Revealing the structure of a biosynthetic gene cluster in a barley variety using Xdrop® and nanopore sequencing

Background

The 5.3-Gb genome of barley (*Hordeum vulgare*) includes several gene clusters involved in the biosynthesis of natural products. Extensive sequencing efforts have generated a barley reference genome and a pan genome,[1,2] and it is known to contain a large proportion (~80%) of transposable elements with repetitive sequences. Most varieties do not have a reference genome.

Xdrop can enrich specific genomic regions, enabling in-depth sequence elucidation. Here, we demonstrate this method using an important gene cluster in a barley variant.

Experimental setup

The Cer-quc gene cluster of the barley variety PLG-1041 Amsbio, which lacks a reference genome, encodes three genes involved in the production of β-diketone polyketides. They form part of the wax layer that protects plants against pests and reduces water loss, among its other functions.[3]

First, we used the Samplix online tool to design three detection sequences within the 100 kb gene cluster based on the reference genome sequence assembly Morex V3. We positioned them as close to the genes as possible (Figure 1). Since the size and sequence of the gene cluster in PLG-1041 Amsbio is unknown, we performed three enrichments. The enriched DNA was sequenced using the Oxford Nanopore® solution, yielding 1.9 Gb of data.

Mapping to the Morex V3 reference genome

We mapped the sequence data to the Morex V3 reference genome using Minimap2 (Figure 2). Coverage is high on the gene regions, highlighting the similarities, but few reads map to the intergenic regions, suggesting substantial differences between the two varieties. We performed de novo assembly to characterize the structure in PLG-1041 Amsbio.

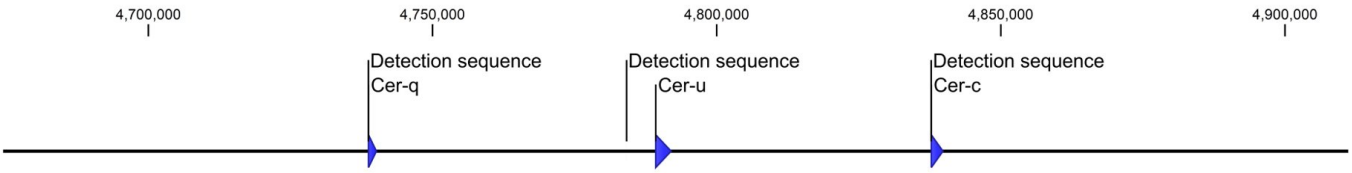

Figure 1. *Overview of the Cer-quc gene cluster on chromosome 2H in the More V3 genome assembly. The locations of the three detection sequences are indicated.*

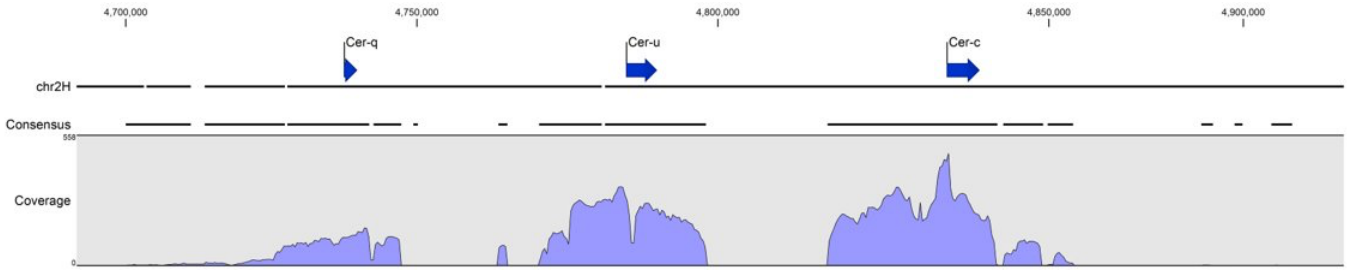

Figure 2. *Mapping of the obtained reads to the Morex V3 reference assembly genome using Minimap2. Genes in the cluster show high coverage, whereas limited coverage breadth is observed for the intergenic regions. De novo assembly should be performed to resolve the sequence.*

De novo assembly

Prior to de novo assembly, we performed a NECAT error correction and SACRA chimera splitting to ensure high quality. The resulting reads were used for de novo assembly with Canu.[3] This resulted in a 57.5-kb contig (tig00000020) that contains the sequences of Cer-q, Cer-u and Cer-c. We subsequently mapped the original reads to the contig, finding that the coverage supports the sequence of the assembled region (Figure 3). Additionally, a LAST[4] alignment between the contig and the reference showed that the contig contains sequences around the gene regions that are similar to Morex V3, but parts of the Morex V3 reference sequence are not found in the de novo assembly (Figure 4).

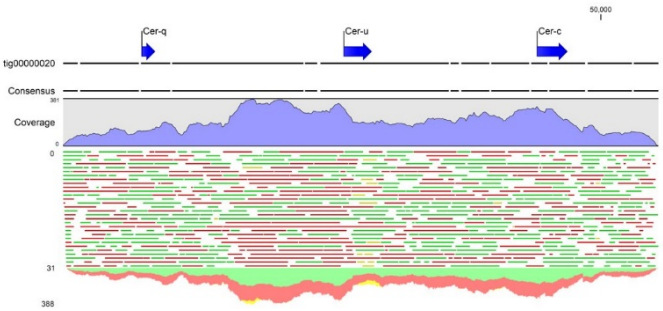

Figure 3. *Mapping of obtained reads to contig tig00000020 containing the Cer-q, Cer-u and Cer-c genes.*

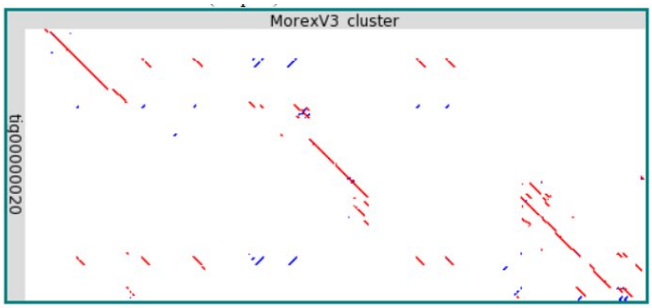

Figure 4. *LAST alignment of the reconstructed contig tig00000020 and the gene cluster in the Morex V3 reference. The diagonals show aligning regions, either forward (red) or reverse (blue).*

Analysis of the gene content

Aligning the genes of the reconstructed contig tig00000020 to the genes of Morex V3 demonstrated that the coding sequences are well conserved between the two varieties with some polymorphisms (Figure 5). These polymorphisms are almost exclusively distributed in the introns of the genes, which should not affect the plant phenotype.

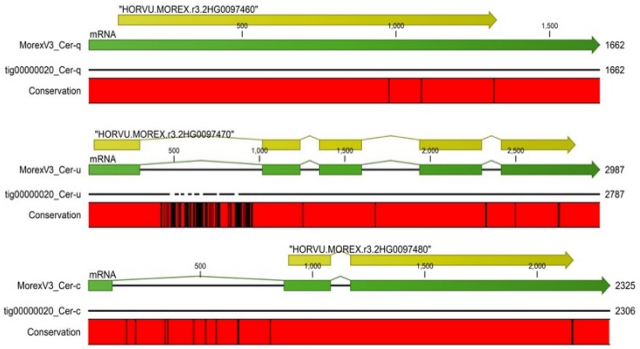

Figure 5. *Genes of Morex V3 and contig tig00000020 aligned for comparison. Annotations of cds (yellow) and mRNA (green) were extracted from Morex V3. Conserved nucleotides are shown in red and polymorphisms as black vertical lines.*

Conclusions

Xdrop can be used to enrich specific genomic regions in plant varieties, enabling in-depth sequence elucidation, even when only limited knowledge about the variety-specific genomic sequence is available.

How Xdrop supports investigations of genomic regions

Xdrop enriches and amplifies a ~100-kb DNA region of interest (ROI) for downstream long- or short-read library preparation and sequencing. The ROI is identified using a ~150-bp detection sequence within or adjacent to it.

First, Xdrop partitions high molecular weight genomic DNA into millions of double emulsion droplets. The droplets containing DNA with the ROI are identified based on the detection sequence. The droplets are then sorted using a standard fluorescence-activated cell sorter (FACS) and the ROI-positive ones are collected. The long DNA fragments are finally amplified in droplets via dMDA to ensure unbiased DNA amplification. Library preparation and sequencing follow.

Learn more about Xdrop at samplix.com/applications and samplix.com/technology.

References and notes

1. Mascher, M., et al. 2017. A chromosome conformation capture ordered sequence of the barley genome. Nature 544: 427. doi: 10.1038/nature22043
2. Jayakodi, M., et al. 2020. The barley pan-genome reveals the hidden legacy of mutation breeding. Nature 588: 284. doi: 10.1038/s41586-020-2947-8
3. S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, A. M. Phillippy, Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. (2017), doi:10.1101/gr.215087.116
4. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. Genome research, 21(3), 487-493. doi: 10.1101/gr.113985.110