Available online at www.ijpab.com

ISSN: 2320 – 7051
Int. J. Pure App. Biosci. **1 (4):** 67-77 (2013)
Research Article

International Journal of Pure & Applied Bioscience

Phylogenetic reconstruction of the narrow-leaf cattail *Typha angustifolia* L., using the chloroplast maturase K (*matK*) gene sequence

A.Sethuraman[1] and K.P. Sanjayan*

[1] Dept. of Plant Biology and Biotechnology
Dept. of Advanced Zoology and Biotechnology, Guru Nanak College, Chennai 600 042
*Corresponding Author Email: kpsanjayan @yahoo.co.in

ABSTRACT

Molecular characterization of the narrow-leaf cattail, Typha angustifolia L., was carried out using the chloroplast maturase K (matK) gene sequence. The matK gene was isolated from the chloroplast and amplified using the primers: forward matK F 5'- CGATCTATTCATTCAATATTTC-3' and reverse matK R 5'- TCTAGCACACGAAAGTCGAAGT-3'. The amplified product was sequenced using the ABI PRISM 3730XL Analyzer. The 879 nucleotide sequences had a GC content of 32.08%. Highlights of the matK gene of T. angustifolia using the data of the alignment of the 6 sequences available in the genebank is provided in terms of the conserved, variable and singleton sites. The nucleotide sequences were also translated into amino acids to compare the patterns of amino acid variation with those of the nucleotide substitutions. The ratio of transition to transversion pairs was 1.25. A Blastn similarity search was conducted and the sequences of other species of Typha namely, T. angustifolia, T. latifolia, T. domingensis, and T.capensis together with 4 species of Sparganium, 2 species of Quesnelia and one each of the genus Neoregelia and Billbergia were used to develop a phylogenetic tree. Analysis of the maximum likelihood of 24 different nucleotide substitution patterns, indicated that the T92+G had the least BIC and AIC scores and therefore considered the best model for the present data set. The evolutionary tree depicts T.capensis, T.aungustifolia and T.latifolia to share a common ancestor among themselves with T. domingensis being a closely related species. All the species of Sparganium were grouped together in terms of their similar number of substitutions per site values and all the other remaining plant species formed a different clade.

Key words: *Typha angustifolia, matK gene, chloroplast, molecular characterization, phylogenetics, cattail, Poales*

INTRODUCTION

Plant species diversity influence ecosytem functions such as net primary productivity[1], nitrogen cycling[2] and hydrological processess[3]. The marsh land ecosystem is unique by way of inflow from the sea and fresh water and therefore has its own heterogenous assemblages of plant species. Knowledge of the distribution of plant species diversity is essential for understanding interactions among organisms and developing sound environment strategies. Historically, plants have been identified based on flower, fruit, leaf and stem morphology covering only one third of the known plant species. Many plant groups, especially the grasses, are difficult to identify because they can be most readily distinguished only by the reproductive structures that are available for a short period each year. Species of *Typha* form the most dominant plant among the order Poales in the marsh land ecosytem. Currently about eleven species of *Typha* have been recorded world wide. The idenfication of these species require highly skilled personnel and time-consuming procedures such as cellular characterization and enormous chemical tests. Under

such circumstances, the future of plant identification lies in the development of DNA-based diagnostic systems for which the life stage or source of tissue is irrelevant. The slow mutational tempo and horizontal gene transfer in plants make mitochondrial barcodes less attractive for barcoding plants. The nuclear ITS regions suffer from practical difficulties associated with the existence of multiple paralogous copies in many plant taxa[4] and therefore of limited utility. The chloroplast geneome has a slow tempo of evolution[5] and is a promising candidate for barcoding plants. The chloroplast coding regions namely, *ndhF, matK* and *rbcL* and the non-coding regions namely, *trnD-trnT* and *rpoB-trnC* are presently being vouched as promising candidates[6,7]. Givnish[8] used the *ndh*F sequence to analyse the phylogenetic relationship of the members of the family Bromeliaceae. The *mat*K gene of chloroplast is 1500 bp long, located within the intron of the *trn*K and codes for maturase like protein, which is involved in Group II intron splicing. The two exons of the *trn*K gene that flank the *mat*K is lost, leaving the gene intact in the event of splicing. The gene contains high substitution rates within the species and is emerging as potential candidate to study plant systematics and evolution[9,10]. A homology search for this gene indicates that the 102 AA at the carboxyl terminus are structurally related to some regions of maturase-like polypeptide and this might be involved in splicing of group II introns[11-14]. It is another emerging gene with potential contribution to plant molecular systematics and evolution[15]. The *mat*K gene has ideal size, high rate of substitution, large proportion of variation at nucleic acid level at first and second codon position, low transition/transversion ratio and the presence of mutationally conserved sectors[16]. These features of *mat*K gene are exploited to resolve family and species level relationships. Here an attempt has been made to study the *matK* gene sequence as molecular marker for identifying *Typha angustifolia* L., and using this gene for reconstructing its phylogeny.

MATERIALS AND METHODS

Genomic DNA isolation

A slightly modified method of Doyle and Doyle[17] was employed. Fresh-leaf tissue (0.2 g) was ground in a 1.5-ml centrifuge tube with a micropestle and preheated freshly prepared 800 µl of CTAB extraction buffer (0.1 M Tris-Cl (pH 9.5), 20 mM EDTA (pH 8), 1.4 M NaCl, CTAB (2%, w/v), b-mercaptoethanol (1%, v/v)) was immediately added to the tube. The tube was incubated at 65°C for 35-45 min, with inversion during incubation. An equal volume of chloroform: isoamyl alcohol (24:1, v/v) was added and then the tubes were inverted 8-10 times and centrifuged at 13,000 rpm for 15 min. The supernatant was placed in a new centrifuge tube and an equal volume of absolute ice-cold isopropanol was added. The tubes were centrifuged at 13,000 rpm for 10 min. The supernatant was discarded and the pellet was washed with 70% (v/v) ethanol. The pellet was air-dried at room temperature and then dissolved in 20 µL TE buffer. The DNA samples were stored at -20°C until further use. The purity of the DNA extracted was checked by recording the absorbance of the sample at 260nm and 280 nm.

Amplification and sequencing of matK gene

PCR was carried out in an Eppendorf Personnel Master Cycler (Germany) at 4°C. The PCR conditions were 94°c for 3minutes (Initial denaturation), 94°c for 30 seconds (Denaturation), 47°c for 1minute (primer annealing), 72°c for 1minute 20 seconds (extension), and further 72°c 7 minutes for final extension. The run had 40 cycles. The primers used for amplification were forward matK F 5'-CGATCTATTCATTCAATATTTC-3' and reverse matK R 5'- TCTAGCACACGAAAGTCGAAGT-3'. This amplified product was sequenced using the ABI PRISM 3730XL Analyzer.

matK gene sequence analysis and phylogenetic reconstruction

A Blastn similarity search was conducted using the sequenced matK gene of *T. angustifolia* as the query sequence. The nucleotide collection database was searched with the organism key as Poales, taxid 38820. The programme was optimized for highly similar sequences (megablast). The top sequences producing significant alignments were selected. These nucleotide sequences were aligned using the ClustalW option present in the MEGA5 software[18]. After computing the alignment, the data menu was opened and the

active data was explored for analysis of various sites such as conserved sites, parsimonious informative sites, variable sites etc., using the highlight section of the sequence data explorer window of the MEGA tool. The statistics of the nucleotide composition was analysed and automatically exported to Microsoft Excel 2007. Further, the aligned sequences were used to find the Best DNA model and to compute the pair-wise distance in order to estimate the evolutionary divergence between the sequences. To construct Phylogenetic trees, the Maximum Likelihood method and the Neighbor-Joining method were employed and the test of phylogeny had 500 bootstrap replications.

RESULTS AND DISCUSSION

The partial nucleotide sequence for the matK gene of *T.angustifolia* is presented in table 1. The sequence had a total of 879 nucleotides. The base statistics for this sequence is presented in table 2. The sequence had the maximum number of T nucleotides followed by the A nucleotide. Also the percentage of A+T was more than that of G+C. The ratio of AT : GC was 2.11 : 1. Guisinger[19] reported that the complete *T. latifolia* plastid genome has a 33.8% GC content. In the current study, for *T. aungustifolia* a GC content of 32.08% was recorded with reference to the *mat*K gene.

Sequence similarity

Using the matK sequence of *T. angustifolia* as the query sequence, a Blastn similarity search conducted resulted in 102 hits that included 82 organisms. Top 12 sequences were selected, aligned and viewed through NCBI sequencer viewer 2.26 (Table 3). These comprised of four species each of *Typha*, and *Sparganium*, 2 species of *Quesnelia* and one each of the genus *Neoregelia* and *Billbergia.* Taxonomically, *Typha,* and *Sparganium* belong to the family Typhaceae, while *Quesnelia, Neoregelia* and *Billbergia* belong to the family Bromeliaceae. The Blast expect values for all these sequences were greater than 1 x 10^{-179} indicating that the similarity between *T. angustifolia* and the other members selected represent some inherent biological relationship. Homology can also be inferred from the good alignments that have been observed here. Table 4 provides a comparative account of the nucleotide frequencies in percentage of the different species with respect to the matK gene. For all species T showed the maximum percentage frequency. Nucleotide frequency for overall codon favoured A+T with a strong bias. At the Ist position, there was a A+T bias, with T ranging from 39-44% at an average of 42%. At the 2nd and 3rd positions also there was an A+T bias with an average T of 32% and 36% respectively. At all the positions, the frequency of A followed that of T in the nucleotide frequency percentage.

The sequence similarity search resulted in 17 hits for *Typha* represented by 4 species, namely *T. angustifolia, T. latifolia, T. domingensis, and T.capensis*. There were six hits for *Typha angustifolia* alone of which Accession AM114723.1 was from Denmark, EU749454.1 and EU749453.1 from Canada, AY952419.1 and GQ434092.1 from China and JN894175.1 from UK. Table 5 provides the highlights of the matK gene of *T. angustifolia* using the data of the alignment of the 6 sequences available in the genebank and the sequence of the present study. The un-translated aligned sequence had 880 conserved, 9 variable and 5 singleton sites. The translated amino acid sequence contained 267 conserved, 6 variable and 2 singleton regions. Conserved site is a site containing the same nucleotide or amino acids in the sequences being studied and is also referred to as constant site. A variable site contains at least two types of nucleotides or amino acids. No parsimony-informative site, ie., a site at which there are at least two different kinds of nucleotides or amino acids, each represented at least twice, was observed in the present set of data. The singleton identified here through the MEGA protocol represents a site where at least three sequences contain unambiguous nucleotides or amino acids. The alignment of the matK sequences of *T. angustifolia* showed that among the 2768 base pairs (bp), 31.79% were conserved, and 0.002% was variable. However, Liang[20] studied the entire matK sequence of eleven plant species and reported that among the 1581 base pairs (bp), 1086 (69%) were variable and 803 (51%) were phylogenetically informative. He observed several small indels, 3-15 bp along the entire length of the coding region that are mostly found in multiples of three nucleotides and opined that in most cases it does not result in a frame shift. The matK sequence of *T. angustifolia* had a very low percentage of variable sites. The section

of the sequence devoid of indels seems to be functionally important. This section of the gene corresponds to what is called "domain X"; a section that *mat*K genes share with group II intron maturases and is believed to reflect an essential function in binding of the intron RNA during reverse transcription and RNA splicing[13,14]. Again, this type of comparison between different species shows that the gene does not represent a homogenous unit in terms of amount of nucleotide variation.

The nucleotide sequences were also translated into amino acids to compare the patterns of amino acid variation with those of the nucleotide substitutions and to further evaluate the functional constraints on the gene. The 2768 nucleotides were translated to 922 amino acids and 267 (28.95%) were conserved and only 0.006% were variable. The percentage conserved sites at the amino acid level is slightly lower than that at the nucleotide level (28.95% Vs 31.79%), while no significant differences were observed for the variable sites. In functionally constrained genes, such as *rbc*L, nucleotide sequences are translated into lower amino acid variation. Johnson and Soltis[21] reported in their study of the Saxifragaceae 5% amino acid variation for the *rbc*L. In the present study the low variation percentage observed indicate that the gene may not be so functionally constrained.

Relative synonymous codon usage (RSCU)

Due to the degeneracy of genetic code, most amino acids are coded by more than one codon (synonymous codon). Studies on the synonymous codon usage can reveal information about the molecular evolution of individual genes. RSCU values are the number of times a particular codon is observed, relative to the number of times that the codon would be observed in the absence of any codon usage bias. In the absence of any codon usage bias, the RSCU value would be 1.00. A codon that is used less frequently than expected will have a value of less than 1.00 and vice versa for a codon that is used more frequently than expected[22]. Table 6 provides the details of the RSCU for the matK gene among the species of *Typha* studied here. The average number of codons observed here is 332. The codons TTA (Leucine), CCA (Proline) and AGA (Arginine) had high usage bias in the sequence.

Nucletotide substitution among the sequences

In the study of molecular evolution, it is important to know the number of nucleotide substitutions per site (d) between DNA sequences. Two important factors that are considered in the estimation of d are the inequality of the rates of transitional and transversional nucleotide substitution (transition-transversion bias) and the deviation of the G+C content from 0.5 (G+C-content bias)[23]. Transitions refer to the substitution of a purine (A or G) by another purine or the substitution of a pyrimidine (T or C) by another pyrimidine; transversions are the substitutions of a purine by a pyrimidine or a pyrimidine by a purine. The nucleotide pair frequencies (Table 7) computed for the data of the different *Typha* species showed an average of 796 identical pairs out of a total average of 798.33 of which there were 302 TT pairs, 245 AA pairs and 115 GG pairs. The ratio of the Transitional pairs versus Transversional pairs was 1.25. When two DNA sequences are derived from a common ancestral sequence, the descendant sequences gradually diverge by nucleotide substitution. A simple measure of the extent of sequence divergence is the proportion of nucleotide sites at which the two sequences are different. This is estimated as the p-distance for nucleotide sequence. It is useful to know the frequenies of different nucleotide pairs between the two sequences. Since there are four nucleotides, there are 16 different types of nucleotide pairs. There are four pairs of identitical nucleotides (AA,TT,CC,GG represented as O), four transition-type pairs (AG,GA,TC,CT represented as P) and remaining 8 transversion-type pairs (represented as Q). The p distance for nucleotide sequence, given by the relationship p=P + Q was calculated to be 3 (ie. 2+1). If nucleotide substitution occurs at random, Q is expected to be about two times higher than P when p is small. This was not the case in the present investigation. In general, transition usually occurs more frequently than transversions. Therefore P may be greater than Q. When the extent of divergence is low, the ratio (R) of transitions to transversions can be estimated from the observed values of P and Q. R is usually 0.2-2 in many nuclear genes, but in mitochondrial DNA it can be as high at 15[24]. In the present study the value of R was 1.25. The analysis of the p value indicate that no synonymous substitution occur

in the first three codons (p for Ist codon is 0.01, 2nd codon could not be calculated, and 3rd codon is 0.01). Transitions are generally more frequent than transversions[25]. Transversions are considered the more reliable type of mutations in constructing phylogenies[25]. Consequently, some workers have assigned more weight to transversions in phylogenetic analyses, or based the analyses on transversions alone, resulting in what is called transversion parsimony[25,26]. Transition/ transversion ratios have been observed to be 2.0 for relatively recently diverged sequences and exceed 0.4 for highly substitution-saturated sequences[27]. The ratio of transition to transversion (nr/nv) ranged from 0.39 between the two pine species to 1.35 between *Pinus contorta* and *Sullavantia sullivantii* (Saxifragaceae). The rate of transversion substitutions must be influenced by other factors, possibly intrinsic genetic or external environmental. The likelihood of transversions was also found to be influenced by the GC content[28,20]. If this is true, then transversions, and subsequently nr/nv values, might not represent particularly conserved characters in phylogenetic studies, but are rather a product of an intrinsic nucleotide composition pattern that characterizes a lineage. This was true for some lineages of the grass family[20].

Best fit DNA model

All methods of phylogenetic inference depend on their underlying models. To have confidence in inferences it is necessary to have confidence in the models[29]. Because of this all methods based on explicit models of evolution should explore which is the model that fits the data best. Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. Table 8 provides details of the maximum Likelihood of 24 different nucleotide substitution models. The T92+G had the least BIC score and therefore considered the best model for the present data set. Another way of selecting the most appropriate model for a data set is to use the Akaike information criterion (AIC)[30] (Akaike 1974), which can be thought of as the amount of information lost when a particular model is used to approximate reality. The AIC implements best-fit model selection by calculating the likelihood of proposed models, and imposing a penalty based on the number of model parameters. Parameter-rich models incur a larger penalty than more simple models so that fitting an excessively complex model is not likely. The best fitting model is the one with the smallest AIC value. The T92+G had the smallest AIC value for the present data set.

Distance matrix

The number of base substitutions per site from between sequences are shown in table 9. Analyses were conducted using the Maximum Composite Likelihood model. The analysis involved 13 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 729 positions in the final dataset. Evolutionary analyses were conducted in MEGA5. The table reveals a low estimate of evolutionary divergence between the sequences among the 13 species of plants studied.

Evolulionary tree using matK gene sequence

The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura 3-parameter model. The tree with the highest log likelihood (-1434.9662) is shown in Fig 1. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.0804)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 13 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 729 positions in the final dataset. *T.capensis, T.aungustifolia* and *T.latifolia* share a common ancestor among themselves with *T. domingensis* being a closely related species. Again, all the species of *Sparganium* were grouped together in terms of their similar number of substitutions per site values and all the other remaining plant species formed a different clade. Graham *et al.*[31] performed a phylogenetic analysis of 17 plastid protein-coding

genes and included taxa from Poaceae, Typhaceae, and eight additional Poales families. They observed that the branch lengths for most members of the Poales were long except for the three earliest diverging families, including Typhaceae.

Fig 1. Molecular Phylogenetic analysis by Maximum Likelihood method

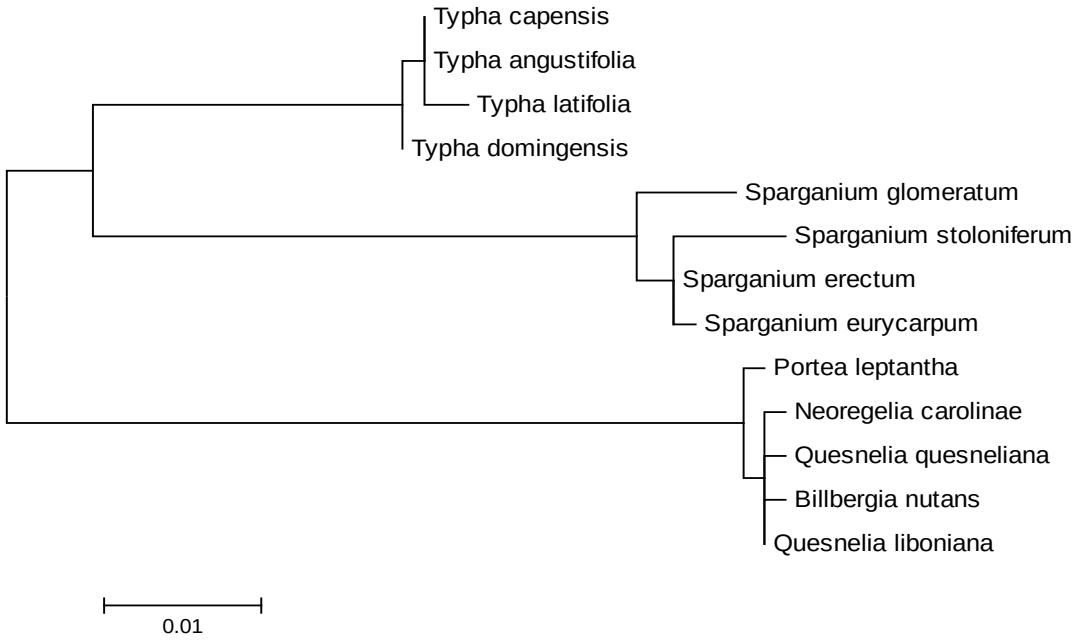

Table 1. matK gene sequence of *T. angustifolia*

1	AACTATATCT	CGGATATACT	AATACCCCAT	CCTATCCATT	TGGAAATCTT	GTTCAAACC
61	TTCAATGTCG	GATCCAAGAT	GTTCCATCTT	TGCATTTATT	GCGATTCTTT	CTCCACGAAT
121	ATCATAATTG	GAATAGTCTC	ATTACTTCGA	AGAAATCTAT	TTACGTTTTT	TCAAAAGAAA
181	ATAAAAGACT	ATTTAGATTA	CTATATAATT	TTTATGTATT	CGAATGTGAA	TTTGTATTCG
241	TTTTTCTTCG	TAAACAATCT	TCTTATTTAC	GATTAACATC	TTTTGGAACT	TTTCTTGAGC
301	GAATACATTT	CTATGGAAAA	ATAGAACATC	TTCTAGTAGT	GTATCGTAAT	TTTTTTAATA
361	AAACCTTATG	GTTCTTCACA	GATCCTTTCA	TGCATTATGT	TCGATATCAA	GGAAAAGCAA
421	TTCTGGCATC	AAAAGGGACT	CATCTTTTTA	TGAAGAAATG	GAAATGTTAC	CTTGTCAATT
481	TCTGGCAATA	TTATTTTCAT	TTTTGGTCTC	AACCGCACAG	GATCCATATA	AACCAATTAT
541	CAAACTATTC	CTTCCATTTT	CTGGGTTATC	TTTCAAGTTT	ACTAAGAAAT	CCTTTGGTGG
601	TAAGGAATCA	AATGCTGGAA	AATTCATATC	TAATAGATAC	TGTTATGACA	AAATTCGATA
661	CCATAGTACC	AGTTGATCCT	CTCATAGGAT	CATTGTCTAA	AGCTAAATTT	TGTACCCTAT
721	TAGGACATCC	TATTAGTAAG	CCGATCTGGA	CCGATTTATC	GGATTGTGAT	ATTATTGATC
781	GCTTTGGTCG	GATATGTAGA	AGTCTTTCTC	ATTATTATAG	TGGATCCTCA	AAAAAACGAA
841	CTTTGTATCG	ATACAGTATA	TACTTCGACT	TTCGTGGTG		

Table 2 Base statistics for the matK gene of T. angustifolia

Nucleotide	Count
T	328
C	152
A	269
G	130
G+C %	32.08
A+T %	67.91

Table 3. Top selected sequences providing significant alignment with the matk gene sequence of *T.angustifolia*

Acc. No.	Species	Aligned bases	Coverage	% Identity	Mismatc-hes	Gaps	unaligned base 5'	unaligned base 3'
HM850522.1	*Typha domingensis*	820	93.2	99.8	1	1	9	-
DQ009587.1	*Typha latifolia*	875	99.3	99.1	6	2	453	211
JQ435570.1	*Sparganium erectum*	876	99.2	95.2	38	4	444	217
HQ180886.1	*Sparganium eurycarpus*	876	99.2	95.2	38	4	522	208
AY952426.1	*Sparganium glomeratum*	876	99.2	94.9	41	4	1369	506
AB088802.1	*Sparganium stoloniferum*	876	99.2	93.8	50	4	447	229
AY950052.1	*Portea leptantha*	866	98.3	93.6	53	2	459	388
AY950049.1	*Billbergia nutans*	866	98.3	93.6	53	2	459	388
AY950048.1	*Quesnelia libonians*	866	98.3	93.6	53	2	459	388
HQ180877.1	*Neoregelia carolinae*	866	98.3	93.5	54	2	540	205
JF295095.1	*Quesnelia quesnelians*	866	98.3	93.5	54	2	449	246

Table 4 . Nucleotide frequencies in percentage for matK gene in selected species of Poales

	T(U)	C	A	G	Total	T-1	C-1	A-1	G-1	Pos #1	T-2	C-2	A-2	G-2	Pos #2	T-3	C-3	A-3	G-3	Pos #3
Typha latifolia	37.1	16.8	31.7	14.4	1539.0	43	11.3	32.7	13.1	513.0	32	22.2	28.7	16.8	513.0	36	16.8	33.7	13.5	513.0
Typha domingensis	37.9	17.2	30.5	14.4	829.0	43	12.7	31.2	13.0	276.0	33	23.5	27.8	15.9	277.0	38	15.6	32.6	14.1	276.0
Typha capensis	37.9	16.5	31.0	14.6	765.0	44	11.4	32.5	12.5	255.0	35	21.6	28.2	15.7	255.0	36	16.5	32.2	15.7	255.0
Sparganium erectum	37.0	16.6	31.7	14.6	1536.0	44	10.5	33.0	12.3	512.0	31	23.2	27.5	17.8	512.0	36	16.0	34.6	13.9	512.0
Sparganium eurycarpum	36.8	16.8	31.7	14.6	1605.0	43	11.4	33.0	12.7	536.0	31	22.8	28.0	17.8	535.0	36	16.3	34.1	13.5	534.0
Sparganium glomeratum	35.5	15.3	33.3	15.9	2749.0	39	12.3	34.6	14.2	917.0	32	18.5	31.7	17.7	917.0	35	15.1	33.6	16.0	915.0
Sparganium stoloniferum	36.5	16.6	31.6	15.3	1538.0	43	10.5	33.0	13.3	512.0	31	23.0	27.9	17.7	513.0	35	16.2	33.9	15.0	513.0
Portea leptantha	36.1	16.5	31.0	16.3	1713.0	40	12.1	31.7	15.8	571.0	33	21.5	27.8	18.0	571.0	35	15.9	33.5	15.2	571.0
Billbergia nutans	36.1	16.6	31.1	16.3	1713.0	40	12.1	31.9	15.6	571.0	32	21.7	27.8	18.0	571.0	35	15.9	33.5	15.2	571.0
Quesnelia liboniana	36.2	16.5	31.0	16.3	1713.0	41	11.7	31.9	15.6	571.0	32	21.7	27.7	18.2	571.0	35	15.9	33.5	15.2	571.0
Neoregelia carolinae	36.3	17.1	30.7	16.0	1611.0	41	12.1	31.3	15.6	537.0	32	22.5	26.8	18.6	537.0	36	16.8	33.9	13.6	537.0
Quesnelia quesneliana	36.6	16.8	30.6	16.0	1561.0	42	11.5	31.5	15.2	520.0	33	22.3	26.9	18.3	520.0	36	16.7	33.4	14.4	521.0
Typha angustifolia	37.3	17.3	30.6	14.8	879.0	42	12.3	31.8	13.4	292.0	33	23.8	27.9	15.6	294.0	37	15.7	32.1	15.4	293.0
Avg.	36.5	16.5	31.4	15.5	1519.3	42	11.7	32.5	14.2	506.4	32	21.9	28.3	17.6	506.6	36	16.1	33.5	14.7	506.3

Table 5 matK gene sequence alignment highlights for *T. angustifolia* data

Un-translated nucleotide	Highlights	Translated sequence
880/2768	**Conserved**	267/922
9/2768	**Variable**	6/922
None	**Parsim-Info**	None
5/2768	**Singleton**	2/922
1023/2768	**Zero fold**	--
340/2768	**Two fold**	--

Table 6 Relative Synonymous codon usage for matK gene

Codon	Count	RSCU	Codon	Count	RSCU	Codon	Count	RSCU	Codon	Count	RSCU
UUU(F)	27	1.44	UCU(S)	13.5	1.78	UAU(Y)	11.8	1.42	UGU(C)	6.5	1.73
UUC(F)	10.5	0.56	UCC(S)	6.3	0.82	UAC(Y)	4.8	0.58	UGC(C)	1	0.27
UUA(L)	10.3	2.18	UCA(S)	10	1.32	UAA(*)	12.3	1.56	UGA(*)	5.8	0.73
UUG(L)	4.5	0.96	UCG(S)	8.3	1.09	UAG(*)	5.5	0.7	UGG(W)	4.8	1
CUU(L)	6.8	1.43	CCU(P)	3	1.17	CAU(H)	4.5	1.57	CGU(R)	2	0.68
CUC(L)	2.3	0.48	CCC(P)	0.3	0.1	CAC(H)	1.3	0.43	CGC(R)	0	0
CUA(L)	2.5	0.53	CCA(P)	6.8	2.63	CAA(Q)	2.5	1.43	CGA(R)	4.5	1.52
CUG(L)	2	0.42	CCG(P)	0.3	0.1	CAG(Q)	1	0.57	CGG(R)	0	0
AUU(I)	16.3	1.51	ACU(T)	4.8	1.04	AAU(N)	8.8	1.32	AGU(S)	5.5	0.73
AUC(I)	10.5	0.98	ACC(T)	3.3	0.71	AAC(N)	4.5	0.68	AGC(S)	2	0.26
AUA(I)	5.5	0.51	ACA(T)	7.8	1.7	AAA(K)	13.5	1.42	AGA(R)	7	2.37
AUG(M)	5.8	1	ACG(T)	2.5	0.55	AAG(K)	5.5	0.58	AGG(R)	4.3	1.44
GUU(V)	1	0.64	GCU(A)	1.5	0.65	GAU(D)	5.3	1.24	GGU(G)	3	1.14
GUC(V)	2	1.28	GCC(A)	1	0.43	GAC(D)	3.3	0.76	GGC(G)	1.3	0.48
GUA(V)	3	1.92	GCA(A)	4.5	1.95	GAA(E)	8.5	1.94	GGA(G)	5	1.9
GUG(V)	0.3	0.16	GCG(A)	2.3	0.97	GAG(E)	0.3	0.06	GGG(G)	1.3	0.48

Table 7 Nucleotide pair frequencies among Typha species.

	ii	**si**	**sv**	**R**	**TT**	**TC**	**TA**	**TG**	**CC**	**CA**	**CG**	**AA**	**AG**	**GG**	**Total**
Avg	796	2	1	1.25	302	0	0	1	134	1	0	245	2	115	798.83
1st	265	0	1	0.60	115	0	0	1	32.	0	0	84	0	34	266.00
2nd	266	0	0	1.00	89	0	0	0	60	0	0	75	0	42	266.83
3rd	265	1	0	3.00	98	0	0	0	42	0	0	86	1	39	266.00

ii = Identical Pairs; si = Transitionsal Pairs; sv = Transversional Pairs; R = si/sv; 1st, 2nd, 3rd Codon position TC AG -Transition; TA TG CA CG - Transversion, TT,CC,AA,GG- Identical pairs

Table. 8 Maximum Likelihood fits of 24 different nucleotide substitution models

Model	**BIC**	**AICc**	**InL**	**Invariant**	**Gamma**	**R**	***f*A**	***fT***	***fC***	***f*G**
T92+G	3108.1	2922.1	-1435	n/a	0.080	1.552	0.341	0.34	0.158	0.16
T92+G+I	3116.4	2923.3	-1434.6	0.659	0.924	1.572	0.341	0.34	0.158	0.15
HKY+G	3122.1	2921.9	-1432.9	n/a	0.082	1.550	0.307	0.37	0.166	0.14
T92	3123.8	2945	-1447.4	n/a	n/a	1.396	0.341	0.34	0.158	0.15
TN93+G	3129.1	2921.8	-1431.8	n/a	0.0690	1.566	0.307	0.37	0.166	0.14
HKY+G+I	3130.4	2923.1	-1432.4	0.656	0.9027	1.574	0.307	0.37	0.166	0.14
T92+I	3131.3	2945.4	-1446.6	0.067	n/a	1.399	0.342	0.34	0.158	0.15
TN93+G+I	3137.6	2923.1	-1431.4	0.677	1.0262	1.583	0.308	0.37	0.166	0.14
HKY	3137.8	2944.7	-1445.3	n/a	n/a	1.395	0.308	0.38	0.166	0.15
HKY+I	3141.3	2941.1	-1442.5	0.223	n/a	1.406	0.308	0.38	0.166	0.15
GTR+G	3144.8	2916	-1425.9	n/a	0.0807	1.254	0.308	0.38	0.166	0.15
TN93	3145.5	2945.3	-1444.6	n/a	n/a	1.39	0.308	0.38	0.166	0.15
TN93+I	3152	2944.6	-1443.2	0.110	n/a	1.395	0.308	0.38	0.166	0.15
GTR+G+I	3153.3	2917.4	-1425.6	0.678	1.1552	1.258	0.307	0.38	0.166	0.15
GTR	3160.2	2938.5	-1438.2	n/a	n/a	1.395	0.307	0.38	0.166	0.15
GTR+I	3167.3	2938.5	-1437.1	0.083	n/a	1.399	0.307	0.38	0.166	0.15
K2+G	3202.9	3024.1	-1487	n/a	0.0613	1.528	0.25	0.25	0.25	0.25
K2+G+I	3212	3026	-1486.9	0.050	0.0649	1.539	0.25	0.25	0.25	0.25
JC+G	3214.4	3042.7	-1497.3	n/a	0.0745	0.5	0.25	0.25	0.25	0.25
K2	3219.3	3047.6	-1499.7	n/a	n/a	1.383	0.25	0.25	0.25	0.25
JC+G+I	3222.7	3043.9	-1496.9	0.641	0.7926	0.5	0.25	0.25	0.25	0.25
K2+I	3225.3	3046.6	-1498.2	0.127	n/a	1.388	0.25	0.25	0.25	0.25
JC	3229.9	3065.4	-1509.6	n/a	n/a	0.5	0.25	0.25	0.25	0.25
JC+I	3239	3067.4	-1509.6	0.00001	n/a	0.5	0.25	0.25	0.25	0.25

Abbreviations: GTR: General Time Reversible; HKY: Hasegawa-Kishino-Yano; TN93: Tamura-Nei; T92: Tamura 3-parameter; K2: Kimura 2-parameter; JC: Jukes-Cantor, G: Gamma distribution; I: evolutionary invariable

Table 9. Estimates of Evolutionary divergence between sequences

	Species	1	2	3	4	5	6	7	8	9	10	11	12
1	Typha_latifolia												
2	Typha_domingensis	0.004											
3	Typha_capensis	0.003	0.001										
4	Sparganium_erectum	0.050	0.048	0.047									
5	Sparganium_eurycarpum	0.051	0.050	0.048	0.001								
6	Sparganium_glomeratum	0.051	0.053	0.051	0.008	0.010							
7	Sparganium_stoloniferum	0.051	0.051	0.050	0.007	0.008	0.013						
8	Portea_leptantha	0.066	0.063	0.064	0.077	0.079	0.077	0.082					
9	Billbergia_nutans	0.066	0.063	0.064	0.079	0.080	0.078	0.084	0.004				
10	Quesnelia_liboniana	0.066	0.063	0.064	0.077	0.079	0.077	0.082	0.003	0.001			
11	Neoregelia_carolinae	0.067	0.064	0.066	0.075	0.077	0.075	0.081	0.004	0.003	0.001		
12	Quesnelia_quesneliana	0.067	0.064	0.066	0.075	0.077	0.075	0.081	0.004	0.003	0.001	0.003	
13	Typha_angustifolia	0.003	0.001	0.000	0.047	0.048	0.051	0.050	0.064	0.064	0.064	0.066	0.066

CONCLUSION

The present study indicates *mat*K gene to be a useful marker in identifying *Typha angustifolia* and could be used for delimiting the species from other members of *Typha.* Also this gene appears useful for separating it from other genera of Poales. For studying the phylogeny of the group, it is suggest that genomic changes and accelerated rates of sequence evolution may not be limited to the Poaceae only, and that a positive correlation between these two phenomena can be shown for lineages leading to the Poaceae. We have here only taken the *mat*k gene sequence for comparative analysis and therefore emphasize that complete genome sequences of a number of Poales are needed to fully understand the evolution of Poales and Poaceae.

REFERENCES

1. Hector,A., Schmid,B., Beierkuhnlein,C., Caldeira,M.C., Diemer,M., Dimitrakopoulos, P.G., Finn, J.A., Freitas,H., Giller,P.S., Good,J., Harris,R., Högberg,P., Huss-Danell, K., Joshi, J. Jumpponen,A., Körner,C., Leadley,P.W., Loreau,M., Minns,A., Mulder,C.P.H., O'Donovan,G., Otway,D.S.J., Pereira,J.S., Prinz,A., Read,D.J., Scherer-Lorenzen,M., Schulze, E.D., Siamantziouras, A.S.D., Spehn, E.M. Terry,A.C., Troumbis,A.Y., Woodward,F.I., Yachi,S., and Lawton, J.H. Plant Diversity and Productivity Experiments in European Grasslands. *Science* **286 (5442):** 1123-1127 (1999)
2. Hooper, D U and Votousek, P.M. The effects of plant composition and diersity on ecosystem processes. *Science,* **277:** 1302-1305 (1997)
3. Jackson, R. B., Carpenter,S.R., Dahm,C.N., McKnight,D.M., Naiman,R.J., Postel,S.L., and Running, S.W. Water in a changing world. *Ecological Applications* **11**:1027-1045 (2001)
4. Alvarez, I. and Wendel, J.F. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* **29**: 417–434 (2003)
5. Cho J-I, Lee S-K, Ko S, Kim H-K, Jun S-H, Lee Y-H, Bhoo SH, Lee K-W, An G, Hahn T-R, Jeon J-S Molecular cloning and expression analysis of the cell-wall invertase gene family in rice (Oryza sativaL.). *Plant Cell Rep* **24**:225–236 (2005)
6. Shaw, J., Lickey,E., Beck, J.T., Farmer,S.B., Liu,W., Miller,J., Siripun,K.C., Inder,C.T.W., Schilling, E.E., and Mall, R.L.S. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* **92**: 142–166 (2005)
7. Shaw J, Lickey,E.B., Edward E.S., Chilling and Randall L. Small Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the Hare. *American Journal of Botany* **94(3):** 275–288 (2007)

8. Givnish T. J., Millam,K.C., Berry, P.E., and Sytsma,K.J. Phylogeny, adaptive radiation, and historical biogeography of Bromeliaceae inferred from *ndh*F sequence data. *Aliso* **23:** 3-26 (2007)
9. Notredame, C., Higgins, D.G., and Heringa, J. T-Coffe: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302 (1):**205-217 (2000)
10. Takundwa M., Chimwamurombe P. M., Kunert K., and Cullis C. A. Developing DNA barcoding (matK) primers for marama bean [*Tylosema esculentum* (Burchell) Schreiber]. *African Journal of Biotechnology* **11(97):** 16305-16312 (2012)
11. Sugita, M., Shinozaiki,K., and Sugiura, M. Tobacco chloroplast tRNA Lys (UUU) gene contains a 2.5-kilobase-pair intron: an open reading frame and a conserved boundary sequence in the intron. *Proceedings of the National Academy of Sciences, USA* **82:** 3557-3561 (1985)
12. Neuhaus, H and G. Link, The chloroplast tRNA LYS (UUU) gene from mustard (Sinapis alba) contains a class II intron potentially coding for a maturase-related polypeptide. *Current Genetics* **11:** 251-257 (1987)
13. Mohr, G., P.S. Perlman and A.M. Lambowitz. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acid Research* **21**: 4991-4997 (1993)
14. Ems, S.C., Morden, C.W., D.K.Dixon, K.H.Wolfe, C.W. Depamphilis and J.D. Palmer. Transcription, splicing and editing of plastid RNAs in the nonphotosynthetic plant *Epifagus virginiana*. *Plant Molecular Biology*. **29:** 721-733 (1995)
15. Khidir,W. H. and Hongping,L. The *mat*K gene: sequence variation and application in plant systematics. *Am. J. Bot.* **19**:830-839 (1997)
16. Selvaraj D., Rajeev Kumar Sarma and Ramalingam Sathishkumar. Phylogenetic analysis of chloroplast *matK* gene from Zingiberaceae for plant DNA barcoding. *Bioinformation* **3(1):** 24-27 (2008)
17. Doyle, J. J. and Doyle,J.L. A rapid total DNA preparation procedure for fresh plant tissue. *Focus* **12:**13-15 (1990)
18. Tamura K, Daniel Peterson, Nicholas Peterson, Glen Stecher, Masatoshi Nei, and Sudhir Kumar MEGA5. Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum, Parsimony Methods. *Mol. Biol.Evol.* **28(10):** 2731-2739 (2011)
19. Guisinger, M.M., Chumley,T.W., Kuehl,J.V., Boore,J.L. and Jansen, R.K. Implications of the Plastid Genome Sequence of Typha (Typhaceae, Poales) for Understanding Genome Evolution in Poaceae. *J Mol Evol.* **70**:149–166 (2010)
20. Liang, H and Hilu,K.W. Application of the matK gene sequences to grass systematics. *Canadian Journal of Botany* **74:** 125-134 (1996)
21. Johnson, L A. and Soltis, D.E. Phylogenetic inference in Saxifragaceae sensu stricto and Gilia (Polemoniaceae) using matK sequences. *Annals of the Mossouri Botanic Garden* **82:** 149-175 (1995)
22. McInerney, J.O. GCUA (General Codon Usage Analysis). *Bioinformatics:* **14(4):** 372-373 (1998)
23. Sanjayan, K.P and Rama, V. Molecular characterization of two species of *Spilostethus* (Insecta: Lygaeidae) and phylogenetic reconstruction using 16S ribosome RNA gene fragment. *Int. J Plant, Animal and Environmental Sciences.* **3(2)** 235-243 (2013)
24. Vigilant, L., Stoneking, M., Harpending,H., Hawkes,K., and Wilson,A.C. African populations and the evolution of human mitochondria] DNA. *Science,* **253:**1503-1507 (1991)
25. Quicke, D.L. J. Principle and techniques of contemporary taxonomy. Chapman & Hall, Glasgow, UK (1993)
26. Lake, J.A. A rate-independent technique for analysis of nucleic acid sequences : evolutionary parsimony. *Molecular Biology and Evolution* **4:** 167-191 (1987)

27. Holmquist, R. Transitions and transversion in evolutionary descent: an approach to understanding. *Journal of Molecular Evolution* **19:** 134-144 (1983)
28. Mortan, B.R. Neighboring base composition and transversion /transition bias in a comparison of rice and maize chloroplast nondcoding regions. *Proceedings of the National Academy of Sciences, USA* **92**: 9717-9721 (1995)
29. Goldman N Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36:**182-198 (1993)
30. Akaike, H. A new look at the statistical model identification. *IEEE Trans Autom Contr*. **19:** 716-723 (1974)
31. Graham S.W., Zgurski J.M., McPherson, M.A, Cherniawsky, D.M., Saarela, J.M., Horne, E.S.C., Smith S.Y, Wong, W.A, O'Brien, H.E., Biron, V.L., Pires, J.C., Olmstead, R.G., Chase, M.W., Rai, H.S. Robust inference of monocot deep phylogeny using an expanded multigene plastid data set. *Aliso* **22**:3–20 (2006)