White Paper

A strategy for evaluating genomic assemblies in QIAGEN® CLC Genomics Workbench

Introduction

Over the last decade, progress in de novo assembly and scaffolding tools has enabled the production of high-quality genomic assemblies, even for complex non-model species. Improvements in long-read technologies, such as PacBio® and Oxford Nanopore®, along with DNA crosslinking technologies, have facilitated large genome, chromosome-scale assemblies. However, in most cases, the assemblies produced by automated pipelines still require manual adjustments before performing downstream analyses.

Genomic assemblies are fundamental constructs that provide the basis for various research and practical applications. Designing a functional assay or genotyping markers with incorrectly assembled or erroneously annotated genes can be costly and time consuming and can generate misleading results. Fortunately, QIAGEN CLC Genomics Workbench provides analysis and visualization tools that help evaluate genomes in a fast and efficient manner. Here we describe a workflow that can identify misassembled areas in annotated regions.

Genome assembly and annotation data

The recently published genome of the cultivated alfalfa plant is an example of a high-quality assembly (1). It consists of 32 allelic chromosomes and was assembled using Illumina reads, Hi-C data, and high-fidelity long reads. A combination of annotation strategies yielded a total of 164,632 protein-coding genes in this allele-aware chromosome-level genome (2n = 4×; 4 alleles for each of the 8 chromosomes).

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, version 20.0.4, using the tracks importer (Figure 1). The genome was imported from a **fasta.gz** file and the annotation from a **gff.gz** file. The total assembly contains 2.738 Gb in 32 super-scaffolds and 419 Mb of unplaced unitigs. The annotation includes 164,632 genes that have only one transcript and one coding sequence (CDS) per gene. Some eukaryotic genes produce multiple transcripts that produce multiple proteins. In the "Improving structural annotation in complex genomes with QIAGEN CLC Genomics Workbench" Application Note, we showed how to improve this preliminary annotation using the annotation tools available in the QIAGEN CLC Genomics Workbench.

Sequencing data

To identify the misassembled areas, we used the following from the dataset used in (1):

- Illumina WGS reads (www.ncbi.nlm.nih.gov/sra/SRR9026574)
- Illumina RNA-seq reads (www.ncbi.nlm.nih.gov/sra/SRX5804124)
- PacBio reads (www.ncbi.nlm.nih.gov/sra/?term=SRR11285798)

We imported the reads into the QIAGEN CLC Genomics Workbench using the corresponding options ('Illumina…', 'PacBio…') in the 'Import' menu (Figure 1). Figure 2 gives an overview of the process.

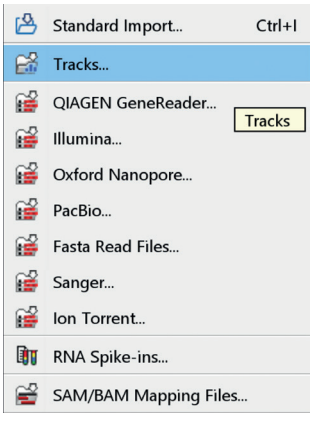

Figure 1. The 'Import' menu of the QIAGEN CLC Genomics Workbench.

▷

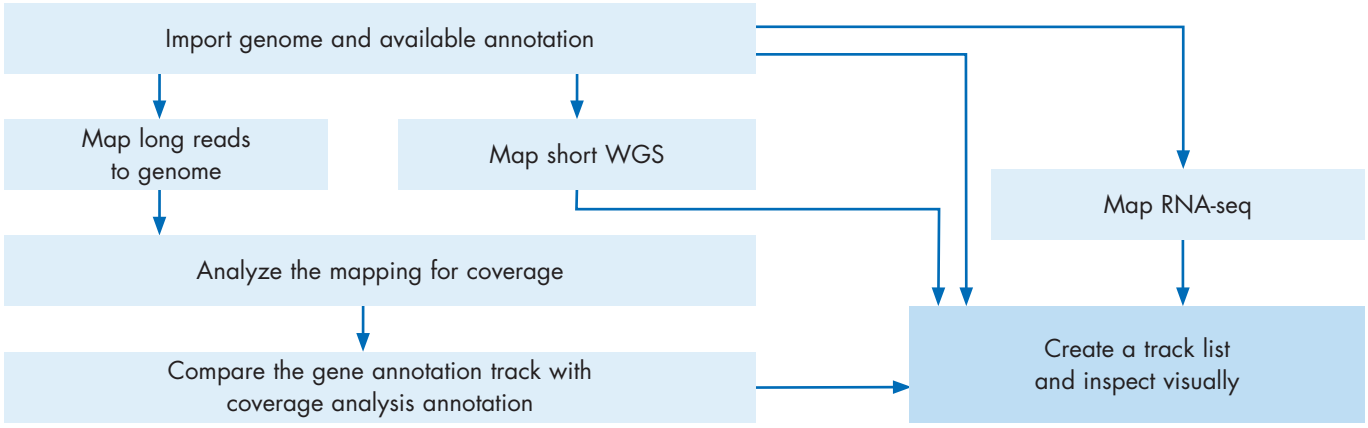

Figure 2. Overview of genome assembly analysis.

Identification of misassembled areas using the QIAGEN CLC Genome Finishing Module

The QIAGEN CLC Genome Finishing Module is a plugin that extends the capabilities of the QIAGEN CLC Genomics Workbench (Figure 3). The module contains multiple tools for assembly analysis and refinement. We used the Analyze Contigs tool to identify misassembled areas in the alfalfa genome. This tool analyses mappings and identifies problematic regions that need further attention. Identified events, such as unaligned read ends or regions with low or high coverage, are annotated and presented in a table. These tables can be converted to tracks which can be compared with other annotation types in the tracks format.

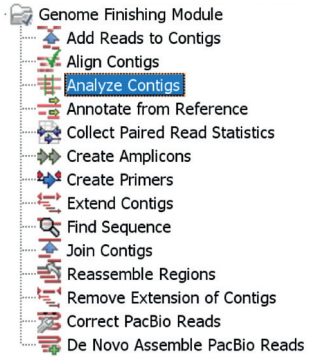

Figure 3. The Analyze Contigs tool in the QIAGEN CLC Genome Finishing Module.

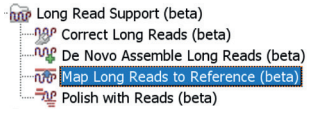

Figure 4. The Map Long Reads to Reference tool in the Long Read Support plugin.

First we needed to create a mapping for the Analyze Contigs tool, and so we align the PacBio reads to the alfalfa genome using the Map Long Reads to Reference tool of the Long Read Support plugin (Figure 4).

From the multiple options for the detection of alignment problems in the Analyze Contigs tool, we chose 'Detect unaligned read ends'. Unaligned read ends are a good indicator of misassembly. When using this tool on large genomes, it is not advisable to select multiple options at once because the processing can be lengthy. In this case, the tool detected 2,124 positions in the genome that contained unaligned read ends (with a minimum of 3 mapped reads and a minimum of 50% reads with unaligned ends).

Identification of misassembled genes

The output of the Analyze Contigs tool is a standalone mapping annotated with misassembly features. For the next analysis steps, this "unaligned ends" annotation was converted to a track using the Convert to Tracks tool (Figure 5). We then used the Filter Based on Overlap tool (Figure 5) to compare the "gene" annotation track, which had been imported earlier (Figure 1), to the "unaligned ends" annotation track. We filtered only for the genes that overlapped with "unaligned ends" annotation and created a new track: "Genes with partially aligned PacBio reads". This track contained only six genes.

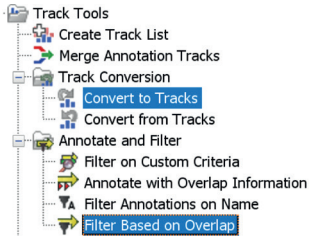

Figure 5. The 'Track Tools' folder provides access to tools for track conversion and filtering.

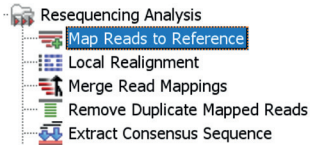

Figure 6. 'Resequencing Analysis' folder with the read mapping tool selected.

We also completed the same workflow using the Analyze Contigs tool with 'Detect low coverage' enabled and 'Low coverage threshold' set to 0. These settings allowed us to identify 14 genes with no coverage. These coverage annotation tracks were later combined into lists of tracks, which were used to visually inspect the misassembled areas.

Mapping Illumina reads

The short read mapping tracks served as supporting data to evaluate the assembly.

WGS Illumina reads were mapped using the Map Reads to Reference tool in the 'Resequencing Analysis' folder (Figure 6). The track of mapped reads can be combined with other tracks originating from the same reference genome. The Map Reads to Reference tool can also be used to map high-quality long reads; however, it is not suitable for error-prone long reads (or reads longer than 99,999 bp). For such reads, the Map Long Reads to Reference tool from the Long Read Support plugin can be used instead (see Figure 4).

To inspect the assembly in the transcribed area, the expression reads were also mapped. RNA-seq reads were mapped using the RNA-Seq Analysis tool (Figure 7). This tool requires the gene and transcript annotation tracks.

Data visualization and navigation of annotation

After completing the above, we had all the necessary data to visualize misassembled regions.

The lists of tracks was created with the Create Track List tool (Figure 5, the first tool in the folder).

The track browser (Track List) allows simultaneous visualization of various experiments and annotations (Figure 8). Starting from the top of the picture, this figure shows the following:

1. The position of the region of the expanded view on chromosome 8.3 (the red double-headed arrow)
2. Transcript annotations (published mRNA) provided with the genome
3. The mapped RNA-seq reads created by the RNA-Seq Analysis tool
4. Unaligned-end annotations created by the Analyze Contigs tool
5. Genes that overlap with the "unaligned ends" annotations created by the Filter Based on Overlap tool
6. PacBio reads mapped by the long read mapper
7. Illumina whole-genome-sequencing reads

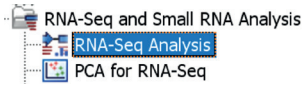

Figure 7. 'RNA-Seq Analysis, the RNA-seq mapping tool.

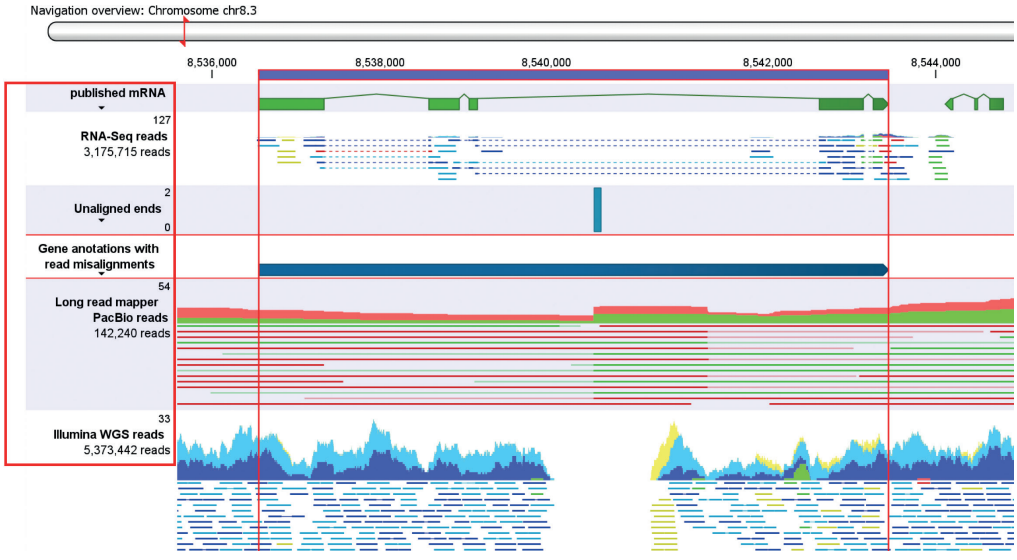

Figure 8. The 'Track List' data browser.

Chromosome	Region	Name
chr1.4	78439534..78445813	MS.gene41424 (merged)
chr7.2	14720044..14728952	MS.gene40406 (merged)
chr8.3	8536543..8543454	MS.gene036588
chr8.3	8536543..8543798	MS.gene036588
22567	complement(7926..9756)	MS.gene001548 (merged)
36327	165..2230	MS.gene001570 (merged)

Figure 9. Table view of the "Gene annotations with read misalignments" annotation track.

The data in the browser can be navigated by opening the relevant annotation tracks in a table view (Figure 9). Double-clicking the track name opens the annotation for the track. The "Gene annotations with read misalignments" track is the result of our analysis. Selecting a row in the annotation table (Figure 9) zooms to the annotation's position in the Track List viewer (Figure 8).

A visual inspection (Figure 10) of the selected MS.gene036588 uncovered the following:

1. Incomplete annotation, as the RNA-seq reads extended beyond the gene and transcript annotation
2. Imperfect alignment of long PacBio reads in the misassembled area
3. Absence of Illumina reads in the misassembled area

We also identified regions with 14 annotated genes that are not supported by long reads nor specific Illumina reads. Figure 11 provides such an example, MS.gene060131. A significant portion of this gene is not covered by PacBio reads and the short Illumina reads aligned to the area are not specific. In the viewer, non-specific reads are indicated in yellow. These short

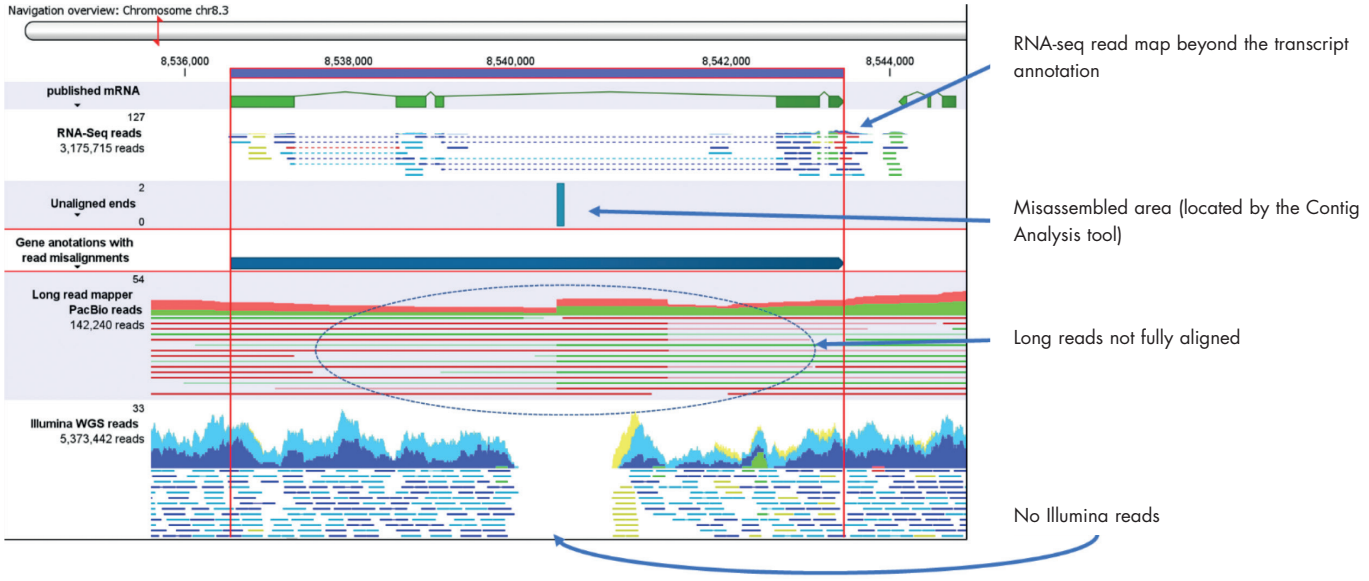

Figure 10. Data browser showing misassembly and annotation problems in one of the genes.

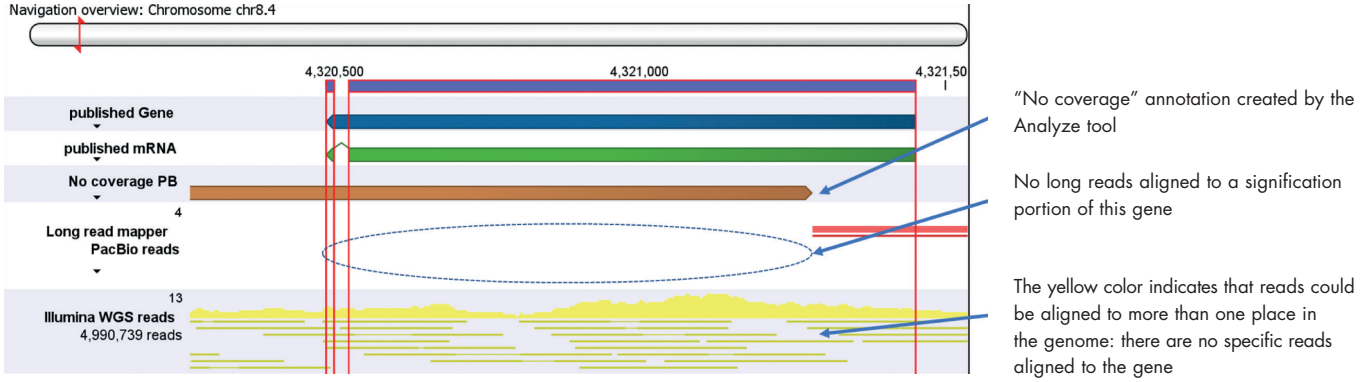

Figure 11. Data browser showing a gene area for which support is not substantial.

reads match equally well elsewhere in the mapping. Thus, none of the read tracks provide clear evidence for the existence of this region.

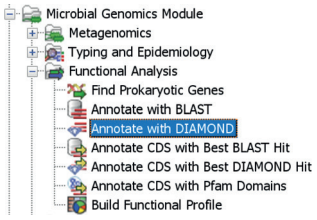

Figure 12. The Annotate with DIAMOND tool in the QIAGEN CLC Microbial Genomics Module.

Sequence motifs or genes of interest can be easily assessed for assembly quality in a similar manner. Our interest was in the quality of assembly in and around the nucleotide binding sequence (NBS) domains of disease resistance genes (R-genes). Whereas we found that many of the NBS domains were misannotated or not annotated, the regions in which they occurred were assembled correctly. The NBS annotation track was created with Annotate with DIAMOND tool using the consensus motifs from https://niblrrs.ucdavis.edu/At_RGenes/HMM_Model/HMM_Model_NBS_Ath.html (2). The Annotate with DIAMOND tool annotates a DNA sequence using a set of known protein reference sequences. This tool can be used on genomic sequences without pre-existing annotation. Annotate with DIAMOND tools are part of the QIAGEN CLC Microbial Genomics Module (Figure 12), but they can also be used for large genomes. This tool produced a CDS annotation track with 478 NBS domain annotations, for which none of the annotations overlapped with the annotations produced by the Analyze Contigs tool (Figure 3). This annotation indicated that no assembly problems were detected. We compared the tracks using the steps described earlier, with the Filter Based on Overlap tool (Figure 5).

All NBS annotations created with the Annotate with DIAMOND tool were supported by specific long reads and specific or non-specific Illumina reads (Figure 13). The example below is an R-gene that was misannotated but correctly assembled. The area that contains the NBS domain (yellow bar) was annotated as an intron (in the green transcript), but it is covered by specific RNA-seq reads. The entire gene is covered by the specific long reads (red and green) and specific paired short reads (blue).

Summary

It is advisable to conduct a rigorous assembly analysis before using an assembled genome in downstream laboratory experiments. The time and reagents spent on erroneously assembled genes can be costly and misleading. The tools in the QIAGEN CLC Genomics Workbench can be used to evaluate genomic assemblies in a fast and efficient manner. These comprehensive, easy-to-use tools enable biologists to detect assembly problems before taking newly assembled genomes to the laboratory for functional assays. The workflow we described can identify and visualize misassembled areas in annotated regions.

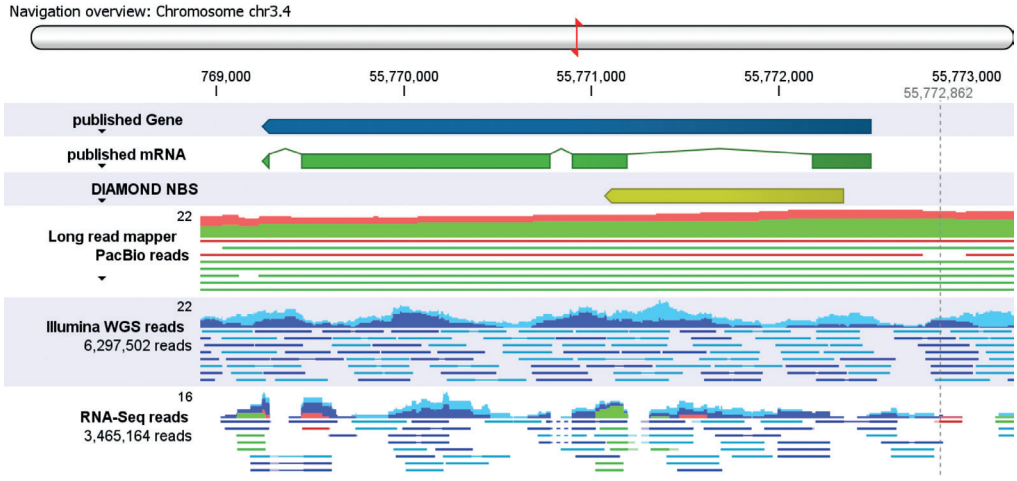

Figure 13. Data browser showing an R-gene with an NBS domain and no apparent problems with assembly.

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. https://doi.org/10.1038/s41467-020-16338-x.

2. Meyers, B., et al. (2003). Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. The Plant Cell **15**, 809. https://doi.org/10.1105/tpc.009308.

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

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.

Contact us to find the right solution for your research needs at **bioinformaticssales@qiagen.com**.

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

Ordering **qiagen.com/bioinformatics** Technical Support **digitalinsights.qiagen.com/support** Website **digitalinsights.qiagen.com/**

1123337 04/2021