The complete sequence of the mitochondrial genome of Butomus umbellatus - a member of an early branching lineage of monocotyledons
Cuenca Navarro, Argelia; Petersen, Gitte; Seberg, Ole

Published in:
PLoS ONE

DOI:
10.1371/journal.pone.0061552

Publication date:
2013

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Cuenca Navarro, A., Petersen, G., & Seberg, O. (2013). The complete sequence of the mitochondrial genome of Butomus umbellatus - a member of an early branching lineage of monocotyledons. PLoS ONE, 8(4), [e61552]. https://doi.org/10.1371/journal.pone.0061552
The Complete Sequence of the Mitochondrial Genome of *Butomus umbellatus* – A Member of an Early Branching Lineage of Monocotyledons

Argelia Cuenca*, Gitte Petersen*, Ole Seberg*

Botanical Garden, Natural History Museum of Denmark, University of Copenhagen, Copenhagen K, Denmark

Abstract

In order to study the evolution of mitochondrial genomes in the early branching lineages of the monocotyledons, i.e., the Acorales and Alismatales, we are sequencing complete genomes from a suite of key taxa. As a starting point the present paper describes the mitochondrial genome of *Butomus umbellatus* (Butomaceae) based on next-generation sequencing data. The genome was assembled into a circular molecule, 450,826 bp in length. Coding sequences cover only 8.2% of the genome and include 28 protein coding genes, four rRNA genes, and 12 tRNA genes. Some of the tRNA genes and a 16S rRNA gene are transferred from the plastid genome. However, the total amount of recognized plastid sequences in the mitochondrial genome is only 1.5% and the amount of DNA transferred from the nucleus is also low. RNA editing is abundant and a total of 557 edited sites are predicted in the protein coding genes. Compared to the 40 angiosperm mitochondrial genomes sequenced to date, the GC content of the *Butomus* genome is uniquely high (49.1%). The overall similarity between the mitochondrial genomes of *Butomus* and *Spirodela* (Araceae), the closest relative yet sequenced, is low (less than 20%), and the two genomes differ in size by a factor 2. Gene order is also largely unconserved. However, based on its phylogenetic position within the core alismatids *Butomus* will serve as a good reference point for subsequent studies in the early branching lineages of the monocotyledons.

Introduction

The mitochondrial genomes of embryophytes or land plants are renowned among eukaryotes for their astonishing complexity, and evolutionary plasticity seems to prevail among angiosperms in particular. The 40 complete angiosperm mitochondrial genomes sequenced as of December 3rd 2012 (GenBank) vary enormously in size, from only ca. 0.2 Mb in *Brassica* [1] to ca. 11.3 Mb in *Silene conica* [2], and in mapping structure. The majority of mitochondrial genomes can be mapped to a single circular molecule, except in *Silene* and *Cucumis* where they were mapped to several individual chromosomes [3,4]. Even when the genomes can be mapped to a master circle [5] this structure is not stable and sub-molecules are formed, facilitated by recombination across repeated sequences [6]. Consequently, gene order is remarkably unconserved even between closely related species e.g., [1,7].

Other peculiarities of angiosperm mitochondrial genomes include substantial variation in gene content, frequent intracellular import of foreign sequences from the plastid and the nuclear genome, export of sequences to the nuclear genome, incorporation of reverse transcribed gene sequences known as “processed paralogs”, frequent RNA editing of both coding and non-coding sequences, vast substitution rate heterogeneity, and postulated ability to import and export sequences horizontally, i.e. across species boundaries (see [4] for a recent review). A striking aspect of mitochondrial evolution in angiosperms is the fact that most of the variable features mentioned above not just vary across the group as a whole but even among closely related taxa; e.g., species of *Silene* [2,8,9], *Brassica* [1], *Pelargonium* [10,11], and *Plantago* [11,12].

Given the huge structural differences sometimes even among closely related species and the limited number of completely sequenced genomes, it is not surprising that the evolutionary mechanisms facilitating these changes are largely unknown. Thus, a much wider taxonomic sampling is needed in order to throw light on other aspects of mitochondrial evolution. Most mitochondrial genomes sequenced to date are from agronomically important species and the taxonomic distribution of the 40 angiosperm genomes is heavily skewed: 12 genomes are from monocotyledons, but of these 10 are grasses, and the remaining 28 genomes are from 14 families of eudicots. Thus, mitochondrial structure and evolution in the early branching angiosperm lineages remains completely unknown though the sequence of the mitochondrial genome of *Amborella*, assumed to be the sister group to all other angiosperms, may be under way [13]. In the monocotyledons the only two non-grass genomes are from *Spirodela*, duckweed, (Araceae) [14] and *Phoenix*, date palm, (Arecaceae) [15].

The present paper is intended as the first in a series devoted to the study of the evolution of whole mitochondrial genomes in the early branching lineages of monocotyledons, i.e., the Acoraceae

Copyright: © 2013 Cuenca et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.


Editor: Luis Herrera-Estrella, Centro de Investigación y de Estudios Avanzados del IPN, Mexico

Received December 9, 2012; Accepted March 10, 2013; Published April 24, 2013

*E-mail: acuenca@snm.ku.dk (AC); gittep@snm.ku.dk (GP); oles@snm.ku.dk (OS)

Competing Interests: The authors have declared that no competing interests exist.

Funding: The research was supported by the Danish Council for Independent Research | Natural Sciences (FNU; http://en.dk/councils-commissions/the-danish-council-for-independent-research/scientific-research-councils/natural-sciences) grant no. 272-06-0436. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

* E-mail: acuenca@snm.ku.dk (AC); gittep@snm.ku.dk (GP); oles@snm.ku.dk (OS)
and the Alismatales. According to the APG III [16], Acoraceae is the sister group to all other monocotyledons and within these Alismatales is the sister group to the remaining. Studies of selected genes have previously shown evolutionary anomalies in parts of the Alismatales [17–19]. In the present paper, we describe the first completely sequenced genome from a member of the core alismatids, i.e., Alismatales, excl. Araceae and Tofieldiaceae. Based on a single genome only few evolutionary questions can be addressed. However, due to its the phylogenetic position within the core alismatids [17], it will serve as an appropriate reference for subsequent comparisons.

Materials and Methods
Mitochondrial Genome Sequencing
Fresh plant material was collected from a single individual (voucher: Seberg et al. C2457 (C)) at St.Vejleå, Ishøj in Denmark (N 55°37.496' E 12°21.939'). No specific permits were required for the collection of this material. The species is not protected by Danish law and it is collected in a public area where no permits are needed. Intact mitochondria were isolated by centrifugation following a modified protocol of Triboush et al. [20], and using DNAase I to digest nuclear and other DNA contaminants. Mitochondrial DNA was extracted using CTAB and a regular chloroform-isomilic DNA isolation protocol. Whole genome amplifications were carried out using the repli-g kit (Qiagen), following the manufacturer’s protocol. To check the identity of the DNA obtained by whole genome amplifications, we attempted to amplify four partial mitochondrial genes (cox1B, mtCO2, nad1, and nad5), one partial plastid gene (rbcL), and the nuclear ITS region. Whereas all mitochondrial amplifications were successful, the plastid and nuclear amplifications were not indicating that the extraction and subsequent whole genome amplification procedure provided clean mitochondrial DNA.

A standard 454 FLX (Roche, USA) shoot gun library was constructed and sequenced in a quarter of a GS PicoTiterPlate according to the manufacturer’s instructions at the National Highthroughput DNA Sequencing Centre, University of Copenhagen.

Sequence Assembly
A total of 37,048 sequences (average size 207 nt) were assembled in Newbler 2.3 (454 Life Sciences Corp, GT, USA) using default settings. This resulted in 572 contigs ranging from 100 to 56,599 nt and 76 were longer than 500 nt (average size 6238 nt). The contigs were extended by blasting the last ca. 75 nt of each contig border against a database of all raw 454 sequence reads. This allowed us to determine the borders of duplications and to identify reads of adjacent contigs. All BLAST analyses were done using the BLASTN program in the stand-alone BLAST ver. 2.2.21 [ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/]. In addition, the consensus sequence of each contig was used as seed sequences and extended using the Short Sequence Assembly by K-mer search and 3’ read Extension program, SSAKE ver. 3.5 [21], with parameters -m 13 -o 2 -r 0.6 -p 0 -t 0 -v 1. In cases where contig extension was not possible, primers were designed and gap closure was done by combinatorial PCR [22] and regular Sanger sequencing. Sequencing data is deposited in DRYAD, DOI: 10.5061/dryad.42gc4. The assembled sequence is deposited in GenBank under accession number KC208619.

Sequence Analyses
Coding regions were identified by BLASTX searches performed against a local database including amino acid sequences for all coding genes from 20 plant mitochondrial genomes available in GenBank The exact gene and exon boundaries were determined by alignment of homologous genes from available annotated plant mitochondrial genes. Similarly, rRNA and tRNA genes were identified by BLASTN searches against a local database including all rRNA genes and a database including all tRNA genes from 20 available land plant mitochondrial genomes. rRNAs were annotated using tRNAscan-SE [23,24]. To identify potential regions of plastid origin the mitochondrial genome of Butomus was blasted against a database of 20 land plant plastid genomes, including Lemna (Araceae) and Acorus (Acoraceae), both closely related to Butomus. Only sequences >50 bp and with a similarity score higher than 90% were considered. To identify potential regions of nuclear origin we primarily searched for repetitive elements using the Repbase Update repetitive element data base [25], but in addition a BLASTN search of three long intergenic regions of the mitochondrial genome were performed against the GenBank Nucleotide Collection filtering for plastid and mitochondrial sequences. To test the overall similarity between the entire Butomus mitochondrial genome and other complete angiosperm mitochondrial genomes a BLASTN search was performed using complete mitochondrial sequences as input.

Phylogenetic Analyses
Sequences of 24 mitochondrial genes from Butomus and 25 seed plant species, for which the complete mitochondrial genome is available, were extracted from GenBank (see Table 1). The genes include all protein coding genes present in Butomus except the ribosomal genes (Table 2). Alignments were generated for each individual gene using MUSCLE [26] integrated in Geneious Pro (ver. 5.3.6; Biomatters Ltd.) with default parameters and concatenated into a matrix of a total of 28,455 characters. The matrix has a few missing entries, viz. the cox2 gene is missing in Vigna and the mtT8 gene is missing in Vitis and Boea. A Maximum Likelihood tree was constructed using the program PhyML [27] with a GTR substitution model and a gamma distribution of substitution rates estimated with four categories.

To investigate substitution rate diversity among rRNA genes individual alignments were also done for each of the three rRNAs universally present in the plant mitochondria (rrn5, rrn18, and rrn26) using MUSCLE integrated in Geneious Pro. Alignments were performed for the same taxa as above, except that Boea was removed because its rrn18 sequence appear quite different and could not be aligned easily. The phylogenetic tree resulting from the analysis of the 24 protein coding genes was used to estimate the substitution rate of the three rRNAs. To estimate substitution rates, the JC+G model was used for rrn5 (119 bp), the TPM1+G (K81) model for rrn18 (2553 bp), and the GTR+G model for rrn26 (4498 bp), as suggested by jModelTest 0.1.1 [28]. All substitution rate were calculated using the program PAML 4.3 [29].

Results and Discussion
The Mitochondrial Genome of Butomus
A total of 63,056 individual sequences, corresponding to 72% of the total number of sequences obtained, were assembled into a circular molecule of 450,026 bp (Fig. 1; GenBank acc. no. KC208619) giving an average coverage of 32×. This circular molecule correspond to the so-called master circle [5] containing all mitochondrial genes of Butomus.

With a total genome size of ca. 451 kb the mitochondrial genome of Butomus is almost twice as large as the ca. 220 kb genome of Spirodela [14], the only other genome from the
Table 1. Comparison of complete mitochondrial genomes of selected seed plants.

<table>
<thead>
<tr>
<th>Species</th>
<th>Refseq</th>
<th>Length</th>
<th>GC%</th>
<th>Coding%</th>
<th>Protein%</th>
<th>tRNA</th>
<th>rRNA</th>
<th>Introns%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycas taitungensis</td>
<td>NC_010303</td>
<td>414,903</td>
<td>46.9</td>
<td>10.0</td>
<td>41</td>
<td>22</td>
<td>3</td>
<td>25 (5)</td>
</tr>
<tr>
<td>Butomus umbellatum</td>
<td>K208619</td>
<td>450,826</td>
<td>49.1</td>
<td>8.2</td>
<td>28</td>
<td>12</td>
<td>4</td>
<td>21 (5)</td>
</tr>
<tr>
<td>Spirodela polyrhiza</td>
<td>NC_017840</td>
<td>228,493</td>
<td>45.7</td>
<td>16.2</td>
<td>35</td>
<td>18</td>
<td>3</td>
<td>21 (6)</td>
</tr>
<tr>
<td>Phoenix dactylifera</td>
<td>NC_016740</td>
<td>715,001</td>
<td>45.1</td>
<td>6.3</td>
<td>38</td>
<td>22</td>
<td>3</td>
<td>24 (4)</td>
</tr>
<tr>
<td>Bambusa oldhamii</td>
<td>EU365401</td>
<td>509,941</td>
<td>44.1</td>
<td>7.8</td>
<td>35</td>
<td>18</td>
<td>3</td>
<td>22 (6)</td>
</tr>
<tr>
<td>Oryza sativa ssp. indica</td>
<td>NC_007886</td>
<td>491,515</td>
<td>43.8</td>
<td>10.2</td>
<td>35</td>
<td>21</td>
<td>3</td>
<td>23 (6)</td>
</tr>
<tr>
<td>Triticum aestivum</td>
<td>NC_007579</td>
<td>452,528</td>
<td>44.4</td>
<td>11.4</td>
<td>33</td>
<td>16</td>
<td>3</td>
<td>22 (6)</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>NC_008360</td>
<td>468,628</td>
<td>43.7</td>
<td>8.2</td>
<td>32</td>
<td>19</td>
<td>3</td>
<td>22 (6)</td>
</tr>
<tr>
<td>Tripsacum dactyloides</td>
<td>NC_008362</td>
<td>704,100</td>
<td>43.9</td>
<td>5.9</td>
<td>32</td>
<td>17</td>
<td>3</td>
<td>22 (6)</td>
</tr>
<tr>
<td>Zea luxians</td>
<td>NC_008333</td>
<td>539,368</td>
<td>43.9</td>
<td>7.2</td>
<td>32</td>
<td>16</td>
<td>3</td>
<td>22 (6)</td>
</tr>
<tr>
<td>Zea mays ssp. mays</td>
<td>NC_007982</td>
<td>569,630</td>
<td>43.9</td>
<td>8.4</td>
<td>32</td>
<td>18</td>
<td>3</td>
<td>22 (7)</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>NC_012119</td>
<td>773,279</td>
<td>44.1</td>
<td>6.3</td>
<td>38</td>
<td>23</td>
<td>3</td>
<td>25 (5)</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>NC_001284</td>
<td>366,924</td>
<td>44.8</td>
<td>11.5</td>
<td>31</td>
<td>17</td>
<td>3</td>
<td>23 (5)</td>
</tr>
<tr>
<td>Brassica napus</td>
<td>NC_008285</td>
<td>221,853</td>
<td>45.2</td>
<td>17.4</td>
<td>32</td>
<td>18</td>
<td>3</td>
<td>24 (5)</td>
</tr>
<tr>
<td>Carica papaya</td>
<td>NC_012116</td>
<td>476,890</td>
<td>45.1</td>
<td>8.6</td>
<td>39</td>
<td>20</td>
<td>3</td>
<td>24 (5)</td>
</tr>
<tr>
<td>Citrullus lanatus</td>
<td>NC_014043</td>
<td>379,236</td>
<td>45.1</td>
<td>10.3</td>
<td>38</td>
<td>20</td>
<td>3</td>
<td>24 (5)+11</td>
</tr>
<tr>
<td>Cucurbita pepo</td>
<td>NC_014050</td>
<td>982,833</td>
<td>42.8</td>
<td>3.9</td>
<td>38</td>
<td>26</td>
<td>3</td>
<td>24 (5)</td>
</tr>
<tr>
<td>Cucumis sativus</td>
<td>NC_016005</td>
<td>1,555,935</td>
<td>44.3</td>
<td>2.9</td>
<td>37</td>
<td>22</td>
<td>3</td>
<td>23 (5)+11</td>
</tr>
<tr>
<td>Lotus japonicus</td>
<td>NC_016743</td>
<td>380,861</td>
<td>45.4</td>
<td>9.9</td>
<td>31</td>
<td>16</td>
<td>3</td>
<td>22 (5)</td>
</tr>
<tr>
<td>Milletia pinnata</td>
<td>NC_016742</td>
<td>425,718</td>
<td>45.0</td>
<td>9.7</td>
<td>33</td>
<td>17</td>
<td>3</td>
<td>23 (5)</td>
</tr>
<tr>
<td>Vigna radiate</td>
<td>NC_015121</td>
<td>401,262</td>
<td>45.1</td>
<td>8.8</td>
<td>31</td>
<td>14</td>
<td>3</td>
<td>22 (5)</td>
</tr>
<tr>
<td>Nicotiana tabacum</td>
<td>NC_006581</td>
<td>430,597</td>
<td>45.0</td>
<td>10.1</td>
<td>37</td>
<td>21</td>
<td>3</td>
<td>23 (6)</td>
</tr>
<tr>
<td>Beta vulgaris</td>
<td>NC_002511</td>
<td>368,801</td>
<td>43.9</td>
<td>11.3</td>
<td>30</td>
<td>21</td>
<td>3</td>
<td>20 (6)</td>
</tr>
<tr>
<td>Silene latifolia</td>
<td>NC_014847</td>
<td>253,413</td>
<td>42.6</td>
<td>14.5</td>
<td>25</td>
<td>9</td>
<td>3</td>
<td>19 (6)</td>
</tr>
<tr>
<td>Daucus carota</td>
<td>NC_017855</td>
<td>281,132</td>
<td>45.4</td>
<td>18.4</td>
<td>34</td>
<td>17</td>
<td>3</td>
<td>24 (5)</td>
</tr>
<tr>
<td>Boea hygrometrica</td>
<td>NC_016741</td>
<td>510,519</td>
<td>43.3</td>
<td>7.6</td>
<td>30</td>
<td>23</td>
<td>3</td>
<td>23 (5)</td>
</tr>
<tr>
<td>Ricinus communis</td>
<td>NC_015141</td>
<td>502,773</td>
<td>45.0</td>
<td>7.6</td>
<td>37</td>
<td>16</td>
<td>3</td>
<td>24 (5)</td>
</tr>
</tbody>
</table>

Alismatales sequences so far, but closer in size to the grass genomes ranging from ca. 453–704 kb (Table 1).

The nucleotide composition of the Butomus genome has the highest GC content, 49.1%, reported so far (Table 1). In other seed plants the GC% content ranges from ca. 43–47% (Table 1). Thus, the three highest percentages are found in Cycas, Butomus, and Spirodela, which vaguely suggest a decline in GC content during angiosperm evolution.

Coding regions constitute only 8.2% of the Butomus genome and with a total gene content of 28 protein-coding genes, 12 tRNA genes, and 3 (+1 CP) rRNA genes (Table 1, 2) the content of coding DNA both in terms of coverage and actual gene numbers is relatively low but not unusual compared to other angiosperms (Table 1). The genes are very unevenly distributed across the circular genome (Fig. 1). Three regions of 33–41 kb are completely devoid of genes or any other recognisable features and in a region spanning ca. 113 kb, i.e. approximately one quarter of the genome, only four genes are located (Fig. 1). Compared to the Spirodela genome, gene order is largely unconserved as reported from other comparative genome studies e.g., [1,7].

Based on BLAST searches the entire mitochondrial genome of Butomus has the highest similarity to the ca. 526 kb genome of Mimulus [30], which matches 23% of the Butomus mitochondrial genome. As the coding sequences of each genome take up only ca. 6 and 5%, respectively (Table 1), similarity clearly extends into the non-coding regions, thus suggesting some level of sequence conservation in those regions of the genome. It might have been expected that the genomes of Butomus and Spirodela, being each others closest sequenced relatives, would have been most similar, but with only 19% overall similarity this is not so. As the Spirodela genome is approximately half the size of the Butomus genome, it may be assumed that sequence loss has reduced overall genome similarity. The overall similarity of the Butomus genome and the 715 kb mitochondrial genome of Phoenix, the only other non-grass monocotyledon, is 22%. The total amount of sequence of the Butomus genome being similar to other Spermatophyta mitochon-
The Mitochondrial Genome of Butomus umbellatus

### Table 2. Gene content of the mitochondrial genome of Butomus umbellatus.

<table>
<thead>
<tr>
<th>Genes of mitochondrial origin</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex I (NADH dehydrogenase)</td>
<td>nad1, nad2, nad3, nad4, nad4L, nad5, nad6, nad7, nad9</td>
<td></td>
</tr>
<tr>
<td>Complex II (succinate dehydrogenase)</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Complex III (cytochrome C reductase)</td>
<td>ccoB</td>
<td></td>
</tr>
<tr>
<td>Complex IV (cytochrome C oxidase)</td>
<td>cox1, cox2, cox3</td>
<td></td>
</tr>
<tr>
<td>Complex V (ATP Synthase)</td>
<td>atp1, atp4, atp6, atp8, atp9</td>
<td></td>
</tr>
<tr>
<td>Cytochrome C biogenesis</td>
<td>ccma8, ccmC1, ccmF6, ccmF7</td>
<td></td>
</tr>
<tr>
<td>Other genes</td>
<td>matR, mttB</td>
<td></td>
</tr>
<tr>
<td>Large subunit ribosomal proteins</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Small subunit ribosomal proteins</td>
<td>rps1, rps3, rps7, rps12</td>
<td></td>
</tr>
<tr>
<td>Transfer RNAs</td>
<td>trnC-GCA, trnD-GUC, trnE-UUC&lt;sup&gt;5&lt;/sup&gt;, trnK-CUU, trnM-CAG&lt;sup&gt;5&lt;/sup&gt;, trnQ-UUG, trnY-GUA</td>
<td></td>
</tr>
<tr>
<td>Ribosomal RNAs</td>
<td>rns16, rns18, rns26</td>
<td></td>
</tr>
<tr>
<td>Pseudogenes, partial</td>
<td>sdh4-ψ, rpl16-ψ</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genes of plastid origin</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ribosomal RNAs</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Transfer RNAs</td>
<td>trnA-UGC, trnH-GUG&lt;sup&gt;5&lt;/sup&gt;, trnL-GAU, trnK-UUU, trnW-CCA</td>
<td></td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0061552.t002

Nuclear DNA Insertions

The mitochondrial genome may also integrate DNA from the nuclear genome [4,33]. However, identification of nuclear sequences may be problematic due to sequence changes over time, a very rare record of fully or even just partially sequenced nuclear genomes, and because sequence transfer between the nuclear and mitochondrial genomes occurs in both directions making it difficult to determine the directionality of transfer events of seemingly featureless DNA sequences [4,33]. Presence of DNA sequences similar to nuclear repetitive elements in mitochondrial genomes has been used as positive evidence of nuclear DNA transfer e.g., [7,34,35].

A similarity search of the mitochondrial genome of Butomus against the Repbase Update repetitive element data base [25] revealed very few good matches (>80% similarity of sequences >50 bp long). Most of these (15) were even very short (<600 bp), the remaining just 71, 75, and 126 bp long, respectively, and in total adding up to only 1076 bp. It is questionable whether these short sequences are indeed remnants of sequences transferred from the nuclear genome, but even if they are they constitute only 0.2% of the mitochondrial genome. This low percentage is consistent with the lack of repetitive elements in the mitochondrial genome of Spirodela [14]. As a supplement to the Repbase Update search we conducted a BLASTN search against GenBank sequences filtering for chloroplast and mitochondrial sequences. In this search we only used the three larger fragments of the Butomus mitochondrial genome devoid of recognizable features (see above) covering approximately a quarter of the genome. The search did not reveal any good matches to sequences of unequivocal nuclear origin. As we did not perform similar searches with the remaining intergenic regions of the mitochondrial genome we cannot rule out the possibility that some of these may include sequences of nuclear origin, but we assume that the general content of nuclear DNA in the Butomus mitochondrial genome is very low indeed.

Protein Coding Genes

The mitochondrial genome of Butomus contains 28 protein coding genes (Table 2, Fig. 2). Eighteen genes encode proteins of the respiratory chain (nine NADH dehydrogenase genes, five ATP
synthase genes, three cytochrome C reductases, one cytochrome C reductase) and four are involved in cytochrome C biogenesis. Together with matR and mttB these 24 genes are almost ubiquitously found in all complete seed plant mitochondria sequenced so far. Exceptions are the loss of cox2 in members of Fabaceae [36], of mttB in Vitis [37] and Boea [15], of a functional copy of ccmFc in Silene conica [2], and the potential loss of atp8 in Allium [38].

Two succinate dehydrogenase genes, sdh3 and sdh4, are present in the mitochondrial genome of Cycas, but based on Southern hybridization Adams et al. [39] suggested that these genes were lost repeatedly among the angiosperms, and in some cases transferred to the nucleus [38]. Among the Alismatales, data from Adams et al. [39] indicated presence of both genes in three genera of Araceae, and absence in two genera of Hydrocharitaceae and one genus of Alismataceae. Thus, the absence of both genes in Butomus is not surprising, though we have identified a partial, 157 bp long, pseudogene-like sequence of sdh4. In the completely sequenced mitochondrial genome of Spirodela (Araceae) sdh4 is present as suggested by Adams et al. [39], but sdh3 is not.

Only four ribosomal protein genes, rps1, rps3, rps7, and rps12, are found in the mitochondrial genome of Butomus. These genes are only rarely lost among angiosperms [Fig. 2] though Adams et al. [38] did report loss of rps7 in the Hydrocharitaceae and Alismataceae. A lower number of ribosomal genes have so far only been recorded for species of Silene [2]. In Spirodela, no less than 10 functional ribosomal genes were detected. In addition to the four complete ribosomal protein genes, the Butomus mitochondrial
Table 3. Sequences of plastid origin in mitochondrial genome of *Butomus umbellatus*.

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Length (bp)</th>
<th>Sequence characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>34.784</td>
<td>34.889</td>
<td>106</td>
<td>ycf2-ndhB intergenic region</td>
</tr>
<tr>
<td>47.701</td>
<td>48.603</td>
<td>903</td>
<td>ndhK-ndhC, partial</td>
</tr>
<tr>
<td>62.015</td>
<td>62.099</td>
<td>85</td>
<td>petD, partial</td>
</tr>
<tr>
<td>171.200</td>
<td>171.359</td>
<td>160</td>
<td>psbD, partial</td>
</tr>
<tr>
<td>215.497</td>
<td>215.653</td>
<td>157</td>
<td>rpoB, partial</td>
</tr>
<tr>
<td>249.922</td>
<td>249.992</td>
<td>71</td>
<td>psbD, partial</td>
</tr>
<tr>
<td>299.456</td>
<td>299.540</td>
<td>85</td>
<td>petD, partial</td>
</tr>
<tr>
<td>368.045</td>
<td>372.941</td>
<td>4.897</td>
<td>rrn16, trnA(ugg), trnI(gau)</td>
</tr>
<tr>
<td>434.652</td>
<td>434.714</td>
<td>63</td>
<td>rps3, partial</td>
</tr>
<tr>
<td>438.532</td>
<td>438.932</td>
<td>401</td>
<td>trnK(uuu)², trnW(cca)</td>
</tr>
</tbody>
</table>

| 6.928 |

¹Repetitive sequence.
²trnP(ugg) in CP genomes, trnK(uuu) in *Butomus*.

doi:10.1371/journal.pone.0061552.t003

genome includes a 398 bp long, pseudogene-like sequence of *rpl16*. Adams et al. [38] did not detect *rpl16* in the families Hydrocharitaceae and Alismataceae, both close relatives of *Butomus*.

The exact boundaries of less conserved protein coding genes may not be precisely determined neither in the present nor in previous investigations, where transcripts have not been studied. In addition to sequence variation, RNA editing which may affect all codons including start and stop codons potentially complicates precise assignments further. Accordingly, identification of the stop codon of *rps1* in the *Butomus* genome may be considered dubious. The 3’-end of the gene has little similarity to other *rps1* sequences and the designated stop codon is 19 bp inside a region supposed to be of plastid origin. Provided that the assignment of the gene and its stop codon, as well as the borders of the plastid insert region, are correct, it is possible that the plastid region has been inserted in the 3’-end of the gene thereby adding new amino acids and a new stop codon to the original *rps1* gene. However, it may be more probable that the true stop codon is located upstream of the currently recognized stop codon, but is not being recognized as such due to RNA editing.

rRNA Genes

The mitochondrial genome of *Butomus* includes three rRNA genes (5S, 18S, and 26S rRNA genes) and as in all other mitochondrial genomes the 5S and 18S genes are located in very close proximity, whereas the larger 26S gene has a distant position (Fig. 1). In addition to the mitochondrial rRNA genes a copy of the plastid 16S rRNA gene is found as part of a 4897 bp fragment of plastid DNA (see above). A complete 16S rRNA gene has previously been observed in the mitochondrial genome of *Boea* [15] and a fragment was observed in *Cucumis* (GenBank acc. no. JF412792). In *Silene latifolia* and *S. vulgaris* Sloan et al. [9] described a possible gene conversion event between a mitochondrial 18S rRNA gene and a plastid 16S rRNA gene. The evidence for conversion was a ca. 50 bp segment of the mitochondrial 18S rRNA gene showing substantially more similarity to plastid 16S rRNA genes than to other mitochondrial 18S rRNA gene. As the mitochondrial genome of *Silene* does not include a plastid 16S rRNA gene or any substantial fragments of it, Sloan et al. [9] hypothesized that the originally transferred plastid gene has been lost from the mitochondrial genome. The plastid 16S rRNA gene found in *Butomus*, is very likely a result of a recent transfer event (see above). Repeated transfer and incorporation into the mitochondrial genome of 16S rRNA is supported by the two occurrences of partial or complete 16S rRNA sequences in *Cucumis* and *Boea*. Thus lending indirect support to the hypothesis of gene conversion in *Silene*.

tRNA Genes

None of the angiosperm mitochondrial genomes sequenced to date has included a full set of tRNA genes, and genome of *Butomus* is no exception. Only 12 different tRNA genes could be found (Table 2, Fig. 2) and five of them (trnA(ugg), trnH(agg), trnG(au), trnK(ugu), trnW(cca)) may be of plastid origin. The trnH(agg) gene has been indicated to be of plastid origin in all previously sequenced angiosperm mitochondrial genomes (except *Triteum* as well as in *Cycas* [40], thus the inferred transfer event appear ancient. In contrast, both trnA(ugg) and trnI(gau) are located in a fragment of plastid DNA which also includes 16S rRNA and apparently represents a very recent transfer (see above). This is consistent with the two tRNA genes being absent from all other mitochondrial genomes sequenced so far. As the only tRNA genes found in *Butomus*, the latter two contain introns of 684 bp (trnA(ugg)) and 937 bp (trnI(gau)), respectively.

With the exception of *Zea* all complete mitochondrial genomes of angiosperms seem to include a copy of trnW(cca) similar to the corresponding plastid gene, whereas *Cycas* has a copy of trnW(cca) more similar to the mitochondrial versions of the gene found in e.g., mosses, liverworts, algae, etc. Consistent with a plastid origin, we find the trnW(cca) gene in a larger region apparently of plastid origin and shared by many angiosperms (see above). Thus, data suggests that the transfer took place either in the early evolution of the angiosperms or even prior to that. The trnK(uuu) gene of *Butomus* is included in the same fragment, potentially of plastid origin, as trnW(cca) (see above), but in plastid genomes the >90% similar sequence encodes trnP(agg). This clearly illustrates the ambiguity in trying to determine the homology of individual tRNA genes. Due to high sequence similarity between many tRNA genes and the automated naming of the gene, which may change due to a single base change in the anticodon region, similar named genes may not be homologous whereas differently named genes, such as here trnK(uuu) and trnP(agg), may be truly homologous. Here we list trnK(uuu) as of plastid origin (Table 2, Fig. 2), but the sequence may be homologous to sequences either listed as mitochondrial trnP's, trnK's, or even other genes in other studied mitochondrial genomes. The wobbly pairing mechanism, which further introduces ambiguity in codon-anticodon recognition [41], makes naming and prediction of tRNAs even more complicated when only raw sequence information is available. Thus, the tRNA gene content listed in figure 2 and similar tables in other paper e.g., [9,42], need not reflect functionality or homology precisely, and the distinction between tRNA genes of plastid and mitochondrial origin is fuzzy and should be considered with great caution.

Introns

In *Butomus* we found a total of 21 group II introns distributed among seven protein coding genes (Table S1). Most of the introns are cis-spliced, but five are trans-spliced (two in nad1, one in nad2, and two in nad5). The cis-spliced introns include a total of 23,946 bp corresponding to ca. 5% of the mitochondrial genome of *Butomus*. Compared to the overview of intron content of selected angiosperms provided by Mower et al. [30], *Butomus* includes the same 12 cis-spliced and five trans-spliced introns indicated as universally present among angiosperms. The second intron of nad4 and the third intron of nad7, which are not universally present, are.
both found in the *Butomus* genome and so far they appear to be present in all monocotyledons. The fourth intron of nad1 which varies between being cis- and trans-spliced, is cis-spliced in *Butomus*, whereas it is trans-spliced in grasses, *Phornix*, and some eudicotyledons [30,42]. We interpret the intron to be trans-spliced in *Spirodela*, even though there are discrepancies between the figure of the mitochondrial genome (Figure 3 in [14]), a table of genes (Table S3 in [14]), and the GenBank record (NC_017840). Compared to the *Butomus* mitochondrial genome and the genomes reviewed by Mower et al. [30] *Spirodela* appears to lack the third of the five exons of nad5 [14]. However, this exon is very short (usually just 21–22 bp) and easily overlooked, but a BLAST search reveals its presence in *Spirodela* with the flanking introns both being trans-spliced. With these corrections regarding nad1 and nad5, the total number of introns (21) and the distribution of cis-spliced (15) and trans-spliced (6) introns given by Wang et al. [14] becomes correct, though tables, figures, and the GenBank file all seem to contain errors.

In *cox2*, most angiosperms have either one or more rarely two introns (see [30]), but *Butomus* lacks both these introns. As in many other plant mitochondria, matR is located within intron 4 of nad1. The group I intron found in *cox1* of many angiosperms is not present in *Butomus* consistent with its reported lack in three species of core alismatids, including two species of Hydrocharitaceae, the sister group to *Butomus* [43].

In addition to the introns located in protein coding genes, we found one intron in each of two tRNA genes, *trnA*(UGC) and

---

**Figure 2. Gene content in selected seed plant mitochondrial genomes.** Dark squares indicate presence of assumed intact, functional genes, lighter gray squares indicate putative pseudogenes, and white squares indicate absence of genes. Numbers below columns are the total numbers of assumed intact, functional genes. The table is modified and expanded from Sloan et al. [2].

doi:10.1371/journal.pone.0061552.g002
trnI(gau), both presumably transferred from the plastid genome (see above).

Repetitive Sequences
Repetitive sequences, often of considerable length, are common in plant mitochondria; the longest being a perfect 87 kb repeat found in *Beta* [44]. Differences in overall mitochondrial genome size can to some extent be explained by presence of such repeats [7].

In *Butomus* we found one long, direct repeat of ca. 9 kb (RR1; Table 4, Fig. 1), and two shorter, inverted repeats of 6.3 kb and 2.8 kb, respectively (RR2 and RR3; Table 4, Fig. 1). Additionally, 40 short repeats (from 50 bp to 1 kb) were identified ranging in copy numbers from two to five (Table 4 [L,200], Table S2 [50<L<200]). Repeats shorter than 50 bp are not listed. The repeated sequences cover approximately 37.8 kb (8.3%) of the mitochondrial genome of *Butomus*. However, some of the repeats are overlapping and the entire fraction of the mitochondrial genome covered by repeated sequences is less than the sum of lengths of all repeat fragments. The amount of repeats and their contribution to genome size falls within the range found in other angiosperm genomes, where the lowest content is 2.3% found in *Phoenix* and the highest 36.4% found in *Tripsacum* [42].

Some of the repeated sequences in *Butomus* include entire genes, which thus occur in duplicate or triplicate. The protein coding *ccmC* gene occurs in three identical copies, and the tRNAs *trnE(uuc)* and *trnH(gug)*, occur in three and two identical copies, respectively.

In particular longer repeats, i.e. >1 kb, are thought to mediate homologous recombination [6], but to what extent repeats found in *Butomus* are involved in recombination is unknown.

**RNA Editing**
In all land plants (Embryophytes), except marchantiid liverworts, transcripts of the mitochondrial genome are known to experience extensive RNA editing, primarily as C-to-U changes [45,46]. Editing mostly affects protein coding gene, breaking the normal 1:1 correspondence between DNA sequence and amino acid sequence of the protein, but non-protein coding sequences,
e.g., introns and tRNAs may also be edited (e.g., [47] and references therein).

The exact number of edited positions in a complete genome is rarely known. This is primarily due to lack of transcriptome data, but also due to the only partial or tissue specific nature of editing [48]. In the best studied angiosperm genomes the total number of edited sites (in the protein coding genes) range from ca. 200–500 positions [7,8,33,48–51]. In Phoenix and Spirodela approximately 600 sites are estimated to be edited [14,42], in Amborella an unpublished survey revealed 835 edited sites [4], and in the only non-angiosperm seed plant investigated to date, Cycas, more than 1000 sites were estimated to be edited [40]. Using PREP-Mt [52,53] we estimate that in Butomus 28 protein coding genes include 557 edited sites.

In an earlier study of selected mitochondrial genes in alismatids we experimentally verified RNA editing in partial atp1 and ccmB sequences of Butomus [18]. Compared to the predictions made here, we found two additional edited sites in atp1. In ccmB we found six extra edited sites, but PREP-Mt predicted a further six sites to be edited, which were actually not edited. Thus, the PREP-Mt predictions occasionally slightly underestimate the actual number of edited sites. Despite using a higher cut-off values than us (0.6 vs. 0.2), thereby lowering the number of predicted edited sites, Wang et al. [14] report that PREP-Mt may potentially produce almost 10% erroneous predictions. However, they do not report the number of edited sites found only by comparisons to RT-PCR sequences of the four investigated genes, i.e. they report only false positives, not false negatives. Regardless, the 557 sites predicted to be edited in Butomus, should not be considered exact, but the figure is in line with the range found in other genomes.

All protein coding genes in Butomus are estimated to have edited sites. In some genera of the Alismataceae and the Hydrocharitaceae previous studies have shown that some mitochondrial genes lack editing completely [17,18]. It was suggested that those genes originated through retrotranscription (“processed paralogs”) and in a number of core alismatids (but not in Butomus) two copies of nad1, one including and one excluding edited sites, respectively, were found [19]. Even though these earlier studies have suggested that the patterns of evolution of both the genes and RNA editing in the alismatids are in need of further exploration, present data from Butomus confirm absence of aberrant traits in this taxon.

**Phylogeny and Substitution Rates**

As expected, a phylogenetic analysis of 24 protein coding genes from 25 angiosperms and Cycas places Butomus as sister to Spirodela (Fig. 3) in accordance with previous phylogenetic studies [54]. Despite the poor taxon sampling the tree is generally in good agreement with the current phylogenies of the angiosperms (see http://www.flnnh.ufl.edu/angiospermATOL/index.html). At family level only the positions of Vitis and Ricinus differ from current views, but within the three families represented by more than two genera the relationships within Fabaceae is resolved as expected [55], whereas the relationships within Cucurbitaceae is not [56], and the resolution within Poaceae is only partially so [57].

Compared to Spirodela the overall substitution rate of the protein coding genes of Butomus is significantly increased and a similar difference in rates can be seen between Phoenix and the grass clade (Fig. 3). Based on data from individual mitochondrial genes Petersen et al. [17] and Cuenca et al. [18] have previously found a highly elevated substitution rate in the core alismatids in comparison not only to members of the Araceae, but to most monocotyledons, and palms were found to have a very low substitution rate. The substitution rate difference between Butomus...
and *Spirodela* differs for individual protein coding genes, but we consistently find a higher rate in *Butomus* (data not shown).

Sloan et al. [9] investigated substitution rates of mitochondrial ribosomal genes and discovered an extremely elevated substitution rate of the SS rRNA gene in *Silene latifolia*. In *Butomus* we observe an elevated but not significantly different substitution rate of all three rRNA genes compared to the genes of all other monocotyledons, but not to all core eudicotyledons (Fig. 4).

**Perspectives**

The number of complete plant mitochondrial genomes is growing rather slowly, primarily due to their complex and labile structure, large amounts of repeats, and their alleged ability to accept alien DNA sequences both through intracellular (viz. from the nucleus and plastids) and horizontal gene transfer (viz. transgressing species boundaries). A potential further complication is RNA editing and the occurrence of processed paralogs.

The limited number of available mitochondrial plant genomes in GenBank is in stark contrast to the number of complete plastid genomes and animal mitochondrial genomes present. However, the majority of animal mitochondrial genomes and plastid genomes are characterized by sharing a rather monotone size and structure, similarities clearly reflected in the number of organelle genomes in GenBank, thus as of December 3rd 2012 there are 72 complete plant (only 41 from Spermatophyta) and 2831 complete metazoan mitochondrial genomes and 223 complete plastid genomes (120 from Spermatophyta).

The vast majority of the sequenced plant mitochondrial genomes are from commercially important crop species and few are selected for their phylogenetic importance. This bias (e.g. of the 40 sequenced genomes from angiosperms, 10 genomes are from grasses) evidently impair our ability to acquire an in-depth understanding of key evolutionary issues such as; how much gene loss is tolerated in the mitochondria of photosynthetic plants or in plants showing different levels of parasitism (e.g. from facultative parasitism to holo-parasitism)? It also weakens our ability to obtain solid evidence for HGT, which is primarily based on conflicts in phylogenetic evidence [30].

The relatively simple nature of animal mitochondria and plant plastids made them ideally suited for extracting phylogenetic information, and sequences from these – usually maternally inherited – organelles have dominated animal and plant phylogeny for two decades. This is most likely not going to change over-night, but these phylogenies will increasingly be challenged by genomic data, from the nucleus, and in plants also from the mitochondrion.

**Supporting Information**

**Table S1** Introns in mitochondrial genes of *Butomus umbellatus*.

**Table S2** Repeated sequences >50 bp in the mitochondrial genome of *Butomus umbellatus*.

**Table S3** Predicted edited sites in protein coding genes in the mitochondrial genome of *Butomus umbellatus*.

**Acknowledgments**

We thank Charlotte Hansen for skillful help with lab work and the staff at the National High-throughput DNA Sequencing Centre, University of Copenhagen, for library construction, 454-sequencing, and initial fragment assembly.

**Author Contributions**

Conceived and designed the experiments: AC GP OS. Performed the experiments: AC. Analyzed the data: AC GP. Wrote the paper: AC GP OS.

**References**


