Genome-Wide Convergence during Evolution of Mangroves from Woody Plants

Shaohua Xu,[†,1] Ziwen He,[†,1] Zixiao Guo,[1] Zhang Zhang,[1] Gerald J. Wyckoff,[2] Anthony Greenberg,[3] Chung-I. Wu,[*,1,4] and Suhua Shi[*,1]

[1]State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Resources, Key Laboratory of Biodiversity Dynamics and Conservation of Guangdong Higher Education Institutes, School of Life Sciences, Sun Yat-Sen University, Guangzhou, Guangdong, China

[2]Molecular Biology and Biochemistry, University of Missouri-Kansas City, Kansas City, MO

[3]Bayesic Research, Ithaca, NY

[4]Department of Ecology and Evolution, University of Chicago, Chicago, IL

[†]These authors contributed equally to this work.

***Corresponding authors:** E-mails: lssssh@mail.sysu.edu.cn; ciwu@uchicago.edu.

Associate editor: Jianzhi Zhang

Abstract

When living organisms independently invade a new environment, the evolution of similar phenotypic traits is often observed. An interesting but contentious issue is whether the underlying molecular biology also converges in the new habitat. Independent invasions of tropical intertidal zones by woody plants, collectively referred to as mangrove trees, represent some dramatic examples. The high salinity, hypoxia, and other stressors in the new habitat might have affected both genomic features and protein structures. Here, we developed a new method for detecting convergence at conservative Sites (CCS) and applied it to the genomic sequences of mangroves. In simulations, the CCS method drastically reduces random convergence at rapidly evolving sites as well as falsely inferred convergence caused by the misinferences of the ancestral character. In mangrove genomes, we estimated ~400 genes that have experienced convergence over the background level of convergence in the nonmangrove relatives. The convergent genes are enriched in pathways related to stress response and embryo development, which could be important for mangroves' adaptation to the new habitat.

Key words: **mangroves, convergent evolution, adaptive evolution, marginal environments, genomes.**

Article

Introduction

Evolutionary convergence at the phenotypic level occurs when independent lineages adapt to the same environment. Phenotypic convergence is common across many taxa (Losos 2011). In recent years, efforts have been made to seek signals of adaptive convergence at the molecular level. In previous studies, candidate genes were first identified based on genetic, physiological and molecular evidence. Given a small subset of genes, sequence analyses were carried out to detect signals of convergence. Examples include *Prestin* in echolocating bats and toothed whales (Li et al. 2010; Liu et al. 2010), the *RNASE* gene in leaf-eating monkeys (Zhang 2006), cardenolide resistance genes in insects (Dobler et al. 2012; Zhen et al. 2012), etc.

With more and more genomes sequenced, detections of convergent evolution from genomic data have been attempted without prior identifications of candidate genes. The results, however, are often controversial (Parker et al. 2013; Foote et al. 2015; Thomas and Hahn 2015; Zou and Zhang 2015a). A main reason for the uncertainty is that molecular convergence is a noisy process. A modest level of noise, undetectable at the genic level, is amplified at the genomic scale and can easily overwhelm true signals.

The noise may come from a variety of sources. First, random changes may occur in parallel without being driven by the same selective pressure, especially at rapidly evolving sites (Zhang and Kumar 1997; Rokas and Carroll 2008; Goldstein et al. 2015; Zou and Zhang 2015b). Second, inference of the ancestral character is not always accurate. When this happens, the sharing of the ancestral character could be misinterpreted as convergence. In other words, conservation would be misconstrued as convergence. This problem has drawn relatively little attention in the literature (see *Random and false convergence across all sites—Simulation of noise*). Third, technical difficulties such as obtaining accurate nucleotide substitution rates and patterns (Zou and Zhang 2015b) could confound the inference of convergence. Removing such noise is essential for studies that rely on genomic sequences.

Mangroves, woody plants that have invaded intertidal zones independently (Ricklefs and Latham 1993; He et al. unpublished data), are good examples of convergent evolution. At the interface between terrestrial and marine environments, habitats are characterized by high salinity, hypoxia, daily fluctuating tides, strong UV light, high temperature, high sedimentation, and muddy anaerobic soils (Giri et al. 2011). To colonize the new habitats, mangroves have evolved a series of highly specialized characters such as salt tolerance,

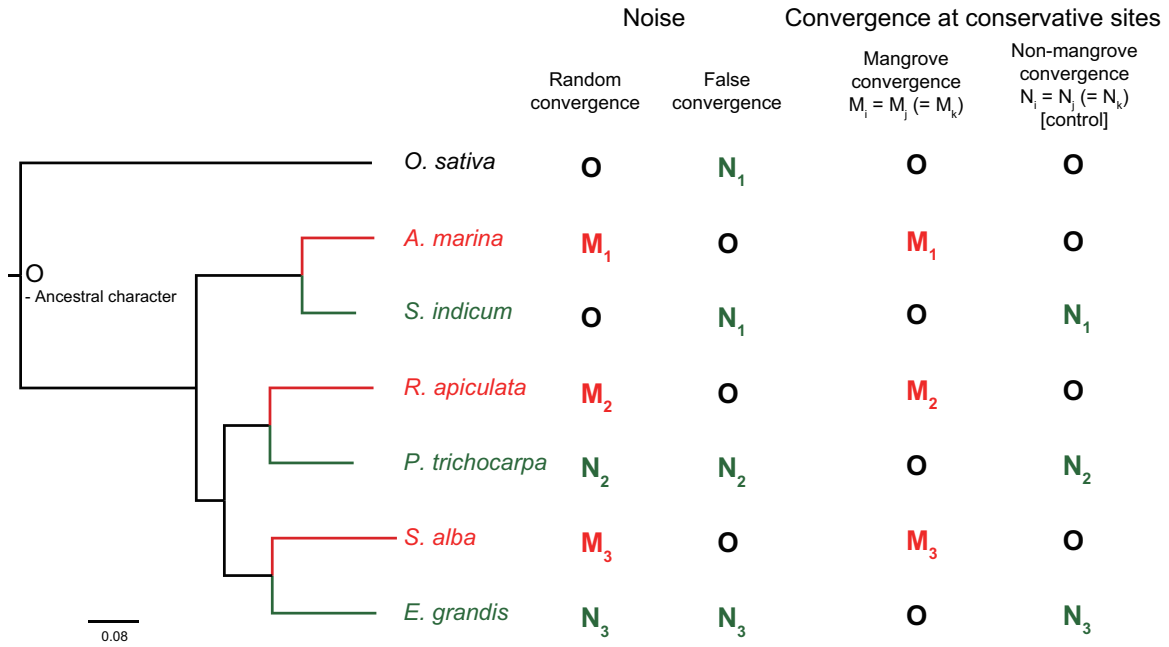


Fig. 1. Convergence noise versus signal at conservative sites. The phylogeny consists of three mangrove—nonmangrove pairs as ingroup species, along with an outgroup species. O indicates the character state of outgroup. M_i and N_i indicate the observed character state in mangrove (red) and nonmangrove (green) species, respectively. First and second column (two examples of noise): In random convergence (first column), two or three M's (or N's) could be the same character by chance, due to the many changes at the same site. The same could occur in nonmangroves. In false convergence (second column), the ancestral state is misinferred to be N_1 due to the parallel substitutions from O to N_1. Hence, all mangroves are misinferred to converge on O from N_1. Third and fourth column: In the CCS method, convergence is inferred only at conservative sites where all three nonmangrove species or all three mangroves shared the same character as the outgroup; i.e., $N_1 = N_2 = N_3 = O$ or $M_1 = M_2 = M_3 = O$. Convergence is inferred when at least two mangroves or two nonmangrove species share the same derived character.

viviparous embryo, and aerial roots (Tomlinson 1986; Parida and Jha 2010). Since few studies have focused on the genetic mechanisms underlying mangroves' adaptation (Shi et al. 2005), we have recently started such characterizations by sequencing several genomes (He et al. unpublished data). Some broad scale convergence was observed that includes genome downsizing, gene family reduction, and changes in amino acid composition (He et al. unpublished data; Lyu et al. unpublished data). These studies suggest that convergence at the molecular level could be discernible.

Results

We wish to identify genes underlying adaptive convergence in the three mangrove species depicted in figure 1 (*Avicennia marina*, *Rhizophora apiculata*, and *Sonneratia alba*), whose whole-genomic sequences have recently been de novo assembled (He et al. unpublished data). For each of the three mangrove species, we choose its closest inland relative with whole genomic sequencing data as the ingroup control (*Sesamum indicum* [Wang et al. 2014], *Populus trichocarpa* [Tuskan et al. 2006] and *Eucalyptus grandis* [Myburg et al. 2014], respectively). Since *A. marina* belongs to asterids, whereas the other two mangroves belong to rosids, we have to choose a basal eudicot or a monocotyledon as outgroup. Considering its high-quality annotation, divergence, and wide usage, we chose *Oryza sativa* (Ouyang et al. 2007) as an appropriate outgroup. This symmetric design with three pairs of mangrove—nonmangrove species provides a means for evaluating the background level of convergence. Across the seven species, 5,353 high confidence coding sequences, with a total of 2.01 million amino acids, were used for convergence detection (see Materials and Methods). The phylogeny and branch lengths were estimated using the PAML package (Yang 2007).

The Results section has two parts. Part I describes computer simulations that assess the level of noise inherent in data sets with similar properties as the one used here. A new method aiming at reducing the noise level to a minimum is then proposed. In Part II, the new method is applied to mangrove data.

I. Computer Simulations

In this section, we first quantify the various sources of noise associated with inference of genic convergence based solely on genomic data. The noise level has not been properly quantified but is in fact substantial. For that reason, one cannot positively conclude convergent evolution for any given gene. Nevertheless, when proper controls are implemented, we can assign a probability (e.g., 50%) that a set of genes have been independently evolving towards similar functional states. Functional tests can then be done on the set of candidates enriched for converging genes.

Random and False Convergence across all Sites—Simulation of Noise

Computer simulations are based on the phylogeny of figure 1 where O designates the character observed in the outgroup

Table 1. Noise Estimated by Simulations—Number of sites of random/false convergence.

Sites Used	Random Convergence	False Convergence	Inference of the Ancestral State Correct: Incorrect (% accuracy)
All sites	18,766	4,645	1,833,311:176,688 (91.2%)
Conservative sites	2,163	17	976,352:1,916 (99.8%)

species. When O is shared with some in-group species, it is usually inferred to be the ancestral character. Two major sources of noise are portrayed—random convergence and falsely inferred convergence (false convergence, for short). First, even in the entire absence of adaptive evolution, there would still be random convergence, especially at rapidly evolving sites. This is portrayed in the first column of figure 1 where $M_i = M_j$ or $N_i = N_j$ may happen by chance, rather than by convergent adaptation. This is a problem akin to the so-called "long-branch" attraction (Bergsten 2005). Second, in false convergence, the ancestral state is misinferred and the retention of the actual ancestral state would be misconstrued as convergence. For example, in the second column of figure 1, the outgroup evolves in parallel with an ingroup species, N_1. As a result, the three mangroves would show false convergence from N_1 to O.

We use computer simulation and Bayesian ancestral inference to estimate both random and false convergence, based on the phylogeny and the genome sequences of species shown in figure 1 (for detail, see Materials and Methods and supplementary fig. S1, Supplementary Material online). Briefly, the *evolver* program in the PAML package (Yang 2007) is used to generate random changes along the tree, in accordance with the rate in our data (see Materials and Methods). The simulation would track the sequence evolution on the phylogeny, including present and ancestral nodes. Then the convergence could be determined if at least two mangrove or nonmangrove species evolved to the same state (supplementary fig. S1, Supplementary Material online). We then used the Bayesian method encoded in the *codeml* program in PAML to statistically infer amino acid changes along the branches. Convergent changes resulting from the incorrect inference of the ancestral state are recorded as false convergence (supplementary fig. S1, Supplementary Material online).

Since the simulations do not assume any adaptive evolution, *all* convergence detected in simulations is noise that consists of false convergence or random convergence as summarized in table 1. The number of random convergence events is very high (18,766 for 2.01 million sites). Furthermore, because the accuracy of ancestral inference by the Bayesian method is only 91.21%, the incorrect inferences of the ancestral state contribute 4,645 additional false convergence events. With such a high noise level, it is not surprising that much of the inferred convergence would be random noise, as previous articles have noticed (Parker et al. 2013; Foote et al. 2015; Thomas and Hahn 2015; Zou and Zhang 2015a). Although the Bayesian estimate can be improved by, for example, raising the cutoff for posterior probability, the risk of filtering out true convergence also increases accordingly (see Discussion).

The CCS Method (Convergence at Conservative Sites) and the Reduction in Noise

In this study, we propose an approach that infers *convergence at conservative sites* (CCS, for short) only in order to filter out noises as stringently as possible. Assuming that conservative sites (in the old habitat) should be more dependent on the environment than nonconservative sites, we reason that the former may be more likely to evolve by convergence when the environment changes than the latter. Conservative sites also permit more accurate inference of the ancestral state (hence avoiding false convergence) while simultaneously reducing the chance of random convergence. The criteria for inferring convergence are given in figure 1, where the inference is symmetric between mangroves and nonmangrove species. At conservative sites, either the three nonmangrove (ingroup) species or the three mangrove species have the same character as the outgroup; i.e., $N_1 = N_2 = N_3 = O$ or $M_1 = M_2 = M_3 = O$. If the criteria are met, we assumed that the ancestral states of all three clades are the same as the outgroup. This assumption is true in nearly all conservative sites, as we show below. With ancestors inferred as O, convergence can be inferred if two (or three) of the three other species share a derived character that is different from the ancestral state, i.e., $M_i = M_j \neq O$ or $N_i = N_j \neq O$.

As shown in table 1, the CSS method reduces the noise level by 93% (from 23,411 to 2,180). The number of random convergent sites decreases from 18,766 to 2,163 and the number of false convergent sites drops from 4,645 to 17. The latter is due to the much greater accuracy in ancestral state inference (99.8% vs. 91.2%). Nevertheless, even when the criteria are as stringent as defined above, the noise level in a genome-wide survey of two million amino acid sites is still non-negligible, necessitating the need for a proper control that is based on actual data. We hence use the level of convergence among nonmangroves as the real world control. Observed convergence in mangroves has to be significantly higher than that in the nonmangrove control to permit the estimation of convergence.

II. Inference of Convergence in Mangrove Species

We now apply the CCS method to the genome sequences of the seven species of figure 1. The objective is to identify candidate genes evolving towards the same state by assuming that many such genes should have multiple convergent sites. Genes with a single such site or mixed sites (convergence in both mangroves and nonmangroves) are excluded because they are less likely to be undergoing convergent evolution for both biological and statistical reasons.

Identification of Candidate Genes Undergoing Convergent Evolution

As a result, we found 5,830 and 4,580 convergent amino acid substitutions in mangroves and nonmangroves, respectively.

Table 2. Number of Genes of Convergence in Mangroves and Nonmangroves.

No. of Convergent Sites per Gene	Mangroves	Nonmangroves	Mangrove/Nonmangrove Ratio
No. of genes with convergent sites in both mangroves and nonmangroves			
	1,860	1,860	1.0
No. of genes with convergent sites in mangroves or nonmangroves only			
≥1 (=1)	1,318 (269)	944 (307)	1.40 (0.88)
≥2 (=2)	1,049 (503)	637 (324)	1.65 (1.55)
≥3 (=3)	546 (314)	313 (196)	1.74 (1.60)
≥4 (=4)	232 (138)	117 (75)	1.98 (1.84)
≥5 (=5)	94 (49)	42 (27)	2.24 (1.81)
≥6 (=6)	45 (24)	15 (12)	3.00 (2.00)
≥7	21	3	7.00
Total number of sites	5,830	4,580	1.27

Table 3. The Convergence/Divergence Test between the Paired Mangroves and the Paired Nonmangroves.

Mangrove Pairs versus Nonmangrove Pairs	Convergent Sites (*C*)	Divergent Sites (*D*)	*C*/*D* ratio	*P*-value
A. marina and *R. apiculata*	18,655	98,332	0.190	5.6×10^{-5}
S. indicum and *P. trichocarpa*	14,225	78,579	0.181	
A. marina and *S. alba*	21,791	112,580	0.194	8.3×10^{-9}
S. indicum and *E. grandis*	16,851	92,727	0.182	
R. apiculata and *S. alba*	18,244	93,992	0.194	6.2×10^{-10}
P. trichocarpa and *E. grandis*	12,893	71,674	0.180	

The convergence level in mangroves is significantly higher than that of nonmangroves ($P < 10^{-33}$ by the χ^2 test). Table 2 shows the number of genes with convergent sites in mangroves as well as nonmangroves. Among them, 1,860 genes contain convergent sites in both mangroves and nonmangroves and are excluded from further analysis. The number of genes that contain convergent sites in mangroves only is 1,318, which is larger than 944 for nonmangroves only. Partitioning these genes by the number of convergent sites, we can discern a trend described below.

Genes with a single convergent site are in fact more common in nonmangrove species than in mangroves (ratio = 0.88). The ratio in the convergence signal between mangroves and nonmangroves is significantly greater than 1 in genes with two or more convergent sites (table 2). Taking the difference between the two numbers, we estimate that ~400 (1,049–637) genes in mangroves bear the signal of convergence. We further examine genes with 4 or more convergent sites in greater detail. With ≥4 sites, the ratio is 232/117 (=1.98, $P < 10^{-9}$ by the χ^2 test), which may be strong enough to reveal the ontology of genes driven by convergent selection.

An important issue is whether the higher level of convergence in mangroves might be due to their faster evolution in the new environment. To address this possibility, we first apply the convergence/divergence, or C/D, test of Castoe et al. (2009) to a pair of mangroves and a corresponding pair of nonmangroves. In this test, divergent sites serve as the control for branch length. The convergent and divergent sites are counted for each pair of mangroves and their nonmangrove counterparts. We use PAML to infer the character state of the common ancestor of the four species. As shown in table 3, all three sets of mangrove—nonmangrove comparisons show significantly higher C/D ratios between mangroves than between nonmangroves.

In a second approach that controls for the difference in substitution rate, we note that our conclusion is based on the highly conservative portion of the genomes. Therefore, the relevant question is whether this portion tends to evolve faster in mangroves than in nonmangroves. We expand the set of conservative sites to include additional sites where at least one mangrove (or nonmangrove) species retains the ancestral character. When we score the number of changes at these sites, the rate of evolution in mangroves and nonmangroves is nearly equal (with a ratio of 1.03, or 208,472 vs. 202,072 changes). In short, among the moderately to strongly conservative sites of the genomes, there is no difference in the rate of amino acid changes between mangroves and nonmangroves. We conclude that the stronger convergence in mangroves is not associated with branch length differences. In addition, the greater concentration of CCS sites in mangroves (i.e., fewer genes with one single site and more genes with multiple sites) than in nonmangroves corroborates this conclusion.

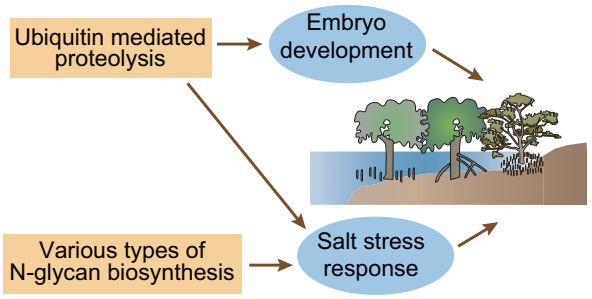
Finally, in the whole-genome analysis, the level of noise as seen in nonmangrove convergence is not trivial. For that reason, the convergence measure in mangroves should be accepted as a probability statement, not a confirmation. Even for genes with ≥4 CCS sites, the probability that such a gene has experienced convergent evolution would only be 50%. This is because nonmangroves also yield a convergence signal at half as many sites as mangroves. The ontology of these candidate genes is hence informative as it represents an approximate functional test.

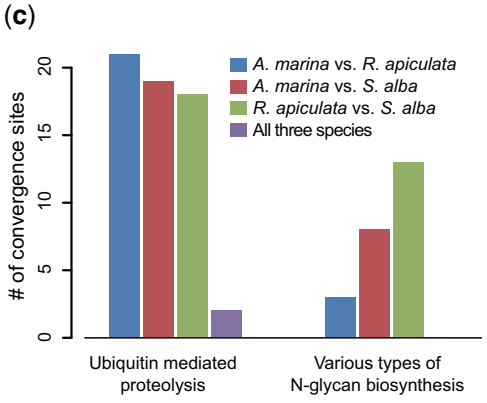
Ontology of Candidate Genes Undergoing Convergent Evolution

We then mapped the 232 candidate genes potentially undergoing convergent evolution to the KEGG PATHWAY

(a)

Pathway	Total No. of genes in the category	No. of convergent genes		P-value
		Mangroves	Non-mangroves	
All	5353	232	117	--
Ubiquitin mediated proteolysis	105	12	0	1.8×10^{-3} (FDR < 0.05)
Various types of N-glycan biosynthesis	19	5	0	1.0×10^{-3} (FDR < 0.05)





(d)

	Domains							Domains						
Site	3	167	335	338	420	520	624	15	39	601	681	706	714	785
O. sativa	S	Q	E	N	S	V	V	L	S	Q	V	K	L	H
A. marina	S	H	E	N	T	I	I	L	P	E	V	N	L	H
S. indicum	S	Q	E	N	S	V	V	L	S	Q	V	K	L	H
R. apiculata	P	H	D	S	T	V	I	I	S	E	I	K	V	R
P. trichocarpa	S	Q	E	N	S	V	V	L	S	Q	V	K	L	H
S. alba	P	Q	D	S	T	I	M	I	P	Q	I	N	V	R
E. grandis	S	Q	E	N	S	V	V	L	S	Q	V	K	L	H
	SAE2							MAN2						

FIG. 2. Two pathways significantly enriched for convergent genes among mangroves. (*a*) The two pathways that are most significantly enriched for convergent genes are shown. Fisher's exact test and FDR (Benjamini–Hochberg correction) probabilities are given. (*b*) A sketch explaining the roles of the two pathways enriched for mangrove convergent genes played in mangroves' adaptation. (*c*) Pairwise convergence among the 82 convergent sites in genes from the two pathways. Note that the convergence rate is higher between *R. apiculata* and *S. alba* in the pathway "various types of N-glycan biosynthesis". (*d*), Two examples of functional domains where convergent sites are clustered (see text). The positions are aligned with the protein sequence of *O. sativa*.

Database (Kanehisa and Goto 2000; see Materials and Methods). Fisher's exact test and Benjamini–Hochberg correction were used to test for statistical significance. Two pathways are significantly enriched for convergent genes in mangroves (Fisher's exact test, $P < 0.01$; FDR < 0.05; fig 2*a* and supplementary table S1, Supplementary Material online); in contrast, no pathway is significant in the nonmangrove list.

The two enriched pathways are "ubiquitin mediated proteolysis" (12 genes) and "various types of N-glycan biosynthesis" (5 genes) (fig. 2*a* and supplementary table S1, Supplementary Material online). In plants, ubiquitination is involved in many biological processes, including embryonic development, photomorphogenesis, organ development and abiotic stress responses, via regulating plant hormone signal transduction (Frugis and Chua 2002; Lyzenga and Stone 2012). Among them, embryo development and abiotic stress response are especially important to mangroves (fig. 2*b*). Abiotic stress tolerance, such as high salinity tolerance, is essential for the survival of mangroves in the intertidal zone, whereas viviparous embryos are protected from the harsh environment. Some of these proteins showing mangrove convergence have been shown to play roles in embryo development or abiotic stress response, including DNA damage binding protein 1 (DDB1), Cullin 3 (CUL3) and small ubiquitin-like modifiers activating enzyme 2 (SAE2) (Thomann et al. 2005; Saracco et al. 2007; Bernhardt et al. 2010; Lyzenga and Stone 2012). Athough in some contexts these genes may seem to be involved in basal functions for a cell or cells, such activities become targets of adaptation when the functions or pathways they are involved in are put under stress. Thus, they may be targets of adaptive evolution. Additionally, N-glycosylation is involved in protein folding and stabilization (Rayon et al. 1998) as well as salinity tolerance (alpha-mannosidase II, MAN2; OST2; WBP1; Koiwa et al. 2003; Kang et al. 2008; Kaulfürst-Soboll et al. 2011).

Within the 17 mangrove convergent genes of the two pathways, there are 84 convergent sites. Two of these sites show convergence in all three mangroves, whereas 82 exhibit the pattern in two out of the three mangroves (fig. 2*c*). The two pan-mangrove sites are located in the genes *SAE2* and *GRR1*, both belonging in "ubiquitin mediated proteolysis". SAE2 conjugates with SAE1 as a heterodimer to activate SUMO, which then attaches to a target protein. SAE2 is essential for viability in *Arabidopsis thaliana*, and SUMOylation plays a key role in stress response (Saracco et al. 2007). Based on 3D structural modeling, the substitution S420T is located in SAE2's adenylation domain, which directly binds and activates SUMO (Lois and Lima 2005; Biasini et al. 2014; supplementary fig. S2, Supplementary Material online). GRR1 encodes a subunit of SCF-type E3 ligase, which directs the degradation of plant hormone transcriptional repressors resulting in stress response and regulation of embryonic

development. In addition, both sites were predicted to increase the stability of protein structure (Predicted with MUpro; Cheng et al. 2006). In a stressful environment, a protein's normal folding and functioning are challenged. Therefore, increasing its stability should be beneficial for mangroves' survival in the intertidal zone. Notably, we found more convergent sites between *R. apiculata* and *S. alba*, mainly in the various types of N-glycan biosynthesis pathway (fig. 2*c*). The greater degree of convergence may correspond to the phenotypic convergence between *R. apiculata* and *S. alba*, both of which exclude salt in roots (Tomlinson 1986). In 11 of the 17 genes, all three types of pairwise convergence shown in Fig. 2*c* can be found, suggesting common adaptive strategies among all mangroves.

Furthermore, most convergent sites (64 of 84) are found in functional domains. Figure 2*d* shows the SAE2 and MAN2 proteins as examples. In both of these cases five of the seven convergent sites are located in domain regions (IPR016040 and IPR028077 in SAE2, IPR028995 and IPR011013 in MAN2).

We have examined in detail only 17 genes in two pathways. There are other pathways also enriched for convergent genes albeit with a lower level of significance (supplementary table S1, Supplementary Material online). One such example is the Glutathione metabolism pathway with five genes showing convergence ($\geq$4 sites per gene) in mangroves and one gene showing convergence in nonmangroves ($P < 0.05$, FDR > 0.05). In this pathway, glutathione is a well-known antioxidant that prevents the important cellular components from damage by reactive oxygen species (Pompella et al. 2003). The five genes in the glutathione metabolism pathway include *OXP1*, *RRM1*, *ODC1*, *G6PD*, and *APX1* (supplementary table S2 and fig. S3, Supplementary Material online). In addition to playing key roles in reactive oxygen scavenging, the gene *APX1* also has been shown to affect embryo development ending in seed dormancy and salt stress tolerance (Noctor and Foyer 1998; Pagnussat et al. 2005). Curiously, genes related to glutathione metabolism have also been found to harbor convergent substitutions in marine mammals (Yim et al. 2014; Foote et al. 2015). More interestingly, gene *ODC1* also carries species-specific amino acid changes in cetaceans (Yim et al. 2014). Perhaps glutathione metabolism is important for switching from a terrestrial to aquatic lifestyle.

Discussion

In this study using the new CCS method, we estimate the total number of convergent sites in three species of mangroves. Our analyses suggest that there are approximately 1,400 such amino acids, distributed among about 400 genes. The estimates are likely conservative due to the need to filter out the noise of random and false convergence, a process that removes some signal as well. However, had we used all sites with based on standard Bayesian inference, there would have been 23,000 sites exhibiting background (nonadaptive) convergence, which would almost certainly overwhelm the signal of only 1,400 sites.

The CCS method uses only highly conservative sites where all mangroves or all nonmangroves are invariant and identical with the outgroup state. The filtering is maximally stringent, but the noise (see Table 1) is still substantial, with ~2,200 amino acids genome-wide. Indeed, any method aiming at detecting convergence has to explicitly factor in the noise level. In this study, we further filter out noise by examining its distribution among genes and retaining those with a high number of convergent sites. The analysis of noise level and noise distribution explains why an earlier attempt at detecting convergence in genomic sequences (Parker et al. 2013) has subsequently been found to have no power at all (Thomas and Hahn 2015; Zou and Zhang 2015a). Since the number of convergent substitutions would decrease as more species are used (Thomas et al. unpublished data), a larger set of species is still the most robust way to detect convergence.

Methods to detect convergence have to maximize noise reduction without purging an undue amount of signal. For example, to reduce false convergence, an alternative is to raise the posterior probability cutoff when using Bayesian inference. As an example, the strongest convergence signals are sites with a configuration of $N_1 = N_2 = N_3 = O$ and $M_1 = M_2 = M_3 \neq O$ as shown in figure 1. Such sites do not yield a high posterior probability by the Bayesian method because both N and M would be given a high probability of being the ancestral state. Indeed, the average posterior probabilities of inferred ancestors in the two three-way convergent sites are only 0.809 and 0.709, respectively. Similar cases of true convergence are also susceptible to being filtered out.

The simplicity of the CCS method frees it from many assumptions and biases. In contrast, Bayesian ancestral reconstruction depends strongly on the accuracy of the parameters of the prior distributions (Zhang and Nei 1997; Zou and Zhang 2015b). These parameters include the amino acid substitution matrix and the equilibrium amino acid frequency. Inaccuracy in estimating these parameters would introduce biases. In general, ancestral sequence reconstruction methods often have systematic biases (Goldstein and Pollock 2006; Matsumoto et al. 2015) which are difficult to ascertain. When the background noise is as large as in the detection of convergence, a simple method like CCS offers the advantage of minimizing and estimating such biases.

The two pathways enriched with mangrove convergent genes are upstream regulation processes, which regulate more than salt stress response and development we emphasized previously. For example, we identified ubiquitin mediated proteolysis, which is involved in plant hormone signal transduction (Frugis and Chua 2002; Kelley and Estelle 2012) and influences other processes of plant development, including embryo development. Protein glycosylation also happens in many different proteins. Although we could not rule out the contributions of other processes on the adaptation of mangroves to high salinity, hypoxia, strong UV, strong wind, etc., salt stress response and embryo development stand out.

In contrast with the more specialized convergence such as echolocation, transitions from terrestrial to aquatic habitats are broad and general. Echolocation likely results from the

convergence of downstream processes regulating a few specific genes. Since mangroves had to have numerous adaptations in their new environment, selection may operate on a large number of genes controlling many traits. We have identified a small portion of genes and pathways that may have evolved by convergence in mangroves. Finally, this study provides a new framework for assessing genomic convergence by explicitly incorporating noise reduction and estimation in the assessment.

Materials and Methods

Orthologous Gene Sets Building

The genome sequences of the three mangrove species (*Avicennia marina*, *Rhizophora apiculata*, and *Sonneratia alba*) were sequenced and assembled using Illumina paired-ends/mate pair short-reads sequencing technology (He et al. unpublished data). The sequences of inland species (*Sesamum indicum*, *Populus trichocarpa*, *Eucalyptus grandis*, and *Oryza sativa*) were downloaded from Phytozome or Sinbase (https://phytozome.jgi.doe.gov; http://ocri-genomics.org/Sinbase/; last accessed December 6, 2016). To build the orthologous groups of the seven species used, we first clustered the genes of the seven species into gene families using OrthoMCL (Li et al. 2003). Then the putative orthologous groups were selected according to the following process: Within each gene family, we chose the most similar pair among the three mangrove:nonmangrove pairs and picked the rice gene most similar to the three pairs as the outgroup. The putative orthologs were aligned using a combination of PAL2NAL (Suyama et al. 2006) and MUSCLE (Edgar 2004). Orthologs shorter than 50 amino acids were removed. This resulted in 5,353 ortholog groups with high confidence were built for convergence detection.

Sequence Simulation

The branch lengths, amino acid frequencies and the best shape parameter for variable rates among sites (alpha) were obtained from the amino acid sequences described above using the *codeml* program from the PAML package (Yang 2007). The JTT+ gamma model was employed as the amino acid substitution model. With these parameters, the amino acid sequence simulation was performed using the program *evolver* from PAML. It first produced a random ancestral sequence according to the given amino acid frequency, then modeled sequences evolution according to the given tree topology and branch lengths. Finally, the simulated amino acid state in each internal node and leaf was obtained. We simulated the same number of amino acids as were found in the real data set to ease comparisons.

Ancestral State Inference

Bayesian ancestral state inference in simulated sequences was performed using the *codeml* program. The amino acid frequency, alpha and amino acid substitution model used in inference were the same as that in the sequence simulation. In the Bayesian framework, the state with the highest posterior probability was retained as the inferred ancestral state. In the CCS method, the state of the outgroup was assumed to be the ancestral state. The assumed ancestral state was compared with the simulated state to get the accuracy of the inference.

KEGG Pathway Mapping

The protein sequences of the three mangrove species were blasted against the KEGG GENES database (*e*-value $\leq 1 \times 10^{-5}$) for KO (KEGG Orthology) assignments and pathway mapping. In each ortholog group, the KEGG assignment of *S. alba* was used to represent the ortholog group.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Y.X. Fu and K. Zeng for suggestions and discussions. This work was supported by the National Natural Science Foundation of China (grant numbers 41130208, 91331202, and 31600182), the 985 Project (grant number 33000-31131105), the Science Foundation of State Key Laboratory of Biocontrol (grant number SKLBC16A35 and SKLBC16A37), the Fundamental Research Funds for the Central Universities (grant number 16lgjc39), the China Postdoctoral Science Foundation (grant numbers 2014M552264 and 2015T80931), and the Chang Hungta Science Foundation of Sun Yat-Sen University.

References

Bergsten J. 2005. A review of long-branch attraction. *Cladistics* 21:163–193.

Bernhardt A, Mooney S, Hellmann H. 2010. Arabidopsis DDB1a and DDB1b are critical for embryo development. *Planta* 232:555–566.

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, et al. 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 42:W252–W258.

Castoe TA, de Koning APJ, Kim H-M, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc Natl Acad Sci U S A.* 106:8986–8991.

Cheng J, Randall A, Baldi P. 2006. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* 62:1125–1132.

Dobler S, Dalla S, Wagschal V, Agrawal A. 2012. Community-wide convergent evolution in insect adaptation to toxic cardenolides by substitutions in the Na,K-ATPase. *Proc Natl Acad Sci U S A.* 109:13040–13045.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.

Foote AD, Liu Y, Thomas GWC, Vinar T, Alfoldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V, et al. 2015. Convergent evolution of the genomes of marine mammals. *Nat Genet.* 47:272–275.

Frugis G, Chua NH. 2002. Ubiquitin-mediated proteolysis in plant hormone signal transduction. *Trends Cell Biol.* 12:308–311.

Giri C, Ochieng E, Tieszen LL, Zhu Z, Singh A, Loveland T, Masek J, Duke N. 2011. Status and distribution of mangrove forests of the world using earth observation satellite data. *Glob Ecol Biogeogr.* 20:154–159.

Goldstein RA, Pollard ST, Shah SD, Pollock DD. 2015. Nonadaptive amino acid convergence rates decrease over time. *Mol Biol Evol.* 32:1373–1381.

Goldstein RA, Pollock DD. 2006. Observations of amino acid gain and loss during protein evolution are explained by statistical bias. *Mol Biol Evol*. 23:1444–1449.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 28:27–30.

Kang JS, Frank J, Kang CH, Kajiura H, Vikram M, Ueda A, Kim S, Bahk JD, Triplett B, Fujiyama K, et al. 2008. Salt tolerance of *Arabidopsis thaliana* requires maturation of N-glycosylated proteins in the Golgi apparatus. *Proc Natl Acad Sci U S A*. 105:7893.

Kaulfürst-Soboll H, Rips S, Koiwa H, Kajiura H, Fujiyama K, Von Schaewen A. 2011. Reduced immunogenicity of Arabidopsis hgl1 mutant N-glycans caused by altered accessibility of xylose and core fucose epitopes. *J Biol Chem*. 286:22955–22964.

Kelley DR, Estelle M. 2012. Ubiquitin-mediated control of plant hormone signaling. *Plant Physiol*. 160:47–55.

Koiwa H, Li F, McCully MG, Mendoza I, Koizumi N, Manabe Y, Nakagawa Y, Zhu J, Rus A, Pardo JM, et al. 2003. The STT3a subunit isoform of the *Arabidopsis* oligosaccharyltransferase controls adaptive responses to salt/osmotic stress. *Plant Cell* 15:2273–2284.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13:2178–2189.

Li Y, Liu Z, Shi P, Zhang J. 2010. The hearing gene Prestin unites echolocating bats and whales. *Curr Biol*. 20:55–56.

Liu Y, Cotton JA, Shen B, Han X, Rossiter SJ, Zhang S. 2010. Convergent sequence evolution between echolocating bats and dolphins. *Curr Biol*. 20:53–54.

Lois LM, Lima CD. 2005. Structures of the SUMO E1 provide mechanistic insights into SUMO activation and E2 recruitment to E1. *EMBO J*. 24:439–451.

Losos JB. 2011. Convergence, adaptation, and constraint. *Evolution* 65:1827–1840.

Lyzenga WJ, Stone SL. 2012. Abiotic stress tolerance mediated by protein ubiquitination. *J Exp Bot*. 63:599–616.

Matsumoto T, Akashi H, Yang Z. 2015. Evaluation of ancestral sequence reconstruction methods to infer nonstationary patterns of nucleotide substitution. *Genetics* 200:873–890.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, et al. 2014. The genome of *Eucalyptus grandis*. *Nature* 510:356–362.

Noctor G, Foyer CH. 1998. Ascorbate and glutathione: keeping active oxygen under control. *Ann Rev Plant Physiol Plant Mol Biol*. 49:249–279.

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, et al. 2007. The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res*. 35:D883–D887.

Pagnussat GC, Yu H-J, Ngo QA, Rajani S, Mayalagu S, Johnson CS, Capron A, Xie L-F, Ye D, Sundaresan V. 2005. Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. *Development* 132:603 LP-614.

Parida AK, Jha B. 2010. Salt tolerance mechanisms in mangroves: a review. *Trees* 24:199–217.

Parker J, Tsagkogeorga G, Cotton J. a, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* 502:1–9.

Pompella A, Visvikis A, Paolicchi A, Tata VD, Casini AF. 2003. The changing faces of glutathione, a cellular protagonist. *Biochem Pharmacol*. 66:1499–1503.

Rayon C, Lerouge P, Faye L. 1998. The protein N-glycosylation in plants. *J Exp Bot*. 49:1463–1472.

Ricklefs RE, Latham RE. 1993. Global patterns of diversity in mangrove floras. In: Species diversity in ecological communities: historical and geographical perspectives. Chicago: University of Chicago Press. p. 215–229.

Rokas A, Carroll SB. 2008. Frequent and widespread parallel evolution of protein sequences. *Mol Biol Evol*. 25:1943–1953.

Saracco S. a, Miller MJ, Kurepa J, Vierstra RD. 2007. Genetic analysis of SUMOylation in *Arabidopsis*: conjugation of SUMO1 and SUMO2 to nuclear proteins is essential. *Plant Physiol*. 145:119–134.

Shi S, Huang Y, Zeng K, Tan F, He H, Huang J, Fu Y. 2005. Molecular phylogenetic analysis of mangroves: independent evolutionary origins of vivipary and salt secretion. *Mol Phylogenet Evol*. 34:159–166.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 34:W609–W612.

Thomann A, Brukhin V, Dieterle M, Gheyeselinck J, Vantard M, Grossniklaus U, Genschik P. 2005. Arabidopsis CUL3A and CUL3B genes are essential for normal embryogenesis. *Plant J*. 43:437–448.

Thomas GWC, Hahn MW. 2015. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. *Mol Biol Evol*. 32:1232–1236.

Thomas GWC, Hahn MW, Hahn Y. unpublished data, http://biorxiv.org/content/early/2016/10/17/081612, posted October 17, 2016.

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

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313:1596–1604.

Wang L, Yu S, Tong C, Zhao Y, Liu Y, Song C, Zhang Y, Zhang X, Wang Y, Hua W, et al. 2014. Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biol*. 15:R39.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.

Yim H-S, Cho YS, Guang X, Kang SG, Jeong J-Y, Cha S-S, Oh H-M, Lee J-H, Yang EC, Kwon KK, et al. 2014. Minke whale genome and aquatic adaptation in cetaceans. *Nat Genet*. 46:88–92.

Zhang J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat Genet*. 38:819–823.

Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. *Mol Biol Evol*. 14:527–536.

Zhang J, Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J Mol Evol*. 44:139–146.

Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012. Parallel molecular evolution in an herbivore community. *Science* 337:1634–1637.

Zou Z, Zhang J. 2015a. No genome-wide protein sequence convergence for echolocation. *Mol Biol Evol*. 32:1237–1241.

Zou Z, Zhang J. 2015b. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Mol Biol Evol*. 32:2085–2096.