Reconstruction of the spinach coding genome using full-length transcriptome without a reference genome

Elizabeth Tseng[1], Hamid Ashrafi[2*], Amanda M. Hulse-Kemp[2], Allen Van Deynze[2]

[1] PacBio, Menlo Park, CA
[2] University of California, Davis, CA
[*]Current address: North Carolina State University, Raleigh, NC

Abstract

The use of single molecule, real-time sequencing from PacBio to sequence transcriptomes (the Iso-Seq method), which produces *de novo*, high-quality, full-length transcripts, has revealed an astonishing amount of alternative splicing in eukaryotic species. With the Iso-Seq method, it is now possible to identify gene families and analyze alternative splicing even without a genome to map against.

We present Cogent, a tool for finding gene families and reconstructing the coding genome without a reference genome. Cogent uses k-mer similarities to first partition the transcripts into different gene families. Then, for each partition, the transcripts are used to build a splice graph. Cogent identifies bubbles resulting from sequencing errors, minor variants, and exon skipping events, and attempts to resolve each splice graph down to the minimal set of reconstructed *contigs*. The contigs can be used to visualize alternative splicing events and compare genome assemblies.

We applied Cogent to the Iso-Seq data for spinach, *Spinacia oleracea*, for which there is both a PacBio-based and an Illumina-based draft genome. Using the PacBio assembly as ground truth, Cogent's gene family partitioning had a recall of 99% and precision of 99%. For the reconstruction, 86% of the partitions were resolved to a single contig by Cogent and was validated to be also a single contig in the PacBio genome. In addition, we identified missing or fragmented portions in the draft genome.

Summary

- Cogent does coding genome reconstruction using full-length transcripts without the need for a reference genome
- Reconstructed contigs can be used for visualization of alternative splicing and help with genome scaffolding

Software available at:
https://github.com/Magdoll/Cogent

Spinach Genome and Transcriptome

Transcriptome

Iso-Seq dataset of mixed tissues

# of Transcripts: 68,263
Transcript Lengths: 528 bp – 6 kb (mean: 2.1 kb)

Genome

VERSION	PACBIO	ILLUMINA
# of contigs	2,881	89,597
# of bases	911,322,488	590,438,824
Contig stats		
Min:	1000 bp	201 bp
Max:	9.8 Mb	426 kb
Median:	23 kb	871 bp
N50:	1.48 Mb	30 kb

Methods

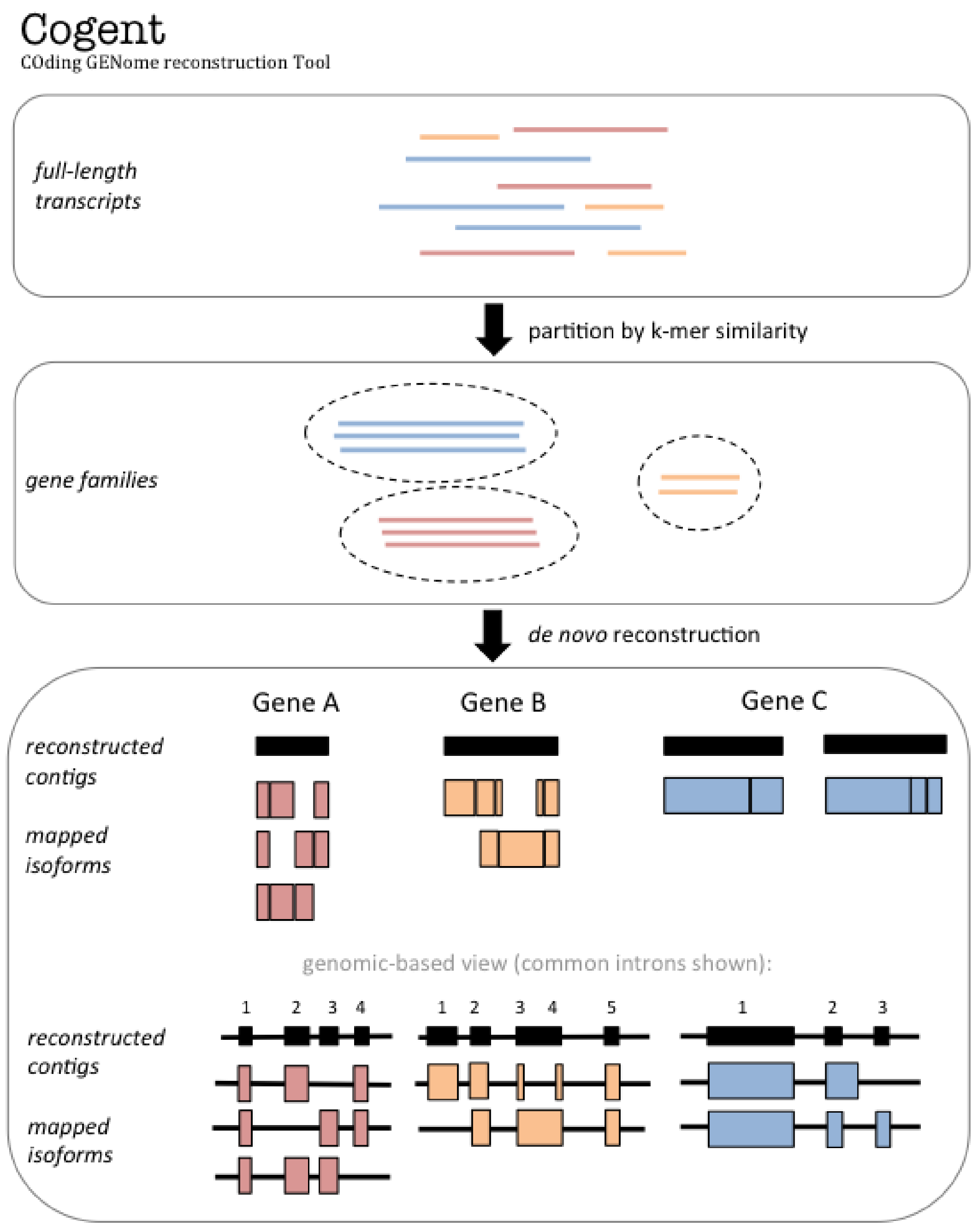

Figure 1. Cogent workflow. Given a set of full-length transcript sequences, Cogent partitions the sequences into gene families, then reconstructs the transcribed regions for each gene by building a de Bruijn graph and simplifying the bubbles caused by errors, minor variants, and exon skipping

Gene Family Partitioning is done by constructing an *k*-mer similarity graph where the edge weights are the proportion of shared k-mers, then partitioning the graph using normalized cut.

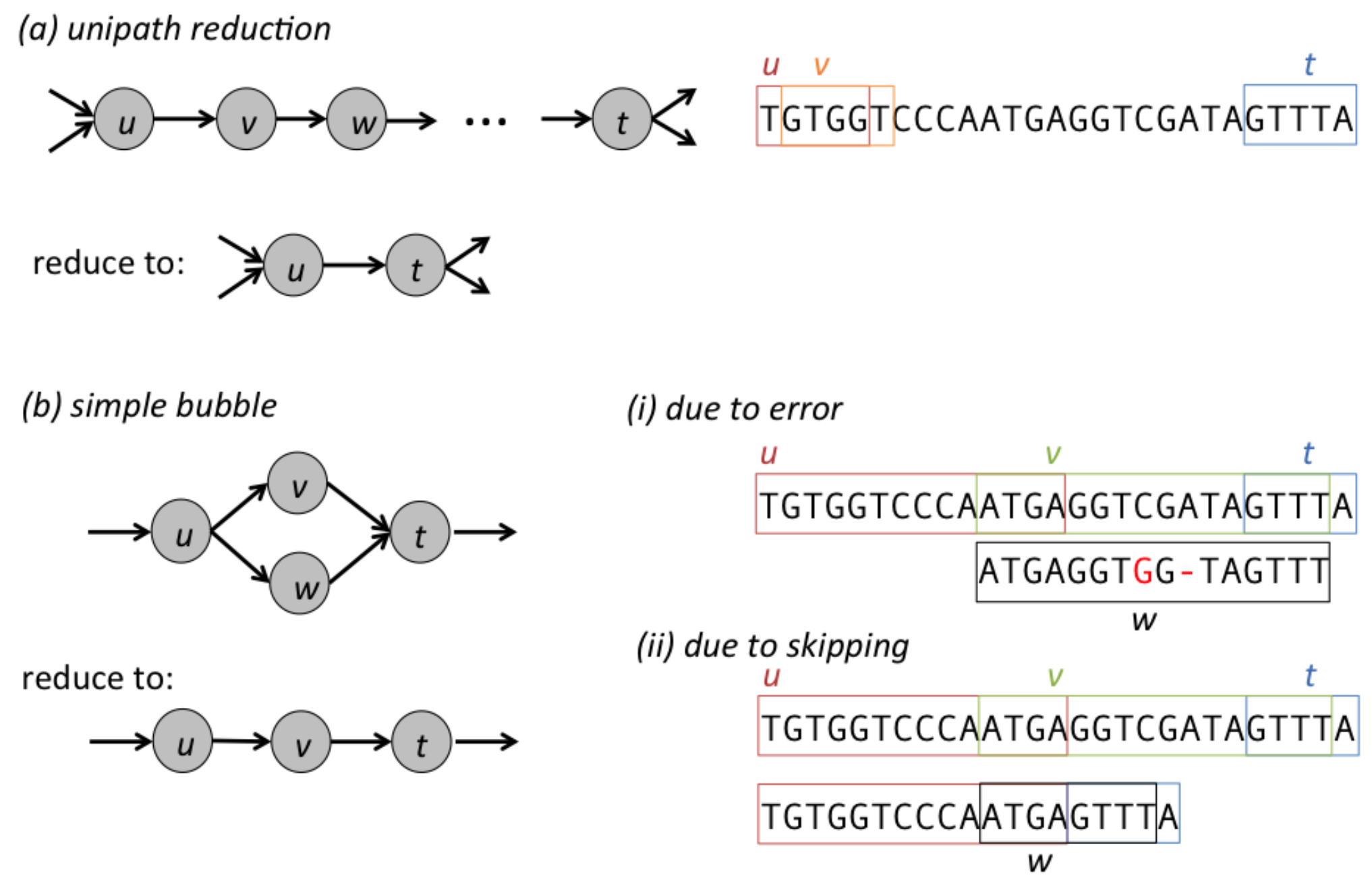

Figure 2. Reducing the de Bruijn graph by collapsing (a) unipaths, which corresponds to transcribed segments shared by all isoforms; and (b) simple bubbles, which can be caused by either errors or exon skipping (or intron retention) events. In the case of errors, either *v* or *w* is removed. In the case of exon skipping, the node containing the extra exon(s) is kept. Note that after removing one of the nodes, $u \rightarrow v \rightarrow t$ is now a unipath that can be reduced.

Runtime Statistics

PROGRAM	PARTITIONING		RECONSTRUCTION	
	Runtime (sec)	Memory (MB)	Runtime (sec)	Memory (MB)
Cogent	5139	1597	46	4326
CD-HIT-EST	10882	1555	NA	NA

Partitioning is run with 12 CPUs on single node.
Reconstruction shows avg. runtime for each partition.

De Novo Gene Family Partitioning

	# OF PARTITIONS		muc RECALL	muc PRECISION
	Size=1	Size>=2		
genome	3214	8381	NA	NA
Cogent	3195	8425	**0.990**	**0.991**
CD-HIT-EST	5013	8664	0.952	0.988

$$mucRecall(genomePartition, denovoPartition) = \frac{\sum_{c\ in\ genomePartition}(size(c) - overlap(c, denovoPartition))}{\sum_{c\ in\ genomePartition}(size(c) - 1)}$$

$$mucPrecision(genomePartition, denovoPartition) = \frac{\sum_{c\ in\ denovoPartition}(size(c) - overlap(c, genomePartition))}{\sum_{c\ in\ denovoPartition}(size(c) - 1)}$$

CD-HIT-EST created more singleton partitions, which resulted in high precision but low recall (inability to find gene families). In contrast, Cogent shows both high recall and precision.

De Novo Genome Reconstruction

IN COGENT…	IN GENOME..	PACBIO GENOME	ILLUMINA GENOME
1 contig	no hit	24	1
1 contig	1 contig	7272	7063
1 contig	> 1 contig	59	291
> 1 contig	1 contig	913	867
> 1 contig	> 1 contig	157	203
TOTAL		8425	8425

Evaluating Cogent's reconstruction using genome:

- 7272 (86%) of the partitions resolved to 1 contig
- 913 (10%) of the partitions resolved to > 1 contig, could be due to lack of exon connectivity information

Comparing genome assemblies:

- PacBio assembly misses 24 gene coding regions, later found to be in p-reads
- PacBio assembly more contiguous than Illumina assembly
- Mapping of Cogent reconstructed contigs back to the genome shows PacBio and Illumina assembly to have comparable accuracy (PacBio: 99.36%, ILLN: 99.56%)

(a) Cogent-based view

(b) Genome-based view

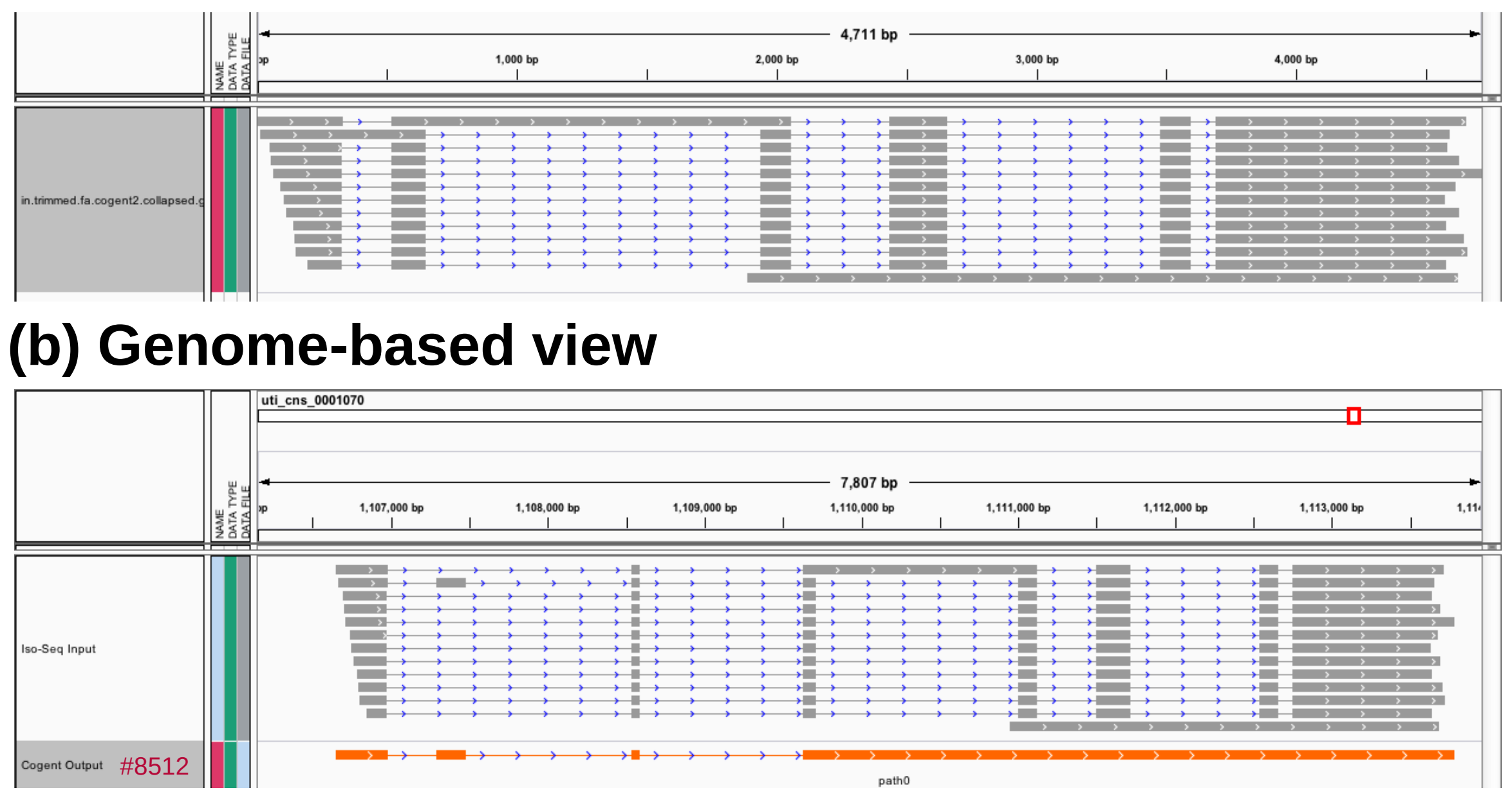

Figure 3. Example of Cogent reconstruction. **(a)** Without the genome, isoforms can be visualized by mapping back to the reconstructed contig. **(b)** Mapping the isoforms (gray) and contig (orange) back to the genome for validation.

Acknowledgements

The Illumina genome assembly was provided by the Spinach Genome Consortium. We also acknowledge Jason Chin, Paul Peluso, David Rank, and Nicole Rapicavoli from PacBio for assisting in the library preparation, sequencing, and assembly process for the PacBio genome assembly.