Complete plastid genome assembly of invasive plant, *Centaurea diffusa*

Kathryn G. Turner and Christopher J. Grassa

University of British Columbia

June 8, 2014

Abstract

New genomic tools are needed to elucidate the evolution of invasive, non-model organisms. Here we present the completed plastome assembly for the problematic invasive weed, *Centaurea diffusa*. This new tool represents a significant contribution to future studies of the ecological genomics of invasive plants, particularly this weedy genus, and studies of the Asteraceae in general.

Introduction

Invasive species offer ample incentive and opportunity to address some of the major questions in evolution, ecology, and genetics. The direct costs of weed control and indirect costs of reduced crop production due to weeds are up to $40 billion annually in North America (Pimentel et al., 2005). Rates of adaptive phenotypic change are high in human-disturbed contexts (Hendry et al., 2008), such as invasion, and more common in introduced relative to native species in the same environment (Buswell et al., 2011). Thus, an ecological genomic approach may reveal key insights into a species' adaptive capacity in the face of human-induced selective pressures.

This work aims to contribute to our genomic knowledge of the largest plant family, Asteraceae, and advance the study of invasion and evolution in *Centaurea diffusa*, a highly invasive weed species in North America. What genetic changes have occurred between native and invasive ranges of this species? Can these genetic changes be associated with adaptive trait shifts in the invaded range? To answer these questions, first several tools must be developed. Here we present one such tool, the complete plastid genome for *C. diffusa*, the first plastome from a genera containing approximately 250 species (Susanna and Garcia-Jacas, 2009), and one of only 10 genera in this speciose family with a complete plastome assembly.

Methods

Study Species

Centaurea species (knapweeds, star thistles) comprise the most abundant noxious weed genus in the western US, and are one of only 15 plant genera in the US significantly more likely to contain weedy species than expected by chance (Lejeune and Seastedt, 2001; Kuester et al., 2014). In the century since *Centaurea diffusa* was first reported there, it has formed dense monocultures, reduced forage quality, and altered soil and water resource availability in invaded grasslands (Lejeune and Seastedt, 2001). Recent work (Turner et al., 2014) has demonstrated the rapid evolution of *C.diffusa* in the invaded range under an array of benign and stressful conditions including drought. Invasive individuals grew larger, performed better, or matured later than native in nearly all tested conditions. Additionally, invasive individuals may have been released from a trade-off between growth and drought tolerance apparent in the native range.

Collection and DNA extraction

Seed was collected from an individual in the native range of *C. diffusa* (TR001-1, Turkey, latitude 41.75111, longitude 27.24778). A voucher specimen of this population is located at the UBC Herbarium, accession number V236765. Seed from this collection was grown in a glasshouse at the University of British Columbia during the summer and fall of 2009. Seeds were germinated on filter paper in 1% plant preservative mixture and distilled H_2O at room temperature. After 12 days seedlings were transplanted into 5 cm diameter cones filled with 80% potting mix and 20% silica sand. After two months, individuals were transplanted into 1 l pots containing potting soil to be used in a crossing experiment ("Maternal common garden", Turner et al., 2014). Young leaf tissue was sampled from a single individual (TR001-1L) and stored at -80° C to be used for plastome assembly.

DNA was extracted from frozen tissue using a modified DNeasy column-less protocol. Concentration and quality was verified by Nanodrop, Qbit high-sensitivity assay, and gel electrophoresis.

Plastome assembly and annotation

This whole genome shotgun library was sequenced at Genome Quebec, using one half lane of Illumina HiSeq 2000 paired-end sequencing. Raw data as well as scripts used for cleaning, assembly, and annotation of this plastome are available on Figshare.com (See Supplementary Materials). Raw reads were quality trimmed and screened for sequencing artifacts using Trimmomatic (Bolger et al., 2014). Clean reads were aligned to the *Lactuca sativa* plastome (Timme et al., 2007) using BWA (Li and Durbin, 2009). Pairs in which both reads aligned to the *L. sativa* plastome were extracted from the Sam files with Picard Tools SamToFastq.jar(Picard, 2009). ALLPATHS-LG (Gnerre et al., 2011) was used to merge overlapping pairs and error-correct the data, which was then assembled with Ray (Boisvert et al., 2010). Ray contigs were aligned to the *L. sativa* plastome with BLAST+ (Altschul et al., 1990) and scaffolded based on synteny using OSLay (Richter et al., 2007). Gaps were filled with GapFiller (Boetzer and Pirovano, 2012) resulting in a sequence containing a single N. Visual inspection indicated that the N separated an erroneous tandem duplication, which was corrected by hand with Vim. IR boundaries were confirmed with aTRAM (Allen and Huang, 2014) assemblies of flanking regions. Reads were aligned to the assembly with BWA and sorted with SAMTOOLS (Li et al., 2009). Visual inspection of the alignment revealed a few small indel and substitution errors, which were hand-corrected with Vim. A final alignment and inspection revealed no errors. The plastome was annotated using DOGMA (Wyman et al., 2004) and validated for NCBI GenBank submission using Sequin 13.05. The NCBI GenBank accession number for this plastome is KJ690264.

Results

Tables

Table 1: Ray assembly output numerics

	Contigs	**Contigs**	**Scaffolds**	**Scaffolds**
Size range	≥ 100 nt	≥ 500 nt	≥ 100 nt	≥ 500 nt
Number	21	14	14	7
Total length	128037	126739	129501	128203
Average	6097	9052	9250	18314
N50	23320	23320	27489	27489
Median	1234	2960	1112	11940
Largest	52890	52890	52890	52890

Figures

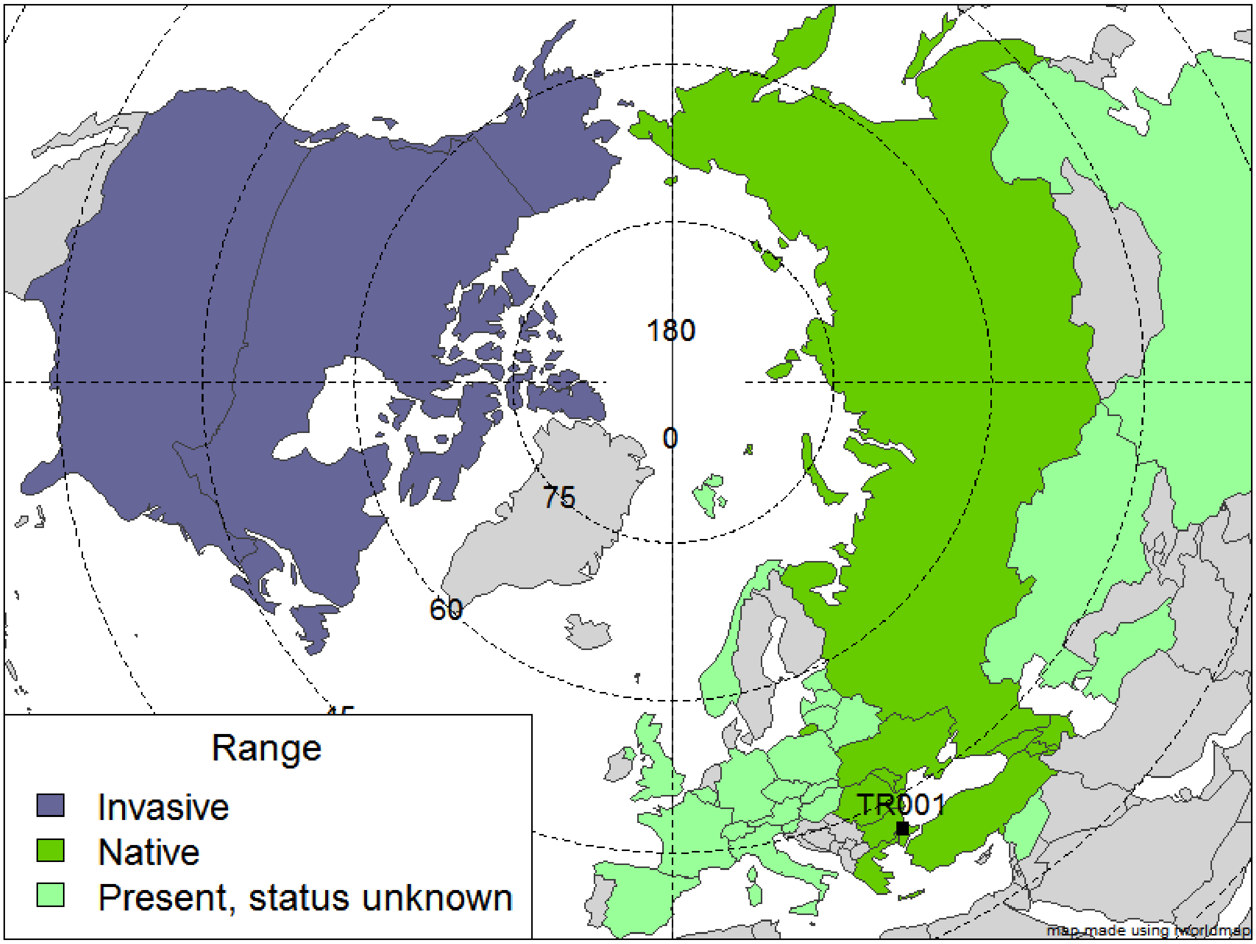

Figure 1: Global range map of *Centaurea diffusa*, by country (modified from Turner et al., 2014). Range status in a particular country is indicated by color. ‘Present, status unknown’ also includes countries where *Centaurea diffusa* is considered naturalized. Degrees of latitude are indicated on dotted lines, and degrees of longitude, solid lines. Seed collection location for the individual used in this assembly is indicated.

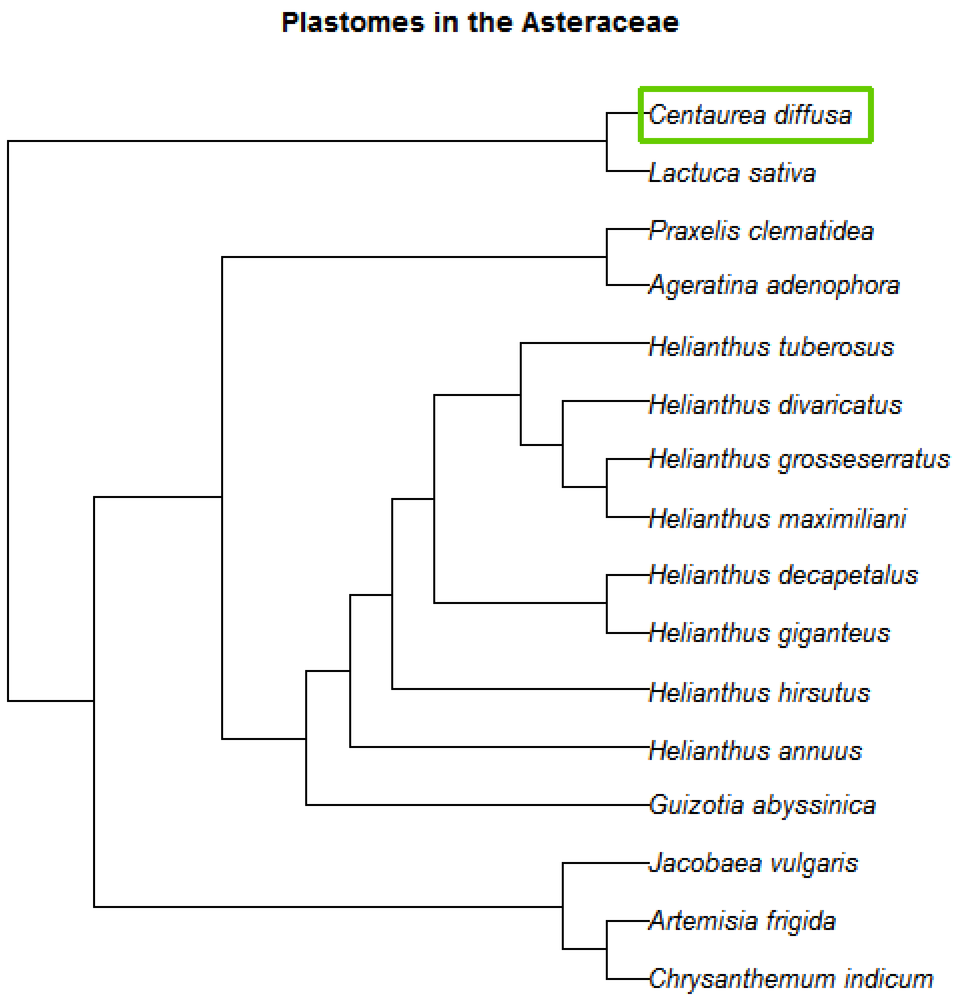

Figure 2: Phylogenetic tree of completed plastomes within the Asteraceae available on NCBI GenBank as of April 17, 2014. Phylogeny based on Smith et al. (2011), using R 3.0.1 (R core team, 2013), taxize (Chamberlain and Szocz, 2013), and Phylomatic (Webb and Donoghue, 2005).

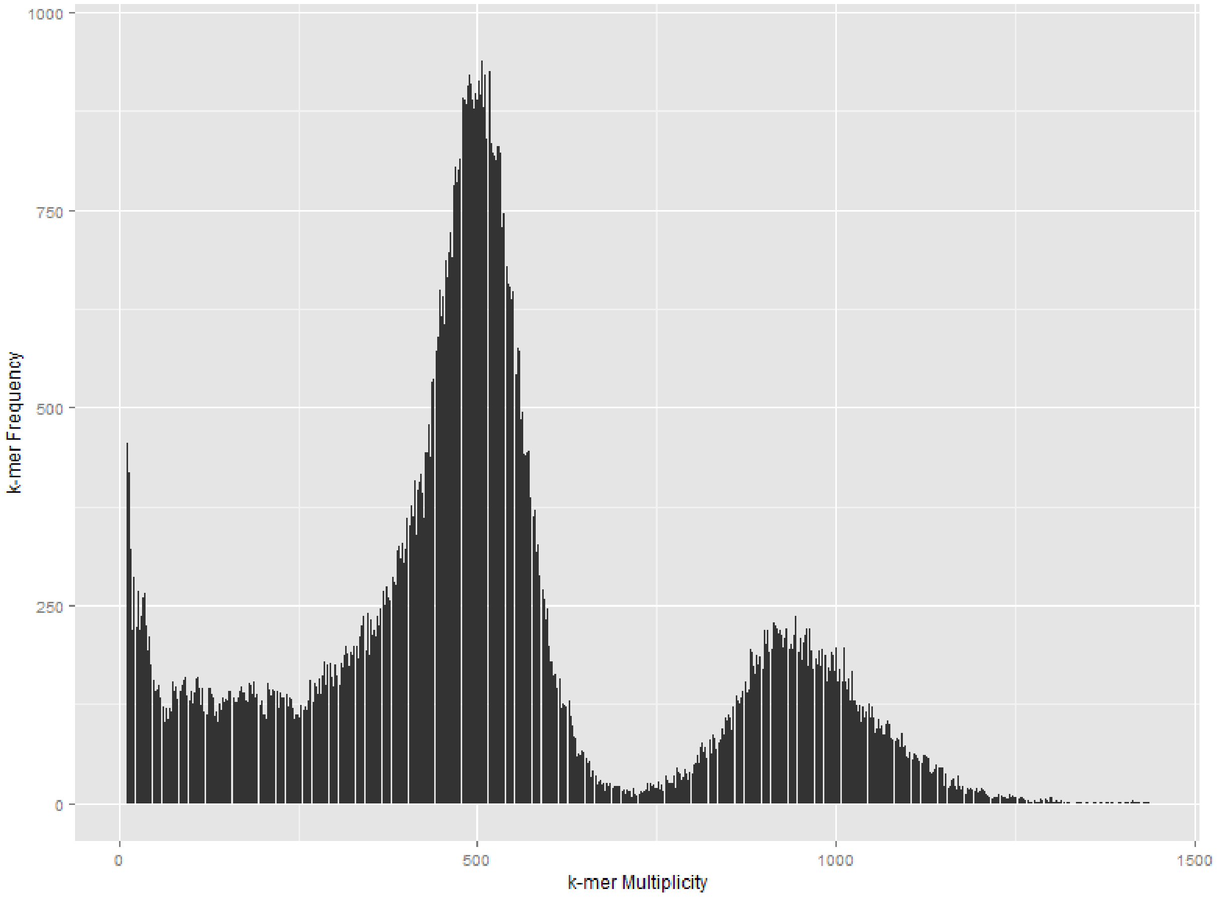

Figure 3: Histogram of 31-mers contained in *Centaurea diffusa* reads which aligned to *Lactuca sativa* plastome and their anchored orphans.

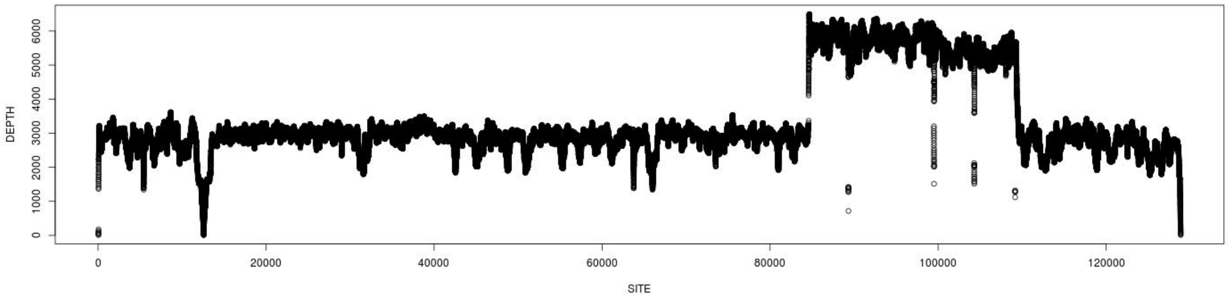

Figure 4: Depth of coverage of raw reads to *Centaurea diffusa* plastome v0.2 after gap filling which suggests: 1) misassembly near supercontig 1 position 12521, and 2) collapse of inverted repeat to a single copy.

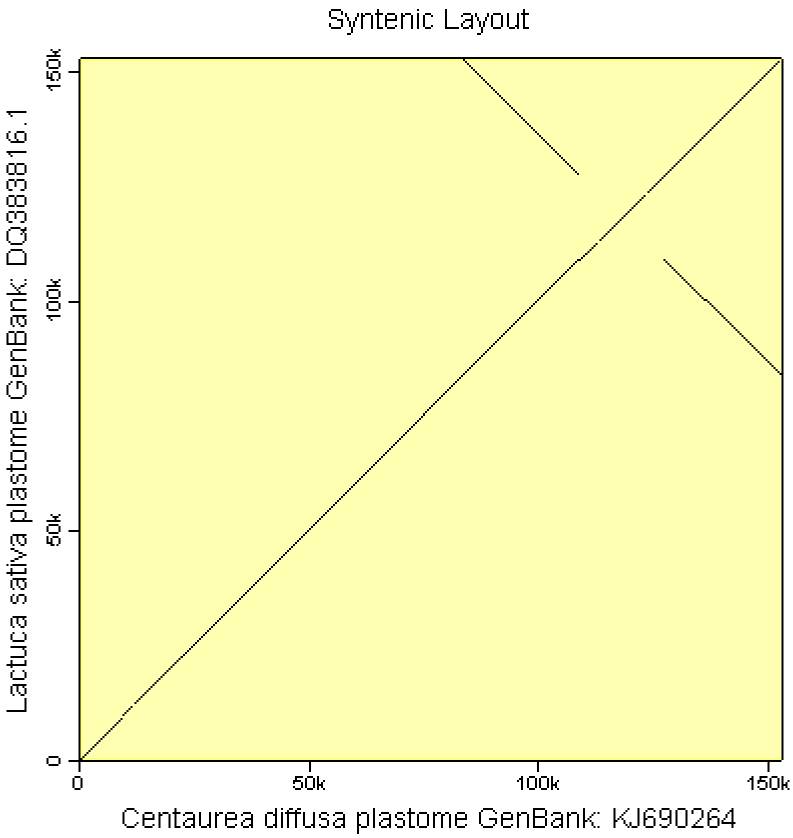

Figure 5: Syntenic alignment between *Lactuca sativa* and *Centaurea diffusa* plastomes using OSLay v1.0 (Richter et al., 2007).

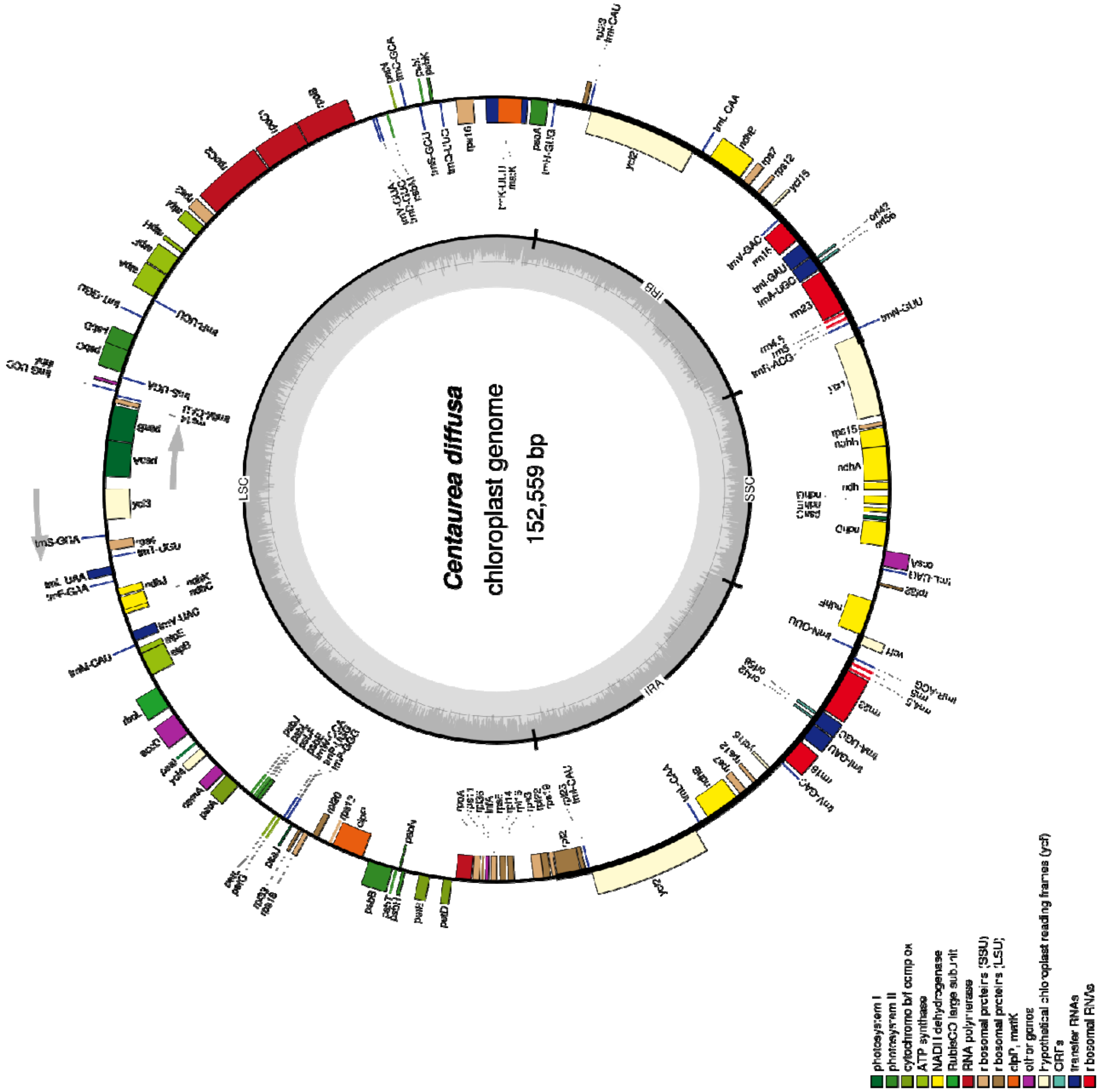

Figure 6: Map of annotated *Centaurea diffusa* plastome, produced using OGDraw (Lohse et al., 2013).

Acknowledgements

We thank A. Guggisberg for seed collection, K. Nurkowski for plant care, D. Huang for assistance with error checking, and R. Timme for help with finalizing the annotation.

References

Allen J, Huang D. 2014. [WWW document and GitHub repository] URL https://github.com/juliema/aTRAM [accessed 25 February 2014]. DOI: 10.5281/zenodo.10431.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. **1990**. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

Boetzer M, Pirovano W. **2012**. Toward almost closed genomes with GapFiller. *Genome Biology* **13**: R56.

Boisvert S, Laviolette F, Corbeil J. **2010**. Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies. *Journal of Computational Biology* **17**: 1519–1533.

Bolger AM, Lohse M, Usadel B. **2014**. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*: btu170.

Buswell JM, Moles AT, Hartley S. **2011**. Is rapid evolution common in introduced plant species? *Journal of Ecology* **99**: 214–224.

Chamberlain S, Szocs E. **2013**. taxize **-** taxonomic search and retrieval in R. *F1000Research* **2**.

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences* **108**: 1513–1518.

Hendry A, Farrugia T, Kinnison M. **2008**. Human influences on rates of phenotypic change in wild animal populations. *Molecular Ecology* **17**: 20–29.

Kuester A, Conner JK, Culley T, Baucom RS. **2014**. How weeds emerge: a taxonomic and trait-based examination using United States data. *New Phytologist* **202**: 1055–1068.

Lejeune KD, Seastedt TR. **2001**. Centaurea species: the forb that won the west. *Conservation Biology* **15**: 1568–1574.

Li H, Durbin R. **2009**. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. **2009**. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lohse M, Drechsel O, Kahlau S, Bock R. **2013**. OrganellarGenomeDRAW--a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research* **41**: W575–W581.

Picard. **2009.** [WWW document] URL http://picard.sourceforge.net [accessed on 29 April 2014].

Pimentel D, Zuniga R, Morrison D. **2005**. Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics* **52**: 273–288.

R Core Team. 2013. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statiscial Computing. [WWW document] URL http://ww.R-project.org/ [accessed on 16 May 2013].

Richter DC, Schuster SC, Huson DH. **2007**. OSLay: optimal syntenic layout of unfinished assemblies. *Bioinformatics* **23**: 1573–1579.

Smith SA, Beaulieu JM, Stamatakis A, Donoghue MJ. **2011**. Understanding angiosperm diversification using small and large phylogenetic trees. *American Journal of Botany* **98**: 404–414.

Susanna A, Garcia-Jacas N. 2009. Chapter 20: Cardueae (Carduoideae). In: Funk VA, Stuessy T, Bayer R, eds. *Systematics, evolution, and biogeography of Compositae*. International Association for Plant Taxonomy Vienna, Austria, 293-313.

Timme RE, Kuehl JV, Boore JL, Jansen RK. **2007**. A comparative analysis of the Lactuca and Helianthus (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. *American Journal of Botany* **94**: 302–312.

Turner KG, Hufbauer RA, Rieseberg LH. **2014**. Rapid evolution of an invasive weed. *New Phytologist* **202**: 309–321.

Webb CO, Donoghue MJ. **2005**. Phylomatic: tree assembly for applied phylogenetics. *Molecular Ecology Notes* **5**: 181–183.

Wyman SK, Jansen RK, Boore JL. **2004**. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**: 3252–3255.

Supplementary materials

Supplementary materials can be found at:

Complete plastid genome assembly of invasive plant, *Centaurea diffusa* **-** supplementary files. Kathryn Turner. fig**share**. Retrieved 22:17, Jun 03, 2014 (GMT) http://dx.doi.org/10.6084/m9.figshare.1044306

Supplementary files contain the following:

All figures.

Dataset S1: Raw reads that align to final assembly can be found in TR001_1L.to.TR001_1L.plastid.0.7.fa.sort.bam and TR001_1L.to.TR001_1L.plastid.0.7.fa.sort.bam.bai.

Scripts S1: All assembly code can be found in Centaurea_diffusa_Assembly_scripts.tar.gz. PLASTOME_MASTER.sh is the master script containing all commands and descriptions of manual steps.

Scripts S2: R code for the production of figures 1, 2, and 3.