**The first fully annotated chloroplast sequence of the large pantropical family Annonaceae.**

Authors: Rita D. Brandão1,\*, Martijn Staats2, Lars W. Chatrou2, Jan W. Maas3, Hans Persoon4, Tatiana Arias5, Roy H.J. Erkens1,6

1. Maastricht Science Programme, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands
2. Biosystematics group, Wageningen University, Wageningen, The Netherlands
3. Ecology and Biodiversity Group, Department of Biology, Utrecht University, Utrecht, The Netherlands
4. Botanical Garden Utrecht, Utrecht University, Utrecht, The Netherlands
5. School of Biological Sciences, The University of Hong Kong, China
6. Naturalis Biodiversity Center, Leiden, The Netherlands

**Abstract**

The pantropical flowering plant family Annonaceae is one of the largest families within the Magnoliidae with c. 2,400 species. The study of its evolution allows to gain insight into the evolution of basal angiosperms, the evolution of rain forests and biodiversification. Phylogenetic analysis have greatly contributed to these studies. However, these are limited to a few DNA markers. Here we present the first chloroplast genome (plastome) sequence from an Annonaceae species: *Uvaria afzelii*.

Illumina paired-end reads were mapped to a closely-related species using iterative alignment steps. A very low percentage of reads (0.5%) was from the chloroplast, and 99,5% from the nuclear and mitochondrial genome. The 167,500 bp plastome of *Uvaria afzelii* contains 139 genes and an inversion within the LSC (57,722 - 66,951 bp) compared to the chloroplasts of *Magnolia kwansiensis* and *Liriodendron tulipifera.*

The sequence of this chloroplast will now aid the analysis of additional Annonaceae plastomes, reveal useful for improved phylogeographic analyses, species-level DNA barcoding and general understanding of magnoliid evolution as well as rain forest evolution.

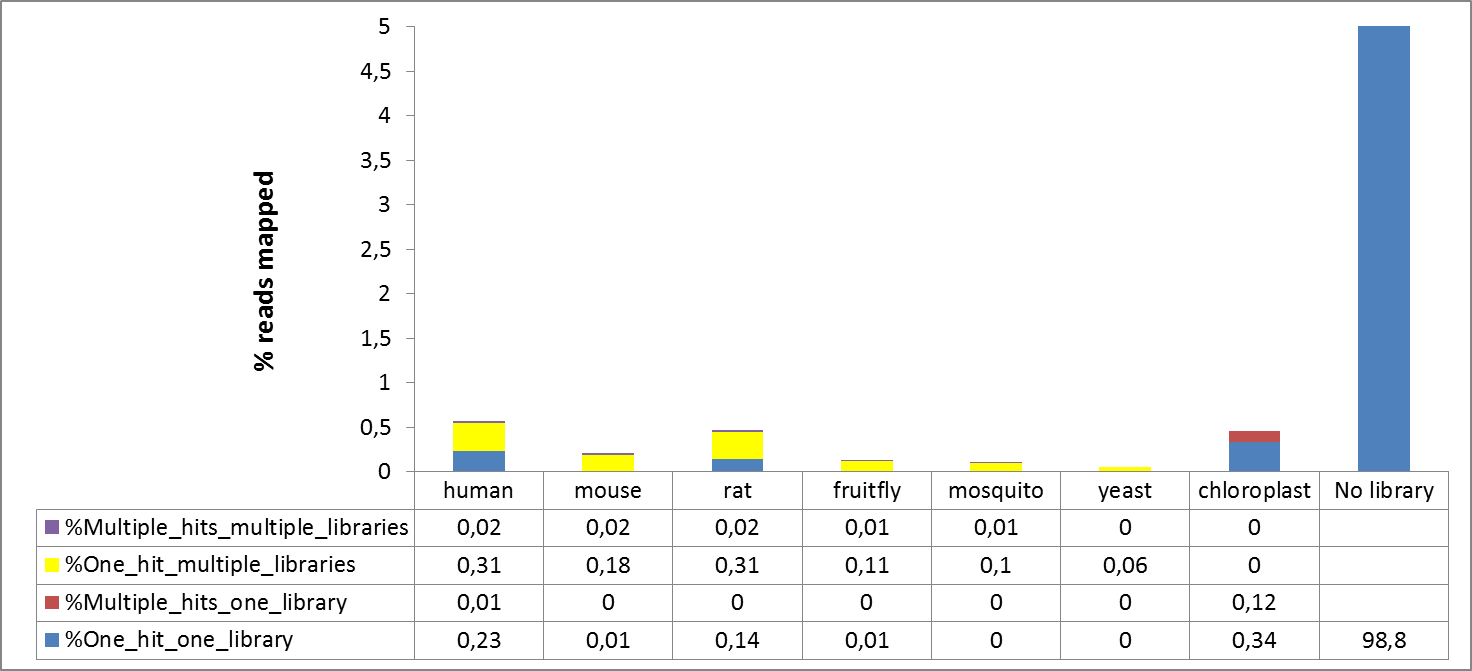
**Introduction**

Understanding the origin of flowering-plant biodiversity is one of the major biological research themes of the moment. A lot of effort has been devoted to studying the monocot and eudicot lineages but less to the other large clade of flowering plants, Magnoliidae, although it contains c. 4% of all plant species. Within Magnoliidae, the pantropically distributed Annonaceae is an extremely genus-rich family, with 108 genera and it currently has a high number of accepted species: *c*. 2,400[1](#_ENREF_1). Understanding the evolution and diversification of Annonaceae will therefore greatly aid in understanding the evolution of basal angiosperms. Additionally, species richness and abundance of individuals within the tropical rainforests indicate Annonaceae may be a key group to understand the evolution of tropical diversity[2-4](#_ENREF_2).

Molecular phylogenetic studies in Annonaceae, as in many other clades, have mostly been conducted using a few chloroplast markers whether it be at the family, subfamily, tribal[5-7](#_ENREF_5), or generic levels[1](#_ENREF_1),[8](#_ENREF_8),[9](#_ENREF_9), with few exceptions[10](#_ENREF_10),[11](#_ENREF_11). The suite of plastid markers available for Annonaceae is limited, mainly due to the fact that previous knowledge of the chloroplast DNA sequence is needed in order to identify the most informative ones at different taxonomic levels. Nowadays, the advent of next generation sequencing (NGS) techniques is revolutionizing the field of phylogenetics. NGS allows us to generate massive amounts of DNA sequencing data in order to revisit challenging evolutionary questions that could not be answered in a time efficient manner using Sanger sequencing (Steel et al., 2012; Straub et al., 2012). So, it is possible to sequence entire plastomes of these plants, including from herbarium material using methods such as published by Staats et al[12](#_ENREF_12). Whole chloroplast assembly using shallow genomic DNA sequencing has become a popular method among botanists for reconstructing plant phylogenies (Steel et al., 2012). NGS sequencing is cheaper and faster than traditional PCR-based methods to obtain large amounts of data. These large amounts of data can also be used to identify more relevant regions for resolving phylogenies of difficult clades[13-16](#_ENREF_13). Currently, in the genome repository GenBank[17](#_ENREF_17) there are almost 650 plant chloroplast genome entries, of which around 500 belong to flowering plants[18](#_ENREF_18). Despite the increasing number of chloroplast genomes every year, until this study, no chloroplast genome was available for Annonaceae. Here, we describe the first plastome of a species of Annonaceae: *Uvaria afzelii* G.Elliott (tribe Uvarieae, subfamily Annonoideae[1](#_ENREF_1)). Small trees or spreading shrubs of *Uvaria afzelii* may grow up to 5 metres tall. This species is distributed from Guinea to southern Nigeria; the fruit is edible and leaves, bark and roots are used for their medicinal properties[19](#_ENREF_19). The chloroplast genome of *Uvaria afzelii* will aid to the reconstruction chloroplast genomes from other Annonaceae species allowing identification of new markers that may help resolving difficult clades, or DNA targeting enrichment for sequencing approaches. Resources like this will contribute to generate more molecular data and facilitate phylogenetic reconstructions in the family and hopefully enabling deeper understanding of the Annonaceae family, the magnoliid clade and, ultimately, of the tropical rain forest evolution.

**Results and discussion**

In this study we report the first completely annotated chloroplast genome of an Annonaceae species (*Uvaria afzelii)*. Results of the quality control freeware Fastq Screen[20](#_ENREF_20) on the raw sequence data show that only a small percentage of all reads belong to the chloroplast sequence (figure 1). Less than 0.5% of the total number of reads was aligned to the unpublished draft plastome of *Miliusa cuneata* (Arias et al., unpublished data), which was used as a reference. The majority of the reads did not correspond to human, mouse or insect DNA, therefore are inferred to be nuclear and/or mitochondrial plant DNA. After assembly of the reads by mapping to the reference chloroplast followed by several iteration mapping steps, the percentage of used reads was 0.7%. Few gaps with coverage below 2X were identified and resolved using *in silico* primer walking and PCR followed by automated Sanger sequencing method.



//

100

80

60

50

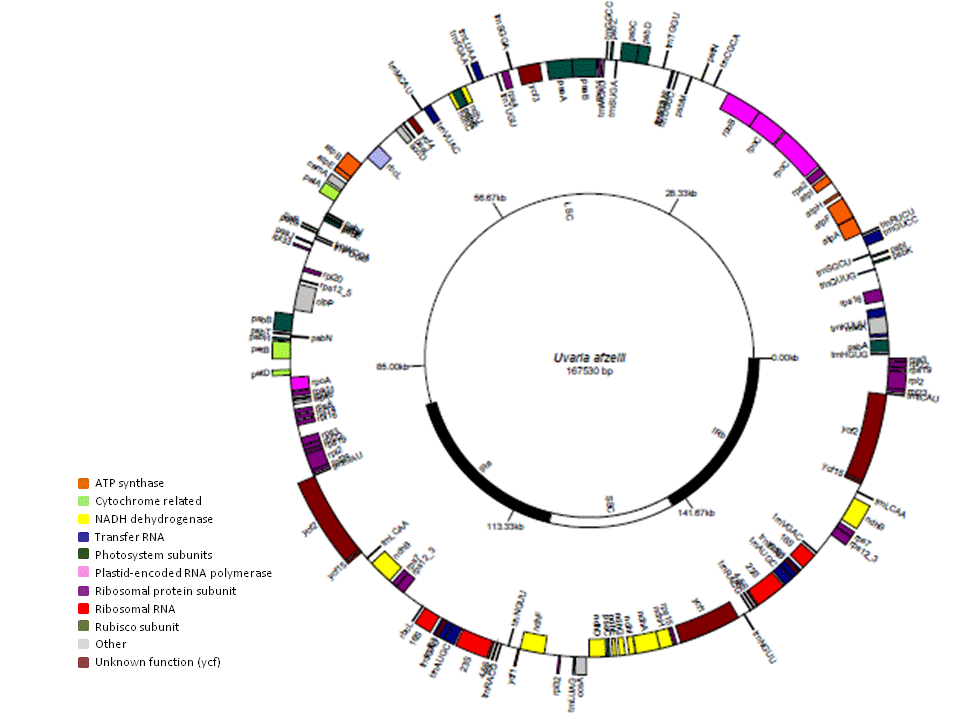
40

30

**Figure 1. Percentage of mapped reads of *Uvaria afzelii* to different genomes.** Genomes used to map the raw reads were the following: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Anopheles gambiae*, and *Saccharomyces cerevisiae*. The chloroplast genome used was from *Miliusa cuneata*. No library refers to all reads that did not map to any of the libraries used.

The complete chloroplast genome of *Uvaria afzelii* (figure 2) is around 167,530 bp. It is noteworthy that small indels, as well as SNPs, were observed in heteroplasmy, so the actual genome size varies slightly. The inverted repeats (IRs) have 28,002 and 28.095 bp respectively and are separated by the small single-copy region (20,010 bp) and the long single-copy region (91,296 bp). We found 139 genes in the chloroplast DNA of *Uvaria afzelii*, 21 of which are completely duplicated and 2 partially duplicated within the IRs. The most noteworthy features of its chloroplast genome is an inversion within the LSC (57,722 - 66,951 bp) compared to the chloroplasts of *Magnolia kwansiensis*[21](#_ENREF_21)and *Liriodendron tulipifera*[22](#_ENREF_22). This region includes the following genes: atpE, atpB, rbcL, accD, psaI and ycf4.

CpDNA heteroplasmy, the condition in which there is more than one organellar haplotype, has been described in several angiosperms genera, such as: *Medicago*[23](#_ENREF_23) (Fabaceae), *Coreopsis*[24](#_ENREF_24)(Asteraceae), *Turnera*[25](#_ENREF_25) (Passifloraceae). This event poses challenges for phylogenetic analyses as bioinformatic tools are not ready to deal with this event. The use of NGS to sequence (parts of) genomes at high coverage allows to easily detect these. So, polymorphism-aware phylogenetic tools should be developed to account for heteroplasmy.. Such tools have already been developed to analyze mutation rates and fixation biases for great apes[26](#_ENREF_26),[27](#_ENREF_27).



Uvaria afzelii 167,530 bp

**Figure 2. Plastome map of *Uvaria afzelii*.** Genes in the outer side of the circle have sense orientation genes, whereas the inner side shows anti-sense oriented genes. Genes have been coloured according to their function.

It is important to realise that despite the chloroplast circular structure representation and the fact that Genbank accessions mention that chloroplast genome is circular, the chloroplast DNA mostly has a complex and branched form, and the circular form is only present in small amounts[28](#_ENREF_28). Probably due to this particularity of the plastome molecule, PCRs over the IRb/LSC boundary, where the chloroplast should circularize, were not successful.

Chloroplast DNA is the most gene-rich molecule of the three genomes of a plant cell, with more than 100 genes, spanning over 100-200 kb, that code for the entire machinery necessary for the photosynthetic process. There is a relatively high degree of conservation in size, structure, gene content, and linear order of the chloroplast genes in land plants[29](#_ENREF_29). In chloroplasts, DNA replication is cell-cycle independent and the number of chloroplast copies present per cell varies. In rapidly dividing leaf tissue there are around hundred copies, but later in leaf development and plant growth, the number of copies decreases to about 20-30[30](#_ENREF_30). In addition, each chloroplast may contain up to 300 copies of the genome[31](#_ENREF_31). Despite the abundance of the chloroplast genome, genome skimming (shallow sequencing) experiments show that only a small percentage of the reads actually map to the chloroplast. So, targeting enrichment sequencing is a very promising technique to sequence the plastome from many different species. However, unlike the standard NGS sequencing protocols, this technique requires some *a priori* knowledge of the genome sequence, such as the genome of closely related species. The sequencing of Annonaceae chloroplast genomes, such as the one presented here will now allow for the construction of (home-made) plastome-target probes for the sequencing of many chloroplast genomes. Parks et al. showed that nearly-complete plastomes of 37 Pinus species allowed for an increased phylogenetic resolution and support values at lower taxonomic levels [32](#_ENREF_32). Similar results were obtained by other using plastome sequences alone [33](#_ENREF_33),[34](#_ENREF_34) or in combination with nuclear ribosomal sequences [35](#_ENREF_35). Therefore, complete Annonaceae plastome analysis might also reveal useful for improved phylogeographic analyses, species-level DNA barcoding and general understanding of magnoliid evolution as well as of rain forest evolution.

**Materials and Methods**

Sample preparation

Fresh leaf material from *Uvaria afzelii* was collected from a green house grown tree at the Botanical garden of Utrecht University (Utrecht, The Netherlands). The chloroplast isolation kit (Sigma, Saint-Louis, USA) was used according to the manufacturer’s instructions. DNA extraction from the chloroplasts was performed with the GenElute plant genomic DNA miniprep kit (Sigma). The obtained DNA was sent to Macrogen (Korea) for paired-end library construction and sequencing on a HiSeq 2000 (Illumina) instrument, following the company’s protocol.

Sequencing analysis

Quality control was performed with FastQC and FastQ Screen (both available at http://www.bioinformatics.babraham.ac.uk/projects/download.html). Geneious version 6.06[36](#_ENREF_36) ([http://www.geneious.com](http://www.geneious.com/)) was used to filter poor quality bases and/or reads and iterative mapping steps. In the first mapping step, reads were mapped to the draft sequence of a chloroplast of *Miliusa cuneata* Craib, an Asian Annonaceae species (kindly provided by Arias et al., unpublished data), to assemble a draft of the *U. afzelii* plastome. In the following 20 steps, consecutive rounds of mapping to the obtained sequences were performed. Remaining gaps were resolved in two ways: a) *in silico* primer walking using the *de novo* assembler PRICE[37](#_ENREF_37) on previously filtered data from the software TrimGalore! (available at http://www.bioinformatics.babraham.ac.uk/projects/download.html); b) PCRs using primers flanking the gaps followed by automated Sanger sequencing.

Gene annotation and chloroplast circular map design

Gene annotation was performed in Geneious software upon alignment of the complete chloroplast sequences of *Liriodendron tulipifera*[22](#_ENREF_22) (genbank accession number: NC\_008326.1) and *Magnolia kwangsiensis*[21](#_ENREF_21) (genbank accession number: NC\_015892.1) using the MAFF package, and in DOGMA[38](#_ENREF_38). The exact boundaries between the inverted repeats and the single copies were determined manually. GenomeVx[39](#_ENREF_39) was used to draw the circular map of the chloroplast genome.

**References**

1 Chatrou, L. W. et al. A new subfamilial and tribal classification of the pantropical flowering plant family Annonaceae informed by molecular phylogenetics. *Botanical Journal of the Linnean Society* **169**, 5-40 (2012).

2 van Gemerden, B. S., Olff, H., Parren, M. P. E. & Bongers, F. The pristine rain forest? Remnants of historical human impacts on current tree species composition and diversity. *Journal of Biogeography* **30**, 1381-1390 (2003).

3 Gentry, A. H. Tree species richness of upper Amazonian forests. *Proceedings of the National Academy of Sciences* **85**, 156-159 (1988).

4 Valencia, R., Balslev, H. & Paz Y Miño C, G. High tree alpha-diversity in Amazonian Ecuador. *Biodivers Conserv* **3**, 21-28 (1994).

5 Pirie, M. D., Chatrou, L. W., Mols, J. B., Erkens, R. H. J. & Oosterhof, J. ‘Andean-centred’ genera in the short-branch clade of Annonaceae: testing biogeographical hypotheses using phylogeny reconstruction and molecular dating. *Journal of Biogeography* **33**, 31-46 (2006).

6 Erkens, R. H. J., Maas, J. W. & Couvreur, T. L. P. From Africa via Europe to South America: migrational route of a species-rich genus of Neotropical lowland rain forest trees (Guatteria, Annonaceae). *Journal of Biogeography* **36**, 2338-2352 (2009).

7 Chaowasku, T. et al. A plastid DNA phylogeny of tribe Miliuseae: Insights into relationships and character evolution in one of the most recalcitrant major clades of Annonaceae. *American Journal of Botany* **101**, 691-709 (2014).

8 Su, Y. & Saunders, R. Evolutionary divergence times in the Annonaceae: evidence of a late Miocene origin of Pseuduvaria in Sundaland with subsequent diversification in New Guinea. *BMC Evolutionary Biology* **9**, 153 (2009).

9 Erkens, R. H. J. et al. A decade of uncertainty: Resolving the phylogenetic position of Diclinanona (Annonaceae), including taxonomic notes and a key to the species. *Taxon* **63**, 1244–1252 (2014).

10 Chatrou, L. W. et al. Flanking regions of monomorphic microsatellite loci provide a new source of data for plant species-level phylogenetics. *Molecular Phylogenetics and Evolution* **53**, 726-733 (2009).

11 Thongpairoj, U.-S. *Taxonomy and molecular phylogeny of Artabotrys R.Brown and palynology of the tribe Unoneae (Annonaceae)* PhD thesis, Chiang Mai University, Thailand, (2008).

12 Staats, M. et al. Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS ONE* **8**, e69189 (2013).

13 Bortiri, E., Coleman-Derr, D., Lazo, G. R., Anderson, O. D. & Gu, Y. Q. The complete chloroplast genome sequence of Brachypodium distachyon: sequence comparison and phylogenetic analysis of eight grass plastomes. *BMC Research Notes* **1**, 61-61 (2008).

14 Huang, H., Shi, C., Liu, Y., Mao, S.-Y. & Gao, L.-Z. Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evolutionary Biology* **14**, 151 (2014).

15 Matsuoka, Y., Yamazaki, Y., Ogihara, Y. & Tsunewaki, K. Whole Chloroplast Genome Comparison of Rice, Maize, and Wheat: Implications for Chloroplast Gene Diversification and Phylogeny of Cereals. *Molecular Biology and Evolution* **19**, 2084-2091 (2002).

16 Shaw, J., Lickey, E. B., Schilling, E. E. & Small, R. L. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany* **94**, 275-288 (2007).

17 Benson, D. A. et al. GenBank. *Nucleic Acids Research* **41**, D36-D42 (2013).

18 Information, N. C. f. B. *GenBank*, <http://www.ncbi.nlm.nih.gov/genbank/> (

19 Burkill, H. M. *The useful plants of West Tropical Africa*. Vol. 1 (Royal Botanic Gardens, 1985).

20 *FastQ Screen*, <http://www.bioinformatics.babraham.ac.uk/projects/fastq\_screen/> (2014).

21 Kuang, D.-Y. et al. Complete chloroplast genome sequence of Magnolia kwangsiensis (Magnoliaceae): implication for DNA barcoding and population genetics. *Genome* **54**, 663-673 (2011).

22 Cai, Z. et al. Complete plastid genome sequences of Drimys, Liriodendron, and Piper: implications for the phylogenetic relationships of magnoliids. *BMC Evolutionary Biology* **6**, 77 (2006).

23 Johnson, L. B. & Palmer, J. D. Heteroplasmy of chloroplast DNA in Medicago. *Plant Molecular Biology* **12**, 3-11.

24 Mason, R. J., Holsinger, K. E. & Jansen, R. K. Biparental Inheritance of the Chloroplast Genome in Coreopsis (Asteraceae). *Journal of Heredity* **85**, 171-173 (1994).

25 Shore, J. S., McQueen, K. L. & Little, S. H. Inheritance of Plastid DNA in the Turnera ulmifolia Complex (Turneraceae). *American Journal of Botany* **81**, 1636-1639 (1994).

26 De Maio, N., Schlötterer, C. & Kosiol, C. Linking Great Apes Genome Evolution across Time Scales Using Polymorphism-Aware Phylogenetic Models. *Molecular Biology and Evolution* **30**, 2249-2262 (2013).

27 De Maio, N., Schrempf, D. & Kosiol, C. PoMo: An Allele Frequency-Based Approach for Species Tree Estimation. *Systematic Biology* **64**, 1018-1031 (2015).

28 Bendich, A. J. Circular Chloroplast Chromosomes: The Grand Illusion. *The Plant Cell Online* **16**, 1661-1666 (2004).

29 Downie, S. & Palmer, J. in *Molecular Systematics of Plants* (eds PamelaS Soltis, DouglasE Soltis, & JeffJ Doyle) Ch. 2, 14-35 (Springer US, 1992).

30 Possingham, J. V. Plastid replication and development in the life cycle of higher plants. *Annual Review of Plant Physiology* **31**, 113-129 (1980).

31 Krupinska, K., Melonek, J. & Krause, K. New insights into plastid nucleoid structure and functionality. *Planta* **237**, 653-664 (2013).

32 Parks, M., Cronn, R. & Liston, A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology* **7**, 84 (2009).

33 Whittall, J. B. et al. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology* **19**, 100-114 (2010).

34 Yang, J.-B., Tang, M., Li, H.-T., Zhang, Z.-R. & Li, D.-Z. Complete chloroplast genome of the genus Cymbidium: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evolutionary Biology* **13**, 84 (2013).

35 Kane, N. et al. Ultra-barcoding in cacao (Theobroma spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany* **99**, 320-329 (2012).

36 Kearse, M. et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649 (2012).

37 Ruby, J. G., Bellare, P. & DeRisi, J. L. PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data. *G3: Genes|Genomes|Genetics* **3**, 865-880 (2013).

38 Wyman, S. K., Jansen, R. K. & Boore, J. L. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**, 3252-3255 (2004).

39 Conant, G. C. & Wolfe, K. H. GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics* **24**, 861-862 (2008).