Shedding light on an extremophile lifestyle through transcriptomics

M. Dassanayake[1], J. S. Haas[2], H. J. Bohnert[1] and J. M. Cheeseman[1]

[1]Department of Plant Biology, University of Illinois, 505 South Goodwin Avenue, Urbana, IL 61801 USA; [2]Office of Networked Information Technologies (ONIT), School of Integrative Biology, University of Illinois, 505 South Goodwin Avenue, Urbana, IL 61801 USA

Author for correspondence:
J. M. Cheeseman
Tel: +217-333-2385
Email: j-cheese@illinois.edu

Received: *19 March 2009*
Accepted: *22 April 2009*

New Phytologist (2009) **183**: 764–775
doi: 10.1111/j.1469-8137.2009.02913.x

Key words: 454 pyrosequencing, convergent evolution, ecogenomics, extremophiles, functional genomics, mangrove, transcriptome profiling.

Summary

- The tropical intertidal ecosystem is defined by trees – mangroves – which are adapted to an extreme and extremely variable environment. The genetic basis underlying these adaptations is, however, virtually unknown. Based on advances in pyrosequencing, we present here the first transcriptome analysis for plants for which no prior genomic information was available. We selected the mangroves *Rhizophora mangle* (Rhizophoraceae) and *Heritiera littoralis* (Malvaceae) as ecologically important extremophiles employing markedly different physiological and life-history strategies for survival and dominance in this extreme environment.
- For maximal representation of conditional transcripts, mRNA was obtained from a variety of developmental stages, tissues types, and habitats. For each species, a normalized cDNA library of pooled mRNAs was analysed using GSFLX pyrosequencing.
- A total of 537 635 sequences were assembled *de novo* and annotated as > 13 000 distinct gene models for each species. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology annotations highlighted remarkable similarities in the mangrove transcriptome profiles, which differed substantially from the model plants *Arabidopsis* and *Populus*.
- Similarities in the two species suggest a unique mangrove lifestyle overarching the effects of transcriptome size, habitat, tissue type, developmental stage, and biogeographic and phylogenetic differences between them.

Introduction

The mangrove ecosystem is defined by a group of halophytes, largely trees, that dominate tropical intertidal zones and estuaries. The term itself refers both to the ecosystem and to the individual trees and tree species. The mangrove environment is an extreme one, characterized by prolonged and sometimes deep flooding, but also by prolonged periods (especially during neap tides) of drying soil, root zone anoxia, high temperatures, hurricane force wind, and high and extremely variable salt conditions in typically resource-poor environments. As organisms evolutionarily adapted to thrive in these extreme conditions, mangroves are true extremophiles (c.f. Inan *et al.*, 2004). The genetic basis for these characters is, however, virtually unknown: mangroves and other extremophiles, indeed most nonmodel plants, are very poorly represented in the plant molecular literature. Thus, they remain an untapped resource for understanding and exploiting plant adaptations to extreme environments.

Despite their common grouping as ‘mangroves’, mangrove taxa are biogeographically and taxonomically diverse. There are several interpretations for the origin of mangroves, but a consensus based on fossil evidence is that mangroves originated during the late Cretaceous near the Sea of Tethys which separated the supercontinents, Laurasia and Gondwanaland (Plaziat *et al.*, 2001; Saenger, 2002). Mangroves today are represented in at least 20 families (Duke *et al.*, 1998), and include ferns, monocots and dicots. With a multitude of structural adaptations reflecting responses to common environmental constraints, the mangrove community exemplifies one of the stronger cases for convergent evolution in the plant kingdom (Tomlinson, 1986; Ellison *et al.*, 1999). Convergent evolution, however, has not led to physiological uniformity. With respect to salt handling, for example, the physiological strategies range from salt excreters (e.g. *Avicennia*, *Aegicaras*, *Sonneratia*), to salt regulators (e.g. *Rhizophora*, *Bruguiera*, *Xylocarpus*) to hyper-excluders (e.g. *Heritiera*) (Scholander *et al.*, 1962; Flowers *et al.*, 1977; Paliyavuth *et al.*, 2004). In the two focal species of this study, this is reflected in the sodium (Na^+) contents of leaves. *Rhizophora mangle* lacks salt glands or other excretory mechanisms, and controls salt entry through the roots, but nevertheless accumulates Na to > 500 mM (tissue water basis), and maintains a Na : potassium (K) ratio of *c.* 4 :1 in full seawater (J. M. Cheeseman, unpublished; Popp, 1984 presents similar data for Australian rhizophoracean mangroves). *Heritiera littoralis*, by contrast, also lacks salt glands, but even in hypersaline conditions, the leaves contain < 50 mM Na with Na : K ratios of 0.5 : 1 or less (Popp, 1984, J. M. Cheeseman, unpublished; Paliyavuth *et al.*, 2004).

Whether the goal is to elucidate the genetic basis for the physiological differences, or to exploit the group's unique genetic resources, much greater genome-level understanding is needed. Particularly in an era of rapid global change, the need for such studies has been increasingly recognized (Reusch & Wood, 2007; Karrenberg & Widmer, 2008). Sequencing of complex genomes remains challenging, however (Pop & Salzberg, 2008), and can be problematic during assembly, even when considering current advancements of sequencing technologies. This applies especially to nonmodel species such as mangroves for which genomic information is scarce. As an alternative, sequencing transcriptomes entails less complexity during assembly while specifically identifying expressed genes. 454/Roche GSFLX pyrosequencing provides a versatile platform with long reads, exceptional accuracy, and ultra-high throughput sequencing compared with older sequencing strategies (Droege & Hill, 2008). Transcriptome analysis with pyrosequencing for model organisms (Weber *et al.*, 2007; Torres *et al.*, 2008) and plant species with extensive expressed sequences tag (EST) data, has demonstrated the suitability of this method in providing deep representation of transcripts (Cheung *et al.*, 2006; Barbazuk *et al.*, 2007; Swaminathan *et al.*, 2007).

Here, we have begun an in-depth analysis of the transcriptomes of *R. mangle* and *H. littoralis*. We chose these species based on their differing physiological strategies, their distinct biogeographic distributions (neotropical and Indo-West Pacific, respectively), and their distinct and evolutionarily distant phylogenetic positions (*R. mangle* is more closely related to *Arabidopsis* and *H. littoralis* is more closely related to *Populus*). We began with RNA collected from a wide variety of tissues and environmental conditions in nature and the glasshouse. We report transcript profiles obtained by 454/Roche GSFLX pyrosequencing and subsequent assembly and global annotation. Similarities are evident between the two mangroves, and the differences in representation of transcript gene ontology (GO) and KEGG (Kyoto Encyclopedia of Genes and Genomes) orthology categories that distinguish them from model plants and point to a unique mangrove 'lifestyle' (*sensu* Melzer *et al.*, 2008).

Materials and Methods

Sample collection

Plant material was harvested from both the glasshouse and the field. Tissue samples were immediately stored in liquid nitrogen or RNAlater (Applied Biosystems/Ambion, Austin, TX, USA) until further processing. RNAlater was used according to the manufacture's instructions with tissue to volume ratios of < 100 mg ml^{-1}.

Rhizophora mangle L. field samples were collected at Twin Cays, Belize, a peat-based mangrove archipelago 12 km from the coast of Belize, just inside the Mesoamerican barrier reef (Feller *et al.*, 2003). Tissues represented included leaves, roots, hypocotyl peels, young and mature propagules (viviparous seedlings still attached to the mother plant) and flower buds of stunted and tall individuals and P-fertilized stunted plants. The propagules for glasshouse plants were obtained from the same field site. The glasshouse samples included young leaf buds and shoot meristems, mature buds, stipules, young leaves, mature leaves, senescing leaves (early stages), young stem, fine roots, old, thickened roots, mature stem bark, and prop root tips from plants growing at salinities ranging from *c.* 2–100% of full seawater. Collectively, 68 different tissue types, growth conditions, and development stages were extracted for *R. mangle*. *Heritiera littoralis* Dryand. tissue samples were taken from young and mature leaves, roots, buds, and young stems of 3-yr-old saplings grown in the glasshouse at *c.* 25% of full seawater salinity. The seeds used originated from an estuarine population on the southwest coast of Sri Lanka.

RNA extraction

Total RNA was isolated using the Plant RNA Isolation Mini Kit (Agilent, Santa Clara, CA, USA). RNA samples were treated with recombinant DNase I (TURBO DNase; Ambion) at 1.5 units μg^{-1} of total RNA, and further processed with Norgen RNA clean-up and concentration kits (Thorold, Ontario, Canada). Equal amounts of mRNA from different tissue types were pooled for each species. Total RNA purity and degradation were checked with 0.8% agarose gels and with the use of an Agilent 2100 Bioanalyzer before proceeding. The RNA samples for each species were pooled for subsequent procedures.

cDNA synthesis and normalization

Approximately 200 µg of total RNA were used to extract mRNA using Oligotex mRNA mini Kits (Qiagen, Valencia, CA, USA). Subsequently, 0.5 µg of mRNA for each species was converted to cDNA using the SMART cDNA synthesis protocol (Clontech, Mountain View, CA, USA). Long poly(A:T) tails in cDNA synthesis have, until recently, led to low quantity and quality sequence reads with the Genome Sequencer FLX system. This limitation was successfully overcome by a combination of modified amplification reactions and primers designed at the WM Keck Center, University of Illinois, Urbana, IL, USA. The modified poly(T) primer includes other nucleotides interspersed in the poly(T): (TAGAGACCGAGGCGGC-CGACATGTTTTGTTTTTTTTTCTTTTTTTTTTVN). For cDNA synthesis, this primer was used in combinations with the 5′ rapid amplification of cDNA ends (RACE) SMARTIV primer (Clontech).

To improve coverage and sequencing of rare transcripts, cDNAs were normalized with a Trimmer Direct Kit (Evrogen, Moscow, Russia). cDNAs were denatured and allowed to self-anneal in a hybridization reaction for a period of 4–6 h. Within this period, most of the abundant transcripts are assumed to pair with their homologs while the unique/rare transcripts and their homologs remain single stranded. After hybridization, duplex/double stranded specific nuclease was added to the reaction to degrade ds-cDNAs. Polymerase chain reaction (PCR) was then used to reamplify the single-stranded transcripts and their homologs, providing the pool of normalized dsDNAs.

Library preparation and sequencing

The cDNAs were nebulized and selected for an average size of 400–500 bp. The FLX specific adapters, AdapterA (GCCT-CCCTCGCGCCATCAG) and AdapterB (GCCTTGCC-AGCCCGCTCAG), were ligated to the cDNA ends after end-polishing reactions, resulting in AdapterA–DNA fragment–AdapterB constructs. The adapter ligated DNAs were then mobilized to library preparation beads to capture the ssDNAs used for clonal amplification in emulsion PCR (emPCR). AdapterA sequences were used as the sequencing primer; AdapterB sequences were used to bind to the homolog sequences present at the surface of the emPCR beads. The emPCR was carried out using emPCR Kit II according to the Roche amplicon procedure. Biotinylated Adapter A, added during ssDNA construction, was used to facilitate capture and recovery of all DNA positive beads using streptavidin-coated magnetic beads. Amplified beads were loaded into a 70 × 75 mm PicoTiterPlate (PTP). Loading was followed by addition of packing beads and enzyme beads, and sequencing was carried out with an LR70 sequencing kit (Roche). The PTP was then placed onto the 454/Roche GSFLX (Roche Applied Science, Indianapolis, IN, USA) and bases (TACG) were sequentially flowed across the plate (100 cycles). A preliminary titration run was performed to determine the optimum reaction conditions; this was followed by a bulk sequencing run.

Contig assembly

Adapter sequences were trimmed using inhouse Perl scripts and any remaining sequences below 20 bp in length were discarded. *De novo* sequence assembly was done combining titration and bulk run sequences for each species. Contigs were assembled with at least 40 bp overlap and 90% identity. Singlets with > 75%, and contigs with > 50% homopolymer regions were discarded. We used sequences selected to be of 'high quality' (> 99.5% accuracy on single base reads) by GSFLX pyrosequencing software to be assembled into contigs. Two programs were tested for contig assembly, the Newbler assembler provided with the GSFLX sequencer, with a quality score threshold set at 40, and the Phrap assembly program (http://www.phrap.org) with quality scores greater than 20. A Phrap score of 20 (Phrap 20) corresponds to 99% accuracy for a given base in an assembled sequence.

Sequence annotation

Sequence annotation was based on a set of sequential BLAST searches (Altschul *et al.*, 1997) designed to find the most descriptive annotation possible for each sequence. The first BLAST search was performed with the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) nonredundant (nr) protein database limited to *Arabidopsis thaliana*, as the *A. thaliana* genome annotation is the most advanced and complete for any higher plant to date. Sequences that did not show a match were then searched against the nr protein database limited to all plants. Next, the sequences were searched using BLASTn, first against *A. thaliana* and then against all plants. For the contigs and singlets above 200 bp, the BLASTX and BLASTn searches were limited to results with e-values lower than 10^{-3} and 10^{-4}, respectively. The BLAST searches for singlets below 200 bp were carried with an e-value cutoff of 10^{-5}. In practice, the e-values for more than 90% of the annotated sequences were $< 10^{-10}$. A final BLASTn search against all sequences in nr was performed for sequences that did not have a match in any of the previous searches. That set was also searched with BLASTn against the NCBI EST and Environmental Samples databases. Inhouse Perl scripts (available on request) were used to parse BLAST outputs. The GO annotations were assigned based on the similarity to *A. thaliana* sequences; KEGG pathway annotations were assigned based on appropriately annotated plant reference genomes in NCBI.

Library preparation for Sanger sequencing and EST mapping

The cDNAs (100 ng) prepared for GSFLX sequencing were cloned nondirectionally into pCRII-Blunt-TOPO vector (Invitrogen, Carlsbad, CA, USA). Plasmid DNA was prepared following heat lysis according to the manufacturer's protocol. Sequencing reactions were carried out in a 1/32 BigDye reaction (Applied Biosystems). Sequencing was performed from the 5′-end of the cDNA in an ABI3730 capillary sequencer using M2 primer (5′-AAGCAGTGGTATCAACGCAG-3′) (Evrogen). Resulting EST sequences were trimmed from vector sequences and compared with GSFLX ESTs using CLUSTALX (Thompson *et al.*, 1997).

Table 1 Number of sequences in contigs and singlets for *Heritiera littoralis* and *Rhizophora mangle*

Sequences	*H. littoralis*	*R. mangle*
Total number of sequences	305 371	232 264
Number of sequences removed owing to low quality	83	1002
Total number of sequences remaining	305 288	231 262
Contigs		
Number of sequences in contigs	228 188	178 110
Number of contigs	31 714	25 535
Number of contigs removed owing to homopolymers	166	149
Number of contigs remaining	31 548	25 386
Average contig size	360	433
Average number of reads per contig	7.1	6.9
Singlets		
Total number of singlets	66 185	42 951
Number singlets removed owing to homopolymers	744	962
Number of singlets remaining	65 441	41 989
Singlets above 200 bp	10 094	11 558

'Sequences' are the raw numbers resulting from GSFLX output. 'Contigs' are longer continuous gene models resulting from Phrap assembly; 'reads per contig' indicates the number of individual but overlapping sequences included in the contig. 'Singlets' are annotated sequences not included in contigs.

Results

GSFLX sequencing

The combined titration and bulk GSFLX runs representing the two normalized cDNA libraries resulted in 232 264 and 305 371 sequence reads for *R. mangle* and *H. littoralis*, respectively (Table 1). After removing low-quality sequences and trimming adapter sequences, the average sequence read length was 208 bp. This was sufficient to circumvent problems of homopolymer generation and to enable annotations with fewer errors (Pop & Salzberg, 2008). All sequences have been deposited at the NCBI, and can be accessed in the Short Read Archive (SRA) under the project accession number SRA002286.3.

As Sanger-type sequencing remains the standard for EST sequencing, to evaluate the accuracy of our GSFLX-derived sequences, the normalized cDNA used to prepare the GSFLX library was used to construct an additional cDNA library. For each species, a random set of 10 cDNAs from this library was selected, cloned and analysed by the dye terminator sequencing method (ABI3730 sequencer), and mapped to the corresponding GSFLX sequences. The alignments between Sanger ESTs and GSFLX ESTs showed 97 ± 0.02% identities with 0.3 ± 0.67 % gaps. This indicated that pyrosequencing now provides a level of accuracy comparable to Sanger sequencing and enabled subsequent sequence annotations.

De novo contig assembly

Common approaches to analyse a GSFLX-generated transcriptome of a virtually unknown genome include mapping the ESTs to a closely-related model genome that has been sequenced, or using existing, extensive, EST databases. In the absence of either of these for mangroves, *de novo* assembly was carried out with Newbler and Phrap assembly programs. Both programs generated comparable contigs, however, for a given contig, where Newbler generated uncalled bases (Ns), Phrap, eliminated uncalled positions with a consensus base. In addition, studies evaluating the accuracy and reliability of Phrap have demonstrated that the program is superior to others in generating homogeneous contigs in nonrepetitive regions (Rieder *et al.*, 1998; Mavromatis *et al.*, 2007; Phillippy *et al.*, 2008). In practice, this means that the potential errors of mis-assemblies and chimeric contigs are minimized when Phrap is used on ESTs. Therefore, Phrap assembly was selected for our analysis (Table 1).

Following assembly, the average contig size was > 350 bp (Fig. 1), which is sufficient to assign functional annotations effectively. A total of 67 375 *R. mangle* and 96 989 *H. littoralis* sequences (contigs *plus* singlets) were used for annotations and further analysis. In *R. mangle*, the longest 10% of the contigs were greater than 830 bp, while only 0.6% were < 100 bp. In *H. littoralis*, the longest 10% were greater than 675 bp and 1.4% were < 100 bp.

Despite normalization, a few ESTs were sequenced and annotated hundreds of times. In *H. littoralis*, for example, the metallothionein 2a annotation was returned 487 times, and in *R. mangle*, a Pfkb-type carbohydrate kinase family protein was annotated 861 times. Within the top 20 most frequently returned annotations there were a number in each species that were totally absent from the data sets for the other (Table 2). Expressed sequence tags matching a mitochondrial transcription termination factor family protein (GI 18415647) and ubiquitin-protein ligase (GI 18395424), by contrast, were sequenced > 100 times in both mangroves. A particularly

Table 2 The most frequently sequenced transcripts of each mangrove species

		Number of ESTs [rank]	Number of sequences	Number of ESTs [rank]	Number of sequences
Best match GI number	Protein/gene name	*Heritiera littoralis*		*Rhizophora mangle*	
Frequently sequenced in *H. littoralis* only					
118489803	Unknown (*Populus trichocarpa* × *Populus deltoides*)	323 [3]	(14,1)	0	0
119224840	*Gossypium barbadense* chloroplast DNA	211 [7]	(27,59)	0	0
30683840	MIOX1 (myo-inisitol oxygenase); oxidoreductase	167 [17]	(5,1)	0	0
Frequently sequenced in *R. mangle* only					
14149114	Bg70 (*Bruguiera gymnorrhiza*)	0	0	756 [2]	(21,1)
76799968	Beta lactamase (synthetic construct)	0	0	242 [4]	(2,0)
18424223	LTP4 (lipid transfer protein 4); lipid binding	0	0	188 [6]	(4,0)
18416327	Phosphorylase family protein (*Arabidopsis thaliana*)	0	0	126 [9]	(5,0)
15237451	PBB2 (20S proteosome beta subunit B2); peptidase	0	0	98 [19]	(2,0)
Frequently sequenced in both species					
18395424	ATUBC2 (ubiquitin-conjugating enzyme 2); ubiquitin-protein ligase	259 [4]	(8,0)	109 [15]	(4,0)
18415647	Mitochondrial transcription termination factor family protein/MTERF family protein	193 [8]	(20,5)	114 [13]	(12,3)

Based on the sequential BLAST annotation,each protein/gene was identified with a unique GI number assigned by National Center for Biotechnology Information (NCBI). The number of expressed sequence tags (ESTs) refers to the number of sequence reads in the transcriptome. Number of sequences refers to the number in the annotated set after assembly and includes both contigs (first value in parentheses) an d singlets (second value). Numbers in square brackets denote the rank with respect to all sequences for the species, i.e. [3] was the third most frequently sequenced.

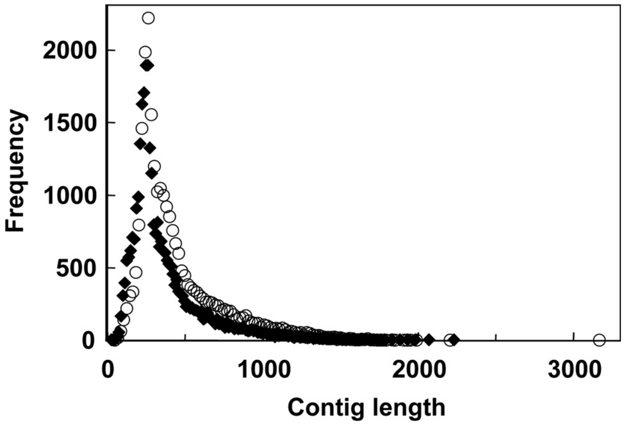

Fig. 1 Distribution of contig lengths for *Heritiera littoralis* (diamonds) and *Rhizophora mangle* (circles) following sequencing and assembly. For *H. littoralis*, the range of lengths was from 28 to 2233 bp; for *R. mangle*, it was from 47 to 3168 bp.

interesting case was Bg70, which was annotated 756 times in *R. mangle*, but absent from *H. littoralis*. This is a gene family of unknown function previously reported only from the mangrove *Bruguiera gymnorrhiza* (Rhizophoraceae) (Banzai *et al.*, 2002). In microarray and real-time PCR studies of this species, the family has been reported as highly expressed in salt-treated plants (Miyama *et al.*, 2006; Miyama & Hanagata, 2007; Liang *et al.*, 2008).

Annotation and classification of sequences into classes

Our annotation approach (see the Materials and Methods section) was based on sequence homology searches and the annotations accompanying them. It aimed to capture the most informative and complete annotation possible. Once completed, we grouped all sequences according to the extent to which they could be reconciled with sequences in public databases (i.e. based on what was 'known' about a given sequence). All sequences which could be assigned functional interpretations were categorized as known-knowns (reconciliation class R1). Overall, 23 843 *R. mangle* and 30 594 *H. littoralis* sequences were assigned to this class (approx. 33%). Second, sequences that had been reported in other species, but designated as unknown, hypothetical, unnamed, predicted or carrying a clone number without further information, were categorized as known-unknowns (class R2). The R2 class comprises 8441 *R. mangle* and 9629 *H. littoralis* sequences (approx. 12%). Finally, all sequences that did not show any similarity to other sequences in GenBank within our e-value criteria were identified as unknown-unknowns (class R3). Nearly 55% of the sequences (35 091 in *R. mangle* and 56 766 in *H. littoralis*) fell into class R3.

Of the sequences identified to classes R1 and R2 based on the NCBI nonredundant (nr) plant database, *c.* 80% were annotated based on a GenBank nucleotide or protein

Table 3 Summary of the annotation sources for mangrove sequences in groups R1 (known knowns) and R2 (known unknowns)

Database	*Heritiera littoralis*	*Rhizophora mangle*
nr_plants	86	91
Arabidopsis	76	80
Vitis	4.3	3.6
Rice	0.9	0.9
Populus	1.7	0.9
Other plants	3.4	5.5
nr_nonplants	1	1
EST	13	8

Values are per cent of total annotations. EST, expressed sequences tag.

annotation for *A. thaliana* (Table 3). Additional gene models were assigned based on the NCBI RefSeq genome database: for *R. mangle,* 13 049 distinct gene models were found for 26 928 sequences, and for *H. littoralis*, 13 598 gene models were found for 31 284 sequences. The remaining sequences with R1 and R2 annotations, 5356 for *R. mangle* and 8939 for *H. littoralis*, did not have a match with any gene models in reference genomes.

For the *R. mangle* R2, 20% of the sequences matched EST sequences from *B. gymnorrhiza*. Thirty-four per cent of the *H. littoralis* class R2 sequences shared similarity with *Gossypium* ESTs; both *Gossypium* and *Heritiera* are in the Malvaceae. In neither case did these sequences share homology with *Arabidopsis* within our annotation threshold parameters in BLAST searches. Interestingly, fewer than 1% showed similarity to sequences from nonplant sources. That the contig pool was free from substantial contamination by small RNAs (e.g. tRNAs, rRNA, plastid RNAs, and possible prokaryotic contaminating RNAs) was also verified by the annotation protocol; such sequences were represented in the overall data set at < 0.003%.

The annotated sequences are available at http://www.mangrove.uiuc.edu.

GO and KEGG analysis

The GO classification system allows descriptions of gene products in terms of their associated biological processes, cellular components and molecular functions. Currently, GO functional interpretations for plants are entirely based on *A. thaliana*. Therefore, even if mangrove sequences share functional similarity with a known plant sequence, if it is not with *Arabidopsis* it would be excluded in the GO functional assignments. In addition, there are sequences (2747 for *R. mangle* and 6587 for *H. littoralis*) that were associated with a function during annotation, but which are not assignable to any GO category (e.g. B-type cyclin (*Nicotiana tabacum*), GI 849074). Overall, 22 596 *R. mangle* and 26 034 *H. littoralis* sequences could be assigned to GO categories. More than half of those, (10 114 *R. mangle* and 11 767 *H. littoralis*) had an assignment in all three GO major categories. Sequences to which GO categories were assigned had the greatest representation in GO 'Molecular Function' (Fig. 2). There were twice as many ESTs shared between 'Molecular Function' and 'Biological Process' as between 'Cellular Component' and either of the other two classes.

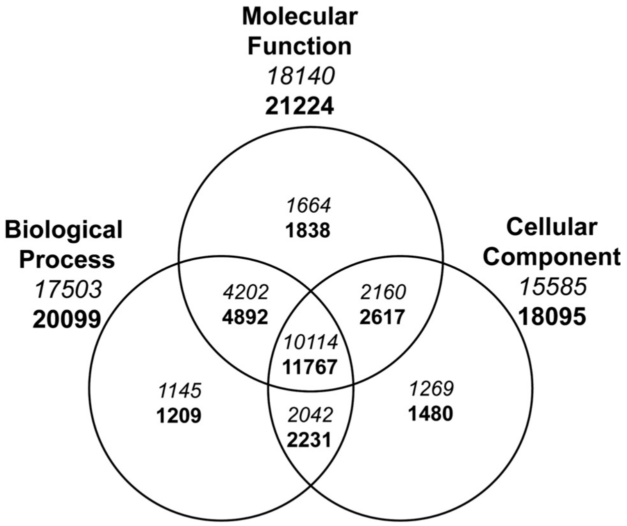

Fig. 2 Venn diagram showing the number of annotations in each gene ontology (GO) category for *Heritiera littoralis* (bold) and *Rhizophora mangle* (italics) and the overlap of their representation. The numbers outside the diagram report the total number of annotations in each GO category.

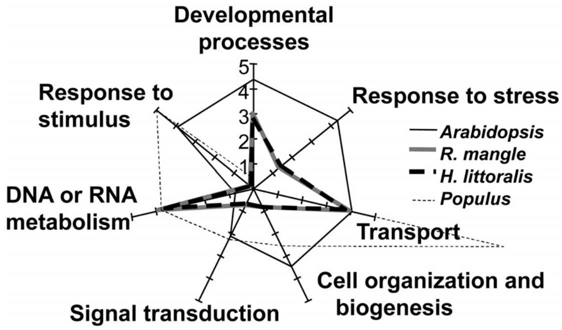

Fig. 3 Transcript profiles for gene ontology (GO) 'Biological Processes' for *Arabidopsis*, *Rhizophora mangle*, *Heritiera littoralis* and *Populus trichocarpa*, comparing relative number of gene models as percentages of the total in each GO subclass. Numbers do not add up to 100% because the process 'undefined', which comprises the greatest number of transcripts in all species, is not included.

For each GO lineage, we compared *R. mangle* and *H. littoralis* sequences with the genome-wide GO assignments for *A. thaliana* (http://www.arabidopsis.org/) and *Populus trichocarpa* (http://genome.jgi-psf.org/cgi-bin/ToGo?species=Poptr1_1). Figure 3 shows the percentage of transcripts/gene models for each species, in the major GO category 'Biological Process'. The two mangroves display almost identical profiles, but are noticeably different from the model plants. Similar results were obtained with transcript profile comparisons for GO 'Cellular Component' and 'Molecular Function' (see the Supporting Information, Fig. S1). Both the striking similarity between the mangroves and their considerable difference from the model plants were even more remarkable given the phylogenetic relationships of these species: *Populus* and *R. mangle* are grouped in the clade eurosids I, while *Arabidopsis* and *H. littoralis* are in the clade eurosids II. Given the environment in which they live, the mangroves also display a notable lack of representation in the response to stress and response to stimuli categories. In the case of *R. mangle* in particular, the extensive sampling from a wide variety of field conditions should have assured that these transcripts, if present, would be well represented.

The KEGG orthology (KO) is a classification system that provides an alternative functional annotation of genes by their associated biological pathways. The KO annotations for the mangroves were based on sequence similarity searches to reference sequence genomes (RefSeq) at NCBI. Overall, 2246 *R. mangle* and 2590 *H. littoralis* sequences were assigned to KOs, of which only 397 *R. mangle* and 468 *H. littoralis* sequences, also had all three GO classes assigned.

Figure 4 compares the mangrove KO annotations with *Populus* genome KO annotations (http://genome.jgi-psf.org/cgi-bin/metapathways?db=Poptr1_1). *P. trichocarpa* provides the second dicot genome to be completely sequenced and is annotated almost to completion. In almost all KO pathways examined, the two mangroves had similar representation in the number of distinct annotations within each subpathway, sometimes with a very different representation from *Populus*. Those differing by more than a factor of two are labeled in the figure, and many of these (marked with asterisks) are pathways related either to energy metabolism, especially in low O_2 environments (e.g. 'synthesis and degradation of ketone bodies') or pathways associated with photoprotection and reactive oxygen species (ROS) scavenging (e.g. 'terpenoid biosynthesis') or repair to components of the light processing systems (e.g. 'photosynthesis').

Discussion

Mangroves occupy a common, extremely challenging and variable habitat, but they are by no means behaviorally or physiologically homogeneous. In this study, we chose two species with distinctly different life history and physiological strategies for stress tolerance. *Rhizophora mangle* (Rhizophoraceae), the neotropical red mangrove, is considered a keystone plant species (Eddy & Faud, 1996; Proffitt & Travis, 2005) as well as an 'ecosystem engineer' (Crooks, 2002). At Twin Cays and throughout the islands of the Mesoamerican reef, the substrate is peat, derived from mangroves, to a depth of 10 m, without any mineral soil component. The islands were constructed of and by the mangroves over a period of the last 10 000 yr (Wooller *et al.*, 2004). *Rhizophora mangle* commonly dominates the landscape, often forming near-monocultures.

Physiologically, *R. mangle* is characterized as a nonsecreting, salt-including halophyte. It also shows vivipary, with the fertilized ovary growing directly into a seedling (its propagule) while remaining attached to the mother plant. One of its major defense strategies against pathogens, herbivores, UV and oxidative stresses is based on the remarkable accumulation of phenylpropanoids, particularly proanthocyanidins that constitute up to 25% of leaf dry weight (Kandil *et al.*, 2004). By contrast, *H. littoralis* (Malvaceae), the Indo-West Pacific looking-glass mangrove, is a eudicot (Chase, 2003). While it is capable of growing in full strength sea water, it is generally found at the terrestrial edge of the mangrove zone and generally isolated and scattered in the community. In contrast to *R. mangle*, *H. littoralis* displays a unique but poorly understood mechanism for extreme salt exclusion from its leaves even in hypersaline substrates.

Our goals in this study were, first, to begin to assemble the most complete representation and annotation of the mangrove transcriptome possible, and second, to use the results to extract characteristics of mangrove transcriptomes by comparing phylogenetically diverse transcript populations that might reflect their evolutionary convergence (i.e. in the sense of Melzer *et al.* (2008), a mangrove 'lifestyle'). We see these goals as an important step and contribution toward the broader goal of interweaving ecological and molecular understanding in nonmodel systems, a need which is receiving increased recognition (Reusch & Wood, 2007; Karrenberg & Widmer, 2008).

By pooling RNA samples from different developmental stages, tissues types and microhabitats, we have now increased the publicly available molecular genetic information for mangroves severalfold. Before this, the only other mangrove for which a significant number of ESTs (20 664) was publicly available was *B. gymnorrhiza*. Smaller libraries have been deposited in public databases for several other species, primarily for *Avicennia* spp. and *Sonneratia* spp.

For our first goal, we selected GSFLX pyrosequencing based on cost effectiveness, rapidly improving accuracy, and technological improvement to the point that read lengths are suitable for *de novo* assembly of sizeable contigs even in the absence of genome-based templates (Cheung *et al.*, 2006; Wicker *et al.*, 2006; Novaes *et al.*, 2008; Vera *et al.*, 2008). The success of this approach is indicated by the fact that, with what amounted to very limited pyrosequencing, we were able to generate, *de novo*, two annotated mangrove transcriptomes to a combined depth of 536 550 ESTs and 17 066 gene models (Table 1). Using estimates based on model plants (Cheung *et al.*, 2006; Weber *et al.*, 2007), we expect that these cover 50% of the transcribed, polyadenylated portion of the genome.

The transcriptome coverage was greatly increased by normalizing the pooled mRNA (Cheung *et al.*, 2006; Bogdanovaa *et al.*, 2008; Bräutigam *et al.*, 2008; Cheung *et al.*, 2008;

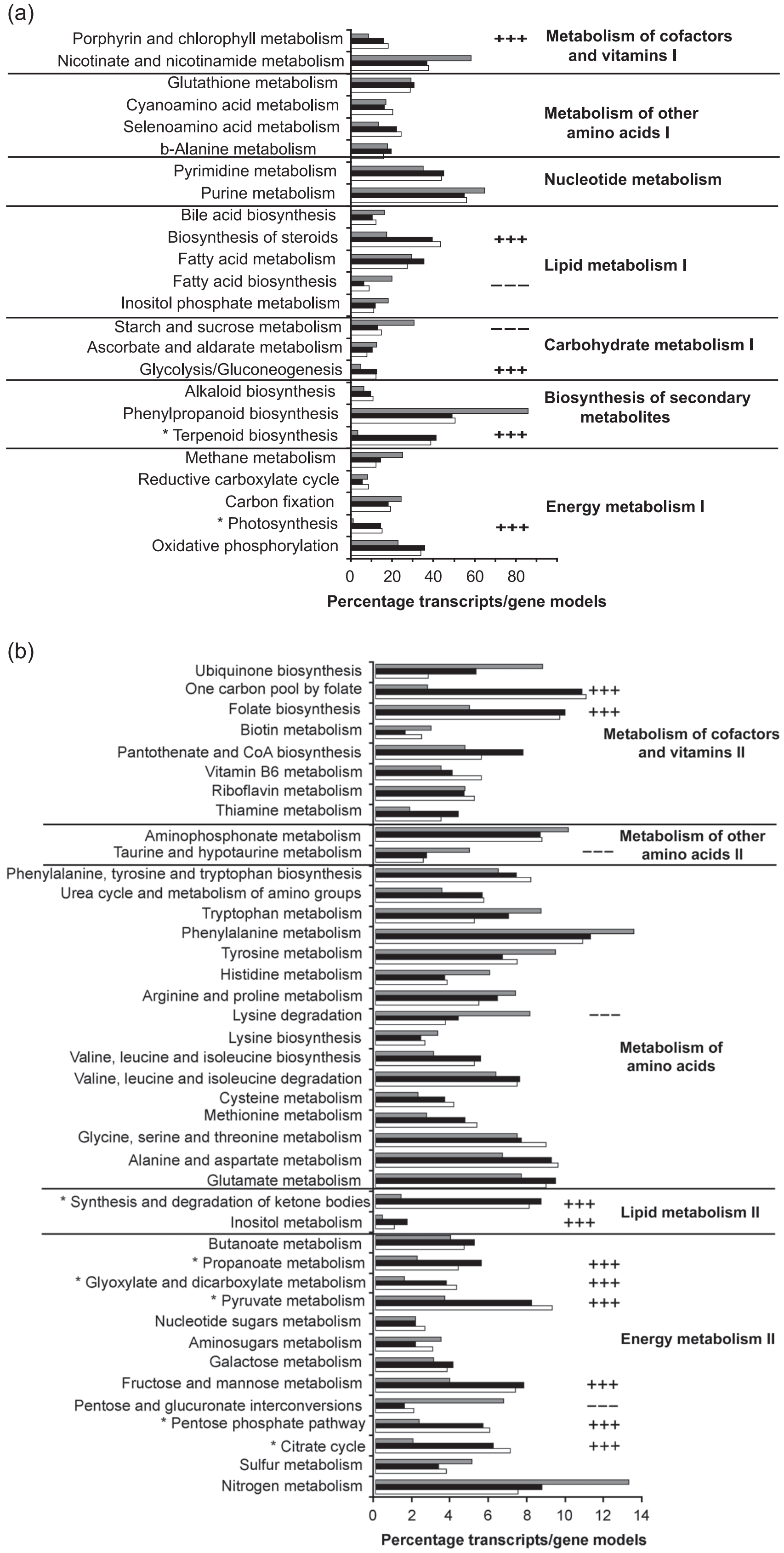

Fig. 4 Comparison of transcript numbers in selected Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways between *Populus trichocarpa* (tinted bars), *Heritiera littoralis* (closed bars) and *Rhizophora mangle* (open bars). The numbers of gene models are compared as percentages in each KEGG pathway, with values in (a) and (b) adding up to 100%. Pathways are grouped in (a) and (b) by their contributions to the total (note differences in *x*-axes). Pathways accounting for < 1% of the total gene models in any species (e.g. 'Biodegradation of xenobiotics') are not included. *Arabidopsis* is not included as genome-wide KO annotations are not available in the TAIR database. +++, Cases in which the pathway representation in mangroves is at least twice that in poplar; −−−, cases in which the pathway representation in mangroves is 50% or less of that in poplar. Pathways with (*) to the left are related either to energy metabolism, especially in low O_2 environments, or associated with photoprotection, reactive oxygen species (ROS) scavenging or repair of components of the light processing systems.

Garcia-Reyero *et al.*, 2008). Nonetheless, a few ESTs were represented an unexpectedly large number of times (Table 3). This has previously been attributed to G + C-rich regions hybridizing more slowly during the normalization process (Poroyko *et al.*, 2005). However, the most highly represented sequences in Table 3 had less than 40% G + C content. Alternatively, the 'excess' copies might represent families of paralogous genes (e.g. pfkB-type carbohydrate kinase, Table 3), or splice variants of sequences that eluded normalization.

Comparison of GSFLX and Sanger-type sequences in this study showed that sequence disparities between the two methods are negligible for the purposes of annotation, a conclusion supported by previous studies (Agaton *et al.*, 2002; Gharizadeh *et al.*, 2006). Each mangrove annotation represented, on average, 10 overlapping ESTs, which would compensate for sequencing errors if they did occur (Huse *et al.*, 2007). In fact, the GSFLX sequencing approach may have increased coverage and accuracy over what is possible by Sanger sequencing as it has previously been reported that some genes which are recalcitrant to cloning in conventional EST sequencing posed no problem for GSFLX pyrosequencing (Weber *et al.*, 2007).

Central to downstream uses of the assembled transcriptome is the annotation process. Our approach was designed to capture the most complete and informative annotation possible. This resulted in the three groupings based on their reconciliation with sequences deposited in public databases. Critical to our success here was the decision to occasionally reject a higher alignment score or a lower e-value, if it allowed us to replace uninformative descriptions, such as 'hypothetical protein', with more functional annotations. This led to 33% of the transcripts being successfully placed in class R1. Using the same approach, we have annotated the *c.* 24 000 *Bruguiera* ESTs in GenBank, successfully converting around 50% of the annotations to R1 status that had previously been identified by clone ID alone. In addition, we were able to annotate 88% of all other publicly available mangrove ESTs to class R1 or R2. Nevertheless, there remained sequences, which despite the sequential BLAST protocol produced no annotation (R3). Clearly, this is not a case limited to mangroves. A single FLX run in *Zea mays*, for example, revealed over 9000 maize-specific 'orphan genes' (maize R3) many of which had not previously been detected with conventional EST libraries (Emrich *et al.*, 2007). Given the fact that mangroves display numerous structural and physiological adaptations to an extreme environment, R3 sequences of mangroves, amounting to 55% of the transcriptome, may play a key role in understanding different physiological strategies utilized by different mangrove species.

Our emphasis on capturing all possible expressed paralogs carries with it a potential error (i.e. that a single gene would be counted multiple times when it is represented as nonoverlapping contigs). While this was a more serious problem in earlier models of the 454 GS sequencers, it has been minimized in GSFLX by improving contig length, and in our study > 12 000 contigs were longer than 500 bp.

With the specific goal of finding examples of this error, we inspected 10 individual annotation groups, for the CAT, DFR, CHS, PIP, TIP, SOS, PAL, NHX, 4CL and AHA gene families. In all cases, the potential error was contraindicated (i.e. multiple contigs overlapped and presented as unique transcripts). Figure S2 in the Supplementary Information, for example, shows a region of *Arabidopsis* catalase 2 (AtCAT2) aligned with the six *H. littoralis* homologs that share identity ranging from 72 to 90%. The *Arabidopsis* genome has three catalase genes, including four splice variants, for a total of seven gene models, while *Populus* has four genes with five gene models. In the mangrove, in the absence of a fully sequenced genome, we do not have sufficient information to distinguish between splice variants. Thus, as Fig. S2 highlights, the eight unique catalase transcripts in *H. littoralis* represent the minimum number of actual gene models.

Is there a genetic basis for a 'mangrove lifestyle'?

Our second goal was to mine the transcriptome database for indications of an essentially mangrove-specific transcript complement that might reflect their evolutionary convergence. Such convergence implies adaptive changes by which distantly related entities come to appear more related than they are (Doolittle, 1994). Convergent evolution has occurred at all levels of biological organization including morphological structures, proteins, gene families, organelle genomes and regulatory gene circuits (Conant & Wagner, 2003; Stiller *et al.*, 2003). In mangroves, phenotypic convergence appears not to be easily discerned at the gene sequence level, but must be instead recognized at higher organization levels.

Parani *et al.* (1998) approached this question earlier using random amplified polymorphic DNA (RAPD) and restriction fragment length polymorphism (RFLP) markers to consider the evolution of mangroves from terrestrial species. They examined 11 species of true mangroves, three species classified as 'minor mangroves' (defined as usually only having limited representation in the community), seven mangrove associates, mangrove parasites and terrestrial salt-tolerant species, and *Solanum esculentum* (as an outgroup). They generated a dendrogram depicting genomic level relationships between the species that, in some cases, suggested relationships far different from those based on systematic classifications. They concluded that mangroves, in evolving multiple times from different lineages, had converged to having significant genetic homogeneity underlying physiological strategies critical to survival in the intertidal.

Our results suggest that gene expression convergence has occurred with respect to the number of transcripts in all of the major categories recognized in GO classifications (i.e. molecular functions, biological processes and cellular localization; Figs 3, S1) as well as those associated with specific metabolic pathways (KO, Fig. 4). With respect to the number of transcripts linked to each GO lineage or KO biological pathway, in all cases, the two mangroves demonstrated remarkable similarities to each other, and fundamental differences to the model species (which also differed strongly from each other).

Clearly, on first discovering patterns as apparent as these, it is essential to consider potential artifacts. First, for example, the remarkable similarity of mangrove transcriptome profiles might reflect their phylogeny. However, each of the mangroves is more closely related to one of the model species than it is to the other mangrove; *Populus* and *R. mangle* are grouped in the clade eurosids I, while *Arabidopsis* and *H. littoralis* are in the clade eurosids II (Soltis *et al.*, 2000). Moreover, the physiological strategies employed by the mangroves to thrive in their extreme environment differ substantially.

Second, the similarities might reflect the similarity of the samplings used in constructing the cDNA libraries. However, more than 60 different tissue types and growth condition combinations, field and glasshouse, were included for *R. mangle* compared with eight glasshouse tissues sampled for *H. littoralis*. Based on that, we would have no reason to expect more similarity between the mangroves than between one of them and one of the models.

Third, the similarity might simply reflect an equal number of sequences being generated for the two mangroves. However, *H. littoralis* was represented by 30% more ESTs than *R. mangle* (Table 1) because of an additional partial GSFLX run during the optimization phase.

Fourth, the similarity of the results for the GO and KO analyses could reflect the dependence of one set of category assignments on the other. However, the functional assignments to GO lineages and KO categories were made independently, and only 18% of the transcripts were assigned both a GO and a KO annotation. Moreover, within the GO lineages, only half of the sequences had assignments in all three categories. Thus, the data sets represented in Figs 3, 4 and S1 were substantially different. Nevertheless, all three GO profiles and the KO profiles support the notion that there is a common pattern in mangroves which is not apparent in the model plants. In addition, although the mangrove transcriptomes are most likely < 50% explored and the collection includes a substantial number of R3 sequences, it is already clear that in certain KO categories (Fig. 4), their gene percentages are greater than in the model plants whose genomes have been sequenced. Considering the nature of the divergent transcripts, these are not simply scattered instances but rather they encode genes whose presence would be expected based on the specific needs of plants growing in extreme environments.

Finally, the contrasts could reflect the fact that model plant genomes were compared with mangrove transcriptomes. However, all of the GO and KO annotations were made by comparison with gene models. In the case of GO, in particular, this was unavoidable: *Arabidopsis* is the only plant species for which independent GO assignments have been made. In addition, the mangrove transcriptomes were sampled to represent multiple conditions as much as possible, in order to give a comparable set of gene models that imitate the gene representation of their genomes.

Therefore, we conclude that the unusual similarities observed in the two mangrove transcriptome profiles suggest a unique mangrove genomic lifestyle overarching the effects of transcriptome size, habitat, tissue type, developmental stage, and the biogeographic and phylogenetic differences that exist between the two species. This strongly favors convergent evolution playing a role at the transcriptome level in two diverse species that evolved separately to fit a common habitat.

Acknowledgements

The authors are indebted to the Smithsonian Caribbean Coral Reef Ecosystem project, and especially to Klaus Ruetzler, Candy Feller, Mike Carpenter and all the Carrie Bow Cay station managers for continued support for the field work and access to Twin Cays. The authors also thank the Vice Chancellor for Research at the University of Illinois at Urbana-Champaign for funding the sequencing itself, Indu Rupassara for *H. littoralis* seed collection, Bette Chapman for invaluable assistance in field sampling, Michelle Harris and Kaley Major for assistance in extracting more than 500 RNA samples, Suniti Karunatillake for programming assistance, and Shahjahan Ali, Jyothi Thimmapuram and Deepika Vulaganthi in the Keck Center for Comparative and Functional Genomics at UIUC for sequencing and assembly. This is contribution number 860 of the Caribbean Coral Reef Ecosystems Program (CCRE), Smithsonian Institution, supported in part by the Hunterdon Oceanographic Research Endowment.

References

Agaton C, Unneberg P, Sievertzon M, Holmberg A, Ehn M, Larsson M, Odeberg J, Uhlén M, Lundeberg J. 2002. Gene expression analysis by signature pyrosequencing. *Gene* **289**: 31–39.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Banzai T, Hershkovits G, Katcoff DJ, Hanagata N, Dubinsky Z, Karube I. 2002. Identification and characterization of mRNA transcripts differentially expressed in response to high salinity by means of differential display in the mangrove, *Bruguiera gymnorrhiza*. *Plant Science* **162**: 499–505.

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. 2007. SNP discovery via 454 transcriptome sequencing. *Plant Journal* **51**: 910–918.

Bogdanovaa EA, Shaginab DA, Lukyanov SA. 2008. Normalization of full-length enriched cDNA. *Molecular BioSystems* **4**: 205–212.

Bräutigam A, Shrestha RP, Whitten D, Wilkerson CG, Carr KM, Froehlich JE, Weber APM. 2008. Low-coverage massively parallel pyrosequencing of cDNAs enables proteomics in nonmodel species: comparison of a species-specific database generated by pyrosequencing with databases from related species for proteome analysis of pea chloroplast envelopes. *Journal of Biotechnology* **136**: 44–53.

Chase M. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society* **141**: 399–436.

Cheung F, Haas B, Goldberg S, May G, Xiao Y, Town C. 2006. Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics* 7: 272.

Cheung F, Win J, Lang JM, Hamilton J, Vuong H, Leach JE, Kamoun S, Levesque CA, Tisserat N, Buell CR. 2008. Analysis of the *Pythium ultimum* transcriptome using Sanger and pyrosequencing approaches. *BMC Genomics* **9**: 542.

Conant GC, Wagner A. 2003. Convergent evolution of gene circuits. *Nature Genetics* **34**: 264–266.

Crooks JA. 2002. Characterizing ecosystem-level consequences of biological invasions: the role of ecosystem engineers. *Oikos* **97**: 153–166.

Doolittle RF. 1994. Convergent evolution: the need to be explicit. *Trends in Biochemical Sciences* **19**: 15–18.

Droege M, Hill B. 2008. The Genome Sequencer FLX(TM) System – longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of Biotechnology* **136**: 3–10.

Duke NC, Ball MC, Ellison JC. 1998. Factors influencing biodivesity and distributional gradients in mangroves *Global Ecology and Biogeography Letters* 7: 27– 47.

Eddy J, Faud T. 1996. Global climate change impacts on habitats: assessing ecological implications of changes in climate. *Bulletin of the Ecological Society of America* 77: 109–122.

Ellison AM, Farnsworth EJ, Merkt RE. 1999. Origins of mangrove ecosystems and the mangrove biodiversity anomaly. *Global Ecology and Biogeography* **8**: 95–115.

Emrich SJ, Barbazuk WB, Li L, Schnable PS. 2007. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research* **17**: 69–73.

Feller IC, McKee KL, Whigham DF, O'Neill JP. 2003. Nitrogen vs. phosphorus limitation across an ecotonal gradient in a mangrove forest. *Biogeochemistry* **62**: 145–175.

Flowers TJ, Troke PF, Yeo AR. 1977. The mechanism of salt tolerance in halophytes. *Annual Review of Plant Physiology* **28**: 89–121.

Garcia-Reyero N, Griffitt RJ, Liu L, Kroll KJ, Farmerie WG, Barber DS, Denslow ND. 2008. Construction of a robust microarray from a nonmodel species largemouth bass, *Micropterus salmoides* (Lacepède), using pyrosequencing technology. *Journal of Fish Biology* **72**: 2354–2376.

Gharizadeh B, Herman ZS, Eason RG, Jejelowo O, Pourmand N. 2006. Large-scale pyrosequencing of synthetic DNA: a comparison with results from Sanger dideoxy sequencing. *Electrophoresis* **27**: 3042–3047.

Huse S, Huber J, Morrison H, Sogin M, Welch D. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**: R143.

Inan G, Zhang Q, Li P, Wang Z, Cao Z, Zhang H, Zhang C, Quist TM, Goodwin SM, Zhu J *et al.* 2004. Salt cress. A halophyte and cryophyte *Arabidopsis* relative model system and its applicability to molecular genetic analyses of growth and development in extremophiles. *Plant Physiology* **135**: 1718–1737.

Kandil FE, Grace MH, Seigler DS, Cheeseman JM. 2004. Polyphenolics in *Rhizophora mangle* L. leaves and their changes during leaf development and senescence. *Trees* **18**: 518–528.

Karrenberg S, Widmer A. 2008. Ecologically relevant genetic variation from a non-*Arabidopsis* perspective. *Current Opinion in Plant Biology* **11**: 156–162.

Liang S, Zhou R, Dong S, Shi S. 2008. Adaptation to salinity in mangroves: implication on the evolution of salt-tolerance. *Chinese Science Bulletin* **53**: 1708–1715.

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M *et al.* 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods* 4: 495–500.

Melzer S, Lens F, Gennen J, Vanneste S, Rohde A, Beeckman T. 2008. Flowering-time genes modulate meristem determinacy and growth form in *Arabidopsis thaliana*. *Nature Genetics* **40**: 1489–1492.

Miyama M, Hanagata N. 2007. Microarray analysis of 7029 gene expression patterns in Burma mangrove under high-salinity stress. *Plant Science* **172**: 948–957.

Miyama M, Shimizu H, Sugiyama M, Hanagata N. 2006. Sequencing and analysis of 14 842 expressed sequence tags of burma mangrove, *Bruguiera gymnorrhiza*. *Plant Science* **171**: 234–241.

Novaes E, Drost D, Farmerie W, Pappas G, Grattapaglia D, Sederoff R, Kirst M. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**: 312.

Paliyavuth C, Clough B, Patanaponpaiboon P. 2004. Salt uptake and shoot water relations in mangroves. *Aquatic Botany* **78**: 349–360.

Parani M, Lakshmi M, Senthikumar P, Ram N, Parida A. 1998. Molecular phylogeny of mangroves V. Analysis of genome relationships in mangrove species using RAPD and RFLP markers. *Theoretical and Applied Genetics* **97**: 617–625.

Phillippy A, Schatz M, Pop M. 2008. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology* **9**: R55.

Plaziat J-C, Cavagnetto C, Koeniguer J-C, Baltzer F. 2001. History and biogeography of the mangrove ecosystem, based on a critical reassessment of the paleontological record. *Wetlands Ecology and Management* **9**: 161–180.

Pop M, Salzberg SL. 2008. Bioinformatics challenges of new sequencing technology. *Trends in Genetics* **24**: 142–149.

Popp M. 1984. Chemical composition of Australian mangroves I. Inorganic ions and organic acids. *Zeitschrift für Pflanzenphysiologie* **113**: 395–409.

Poroyko V, Hejlek LG, Spollen WG, Springer GK, Nguyen HT, Sharp RE, Bohnert HJ. 2005. The maize root transcriptome by serial analysis of gene expression. *Plant Physiology* **138**: 1700–1710.

Proffitt CE, Travis SE. 2005. Albino mutation rates in red mangroves (*Rhizophora mangle* L.) as a bioassay of contamination history in Tampa Bay, Florida, USA. *Wetlands* **25**: 326–334.

Reusch T, Wood T. 2007. Molecular ecology of global change. *Molecular Ecology* **16**: 3973–3992.

Rieder MJ, Taylor SL, Tobe VO, Nickerson DA. 1998. Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Research* **26**: 967–973.

Saenger P. 2002. *Mangrove ecology, silviculture and conservation*. Berlin, Germany: Springer.

Scholander PF, Hammel HT, Hemmingsen E, Garey W. 1962. Salt balance in mangroves. *Plant Physiology* **37**: 722–729.

Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolainen V, Hahn WH, Hoop SB, Fay MF *et al.* 2000. Angiosperm phylogeny inferred from 18S rDNA, vbcL, and atpB sequences. *Botanical Journal of the Linnean Society* **133**: 381–461.

Stiller JW, DeEtte C, Jeffrey R, Johnson C. 2003. A single origin of plastids revisited: convergent evolution in organellar genome content. *Journal of Phycology* **39**: 95–105.

Swaminathan K, Varala K, Hudson M. 2007. Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genomics* **8**: 132.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**: 4876–4882.

Tomlinson PB. 1986. *The botany of mangroves*. Cambridge, UK: Cambridge University Press.

Torres T, Metta M, Ottenwälder B, Schlötterer C. 2008. Gene expression profiling by massively parallel sequencing. *Genome Research* **18**: 172–177.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* **17**: 1636–1647.

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. 2007. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiology* **144**: 32– 42.

Wicker T, Schlagenhauf E, Graner A, Close T, Keller B, Stein N. 2006. 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* 7: 275.

Wooller MJ, Behling H, Smallwood B, Fogel M. 2004. Mangrove ecosystem dynamics and elemental cycling at Twin Cays, Belize, during the Holocene. *Journal Of Quaternary Science* **19**: 1–9.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Transcript profiles for (a) Gene ontology (GO) molecular function and (b) GO cellular component for *Arabidopsis*, *Rhizophora mangle*, *Heritiera littoralis* and *Populus trichocarpa*.

Fig. S2 Alignment of *Arabidopsis* catalase 2 (AtCAT2) with the six CAT2 homologs annotated in *Heritiera littoralis*.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.