Gap closing with Xdrop™: Complete sequencing of the falcarindiol gene cluster in tomato

Background

The genome of the inbred tomato cultivar Heinz 1706 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.

Gene clusters are prominent in plant genomes.[2] With knowledge of the mRNA sequence of one protein in a biosynthetic pathway, you can identify the corresponding genomic location and the rest of the pathway genes based on their proximity. However, limited knowledge of the region of interest, repetitive sequences, and large genome size are all factors that make it difficult to sequence the context surrounding key genes in a pathway. Enrichment with Xdrop™ solves this issue.

The Xdrop™ Technology

The Xdrop™ technology combines high-resolution droplet PCR (dPCR) with droplet sorting and Multiple Displacement Amplification in droplets (dMDA).

Firstly, Xdrop™ partitions the DNA into millions of double emulsion droplets. Droplets containing the target DNA molecules are identified by a ~150 bp targeted dPCR, specific to a Detection Sequence within or adjacent to the region of interest.

The detection and sorting of droplets are performed using a standard cell sorter, which allows the PCR positive droplets containing the ROI to be collected. The sorted long DNA fragments are finally amplified in droplets (dMDA) to ensure unbiased DNA amplification.

The Xdrop™ enrichment and amplification technology are compatible with both long- and short-read library preparation and sequencing.

The falcarindiol gene cluster in tomato (*Solanum lycopersicum*)

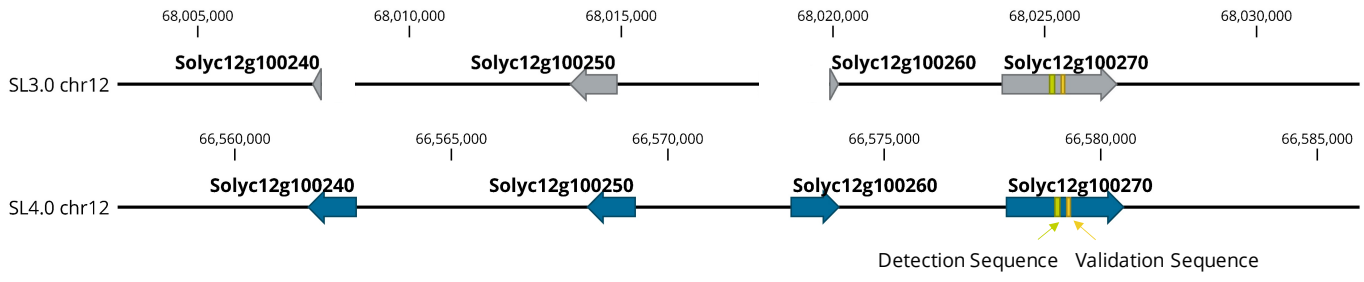
Falcarindiol is a modified lipid involved in pathogenic defense. Recent metabolic and mRNA analyses identified candidate genes in a ~20 kb cluster encoding four biosynthetic genes.[3] The data from these analyses, however, did not correspond to the reference genome build SL3.0. Whole genome sequencing using Oxford nanopore technology (ONT) and Sanger sequencing revealed two full-length duplicated genes (Solyc12g100240 and Solyc12g100260), only partially present in the SL3.0 build. Since that study, a newer build (SL4.0) has been completed which is congruent with the data (see figure below).

At the time of their study, Jeon and colleagues had to sequence the whole genome and complement the finding using PCR and Sanger sequencing. This approach was laborious and resulted in only 3 long ONT reads covering the region of interest among almost 4 Gb of data. Our goal in this Application Note is to demonstrate that Xdrop™ can enrich the entire gene cluster for in depth examination via long-read sequencing.

Sample Preparation, Xdrop™ and Sequencing Setup

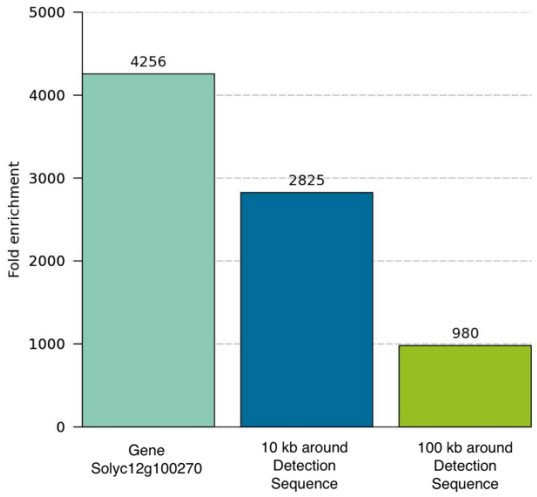
Xdrop™ primers were designed to target Solyc12g100270 by Niraj Mehta (Department of Chemistry, Stanford University). Though not centrally located in the gene cluster, Xdrop™ captured roughly 50 kb on each side of the Detection Sequence. DNA of high molecular weight (>60 kb) was extracted from tomato leaves (*Solanum lycopersicum*, cultivar VF36) and also provided to Samplix by Niraj Metha. Only 2.4 ng were needed for the Xdrop™ workflow. Long-read sequencing was performed using Oxford Nanopore Technologies.

Falcarindiol gene cluster in the SL3.0 and SL4.0 builds



Enrichment Estimate

We obtained a 4,256-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 980-fold.



De novo Assembly

We performed *de novo* assembly of the reads obtained from the ONT run to demonstrate the efficiency of Xdrop™ in closing gaps when having only partial genomic information.

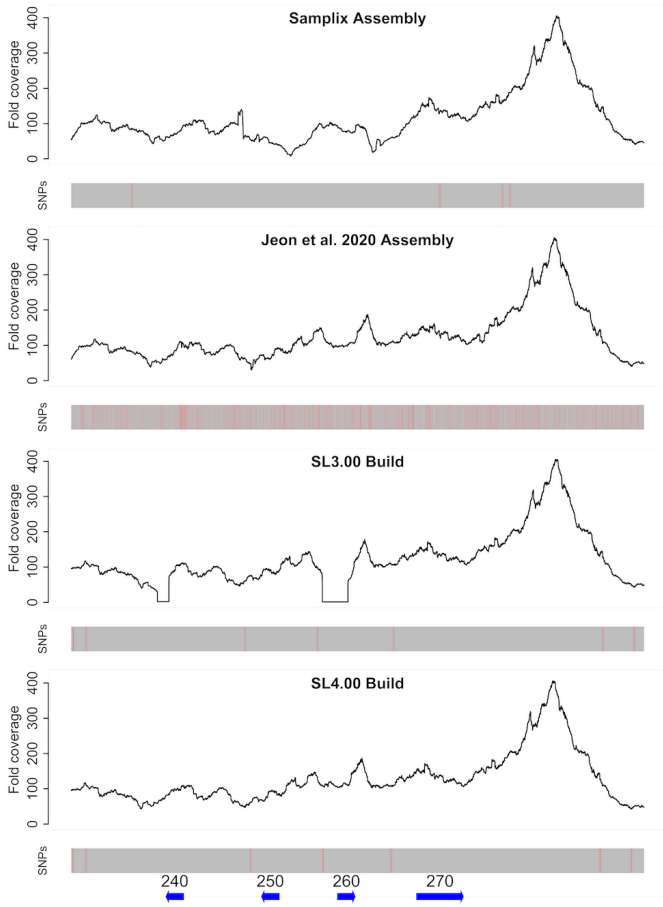
The data consisted of 73,728 reads totaling 0.39 Gb. We used Pacasus[4] to split chimeric reads (inverted repeats of the same read) introduced during dMDA and assembled the resulting reads with Canu v.1.9[5] using the default settings.

The raw assembly produced by Canu, was exposed to 3 iterations of the high-quality consensus sequences generation by Racon[6] to create the Samplix Assembly. Using minimap2 and the default settings, the split reads were mapped back to the *de novo* assembled region ("Samplix Assembly"). The same reads were also mapped against the assembly reconstructed in Jeon et al. 2020 Heinz and the tomato genome builds SL3.0 and SL4.0 (see figure coverage distribution in figure below).

SNP variants were called using freeBayes[7] (see the grey bar in the figure below) . Notice that the assembly based on Xdrop™ enrichment (Samplix) displays many fewer SNPs (a proxy for errors) compared to the assembly from Jeon et al. 2020, based only on 3 ONT reads. The increased assembly accuracy was possible thanks to the > 2000-fold enrichment, which allowed more data to be available to reconstruct the cluster region.

Mapping to the Reference Genome

We also mapped the reads to the SL4.0 reference genome. 56,454 reads mapped to the genome, where 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 are filled in the newer SL4.0 build, where we could obtain full coverage for the 100 kb region (see figure below).



Conclusions

Xdrop™ opens new possibilities to explore gene clusters in plant genomes with minimal knowledge of the region of interest. The resulting sequencing coverage and enrichment are high, enabling in-depth characterization of functionally and spatially related genes.

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 has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 848497.

References

1. Sato, S., et al. (2012) Nature 485, 635–641.
2. Nützmann, H.-W., et el. (2017) New Phytologist 211: 771.
3. Jeon, J.E., et al. (2020) Cell 180: 176.
4. Warris, S., et al. (2018) BMC Genomics 19: 798.
5. Koren, S., et al. (2017) Genome Res. 27: 722.
6. Vaser R, et al. (2017) Genome Res. 27(5):737-746.
7. Garrison, E., et al. (2012) arXiv:1207.3907.