Software Impacts 10 (2021) 100163

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

JCAST: Sample-specific protein isoform databases for mass spectrometry-based proteomics experiments

R.W. Ludwig, Edward Lau *

Department of Medicine, Consortium for Fibrosis Research & Translation, University of Colorado School of Medicine, Aurora, CO, USA

ARTICLE INFO

Keywords:
Proteomics
Mass spectrometry
RNA sequencing
Alternative splicing
Protein isoforms
Proteoforms
Proteogenomics

ABSTRACT

JCAST is an open-source Python software tool that allows users to easily create custom protein sequence databases for proteogenomic applications. JCAST takes in RNA sequencing data containing alternative splicing junctions as input, models the likely translatable protein isoform sequences within a particular sample, performs in silico translation using annotated open reading frames, and outputs sample-specific protein sequence databases in FASTA format to support downstream mass spectrometry data analysis of protein isoforms. This article describes the functionality and usage of the JCAST software and documents a stable code repository for user access.

Code metadata

Current Code version	*v.0.3.3*
Permanent link to code/repository used of this code version	https://github.com/SoftwareImpacts/SIMPAC-2021-131
Permanent link to reproducible capsule	https://codeocean.com/capsule/6293191/tree/v1
Legal Code License	*MIT*
Code Versioning system used	*git*
Software Code Language used	*Python*
Compilation requirements, Operating environments & dependencies	*biopython, gtfparse, pandas, requests, tqdm, scipy, scikit-learn, matplotlib, pomegranate*
If available Link to developer documentation/manual	https://github.com/ed-lau/jcast
Support email for questions	edward.lau@cuanschutz.edu

1. Introduction (background and problem)

A common task in biomedical research is to determine the abundances of protein species in a sample, from which one can discover correlates between protein levels and physiological states. Multiple protein isoforms with distinct amino acid sequences can be created from a single gene through various biological processes including alternative splicing, which combines exonic segments in a protein coding gene in manners different from that in the primary, canonical gene product containing constitutively spliced exons. Mass spectrometry-based proteomics is commonly employed to identify and quantitate proteins on a large scale, but the discovery of non-canonical protein isoforms currently remains challenging due in part to informatics challenges.

The typical computational workflow to analyze mass spectrometry data involves matching acquired experimental spectra to theoretical spectra generated from a compilation of known protein sequences using a database search engine. Known protein sequences are typically retrieved from sequence databases including UniProt [1] and RefSeq [2]. Due to incomplete annotations, these databases frequently contain only a subset of all true protein isoform sequences in a sample, leading to the

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail address: edward.lau@cuanschutz.edu (E. Lau).

https://doi.org/10.1016/j.simpa.2021.100163
Received 5 October 2021; Received in revised form 22 October 2021; Accepted 24 October 2021

non-identification of omitted isoform sequences. This problem is further exacerbated in protein isoforms with specific spatiotemporal expression patterns (appearing only in specific tissues or cell states) and those in poorly annotated non-human organisms, forming a significant barrier to the characterization of protein isoform function in physiological and disease settings.

Proteogenomics approaches attempt to overcome this problem by creating sample-specific protein sequence databases, such as by performing in silico translation of a sample-specific transcriptome from RNA sequencing data into custom protein sequences. This has emerged as an alternative solution for identifying non-canonical gene products at the protein level. Despite progress however, challenges remain for producing sample-specific protein isoform databases. Not all isoform transcript carries equal protein coding potential, hence methods are needed to identify and prioritize the isoforms that are more likely to produce stable protein. The translated databases may also contain sequences not physically present in a particular sample, under certain scenarios of which an oversized database can inflate false positive identifications.

JCAST provides a software tool for basic and clinical researchers to support the protein isoform identification task, by allowing easy creation of protein isoform sequences in a sample-specific protein sequence database for the analysis of mass spectrometry data. JCAST implements several methodological advances, e.g.: (i) JCAST implements a mixture model to predict the translatable isoform transcripts from splice junction read distributions; (ii) JCAST strictly avoids of premature termination codons to reduce database sizes; and (iii) JCAST classifies output sequences by confidence tiers based on frame and alignment to full-length canonical sequences.

2. Functionality and usage overview

JCAST v.0.3.3 is provided as an open source Python module, and can be acquired directly at GitHub or from PyPI via pip. JCAST can be run standalone in the command line through `python -m jcast`. The basic architecture of JCAST is shown in Fig. 1. The main function outputs custom protein isoform sequences in FASTA format and requires three inputs. The first is sample-specific RNA sequencing data, which is pre-processed through an upstream alignment and junction counting pipeline consisting of existing third-party tools STAR [3] followed by rMATS [4]. The second input is a genome annotation gtf file from Ensembl or GENCODE containing open reading frame information including transcript translation start, end, and phase. The third input in a genome FASTA file which is used to match exon coordinates to genetic sequences.

JCAST v.0.3.3 takes in the following options: `-o --out` specifies the path of the output folder; the `-c --canonical` flag controls whether JCAST outputs canonical sequences even if the alternative splice junction is not translated; `-r --read` specifies the minimal summed skipped junction read count for a junction to be considered for translation [default: 1] and can be overridden by the -m flag (see below); `-q --qvalue` specifies the lower and upper range of FDR-adjusted P values between two biological replicates for a junction to be considered for translation [defaults: 0 1]. Since the first release, JCAST has undergone several recent improvements, including improved compatibility with GENCODE gtf files, more detailed logging, and the implementation of a read count model directly in the Python module through the `-m --model` flag. If set, the model supersedes the prior user-defined minimal read count values that had to be manually identified for each dataset.

JCAST first reads in the RNA-seq data and finds all junctions falling within one of the five rMATS alternative splicing types. It then performs filtering of junctions based on the `-q`, `-r`, and/or `-m` options. JCAST applies a power transformation to the sum of the skipped junction read counts for a splice event across all biological and technical replicates. It then fits a gamma/Gaussian two-component mixture model to determine the minimal read count for a junction to be predicted as belonging to the high-read, likely translatable population of isoform transcripts (Fig. 2). Each qualifying junction is represented as a slice of the upstream, alternative, and downstream exons. Here a slice is a partial DNA or protein sequence corresponding to one or more alternative splice junctions, which can be combined to recover perspective isoform sequence. The junctions are trimmed by translation starts and ends in the GTF file, and the translation phase from the upstream exon is retrieved. JCAST then reads the genome file in memory and retrieves nucleotide sequences, and attempts in silico translation using the retrieved phase.

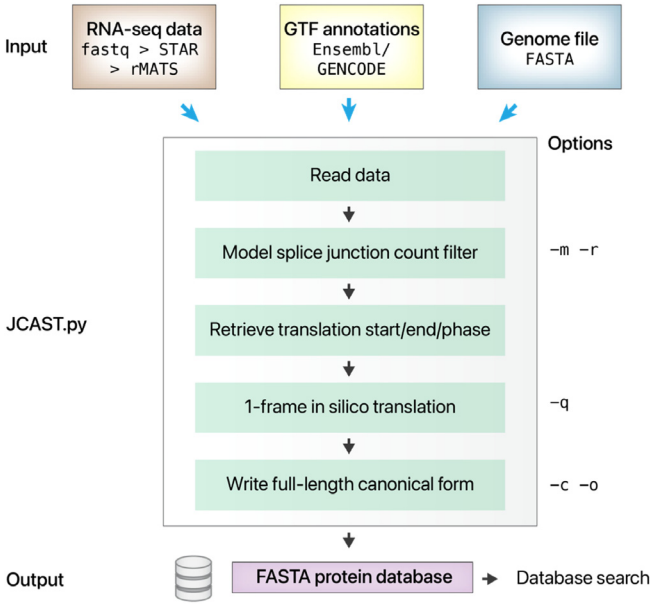

Fig. 1. JCAST functionality and usage. JCAST takes in RNA sequencing data, genome annotations, and genome sequences, and outputs custom protein sequence databases.

JCAST enforces one-frame translation and groups all non-canonical sequences translated from splice junctions into four confidence tiers in separate FASTA output files. Tier 1 junctions are translated in-frame according to the annotated frames, and do not result in a frameshift or premature stop codon. Tier 2 junctions are translated in frame according to annotated translation frames and do not encounter premature stop codons, but have encountered a possible frameshift (length differences in alternative slices that are not multiples of 3). Tier 3 junctions encounter a premature stop codon under the retrieved translation frame, but may be translated fully without encountering a stop codon in another frame. Finally, Tier 4 junctions encounter a premature stop codon in at least one of the two alternative junction slices in any reading frame. They are written into a FASTA file if they can be translated using one of three frames into a peptide fragment at least a certain proportion in length as the successfully translated slice (default as 0.33 and can be changed in `params.py`. The Tier 4 low-confidence sequences are provided for reference, but should either be excluded from database search or interpreted with caution.

For each confidence tier, JCAST further attempts to recover hypothetical full-length alternatively sliced protein sequences by joining the translated slice and judging the quality of their sequence alignment to canonical sequences in a database. To do so, JCAST makes a call to the UniProt web API as needed and caches any retrieved canonical sequence locally. Note that the joined sequences may or may not represent biological full-length protein isoforms due to the nature of short-read sequencing, as actual isoforms may contain multiple

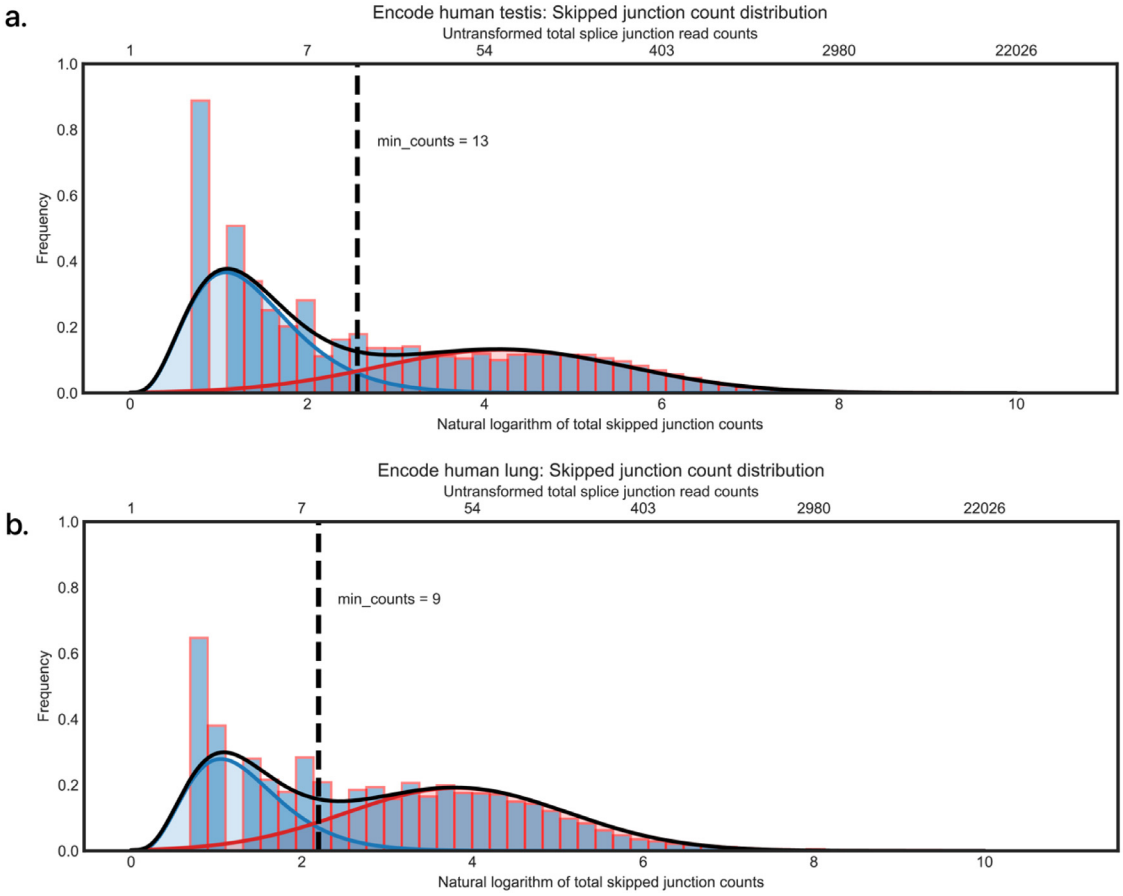

Fig. 2. Junction count model. The majority of alternative splice isoform transcripts are likely to be untranslated. JCAST makes the assumption that high-abundance transcript population is more likely to be translatable or detectable in mass spectrometry experiments, and uses a gamma/Gaussian mixture model to automatically select the read count threshold for translation. The best-fit models for ENCODE human (a) testis and (b) lung datasets are shown.

alternative splice sites or alternative translation starts and ends in conjunction. Sequences that do not align back to the canonical protein are designated as orphan sequences and output to separate files. A total of up to nine FASTA files are created (T1–T4 full-length proteins, T1–T4 orphan proteins, and canonical sequences identical to UniProt SwissProt entries). The output FASTA files can be used in combination or alone for downstream analyses. The protein sequence databases are compatible with major database search algorithms commonly utilized to analyze mass spectrometry based proteomics experiments to identify proteins, such as Comet [5], MSFragger [6], and MaxQuant [7].

JCAST currently has the several limitations. It does not model novel open reading frames (ORFs) or transcripts containing coding single nucleotide/amino acid variants (SNVs/SAAVs). These usages are addressed by other existing software tools. Secondly, JCAST is currently limited to short-read sequencing data only, hence full-length transcript isoforms are not explicitly modeled. Thirdly, a connection to UniProt is required to retrieve canonical sequences.

3. Impact overview

Several software tools and packages allow the translation of custom protein databases, including ProteomeGenerator [8], customProDB [9], and Galaxy-P [10]. These existing tools primarily focus on translating single amino acid variants differing from the reference genome or finding novel open reading frames. JCAST is distinguished from comparable tools by providing a tool targeted for alternative splicing derived isoform sequences. It also enforces one-frame translation and premature termination codon avoidance to avoid the inflation of database size that can lead to false positive protein identifications.

Since its release, JCAST has been used by us and others to examine the biology of alternative splicing-derived protein isoforms. Our team and collaborators applied an experimental workflow supported by JCAST to examine the shifts in protein isoform abundance during human induced pluripotent stem cell (hiPSC) differentiation into cardiomyocytes [11], as well as to perform in silico translation of potential alternative splicing products in order to develop targeted mass spectrometry assays [12]. JCAST has been used by other researchers to analyze the sequence features of predicted translatable isoforms. Kelly et al. analyzed the isoforms in hiPSC-derived cardiomyocyte transcriptome and found thousands of putative N-glycosylation sites that may be gained, lost, or shifted between canonical and alternative isoforms, suggesting a potential biological function of alternative splicing may be to regulate the availability of N-glycosylation substrates during development and diseases [13]. The logic behind JCAST to adjudicate protein isoform detectability by splice junction read counts has also been successfully adopted and cited by other groups [14].

4. Conclusion

Proteogenomics is a developing field where customized, sample-specific protein sequence databases are created to interrogate novel or non-canonical protein gene products. JCAST incorporates several contemporary concepts to create custom sequence databases, including modeling the likely detectability of protein isoforms through RNA-seq junction read counts, and constraints on translation frames and premature termination codons. The software and workflow are compatible with short-read RNA sequencing and proteomics data in any tissue or organisms where alternative splicing is of interest. Ongoing work aims

to incorporate joint modeling of long-read sequencing and ribosome footprint profiling to improve database accuracy.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by NHLBI award R00-HL144829 and NIH Office of the Director award R03-OD032666 to E.L., and University of Colorado Consortium for Fibrosis Research and Translation (CFReT) funding.

References

[1] UniProt Consortium, UniProt: The universal protein knowledgebase in (2021), Nucleic Acids Res. 49 (2021) D480–D489.
[2] N.A. O'Leary, et al., Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation, Nucleic Acids Res. 44 (2016) D733–745.
[3] A. Dobin, et al., STAR: Ultrafast universal RNA-seq aligner, Bioinforma. Oxf. Engl. 29 (2013) 15–21.
[4] S. Shen, et al., rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-seq data, Proc. Natl. Acad. Sci. USA 111 (2014) E5593–5601.
[5] J.K. Eng, et al., A deeper look into Comet–implementation and features, J. Am. Soc. Mass Spectrom. (2015) 1865–1874.
[6] A.T. Kong, F.V. Leprevost, D.M. Avtonomov, D. Mellacheruvu, A.I. Nesvizhskii, MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics, Nature Methods 14 (2017) 513–520.
[7] S. Tyanova, T. Temu, J. Cox, The MaxQuant computational platform for mass spectrometry-based shotgun proteomics, Nat. Protoc. 11 (2016) 2301–2319.
[8] P. Cifani, et al., ProteomeGenerator: A framework for comprehensive proteomics based on de novo transcriptome assembly and high-accuracy peptide mass spectral matching, J. Proteome Res. 17 (2018) 3681–3692.
[9] X. Wang, B. Zhang, customProDB: An R package to generate customized protein databases from RNA-seq data for proteomics search, Bioinforma. Oxf. Engl. 29 (2013) 3235–3237.
[10] G.M. Sheynkman, et al., Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations, BMC Genomics 15 (2014) 703.
[11] E. Lau, et al., Splice-junction-based mapping of alternative isoforms in the human proteome, Cell Rep. 29 (2019) 3751–3765.e5.
[12] Y. Han, et al., Computation-assisted targeted proteomics of alternative splicing protein isoforms in the human heart, J. Mol. Cell. Cardiol. 154 (2021) 92–96.
[13] M.I. Kelly, et al., Importance of evaluating protein glycosylation in pluripotent stem cell-derived cardiomyocytes for research and clinical applications, Pflugers Arch. 473 (2021) 1041–1059.
[14] B. Salovska, et al., Isoform-resolved correlation analysis between mRNA abundance regulation and protein level degradation, Mol. Syst. Biol. 16 (2020) e9170.