Cogent: Reconstructing the coding genome from full-length transcriptome sequences

Elizabeth Tseng[1], Xiang Qin[2], Muthuswamy Raveendran[2], Yue Liu[2], Shwetha C. Murali[2], Yi Han[2], Kim C. Worley[2], Jeffrey Rogers[2], Ting Hon[1], Tyson Clark[1], Nathan Tublitz[3]

[1] Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025
[2] Baylor College of Medicine, Houston, TX [3] University of Oregon, Eugene, OR

Introduction

For highly complex and large genomes, a well-annotated genome may be computationally challenging and costly, yet the study of alternative splicing events and gene annotations usually rely on the existence of a genome. Long-read sequencing technology provides new opportunities to sequence full-length cDNAs, avoiding computational challenges that short-read transcript assembly brings. The use of single molecule, real-time sequencing from Pacific Biosciences to sequence transcriptomes (the Iso-Seq^{TM} 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 reconstruct the transcribed regions of the genome using just the transcripts themselves.

We present Cogent, a tool for finding gene families and reconstructing the coding genome in the absence of a reference genome.

We first apply Cogent to a proof-of-concept dataset of nine human genes. We show that Cogent successfully identifies the gene families and reconstructs the coding genome. We then apply Cogent a cuttlefish dataset, for which there is a highly fragmented, Illumina-based draft genome assembly and little annotation. We show that Cogent successfully discovers gene families and can reconstruct the coding region of gene loci. The reconstructed *contigs* can then be used to visualize alternative splicing events, identify minor variants, and even be used to improve genome assemblies.

Tool & Dataset Information

Input to Cogent are Iso-Seq datasets which are full-length transcript sequences with accuracy ≥ 99%.

	Input Sequences	Sequence Lengths	Expected # of Gene Families
Human Heart	204	500 bp – 7.4 kb (avg: 3.5 kb)	9
Cuttlefish Brain	24,313	400 bp – 7 kb (avg: 2.9 kb)	Unknown

An unofficial version of Cogent is available for download at: https://github.com/Magdoll/Cogent

Cogent: Coding Genome Reconstruction

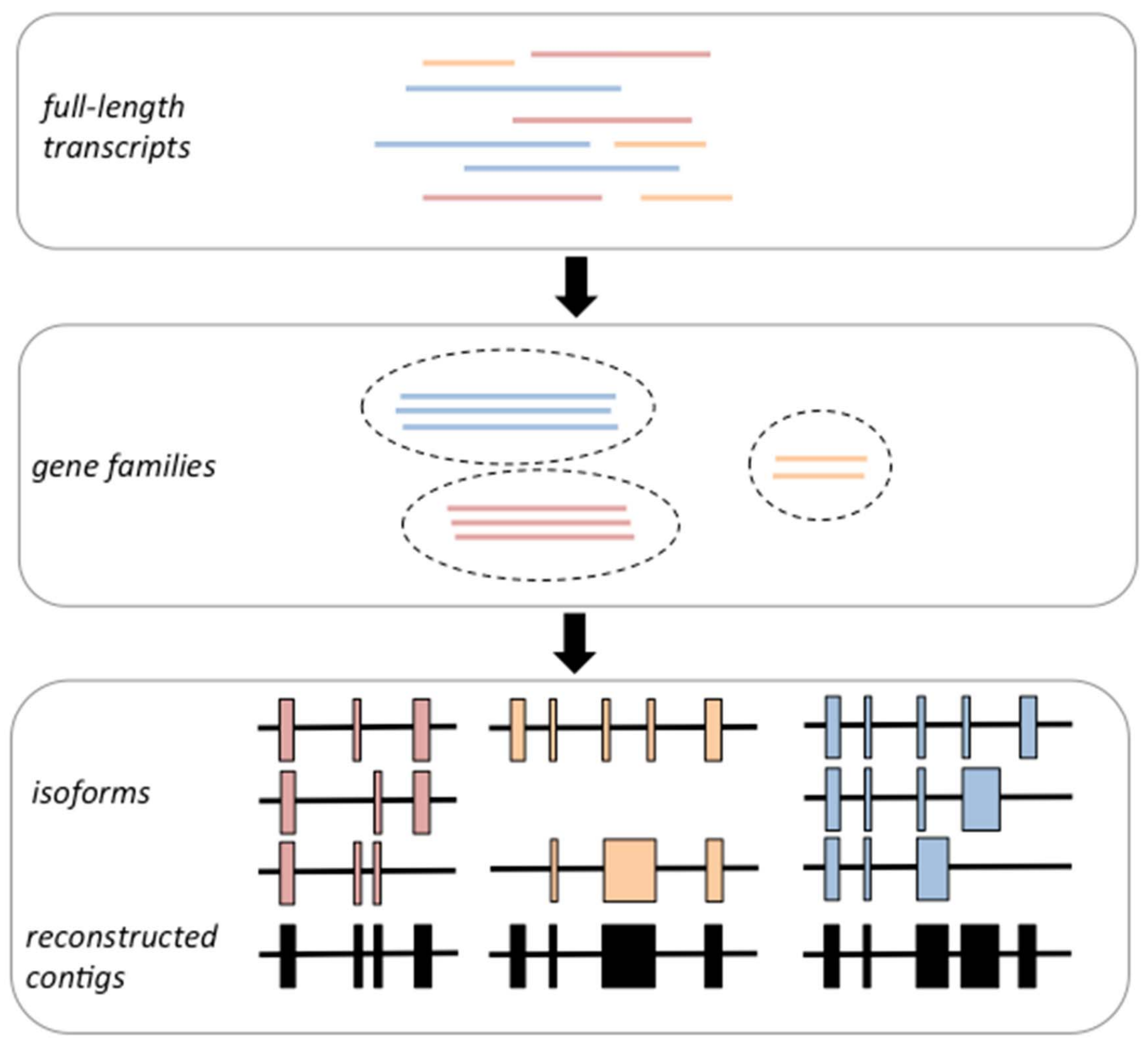

Figure 1. Cogent workflow. Given a set of full-length transcript sequences, Cogent first partitions the sequences into gene families, then reconstructs the transcribed regions for each gene by building a de Bruijn graph and simplifying the bubbles caused by errors, minor variants, and exon skipping. Reconstructed contigs are shown here with introns depicted, but the actual output will have all common introns spliced out.

Family Finding 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 a normalized cut.

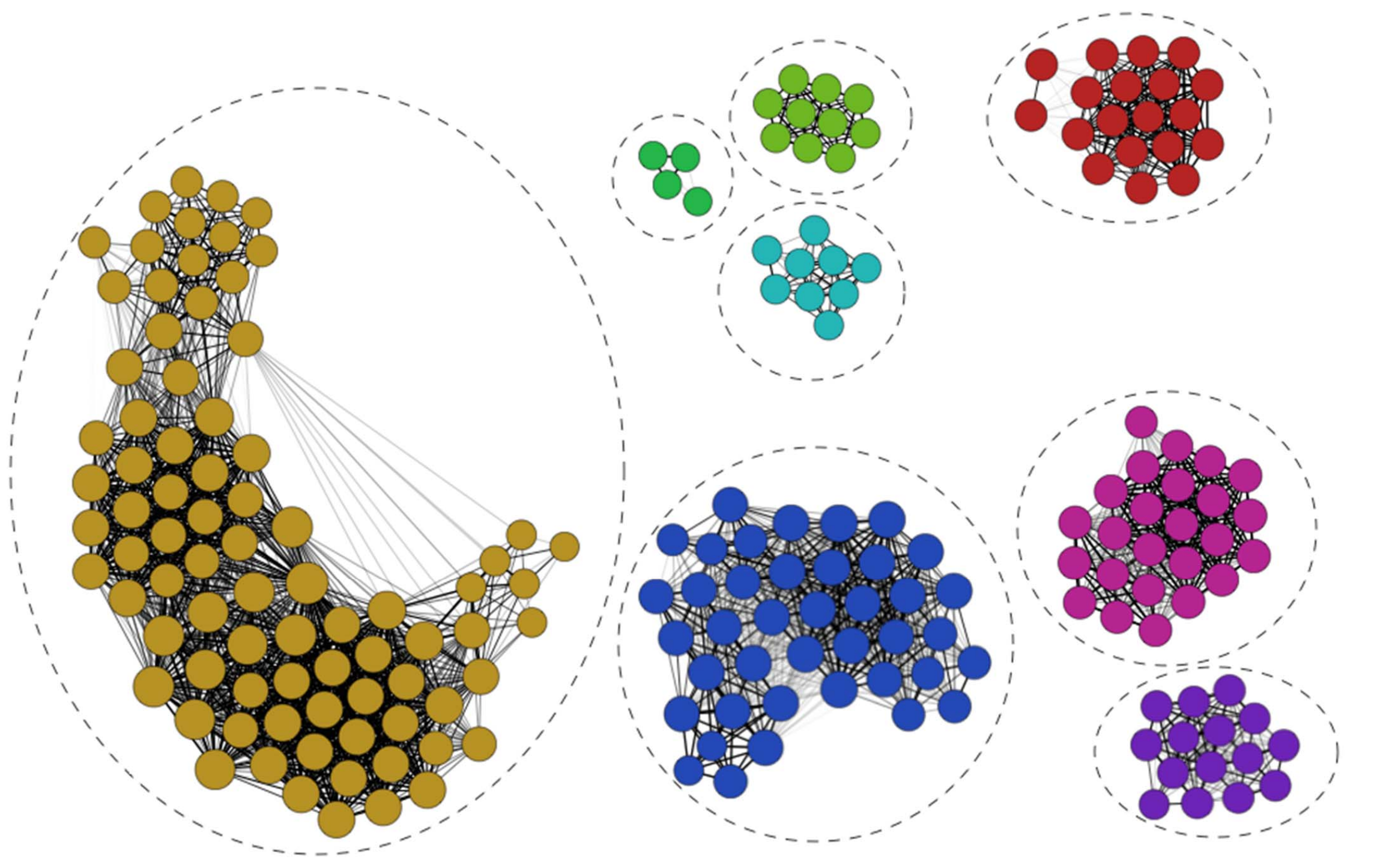

Figure 2. Gene family finding using k-mer similarity on human heart genes. Different node colors indicate transcripts from different genes. Edges connecting nodes indicate a shared k-mer profile above 5% similarity. Dashed circles is the Cogent output partitioning. Seven of the nine genes were perfectly grouped. The remaining two genes, MYH6 and MYH7, share several identical exons and were thus grouped together (blue).

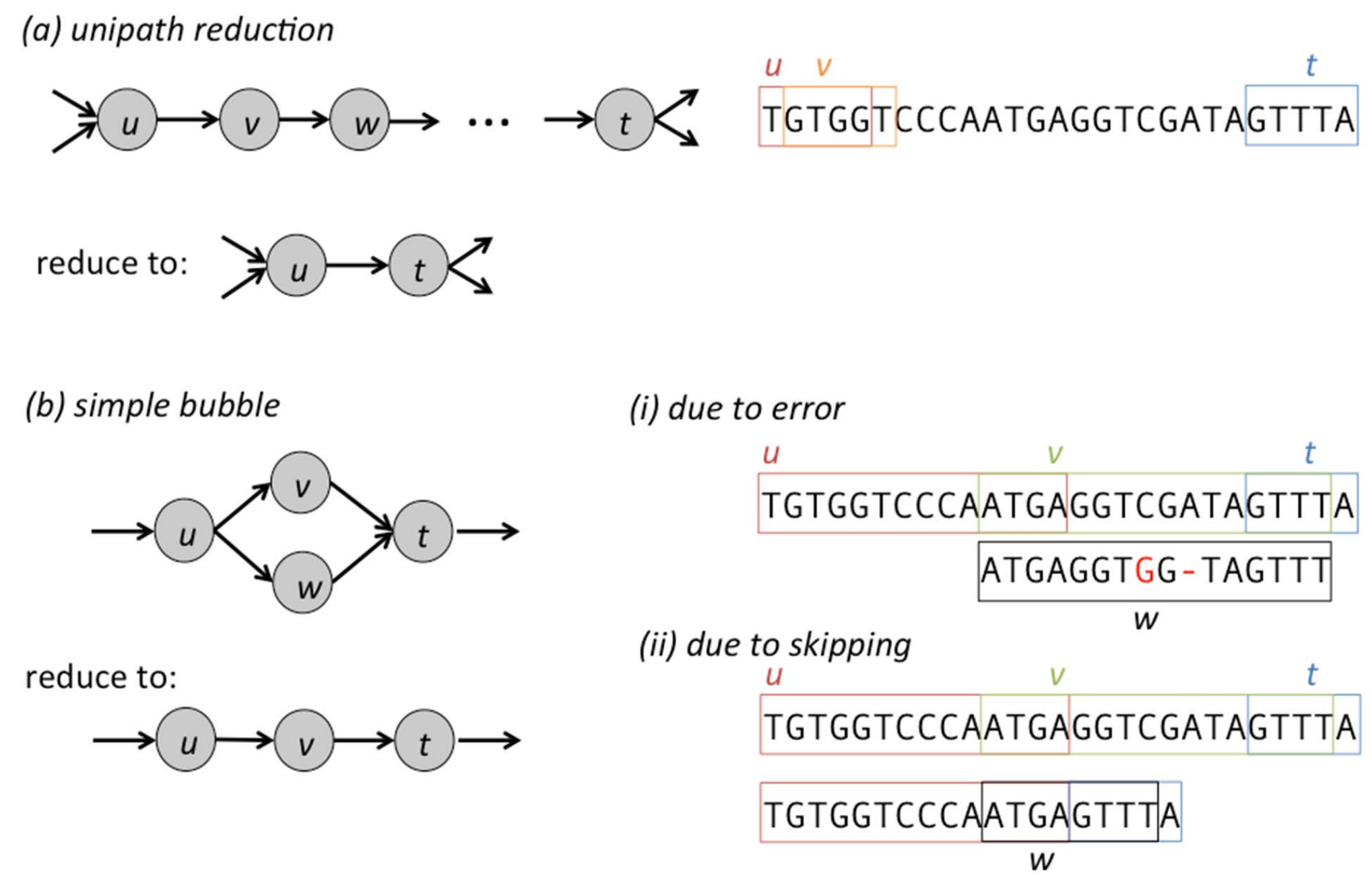

Figure 3. Reducing the de Bruijn graph by collapsing (a) unipaths, which corresponds to transcribed segments shared by all isoforms; and (b) simple bubbles, which can be caused by either errors or exon skipping (or intron retention) events. In the case of errors, either *v* or *w* is removed. In the 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.

Validation Against Known Human Genes

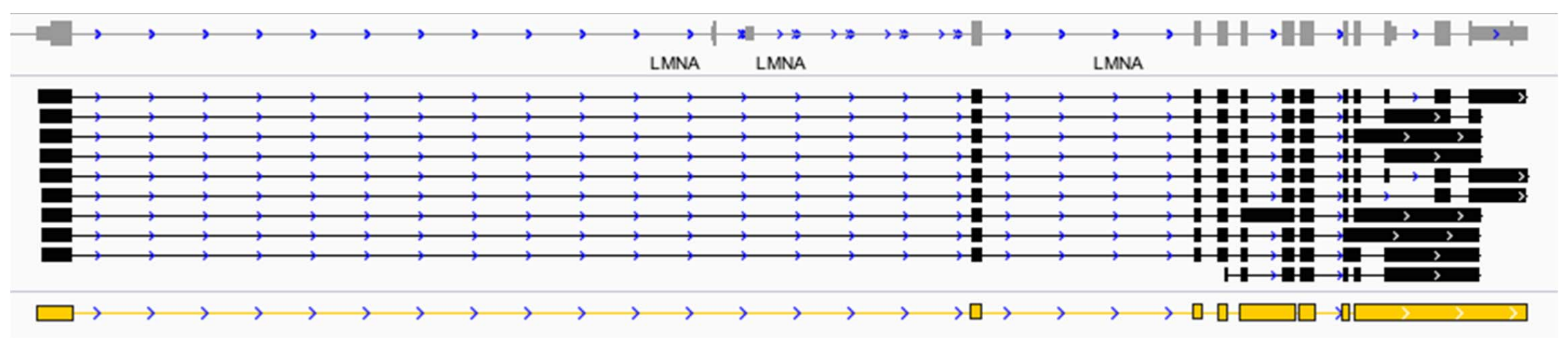

Figure 4. Cogent reconstructs the coding portion (yellow) of the LMNA gene given 10 input isoforms (black). Results were mapped back to the genome to show common introns. Input sequences aligned to hg19 with 98-100% accuracy. The reconstructed contig had an accuracy of 98.6%.

Three of seven gene families were resolved to one contig. The other four families, due to their high splicing complexity, with 15 – 82 isoforms, resolved to between 4 – 11 contigs.

Reconstructed Cuttlefish Coding Contigs Aids in Alternative Splicing and Genome Scaffolding

Of the 24,313 input transcripts, Cogent found 3,542 gene families with 2 or more isoforms (total: 16,860 transcripts).

To fully visualize the 3,542 gene families, Cogent reconstructed 4,938 contigs. Some gene families required more than one contig to represent all isoforms.

The same transcripts were mapped back to the Illumina genome assembly. Of the 16,860 transcripts, 444 failed to align at all, and 4578 mapped to multiple scaffolds (ex: Figure 4). Many only partially mapped (ex: Figure 5)

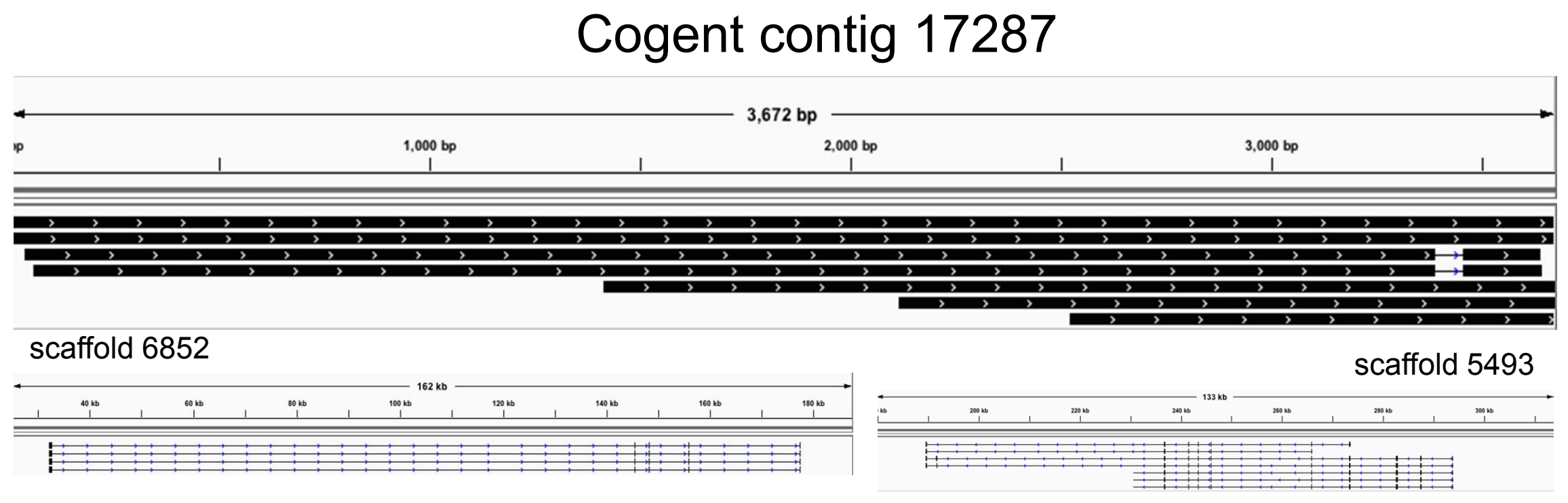

Figure 4. A seven-isoform gene locus visualized via the reconstructed contig (top). Mapping the Cogent contig back to the genome shows that scaffold_6852 should precede scaffold_5493, where the isoforms (blue) map the first and second half. The transcripts partially align to predicted UPF0577 protein.

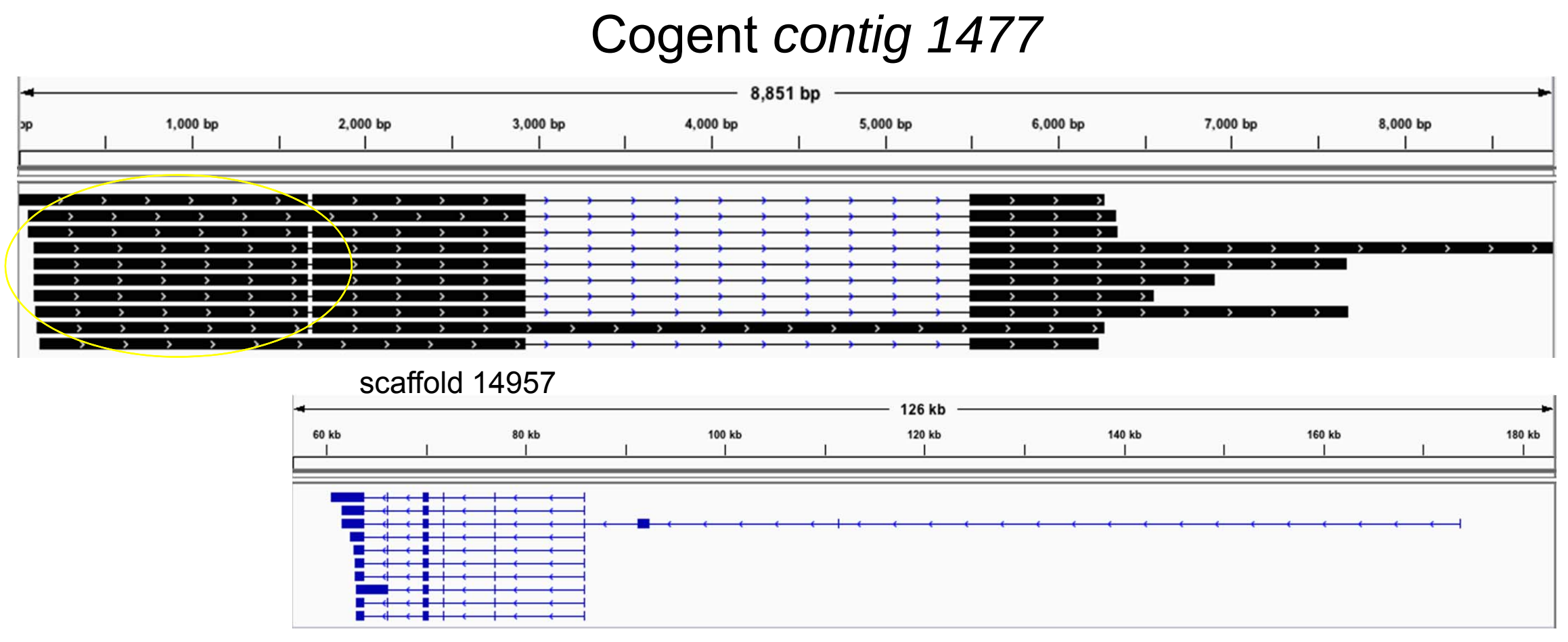

Figure 5. A 10-isoform Cuttlefish gene family reconstructed. (top) The Cogent reconstructed contig. (bottom) The 10 isoforms map to genome scaffold_14957 but is missing the first 1.8 kb (first exon, yellow circle). Pfam search of the transcripts hit the FAM91 domain.

Conclusion

- Cogent does *de novo* coding genome reconstruction using full-length transcripts
- Reconstructed contigs can be used for visualization of alternative splicing and help with genome scaffolding