

Supplementary Methods, Figures and Tables

Gene Body Architecture Predicts Isoform Diversity Across Eukaryotic Genomes

Quinton Brohart^{1,2}, Arvind Mer^{1,2,3,4*}

¹Department of Biochemistry, Microbiology & Immunology, University of Ottawa, Ottawa, Canada

²Ottawa Institute of Systems Biology, Ottawa, Ontario, Canada

³School of Electrical Engineering & Computer Science, University of Ottawa, Ottawa, Canada

⁴Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

Supplementary Methods

Exon and Intron Region Length Calculations

The GRCh38.p14 human reference genome was imported from the NCBI RefSeq database¹, with gene and transcript data extracted, then converted from GRanges objects to data frames. Exons were grouped and merged by gene for length and count calculations. Introns were inferred from exon spacing, and total exon-intron lengths and counts were normalized by gene length. When determining gene length, the coordinates of each gene were considered, and length was calculated based on the total number of base pairs (bp) within those regions. To validate whether intron and exon length and isoform diversity were confounded by exon number, we fit multiple linear regression models. These were utilized to predict isoform count from gene-normalized intronic or exonic length proportion while adjusting for exon count as a covariate. One model utilized intron proportion in combination with exon count to predict isoform count, and the other used exon proportion. This approach allowed the independent contribution of length proportion to be assessed after accounting for the linear effect of exon number.

3' and 5' Untranslated Region (UTR) Length Determination

To examine the relationship between isoform diversity and both 5' and 3' UTR length, transcript-level UTR data were integrated with a gene-level isoform dataset. For each gene, total UTR length was computed by summing transcript-specific UTR lengths. Gene-level average UTR length was then calculated by normalizing the total UTR length by relative gene length. Genes lacking UTR length data were removed. A subset of NC transcripts had an annotated CDS. These transcripts were primarily nonsense-mediated decay (NMD), retained-intron biotypes, which represent alternatively spliced products of loci with translation potential that are flagged as non-productively spliced at the transcript level. Non-coding biotypes lacking any annotated open reading frame were not utilized in this analysis. This analysis was performed for all four species of study.

GC Content and CpG Island Density

Genomic regions with high CpG dinucleotide frequencies were obtained from the UCSC Genome Table Browser^{2,3} and converted into GRanges objects using the GenomicRanges package (version 1.60.0). The countOverlaps function was used to quantify CpG islands within each gene's length, producing per-gene genomic coordinates, GC content, gene length, and associated CpG island counts. CpG island data were analyzed for human and mouse genomes only due to data availability. Gene-level GC content was estimated using genome annotations retrieved from Ensembl. For genes lacking GC content estimates, exon-intron genomic sequences were retrieved, and GC content was calculated as the percentage of guanine and cytosine nucleotides within each gene sequence. GC content data was analyzed for all four species of study.

Protein-Protein Interaction Networks

Protein interaction networks were generated using the STRINGdb package (version 12.0; min score = 400) for human (species 9606) and mouse (species 10090) based on mapped gene names. Isoform counts were linked to both source and partner proteins, and metrics including the total number of isoforms among a gene's interaction partners and average isoforms per interaction.

Aggregated Isoform Count Plotting

To visualize the overall relationship between isoform count and the corresponding variables, LOESS-smoothed trend lines were generated using method = "loess" in the R package ggplot2. This visualization method was applied to intron/exon and GC content data.

Cross Species Random Forest Modelling

To assess the robustness of cross-species generalization beyond a single data partition, we extended the four-species random forest framework to a repeated-validation design using the reduced set of six features shared across all species (gene length, 5' and 3' UTR length, exon length, intron length, and GC content). For each species, the dataset was randomly partitioned into training (80%) and held-out (20%) subsets across four independent iterations, each with an independently seeded split. A random forest model was trained on each iteration's training partition and evaluated for within-species performance on its corresponding held-out subset. To evaluate cross-species generalization, each of the four models trained per species was subsequently applied to the entire dataset of every other species. Cross-species predictions were never restricted to a held-out subset of the target species, so as to maximize statistical power for external validation. For every train-test pair and iteration, we computed Pearson's r , Spearman's ρ , RMSE, raw R^2 , and a linearly recalibrated R^2 correcting for species-level differences. Metrics were summarized across the four iterations as mean $\pm$ SD.

Supplementary Figures

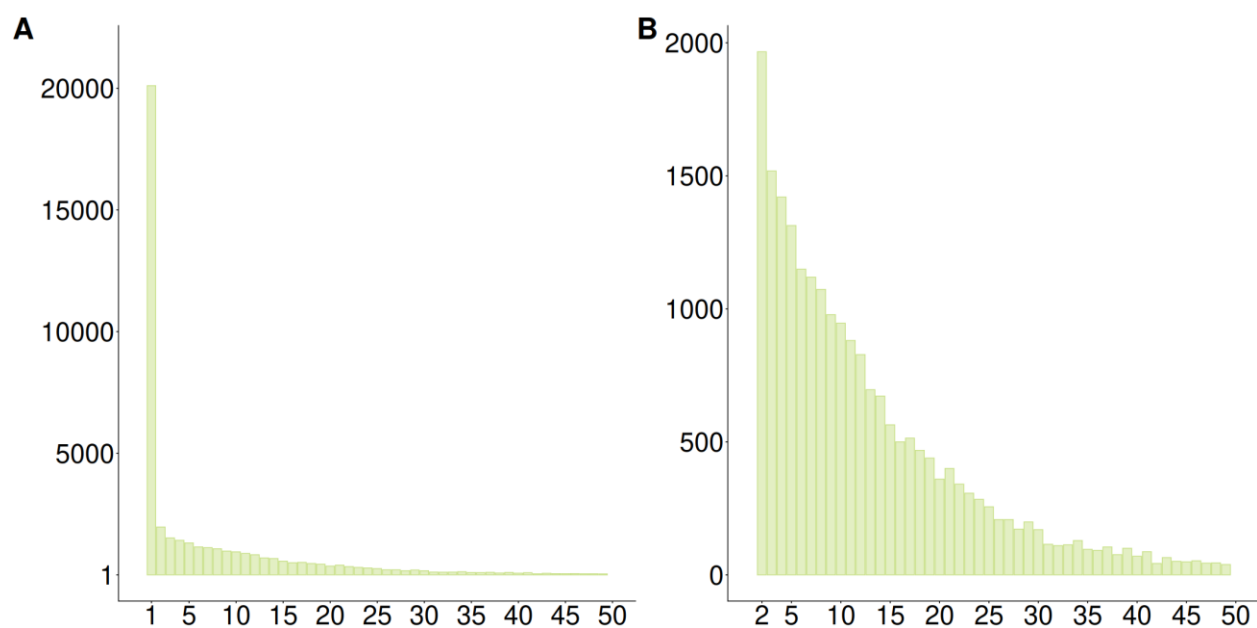


Figure 1: Gene totals by respective isoform count are shown for the human genome. Panel (A) displays total genes across all isoform counts, while panel (B) highlight genes with more than two isoforms. Isoform counts are capped at 50 to best illustrate species-specific isoform distributions.

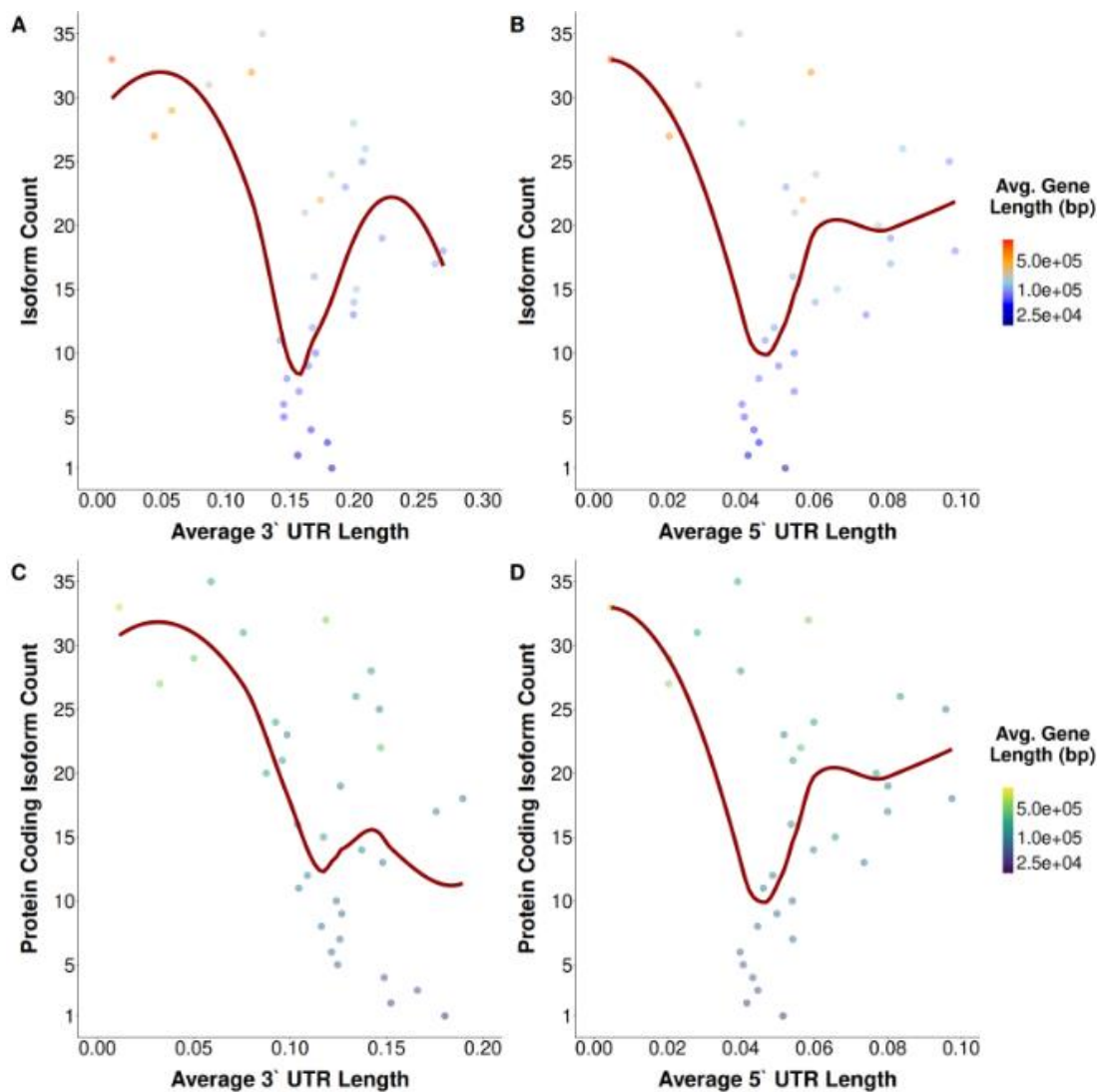


Figure 2: Isoform counts per gene in relation to 3' and 5' UTRs across the mouse genome. Panels (A-B) show normalized 3' (A) and 5' (B) UTR length (relative to gene length), averaged by respective isoform count. Panels (C-D) display the relationship between protein coding isoform counts and 3' (C) and 5' (D) UTR lengths. Red trend lines were generated via the local polynomial regression fitting (loess) smoothing function to demonstrate the relationship between isoform count and respective variables.

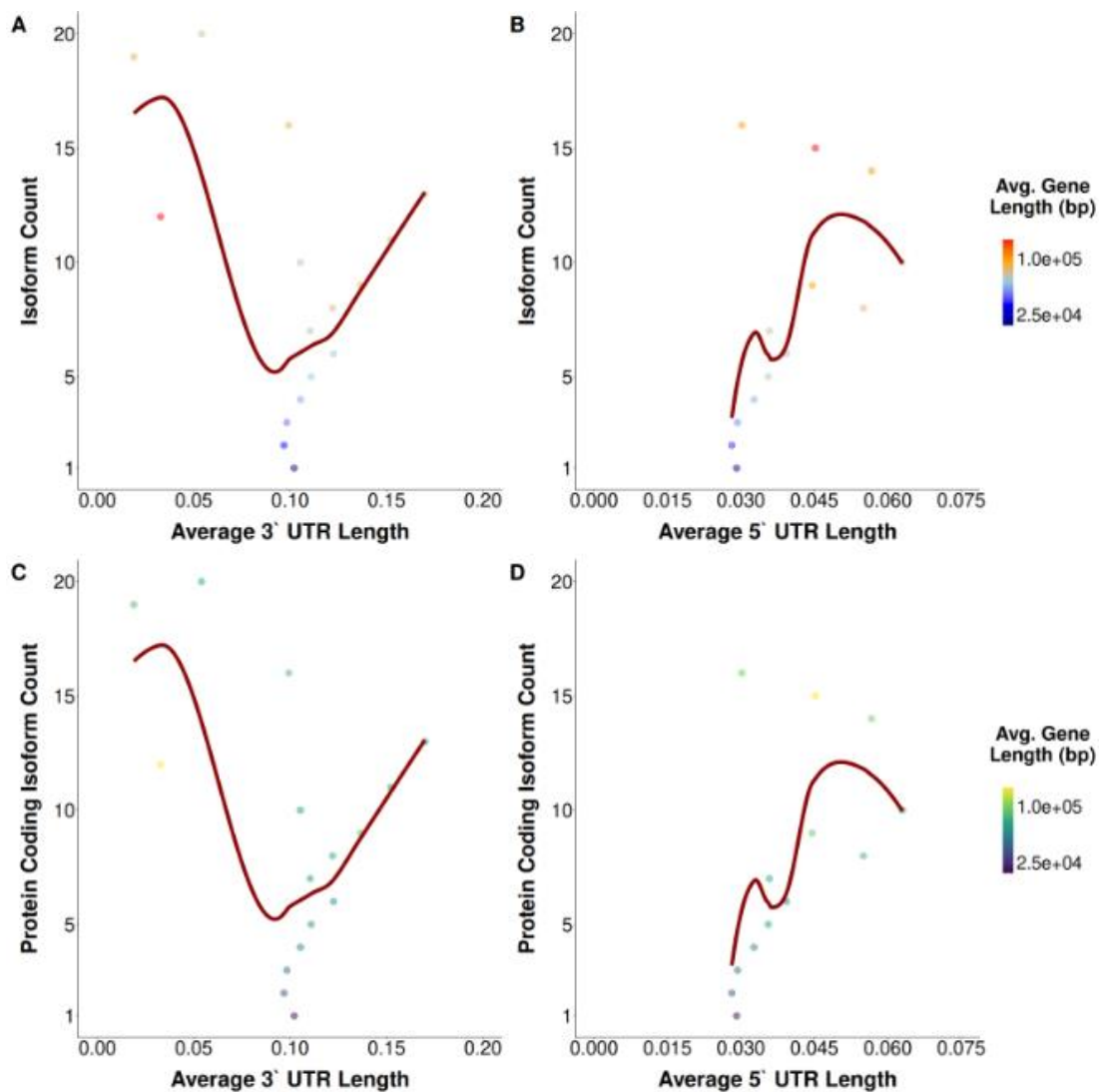


Figure 3: Isoform counts per gene in relation to 3' and 5' UTRs across the zebrafish genome. Panels (A-B) show normalized 3' (A) and 5' (B) UTR length (relative to gene length), averaged by respective isoform count. Panels (C-D) display the relationship between protein coding isoform counts and 3' (C) and 5' (D) UTR lengths. Red trend lines were generated via the local polynomial regression fitting (loess) smoothing function to demonstrate the relationship between isoform count and respective variables.

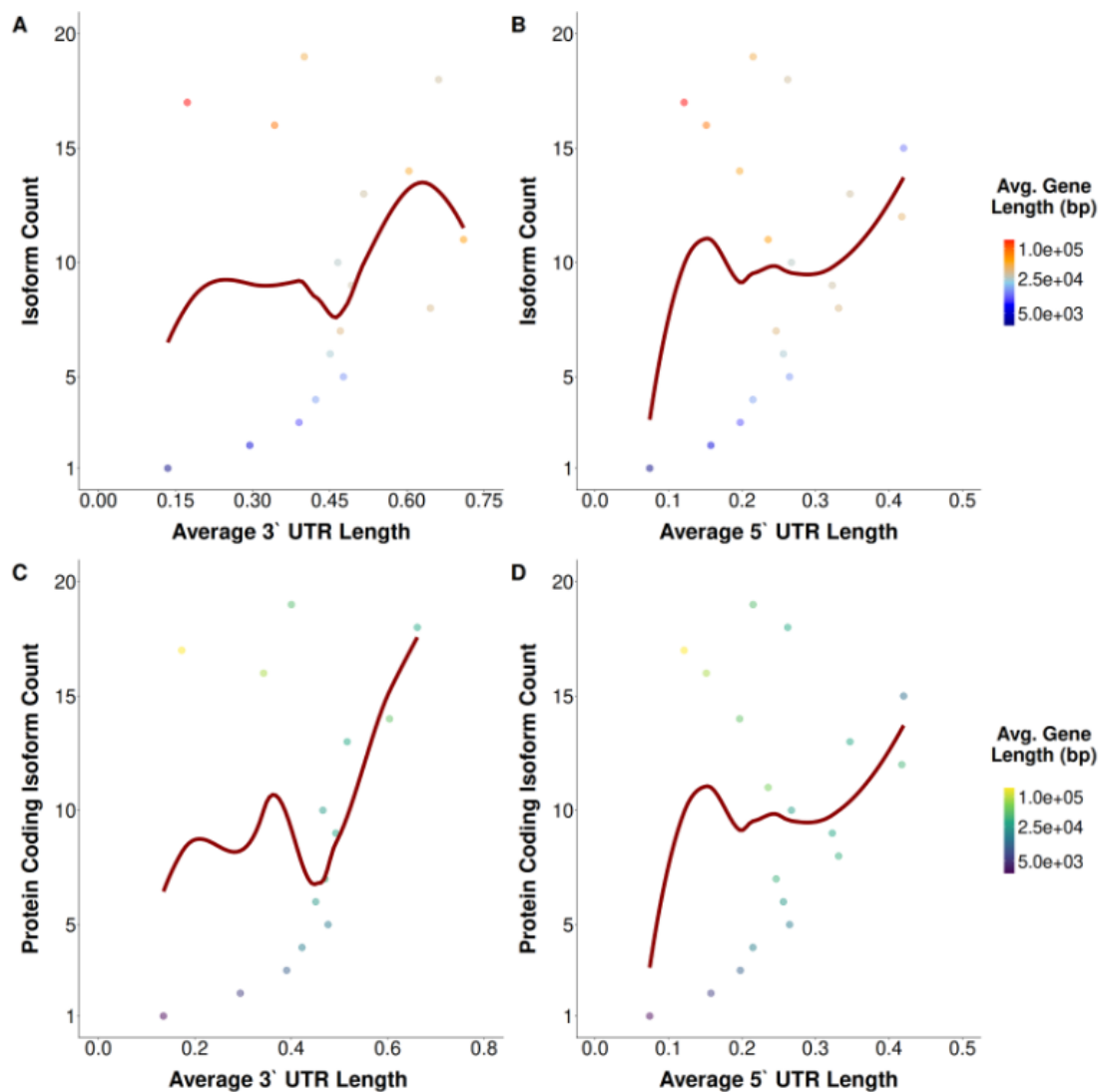


Figure 4: Isoform counts per gene in relation to 3' and 5' UTRs across the *Drosophila* genome. Panels (A-B) show normalized 3' (A) and 5' (B) UTR length (relative to gene length), averaged by respective isoform count. Panels (C-D) display the relationship between protein coding isoform counts and 3' (C) and 5' (D) UTR lengths. Red trend lines were generated via the local polynomial regression fitting (loess) smoothing function to demonstrate the relationship between isoform count and respective variables.

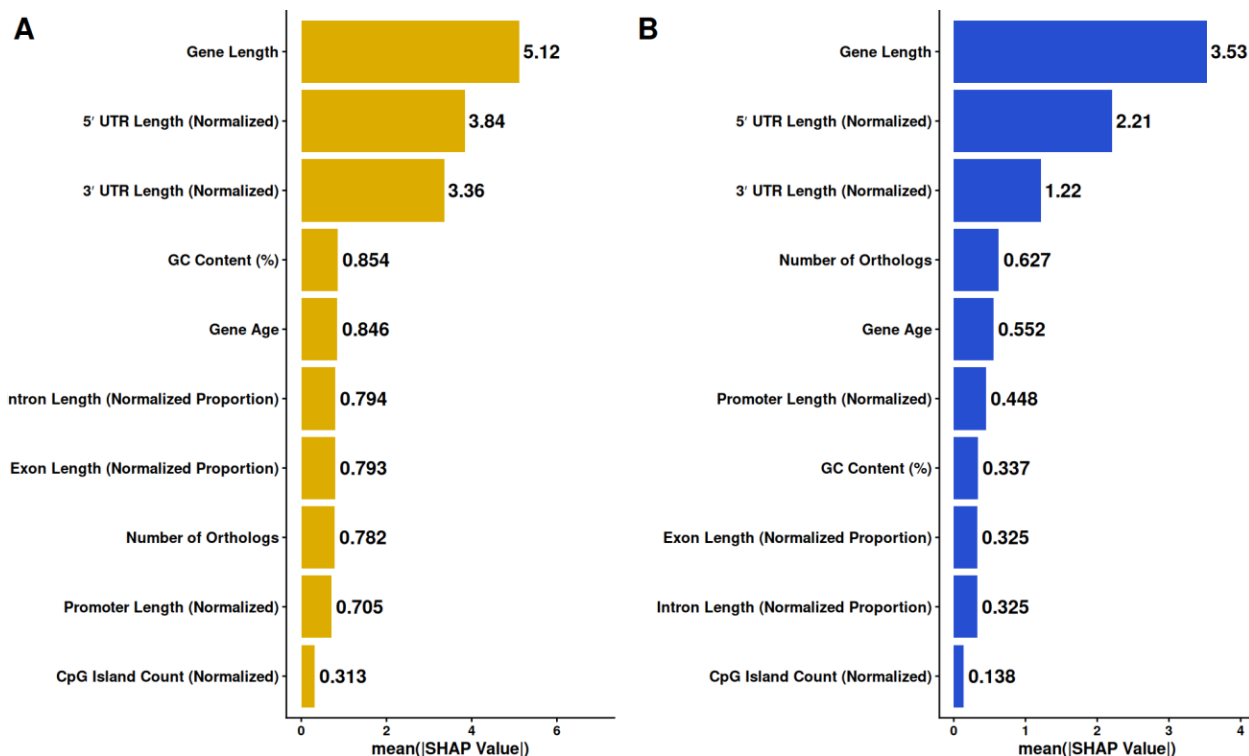


Figure 5: SHAP value plots showing the contribution of individual features to the random forest model used to predict isoform count across the human (A) in yellow and mouse (B) in blue, datasets. Each bar represents a feature used in the model to predict isoform count. SHAP value reflects the global importance of each respective feature on the predictive capacity of the model, reflecting the mean absolute value of the respective feature. Gene length and 5' and 3' UTR lengths showed the greatest overall contributions to the model.

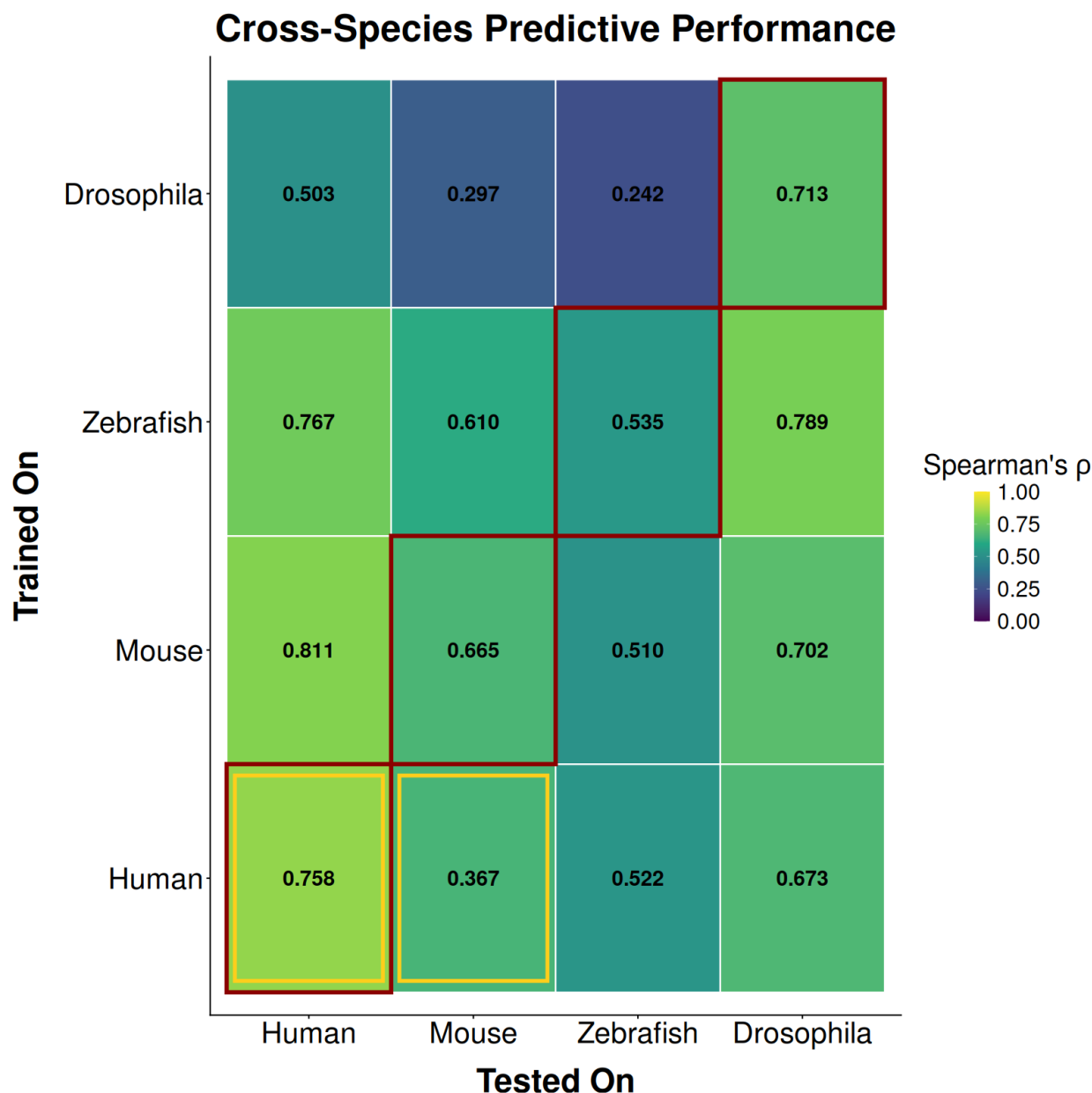


Figure 6: Cross-species predictive performance of a smaller random forest models trained on gene structural and compositional features. Tiles show the performance of a random forest model trained on one species and evaluated on another. Diagonal tiles (outlined in dark red) show within-species performance on a held-out 20% partition, while off-diagonal tiles show cross-species performance evaluated on the complete dataset of the target species. For cross-species analysis, models were trained on common features between species. Tile color reflects the mean Spearman's ρ between predicted and observed isoform counts across four independent training/held-out splits per species. Numeric labels give the range of the Spearman correlation p-value observed across those same four splits.

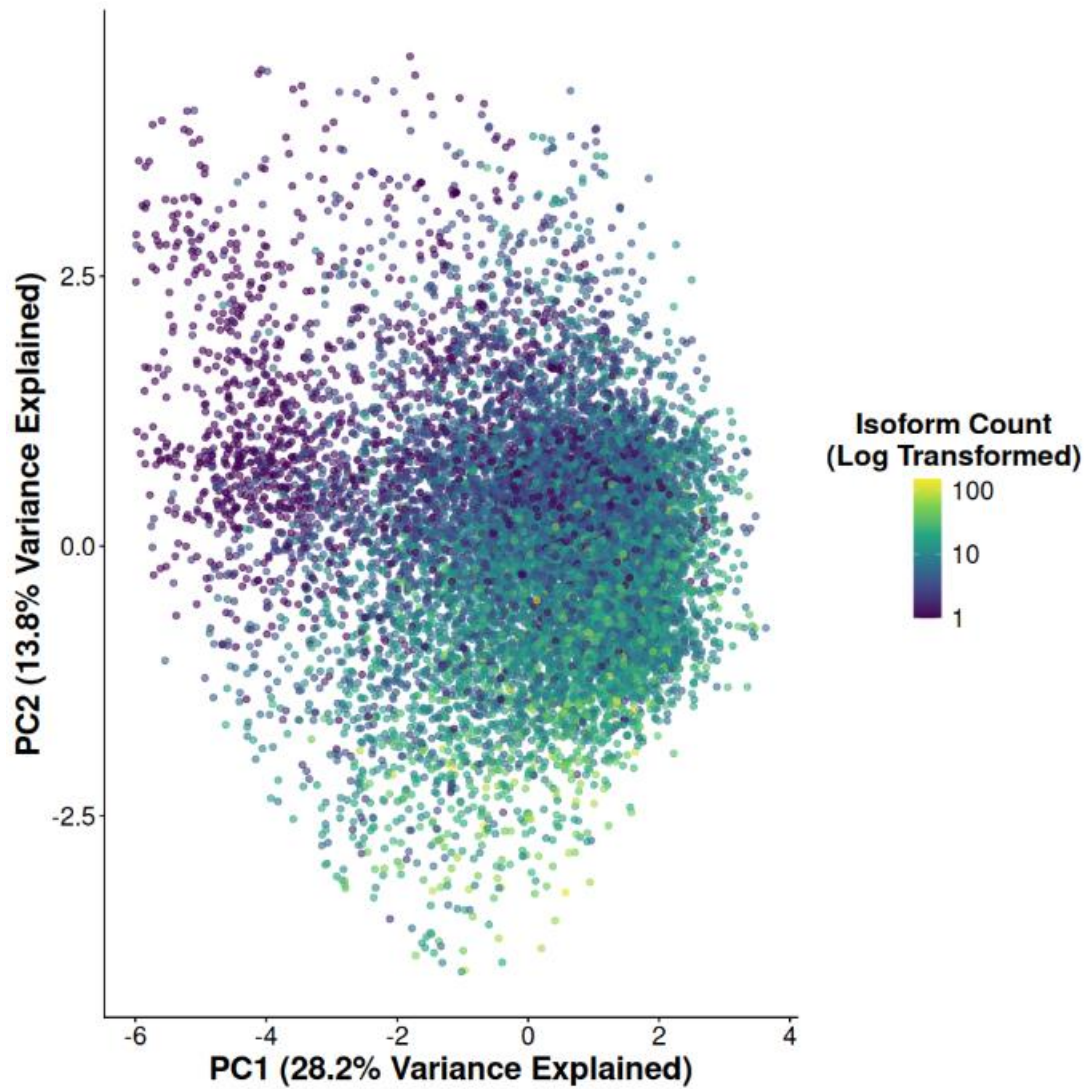


Figure 7: PCA of gene features related to isoform diversity. The PCA plot displays genes projected onto two principal components (PC1 and PC2), summarizing the variance in structural, regulatory, and evolutionary features. Each point represents a gene, colored by isoform count, which has been log transformed. The plot reveals distinct subgroups, with a majority cluster of genes with 1 isoform and smaller clusters corresponding to genes with higher isoform counts.

Supplementary Tables

Table 1: Gene and transcript count across the human (*Homo sapiens*), mouse (*Mus musculus*), *Drosophila* (*Drosophila melanogaster*), and zebrafish (*Danio rerio*) genomes. The number of genes per respective isoform count is presented in both raw counts and percentages

	Gene Total	Transcript Total	Genes with 1 Isoform	Genes with ≥ 2 Isoforms	Genes with > 2 Isoforms	Genes with > 5 Isoforms
<i>Homo sapiens</i>	42,389	357,675	20107	22282	20316	16065
			47.43 %	52.57 %	47.93 %	37.90 %
<i>Mus musculus</i>	77,097	273,169	43,050	34,047	25,069	12,630
			55.84 %	44.16 %	32.52 %	16.38 %
<i>Drosophila melanogaster</i>	23,723	41,068	16,340	7,428	3,448	823
			68.69 %	31.31 %	14.53 %	3.47 %
<i>Danio rerio</i>	29,250	55,746	15,597	13,653	6,352	812
			53.32 %	46.68 %	21.72 %	2.78 %

Table 2: Statistical relationship between isoform count and multiple cellular and biomolecular properties present within the human genome. Data has been averaged (Avg.) and aggregated (Agg.) across the properties to demonstrate the association per isoform count.

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
Avg. Protein-Protein Interactions (PPIs)	0.2974363 (0)	0.4120701 (0)	0.2887855 (0)	28.19 ($<2e-16$)	0	6.8777 ($<2.2e-16$)
Avg. Protein-Protein Interactions (PPIs) (Agg.)	0.3201288 (0.0001221422)	0.3556236 (1.97406e-05)	0.3024711 (1.273963e-07)	NaN (NaN)	0.4839898	NaN (NaN)
Avg. CpG Islands within Gene Regions (Agg.)	-0.204001 (0.007992997)	-0.5998046 (0)	-0.4488167 (5.823301e-18)	NaN (NaN)	0.485	NaN (NaN)

Avg. GC Content within Gene Regions (Agg.)	-0.1692849 (0.0212469)	-0.2755676 (0.0001553158)	-0.2065805 (2.977567e-05)	NaN (NaN)	0.486	NaN (NaN)
3' UTR Sequence Avg. Length (Agg)	0.2555888 (0.002304899)	0.5171634 (6.066093e-11)	0.44111 (1.057179e-14)	NaN (NaN)	0.484	NaN (NaN)
5' UTR Sequence Avg. Length (Agg)	0.2578092 (0.002103267)	0.4447793 (3.678895e-08)	0.3786228 (3.204567e-11)	NaN (NaN)	0.484	NaN (NaN)
Intron Regions Avg. Counts (Agg)	0.1622055 (0.02913903)	0.05794489 (0.4384465)	0.0665151 (0.1850807)	NaN (NaN)	0.486	NaN (NaN)
Exon Regions Avg. Counts (Agg)	0.1041588 (0.162894)	0.003483261 (0.9628815)	0.03215095 (0.5218123)	NaN (NaN)	0.486	NaN (NaN)

Table 3: Statistical relationship between isoform count and multiple cellular and biomolecular properties present within the human genome. Data has been averaged (Avg.) and aggregated (Agg.) across the properties to demonstrate the association per isoform count, subset by biotype: protein coding [PC] and non-protein coding [NC].

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
[PC] Avg. Protein-Protein Interactions	0.3022586 (0)	0.4153316 (0)	0.2909355 (0)	29.5 (<2e-16)	0	615 (< 2.2e-16)
[C] Avg. Protein-Protein Interactions (PPIs) (Agg.)	0.3175681 (0.0001393301)	0.3593235 (1.5984e-05)	0.3037222 (1.130418e-07)	NaN (NaN)	0.4839898	NaN (NaN)
[NC] Avg. Protein-Protein Interactions	0.0003801812 (0.998695)	0.2794503 (0.2198937)	0.2548236 (0.1899633)	0.083 (0.000659)	0.27	0.211 (0.9673)
[C] Avg. Protein-Protein Interactions (Agg.)	-0.3378838 (0.4585652)	-0.5225437 (0.2288779)	-0.19518 (0.5434239)	NaN (NaN)	0.4231901	NaN (NaN)
[PC] Avg. CpG islands within Gene Regions (Agg.)	-0.182562 (0.03838128)	-0.4492062 (1.298106e-07)	-0.3415698 (9.397151e-09)	NaN (NaN)	0.483	NaN (NaN)

NC] Avg. CpG islands within Gene Regions (Agg.)	-0.1651015 (0.04348363)	-0.4984168 (1.05086e-10)	-0.3544519 (1.222284e-10)	NaN (NaN)	0.485	NaN (NaN)
PC] Avg. GC Content within Gene Regions (Agg.)	-0.1492897 (0.09259003)	-0.1573773 (0.07607613)	-0.1227854 (0.03982059)	NaN (NaN)	0.483	NaN (NaN)
NC] Avg. GC Content within Gene Regions (Agg.)	-0.007758323 (0.9256946)	-0.1280629 (0.1220708)	-0.1096822 (0.04869512)	NaN (NaN)	0.484	NaN (NaN)
[PC] 3' UTR Sequence Avg. Length (Agg.)	0.2556046 (0.002303408)	0.5175526 (5.836044e-11)	0.4413155 (1.027662e-14)	NaN (NaN)	0.484	NaN (NaN)
[NC] 3' UTR Sequence Avg. Length (Agg.)	-0.3669501 (0.6330499)	-0.4 (0.6)	-0.3333333 (0.4969058)	NaN (NaN)	0.392	NaN (NaN)
[PC] 5' UTR Sequence Avg. Length (Agg.)	0.2577816 (0.002105671)	0.4447793 (3.678895e-08)	0.3786228 (3.204567e-11)	NaN (NaN)	0.484	NaN (NaN)
[NC] 5' UTR Sequence Avg. Length (Agg.)	-0.3669501 (0.6330499)	0.4447793 (3.678895e-08)	NaN (NaN)	NaN (NaN)	0.484	NaN (NaN)

Table 4: Statistical data and relationship between isoform count and multiple cellular and biomolecular properties present within the human genome. Data is either raw or normalized by gene length, depending on the property.

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
Raw Protein-Protein Interactions	0.221911 (1.009654e-210)	0.362455 (0)	0.2511027 (0)	12.18 (<2e-16)	0	6.7746 (< 2.2e-16)
3' UTR Sequence Length	0.3660974 (0)	0.3279191 (0)	0.2269565 (0)	83.92 (<2e-16)	0	14.667 (< 2.2e-16)
5' UTR Sequence Length	0.3344119 (0)	0.3367959 (0)	0.2317197 (0)	43.93 (<2e-16)	0	17.393 (< 2.2e-16)

Intron Regions - Lengths	0.3543661 (0)	0.5694303 (0)	0.4181802 (0)	192.9 (<2e-16)	0	8.0551 (< 2.2e-16)
Intron Regions - Raw Counts	0.3486774 (0)	0.7239294 (0)	0.5652179 (0)	63.51 (< 2e-16)	0	23.624 (< 2.2e-16)
Exon Regions - Lengths	-0.3236068 (0)	-0.5733808 (0)	-0.4206762 (0)	138.3 (<2e-16)	0	5.894 (< 2.2e-16)
Exon Regions - Raw Counts	0.3491856 (0)	0.7307922 (0)	0.5703224 (0)	66.01 (<2e-16)	0	25.371 (< 2.2e-16)
CpG Islands within Gene Regions	-0.08640544 (2.152804e-22)	-0.2682764 (2.119587e-207)	-0.1855676 (4.303165e-208)	11.07 (<2e-16)	1.2e-219	10.55 (< 2.2e-16)
GC Content within Gene Regions	-0.06006354 (4.091866e-30)	-0.1055639 (1.166774e-89)	-0.07398131 (1.980846e-88)	4.426 (<2e-16)	4.95e-88	3.5619 (< 2.2e-16)

Table 5: Statistical data and relationship between isoform count and multiple cellular and biomolecular properties present within the human genome. Data is either raw or normalized by gene length, depending on the property, and subset by biotype: protein coding [PC] and non-protein coding [NC].

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
[PC] Raw Protein-Protein Interactions (PPIs)	0.1764939 (2.280481e-132)	0.2826785 (0)	0.1941716 (0)	7.91 (<2e-16)	1.12e-251	4.5114 (< 2.2e-16)
[NC] Raw Protein-Protein Interactions (PPIs)	-0.1504069 (0.5151987)	-0.1173157 (0.6125454)	-0.1087321 (0.5941207)	0.454 (0.831)	0.619	0.3678 (0.8874)
[PC] 3' UTR Sequence Length	0.3661195 (0)	0.3281542 (0)	0.2271368 (0)	83.87 (<2e-16)	0	14.655 (< 2.2e-16)
[NC] 3' UTR Sequence Length	-0.2535391 (0.5103685)	-0.2568915 (0.5045974)	-0.2333333 (0.4168451)	0.422 (0.746)	0.234	0.2712 (0.8441)
[PC] 5' UTR Sequence Length	0.3343447 (0)	0.3369007 (0)	0.2317934 (0)	43.9 (<2e-16)	0	17.379 (< 2.2e-16)

[NC] 5' UTR Sequence Length	NaN (NaN)	NaN (NaN)	NaN (NaN)	0.422 (0.746)	0.234	0.2712 (0.8441)
[PC] CpG Islands in Gene Regions	-0.1086295 (9.050461e-27)	-0.1416262 (1.705463e-44)	-0.0964483 (1.537124e-44)	8.032 (<2e-16)	8.53e-38	7.1624 (< 2.2e-16)
[PC] GC Content in Gene Regions	-0.0198986 (0.05247605)	-0.02452572 (0.01683621)	-0.01677825 (0.01578458)	1.486 (0.000347)	0.00539	1.4387 (0.0009728)
[NC] CpG Islands in Gene Regions	-0.121061 (3.350714e-11)	-0.5499433 (3.307331e-235)	-0.3983877 (1.716304e-217)	3.393 (<2e-16)	1.86e-141	3.3078 (< 2.2e-16)
[NC] GC Content in Gene Regions	-0.1225423 (2.172199e-11)	-0.2008935 (2.308792e-28)	-0.1392832 (5.113043e-28)	1.814 (2.2e-08)	5.34e-07	2.2681 (3.634e-15)

Table 6: Multiple linear regression of isoform count on intronic proportion (Model 1), adjusted for exon count. Beta = unstandardized regression coefficient; CI = confidence interval.

Model 1: Intron Proportion			
Characteristic	Beta	95% CI	p-value
Intron Proportion	11	10, 11	<0.001
Exon Count	0.37	0.35, 0.39	<0.001
Abbreviation: CI = Confidence Interval			
R² = 0.167; Adjusted R² = 0.167; N = 31,543			

Table 7: Multiple linear regression of isoform count on exonic proportion (Model 2), adjusted for exon count. Beta = unstandardized regression coefficient; CI = confidence interval.

Model 2: Exon Proportion

Characteristic	Beta	95% CI	p-value
----------------	------	--------	---------

Exon Proportion	-8.2	-8.7, -7.8	<0.001
-----------------	------	------------	--------

Exon Count	0.41	0.39, 0.42	<0.001
------------	------	------------	--------

Abbreviation: CI = Confidence Interval

$R^2 = 0.157$; Adjusted $R^2 = 0.157$; N = 31,543

Table 8: Statistical relationship between isoform count and multiple cellular and biomolecular properties present within the mouse genome. Data has been averaged (Avg.) and aggregated (Agg.) across the properties to demonstrate the association per isoform count.

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
Avg. Protein-Protein Interactions (PPIs)	0.3102822 (0)	0.3812323 (0)	0.2751124 (0)	68.04 (<2e-16)	0	19.296 (< 2.2e-16)
Avg. Protein-Protein Interactions (PPIs) (Agg.)	0.4951826 (0.0007355408)	0.5711265 (8.519419e-05)	0.4839424 (1.930704e-06)	NaN (NaN)	0.4709744	NaN (NaN)
3' UTR Sequence Avg. Length (Agg)	0.1817311 (0.2434859)	-0.1347025 (0.3891225)	-0.04318937 (0.6831625)	NaN (NaN)	0.471	NaN (NaN)
5' UTR Sequence Avg. Length (Agg)	0.2353195 (0.1287513)	-0.1138629 (0.4672157)	-0.00110742 (0.99165)	NaN (NaN)	0.471	NaN (NaN)
Intron Regions - Avg. Counts (Agg)	-0.01394081 (0.8692088)	-0.2670182 (0.001316783)	-0.2045098 (0.0003463329)	NaN (NaN)	0.484	NaN (NaN)
Exon Regions - Avg. Counts (Agg)	-0.1124776 (0.1826265)	-0.2931002 (0.0004002927)	-0.2182972 (0.0001353421)	NaN (NaN)	0.484	NaN (NaN)
Avg. CpG Islands within Gene Regions (Agg.)	-0.2135121 (0.02874674)	-0.6597657 (0)	-0.5091575 (1.387387e-14)	NaN (NaN)	0.482	NaN (NaN)

Avg. GC Content within Gene Regions (Agg.)	-0.3093742 (0.0002344966)	-0.3695792 (1.022196e-05)	-0.2769429 (1.579414e-06)	NaN (NaN)	0.484	NaN (NaN)
--	------------------------------	------------------------------	------------------------------	-----------	-------	-----------

Table 9: Statistical relationship between isoform count and multiple cellular and biomolecular properties present within the mouse genome. Data has been averaged (Avg.) and aggregated (Agg.) across the properties to demonstrate the association per isoform count, subset by biotype: protein coding [PC] and non-protein coding [NC].

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
[PC] Avg. Protein-Protein Interactions (PPIs)	0.3208414 (0)	0.3825407 (0)	0.2753745 (0)	71.69 (<2e-16)	0	16.912 (< 2.2e-16)
[PC] Avg. Protein-Protein Interactions (PPIs) (Agg.)	0.401453 (0.007625737)	0.5792812 (6.44498e-05)	0.4972315 (9.260034e-07)	NaN (NaN)	0.4709744	NaN (NaN)
[NC] Avg. Protein-Protein Interactions (PPIs)	0.3779649 (1.280873e-05)	0.06338296 (0.4807538)	0.05251701 (0.4819253)	NaN (NaN)	0.0671	NaN (NaN)
[NC] Avg. Protein-Protein Interactions (PPIs) (Agg.)	0.7600671 (0.01072747)	0.3751832 (0.2853823)	0.3103164 (0.2280917)	NaN (NaN)	0.4372742	NaN (NaN)
[PC] 3' UTR Sequence Avg. Length (Agg)	0.2116703 (0.1730048)	-0.3245243 (0.03373882)	-0.2580288 (0.01475069)	NaN (NaN)	0.471	NaN (NaN)
[NC] 3' UTR Sequence Avg. Length (Agg)	-	-	-	-	-	-
[PC] 5' UTR Sequence Avg. Length (Agg)	0.2351417 (0.1290488)	-0.1090305 (0.4864463)	0.005537099 (0.9582681)	NaN (NaN)	0.471	NaN (NaN)
[NC] 5' UTR Sequence Avg. Length (Agg)	-	-	-	-	-	-

[PC] Avg. CpG Islands within Gene Regions (Agg.)	0.06153884 (0.7060087)	-0.8268293 (0)	-0.7205128 (1.51686e-13)	NaN (NaN)	0.47	NaN (NaN)
[NC] Avg. CpG Islands within Gene Regions (Agg.)	-0.2319128 (0.01728899)	-0.7056396 (0)	-0.5468864 (1.363328e-16)	NaN (NaN)	0.482	NaN (NaN)
[PC] Avg. GC Content within Gene Regions (Agg.)	-0.7969505 (2.145264e-09)	-0.8325856 (1.780358e-08)	-0.655761 (2.250489e-10)	NaN (NaN)	0.469	NaN (NaN)
[NC] Avg. GC Content within Gene Regions (Agg.)	-0.3626556 (0.0002251709)	-0.226444 (0.02440193)	-0.1622346 (0.0173499)	NaN (NaN)	0.481	NaN (NaN)

Table 10: Statistical data and relationship between isoform count and multiple cellular and biomolecular properties present within the mouse genome. Data is either raw or normalized by gene length depending on the property

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
Raw Protein- in Interactions (PPIs)	61779 (4.516963e-59)	0.2135689 (0.035001e-198)	05689 (6.872268e-195)	27 (<2e-16)	574495e-168	93 (< 2.2e-16)
3' UTR ence Length	891531(0.4775938)	0.01017336 (0.1396619)	836527(0.1049515)	309 (0.0869)	9.33e-06	2055 (0.17)
5' UTR ence Length	0.02030219 (0.003261902)	0.02316195 (0.0007895983)	0.0164839 (0.0006642469)	111 (0.287)	1.24e-12	167 (0.6253)
Intron Regions - Lengths	0.2237899 (0)	4551705 (0)	0.3352664 (0)	9.8 (<2e-16)	0	61 (< 2.2e-16)
Intron Regions - aw Counts	45941 (6.273242e-169)	5766577 (0)	0.4402949 (0)	63 (<2e-16)	0	42 (< 2.2e-16)

Exon Regions - Lengths	-0.2221911 (0)	4563059 (0)	-0.3359603 (0)	9.1 (<2e-16)	0	17 (< 2.2e-16)
Exon Regions - Raw Counts	32685 (7.498446e-166)	5793781 (0)	0.4424573 (0)	0.45 (<2e-16)	0	12 (< 2.2e-16)
CpG Islands Gene Regions	74732 (2.059644e-13)	3829811 (0)	-0.272355 (0)	41 (<2e-16)	7.06e-279	01 (< 2.2e-16)
GC Content Gene Regions	-0.02632129 1.526554e-11)	0.02510859 239018e-10)	-0.01803212 2.805245e-10)	0.82 (<2e-16)	1.37e-12	84 (< 2.2e-16)

Table 11: Statistical data and relationship between isoform count and multiple cellular and biomolecular properties present within the mouse genome. Data is either raw or normalized by gene length, depending on the property, and subset by biotype: protein coding [PC] and non-protein coding [NC].

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
[PC] Raw Protein-Protein Interactions (PPIs)	0.1155341 (1.234913e-57)	0.2096437 (2.86434e-188)	0.147548 (7.938179e-185)	10.03 (<2e-16)	1.13e-159	4.6663 (< 2.2e-16)
[NC] Raw Protein-Protein Interactions (PPIs)	-0.1334369 (0.1363379)	-0.1762354 (0.04838165)	-0.1514417 (0.04857346)	0.466 (0.895)	0.743	0.5952 (0.7988)
[PC] 3' UTR Sequence Length	-0.02675353 (0.0001121787)	-0.08302119 (3.344502e-33)	-0.05761985 (1.954264e-32)	2.115 (3.42e-05)	4.57e-23	1.0262 (0.4241)
[NC] 3' UTR Sequence Length	0.02053136 (0.00293406)	NaN (NaN)	NaN (NaN)	1.108 (0.292)	1.21e-12	0.9127 (0.6326)
[PC] 5' UTR Sequence Length	0.02053136 (0.00293406)	0.02320772 (0.0007726551)	0.01651515 (0.0006501479)	1.108 (0.292)	1.21e-12	0.9127 (0.6326)

[NC] 5' UTR Sequence Length	0.02053136 (0.00293406)	NaN (NaN)	NaN (NaN)	1.108 (0.292)	1.21e-12	0.9127 (0.6326)
[PC] CpG Islands in Gene Regions	-0.0734214 (5.505701e-09)	-0.222804 (1.25363e-71)	-0.1565867 (8.843476e-72)	8.942 (<2e-16)	3.22e-57	8.3351 (< 2.2e-16)
[PC] GC Content in Gene Regions	-0.07115459 (1.860318e-08)	-0.06481286 (3.021296e-07)	-0.04516767 (2.701726e-07)	2.075 (0.000144)	0.00174	1.7069 (0.004847)
[NC] CpG Islands in Gene Regions	-0.09912642 (9.328573e-09)	-0.5563254 (6.308497e-271)	-0.4127917 (3.529545e-250)	2.95 (<2e-16)	5.88e-161	2.6637 (< 2.2e-16)
[NC] GC Content in Gene Regions	-0.09242646 (9.85667e-08)	-0.1172167 (1.309099e-11)	-0.08300377 (1.377712e-11)	1.374 (0.00938)	0.0259	2.1954 (2.639e-10)

Table 12: Statistical relationship between isoform count and multiple cellular and biomolecular properties present within the zebrafish genome. Data has been averaged (Avg.) and aggregated (Agg.) across the properties to demonstrate the association per isoform count.

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
Avg. GC Content within Gene Regions (Agg.)	0.4777829 (0.01821273)	0.5521739 (0.005856954)	0.3478261 (0.01725668)	NaN (NaN)	0.461	NaN (NaN)
3' UTR Sequence Avg. Length (Agg.)	-0.155376 (0.5253206)	-0.2017544 (0.4075051)	-0.04093567 (0.8065346)	NaN (NaN)	0.456	NaN (NaN)
5' UTR Sequence Avg. Length (Agg.)	-0.02678582 (0.9133222)	0.06315789 (0.7972829)	0.0994152 (0.5520074)	NaN (NaN)	0.456	NaN (NaN)
Intron Regions - Avg. Counts (Agg.)	0.6789938 (0.001389911)	0.7210526 (0.0004947295)	0.6023392 (0.000313944)	NaN (NaN)	0.456	NaN (NaN)

Exon Regions - Avg. Counts (Agg.)	0.6903615 (0.001068781)	0.7210526 (0.0004947295)	0.6023392 (0.000313944)	NaN (NaN)	0.456	NaN (NaN)
-----------------------------------	----------------------------	-----------------------------	----------------------------	-----------	-------	-----------

Table 13: Statistical relationship between isoform count and multiple cellular and biomolecular properties present within the zebrafish genome. Data has been averaged (Avg.) and aggregated (Agg.) across the properties to demonstrate the association per isoform count, subset by biotype: protein coding [PC] and non-protein coding [NC].

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
[PC] Avg. GC Content in Gene Regions (Agg.)	0.04568269 (0.852677)	-0.09298246 (0.7047526)	-0.05263158 (0.7825416)	NaN (NaN)	0.456	NaN (NaN)
[NC] Avg. GC Content in Gene Regions (Agg.)	0.4379818 (0.069078)	0.6181631 (0.007428688)	0.4117647 (0.01720772)	NaN (NaN)	0.454	NaN (NaN)
[PC] 3' UTR Sequence Avg. Length (Agg.)	-0.1554218 (0.5251972)	-0.2017544 (0.4075051)	-0.04093567 (0.8065346)	NaN (NaN)	0.456	NaN (NaN)
[NC] 3' UTR Sequence Avg. Length (Agg.)	0.6353413 (0.3646587)	0.4 (0.6)	0.3333333 (0.4969058)	NaN (NaN)	0.392	NaN (NaN)
[PC] 5' UTR Sequence Avg. Length (Agg.)	-0.02683285 (0.9131706)	0.06315789 (0.7972829)	0.0994152 (0.5520074)	NaN (NaN)	0.456	NaN (NaN)
[NC] 5' UTR Sequence Avg. Length (Agg.)	-0.1133427 (0.8866573)	-0.2 (0.8)	0 (1)	NaN (NaN)	0.392	NaN (NaN)

Table 14: Statistical data and relationship between isoform count and multiple cellular and biomolecular properties present within the zebrafish genome. Data is either raw or normalized by gene length depending on the property.

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
--	----------------	-------------------	------------------	--------------	--------	------------------

GC Content within Gene Regions	0.3408138 (0)	-0.04965961 (3.758567e-16)	-0.03860408 (4.124399e-17)	265.8 (<2e-16)	2.26e-294	54.842 (< 2.2e-16)
3' UTR Sequence Length	-0.01374183 (0.04318706)	-0.0178227 (0.008730843)	-0.01336223 (0.008977297)	3.009 (1.77e-05)	2.1e-07	2.51 (0.0003942)
5' UTR Sequence Length	0.01876236 (0.005873427)	0.01463587 (0.03164844)	0.01117944 (0.02910789)	5.523 (3.07e-13)	5.16e-13	4.6333 (2.281e-10)
Intron Regions - Lengths	0.2062286 (3.827449e-204)	0.2247341 (5.02186e-243)	0.1707053 (7.956238e-241)	80.43 (<2e-16)	7.21e-223	66.713 (< 2.2e-16)
Intron Regions - Raw Counts	0.1980872 (3.46068e-188)	0.3100432 (0)	0.2412058 (0)	55.29 (<2e-16)	0	6.3483 (5.619e-16)
Exon Regions - Lengths	-0.2113053 (1.90374e-214)	-0.2519404 (6.063675e-307)	-0.1914832 (1.550984e-302)	999.6 (<2e-16)	2.58e-282	70.046 (< 2.2e-16)
Exon Regions - Raw Counts	0.2548589 (2.738157e-314)	0.3522866 (0)	0.2744153 (0)	95.74 (<2e-16)	0	16.425 (< 2.2e-16)

Table 15: Statistical data and relationship between isoform count and multiple cellular and biomolecular properties present within the zebrafish genome. Data is either raw or normalized by gene length depending on the property, and subset by biotype: protein coding [PC] and non-protein coding [NC].

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
[PC] GC Content in Gene Regions	-0.1084693 (2.731492e-62)	-0.1061959 (9.089781e-60)	-0.08026887 (6.242725e-60)	29.43 (<2e-16)	1.32e-49	50.823 (< 2.2e-16)
[NC] GC Content in Gene Regions	0.6854988 (0)	0.6912192 (0)	0.5189903 (0)	636.5 (<2e-16)	0	27.534 (< 2.2e-16)

[PC] 3' UTR Length	-0.01380807 (0.04223249)	-0.01794156 (0.00830627)	-0.01345176 (0.008541163)	3.012 (1.73e-05)	2.03e-07	2.5128 (0.0003878)
[NC] 3' UTR Length	0.5268986 (0.1449727)	0.2935903 (0.4432145)	0.2333333 (0.4168451)	3.684 (0.0972)	0.279	0.329 (0.8054)
[PC] 5' UTR Length	0.01878851 (0.005816237)	0.0147006 (0.03094122)	0.01122621 (0.02847628)	5.516 (3.24e-13)	5.49e-13	4.6263 (2.402e-10)
[NC] 5' UTR Length	-0.1727669 (0.6824504)	-0.05455447 (0.8979131)	0 (1)	0.169 (0.912)	0.675	0.3826 (0.772)

Table 16: Statistical relationship between isoform count and multiple cellular and biomolecular properties present within the *Drosophila* genome. Data has been averaged (Avg.) and aggregated (Agg.) across the properties to demonstrate the association per isoform count.

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
Avg. GC Content within Gene Regions (Agg.)	-0.3051425 (0.617597)	0.3 (0.6833333)	0.2 (0.8166667)	NaN (NaN)	0.406	NaN (NaN)
3' UTR Sequence Avg. Length (Agg.)	0.308209 (0.1524889)	0.5335968 (0.008735939)	0.438735 (0.003372635)	NaN (NaN)	0.46	NaN (NaN)
5' UTR Sequence Avg. Length (Agg.)	0.1110742 (0.6138732)	0.1482213 (0.4997097)	0.1383399 (0.3552958)	NaN (NaN)	0.46	NaN (NaN)
Intron Regions - Avg. Counts (Agg.)	0.8933073 (1.603448e-10)	0.9759168 (1.017115e-18)	0.8994709 (1.852285e-11)	NaN (NaN)	0.464	NaN (NaN)
Exon Regions - Avg. Counts (Agg.)	0.8933073 (1.603448e-10)	0.9759168 (1.017115e-18)	0.8994709 (1.852285e-11)	NaN (NaN)	0.464	NaN (NaN)

Table 17: Statistical data and relationship between isoform count and multiple cellular and biomolecular properties present within the *Drosophila* genome. Data is either raw or normalized by gene length depending on the property.

	Pearson (r, p)	Spearman (rho, p)	Kendall (tau, p)	ANOVA (F, p)	KW (p)	Levene (stat, p)
GC Content within Gene Regions	0.03443212 (0.1871805)	0.08908507 (0.0006301293)	0.07246336 (0.0006591894)	2.418 (0.0468)	0.0168	0.9209 (0.4508)
3' UTR Sequence Length	0.3882028 (5.72094e-82)	0.4969499 (6.704129e-141)	0.3753752 (3.458941e-127)	35.21 (<2e-16)	1.27e-108	22.437 (< 2.2e-16)
5' UTR Sequence Length	0.3189478 (2.127199e-54)	0.4802411 (3.176263e-130)	0.3620945 (2.35245e-118)	25.08 (<2e-16)	2.8e-99	15.493 (< 2.2e-16)
Intron Regions - Lengths	0.4374406 (0)	0.5052951 (0)	0.4082002 (0)	245.8 (<2e-16)	0	53.367 (< 2.2e-16)
Intron Regions - Raw Counts	0.663427 (0)	0.5658304 (0)	0.4845366 (0)	559.3 (<2e-16)	0	83.722 (< 2.2e-16)
Exon Regions - Lengths	-0.437446 (0)	-0.5054048 (0)	-0.4082913 (0)	245.8 (<2e-16)	0	53.39 (< 2.2e-16)
Exon Regions - Raw Counts	0.663427 (0)	0.5658304 (0)	0.4845366 (0)	559.3 (<2e-16)	0	83.722 (< 2.2e-16)

References

1. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res.* 53, D20–D29 (2025).
2. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32, D493–D496 (2004).
3. Perez, G. *et al.* The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res.* 53, D1243–D1249 (2025).