2003

Nuclear Genes That Encode Mitochondrial Proteins for DNA and RNA Metabolism Are Clustered in the Arabidopsis Genome

Annakaisa Elo
University of Nebraska-Lincoln

Anna Lyznik
University of Nebraska-Lincoln

Delkin O. Gonzalez
University of Nebraska - Lincoln

Stephen D. Kachman
University of Nebraska-Lincoln, steve.kachman@unl.edu

Sally A. Mackenzie
University of Nebraska-Lincoln, sam795@psu.edu

Follow this and additional works at: http://digitalcommons.unl.edu/statisticsfacpub

Part of the Other Statistics and Probability Commons

Elo, Annakaisa; Lyznik, Anna; Gonzalez, Delkin O.; Kachman, Stephen D.; and Mackenzie, Sally A., "Nuclear Genes That Encode Mitochondrial Proteins for DNA and RNA Metabolism Are Clustered in the Arabidopsis Genome" (2003). Faculty Publications, Department of Statistics. 49.

http://digitalcommons.unl.edu/statisticsfacpub/49

This Article is brought to you for free and open access by the Statistics, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Faculty Publications, Department of Statistics by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Nuclear Genes That Encode Mitochondrial Proteins for DNA and RNA Metabolism Are Clustered in the Arabidopsis Genome

Annakaisa Elo, Anna Lyznik, Delkin O. Gonzalez, Stephen D. Kachman, and Sally A. Mackenzie

The plant mitochondrial genome is complex in structure, owing to a high degree of recombination activity that subdivides the genome and increases genetic variation. The replication activity of various portions of the mitochondrial genome appears to be nonuniform, providing the plant with an ability to modulate its mitochondrial genotype during development. These and other interesting features of the plant mitochondrial genome suggest that adaptive changes have occurred in DNA maintenance and transmission that will provide insight into unique aspects of plant mitochondrial biology and mitochondrial-chloroplast coevolution. A search in the Arabidopsis genome for genes involved in the regulation of mitochondrial DNA metabolism revealed a region of chromosome III that is unusually rich in genes for mitochondrial DNA and RNA maintenance. An apparently similar genetic linkage was observed in the rice genome. Several of the genes identified within the chromosome III interval appear to target the plastid or to be targeted dually to the mitochondria and the plastid, suggesting that the process of endosymbiosis likely is accompanied by an intimate coevolution of these two organelles for their genome maintenance functions.

INTRODUCTION

The plant mitochondrial genome is characterized by unusual structural complexity. A multipartite genome structure derives from both high-frequency recombination within defined repeated sequences in the genome and low-frequency ectopic recombination events (Fauron et al., 1995). Evidence of both types of activity is extensive in many higher plants (Mackenzie and McIntosh, 1999). Although plant mitochondrial genomes map as circular molecules, they appear to be predominantly linear in vivo (Bendich and Smith, 1990; Backert et al., 1997; Backert and Börner, 2000; Oldenburg and Bendich, 2001), and the role of recombination in their replication is not yet known. Although a rolling-circle replication mechanism has been speculated based on electron microscopic analyses of DNA structures (Backert and Börner, 2000), virtually none of the mitochondrial DNA metabolism apparatus has been well defined genetically or biochemically in plants.

One particularly intriguing behavior observed in plant mitochondria is a dynamic system of genome copy number modulation. With a multipartite genome structure, the plant mitochondrion contains a highly redundant gene assemblage organized within subgenomic DNA molecules. The copy number of at least some of these subgenomic molecules appears to be regulated independently (reviewed by Mackenzie and McIntosh, 1999). The phenomenon termed substoichiometric shifting refers to a process that permits dramatic copy number suppression of particular subgenomic DNA molecules to nearly undetectable levels during plant development. This process, first discovered in maize (Small et al., 1987), appears to be widespread in plants and may constitute a means of maintaining mitochondrial genetic variation in a silenced but retrievable form (Small et al., 1989; Janska et al., 1998). Substoichiometric shifting of specific portions of the genome appears to engage replicative and/or recombinational mechanisms that are uncharacterized to date.

An argument has been made for the endosymbiotic origin of mitochondria from a common ancestor of the rickettsial subdivision of the α-proteobacteria (Andersson et al., 1998; Andersson et al., 1999; Emelyanov, 2001). Evidence suggests that during eukaryotic evolution, the protomitochondrial form relinquished much of its genetic complement to the nucleus by a process of interorganellar gene transfer. More recently, these events apparently occurred via processed RNA intermediates that must acquire, after integration to the nuclear genome, a promoter and targeting presequence for nuclear function (reviewed by Brennicke et al., 1993; Martin and Herrmann, 1998; Adams et al., 1999; Palmer et al., 2000). The additional components likely are acquired by intergenic recombination (Kadowaki et al., 1996).

The identification of nuclear genes that encode mitochondrial proteins has allowed more detailed investigation of their origins and likely evolutionary paths. Although a relatively small number of genes have been studied in detail, several examples exist of proteins capable of dual targeting to mitochondria and chloroplast compartments (Small et al., 1998; Peeters and Small, 2001) and of interorganellar gene substitution. In the latter cases, nuclear genes of plastid origin have been found that now encode mitochondrial (Adams et al., 2002) or cytosolic (Krepinsky et al., 2001) proteins. Similarly, nuclear genes of mi-
RESULTS

Identification of Mitochondrial DNA and RNA Metabolism Loci on Chromosome III in Arabidopsis

An extensive survey of the Arabidopsis genome for genes that might be involved in mitochondrial genome maintenance functions revealed the presence of a number of genes on chromosome III that appeared to encode mitochondrial proteins based on their prokaryotic sequence homology and targeting capacity. Our comparisons with known rickettsial and yeast mitochondrial DNA and RNA metabolism proteins suggested that several genes in this interval were involved in similar functions. A list of the genes identified is provided in Table 1.

In addition, a number of genes that encode mitochondrial targeted proteins of yet unknown function were present (data not shown). Conservatively, we have identified >50 genes within a 6-Mbp interval, both annotated and unannotated for function, that appear to encode proteins with mitochondrial targeting capacity. This estimate does not include the >100 genes predicted to encode plastid proteins within the same interval and several that appear to be of rickettsial origin by homology with no clear targeting features annotated (data not shown). The association of several genes from this interval with organellar DNA and RNA maintenance suggests that at least some of the genes that lack functional annotation also might participate in related functions or their regulation.

We also surveyed the Arabidopsis genome for plant equivalents of proteins involved in eukaryotic mitochondrial DNA maintenance functions on the basis of published sequences in yeast and human (Kang and Hamasaki, 2002; MITOP database at http://mips.gsf.de/services/genomes). Nearly all of the plant equivalents for mitochondrial DNA repair are uncharacterized. Several of the identified candidates were located on chromosome III (Table 1). A complete list of the putative Arabidopsis homologs can be found in the supplemental data online (http://psweb.unl.edu/mackenzie/table1.html).

The chromosome III interval that encompasses the identified mitochondrial genes can be estimated to extend as much as 26 Mbp, although mitochondrial gene density appeared to be greatest within the 6.5- to 9.7-Mbp interval, with a second, smaller cluster at 17.0 to 22.0 Mbp on the chromosome III map. Another statistically significant cluster was found on chromosome V. To determine whether the apparent clustering of related organellar genes in Arabidopsis was a conserved tendency in higher plants, we investigated the genomic locations of the homologous sequences in rice. Although in some cases it was not possible to identify a corresponding gene candidate from the annotated rice sequences currently available, a surprisingly large proportion of identified homologous genes were located on either chromosome I (37%) or chromosome IV (23%) in the rice genome (Table 1). This result suggests that linkage of genes for mitochondrial DNA and RNA metabolism exists within the rice genome.

The Organization of Rickettsia-Homologous Genes for DNA Metabolism in the Arabidopsis Genome

Because Rickettsia americana has been postulated to represent one of the closest living relatives of present-day mitochondria (Andersson et al., 1998), we surveyed the Rickettsia genome for genes involved in DNA maintenance and RNA metabolism functions. Of 43 genes identified, 25 (58%) homologous sequences were identified in the Arabidopsis genome, as shown in the supplemental data online (http://psweb.unl.edu/mackenzie/table2a.html and http://psweb.unl.edu/mackenzie/table2b.html). Of the genes in Arabidopsis predicted to be analogous to Rickettsia DNA metabolism loci, 14 (56%) were located on chromosome III, and all but 3 encoded organelle-targeting proteins. For those Rickettsia homologs located elsewhere in the genome, seven (64%) were predicted to encode proteins with ambiguous organelle-targeting discrimination values. Remarkably, of all the Rickettsia-homologous loci predicted to encode organelle-targeting proteins, half appeared to be dual targeting or plastid targeting (see supplemental data online). Together, these observations suggest an intimate coevolution of mitochondrial and plastid DNA maintenance functions during the evolution of higher plants and a striking trend toward the clustering of organellar DNA maintenance components.

We were surprised to find that a relatively large proportion of the Rickettsia-homologous loci (19%) encoded proteins with ambiguous or undetected organelle-targeting predissequences. Some of these genes appeared to encode proteins that function within the nucleus, but others may represent organellar proteins targeted by a distinct mechanism or genes misannotated for translation start site.

Organellar DNA and RNA Metabolism Loci Appear to Be Clustered in Two Genomic Intervals in Arabidopsis

A permutation test for the consolidation of organellar DNA and RNA maintenance genes was constructed by comparing the observed clustering with the pattern of distribution that would be expected as a result of random rearrangement and duplication of Arabidopsis organellar genes. A total of 791 genes predicted to target mitochondria were identified using the Predotar program to survey the Arabidopsis genome database (http://www.inra.fr/Internet/Produits/Predotar/Arabidopsis_mit.html). From within this list, genes were classified by whether or not they had a putative DNA maintenance function. The chromosome location of each gene also was determined. From chro-
<table>
<thead>
<tr>
<th>Locus</th>
<th>Location (Mbp)</th>
<th>Annotation</th>
<th>TargetP</th>
<th>MitProt</th>
<th>Predator</th>
<th>Confirmed Localization</th>
<th>ESTs or cDNA</th>
<th>Homology/Gene</th>
<th>Rice Homology and P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>At3g10140</td>
<td>3.134</td>
<td>Putative RecA</td>
<td>0.861</td>
<td>0.9384</td>
<td>0.998</td>
<td></td>
<td></td>
<td></td>
<td>DNA metabolism loci</td>
</tr>
<tr>
<td>At3g10270</td>
<td>3.17</td>
<td>Putative DNA gyrase subunit B</td>
<td>0.801</td>
<td>0.9997</td>
<td>0.997</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At3g10990</td>
<td>3.34</td>
<td>Putative DNA gyrase subunit A</td>
<td>ct 0.975/mt 0.879</td>
<td>0.9</td>
<td>p 0.998/mt 0.967</td>
<td>mTEL, Headwood, A.Gout, J.M. Whelan, D.A. Day, and A.H. Miller, unpublished data</td>
<td></td>
<td></td>
<td>1 YGRB, Nostoc punctiforme</td>
</tr>
<tr>
<td>At3g18580</td>
<td>6.397</td>
<td>Hypothetical protein</td>
<td>0.717</td>
<td>0.9948</td>
<td>0.544</td>
<td>mt</td>
<td></td>
<td></td>
<td>Single-stranded DNA binding protein, Escherichia coli</td>
</tr>
<tr>
<td>At3g20390</td>
<td>7.11</td>
<td>Translational inhibitor, putative DNA polymerase, putative</td>
<td>p 0.497</td>
<td>0.795</td>
<td>mt 0.617</td>
<td>mt</td>
<td></td>
<td></td>
<td>MFM1p-mitochondrial DNA maintenance</td>
</tr>
<tr>
<td>At3g20540</td>
<td>7.168</td>
<td>DNA polymerase, mitochondrial</td>
<td>0.741</td>
<td>0.8028</td>
<td>0.776</td>
<td>mt/ct (this report)</td>
<td></td>
<td></td>
<td>Chromosome 8/4, 5.6e-207/7.7e-178</td>
</tr>
<tr>
<td>At3g24340</td>
<td>8.831</td>
<td>Hypothetical protein</td>
<td>0.338</td>
<td>0.2168</td>
<td>0.77</td>
<td>1</td>
<td></td>
<td></td>
<td>RAD26 DNA repair and recombination, helicase</td>
</tr>
<tr>
<td>At3g14990</td>
<td>5.008</td>
<td>Putative DNA nick sensor</td>
<td>ct 0.576/mt 0.269</td>
<td>0.8207</td>
<td>ct 0.922</td>
<td>mTEL (L. McIntosh, unpublished data)</td>
<td>AF453835</td>
<td>3 DNA phosphatase</td>
<td>Chromosome 1, 6.2e-60</td>
</tr>
<tr>
<td>At3g18630</td>
<td>6.413</td>
<td>Putative uracil-DNA glycosylase</td>
<td>ct 0.928</td>
<td>0.3255</td>
<td>mt 0.933</td>
<td>mTEL (Kang and Hamasaki, 2002)</td>
<td></td>
<td></td>
<td>Ung1, human</td>
</tr>
<tr>
<td>At3g24320</td>
<td>8.823</td>
<td>Putative mismatch repair protein</td>
<td>0.857</td>
<td>0.961</td>
<td>0.943</td>
<td>mt</td>
<td></td>
<td></td>
<td>Chromosome 4, 2.1e-70</td>
</tr>
<tr>
<td>At3g26690</td>
<td>9.804</td>
<td>Mut4-like protein</td>
<td>dUTP</td>
<td>ct 0.79</td>
<td>0.5745</td>
<td>ct 0.151</td>
<td>mt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At3g46940</td>
<td>17.60</td>
<td>Hypothetical protein</td>
<td>ct 0.305</td>
<td>None</td>
<td>ct 0.985</td>
<td>mt</td>
<td></td>
<td></td>
<td>Chromosome 3, 3.000254</td>
</tr>
<tr>
<td>At3g50880</td>
<td>19.22</td>
<td>DNA helicase homolog</td>
<td>ct 0.843</td>
<td>0.1253</td>
<td>None</td>
<td>mt</td>
<td></td>
<td></td>
<td>Yeast MAG, human MPG</td>
</tr>
<tr>
<td>At3g51690</td>
<td>19.49</td>
<td>DNA helicase homolog</td>
<td>ct 0.253</td>
<td>0.017</td>
<td>ct 0.188</td>
<td>mt</td>
<td></td>
<td></td>
<td>Chromosome 1, 1.1e-66</td>
</tr>
<tr>
<td>At3g51880</td>
<td>19.56</td>
<td>HMG2-like protein</td>
<td>mt 0.121</td>
<td>0.0851</td>
<td>ct 0.716</td>
<td>mt</td>
<td></td>
<td></td>
<td>Yeast AFB2, human TFAM</td>
</tr>
<tr>
<td>At3g01410</td>
<td>1.536</td>
<td>RNase H</td>
<td>0.651</td>
<td>0.938</td>
<td>0.962</td>
<td></td>
<td></td>
<td></td>
<td>Chromosome 8, 2.6e-32</td>
</tr>
<tr>
<td>At3g09210</td>
<td>2.825</td>
<td>Unknown protein</td>
<td>0.536</td>
<td>0.571</td>
<td>0.583</td>
<td></td>
<td></td>
<td></td>
<td>Chromosome 3, 7.3e-45</td>
</tr>
<tr>
<td>At3g18090</td>
<td>6.195</td>
<td>DNA-directed RNA polymerase II</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>mt</td>
<td></td>
<td></td>
<td>DNA-directed RNA polymerase II, second largest chain</td>
</tr>
<tr>
<td>At3g22310</td>
<td>7.887</td>
<td>Putative RNA helicase</td>
<td>0.771</td>
<td>0.9552</td>
<td>0.982</td>
<td>mTEL (L. McIntosh, unpublished data)</td>
<td></td>
<td></td>
<td>DEAD-box family helicases</td>
</tr>
<tr>
<td>At3g22330</td>
<td>7.892</td>
<td>Putative RNA helicase</td>
<td>0.635</td>
<td>0.9406</td>
<td>0.943</td>
<td>mTEL (Kang and Hamasaki, 2002)</td>
<td></td>
<td></td>
<td>Chromosome 4, 1.6e-127</td>
</tr>
<tr>
<td>At3g23780</td>
<td>8.568</td>
<td>DNA-directed RNA polymerase, putative</td>
<td>0.42</td>
<td>0.253</td>
<td>0.92</td>
<td>mt</td>
<td></td>
<td></td>
<td>DNA-directed RNA polymerase, Arabidopsis, yeast</td>
</tr>
<tr>
<td>At3g23830</td>
<td>8.607</td>
<td>Gly-rich RNA binding protein</td>
<td>0.504</td>
<td>0.8331</td>
<td>ct 0.920</td>
<td>mt</td>
<td></td>
<td></td>
<td>Chromosome 1, 1.6e-18</td>
</tr>
<tr>
<td>At3g25430</td>
<td>9.22</td>
<td>Hypothetical protein</td>
<td>0.464</td>
<td>0.9284</td>
<td>0.982</td>
<td></td>
<td></td>
<td></td>
<td>Poly(A)-specific ribonuclease</td>
</tr>
<tr>
<td>At3g05780</td>
<td>1.714</td>
<td>LON Ser-type protease</td>
<td>mt 0.587</td>
<td>0.9110</td>
<td>None</td>
<td>mTEL (L. McIntosh, unpublished data)</td>
<td></td>
<td></td>
<td>Yeast mitochondrial DNA binding affinity in yeast</td>
</tr>
<tr>
<td>At3g05790</td>
<td>1.720</td>
<td>LON Ser-type protease</td>
<td>ct 0.589</td>
<td>0.8992</td>
<td>ct 0.852</td>
<td>0</td>
<td>0</td>
<td></td>
<td>Mitochondrial DNA binding affinity in yeast</td>
</tr>
<tr>
<td>At3g07200</td>
<td>2.291</td>
<td>RING Zn finger protein, putative</td>
<td>0.84</td>
<td>0.87</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>Zinc finger, SHC4</td>
<td></td>
</tr>
<tr>
<td>At3g10010</td>
<td>3.111</td>
<td>DML2</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>mt</td>
<td></td>
<td></td>
<td>Chromosome 1, 3.2e-64</td>
</tr>
<tr>
<td>At3g10110</td>
<td>3.116</td>
<td>TIM22-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>mt</td>
<td></td>
<td></td>
<td>Chromosome 3, 3.0023</td>
</tr>
<tr>
<td>At3g10810</td>
<td>3.38</td>
<td>Ring Zn finger, putative</td>
<td>Expressed protein</td>
<td>ct 0.914</td>
<td>0.990</td>
<td>mt 0.998</td>
<td>mt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At3g15000</td>
<td>5.050</td>
<td>Cd45-like</td>
<td>0.087</td>
<td>0.131</td>
<td>0.998</td>
<td>0</td>
<td>0</td>
<td>COQ45 minichromosome, DNA replication</td>
<td></td>
</tr>
<tr>
<td>At3g20000</td>
<td>6.387</td>
<td>TOM40</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>mt</td>
<td></td>
<td></td>
<td>Chromosome 9/6, 3e-21</td>
</tr>
<tr>
<td>At3g23990</td>
<td>6.868</td>
<td>Hsp60</td>
<td>mt 0.843</td>
<td>0.9897</td>
<td>0.999</td>
<td>mt</td>
<td></td>
<td></td>
<td>Membrane import protein</td>
</tr>
<tr>
<td>At3g25100</td>
<td>9.144</td>
<td>Cd45-like</td>
<td>0.087</td>
<td>0.131</td>
<td>0.998</td>
<td>0</td>
<td>0</td>
<td>COQ45 minichromosome, DNA replication</td>
<td></td>
</tr>
<tr>
<td>At3g26480</td>
<td>9.687</td>
<td>WD-repeat protein</td>
<td>0.457</td>
<td>0.736</td>
<td>0.805</td>
<td>2</td>
<td>2</td>
<td>DIP2, CRN1, melosis</td>
<td></td>
</tr>
<tr>
<td>At3g27070</td>
<td>9.982</td>
<td>TOM 20-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>mt</td>
<td></td>
<td></td>
<td>Chromosome 12, 7.4e-58</td>
</tr>
<tr>
<td>At3g27080</td>
<td>9.984</td>
<td>TOM 20-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>mt</td>
<td></td>
<td></td>
<td>Chromosome 1, 4.4e-10</td>
</tr>
<tr>
<td>At3g27360</td>
<td>10.129</td>
<td>Histone H3</td>
<td>ct 0.765</td>
<td>0.992</td>
<td>mt 0.995</td>
<td>mt</td>
<td></td>
<td></td>
<td>Mitochondrial import receptor</td>
</tr>
<tr>
<td>At3g28810</td>
<td>21.98</td>
<td>Kefol acid reductoisomerase</td>
<td>ct 0.980</td>
<td>0.963</td>
<td>mt 0.995</td>
<td>mt</td>
<td></td>
<td></td>
<td>Chromosome 1, 1.8e-55</td>
</tr>
</tbody>
</table>

ct, chloroplast; mt, mitochondria; None, no significant targeting value; p, plastic.
aReciprocal best hit.
bLocalization confirmed by proteomics data.
A homologous protein is mitochondrial in other species.
 Localization confirmed by experimental data.
*This protein does not have a target peptide.
Patterns of Genomic Duplication Suggest Selection for the Present-Day Arrangement and Interorganellar Substitutions

As is common for many regions of the Arabidopsis genome (Blanc et al., 2000), extensive duplication has occurred for the region of interest on chromosome III. In some cases, duplication has involved large intervals, whereas in other instances, only one or a few genes were involved. The duplication pattern of the chromosome III interval is shown in Figure 1. The extent of duplication implies that the chromosome III genomic interval has undergone significant rearrangement during the evolution of the Arabidopsis genome. This extent of rearrangement and duplication likely precludes the conservation of any ancient gene clusters that might have resulted during the process of endosymbiosis. Likewise, the gene order of the linked loci on Arabidopsis chromosome III is rearranged dramatically relative to the homologous loci on rice chromosome I (Figure 1B). No evidence of recombination suppression was detected within the chromosome III interval (data not shown).

Analysis of duplicated loci indicates that functional divergence has occurred by two routes. In some cases, gene duplication may have resulted in functional divergence of the gene. This appears to be the case for a mitochondrial single-stranded DNA binding (SSB) protein gene, as shown in Figure 2. A mitochondrial SSB (mtSSB) locus present in higher plants apparently was duplicated early during plant evolution, resulting in two diverged copies in all plant species surveyed. In Arabidopsis, one copy of the gene resides within the chromosome III interval and another resides on chromosome IV. The pattern of sequence diversification after gene duplication results in two gene clusters (Figure 2B) and suggests that mtSSBs may have evolved for an additional plant-specific function. All identified plant mtSSB genes identified were predicted to encode mitochondrially targeted proteins (data not shown).

A second fate of duplications appears to involve expanded protein-targeting capacity. In several cases, gene duplications result in proteins predicted to target mitochondria, plastids, or both organelles dually (Figure 1). These data, combined with the observed targeting features of the identified Rickettsia homologs (see supplemental data online [http://psiweb.unl.edu/mackenzie/table2a.html and http://psiweb.unl.edu/mackenzie/table 2b. html]), suggest that the interchange of mitochondrial and plastid components for DNA maintenance has been extensive. We generally used three computer programs designed to predict protein targeting within the plant cell (see Methods). It is not possible to conclude definitively a protein’s targeting capacity by computer prediction programs. However, good correspondence generally is found between computer-assisted predictions and in vivo observations for those proteins for which correspondence is seen among prediction programs (I. Small, personal communication; A. Lyznik, A. Elo, and S. Mackenzie, unpublished data).

Evidence of Mitochondrial and Plastid Coevolution for DNA Metabolism Functions

Our observations suggest that the duplication of genes involved in organelar DNA and RNA maintenance may have facilitated functional specialization and organellar coevolution.

Table 2. Statistical Test of Gene Clustering for Genes Involved in Organelar DNA and RNA Maintenance Functions

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Total Number of Genes</th>
<th>Observed DNA and RNA Maintenance Genes</th>
<th>Standard Deviation (s)</th>
<th>Consolidation Score (Zc)</th>
<th>Estimated P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>219</td>
<td>13</td>
<td>17.5</td>
<td>2.9</td>
<td>1.6</td>
</tr>
<tr>
<td>II</td>
<td>180</td>
<td>9</td>
<td>14.9</td>
<td>2.8</td>
<td>2.1</td>
</tr>
<tr>
<td>III</td>
<td>148</td>
<td>26</td>
<td>12.9</td>
<td>2.7</td>
<td>4.9</td>
</tr>
<tr>
<td>IV</td>
<td>173</td>
<td>7</td>
<td>14.5</td>
<td>2.8</td>
<td>2.7</td>
</tr>
<tr>
<td>V</td>
<td>71</td>
<td>13</td>
<td>6.6</td>
<td>2.2</td>
<td>2.9</td>
</tr>
<tr>
<td>Maximum consolidation score (Zmax)</td>
<td></td>
<td></td>
<td></td>
<td>4.9</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Given the prediction of both mitochondrial and plastid targeting of gene products within the chromosome III interval, we investigated the fate of targeting presequence information in the evolution of genes that are involved in mitochondrial DNA and RNA metabolism. Other groups have shown that the transfer of organelar genes to the nucleus must be accompanied by the acquisition of mitochondrial or plastid protein-targeting information to allow the gene to become functional (reviewed by Martin and Herrmann, 1998). This targeting information apparently can be acquired via nuclear intergenic recombination.

Amino acid sequence homology searches revealed that, in several cases, some genes within the chromosome III interval share protein presequence identity with functionally unrelated organelle-targeting proteins, as shown in Table 3. Putative presequences were searched using BLAST (Basic Local Alignment Search Tool) against the Arabidopsis database. The short amino acid sequences used in these searches resulted in relatively high e-values. To identify homologous presequences, the following criteria were used. Evidence of presequence exchange involved e-values of $<1.0$ for two sequences located at the immediate N termini of proteins, demonstrating organelle-targeting capacity. Several homologous sequences with significant similarity and low e-values were excluded by their localization other than at the N terminus. For this reason, it is possible that we missed those examples of presequence exchange that involve exon shuffling.

Analysis of Arabidopsis mitochondrial proteome data (J.L. Heazlewood, A. Gout, J.M. Whelan, D.A. Day, and A.H. Millar, unpublished data; L. McIntosh, personal communication), as well as studies by others, confirmed that several of the identified genes postulated to have shared mitochondrion-targeting presequences encode mitochondrially localized proteins (Tables 1 and 3). From this observation, we assume that intergenic recombination is the predominant mechanism involved in the acquisition of organelle-targeting capacity for genes in this region. In some cases, the intergenic exchanges occurred within the chromosome III interval, as shown in Figure 3B. The Arabidopsis nuclear rps10 gene (At3g22300; Figure 3) represents an inde-
Figure 2. Sequence Comparison of Two Arabidopsis Genes Predicted to Encode SSB-Like Proteins.

(A) Alignment of SSB-like sequences. The N-terminal portions of the sequences, which encode the mitochondrion-targeting presequence, were truncated before alignment using CLUSTAL X. Identical residues are shown by reverse contrast and similar residues with shading, with dark gray indicating a block of similar and light gray weakly similar amino acid residues. In the interest of space, only a portion of the identified genes are shown.

(B) Phylogenetic tree developed from SSB-like sequences reveals two plant forms of the gene. The phylogenetic tree was constructed using the NJ method (Saitou and Nei, 1987) based on a distance matrix computed using the PROTDIST program in the PHYLIP software package version 3.6 (Felsenstein, 2002). Estimated statistical confidence was generated from 100 bootstrapped data sets using the SEQBOOT program, and the consensus tree was generated using the CONSENSE program in the PHYLIP package (Felsenstein, 2002). EST sequences were derived from BLAST searches. Two distinct ESTs were identified for nearly every plant species surveyed, providing evidence of SSB gene duplication.
pendent and recent gene transfer event from the mitochondrion during Brassicaceae evolution. The gene remains within the mitochondrial in several plant species (Adams et al., 2000). The \textit{rps10} gene apparently acquired its targeting presequence from an adjacent RNA helicase locus on chromosome III (Figure 3).

In several instances, gene duplications have resulted in mitochondrial, plastid, and dual targeting forms of a nearly identical gene identified on chromosome III (At3g20540), designated \textit{AtPol} \text{H9253}\textsubscript{1}, and its duplicate on chromosome I (At1g50840), designated \textit{AtPol} \text{H9253}\textsubscript{2}, both genes contain features similar to those of an adjacent RNA helicase locus on chromosome III (Figure 3).

Table 3. Evidence of Protein Targeting of Presequence Exchanges in Arabidopsis

<table>
<thead>
<tr>
<th>Locus</th>
<th>Putative Protein Function</th>
<th>Presequence Homology with</th>
<th>Putative Protein Function</th>
<th>Predicted Localization</th>
<th>Confirmed Localization</th>
</tr>
</thead>
<tbody>
<tr>
<td>At3g20540 DNA polymerase (mt)</td>
<td>At1g50840 DNA polymerase</td>
<td>ct</td>
<td>ct\textsuperscript{a} (this report)</td>
<td>61/64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At3g14530 GGPLS3</td>
<td>ct</td>
<td>ct\textsuperscript{b} (Okada et al., 2000)</td>
<td>34/52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At3g32040 GGPLS, putative</td>
<td>ct</td>
<td>37/59</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>At2g47270 Unknown protein</td>
<td>ct</td>
<td>30/53</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>At3g18420 Unknown protein</td>
<td>ct</td>
<td>36/53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At3g22310 RNA helicase (mt)</td>
<td>At3g22330 RNA helicase</td>
<td>mt</td>
<td>mt\textsuperscript{a} (Miliar et al., 2001)</td>
<td>66/83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At3g22300 40S ribosomal protein S10</td>
<td>mt</td>
<td>mt\textsuperscript{a} (Wischmann and Schuster, 1995)</td>
<td>51/68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At3g27620 Alternative oxidase 1c</td>
<td>mt</td>
<td>mt\textsuperscript{a} (Saisho et al., 1997)</td>
<td>50/66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At1g60950 Ferredoxin precursor</td>
<td>mt</td>
<td>mt\textsuperscript{a} (Peltier et al., 2000)</td>
<td>45/64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At1g10960 Ferredoxin precursor ct</td>
<td>72/81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>At5g38430 Rubisco small subunit 1b</td>
<td>ct</td>
<td>ct\textsuperscript{a} (Peltier et al., 2000)</td>
<td>56/67</td>
<td></td>
</tr>
<tr>
<td>At3g23830 RNA binding protein (mt)</td>
<td>At4g13850 AtGRP2-like</td>
<td>mt</td>
<td>mt\textsuperscript{b} (Kuff et al., 2001)</td>
<td>72/81</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At5g47320 40S ribosomal protein S19</td>
<td>mt</td>
<td>mt\textsuperscript{a} (Sanchez et al., 1996)</td>
<td>66/72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At5g49210 Unknown protein</td>
<td>mt</td>
<td>58/67</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>At5g61030 RNA binding-like</td>
<td>mt</td>
<td>36/53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g66660 DNA polymerase (ct)</td>
<td>At14a-like</td>
<td>ct</td>
<td>35/48</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{a}Localization confirmed by experimental data.
\textsuperscript{b}Localization confirmed by proteomics data.

dotar value of 0.96, TargetP value of 0.93). We investigated the in vivo targeting properties of both genes by developing gene constructs that fused the predicted targeting presequence information from either locus (763 bp for chromosome III or 752 bp for chromosome I) to the enhanced \textit{GREEN FLUORESCENT PROTEIN} (GFP) reporter gene. The gene constructs were used for particle bombardment of Arabidopsis leaf tissues.

Confocal analysis suggested that the product from the chromosome I \textit{AtPol} \text{H9253}\textsubscript{1} gene was targeted to the plastid, whereas the chromosome III \textit{AtPol} \text{H9253}\textsubscript{2} gene product was dual targeted to the mitochondria and the plastid (Figure 4B). Dual targeting of at least one gene product would be consistent with biochemical data for similar activities of mitochondrial and plastid DNA polymerases reported by Sakai (2001) and Kimura et al. (2002). Whether the two enzyme forms can reciprocally compensate for function is being investigated.

DISCUSSION

Our investigations revealed a nuclear genomic interval in Arabidopsis that contains a large number of genes involved in DNA and RNA metabolism functions. Several of these genes have never been characterized in higher plants. A large number of unidentified genes predicted to encode mitochondrial proteins also were found in this interval. For example, two C3HC4 zinc RING finger proteins were identified with compelling discrimination values predicting mitochondrial targeting (At3g07200 and At3g10810; Table 1). This RING finger motif has been associated functionally with E3 ubiquitin protein ligase activity in a significant number of proteins (Joaizeiro et al., 1999; Freemont, 2000). Interestingly, a process involving ubiquitination has been
described to influence mitochondrial inheritance in yeast (Fisk and Yaffe, 1999) and the maternal transmission of mitochondria in mammals (Sutovsky et al., 1999, 2000). An intriguing possibility is that the identified Arabidopsis proteins might represent components of the maternal mitochondrial inheritance apparatus of higher plants.

We suspect that more intensive mining of this genomic interval will provide valuable insights into several similarly underdeveloped aspects of higher plant organelle biology. Some of the loci listed in Table 1 do not appear, by functional annotation, to be involved in aspects of genome maintenance functions. However, it has become increasingly clear that nucleus-encoded mitochondrial proteins can serve dual functions (Bateman et al., 2002), and mutation of such proteins can result in the destabilization of the mitochondrial genome in yeast (Suzuki et al., 1994). These observations in yeast and other species were considered in our analysis.

Until recently, the genetic clustering of genes for related functions was considered unusual in higher plant genomes. With the exception of disease resistance gene clusters (Richter and Ronald, 2000; Hulbert et al., 2001) and other gene families arising by duplication, genes associated with the same pathway or developmental process more often appeared to be unlinked. In fact, we were unable to detect any apparent genomic association pattern among most nuclear genes involved in mitochondrial processes for electron transport, energy transduction, biosynthetic pathways, and protein translation in Arabidopsis.

However, more extensive whole-genome quantitative trait loci analysis in maize has suggested that particular functional gene clusters may exist (Khavkin and Coe, 1997). Several loci involved in the regulation of maize development have been shown to link to quantitative trait loci associated with related developmental processes. Likewise, clustering of classes of housekeeping

---

**Figure 3.** Evidence of Targeting of Presequence Exchange during the Evolution of Mitochondrial Genes in the Chromosome III Organellar Gene-Rich Interval in Arabidopsis.

**A** Amino acid sequence alignment of the Arabidopsis proteins sharing common N-terminal amino acid sequence with the DNA polymerase-like protein (At3g20540), the RNA helicase-like protein (At3g22310), or the RNA binding protein (At3g28830). Identical residues are shown by reverse contrast and similar residues with shading, with dark gray indicating a block of similar and light gray weakly similar amino acid residues.

**B** The predicted mitochondrial presequence from the RNA helicase (At3g22310; underlined) was used in a BLAST search to identify homologous sequences in the database. The three most significant hits were obtained for mitochondrial presequences of three neighboring loci within the same chromosomal interval, with amino acid identities ranging from 50 to 66%.
Figure 4. Organellar Targeting of a Putative γ-Like DNA Polymerase.

(A) Amino acid sequence alignment reveals conserved motifs within exonuclease and polymerase domains of the Pol1-like DNA polymerase in both chromosome I and III versions of the Arabidopsis genes. The alignment presented here incorporates data taken directly from Lewis et al. (1996) for comparison.

(B) Arabidopsis leaf cell bombardment was used to test for organellar targeting of the putative DNA polymerase I-like gene products encoded on chromosome III (AtPolγ1; At3g20540) and chromosome I (AtPolγ2; At1g50840). Plastids are shown red as a result of autofluorescence. Enhanced GFP was used as the reporter gene. Although the product from AtPolγ2 appeared to target plastids, the product from AtPolγ1 targeted both plastids and mitochondria (the white arrow indicates mitochondria, and the green arrows indicate plastids).

Genes is evident in the human genome (Lercher et al., 2002). In Arabidopsis, terpenoid synthase genes are clustered and associated with genes that encode consecutive steps in terpenoid biosynthesis (Aubourg et al., 2002). In the case of genes associated with mitochondrial DNA maintenance, genetic linkage is observed not only in the higher plant genome but apparently also within the genome of Drosophila melanogaster (Lefai et al., 2000). Five D. melanogaster loci are found in close association, and their clustering is assumed to be a vestige of their cotranslocation from the mitochondrion to the nuclear genome.
What accounts for the presence of a large number of mitochondrial genes of related function within a single plant genomic interval? The observed linkage could result from the preferential integration of genes transferred from the mitochondrion to this region of the nuclear genome, either as gene blocks or as individual genes. Perhaps subsequent gene evolution, including the acquisition of regulatory and target sequence information, was facilitated by the presence of like genes in physical proximity, as appears to be the case for rps10. Most mitochondrial genes successfully transferred and expressed in the nucleus in recent evolutionary time appear to have been transferred individually via RNA intermediates (Martin and Herrmann, 1998). However, it is likely that many of the genes involved in organellar genome maintenance were transferred before the advent of organelle RNA editing; therefore, they may have integrated as multigene segments. The presence on Arabidopsis chromosome II of a 620-kb segment representing the entire mitochondrial genome with additional duplications (Stupar et al., 2001) demonstrates that large multigene DNA transfers are feasible (reviewed by Henze and Martin, 2001).

Several observations in our study suggest that regardless of whether the linkage association on chromosome III is a consequence of the original integration pattern, the present-day gene linkages likely were maintained by selection. Organellar genes found on chromosome III do not appear to have been integrated to the nuclear genome simultaneously. The DNA and RNA metabolism genes identified on chromosome III appeared to be derived from both cyanobacterial and rickettsial origins. No evidence was found for conserved order between the Rickettsia or Reclinomonas genomes and the genes found in the Arabidopsis interval (data not shown). This is not surprising, given the degree of genome reduction that has occurred within Rickettsia (Lang et al., 1999; Gray et al., 2001). Gene order also differed dramatically between genes linked on Arabidopsis chromosome III and their homologs on rice chromosome I. These details, and the remarkable degree of genome rearrangement that has occurred during the evolution of the Arabidopsis genome (Blanc et al., 2000), make it difficult to envision the conservation of gene linkages established early in the endosymbiotic process without their selection.

Although the selective advantage for retaining a significant number of organellar DNA and RNA metabolism loci within a single linkage group is not clear, the observed genome synteny for this interval between two diverse plant species is intriguing. We have likely underestimated the degree of Arabidopsis chromosome III/rice chromosome I synteny, because the assembly of the rice genomic data is still in progress. Comparisons between rice and Arabidopsis by other groups have revealed similar colinearity between the two genomes (Liu et al. 2001).

Unlike the many other functions required by plant mitochondria, genome maintenance and transmission is one in which maternal and paternal nuclear contributions likely differ. In most plant species, mitochondria are strictly maternally inherited, and exceptions to this rule are rare (Mogensen, 1996). It is possible that gene imprinting might influence the mitochondrial genome replication and transmission process within the gametes during fertilization to ensure nuclear-mitochondrial compatibility. Coordinated gene regulation via epigenetic mechanisms or trans-acting regulators is thought to be facilitated by physical linkage of the coregulated genes (Boutanaev et al., 2002; Oliver et al., 2002; Spellman and Rubin, 2002). At least one gene that has been shown to be epigenetically regulated by hypomethylation, the SUPERMAN locus (Jacobsen and Meyerowitz, 1997), is known to reside within the chromosome III interval. As one test of this hypothesis, we are currently investigating whether the expression of organellar DNA maintenance loci on chromosome III is predominantly maternal during ovule and/or embryo development.

The role of nuclear gene duplication in gene function is of particular interest to studies of plant organellar biology, because plant mitochondrial genomes display many novel functions and genomic features. One means by which new traits originate is the evolution of novel gene functions in sequences that arise by duplication (Mitchell-Olde and Clauss, 2002). The case of mtSSB, with observed divergence of two different forms in higher plants, is intriguing because it allows for the possibility that one form of the SSB might now play a novel role. This hypothesis is consistent with what has been observed for certain fungi (Tomáška et al., 2001). In Candida species in which the mitochondrial genome has linearized, a divergent mtSSB now appears to function as a telomere binding protein (Nosek et al., 1995, 1998). With the apparent linear arrangement of the plant mitochondrial genome, that could be the situation in plants as well.

The chromosome III interval encompassed both cyanobacteria-related and Rickettsia-related sequences, with gene products predicted to target mitochondria, plastids, or both. Such intimate coevolution of mitochondria and plastids also is evident for organellar RNA polymerases (Hedtke et al., 2000), aminoacyl tRNA synthases (Small et al., 1998; Peeters et al., 2000), and ribosomal proteins (Adams et al., 2002). As with these previously reported cases, several original components of DNA and RNA metabolism reported here apparently have been supplanted by interorganellar protein substitutions. Indeed, it appears from our analysis that these two subcellular organelles may share a surprisingly large proportion of their DNA and RNA metabolism apparatus. From these observations, we conclude that beyond the acquisition of important energy metabolic capacity, the endosymbiotic process has provided important elements of genomic plasticity to higher plant genomes.

METHODS

Bioinformatics

The search for N termini as well as whole amino acid sequence similarities was performed using BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990, 1997) through the GenBank and MatDB (http://mips.gsf.de/proj/thal/db/index.html) (Schoof et al., 2002) databases. Rice-homologous sequence information was obtained by BLASTX analysis of the TIGR Rice Genome Database. All sequence alignments were performed using CLUSTAL X (Thompson et al., 1997).

The transit peptide predictions were performed using Predotar version 0.5 (http://www.inra.fr/Internet/Produits/Predotar/), TargetP (Emanuelsson et al., 2000; http://www.cbs.dtu.dk/services/TargetP/), and MitoProt.
Particle Bombardment for Presequence Targeting Analysis

The targeting presequence derived from the putative DNA polymerase I homolog on chromosome III (At3g20540) was amplified by PCR using the primers 17401MF (5'-GGGATATGGCTTTACCTACT-3') and 18142MR (5'-GGCCATGGGCTTTACCTACT-3'). The targeting presequence from the putative DNA polymerase I homolog on chromosome I (At1g50840) was amplified by PCR using the primers 28431CF (5'-TCTCCGATCTGTCCGTCCT-3') and 29162CR (5'-GGCCATGGGCTTTACCTACT-3'). PCR products then were ligated to the TOPO cloning vector (Invitrogen, Carlsbad, CA) and digested using the resulting Ncol site to release the insert. Fragments then were ligated to the pCAMBIA 1302 vector (http://www.cambia.org) and the resulting Ncol site to release the insert. Fragments then were ligated to the TOPO cloning vector (Invitrogen, Carlsbad, CA) and digested using the resulting Ncol site to release the insert. Fragments then were ligated to the pCAMBIA 1302 vector (http://www.cambia.org) within the Ncol site that resides at the start codon for the gene for enhanced GFP.

In separate experiments, 7 μg of DNA from the respective constructs was delivered into 4-week-old leaves of Arabidopsis thaliana (Columbia) using tungsten particles and the Bio logic PDS-1000/He system (Bio-Rad). Particles were bombarded into Arabidopsis leaves using 900-p.s.i. rupture discs under a vacuum of 26 inches of Hg. After the bombardment, Arabidopsis leaves were allowed to recover for 18 to 22 h on Murashige and Skoog (1962) medium plates at 22°C in 16 h of daylight. Localization of GFP expression was conducted by confocal laser scanning microscopy with Bio-Rad 1024 MRC-ES with excitation at 488 nm (red/chlorophyll).

Upon request, all novel materials described in this article will be made available in a timely manner for noncommercial research purposes.

ACKNOWLEDGMENTS

We thank Natalya Nersesian and the Center for Biotechnology Core Microscopy Facility for assistance in tissue imaging. We also thank Michael Gray for valuable discussions and Alan Christensen for critical review of the manuscript. This work was supported by grants to S.A.M. from the National Science Foundation (MCB-0090813) and the Department of Energy (DE-FG03-00ER15075).

Received December 16, 2002; accepted May 13, 2003.

REFERENCES


Gene Clustering in Arabidopsis


