



HAL
open science

A High Quality Draft Assembly of the Diploid 'Regina' Sweet Cherry Genome

Loick Le Dantec, Nabil Girollet, Mathieu Fouche, Jerome Gouzy, Erika Sallet,
José Quero-Garcia, Teresa Barreneche, Bénédicte Wenden, Elisabeth
Dirlewanger

► **To cite this version:**

Loick Le Dantec, Nabil Girollet, Mathieu Fouche, Jerome Gouzy, Erika Sallet, et al.. A High Quality Draft Assembly of the Diploid 'Regina' Sweet Cherry Genome. Plant & Animal Genome Conference XXVI, Jan 2018, San Diego, United States. hal-01865005

HAL Id: hal-01865005

<https://hal.science/hal-01865005v1>

Submitted on 30 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A High Quality Draft Assembly of the Diploid 'Regina' Sweet Cherry Genome

Loïck Le Dantec^{1*}, Nabil Girollet², Mathieu Fouché¹, Jérôme Gouzy³, Erika Sallet³, Jose Quero-Garcia¹, Teresa Barreneche¹, Bénédicte Wenden¹ and Elisabeth Dirlewanger¹

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

*Please send correspondence to loick.le-dantec@inra.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 and assembled the genome of the 'Regina' sweet cherry variety, which is a late blooming cultivar. Here we present a draft phased sweet cherry genome assembly using a combination of sequencing strategies (long reads sequencing with PacBio RSII and optical mapping with BioNano whole-genome maps).

PacBio Sequencing
RSII P6-C4 chemistry
30 SMRT cells

PacBio Assembly
Falcon and FalconUnzip

Optical Mapping
Bionano

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

	PacBio assembly Falcon (draft)	PacBio assembly FalconUnzip (polished primary contigs)	PacBio assembly FalconUnzip (polished haplotigs)
# contigs	2019	1078	1289
N50 (Mb)	1,2	1,23	0,35
L50	77	68	150
Max Size (Mb)	10,7	10,5	1,83
total Size (Mb)	339	323	177
% genome coverage (338 Mb)	100 %	95 %	52 %

Nicking enzyme	BssS1
Theoretical nb labels/100kb	13,6
Real nb labels/100kb	11
Filtered data >100kb (Gb)	156(459X)
Molecules N50 (kb)	178

Bionano Assembly	
# contigs	433
Median length (Mb)	0,809
N50 (Mb)	1,7
Max Size (Mb)	9,9
total Size (Mb)	535
% genome coverage (338 Mb)	158

Hybrid Assembly: Scaffolding

Assessment of genome assembly : BUSCO (V2.0)

Contig set	PacBio FalconUnzip Primary + Haplotigs	Hybrid Assembly scaffolds	Hybrid Assembly scaffolds + not scaffolded PacBio Contigs	Hybrid Assembly scaffolds + not scaffolded PacBio Contigs + FalconUnzip haplotigs
Complete	95,6%	90,9%	93,6%	95,9%
Duplicated	48,5%	2,9%	5%	46,5%
Fragmented	1,3%	2,1%	1,7%	1,1%
Missing	3,1%	7%	4,7%	3%

PacBio polished primary contigs + bionano assembly

# 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%

Hybrid scaffolds + not scaffolded PacBio contigs : 338Mb

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

PacBio long reads were de novo assembled using FALCON assembler and phased using FALCON UNZIP. BioNano optical whole genome maps were used for further scaffolding. Our de novo genome assembly resulted in a genome of 279 Mb (83 % of estimated genome size), 92 scaffolds and a largest scaffold of 16,3 Mb. The assembly has high contiguity (contig N50 1,23Mb, scaffold N50 ~6Mb) and good completeness with 95,9% BUSCO genes complete (scaffolds+ unscaffolded contigs). We are now in the process of reconstructing the pseudo molecules using high density genetic linkage maps and GBS data. Structural annotation is on the way using the integrative EuGene platform with different sources of evidence available for gene prediction including transcriptomic data from Regina obtained previously in the lab (Unigene set and RNASeq).

