Noncanonical GA and GG 5’ Intron Donor Splice Sites Are Common in the Copepod Eurytemora affinis

Hugh M. Robertson
Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801

ABSTRACT
The noncanonical 5’ intron donor splice sites GA and GG are exceedingly rare in described eukaryotic genomes; however, they are present in ~12% of introns in the genome of the copepod Eurytemora affinis. Failure to recognize the high frequency of these donor sites compromised the modeling of genes in this newly sequenced genome, including 10 conserved ionotropic glutamate receptor (GluR) family genes curated herein. These introns appear to have been acquired recently, along with many additional idiosyncratic introns. Their high frequency implies the evolution of modified intron donor splice site recognition in this copepod.

KEYWORDS
noncanonical intron donor splice sites copepod genome Eurytemora intron evolution GA donors

The canonical sequence for 5’ intron donor splice sites in eukaryotes has an obligate G to start the intron followed by U, or occasionally C, as part of a larger consensus splice site sequence, with DNA sequence AG/GTAAGT (Mount 1982). For example, among 222,263 introns in the human genome, Parada et al. (2014) found only 184 noncanonical splice sites, of which only 14 were GA and 32 were GG donors (0.006 and 0.014%, respectively). Most of these are sites of alternative splicing, as are the most intensively studied GA donors, belonging to the vertebrate fibroblast growth factor receptor 1–3 genes (Brackenridge et al. 2003).

Eyun et al. (2017) examined the evolution of arthropod chemosensory genes, focusing on crustaceans and particularly the genome of the copepod Eurytemora affinis, one of the first copepod genomes to be sequenced. Among others, the authors reported nine Ionotropic Receptor (IR) genes and five members of the ionotropic GluR family, from which the IRs evolved (Benton et al. 2009); however, the amino acid sequences they provided for these proteins are almost all truncated at one or both ends. I have completed the gene models for five conserved members of the IR family and the five GluR family members, and found that they contain an unusually high frequency of noncanonical GA and GG 5’ intron donor splice sites. A sample of 26 other large genes indicates that this unusual phenomenon is likely to be genome-wide.

MATERIALS AND METHODS
Gene models were built manually in the text editor TEXTWRANGLER using genomic sequences from the assembly published in Eyun et al. (2017) and presented in the i5k Workspace@NAL genome browser (Poelchau et al. 2015; https://i5k.nal.usda.gov/). RNAseq reads spanning each intron were either obtained from the i5k genome browser if there was RNAseq mapped across an intron, or from the Short Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) using BLASTN searches with flanking exon sequences as queries and default parameters. Exons missing in assembly gaps were recreated using raw genome reads from the SRA, using the RNAseq reads obtained above as queries. Models were compared with the protein sequences reported by Eyun et al. (2017); however, these appear to have been derived from a transcriptome rather than from the genome itself, because they contain descriptors that indicate they were derived from transcripts from CUFFLINKS or another transcriptome assembly (e.g., EaffNMDAR2-1 comp31752_c0_seq1 CUFF.1933.3). They also commonly span both GA and GG donor introns as well as misassembled exons and exons missing from the assembly (see, for example,

Copyright © 2017 Robertson
doi: https://doi.org/10.1534/g3.117.300189
Manuscript received August 23, 2017; accepted for publication October 16, 2017; published Early Online October 27, 2017.
This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Supplemental material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.117.300189/-/DC1.
1Address for correspondence: Department of Entomology, University of Illinois at Urbana-Champaign, 320 Morrill Hall, 505 S. Goodwin Ave., Urbana, IL 61801. E-mail: hughrobe@uiuc.edu
The five conserved members of the IR family (IR8a, 21a, 25a, 76b, and 93a) and the five GluR members (GluR1 and 2, and NMDAR1, 2-1, and 2-2) were successfully extended by employing RNAseq and genome reads to full-length genes (sequences available in File S1). Full-length status was confirmed by comparison with orthologs from other insects (e.g., Benton et al. 2009; Croset et al. 2010; Terrapon et al. 2014; Eyun et al. 2017). Suffix “F” after gene name indicates that the genome assembly had to be repaired for a complete gene model to be built (details of each gene model are provided in File S1). Lengths are from start to stop codon in large scaffolds, excluding exons present on short separate scaffolds or those that were built de novo, both of which presumably belong in sequence gaps in the large scaffolds, the lengths of which are included in these counts. Only coding exons are included (IR76b and NMDAR2-2 have single noncoding 5’ exons). Models are the number of models in the automated gene set available at the i5k Workspace@NAL genome browser (EAFF_v0.5.3), and usually not all exons are modeled. Numbers in parentheses are for the proteins reported in Eyun et al. (2017).
phases of these GA and GG introns for the copepod. Alignment of the positions and phases of these GA and GG introns for the five ionotropic receptors with those of homologous genes in T. californicus, as well as two other available crustaceans, H. azteca (also available from the i5k pilot consortium) and D. pulex (Croset et al. 2010), as well as various insects (e.g., Benton et al. 2009; Terrapon et al. 2014; Ioannidis et al. 2017) and other arthropods like a tick and mite (Gulia-Nuss et al. 2016; Hoy et al. 2016), reveals that they are unique introns in Eurytemora, along with many additional idiosyncratic introns with canonical donors. For example, of the 33 introns in IR25a in E. affinis, nine are shared with other arthropods and the remaining 24, including the six GA or GG donors, are idiosyncratic to it. It appears that this copepod underwent an explosion of intron gains, including those with noncanonical donors.

Alternative splicing of the GA donor sites in the vertebrate fibroblast growth factor receptor 1–3 genes is a complicated process involving a nearby consensus sequence (Brackenridge et al. 2003), but no such sequence was noticed in these copepod GA or GG donor introns, which are also not alternatively spliced but rather required for their genes to encode full-length proteins. The high frequency of these noncanonical donors implies the evolution of modified 5′ intron donor site recognition in this copepod. The only intact U1 snRNA in the genome assembly has the same highly conserved 5′ end with sequence complementing the 5′ intron donor splice consensus of AG/GTAAGT common to animals; however, recognition of the 5′ donor site is affected by other components of the snRNPs, so it is unclear how these noncanonical donor sites are recognized.

ACKNOWLEDGMENTS

This work was supported by a Romano Professorial Scholarship from the University of Illinois at Urbana-Champaign.

LITERATURE CITED

Benton, R., K. S. Vannice, C. Gomez-Diaz, and L. B. Voshall, 2009 Variant ionotropic glutamate receptors as chemosensory receptors in Drosophila. Cell 136: 149–162.

Brackenridge, S., A. O. Wilkie, and G. R. Scretton, 2003 Efficient use of a ‘dead-end’ GA 5′ splice site in the human fibroblast growth factor receptor genes. EMBO J. 22: 1602–1631.

Croset, V., R. Rytz, S. F. Cummins, A. Budd, D. Brawand et al., 2010 Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genet. 6: e1001064.

Eyun, S. I., H. Y. Soh, M. Posavi, J. B. Munro, D. S. T. Hughes et al., 2017 Evolutionary history of chemosensory-related gene families across the Arthropoda. Mol. Biol. Evol. 34: 1838–1862.

Gulia-Nuss, M., A. B. Nuss, J. M. Meyer, D. E. Sonenshine, R. M. Roe et al., 2016 Genomic insights into the Ixodes scapularis tick vector of Lyme disease. Nat. Commun. 7: 10507.

Hoy, M. A., R. M. Waterhouse, K. Wu, A. S. Estep, P. Ioannidis et al., 2016 Genome sequencing of the phytoseiid predatory mite Metaseius occidentalis reveals completely atomized Hox genes and superdynamic intron evolution. Genome Biol. Evol. 8: 1762–1775.

Ioannidis, P., F. A. Simao, R. M. Waterhouse, M. Manni, M. Seppey et al., 2017 Genomic features of the damselﬂy Calopteryx splendens representing a sister clade to most insect orders. Genome Biol. Evol. 9: 415–430.

Kang, S., D. H. Ahn, J. H. Lee, S. G. Lee, S. C. Shin et al., 2017 The genome of the Antarctic-endemic copepod, Tigriopus kingsejongensis. Gigascience 6: 1–9.

Mount, S. M., 1982 A catalogue of splice junction sequences. Nucleic Acids Res. 10: 459–472.

Parada, G. E., R. Munita, C. A. Cerda, and K. Gyasing, 2014 A comprehensive survey of non-canonical splice sites in the human transcriptome. Nucleic Acids Res. 42: 10564–10578.

Poehlchau, M., C. Childers, G. Moore, V. Tsavatapalli, J. Evans et al., 2015 The i5k Workspace@NAL—enabling genomic data access, visualization and curation of arthropod genomes. Nucleic Acids Res. 43: D714–D719.

Terrapon, N., C. Li, H. M. Robertson, L. Ji, X. Meng et al., 2014 Molecular traces of alternative social organization in a termite genome. Nat. Commun. 5: 3636.

Communicating editor: H. Salz