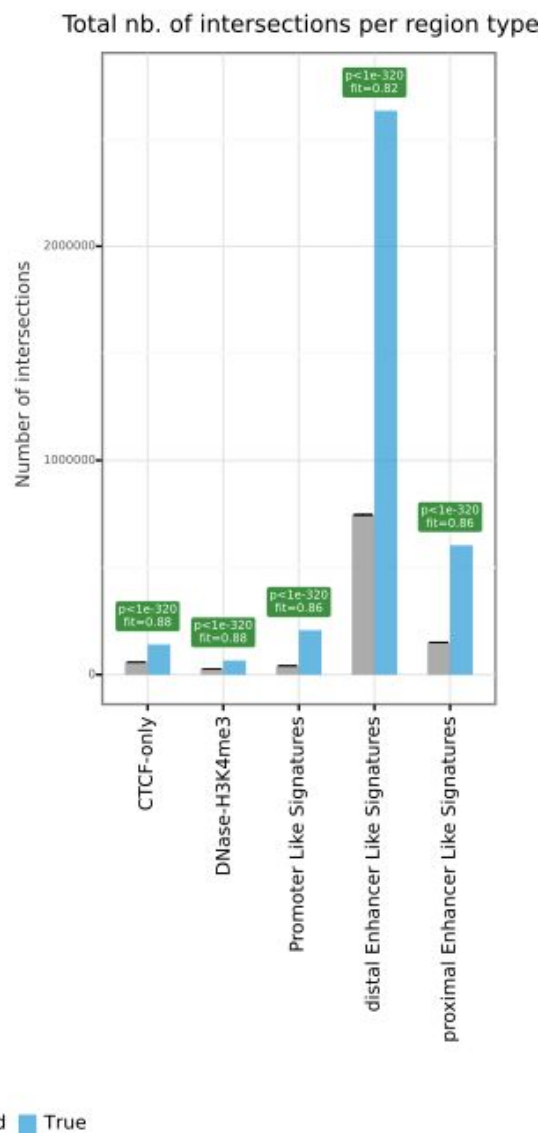
Supplementary Figure 24 (continued). Genomic distribution of *M. musculus* TFBSs. Distribution of the proportion of *M. musculus* UniBind robust TFBSs overlapping with different types of genomic regions (columns; see legend) across TFs (rows).



Supplementary Figure 25. Genomic distribution of *R. norvegicus* TFBSs. Distribution of the proportion of *R. norvegicus* UniBind robust TFBSs overlapping with different types of genomic regions (columns; see legend) across TFs (rows).

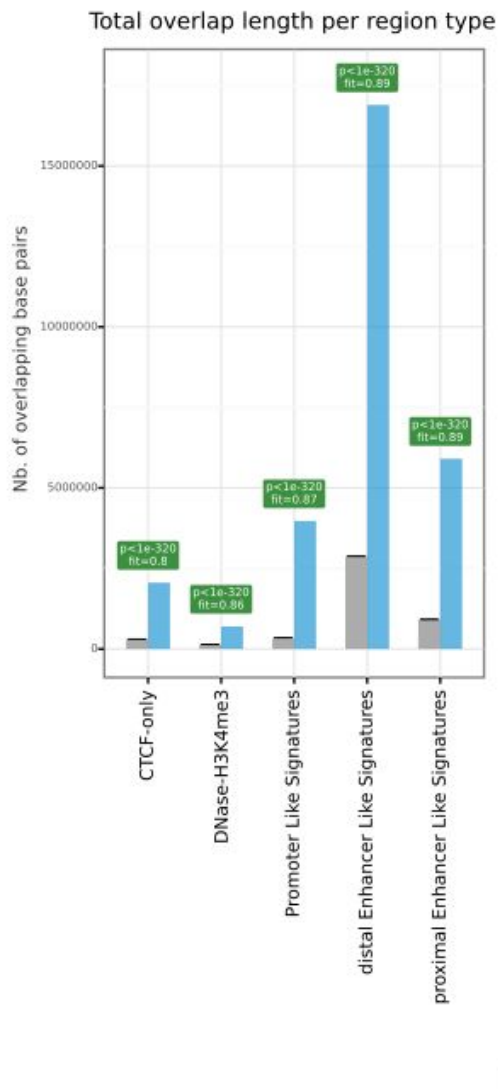


Supplementary Figure 26. Genomic distribution of *S. cerevisiae* TFBSs. Distribution of the proportion of *S. cerevisiae* UniBind robust TFBSs overlapping with different types of genomic regions (columns; see legend) across TFs (rows).

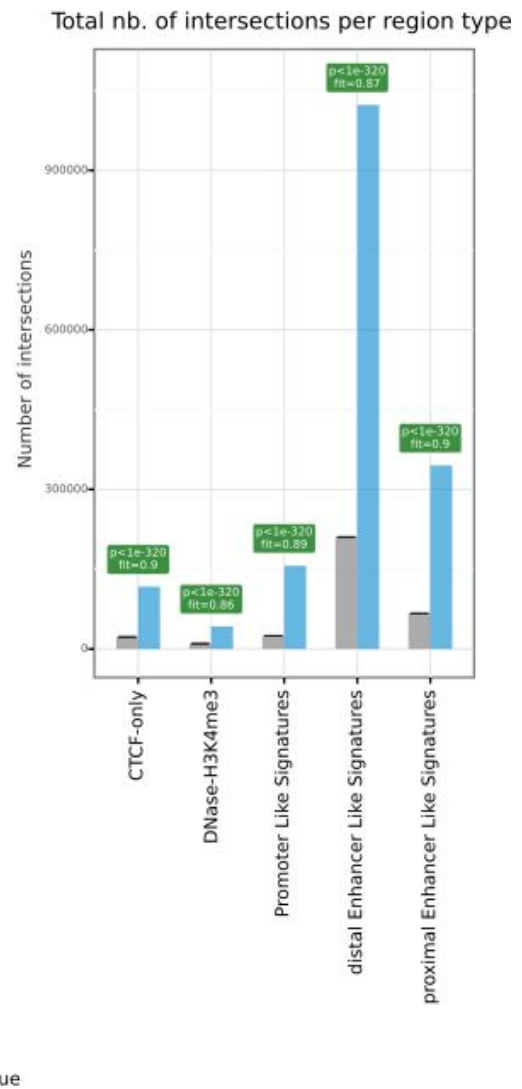
A**B**

Supplementary Figure 27. Enrichment analysis for *H. sapiens* TFBSs in ENCODE cCREs. Barplots representing the expected (grey bars) versus observed (blue bars) overlap lengths (**A**) or number of intersections (**B**) between *H. sapiens* TFBSs from the robust collection and ENCODE cCREs (x-axis). The plots and computed p-values (green: enrichment; orange: depletion) were obtained using the OLOGRAM command of the GTF toolkit.

A

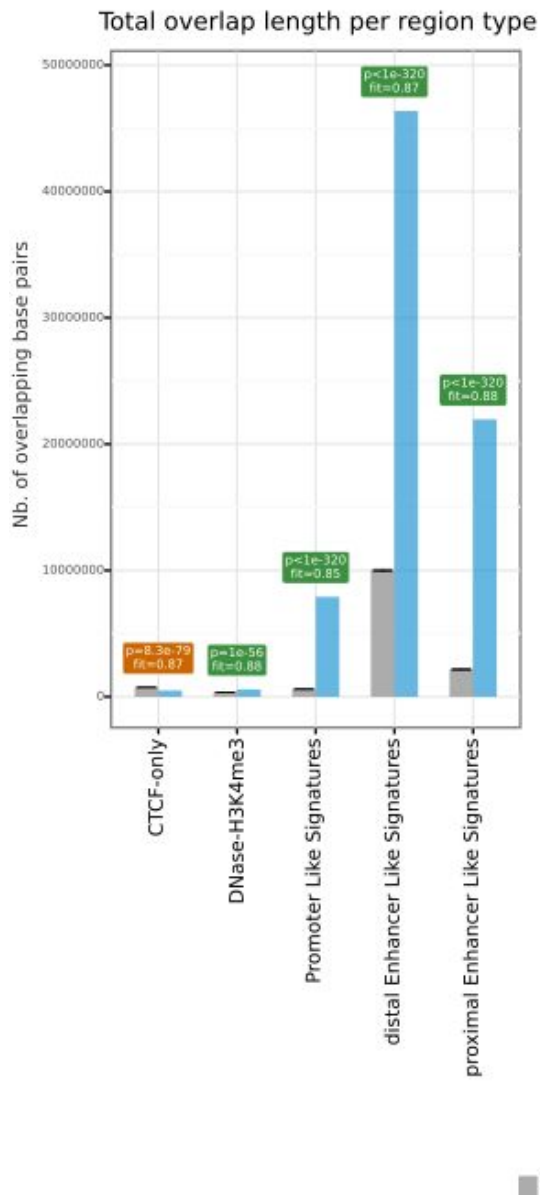


B

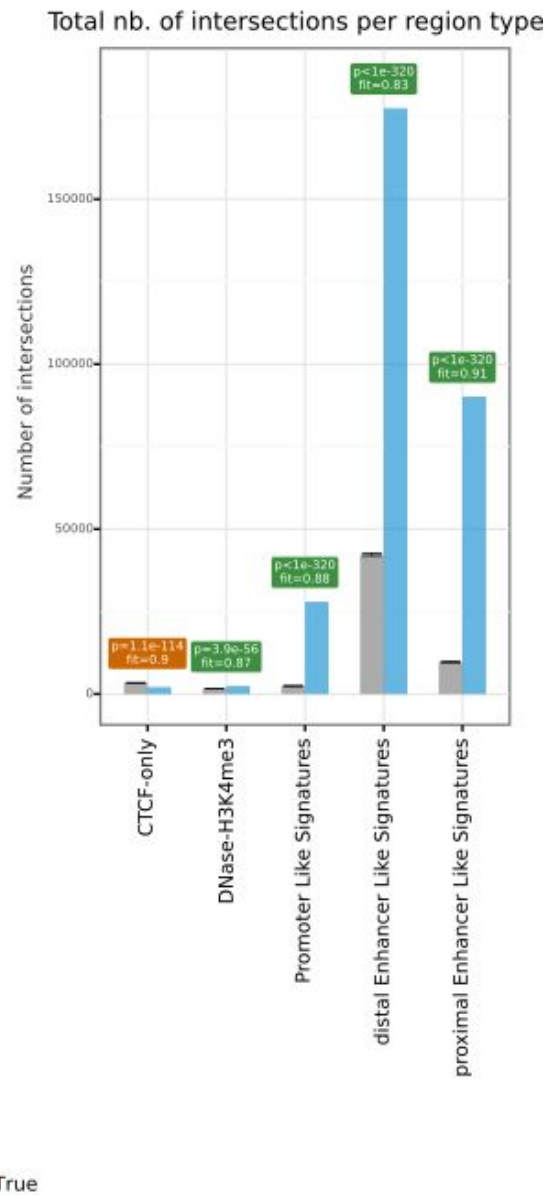


Supplementary Figure 28. Enrichment analysis for *M. musculus* TFBSs in ENCODE cCREs. Barplots representing the expected (grey bars) versus observed (blue bars) overlap lengths (A) or number of intersections (B) between *M. musculus* TFBSs from the robust collection and ENCODE cCREs (x-axis). The plots and computed p-values (green: enrichment; orange: depletion) were obtained using the OLOGRAM command of the GTF toolkit.

A

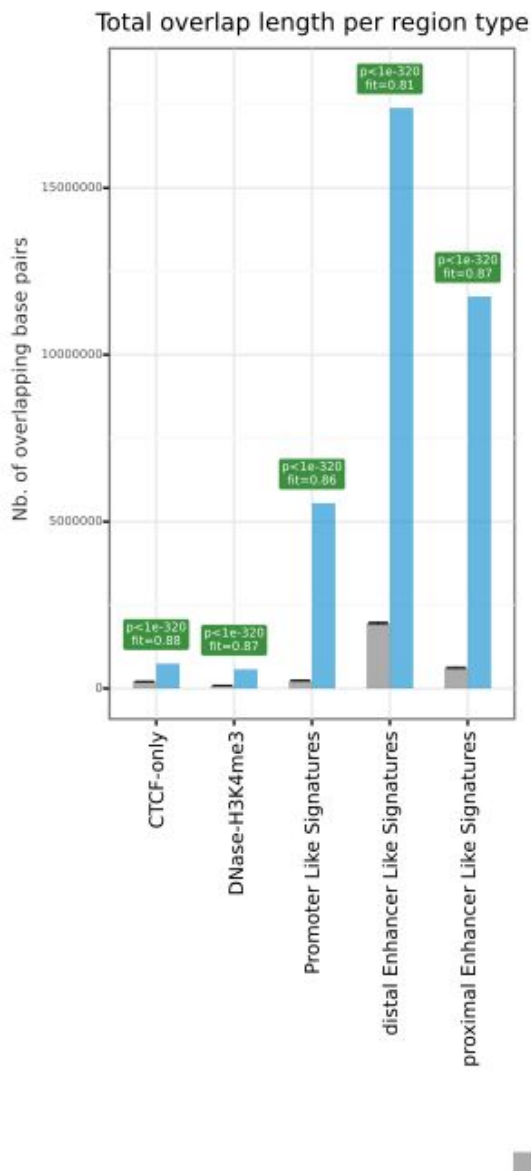


B

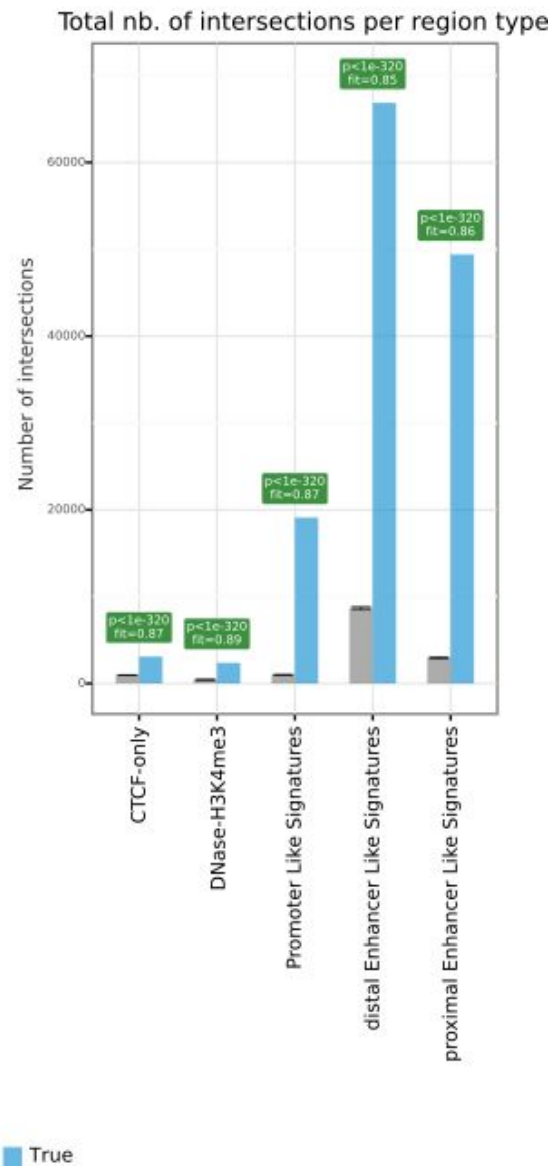


Supplementary Figure 29. Enrichment analysis for *H. sapiens* CRMs in ENCODE cCREs. Barplots representing the expected (grey bars) versus observed (blue bars) overlap lengths (A) or number of intersections (B) between *H. sapiens* CRMs from the robust collection and ENCODE cCREs (x-axis). The plots and computed p-values (green: enrichment; orange: depletion) were obtained using the OLOGRAM command of the GTF toolkit.

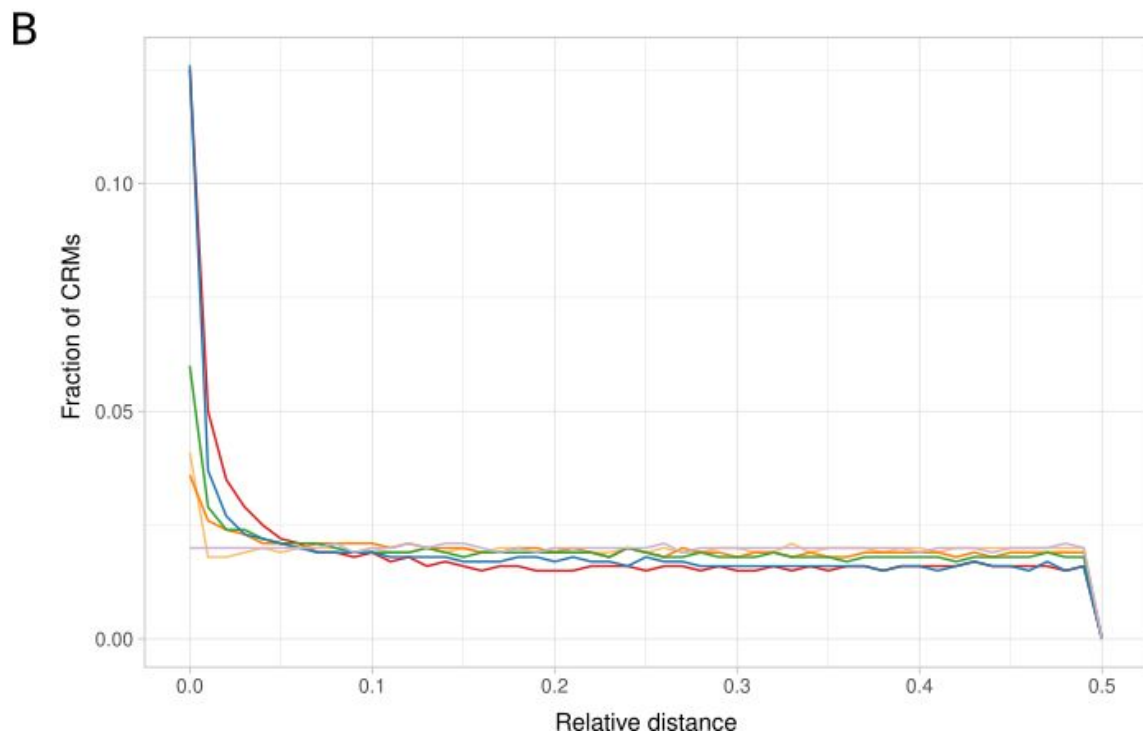
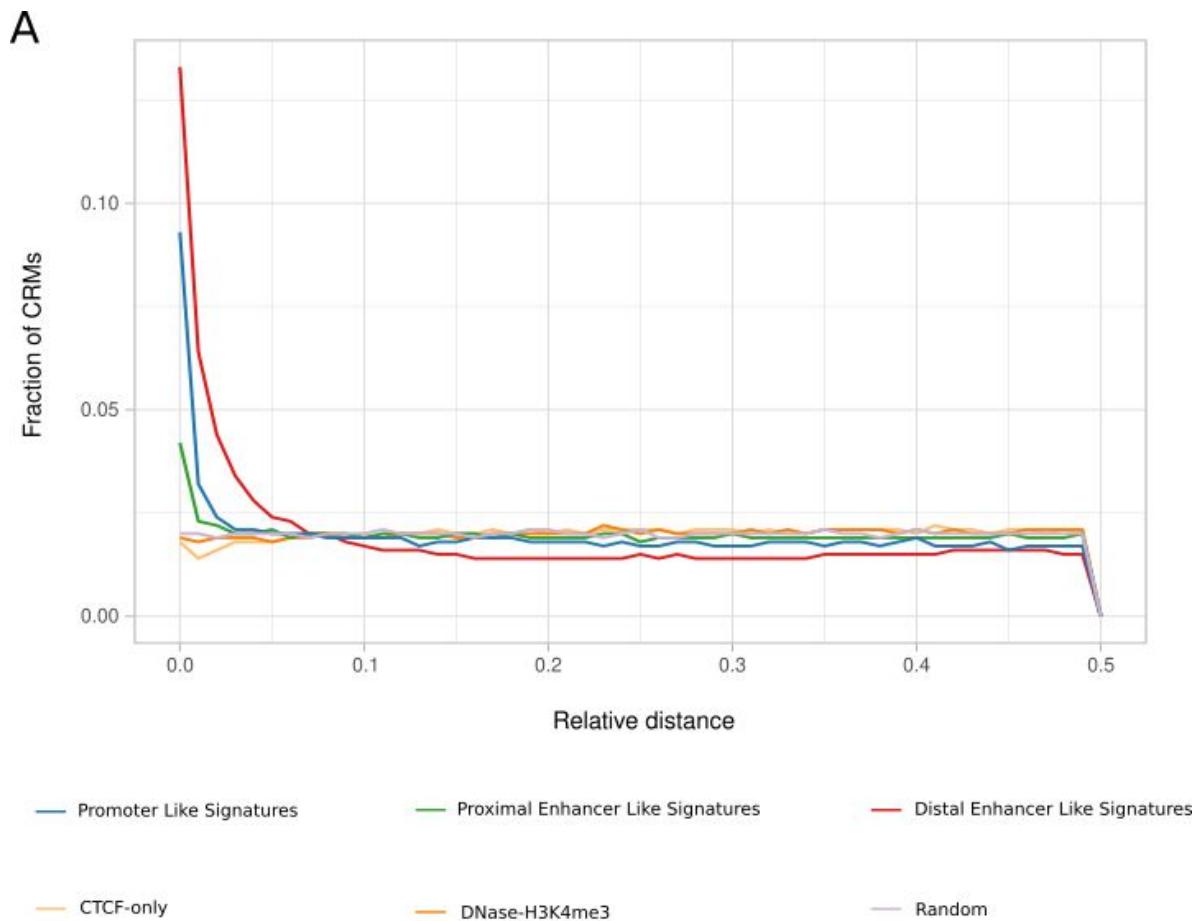
A



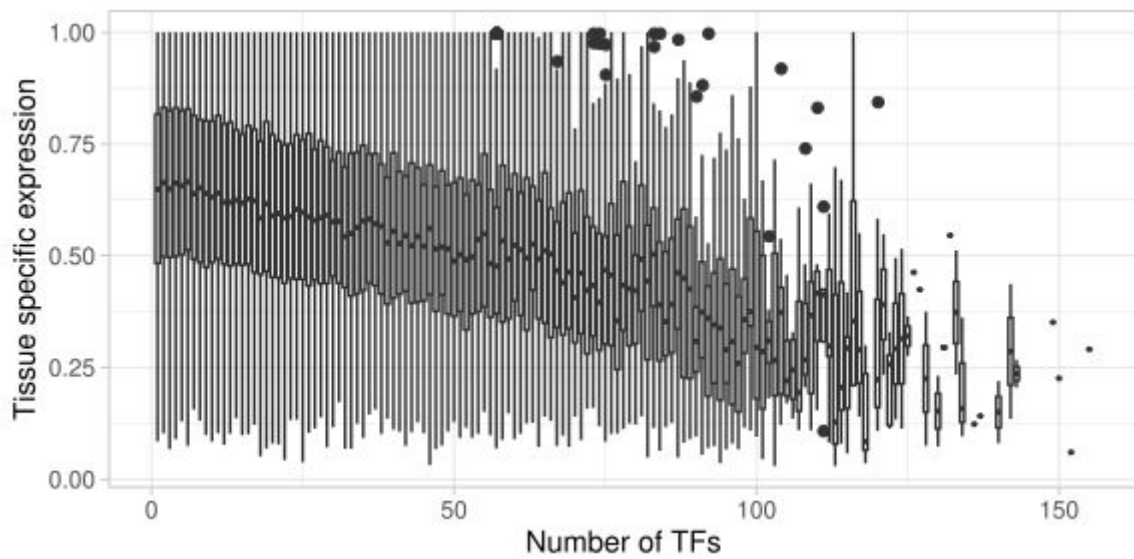
B



Supplementary Figure 30. Enrichment analysis for *M. musculus* CRMs in ENCODE cCREs. Barplots representing the expected (grey bars) versus observed (blue bars) overlap lengths (A) or number of intersections (B) between *M. musculus* CRMs from the robust collection and ENCODE cCREs (x-axis). The plots and computed p-values (green: enrichment; orange: depletion) were obtained using the OLOGRAM command of the GTF toolkit.



Supplementary Figure 31. Relative distance distributions between CRMs and ENCODE cCREs. Fraction of CRMs in the UniBind robust collection (y-axis) with respect to increasing relative distances (x-axis) from ENCODE cCREs computed using the *bedtools reldist* command for human (A) and mouse (B). When two genomic tracks are not spatially related, one expects the fraction of relative distance distribution to be uniform.



Supplementary Figure 32. Correlation between enhancer activity and TF binding. For each enhancer predicted using Cap Analysis of Gene Expression (CAGE) by the FANTOM5 consortium, we computed the number of TFs with overlapping TFBSs in the robust collection of UniBind (x-axis). The figure provides, for each value of the number of TFs, a bar plot of the distribution of tissue specific activity of these enhancers. The expression measures were derived from CAGE (capturing enhancer RNA expression). The tissue specificity of activity (y-axis) is provided within the [0; 1] range with 0 representing ubiquitous enhancer activity and 1 exclusive expression activity.