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

**Running title: The first Annonaceae chloroplast genome**

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, 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

\*Corresponding author:

rita.brandao@maastrichtuniversity.nl; PO Box 616, 6200 MD Maastricht; tel. +31 043 3881350

**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 so far limited to few DNA markers. Here we present the first fully-annotated 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 167,530 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 information for improved phylogeographic analyses, species-level DNA barcoding and general understanding of magnoliid evolution as well as rain forest evolution.

Keywords: Annonaceae, next generation sequencing, plastome, complete chloroplast sequence, complete genome

**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 ([Chatrou et al. 2012](#_ENREF_9)). 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 ([Gentry 1988](#_ENREF_15); [Valencia et al. 1994](#_ENREF_37); [van Gemerden et al. 2003](#_ENREF_38)).

Molecular phylogenetic studies in Annonaceae, as in many other clades, have mostly been conducted using few chloroplast markers whether it be at the family, subfamily, tribal ([Pirie et al. 2006](#_ENREF_25); [Erkens et al. 2009](#_ENREF_14); [Chaowasku et al. 2014](#_ENREF_7)), or generic levels ([Su and Saunders 2009](#_ENREF_34); [Chatrou et al. 2012](#_ENREF_9); [Erkens et al. 2014](#_ENREF_12)), with few exceptions ([Thongpairoj 2008](#_ENREF_36); [Chatrou et al. 2009](#_ENREF_8)). 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 ([Steele et al. 2012](#_ENREF_32); [Straub et al. 2012](#_ENREF_33)). So, it is possible to sequence entire plastomes of these plants, including from herbarium material using methods such as published by Staats et al.([Staats et al. 2013](#_ENREF_31)) Whole chloroplast assembly using shallow genomic DNA sequencing has become a popular method among botanists for reconstructing plant phylogenies ([Steele et al. 2012](#_ENREF_32)). 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 ([Matsuoka et al. 2002](#_ENREF_23); [Shaw et al. 2007](#_ENREF_29); [Bortiri et al. 2008](#_ENREF_4); [Huang et al. 2014](#_ENREF_16)). Currently, in the RefSeq Database ([Pruitt et al. 2002 [updated 2012]](#_ENREF_27)) there are almost 900 plant chloroplast genome entries, of which 685 belong to flowering plants (Genbank, accessed May 2016). Basal angiosperms (ANITA clade, Magnoliidae clade, Choranthaceae and Ceratophyllaceae) are only represented by 29 plastomes, which represents 4% of all angiosperms' plastomes. 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 ([Chatrou et al. 2012](#_ENREF_9))). 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 ([Burkill 1985](#_ENREF_5)). The chloroplast genome of *Uvaria afzelii* will aid to the reconstruction of chloroplast genomes from other Annonaceae species since it can be used as a closely related reference genome for assembly. Furthermore, it allows identification of new markers that may help resolving difficult clades, or DNA targeting enrichment for sequencing approaches.

**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 ([Kearse et al. 2012](#_ENREF_19)) ([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 TI ([Ruby et al. 2013](#_ENREF_28)) 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* ([Cai et al. 2006](#_ENREF_6)) (genbank accession number: NC\_008326.1) and *Magnolia kwangsiensis* ([Kuang et al. 2011](#_ENREF_21)) (genbank accession number: NC\_015892.1) using the MAFF package, and in DOGMA ([Wyman et al. 2004](#_ENREF_40)). The exact boundaries between the inverted repeats and the single copies were determined manually. GenomeVx ([Conant and Wolfe 2008](#_ENREF_10)) was used to draw the circular map of the chloroplast genome. The complete chloroplast genome with annotations was submitted to Genbank (accession number XXXXXXX).

**Results and discussion**

Results of the quality control freeware Fastq Screen([2014](#_ENREF_2)) on the raw sequence data show that only a small percentage of all 20,636,570 reads belong to the chloroplast sequence (figure 1). Less than 0.6% of the total number of reads was aligned to an 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 mitochondrial plant DNA. After assembly of the reads by mapping to the reference chloroplast, several iteration steps were performed, increasing the percentage of used reads to 0.7%. Few gaps with coverage below 2X were identified and resolved using both *in silico* primer walking and PCR followed by automated Sanger sequencing method.

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* ([Kuang et al. 2011](#_ENREF_21))and *Liriodendron tulipifera* ([Cai et al. 2006](#_ENREF_6)). 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* ([Johnson and Palmer](#_ENREF_17)) (Fabaceae), *Coreopsis* ([Mason et al. 1994](#_ENREF_22))(Asteraceae), *Turnera* ([Shore et al. 1994](#_ENREF_30)) (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 easily detection of these regions. So, polymorphism-aware phylogenetic tools must be implemented to account for heteroplasmy.

It is important to realise that despite the chloroplast circular structure representation and the fact that Genbank accessions mention that the chloroplast genome is circular, the chloroplast DNA mostly has a complex and branched form, and the circular form is only present in small amounts ([Bendich 2004](#_ENREF_3)). 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 a very gene-rich molecule, with more than 100 genes, spanning over 100-200 kb comparing plant mitochondrial genome (around 60 genes in a variable size with a minimum of 200kb). Chloroplast genes 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 ([Downie and Palmer 1992](#_ENREF_11)). 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 ([Possingham 1980](#_ENREF_26)). In addition, each chloroplast may contain up to 300 copies of the genome ([Krupinska et al. 2013](#_ENREF_20)). 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 (homemade) 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 ([Parks et al. 2009](#_ENREF_24)). Similar results were obtained by other using plastome sequences alone ([Whittall et al. 2010](#_ENREF_39); [Yang et al. 2013](#_ENREF_41)) or in combination with nuclear ribosomal sequences ([Kane et al. 2012](#_ENREF_18)). Therefore, complete Annonaceae plastome analysis might also reveal useful for improved phylogeographic analyses to elucidate relationships within, for instance, recently radiated genera ([Erkens et al. 2007](#_ENREF_13); [Tang et al. 2015](#_ENREF_35)) or ill-resolved branching events in the Annonaceae family phylogeny ([Erkens et al. 2009](#_ENREF_14)). Lastly, species-level DNA barcoding and general understanding of magnoliid evolution as well as investigations of rain forest evolution greatly benefit from this data.

**References**

Genbank. Available at http://www.ncbi.nlm.nih.gov/nuccore. Accessed 16 May 2016.

(2014). FastQ Screen. Available at http://www.bioinformatics.babraham.ac.uk/projects/fastq\_screen/.

Bendich, A. J. (2004) Circular Chloroplast Chromosomes: The Grand Illusion. The Plant Cell Online 16: 1661-1666. doi: 10.1105/tpc.160771

Bortiri, E., D. Coleman-Derr, et al. (2008) The complete chloroplast genome sequence of Brachypodium distachyon: sequence comparison and phylogenetic analysis of eight grass plastomes. BMC Research Notes 1: 61-61. doi: 10.1186/1756-0500-1-61

Burkill, H. M. (1985). The useful plants of West Tropical Africa. Kew, Richmond, United Kingdom, Royal Botanic Gardens.

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

Chaowasku, T., D. C. Thomas, et al. (2014) 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. doi: 10.3732/ajb.1300403

Chatrou, L. W., M. P. Escribano, et al. (2009) Flanking regions of monomorphic microsatellite loci provide a new source of data for plant species-level phylogenetics. Molecular Phylogenetics and Evolution 53: 726-733. doi: http://dx.doi.org/10.1016/j.ympev.2009.07.024

Chatrou, L. W., M. D. Pirie, et al. (2012) 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. doi: 10.1111/j.1095-8339.2012.01235.x

Conant, G. C. and K. H. Wolfe (2008) GenomeVx: simple web-based creation of editable circular chromosome maps. Bioinformatics 24: 861-862. doi: 10.1093/bioinformatics/btm598

Downie, S. and J. Palmer (1992). Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. Molecular Systematics of Plants. P. Soltis, D. Soltis and J. Doyle, Springer US**:** 14-35.

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

Erkens, R. H. J., L. W. Chatrou, et al. (2007) A rapid diversification of rainforest trees (*Guatteria*; Annonaceae) following dispersal from Central into South America. Molecular Phylogenetics and Evolution 44: 399-411. doi: http://dx.doi.org/10.1016/j.ympev.2007.02.017

Erkens, R. H. J., J. W. Maas, et al. (2009) 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. doi: 10.1111/j.1365-2699.2009.02162.x

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

Huang, H., C. Shi, et al. (2014) Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. BMC Evolutionary Biology 14: 151. doi:

Johnson, L. B. and J. D. Palmer Heteroplasmy of chloroplast DNA in *Medicago*. Plant Molecular Biology 12: 3-11. doi: 10.1007/bf00017442

Kane, N., S. Sveinsson, et al. (2012) Ultra-barcoding in cacao (Theobroma spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. American Journal of Botany 99: 320-329. doi: 10.3732/ajb.1100570

Kearse, M., R. Moir, et al. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647-1649. doi: 10.1093/bioinformatics/bts199

Krupinska, K., J. Melonek, et al. (2013) New insights into plastid nucleoid structure and functionality. Planta 237: 653-664. doi: 10.1007/s00425-012-1817-5

Kuang, D.-Y., H. Wu, et al. (2011) Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. Genome 54: 663-673. doi: 10.1139/g11-026

Mason, R. J., K. E. Holsinger, et al. (1994) Biparental inheritance of the chloroplast genome in *Coreopsis* (Asteraceae). Journal of Heredity 85: 171-173. doi:

Matsuoka, Y., Y. Yamazaki, et al. (2002) 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. doi:

Parks, M., R. Cronn, et al. (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology 7: 1-17. doi: 10.1186/1741-7007-7-84

Pirie, M. D., L. W. Chatrou, et al. (2006) ‘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. doi: 10.1111/j.1365-2699.2005.01388.x

Possingham, J. V. (1980) Plastid replication and development in the life cycle of higher plants. Annual Review of Plant Physiology 31: 113-129. doi: doi:10.1146/annurev.pp.31.060180.000553

Pruitt, K., G. Brown, et al. (2002 [updated 2012]). The Reference Sequence (RefSeq) Database.In The NCBI Handbook [Internet], National Center for Biotechnology Information (US), Bethesda (MD), "18" 18. Available at http://www.ncbi.nlm.nih.gov/books/NBK21091/

Ruby, J. G., P. Bellare, et al. (2013) PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data. G3: Genes|Genomes|Genetics 3: 865-880. doi: 10.1534/g3.113.005967

Shaw, J., E. B. Lickey, et al. (2007) 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. doi: 10.3732/ajb.94.3.275

Shore, J. S., K. L. McQueen, et al. (1994) Inheritance of plastid DNA in the *Turnera ulmifolia* complex (Turneraceae). American Journal of Botany 81: 1636-1639. doi:

Staats, M., R. H. J. Erkens, et al. (2013) Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. PLoS ONE 8: e69189. doi: 10.1371/journal.pone.0069189

Steele, P. R., K. L. Hertweck, et al. (2012) Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. American Journal of Botany 99: 330-348. doi: 10.3732/ajb.1100491

Straub, S. C. K., M. Parks, et al. (2012) Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. American Journal of Botany 99: 349-364. doi: 10.3732/ajb.1100335

Su, Y. and R. Saunders (2009) 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. doi:

Tang, C. C., D. C. Thomas, et al. (2015) Molecular phylogenetics of the species-rich angiosperm genus *Goniothalamus* (Annonaceae) inferred from nine chloroplast DNA regions: Synapomorphies and putative correlated evolutionary changes in fruit and seed morphology. Molecular Phylogenetics and Evolution 92: 124-139. doi: http://dx.doi.org/10.1016/j.ympev.2015.06.016

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

Valencia, R., H. Balslev, et al. (1994) High tree alpha-diversity in Amazonian Ecuador. Biodivers Conserv 3: 21-28. doi: 10.1007/bf00115330

van Gemerden, B. S., H. Olff, et al. (2003) The pristine rain forest? Remnants of historical human impacts on current tree species composition and diversity. Journal of Biogeography 30: 1381-1390. doi: 10.1046/j.1365-2699.2003.00937.x

Whittall, J. B., J. Syring, et al. (2010) Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. Molecular Ecology 19: 100-114. doi: 10.1111/j.1365-294X.2009.04474.x

Wyman, S. K., R. K. Jansen, et al. (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252-3255. doi: 10.1093/bioinformatics/bth352

Yang, J.-B., M. Tang, et al. (2013) Complete chloroplast genome of the genus Cymbidium: lights into the species identification, phylogenetic implications and population genetic analyses. BMC Evolutionary Biology 13: 84. doi:

**Figure Captions**

**Fig. 1** Percentageof 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

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

**Conflict of Interest:** The authors declare that they have no conflict of interest.

**Funding:** This work was supported in part by an Innovational Research Incentives Scheme grant (VENI, nr. 863.09.017; NWO-ALW) to RHJE.