High Quality Genome Annotation Services Through Dovetail Genomics

Technical Note

Introduction

Virtually all biological analyses can benefit from a high quality, well annotated genome assembly of the species under study. Once a *de novo* assembly of the genome has been built, the typical next step is genome annotation. The genome annotation approach is critical, as it defines the location of genes and their underlying structure (exon/intron boundaries) throughout the genome. The accurate prediction of exon-intron junctions determines translational coding frame and, hence, resulting amino acid composition **(Figure 1)**. Moreover, the quality of many downstream analyses, such as variant detection and differential expression, rely on the accuracy of these predictions. For example, to reduce complexity, time and cost, variant analyses commonly exclude non-coding sequence. As illustrated in **Figure 2**, a potentially crucial variant in the second exon would be missed if gene annotation is erroneous.

The process of accurate genome annotation can be daunting as it is computationally intensive and potentially error prone. A unique gene model for your species must be built from scratch and iteratively refined to improve prediction sensitivity and specificity. As the amount of evidence used in the prediction increases, so do the computational resources needed to convert data into evidence that will increase prediction accuracy (see Step 3 below). Notwithstanding the complexity of data and model preparation, gene prediction is potentially error prone as species-specific gene patterns may not be learned during the training process. In addition, regions of the genome may also lack sufficient evidence to help guide gene prediction and are thus predicted incorrectly. Every genome is unique and, as such, the annotation process must be tailored to the nuances of the genome under study. With this in mind, Dovetail Genomics has developed an accurate, high-throughput, bespoke genome annotation pipeline that takes into account (i) the individual nuances and attributes of your genome and (ii) your specific research goals and genes of interest.

Figure 1 – Gene annotation fundamentals.

(A) *Gene annotation fundamentally describes the location of exons in the reference genome.* ***(B)*** *An example of a three exon gene in the forward frame as shown in the Artemis genome browser[1]. Cyan boxes and lines represent exons and their relationships to eachother. The top and bottom three horizontal tracks represent the forward and reverse translation frames of the sequence respectively. Black vertical bars represent locations of translational stop codons. The bottom blue box shows a blown-up section of the exon-intron junction for exon 1.*

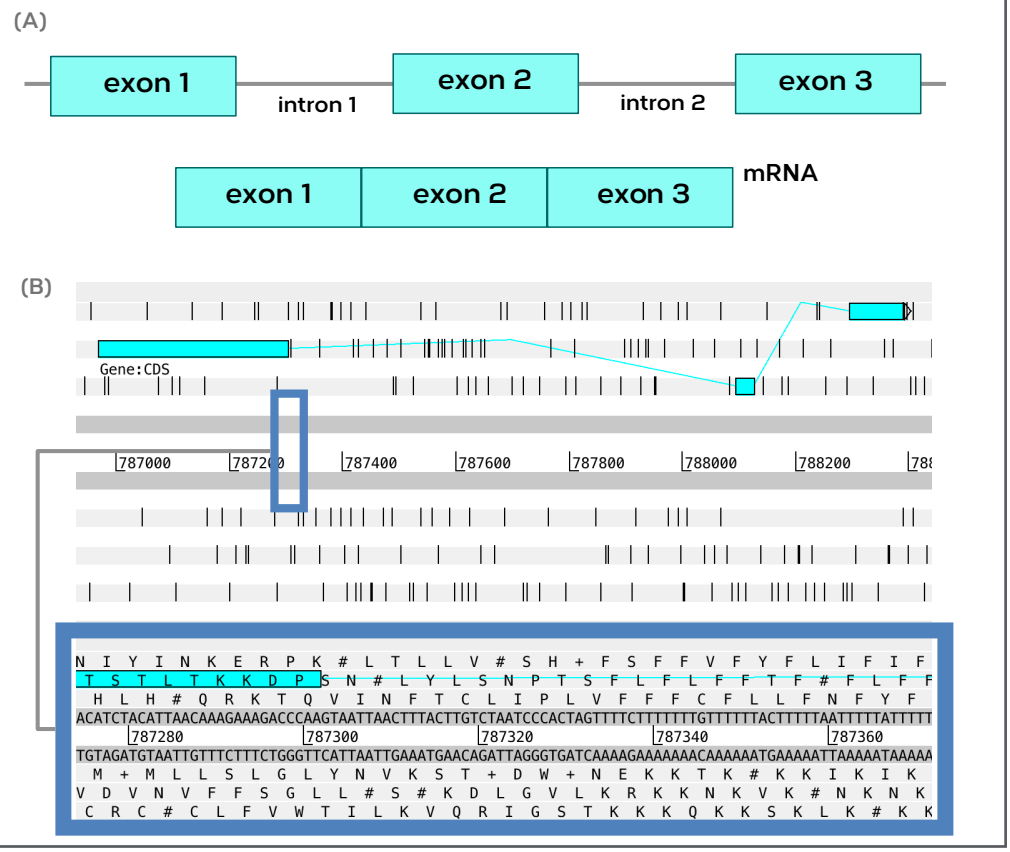


Figure 2 – An accurate annotation depends on a high-quality assembly.

(A) *Manually curated annotation showing aligned RNAseq reads (blue horizontal lines) across the genome. Vertical red lines represent SNPs identified by RNAseq. Cyan boxes represent exons within the negatively transcribed gene. SNPs are identified on the 1st and 2nd exons.*
(B) *Depiction of the in-silico gene annotation prior to manual correction. The second exon was missed by the ab initio annotation process but was identified by a manual inspection of this region of the genome. Figure 2a depicts the correct structure of this gene and the SNP in exon 2 would therefore be included in subsequent analyses.*

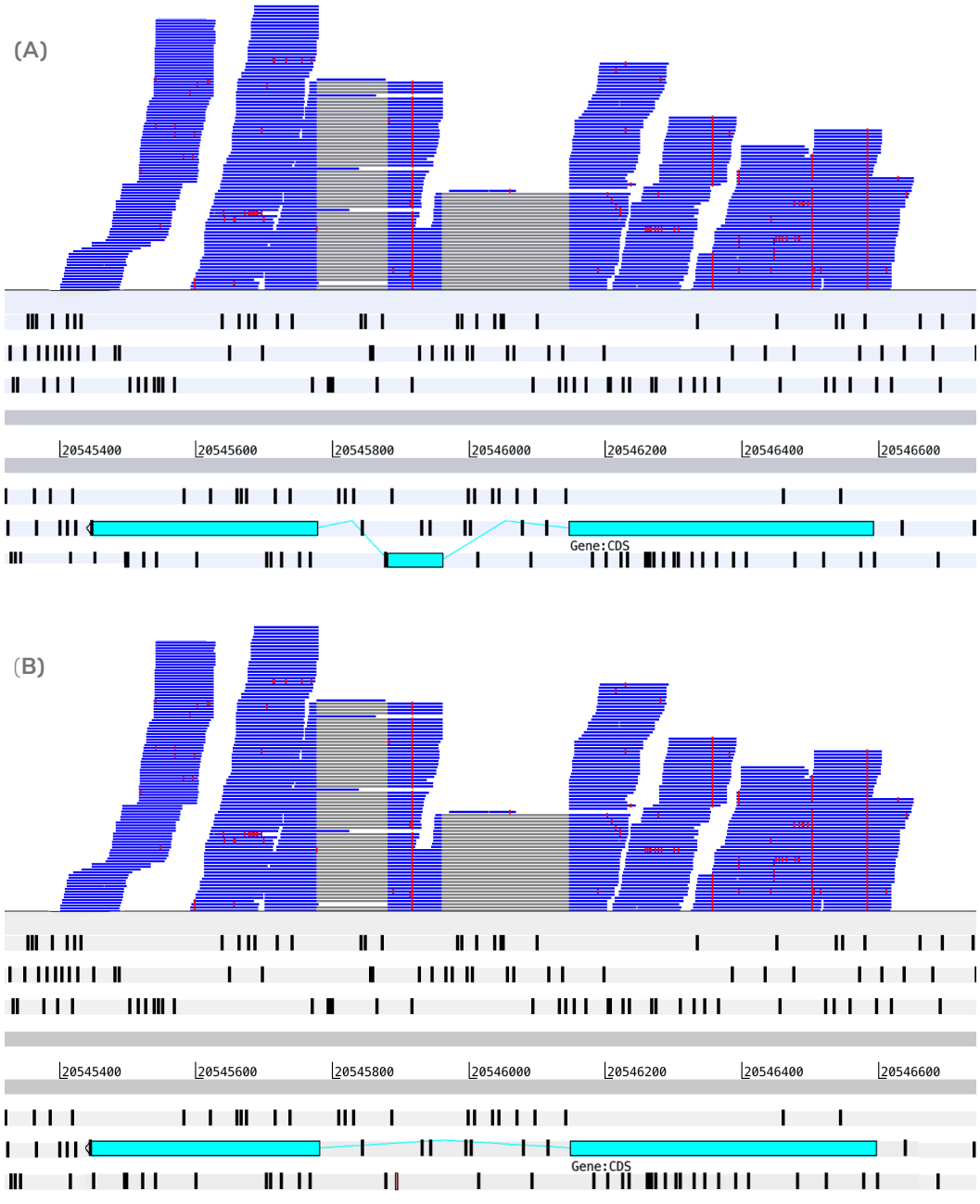
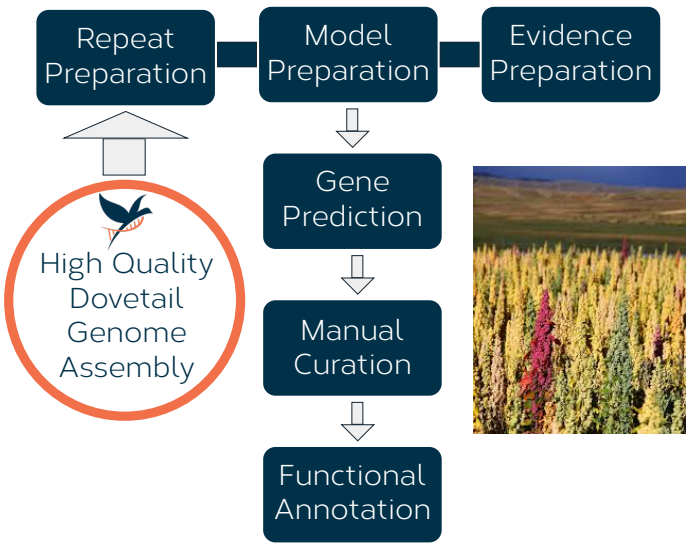


Figure 3 – Dovetail Genomics Genome Annotation Service workflow.



Six Step Process for Accurate Genome Annotation

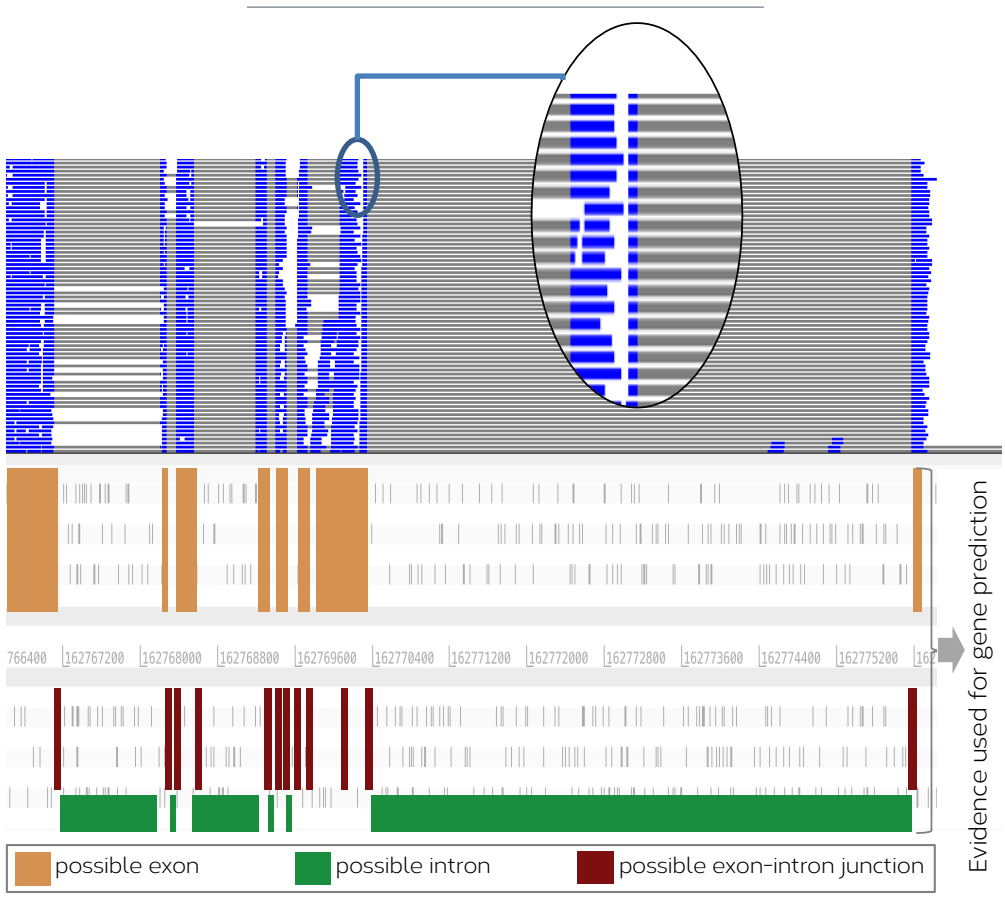
The Dovetail genome annotation pipeline is made up of six distinct steps **(Figure 3)**. A detailed description of the analysis pipeline and methodology can be found in Jarvis *et. al.*[2] .

Step 1: Repeat Preparation

Repetitive sequences can account for up to 80% of a genome. These repeats are problematic for *ab initio* prediction algorithms during genome annotation as they can cause false positive hits. To eliminate this, a species-specific repeat model is created based on your reference genome. This custom-tailored repeat model is then used to mask out repetitive regions. Repeat statistics for each repeat category are provided as part of the publication-ready delivery package.

Figure 4 – Transcriptome evidence provides gene structure evidence.

The extraction of gene structure evidence from RNA-seq read alignment. Locations where RNA-seq reads are mapped provide crucial evidence for possible exon locations. Using a gap aligner during the mapping stage, gaps within the map enables us to determine possible exon-intron junctions. This evidence is later used in the prediction algorithm to help refine the final gene structure.



Step 2: Model Preparation

This is the process of developing a species-specific Hidden Markov Model (HMM) that will describe how genes are encoded in the genome. A dedicated Project Manager will work with you to identify related genomes that are used to develop the model. Protein sequences from related species will be separated equally into training and validation sets. The HMM is then optimized by multiple iterations of training and subsequently used by the gene prediction algorithm in Step 4.

Step 3: Evidence Preparation

This is the process of generating gene structure evidence to improve genome annotation accuracy. This process utilizes transcriptome data (RNA-seq or Iso-Seq) or protein sequences. When projected onto the genome using a gap-aware aligner, the gene structure evidence highlights the location of exons, introns and exon-intron junctions. **(Figure 4)**. These are later used in the gene prediction algorithm to help improve the accuracy of gene boundaries.

When preparing evidence for downstream gene prediction, we recommend our customers prioritize transcript diversity (multiple libraries from multiple tissues and/or cell types) over deeper sequence coverage. A single RNA-seq library representing transcription at a single time point, for a particular tissue and specific biological perturbation event will only capture expression from a subset of genes. By providing Dovetail Genomics with transcriptome samples from a variety of developmental stages, tissues, time points and perturbation events (especially those relevant to your research), a greater number of genes will make it into the final annotated assembly.

Additionally, including transcript evidence from a variety of sequencing technologies can bolster evidence hints. For example, where regions of the genome are not well sequenced or difficult to map using RNA-seq, the availability of full-length transcript sequence (e.g. Iso-Seq; **Figure 5**) can provide additional data that can further refine an accurate gene structure.

Figure 5 – Full-length transcriptome data provides high accuracy gene structure evidence.

*Mapping of RNA-seq data **(A)** and full-length transcript data **(B)** onto the genome. In this example, full-length transcript data is able to provide a better resolution of exons, introns and exon-intron junctions when compared to RNA-seq data.*

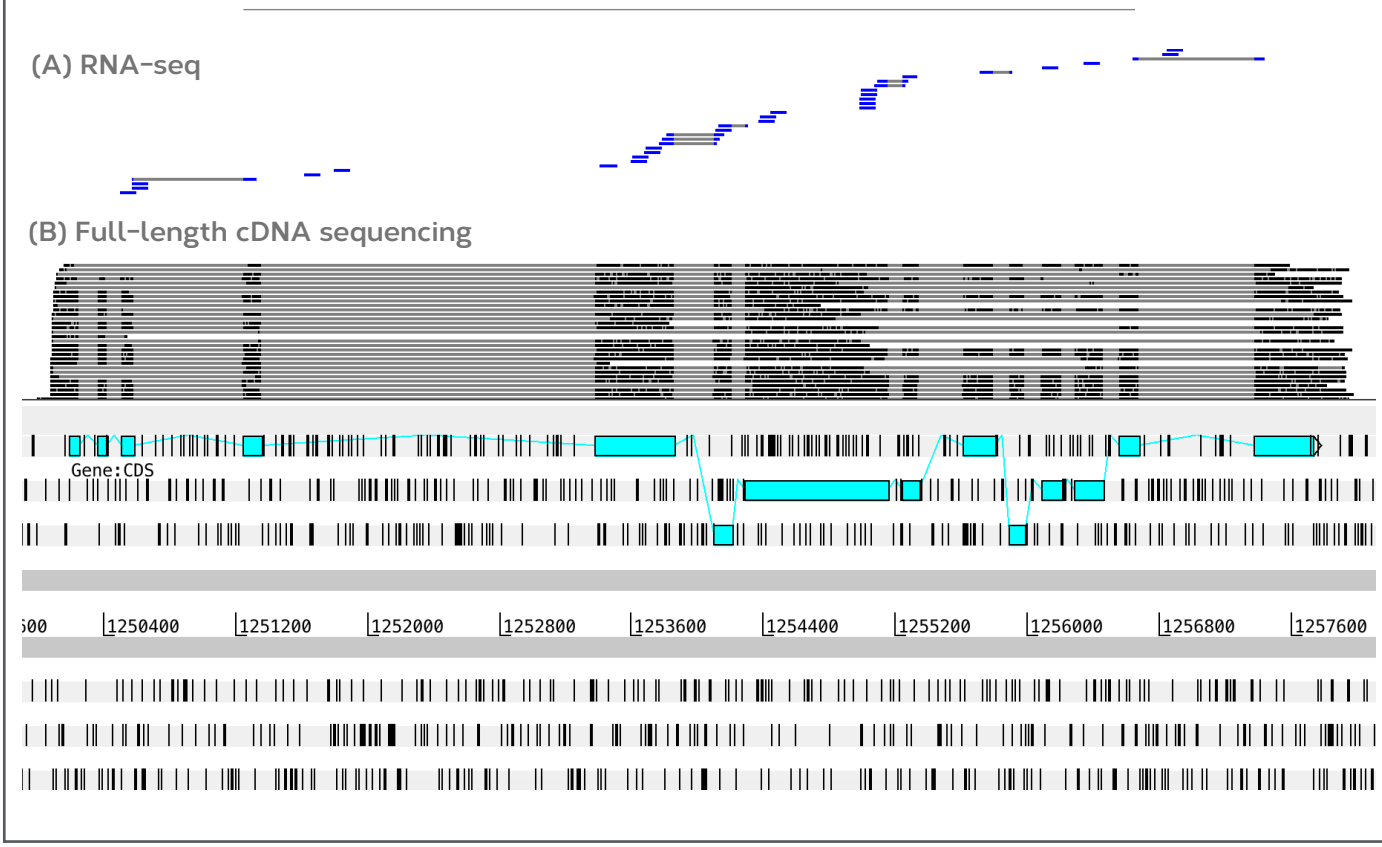
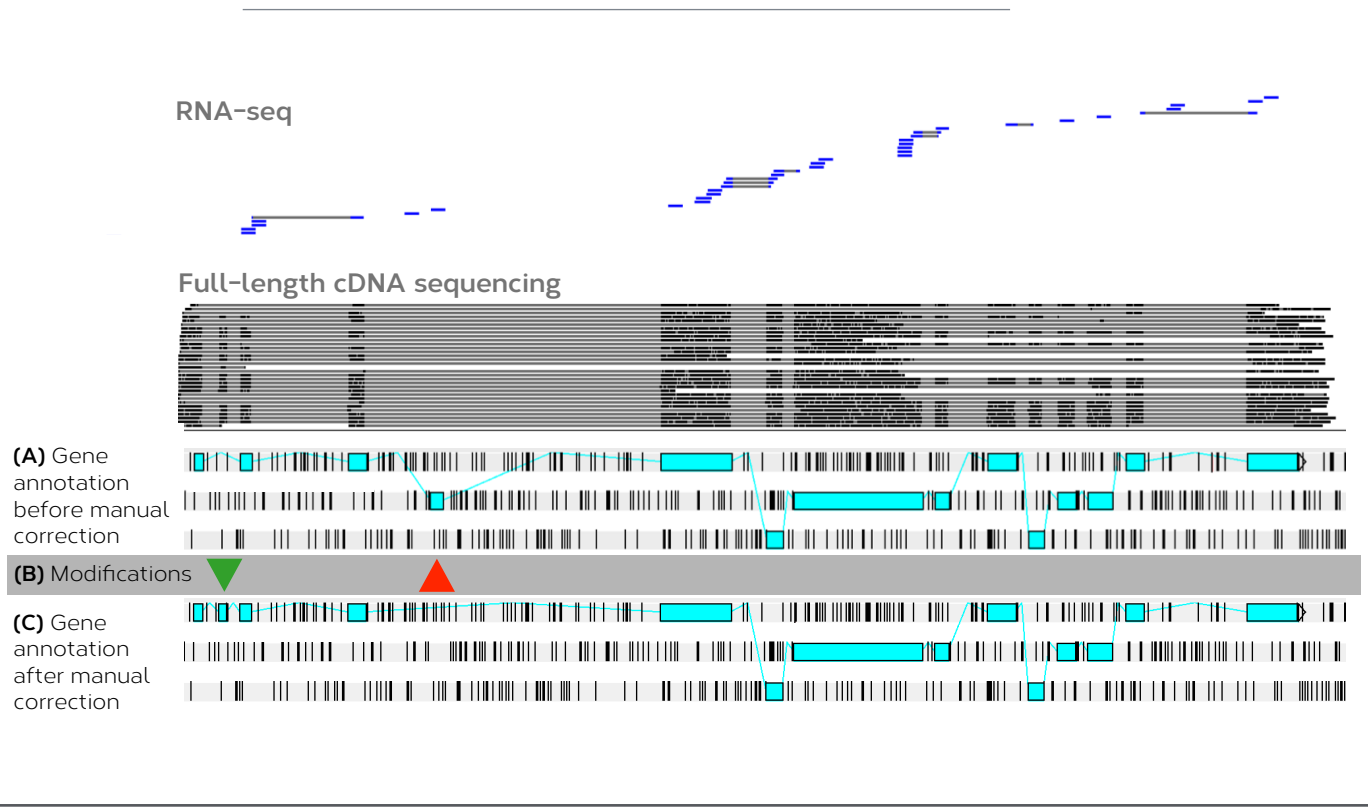


Figure 6 – Manual curation corrects annotation errors.

Illustration of manual correction performed on an erroneous ab initio predicted gene structure by leveraging transcript evidence. ***(A)*** *Initial gene structure as predicted from the ab initio pipeline. Blue cyan boxes represent exons.* ***(B)*** *Modifications to the initial gene prediction based on full length cDNA evidence. The green and red arrows represent exons that were added or removed, respectively.* ***(C)*** *The final gene structure following manual curation showing a higher level of concordance with the transcript evidence.*



Step 4: High Throughput Genome Annotation

This process uses the inputs from the first three steps to predict gene structures genome wide. A genome annotation project can take months to complete using a moderately sized high-performance cluster. The proprietary Dovetail Genomics workflow scales up computational resources to thousands of CPUs/Gbase when needed, significantly reducing overall project turnaround time.

Step 5: Manual Curation

This is an important step as it ensures that the structure of genes of specific interest to your research are verified and corrected, if necessary. Since algorithmic genome annotation is never 100% accurate, we have developed a manual curation process that is applied to a select list of genes most relevant to your research. In the example displayed in **Figure 6**, an exon was missed and an exon erroneously inserted by the *ab initio* pipeline. Using full-length transcript evidence as a guide, the Dovetail Genomics manual curation process was able to correct this error, reflecting the gene's true structure.

Step 6: Functional Annotation

This is the final step in the Dovetail Genomics workflow. Gene nucleotide sequences are queried against current public databases using the BLAST[3] algorithm and gene function descriptions are transferred onto the assembly.

Summary

Accurate genome annotation is an essential step in producing a publishable genome assembly. The Dovetail Genomics Genome Annotation Service provides researchers with a fast, reliable and proven workflow for highly accurate annotation fully tailored to the unique attributes of the genome under study.

Genome Annotation Deliverables

Contents of the Delivery Package

A comprehensive report (see "Sample Annotation Delivery Report" for example) including:

- Genome annotation file in GFF3 format
- Predicted gene CDS sequences in FASTA format
- Predicted gene peptide sequences in FASTA format
- Alignment files from RNA-seq and Iso-Seq data in BAM format (if evidence data is provided)
- Repeat annotation file in GFF3 format
- Visualization of structures of manually curated genes

References

[1]Carver, Tim, *et al.* Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28.4**, 464-469 (2011)

[2]Jarvis, David E., *et al.* The genome of Chenopodium quinoa.*Nature* **542.7641**, 307 (2017)

[3]Lipman, D.J. and Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* **227**, 1435-1441. (1985)