Haplotype Resolved PacBio HiFi Assembly of Potato Dihaploids Capturing North American Allelic Diversity of Tetraploid Parents

Brieanne Vaillancourt[1], Kathrine Mailloux[1], John P. Hamilton[1,2], Xiaoxi Meng[3], Laura Shannon[3], and C. Robin Buell[1,2,4]

[1]Center for Applied Genetic Technologies, University of Georgia, Athens, GA 30602, USA
[2]Department of Crop & Soil Sciences, University of Georgia, Athens, GA 30602, USA
[3]Department of Horticultural Science, University of Minnesota, St. Paul, MN 55108, USA
[4]Institute of Plant Breeding, Genetics & Genomics, University of Georgia, Athens, GA 30602, USA

Grant 2019-51181-30021

Introduction

Potato (*Solanum tuberosum*) is the third most important food crop in the world with a value of over $3.9 billion in the United States. Cultivated potato, *Solanum tuberosum*, is a highly heterozygous tetraploid (2n = 4x = 48) and as a consequence, breeding is primarily phenotypic based with the breeder selecting an F1 from a cross of two tetraploid parents. Diploid (2n = 2x = 24) breeding of inbred/F1 hybrids has the potential to revolutionize potato breeding due to the simpler genetics and the ability to use true seed instead of seed tubers. Sources of diploid germplasm include a limited number of diploid cultivars and landraces, wild potato species which would require domestication, and dihaploids generated from cultivated tetraploids via genome reduction. Advantages of dihaploids from elite cultivars are that breeders have already selected alleles for agronomic traits including yield, tuber shape, tuber size, disease resistance, and market quality traits. In this project, we are generating chromosome-scale, haplotype-resolved genome assemblies for 20 dihaploids that represent the allelic diversity of North American elite cultivars, a key component of converting US potato breeding to a diploid/F1 hybrid approach.

Potato 2.0 Project

This project is part of the overall USDA-funded Potato 2.0 project with four objectives:

- Determine the genetic basis and environmental stability of self-fertility.
- Develop self-fertile, diploid germplasm that captures the allelic diversity of tetraploid cultivars.
- Create inbred lines that are fixed (homozygous) for key traits.
- Develop agronomic and economic frameworks for incorporating true potato seed into the potato production system and assessing its impacts.

https://potatov2.github.io/

Tetraploids and Derived Dihaploids

Dihaploids for this project were created by crossing tetraploid female parents with pollen from haploid inducer male parent IVP101 or IVP48 (*Solanum tuberosum* group phureja). Female fertile dihaploids were checked for aneuploidy using Illumina whole genome shotgun sequencing data. Based on aneuploidy results, vigor, and genetic diversity, a set of 20 dihaploids were selected for chromosome-scale long read assembly.

Table 1. List of dihaploids and their tetraploid parent.

Dihaploid	Tetraploid Parent	Dihaploid	Tetraploid Parent
A07061-6-DH07	A07061-6Rus	PayetteRusset-DH04	Payette Russet
CaribouRusset-DH42	Caribou Russet	Pike-DH04	Pike
CO99076-6R-DH02	CO99076-6R	RangerRusset-DH10	Ranger Russet
Defender-DH45	Defender	RedNorland-DH0205	Red Norland
Kalkaska-DH31	Kalkaska	VanguardRusset-DH26	Vanguard Russet
Lelah-DH12	Lelah	W14NYQ29-5-DH24	W14NYQ29-5
MercuryRusset-DH06	Mercury Russet	W14NYQ9-2-DH137	W14NYQ9-2
Missaukee-DH157	Missaukee	W8890-1R-DH02	W8890-1R
MSJ147-1-DH1	MSJ147-1	W9426-3R-DH05	W9426-3R
MSR127-2-DH01	MSR127-2	W9968-5-DH84	W9968-5

Haplotype-Resolved Genome Assemblies

Dihaploids were sequenced to 32-35x coverage per haplotype using PacBio HiFi and assembled using hifiasm to generate phased, haplotype-resolved assemblies.

Table 2. Dihaploid assembly statistics before filtering.

	Haplotype	Number of Contigs	Assembly Size (Gbp)	N50
Kalkaska-DH31	1	906	0.84	33,661,863
	2	456	0.78	17,613,795
Missaukee-DH157	1	840	0.80	17,361,192
	2	285	0.82	20,492,766
PayetteRusset-DH04	1	1,623	0.87	53,299,128
	2	483	0.82	28,593,982
RedNorland-DH205	1	804	0.80	21,481,748
	2	331	0.78	14,054,775

Figure 1.

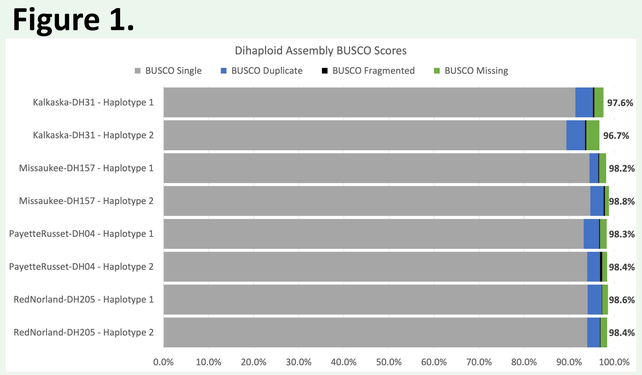


Figure 1. BUSCO scores of dihaploid assemblies before filtering. Values next to bar indicate the complete BUSCO score.

Figure 2.

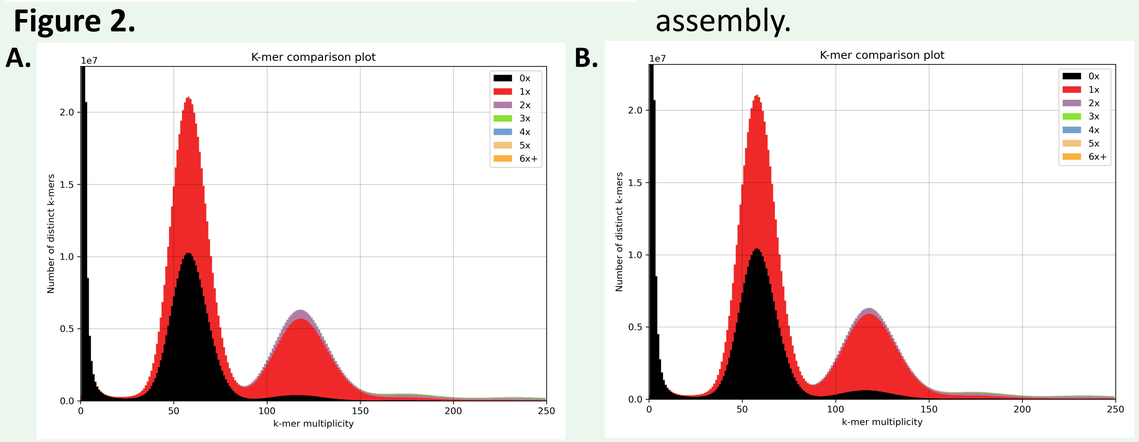


Figure 2. K-mer spectrum of resolved haplotypes 1 (A) and 2 (B) of Kalkaska-DH31 using gDNA Illumina sequencing data and the respective hifiasm assembly.

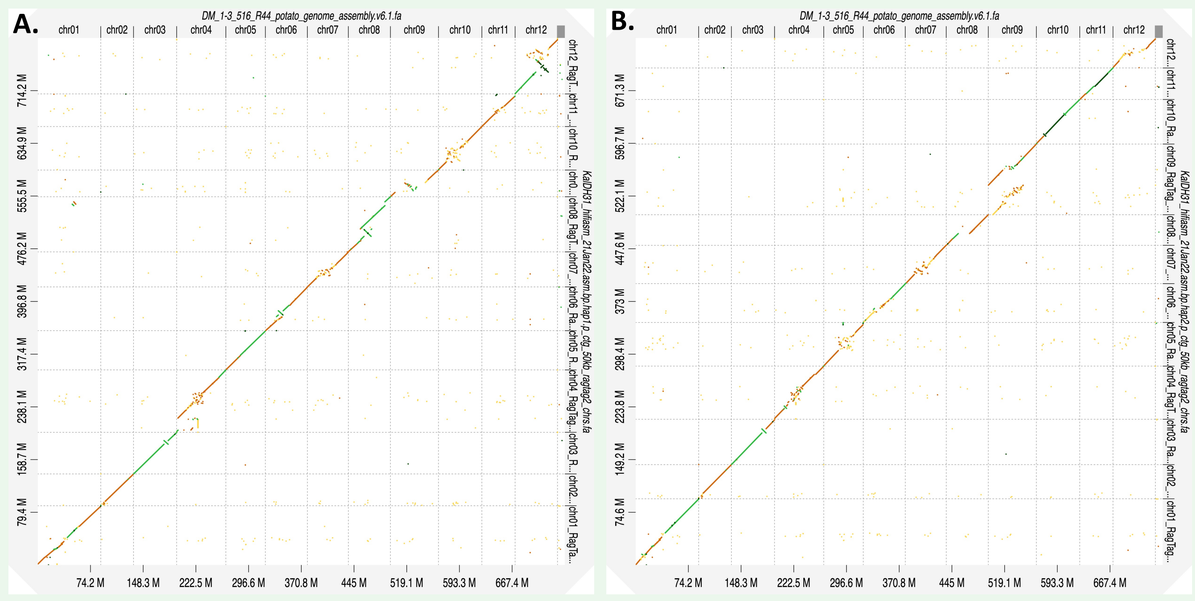
Chromosome-Scale Assembly of Dihaploid Genomes

Ragtag was used to generate 24 pseudomolecules for each dihaploid. We are currently developing a computational pipeline for validation and refinement of the Ragtag-derived pseudomolecules using Hi-C contacts with Juicer and 3D-DNA.

Table 3. Kalkaska-DH31 dihaploid assembly statistics after two rounds of Ragtag.

	Number of Pseudomolecules	Assembly Size (Gbp)	N50 (Kb)
Haplotype 1	12	0.79	65,931,505
Haplotype 2	12	0.75	63,039,504

Figure 3. D-Genies plot of Kalkaska-DH31 haplotype 1 (A) and haplotype 2 (B) assemblies against DM v6.1 (http://spuddb.uga.edu/).



Haplotype-Specific Gene Annotation

Each dihaploid will be annotated using dihaploid-specific transcript evidence. We have generated Oxford Nanopore Technologies full-length cDNA sequences of young leaf and tuber.

On-Going Efforts

- Finish sequencing dihaploids
- Finalize HiC pipeline for pseudochromosome assembly
- Annotate all dihaploids
- Assess genetic diversity
- Release data on SpudDB (spudb.uga.edu)

Contributions

Generation of dihaploids for this project:
Paul Bethke, USDA-ARS; Walter de Jong, Cornell University; David Douches, Michigan State University; Jeffrey Endelman, University of Wisconsin-Madison; Shelley Jansky, USDA-ARS; Sagar Sathuvalli, Oregon State University; Ek Han Tan, University of Maine