SUPPLEMENTARY MATERIAL

Genome-wide analysis of lncRNA and mRNA transcript complexity in human and mouse

KOUSHIKI BASU[1] and MANJARI KIRAN[1]*

[1]*Department of Systems and Computational Biology, School of Life Sciences, University of Hyderabad, Hyderabad, Telangana, 500 046, India.*

Organism	**Human (GENCODE v38)**		**Mouse (GENCODE M28)**	
Type of Gene	**lncRNA**	**mRNA**	**lncRNA**	**mRNA**
Number of Genes	16877	19950	9949	21862
Number of Transcripts	25279	58402	11992	43573
Number of Exons	74188	285138	35366	255166
Transcript Per Gene (TPG)	1.50	2.93	1.21	1.99

Table S1. Number of genes, transcripts, and exons used in the analyses (from GENCODE dataset) for human and mouse. Transcript per Gene (TPG) is calculated by dividing total number of transcripts by total number of genes.

TSL included	Median of TC		Correlation between #transcripts and #exons	
	lncRNA	mRNA	lncRNA	mRNA
Human (GENCODE v38)				
TSL1	0.33	0.25	0.35	0.45
TSL1 + TSL2	0.4	0.25	0.45	0.51
TSL1 + TSL2 + TSL3	0.4	0.25	0.54	0.52
TSL1 + TSL2 + TSL3 + TSL4	0.4	0.25	0.57	0.52
TSL1 + TSL2 + TSL3 + TSL4 + TSL5	0.4	0.23	0.62	0.59
TSL1 + TSL2 + TSL3 + TSL4 + TSL5 + TSLNA	0.5	0.25	0.57	0.62
Without any TSL included	0.5	0.25	0.69	0.66
Mouse (GENCODE M28)				
TSL1	0.33	0.2	0.44	0.40
TSL1 + TSL2	0.33	0.2	0.47	0.40
TSL1 + TSL2 + TSL3	0.33	0.2	0.53	0.40
TSL1 + TSL2 + TSL3 + TSL4	0.33	0.2	0.53	0.40
TSL1 + TSL2 + TSL3 + TSL4 + TSL5	0.33	0.2	0.55	0.50
TSL1 + TSL2 + TSL3 + TSL4 + TSL5 + TSLNA	0.33	0.21	0.52	0.47
Without any TSL included	0.33	0.21	0.53	0.48

Table S2. Median of TC and result of Spearman Correlation test (p-value < 2.2e-16) calculated for genes with different TSLs for human and mouse.

TSL included	Human (GENCODE v38)		Mouse (GENCODE M28)	
	lncRNA	mRNA	lncRNA	mRNA
TSL1	1521	16438	3842	17188
TSL1 + TSL2	4144	17320	4482	17665
TSL1 + TSL2 + TSL3	7947	17511	5869	17982
TSL1 + TSL2 + TSL3 + TSL4	9036	17569	5869	17984
TSL1 + TSL2 + TSL3 + TSL4 + TSL5	10831	18389	7692	20277
TSL1 + TSL2 + TSL3 + TSL4 + TSL5 + TSLNA	13149	19550	8325	21643

Table S3. Number of genes with different TSLs used in the analyses (from GENCODE dataset) for human and mouse.

TSL included	W value	p value	Accepts H0/rejects H0
Human (GENCODE v38)			
TSL1	8821177	<2.2e-16	Rejects
TSL1 + TSL2	20612047	<2.2e-16	Rejects
TSL1 + TSL2 + TSL3	37424573	<2.2e-16	Rejects
TSL1 + TSL2 + TSL3 + TSL4	42676264	<2.2e-16	Rejects
TSL1 + TSL2 + TSL3 + TSL4 + TSL5	46229885	<2.2e-16	Rejects
TSL1 + TSL2 + TSL3 + TSL4 + TSL5 + TSLNA	60155255	<2.2e-16	Rejects
Without any TSL included	80801453	<2.2e-16	Rejects
Mouse (GENCODE M28)			
TSL1	15858164	<2.2e-16	Rejects
TSL1 + TSL2	19142082	<2.2e-16	Rejects
TSL1 + TSL2 + TSL3	25326427	<2.2e-16	Rejects
TSL1 + TSL2 + TSL3 + TSL4	25335562	<2.2e-16	Rejects
TSL1 + TSL2 + TSL3 + TSL4 + TSL5	35772481	<2.2e-16	Rejects
TSL1 + TSL2 + TSL3 + TSL4 + TSL5 + TSLNA	48492365	<2.2e-16	Rejects
Without any TSL included	56618539	<2.2e-16	Rejects

Table S4. Result of Wilcoxon rank-sum test calculated for genes with different TSLs with the null hypothesis that the median of TC for lncRNA is less than mRNA (p-value < 0.05) for human and mouse.

Organism	Human (GENCODE v38)				Mouse (GENCODE M28)			
Type of Gene	lncRNA		mRNA		lncRNA		mRNA	
Average Intron Length	45756.10		159775		21354.10		101853.6	
Average Exon Length	2134.04		9204.36		1568.05		6043.88	
Low/High TC	Low	High	Low	High	Low	High	Low	High
Intron Length (Mean)	89059.39	12170.43	177808.6	108058.3	40292.84	14970.76	130264.8	56171.15
Intron Length (Median)	12140.5	142	29185.5	1135.5	15117.5	2994	35877	6121
Wilcoxon rank-sum test (Intron Length)	**W=52145000, p-value is <2.2e-16 (Rejects)**		**W=49382000, p-value is <2.2e-16 (Rejects)**		**W=13888000, p-value is <2.2e-16 (Rejects)**		**W=85153000, p-value is <2.2e-16 (Rejects)**	
Exon Length (Mean)	28888.34	1549.01	9547.81	8219.43	1724.59	1515.29	6726.55	4946.22
Exon Length (Median)	1467.5	707	6302	4155.5	1357	904	4373	3312
Wilcoxon rank-sum test (Exon Length)	**W=46681000, p-value is <2.2e-16 (Rejects)**		**W=46554000, p-value is <2.2e-16 (Rejects)**		**W=11176000, p-value is <2.2e-16 (Rejects)**		**W=67659000, p-value is <2.2e-16 (Rejects)**	

Table S5. Dataset of intron and exon length for human and mouse. Result of Wilcoxon rank-sum test calculated for genes with null hypothesis that the median of intron and exon length for gene with low TC is less than with high TC (p-value < 0.05) for human and mouse.

Organism	lncRNA			mRNA		
5’ Splice Site	AT	GC	GT	AT	GC	GT
Low Transcript Complexity						
Human (GENCODE v38)	0.02	1.21	62.19	0.14	0.74	81.39
Mouse (GENCODE M28)	0.01	0.96	43.19	0.12	0.70	84.54
High Transcript Complexity						
Human (GENCODE v38)	0.01	1.17	35.29	0.03	0.14	17.40
Mouse (GENCODE M28)	0.02	1.38	53.81	0.02	0.14	14.36

3’ Splice Site	AC	AG	AC	AG
Low Transcript Complexity				
Human (GENCODE v38)	0.01	63.38	0.13	82.12
Mouse (GENCODE M28)	0.04	44.12	0.10	85.24
High Transcript Complexity				
Human (GENCODE v38)	0.00	36.48	0.03	17.54
Mouse (GENCODE M28)	0.19	55.18	0.01	14.50

Table S6. Percentage of introns having different 5’ and 3’ splice sites with low and high TC for human and mouse.

5’ Splice Site Strength				
Scoring Model	**Sample 1**	**Sample 2**	**Kolmogorov-Smirnov test**	**Wilcoxon rank-sum test**
Maximum Entropy Model (MAXENT)	**lncRNA (low TC)**	**lncRNA (high TC)**	D = 0.06, p-value is <2.2e-16 (Rejects)	W=309670000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**lncRNA (low TC)**	D = 0.04, p-value is <2.2e-16 (Rejects)	W=2.98e+09, p-value is <2.2e-16 (Rejects)
	mRNA (high TC)	**lncRNA (low TC)**	D = 0.04, p-value is <2.2e-16 (Rejects)	W=501910000, p-value is 7.266e-11 (Rejects)
	mRNA (low TC)	**lncRNA (high TC)**	D = 0.1, p-value is <2.2e-16 (Rejects)	W=2105600000, p-value is <2.2e-16 (Rejects)
	mRNA (high TC)	**lncRNA (high TC)**	D = 0.08, p-value is <2.2e-16 (Rejects)	W=354620000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**mRNA (high TC)**	D = 0.03, p-value is <2.2e-16 (Rejects)	W=3228800000, p-value is 9.782e-05 (Rejects)
Maximum Dependence Decomposition Model (MDD)	**lncRNA (low TC)**	**lncRNA (high TC)**	D = 0.06, p-value is <2.2e-16 (Rejects)	W=310210000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**lncRNA (low TC)**	D = 0.04, p-value is <2.2e-16 (Rejects)	W=2939700000, p-value is <2.2e-16 (Rejects)
	mRNA (high TC)	**lncRNA (low TC)**	D = 0.00, p-value is <2.2e-16 (Rejects)	W=493860000, p-value is 0.002245 (Rejects)
	mRNA (low TC)	**lncRNA (high TC)**	D = 0.09, p-value is <2.2e-16 (Rejects)	W=2084800000, p-value is <2.2e-16 (Rejects)
	mRNA (high TC)	**lncRNA (high TC)**	D = 0.07, p-value is <2.2e-16 (Rejects)	W=350420000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**mRNA (high TC)**	D = 0.02, p-value is 2.276e-14 (Rejects)	W=3239700000, p-value is 1.161e-06 (Rejects)
First-order Markov Model (MM)	**lncRNA (low TC)**	**lncRNA (high TC)**	D = 0.05, p-value is <2.2e-16 (Rejects)	W=305410000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**lncRNA (low TC)**	D = 0.05, p-value is <2.2e-16 (Rejects)	W=3007900000, p-value is <2.2e-16 (Rejects)
	mRNA (high TC)	**lncRNA (low TC)**	D = 0.05, p-value is <2.2e-16 (Rejects)	W=514570000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**lncRNA (high TC)**	D = 0.1, p-value is <2.2e-16 (Rejects)	W=2093600000, p-value is <2.2e-16 (Rejects)

	mRNA (high TC)	**lncRNA (high TC)**	D = 0.09, p-value is <2.2e-16 (Rejects)	W=357730000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**mRNA (high TC)**	D = 0.02, p-value is 3.897e-14 (Rejects)	W=3173700000, p-value is 0.9029 (Accepts)
Weight Matrix Model (WMM)	**lncRNA (low TC)**	**lncRNA (high TC)**	D = 0.05, p-value is <2.2e-16 (Rejects)	W=307010000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**lncRNA (low TC)**	D = 0.03, p-value is <2.2e-16 (Rejects)	W=2946100000, p-value is <2.2e-16 (Rejects)
	mRNA (high TC)	**lncRNA (low TC)**	D = 0.03, p-value is 1.166e-10 (Rejects)	W=498420000, p-value is 5.742e-07 (Rejects)
	mRNA (low TC)	**lncRNA (high TC)**	D = 0.08, p-value is <2.2e-16 (Rejects)	W=2066300000, p-value is <2.2e-16 (Rejects)
	mRNA (high TC)	**lncRNA (high TC)**	D = 0.06, p-value is <2.2e-16 (Rejects)	W=349590000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**mRNA (high TC)**	D = 0.02, p-value is 1.157e-13 (Rejects)	W=3160600000, p-value is 0.9938 (Accepts)
3’ Splice Site Strength				
Scoring Model	**Sample 1**	**Sample 2**	**Kolmogorov-Smirnov test**	**Wilcoxon rank-sum test**
Maximum Entropy Model (MAXENT)	**lncRNA (low TC)**	**lncRNA (high TC)**	D = 0.04, p-value is <2.2e-16 (Rejects)	W=300070000, p-value is 1.589e-05 (Rejects)
	mRNA (low TC)	**lncRNA (low TC)**	D = 0.09, p-value is <2.2e-16 (Rejects)	W=319240000, p-value is <2.2e-16 (Rejects)
	mRNA (high TC)	**lncRNA (low TC)**	D = 0.11, p-value is <2.2e-16 (Rejects)	W=556460000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**lncRNA (high TC)**	D = 0.11, p-value is <2.2e-16 (Rejects)	W=2176500000, p-value is <2.2e-16 (Rejects)
	mRNA (high TC)	**lncRNA (high TC)**	D = 0.12, p-value is <2.2e-16 (Rejects)	W=378820000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**mRNA (high TC)**	D = 0.04, p-value is <2.2e-16 (Rejects)	W=3091800000, p-value is 1 (Accepts)
First-order Markov Model (MM)	**lncRNA (low TC)**	**lncRNA (high TC)**	D = 0.04, p-value is <2.2e-16 (Rejects)	W=302080000, p-value is 4.316e-12 (Rejects)
	mRNA (low TC)	**lncRNA (low TC)**	D = 0.07, p-value is <2.2e-16 (Rejects)	W=3099600000, p-value is <2.2e-16 (Rejects)

	mRNA (high TC)	**lncRNA (low TC)**	D = 0.09, p-value is <2.2e-16 (Rejects)	W=542440000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**lncRNA (high TC)**	D = 0.1, p-value is <2.2e-16 (Rejects)	W=2131500000, p-value is <2.2e-16 (Rejects)
	mRNA (high TC)	**lncRNA (high TC)**	D = 0.12, p-value is <2.2e-16 (Rejects)	W=372170000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**mRNA (high TC)**	D = 0.04, p-value is <2.2e-16 (Rejects)	W=3293400000, p-value is <2.2e-16 (Rejects)
Weight Matrix Model (WMM)	**lncRNA (low TC)**	**lncRNA (high TC)**	D = 0.04, p-value is 2.642e-14 (Rejects)	W=301150000, p-value is 2.348e-10 (Rejects)
	mRNA (low TC)	**lncRNA (low TC)**	D = 0.06, p-value is <2.2e-16 (Rejects)	W=3032200000, p-value is <2.2e-16 (Rejects)
	mRNA (high TC)	**lncRNA (low TC)**	D = 0.08, p-value is <2.2e-16 (Rejects)	W=534460000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**lncRNA (high TC)**	D = 0.08, p-value is <2.2e-16 (Rejects)	W=2083400000, p-value is <2.2e-16 (Rejects)
	mRNA (high TC)	**lncRNA (high TC)**	D = 0.1, p-value is <2.2e-16 (Rejects)	W=366320000, p-value is <2.2e-16 (Rejects)
	mRNA (low TC)	**mRNA (high TC)**	D = 0.04, p-value is <2.2e-16 (Rejects)	W=3311800000, p-value is <2.2e-16 (Rejects)

Table S7. Dataset of 5’ and 3’ splice site strength using different scoring models (MAXENT, MDD, MM, WMM) for human. Result of Kolmogorov-Smirnov test calculated for genes with null hypothesis that the cumulative distribution function for sample 1 and sample 2 are same ($p < 0.05$) for human. Result of Wilcoxon rank-sum test calculated for genes with null hypothesis that the median of splice site strength for gene in sample 1 is greater than with sample 2 (p-value < 0.05) for human.

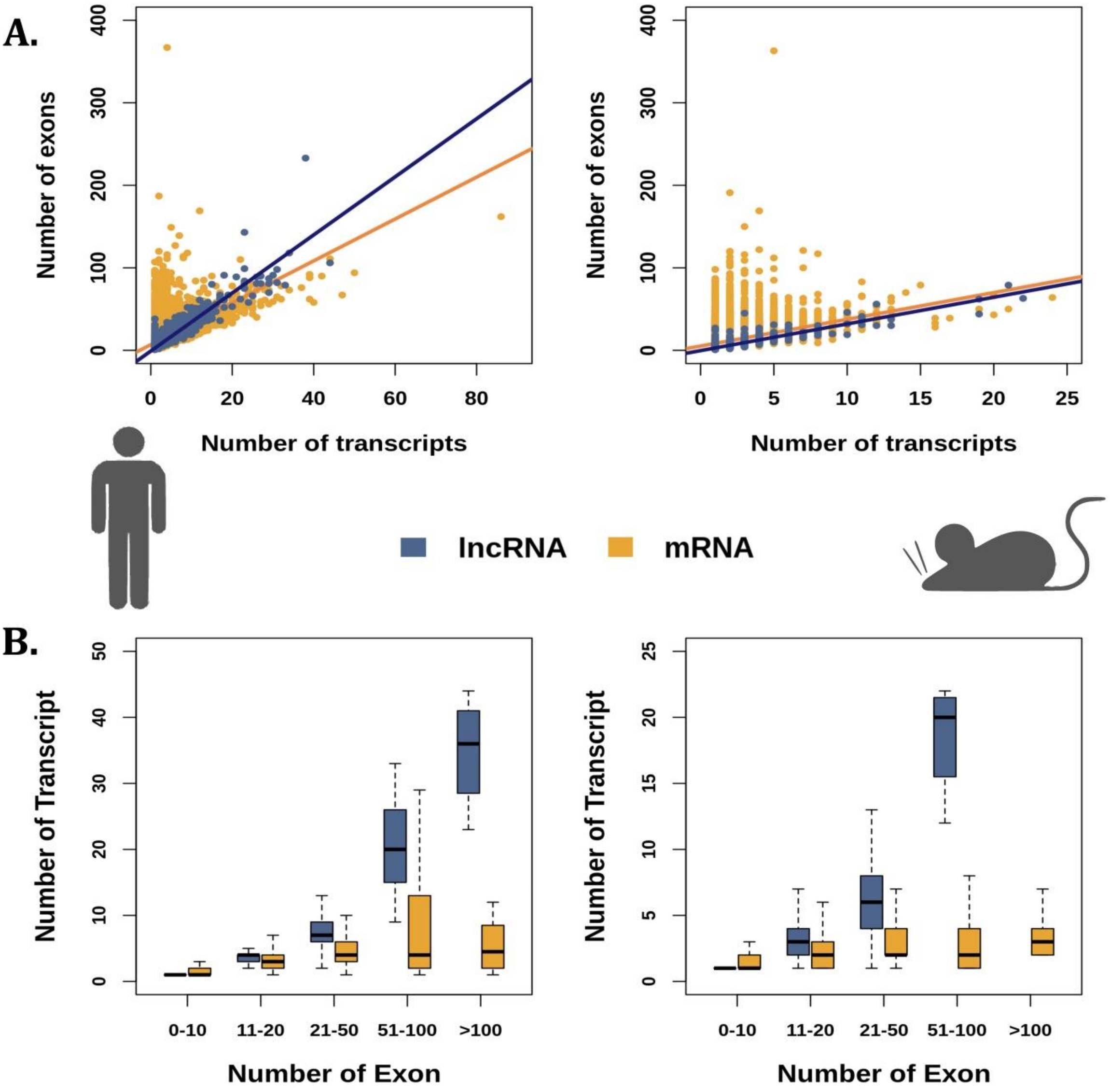

Figure S1. Depicts the correlation between Number of Transcripts and Number of Exons in lncRNA and mRNAs (A) as scatterplot and (B) as boxplot for Human (left hand side) and Mouse (right hand side)

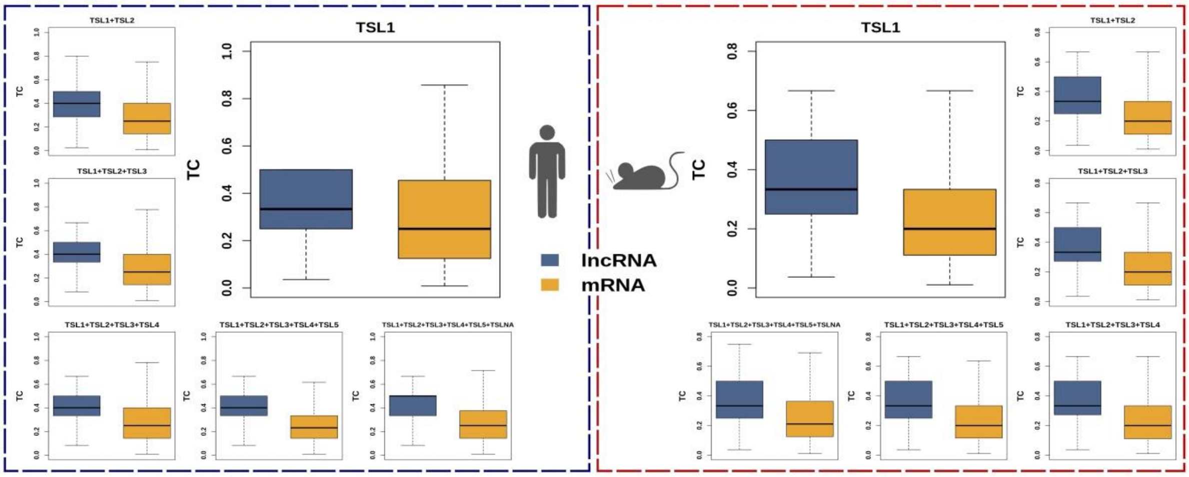

Figure S2. Representation of distribution of TC with different Transcript Support Levels for lncRNAs and mRNAs humans (left hand side or blue dashed box) and mouse (right hand side or red dashed box)

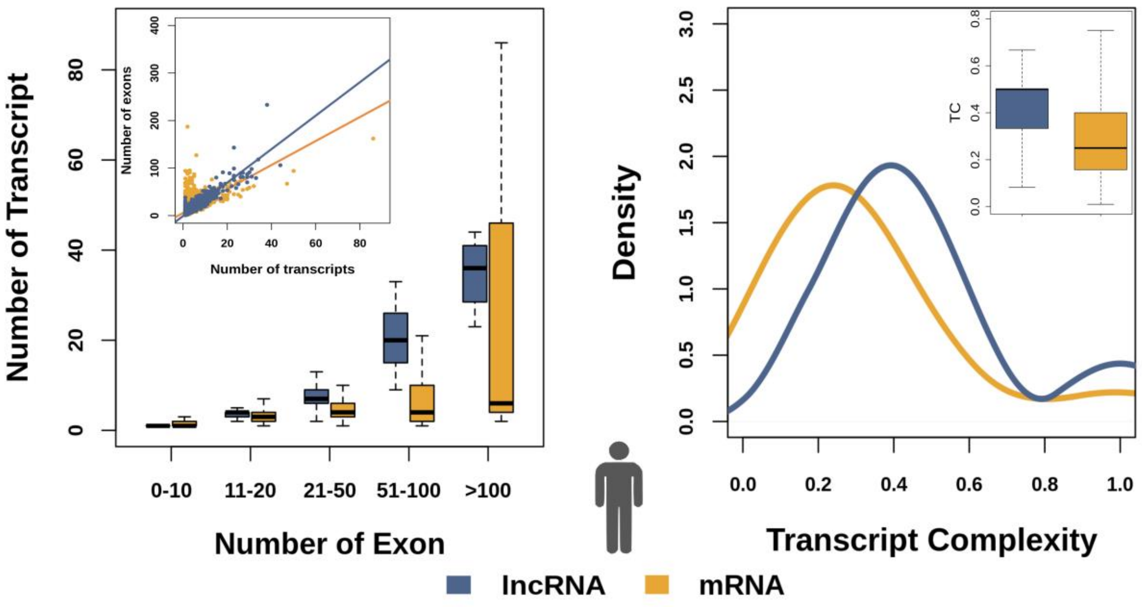

Figure S3. Depicts (A) the correlation between Number of Transcripts and Number of Exons in mRNAs (left hand side) (B) distribution of Transcript Complexity (TC) in mRNAs (right hand side) for human [Data consists of 102 eCLIP experiments in biological duplicate for a diverse collection of 74 RBPs in HepG2 and K562 cells (GSE80039_RAW.tar)]