Ohio Agricultural Research and Development Center

The Draft Genome of *Taraxacum kok-saghyz*, an Alternative Natural Rubber Resource

Xiaofeng Zhuang, Zinan Luo, Brian Iaffaldano, Katrina Cornish
Department of Horticulture and Crop Science, Ohio State University

ABSTRACT

Kazak dandelion (*Taraxacum kok-saghyz Rodin*, TK) is being developed as an alternative natural rubber resource in response to the increasing global rubber market demand. However, TK is still genome resource-poor in part due to the large genome size and large fraction of highly repetitive DNA. Here, we present a de novo draft genome assembly of *T. kok-saghyz*. The genome consists of 188,608 scaffolds covering 1,440 Mb, with an N50 scaffold size of 55.2 Kb. Preliminary whole genome annotation predicted 23,760 protein-coding genes (with Start/Stop codon), in which more than 80% of these genes were further confirmed by known plant genes in GenBank. The whole genome sequencing of TK will provide a needed platform to accelerate the improvement of this economically important, new industrial crop.

INTRODUCTION

Natural rubber (*cis*-1,4-polyisoprene) is a high molecular weight polymer (i.e., >1000 kilograms per mole (kg/mol)) that cannot be replaced by synthetics in most applications (Collins-Silva et al., 2012).

Taraxacum kok-saghyz (TK) is an alternative rubber resource under commercial development because the production of natural rubber, currently harvested from the Para rubber tree (*Hevea brasiliensis*), is faced with serious challenges, such as potential shortages of supply and pathogen threats (van Beilen and Poirier, 2007). Also, TK's short life cycle, high quality rubber and adaptation to diverse environments makes it a potentially ideal rubber-producing crop. However, difficulties with conventional breeding, along with limited genome-based information, have impeded efficient crop improvement of TK. Marker assisted selection can improve the efficiency of breeding by enabling the direct selection of targeted genotypes (Ribaut and Hoisington, 1998). Analysis of genetic linkages among markers, and identification of the genetic locations of desirable phenotypes, would further improve selection accuracy.

Here, we describe progress in assembling the reference genome of TK, which will aid in the future development of high-yielding clones to keep up with the increasing need for nature rubber.

METHODS

- ➢ **Whole genome sequencing**: A total of 327 TK plants with big roots were collected from the field. After quantifying the rubber content, TK-009, a high rubber plant (115.7 mg/gdw root), was selected for whole genome sequencing using NGS (Next-Generation Sequencing) method (Fig 1).

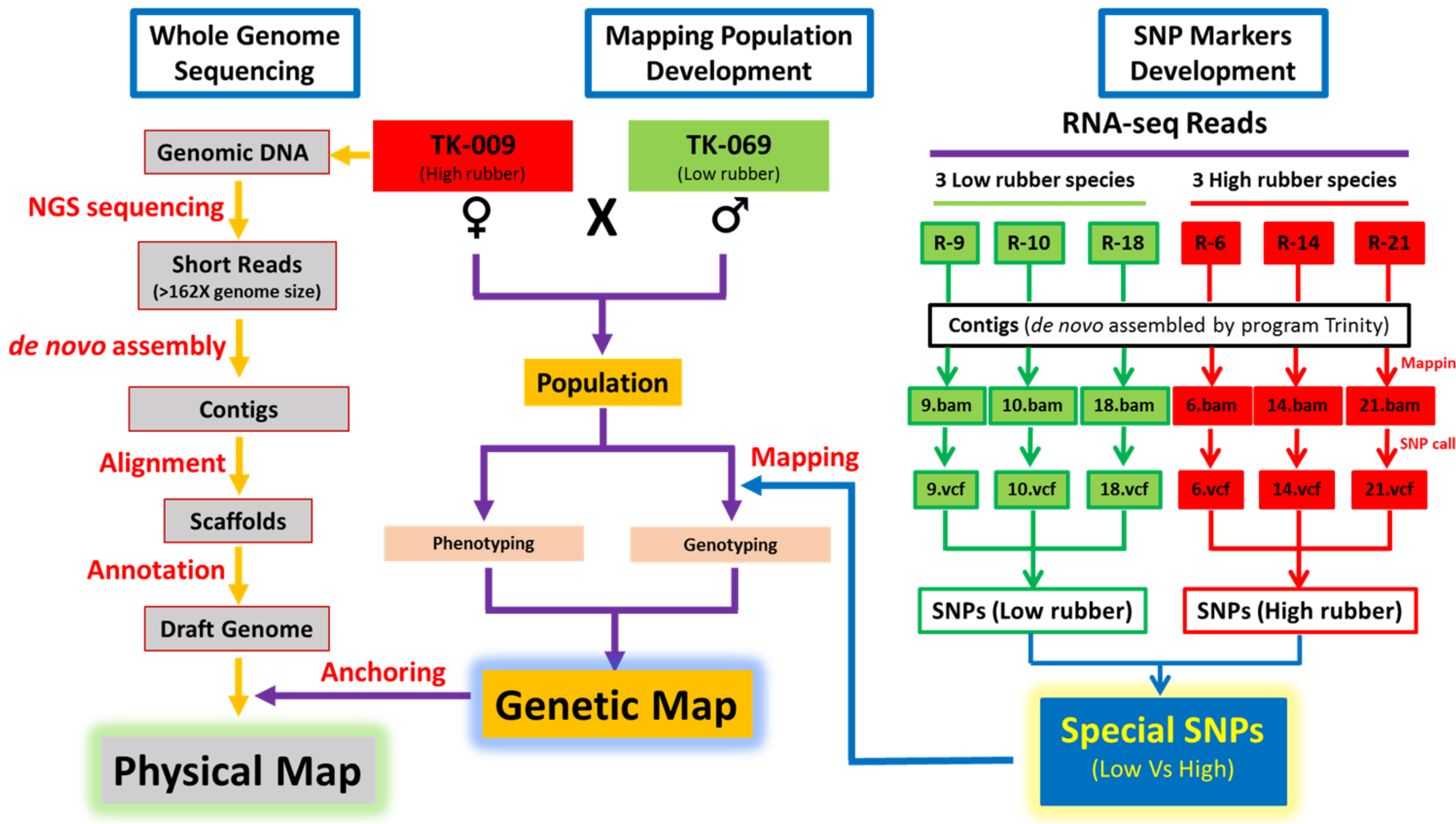

Fig 1. Workflow for TK whole genome sequencing, mapping population development, and SNP markers development

- ➢ **Mapping population development:** A high rubber plant (TK-009) was crossed with a low rubber plant (TK-069) for mapping population construction.
- ➢ **SNP markers development:** A total of 24 *T. kok-saghyz* plants were collected randomly from the field. Three plants with the highest rubber concentration (49.5, 42.3, 41.3 mg/gdw root, respectively) and three with the lowest (13.2, 13.3, 18.6 mg/gdw root, respectively) were chosen for RNA sequencing and SNP markers development.

RESULTS

➢ **Whole genome sequencing**

- ❖ More than 2 billion high quality reads (sequencing depth 192Ｘ) were produced by the Illumina HiSeq2500 platform (Table 1).

Table 1. High quality reads generated from TK-009 by Next Generation Sequencing

Library (insert size)	Clean Reads	Clean bases	Read length (bp)	Q20(%)
170 bp	725,363,346	90,670,418,250	125	95.27%
500 bp	431,542,664	53,942,833,000	125	90.68%
800 bp	325,866,654	40,733,331,750	125	89.61%
2 kb	309,682,758	38,710,344,750	125	92.91%
5 kb	215,861,616	26,982,702,000	125	92.80%
10 kb (1st run)	76,214,390	9,526,798,750	125	88.70%
10 kb (2nd run)	34,283,464	3,428,346,400	100	91.18%
20 kb (1st run)	55,364,098	6,920,512,250	125	90.04%
20 kb (2nd run)	11,370,982	1,137,098,200	100	91.95%
Total	2,185,549,972	272,052,385,350	-	-

- ❖ After *de novo* assembly by program SOAPdenovo2 (Luo et al., 2012), a total of 188,608 scaffolds were generated, covering 1.4 Gb, with an N50 scaffold size of 55.2 Kb (Table 2).

Table 2. Comparison of de novo assembly results between TK and Rubber tree (Rahman et al., 2013)

	T. kok-saghyz	Rubber tree
Scaffold number	188,608	608,017
Total scaffold length	1,4 Gb	1.1 Gb
Average scaffold length	7,631	1,840
Longest scaffold	1,077,596	531,465
N50	55,183	2,972

- ❖ A total of 23,760 protein-coding genes (with Start/Stop codon) were predicted, by preliminary whole genome annotation, using program Augustus-3.2.1 (Stanke et al., 2006). in which more than 80% of these genes were further confirmed by known plant genes in GenBank.

➢ **Mapping population development**

- ❖ At present, we have obtained 95 progeny after crossing (TK-009 x TK-069). Clones of these parents are being interbred to get at least 200 progeny.

➢ **SNP markers development**

- ❖ RNAseq produced more than 350 million of high quality reads from 3 high and 3 low rubber plants.
- ❖ After *de novo* assembly, around 55,532 contigs (length larger than 200 bp) were produced and are being used as a reference for SNP development.
- ❖ By SNP calling, 92,150 SNPs were detected in all three low rubber plants, 78,407 in all three high rubber plants (Fig. 2). Finally, 16,891 candidate SNP markers were developed with read depths of more than 10 in each sample.

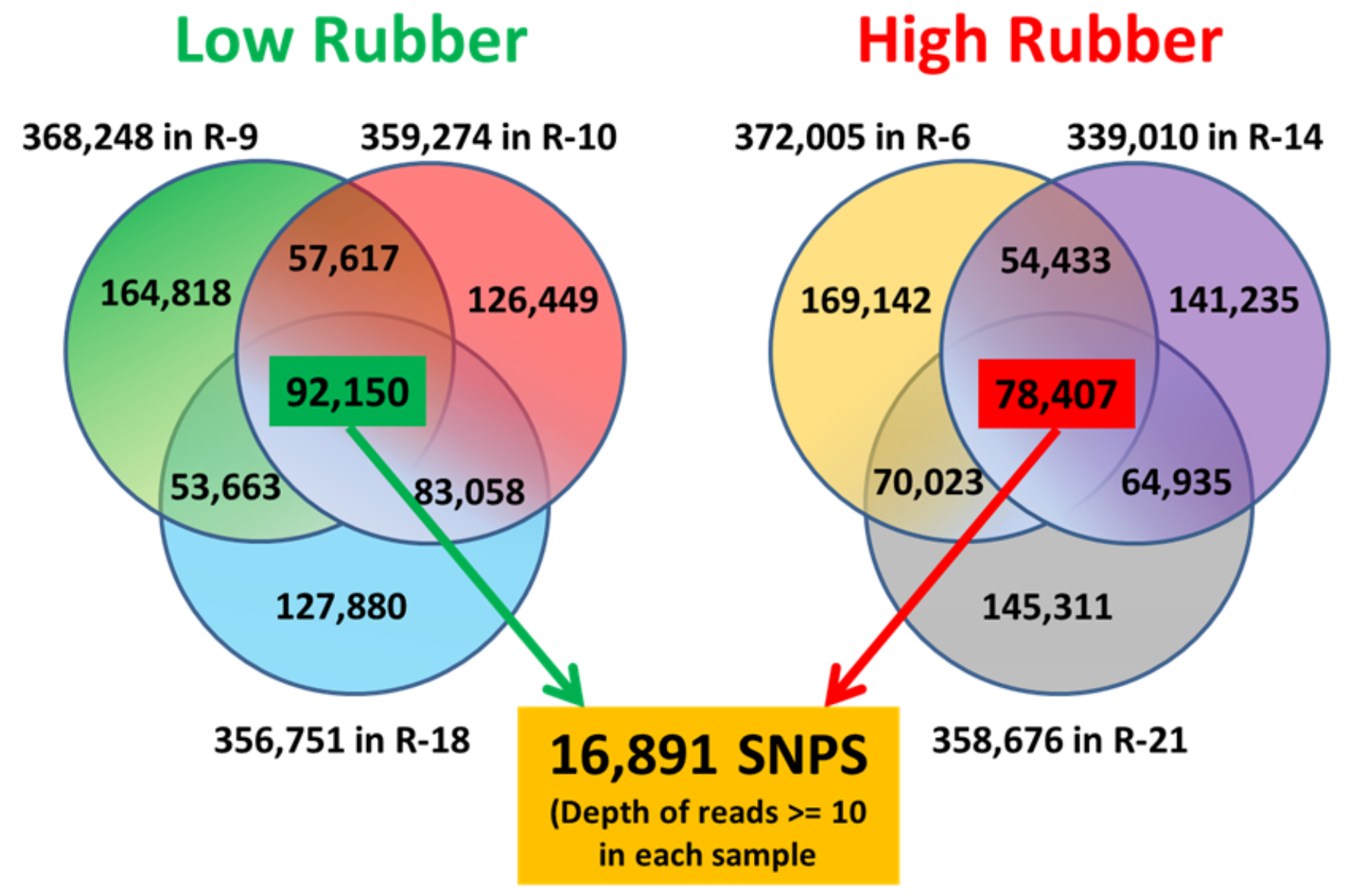

Fig 2. Candidate SNP markers were developed in high rubber/low rubber TK plants.

- ❖ To validate these markers, 20 SNP markers were selected randomly, 16 of which were validated in the parental population consisting of 30 individuals (rubber ranging from 22 to 150 mg/gdw root, Table 3).

Table 3. Validation of 20 SNP markers in 30 individuals TK plants.

SNP Name	Expected Length	Amplicon Size	AA	Aa	aa	Polymorphic/ monomorphic
SNP00052	150	~250	1	27	1	monomorphic
SNP00189	306	--	--	--	--	--
SNP00197	204	~200	10	10	8	polymorphic
SNP00213	223	~450	25	1	3	polymorphic
SNP00215	328	~650	3	20	6	polymorphic
SNP00375	250	~380	4	24	2	polymorphic
SNP00387	250	~380	--	--	--	--
SNP00406	412	~480	8	2	3	polymorphic
SNP00503	155	~150	0	30	0	monomorphic
SNP00521	311	~320	3	23	3	polymorphic
SNP00536	461	--	--	--	--	--
SNP00635	363	~600	6	16	7	polymorphic
SNP00822	451	~451	0	30	0	monomorphic
SNP00854	252	~252	2	22	3	polymorphic
SNP00915	216	~300	0	30	0	monomorphic
SNP01012	142	~250	6	9	9	polymorphic
SNP01024	393	~400	8	14	7	polymorphic
SNP01113	301	~520	4	8	6	polymorphic
SNP01158	301	~520	2	21	2	polymorphic
SNP01193	341	--	--	--	--	--

CONCLUSIONS & DISCUSSION

- ➢ Our results represent the first study of whole genome sequencing and EST-derived SNPs development in *Taraxacum kok-saghyz*, an alternative rubber crop.
- ❖ Through next generation sequencing, billions of high quality reads were produced from the high rubber plant TK-009. After *de novo* assembly, a draft genome of *T. kok-saghyz* was generated, which including 188,608 scaffolds, covering 1.4 Gb.
- ❖ By SNP calling, using RNAseq data from 3 high rubber TK plants and 3 low rubber plants, a total of 16,891 candidate SNP markers were developed.
- ❖ A small mapping population of 95 progeny was produced by crossing a high rubber plant TK-009 with a low rubber plant TK-069.
- ➢ A large mapping population is under construction.
- ➢ A high-density genetic map/physical map will be constructed.

REFERENCES

- **Collins-Silva, J. et al.** (2012). Phytochemistry **79,** 46-56.
- **Luo, R.B., et al.** (2012). Gigascience **1**:18, doi:10.1186/2047-217-X-1-18.
- **Rahman, A.Y. et al.** (2013). BMC genomics **14,** 75, doi: 10.1186/1471-2164-14-75.
- **Ribaut, J.M., and Hoisington, D.** (1998). Trends Plant Sci **3,** 236-239.
- **Stanke, M.A., and Morgenstern, B.** (2006). Genome Biology **7,** S11, doi:10.1186/gb-2006-7-s1-s11.
- **van Beilen, J.B., and Poirier, Y.** (2007). Critical reviews in biotechnology **27,** 217-231.

ACKNOWLEDGEMENTS

This work was funded by **CAPS** (Center for Applied Plant Sciences, Ohio State University). Thanks to OSC (Ohio Supercomputer Center) for providing hardware support for a partial of bioinformatics analysis.

THE OHIO STATE UNIVERSITY
COLLEGE OF FOOD, AGRICULTURAL, AND ENVIRONMENTAL SCIENCES