Available online at www.sciencedirect.com

ScienceDirect

Genomics 88 (2006) 372–380

GENOMICS

www.elsevier.com/locate/ygeno

Construction, database integration, and application of an *Oenothera* EST library

Jaroslav Mráček [a,1], Stephan Greiner [a,1], Won Kyong Cho [a,1], Uwe Rauwolf [a,1], Martha Braun [a], Pavan Umate [a], Johannes Altstätter [a], Rhea Stoppel [a], Lada Mlčochová [a], Martina V. Silber [a], Stefanie M. Volz [a], Sarah White [a], Renate Selmeier [a], Stephen Rudd [b], Reinhold G. Herrmann [a], Jörg Meurer [a,*]

[a] *Department Biologie I, Botanik, Ludwig-Maximilians-Universität München, Menzinger Strasse 67, 80638 München, Germany*
[b] *Centre for Biotechnology, Tykistökatu 6, FIN-20521 Turku, Finland*

Received 17 November 2005; accepted 30 May 2006
Available online 10 July 2006

Abstract

Coevolution of cellular genetic compartments is a fundamental aspect in eukaryotic genome evolution that becomes apparent in serious developmental disturbances after interspecific organelle exchanges. The genus *Oenothera* represents a unique, at present the only available, resource to study the role of the compartmentalized plant genome in diversification of populations and speciation processes. An integrated approach involving cDNA cloning, EST sequencing, and bioinformatic data mining was chosen using *Oenothera elata* with the genetic constitution nuclear genome AA with plastome type I. The Gene Ontology system grouped 1621 unique gene products into 17 different functional categories. Application of arrays generated from a selected fraction of ESTs revealed significantly differing expression profiles among closely related *Oenothera* species possessing the potential to generate fertile and incompatible plastid/nuclear hybrids (hybrid bleaching). Furthermore, the EST library provides a valuable source of PCR-based polymorphic molecular markers that are instrumental for genotyping and molecular mapping approaches.

Keywords: Eukaryotic genome evolution; Genome/plastome incompatibility; Interspecific genome/plastome hybrids and cybrids; EST database; *Oenothera*; Expression profiling; Codominant molecular marker

The genetic compartments of the plant cell, nucleus/cytosol, plastid, and mitochondrion, are descendents of an endosymbiotic conglomerate of three cells, a host cell, a cyanobacterium (oxygen-producing), and an α-proteobacterium (oxygen-consuming), respectively. The endocytobioses of these once free-living cells generating eukaryotism occurred more than 1.6 billion years ago. This led to an integrated compartmentalized genetic system with a common metabolism and a common inheritance [1,2] in which both energy-transducing organelles preserved remnants of their archetype eubacterial genomes [3]. The processes that shaped eukaryotic genomes rest on an enormous capacity of restructuring and streamlining the potentials of the three genetic compartments [1]. They included intracellular transfer, loss and gain of genetic information, and the development of a sophisticated integration of spatially separated gene expression systems into a common regulatory network. Indicative of this compartmental coevolution are disorders of plant development and function when organelles even with an identical gene content are exchanged between closely related species. Recent studies of such interspecific genome/plastome hybrids or cybrids between the Solanacean species *Atropa belladonna* and *Nicotiana tabacum* have revealed that an impairment in recognizing species-specific editing sites could be involved in the manifestation of genome/

Abbreviations: EST, expressed sequence tag; CAPS, cleavable amplified polymorphic sequence; ORF, open reading frame; HPT2, Hashed Position Tree2.

* Corresponding author. Fax: +49 89 1782274.
E-mail address: joerg.meurer@lrz.uni-muenchen.de (J. Meurer).

[1] These authors contributed equally to this work.

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.05.010

plastome incompatibility [4], but compartmental coevolution presumably occurs at other levels as well.

Like no other experimental plant model currently en vogue, the genus *Oenothera* (evening primrose) has been genetically studied for more than a century and provides a unique resource for investigating functional as well as phylogenetic aspects of genome compartmentation [5,6]. It possesses wide interspecific fertility in sexual crosses, a unique and intriguing genome structure (partial or terminal complex—or translocation heterozygosity), and transmits plastids biparentally. This allows the exchange of chloroplasts and/or chromosome pairs between species by simple genetic crosses. Collectively, these features combined with substantial taxonomic, genetic, and cytogenetic records make *Oenothera* an appealing model to probe the basic aspects of the ontogenetic and phylogenetic integration of organelles, in particular the plastid, into the eukaryotic cell in general and into the intricate biology of this genus in particular [5,7–9].

In the subsection *Oenothera* (=*EuOenothera*), three types of basic genomes occurring in homozygous (AA, BB, CC) or heterozygous (AB, AC, BC) constellations and five basic plastomes (I–V) are found in nature. In nature, genomes are found only in distinct and compatible combinations with plastomes. Loss of or reduced homologous recombination and free segregation of chromosomes in complex heterozygous combinations (reciprocal translocations of all chromosome arms) allows the generation of interspecific genome/plastome hybrids or cybrids, which are generally viable and, even if they are genetically incompatible, also fertile [6,10,11]. This incompatibility is indicative of a specific interplay between the genetic compartments of the cell [12]. Nucleo-organelle incompatibility is not restricted to *Oenothera.* It is found in a wide range of microbial, plant, and animal species, implying that such compartmental coevolution is a general principle of speciation processes [13–18]. The phenomenon is, therefore, more widespread and probably more important than currently assumed.

Knowledge of nuclear genes is of intrinsic interest in understanding processes of eukaryotic genome evolution in wider perspective and will doubtlessly have a significant impact on the field of plant genomics, not least because one-fourth or more of the nuclear gene complement appears to be involved in the management of the energy-transducing organelles [1,19]. Affordable access to nuclear genes from plants from which genomic sequence information is not available is provided by expressed sequence tag (EST) databases, which select gene-rich sequences and provide substantial genome information in a relatively short time. Applying bioinformatic tools to such sequence information enables one to predict the primary amino acid sequence of the corresponding gene products, to group the postulated gene products into functional categories, to calculate redundancy of EST data sets, to estimate expression levels of individual genes [20,21], and to integrate the available classifications into databases [22]. ESTs have proven an invaluable resource for studies on gene expression profiles using array technology and on gene discovery, including those of model plants (e.g., *Arabidopsis, Chlamydomonas, Physcomitrella*), crop plants (e.g., tomato, potato, sorghum, and rice and other cereals), and forest trees (pine, poplar) [23]. In addition, ESTs are useful in the development of codominant genetic markers, including single nucleotide polymorphisms (SNPs), which represent the most abundant class of genetic variation found in eukaryotic genomes [24].

There are no reported EST collections from any plant that can provide access to aspects of cellular genome compartmentation, in particular its intriguing coevolution. Since *Oenothera* represents an obvious choice for this, we have developed the cell and molecular biology of that genus by initiating a series of molecular genetic approaches with the subsection *EuOenothera* (=*Oenothera*), the best studied of the genus, including amplified fragment length polymorphism genotyping for genome analysis; sequencing of the chromosomes of the five basic, genetically distinguishable plastid types ([6,25] and data not shown); as well as cytogenetic and phylogeographic studies. Concomitantly, we have established tissue and protoplast culture as well as transformation techniques [26,27] and developed diverse strategies to investigate the physiological and molecular causes of genome/plastome incompatibility in different combinations (data not shown).

In the present study, we report the construction and pilot application of the first, nonnormalized cDNA library, derived from *Oenothera elata* subsp. *hookeri* strain hookeri de Vries (genetic constitution AA-I). Primary goals of this long-term EST project are (1) to build an *Oenothera* database, (2) to provide a well-characterized, nonredundant EST resource for advanced genomics, (3) to generate arrays for expression studies [27] with the view to investigate the expression of genes involved in genome/plastome interaction in naturally occurring species, and (4) to develop codominant markers for genotyping and gene assignments on the seven chromosomes of *Oenothera.*

Results

EST sequencing and distribution of GC contents

A total of 3532 cDNAs including 1648 inserts from fraction 1, 1577 from fraction 2, and 307 from fraction 3 were randomly selected and single-pass sequenced from their 5′ ends. The mean sizes of cDNA fragments were 1.41 kb in fraction 1, 1.36 kb in fraction 2, and 0.90 kb in fraction 3. Detailed information of the mean sizes of cDNAs is presented in Table 1. Altogether the inserts produced high-quality sequence information for 1,774,576 nucleotides with an average read length of more than 500 nucleotides per insert after vector trimming. All clones that had been sequenced from the 3′ region contained

Table 1
The mean sizes of cDNA fragments in the *Oenothera* EST library

Size of cDNA insert	Fraction 1 (%)	Fraction 2 (%)	Fraction 3 (%)
<1 kb	9.7	10.9	36.8
1–2 kb	50.5	58.9	54.0
2–3 kb	28.0	25.9	8.1
>3 kb	11.8	4.3	1.1

poly(A) tails. Therefore, a comparison with the NCBI protein database revealed that 75% of the sequenced ESTs corresponded to full-length clones. Sequences have been assembled into groups on the basis of their GC content. Most of the ESTs contain a GC content ranging from 45 to 50% (Supplemental Fig. 1).

Cluster analysis of the EST library

The EST collection was computationally clustered and assembled to produce a nonredundant (unigene) sequence set. The resulting unigene set contains a total of 1621 unigenes (nonredundant sequence assemblies). This includes 1133 singletons and 488 multimember EST clusters. A singleton is defined as an image clone whose EST does not coherently overlap with any other clone and that contains a minimum of 50 consecutive base pairs of nonrepetitive sequence. A multimember unigene is defined as a candidate gene cluster containing sequences from more than one clone. A total of 875,940 nonredundant nucleotides were found after clustering. The average unigene length was 559 nucleotides. Most of the 488 multimember unigenes contain fewer than 5 ESTs. Only 3.1% of the clusters contain more than 10 ESTs and 1.2% include more than 20 sequences (Supplemental Tables S1 and S2).

Annotation of EST sequences

The available ESTs have been annotated with respect to predicted functions of their products on the basis of the MIPS Functional Catalogue (http://www.mips.gsf.de/projects/funcat). An *E* value of 1e−10 was chosen to define a significant database match for all analyses. As a result, 848 (~50%) unigenes of the 1621 nonredundant ESTs showed sequence similarity to genes registered in the public database (Table 2). ESTs were grouped into 17 functional categories as outlined in Fig. 1. Of the nonredundant ESTs by sequence homology 29.9% correspond to the category unclassified proteins (including classification “not yet clear-cut”). From genes with known function, the most underlying EST sequences were in the functional categories metabolism (13.6%), cellular organization (11.2%), protein synthesis (7.5%), energy metabolism (7.5%), cell rescue/aging (6.2%), communication/signal transduction (4.5%), transcription (3.6%), and transport (3.6%) (Fig. 1). The representation of functional classes within Fig. 1 is typical of a nonnormalized plant cDNA collection. Nuclear-encoded chloroplast and mitochondrial proteins are represented by 21.1 and 10.6% of the ESTs, respectively.

Of 1621 unigene sequences there are 1281 *Oenothera* sequences that match the *Arabidopsis* protein set and 197 of these sequences are found only in the *Arabidopsis* data set and not elsewhere within plant genome or EST databases. One hundred ninety unigene sequences of significant length, and with a probability of encoding a polypeptide, are unique to this data set and thus may represent putative *Oenothera*-specific genes.

Table 2
Similarity search of all ESTs obtained against public databases

Similarity	Number of unigenes	Number of EST sequences
Genes of known function	632	1823
Unclassified proteins [a]	387	642
No similarity [a]	773	1133
Total [b]	1621	3532

[a] Some genes are present in both classes, causing redundancy.
[b] In the sum the redundancy has been eliminated.

Comparison of the expression profiles of nuclear genes for chloroplast proteins in three different Oenothera species

Differences due to the coevolution of plastome and genome could result in variation in expression profiles even when closely related species are compared. They, therefore, represent the current branchpoints of speciation processes. To address the question of coregulated clusters and to what extent transcriptomes differ between naturally occurring species residing in different habitats, macroarrays were equipped with EST-derived probes of 187 nuclear genes that contribute to known and unknown chloroplast functions. The expression profiles of leaves from three different *Oenothera* species, *Oe. elata* subsp. *hookeri* strain johansen (AA-I), *Oe. grandiflora* strain tuscaloosa (BB-III), and *Oe. argillicola* strain douthat 1 (CC-V), kept under the same physiological conditions, were investigated. For precise quantification, the spots on each filter contained 112, 28, and 7 ng of each PCR product in duplicate. A representative array is illustrated in Fig. 2. The hybridization signals were statistically normalized using standard methods (Data Range, AIDA Array Compare program, version 4.0; Raytest Isotopenmeßgeräte GmbH). Data Range is a global normalization method, normalizing all spots using the same reference value. After normalization the hybridization signals were compared to one another (BB-III versus AA-I, CC-V versus AA-I, and CC-V versus BB-III) (Supplemental Table 3). The selected nuclear genes were grouped into 10 different major functional categories of the chloroplast, including amino acid metabolism, carbohydrate metabolism, photosynthesis (light and dark reactions), protein modification and fate, protein biosynthesis, secondary metabolism, transcription, unknown proteins, and others. The proportions of genes differentially or identically expressed were determined and histograms of the corresponding categories were generated (Figs. 3A and B).

To validate the expression data obtained by the macroarray-based approach mRNA levels of three gene clusters were quantified by real-time RT-PCR. A comparison of the expression levels of three gene clusters between the three genomes confirmed the differential expression of the genes and highlighted gene clusters that were more highly and less expressed in the respective species (Supplemental Fig. 2). However, the extent of the ratios, irrespective of whether higher or lower than 1, differed somehow depending on the gene cluster and the methods used. In summary, the data obtained revealed distinct expression signatures in the different *Oenothera* wild-type species, demonstrating the applicability

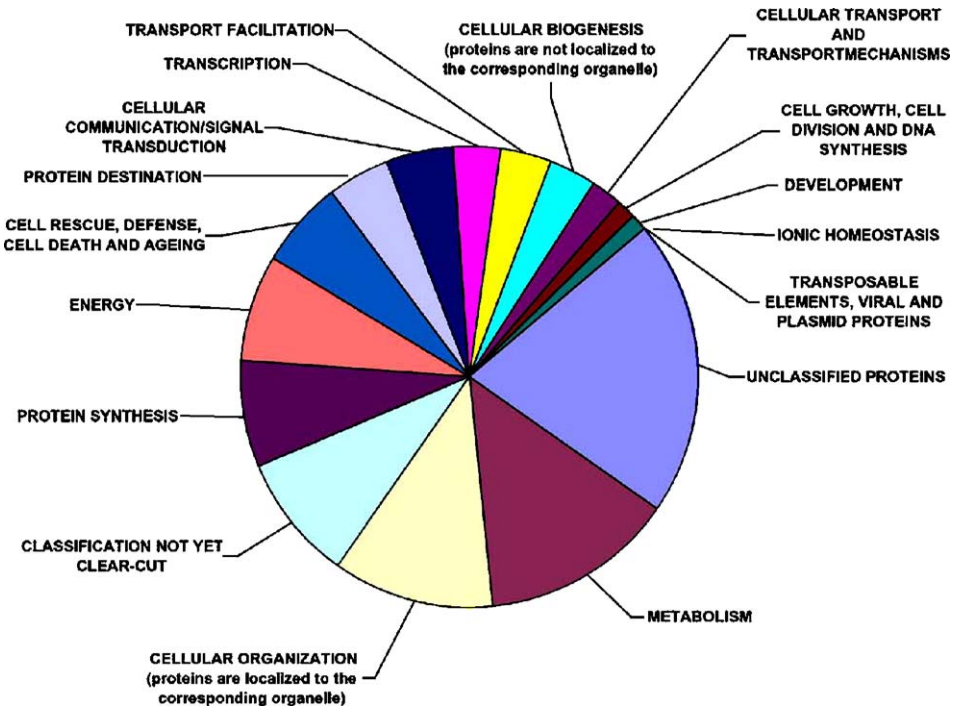
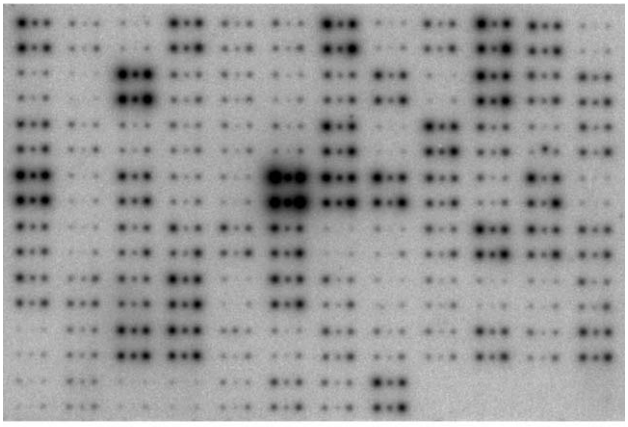


Fig. 1. Distribution of 17 functional categories of genes represented by the *Oenothera* EST library.

of the array-based technology in judging stationary RNA levels and in analyzing the interspecific variance of nuclear gene expression. Clusters of nuclear genes that are potential targets for species-specific coregulation could be highlighted. While the stationary RNA levels of a subset of genes from all categories were unchanged, mutual comparison of all hybridization signals revealed that about 50% of total genes were differentially expressed. Remarkably, most differences could be found in the category photosynthesis, light as well as dark reaction, ranging from 61 to 84% differentially expressed genes. However, less than 22% of genes for the secondary metabolism were differentially expressed. The data obtained show that transcript levels of differentially expressed genes in AA-I were generally somewhat higher than those in BB-III or CC-V. In addition, the general expression level of genes of almost all categories was significantly lower in CC-V compared to BB-III (Fig. 3A). When AA-I was compared with either BB-III or CC-V, maximal expression differences of genes were up to 19-fold higher and 25-fold lower (Fig. 3B). The extent of differential expression was less striking between CC-V and BB-III than between AA-I and any of the other two genomes.

Generation of EST-derived codominant molecular markers from Oenothera

The EST library was also utilized to validate its usefulness for mapping strategies. 5′-end sequences of selected cDNAs from the AA-I-EST library were used to generate primers and to amplify PCR products from genomic DNA of the two species *Oe. elata* subsp. *hookeri* strain johansen (AA-I) and *Oe. grandiflora* strain tuscaloosa (BB-III) (Supplemental Table 4). Forty-nine selected primer pairs generated from the hookeri de Vries library successfully amplified 39 products from AA-I (strain johansen) and BB-III, demonstrating the close relationship between the two species. The PCR products were sequenced and compared between the A and the B genotypes. The mean length of the products was 277 bp and 10 products contained an intron. Among them 7 product pairs showed size



112	7	28
28	7	112

Fig. 2. Representative autoradiogram of an array filter representing a subset of nuclear genes for chloroplast function. Order and quantities (ng/spot) of immobilized probes are given below. Signals were generated by hybridizing labeled cDNAs from the AA genome.

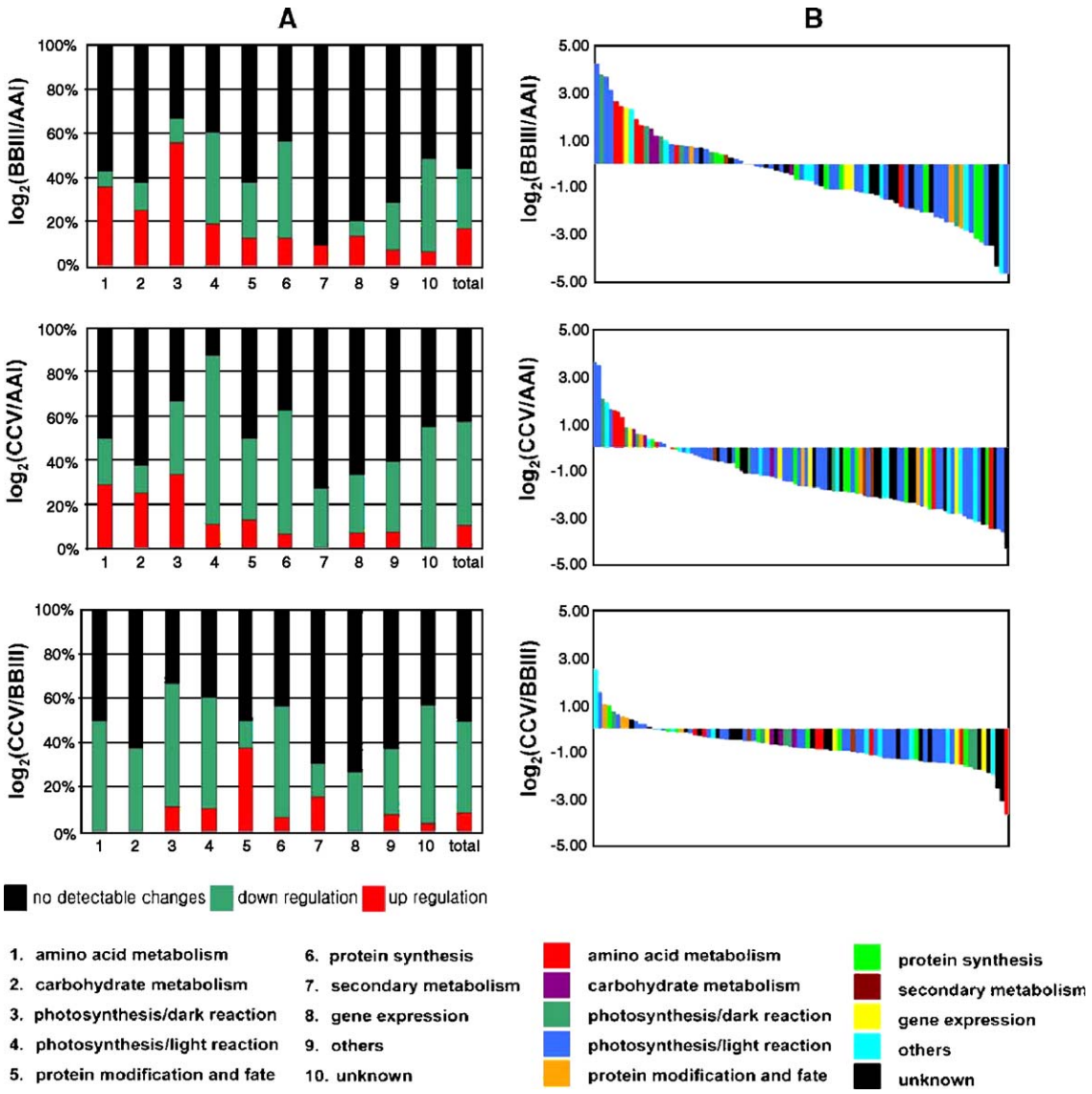


Fig. 3. Comparison of the transcript expression profiling of 187 nuclear genes for chloroplast function in the three naturally occurring genome/plastome combinations AA-I, BB-III, and CC-V. All expression ratios are converted to $\log_2$ for simplicity. (A) The histograms of 10 major functional categories show the proportions of identically expressed (black), more highly expressed (red), and less expressed (green) genes. (B) Histograms of average ratios of transcripts from plants of the three species. The ratio for each spot represents the average of six independent quantities per gene and two replicates. The colors of the bars represent the categories as indicated.

polymorphisms. Twenty-four of the investigated PCR products gave at least one cleavable amplified polymorphic sequence (CAPS) marker appropriate to distinguish both genotypes by analyzing cleaved products on agarose gels. In 27 products a minimum of one SNP could be detected (Supplemental Tables 4 and 5). Supplemental Tables 4 and 5 can be used to retrieve primer and sequence information for the genomes A and B necessary to generate molecular markers. The less conserved, sometimes intron-containing 5′ UTR and 5′ coding regions of the genes of the AA and BB nuclear genomes used in this study displayed 98.41% sequence identity. The sequences of the corresponding regions of the subspecies *Oe. elata* subsp. *hookeri* strain hookeri de Vries display 99.7% identity with the strain johansen, indicating a much closer relationship between the two A subgenomes (data not shown).

Discussion

Since traditional work on the model *Oenothera* has been restricted to classical genetic and cytogenetic approaches, we initiated the first EST program aimed at investigating the phylogenetic dynamics of genome/plastome interaction on a microevolutionary scale and its role in speciation processes. To date, phylogenetic work on the genus *Oenothera* or other representatives of the Onagracean family has been restricted to a few genes, notably to nuclear ribosomal and plastid genes, including genes encoding polypeptides of the photosynthetic machinery that are variant between the *EuOenothera* genomes or plastomes [28,29]. However, it is doubtful whether an exact genealogy of the highly complex speciation processes of this genus can be reconstructed with confidence, using data from

single genes. Understanding expression patterns of groups of genes that control basic organelle development and metabolic pathways of the plant cell with respect to speciation processes is, therefore, of general prime interest, not only for an *Oenothera* project. In this communication, we sampled young leaf tissue as a source for EST generation to obtain a mRNA spectrum relevant to chloroplast development and related processes and present the first EST database for *Oenothera,* including the coding sequences of 3532 cDNAs derived from 9-week-old *Oenothera* plantlets, of which 75% contain complete ORFs, and evaluate the appropriateness of this data set in two ways, i.e., (i) monitor whether arrays from this material can be used to study adaptive processes of genome evolution and (ii) generate molecular markers for genotyping, gene clustering, and comparative mapping studies.

Several so far unresolved technical problems had to be addressed. Application of molecular approaches is not trivial for *Oenothera* because of relatively high amounts of mucilage and polyphenols that adversely affect isolation of cellular components and enzymatic reactions. Optimized protocols described here allow consistently an efficient isolation of high-quality DNA and RNA from a wide range of tissues in various scales. Furthermore, generation of larger sets of cDNAs in an expression-ready form that could serve as an important source for functional genome analyses often suffers from size bias with a low representation of long cDNAs. This problem is quite serious since long (>3.5 kb) transcripts frequently encode biologically significant multidomain proteins. Size fractionation of the cDNAs into different fractions results usually in representative cDNA libraries with an appropriate variety [30].

By EST clustering, the aspects of redundancy, quality, and data handling can be addressed. This includes the grouping of ESTs into a set of "gene indices" on the basis of sequence similarity. Various clustering algorithms have been developed. In this article, the recently established Sputnik interphase was used for bioinformatic annotation of all ESTs [31]. The annotation of the deduced *Oenothera* protein data set led to classification of ESTs into relatively broad functional categories. Among the clones identified, cDNAs encoding various ribosomal proteins and metabolism-related genes were prevalent. Other, more prominently expressed, gene clusters were related to the "energy group," which includes functions for photosynthesis, carbon sink processes, energy transport, and membrane-associated energy conversion. Approximately 5.9% of the small subunit of the ribulose 1,5-bisphosphate carboxylase/oxygenase, 3.1% of chlorophyll *a/b* binding protein, and 1.3% of fructose bisphosphate aldolase sequence matches represent the most conspicuously expressed genes. As far as cellular organization is concerned, various genes involved in development and structural aspects could also be identified and separately clustered. It is conceivable that the coordinated expression of members of such gene categories determines specific cellular characteristics of *Oenothera* in particular, which, in turn, direct morphogenesis. The prominent group of genes from the cellular biogenesis cluster is related to cytoskeleton, cell wall, and chloroplast organization. Identification of latter ESTs in *Oenothera* is of especial interest and relevance for the comparison of gene functions during organelle development in compatible and incompatible interspecific genome/plastome combinations. Approximately one-third of the cDNAs studied correspond to sequences that share homology with hypothetical proteins with unknown functions. The roles of such genes that align with ESTs from other genome projects or display highly conserved regional sequence domains may contribute to identifying functions and classifying anonymous genes. The relatively high percentage of ESTs (30%) encoding mitochondrial and chloroplast proteins reinforces the importance of the organelles in plant metabolism and development as well as the large fraction of nuclear genes required for their management.

Clustering of nuclear genes for chloroplast functions has been shown to be useful to describe the metabolic and developmental status of the chloroplast [31–33]. Therefore, array-based gene expression and cluster analysis should provide valuable information for the nuclear–chloroplast network, if compared between naturally occurring, closely related species harboring genetically distinguishable chloroplasts and presumably in incompatible interspecific genome/plastome hybrids or cybrids. Global expression patterns of genes associated with regulating compartmental genetic interplay in a phylogenetic context have not been considered so far. The cDNA library can also be used to address the question whether genes for pathways that are specific for chloroplast-related functions are differently expressed and positionally clustered.

Array-based pilot expression profiles were generated to investigate an organelle-relevant transcriptome of the related *EuOenothera* genotypes with the genetic constitution AA-I, BB-III, and CC-V from which fertile plastid/nuclear hybrids can be generated. The obtained data demonstrate that the expression levels of various genes differ significantly between the chosen species. The expression data obtained are consistent with previous reports showing that a substantial portion of genes were similarly expressed, but up to 40% of the nuclear transcriptome showed natural variation among different *Arabidopsis thaliana* accessions [34,35]. Heterologous hybridization to potato cDNA arrays has assessed unique expression characteristics within different Solanacean species as well [36]. These results suggest that diverse gene expression patterns can be used not only to identify expression level polymorphisms [37,38] but also to identify sequence diversity among species [35]. The comparative studies show that the transcript profiles in the three species differ significantly and may serve to distinguish genes that are differentially expressed. This may be particularly relevant when regulation of plastid and nuclear mRNA levels have to be intimately coordinated to allow stoichiometric delivery of chloroplast-destined components. Therefore, these differences could potentially reflect species-specific expression patterns and genome/plastome interactions.

To compare protein levels with those of transcripts immunological analyses have been performed with all three species, AA-I, BB-III, and CC-V, using antibodies raised against various nuclear encoded photosynthetic thylakoid membrane proteins (data not shown). It appeared that there are no

significant differences in the accumulation of proteins. Although the macroarray expression data appear to show differential expression of certain genes in the three different species, there are clearly other levels of control (e.g., translational) that result in a fairly constant abundance of the protein components of photosynthesis [34].

An EST library in *Oenothera* may not only be instrumental in deciphering distinct regulatory networks in eukaryotic cells but also can serve as a useful base for the generation of codominant markers for genotyping, molecular mapping, and phylogenetic studies. Surprisingly, already minor divergence at the nucleotide level was appropriate to generate CAPS markers from almost all genes studied (~60% comparing only 277 bp on average). Molecular mapping of EST-derived genetic markers using the recently developed mapping population comprising 600 individual F2 and F3 families (data not shown) will allow the assignment of their chromosomal locations, distribution, and density along the reciprocally translocated chromosome arms in the *Oenothera* genomes. Furthermore, the generation of the first genetic map and the assignment of loci causing compartmental incompatibility in *Oenothera* by molecular marker technology, in turn, will allow the investigation of genome/plastome interactions on an evolutionary time scale and in a continental dimension [5].

Materials and methods

Plant growth conditions

The *Oenothera* species *Oe. elata* subsp. *hookeri* strains johansen and hookeri de Vries (genetic constitution AA-I) [39,40], *Oe. grandiflora* strain tuscaloosa (BB-III) [41], and *Oe. argillicola* strain douthat 1 (CC-V) [42] were grown either under standardized greenhouse conditions or in growth chambers at 24°C with 8/16-h dark/light cycles at 100 μmol photons m^{-2} s^{-1} (Osram L85W/25 Universal white fluorescent lamps). Plants were grown from seeds on 1/2 MS medium [43] containing vitamins (100 mg/L myoinositol, 10 mg/L thiamine HCl, 1 mg/L pyridoxine HCl, 1 mg/L nicotinic acid) and 3% glucose.

Isolation of total RNA and poly(A)$^+$ mRNA

Total RNA was prepared from the first two leaves of 9-week-old plants of all *Oenothera* species and subspecies using phenol extraction and LiCl precipitation [44] with a modified homogenization buffer containing 0.33 M sorbitol, 0.2 M Tris–HCl, pH 9.0, 0.3 M NaCl, 10 mM EDTA, 10 mM EGTA, and 2% (w/v) SDS. Nucleic acids were first precipitated with isopropanol and resuspended in Tris–borate buffer (80 mM Tris–borate, 10 mM EDTA, pH 8.0) before the RNA was selectively precipitated on ice in the presence of 2 M LiCl. Poly(A)$^+$ mRNA was prepared from total RNA using the Dynabeads mRNA DIRECT kit (Dynal Biotech, Oslo, Norway).

Construction of cDNA libraries

The cDNA libraries were constructed with the SuperScript Plasmid System for cDNA synthesis and plasmid cloning (Invitrogen, Carlsbad, CA, USA). Double-stranded cDNA (127 ng) was synthesized from 4 μg poly(A)$^+$ mRNA from the two youngest leaves of 9-week-old plants of *Oe. elata* subsp. *hookeri* strain hookeri de Vries (AA-I). The outlined protocol for RNA isolation was optimized efficiently to isolate high-quality RNA as evident from the high percentage of full-length cDNAs even from the high-molecular-weight mRNA fraction (see below). The cDNAs were inserted unidirectionally into the *Sal*I and *Not*I restriction sites of the vector pSPORT1. Transformation was performed with MAX Efficiency DH5α competent cells (Invitrogen). cDNAs were size fractionated by using Sephacryl S-500 HR resin column chromatography. Cloning efficiency was 12,000 colonies/ng cDNA without amplification, corresponding to 380,000 colonies/μg of poly(A)$^+$ mRNA. Fourteen thousand recombinant strains (7000 colonies of fraction 1 and 3500 colonies each of fractions 2 and 3) were randomly selected and stored at −70°C in 96-well microtiter plates in LB medium containing 20% glycerol. From 4 μg of poly(A)$^+$ mRNA, 127 ng cDNA was recovered and separated into three fractions by column chromatography. The library was used to build an *Oenothera* sequence database.

Plasmid isolation, PCR, and purification

Plasmids were isolated and the PCR assays were performed with *Taq* polymerase (Qiagen, Hilden, Germany) using the vector primers M13 forward (5′-CCCAGTCACGACGTTGTAAAACG-3′) and M13 reverse (5′-AGCGGATAACAATTTCACACAGG-3′) (MWG Biotech AG, Ebersberg, Germany). Standard conditions for PCR were 30 s at 93°C, 30 s at 57°C, and 3 min at 72°C. Cycles were repeated 30 times. PCR products were purified by precipitation with 75% ethanol in the presence of 0.6 M ammonium acetate and checked electrophoretically in 0.8% agarose gels for size and quality.

cDNA sequencing

Two different kinds of templates, PCR products and plasmid DNA, were applied for 5′-end cDNA sequencing using the T7 promoter-specific primer (5′-GCTCTAATACGACTCACTATAGGG-3′) with the ABI Prism 377 DNA sequencer (Applied Biosystems, Foster City, CA, USA). The DNA was labeled with the Thermo Sequenase DYEnamic direct sequencing kit (Amersham Biosciences, Uppsala, Sweden).

Computer analyses

Processed sequence data were entered into the Microsoft Access database (Microsoft, Redmond, WA, USA). EST clustering and assembly, peptide predictions, and sequence annotations were performed within the openSputnik sequence analysis pipeline [21]. EST sequences were clipped of any vector or polylinker sequence using the Crossmatch application (unpublished data) and the NCBI UniVec database was adapted to include the vector and polylinker sequences used during cDNA preparation. Simple repeats and regions of low complexity were masked using the RepeatBeater software [45] and the ESTs were clustered using the Hashed Position Tree2 (HPT2) algorithm (Biomax Informatics AG). HPT2 was optimized for overclustering by imposing a similarity threshold of 0.7 and 300 iterations for cluster classification. The derived cluster sequences were assembled using the repeat-unmasked EST sequences with the CAP3 algorithm using the default settings [46]. BLASTX (http://www.ncbi.nlm.nih.gov) was performed against a nonredundant protein sequence database and filtered at 1e−10 to identify stretches of *Oenothera* sequences that correspond to a presumed coding sequence (CDS). The aggregated CDSs were used to train the ESTScan application for the nuance of *Oenothera* codon usage. The subsequent *Oenothera* ESTScan model was used to derive peptide sequences from the unigene set. Unigene sequences were annotated for function using the MIPS Funcat [47] and Gene Ontology assignments [9]. Homology (BLAST) methods were used to place the sequences within the context of both the rice and the *Arabidopsis* genome scaffold and to place the sequences within the context of other plant EST collections. Peptide sequences were annotated for Interpro domains, subcellular and organellar localization, and transmembrane domain content. Sequence information is available in a relational database and presented through the World Wide Web at http://www.sputnik.btk.fi.

Generation and application of macroarrays in three Oenothera species

The three related species mentioned above with the genetic constitutions AA-I, BB-III, and CC-V, from which fertile genome/plastome hybrids and cybrids can be produced, were chosen to compare the individual expression profiles. The universal vector primers M13for and M13rev were used to amplify PCR products

of a subset of 187 selected cDNAs known or predicted to encode chloroplast proteins. The corresponding *Arabidopsis* gene designation for these probes is indicated in Supplemental Table 3. cDNAs were amplified using standard and optimized PCR conditions with *Taq* DNA polymerase (Qiagen). All PCR products were subjected to agarose gel electrophoresis to confirm the size and the amplification quality. Once checked, each individual amplicon was then adjusted to three different concentrations of 3.5, 14, and 56 ng/μl. Each dilution was spotted in duplicate onto a 7.8 × 11.9-cm positively charged nylon membrane (Hybond N^+; Amersham, UK) by 20-fold repetition to the same points using robotics equipped with a 0.4-mm 96-pin gridder (BioRobotics, UK). As a negative control the pBluescript vector (Stratagene, La Jolla, CA, USA) was also spotted onto the filters. After spotting, filters were denatured in 1.5 M NaCl, 0.5 M NaOH and neutralized in 0.5 M Tris, pH 7.2, 1 M NaCl. After drying, filters were cross-linked with 120 mJ of 302-nm UV light (UV-Stratalinker 1800). [^{32}P]dCTP-labeled cDNA probes were synthesized from 10 μg total RNA with random hexanucleotides (Roche, Mannheim, Germany) using the Super-Script III RNase H^- reverse transcriptase (Invitrogen). The labeled cDNAs were incubated for 20 min at 37°C with RNase H (Invitrogen) to remove RNA. The labeled cDNAs were purified using MicroSpin G-50 columns (Amersham, Freiburg, Germany). The arrays were prehybridized for 2 h at 60°C in phosphate buffer (0.25 mM Na_2HPO_4, 2.5 mM EDTA, 7% SDS, pH 7.2). The labeled cDNAs were hybridized to a filter overnight at 60°C. Filters were washed twice at 60°C in 2× SSC, 0.1% SDS and finally twice in 1× SSC and 0.1% SDS. Filters were exposed on imaging plates (Fuji Film). The radioactive images were obtained with the FLA-3000 phosphoimager (Fuji, Tokyo, Japan). Array images were imported into the AIDA Image Analyzer program (version 4.0; Raytest) and signals were deduced. For normalization, the mean value of the selected background within each subgrid was averaged and subtracted to calculate the intensity of all spots. The duplicate signals from three different concentrations were averaged and the obtained expression profiles were compared to calculate the ratios with the AIDA Array Compare program (version 4.0; Raytest). Histograms were generated using the Microsoft Excel 2002 program. The original macroarray data can be accessed from Supplemental Table 3.

Real-time RT-PCR analysis

The primer pairs M34for/M34rev (5′-GAGACTCCTGTCTGACGCCAG-3′ and 5′-CCATGGCGTGTTCACGGACAC-3′), M60for/M60rev, and M75for/M75rev were used for clusters C_936-9-B11 (transketolase), C_2590-26-F11 (phosphoribulokinase), and C_4066-89-H09 (chlorophyll *a/b* binding family, Elip2), respectively (for sequences of primers M60 and M75 see Supplemental Table 4). AA-I, BB-III, and CC-V cDNAs were normalized using a primer pair for actin (cluster S_2275-22-F04, *Arabidopsis* Accession No. At5g09810, primers M101for, 5′-GTGCTTCTAAGTGTGGAGCAACA-3′, and M101rev, 5′-CATCAGACCTTTCTTCTCCATACAGA-3′). Real-time RT-PCR experiments were performed essentially as described [48]. Due to the sequence dissimilarity (1.59%) among the different *Oenothera* species and the resulting mismatch of the A-genome-specific primers with the genomic sequences of B and C, some primer pairs chosen did not amplify a product in all species. This was also evident by the temperature shifts of the melting curves.

Development of codominant genetic markers

Gene-specific primers from the 5′ end were designed from ESTs encoding predicted chloroplast proteins and used to amplify products from isolated DNA of *Oe. elata* subsp. *hookeri* strains johansen and hookeri de Vries (AA-I) and *Oe. grandiflora* strain tuscaloosa (BB-III). DNA was isolated with the DNeasy Plant Mini Kit (Qiagen). Genomic sequences obtained from PCR products were compared between the AA-I and the BB-III genotypes. Appropriate restriction enzymes were chosen to detect polymorphisms and to generate CAPS markers for fingerprint analysis.

Acknowledgments

We thank Ingrid Duschanek and Elisabeth Gerick for excellent technical assistance. This research was supported by the Deutsche Forschungsgemeinschaft (SFB TR1), the Deutscher Akademischer Austauschdienst, and the Hanns-Seidel-Stiftung supported by the Bundesministerium für Bildung und Forschung.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2006.05.010.

References

[1] R.G. Herrmann, Eukaryotism, towards a new interpretation, in: H.E.A. Schenk, R.G. Herrmann, K.W. Jeon, N.E. Müller, W. Schwemmler (Eds.), Eukaryotism and Symbiosis, Springer, Heidelberg, New York, 1997, pp. 73–118.
[2] W. Martin, R.G. Herrmann, Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol. 118 (1998) 9–17.
[3] W. Martin, M.J. Russell, On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells, Philos. Trans. R. Soc. London B Biol. Sci. 29 (2003) 59–83.
[4] C. Schmitz-Linneweber, et al., Deficiency in nightshade/tobacco cybrids is caused by the failure to edit the plastid ATPase {alpha}-subunit mRNA, Plant Cell 17 (2005) 1815–1828.
[5] R.E. Cleland, *Oenothera:* Cytogenetics and Evolution, Academic Press, London, 1972.
[6] W. Stubbe, *Oenothera*—An ideal system for studying the interaction of genome and plastome, Plant Mol. Biol. Rep. 7 (1989) 245–257.
[7] D.G. Catcheside, Genetics of *Oenothera*, Nature 245 (1973) 439.
[8] W. Dietrich, W.L. Wagner, P.H. Raven, Systematics of *Oenothera* section *Oenothera* subsection *Oenothera* (Onagraceae), in: C. Anderson (Ed.), Systematic Botany Monographs, Am. Soc. Plant Taxonomists, Vol. 50, 1991.
[9] C. Harte, *Oenothera*—Contributions of a plant to biology, in: R. Frankel, M. Grossmann, M. Maliga (Eds.), Monographs on Theoretical and Applied Genetics, Springer–Verlag, Berlin, 1994.
[10] R.E. Glick, B.B. Sears, Genetically programmed chloroplast dedifferentiation as a consequence of plastome–genome incompatibility in *Oenothera*, Plant Physiol. 106 (1994) 367–373.
[11] M. Goldschmidt-Clermont, Coordination of nuclear and chloroplast gene expression in plant cells, Int. Rev. Cytol. 177 (1998) 115–180.
[12] R.G. Herrmann, R.M. Maier, C. Schmitz-Linneweber, Eukaryotic genome evolution: rearrangement and coevolution of compartmentalized genetic information, Philos. Trans. R. Soc. London B Biol. Sci. 358 (2003) 87–97.
[13] R.A.E. Tilney-Bassett, The inheritance and genetic behaviour of plastids, in: J.T.O. Kirk, R.A.E. Tilney-Bassett (Eds.), The Plastids: Their Chemistry, Structure, Growth and Inheritance, Elsevier/North-Holland Biochemical, Amsterdam, 1978, pp. 251-524.
[14] A. Perl, D. Aviv, E. Galun, Nuclear–organelle interaction in Solanum: interspecific cybridizations and their correlation with a plastome dendrogram, Mol. Gen. Genet. 228 (1991) 193–200.
[15] S. Kushnir, et al., Nucleo-cytoplasmic incompatibility in cybrid plants possessing an *Atropa* genome and a *Nicotiana* plastome, Mol. Gen. Genet. 225 (1991) 225–230.
[16] A.A. Wolters, A.C. Vergunst, F. van der Werff, M. Koornneef, Analysis of nuclear and organellar DNA of somatic hybrid calli and plants between *Lycopersicon* spp. and *Nicotiana* spp, Mol. Gen. Genet. 241 (1993) 707–718.
[17] L. Kenyon, C.T. Moraes, Expanding the functional human mitochondrial DNA database by the establishment of primate xenomitochondrial cybrids, Proc. Natl. Acad. Sci. USA 94 (1997) 9131–9135.
[18] M.K. Zubko, et al., Extensive developmental and metabolic alterations in cybrids *Nicotiana tabacum* (+ *Hyoscyamus niger*) are caused by complex nucleo-cytoplasmic incompatibility, Plant J. 25 (2001) 627–639.
[19] W. Martin, et al., Evolutionary analysis of *Arabidopsis,* cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of

cyanobacterial genes in the nucleus, Proc. Natl. Acad. Sci. USA 99 (2002) 12246–12251.
[20] A. Ruepp, et al., The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, Nucleic Acids Res. 32 (2004) 5539–5545.
[21] S. Rudd, Expressed sequence tags: alternative or complement to whole genome sequences? Trends Plant Sci. 8 (2005) 321–329.
[22] M.A. Harris, et al., The Gene Ontology (GO) database and informatics resource, Nucleic Acids Res. 1 (2004) 258–261.
[23] K. Mayer, H.W. Mewes, How can we deliver the large plant genomes? Strategies and perspectives, Curr. Opin. Plant Biol. 5 (2002) 173–177.
[24] L. Picoult-Newberg, et al., SNPs from EST databases, Genome Res. 9 (1999) 167–174.
[25] H. Hupfer, et al., Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome I of the five distinguishable *EuOenothera* plastomes, Mol. Gen. Genet. 263 (2000) 581–585.
[26] Mehra-Palta, et al., Tissue culture of wild-type, interspecific genome/plastome hybrids and plastome mutants of evening primrose (*Oenothera*): controlled morphogenesis and transformation, Plant Cell Rep. 17 (1998) 605–611.
[27] N. Kuchuk, R.G. Herrmann, H.U. Koop, Plant regeneration from leaf protoplasts of evening primrose (*Oenothera hookeri*), Plant Cell Rep. 17 (1998) 601–604.
[28] R.A. Levin, et al., Family-level relationships of Onagraceae based on chloroplast *rbcL* and *ndhL* data, Am. J. Bot. 90 (2003) 107–115.
[29] R.A. Levin, et al., Paraphyly in Tribe Onagreae: insights into phylogenetic relationships of Onagraceae based on nuclear and chloroplast sequence data, Syst. Bot. 29 (2004) 147–164.
[30] M.P. Draper, P.R. August, T. Connolly, B. Packard, K.M. Call, Efficient cloning of full-length cDNAs based on cDNA size fractionation, Genomics 79 (2002) 603–607.
[31] S. Rudd, H.W. Mewes, K.F. Mayer, Sputnik: a database platform for comparative plant genomics, Nucleic Acids Res. 31 (2003) 128–132.
[32] E. Richly, et al., Covariations in the nuclear chloroplast transcriptome reveal a regulatory master-switch, EMBO Rep. 4 (2003) 491–498.
[33] A. Biehl, E. Richly, C. Noutsos, F. Salamini, D. Leister, Analysis of 101 nuclear transcriptomes reveals 23 distinct regulons and their relationship to metabolism, chromosomal gene distribution and co-ordination of nuclear and plastid gene expression, Gene 344 (2005) 33–41.
[34] J. Lempe, et al., Diversity of flowering responses in wild *Arabidopsis thaliana* strains, PLoS Genet. 1 (2005) 109–118.
[35] D.J. Kliebenstein, et al., Genomic survey of gene expression diversity in *Arabidopsis thaliana*, Genetics 172 (2006) 1179–1189.
[36] W.A. Rensink, et al., Comparative analyses of six *Solanaceous* transcriptomes reveal a high degree of sequence conservation and species-specific transcripts, BMC Genom. 6 (2005) 124–138.
[37] J.D. Werner, et al., Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation, Proc. Natl. Acad. Sci. USA 102 (2005) 2460–2465.
[38] S.B. Carrol, Endless forms: the evolution of gene regulation and morphological diversity, Cell 101 (2000) 577–580.
[39] R.E. Cleland, Cyto-taxonomic studies on certain *Oenothera*s form Carlifornia, Proc. Am. Philos. Soc. 75 (1935) 339–429.
[40] R.E. Cleland, A.F. Blakeslee, Segmental interchange, the basis of chromosomal attachment in *Oenothera*, Cytologia 2 (1931) 175–233.
[41] E. Steiner, W. Stubbe, A contribution to the population biology of *Oenothera grandiflora* L'Her, Am. J. Bot. 71 (1984) 1293–1301.
[42] H.T. Stinson, Cytogenetics and phylogeny of *Oenothera argillicola* MacKenz, Genetics 38 (1953) 389–406.
[43] T. Murashige, F. Skoog, A revised medium for rapid growth and bio assays with tobacco tissue cultures, Physiol. Plant 15 (1962) 473–497.
[44] P.M. Lizardi, Methods for the preparation of messenger RNA, Methods Enzymol. 96 (1983) 24–38.
[45] K. Schneeberger, K. Malde, E. Coward, I. Jonassen, Masking repeats while clustering ESTs, Nucleic Acids Res. 33 (2005) 2176–2180.
[46] X. Huang, A. Madan, CAP3: a DNA sequence assembly program, Genome Res. 9 (1999) 868–877.
[47] H.W. Mewes, et al., MIPS: a database for genomes and protein sequences, Nucleic Acids Res. 30 (2002) 31–34.
[48] L. Lezhneva, J. Meurer, The nuclear factor HCF145 affects chloroplast *psaA-psaB-rps14* transcripts abundance in *Arabidopsis*, Plant J. 38 (2004) 740–753.