Application Note

Improving structural annotation in complex genomes with QIAGEN® CLC Genomics Workbench

This Application Note describes how annotation tools from QIAGEN CLC Genomics Workbench, such as *ab initio* Transcript Discovery and Annotate with DIAMOND, can improve the structural annotation of genomes.

Introduction

Creating comprehensive gene annotation before using a new genome assembly for downstream experiments is fundamental. However, many assembled and annotated genomes may have significant shortcomings that affect the performance of laboratory assays: The genes of interest could be incorrectly annotated or not annotated at all. Here, we show how to complement and add annotation before designing functional or genotyping assays. The focus of this Application Note is on annotating plant disease resistance genes (R-genes), as they are often mis-annotated in plant genomes. R-genes may contain repetitive elements and are often filtered out by annotation pipelines that mask those repeats. We tested the completeness of R-gene annotation in the recently published genome of alfalfa.

Results

Transcript Discovery is a free plugin available with QIAGEN CLC Genomics Workbench. This tool produced more comprehensive structural gene annotation than the previously published annotation (Chen et al., 2020). The Annotate with DIAMOND tool, which is part of QIAGEN CLC Microbial Genomics Module and available with QIAGEN CLC Genomics Workbench Premium, located all R-genes when using nucleotide binding sequence (NBS) motif sequences as the protein reference. We identified 100 expressed R-genes which were not previously annotated. The red box in Figure 1 is an example of one such finding. For the selected NBS annotation, there were no annotations in the published genome; however, the Transcript Discovery tool produced the gene, transcript and coding sequence (CDS) annotations shown below. The NBS annotation was added by the Annotate with DIAMOND annotation plugin.

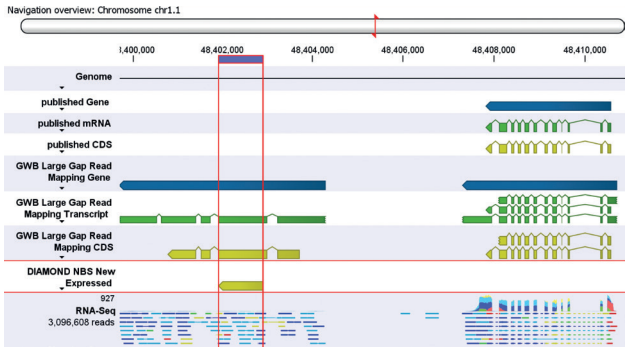


Figure 1. Navigation of the genome, annotations and mapping tracks in QIAGEN CLC Genomics Workbench.

Data

The genome assembly used here is an allele-aware chromosome-level genome assembly for cultivated alfalfa (Chen et al, 2020). The assembly and annotation were downloaded from **https://figshare.com/projects/whole_genome_sequencing_and_assembly_of_Medicago_sativa/66380** and imported to QIAGEN CLC Genomics Workbench using the tracks importer.

For *ab initio* annotation using the Transcript Discovery plugin, we downloaded the pooled alfalfa mRNA-seq sample data from **https://www.ncbi.nlm.nih.gov/sra/SRX5804124**. This Illumina reads file is a part of the dataset from Chen et al., 2020. The file was imported to QIAGEN CLC Genomics Workbench using the Illumina importer.

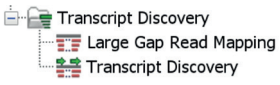
To identify R-genes using the Annotate with DIAMOND tool, we used two NBS subdomains: NBS_SubDomA and NBS_SubDomB. These protein sequence motifs were derived from multiple alignments of NBS domains from *Arabidopsis* disease resistance genes (Meyers et al, 2003). For the NBS subdomains, we created a sequence file containing both motifs from **https://niblrrs.ucdavis.edu/At_RGenes/HMM_Model/HMM_Model_NBS_Ath.html**. The genome was compared with these sequences, as the NBS domains are an essential part of every plant disease resistance gene.

Genome assembly data

Most types of genomic data can be easily imported to QIAGEN CLC Genomics Workbench. The assembly was imported from a *fasta.gz* file and the annotation from a *gff.gz* file. The total assembly contains 2738 Gb in 32 super-scaffolds and 419 Mb of unplaced unitigs. This is an allele-aware chromosome-level assembly of an autotetraploid species. The annotation includes 164,632 genes, with one transcript and one CDS per gene. This is a publicly available preliminary annotation that will require improvement because many eukaryotic genes contain multiple transcripts producing multiple proteins.

Annotation with the Transcript Discovery plugin

This mapping tool does not require existing annotation and can stretch the sequencing reads over introns. We used the Large Gap Read Mapping tool to align 96 million paired-end reads of 150 nt (from the pooled mRNA sample) to the reference genome.



The output of Large Gap Read Mapping is then used by the Transcript Discovery plugin, which relies heavily on reads mapped with a gap as evidence for spliced transcripts. The tool produces a set of gene, transcript and CDS annotations based on gene expression data. Here, the tool annotated 173,589 genes, 274,031 transcripts and 114,054 CDS. An example annotation created by these two tools is shown in Figure 2, in which the shown gene contains two exons, three transcripts and two CDS. The upper track in Figure 2 is output from Large Gap Read Mapping, and the three lower tracks are the structural annotations produced by the Transcript Discovery plugin.

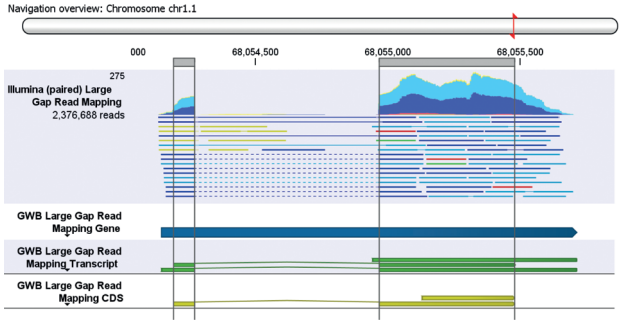
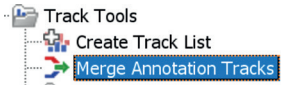


Figure 2. *Ab initio* annotations produced by QIAGEN CLC Genomics Workbench tools.



This analysis missed some published annotations, most likely due to the absence of transcription in the tissues used for the mRNA-seq pooled library. A combination of various annotation approaches would produce the most comprehensive annotation. The Merge Annotation Tracks tool produced 263,684 CDS annotations after combining both the published and QIAGEN CLC Genomics Workbench CDS tracks into one track. Some of the published annotations do not appear accurately when placed together with the output from the Large Gap Read Mapping tool (Figure 3), as the first exon is apparently truncated in the displayed gene.

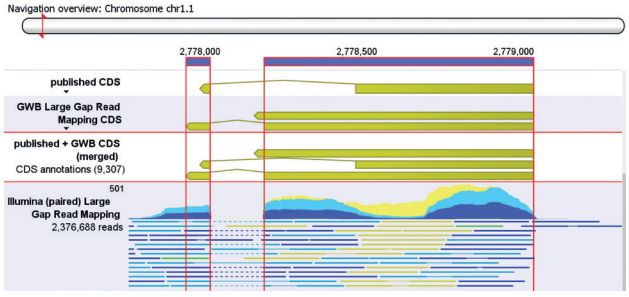
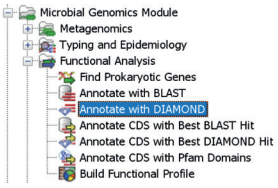


Figure 3. Visualization of the published CDS track (top), *ab initio* created QIAGEN CLC Genomics Workbench annotation (2 CDS) and the merged track with three CDS. The bottom track is the Read Mapping support for the QIAGEN CLC Genomics Workbench predicted transcripts.

Annotation with DIAMOND hits

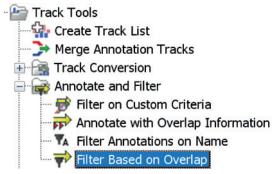


R-genes can be difficult to annotate using automated pipelines, as they are often filtered out by repeat-masking pipelines. However, each R-gene contains an NBS motif, which we utilized for annotation with DIAMOND. The Annotate with DIAMOND plugin annotates a DNA sequence using a set of known protein reference sequences. This tool can be used on genomic sequences without pre-existing annotation. Even though DIAMOND tools are part of QIAGEN CLC Microbial Genomics Module, they can also be used for large genomes.

After running the Annotate with DIAMOND tool we introduced 478 NBS domain annotations, and 249 of these did not overlap with the annotations supplied for this genome. A comparison of annotations can be easily performed using the Filter Based on Overlap tool in the Track Tools folder.

When the novel NBS annotations were compared with the Large Gap RNA-seq Read Mapping track using the Filter Based on Overlap tool, we detected 100 NBS domains with expression support (Figure 1 shows an example).

Figure 4 shows an example of an R-gene CDS that was truncated just before the duplicated area in the gene. Although it was partially detected by the QIAGEN CLC Genomics Workbench transcript discovery tools, the gene could be further properly annotated when the DIAMOND data was included.



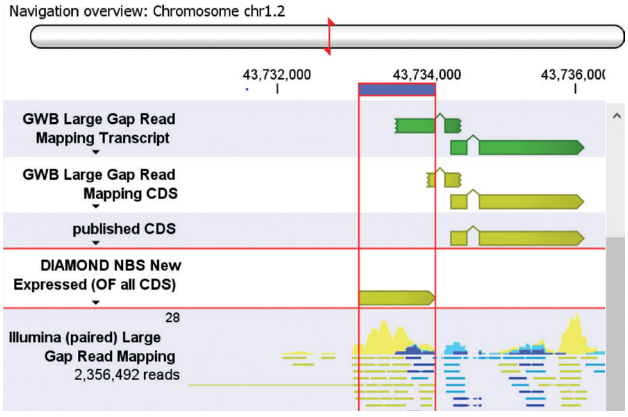


Figure 4. Most annotation tools struggle to properly annotate repeats containing CDS. The yellow reads in the lowest track are non-specifically placed matches on duplicated regions.

Summary

When working with genomic assemblies, a combination of annotation tools should be used for the production and refinement of gene models. QIAGEN CLC Genomics Workbench offers multiple annotation tools, two of which are described here: *Ab initio* Transcript Discovery and Annotate with DIAMOND. These comprehensive, easy-to-use tools enable biologists to take the newly assembled genomes back to the laboratory for functional assays.

References

1. Chen, H., et al. (2020) Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat Commun **11**, 2494.
2. Meyers, B.C., et al. (2003). Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. The Plant Cell **15**, 809.

QIAGEN CLC Genomics products are intended for molecular biology applications. These products are not intended for the diagnosis, prevention or treatment of a disease.

Get the right bioinformatics analysis tool for your research needs. Contact us to find the right solution for your research. **bioinformaticssales@qiagen.com**

Learn more and request a consultation at **digitalinsights.qiagen.com/CLC**

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN CLC Genomics product site. Further information can be requested from **ts-bioinformatics@qiagen.com** or by contacting your local account manager.

Ordering **www.qiagen.com/shop/analytics-software** | Technical Support **digitalinsights.qiagen.com/support**
Website **digitalinsights.qiagen.com**