Nucleic Acids Research, 2024, **52**, e28
https://doi.org/10.1093/nar/gkae056
Advance access publication date: 10 February 2024
Methods

Nucleotide-level distance metrics to quantify alternative splicing implemented in *TranD*

Adalena Nanni[1,2], James Titus-McQuillan[3], Kinfeosioluwa S. Bankole[1,2], Francisco Pardo-Palacios[4], Sarah Signor[5], Srna Vlaho[6], Oleksandr Moskalenko[7], Alison M. Morse[1,2], Rebekah L. Rogers[3,*], Ana Conesa[4] and Lauren M. McIntyre[1,2,*]

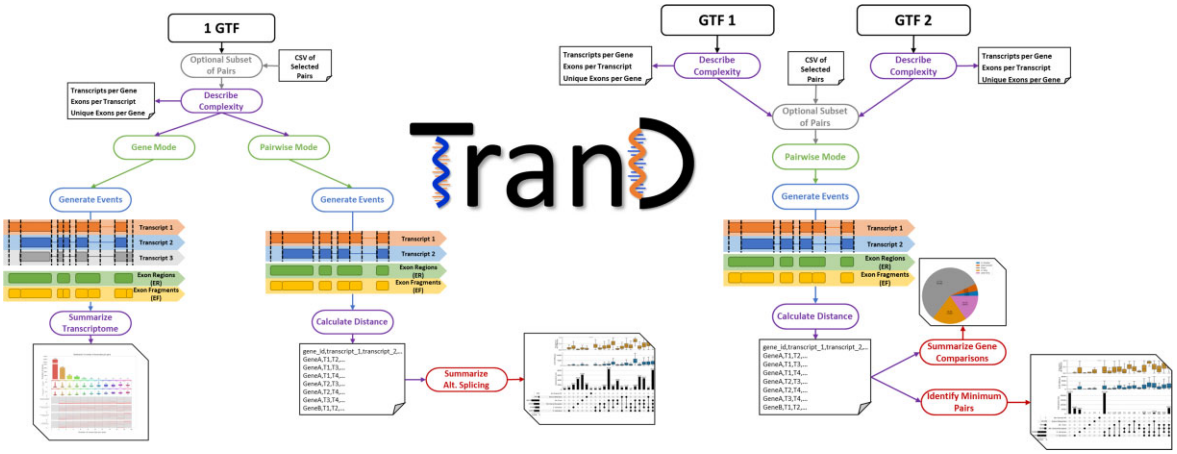
[1]Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL 32611, USA
[2]University of Florida Genetics Institute, University of Florida, Gainesville, FL 32611, USA
[3]University of North Carolina at Charlotte Department of Bioinformatics and Genomics Charlotte, NC, USA
[4]Institute for Integrative Systems Biology. Spanish National Research Council, Paterna, Spain
[5]Department of Biological Sciences, North Dakota State University, Fargo, ND, USA
[6]Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA
[7]University of Florida Research Computing, University of Florida, Gainesville, FL 32611, USA

[*]To whom correspondence should be addressed. Tel: +1 352 273 8024; Fax: +1 352 273 8284; Email: mcintyre@ufl.edu
Correspondence may also be addressed to Rebekah L. Rogers. Tel: +1 704 687 1321; Email: rebekah.rogers@uncc.edu

Abstract

Advances in affordable transcriptome sequencing combined with better exon and gene prediction has motivated many to compare transcription across the tree of life. We develop a mathematical framework to calculate complexity and compare transcript models. Structural features, i.e. intron retention (IR), donor/acceptor site variation, alternative exon cassettes, alternative 5′/3′ UTRs, are compared and the distance between transcript models is calculated with nucleotide level precision. All metrics are implemented in a PyPi package, *TranD* and output can be used to summarize splicing patterns for a transcriptome (1GTF) and between transcriptomes (2GTF). *TranD* output enables quantitative comparisons between: annotations augmented by empirical RNA-seq data and the original transcript models; transcript model prediction tools for longread RNA-seq (e.g. FLAIR versus Isoseq3); alternate annotations for a species (e.g. RefSeq vs Ensembl); and between closely related species. In *C. elegans, Z. mays, D. melanogaster, D. simulans* and *H. sapiens*, alternative exons were observed more frequently in combination with an alternative donor/acceptor than alone. Transcript models in RefSeq and Ensembl are linked and both have unique transcript models with empirical support. *D. melanogaster* and *D. simulans,* share many transcript models and long-read RNAseq data suggests that both species are under-annotated. We recommend combined references.

Graphical abstract



Introduction

Advances in sequencing technology have facilitated an explosion of genomic data, with an ever-increasing number of species having genomes sequenced and annotated (e.g. (1–4). This wealth of data offers tremendous potential for exploring fundamental questions in biology. Many eukaryotic genes are transcribed into multiple mRNAs through alternative splicing (AS) leading to differing functional properties (e.g. (5))

Received: July 21, 2023. **Revised:** November 29, 2023. **Editorial Decision:** January 6, 2024. **Accepted:** January 18, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

and/or encoded proteins from the same gene (e.g. (6–9). This differential processing has been hypothesized to be a primary mechanism of protein diversity in a variety of eukaryotes (e.g. (6,10–13). There is a positive relationship between organismal complexity and splicing (14). Dramatic differences in splicing events among tissues, environments, cell types, and developmental stages within a single species are well documented across the tree of life (e.g. (15–25). Evidence of interspecific splicing has also been described across a variety of plant (e.g. (26), invertebrate (e.g. (27,28), and vertebrate (e.g. (29)) species, but there are many challenges associated with cross-species transcriptome comparisons that may lead to inaccurate interpretations (e.g. (30)).

Multiple different molecular mechanisms contribute to the process of AS (e.g. (31–33). The spliceosome is a large complex with some conserved splicing factors (34–36) and a plethora of context specific factors (36), some of which may be shared along specific evolutionary branches (37). More evolutionary distant species (such as plants and mammals) have differences in splicing patterns (e.g. (38–41) that can be due to underlying mechanistic changes in splicing,. While we can visually separate alternative splicing patterns for each gene, we lack the ability to computationally group similar transcripts and identify patterns among transcript models.

Recent advancements in sequencing technologies have shed light on the importance of AS in adaptation and ecological speciation (reviewed in (42)). Through the inclusion/exclusion of exons or alterations in donor/acceptor sites, AS can generate variation in protein sequences. There is selection in alternative exons that is correlated with inclusion/exclusion rates (43), and alternative splicing and gene duplication appear to be inversely correlated evolutionary mechanisms in mammals (44). Additionally, AS plays a role in phenotypic plasticity, allowing genotypes to exhibit diverse phenotypes in varying environments (e.g. (45)), as well as ecological adaptations (e.g. (46,47)). Alternative splicing profiles can be species-specific (e.g. (29)), sex-specific (e.g. (48–50), Drosophila sex-determination pathway (e.g. (51), and reviewed in (52)), and may contribute to the resolution of conflicts between sexes during speciation (28,53). Although challenges remain in studying alternative splicing, long-read sequencing offers promising opportunities for further research in understanding how alternative splicing facilitates evolution.

Genomic approaches comparing transcript models from a 'new' species to a reference annotation from a similar species are usually conducted via BLAST (e.g. (54,55)). These annotations are evaluated for 'completeness' primarily by examining the predicted protein composition of a transcriptome (56). Building annotations for a new species is an ongoing effort in tool development. Popular genome annotation tools such as MAKER (57), BRAKER (58,59), AUGUSTUS (60,61), StringTie (62,63), and PASA (64) have emerged as powerful and widely used tools for genome annotation, leveraging both *ab initio* and evidence-based methods to annotate genes, assemble transcripts, and identify functional elements. Furthermore, these tools can help identify regulatory regions, structural variation, and other genomic features that are essential for understanding the biology of the organism under study. Given the widespread interest in genome annotation and its importance for a wide range of applications, we anticipate continued growth in the development and refinement of genome annotation tools for new species in the coming years. However, we lack tools that enable us to evaluate the impact of bioinformatic choices on the resulting structure of the transcript models.

A benefit to long-read sequencing is the ability to directly sequence full length mRNA molecules, providing an opportunity to directly observe splicing patterns. As with any new technology, there are challenges as well as opportunities in the use of lrRNA-seq. The challenges of data processing, annotation, and interpretation of long-read data have inspired many computational tools. A database has been set up to track and describe the efforts in this area (65). As of this submission, over 720 tools were listed in this database. Of the tools developed for lrRNA-seq, most produce transcript models. There are at least 36 analysis tools for processing lrRNA-seq and estimating transcript models (66). Benchmarking is performed by calculating 'accuracy', 'sensitivity' and 'specificity' relative to some objective truth, usually determined by simulation or by spike-in synthetic standards (67,68). Comparisons between approaches are based on comparisons of the performance of these metrics between tools. Many lrRNA-seq experiments will also compare the *de novo* transcript models to a reference annotation and identify exact matches for splice-junctions (full-splice matches, FSM) using tools such as SQANTI (69), GFFcompare (70), TALON (Wyman *et al.* 2020), FLAMES (71) and IsoTools (72). However, when the structure of transcript models does not match the reference there are no tools for understanding the structural differences, or the distance, between the transcript models and the reference.

We have developed a set of distance metrics, and an accompanying software *TranD* that can be used to calculate distance metrics between transcript models. These metrics can be calculated between transcript models for each gene in an annotation (1GTF) and between pairs of transcript models for the same gene present between two annotations (2GTF). With these metrics we can group transcript models, pinpoint nucleotide level differences in alternative splicing between pairs of transcript models for all possible pairs, describe patterns of alternative splicing within and between transcriptomes. We illustrate the utility of these metrics and the resulting patterns in the examination of short and long read RNA-seq data in 6 different species.

Materials and methods

Starting from a GTF file, we describe alternative splicing patterns for each gene: alternative exons, intron retention, alternative donor/acceptors and alternative 5′, 3′ variation in terms of the number of nucleotides that differ using distance metrics. Our distance metrics focus on the structural differences between transcript models and report distance based on genomic positions of sequences, not the sequence itself (A,T,G,C) or on percent sequence identity. Exons that correspond to a single region that do not overlap with an exon in other transcript models are alternative exon cassettes. An exon region (ER) is defined by the boundaries of the 5′ most and 3′ most exon coordinates when exon annotation overlaps among transcript models. The exonic space is the union of all exon regions. Alternative splicing (AS) categories are associated with ERs (Figure 1A) and the number of nucleotides associated with each AS category are counted: (i) 5′ transcript length variation, (ii) 3′ transcript length variation, (iii) alternative donor/acceptor, (iv) alternative exon cassette, (v) intron retention (IR) and (vi) non-overlapping (referred to here as 'No Shared Nucleotides'). The presence/absence of each

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

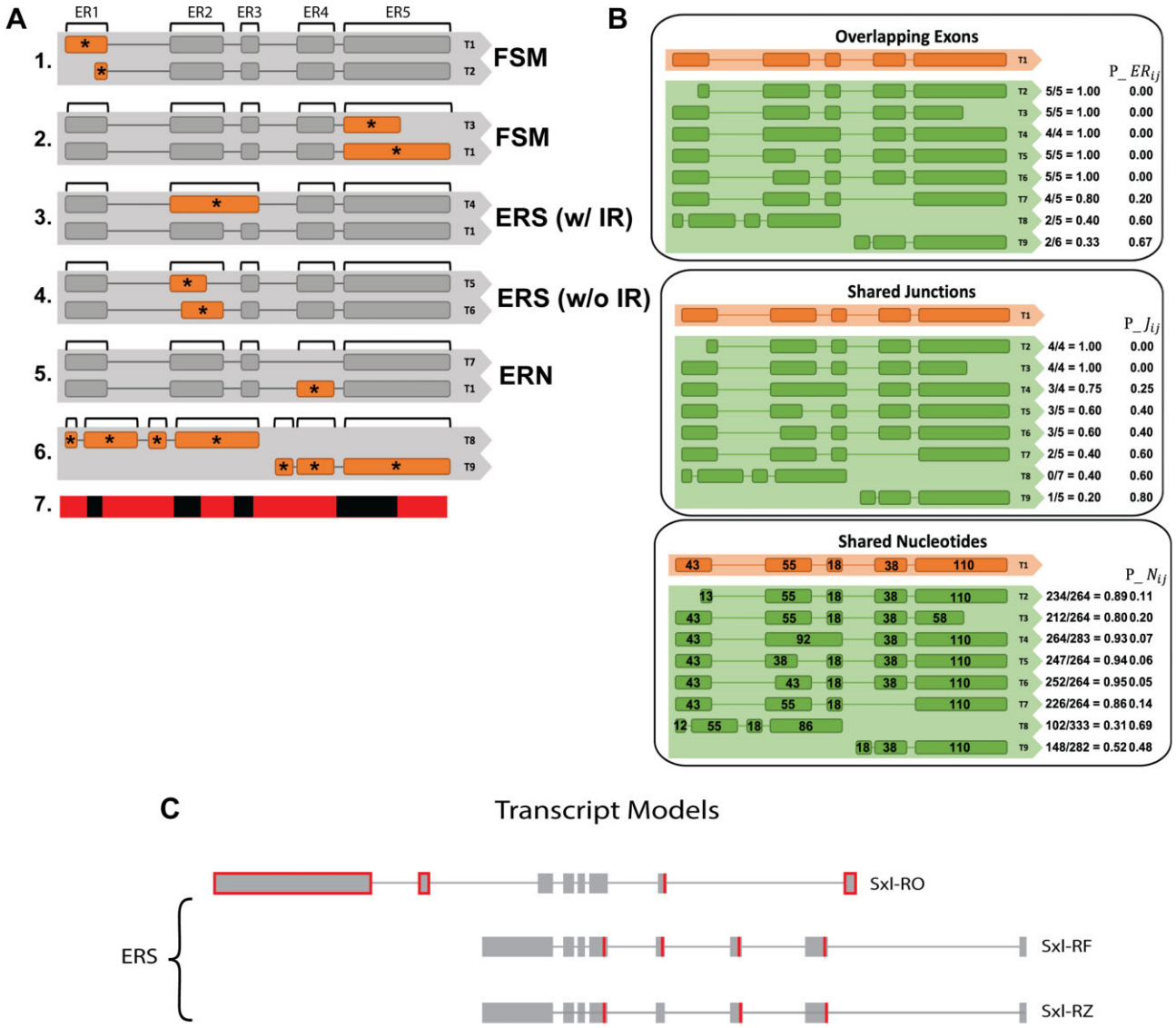


Figure 1. Alternative splicing is relative (**A**). 1. Two transcript models that differ at the transcription start site in the first exon region (ER1), and match on all junctions (Full Splice Match; FSM). 2. Two transcript models that differ at the transcription termination site in the last exon region (ER5); and FSM. 3. Two transcript models that differ but whose exons overlap (ERS) with an intron retention. 4. Two transcript models that differ but ERS with alternative donor/acceptor. 5. Two transcript models that differ with an alternative exon, so the exon regions do not overlap (ERN). 6. Two transcript models that differ with no exon overlap (no shared nucleotides). 7. Positions with variable annotations are indicated in red. (**B**) Three metrics for comparing transcript models. Each green transcript model is compared pairwise to the orange model and metrics are reported for (i) number of overlapping exons, percent of overlapping exons, (ii) number of shared junctions, percent of shared junctions (iii) number of overlapping nucleotides, percent of shared nucleotides. (**C**) Transcript models can be grouped based on overlapping exons. Three transcript models from *D. melanogaster* Sxl RF, RO and RZ. Red indicates a difference in nucleotide position. Models RF and RZ having overlapping exon regions (ERS) and can be grouped together into an exon-region group (ERG).

category can be used to summarize the patterns of alternative splicing across genes and describe the structural variation in the transcriptome. In addition, metrics for the number of exons per gene (EpG), transcripts per gene (TpG), and the number of exons per transcript (EpT) (73–75) are calculated and output as transcriptome_complexity_counts.csv and plots of the distributions for these metrics are output as complexity_plots.png (Table 1).

Distance metrics based on Jaccard Distances (76) identify the proportion of the exon space that varies for each of the 6 categories of alternative splicing and are provided as well as summary plots in the output of the open source *PyPi* package, *TranD*. We provide details and examples, with additional scripts for processing the *TranD* output on our github page (https://github.com/McIntyre-Lab/TranD).

Distance metrics at the gene level (TranD 1GTF gene)

With 1 GTF file as input, for each gene g, with t transcript models ($t > 1$), the ER space is defined by the boundaries of the 5′ most and 3′ most exon coordinates across the t transcript models. We calculated the following quantitative metrics:

$$P\_ER_g = \frac{(number\ of\ ER\ shared\ between\ transcripts)}{total\ number\ of\ ER\ present\ in\ either\ transcript}$$

$$P\_N_g = \frac{(num\ nucleotides\ annotated\ as\ exonic\ in\ any\ transcript\ model)}{total\ number\ of\ nucleotides}$$

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

Table 1. Primary distance metrics. The descriptions and abbreviations of primary distance metrics used. All metrics are described in detail on the TranD github (https://github.com/McIntyre-Lab/TranD/wiki/Output-File-Column-Descriptions)

Metric	Abbreviation	Description
Exons per Gene	EpG	number of exons annotated within a gene
Transcripts per Gene	TpG	number of transcripts annotated within a gene
Exons per Transcript	EpT	number of exons annotated within a transcripts
Proportion of exon regions shared	P_{ER_g}	(number of ER shared between transcripts)/(total number of ER present in either transcript)
Proportion of nucleotides shared	P_{N_g}	(number nucleotides annotated as exonic in any transcript model)/(total number of nucleotides)
Proportion of variable exon regions	$P_{ER_{ij}}$	$(num_{ER_i} + num_{ER_j})/total_{ER_{ij}}$, where num_{ER_i} is the number of exon regions only annotated in T_i, num_{ER_j} is the number of exon regions only annotated in T_j, and $total_{ER_{ij}}$ is the union of unique exon regions annotated in either transcript model T_i or T_j
Proportion of variable junctions	P_J_{ij}	$(num_{J_i} + num_{J_j})/total_{J_{ij}}$, where num_{J_i} is the number of junctions only annotated in T_i, where num_{J_j} is the number of junctions only annotated in T_j, and $total_{N_{ij}}$ is the union of unique junction coordinates annotated in either T_i or T_j
Proportion of variable nucleotides	$P_{N_{ij}}$	$(num_{N_i} + num_{N_j})/total_{N_{ij}}$, where num_{N_i} is the number of nucleotide co-ordinates annotated as part of T_i only, num_{N_j} is the number of nucleotide co-ordinates annotated as part of T_j only, and $total_{N_{ij}}$ is the number of nucleotide co-ordinates annotated in either T_i or T_j
Percent of annotated nucleotides	$1 - P_{N_{ij}}$	1 minus the proportion of variable nucleotides $P_{N_{ij}}$

For each gene, the P_ER and P_N (all_gene_prop_nt_variability.csv) as well as the number of exons per gene (uniq_exons_per_gene.csv), are output and summary statistics are reported in output transcriptome_complexity_counts.csv. These metrics provide insights into the complexity of the annotation. Species can be compared for complexity without needing to refer to the genome coordinates (30). By default, *TranD* 1GTF produces a summary for each gene of the presence/absence of each of the 6 AS categories (pairwise_distance.csv), a graphical summary of the complexity in transcriptome_summary_plot.png; all_gene_prop_nt_variability.png (the distribution of P_N_g), and catalog files for junctions (junction_catalog.csv), exon regions (event_analysis_er.csv), and a list of transcript models with IR (ir_transcripts.csv).

Distance metrics for pairs of transcript models (TranD 1GTF pairwise-mode)

Given any pair of transcript models T_i and T_j in gene g, the proportion of variable exon regions is $P_{ER_{ij}} = (num_{ER_i} + num_{ER_j})/total_{ER_{ij}}$, where num_{ER_i} is the number of exon regions only annotated in T_i, num_{ER_j} is the number of exon regions only annotated in T_j, and $total_{ER_{ij}}$ is the union of unique exon regions annotated in either transcript model T_i or T_j (Figure 1B). The $P_{ER_{ij}}$ metric signifies the relative proportion of exons between the transcript models that vary in inclusion/exclusion and therefore represents the proportion of alternative exons present.

Junction coordinates, defined by the coordinates of the donor and acceptor associated with the junction, and exon regions are tracked between T_i and T_j. The proportion of variable junction coordinates between the transcript models is calculated as $P\_J_{ij} = (num_{J_i} + num_{J_j})/total_{J_{ij}}$, where num_{J_i} is the number of junctions only annotated in T_i, where num_{J_j} is the number of junctions only annotated in T_j, and $total_{J_{ij}}$ is the union of unique junction coordinates annotated in either T_i or T_j (Figure 1B). The P_J_{ij} metric indicates the proportion of junctions that vary and therefore represents the relative amount of junction variability (due to alternative donor/acceptors, alternative exons, or intron retentions IR).

Between a pair of transcript models T_i and T_j, $P_{N_{ij}} = (num_{N_i} + num_{N_j})/total_{N_{ij}}$ is a the proportion of nucleotides whose genomic position varies between the two transcript models, where num_{N_i} is the number of nucleotide co-ordinates annotated as part of T_i only, num_{N_j} is the number of nucleotide co-ordinates annotated as part of T_j only, and $total_{N_{ij}}$ is the number of nucleotide co-ordinates annotated in either T_i or T_j (Figure 1B). The proportion of the co-ordinates that overlap is $1 - P_{N_{ij}}$ (percent annotated nucleotides). The $P_{N_{ij}}$ metric represents the relative number of co-ordinates that vary between the transcript models due to differences in splicing between T_i or T_j.

We classify the distance between the transcript model pair based on the structural relationships between T_j and T_k: FSM, ERS_noIR, ERS_wIR, ERN (Figure 1A). All metrics are output along with binary indicators (pairwise_transcript_distance.csv). Metrics are summarized using upset plots for transcript pairs (transcript_pair_AS_upset_nt_boxplot.png) and genes (genes_AS_upset.png). Plots have automatically generated legends that report the total number of pairs/genes in the plot.

Comparing transcript models between annotations (TranD 2GTF pairwise-mode)

Given two transcriptome annotation files, GTF1 and GTF2, annotated on the same genome coordinates, we ascertain the number of genes in both annotations, the number of genes only in GTF1 and the number of genes only in GTF2. We match based on gene name, and a step may need to be taken to facilitate this comparison. For each gene *g*, that is present in both GTF1 and GTF2 all transcript models *j*, from GTF1 are compared pairwise to transcript models 1 to *k* in GTF2 and $P_{ER_{jk}} P_{ER_{jk}}$, $P_{J_{jk}}$ and $P_{N_{jk}}$ are calculated (output in pairwise_transcript_distance.csv).

We define the minimum distance min_{TD_j} between transcript model *j* and the set of models [1,…,*k*] with the following

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

sequential procedure: 1) select all transcript models with min ($P_{ER_{jk}}$), 2) from the minimum transcript models in step 1, select all transcript models with min ($P_{J_{jk}}$), 3) from the minimum transcript models in step 2, select all models with min ($P_{N_{jk}}$). The pair with the min_{TD_j} also has a classification of the based on the structural relationship of the pair: FSM, ERS_noIR, ERS_wIR, ERN. There may be more than 1 transcript pair with the same minimum value (ties). For example, when all transcripts are similar and differ by only donor/acceptor variation. If there is a tie, an indicator variable (*flag_[d1]_tie* = 1 or *flag_[d2]_tie* = 1, where d1 and d2 are the names of GTF 1 and GTF 2); the first T_k (as defined by the alphabetical order of *transcript_id* values) is listed as min_{TD_j}, and the set of transcript_ids that are ties are indicated *([d1]_distance_ties* or *[d2]_distance_ties*). The minimum distance pair for transcripts in GTF1 is indicated with the variable flag_minimum_distance_transcript_GTF1 = 1; and the minimum distance pair for transcripts in GTF2 is indicated with the variable flag_minimum_distance_transcript_GTF2 = 1. All metrics are present in the pairwise_transcript_distance.csv file, enabling other versions of the minimum distance between two transcripts to be readily calculated. For all genes present in both GTF files, the default is to output distances for all pairs, to output only the minimum distance pairs use the option '-p both'.

An often-overlooked computational challenge in the comparison of isoforms is the existence of non-transitive relationships (where A is most similar to B, B is most similar to C, but A and C may not be similar at all). For transitive relationships, $min_{TD_j} = min_{TD_k}$ and when this condition is met, we classify this relationship as a *reciprocal minimum pair* (RMP, *flag_RMP* = 1). Care must be taken in the interpretation of RMP. If there is only 1 transcript model for a gene in either GTF1 or GTF2, by definition there will be an RMP. If there is a transcript *j* or *k* for which there is no reciprocal minimum pair, the transcript is classified as no reciprocal match (NRM). The binary presenence/absence variables: flag_FSM, flag_ERS_wIR, flag_ERS_noIR, flag_RMP, flag_NRM are also present in the output.

Exon region groups

For any pair of transcripts T_i and T_j, when $P_{ER_{ij}} = 0$ all exon regions overlap (Figure 1A.1–A.4). An exon region group (ERG) is defined when $P_{ER_{ij}} = 0$ (Figure 1C). ERG can be constructed allowing for intron retention (IR) containing models ('–includeIR Y') or not including models with IR ('–includeIR N', default). ERGs are identified by running the utility id_ERG.py on the pairwise_distance.csv file that is output from *TranD*. In addition to a label identifying the sets of transcripts in the same exon region group, a GTF file with a single representative transcript for each exon region group is also generated.

Comparing transcript models (transcript model maps)

When a transcript pair is a full-splice match (FSM), the two transcripts share the same junction chain, and we annotate and output a single representative of the unique junction chain (UJC) in the union GTF. Otherwise, both transcripts are present in the union GTF. The union_UJC_ID.csv contains gene, transcript, and junction identifiers for all transcript_id's (Figure 2A). We then identify the exon region groups using the id_ERG.py script. The resulting output ERG.csv is merged to the union_UJC_ID to form the file union_UJC_transcript_map.csv. The transcript model map is defined as: (i) a GTF file, with one transcript model representing each unique junction chain (UJC) in the union of the two annotations and the accompanying union_UJC_ID.csv; (ii) an exon Region Group (ERG) GTF file, with one transcript model representing each unique ERG, or set of overlapping exons, in the union of the annotations and accompanying ERG.csv that links the individual transcriptIDs to the ERG; (iii) a pairwise_distance.csv file that is the output of TranD 2 GTF for genes shared in GTF1 and GTF2; (iv) for genes that are present in only 1 of the 2 GTF files TranD 1 GTF output.

Comparing transcript models between species

Transcriptomes from closely related species can be compared if the transcript models from one species can be mapped to the genomes coordinates of the other species (e.g. *D. simulans* and *D. melanogaster*). For each set of coordinates, the transcript models from one species are mapped onto the coordinates of the other species. When a pair of transcripts are RMP on both sets of coordinates and that pair is an FSM, or an ERS_noIR with a 'small' number of nucleotide positions different the two transcript models ($\hat{T}_{sp1}$, $\hat{T}_{sp2}$) we consider the transcript models between the two species structurally similar, and we label these as a single transcript model ($\hat{T}$) (Figure 3). The pair is indicated by *flag_T_hat* = 1 (a binary 0/1) in the cross_species.csv (Supplementary File 1). We note that there is no phylogenetic relationship demonstrated here. This is intended identify transcript models that are structurally similar in both species, while allowing for variation due to technical and biological issues in comparing genomes. In addition, we identify transcript models in one species that are not annotated in the other species. The definition of 'small' should depend on the particular biological comparison and the quality of the genomes and annotations being compared. For the *D. melanogaster* and *D. simulans* comparison, we consider the distance between the structure of the two transcript models to be 'small' when there are fewer than 15 nucleotide positions different in the splice junctions. This is the average number of exons per gene (5) multiplied by the length of a codon (3) (Supplementary Table S1). This criterium can be easily modified as the code allows a variable input and the distance file retains the nucleotide distance information.

Running TranD

In all modes of *TranD*, the output includes complexity metrics (EpG, TpG, EpT). Input files are in the GTF file format. A pairwise_distance.csv file contains a line for each pair of transcripts compared, the distance metrics for that particular pair of transcripts and then the classifications of the pair. A series of summary plots that visualize splicing patterns are also output by default.

We have developed utilities that use the pairwise_distance.csv file as input to: subset the file (subset_*TranD*_pairwise_transcript_distance.py), generate plots (plot_*TranD*_from_output_files.py), make the csv files for union annotation (make_union_key.py) and construct the transcript model map (Make_transcript_map.py).

We compare transcript models with the same gene identifier and not within a specific region due to overlap of genes in many organisms' annotations (e.g. (77–79) and reviewed in

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

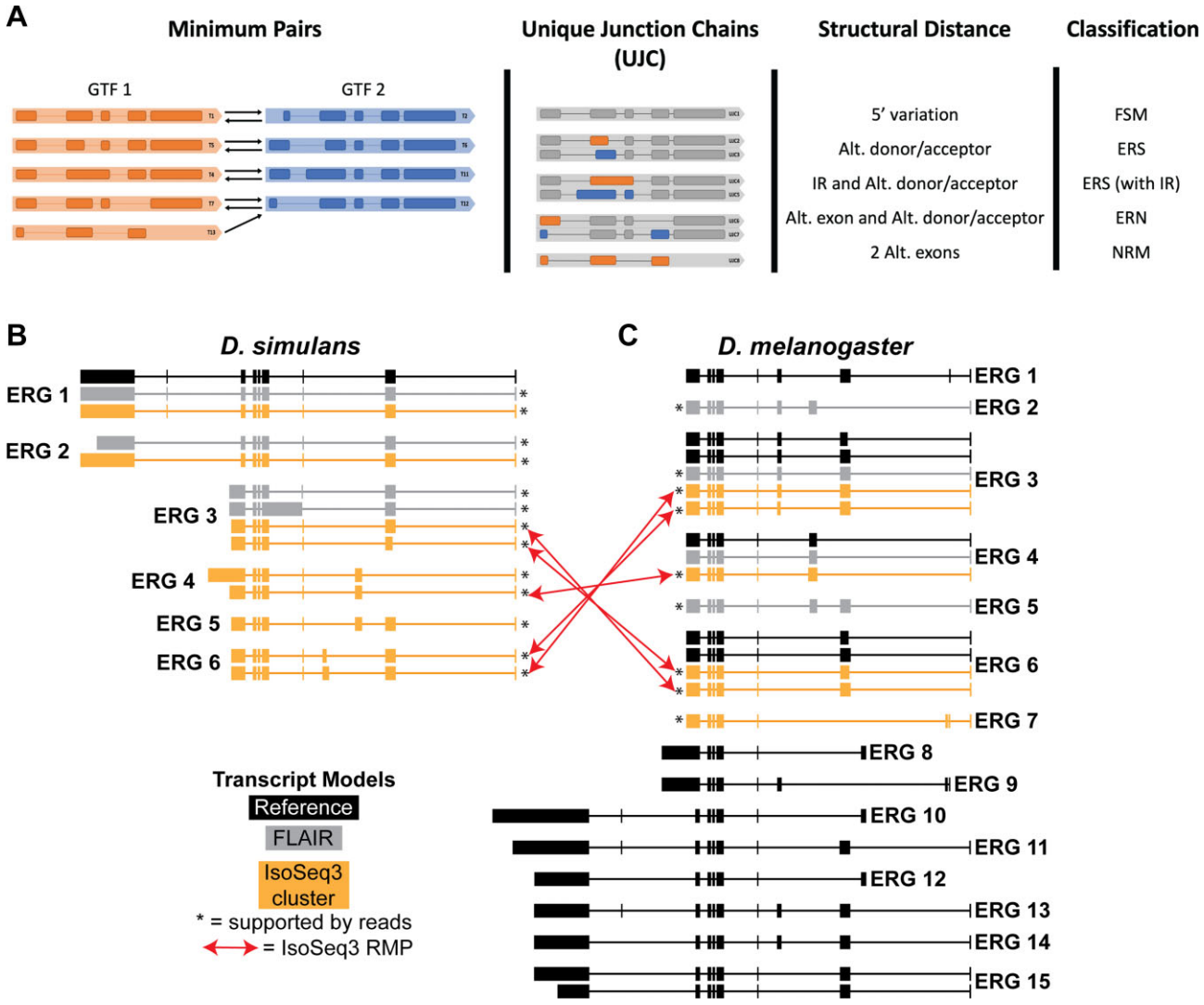


Figure 2. Union GTF and exon region group (ERG) example. (**A**) A union file. When the transcript model in GTF1 is an FSM with the transcript model in GTF2 there is a single unique junction chain (UJC) for both transcript models in the Union GTF and the corresponding csv file. In all other cases the Union GTF file has 2 UJC. (**B**) *D. simulans Sxl* gene (FBgn0016343) transcript models from the reference (black), FLAIR (gray) and IsoSeq3 cluster (yellow) that have been assigned exon region groups (ERGs) based on the distance output of *TranD* 2GTF. (**C**) *D. melanogaster Sxl* (FBgn0264270) FLAIR and IsoSeq3 cluster transcript models assigned ERG. Red arrows show the reciprocal minimum distance between species transcript models. We note not all transcript models are displayed.

(80)). To allow more flexibility and functionality in pairwise mode (1 GTF or 2 GTF), an optional argument (*–subset-pairs*) can be utilized to specify the pairwise comparisons that will be calculated. If one is interested in a specific subset of transcript model pairs, either within one annotation or between two annotations, this argument can be used. For example, to make comparisons based on specific genome regions. In addition, this option can greatly reduce computational resources required and time to process. The input file for this argument is a text file with no header row and each row containing the transcript model identifiers (or transcript_id values in the GTF) for each pair to be assessed for distance calculations. For example if transcript model T1 and T5 are to be compared, the input file would have one row that contains 'T1,T5'. For the two GTF file input the transcript_id from GTF1 is the first column and the transcript_id from GTF2 is the second column. Note that when the subset pair input is used, *TranD* will not calculate minimum distance metrics.

In genes with many transcripts there may be excessively large numbers of pairwise comparisons. In these situations, runtime becomes a concern. It is advisable to check the number of transcripts per gene before running TranD to calculate distances by using the TranD with the *-c/–complexityOnly* option. We also recommend splitting 2GTF comparisons by chromosome.

Long-read data

Head tissue of males ($n = 2$) and females ($n = 2$) from isogenic lines of both *D. melanogaster* (R153) and *D. simulans* (Sz12), for a total of eight independent samples, were collected (by authors SS and SV). Additional details about the experiment and long-read pre-processing can be found in Supplementary Materials Section 2. We also use data publicly available from *C. elegans* (81), *Z. mays* (82) and human cell lines (83). Further details for these data are in Supplementary Table S2.

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

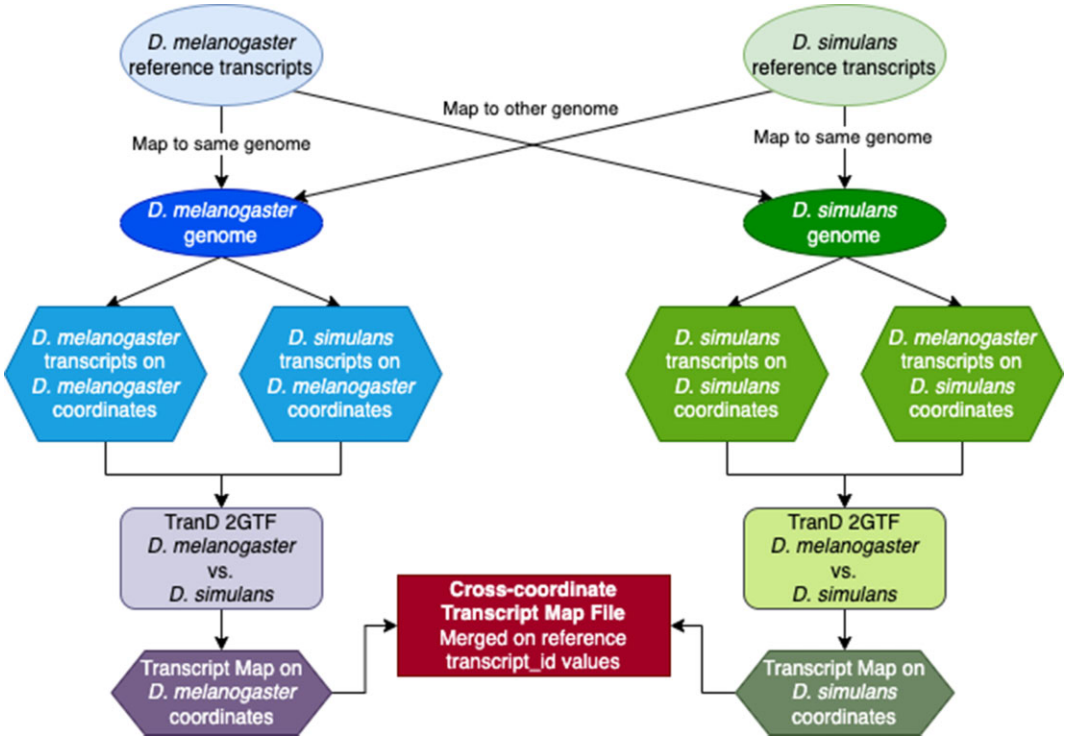


Figure 3. Workflow for generation of cross-species Transcript model Map. We use *D. melanogaster* FlyBase 6.17 and *D. simulans* FlyBase 2.02 annotations to demonstrate the construction of a cross-species Transcript Model Map. Transcript sequences from both species are mapped to both sets of genome coordinates. Transcript models from the mapping of the two species are compared using *TranD* for each set of coordinates. Using the calculated distance metrics and minimum distance associations, a Transcript Map file is generated for each set of coordinates. The cross-species Transcript Model Map file links annotations across coordinates. Note that not all transcripts in either species will have a pair in the other species, and that in addition to identifying annotated pairs, potential missing annotations are also identified.

TranD 1 GTF examples

We illustrate the utility of distance metrics and how to interpret results first with a single gene from *D. melanogaster Sxl* (Figure 2C). There are experimentally validated isoforms with a male-specific exon cassette in *D. simulans* (16,84–87) and, although these transcript models are not present in the *D. simulans* annotation, they are present in the annotation of the sister species *D. melanogaster*. From the IsoSeq3/FLAIR transcript models derived from lrRNA-seq data we observe more splicing patterns than are annotated in *D. simulans* (Figure 2B). We grouped the IsoSeq3/FLAIR transcript models according to their overlapping exons into exon region groups (ERGs). For *D. melanogaster* although there are model transcript models annotated, we still observe a novel set of exons for FLAIR (ERG2) and Isoseq3 (ERG7) (Figure 2C). ERG with a similar structure in both species are linked by red arrows (Figure 2B, C).

We ran *TranD* 1 GTF using GTF input files from *C. elegans* WBcel235 annotation (88,89), two versions of the *Z. mays* annotation v4 B73 and Mo17 (90,91), *H. sapiens* RefSeq GRch38p14, *H. sapiens* Ensembl GRch38.104, *D. melanogaster* FlyBase r6.17 (92–94) and *D. simulans* FlyBase r2.02 (92–94).

TranD 2 GTF examples

We demonstrate how TranD 2 GTF can be used by comparing GTF files from reference annotation to augmented annotation in non-model *D. yakuba* by comparing an augmented annotation based on RNA-seq data (95) to the *D. yakuba* FlyBase r1.05 (here on referred to as dyak-FB105). We used each of these two GTF files as input into *TranD* 2GTF and summarized the output with the plotting utility. Similarly, we compared the RNA-seq data augmented *D. simulans* w^{XD1} (21 562 transcript models) (96) constructed using MAKER to the current FlyBase *D. simulans* r2.02 (26 261 transcript models). We mapped *D. simulans* r2.02 GTF annotation and *D. simulans* w^{XD1} GTF to the *D. simulans* w^{XD1} genome. We used SQANTI3 QC (69) to identify FSM pairs between the annotated positions and the mapped positions. 26065 transcript models from *D. simulans* r2.02 and 19 188 transcript models from *D. simulans* w^{XD1}.

There are many methods for processing and estimating transcript models from empirical data. For simplicity we focus on two approaches to lrRNA-seq but as the input to *TranD* is a GTF file, output from short read assemblers could also be evaluated. Since our goal here is not to debate long-read versus short read assembly, but to highlight the distance-based approach used in *TranD* we focus on two relatively new lrRNA-seq methods to demonstrate how *TranD* 2GTF aids in the analysis of the differences. We compare a reference-based method, FLAIR (v1.5) (97), with a *de novo* approach, IsoSeq3 cluster (https://github.com/PacificBiosciences/IsoSeq). For each of the five species (Supplementary Table S2) we implemented the FLAIR and IsoSeq3 cluster protocols from the same set of starting reads and compared the resulting transcript models (Supplementary Materials Section 6, https://github.com/McIntyre-Lab/TranD/wiki/Long-read-Method-Comparison-(Drosophila-PacBio)).

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

We use *TranD* 2GTF output to compare the human annotations for hg38 Ensembl (release 104) and RefSeq (p13). For transcripts that share all their junctions (FSM) we include only a single representative unique junction chain (UJC) in the GTF file for comparison using the utility ID_UJC.py (a reduced reference). This step avoids ties in minimum distances due to exact matches at the junctions within a GTF. We use the reduced reference GTF files from each of these annotations as input to *TranD* 2GTF. We use the pairwise distance output of *TranD* 2GTF to identify exon region groups (ERG) for genes shared in both annotations, and separately for the genes found only in one of the two GTF files using the utility ERG_id.py. Full details including the exact options used are on in the Supplementary Materials and all scripts are provided (https://github.com/McIntyre-Lab/TranD/docs/).

Comparing *D. melanogaster* and *D. simulans*

We linked the *D. melanogaster* and *D. simulans* results across species by comparing the minimum distance pairs on both sets of co-ordinates. We only linked transcript models when the results were concordant.

Results

Distance metrics can be applied in a wide variety of scenarios. We illustrate some of the anticipated common applications of *TranD*. We highlight how scientists who either seek to improve annotations or wish to quantify transcript level differential expression can gain insights using the formal distance-based approach developed in *TranD*. We also provide on our wiki page the code for all of the examples we use for illustration (https://github.com/McIntyre-Lab/TranD/wiki).

Quantifying splicing patterns (1GTF)

TranD can be used to explore and compare the complexity of transcriptomes without the need of matching transcript IDs in the 1GTF analysis mode. We compare transcript models from *C. elegans* (88) (WBcel235 annotation, Supplementary Figure S1A), *D. melanogaster* (FB6.17), *D. simulans* (FB 2.02), *H. sapiens* Hg38 Refseq and Ensembl and *Z. mays* B73 v4 (90) (Mo17 YAN annotation, Supplementary Figure S1B) and quantify the splicing patterns using the distance metrics in *TranD* 1GTF pairwise. *Z. mays* has greater structural complexity in the transcriptome annotation than *C. elegans* with on average a greater number of transcripts per gene (TpG: ∼2.98 versus ∼1.36 in *C. elegans*), more exons per gene (EpG: ∼8.80 versus ∼4.29 in *C. elegans*), and exons per transcript (EpT: ∼8.75 versus ∼4.45 in *C. elegans*). For genes with multiple transcripts, we can immediately identify differences in splicing patterns between these species using the *TranD* 1GTF output summaries (Figure 4). *Z. mays* has more annotated intron retentions than *C. elegans* consistent with previous literature describing IR as the most prevalent class of alternative splicing in plants (reviewed in (39,98). While in *C. elegans*, the use of alternative exon cassettes is more prevalent than in *Z. mays*. Of note, in both species, alternative exons occur more frequently together with alternative donors/acceptors than either event individaully.

We use *TranD* 1GTF to quantify the splicing patterns for human hg38 Ensembl (release 104) and RefSeq (p13) annotations (Supplementary Figure S6). There were 162 865 RefSeq transcript models and 234 201 Ensembl transcript models. Additionally, we can see that Ensembl contains more transcripts that differ only at the 5′/3′ end, and has both more alternative donors/acceptors, and more alternative exons than RefSeq (Figure 4, Supplementary Figure S6). The Ensembl annotation contains an average TpG of 3.86 ± 6.84, EpT of 6.34 ± 6.95, and EpG of 10.71 ± 17.61. The RefSeq annotation has similar TpG (4.28 ± 6.90), and EpG (10.28 ± 11.85) but larger EpT (11.78 ± 11.25) (Figure 4) and Ensembl has a higher proportion of non-overlapping transcripts (∼16%) compared to RefSeq (∼3%).

When transcripts differ by only donor/acceptors all exonic regions overlap (ERS). Interestingly, donor/acceptor variation is more likely to occur in conjunction with alternative exons than alone and the combination is more frequent than expected by chance in all of the species we examined (Figure 4). This underscores the importance of whole transcript evaluation in understanding the impact of alternative splicing on the complexity of the transcriptomes.

Data driven annotation augmentation (2GTF)

It is common practice to support and extend annotations with empirical data from RNA-seq short reads using tools such as MAKER (57), BRAKER (58,59) and AUGUSTUS (60,61). The default output of these tools does not pinpoint changes in transcript models made as a result of the data inputs. It can be difficult to know when transcript models have been modified by 'small' versus 'large' amounts and how the experimental data improves the quality of the annotation.

TranD 2GTF mode can be used to pinpoint differences in transcript models between the starting and the augmented reference. For example, in the non-model species *D. yakuba,* an augmented annotation was published that leveraged RNA-seq data (95). This annotation (here on referred to as dyak-RR-revised) was compared to the *D. yakuba* FlyBase r1.05 (here on referred to as dyak-FB105). The dyak-RR-revised has fewer transcripts per gene (TpG) but more exons per gene (EpG) than dyak-FB105 (Supplementary Figure S3). Although genes exclusive to the dyak-FB105 (7907 gene loci) and dyak-RR-revised (5168 gene loci) are easily identified by a variety of tools (Supplementary Figure S4A), *TranD* provides this information in addition to metrics designed to quantitatively compare transcript models in the 8162 genes present in both annotations. There were 5241 transcript models with new exons or new exon combinations in dyak-RR-revised compared to the closest transcript model in dyak-FB105. The vast majority of these (4464 transcript models) also differ by an alternative donor/acceptor (Supplementary Figure S4D). The *TranD* output pinpoints the structural differences with nucleotide-level resolution between each of the dyak-RR-revised transcript models and its closest dyak-FB105 transcript (Supplementary Figure S4D, Supplementary File 2). This enables the scientist to pinpoint not only the new exons, but the changes in the donor/acceptors that accompany that exon in the transcript model.

The same general trends were observed when we compared the RNA-seq data augmented *D. simulans* w^{XD1} genome and annotation (21 562 transcript models) (96) to the FlyBase *D. simulans* r2.02 annotation (26261 transcript models) using *TranD* 2GTF (Supplementary Figure S5). There were 3103 transcript models with large (hundreds of nucleotides) differences compared to the r2.02 model. As with the *D. yakuba* data augmentation, the *D. simulans* w^{XD1} MAKER aug-

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

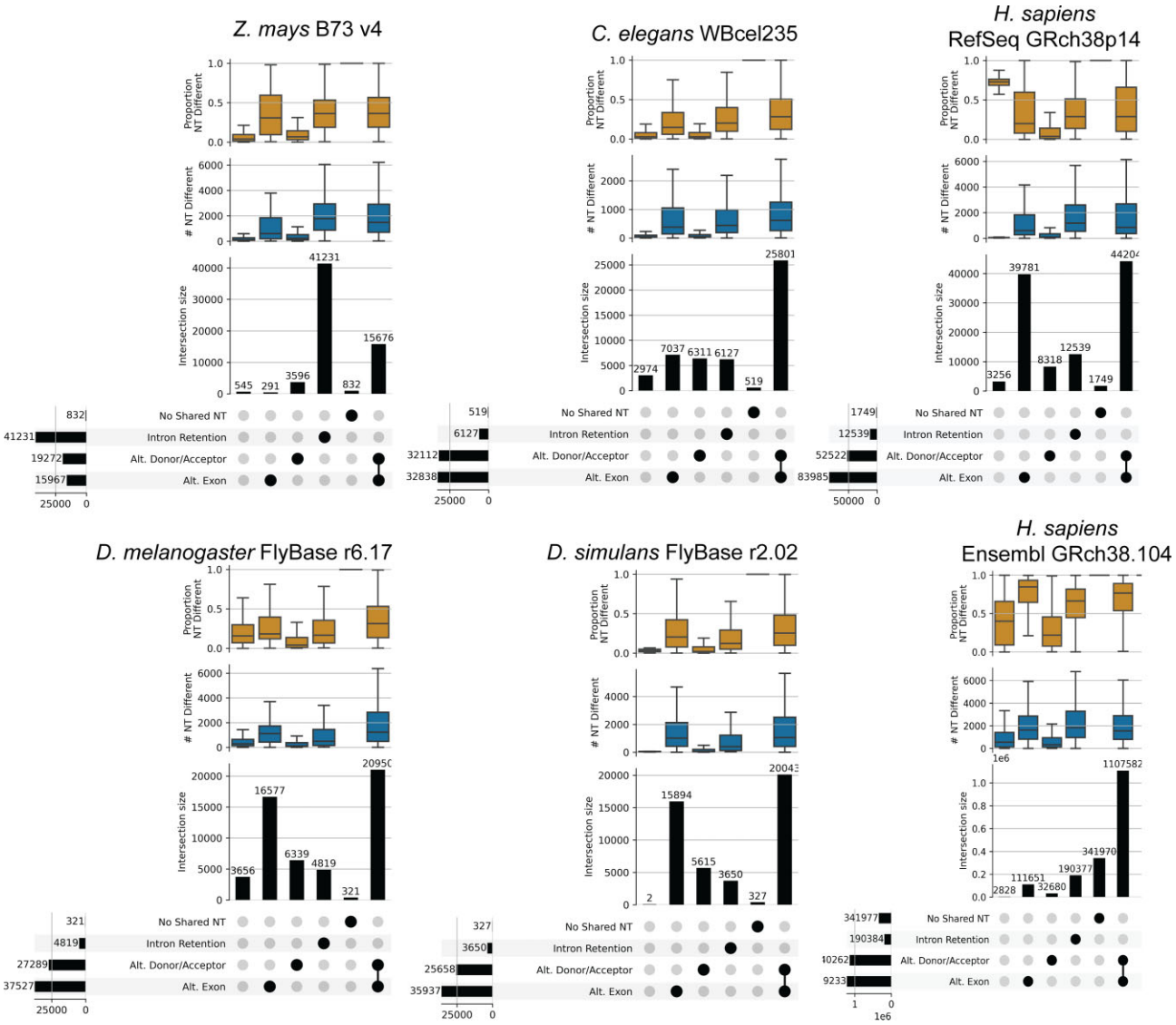


Figure 4. Splicing patterns in reference annotations of five species. *TranD* 1 GTF pairwise was run on the following annotations: H. sapiens RefSeq GRch38p14, H. sapiens Ensembl GRch38.104, *Z. mays* B73 v4, *C. elegans* WBcel325, *D. melanogaster* FlyBase r6.17 and *D. simulans* FlyBase r2.02. The pairwise_distance.csv output file from each genome was used *ignore_AS_type.py*, with the options '-i3 -i5'.

mented transcript annotation makes significant additions to the structure of the transcript models altering donor/acceptor sites in combination with alternative exon structure. TranD distance metrics pinpoint all of the changes in the structure of the transcript and enable a transcriptome level summary of the impact of data augmentation on the reference annotation (Supplementary File 3).

Comparison of transcript model estimation methods

With the advent of long reads, there are a number of tools for estimating transcript models. To illustrate how TranD can be used to provide a quantitative assessment of the differences in the transcript models estimated between different tools, we compare FLAIR transcript models (GTF1) to IsoSeq3 cluster transcript models (GTF2) using *TranD* 2GTF. We use 5 independent sets of data from 5 different species representing a range of complexity in the reference transcriptomes: (i) PacBio long read cDNA from a human WTC11 cell line from LRGASP (83), (ii) *Z. mays* B73 root tissue (82), (iii) ONT from *C. elegans* L1 larval stage (81), (iv) PacBio IsoSeq data from *D. melanogaster* head tissue and (v) PacBio IsoSeq data from *D. simulans* head tissue. Each of these datasets differ in computational complexity and accuracy with respect to different types of isoform variation.

The LRGASP project reported large difference between bioinformatic methods when inferring transcript models from long-read data (83) and focused many evaluations on differences from matching a known, spliced spike-in control or performance using simulated long-read data. LRGASP identified overlap among transcript models but did not identify similar models between approaches, limiting our understanding of the differences between tools in complex scenarios when there is no single objective truth. TranD can be used for this purpose.

We compared FLAIR and IsoSeq3 for 5 different datasets. For each dataset we compared transcript models using *TranD* 2 GTF. We see some general trends for all the datasets analyzed (Figure 5A, B). Isoseq3 (Orange) estimates more transcript models than FLAIR (Gray) (Figure 5A). When transcript

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

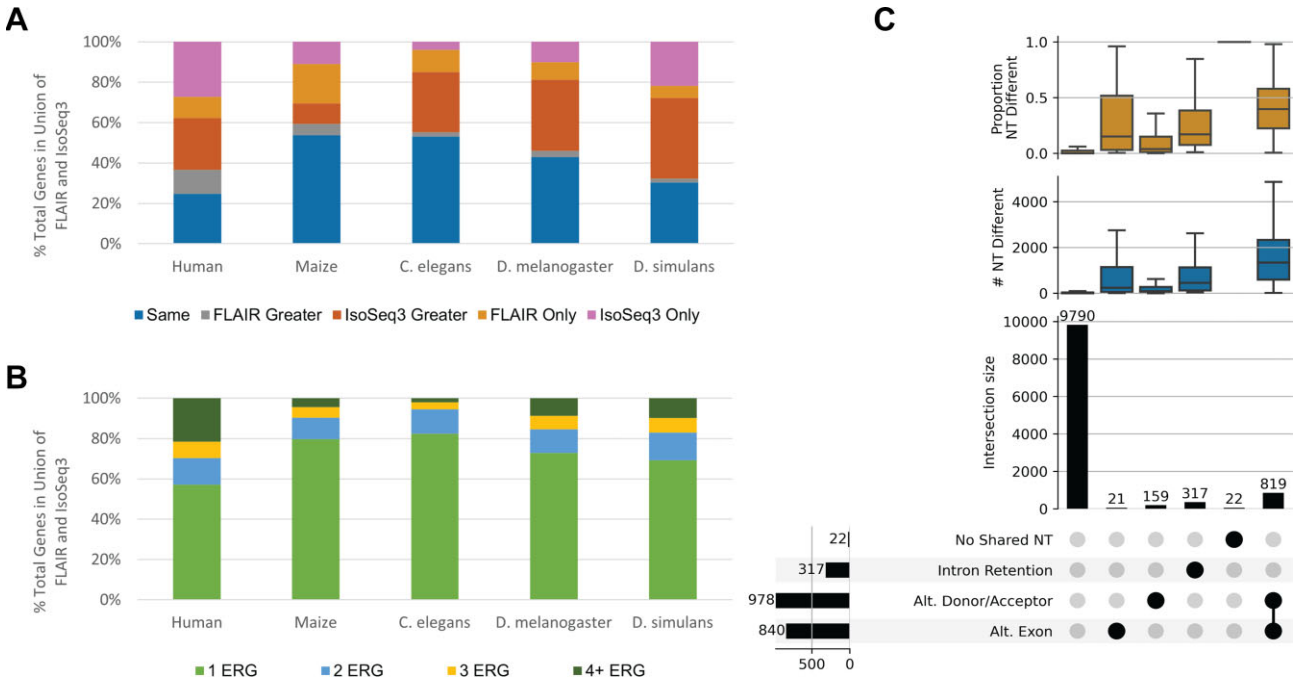


Figure 5. Comparison of transcript model estimation with FLAIR or IsoSeq3 cluster. (**A**) Percent of genes identified in each dataset by either FLAIR or IsoSeq3 cluster, colored by the relationship of the number of estimated transcript models identified in each method. Genes with the same number of models are blue, genes with more models in FLAIR compared to IsoSeq3 are gray, genes with the reverse are orange. (**B**) Percent of genes with 1, 2, 3 or 4+ exon region groups (ERG) present across all IsoSeq3 cluster and FLAIR estimated transcript models. (**C**) Reciprocal minimum transcript model pairs ($n = 11\ 128$) of FLAIR vs. IsoSeq3 cluster in the *D. simulans* dataset are plotted results including 5′/3′ variation is in Supplementary Figure S2A.

models are estimated in FLAIR and Isoseq3, both approaches identify the same set of exonic regions, a *single* exon region group (ERG), for a majority of the genes (Figure 5B). The majority of transcript models also agree on the splice junctions. For example, in *D. simulans*, ∼87% of the transcript models are FSM (9790/11218, Figure 5C) but only ∼26% are identical (matching also both 5′ and 3′ ends) between the two methods (Supplementary Figure S2A). Variation at the 5′/3′ ends of the transcript is common in all the datasets examined and underscores the difficulties in algorithmically identifying the 5′/3′ ends from lrRNA-seq data.

For *D. simulans*, there were transcript models for 670 genes in FLAIR only and 2475 genes in IsoSeq3 cluster only. There were 8206 genes with transcript models in both methods but only ∼42% (3447) had the same number of transcript models. For these 3447 genes, 3127 had only one transcript model. Interestingly, when the number of transcript models is greater in one of the methods, the transcript models from the method with fewer transcript models are often an RMP subset of the other approach. These patterns are similar for the other 4 species (Supplementary Table S3). We note that the RMP that are not FSM are likely to differ both alternative exon and alternative donor/acceptor (*D. simulans* Figure 5C, Supplementary Figure S2). This pattern is consistent for all the datasets examined (Supplementary Materials Section 6.2).

We determined how many of the transcript models from FLAIR and IsoSeq3 cluster were supported by the long-read data that were used as input (Supplementary Table S3). Most of the transcript models were supported by the data, including all transcript models identified by FLAIR that were not found in IsoSeq3. However, some IsoSeq3 transcript models not identified by FLAIR lacked support from the original reads (Supplementary Table S3, Supplementary Materials Section 6). The concordance between FLAIR and Isoseq3 was much lower for the WTC11 PacBio human data compared to the other four species, suggesting that additional caution is needed in estimating transcript models as transcriptome complexity increases.

Comparison of alternative annotations (transcript models) for the same genome

The choice of a particular annotation affects downstream analyses and impacts biological interpretations (99,100). We used *TranD* 2 GTF to compare the transcript models in RefSeq and Ensemble for *H. sapiens* hg38 genome (Supplementary Figure S7). Differences between RefSeq and Ensembl are a known issue, and as a result a consortium to resolve these annotations has been working for several years resulting in a set of ∼19 000 manually curated transcripts (101). We verified that the TranD 2GTF output from the comparison of RefSeq to Ensembl contained 19 316 of the MANE transcripts (a handful of transcripts were missing from the version of Ensembl we used). The similarity between the two annotations is higher when only protein coding genes are considered. There are 19 417 protein coding genes in both annotations. For all of these genes there is a RMP for at least 1 transcript. For protein coding genes there were 52 567 RMPs with 46 712 FSM. Overall, there were 55 538 FSM transcript pairs. We note that the MANE consortium has developed criteria for linking transcripts, and are not suggesting that exact junction matches supplant the MANE criteria.

The majority of genes are different between the RefSeq and Ensembl annotations (35 641/66 921 = 53% of genes, Table 2). Genes unique to one annotation had fewer transcript models than those annotated by both approaches. Transcript

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

Table 2. RefSeq GRch38p14, compared to *H. sapiens* Ensembl GRch38.104 overall (protein coding gene)

	Genes
RefSeq	6316 (343)
Ensembl	29 325 (611)
both	31 281 (19 417)

models in Ensembl only had 1.12 TpG(transcripts per gene) and while those in RefSeq only had 1.6 TpG. In contrast there were an average of 9.1 TpG in genes that are annotated in both references.

We hypothesized that lrRNA-seq data could provide insight into whether one of these two annotations had better empirical support. We used publicly available lrRNA-seq data from PacBio and ONT from human WTC11 cells to evaluate the long-read support for the union of the RefSeq and Ensembl annotations (Figure 6). We mapped the data to hg38 and used SQANTI3 (69) to identify FSM and incomplete splice matches (ISM; reads match a continuous subset of junctions) (69) between reads and annotated UJC for the PacBio data (Figure 6A) and the ONT data (Figure 6C). The two technologies provide consistent support of the same transcript models even with more ONT reads overall. In expressed genes, 70% of the transcript models from both RefSeq and Ensembl have read support for both PacBio (Figure 6B) and ONT (Figure 6D). RefSeq transcript models had higher levels of read support compared to Ensembl but the number of transcript models supported is similar.

Comparative genomics leveraging annotation across species

We note that our linking of transcript models between species (Figure 3) does not depend on the annotation of the genes as an ortholog but on the transcript models being both structurally similar and reciprocally mapping to the same positions. For example, we find a single transcript model $\hat{T}$ between *D. melanogaster Ubx-RE* and *D. simulans Ubx-RB*. While in Fly Base OrthoDB release FB2022_01 (102), Dmel/Ubx is identified as an ortholog to Dsim/Ubx, we caution that $\hat{T}$cannot be interpreted as an orthologous transcript without the additional work necessary to confirm the evolutionary history.

In the 10 542 genes with a single annotated transcript model in both *D. melanogaster* and *D. simulans*, most meet the definition for a single transcript model $\hat{T}$ (Supplementary Figure S8A). For genes with 2 transcripts in each species (1375), ∼81% have a single $\hat{T}$ for both transcripts (Supplementary Figure S8B). When the number of transcript models is the same between the species for larger numbers of transcripts, ∼62% have a full complement of $\hat{T}$ (Supplementary Figure S8C) and when the number of transcripts is larger in one species, most genes present with the transcripts in the species with the smaller number of transcripts are a subset of $\hat{T}$ to the transcript models in the other species (Supplementary Figure S8D,E) suggesting that the 'extra' transcript models may be potentially missing annotation. Overall, we identify 14800 structurally similar $\hat{T}$. Of these 14 669 are in 12 863 one-one gene pairs. In head tissue we find empirical support for the majority of these transcript models in lrRNA-seq data for both species (Figure 7A). Since this is a single tissue, we do not expect to find support for all genes and all models. We also find evidence in lrRNA-seq data in *D. melanogaster* for *D. simulans* annotated transcripts and *vice versa* (Figure 7B, C) indicating that both species may be under-annotated.

Discussion

We develop a set of distance metrics that identify and quantitate the relative structure of transcript model pairs. We propose a set of sequential rules that prioritize minimizing the number of overlapping exons, then the number of shared junctions, and finally the number of overlapping nucleotides to identify minimum distance pairs. In addition to distance metrics, we also develop complexity metrics for transcriptomes based on the number of annotated genes, transcripts, and exons (EpT, TpG, EpG). We developed *TranD* to calculate distance and complexity metrics for a single annotation (1GTF) and between two annotations (2GTF). *TranD* output can be used to group transcript models, by structural components; describe alternative splicing patterns within and between transcriptomes; and pinpoint variation in transcripts with nucleotide-level precision- for an entire transcriptome. This enables the scientist to quickly identify sets of transcript models that are structurally similar and compare transcript structure across the tree of life (30). We have provided examples for 5 species (https://github.com/McIntyre-Lab/TranD/wiki/Precomputed-Files).

The choice of annotation impacts the quantification of gene expression for the human genome (100). There has been an effort to unify annotations (e.g. 101) with a single representative transcript per gene, however an extensive process was necessary, ending with manual curation that selects a high quality intersection of the two annotations. Using distance metrics we are able to reproducibly and automatically: (i) identify the transcripts that share all junctions in both annotations and seamlessly link the individual RefSeq and Ensembl identifiers to the same set of junctions, (ii) make a combined/union reference that retains all individual transcript models from both annotations, (iii) identify all transcript models in both annotations that contain overlapping exon regions (shared transcript structure), (iv) generate a representative transcript model for sets of overlapping transcripts and (v) pinpoint structural differences among all of the transcript models at nucleotide resolution. We demonstrate how to use the a transcript model map between two annotations to map lrRNA-seq data from WTC11 lrRNA-seq (83). Transcript models were inadequately described by RefSeq / Ensembl annotations individually, but the union annotation adequately captures the variability in the data, with 75% of the PacBio reads and 90% of the ONT reads associated with an annotated splicing pattern. (https://github.com/McIntyre-Lab/TranD/wiki). We suggest that investigators using the union reference (https://data.rc.ufl.edu/pub/mcintyre/trand/tmm/human_tmm/) as a way of capturing the combined expertise of RefSeq and Ensembl.

TranD distance metrics can be used to compare any two annotations for a single genome. For example to compare different methods, FLAIR and IsoSeq3, for estimating transcript models from long-read data on 5 different datasets. For *H. sapiens* data we evaluate both methods for two technologies, PacBio and ONT. The technologies resulted in very similar splicing patterns for each of the two methods. Overall FLAIR and IsoSeq3 produce similar transcript models when genes have one predominant splicing pattern per gene. However, for

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

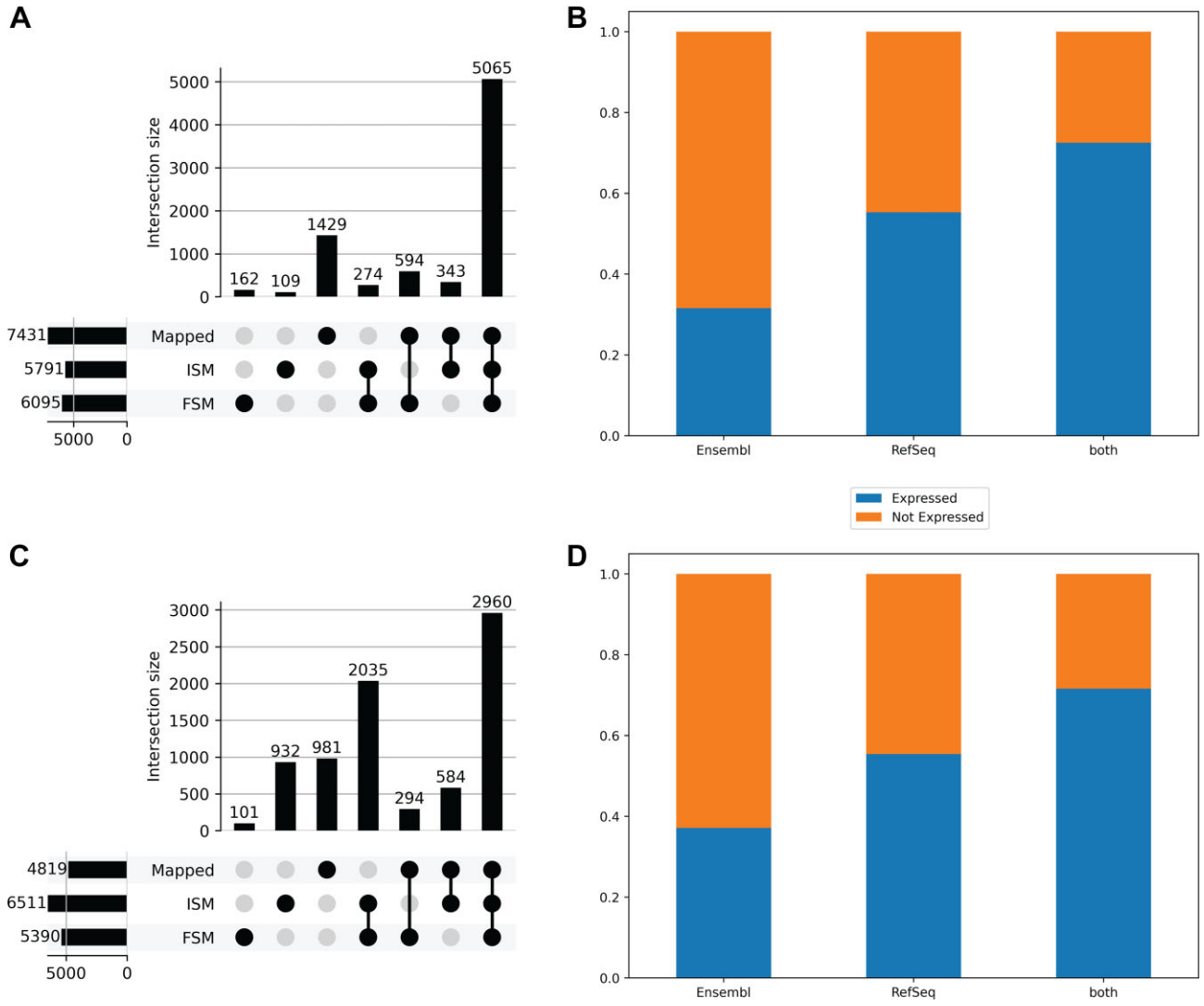


Figure 6. Long-read support for hg38 RefSeq and Ensembl union annotation. (**A**) Number of genes annotated in both RefSeq and Ensembl with long-read evidence in WTC11 PacBio data that are FSM, ISM or mapped to an annotated transcript model. (**B**) Exon region groups (ERG) in expressed genes with transcript models in both annotations are separated by ERG exclusive to Ensembl, RefSeq or in both annotations that are expressed (blue) or not expressed in the WTC11 PacBio data. (**C**) Number of genes annotated in both RefSeq and Ensembl with long-read evidence in WTC11 ONT data that are FSM, ISM, or mapped to an annotated transcript model. (**D**) Exon region groups (ERG) in expressed genes with transcript models in both annotations are separated by ERG exclusive to Ensembl, RefSeq, or in both annotations that are expressed (blue) or not expressed in the WTC11 ONT data. For panels B and D the number of ERG represented is indicated within each bar and the total ERG (*n*) below each.

the *H. sapiens* data, where there is empirical evidence for multiple transcript models per gene, the two methods diverged. FLAIR transcript models were frequently, but not always, a subset of the Isoseq3 transcript models. While Isoseq3 generally produced more transcript models, not all of these models were validated by individual reads. We recommend that users of any transcript model estimation tool check whether there is empirical support for each reported transcript model. In addition we note that empirically observed transcript models not present in the reference annotation can be included in a union annotation using *TranD*.

TranD distance metrics are based on a GTF file, and can be used to compare annotations. For example between an existing reference and a reference improved with short read data. We illustrate this *by* quantifying the data-driven annotation updates for *D. yakuba* and *D. simulans*. We observed that for both of these empirical processes, the inclusion of an alternative exon was more often accompanied by a change in a donor/acceptor than expected due to chance. Intriguingly, in five different species (*C. elegans*, *D. melanogaster*, *D. simulans*, *H. sapiens* and *Z. mays*) an analysis of the reference transcriptomes showed the same pattern: when a pair of transcript models differ by an alternative exon, they are more likely to also differ by a donor/acceptor.

For each of these five species we also examined lrRNA-seq data. We observed in the lrRNA-seq data of all five species that when there are multiple transcript models represented for a gene and when a pair of transcripts differ by an alternative exon, they are more likely to also differ by a donor/acceptor. This empirical observation supports the pattern identified in the reference annotations. Although we do not yet have a mechanistic insight for this observation, we suggest that this and many other observations will be facilitated by deploying the *TranD* tool.

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

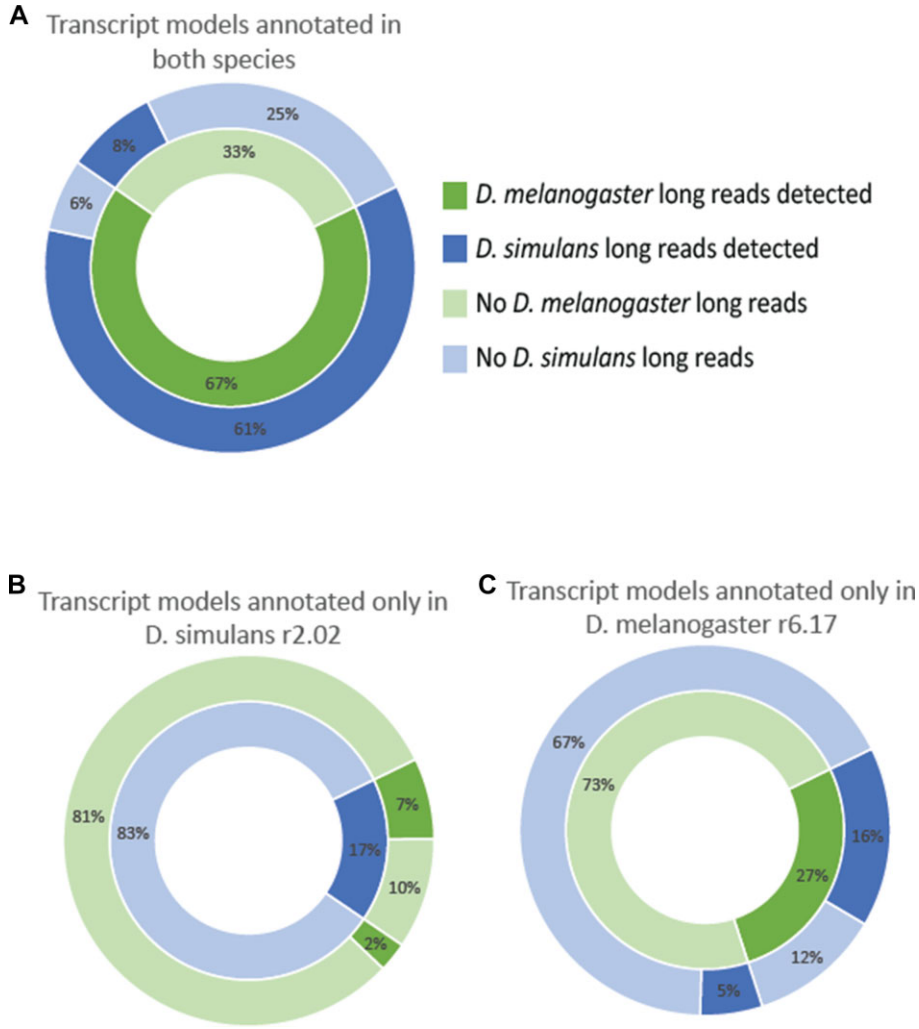


Figure 7. Read support for transcript models. (Panel **A**) *D. melanogaster* and *D. simulans* PacBio long read support for $\hat{T}$. (Panel **B**) *D. melanogaster* and *D. simulans* PacBio long read support for transcript models annotated only in *D. simulans* r2.02. (Panel **C**) *D. melanogaster* and *D. simulans* PacBio long read support for transcript models annotated only in *D. melanogaster* r6.17.

There are many scenarios where *TranD* structural comparisons provide valuable distance and complexity metrics that can enhance biological interpretations, generate hypotheses and advance our understanding of transcriptome evolution. We have previously demonstrated the utility of some of these metrics in phylogenetic comparative studies (56). *TranD* has been designed to be general and flexible to implement to allow for more broad applications of the transcript distance metrics and potentially further studies incorporating these metrics with phylogenetic comparative methods.

Data availability

All code, examples, results and documentation are available in Zenodo at https://doi.org/10.5281/zenodo.10475517. These data are also available in Github (https://github.com/McIntyre-Lab/TranD/wiki). There is a PyPi package and instructions for installation using a conda environment. Drosophila data are deposited to the SRA BioProject PRJNA737411.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

The Department of Molecular Genetics and Microbiology, The University of Florida Genetics Institute, University of Florida Research Computing, HiPerGator, Jeremy R.B. Newman for questions and support. Anna Yang for help with Figure 1. Netanya Keil for helpful discussion. The scientific community for feedback, questions and encouragement at poster sessions and in the hallways at Dros22, Dros23, and PEQG22.

Funding

National Institute of General Medical Sciences [R01GM128193, R01GM1374430, R35 GM133376]; Department of Molecular Genetics and Microbiology, the University of Florida Genetics Institute, University of Florida Research Computing, University of Florida. Funding for open access charge: University of Florida Genetics Institute and Department of Molecular Genetics and Microbiology University of Florida

Conflict of interest statement

None declared.

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

References

1. Lewin,H.A., Robinson,G.E., Kress,W.J., Baker,W.J., Coddington,J., Crandall,K.A., Durbin,R., Edwards,S.V., Forest,F., Gilbert,M.T.P., *et al.* (2018) Earth BioGenome Project: sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 4325–4333.
2. Rhie,A., McCarthy,S.A., Fedrigo,O., Damas,J., Formenti,G., Koren,S., Uliano-Silva,M., Chow,W., Fungtammasan,A., Kim,J., *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, **592**, 737–746.
3. Formenti,G., Theissinger,K., Fernandes,C., Bista,I., Bombarely,A., Bleidorn,C., Ciofi,C., Crottini,A., Godoy,J.A., Höglund,J., *et al.* (2022) The era of reference genomes in conservation genomics. *Trends Ecol. Evol.*, **37**, 197–202.
4. Kovaka,S., Ou,S., Jenike,K.M. and Schatz,M.C. (2023) Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. *Nat. Methods*, **20**, 12–16.
5. Roretz,C. and Gallouzi,I.E. (2008) Decoding ARE-mediated decay: is microRNA part of the equation?*J. Cell Biol.*, **181**, 189–194.
6. Gilbert,W. (1978) Why genes in pieces?*Nature*, **271**, 501–501.
7. Gilbert,W., de Souza,S.J. and Long,M. (1997) Origin of genes. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 7698–7703.
8. Vibranovski,M.D., Sakabe,N.J., de Oliveira,R.S. and de Souza,S.J. (2005) Signs of ancient and modern exon-shuffling are correlated to the distribution of ancient and modern domains along proteins. *J. Mol. Evol.*, **61**, 341–350.
9. Frankish,A., Mudge,J.M., Thomas,M. and Harrow,J. (2012) The importance of identifying alternative splicing in vertebrate genome annotation. *Database.*, **2012**, bas014.
10. Hirai,M.Y., Yano,M., Goodenowe,D.B., Kanaya,S., Kimura,T., Awazuhara,M., Arita,M., Fujiwara,T. and Saito,K. (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc. Nat. Acad. Sci. U.S.A.*, **101**, 10205–10210.
11. McGuire,A.M., Pearson,M.D., Neafsey,D.E. and Galagan,J.E. (2008) Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol.*, **9**, R50.
12. Lu,T.T., Lu,G.J., Fan,D.L., Zhu,C.R., Li,W., Zhao,Q.A., Feng,Q., Zhao,Y., Guo,Y.L., Li,W.J., *et al.* (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.*, **20**, 1238–1249.
13. Mudge,J.M., Frankish,A., Fernandez-Banet,J., Alioto,T., Derrien,T., Howald,C., Reymond,A., Guigo,R., Hubbard,T. and Harrow,J. (2011) The origins, evolution, and functional potential of alternative splicing in vertebrates. *Mol. Biol. Evol.*, **28**, 2949–2959.
14. Chen,L., Bush,S.J., Tovar-Corona,J.M., Castillo-Morales,A. and Urrutia,A.O. (2014) Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. *Mol. Biol. Evol.*, **31**, 1402–1413.
15. Akam,M.E. and Martinez-Arias,A. (1985) The distribution of ultrabithorax transcripts in Drosophila embryos. *EMBO J.*, **4**, 1689–1700.
16. Bell,L.R., Maine,E.M., Schedl,P. and Cline,T.W. (1988) Sex-lethal, a Drosophila sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins. *Cell*, **55**, 1037–1046.
17. Bermingham,J.R. and Scott,M.P. (1988) Developmentally regulated alternative splicing of transcripts from the Drosophila homeotic gene antennapedia can produce four different proteins. *EMBO J.*, **7**, 3211–3222.
18. O'Connor,M.B., Binari,R., Perkins,L.A. and Bender,W. (1988) Alternative RNA products from the ultrabithorax domain of the bithorax complex. *EMBO J.*, **7**, 435–445.
19. Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.
20. Celniker,S.E., Dillon,L.A.L., Gerstein,M.B., Gunsalus,K.C., Henikoff,S., Karpen,G.H., Kellis,M., Lai,E.C., Lieb,J.D., MacAlpine,D.M., *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.
21. Levin,M., Hashimshony,T., Wagner,F. and Yanai,I. (2012) Developmental milestones punctuate gene expression in the Caenorhabditis embryo. *Dev. Cell*, **22**, 1101–1108.
22. Klepikova,A.V., Kasianov,A.S., Gerasimov,E.S., Logacheva,M.D. and Penin,A.A. (2016) A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J.*, **88**, 1058–1070.
23. Newman,J.R.B., Conesa,A., Mika,M., New,F.N., Onengut-Gumuscu,S., Atkinson,M.A., Rich,S.S., McIntyre,L.M. and Concannon,P. (2017) Disease-specific biases in alternative splicing and tissue-specific dysregulation revealed by multitissue profiling of lymphocyte gene expression in type 1 diabetes. *Genome Res.*, **27**, 1807–1815.
24. Xiong,J.Y., Jiang,X., Ditsiou,A., Gao,Y., Sun,J., Lowenstein,E.D., Huang,S.Y. and Khaitovich,P. (2018) Predominant patterns of splicing evolution on human, chimpanzee and macaque evolutionary lineages. *Hum. Mol. Genet.*, **27**, 1474–1485.
25. Aguet,F., Barbeira,A.N., Bonazzola,R., Brown,A., Castel,S.E., Jo,B., Kasela,S., Kim-Hellmuth,S., Liang,Y.Y., Parsana,P., *et al.* (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.
26. Ner-Gaon,H., Leviatan,N., Rubin,E. and Fluhr,R. (2007) Comparative cross-species alternative splicing in plants. *Plant Physiol.*, **144**, 1632–1641.
27. Barbosa-Morais,N.L., Irimia,M., Pan,Q., Xiong,H.Y., Gueroussov,S., Lee,L.J., Slobodeniuc,V., Kutter,C., Watt,S., Colak,R., *et al.* (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**, 1587–1593.
28. Gibilisco,L., Zhou,Q., Mahajan,S. and Bachtrog,D. (2016) Alternative splicing within and between drosophila species, sexes, tissues, and developmental stages. *PLoS Genet.*, **12**, e1006464.
29. Singh,P., Börger,C., More,H. and Sturmbauer,C. (2017) The role of alternative splicing and differential gene expression in Cichlid adaptive radiation. *Genome Biol. Evol.*, **9**, 2764–2781.
30. Titus-McQuillan,J.E., Nanni,A.V., McIntyre,L.M. and Rogers,R.L. (2023) Estimating transcriptome complexities across eukaryotes. *Bmc Genomics (Electronic Resource)*, **24**, 254.
31. Tolstrup,N., Rouze,P. and Brunak,S. (1997) A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites. *Nucleic Acids Res.*, **25**, 3159–3163.
32. Lorkovic,Z.J. and Barta,A. (2002) Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA-binding proteins from the flowering plant Arabidopsis thaliana. *Nucleic Acids Res.*, **30**, 623–635.
33. Zhu,W., Schlueter,S.D. and Brendel,V. (2003) Refined annotation of the Arabidopsis genome by complete expressed sequence tag mapping. *Plant Physiol.*, **132**, 469–484.
34. Kreivi,J.P. and Lamond,A.I. (1996) RNA splicing: unexpected spliceosome diversity. *Curr. Biol.*, **6**, 802–805.
35. Collins,L. and Penny,D. (2005) Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.*, **22**, 1053–1066.
36. Jangi,M. and Sharp,P.A. (2014) Building robust transcriptomes with master splicing factors. *Cell*, **159**, 487–498.
37. McManus,C.J., Coolon,J.D., Eipper-Mains,J., Wittkopp,P.J. and Graveley,B.R. (2014) Evolution of splicing regulatory networks in Drosophila. *Genome Res.*, **24**, 786–796.
38. Reddy,A.S. (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu. Rev. Plant Biol.*, **58**, 267–294.
39. Barbazuk,W.B., Fu,Y. and McGinnis,K.M. (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res.*, **18**, 1381–1392.
40. Zhiguo,E., Wang,L. and Zhou,J. (2013) Splicing and alternative splicing in rice and humans. *BMB Rep*, **46**, 439–447.

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

41. Martin,G., Marquez,Y., Mantica,F., Duque,P. and Irimia,M. (2021) Alternative splicing landscapes in Arabidopsis thaliana across tissues and stress conditions highlight major functional differences with animals. *Genome Biol.*, **22**, 35.
42. Singh,P. and Ahi,E.P. (2022) The importance of alternative splicing in adaptive evolution. *Mol. Ecol.*, **31**, 1928–1938.
43. Xing,Y. and Lee,C. (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 13526–13531.
44. Xing,Y. and Lee,C. (2006) Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.*, **7**, 499–509.
45. Jakšić,A.M. and Schlötterer,C. (2016) The interplay of temperature and genotype on patterns of alternative splicing in Drosophila melanogaster. *Genetics*, **204**, 315–325.
46. Tovar-Corona,J.M., Castillo-Morales,A., Chen,L., Olds,B.P., Clark,J.M., Reynolds,S.E., Pittendrigh,B.R., Feil,E.J. and Urrutia,A.O. (2015) Alternative splice in Alternative lice. *Mol. Biol. Evol.*, **32**, 2749–2759.
47. Smith,C.C.R., Tittes,S., Mendieta,J.P., Collier-Zans,E., Rowe,H.C., Rieseberg,L.H. and Kane,N.C. (2018) Genetics of alternative splicing evolution during sunflower domestication. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 6768–6773.
48. Telonis-Scott,M., Kopp,A., Wayne,M.L., Nuzhdin,S.V. and McIntyre,L.M. (2009) Sex-specific splicing in Drosophila: widespread occurrence, tissue specificity and evolutionary conservation. *Genetics*, **181**, 421–434.
49. Ray,M., Conard,A.M., Urban,J., Mahableshwarkar,P., Aguilera,J., Huang,A., Vaidyanathan,S. and Larschan,E. (2023) Sex-specific splicing occurs genome-wide during early. *eLife*, **12**, e87865.
50. Singh,A. and Agrawal,A.F. (2023) Two forms of sexual dimorphism in gene expression in drosophila melanogaster: their coincidence and evolutionary genetics. *Mol. Biol. Evol.*, **40**, msad091.
51. Nanni,A.V., Martinez,N., Graze,R., Morse,A., Newman,J.R.B., Jain,V., Vlaho,S., Signor,S., Nuzhdin,S.V., Renne,R., *et al.* (2023) Sex-biased expression is associated with chromatin state in Drosophila melanogaster and Drosophila simulans. *Mol. Biol. Evol.*, **40**, msad078.
52. Salz,H.K. and Erickson,J.W. (2010) Sex determination in Drosophila: the view from the top. *Fly (Austin)*, **4**, 60–70.
53. Rogers,T.F., Palmer,D.H. and Wright,A.E. (2021) Sex-specific selection drives the evolution of alternative splicing in birds. *Mol. Biol. Evol.*, **38**, 519–530.
54. Clark,A.G., Eisen,M.B., Smith,D.R., Bergman,C.M., Oliver,B., Markow,T.A., Kaufman,T.C., Kellis,M., Gelbart,W., Iyer,V.N., *et al.* (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.
55. Scott,J.G., Warren,W.C., Beukeboom,L.W., Bopp,D., Clark,A.G., Giers,S.D., Hediger,M., Jones,A.K., Kasai,S., Leichter,C.A., *et al.* (2014) Genome of the house fly, Musca domestica L., a global vector of diseases with adaptations to a septic environment. *Genome Biol.*, **15**, 466.
56. Simao,F.A., Waterhouse,R.M., Ioannidis,P., Kriventseva,E.V. and Zdobnov,E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
57. Cantarel,B.L., Korf,I., Robb,S.M., Parra,G., Ross,E., Moore,B., Holt,C., Sánchez Alvarado,A. and Yandell,M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, **18**, 188–196.
58. Brůna,T., Hoff,K.J., Lomsadze,A., Stanke,M. and Borodovsky,M. (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform*, **3**, lqaa108.
59. Hoff,K.J., Lange,S., Lomsadze,A., Borodovsky,M. and Stanke,M. (2016) BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, **32**, 767–769.
60. Nachtweide,S. and Stanke,M. (2019) Multi-genome annotation with AUGUSTUS. *Methods Mol. Biol.*, **1962**, 139–160.
61. Stanke,M., Steinkamp,R., Waack,S. and Morgenstern,B. (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.*, **32**, W309–W312.
62. Kovaka,S., Zimin,A.V., Pertea,G.M., Razaghi,R., Salzberg,S.L. and Pertea,M. (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.*, **20**, 278.
63. Shumate,A., Wong,B., Pertea,G. and Pertea,M. (2022) Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS Comput. Biol.*, **18**, e1009730.
64. Haas,B.J., Delcher,A.L., Mount,S.M., Wortman,J.R., Smith,R.K., Hannick,L.I., Maiti,R., Ronning,C.M., Rusch,D.B., Town,C.D., *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**, 5654–5666.
65. Amarasinghe,S.L., Ritchie,M.E. and Gouil,Q. (2021) long-read-tools.Org: an interactive catalogue of analysis methods for long-read sequencing data. *Gigascience*, **10**, giab003.
66. Amarasinghe,S.L., Su,S., Dong,X.Y., Zappia,L., Ritchie,M.E. and Gouil,Q. (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.*, **21**, 30.
67. Burset,M. and Guigó,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.
68. Nanni,A., Titus-McQuillan,J., Moskalenko,O., Pardo-Palacios,F., Liu,Z., Conesa,A., Rogers,R.L. and McIntyre,L.M. (2021) The evolution of splicing: transcriptome complexity and transcript distances implemented in *TranD*. bioRxiv doi: https://doi.org/10.1101/2021.09.28.462251, 28 September 2021, preprint: not peer reviewed.
69. Tardaguila,M., de la Fuente,L., Marti,C., Pereira,C., Pardo-Palacios,F.J., Del Risco,H., Ferrell,M., Mellado,M., Macchietto,M., Verheggen,K., *et al.* (2018) SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.*, **28**, 396–411.
70. Pertea,G. and Pertea,M. (2020) GFF utilities: gffRead and GffCompare (version 2; peer review: 3 approved). *F1000Research*, **9**, ISCB Comm J-304.
71. Holmqvist,I., Backerholm,A., Tian,Y.R., Xie,G.J., Thorell,K. and Tang,K.W. (2021) FLAME: long-read bioinformatics tool for comprehensive spliceome characterization. *RNA*, **27**, 1127–1139.
72. Lienhard,M., van den Beucken,T., Timmermann,B., Hochradel,M., Boerno,S., Caiment,F., Vingron,M. and Herwig,R. (2022) IsoTools – a flexible workflow for long-read transcriptome sequencing analysis. https://doi.org/10.21203/rs.3.rs-1952129/v1.
73. Sakharkar,M.K., Chow,V.T. and Kangueane,P. (2004) Distributions of exons and introns in the human genome. *In Silico Biol.*, **4**, 387–393.
74. Spieth,J. and Lawson,D. (2006) In: *Overview of Gene Structure*. WormBook, pp. 1–10.
75. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F., *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
76. Jaccard,P. (1912) The distribution of the flora in the Alpine Zone. *New Phytol.*, **11**, 37–50.
77. Cock,P.J. and Whitworth,D.E. (2010) Evolution of relative reading frame bias in unidirectional prokaryotic gene overlaps. *Mol. Biol. Evol.*, **27**, 753–756.
78. Assis,R., Kondrashov,A.S., Koonin,E.V. and Kondrashov,F.A. (2008) Nested genes and increasing organizational complexity of metazoan genomes. *Trends Genet.*, **24**, 475–478.
79. Williams,B.A., Slamovits,C.H., Patron,N.J., Fast,N.M. and Keeling,P.J. (2005) A high frequency of overlapping gene

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025

expression in compacted eukaryotic genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 10936–10941.
80. Wright,B.W., Molloy,M.P. and Jaschke,P.R. (2022) Overlapping genes in natural and engineered genomes. *Nat. Rev. Genet.*, **23**, 154–168.
81. Roach,N.P., Sadowski,N., Alessi,A.F., Timp,W., Taylor,J. and Kim,J.K. (2020) The full-length transcriptome of C. elegans using direct RNA sequencing. *Genome Res.*, **30**, 299–312.
82. Wang,B., Tseng,E., Baybayan,P., Eng,K., Regulski,M., Jiao,Y.P., Wang,L.Y., Olson,A., Chougule,K., Van Buren,P., *et al.* (2020) Variant phasing and haplotypic expression from long-read sequencing in maize. *Commun. Biol.*, **3**, 78.
83. Pardo-Palacios,F., Reese,F., Carbonell-Sala,S., Diekhans,M., Liang,C., Wang,D., Williams,B., Adams,M., Behera,A., Lagarde,J., *et al.* (2021) Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. Res. Square doi: https://doi.org/10.21203/rs.3.rs-777702/v1, 03 August 2021, preprint: not peer reviewed.
84. Samuels,M.E., Schedl,P. and Cline,T.W. (1991) The complex set of late transcripts from the Drosophila sex determination gene sex-lethal encodes multiple related polypeptides. *Mol. Cell. Biol.*, **11**, 3584–3602.
85. Keyes,L.N., Cline,T.W. and Schedl,P. (1992) The primary sex determination signal of Drosophila acts at the level of transcription. *Cell*, **68**, 933–943.
86. Bopp,D., Calhoun,G., Horabin,J.I., Samuels,M. and Schedl,P. (1996) Sex-specific control of Sex-lethal is a conserved mechanism for sex determination in the genus Drosophila. *Development*, **122**, 971–982.
87. Bhadra,M.P., Bhadra,U. and Birchler,J.A. (2006) Misregulation of sex-lethal and disruption of male-specific lethal complex localization in drosophila species hybrids. *Genetics*, **174**, 1151–1159.
88. elegans Sequencing Consortium,C.. (1998) Genome sequence of the nematode C-elegans: a platform for investigating biology. *Science*, **282**, 2012–2018.
89. Davis,P., Zarowiecki,M., Arnaboldi,V., Becerra,A., Cain,S., Chan,J., Chen,W.J., Cho,J., da Veiga Beltrame,E., Diamantakis,S., *et al.* (2022) WormBase in 2022-data, processes, and tools for analyzing Caenorhabditis elegans. *Genetics*, **220**, iyac003.
90. Yang,N., Xu,X., Wang,R., Peng,W., Cai,L., Song,J., Li,W., Luo,X., Niu,L., Wang,Y., *et al.* (2017) Contributions of Zea mays subspecies Mexicana haplotypes to modern maize. *Nat. Commun.*, **8**, 1874.
91. Woodhouse,M.R., Cannon,E.K., Portwood,J.L., Harper,L.C., Gardiner,J.M., Schaeffer,M.L. and Andorf,C.M. (2021) A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biol.*, **21**, 385.
92. Gramates,L.S., Agapite,J., Attrill,H., Calvi,B.R., Crosby,M.A., Dos Santos,G., Goodman,J.L., Goutte-Gattat,D., Jenkins,V.K., Kaufman,T., *et al.* (2022) FlyBase: a guided tour of highlighted features. *Genetics*, **220**, iyac035.
93. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
94. Aleksander,S.A., Balhoff,J., Carbon,S., Cherry,J.M., Drabkin,H.J., Ebert,D., Feuermann,M., Gaudet,P., Harris,N.L., Hill,D.P., *et al.* (2023) The gene ontology knowledgebase in 2023. *Genetics*, **224**, iyad031.
95. Rogers,R.L., Shao,L., Sanjak,J.S., Andolfatto,P. and Thornton,K.R. (2014) Revised annotations, sex-biased expression, and lineage-specific genes in the Drosophila melanogaster group. *G3- Genes Genomes Genetics*, **4**, 2345–2351.
96. Chakraborty,M., Chang,C.H., Khost,D.E., Vedanayagam,J., Adrion,J.R., Liao,Y., Montooth,K.L., Meiklejohn,C.D., Larracuente,A.M. and Emerson,J.J. (2021) Evolution of genome structure in the *Drosophila simulans* species complex. *Genome Res.*, **31**, 380–396.
97. Tang,A.D., Soulette,C.M., van Baren,M.J., Hart,K., Hrabeta-Robinson,E., Wu,C.J. and Brooks,A.N. (2020) Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.*, **11**, 1438.
98. Marquez,Y., Brown,J.W.S., Simpson,C., Barta,A. and Kalyna,M. (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res.*, **22**, 1184–1195.
99. Frankish,A., Uszczynska,B., Ritchie,G.R., Gonzalez,J.M., Pervouchine,D., Petryszak,R., Mudge,J.M., Fonseca,N., Brazma,A., Guigo,R., *et al.* (2015) Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *Bmc Genomics (Electronic Resource)*, **16**(Suppl. 8), S2.
100. Zhao,S. and Zhang,B. (2015) A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *Bmc Genomics (Electronic Resource)*, **16**, 97.
101. Morales,J., Pujar,S., Loveland,J.E., Astashyn,A., Bennett,R., Berry,A., Cox,E., Davidson,C., Ermolaeva,O., Farrell,C.M., *et al.* (2022) A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, **604**, 310–315.
102. Kuznetsov,D., Tegenfeldt,F., Manni,M., Seppey,M., Berkeley,M., Kriventseva,E.V. and Zdobnov,E.M. (2023) OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.*, **51**, D445–D451.

Received: July 21, 2023. **Revised:** November 29, 2023. **Editorial Decision:** January 6, 2024. **Accepted:** January 18, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article/52/5/e28/7606259 by guest on 18 April 2025