A phase-resolved, chromosome-scale de novo assembly of the diploid 'Regina' sweet cherry genome.

Loïck Le Dantec, Nabil Girollet, Jérôme Gouzy, Erika Sallet, Armel Donkpegan, Shawn Sullivan, Hayley Mangelson, Kyle Langford, Mathieu Fouché, Jose Quero-Garcia, et al.

A PHASE-RESOLVED, CHROMOSOME-SCALE DE NOVO ASSEMBLY OF THE DIPLOID 'REGINA' SWEET CHERRY GENOME

Loïck Le Dantec[1*], Nabil Girollet[2], Jérôme Gouzy[3], Erika Sallet[3], Armel Donkpegan[1], Shawn Sullivan[4], Hayley Mangelson[4], Kyle Langford[4,] Mathieu Fouché[1], Jose Quero-Garcia[1], and Elisabeth Dirlewanger[1]

[1]INRA-UMR 1332 Biologie du Fruit et Pathologie, Villenave d'Ornon, France, [2]UMR EGFV 1287 – Ecophysiologie et Génomique Fonctionnelle de la Vigne, Villenave d'Ornon, France, [3]Laboratoire des Interactions Plantes Micro-organismes (LIPM) -INRA/CNRS, Castanet-Tolosan, France, [4]Phase Genomics Inc, Seattle, WA, United States

*Please send correspondence to loick.le-dantec@inrae.fr

SUMMARY

Sweet cherry (Prunus avium L.) is a diploid species with an estimated genome size of 338 MB (Arumuganathan et al. 1991). It is mostly self-incompatible, and therefore has a heterozygous genetic background. We have sequenced the genome of the 'Regina' sweet cherry variety, which is a late blooming cultivar. Here we present whole-chromosomes assembly and phasing using a combination of long read sequencing, proximity ligation (HiC) data and GBS markers.

1 PacBio Sequencing
RSII P6-C4 chemistry
30 SMRT cells

	Total (30SMRT Cells)
Number of bases	27,9 billions
N50	17,5kb
Depth	82,5 x

2 PacBio Assembly

	PacBio assembly FalconUnzip (polished primary contigs)	PacBio assembly FalconUnzip (polished haplotigs)
# contigs	1078	1289
N50 (Mb)	1,23	0,35
L50	68	150
Max Size (Mb)	10,5	1,83
total Size (Mb)	323	177
% genome coverage (338 Mb)	95 %	52 %

3 Optical Mapping Scaffolding

Hybrid Assembly:	bionano assembly of PacBio polished primary contigs
# scaffolds	92
Median length (Mb)	1,9
N50 (Mb)	5,96
Max Size (Mb)	16,3
total Size (Mb)	279
% genome coverage (338 Mb)	83%

Assessment of genome assembly

Contig/Scaffold set	Complete Busco
Primary contigs + haplotigs	95,6%
Hybrid Assembly scaffolds	90,9%

4 Regina pseudomolecules assembly with GBS markers

	LG size(cM)	SNP used in map construction	Scaffolds anchored	coverage of cherry scaffolds (bp)
chr1	112.0	322	11	54,172,154
chr2	99.4	276	13	39,341,669
chr3	82.0	253	11	30,425,033
chr4	79.3	326	7	28,891,835
chr5	83.1	205	3	21,168,142
chr6	93.4	244	10	34,796,725
chr7	78.2	225	10	29,363,141
chr8	87.3	279	7	26,685,109
Total	714.6	2131	72	**264,843,808**

5 Structural annotation with Eugene
(http://eugene.toulouse.inra.fr/)

Number of nucleotides (without 'N') 323361974
Per cent GC 38.25
Total number of genes **39180**
Total nucleotides (bp) 93957457

** Protein coding genes
Number of protein coding genes **34591**
Mean gene length (bp) 2650.79

** Non protein coding genes
Number of non protein coding genes **4589**
Mean ncRNA gene length (bp) 493.33

** Intergenic (inter protein-coding genes)
Mean length 6420.76
GC per cent 34.59

Total repeat regions Repeat region number = 50182
Repeat region length = 135602147 nt (**39.2% of genomic sequences**)

HiC sequencing
(Proximo HiC sequencing kit)
250000 read pairs

GBS markers
2041 SNP issued from GBS analyses of 116 Regina x Garnet hybrids

6 Falcon-Phase
Haplotype phasing

Proximo scaffolding

Regina genome Version 1 (V1) (consensus)

https://doi.org/10.15454/KEW474
Portail Data INRAE, V1

Regina genome Version 2 (V2) (Haplotypes P0 and P1)

	P0	P1
chr1	56,691,622	56,442,611
chr2	45,878,335	46,651,641
chr3	37,108,384	37,465,270
chr4	32,078,273	32,266,274
chr5	24,269,681	24,238,735
chr6	38,593,614	39,404,943
chr7	34,940,866	33,814,143
chr8	33,900,371	33,469,941
Total	303,461,146	303,753,558
genome coverage (338 Mb)	89,78%	89,87%
N50 (bp)	**37,108,384**	**37,465,270**
Complete BUSCO	93,10%	93,60%
Complete BUSCO	**P0+P1: 95,6%**	

Acknowledgements

The sweet cherry genome sequencing was financed by the INRA 'Plant Biology and Breeding' division in the 'CherryGeno' project.

The PacBio sequencing was done by the GeT-PlaGe INRA platform (http://get.genotoul.fr/) and the optical map was performed by INRA CNRGV (http://cnrgv.toulouse.inra.fr/). We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing and storage resources.

CONCLUSION

Regina PacBio long reads were de novo assembled using FALCON UNZIP. A first assembly using optical mapping and high density genetic linkage maps with GBS data resulted in a genome of 264,8 Mb (78,4 % of estimated genome size) distributed on 8 linkage groups with a scaffold N50 ~6Mb (Regina V1). A new assembly (ReginaV2) combining Falcon Unzip assembly of PacBio long reads, proximity ligation (HiC) data and GBS markers has been done with Falcon Phased and the Proximo pipeline of Phase Genomics. It produced complete, fully phased, diploid chromosome scale scaffolds with a genome coverage of ~90%, a N50 of ~37 Mb and a complete BUSCO score of 95,6%. Structural and functional annotations of Regina V2 are in progress. Obtaining these phased chromosomes will allow very detailed studies of the allelic effects on the traits studied, in particular by making it possible to study the allelic expressions of candidate genes.

10th ROSACEAE GENOMICS CONFERENCE RGC10 VIRTUAL 2020 9-10-11 December 16-17-18 December