Supplemental analysis: Data reanalysis DNA Transposons Favor De Novo Transcript Emergence Through Enrichment of Transcription Factor Binding Motifs

Marie Kristin Lebherz, Bertrand Fouks, Erich Bornberg-Bauer and Anna Grandchamp

March 10, 2026

1. TE overlap in *de novo* transcripts and surrounding regions

Table 1: Summary of the statistical Model (glmmTMB): TE_overlap ~ Type + (1—Population), Zero-inflation:Type + (1—Population), Gamma family

	TransposonUltimate	**RepeatModeler**	**MCHelper**	**EarlGrey**	**FasTE**
Model Statistics					
AIC	38005.1	77019.9	77373.0	77890.2	80883.6
BIC	38104.0	77119.6	77472.7	77989.9	80983.3
logLik	-18991.5	-38498.9	-38675.5	-38934.1	-40430.8
$-2\log(L)$	37983.1	76997.9	77351.0	77868.2	80861.6
df.resid	59308	63873	63873	63873	63873
Random Effects (Variance / Std. Dev.)					
Population (Conditional model, Intercept)	0.005522 / 0.07431	0.0003183 / 0.01784	0.000571 / 0.0239	0.000928 / 0.03046	0.002373 / 0.04871
Population (Zero-inflation model, Intercept)	0.04597 / 0.2144	0.01132 / 0.1064	0.01174 / 0.1084	0.007311 / 0.0855	0.0098 / 0.099
Conditional Model (Estimate / $P(>\|z\|)$)					
(Intercept)	2.81461 / $< 2 \times 10^{-16}$ ***	1.49262 / $< 2 \times 10^{-16}$ ***	1.45617 / $< 2 \times 10^{-16}$ ***	1.41648 / $< 2 \times 10^{-16}$ ***	1.69794 / $< 2 \times 10^{-16}$ ***
Type: Intergenic	1.58675 / $< 2 \times 10^{-16}$ ***	1.20586 / $< 2 \times 10^{-16}$ ***	1.24287 / $< 2 \times 10^{-16}$ ***	1.24000 / $< 2 \times 10^{-16}$ ***	1.27632 / $< 2 \times 10^{-16}$ ***
Type: Transcript	-0.30448 / 1.4×10^{-9} ***	-0.00817 / 0.866	0.02333 / 0.629	0.00001729 / 1.000	-0.12509 / 0.00874 **
Type: Upstream	0.01589 / 0.758	0.02574 / 0.595	0.02333 / 0.629	-0.00941	-0.0000487 / 0.99919
Pairwise Comparisons (Bonferroni-adjusted p-values)					
Downstream / Intergenic	2.48×10^{-20}	2.57×10^{-12}	1.61×10^{-8}	5.28×10^{-7}	4.29×10^{-15}
Downstream / Transcript	2.58×10^{-20}	4.93×10^{-4}	2.47×10^{-4}	6.93×10^{-5}	4.78×10^{-1}
Downstream / Upstream	1.00	1.00	1.00	1.00	0.801
Intergenic / Transcript	0.049	2.68×10^{-39}	2.01×10^{-32}	1.33×10^{-31}	5.02×10^{-27}
Intergenic / Upstream	1.22×10^{-18}	7.31×10^{-10}	9.82×10^{-8}	9.85×10^{-5}	2.42×10^{-8}
Transcript / Upstream	4.38×10^{-19}	5.37×10^{-5}	1.17×10^{-4}	1.74×10^{-6}	6.50×10^{-3}

2. *De novo* transcripts and Motifs

a. Comparison of Core Motifs

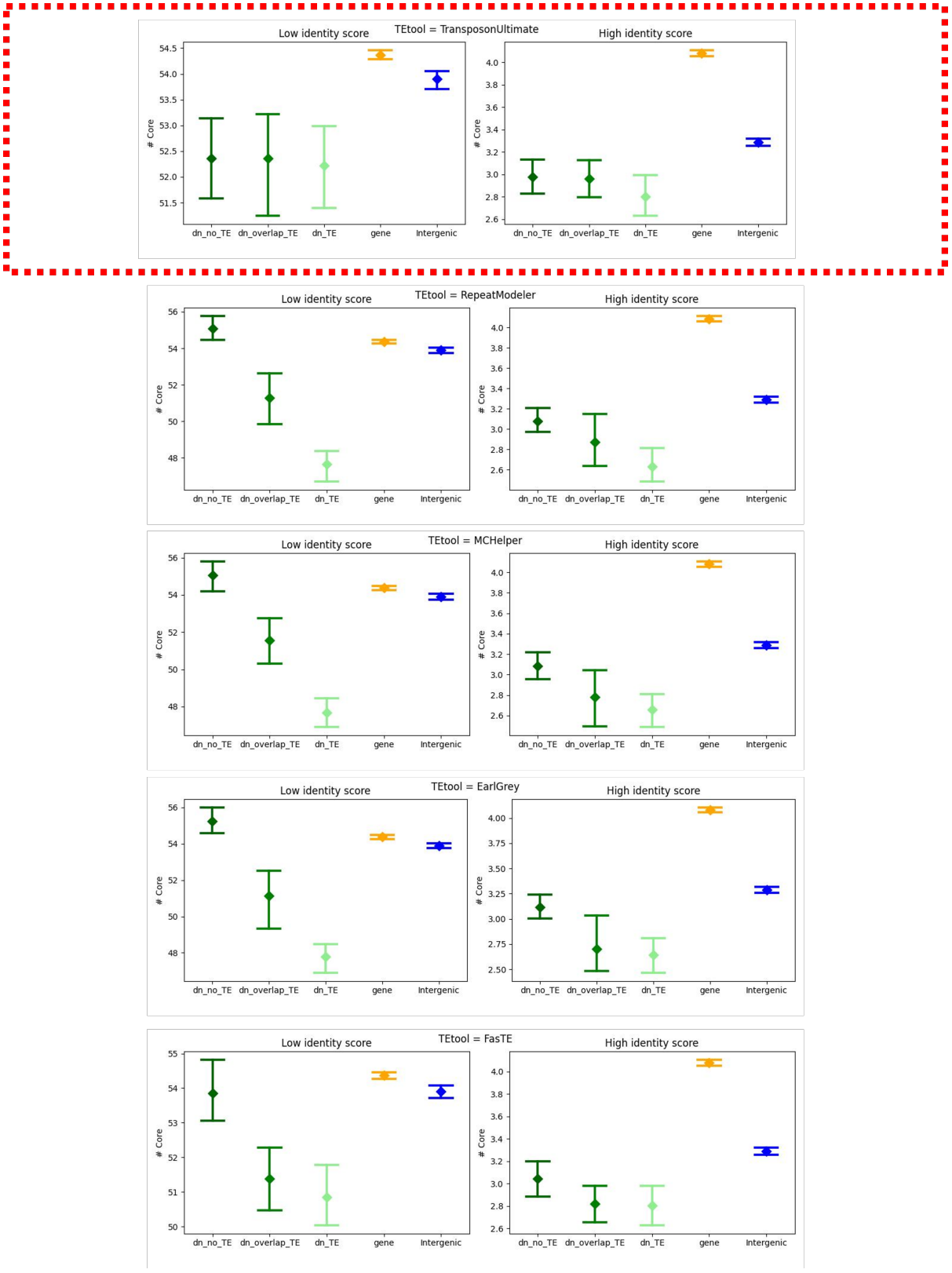

Figure 1: Number of core motifs per upstream region (-200/+100) based on the TE overlap of the sequence using both a low (left) and high (right) threshold. Shown is Figure 4 from the original paper. In addition to TransposonUltimate (= used in the paper, red box) we ran four additional TE detection and annotation pipelines, namely RepeatModeler, MCHelper (RepeatModeler library), EarlGrey and FasTE (combines EDTA and DeepTE)

Table 2: Summary of the statistical Model: NumberCore ~ Type + (1—population) (family = poisson)

	TransposonUltimate	RepeatModeler	MCHelper	EarlGrey	FasTE
Model Statistics					
AIC	1193124.2	1191990.8	1191988.0	1191916.8	1192936.4
BIC	1193183.8	1192050.4	1192048.0	1191976.5	1192996.0
logLik	-596556.1	-595989.4	-595988.0	-595952.4	-596462.2
$-2\log(L)$	1193112.2	1191978.8	1191976.0	1191904.8	1192924.4
df.resid	153461	153461	153461	153461	153461
Random Effects					
Population (Intercept: Var/SD)	1.747e-11 / 4.18e-06	1.754e-11 / 4.189e-06	2.049e-11 / 4.527e-06	2.242e-11 / 4.735e-06	1.734e-11 / 4.164e-06
Model Summary (Estimate / $P(>\|z\|)$)					
(Intercept)	3.958 / $< 2 \times 10^{-16}$ ***	4.008996 / $< 2 \times 10^{-16}$ ***	4.008370 / $< 2 \times 10^{-16}$ ***	4.011425 / $< 2 \times 10^{-16}$ ***	3.986220 / $< 2 \times 10^{-16}$ ***
Type: Overlap TE	-1.926e-05 / 0.997	-0.071678 / $< 2 \times 10^{-16}$ ***	-0.065711 / $< 2 \times 10^{-16}$ ***	-0.077195 / $< 2 \times 10^{-16}$ ***	-0.046919 / $< 2 \times 10^{-16}$ ***
Type: TE	-0.002604 / 0.555	-0.144951 / $< 2 \times 10^{-16}$ ***	-0.143761 / $< 2 \times 10^{-16}$ ***	-0.144535 / $< 2 \times 10^{-16}$ ***	-0.057237 / $< 2 \times 10^{-16}$ ***
Type: Gene	0.03772 / $< 2 \times 10^{-16}$ ***	-0.013176 / 1.54×10^{-7} ***	-0.012551 / 5.21×10^{-7} ***	-0.015605 / 4.35×10^{-10} ***	0.009600 / 0.000825 ***
Type: Intergenic	0.02914 / $< 2 \times 10^{-16}$ ***	-0.021751 / $< 2 \times 10^{-16}$ ***	-0.021125 / $< 2 \times 10^{-16}$ ***	-0.024179 / $< 2 \times 10^{-16}$ ***	0.001025 / 0.723
Pairwise Comparisons (Bonferroni-adjusted p)					
No TE / Overlap TE	1	$< .0001$	$< .0001$	$< .0001$	$< .0001$
No TE / TE	1	$< .0001$	$< .0001$	$< .0001$	$< .0001$
No TE / Gene	$< .0001$	$< .0001$	$< .0001$	$< .0001$	0.0083
No TE / Intergenic	$< .0001$	$< .0001$	$< .0001$	$< .0001$	1
Overlap TE / TE	1	$< .0001$	$< .0001$	$< .0001$	0.4313
Overlap TE / Gene	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
Overlap TE / Intergenic	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
TE / Gene	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
TE / Intergenic	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
Gene / Intergenic	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$

Table 3: Summary of the statistical Model: NumberCoreHigh ~ Type + (1—population) (family = poisson)

	TransposonUltimate	RepeatModeler	MCHelper	EarlGrey	FasTE
Model Statistics					
AIC	687742.2	687682.1	687681.9	687658.6	687729.6
BIC	687801.8	687741.7	687741.5	687718.2	687789.2
logLik	-343865.1	-343835.0	-343834.9	-343823.3	-343858.8
$-2\log(L)$	687730.2	687670.1	687669.9	687646.6	687717.6
df.resid	153461	153461	153461	153461	153461
Random Effects					
Population (Intercept: Var/SD)	3.955e-10 / 1.989e-05	2.475e-10 / 1.573e-05	3.194e-10 / 1.787e-05	3.36e-10 / 1.833e-05	2.545e-10 / 1.595e-05
Model Summary (Estimate / $P(>\|z\|)$)					
(Intercept)	1.090720 / $< 2 \times 10^{-16}$ ***	1.12365 / $< 2 \times 10^{-16}$ ***	1.12647 / $< 2 \times 10^{-16}$ ***	1.13594 / $< 2 \times 10^{-16}$ ***	1.11366 / $< 2 \times 10^{-16}$ ***
Type: Overlap TE	-0.005165 / 0.797	-0.06900 / 0.00409 **	-0.10372 / 5.09×10^{-5} ***	-0.14188 / 3.49×10^{-7} ***	-0.07650 / 6.17×10^{-5} ***
Type: TE	-0.060509 / 0.00129 **	-0.15556 / $< 2 \times 10^{-16}$ ***	-0.14879 / 3.64×10^{-16} ***	-0.16420 / $< 2 \times 10^{-16}$ ***	-0.08314 / 2.74×10^{-5} ***
Type: Gene	0.315543 / $< 2 \times 10^{-16}$ ***	0.28261 / $< 2 \times 10^{-16}$ ***	0.27979 / $< 2 \times 10^{-16}$ ***	0.27033 / $< 2 \times 10^{-16}$ ***	0.29260 / $< 2 \times 10^{-16}$ ***
Type: Intergenic	0.099363 / 4.88×10^{-15} ***	0.06643 / 5.97×10^{-10} ***	0.06361 / 2.51×10^{-9} ***	0.05415 / 3.57×10^{-7} ***	0.07642 / 3.34×10^{-10} ***
Pairwise Comparisons (Bonferroni-adjusted p)					
No TE / Overlap TE	1	0.0409	0.0005	$< .0001$	0.0006
No TE / TE	0.0129	$< .0001$	$< .0001$	$< .0001$	0.0003
No TE / Gene	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
No TE / Intergenic	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
Overlap TE / TE	0.088	0.0111	1	1	1
Overlap TE / Gene	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
Overlap TE / Intergenic	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
TE / Gene	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
TE / Intergenic	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
Gene / Intergenic	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$

b. Comparison of TFSB Motifs

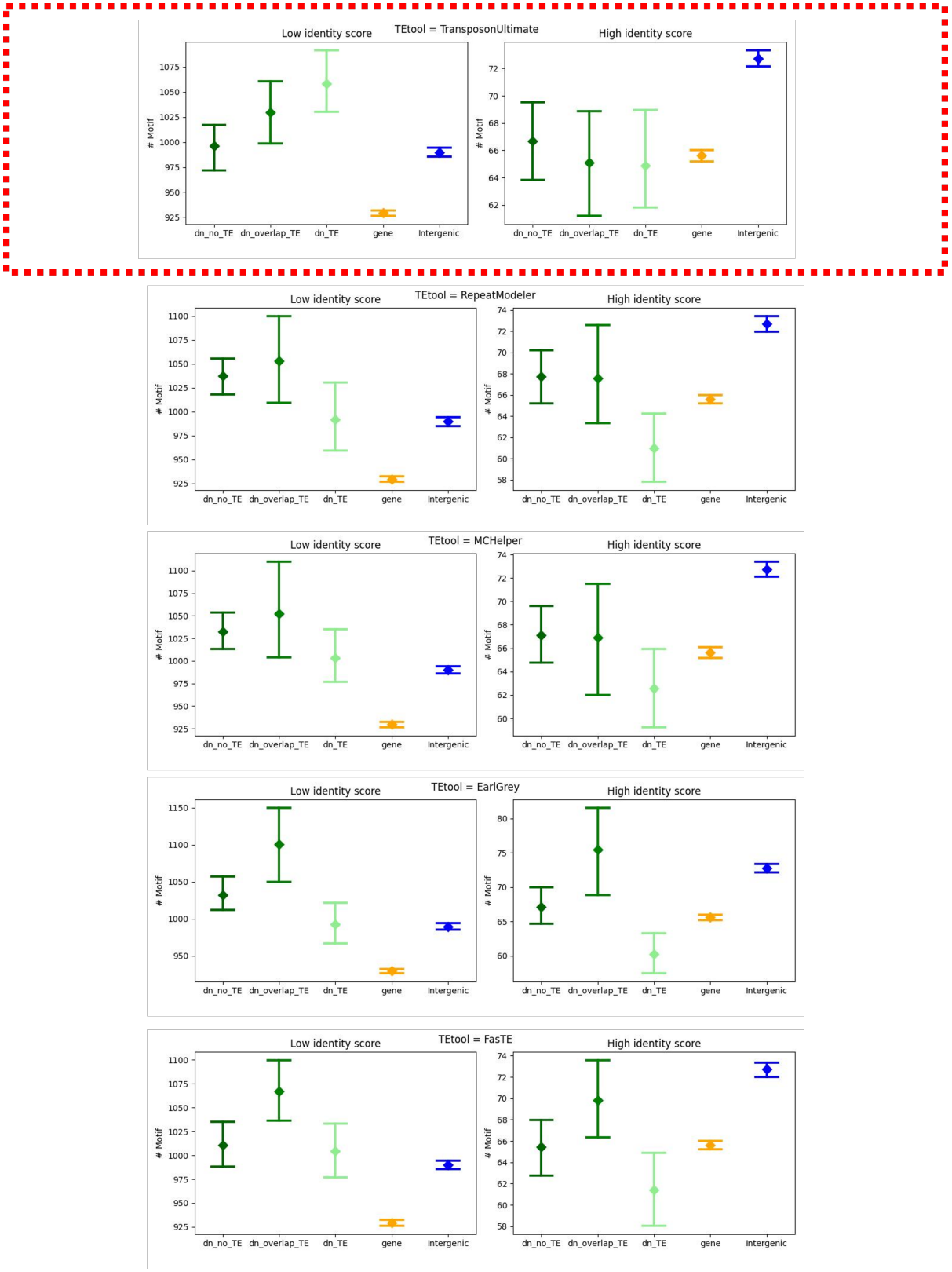

Figure 2: Number of promotor/enhancer motifs per upstream region (-1000/+100) based on the TE overlap of the sequence using both a low (left) and high (right) threshold. Shown is Figure 4 from the original paper. In addition to TransposonUltimate (= used in the paper, red box) we ran four additional TE detection and annotation pipelines, namely RepeatModeler, MCHelper (RepeatModeler library), EarlGrey and FasTE (combines EDTA and DeepTE)

Table 4: Summary of the statistical Model: NumberMotif ~ Type + (1—population) (family = poisson)

	TransposonUltimate	RepeatModeler	MCHelper	EarlGrey	FasTE
Model Statistics					
AIC	14947764	14948696	14950062	14946507	14947676
BIC	14947824	14948756	14950121	14946567	14947736
logLik	-7473876	-7474342	-7475025	-7473248	-7473832
$-2\log(L)$	14947752	14948684	14950050	14946495	14947664
df.resid	153459	153459	153459	153459	153459
Random Effects					
Population (Intercept: Var/SD)	1.314e-06 / 0.001146	1.41e-06 / 0.001187	1.422e-06 / 0.001192	1.399e-06 / 0.001183	1.425e-06 / 0.001194
Model Summary (Estimate / $P(>\|z\|)$)					
(Intercept)	6.9040831 / $< 2 \times 10^{-16}$ ***	6.9445219 / $< 2 \times 10^{-16}$ ***	6.9396559 / $< 2 \times 10^{-16}$ ***	6.9389725 / $< 2 \times 10^{-16}$ ***	6.9181389 / $< 2 \times 10^{-16}$ ***
Type: Overlap TE	0.0330892 / $< 2 \times 10^{-16}$ ***	0.0150026 / $< 2 \times 10^{-16}$ ***	0.0193734 / $< 2 \times 10^{-16}$ ***	0.0645010 / $< 2 \times 10^{-16}$ ***	0.0546187 / $< 2 \times 10^{-16}$ ***
Type: TE	0.0602856 / $< 2 \times 10^{-16}$ ***	-0.0450370 / $< 2 \times 10^{-16}$ ***	-0.0284142 / $< 2 \times 10^{-16}$ ***	-0.0384727 / $< 2 \times 10^{-16}$ ***	-0.0058326 / 4.03×10^{-8} ***
Type: Gene	-0.0695211 / $< 2 \times 10^{-16}$ ***	-0.1099604 / $< 2 \times 10^{-16}$ ***	-0.1050946 / $< 2 \times 10^{-16}$ ***	-0.1044107 / $< 2 \times 10^{-16}$ ***	-0.0835768 / $< 2 \times 10^{-16}$ ***
Type: Intergenic	-0.0066637 / $< 2 \times 10^{-16}$ ***	0.0471030 / $< 2 \times 10^{-16}$ ***	-0.0422393 / $< 2 \times 10^{-16}$ ***	-0.0415523 / $< 2 \times 10^{-16}$ ***	-0.0207182 / $< 2 \times 10^{-16}$ ***
Pairwise Comparisons (Bonferroni-adjusted p)					
No TE / Overlap TE	< .0001	< .0001	< .0001	< .0001	< .0001
No TE / TE	< .0001	< .0001	< .0001	< .0001	< .0001
No TE / Gene	< .0001	< .0001	< .0001	< .0001	< .0001
No TE / Intergenic	< .0001	< .0001	< .0001	< .0001	< .0001
Overlap TE / TE	< .0001	< .0001	< .0001	< .0001	< .0001
Overlap TE / Gene	< .0001	< .0001	< .0001	< .0001	< .0001
Overlap TE / Intergenic	< .0001	< .0001	< .0001	< .0001	< .0001
TE / Gene	< .0001	< .0001	< .0001	< .0001	< .0001
TE / Intergenic	< .0001	0.1016	< .0001	0.0006	< .0001
Gene / Intergenic	< .0001	< .0001	< .0001	< .0001	< .0001

Table 5: Summary of the statistical Model: NumberMotifHigh ~ Type + (1—population) (family = poisson)

	TransposonUltimate	RepeatModeler	MCHelper	EarlGrey	FasTE
Model Statistics					
AIC	4707456	4706729	4707156	4705832	4706694
BIC	4707515	4706788	4707215	4705892	4706754
logLik	-2353722	-2353358	-2353572	-2352910	-2353341
$-2\log(L)$	4707444	4706717	4707144	4705820	4706682
df.resid	153459	153459	153459	153459	153459
Random Effects					
Population (Intercept: Var/SD)	1.156e-05 / 0.0034	1.155e-05 / 0.003399	1.16e-05 / 0.003405	1.16e-05 / 0.003407	1.166e-05 / 0.003414
Model Summary (Estimate / $P(>\|z\|)$)					
(Intercept)	4.199720 / $< 2 \times 10^{-16}$ ***	4.215585 / $< 2 \times 10^{-16}$ ***	4.206292 / $< 2 \times 10^{-16}$ ***	4.205884 / $< 2 \times 10^{-16}$ ***	4.180648 / $< 2 \times 10^{-16}$ ***
Type: Overlap TE	-0.023533 / 3.55×10^{-8} ***	-0.002972 / 0.551	-0.002799 / 0.595	0.117474 / $< 2 \times 10^{-16}$ ***	0.065517 / $< 2 \times 10^{-16}$ ***
Type: TE	-0.027343 / 3.81×10^{-12} ***	-0.105165 / $< 2 \times 10^{-16}$ ***	-0.069929 / $< 2 \times 10^{-16}$ ***	-0.107519 / $< 2 \times 10^{-16}$ ***	-0.063193 / $< 2 \times 10^{-16}$ ***
Type: Gene	-0.015986 / 1.98×10^{-9} ***	-0.031856 / $< 2 \times 10^{-16}$ ***	-0.022563 / $< 2 \times 10^{-16}$ ***	-0.022155 / $< 2 \times 10^{-16}$ ***	0.003083 / 0.237
Type: Intergenic	0.086875 / $< 2 \times 10^{-16}$ ***	0.071009 / $< 2 \times 10^{-16}$ ***	0.080298 / $< 2 \times 10^{-16}$ ***	0.080714 / $< 2 \times 10^{-16}$ ***	0.105951 / $< 2 \times 10^{-16}$ ***
Pairwise Comparisons (Bonferroni-adjusted p)					
No TE / Overlap TE	< .0001	1	1	< .0001	< .0001
No TE / TE	< .0001	< .0001	< .0001	< .0001	< .0001
No TE / Gene	< .0001	< .0001	< .0001	< .0001	1
No TE / Intergenic	< .0001	< .0001	< .0001	< .0001	< .0001
Overlap TE / TE	1	< .0001	< .0001	< .0001	< .0001
Overlap TE / Gene	0.2572	< .0001	0.0004	< .0001	< .0001
Overlap TE / Intergenic	< .0001	< .0001	< .0001	< .0001	< .0001
TE / Gene	0.0012	< .0001	< .0001	< .0001	< .0001
TE / Intergenic	< .0001	< .0001	< .0001	< .0001	< .0001
Gene / Intergenic	< .0001	< .0001	< .0001	< .0001	< .0001

c. Comparison of motifs enrichments between de novo transcripts and their non transcribed homologous sequences

Table 6: Anova test (type = "III") for the model (Transcription_status ~ Coverage (= rel. TE overlap) + Number (= number TE) + Number_core + binO (= presence TE) :Number_motifs + (1 — Number_Orthogroup) + (1 — Population)).

Tool	Term	Chisq	Df	Pr(>Chisq)
TransposonUltimate	(Intercept)	69.718892	1	6.84×10^{-17}
	Coverage	278.063131	1	1.98×10^{-62}
	Number	32.058692	1	1.50×10^{-8}
	Number_core	4.576502	1	3.24×10^{-2}
	binO:Number_motifs	17.731103	1	2.54×10^{-5}
RepeatModeler	(Intercept)	91.40412103	1	1.17×10^{-21}
	Coverage	10.73471437	1	1.05×10^{-3}
	Number	22.80459691	1	1.79×10^{-6}
	Number_core	3.38218145	1	6.59×10^{-2}
	binO:Number_motifs	0.06602595	1	7.97×10^{-1}
MCHelper	(Intercept)	89.0769015	1	3.80×10^{-21}
	Coverage	7.0254064	1	8.04×10^{-3}
	Number	23.5274094	1	1.23×10^{-6}
	Number_core	3.6741814	1	5.53×10^{-2}
	binO:Number_motifs	0.2908766	1	5.90×10^{-1}
EarlGrey	(Intercept)	89.9397236	1	2.46×10^{-21}
	Coverage	11.95692048	1	5.44×10^{-4}
	Number	26.40619463	1	2.77×10^{-7}
	Number_core	3.467874863	1	6.26×10^{-2}
	binO:Number_motifs	0.005539105	1	9.41×10^{-1}
FasTE	(Intercept)	102.9667483	1	3.41×10^{-24}
	Coverage	11.1987402	1	8.19×10^{-4}
	Number	17.8812966	1	2.35×10^{-5}
	Number_core	2.1835251	1	1.39×10^{-1}
	binO:Number_motifs	0.5561409	1	4.56×10^{-1}

Table 7: Summary of the statistical Model: Transcription_status ~ Coverage (= rel. TE overlap) + Number (= number TE) + Number_core + binO (= presence TE) :Number_motifs + (1 — Number_Orthogroup) + (1 — Population)

Tool / Variable	**Term**	**Estimate / Variance**	**Pr(>—z—)**
TransposonUltimate			
AIC=9984.7, BIC=10036.4, logLik=-4985.4, $-2\log(L) = 9970.7$, df.resid=11788			
Random effects (Variance / Std.Dev)			
NumberOrthogroups	(Intercept)	1.157×10^{-9} / 3.401×10^{-5}	
Population	(Intercept)	4.286×10^{-2} / 2.070×10^{-1}	
Conditional model			
	(Intercept)	-1.21	$< 2 \times 10^{-16}$ ***
	Coverage	-2.82	$< 2 \times 10^{-16}$ ***
	Number	1.45×10^{-1}	1.50×10^{-8} ***
	Number_core	-4.80×10^{-3}	3.24×10^{-2} *
	binO:Number_motifs	3.24×10^{-4}	2.54×10^{-5} ***
RepeatModeler			
AIC=10482.4, BIC=10534.0, logLik=-5234.2, $-2\log(L) = 10468.4$, df.resid=11788			
Random effects (Variance / Std.Dev)			
NumberOrthogroups	(Intercept)	5.705×10^{-10} / 2.389×10^{-5}	
Population	(Intercept)	3.728×10^{-2} / 1.931×10^{-1}	
Conditional model			
	(Intercept)	-1.40	$< 2 \times 10^{-16}$ ***
	Coverage	-3.05×10^{-1}	1.05×10^{-3} **

Tool / Model component	Term	Estimate / Variance	Pr(>—z—)
	Number	9.22	1.79×10^{-6} ***
	Number_core	-4.17×10^{-3}	6.59×10^{-2} .
	binO:Number_motifs	2.13×10^{-5}	7.97×10^{-1}
MCHelper			
AIC=10483.1, BIC=10534.7, logLik=-5234.5, $-2\log(L) = 10469.1$, df.resid=11788			
Random effects (Variance / Std.Dev)			
NumberOrthogroups	(Intercept)	7.553×10^{-10} / 2.748×10^{-5}	
Population	(Intercept)	3.832×10^{-2} / 1.958×10^{-1}	
Conditional model			
	(Intercept)	-1.39	$< 2 \times 10^{-16}$ ***
	Coverage	-2.49×10^{-1}	8.04×10^{-3} **
	Number	1.21×10^{-1}	1.23×10^{-6} ***
	Number_core	-4.35×10^{-3}	5.53×10^{-2} .
	binO:Number_motifs	-4.73×10^{-5}	5.90×10^{-1}
EarlGrey			
AIC=10477.7, BIC=10529.3, logLik=-5231.8, $-2\log(L) = 10463.7$, df.resid=11788			
Random effects (Variance / Std.Dev)			
NumberOrthogroups	(Intercept)	1.052×10^{-9} / 3.243×10^{-5}	
Population	(Intercept)	3.799×10^{-2} / 1.949×10^{-1}	
Conditional model			
	(Intercept)	-1.40	$< 2 \times 10^{-16}$ ***
	Coverage	-3.23×10^{-1}	5.44×10^{-4} ***
	Number	1.66×10^{-1}	2.77×10^{-7} ***
	Number_core	-4.23×10^{-3}	6.26×10^{-2} .
	binO:Number_motifs	-6.73×10^{-6}	9.41×10^{-1}
FasTE			
AIC=10487.5, BIC=10539.1, logLik=-5236.7, $-2\log(L) = 10473.5$, df.resid=11788			
Random effects (Variance / Std.Dev)			
NumberOrthogroups	(Intercept)	6.802×10^{-10} / 2.608×10^{-5}	
Population	(Intercept)	3.784×10^{-2} / 1.945×10^{-1}	
Conditional model			
	(Intercept)	-1.46	$< 2 \times 10^{-16}$ ***
	Coverage	-2.89×10^{-1}	8.19×10^{-4} ***
	Number	5.47×10^{-2}	2.35×10^{-5} ***
	Number_core	-3.28×10^{-3}	1.39×10^{-1}
	binO:Number_motifs	5.12×10^{-5}	4.56×10^{-1}

References