Chloroplast genome of serrated tussock (*Nassella trichotoma*): structure and evolution

Aisuo Wang[1,2], Hanwen Wu[1,2] and David Gopurenko[1,2]
[1] NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, PMB, Wagga Wagga, New South Wales 2650, Australia
[2] Graham Centre for Agricultural Innovation, Locked Bag 588, Wagga Wagga, New South Wales 2678, Australia
(aisuo.wang@dpi.nsw.gov.au)

Summary The genus *Nassella* contains many economically and environmentally important weed species in Australia, such as *Nassella trichotoma* (Nees) Hack. ex Arechav. (serrated tussock) and *Nassella neesiana* (Trin. & Rupr.) Barkworth (Chilean needle grass), which are both Weeds of National Significance (WoNS). Availability of *Nassella* chloroplast genome sequences can provide a versatile tool for identifying novel gene targets for DNA barcoding of these weeds species. We report here the first chloroplast genome of *Nassella* (*N. trichotoma*) obtained through de novo assembly of Illumina paired-end reads produced by total DNA sequencing. The *N. trichotoma* chloroplast genome is 112,102 bp in size, encodes 140 genes including 99 protein-coding genes, 37 tRNA genes and 4 ribosofmal RNA genes. The total size of intergenic regions within the genome is up to 48,764 bp. The total GC content of the genome is 37.88%, which is relatively lower than that of the reported genomes within the *Stipeae* tribe. A comparison of this chloroplast genome with other *Stipeae* chloroplast genomes (*Oryzopsis asperifolia* Michx., *Piptochaetium avenaceum* (L.) Parodi, *Achnatherum hymenoides* Roem. & Schult., *Stipa lipskyi* Roshev. and *Stipa purpurea* Griseb.) provides new insights into their chloroplast genome evolution. Our future work will report genomes of other *Nassella* species as a means to identify novel DNA barcode regions useful for distinguishing these species.

Keywords DNA barcoding, invasive weeds, *Stipeae*, Next Generation Sequencing, Illumina.

INTRODUCTION

The genus *Nassella* contains many economically and environmentally important weeds species in Australia, such as *Nassella trichotoma* (Nees) Hack. ex Arechav. (Serrated tussock) and *Nassella neesiana* (Trin. & Rupr.) Barkworth (Chilean needle grass). Effective management of these invasive weeds relies on correct identification of them at all growth stages. Traditional morphological identification plays a crucial role in identifying these weeds species, but relies on the availability of flowering materials of the weeds and expert taxonomic knowledge. A genetic approach such as DNA barcoding provides a hope for identifying these weeds at all growth stages (Syme *et al.* 2013, Wang *et al*. 2014). Nevertheless, a lack of universal and robust markers has been hindering the development of DNA barcoding technology. Screening ideal DNA barcoding markers to identify *Nassella* species and other invasive weeds thus becomes a necessity.

The chloroplast genome is a major source for markers applied in plant DNA barcoding technology. Most of the well-known DNA barcoding markers for plants such as *rbcL, matK, psbK–psbI, trnH–psbA, atpF–atpH, rpoB* and *rpoC1* (CPW Group 2009) were selected from sequences of chloroplast genome of plants. With the fast development of next-generation sequencing technology (NGS) in recent years, it becomes increasingly practical and cheaper to sequence a whole grass chloroplast genome. The availability of whole chloroplast genome sequences of invasive weeds will undoubtedly facilitate the marker selection process that has been a bottleneck for weeds DNA barcoding technology.

Here we report the chloroplast genome sequences of *N. trichotoma*, the first chloroplast genome for the *Nassella* genus, obtained through de novo assembly of Illumina paired-end reads. Information on this *Nassella* chloroplast genome sequence will be instrumental for DNA barcoding, phylogeny analysis and other related research for weeds of the *Nassella* genus.

MATERIALS AND METHODS

Sample collection Fresh *N. trichotoma* leaves were collected from Wagga Wagga NSW in Australia (34° 58' 40.7", 147° 26' 19.5"). The voucher samples have been deposited into Wagga Wagga Agricultural Institute (WWAI) (voucher number: ww19856).

Genomic DNA extraction and sequencing Total genomic DNA was extracted from fresh leaves of *N. trichotoma* using traditional phenol chloroform protocol (Sambrook and Russell 2006) with modifications.

Basically the plant tissue was digested in CTAB extraction buffer (100mM Tris-HCl (pH 7.5), 25mM EDTA, 1.5 M NaCl, 2% (w/v) CTAB) at 55°C overnight before being extracted twice with chloroform: isoamyl alcohol 24:1. About 10% volume of 5M NaCl was added into the supernatant before being precipitated with 100% ethanol. DNA quality was measured by running gel electrophoresis (1% agarose, Bioline) and using NanoDrop 2000 Spectrophotometer (Thermo Scientific, Australia). Only DNA samples that meet the quality check criteria (sample concentration: ≥30 ng μL^{-1}; total DNA quantity ≥3 μg; no or partial degradation; DNA absorbance ratios OD260/280:1.8-2.0 and OD260/230: 2.0-2.2) were sent to Beijing Genomics Institute (BGI) in Hong Kong for NGS sequencing. At BGI, high quality DNA samples were applied in library construction (insert size 500 bp) before being sequenced in an Illumina HiSeq2000 platform.

Chloroplast genome assembly and annotation Illumina sequencing data were de novo assembled using SOAPdenovo-Trans (Xie *et al.* 2014). Annotation of the *N. trichotoma* chloroplast genome was performed using DOGMA with default settings (Wyman *et al.* 2004). In addition, all tRNA genes were further verified online using tRNAscan-SE search server (Lowe and Eddy 1997) (http://lowelab.ucsc.edu/tRNAscan-SE/). The circular *N. trichotoma* chloroplast genome map was drawn using OGDraw v1.2 (Lohse *et al.* 2007).

Genome analysis The chloroplast genome of *N. trichotoma* was comparatively studied with the chloroplast genomes of other species of the *Stipeae* tribe, including *Stipa purpurea* (NC_029390.1), *Stipa lipskyi* (NC_028444.1), *Piptochaetium avenaceum* (NC_027483.1), *Oryzopsis asperifolia* (NC_027479.1) and *Achnatherum hymenoides* (NC_027464.1) (originally defined as *Stipa hymenoides*). All of these chloroplast genomes were re-analysed in DOGMA for better comparison with the results of *N. trichotoma*. GMATo (Wang *et al.* 2013) was applied to search Simple Sequence Repeats (SSR) existing in the genomes. Whole genomes across these six species were aligned using progressive MAUVE implemented by MAUVE v2.3.1 software (Darling *et al.* 2004).

RESULTS AND DISCUSSION

General features of the *N. trichotoma* chloroplast genome The obtained *N. trichotoma* chloroplast genome in present study is 112,102 bp in size, encodes 118 unique genes including 85 protein-coding genes, 29 tRNA genes and 4 ribosomal RNA genes. Some genes were duplicated in the genome, including *atpF, ndhA , ndhB , ndhH , orf56 , rpl2 , rpl23 , ycf1 , ycf2, ycf3, trnA-UGC , trnI-GAU , trnL-UAA , trnM-CAU , trnT-GGU* and *trnV-UAC*. This increases the total gene number of *N. trichotoma* chloroplast genome to 140. The total GC content of the genome is 37.88%, which is relatively lower than that of the reported genomes within the *Stipeae* tribe (Table 1).

AT-rich regions in the *N. trichotoma* chloroplast genome are intergenic (66.17%) and protein-coding (60.24%), while rRNAs (45.3%) and tRNAs (48.53%) have a much lower AT content. This pattern is similar to other five *Stipeae* species (Table 1). AT content in genomic regions were reported to be associated with the dynamics of repeats, the codon bias of chloroplast protein-coding genes and the regulation of gene expression (Rouwendal *et al.* 1997, Morton 2003). Further analyses are thus needed to investigate the links between AT content and its significance to the functions of *N. trichotoma*.

The annotated chloroplast genome of *N. trichotoma* is shown in (Figure 1). The genome is similar to that of other *Stipeae* species but shorter in size (25,752 bp shorter than the largest chloroplast genome of *Stipa lipskyi*). Only one inverted repeat (IRA) was identified to separate the long single copy section (LSC) and the short single copy section (SSC) in the genome, which is different from the remaining *Stipeae* species.

Genome comparison The *N. trichotoma* chloroplast genome was aligned with the chloroplast genomes of other *Stipeae* species to compare the organization of these genomes. Three major locally collinear blocks (LCBs) across these genomes were identified (Figure 2). These blocks suggest a high level of similarity in genome organization of *S. purpurea, A. hymenoides, P. avenaceum* and *O. asperifolia*. By comparison, the gene arrangements of *S. lipskyi* and *N. trichotoma* are different from these four chloroplast genomes as several inversions and translocations events occurred through these two chloroplast genomes when compared with that of the other four *Stipeae* species. The alignment appears to suggest a close evolutionary relationship between *N. trichotoma* and *S. lipskyi*, and a more distant relationship between *N. trichotoma* and the remaining *Stipeae* species. Nevertheless, further research is required to verify this conclusion.

It was noted that the gene content of the five *Stipeae* species chloroplast genomes are different from those previously published in NCBI (Table 1), which are probably due to differences in annotation methodology. There were three major differences between *N. trichotoma* and the remaining *Stipeae* species. Firstly, the copies of duplicated genes in *N. trichotoma* were only half of the corresponding gene

Table 1. Genome features of the chloroplast genome of six species from the *Stipeae* tribe (gene numbers shown in parentheses are the results of DOGMA re-analyses).

Characteristics	*Nassella trichotoma*	*Oryzopsis asperifolia*	*Piptochaetium avenaceum*	*Achnatherum hymenoides*	*Stipa lipskyi*	*Stipa purpurea*
GenBank Accession No.	TBA	NC_027479	NC_027483	NC_027464	NC_028444	NC_029390.1
Size (bp)	112,102	134,281	137,701	137,742	137,854	137,370
GC content (%)	37.88	38.8	38.8	38.8	38.8	38.8
Total number of genes	140	130 (175)	129 (179)	130 (181)	129 (181)	130 (181)
Total number of unique genes	118	119	119	119	119	119
Protein-coding genes	99	84 (118)	83 (122)	84 (124)	81 (124)	84 (124)
Ribosomal RNAs	4	8	8	8	8	8
Transfer RNAs	37	38 (49)	38 (49)	38 (49)	40 (49)	38 (49)
Protein-coding genes (bp)	58974	63717	66447	66510	66201	66231
Ribosomal RNAs (bp)	4596	9192	9192	7972	9192	9192
Transfer RNAs (bp)	2442	3110	3110	3112	3110	3110
Intergenetic regions (bp)	48764	59333	60234	60240	60198	59942
AT content (%)						
Genome	62.04	61.23	61.23	61.18	61.20	61.21
Protein-coding genes	60.24	60.32	60.35	60.22	60.25	60.30
Ribosomal RNAs	45.3	45.3	45.28	45.48	45.28	45.28
Transfer RNAs	48.53	48.04	48.07	48.04	47.91	47.91
Intergenic regions	66.17	65.3	65.25	65.26	65.23	65.26

copies in the five *Stipeae* species. These include 11 protein-coding genes (*ndhB, orf42, orf56, rpl2, rps12, rps12_3end, rps15, rps19, rps7, ycf1, ycf15),* eight tRNA genes *(trnA-UGC, trnH-GUG, trnI-CAU, trnI-GAU, trnL-CAA, trnN-GUU, trnR-ACG, trnV-GAC*) and all rRNA genes (*rrn16, rrn23, rrn4.5, rrn5*). Secondly, two genes were not present in the five *Stipeae* species but were found in the chloroplast genome of *N. trichotoma* (*psbA* and *psbG*). Thirdly, another two genes (*matK* and *trnK-UUU*) were not found in *N. trichotoma* but were present in the five chloroplast genomes of *Stipeae* species.

Simple Sequence Repeats Three Simple Sequence Repeats (SSR) were identified in the chloroplast genome of *N. trichotoma*, including a 'TC' 5× repetitions between the range of 12636 and 12645, a 'TA' 5× repetitions between the range of 30178 and 30187 and a 'AT' 5× repetitions between the range of 89616 and 89625. The SSR numbers in *N. trichotoma* was less than that of the five *Stipeae* species [*A. hymenoides* (4), *O. asperifolia* (7), *P. avenaceum* (7), *S. lipskyi* (7), and *S. purpurea* (7)].

Intergenic regions *Nassella trichotoma* contains 125 intergenic regions, which is less than other *Stipeae* species (160 in *O. asperifolia*; 162 in *P. avenaceum* and *A. hymenoides*; 163 in *S. lipskyi* and 164 in *S. purpurea*). Among them, seven intergenic regions are greater than 1000 bp in size, whist 29 other intergenic regions have sequences greater than 500 bp in size. This provides a valuable resource for selecting robust DNA barcodes for differentiation of *N. trichotoma* from other grasses because intergenic regions have been widely used as plastid barcodes for species differentiation (Dong *et al.* 2012, Suzuki *et al.* 2014), particularly for those with highly variable sequence areas and in good size.

In summary, the present study represents the first attempt to sequence the chloroplast genome of *Nassella* using NGS technology. In comparison to chloroplast genomes of five published *Stipeae* species, the available data indicates that gene re-arrangements and inversions have likely occurred in the evolution of *N. trichotoma* chloroplast genome. Our data provided valuable resources for selecting robust markers for DNA barcoding of *Nassella* weeds species, and

made it easier to sequence the chloroplast genomes of other *Nassella* species (as the current data could be used as a reference genome for genome assembly). Nevertheless, more research is required for the chloroplast genome of *N. trichotoma* as the evidences from the present study (e.g. shorter genome size and reduced gene numbers of the genome) suggested that the obtained chloroplast genome of *N. trichotoma* is likely to be incomplete.

ACKNOWLEDGMENTS

The authors would like to thank Dr Shanlin Liu and Dr Chentao Yang at Beijing Genomics Institute (BGI) for technical support of the NGS sequencing; the authors would also like to acknowledge Mr Jeremy Crocker at Wagga Wagga City Council for assisting to collect *N. trichotoma* samples.

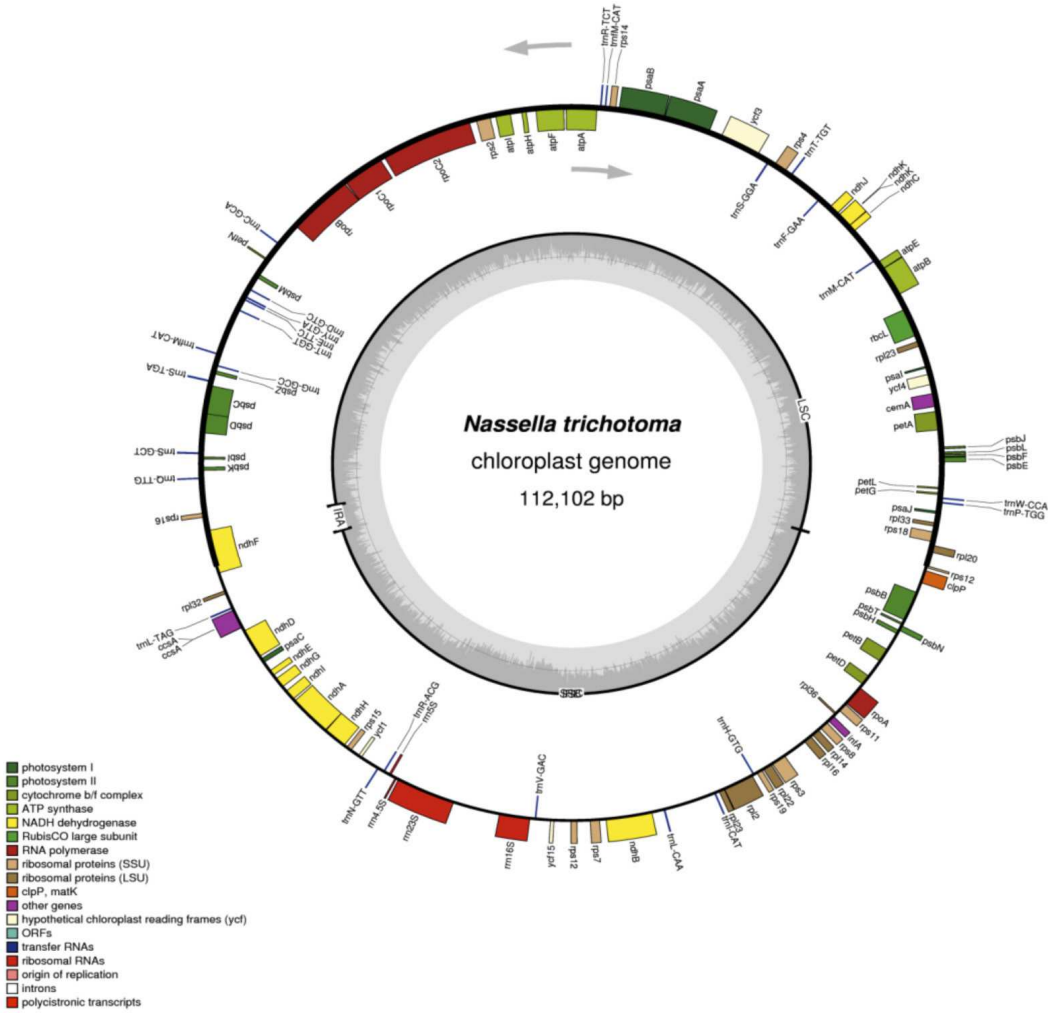

Figure 1. Sequence map of the *Nassella trichotoma* chloroplast genome. Genes drawn inside of the circle are transcribed clockwise, while genes shown outside of the circle are transcribed counter-clockwise. Genes belonging to different functional groups are colour-coded. The darker grey in the inner circle indicates GC content, while the lighter grey corresponds to AT content.

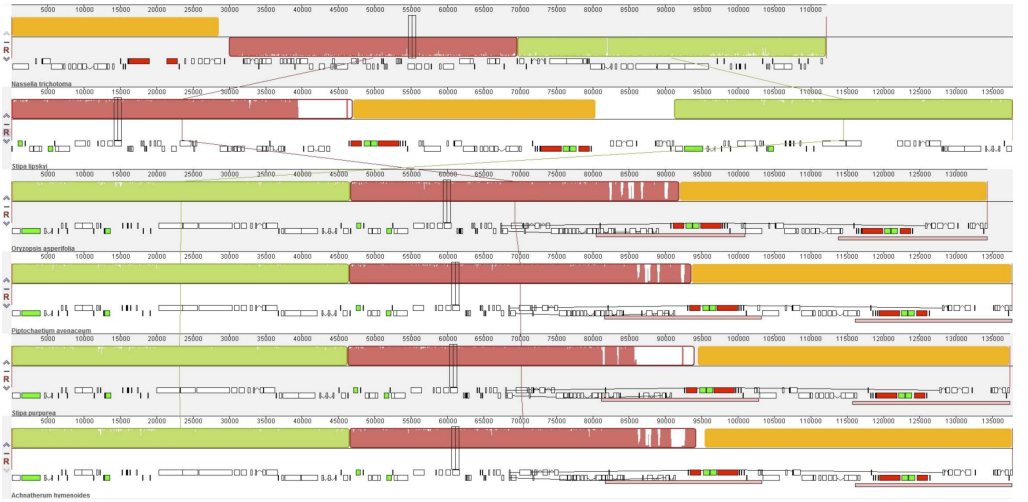

Figure 2. Alignment of *N. trichotoma* chloroplast genome with the chloroplast genomes of five *Stipeae* species (*Stipa purpurea, Stipa lipskyi, Piptochaetium avenaceum, Oryzopsis asperifolia* and *Achnatherum hymenoides*). Three major locally collinear blocks (LCBs) were labelled with three different colours (orange, red and green).

REFERENCES

Darling, A.C., Mau, B., Blattner, F.R. and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* 14, 1394-1403.

Dong, W.P, Liu, J., Yu, J., Wang, L. and Zhou, S.L. (2012). Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS One* 7 (4): e35071. doi: 10.1371/journal.pone.0035071

Group CPW. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences U.S.A.* 106, 12794-97.

Lohse, M., Drechsel, O. and Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics* 52, 267-74.

Lowe, T.M. and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* 25, 955-64.

Morton, B.R. (2003). The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. *Journal of Molecular Evolution* 56, 616-29.

Rouwendal, G.J., Mendes, O., Wolbert, E.J. and Douwe de Boer, A. (1997). Enhanced expression in tobacco of the gene encoding green fluorescent protein by modification of its codon usage. *Plant Molecular Biology* 33, 989-99.

Sambrook, J. and Russell, D.W. (2006). Purification of nucleic acids by extraction with phenol: chloroform. *Cold Spring Harbor Protocols* 2006.

Suzuki, J.Y., Matsumoto, T.K., Keith, L.M. and Myers, R.Y. (2014). The chloroplast psbK-psbI intergenic region, a potential genetic marker for broad sectional relationships in *Anthurium*. *Hortscience* 49, 1244-52.

Syme, A.E., Udovicic, F., Stajsic, V. and Murphy, D.J. (2013). A test of sequence-matching algorithms for a DNA barcode database of invasive grasses. *DNA Barcodes* 1, 19-26.

Wang, A, Gopurenko, D., Wu, H., Stanton, R. and Lepschi, B. (2014). DNA barcoding for identification of exotic grass species present in eastern Australia. Proceedings of the 19th Australasian Weeds Conference – Science, Community and Food Security: the Weed Challenge, 1–4 September: Hobart, Tasmania, pp. 444-7.

Wang, X., Lu, P. and Luo, Z. (2013). GMATo: a novel tool for the identification and analysis of microsatellites in large genomes. *Bioinformation* 9, 541-4.

Wyman, S.K., Jansen, R.K. and Boore, J.L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20, 3252-5.

Xie, Y., Wu G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., *et al.* (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30, 1660-6.