RESEARCH

Open Access

Estimating transcriptome complexities across eukaryotes

James E. Titus-McQuillan[1*], Adalena V. Nanni[2,3], Lauren M. McIntyre[2,3] and Rebekah L. Rogers[1]

Abstract

Background Genomic complexity is a growing field of evolution, with case studies for comparative evolutionary analyses in model and emerging non-model systems. Understanding complexity and the functional components of the genome is an untapped wealth of knowledge ripe for exploration. With the "remarkable lack of correspondence" between genome size and complexity, there needs to be a way to quantify complexity across organisms. In this study, we use a set of complexity metrics that allow for evaluating changes in complexity using TranD.

Results We ascertain if complexity is increasing or decreasing across transcriptomes and at what structural level, as complexity varies. In this study, we define three metrics – TpG, EpT, and EpG- to quantify the transcriptome's complexity that encapsulates the dynamics of alternative splicing. Here we compare complexity metrics across 1) whole genome annotations, 2) a filtered subset of orthologs, and 3) novel genes to elucidate the impacts of orthologs and novel genes in transcript model analysis. Effective Exon Number (EEN) issued to compare the distribution of exon sizes within transcripts against random expectations of uniform exon placement. EEN accounts for differences in exon size, which is important because novel gene differences in complexity for orthologs and whole-transcriptome analyses are biased towards low-complexity genes with few exons and few alternative transcripts.

Conclusions With our metric analyses, we are able to quantify changes in complexity across diverse lineages with greater precision and accuracy than previous cross-species comparisons under ortholog conditioning. These analyses represent a step toward whole-transcriptome analysis in the emerging field of non-model evolutionary genomics, with key insights for evolutionary inference of complexity changes on deep timescales across the tree of life. We suggest a means to quantify biases generated in ortholog calling and correct complexity analysis for lineage-specific effects. With these metrics, we directly assay the quantitative properties of newly formed lineage-specific genes as they lower complexity.

Keywords OrthoDB, Transcriptome complexity, Evolutionary rates, Orthologs, Novel genes, Effective exon number, TranD, Transcript model

Background

Transcriptome complexity is the product of evolutionary, biophysical, and molecular constraints that depend on the environmental context of an organism [1]. Alternative splicing allows more diverse variations of proteins from a single DNA locus, increasing the number of structures that can arise from a given genomic region [2–4]. Moreover, the combination of different exons and splicing action may change regulatory profiles for transcripts and therefore proteins produced [5]. This isoform complexity

*Correspondence:
James E. Titus-McQuillan
jmcquil2@uncc.edu
[1] Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
[2] Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL 32611, USA
[3] University of Florida Genetics Institute, University of Florida, Gainesville, FL 32611, USA

increases information content stored in a single genomic region [6] often with functional consequences [7, 8]. The ways that global genome-wide transcript structures change across species influence the scope and consequences of genetic diversity and genetic architecture [9].

Alternative splicing is a distinct but pervasive product of Eukaryotic evolution. Among different organisms, patterns of alternative splicing and transcriptome complexity may shift, producing differences in exon/intron boundaries, length, and number [10]. Some of these changes are linked directly to genetic and biochemical changes in spliceosome machinery including U11 and U12 spliceosomes [11], long intron splicing dynamics [12], and specific tissue regulation [13, 14]. Others may be the product of selection for specific genetic features and functions within the cell [15–17]. Complexity is driven by many factors such as tissue type [18], sex-specificity [19, 20], development [9, 13, 21], phenotype [22], and biochemical constraints [23]. These factors are associated with multiple modes of genetic evolution across the tree of life including sex with reproduction, body formation and tissue complexity, and pathogens/symbioses [24, 25]. Intron density and distribution can vary by orders of magnitude across unicellular and multicellular taxa [26, 27], sometimes including complete spliceosome loss or reemergence of splicing function [28]. Modes of exon use, levels of alternative splicing, and isoforms generated may change as spliceosome structures evolve. Given the pervasiveness and consequences of alternative splicing, analysis of genome-wide patterns for alternative splicing may reveal biological variation that shapes molecular and evolutionary biology.

One of the challenges of comparative transcriptome and splicing analysis across the tree of life has been the (seemingly) "necessary evil" of assigning orthology across distantly related taxa [24, 28–31]. Over time, the evolution of novel genes results in lower concordance in gene content across organisms. Heterogeneity in gene content has been highlighted in phylogenetics as an analytical complication, but it is also a biological contributor to changes in species complexity [32, 33]. Novel genes are associated with fewer numbers of exons, shorter exons, and lower biochemical complexity, along with increasing sequence ambiguity [34–36]. Failure to include novel genetic elements may therefore alter estimates of complexity as a genomic trait. Moreover, as genetic distance increases, sequence alignment and ortholog calling becomes more challenging when generating additional sources of uncertainty and potential bias [37]. As a result of these many analytical complications, studies of intron–exon placement, complexity, and splicing patterns with ortholog conditioning may be more affected across vast evolutionary distance. Using subsets of constrained genome sequences may portray evolutionary processes for single genes that disagree with the evolutionary history of the species due to incomplete lineage sorting (ILS), introgression, and gene duplication and loss (GDL) [38]. Sets of proteins that are highly conserved across distantly related taxa may make inference simpler [39–41], but selections of specific proteins may introduce biases against rapidly evolving genes [42–46].

Orthologous sequences are shared genetic code between species that share an ancestor, separated by a speciation event [39]. Orthologs are a central tenet of comparative studies including comparative genomics, phylogenetics, protein function annotation, genome rearrangement and structure. Using orthologs for comparisons are robust, and yield results that are applicable for one-to-one comparisons across evolutionary time scales. However, only focusing on orthologs limits understanding of novel genetic elements and does not explain the whole evolutionary history. Given that novel genetic elements are important to a lineage's genetic history and adaptability [35, 47–56], ortholog-only comparisons will miss important variation that generates gene turnover and modifies transcriptome content and complexity over time. Complexity metrics from TranD are agnostic to orthology, as they present complexity independently from taxa being compared, using only within-species analysis on a single reference. Here, we compare the dynamics of conditioning on orthologs versus whole-transcriptome complexity metrics.

In the face of these challenges, clear solutions for complexity analysis are needed that are robust and free from biases of ortholog conditioning. While single gene analysis on close evolutionary distance may be best served by the precision of cross-species alignment of introns and exons, whole-genome transcriptome splicing may be assayed as a global phenotype that can shift across species. Comparing complexity within a single species' transcriptome, we can estimate the properties of all genes at once in an ortholog-free framework. These global, whole-genome properties may provide more complete information about transcriptome architecture than previous approaches comparing gene-by-gene. A new program for transcript model analysis, TranD [57] (https://doi.org/10.1101/2021.09.28.462251; https://github.com/McIntyre-Lab/TranD), enables comparisons of complexity across a wide array of genomes across the eukaryotic tree of life. We define complexity with these principles in mind and quantify complexity metrics from TranD here as variation among genetic regions with respect to unique exons per gene (EpG), exons per transcript (EpT), and transcripts per gene (TpG). Additionally, we add one previously developed measure of transcript model complexity, the effective exon number (EEN). In alternative splicing, EpG defines the total number of genetic elements available to generate unique combinations, while TpG is the product of the total of these combinations. EpT governs the number of introns that must be spliced, influencing transcript regulation. Finally, EEN is the product of splice junction spacing reflecting biochemical constraints of splicing factors. Using these metrics, we quantify complexity across phylogenies for major metazoan taxa: Deuterostomes, *Drosophila*, Plants, and Fungi. A glossary of terms can be found in Table 1. We focus on these well-annotated taxa as examples for how evolutionary analysis of transcriptome complexity may guide inference of evolutionary processes.

The analyses presented here offer an important case study for comparative evolutionary analyses in model and emerging non-model systems. Quantifying complexity as a set of metrics allows for evaluation of changes in complexity, ascertain if complexity is increasing or decreasing across transcriptomes and at what structural level. Given that there are a variety of levels to complexity, one metric does not fit all, therefore, we present three distinct complexity metrics defined above. Here we compare results from 1) whole-genome annotations, 2) a filtered subset of orthologs, and 3) novel genes to elucidate the impacts of orthology and novel genes in transcript model analysis. We suggest a means to quantify biases generated in ortholog calling and correct complexity analysis for lineage-specific effects. With these metrics, we directly assay the quantitative properties of newly formed lineage-specific genes as they lower complexity in transcriptomes. We implement evolutionary rate analyses from complexity changes across diverse lineages with greater precision and accuracy than previous cross-species comparisons under ortholog conditioning. These analyses represent a step forward toward whole-transcriptome analysis in the emerging field of non-model evolutionary genomics, with key insights for evolutionary inference of complexity changes on deep timescales across the tree of life.

Results

Complexity metrics

TranD empirically estimates complexity by calculating metrics that are immediately comparable across species with rigorous and repeatable analytical criteria [57] (https://doi.org/10.1101/2021.09.28.462251). Each species is estimated independently from all others, generating complexity metrics for each annotation. These metrics offer complexity measures as genetic traits that can be compared across a phylogeny. Because each complexity is derived from a single branch in the tree of life, there is breadth to compare deep-time divergence between and among distantly related species. We use four major clades as test cases to assay genome complexity across well annotated phylogenies: Deuterostomia, *Drosophila,* Plantae, and Fungi. Each of these clades spans different timescales (from ~ 50 mya in the *Drosophila* group to ~ 1 billion years ago in the fungi group) and unique biological features. These examples show how analysis of transcript model complexity can clarify unique genetic properties of species within clades. We show how such inference is affected by orthology (and genetic novelty) across highly divergent phylogenetic groups.

Transcript model complexities vary across organisms from distantly related lineages (Fig. 1). Among the deuterostome clade mean TpG is between 1.00–3-97. The means between EpT and EpG are more concordant with mean EpT 7.10–13.31 and mean EpG is 7.10–11.09. *Drosophila* have smaller transcript complexity metrics than those of deuterostome lineages. This is concordant with genome size disparities between the two groups. The range of EpT across *Drosophila* species are 3.38–6.08 and the means of EpG are 3.38–5.01, which do not overlap deuterostome EpT and EpG mean metrics. TpG

Table 1 Glossary of complexity terms

Alternative Splicing (AS): Molecular mechanism that modifies pre-mRNA constructs prior to translation, which produces a diverse set of mRNA from a single gene.

Transcripts per Gene (TpG): The number of transcripts that can be constructed from a single gene that have been annotated.

Exons per Transcript (EpT): The exons that are annotated within single transcript produced from a single gene.

Exons per Gene (EpG): The union of unique exons annotated within a single gene.

Effective Exon Number (EEN): A distribution of exons across a complexity metric (or gene), given the relative length of the exon to the total length of the region (L_e).

Ortholog: One of a set of homologous genes that have diverged from an ancestor as a consequence of speciation.

Novel Gene: Genes that have emerged inside a defined time frame are novel genes. Novel genes are will be classified by their age. The time frame is not fixed and needs to be defined for each study.

Evolutionary Rate: As defined through Rabosky et al. (2014) interpretation. A macroevolutionary rate dynamic that applies to some part of a phylogenetic tree. All lineages that share a common regime have exactly the same macroevolutionary rate at a given point in time.

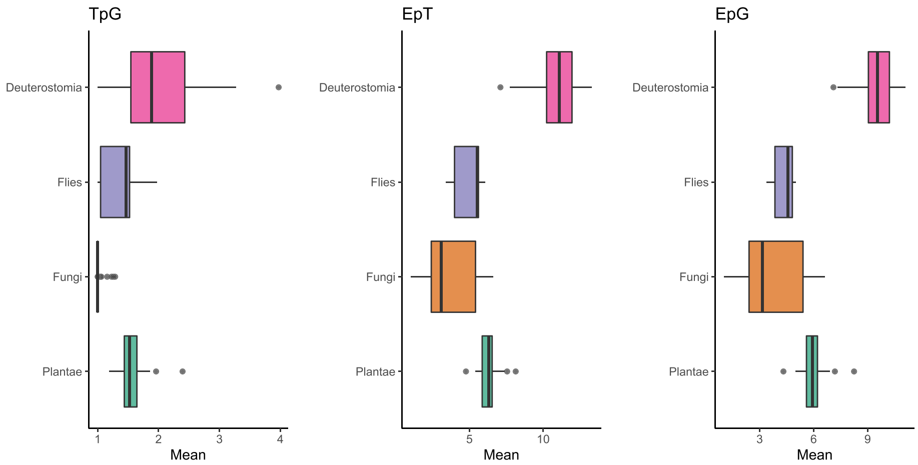

Fig. 1 Box and whisker plots showing complexity metrics for TpG, EpT, and EpG for broad taxonomic groupings. Black lines indicate median metrics with the first and third quartiles at the borders of the box. The minimum and maximum range indicated by black lines. Taxonomic groups consist of Deuterostomia, *Drosophila*, Plantae, and Fungi

means are lower too at 1.00–1.97 TpG. Among taxa in the Plantae division Viridiplantae the mean TpG range is 1.19–2.39, mean EpT range is 4.76–8.14, and mean EpG range is 4.32–8.25. Fungi lineages have a much lower distribution in complexity compared to the other higher-level taxonomic lineages. Fungi mean TpG range is 1.00–1.29. Fungi EpT mean ranges from 1.03–6.62. Finally, the Fungi mean EpG is of 1.03–6.63. Across all higher lineages: Deuterostomia, *Drosophila*, Plantae, and Fungi, mean TpG is significantly different across groups, KW $Chi^2 = 148.4$, P = 5.777e-32. Both EpT and EpG too are significantly different across higher lineages, for EpT KW $Chi^2 = 160.9$, P = 1.19e-34 and EpG KW $Chi^2 = 155.1$, P = 2.08e-33 (Supp. Mat. Mean Complexity Metrics.pdf). Hence, each clade portrays a characteristic complexity that is the unique product of biochemical, molecular, and evolutionary properties of the organisms. See Supp. Figure 1 for a complete list of complexity metrics across each organism.

Complexity whole-transcriptome and orthologs

Annotations partitioned by ortholog have higher complexities across most metrics compared to the whole-transcriptome (Fig. 2; Supp. Mat. Whole-Transcriptome Vs Ortholog Density Plots). In the Plantae group, for TpG, 36.4% (16/44) taxa have ortholog complexities lower than the whole-transcriptome complexity metrics and 4.5% (2/44) for EpT. Fungi have 3.9% (3/77) taxa where TpG complexity metrics are higher in whole-transcriptomes than in the ortholog set. Finally, 9.1% (7/77) taxa across EpT and EpG in all the same taxa show the same complexity patterns above. The shift from higher complexity in ortholog datasets are due to exclusion of lineage-specific (non-orthologous) genes that are present in the whole-transcriptome annotations. Novel gene complexities are substantially lower than genes with orthologs (Fig. 3; Supp. Mat. Novel Density Plots), consistent with prior work showing novel genes are less complex [35, 58–61]. This observation shows that many of the novel genetic elements are single-transcript and single-exon genes compared to that of the older orthologous genetic elements representing higher complexity transcripts. Whole transcriptome annotations contain both novel genetic elements and orthologous genetic elements. Hence, the less complex novel elements drive differences in complexity compared with orthologs for all complexity metrics (EpT, EpG, TpG). Four well annotated species serve as examples for these trends (one species per clade). Distributions for *Homo sapiens*, *Drosophila melanogaster*, *Zea mays*, and *Neocallimastix californiae* all show significant differences, with p-values zero or near zero, between whole-transcriptome and partitioned orthologous annotations (Fig. 2). Interestingly, novel genetic elements across all taxa shift to lower EpT compared to orthologous genes as most new genetic elements have lower transcript model complexity than older orthologs (Fig. 3). An artifact observed in our data, between novel and orthologous datasets, is the proportion of novel genes greater than 50% for *Gorilla gorilla* (western lowland gorilla) [GCF_008122165.1_Kamilah_GGO_v0], *H. sapiens* [GCF_000001405.39_GRCh38.p13], and *Musa acuminata* (wild Malaysian banana) [GCF_000313855.2_ASM31385v2] (Supp. Figure 2). See Supp. Mat. Whole-Transcriptome Vs Ortholog Density Plots and Supp. Mat. Novel Density Plots for a list of all organisms used in this study.

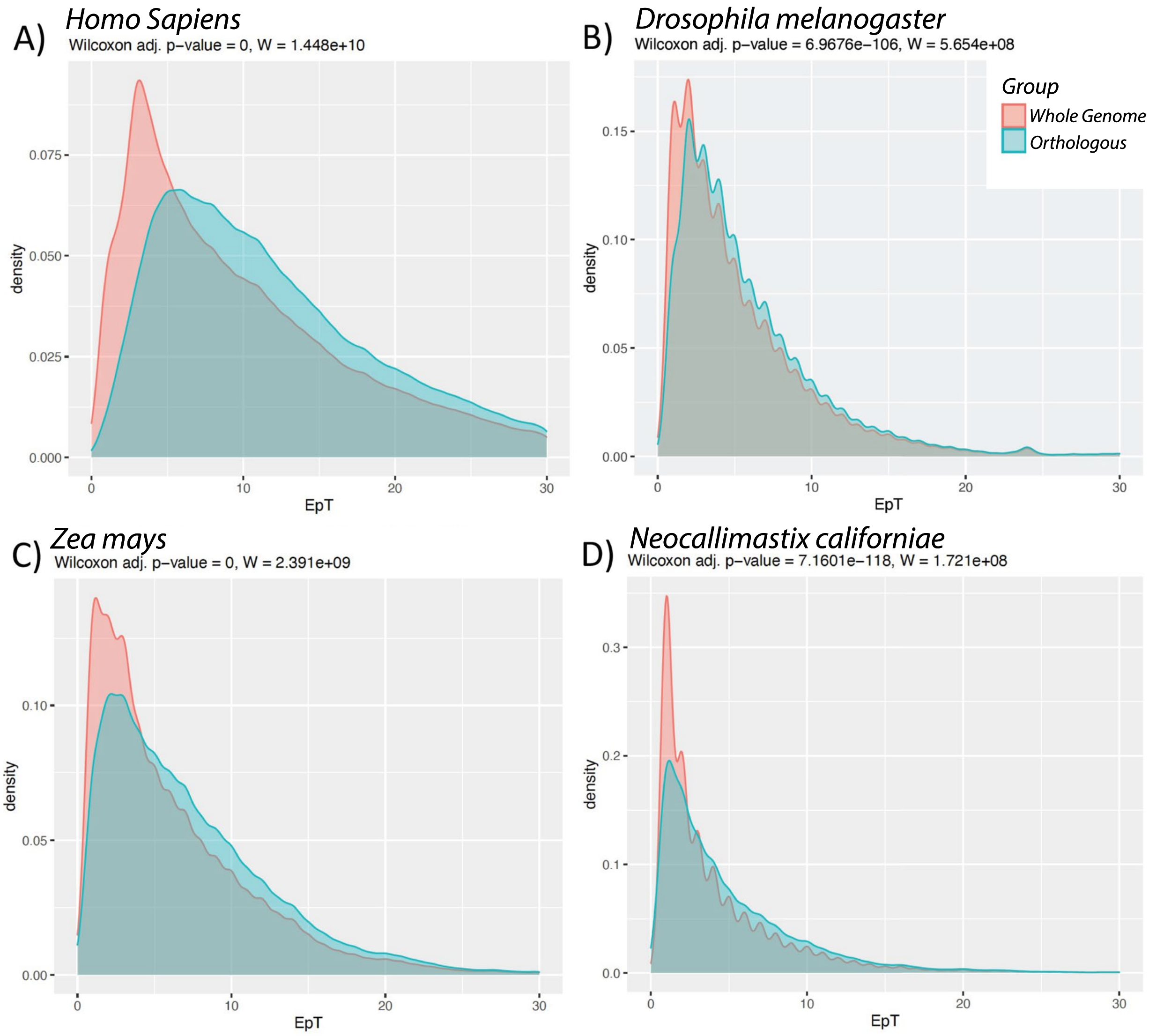

Fig. 2 Density plots between whole genome (red) and orthologs (light blue) for Exons per Transcript (EpT) complexity metrics with densities on the y-axis and the number of exons per transcript (EpT) on x-axis. Panels are truncated to 30 exons for purposes of visualization. We observe a significant difference in EpT for whole -transcriptome annotations and partitioned by ortholog annotations in each of four example species using Wilcoxon rank sum test. A) H. sapiens (GCF_000001405.39_GRCh38.p13), W = 1.448 × 1010, *P* = 0, B) Drosophila melanogaster (dmel-all-r6.07), W = 5.654 × 109, *P* = 6.9676 × 10–106, C) Zea mays (GCF_902167145.1_Zm-B73-REFERENCE-NAM-5.0), W = 2.391 × 109, *P* = 0, and D) Neocallimastix californiae (GCA_002104975.1_Neocallimastix_sp._G1_v1.0_genomic), W = 1.721 × 1010, *P* = 7.166 × 10–118. Across all taxa, ortholog conditioning results in data that are less likely to include low complexity transcripts

Across the Deuterostomia, mean TpG metrics between whole-transcriptome and ortholog annotations are significantly different in 66 of the 68 evaluated species. Only the Crested Ibis [GCF_000708225.1_ ASM70822v1] and the Florida lancelet [GCF_000003815.1_Version_2] have no significant differences for TpG metrics. When measuring complexity with mean EpT and mean EpG, all species showed significant differences between orthologs and whole-transcriptomes.

Across *Drosophila,* we observe a parallel pattern in complexity between ortholog and whole-transcriptome data. Of the 11 species evaluated, three do not show significant differences for TpG under ortholog conditioning. These include *D. grimshawi* [dgri − all − r1.3], *D. persimilis* [dper − all − r1.3], and *D. sechellia* [dsec − all − r1.3]. These results could be an artifact of annotation performance and quality compared to more well-assembled species with molecular support for gene features, as many *Drosophila* lineages from the

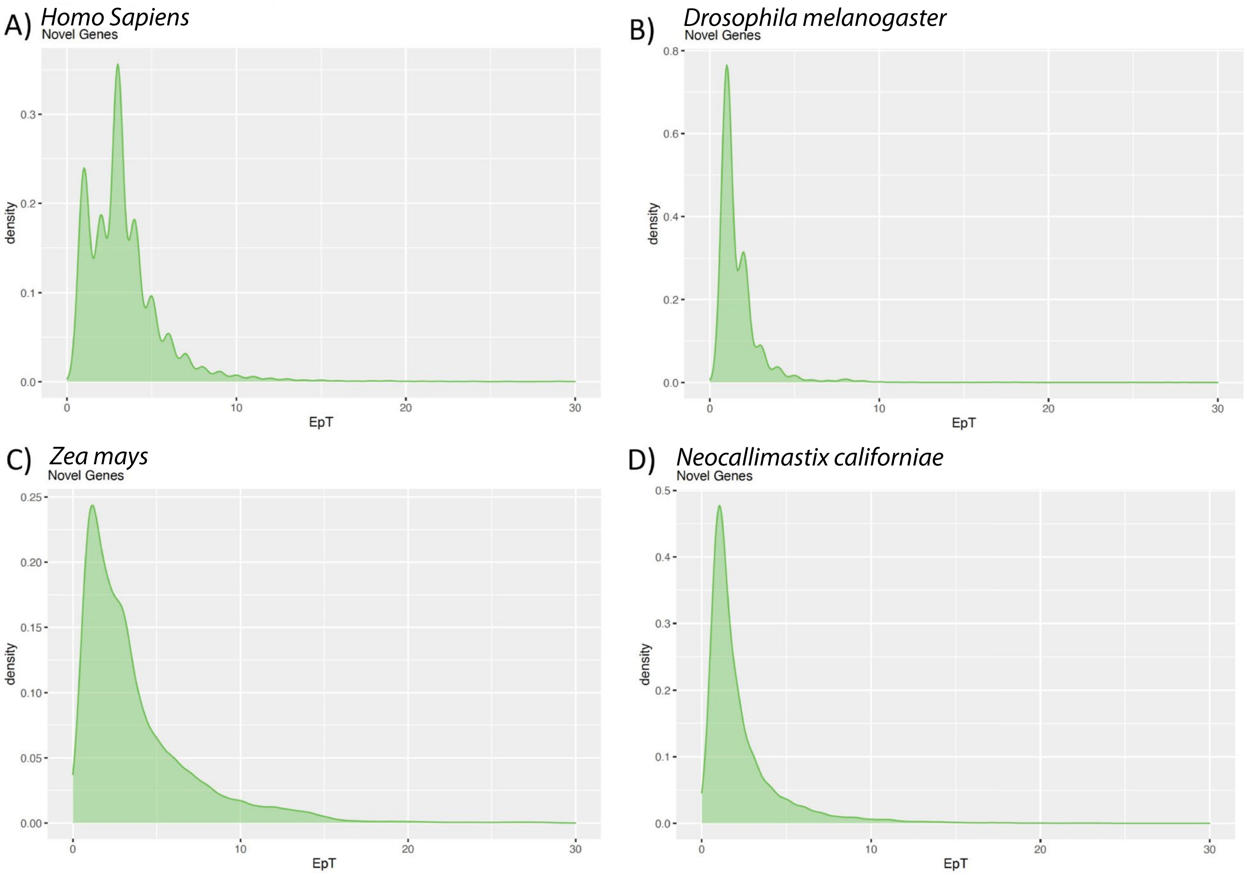

Fig. 3 Density plots of EpT complexity metrics for novel lineage specific genes with densities on the y-axis and number of exons plotted on x-axis. Panels are truncated to 30 exons for visualization. Each panel A) H. sapiens (GCF_000001405.39_GRCh38.p13) show a shift in the number of exons where most new genes have fewer than 10 exons per transcript with a mean EpT of 3.431. B) D. melanogaster (dmel-all-r6.07) concentrates most novel genes between one (~61.2%) and two (~24.7%) exons per transcript with a mean EpT of 1.863. EpT has few transcripts with complexity higher than two exons per transcript. C) Z. mays (GCF_902167145.1_Zm-B73-REFERENCE-NAM-5.0) follows a similar pattern with a shift left to lower numbers of exons per transcripts with a mean EpT at 3.822. Only ~27.7% of novel genes in maize are five EpT and higher, while under five is ~72.3% of all novel genes. Finally, D) Neocallimastix californiae (GCA_002104975.1_Neocallimastix_sp._G1_v1.0_genomic) novel genes show exon per transcript counts have a mean EpT of 2.321. Most of the distribution is concentrated in below five EpT (~88.4%). Values are substantially lower than observed for Orthologs and Whole-Transcriptome

12 Genomes Consortium were annotated by identifying ORFs not tuned to a specific lineage [62]. Lineages that are not significantly different between whole-transcriptome and conditioning on orthologs vary in annotation completeness compared with other taxa, with a maximum TpG = 2. This is a noticeable departure from the other lineages that range from 45 maximum TpG, *D. simulans* [dsim − all − r2.01] to 75 maximum TpG in *D. melanogaster* [dmel − all − r6.07]. In these instances, we suggest that the robustness of annotations varies between lineages and collapsed annotations combining exons from multiple transcripts have artificially skewed apparent transcript model complexity [63]. In contrast, EpT and EpG metrics are all significantly different between whole-transcriptome and orthologs for all taxa.

In Plantae, across both EpT and EpG metrics, all samples (44/44 species) are significantly different when conditioning on orthologous. For TpG, 34/44 species show significant differences between whole-transcriptome complexity and complexity for orthologous sequences.

Only 1 fungus of 77 shows a significant difference between whole-transcriptome and ortholog TpG, *Bactrachochytrium salamandrivorans* [GCA_002006685.1_Batr_sala_BS_V1]. All other fungi either do not deviate between whole-transcriptome and ortholog TpG or only have TpG metrics that equal one for all genes in whole-transcriptomes and subsequent ortholog partitions, i.e., no differences to compare. There are significant differences in EpT between orthologs and whole-transcriptome annotations for $n=53/77$, with EpG having the same dynamics as EpT, where the same taxa (53/77) are significantly different between orthologs and whole-transcriptome annotations.

Conditioning on orthologs yields transcript model complexity that deviates from the whole-transcriptome complexity. Of the 200 species used in this project 185/200 had significant differences under ortholog conditioning for TpG and 176/200 for EpT and EpG. These lower numbers are driven by fungi EpT and EpG having nearly identical mean EpT and EpG metrics. Across all lineages from each taxonomic group, a significant impact on complexity when orthology is required under all three metrics, except for mean TpG for fungi ($|T|=47.333$, Fisher's combined $P=0.840$, Supp. Table 1).

From our Jackknife cross-validation, 197/200 samples validate. Only 3 plant species differ with *Musa acuminata* (wild Malaysian banana) [GCF_000313855.2_ASM31385v2] not validating for EpT, and TpG not validating for *Pyrus x bretschneideri* (Chinese white pear) [GCF_000315295.1_Pbr_v1.0] and *Oryza sativa*, Japonica Group, (Japanese rice) [GCF_001433935.1_IRGSP-1.0].

Across lineages, we expect there to be some transcript overlap within an GTF as polycistronic mRNA exists in eukaryotes. We calculated the transcript overlap between genes of an organism's GTF to see if there is any elevation for a complexity metric given an overlapping transcript. In most taxa from deuterostomes, flies, and plants, fewer than 2.6% of transcripts are overlapping. Exceptions are the tunicate Ciona (4.9%), Club Moss (6.4%), and Brassica rapa (3.7%). Fungi, however, have fewer genes and fewer isoforms, generally speaking. Schizosaccharomyces octosporus yFS286 (GCF_000150505.1_SO6) have up to 11.1% of exons overlapping. With 29/77 having no overlapping transcripts (Supp. Table 2). We reiterate that care should be taken in the annotation quality when conducting these analyses.

Relative performance of complexity statistics

All correlations are significant and positive (Fig. 4). TpG has the highest residuals from the regression line (R^2) when compared to the EpT and EpG metrics. While EpT and EpG metrics are more closely correlated. R^2 values are nearly one for both *Drosophila* ($R^2=0.962$, $P<0.01$) and Plantae ($R^2=0.911$, $P<0.01$). Deuterostomes show more moderate correlations for EpT vs EpG ($R^2=0.572$, $P<0.01$), however, it is the highest R^2 value for the group. Deuterostomes have the largest variance between plotted metrics with TpG vs EpT ($R^2=0.370$, $P<0.01$) and TpG vs EpG ($R^2=0.0964$, $P<0.01$). *Drosophila* also show a significant correlation TpG vs EpT ($R^2=0.741$, $P<0.01$) and TpG vs EpG ($R^2=0.831$, $P<0.01$). Plants are moderately aligned closer to one in the R^2 values: TpG vs EpT ($R^2=0.603$, $P<0.01$) and TpG vs EpG ($R^2=0.582$, $P<0.01$). One non-vascular plant constituent, (Bryophta, *Physcomitrium patens*) the spreading earthmoss, is not clustered with our vascular plants, having the highest values for all metrics comparison within the group (See Supp. Figure 3).

Most fungi annotations are single transcript genes and are therefore not amenable for similar analysis (Supp. Figure 4). In current annotations, given that a significant majority of genes are single transcript, EpT and EpT metrics are highly correlated ($R^2=0.999$).

Comparison to broken stick is robust to orthology

Differences in complexity for orthologs and whole-transcriptome analyses are driven largely by a bias in novel genes toward low complexity genes with few exons and few alternative transcripts. The Effective Exon Number (EEN) offers a metric that can compare the distribution of exon sizes within transcripts against random expectations of uniform exon placement, and accounts for differences in exon sizes [64]. If EEN = EpT, splice junctions are evenly spaced, with all exons of equal size. Where EEN is far less than EpT, splice junctions are more uneven than expected under a uniform distribution, with over-dispersed exon sizes. While absolute patterns of EEN may be informative, comparisons of EEN to EpT characterize the deviation from a Broken Stick Model (Supp. Figure 17) of randomly scattered intron positions. Across all species of chordates, *Drosophila*, and plants, EEN comparisons with the Broken Stick Model remain robust to conditioning on orthology and exclusion of lineage specific genes. We observe no difference in deviations from the Broken Stick Model when considering orthologs compared with whole-transcriptome data. Only a minor shift upward is apparent in the lowest EpT values for Deuterostomes, *Drosophila*, Plants, and most Fungi (Supp. Figure 18). Because the Broken Stick Model conditions on EpT, the effects of removing less complex lineage specific genes are largely mitigated so long as the remaining genes follow similar patterns of exon sizes.

Out of 68 deuterostome species, 62 have mean ± 2SE EEN below the bound of the Broken Stick Model (Fig. 5). Humans, whose annotations are extensively validated and supported by abundant molecular evidence, lie among some of the lowest values suggesting more clustered intron breakpoints than the Broken Stick Model. These results suggest that molecular or evolutionary constraints on splicing processes or differences in annotation models are producing more tightly spaced intron breaks than one would expect based on random intron placement drawn from a uniform distribution. Human annotations are

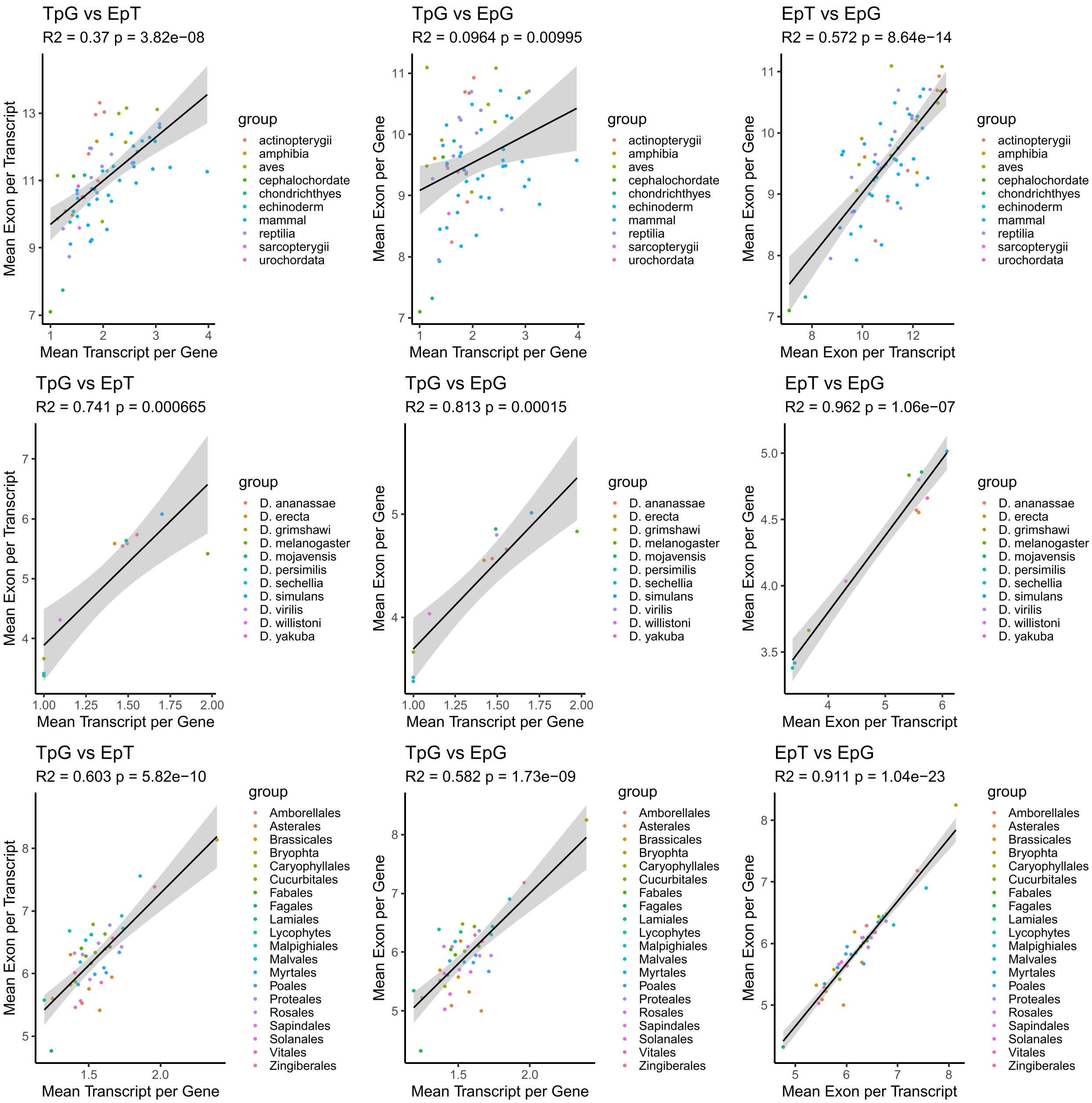

Fig. 4 Pair plots comparing complexity metrics. Metrics are compared among TpG, EpT, and EpG. The *P*-value and R^2 derived from calculating the Pearson correlation coefficient. Taxonomic groups consist of Deuterostomes, *Drosophila*, and Plantae

supported by exceptional molecular evidence and more well-developed curation efforts.

The remaining six species of chordates that show elevated EEN encompass diverse species that are not related phylogenetically (duck, turtle, lancelet, hedgehogs). Results do not shift significantly when conditioning on orthologs and excluding lineage specific genes. Hence, these metrics, unlike EpT, EpG, and TpG, appear to be robust in the face of phylogenetic comparisons that require orthology. However, EEN conditions on the exon number and compares to expectations for randomly placed junctions. These are not sensitive to the subset of lineage specific genes that orthology excludes, which have few introns.

Drosophila show a different pattern. Transcripts with 7 exons or less show EEN greater than or equal to a Broken Stick model across all species. However, when transcripts have 12 exons or more, we begin to observe

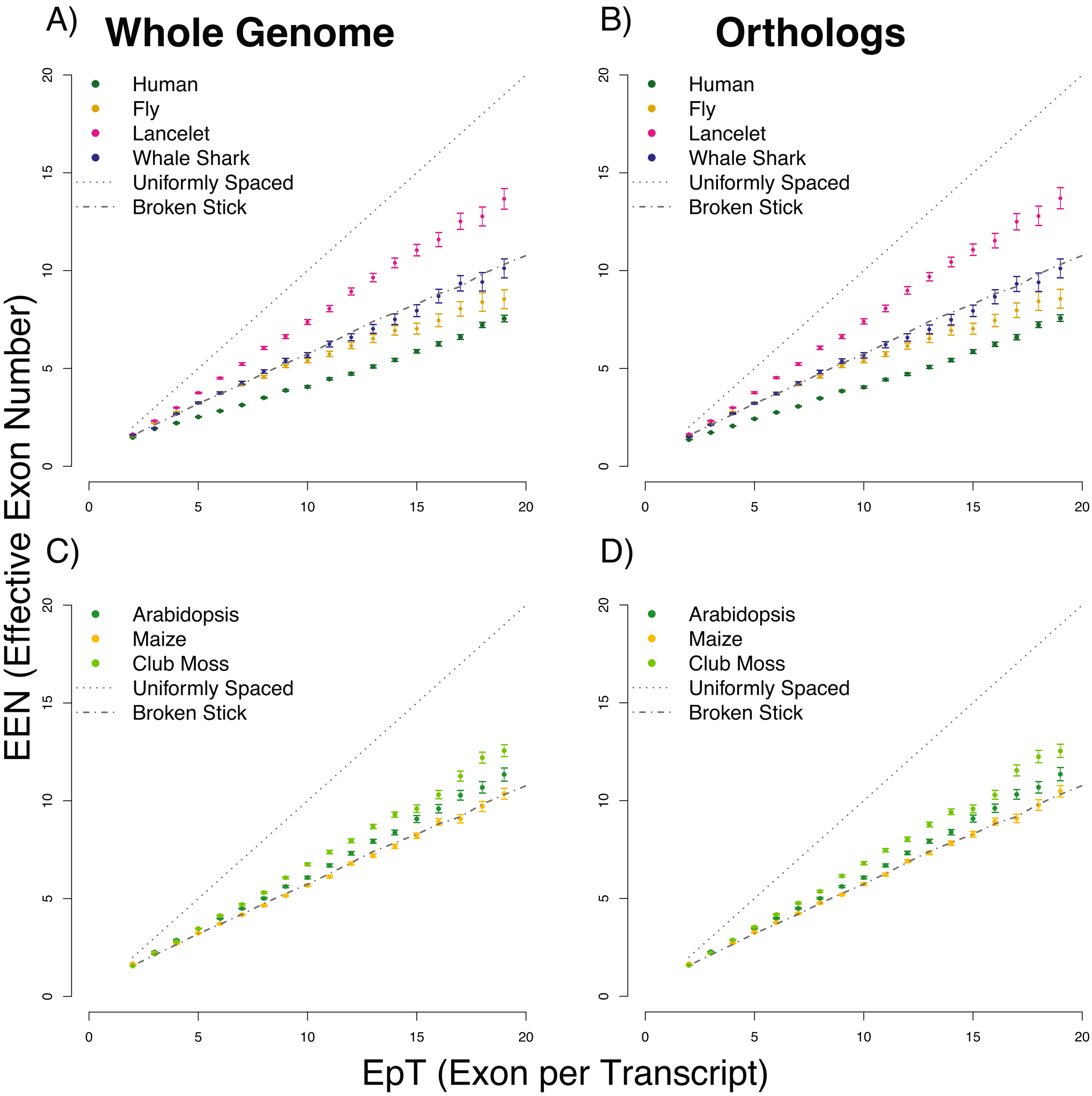

Fig. 5 Effective Exon Number (EEN) vs Exons per Transcript (EpT) (mean EEN ± 2*SE) in 4 species of animals and three species of plants. A/C) Whole-Transcriptome data and B/D) after conditioning on Orthologs being present in at least one species in the phylogeny. Ortholog data excludes lineage specific genes. EEN is expected to follow a Broken Stick model if intron bounds are randomly drawn from a uniform distribution. Humans and *D. melanogaster* show lower EEN with a bigger effect at transcripts with high EpT values. Whale sharks show EEN fully consistent with the null expectation. Lancelets show elevated EEN, the highest of any Deuterostome, suggesting more evenly distributed exon sizes than random. While orthologs show a nominal shift in the smallest EpT values, requiring orthology does not alter comparisons to a broken stick model. Hence, Broken Stick Model comparisons are likely to be robust to effects of orthologs and exclusion of lineage specific genes in evolutionary analysis

a departure with significantly lower EEN in 5/11 *Drosophila* transcript models. Such metrics indicate more tightly clustered intron breaks than expected if intron breaks are randomly chosen from a uniform distribution. This pattern holds true even in the well validated model organism *Drosophila melanogaster.* Such results suggest that there may be different biochemical or evolutionary constraints on exon junction placement and associated splicing processes for genes with large EpT. There are 4/11 *Drosophila* that show elevated EEN, suggesting a more even distribution of intron breaks than expected from a random uniform distribution.

Conversely, *D. sechellia*, only 0.5 million years divergent from *D. simulans*, shows elevated EEN compared to the null model. *D. sechellia* annotations showed a lower density of transcripts with higher EpT values (Supp. Mat. Whole-Transcriptome Vs Ortholog Density Plots), and often are annotated with only one transcript per gene. Annotations that collapse isoforms with alternate exons into a single transcript will obscure the true distribution of EEN, biasing statistics toward higher values. Whole genome divergence in transcriptome complexity across such short timescales would be surprising. Whether the ultimate variation of EEN reflect biology or artifacts, these results suggest that cross-species comparisons of complexity require normalization for whole genome differences (Supp. Figure 19).

The majority of plant transcript models lies above the Broken Stick model, with the highest EEN values in a club moss, *Selaginella moellendorffii*. Maize (*Z. mays*) aligns well with the Broken Stick model, but *Arabidopsis thaliana* lies well above. Sorghum agrees with maize (with overlapping error bars), except at the highest values of EpT (between 17–20 EpT) where it converges with *Arapbidopsis*.

Fungi show unusual variation in EEN compared with other clades. Most species show distributions of EEN centered above the Broken Stick model when EpT < 10. However, above 15 EpT, EEN increases toward more uniform spacing. These results appear more similar to patterns of exon distribution in plants than in animals. However, some fungi show atypical patterns as clear outliers in comparison with the Broken Stick model. A few unicellular fungi such as *S. cerevisiae* have few exons per transcript genome-wide, resulting in low variation in EEN, with little information via this metric. Fungi remain less well annotated in comparison with other clades, a gap that can now be addressed as cost and infrastructure for genome sequencing improves.

These results portray yet another class of variation among transcript model complexity across taxa, which is fortunately robust to the effects of ortholog calling. Analysis of splicing patterns using EEN and comparison to models like the Broken Stick may be informative as annotation projects assess quality and compare molecular evolutionary variation (Supp. Figure 20).

Evolutionary rate analysis

Understanding genetic complexity across the tree of life is essential to understand the processes that influence form and function of life. Analyzing genetic complexity in a phylogenetic context is important because biased estimation can lead to conceptually flawed interpretations of genetic function [65–67]. Furthermore, understanding the rate shift dynamics of complexities between orthologs and whole-transcriptomes gives empirical estimations to observe the variation between lineages' annotations containing novel genetic elements or shared genetic elements (orthologs). Here, we perform evolutionary analysis using transcript model complexity as a genetic trait to identify shifts in complexity across taxa and to explore biases introduced in ortholog conditioning.

When estimating rates of evolutionary complexity across phylogenies, we find that in some cases conditioning on orthologs causes a significant shift in rate estimates, depending on the taxonomic group and metric being used. Here we used Phytools to perform posthoc tests of evolutionary rate changes between whole-transcriptome and orthologous rates. All metrics in the deuterostome group are significantly different: TpG (t = -3.89, P = 2.00E-04), EpT (t = -2.60, P = 0.0108), and EpG (t = 4.48, P = 0). We find no significant difference in the *Drosophila* group between complexity metrics: TpG (t = -0.564, P = 0.580), EpT (t = -0.123, P = 0.904), and EpG (t = -0.314, P = 0.7577). While in other groups, we find some metrics are significantly different while others are not. Plantae orthologs compared to whole-transcriptome metrics are as follows TpG (t = -2.18, P = 0.0335), EpT (t = 0.423, P = 0.667), and EpG (t = -1.68, P = 0.0972). Where mean TpG is significantly different between ortholog and whole-transcriptome metrics, but non-significant for EpT and EpG. The fungi group, among classes, had no significant differences between orthologs and whole-transcriptome for all metrics: TpG (t = -0.0161, P = 0.9872), EpT (t = -0.461, P = 0.648), and EpG (t = -0.463, P = 0.647).

By estimating credible rate shifts (each defined as an event) from the posterior probabilities (PP), we observe how rates differ between whole-transcriptome and ortholog complexities across a phylogeny for EpT (Fig. 6; See Supp. Figs. 21, 22, 23, 24 for MCMC convergence). In deuterostomes we identify a shift (e.g., warmer heat color) for taxa related to humans. At the simian node [from *Callithrix jacchus* (common marmoset) to *Homo spaiens*] in the phylogeny containing whole-transcriptome EpT, we observe an elevated evolutionary rate [PP = 0.0972]. While the ortholog EpT evolutionary rate is also elevated at crown primates [PP = 0.149] There is also an elevated rate for the branch of *Branchiostoma floridae*. The rate is lower in whole-transcriptome EpT [PP = 0.0679] but higher, given the edge taxa, in the orthologous dataset [PP = 0.108]. Outside of *Branchiostoma floridae*, the rates on the phylogeny illustrates that all chordates have similar rates, outside of primates. Some of these rate shifts may be due to the artifact of human and primate genomes having more complete annotation with greater molecular evidence to support isoform detection.

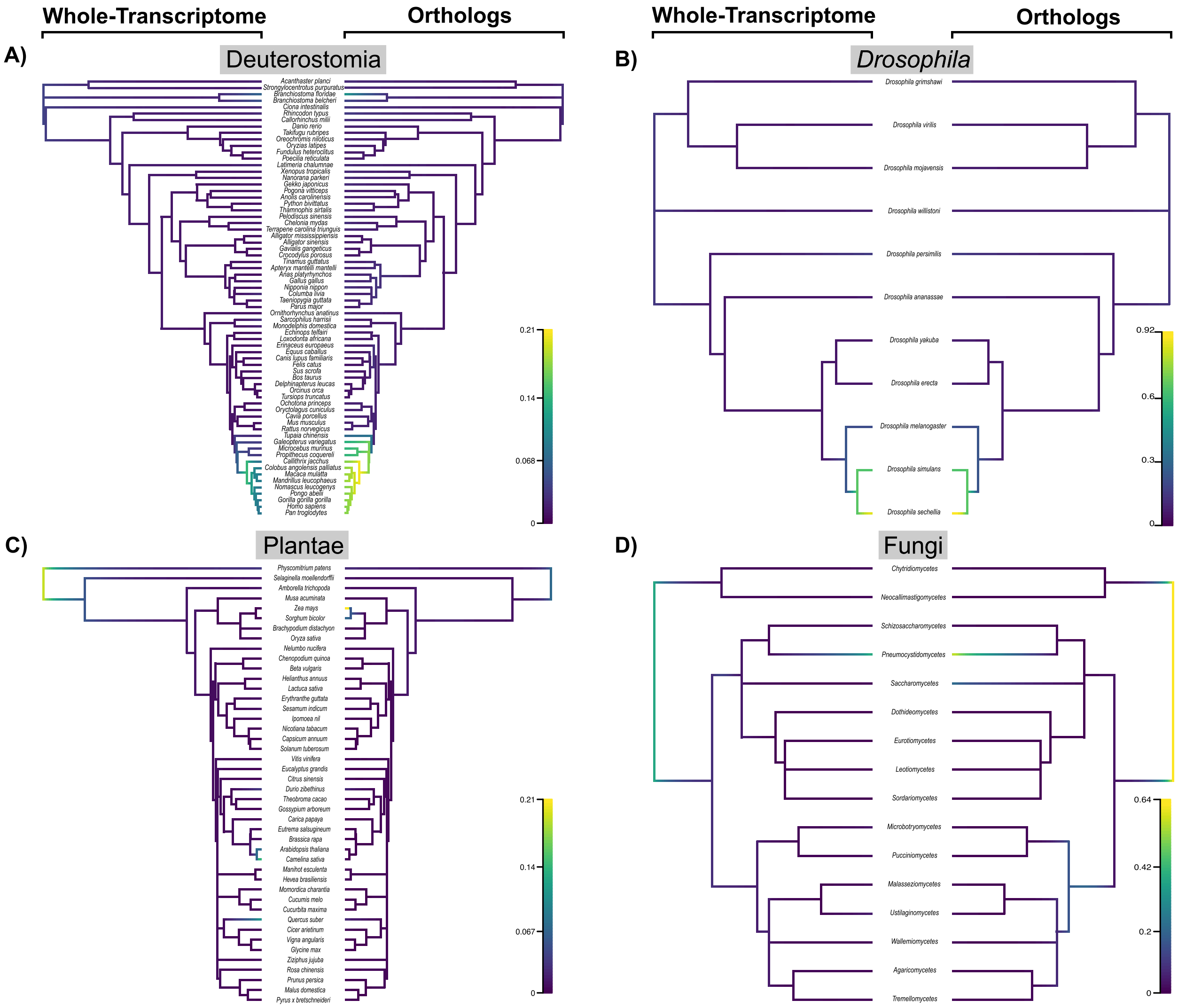

Fig. 6 Evolutionary rates of EpT complexities across phylogenies for A) Deuterostomia, B) *Drosophila*, C) Plantae, and D) Fungi. Fungi phylogenies are collapsed to classes. Phylogenies were generated from the TimeTree portal. Rates were calculated in BAMM with 10,000,000 generations with a 10% burn-in. Viridis color pallet illustrates higher marginal rate-shift probabilities, with warm colors signifying higher rates shift and cooler colors being closer to no rate-shift toward zero. Each facet is analyzed independently and not relative to each other. Plots were generated in BAMMTools R package. Evolutionary rates compare whole-transcriptomes (left trees) to orthologous genes (right trees). We observe a significant difference in evolutionary rates for Deuterostomia when conditioning on orthology, but not in other clades. Phylogenetic trees are scaled within each clade, with yellow being higher and blue being lower rates

Drosophila have an elevated rate of EpT at the node between *D. sechellia* and *D. simulans* for both whole-transcriptome [PP = 0.643] and orthologs [PP = 0.649], consistent with the *D. sechellia* annotation containing transcript models with a union of exons from multiple transcripts within a gene [63]. Elevated rate shifts within plants and deuterostomes occur at single branches of organisms that are highly annotated. Phylogenetic branch tips that have high-rate shifts include *Arabadopsis thalana* and *Camelina sativa* (false flax) [PP = 0.0592], and *Quercus suber* (cork oak) [PP = 0.110]. These plants are highly studied cash crops or, in the case of *Arabadopsis*, a genetic model system. While orthologous genes have a rate shift in *Z. mays* (maize) [PP = 0.278] with a higher mean EpT than other Potales, no shift is observed in whole-transcriptome annotations. There is a rate shift at the basal portion of the tree between vascular and non-vascular plants. This rate shift is lower in ortholog genes (PP = 0.00852), than it is whole-transcriptomes (PP = 0.0162). With larger genome size, plants contain more genes and a plethora of extra functions evolved in vascular plants compared to that of mosses [68]. Hence, we suggest these differences are likely biological.

There are rate shifts within fungi lineages, including a split between Chytridiomycota and all other fungi phyla (Basidiomycota and Ascomycota), PP = 0.0301 in whole-transcriptomes and PP = 0.0427 in orthologs. Another shift is observed in the class branch Pneumocystidomycetes, PP = 0.396 in whole-transcriptomes and PP = 0.570 in orthologs. The rate shifts are higher in orthologous genes than in whole-transcriptomic genes.

Biologically, complexity rate shifts are relatively rare across the tree of life (Fig. 6). We only observe high shifts in deep time or shift among highly studied biological systems. Shifts in complexity among branches may stem from annotation quality or vigor, given the specific taxa where shifts occur. The difference between whole-transcriptomes and orthologs have higher probabilities in rate shifts among orthologous dataset than of whole-transcriptomes. Overall, these observations suggest that analysis of evolutionary rates is not severely impacted by ortholog conditioning, unlike species-level complexity analyses.

Discussion

Analysis of transcript complexity

Over 95% of multi-exon genes are subject to alternative splicing in eukaryotes [69, 70]. As the genomic field expands, we have found that a tremendous amount of complexity is not simply found in how many single genes a genome possesses, but instead of driven by genetic machinery "cutting" and arranging genetic units for specific tasks. These dynamics are observed in the human genome with only ~ 25 k genes coding upwards of ~ 90 k proteins [71, 72]. These genetic mechanisms and genomic dynamics add extra layers to understand organismal complexity. Understanding complexity is a central tenet to evolutionary biology. As mutations accumulate, novel genes form and gain function, where speciation events occur, yielding new species. Understanding the processes that drive complexity across the tree of life, inherently, has the potential to illuminate biological diversity that we observe in nature.

Transcriptome complexity is influenced by many biological factors. Previous work has observed exceptional splicing patterns in the testes and heads of *Drosophila*, mice, and humans (reviewed in [73–76]. Alternative splicing in males and females produces functional differences in sex determination pathways (reviewed in [77]). The use of different transcripts across timepoints and tissues influences animal development and complexity of body forms [78]. Changes in alternative splicing and the addition of complex combinations in isoforms can allow for greater functional diversity and drive evolutionary innovation [2]. Genes expressed at different times, in different tissues, and developmental stages could have complexity bias as well [21], especially given the difference in U11/U12 spliceosome use. In taxa with many differentiated cell types, separate sexes, or life stages, total complexity among annotations may be amplified compared with taxa that exhibit uniformity with fewer cell types. Moreover, whole organism isoform use may be dynamic in comparison with the transcript diversity present in any single cell. We also find that the quality of the assembly has a small effect on transcript diversity (Supp. Figs. 25 and 26).

We observe different dynamics with fungi from other groups in regards to TpG being collapsed on single transcript genes. Fungal genome sequencing and annotation is a burgeoning field and still has artifacts as genomic sequencing remains challenging [79] and annotation of isoforms in many species appears incomplete [80] (Supp. Information, Supp. Figs. 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, and 16). Hence, a clear need exists for metrics that elucidate whether and how evolution has reshaped transcriptome complexity in different organisms. Furthermore, polycistronic gene express is a dynamic known to occur in eukaryotes, even though its potentially rare [81]. It is unclear how many of these might be candidates for combination into the same gene, but we expect that, overall, the numbers are not high enough to skew results in any meaningful way. Along with ways to improve annotation across the tree of life. Here, we use new metrics generated by TranD to measure trends of complexity accurately and precisely across taxa. Understanding complexity is quite important to understanding the dynamics that facilitate the machinery of life [82, 83].

These metrics, presented here, are robust since they are agnostic in the way they are generated from the data. These metrics also allow for independent diagnosis of complexity for a single organism along with the ability to compare each metric. Caveats for use cases are largely driven by the quality of annotation input data. TranD will collapse and collate transcript model complexity across genes only as accurately as the annotation data at hand. We observe such effects in a few constituents, such as in Deuterostomia, the primate group has higher than average metrics. This is likely a product of more thorough annotation of isoforms in the human genome based on extensive molecular data. In such a case, homology-based annotation in primates may use human genomes as a resource for more complete annotation than more distantly related organisms. However, it is also possible that real biological rate-shifts exist in primates, in addition to methodological factors. In contrast, *Drosophila sechellia, B. floridae* and whale sharks may show signs of collapsed isoforms among annotations, which yield unusual complexity results (Supplementary Information). However, this is not to say that these are not biological results. Do be aware that evolutionary inference is dependent on the completeness and accuracy of an annotation for an organism.

Ortholog-free complexity comparisons capture genetic novelty

Estimating complexity metrics that are conditioned on orthologous genetic elements, in our findings, induces bias for more complexity in older more studied genetic elements. Which do not take into account the novel gene formations that are found in whole-transcriptome annotations. This bias of older genetic elements will have complexity metrics that capture more well developed AS schemes, in turn, being more complex from the whole-transcriptome complexities that also take into account all the less complex novel elements. Inducing a shift in a higher frequency of lower complexity. These novel genetic elements arise from either re-organization of pre-existing genes or are de novo [35, 84]. Since novel genetic elements are, indeed, "new" less time has allowed for genes to accumulate more AS schema yielding more isoforms, hence more complexity in this study's metrics. Using orthologous genes in phylogenetic comparisons allows for one-to-one comparisons given a speciation event [85, 86] dubbed the "ortholog conjecture" [87]. This methodology is a "necessary evil" in the genomic age, because much of the data generated may not adhere to all data being of orthologous origins. Even in relatively recent studies, across multiple taxa using many genomes, have lacked the resolution for concordance of a single gene copy [88]. This is also not a unique occurrence in whole-genome phylogenetic studies as more taxa are incorporated [33, 89].

Annotations can be facilitated by homology from better studied species, where the differences in metrics between orthologous partitions and whole-transcriptome data are not significant. From the novel genetic element conditioning, we see that the variation of almost all species, is that ortholog metrics shift to higher complexities, while novel lineage specific genes have lower complexity. Hence, whole-transcriptome data will show different values than either of these subsets. Future phylogenetic analysis must be aware that there will be biases from annotations, whether from highly rigorously annotated species or annotations from understudied organisms. Studies should consider how their data performs, and potentially correct for ortholog bias based on the questions and aims that are at hand.

More studies have tested the notion if discarding paralogs is indeed the best practice [33, 86, 90]. Stamboulian et al. (2020) observed that paralogs still aid in functional prediction and throwing away huge swaths of data can often lead to the lack of predictions because no orthologs are available. Recent work has demonstrated that paralogous data should not be as feared as once thought for phylogenomic inference [33, 91]. Of course, this paper uses EEN to normalize complexity discrepancy between whole-transcriptome and orthologous complexity metrics. Using Hong et al.'s (2006) effective exon lengths model [64], EEN allows for whole-transcriptome and orthologous metrics to me normalized by the same scale. Thus, making a more concordant comparison between various datasets without the need to discard non-orthologous data (See Results: *Comparison to broken stick is robust to orthology*). Even with corrections to mitigate biases and loss of potentially informative data, we still recommend researchers use best practices for their experimental designs be it including whole-transcriptomes, condition on orthologs, or incorporate the EEN correction.

Novel lineage specific genes frequently appear with fewer exons and lower protein complexity than background genomic properties [92]. As new genes are added to the transcriptome, these sources of innovation may therefore add sequences that have unusual complexity compared with long-standing sequences. Our analysis confirms this hypothesis, and these effects are observed in comparisons of orthologs and ortholog-free whole-transcriptome analysis. We find that novel genes have lower complexity with respect to every metric for analysis. If complexity analysis only uses orthologous sequences, the unique properties of these novel sequences will be obscured. Moreover, bias in complexity may offer a false portrait of whole-transcriptome complexity in different organisms. However, in other use cases, such as comparisons of EEN to Broken Stick, novel genes do not alter results significantly as they condition on baseline complexity metrics of EpT. In future applications, users will undoubtedly wish to assess the impact of ortholog conditioning on the questions at hand, potentially with analytical corrections for biases. With quantitative, precise descriptions of complexity biases for orthologs and new genes, we can move forward with evolutionary analysis with greater power than if accepting orthologs solely as a "necessary evil." The working examples from analyses presented here show how transcript model analysis can be implemented for future insights as the number of successfully sequenced and annotated organisms expands, especially in previously non-model systems.

Metrics and use cases for future evolutionary analysis

Understanding how metrics perform in whole-transcriptome data is essential to compare performance in evolutionary analysis of transcript model complexity for analytical applications. If the number of isoforms depends on the number of exons, then we may expect correlations between EpT, EpG, and TpG. However, if isoform combinatorics decouple the number of exons and the number of isoforms, then we may expect distinct patterns from each metric as species diverge. In such a case, analysis with one metric could suggest evolutionary stability, masking real variation that could be apparent using alternative metrics. To diagnose these differences between our complexity metrics we estimated Pearson R correlation coefficients among each group.

This study offers examples for how metrics of transcript model complexity may reveal differences in genomes for divergent taxa. These complexity metrics can be utilized further than this current study to analyze newly annotated genomes in emerging systems. With the appearance of well annotated genomic datasets, it may soon be possible to determine how biological factors influence isoform variation and alternative splicing patterns. TranD is not only limited to publicly available annotations but can calculate complexity metrics on de novo annotations with a pipeline of choice like BRAKER, MAKER, or Augustus and validation with a BUSCO analysis [93–96]. It is, however, advised to pay close attention to annotation quality and completeness when using TranD to estimate complexity metrics, since the results are highly sensitive to the data given. We have observed markedly different modes of transcript model complexity among and between lineages across the tree of life.

From high complexity in model systems compared to their sister taxa as seen in *Z. mays* and humans to the simpler transcript model forms identified in many fungi, *Arabidopsis*, and *Drosophila*, potentials for biases in annotation of highly studied organisms compared to those less studied. Furthermore, the ploidy of an organism may have a role on these transcript complexity models. The ploidy levels of plants vary tremendously among the group, and annotation comparisons may not directly be comparable. Understanding the molecular and cellular underpinnings for such variation may reveal insights into fundamental biology in future studies. Similarly, new analyses similar to our current study may help us characterize observable phenomena in transcriptome evolution, with potential to illuminate processes, mechanisms, and dynamics of evolution in the tree of life.

As fields of biology continue to progress and generate more precise and accurate genomic sequences, these metrics will hold even more power to understand the role complexity has on the tree of life. These metrics are broad in nature, being able to be applied to any organism that possess a genome. For example, multicellular organisms are assumed to be more complex than unicellular representatives. Multicellular organisms have larger sizes and tissue types likely a caused by novel adaptation giving rise to new functional outlets for the transcriptome to evolve [18, 97]. Our metrics can be used to guide further research to understand the phenomenon of why lineages with less complex body plans and traits may have more complex genomes. As future genomic resources emerge, it may be possible to compare single celled and multicellular relatives to infer how transcriptomes evolve as novel body forms emerge. Similarly, we may ask whether genomic complexity correlates with transcriptome complexity, or how sexual reproduction compared to asexuality influences complexity fluctuations. Understanding complexity systematically can help to expound patterns in evolutionary tracts across the tree of life where different dynamics have a role in evolution.

Conclusions

This study illustrates the power and utility using TranD transcript model complexity metrics as both a comparative method and to independently validate transcriptome complexity across any and all lineage of life and viruses given an annotation file. Care must be conducted on the annotation data being fed into TranD, because any errors will be incorporated into the results. This work is a first step to elucidate complexity patterns and validate using whole-transcriptome sequences versus conditioning on orthologous genetic elements across the tree of life.

Throughout our observations and analyses of transcriptomes, using TranD's transcript models, we found some interesting phenomena. First is that complexity metrics for Fungi, in regards to transcripts per gene (TpG), were mainly single transcript genes across the entire group. Prompting us to ask, is this a biological phenomenon in how the group's translational machinery works, or is this an artifact in how fungus is annotated? Maybe more care needs to be taken in uncovering AS dynamics in the group. This study cannot say either way, just that this pattern is seen. We also find that in our early analyses that the chicken (*Gallus gallus*) annotation was missing a well annotated gene *Titin* [TTN] in the previous RefSeq annotation, GCF_000002315.5_GRCg6a. This annotation was indeed updated over the course of writing this manuscript and is updated here, GCF_016699485.2_bGalGal1.mat.broiler.GRCg7b. This new annotation does include *Titin*, showing that TranD can illuminate potential errors in annotation. Given that *Titin* is a highly studied and well-known gene, being the largest gene in vertebrates, and being present across the deuterostome phylogeny, it was obviously missed. This is luckily rectified in updated annotations.

Mosses are also a curious case. Given that over evolutionary time and structurally, Bryophta is considered some the least complex group of contemporary plants. However, in our study we found them to have the highest complexities within the Viridiplantae group. The club moss also has a high rate of transcript overlap compared to other eukaryotes ate 6.4%. Club moss being higher than expected may not be unexpected as green algae has high rights of polcistronic gene expression [98].

Observations of evolutionary rate shifts in EpT (Fig. 6) show patterns of elevated shifts among model organisms within their respective groups. Plants have elevated rates for whole-transcriptomes in the model system *Arabidopsis* and elevated rates in orthologs in *Z. mays*. This trend is also seen within the primate groups. One hypothesis is that rates here are driven by annotations being driven by homology from model systems, especially in the primates given our sampling. In *Arabidopsis*, being a highly studied genetic system, may simply have more resources poured into annotation compared to other plants. *Z. mays* has a high-rate shift within the orthologous annotation filters. This may be attributed to its extensive domestication of traits and genes. The consistency of rate shift appearing in model systems requires future investigation.

Further research should be conducted on complexity to understand the tempo of evolution. Co-evolutionary patterns between host and parasite genomes, tempo of evolution in specific groups, wild-type vs domesticated genome complexity, and unicellular verses multicellular dynamics are all interesting questions that can be addressed using these metrics of transcriptome complexity. Furthermore, the ability to validate annotation files for specific studies has utility downstream. The creativity of the scientific question at hand and the annotation available are the only bottlenecks when using these agnostic transcriptome complexity metrics.

Methods

TranD complexity calculations

We used the new transcript model analysis software TranD to calculate three metrics of transcript model complexity within species: TpG, EpT, and EpG [57] (https://doi.org/10.1101/2021.09.28.462251; https://github.com/McIntyre-Lab/TranD/wiki). The exons per gene (EpG) metric was initially derived from Spieth & Lawson (2005) [99]. These three generalized metrics describe the global phenotypic structure of transcriptomes and are used to quantify transcript complexity, from annotations, of lineages across the tree of life. TranD derives these metrics by consolidating exons and transcripts into their parent genes (for both exons and transcripts) and into specific transcripts (for exons only) to illustrate the dynamics between various coordinate systems, transcriptome structure, and alternative splicing (AS) among and between lineages. We used describe_transcriptome_complexity_GTF.py script to quantify transcript model complexity for each organism using GTF files obtained from NCBI RefSeq and GeneBank. Each organism was then consolidated to one file, per each group of interest [Deuterostome, *Drosophila*, Plantae, and Fungi], using merge_species_transcriptome_info_counts.py script. Complexities were described for partitions for orthologs filters and novel gene filters (see ortholog identification).

First we tested for normality using the Shapiro–Wilk Normality Test using the shapiro.test() function in R [100]. To evaluate the interdependence between our metrics in the data we present, we used the Pearson correlation statistic using the lm() function in R [101, 102]. If the data set was normally distributed, we used the Pearson correlation, if it was not the Spearman's rank correlation coefficient was used for nonparametric data (Supplementary Information).

Effective exon number

We estimated the "Effective Exon Number" for each transcript according to models previously developed by Hong et al. (2006) [64]. For each transcript, the Effective Exon Number (EEN, formerly reported as N_e) is given by $EEN = 1/(\sum_{i=1}^{EpT}(1/(L_e^2))$ where i goes from one to EpT, the total number of Exon per Transcript, and L_e is the exon length scaled to a proportion of total transcript length. EEN is naturally bounded by [0, EpT].

EEN depends directly on the distribution of exon lengths and the dispersion or clustering of intron positions in cDNA transcripts. EEN is equal to the number of exons (EEN = EpT) if a transcript contains evenly spaced introns (overdispersion) with equal exon lengths throughout the transcript. Lower values (EEN << EpT) represent more clustered intron positions and under dispersed distribution of exon lengths [64]. Theoretical predictions would suggest that exon fragment lengths should follow a Broken Stick model [103]. Where the model takes unit of length and randomly (and simultaneously) selecting break points from a uniform distribution breaking it into N pieces. Values above the Broken Stick model suggest biological or analytical factors that create more evenly spaced intron breaks than the null. Values below this null model suggest factors that create more tightly clustered distributions of intron breakpoints across the transcript.

Ortholog identification

We used TranD to analyze genomic annotations from clades of eukaryotes where reference genomes and whole-transcriptome annotations were available in RefSeq and OrthoDB. Organism annotation data was selected by covering tractable data from well-studied phylogenetic lineages in the Eukarya tree of life, where possible. Reference files were procured from the Refseq database in GTF format (http://www.ncbi.nlm.nih.gov/refseq/). Guidelines for proper annotation acquisition are as follows:

> i) Annotations must be available on RefSeq and OrthoDB 10v1 (or v9.1 for *Drosophila* only). RefSeq is a standardized public database that is actively curated providing the most comprehensive and rich annotations of the tree life available. This allows for the highest quality annotation among lineages to pull from in our analyses.
> ii) The reference organisms must be annotated and have an assembled and annotated reference genome. Again, to facilitate the best annotated sequence products available within lineages.
> iii) The reference must have an NCBI release. All RefSeq genome annotations have an NCBI release, meaning they were processed by biological experts using the RefSeq processing pipeline [104, 105].

Complete references were gathered where available. Some key lineages are not available on RefSeq at high quality (at the time of publication) and so were omitted, e.g., Myxini, Tardigrada, Onychophora, etc. For some organisms the Fungi group only GenBank (GCA) reference ($n = 16/77$) were available. To have the adequate phylogenetic representation required, in the fungi group, for robust estimation, we choose to use GCA references for some constituents with $n = 61/77$ having RefSeq releases (GCF).

Orthologs were gathered from the OrthoDB v10.1 portal, where taxon organism IDs were collected. Orthologs were considered by selecting the most recent common ancestor (MRCA) for the groups (level on OrthoDB) and follow a 1:Multiple selection of orthologs, which are genes in one species that have multiple orthologs in another species due to gene duplication events. Input files needed to parse novel and orthologous genes, downloaded from OrthoDB's data section (https://www.orthodb.org/?page=filelist), include < odb10v1_OG2genes.tab.gz > and < odb10v1_gene_xrefs.tab.gz >. Our code translates ortholog group IDs from OrthoDB to xrefs and parses NCBI reference GTFs generating a total of three GTF files – [original] whole-transcriptome GTF, ortholog GTF, and novel gene GTF. Ortholog selection was conducted through a series of python scripts to filter OrthoDB orthologs from novel lineage specific genes NCBI GTF files. Full descriptions with examples for using code can be found in on GitHub: https://github.com/jemcquillan/OrthoDB_Parser.

To estimate whether there are significant differences between the whole-transcriptome, orthologous, and novel genetic data we ran a two-sample Wilcoxon Rank Sum (Mann-Whittney Test) using the wilcoxon.test() function in R [106, 107]. Estimated *P*-values were adjusted with Bonferroni using the p.adjust() function in R [108]. We estimated if there are biases in complexity metric variances between orthologs and whole-transcriptomes by cross-validating with a Jackknife resampling with 10,000 replicates to ensure that differences ortholog subsets of the transcriptome are not biased by lack of independence. Given that the distributions of EpT, EpG, and TpG are independent per organism, a Fisher's combined probability test was conducted for each complexity metric using the R package poolr's fisher() function [109–112].

Evolutionary rate analysis

If complexity metrics differ for genes that have orthologs across phylogenies compared with complex datasets that include lineage specific genes, conditioning on orthology could introduce biases in evolutionary rate analysis of transcriptomes [26, 28]. To understand how ortholog complexities behave compared to whole-transcriptome complexities in downstream analyses, we ran evolutionary rates on complexity traits estimated using whole-transcriptome data and conditioning on genes with orthologs across the entire phylogeny. We compare evolutionary rates using PhyTools [113] and BAMM [114] to determine whether there are significant differences when conditioning on the presence of orthologs across taxa. We used PhyTools ratebytree() function [113, 115–117] to ascertain if the evolutionary rate between orthologs and whole-transcriptome data has any bearing on which dataset to use for a specific group. BAMM was used to find mean phylorate shifts across both whole-transcriptome complexity metrics and orthologous gene complexity metrics to compare variation in complexity rate shifts across the phylogeny. This was conducted to understand the dynamics between complexity between whole-transcriptome annotations vs. annotations conditioned on orthologous genes.

Phylogenetic analysis was calibrated using divergence times from the Time Tree of Life (TToL) [118, 119]. The TToL constructs phylogenetic relationships through meta-analysis of currently published time-calibrated phylogenies. Each broad group (Deuterostomia, *Drosophila*, Plantae, and Fungi) had a phylogeny constructed from the TimeTree portal. If the TimeTree database did not have a specific individual that had criteria for annotation for selection, we picked closely related sister taxa present in the tree as comparable divergence times at the same relative tip placement of the phylogeny. For Fungi, phylogenies are still ongoing for proper placement of many taxa, so metrics were consolidated by class, the taxonomic grouping currently offered in the TimeTree database.

We used Phytools ratebytree() function for continuous traits, using the "OU" model of trait evolution to see if the rate is equal among all trees, or if the rates or regimes can differ between trees. The phylogenies were identical for each group. We used mean TpG, EpT, and EpG independently as traits. Then having the evolutionary rate estimated by complexity trait for ortholog metrics and whole-transcriptome complexities. The posthoc() function [113] in PhyTools was then conducted to test if there was a significant difference between orthologous partitioned complexity metrics compared to whole-transcriptome complexity metrics.

BAMM was run across 10,000,000 generations with a sampling frequency every 1000 Markov Chains. Burnins, for all groups, were set to a 10% burnin (a total of 1000 Markov Chains). Plotting of the BAMM and downstream analyses were conducting using BAMMTools R package [120]. We checked effective sampling sizes (ESS) and convergence of MCMC runs. All our runs ran to convergence, with ESS > 200 in most cases. To assess if the ESS numbers could be higher, we ran multiple runs with longer generate time, and combined runs. Here the ESS numbers did not change, yet still ran to convergence. Given our analysis did converge and ESS numbers did not dramatically change, we proceeded given the data at hand. Mean phylorate shift was calculated from the getEventData() function in BAMMTools R package. Rate shift probabilities were gathered using BAMMTools built-in functions to extract posterior-probabilities of a species [getTipRates()], a monophyletic gorup [getCladeRates()], or a branch in the phylogeny [getMarginalBranchRateMatrix()].

Data acquisition and code

All data, links to data used, and code can be found at https://github.com/jemcquillan/Transcriptome_Complexity and https://github.com/jemcquillan/OrthoDB_Parser.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09326-0.

Additional file 1.

Acknowledgements
We would like to thank Dr. T. Rhyker Ranallo-Benavidez and Dr. Farnaz Fouladi for their help and knowledge providing insights into code and code accuracy. Furthermore, conversations with both Dr. Denis Machado and Dr. Alex Dornburg helped to provide proper statistical analysis for phylogenetic methods. Their help is much appreciated.

Conflict of interest
There are not conflicts of interest to declare.

Authors' contributions
JET and RLR designed, executed, and developed code for the analyses within the manuscript. AVN and LMM developed code designed to estimate complexity metrics. The primary draft was written by JET and RLR. All authors edited the manuscript, designed the scope of the project, and had intellectual input throughout the process. The author(s) read and approved the final manuscript.

Funding
This study was funding with NIH NIGMS R35GM133376 (RLR) and GM128193 (LMM).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors have no competing interests to declare.

Received: 2 February 2023 Accepted: 20 April 2023
Published online: 11 May 2023

References

1. Lynch M, Conery JS. The origins of genome complexity science. 2003;302(5649):1401–4.
2. Gilbert W. Why genes in pieces? Nature. 1978;271(5645):501–501.
3. C. elegans Sequencing Consortium*. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998;282(5396):2012–8. https://www.science.org/doi/full/10.1126/science.282.5396.2012.
4. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931–45.
5. Black DL. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell. 2000;103(3):367–70.
6. Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. Genome-wide analysis of transcript isoform variation in humans. Nat Genet. 2008;40(2):225–31.
7. Kalsotra A, Cooper TA. Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet. 2011;12(10):715–29.
8. Kragh-Hansen U, Minchiotti L, Galliano M, Peters T Jr. Human serum albumin isoforms: genetic and molecular aspects and functional consequences. Biochim Biophys Acta. 2013;1830(12):5405–17.
9. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010;463(7280):457–63.
10. Furlanis E, Scheiffele P. Regulation of neuronal differentiation, function, and plasticity by alternative splicing. Annu Rev Cell Dev Biol. 2018;34:451.
11. Moyer DC, Larue GE, Hershberger CE, Roy SW, Padgett RA. Comprehensive database and evolutionary dynamics of U12-type introns. Nucleic Acids Res. 2020;48(13):7066–78.
12. Shepard S, McCreary M, Fedorov A. The peculiarities of large intron splicing in animals. PLoS ONE. 2009;4(11): e7853.
13. Sánchez L. Sex-determining mechanisms in insects. Int J Dev Biol. 2004;52(7):837–56.
14. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engström PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B, Wahlestedt C.

Antisense transcription in the mammalian transcriptome. Science. 2005;309(5740):1564–6.
15. Hodges D, Bernstein SI. Genetic and biochemical analysis of alternative RNA splicing. Adv Genet. 1994;31:207–81.
16. Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 2005;6(5):386–98.
17. Schad E, Tompa P, Hegyi H. The relationship between proteome size, structural disorder and organism complexity. Genome Biol. 2011;12(12):1–13.
18. Sealfon RS, Wong AK, Troyanskaya OG. Machine learning methods to model multicellular complexity and tissue specificity. Nat Rev Mater. 2021;6(8):717–29.
19. Maine EM, Salz HK, Cline TW, Schedl P. The Sex-lethal gene of Drosophila: DNA alterations associated with sex-specific lethal mutations. Cell. 1985;43(2):521–9.
20. Schutt C, Nothiger R. Structure, function and evolution of sex-determining systems in Dipteran insects. Development. 2000;127(4):667–77.
21. Coronado-Zamora M, Salvador-Martínez I, Castellano D, Barbadilla A, Salazar-Ciudad I. Adaptation and conservation throughout the Drosophila melanogaster life-cycle. Genome Biol Evol. 2019;11(5):1463–82.
22. Bonner JT. The evolution of complexity. Princeton, NJ: Princeton Univ. Press; 1988.
23. Qian H, Shi PZ, Xing J. Stochastic bifurcation, slow fluctuations, and bistability as an origin of biochemical complexity. Phys Chem Chem Phys. 2009;11(24):4861–70.
24. Holland PW. Gene duplication: past, present and future. Semin Cell Dev Biol. 1999;10(5):541–7. https://www.sciencedirect.com/science/article/pii/S108495219990335X.
25. Adami C. What is complexity? BioEssays. 2002;24(12):1085–94.
26. Roy SW, Irimia M. Splicing in the eukaryotic ancestor: form, function and dysfunction. Trends Ecol Evol. 2009;24(8):447–55.
27. Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Archibald JM. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature. 2012;492(7427):59–65.
28. Irimia M, Roy SW. Origin of spliceosomal introns and alternative splicing. Cold Spring Harb Perspect Biol. 2014;6(6): a016071.
29. Shimeld SM, Holland PW. Vertebrate innovations. Proc Natl Acad Sci. 2000;97(9):4449–52.
30. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol. 2003;13(17):1512–7.
31. Kumar A. An overview of nested genes in eukaryotic genomes. Eukaryot Cell. 2009;8(9):1321–9.
32. Bravo GA, Antonelli A, Bacon CD, Bartoszek K, Blom MP, Huynh S, Knowles LL, Lamichhaney S, Marcussen T, Morlon H, Nakhleh LK, Oxelman B, Pfeil B, Schliep A, Wahlberg N, Werneck FP, Wiedenhoeft J, Willows-Munro S, Edwards SV. Embracing heterogeneity: coalescing the tree of life and the future of phylogenomics. PeerJ. 2019;7: e6399.
33. Smith ML, Hahn MW. New approaches for inferring phylogenies in the presence of paralogs. Trends Genet. 2021;37(2):174–87.
34. Long M. Evolution of novel genes. Curr Opin Genet Dev. 2001;11(6):673–80.
35. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010;20(10):1313–26.
36. Willis S, Masel J. Gene birth contributes to structural disorder encoded by overlapping genes. Genetics. 2018;210(1):303–13.
37. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16(1):1–14.
38. Maddison WP. Gene trees in species trees. Syst Biol. 1997;46(3):523–36.
39. Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool. 1970;19(2):99–113.
40. Day WH. Properties of the nearest neighbor interchange metric for trees of small size. J Theor Biol. 1983;101(2):275–88.
41. De Queiroz K, Donoghue MJ. Phylogenetic systematics and the species problem. Cladistics. 1988;4(4):317–38.
42. Dover GA. DNA turnover and the molecular clock. J Mol Evol. 1987;26(1):47–58.
43. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39(1):309–38.
44. Ellegren H, Parsch J. The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet. 2007;8(9):689–98.
45. O'Toole ÁN, Hurst LD, McLysaght A. Faster evolving primate genes are more likely to duplicate. Mol Biol Evol. 2018;35(1):107–18.
46. Begum T, Robinson-Rechavi M. Special care is needed in applying phylogenetic comparative methods to gene trees with speciation and duplication nodes. Mol Biol Evol. 2021;38(4):1614–26.
47. Ohno S. From So much "junk" DNA in our genome. In Evolution of Genetic Systems. Brookhaven Symp Biol. 1972;(23):366-70. https://www.scirp.org/(S(lz5mqp453edsnp55rrgjct55))/reference/ReferencesPapers.aspx?ReferenceID=1834025.
48. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999;151(4):1531–45.
49. Emes RD, Goodstadt L, Winter EE, Ponting CP. Comparison of the genomes of human and mouse lays the foundation of genome zoology. Hum Mol Genet. 2003;12(7):701–9.
50. Marques AC, Vinckenbosch N, Brawand D, Kaessmann H. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. Genome Biol. 2008;9(3):1–12.
51. Conant GC, Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 2008;9(12):938–50.
52. Demuth JP, Hahn MW. The life and death of gene families. BioEssays. 2009;31(1):29–39.
53. Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet. 2009;10(1):19–31.
54. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 2010;11(2):97–108.
55. Rogers RL, Bedford T, Lyons AM, Hartl DL. Adaptive impact of the chimeric gene Quetzalcoatl in Drosophila melanogaster. Proc Natl Acad Sci. 2010;107(24):10943–8.
56. Rogers RL, Hartl DL. Chimeric genes as a source of rapid evolution in Drosophila melanogaster. Mol Biol Evol. 2012;29(2):517–29.
57. Nanni AV, Titus-McQuillan JE, Moskalenko O, Pardo-Palacios F, Liu Z, Conesa A, Rogers RL, McIntyre LM. The evolution of splicing: transcriptome complexity and transcript distances implemented in TranD. BioRxiv. 2021:09.28.462251. https://www.biorxiv.org/content/10.1101/2021.09.28.462251v1.full.
58. Jacob F. Evolution and tinkering. Science. 1977;196(4295):1161–6.
59. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 2009;25(9):404–13.
60. Siepel A. Darwinian alchemy: Human genes from noncoding DNA. Genome Res. 2009;19(10):1693–5.
61. Jarosz DF, Taipale M, Lindquist S. Protein homeostasis and the phenotypic manifestation of genetic diversity: principles and mechanisms. Annu Rev Genet. 2010;44(1):189–216.
62. Yang H, Jaime M, Polihronakis M, Kanegawa K, Markow T, Kaneshiro K, Oliver B. Re-annotation of eight Drosophila genomes. Life Science Alliance. 2018;1(6):e201800156.
63. Shiao MS, Chang JM, Fan WL, Lu MYJ, Notredame C, Fang S, Kondo R, Li WH. Expression divergence of chemosensory genes between Drosophila sechellia and its sibling species and its implications for host shift. Genome Biol Evol. 2015;7(10):2843–58.
64. Hong X, Scofield DG, Lynch M. Intron size, abundance, and distribution within untranslated regions of genes. Mol Biol Evol. 2006;23(12):2392–404.
65. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. Universal trees based on large combined protein sequence data sets. Nat Genet. 2001;28(3):281–5.
66. Bininda-Emonds OR. The evolution of supertrees. Trends Ecol Evol. 2004;19(6):315–22.
67. Gevers D, Vandepoele K, Simillion C, Van de Peer Y. Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. 2004;12(4):148–54.
68. Bainard JD, Newmaster SG, Budke JM. Genome size and endopolyploidy evolution across the moss phylogeny. Ann Bot. 2020;125(4):543–55.

69. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40(12):1413–5.
70. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470–6.
71. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Kalush F. The sequence of the human genome. Science. 2001;291(5507):1304–51.
72. Valdivia HH. One gene, many proteins: alternative splicing of the ryanodine receptor gene adds novel functions to an already complex channel protein. Circ Res. 2007;100(6):761–3.
73. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Çolak R, Kim T, Misquitta-ali CM, Wilson MD, Kim PM, Odom DT, Frey, BJ, Blencowe BJ. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;338(6114):1587–93.
74. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012;338(6114):1593–9.
75. Gibilisco L, Zhou Q, Mahajan S, Bachtrog D. Alternative splicing within and between Drosophila species, sexes, tissues, and developmental stages. PLoS Genet. 2016;12(12): e1006464.
76. Naro C, Cesari E, Sette C. Splicing regulation in brain and testis: common themes for highly specialized organs. Cell Cycle. 2021;20(5–6):480–9.
77. Salz HK. Sex determination in insects: a binary decision based on alternative splicing. Curr Opin Genet Dev. 2011;21(4):395–400.
78. Gustincich S, Sandelin A, Plessy C, Katayama S, Simone R, Lazarevic D, Hayashizaki Y, Carninci P, Carninci P. The complexity of the mammalian transcriptome. J Physiol. 2006;575(2):321–32.
79. Wang Z, Nilsson, RH, James TY, Dai Y, Townsend JP. Future perspectives and challenges of fungal systematics in the age of big data. Biol Microfungi. 2016:25-46. https://link.springer.com/chapter/10.1007/978-3-319-29137-6_3.
80. Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, Kang D, Underwood J, Grigoriev IV, Figueroa M, Schilling JS, Chen F, Wang Z. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. PLoS ONE. 2015;10(7): e0132628.
81. Lee SJ. Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure. Proc Natl Acad Sci. 1991;88(10):4250–4.
82. McShea DW, Brandon RN. Biology's first law: the tendency for diversity and complexity to increase in evolutionary systems. University of Chicago Press; 2010.
83. Day T. Computability, Gödel's incompleteness theorem, and an inherent limit on the predictability of evolution. J R Soc Interface. 2012;9(69):624–39.
84. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011;12(10):692–702.
85. Rentzsch R, Orengo CA. Protein function prediction–the power of multiplicity. Trends Biotechnol. 2009;27(4):210–9.
86. Studer RA, Robinson-Rechavi M. How confident can we be that orthologs are similar, but paralogs differ? Trends Genet. 2009;25(5):210–6.
87. Nehrt NL, Clark WT, Radivojac P, Hahn MW. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol. 2011;7(6): e1002073.
88. Thomas GW, Dohmen E, Hughes DS, Murali SC, Poelchau M, Glastad K, Richards S. Gene content evolution in the arthropods. Genome Bio. 2020;21(1):1–14.
89. Emms DM, Kelly S. STAG: species tree inference from all genes. BioRxiv. 2018:267914. https://www.biorxiv.org/content/10.1101/267914v1.full.
90. Stamboulian M, Guerrero RF, Hahn MW, Radivojac P. The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. Bioinformatics. 2020;36(Supplement_1):i219–26.
91. De Oliveira Martins L, Mallo D, Posada D. A Bayesian supertree model for genome-wide species tree reconstruction. Syst Biol. 2016;65(3):397–416.
92. Wang W, Zheng H, Yang S, Yu H, Li J, Jiang H, Su J, Yang L, Zhang J, McDermott J, Samudrala R, Wang J, Yang H, Yu J, Kristiansen K, Wong GK, Wang J. Origin and evolution of new exons in rodents. Genome Res. 2005;15(9):1258–64.
93. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Yandell M. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18(1):188–96.
94. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637–44.
95. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. Gene prediction: methods and protocols. 2019;65-95. https://link.springer.com/protocol/10.1007/978-1-4939-9173-0_5.
96. Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: assessing genomic data quality and beyond. Current Protocols. 2021;1(12): e323.
97. Levine M, Tjian R. Transcription regulation and animal diversity. Nature. 2003;424(6945):147–51.
98. Gallaher SD, Craig RJ, Ganesan I, Purvine SO, McCorkle SR, Grimwood J, Merchant SS. Widespread polycistronic gene expression in green algae. Proc Natl Acad Sci. 2021;118(7):e2017714118.
99. Spieth J, Lawson D. From Overview of gene structure (in press). In WormBook. 2005. http://www.wormbook.org/chapters/www_overviewgenestructure/overviewgenestructure.html.
100. Royston JP. An extension of Shapiro and Wilk's W test for normality to large samples. J Roy Stat Soc: Ser C (Appl Stat). 1982;31(2):115–24.
101. Wilkinson GN, Rogers CE. Symbolic description of factorial models for analysis of variance. J Roy Stat Soc: Ser C (Appl Stat). 1973;22(3):392–9.
102. Chambers JM, Hastie TJ. Linear models. From Chapter 4. In Statistical Models. in S. Wadsworth & Brooks/Cole; 1992. https://www.taylorfrancis.com/chapters/edit/10.1201/9780203738535-4/linear-models-john-chambers.
103. Holst L. On the lengths of the pieces of a stick broken at random. J Appl Probab. 1980;17(3):623–34.
104. Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res. 2009;37(suppl_1):D32–6.
105. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.
106. Bauer DF. Constructing confidence sets using rank statistics. J Am Stat Assoc. 1972;67(339):687–90.
107. Hollander M, Wolfe DA. From Nonparametric Statistical Methods. In Series In Probability And Mathematical Statistics. New York: Wiley; 1972. p. 27–33 (one-sample), 68–75 (two-sample). https://www.wiley.com/en-us/Nonparametric+Statistical+Methods,+3rd+Edition-p-9780470387375.
108. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ. 1995;310(6973):170.
109. Fisher RA. Statistical Methods for Research Workers. 4th ed. Edinburgh: Oliver and Boyd; 1932.
110. Brown MB. 400: A method for combining non-independent, one-sided tests of significance. Biometrics. 1975;31(4):987–92.
111. Higham NJ. Computing the nearest correlation matrix: A problem from finance. IMA J Numer Anal. 2002;22(3):329–43.
112. Cinar O, Viechtbauer W. The poolr package for combining independent and dependent p values. J Stat Softw. 2022;101(1):1–42.
113. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012;2:217–23.
114. Rabosky DL. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. PLoS ONE. 2014;9(2): e89543.
115. O'Meara BC, Ané C, Sanderson MJ, Wainwright PC. Testing for different rates of continuous trait evolution using likelihood. Evolution. 2006;60(5):922–33.
116. Adams DC. Comparing evolutionary rates for different phenotypic traits on a phylogeny using likelihood. Syst Biol. 2012;62:181–92.
117. Revell LJ, González-Valenzuela LE, Alfonso A, Castellanos-García LA, Guarnizo CE, Crawford AJ. Comparing evolutionary rates between trees, clades and traits. Methods Ecol Evol. 2018;9(4):994–1005.
118. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 2006;22(23):2971–2.

119. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol. 2017;34(7):1812–9.
120. Rabosky DL, Grundler M, Anderson C, Title P, Shi JJ, Brown JW, Huang H, Larson JG. BAMM tools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. Methods Ecol Evol. 2014;5(7):701–7.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.