Ranked Choice Voting for Representative Transcripts with TRaCE

Andrew J. Olson[1] and Doreen Ware[1,2]
[1] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11768 USA
[2] USDA ARS Robert W. Holley Center for Agriculture and Health Cornell University, Ithaca, New York 14853, USA

Abstract

Summary: Genome sequencing projects annotate protein-coding gene models with multiple transcripts, aiming to represent all of the available transcript evidence. However, downstream analyses often operate on only one representative transcript per gene locus, sometimes known as the canonical transcript. To choose canonical transcripts, TRaCE (Transcript Ranking and Canonical Election) holds an 'election' in which a set of RNA-seq samples rank transcripts by annotation edit distance. These sample-specific votes are tallied along with other criteria such as protein length and InterPro domain coverage. The winner is selected as the canonical transcript, but the election proceeds through multiple rounds of voting to order all the transcripts by relevance. Based on the set of expression data provided, TRaCE can identify the most common isoforms from a broad expression atlas or prioritize alternative transcripts expressed in specific contexts.

Availability and Implementation: Transcript ranking code can be found on GitHub at {{https://github.com/warelab/TRaCE}}

Contact: olson@cshl.edu, ware@cshl.edu

Supplementary information: Additional data are available in the GitHub repository.

Introduction

Genome sequencing projects often use complex, automated annotation pipelines to build reference sets of gene models. These pipelines mask repeats in the assembled genome, align protein and transcript evidence, and build gene models by aggregating overlapping alignments that adhere to known or inferred splice site patterns (Hoff et al. 2019; Campbell et al. 2014; Haas et al. 2003). Before a project releases a set of high-confidence gene models, additional filtering steps may remove transcript models that lack homology or are subject to nonsense-mediated degradation (NMD).

Alternative splicing contributes to the functional diversity of a genome (Black 2003); and new sequencing technology such as PacBio IsoSeq can capture splice variants at an unprecedented scale (Wang et al. 2016; Zhang et al. 2019; Bruijnesteijn et al. 2018). However, this heightened sensitivity can lead to the detection of transcriptional noise, which can be misreported by gene builders as biologically relevant splice variants. Furthermore, it is possible for partially processed transcripts containing retained introns that neither disrupt the reading frame nor introduce stop codons to be promoted to canonical transcripts (Figure 1).

Comparative gene tree analysis platforms such as Ensembl Compara (Herrero et al. 2016) operate on a single canonical transcript for each gene locus. In the absence of a curated canonical transcript, this is usually defined as the longest transcript with the longest translation, but this definition does not necessarily select the best representative transcript for a gene locus. Subsequently developed techniques have defined canonical isoforms based on expression level, sequence conservation, annotation of functional domains, or some combination of these features (Li et al. 2014; Pruitt et al. 2012; Rodriguez et al. 2018; The UniProt Consortium et al. 2016). For example, NCBI's RefSeq Select dataset uses an evidence hierarchy to identify a transcript in each protein-coding human and mouse gene model. The Matched Annotation from NCBI and EMBL-EBI (MANE) project has the goal of providing a unified set of human protein-coding gene annotations, but it is not known if and when such efforts will be applied to other species.

We developed TRaCE (Transcript Ranking and Canonical Election) to choose canonical transcripts based on data typically available at the time of a new genome annotation. In this approach, transcripts are ranked by length, domain coverage, and how well they represent a diverse population of transcriptome RNA-seq data. An 'election' based on ranked-choice voting selects a canonical transcript that is the first- or second-choice transcript for the majority of samples. The election proceeds through multiple rounds, effectively sorting all transcripts by relevance. Here we present the TRaCE algorithm and results obtained by running TRaCE on *Zea mays* and *Homo sapiens* gene annotations. In addition, we describe validation of TRaCE predictions by manual curation (Tello-Ruiz et al. 2019) and compare TRaCE to RefSeq/MANE Select and APPRIS (Rodriguez et al. 2018) human transcript classifications.

Methods

The first step in preparing to run TRaCE is to gather a diverse set of RNA-seq expression data covering a wide variety of tissues or conditions to act as 'voters' in the upcoming elections. The next step is to align the reads, assemble sample-specific transcripts, and quantify their expression. Each reference gene model with multiple transcripts (candidates) will hold an election to sort the reference transcripts by relevance (Figure 2).

In each election, samples rank the candidate transcripts based on the annotation edit distance (AED) to the most highly expressed overlapping sample-specific transcripts (Eilbeck et al. 2009). AED scores range from 0 (perfect agreement) to 1 (no overlap) and are calculated from the pairwise similarity of reference transcripts and aligned evidence based on the proportion of exonic overlap. Because there may be insufficient data to assemble full-length transcripts from samples in which the gene is expressed at low levels, the AED score calculation is restricted to overlapping portions of candidate transcripts. A maximum AED score cutoff (default, 0.5) prevents samples from voting for candidate transcripts with very little similarity. There are also cutoff parameters for minimum expression level (default TPM, 0.5) and proportion overlapping (default, 0.5) to filter out some noise in the sample transcriptome data. The election includes additional voters that rank transcripts based on domain coverage, protein length, and transcript length. To avoid overwhelming the length-based voters when running TRaCE with many samples, sample votes are weighted to balance the electorate. Default weights were selected to prioritize functional domain coverage over protein length and total transcript length.

Once each sample voter and the length-based voters have ranked the transcripts, the election proceeds in multiple rounds selecting winners until no candidates remain. In each round, TRaCE tallies votes for top-ranked candidates; and so long as there is a tie for first place, votes for the subsequent rankings are added to the tally.

Results

We ran TRaCE on a pre-release set of *Zea mays* B73 gene models with the set of 10 RNA-seq samples that had already been aligned to the genome as part of the evidence-based gene annotation pipeline (Hufford et. al. 2021). The samples were derived from shoot, root, embryo, endosperm, ear, tassel, anther, and three leaf sections (base, middle, and tip). StringTie version 1.3.5 (with the --rf flag) was used for transcript assembly and quantification (Pertea et al. 2016) and InterProScan version 5.38-76.0 was run to identify Pfam domains (Mulder and Apweiler 2007). The *Zea mays* B73 V5 annotation set (Zm00001eb) has 15,162 multi-transcript protein-coding gene models; for 5,616 of these (37%), the canonical transcript chosen by TRaCE was not the longest isoform. TRaCE selected canonical transcripts for the genome annotations of 25 additional maize accessions, 33-38% of which were not the longest isoform (Suppl Table 1).

We used two approaches to validate TRaCE's predictions on maize genes. First, we modified an interactive gene tree viewer, designed to flag problematic gene models by visual inspection of the multiple sequence alignment and domain annotations (Tello-Ruiz et al. 2020). We used this interface to compare maize B73 V5 canonical transcripts (Zm00001eb) selected by TRaCE with the prior set of maize V4 canonical transcripts (Zm00001d) selected by length criteria alone. A random selection of 173 pairs of genes for which the TRaCE canonical was not the longest transcript were evaluated in the gene tree viewer and flagged if the alignment was inconsistent with outgroup orthologs. Genes were flagged if there was a relative gain or loss of conserved sequence within the transcript or at either end. Of these gene pairs, 32% were flagged as problematic in Zm00001d only, 4% in Zm00001eb only, and 5% in both versions (Suppl Table 2). The most common issue in the flagged Zm00001d gene models was gain of sequence due to an intron retention. Thus, according to this approach, TRaCE was selecting better-conserved isoforms than the prior length-based algorithm.

In the second approach, TRaCE predictions were validated by student curators who were given a subset of 48 gene models with two to five transcripts, for which TRaCE's top-ranked isoform was not the longest isoform. The students, who were not aware of TRaCE's output, were asked to rate transcripts as best, good, or poor, based on viewing the gene structure and expression evidence in the Apollo genome browser (Dunn et al. 2019). Each gene model was curated by at least 3 different students. The transcript ratings were mapped to a score (best 2, good 1, poor -1). Transcript rankings from TRaCE and rankings based on length alone were compared to rankings based on curator scores. For each rank (1-5), we calculated the sum of the curator scores for the associated transcripts. The correlation of these sums between the length-based ranking and the curator-based ranking was 0.917, whereas the TRaCE and curator ranking sums had a higher correlation coefficient of 0.985 (Suppl Table 3).

We also ran TRaCE on human GRCh38 annotations (Frankish et al. 2019) with a diverse panel of 127 samples of human RNA-seq data covering the development of seven major organs (brain, cerebellum, heart, kidney, liver, ovary and testis) from 4 weeks post-conception to adulthood (https://www.ebi.ac.uk/gxa/experiments/E-MTAB-6814/Results). Reads were aligned with hisat2 version 2.1.0 (--dta --reorder), transcripts were assembled and quantified with stringtie version 2.1.4 (--conservative), and protein-coding reference transcripts were annotated with Pfam domains using InterProScan version 5.38-76.0 (Pertea et al. 2016; Mulder and Apweiler 2007).

The GRCh38 annotation set has 13,848 multi-transcript protein-coding gene models that were classified by both APPRIS and MANE Select. The TRaCE canonical was not the longest isoform in 3,717 (27%) of these gene models. For comparison, the principal isoform according to APPRIS and the MANE Select transcript was not the longest isoform in 3,061 (22%) and 4,292 (31%) of gene models, respectively. There are 1,202 gene models where APPRIS and MANE Select disagree. In these cases, TRaCE agrees with APPRIS on 408 (34%) genes, MANE Select on 597 (50%) genes, and neither on 197 (16%) genes. On the 12,646 multi-transcript gene models where APPRIS and MANE Select agree, TRaCE gives 10,677 (84%) transcripts rank 1, 1470 (12%) rank 2, 351 (3%) rank 3, and 148 (1%) rank 4 or higher. To assess TRaCE's performance on gene models with many transcripts, we compared TRaCE to APPRIS and MANE Select on the 90% of genes with 2-10 transcripts and the remaining 10% of human protein-coding gene models with 11-151 transcripts. There are 1,399 genes with many transcripts where APPRIS and MANE Select agree. In these cases, TRaCE selects 1,021 (73%) of these as the canonical transcript, 215 (15%) have rank 2, 92 (7%) have rank 3, and 71 (5%) have rank 4 or higher. On the 11,247 genes with fewer transcripts where APPRIS and MANE Select agree TRaCE assigns 9,656 (86%) rank 1, 1,255 (11%) rank 2, 259 (2%) rank 3, and 84 (1%) rank 4 or higher. For the initial release of TRaCE, we manually tuned the weights on TRaCE's length-based votes, but future versions may benefit from an automated parameter sweep to minimize these differences.

Acknowledgements

The authors thank Ware lab summer students Sarina Awatramani and Pragati Muthukumar for their preliminary research on intron retentions in Zm00001d, collaborators on the maize NAM project for feedback on TRaCE selected canonical transcripts, Kevin Ahern and students from his genetics course at Cornell University for curation, Bruno Contreras-Moreira and reviewers for suggesting comparisons with APPRIS and RefSeq Select, Cristina Marco and Marcela K. Tello-Ruiz for running curation jamborees and testing the algorithm, Liya Wang, Marcela K. Tello-Ruiz, and Christopher Patil for reviewing and editing the manuscript and Augusto Lima Diniz for the TRaCE acronym.

Funding

This work was supported by the USDA [1907-21000-030-00D].

References

Black, Douglas L. 2003. “Mechanisms of Alternative Pre-Messenger RNA Splicing.” *Annual Review of Biochemistry*. https://doi.org/10.1146/annurev.biochem.72.121801.161720.

Bruijnesteijn, Jesse, Marit K. H. van der Wiel, Wendy T. N. Swelsen, Nel Otting, Annemiek J. M. de Vos-Rouweler, Diënne Elferink, Gaby G. Doxiadis, et al. 2018. “Human and Rhesus Macaque Haplotypes Defined by Their Transcriptomes.” *Journal of Immunology* 200 (5): 1692–1701.

Campbell, Michael S., Carson Holt, Barry Moore, and Mark Yandell. 2014. “Genome Annotation and Curation Using MAKER and MAKER-P.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 48 (December): 4.11.1–39.

Dunn, Nathan A., Deepak R. Unni, Colin Diesh, Monica Munoz-Torres, Nomi L. Harris, Eric Yao, Helena Rasche, Ian H. Holmes, Christine G. Elsik, and Suzanna E. Lewis. 2019. “Apollo: Democratizing Genome Annotation.” *PLoS Computational Biology* 15 (2): e1006790.

Eilbeck, Karen, Barry Moore, Carson Holt, and Mark Yandell. 2009. “Quantitative Measures for the Management and Comparison of Annotated Genomes.” *BMC Bioinformatics* 10 (February): 67.

Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, et al. 2019. “GENCODE Reference Annotation for the Human and Mouse Genomes.” *Nucleic Acids Research* 47 (D1): D766–73.

Haas, Brian J., Arthur L. Delcher, Stephen M. Mount, Jennifer R. Wortman, Roger K. Smith Jr, Linda I. Hannick, Rama Maiti, et al. 2003. “Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies.” *Nucleic Acids Research* 31 (19): 5654–66.

Herrero, Javier, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J. Vilella, et al. 2016. “Ensembl Comparative Genomics Resources.” *Database*. https://doi.org/10.1093/database/bav096.

Hoff, Katharina J., Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke. 2019. “Whole-Genome Annotation with BRAKER.” *Methods in Molecular Biology* 1962: 65–95.

Hufford, Matthew B., Arun S. Seetharam, Margaret R. Woodhouse, Kapeel M. Chougule, Shujun Ou, Jianing Liu, William A. Ricci, et al. 2021. “De Novo Assembly, Annotation, and Comparative Analysis of 26 Diverse Maize Genomes.” *bioRxiv*. https://doi.org/10.1101/2021.01.14.426684.

Li, Hong-Dong, Rajasree Menon, Gilbert S. Omenn, and Yuanfang Guan. 2014. “Revisiting the Identification of Canonical Splice Isoforms through Integration of Functional Genomics and Proteomics Evidence.” *Proteomics* 14 (23-24): 2709–18.

Mulder, Nicola, and Rolf Apweiler. 2007. “InterPro and InterProScan: Tools for Protein Sequence Classification and Comparison.” *Comparative Genomics*. https://doi.org/10.1385/1-59745-515-6:59.

Pertea, Mihaela, Daehwan Kim, Geo M. Pertea, Jeffrey T. Leek, and Steven L. Salzberg. 2016. “Transcript-Level

Expression Analysis of RNA-Seq Experiments with HISAT, StringTie and Ballgown.” *Nature Protocols* 11 (9): 1650–67.

Pruitt K, Brown G, Tatusova T, et al. The Reference Sequence (RefSeq) Database. 2002 Oct 9 [Updated 2012 Apr 6]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 18. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21091/

Rodriguez, Jose Manuel, Juan Rodriguez-Rivas, Tomás Di Domenico, Jesús Vázquez, Alfonso Valencia, and Michael L. Tress. 2018. “APPRIS 2017: Principal Isoforms for Multiple Gene Sets.” *Nucleic Acids Research* 46 (D1): D213–17.

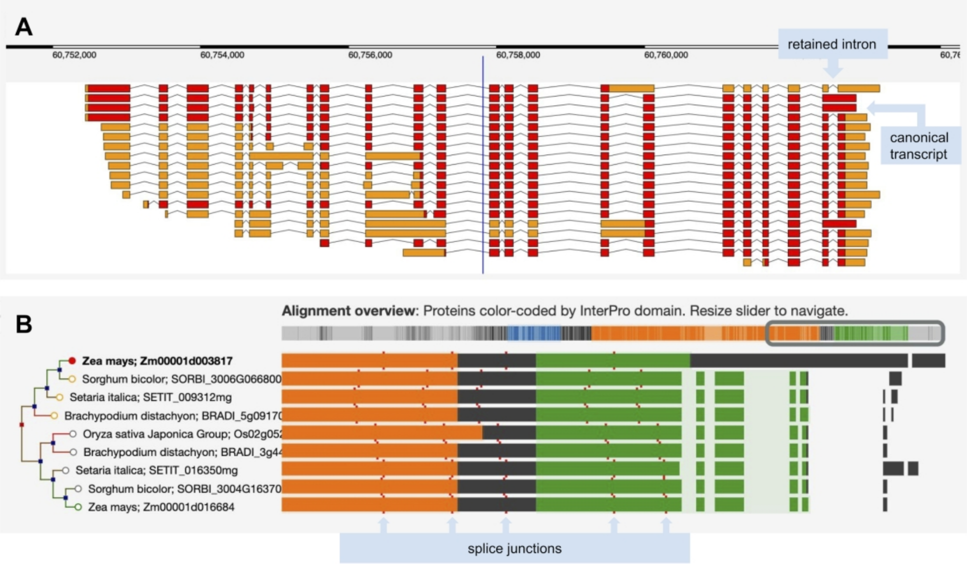
Tello-Ruiz, Marcela K., Cristina F. Marco, Fei-Man Hsu, Rajdeep S. Khangura, Pengfei Qiao, Sirjan Sapkota, Michelle C. Stitzer, et al. 2019. “Double Triage to Identify Poorly Annotated Genes in Maize: The Missing Link in Community Curation.” *PloS One* 14 (10): e0224086.

Tello-Ruiz, Marcela K., Sushma Naithani, Parul Gupta, Andrew Olson, Sharon Wei, Justin Preece, Yinping Jiao, et al. 2020. “Gramene 2021: Harnessing the Power of Comparative Genomics and Pathways for Plant Research.” *Nucleic Acids Research*, November. https://doi.org/10.1093/nar/gkaa979.

The UniProt Consortium, Alex Bateman, Maria Jesus Martin, Claire O’Donovan, Michele Magrane, Emanuele Alpi, Ricardo Antunes, et al. 2016. “UniProt: The Universal Protein Knowledgebase.” *Nucleic Acids Research* 45 (D1): D158–69.

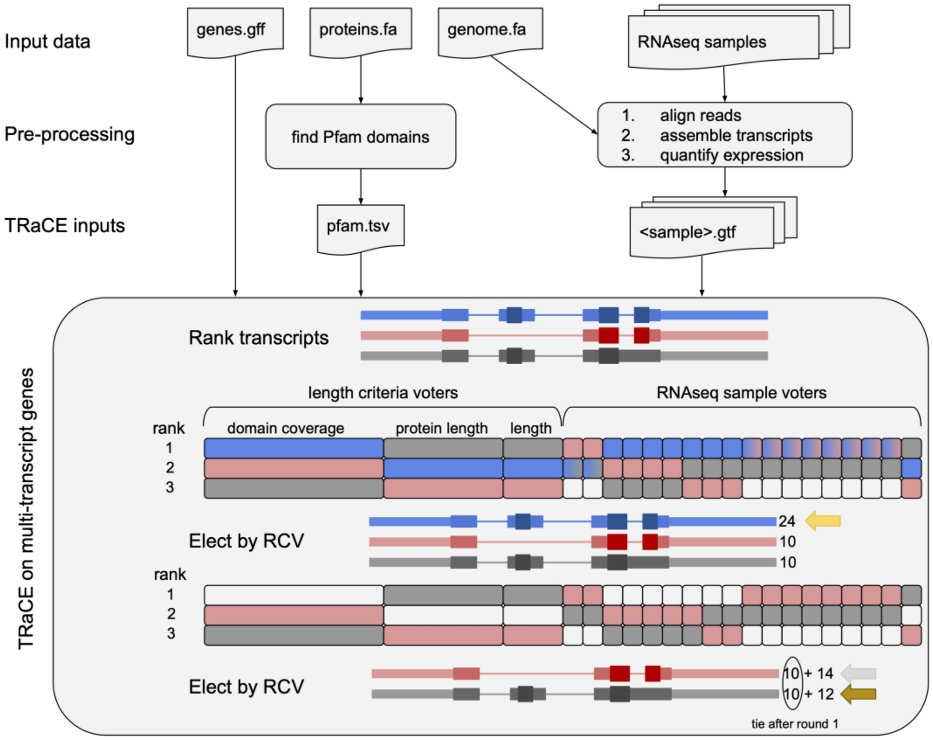
Wang, Bo, Elizabeth Tseng, Michael Regulski, Tyson A. Clark, Ting Hon, Yinping Jiao, Zhenyuan Lu, Andrew Olson, Joshua C. Stein, and Doreen Ware. 2016. “Unveiling the Complexity of the Maize Transcriptome by Single-Molecule Long-Read Sequencing.” *Nature Communications* 7 (June): 11708.

Zhang, Guoqiang, Min Sun, Jianfeng Wang, Meng Lei, Chenji Li, Duojun Zhao, Jun Huang, et al. 2019. “PacBio Full-Length cDNA Sequencing Integrated with RNA-Seq Reads Drastically Improves the Discovery of Splicing Transcripts in Rice.” *The Plant Journal: For Cell and Molecular Biology* 97 (2): 296–305.


A
60,752,000
60,754,000
60,756,000
60,758,000
60,760,000
60,76
retained intron
canonical transcript
B
Alignment overview: Proteins color-coded by InterPro domain. Resize slider to navigate.
Zea mays; Zm00001d003817
Sorghum bicolor; SORBI_3006G066800
Setaria italica; SETIT_009312mg
Brachypodium distachyon; BRADI_5g0917
Oryza sativa Japonica Group; Os02g052
Brachypodium distachyon; BRADI_3g44
Setaria italica; SETIT_016350mg
Sorghum bicolor; SORBI_3004G16370
Zea mays; Zm00001d016684
splice junctions

A) The complex set of transcript models for the Zea mays B73 gene sbe4 (starch branching enzyme4). Red blocks show the predicted coding regions, and orange blocks are untranslated regions. The longest translation contains a retained intron and was selected as the canonical transcript for Compara gene tree analysis. B) The left side shows a portion of the gene tree focused on this maize gene and displaying homologs from Sorghum bicolor, Setaria italica, Brachypodium distachyon, and Oryza sativa Japonica. The right side shows regions of protein sequences participating in the multiple sequence alignment, color coded by InterPro domain. The first row shows a unique region relative to other species that derives from the retained intron.

254x190mm (300 x 300 DPI)



Flowchart of preparation of TRaCE inputs and a schematic of the rank-choice voting (RCV) approach to select transcripts for an example gene with three transcripts (blue, red, gray). Exon thickness corresponds to non-coding, coding, and functional regions with Pfam domains. Voters are represented by rectangles, and rank transcripts by length criteria (9, 6, or 3 votes) or AED (1 vote per sample). Eight of the samples rank the red and blue transcripts equally (blue-red gradient), so both get tallied in round 1. RCV selects the blue transcript first with 24 rank 1 votes. After removing the blue votes from consideration, the red and gray transcripts tie with 10 rank 1 votes, but the red transcript is elected with 14 rank 2 votes.

254x190mm (300 x 300 DPI)