Vitis **55**, 95–98 (2016) DOI: 10.5073/vitis.2016.55.95-98

Assembly and analysis of the chloroplast genome of the North American hybrid grape 'Norton'

S. Howard[1)], J. Bouck[2)], and Wenping Qiu[1)]

[1)] Center for Grapevine Biotechnology, W. H. Darr School of Agriculture, Missouri State University, Mountain Grove, MO, USA
[2)] Creekside Computing, PLC, Mountain Home, MO, USA

Summary

'Norton' is a hybrid grape cultivar of North American *Vitis aestivalis*. In this report, the whole chloroplast (cp) genome was assembled and analyzed. Total length of the 'Norton' cp genome is 160,913 bp in comparison to 160,928 bp of the cp genome of *V. vinifera* 'Maxxa'. The sequence is highly conserved between the two cp genomes, with a sequence identity of 99 % and identical haplotypes for the three best-studied cpSSR markers (cpSSR3, cpSSR5 and cpSSR10). A total of 73 indels and 147 single nucleotide polymorphisms (SNPs) were found between the two genomes. Amino acid changes were identified in 10 coding regions. Six DNA fragments with predicted changes between the two cp genomes were amplified from 'Norton' and wild *V. aestivalis* by polymerase chain reaction (PCR). The sequencing of these DNA fragments verified assembled sequences except for one region that has low coverage in the assembly. Comparison of indels and SNPs in the verified regions indicated that 'Norton' cp genome contains signature variants that are found only in the cp genome of wild *V. aestivalis*. These sequence variants can be used as genetic markers for distinguishing grape cultivars and in breeding new cultivars.

Key words: grapevine, chloroplast, genome, variant, indel, SNP.

The wine grape cultivar 'Norton' has been grown in the Midwestern to Southeastern regions of the US for over 180 years. 'Norton' has robust resistance to major fungal pathogens (Fung *et al.* 2008) and tolerance to freezing and fluctuating temperatures, and thus 'Norton' requires lower inputs of fungicides and sustains profits under conditions not suitable for growing *Vitis vinifera* cultivars. 'Norton' berries have a high anthocyanin content and unique profiles of anthocyanins and proanthocyanidins (Hogan *et al.* 2009, Ali *et al.* 2011). 'Norton' is considered to have a genetic background of *Vitis aestivalis* and *V. vinifera* based on historic records and ampelography (Ambers 2012), the identity of its male and female parents, however, has yet to be determined. Understanding the chloroplast and nuclear genomic background of 'Norton' will help discover its parentage as well as origins of genetic loci of enabling disease resistance. The chloroplast (cp) genome is inherited in a non-meiotic mode from the maternal parent (Arroyo-García *et al.* 2002), and evolves slower than the nuclear genome. It consists of a single circular DNA molecule that has four defined regions: a large single copy (LSC) and a small single copy (SSC) region, and two copies of inverted repeats (IRa and IRb). Variations of cp genome sequences are used in genetic studies and phylogenetic analysis that provide evidence on the genetic background of economically important crops and valuable information for breeding new cultivars. To date, complete cp genomes of *V. vinifera* 'Maxxa' (Jansen *et al.* 2006), *V. vinifera* 'Meskhuri Mtsvane', 'Rkatsiteli', 'Saperavi' in Georgia (Tabidze *et al.* 2014), and *V. rotundifolia* (GenBank accession no: NC023790) have been published. However, only 63 partial cp sequences of *V. aestivalis* are available in the GenBank. Since the genome of *V. vinifera* 'Maxxa' was published first, it is used in this study as a reference genome for comparative genomics analysis. Its cp genome is 160,928 bp in length (Jansen *et al.* 2006). LSC and SSC regions contain 89,147 bp and 19,065, respectively, that are separated by two IRs of 26,358 bp.

The cp genome is highly conserved among *Vitis* species (Nicole *et al.* 2013), and possesses genetic differences that have been used to find ancestral parents of grape cultivars. Polymorphisms in simple sequence repeats (SSR) and intergenic regions of chloroplast genome allow development of SSR and single nucleotide polymorphisms (SNPs) markers for differentiating *Vitis* species (Arroyo-García *et al.* 2002). These molecular markers have been employed in analyzing ancestry and origins of grape cultivars (Snoussi *et al.* 2004, Arroyo-García 2006), and verifying cultivar identity (Castro *et al.* 2013). Four haplotypes at three cpSSR loci with seven alleles were present in the 92 *Vitis* cultivars samples (Arroyo-García *et al.* 2002). The majority of North American *V. berlandieri*, *V. riparia*, and *V. rupestris* accessions have haplotype B that is proposed to be an ancestral haplotype in current grape hybrid grape cultivars, suggesting their low evolutionary rate and close relatedness in maternal inheritance of cp genome. In this research note, we report the assembly and sequence analysis of the cp genome of 'Norton' grape. We present the comparative analysis of the 'Norton' cp genome with the reference cp genome of *V. vinifera* 'Maxxa' (GenBank accession no: NC007957).

Correspondence to: Dr. Wenping Qiu, Center for Grapevine Biotechnology, William H. Darr School of Agriculture, Missouri State University, 9740 Red Spring Road, Mountain Grove, MO 65711. Fax: +1-417-547-7540. E-mail: wenpingqiu@missouristate.edu

We describe genetic variables that can be used in the genetic and phylogenetic studies of grapevines and in the genotyping and breeding of grape cultivars.

Total DNA was extracted from young leaves of a single 'Norton' vine ('Norton-MSU1') in the Foundation Vineyard of the Missouri State Fruit Experiment Station (Mountain Grove, Missouri, USA). The DNA was sent to the SeqWright Inc. (Houston, TX) for the 454 sequencing that produced a total of 13,084,142 reads with an average read length of 314 bases. The total base count was 4,102,935,547, corresponding to an 8X coverage of the *Vitis* genome. The DNA also was sent to the University of Missouri DNA Core Facility (Columbia, MO) for Solexa sequencing on a HiSeq 2500 machine. Adaptor sequences were removed from all reads and low quality reads were filtered out by using the NGS QC Toolkit (McKenna *et al.* 2010). All reads then were aligned to the *V. vinifera* 'Maxxa' cp reference sequence using BWA (Li and Durbin 2010). The resulting alignment files were filtered to select only those reads aligning to the cp sequence using Samtools (Li *et al.* 2009), which were then converted back to fastq files using Picard tools (http://picard.sourceforge.net.). These 733,806 selected reads were assembled *de novo* into contigs using ABySS (Simpson *et al.* 2009). They were also used to produce reference guided contigs using Velvet Columbus (Zerbino *et al.* 2008). While contig fold coverage of the reference sequence varied between the *de novo* and template-guided assembly methods (44x and 77x), each set of contigs covered 100 % of the reference sequence and showed similar distribution patterns of contigs and repeat regions. Both sets of assembled contigs were visualized using the Integrated Genomic Viewer (IGV) (https://www.broadinstitute.org/igv/) and a consensus sequence with a length of 160,928 bp was exported (Norton-Consensus).

Variant calling was performed using the filtered reads as input to the Genome Analysis ToolKit (GATK, https://www.broadinstitute.org/gatk/) and the tool "Unified Genotyper" with –glm switch to find both SNPs and indels in 'Norton' cp genome in comparison to the 'Maxxa' cp reference sequence. The unfiltered output contained 426 variants, 73 of which are indels. GATK uses a modified phred quality score included in the variant output file, and a cutoff value of 1,000 was used to avoid false positive variant calls. This reduced the number of SNPs to 147, and did not affect the number of indels. Ninety nine of the 147 SNPs fall into intergenic regions, 48 are located in coding regions. About half of the 73 indels were 1 bp long, only 2 indels were longer than 10 bp. Overall the predicted variants are located only in the SSC and LSC regions of the chloroplast genome as shown in the Figure, which was produced with the program GenomeVx (http://wolfe.ucd.ie/GenomeVx/). Similar distribution of variants has been shown in other plant species, including switchgrass (Young *et al.* 2011) and poplar (Kersten *et al.* 2016). Using the tool "Fasta Alternate Reference Maker" (GATK, Broad Institute), the predicted variants were incorporated into the assembly sequence. The resulting 'Norton' chloroplast sequence is 160,903 bp long.

In a phylogenetic relationship study of 89 wild *Vitis* species using four chloroplast intergenic spacer regions (*trnH-psbA, trnK-rps16, trnF-nahJ*, and *rpl32-trnL*) and the nuclear gene *RPB2-I*, Zecca *et al*. (2012) found that the Asian *Vitis* species are closely related to *V. vinifera* subsp. *sylvestris* while the American species form three major clusters (Zecca *et al.* 2012), suggesting that geographical isolation plays a major role in speciation in the *Vitis* genus. By using the newly sequenced 'Norton' cp genome, we investigated the phylogenetic relationship of 'Norton' among the wild *Vitis* species. We retrieved nucleotide popsets 339648230, 339648402, 339648068 and 339647980 of the *trnH-psbA*, *trnK-rps16*, *trnF-ndhJ* and *rpl32-trnL* regions from the GenBank. Each set of sequences was processed by eliminating the sequences of the non-*Vitis* species and duplicate sequences. The corresponding sequences of the four intergenic cp spacers of 'Norton' and 'Maxxa' were added to the data set. Four phylogenetic trees were generated using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/). We found that 'Norton' is clustered closely to with wild *V. aestivalis* based on the sequences of two *rpl31-trn* and *trnF-ndhJ* intergenic regions, but closely to *V. vinifera* subsp. *sylvestris* for the intergenic regions *trnH-psbA* and *trnK-rps16*.

The 'Maxxa' reference cp genome sequence and the 'Norton' cp genome sequence were aligned to each other using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) to show the locations of SNPs and indels (Figure). Both sequences were annotated using DOGMA (http://dogma.ccbb.utexas.edu/), followed by manual annotation. The order of 113 genes was conserved with their location differing by a maximum of 30 bp, mostly due to intergenic indels. Coding regions of cp genes are conserved with the exception of one shorter ycf1 gene copy. Overall there are 10 coding regions in which amino acid substitutions occurred in comparison with the reference cp sequence of 'Maxxa': matK, atpF, rpoC2, psbC, atpB, rbcL, psbT, petD, ndhF, and ycf1.

To verify some of the predicted variants and confirm the assembled sequence, we selected six regions along the entire sequence, and amplified the DNA fragments from 'Norton' and wild *V. aestivalis* at these loci by PCR and sequenced these regions again. Sequences of the six DNA fragments verified variants between 'Maxxa', 'Norton' and *V. aestivalis* cp genomes (Table). Of the 5 SNPs, two are identical between 'Maxxa' and wild *V. aestivalis*, three are the same between 'Norton' and *V. aestivalis*. The insert of 'AAA' at locus 4,424 and 'TTGACTATAA' at locus 115,461 are found in both 'Norton' and *V. aestivalis*, while the insert 'GCCTGTG' (at locus 72,702) is unique to 'Maxxa', and the insert 'TTATTT' at 13,325 is found in both *V. aestivalis* and 'Maxxa'. These sequences can be used as signatures in identifying grape cultivars. The region 115,400 to 115,600 shows high variability, including several SNPs, and one long variant of 17 bp that includes a 1 bp indel. For the first 11 bp at locus 115,458 "GGTTTGGGGAT" is found only in "Maxxa" while "TTGACTATAA" is in both 'Norton' and *V. aestivalis*, but absent in the assembled sequence. This is one of three regions with coverages of below 30X in the assembly. The remaining 6 bp of the long variant were 'TTGCAT' in 'Maxxa', 'TTGAAT' in the assembled sequence and 'ACGATG' in both Norton and *V. aestivalis*. Since 'TTGACTATAA' and 'ACGATG' were confirmed by sequencing the PCR-amplified DNA fragment, we incorporated the confirmed sequences into the assembled Norton

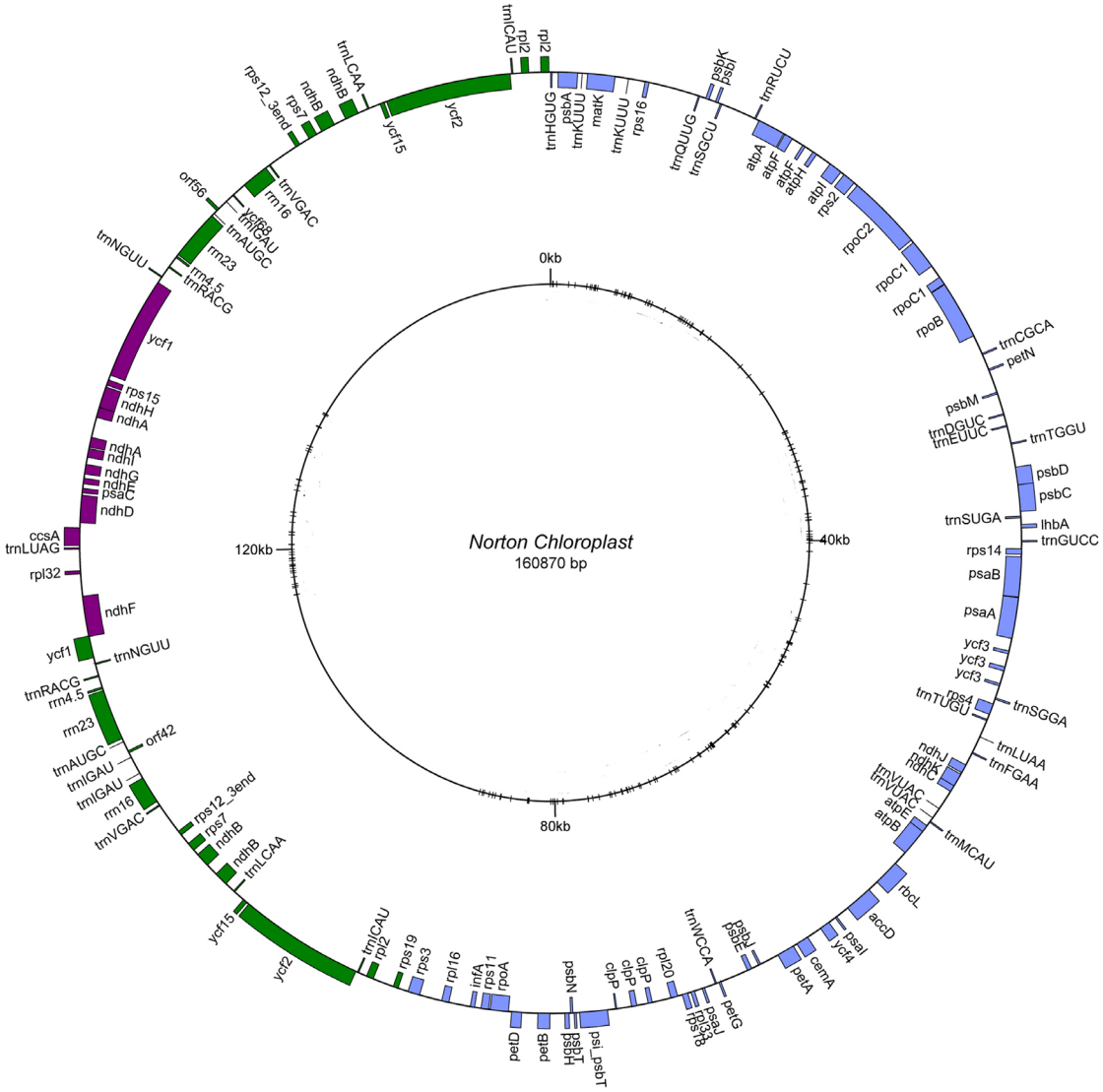

Figure: Graphic representation of the 'Norton' chloroplast genome. Locations of indels and SNPs are showed on the inner circle. Large single copy (LSC) region is in blue, inverted repeat (IRb and IRa) regions are in green, and small single copy (SSC) region is in purple. Note: A list of 28 genes were not included on the graph for readability. These genes are trnG-GCC, trnY-GUA, trnM-CAU, trnfM-CAU, trnL-UAA, trnT-GGU, psbL, psbF, petL, trnP-UGG, trnP-GGG, rps12, rpl36, rps8, rpl14, rpl22, rpl2, rpl23, ycf68, trnA-UGC, orf56, rrn5, orf188, rrn5, orf42, trnA-UGC, trnI-GAU, rpl2.

Table

Verification of genetic variants in the six PCR amplified DNA fragments of chloroplast genome of 'Norton' and wild *Vitis aestivalis* in comparison with the corresponding fragments in the chloroplast genome of *Vitis vinifera* 'Maxxa'

PCR DNA fragment	Locations of fragment	Locus	Variant type	Maxxa	Norton		*V. aestivalis*
					assembled	verified	
I	4200-5022	4424	indel	---	AAA	AAA	AAA
		4480	SNP	A	G	G	G
II	13200-13670	13289	indel	--	--	--	AA
		13325	indel	TTATTT	------	------	TTATTT
		13486	SNP	T	A	A	T
		13622	SNP	G	A	A	G
III	52660-52990	52690	indel	--	TT	TT	--
IV	72500-72900	72702	indel	GCCTGTG	-------	-------	-------
V	115350-115650	115461	variants	GGTTTGGGGAT	----------	TTGACTAT-AA	TTGACTAT-AA
		115464	indel	G	-	-	-
VI	118150-118450	118286	SNP	A	C	C	C
		118354	SNP	C	T	T	T

cp genome to produce a total of 160,913 bp that has been submitted to GenBank. The analysis of sequences indicates that the 'Norton' cp genome has unique sequences 'AAA' and 'TTGACTATAA' that are found only in wild *V. aestivalis*, suggesting that the 'Norton' cp genome might be derived from a female *V. aestivalis* parent.

Conclusions

The whole cp genome of 'Norton' grape was assembled and analyzed, it contains 160,913 bp that shared 99 % identity with the cp genome of *V. vinifera* 'Maxxa'. Unique genetic variants of indels and SNPs were discovered on the cp genome of 'Norton' that can be used as genetic markers in the breeding program of new grape cultivars and also for identifying genetic background of current grape cultivars. Sequencing of the whole cp genome of a wild *V. aestivalis* will help verify the parentage of a very important North American grape cultivar 'Norton' (GenBank Accession Number: NC029454).

References

Ali, M.; Howard, S.; Chen, S.; Wang, Y.; Yu, O.; Kovacs, L.; Qiu, W.; 2011: Berry skin development in Norton grape: Distinct patterns of transcriptional regulation and flavonoid biosynthesis. BMC Plant Biol. **11**, 7.

Ambers, C. P.; 2012: A historical hypothesis on the origin of the Norton grape. J. Wine Res. **24**, 85-95.

Arroyo-García, R.; Lefort, F.; de Andrés, M. T.; Ibáñaez, J.; Borrego, J.; Jouve, N.; Cabello, F.; Martínez-Zapater, J. M.; 2002: Chloroplast microsatellite polymorphisms in *Vitis* species. Genome **45**, 1142-1149.

Arroyo-García, R.; Ruiz-García, L.; Bolling, L.; Ocete, R.; López, M. A.; Arnold, C.; Ergul, A.; Söylemezoğlu, G.; Uzun, H.I.; Cabello, F.; Ibáñez, J.; Aradhya, M. K.; Atanassov, A.; Atanassov, I.; Balint, S.; Cenis, J. L.; Costantini, L.; Goris-Lavets, S.; Grando, M. S.; Klein, B. Y.; McGovern, P. E.; Merdinoglu, D.; Pejic, I.; Pelsy, F.; Primikirios, N.; Risovannaya, V.; Roubelakis-Angelakis, K. A.; Snoussi, H.; Sotiri, P.; Tamhankar, S.; This, P.; Troshin, L.; Malpica, J. M.; Lefort, F.; Martinez-Zapater, J. M.; 2006: Multiple origins of cultivated grapevine (*Vitis vinifera* L. ssp. *sativa*) based on chloroplast DNA polymorphisms. Mol. Ecol. **15**, 3707-3714.

Castro, I.; Pinto-Carnide, O.; Ortiz, J.; Martín, J.; 2013: Chloroplast genome diversity in Portuguese grapevine (*Vitis vinifera* L.) cultivars. Mol. Biotechnol. **54**, 528-540.

Fung, R. W. M.; Gonzalo, M.; Fekete, C.; Kovacs, L. G.; He, Y.; Marsh, E.; McIntyre, L. M.; Schachtman, D. P.; Qiu, W. P.; 2008: Powdery mildew induces defense-oriented reprogramming of the transcriptome in a susceptible but not in a resistant grapevine. Plant Physiol. **146,** 236-249

Hogan, S.; Zhang, L.; Li, J.; Zoecklein, B.; Zhou, K.; 2009: Antioxidant properties and bioactive components of Norton (*Vitis aestivalis*) and Cabernet Franc (*Vitis vinifera*) wine grapes. LWT-Food Sci. Technol. **42**, 1269-1274.

Jansen, R.; Kaittanis, C.; Saski, C.; Lee, S. B.; Tomkins, J.; Alverson, A.; Daniell, H.; 2006: Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol. Biol. **6**, 32.

Kersten, B.; Faivre Rampant, P.; Mader, M.; Le Paslier, M. C.; Bounon, R.; Berard, A.; Vettori, C.; Schroeder, H.; Leplé, J. C.; Fladung, M.; 2016: Genome sequences of *Populus tremula* chloroplast and mitochondrion: Implications for holistic poplar breeding. PLoS ONE **11**, e0147209.

Li, H.; Durbin, R.; 2010: Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics **26**, 589-595.

Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Genome Project Data Processing, Subgroup; 2009: The sequence alignment/map format and SAM tools. Bioinformatics **25**, 2078-2079.

McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; DePristo, M. A.; 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. **20**, 1297-1303.

Nicole, S.; Barcaccia, G.; Erickson, D.; Kress, J.; Lucchin, M.; 2013: The coding region of the UFGT gene is a source of diagnostic SNP markers that allow single-locus DNA genotyping for the assessment of cultivar identity and ancestry in grapevine (*Vitis vinifera* L.). BMC Res. Notes **6**, 502.

Simpson, J. T.; Wong, K.; Jackman, S. D.; Schein, J. E.; Jones, S. J. M.; Birol, İ.; 2009: ABySS: A parallel assembler for short read sequence data. Genome Res. **19**, 1117-1123.

Snoussi, H.; Slimane, M. H. B.; Ruiz-García, L.; Martínez-Zapater, J. M.; Arroyo-García, R.; 2004: Genetic relationship among cultivated and wild grapevine accessions from Tunisia. Genome **47**, 1211-1219.

Tabidze, V.; Baramidze, G.; Pipia, I.; Gogniashvili, M.; Ujmajuridze, L.; Beridze, T.; Hernandez, A. G.; Schaal, B.; 2014: The complete chloroplast DNA sequence of eleven grape cultivars. Simultaneous resequencing methodology. J. Int. Sci. Vigne Vin **48**, 99-109.

Young, H. A.; Lanzatella, C. L.; Sarath, G.; Tobias, C. M.; 2011: Chloroplast genome variation in upland and lowland switchgrass. PLoS ONE **6**, e23980.

Zecca, G.; Abbott, J. R.; Sun, W. B.; Spada, A.; Sala, F.; Grassi, F.; 2012: The timing and the mode of evolution of wild grapes (*Vitis*). Mol. Phylogenet. Evol. **62**, 736-747.

Zerbino, D. R.; Birney, E.; 2008: Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. **18**, 821-829.

Received November 13, 2015