Bioinformatics, 2017, 1–3
doi: 10.1093/bioinformatics/btx527
Advance Access Publication Date: 26 August 2017
Applications Note

Genome analysis

CESAR 2.0 substantially improves speed and accuracy of comparative gene annotation

Virag Sharma[1,2,†], Peter Schwede[1,2,†] and Michael Hiller[1,2,*]

[1]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden 01307, Germany and [2]Max Planck Institute for the Physics of Complex Systems, Dresden 01187, Germany

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint first authors.

Associate Editor: Janet Kelso

Received on April 3, 2017; revised on July 23, 2017; editorial decision on August 15, 2017; accepted on August 23, 2017

Abstract

Motivation: Homology-based gene prediction is a powerful concept to annotate newly sequenced genomes. We have previously demonstrated that whole genome alignments can be utilized for accurate comparative coding gene annotation.

Results: Here we present CESAR 2.0 that utilizes genome alignments to transfer coding gene annotations from one reference to many other aligned genomes. We show that CESAR 2.0 is 77 times faster and requires 31 times less memory compared to its predecessor. CESAR 2.0 substantially improves the ability to align splice sites that have shifted over larger distances, allowing for precise identification of the exon boundaries in the aligned genome. Finally, CESAR 2.0 supports entire genes, which enables the annotation of joined exons that arose by complete intron deletions. CESAR 2.0 can readily be applied to new genome alignments to annotate coding genes in many other genomes at improved accuracy and without necessitating large-computational resources.

Availability and implementation: Source code is freely available at https://github.com/hillerlab/CESAR2.0

Contact: hiller@mpi-cbg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

An essential step after sequencing a new genome is the annotation of coding genes. Homology-based gene annotation approaches exploit the fact that coding genes are typically well conserved between related species and are an essential part of many gene annotation pipelines (Cantarel *et al.*, 2008; Curwen *et al.*, 2004; Haas *et al.*, 2008; Stanke *et al.*, 2006). One powerful homology-based approach uses alignments of entire genomes to transfer gene annotations from one well-annotated reference genome to many other aligned query genomes (Sharma *et al.*, 2016; Sharma and Hiller, 2017; Stanke *et al.*, 2008; Zhu *et al.*, 2007).

To utilize genome alignments to transfer annotations of coding genes, we have previously developed the Coding Exon-Structure Aware Realigner (CESAR) method (Sharma *et al.*, 2016). CESAR uses a Hidden Markov Model (HMM) that captures the exon's reading frame and splice site annotation of the reference species and uses the Viterbi algorithm to 'realign' the putative exon sequence provided by a genome alignment. CESAR aims at finding a new alignment with an intact reading frame and consensus splice sites for exons that are conserved in the query genome. We have shown that this 're-alignment' step avoids numerous spurious frameshift mutations in truly conserved genes. In addition, the new alignment detects splice sites that have been shifted in the query species, which is important to precisely identify the boundaries of intact exons. The identified splice site coordinates are then used as exon coordinates to annotate exons of all coding genes in the query genome.

CESAR has the following three limitations. First, its runtime is relatively slow, especially for long exons where the re-alignment step can take several hours or even days (Fig. 1A). Thus, large computer resources were necessary to annotate genes in many genomes.

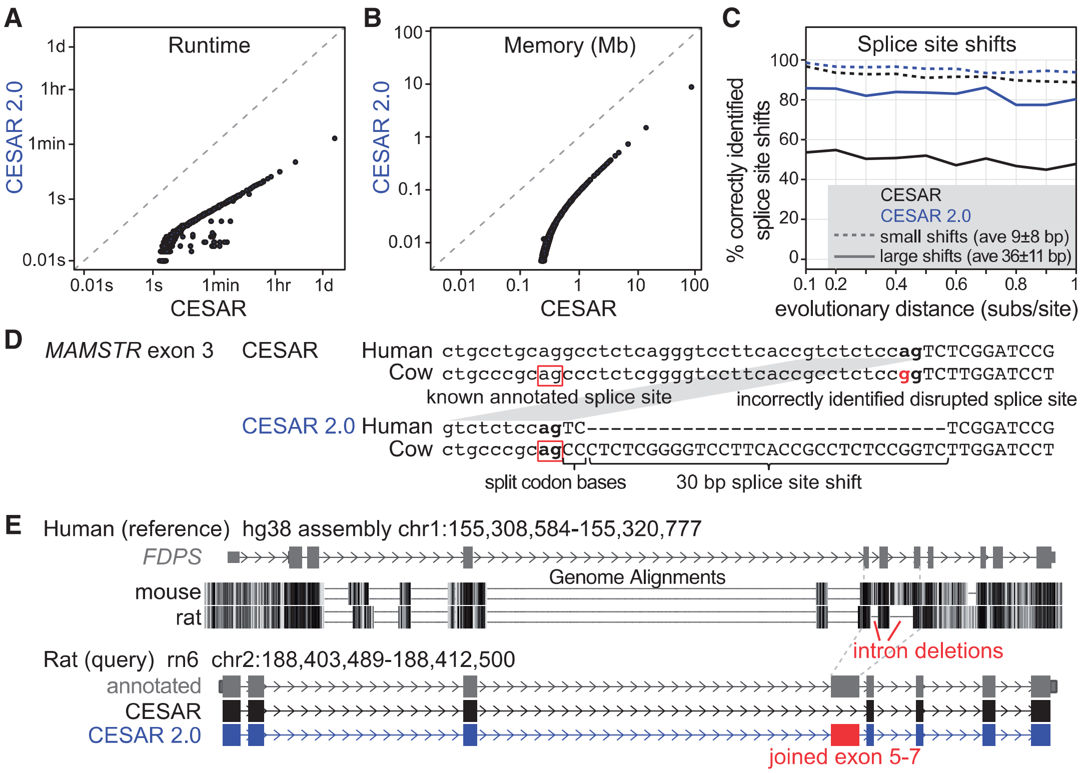

Fig. 1. CESAR 2.0 is faster, requires less memory and has improved accuracy in annotating intact exons. (**A**, **B**) Aligning 5977 exons shows that on average CESAR 2.0 achieves a speedup of 77X (A) and requires 31 times less memory (B). (**C**) CESAR 2.0 (blue line) has a substantially improved ability to correctly identify both splice sites in exon datasets where we shifted one splice site into the intron or into the exon. Dashed and solid lines show datasets that differ in the average distance of the shift. Each dataset consists of 500 exons that we evolved *in silico* for 0.1–1 substitutions per site to sample 10 different evolutionary distances (Sharma *et al.*, 2016). (**D**) Exon 3 of the *MAMSTR* gene shows an acceptor mutation (GG dinucleotide) and a 30 bp acceptor shift in cow. Only CESAR 2.0 but not the previous method identifies the shifted splice site and thus the correct exon start in cow. Intronic bases are in lower case, exonic bases are in upper case. (**E**) Example of intron deletions: The genome alignment visualization shows that two introns of the *FDPS* gene are completely deleted in rat (but not mouse), probably by recombination with a processed transcript of the same gene (Coulombe-Huntington and Majewski, 2007). The new gene mode of CESAR 2.0 models these events and correctly annotates the large exon in rat (red box) that consists of the joined human exons 5–7

Second, while it achieved a high accuracy of 91% in correctly aligning shifted splice sites, it typically failed to align a splice site that has shifted over larger distances. Third, CESAR is only able to realign individual exons. This is problematic for gene structure changes such as complete intron deletions (Coulombe-Huntington and Majewski, 2007), where CESAR reports both the deletion of the donor and acceptor splice site and thus fails to annotate two or more exons. Here, we present CESAR 2.0, which is much faster, substantially improves the accuracy in detecting distal splice site shifts and provides a new gene mode to align entire genes while considering the possibility of intron deletions.

2 Methods

We implemented CESAR 2.0 from scratch in C for the following two reasons. First, the previous python implementation based on the generic HMM library YAHMM was relatively slow, especially for long exons. Second, YAHMM requires that all HMM states have to emit the same number of characters. This does not allow to directly model frameshifting deletions as states that emit a partial 1 or 2 bp codon. As a workaround, the previous method modified the original input sequence by replacing each base i by the triplet (i, i + 1, i + 2) and implemented each codon unit as three triplet-emitting states with transitions modeling frameshifting deletions. This additional model complexity further increased runtime. To simplify the model structure and to directly support partial codon emissions, the new HMM models each codon unit as a state that emits a full codon [according to the codon substitution probabilities (Schneider *et al.*, 2005) for the respective reference codon], and two states that emit partial codons [(di)nucleotides with a uniform distribution] to allow for frameshifting deletions (Supplementary Fig. S1).

The HMM implementation underlying CESAR 2.0 is optimized for speed and efficient memory usage, which we achieved by using pointers to emission tables that hold the same probabilities, instead of storing tables with the same content repeatedly.

To improve the accuracy in detecting such large splice site shifts, CESAR 2.0 implements a higher probability for allowing frame-preserving insertions and deletions right downstream of the acceptor and right upstream of the donor splice site states. Apart from these changes, CESAR 2.0 uses the same parameters as the previous implementation, resulting in the same or an even higher accuracy.

While CESAR was limited to align each exon separately, CESAR 2.0 implements a new gene mode that is able to align all exons of a coding gene to the genomic locus in the query.

3 Results

We first compared the runtime of CESAR 2.0 to its previous implementation by re-aligning 5977 exons (all coding exons of genes located on mouse chromosome 19) of different lengths. On average, CESAR 2.0 achieved a speedup of 77X and required 31 times less memory (Fig. 1A and B, Supplementary Table S1). Notably, for exons longer than 500 bp, the average speedup is even 178X. To further compare the runtime on a realistic application, we mapped all 19, 919 human genes (196 259 exons) to the mouse genome. CESAR 2.0 required 7 h using a single core on a desktop computer, which is 126 times faster than the previous implementation. We conclude that CESAR 2.0 is much faster and enables large-scale comparative gene annotation without large computational resources.

Next, we tested the ability of CESAR 2.0 to detect distal splice site shifts. We used ten simulated datasets of exons covering ten different evolutionary distances, where we shifted the splice sites as described in (Sharma *et al.*, 2016). Then we determined the percentage of correctly identified shifted splice sites. We found that CESAR 2.0 consistently improves the ability to detect splice site shifts (Fig. 1C). In particular, CESAR 2.0 achieved a ~33% higher accuracy for the dataset where splice sites were shifted over a large average distance of 36 bp (Fig. 1C). A real example of a 30 bp splice site shift is shown in Figure 1D. To test if the improved ability to detect splice site shifts translates into an improved comparative gene annotation, we used mouse Ensembl genes to compare the ability of CESAR 2.0 and its predecessor in annotating genes in mouse. Applied to all 19, 919 human protein-coding genes, CESAR 2.0 correctly detected both boundaries of 1, 396 additional exons that had incorrect boundaries in the CESAR alignment (Supplementary Fig. S2). This shows that CESAR 2.0 is able to annotate numerous exons that are missed by CESAR, which increases the precision of correctly annotated exons from 95.9 to 96.3%. We conclude that the special insertion/deletion probabilities implemented in CESAR 2.0 substantially improve the ability to correctly align distal shifted splice sites, which increases the completeness of comparative gene annotations.

Finally, we tested the gene mode of CESAR 2.0, which is able to model genomic deletions of entire introns, an evolutionary event that affects over 100 mammalian genes (Coulombe-Huntington and Majewski, 2007). As exemplified by the *FDPS* gene (Fig. 1E), both introns 5 and 6 are completely deleted in the rat, resulting in a larger rat exon that comprises the human exons 5, 6 and 7. The single-exon mode of CESAR fails to annotate these three exons as the splice sites are deleted. In contrast, the gene mode of CESAR 2.0 recognizes the deletion of two introns and precisely annotates the larger single exon (Fig. 1E). Thus, CESAR 2.0 increases the sensitivity of comparative coding gene annotation by annotating joined exons that arose by intron deletion events.

4 Discussion

Genome alignments are highly useful to map gene annotations from a well-annotated reference to other aligned genomes. The new CESAR 2.0 substantially enhances the utility of genome alignments for comparative gene annotation by (i) being significantly faster and more memory efficient, which allows routine application without large computer resources, (ii) improving the ability to identify distal splice site shifts, which increases the accuracy of the gene annotation and (iii) providing a new gene mode that is able to detect complete intron deletions and can be used to annotate entire genes instead of individual exons. We also provide a workflow that allows users to readily apply CESAR 2.0 to quickly annotate coding genes in many other genomes at improved accuracy (https://github.com/hillerlab/CESAR2.0).

Acknowledgements

We are grateful to Computer Service Facilities of the MPI-CBG and MPI-PKS for their support and Katrin Sameith for help with the figure.

Funding

This work was supported by the Max Planck Society and the German Research Foundation (HI 1423/3-1).

Conflict of Interest: none declared.

References

Cantarel,B.L. *et al.* (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, **18**, 188–196.

Coulombe-Huntington,J. and Majewski,J. (2007) Characterization of intron loss events in mammals. *Genome Res.*, **17**, 23–32.

Curwen,V. *et al.* (2004) The Ensembl automatic gene annotation system. *Genome Res.*, **14**, 942–950.

Haas,B.J. *et al.* (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.*, **9**, R7.

Schneider,A. *et al.* (2005) Empirical codon substitution matrix. *BMC Bioinformatics*, **6**, 134.

Sharma,V. *et al.* (2016) Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. *Nucleic Acids Res.*, **44**, e103.

Sharma,V. and Hiller,M. (2017) Increased alignment sensitivity improves the usage of genome alignments for comparative gene annotation. *Nucleic Acids Res.*, **45**, 8369–8377.

Stanke,M. *et al.* (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, **24**, 637–644.

Stanke,M. *et al.* (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.

Zhu,J. *et al.* (2007) Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput. Biol.*, **3**, e247.