Samplix

Closing gaps in the sequence of a biosynthetic gene cluster in a tomato variety using Xdrop®

Lea Møller Jagd, Cristina Gamba, and Peter Mouritzen
Samplix ApS, Birkerød, Denmark

Summary

- Whole genome sequencing is not the ideal method to close gaps in gene cluster sequences.
- Xdrop supports targeted and efficient gap closing.
- Here, Xdrop is used to close gaps in the structure of a biosynthetic gene cluster in a tomato variety with excellent coverage and accuracy.

Introduction

The genome of the Heinz 1706 cultivar of tomato (*Solanum lycopersicum*) was published by The Tomato Genome Consortium in 2012[1] and has been improved several times since then. The genome is approximately 900 Mb in size and consists of 12 chromosomes.

Falcarindiol is a modified lipid found in tomato, where it is involved in defense against pathogens. Candidate genes for its biosynthesis were identified in a ~20 kb gene cluster,[2] but the data did not correspond to the SL3.0 reference genome build. Whole genome sequencing (Oxford Nanopore® Technology and Sanger sequencing) revealed two full-length duplicated genes that were only partially present in the SL3.0 build: Solyc12g100240 and Solyc12g100260 (Figure 1).

At the time of their study, Jeon et al. had to sequence the whole genome and complement the findings using PCR and Sanger sequencing. This laborious approach resulted in only 3 long reads covering the region of interest among almost 4 Gb of data. After that study, the SL4.0 build was completed. It proved congruent with the data for the candidate genes (Figure 1).

Here, we show that Xdrop can enrich the entire gene cluster for in-depth examination via long-read sequencing, significantly streamlining the workflow.

Experimental setup

Niraj Mehta (Department of Chemistry, Stanford University) designed the detection sequence primers to target Solyc12g100270 and extracted DNA of high molecular weight from the leaves of *S. lycopersicum*, cultivar VF36. Only 2.4 ng were needed for the Xdrop workflow (see How Xdrop supports gap closing for details). Long-read sequencing was performed using Oxford Nanopore Technologies.

Enrichment

We obtained a 1,029-fold enrichment of the gene Solyc12g100270, using the SL4.0 build as the reference genome. The broader 100 kb region around the detection sequence was enriched 237-fold.

De novo assembly and mapping

The data consisted of 73,728 reads totaling 0.39 Gb. Prior to de novo assembly, we used Pacasus[3] to split chimeric reads (inverted repeats of the same read) to ensure high quality. The resulting reads were assembled with Canu v.1.9[4] using the default settings. The raw assembly produced by Canu, was exposed to 3 iterations of the highquality consensus sequences generation by Racon[5] to create the Samplix Assembly.

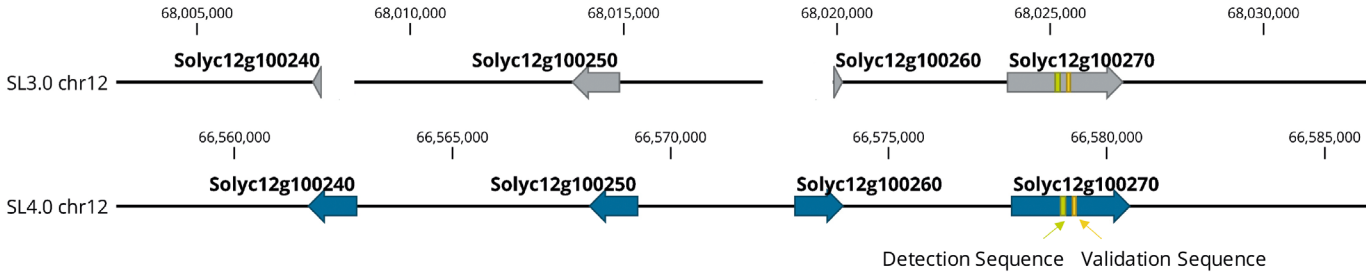

Figure 1. The falcarindiol gene cluster in the SL3.0 and SL4.0 builds of the Heinz 1706 tomato cultivar. The locations of the detection and validation sequences for the Xdrop target enrichment are shown.

Using Minimap2 and the default settings, the split reads were mapped back to the Samplix Assembly. The same reads were also mapped against the assembly reconstructed by Jeon et al. as well as the SL3.0 and SL4.0 builds (Figure 2).

SNP variants were called using freeBayes[6] (grey bar, Figure 2). Note that the Samplix Assembly has considerably fewer SNPs (a proxy for errors) than the assembly from Jeon et al., which was based on only 3 Oxford Nanopore Technology reads. The increased assembly accuracy was possible thanks to the high enrichment, which made more data available to reconstruct the cluster region.

We also mapped the reads to the SL4.0 reference genome, finding that 56,454 reads mapped to the genome and 1,711 reads aligned with the 100 kb region around the detection sequence. The two gaps visible in the coverage graph for SL3.0 correspond to segments of Solyc12g100240 and Solyc12g100260 that are not represented in this reference build. These gaps are closed in the newer SL4.0 build, where we could obtain full coverage for the 100 kb region (Figure 2).

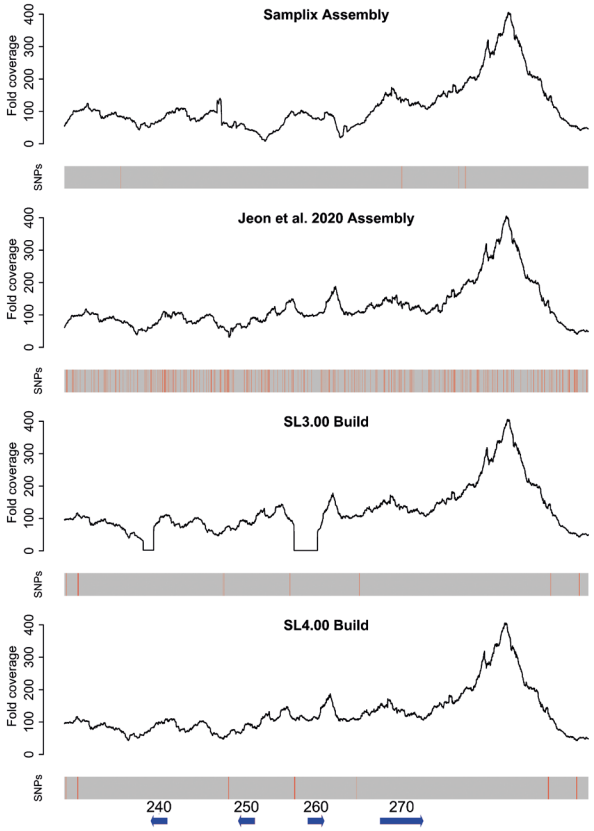

Figure 2. Mapping of the Samplix Assembly to the assembly from Jeon et al. and the tomato genome SL3.0 and 4.0 builds.

Conclusion

Xdrop can be used to enrich specific genomic regions in plant varieties, enabling highly efficient gap closing without laborious whole genome sequencing.

How Xdrop supports gap closing

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 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.

Acknowledgements

We would like to thank Niraj Mehta (Department of Chemistry, Stanford University) for providing samples, primers, and his valuable contribution to this project. This project received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 848497.

References and notes

1. The Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution Nature 485: 635–641.
2. Jeon, J.E., et al. 2020. A Pathogen-Responsive Gene Cluster for Highly Modified Fatty Acids in Tomato Cell 180(1): 176–187.
3. Warris, S., et al. 2018. Correcting palindromes in long reads after whole-genome amplification. BMC Genomics 19: 798.
4. Koren, S., et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27: 722.
5. Vaser R, et al. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27(5): 737–746.
6. Garrison, E. and Marth, G. 2012. Haplotype-based variant detection from short-read sequencing arXiv:1207.3907.