Plant Mol Biol Rep
DOI 10.1007/s11105-015-0878-6

ORIGINAL PAPER

Comparative Transcriptome Analysis of Two Species of *Curcuma* Contrasting in a High-Value Compound Curcumin: Insights into Genetic Basis and Regulation of Biosynthesis

T. E. Sheeja[1] · K. Deepa[1] · R. Santhi[1] · B. Sasikumar[1]

Abstract Turmeric (*Curcuma longa* L), one of the widely used spices and herbal medicines, is rich in biologically active compound curcumin. Efficacy of curcumin in treating various diseases has been established through over 1000 published in vivo and in vitro studies. Curcumin content is reported to vary between accessions and subject to agro-climatic conditions, and the reasons are unexplored. Gaining an understanding on the molecular mechanism underlying curcumin biosynthesis will help in improving its content and maintaining stability under all conditions of cultivation. To investigate the genes involved in curcuminoid biosynthesis, we used Illumina sequencing platform and generated a substantial amount of expressed sequence tag (EST) dataset from two species viz., *C. longa* and its wild relative *Curcuma aromatica* Salisb. contrasting in curcumin content. The data reads of Illumina sequencing have been deposited in the National Center for Biotechnology Information Sequence Read Archive under BioProject ID PRJNA270561 and PRJNA277549 for *C. longa* and *C. aromatica*, respectively. De novo assembly produced a total of 99,482 and 104,514 contigs with an average length of 367 and 359 bp for *C. longa* and *C. aromatica*, respectively. The transcripts were assembled by BLAST similarity searches and annotated with Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology identifiers. Bioinformatics analysis using BLASTX revealed that both the transcriptomes contained all the ten genes putatively involved in curcuminoid biosynthesis. Two novel polyketide synthase genes showing similarity to that of *Musa accuminata* (*clpks1* and *clpks2*) were found to be up-regulated in *C. longa* in comparison with *C. aromatica*. Transcription factors with putative roles in phenylpropanoid biosynthetic pathway were also reported for the first time in *Curcuma* spp. We also identified 5488 and 5620 putative simple sequence repeats (SSRs) for *C. longa* and *C. aromatica*, respectively. By using the assemblies as reference, 190 single-nucleotide polymorphisms (SNPs) for *C. longa* and 108 SNPs for *C. aromatica* in the candidate genes of curcuminoid biosynthetic pathway were identified. We could also detect 68 and 64 transcripts as microRNA (miRNA) targets from *C. longa* and *C. aromatica*, respectively. This study generated an extensive transcriptome data of *C. longa* and *C. aromatica* using deep sequencing. The candidate genes for enzymes involved in curcuminoid biosynthesis were identified from both the species. Differentially expressed genes, novel polyketide synthases, transcription factors, SSRs, SNPs and miRNA targets were identified from the transcriptomes. These EST sequences are a valuable public information platform for functional studies in turmeric and form a rich resource for studies on marker development and turmeric breeding.

Keywords Curcumin · Transcriptome · Curcuma · SSR · SNP · miRNA · Gene expression

Introduction

Curcuma longa L. (Turmeric), a rhizomatous herbaceous perennial plant belonging to the Zingiberaceae family, is a native to Southeast Asia and is now widely cultivated in the tropical and subtropical regions of the world. The use of turmeric dates

T. E. Sheeja holds a Ph.D. degree, Indian Institute of Spices Research.
K. Deepa holds a M. Sc degree, Indian Institute of Spices Research.
R. Santhi holds a M. Sc degree, Indian Institute of Spices Research.
B. Sasikumar holds a Ph.D. degree, Indian Institute of Spices Research.

✉ T. E. Sheeja
sheeja@spices.res.in

[1] Division of Crop Improvement and Biotechnology, Indian Institute of Spices Research, Calicut, Kerala 673 012, India

Published online: 26 March 2015

back to nearly 4000 years to the Vedic culture in India, where it was used as a culinary spice and also had some religious significance. Because of its brilliant yellow colour, turmeric is also known as 'Indian saffron'. Due to its inherent qualities and high content of the important bioactive compound curcumin, Indian turmeric is considered to be best in the world. India is a leading producer and exporter (approximately 90 %) of turmeric in the world.

Turmeric powder obtained from rhizomes of *C. longa* is extensively used as a spice, food preservative, natural dye in food industry and in cosmetics and drugs (Sasikumar 2005; Krup et al. 2013). Ayurveda, Unani, Siddha and Chinese medicines recommend turmeric in treatment for a large number of disorders and diseases (Prasad and Aggarwal 2011). Curcuminoid, a phenylpropanoid derivative, is a mixture of curcumin (50–60 % of the curcuminoids), demethoxycurcumin and bisdemethoxycurcumin that imparts yellow colour to turmeric (Elizabeth et al. 2011). The medicinal properties of curcuminoids such as anti-inflammatory, anti-oxidant, anti-mutagenic, anti-diabetic, anti-bacterial, hepatoprotective and expectorant are reported extensively. It is also well known in treating conditions ranging from arthritis and inflammation to Alzheimer's disease and cancer (Krup et al. 2013).

Studies in the past have indicated that curcumin content varies within accessions of *C. longa* (Elizabeth et al. 2011) and from place to place due to the influence of environment and agro-climatic conditions (Anandaraj et al. 2014; Singh et al. 2013), which is a major concern of the spice industry. This observed difference in curcumin content is mainly due to the expression levels of genes encoding important enzymes of the pathway (Katsuyama et al. 2009a). An understanding of the genes involved in the biosynthesis of curcuminoids is of significance in this context.

The putative curcuminoid biosynthetic pathway in *C. longa* has been unravelled (Fig. 1) based on previous studies of Schröder (1997), del Ramirez-Ahumada et al. (2006), Koo et al. (2013) and Katsuyama et al. (2009b). The transcriptome of rhizome of *C. longa* L. has already been sequenced using Illumina platform to reveal the novel transcripts related to anti-cancer and anti-malarial terpenoids (Annadurai et al. 2013). However, molecular mechanism underlying curcuminoid metabolism in turmeric still requires elucidation, especially with respect to the full set of genes involved in this pathway and transcription factors regulating these genes. Mining of genes of the biosynthetic pathway has been done in several crops, and next-generation sequencing (NGS) has emerged as a cost-effective method for the detection and quantification of genes and low-abundant transcripts involved in specific biological processes (Vaidya et al. 2013; Kalra et al. 2013). Identification of factors regulating the gene expression also assumes importance, and it is reported that transcription factors play a major role in regulating almost all biological processes by regulating mRNA synthesis. Several studies have proved that overexpression of these transcription factors in transgenic plants could regulate the expression of phenylpropanoid pathway genes (Omer et al. 2013; Huang et al. 2012b). In the present study, we applied paired-end Illumina GAII sequencing to non-normalized complementary DNAs (cDNAs) of *C. longa* L. and *Curcuma aromatica* Salisb. for a comprehensive characterization of the de novo transcriptome. Important candidate genes and potential transcription factors involved in curcuminoid biosynthetic pathway were identified, and their expression patterns were validated. Another objective of our study was to detect a large number of informative single-nucleotide polymorphisms (SNPs), simple sequence repeats (SSRs) and microRNA (miRNA) targets as resource for marker development in this non-model crop.

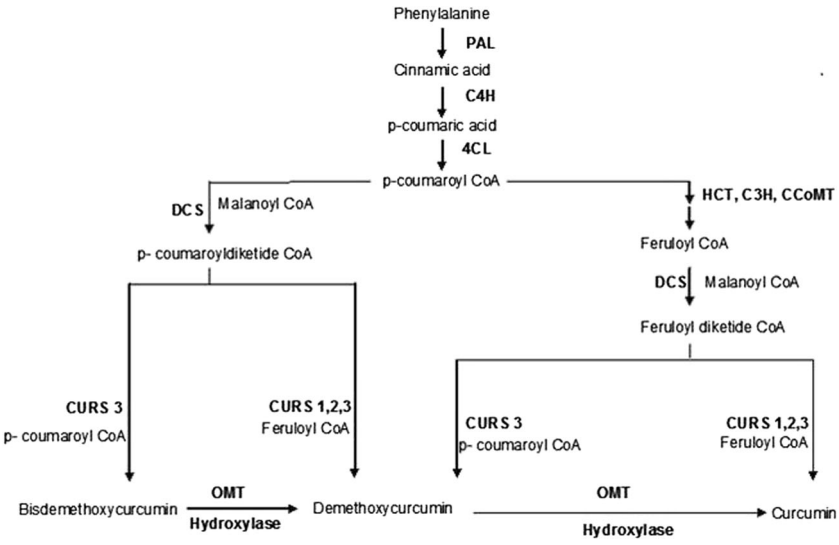

Fig. 1 Biosynthetic pathway of curcuminoids in turmeric. *PAL* phenylalanine ammonia lyase, *C4H* cinnamate 4-hydroxylase, *4CL* 4-coumarate-CoA ligase, *HCT* hydroxycinnamoyl CoA:shikimate/quinate hydroxycinnamoyltransferase, *C3H p*-coumarate 3-hydroxylase, *CCOMT* caffeoyl CoA O-methyltransferase, *DCS* diketide CoA synthase, *CURS* curumin synthase, *OMT* O-methyl transferase

Materials and Methods

Plant Material

The rhizomes from 4-month-old plants of a high-curcumin-containing *C. longa* accession Mega turmeric and a closely related species *C. aromatica* (wild turmeric) which is practically devoid of curcumin were collected from Indian Institute of Spices Research (IISR) Experimental Farm, Peruvannamuzhi, Kozhikode, Kerala.

cDNA Library Construction and Illumina Sequencing

Total RNA from rhizomes of 4-month-old turmeric variety Mega turmeric and *C. aromatica* was isolated to acquire a high number of candidate genes involved in curcuminoid biosynthesis according to the method described by Ghawana et al. (2011). The purity of total RNA was analyzed with Agilent 2100 Bioanalyzer, and the RNA integrity number (RIN) value was found to be 8.2 and 6.5 for *C. longa* and *C. aromatica*, respectively. cDNA library construction and sequencing was carried out at Sandor Proteomics, Hyderabad. Transcriptome library was constructed using Illumina's TrueSeq RNA sample preparation kit as per the manufacturers' instructions and was subjected to sequencing on Illumina HiSeq 2000 platform.

Transcript Assembly, Annotation and Differential Gene Expression Analysis

The bioinformatics analysis was carried out at SciGenom Labs Pvt Ltd, Cochin, Kerala. The fastq files were trimmed before performing assembly to avoid sequence-specific bias and low-quality bases. The reads having average quality score less than 20 in any of the paired end and those contaminated with Illumina adapter were filtered out. The trimmed reads were then assembled individually using SOAPdenovo31mer algorithm with default options. The transcripts whose length ≥150 bp were only used for the estimation of gene expression using Bowtie2 parser program, and the annotation was given to only those transcripts having fragments per kilobase of exon model per million fragments mapped (FPKM) ≥1. Differential gene expression analysis was performed using DESeq program. The assembled transcripts were annotated using in-house pipeline programs for de novo transcriptome assembly. The assembled transcripts were compared with National Center for Biotechnology Information (NCBI) non-redundant protein database using BLASTX program. Matches with *E*-value $\leq 10^{-5}$ and similarity score ≥40 % were retained for further annotation. The predicted proteins from BLASTX were annotated against UniProt databases. With nr annotations, an in-house program was performed to categorize and classify the unigenes to GO terms such as molecular function, biological processes and cellular components. KEGG annotation was carried out using an automatic server KEGG Automatic Annotation Server (KASS), which assigned KEGG pathway terms to the transcripts. The transcripts were also aligned against *Arabidopsis* Gene Regulatory Information Server (AGRIS) database to identify transcripts encoding transcription factors in all unigenes (identity >80 %). Sequencing data have been deposited in the NCBI Sequence Read Archive [http://www.ncbi.nlm.nih.gov/sra] under BioProject ID PRJNA270561 and PRJNA277549 for *C. longa* and *C. aromatica*, respectively.

Gene Validation and Expression Analysis

Total RNA was isolated from 4-month-old rhizomes of turmeric and *C. aromatica* as per the protocol developed later on in our laboratory (Deepa et al. 2014). Approximately, 1 μg of total RNA was treated with 1 U of RNase-free DNase I (Thermo Scientific) and converted to first-strand cDNA in the presence of Oligo-$(dT)_{18}$ primer and RevertAid Reverse transcriptase (Thermo Scientific) as described in supplier's instruction.

The quality of sequence data was checked by amplifying full-length coding sequence of curcumin synthase 3 (*curs3*), an important downstream multifunctional enzyme using gene-specific primers (Katsuyama et al. 2009a) from both the species. The PCR mixture was initially denatured at 94 °C for 2 min and then subjected to 30 cycles with the following conditions: 94 °C for 45 s, 62 °C for 1 min and 72 °C for 2 min with a final extension at 72 °C for 10 min. The amplicons were purified, ligated in pGEMT-vector, cloned in *E. coli* JM109 and sequenced using the Sanger method.

Based on differential gene expression analysis, 14 unigenes which were up-regulated in *C. longa* were chosen for validation using real-time qPCR. This includes ten unigenes with potential roles in curcuminoid biosynthetic pathway (*pal*, *c4h*, *4 cl*, *hct*, *c3h*, *comt*, *dcs*, *curs1*, *curs2* and *curs3*), two transcription factors (*myb4* and *wrky*) and two polyketide synthase genes showing similarity with that of *Musa accuminata* (*clpks2* and *clpks4*). The list of primers is given in Table 1. Quantitative PCR was carried out with three technical replicates with QuantiFast SYBR Green PCR kit (Qiagen) on Rotor-Gene Q (Qiagen). The reaction mixture (20 μl) contained 2× QuantiFast SYBR Green PCR Master Mix (10 μl), 1 μM each of the forward and the reverse primers and tenfold diluted cDNA template. PCR amplification was performed under the following conditions: 95 °C for 10 s, followed by 40 cycles of 95 °C for 10 s and 60 °C for 45 s. A melt curve program of 55—99 °C was included to check the specificity of PCR products. Triplicates of each reaction were performed, and the gene expression of all genes was normalized against an internal reference gene, elongation factor 1α (*ef1α*). A reverse transcription negative control (without reverse transcriptase) and a non-template negative control were

Table 1 Gene-specific primers used for gene expression analysis by quantitative real-time PCR

Gene	Forward primer (5′–3′)	Reverse primer (5′–3′)
pal	GTACAGCGGGTACGACCTA	GCTTTCGAGGAAGAGCTCAA
c4h	TTACTTGCAGGCGGTGATC	AGGCGTTGACCAGTATCTTG
4 cl	GTGGATTCTTCAGGGAAGAGAG	GGCGATTGTCGATGTGAATG
c3h	GATGGTCACCTTCATGCATACT	GGAAGAAGCTGCAGGAGTAATAG
hct	ATCGGGAAGCGGTTGGAG	CTGCTTCCCGCTAGGCAAC
ccomt	TGCCGCAGGTCAAATG	ACCAGAAACTCAACTGAAAGGG
dcs	AAGTCGGGATAAGCGTTCTG	CGTCGTTTCTGTGACCTTCT
curs1	TCAGCTCATCCATCACGAAGTACAC	CATCATTGACGCCATCGAAGC
curs2	TGTTGCCGAACTCGGAGAAGAC	TCGGGATCAAGGACTGGAACAAC
curs3	CCCATTCCTTGATCCCTTTTCC	TGGAGCCCTCCTTCGACGACC
clpks1	TTCCTCACGTCGTCCATCAC	AGGCACGTCTTCAGCGAGTT
clpks2	GCCGACGTTGTAGAGCATCA	TCACCCACCTCGTATTCAGC
myb4	CAAGGGAGTTTGGACCAAGG	AGCAAGCCAGCAGCTTTAGG
wrky	GGCTATCAGGCAGGTTCAGG	TTTGAGGAACCAAGGGAGGA
ef1α	GCTGACTGTGCTGTTCTCATTAT	CTCGTGTCTGTCCATCCTTTGAA

also run for each primer master mix. The relative gene expression was calculated using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen 2001) with the expression level of gene in *C. aromatica* set to '1'.

SSR and SNP Identification

MISA tool (http://pgrc.ipk-gatersleben.de/misa/) was used to identify SSRs. The minimum repeat number was six for dinucleotide and five for trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide, and the maximal distance interrupting two SSRs in a compound microsatellite was 100 bp. SNPs in the candidate genes of curcuminoid pathway between *C. longa* and *C. aromatica* were identified using SAMtools and BCFtools.

Identification of Target Site for miRNAs

The potential target binding site for miRNAs was identified using the psRNATarget (http://plantgrn.noble.org/psRNATarget/) program with default parameters. In order to reduce the rate of false-positive target predictions, we used stringent criteria as described previously (Allen et al. 2005; Schwab et al. 2005). In brief, (1) no more than four mismatches between miRNA and target, (2) no more than two adjacent mismatches in the miRNA/target duplex, (3) no adjacent mismatches in positions 2–12 of the miRNA/target duplex starting from the 5′ end of miRNA, (4) no mismatches in positions 10–11 of miRNA/target duplex, (5) no more than 2.5 mismatches in positions 1–12 of the miRNA/target duplex starting from the 5′ end of miRNA and (6) minimum free energy (MFE) of the miRNA/target duplex should be>= 75 % of the MFE of the miRNA bound to its perfect compliment.

Results and Discussion

Sequencing, De Novo Assembly and Transcript Annotation

In the present study, rhizome-specific transcriptomes from *C. longa* and *C. aromatica* were constructed on Illumina HiSeq 2000 sequencing platform as an invaluable resource for understanding the molecular mechanism underlying curcuminoid biosynthesis. Pair-end (PE) read sequencing with Illumina platform has been reported to not only increase the depth of sequencing but also improve de novo assembly efficiency (Huang et al. 2012a). The tissue-specific transcriptome will serve as a good reference data for gene expression analysis, especially in non-model plants (Zhou et al. 2012). We have chosen two related genotypes contrasting in curcumin content for the study, to get a better insight into the biosynthetic mechanism of curcumin synthesis. A total of 3.30- and 3.71-Gb paired-end short reads of 100 nt in length were generated from rhizomes of *C. longa* and *C. aromatica*, respectively. After removing the adapter and low-quality sequences from the raw data, 2.87-Gb (86 bp×2) and 3.16-Gb (85 bp×2) high-quality reads were obtained for C. *longa* and *C. aromatica*, respectively. The average GC content for *C. longa* and *C. aromatica* sequences is 46.24 and 46.95 %, respectively, showing that the transcripts were marginally AT rich, which is comparable with the transcriptome of other plants (Annadurai et al. 2013; Garg et al. 2011). The transcriptome assembly result is summarized below in Table 2.

Table 2 Summary of *de novo* transcriptome assembly

Descriptions	*C. longa*	*C. aromatica*
Number of paired-end reads	33,384,838	37,135,009
Number of transcripts ≥150 bp	99,482	104,514
N50 size (in bases)	424	410
Longest transcript length (in bases)	7490	5145
Minimum transcript length (in bases)	150	150
Average transcript length (in bases)	366.65	358.60
GC %	49.39	49.68

After assembly, 343,144 transcripts and 381,377 transcripts were identified for *C. longa* and *C. aromatica*, respectively. Among the assembled transcripts, 28.99 % of *C. longa* and 26.4 % of *C. aromatica* have length ≥150 bp. Contigs with lengths between 200 and 500 bp were overrepresented, in both *C. longa* and *C. aromatica* (Fig. 2). The trimmed high-quality reads were aligned back to the assembled transcriptome to get its expression value (FPKM) distribution. We obtained 75,468 unique transcripts for *C. longa* and 79,728 unique transcripts for *C. aromatica* having expression ≥1 FPKM. Based on expression levels, the transcripts were divided into seven groups (Table 3).

In BLASTX homology search, only 42,310 transcripts (56.06 %) of *C. longa* and 45,651 transcripts (57.26 %) of *C. aromatica* gave hits; thus, nearly half of the transcriptome remains unmappable. Previous report on turmeric transcriptome study by Annadurai et al. (2013) utilized six different databases (Genbank-NT, KOG, PlantCyc, Swiss-Prot, TrEMBL and Pfam) and could annotate 54.6 % transcripts only. This may be due to lack of genome information, poor quality reads and sequencing artefacts. De novo transcriptome assembly of other non-model plants also reported many transcripts without BLAST hit (Iorizzo et al. 2011). These sequences without a homologous hit may represent novel genes specifically expressed in turmeric rhizome, or they could be attributed to other technical or biological biases such as

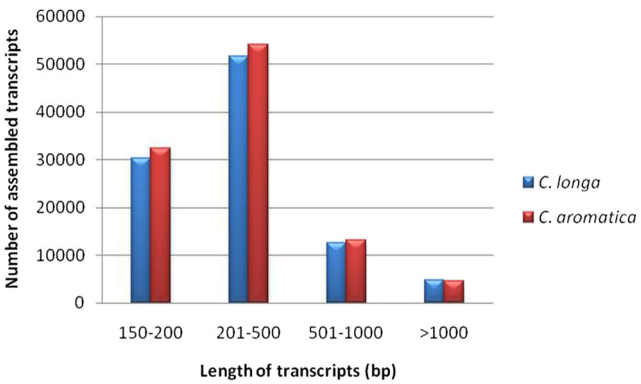

Fig. 2 Length distributions of assembled transcripts of two *Curcuma* species

Table 3 Transcript expression and total number of transcripts (length ≥150 bp)

	FPKM *C.longa*	Number of transcripts *C. aromatica*
1.0–2.0	15,300	15,702
2.0–5.0	16,753	18,390
5.0–10.0	10,157	11,298
10.0–20.0	6259	7392
20.0–50.0	4094	4797
50.0–100.0	1330	1520
≥100.0	1035	1120
Total	54,928	60,219

assembly parameters. It might be also because some of the cDNAs are non-coding, lineage-specific or highly variable (Logacheva et al. 2011; Ferreira de Carvalho et al. 2013). For both *C. longa* and *C. aromatica*, the most frequent organisms in the BLAST top hits were *Vitis vinifera*, *Zea mays* and *Oryza sativa* (Fig. 3). Among the total significant BLASTX hit transcripts, 19,583 transcripts (25.95 %) and 21,116 transcripts (26.49 %) were annotated using UniProt database for *C. longa* and *C. aromatica*, respectively.

The Gene Ontology (GO) terms were assigned to 42,310 transcripts for *C. longa* and 45,651 transcripts for *C. aromatica* and were classified into three main GO categories: biological processes, molecular functions and cellular components. The terms ATP binding, metabolic process and integral component of membrane represented the most represented for molecular function, biological process and cellular component categories, respectively, for both the species. The top 25 terms in each category for both the species are shown in Fig. 4. In both the species, the maximum transcripts fell into the category of metabolic process, suggesting the presence of novel genes involved in the secondary metabolite synthesis pathways. All the annotated contigs were compared against the KEGG database with default bit score and expected threshold. It assigned EC number to 1443 transcripts of *C. longa* and 1541 transcripts of *C. aromatica* and was mapped to 277 and 287 KEGG pathways, respectively. The results showed that the largest pathway groups belonged to the category *metabolism* and *biosynthesis of secondary metabolites* in both the transcriptomes, revealing the presence of vast repertoire of secondary metabolites present in the rhizomes of both *C. longa* and *C. aromatica*. Interestingly, 24 unigenes of *C. longa* and 26 unigenes of *C. aromatica* involved in secondary metabolism were identified (Table 4), and among them, the cluster for 'phenylpropanoid biosynthesis [PATH: ko00940]' represents the largest subgroup followed by 'stilbenoid, diarylheptanoid and gingerol biosynthesis [PATH: ko00945] in both the transcriptomes. These genes involved in the enhancement of curcuminoid-biosynthesis-related pathways would greatly

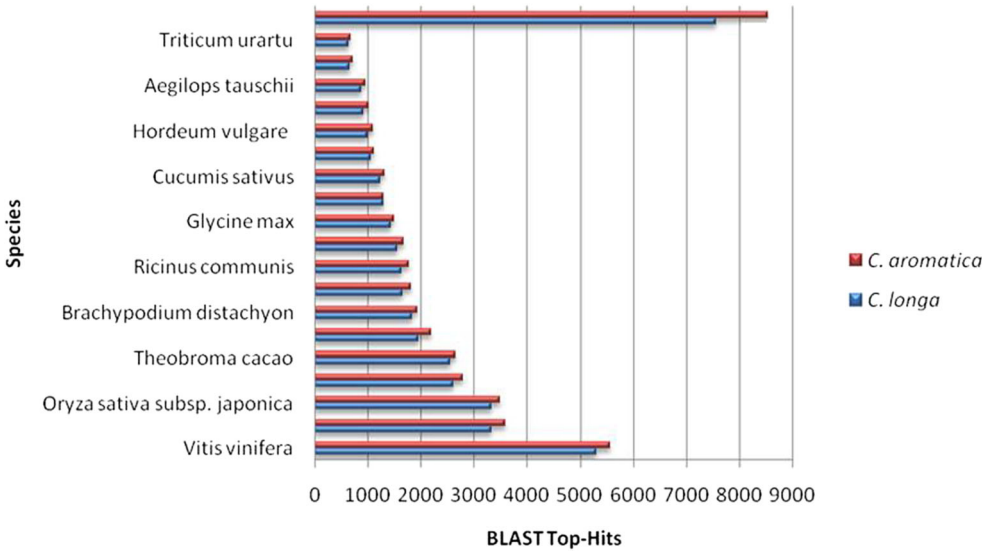

Fig. 3 Top-hit species distribution

enhance the potential for developing transgenics with elevated levels of curcuminoids.

Identification and Differential Expression Analysis of Sequences Related to Candidate Genes of Curcuminoid Biosynthesis

Radiotracer feeding studies suggested that curcumin is derived from the intermediates in the phenylpropanoid pathway (Kita et al. 2008). Schröder (1997) proposed that enzymes similar to polyketide synthases may be responsible for the formation of curcumin with diketide as an intermediate product. del Ramirez-Ahumada et al. (2006) evaluated the activity of phenylalanine ammonia lyase, p-coumaroyl shikmate transferase, p-coumaroyl quinate transferase, caffeic acid O-methyl transferase, caffeoyl CoA O-methyltransferase and polyketide synthases in the protein crude extracts of leaf, shoot and rhizome tissues. All the extracts possessed activity for all the enzymes. Katsuyama et al. (2009a, b) isolated and characterized four polyketide synthases from turmeric, one diketide CoA synthase (dcs) and three isoforms of curcumin synthase (curs1, curs2 and curs3), and proposed the pathway for curcumin biosynthesis. Koo et al. (2013) constructed a cDNA library from rhizomes of turmeric and suggested that large arrays of polyketide synthases, specific reductases, hydroxylases and MYB transcription factors represented in the library are potential candidate genes for further research in elucidating curcuminoid biosynthetic pathway. In the present study, all the reported genes involved in the putative curcuminoid biosynthesis pathway could be detected (Table 5) from the rhizome-based transcriptomes of both *C. longa* and *C. aromatica*, namely phenylalanine ammonia lyase (PAL), cinnamate 4-hydroxylase (C4H), 4-coumarate-CoA ligase (4CL), hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltransferase (HCT), coumarate 3-hydroxylase (C3H), caffeoyl CoA 3-O-methyltransferase (COMT), diketide CoA synthase (DCS), curcumin synthase 1 (CURS1), curcumin synthase 2 (CURS2) and curcumin synthase 3 (CURS3). In many cases, more than one unique sequence was annotated as encoding the same enzyme, probably because they represent fragments of a single transcript, different members of a gene family or both (Hyun et al. 2012).

Table 4 The unigenes related to secondary metabolites

Biosynthesis of secondary metabolites	Unigene numbers	
	C.longa	*C.aromatica*
Phenylpropanoid biosynthesis	10	14
Stilbenoid, diarylheptanoid and gingerol biosynthesis	3	3
Flavonoid biosynthesis	3	3
Isoquinoline alkaloid biosynthesis	2	1
Tropane, piperidine and pyridine alkaloid biosynthesis	2	1
Streptomycin biosynthesis	2	2
Butirosin and neomycin biosynthesis	2	2
Total	24	26

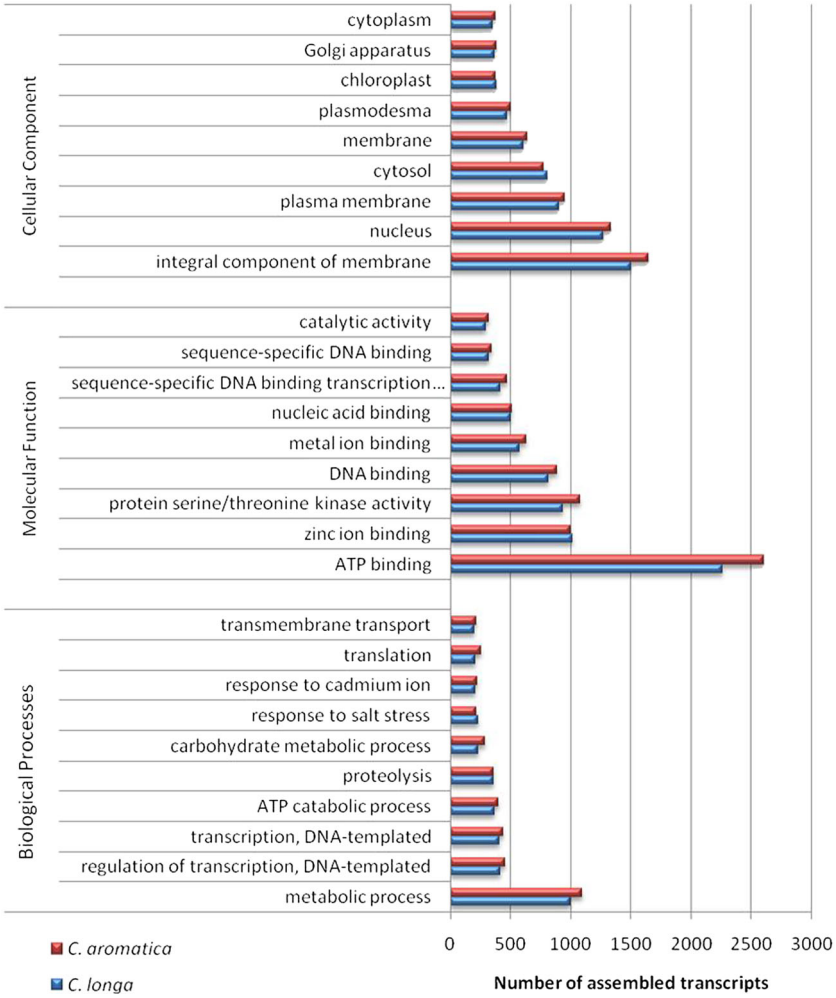

Fig. 4 Representation of Gene Ontology categories in the transcriptomes of C. *longa* and *C. aromatica*

Differential expression analysis identified 192 up-regulated and 317 down-regulated transcripts in *C. longa* and 397 up-regulated and 210 down-regulated transcripts in *C. aromatica*. The list of top 20 up-regulated transcripts of both the species is shown in Fig. 5. Among these, two novel polyketide synthase genes (*clpks1* and *clpks2*) showing similarity to *Musa acuminata* polyketide synthase type 2 (MaPKS2) and *M. acuminata* polyketide synthase type 4 (MaPKS4) were found to be up-regulated in *C. longa*. A previous study by Jitsaeng (2009) in *M. acuminata* reported that MaPKS2 and

Table 5 Statistics of putative genes involved in curcuminoid biosynthesis

Enzyme code	Name of enzyme (abbreviation)	Number of unigenes	
		C.longa	*C.aromatica*
4.3.1.24	Phenylalanine ammonia lyase (PAL)	3	9
1.14.13.11	Cinnamate 4-hydroxylase (C4H)	8	9
6.2.1.12	4-Coumarate:coenzyme A ligase (4CL)	28	27
1.14.14.9	Coumarate 3-hydroxylase (C3H)	2	1
2.3.1.133	Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase (HCT)	5	7
2.1.1.104	Caffeoyl CoA O-methyltransferase (COMT)	3	3
2.3.1.211	Diketide CoA synthase (DCS)	3	3
2.3.1.217	Curcumin synthase 1 (CURS1)	1	2
2.3.1.217	Curcumin synthase 2 (CURS2)	2	3
2.3.1.217	Curcumin synthase 3 (CURS3)	3	4

MaPKS4 were closely related to curcumin synthase and DCS, respectively. But, no significant differences in the expression of both DCS and curcumin synthase were detected between these species at this stage of study. This is interesting since in spite of the fact that *C. aromatica* is practically devoid of curcumin (Sajitha et al. 2014), two major downstream genes DCS and curcumin synthase, are expressed in this species in high levels. Katsuyama et al. (2007) also reported curcuminoid synthase from *O. sativa*, which also produces no detectable curcuminoids. This implies a strong regulation of these genes in deciding curcumin levels at a particular site.

Identification of Transcripts Encoding Transcription Factors

Aligning of the annotated transcripts of *C. longa* and *C. aromatica* to AGRIS database resulted in the identification of unigenes belonging to 39 transcription factor families in *C. longa* and 36 transcription factor families in *C. aromatica* (Fig. 6). The members of AP2-EREBP, MYB, C2H2, C2C2-Dof, WRKY, bHLH, HOMEOBOX, TCP, NAC and bHLH families were dominant in both the species. The transcription factors belonging to MYB, bHLH, AP2-EREBP, WRKY and NAC families were already reported to regulate secondary metabolic pathways (Yang et al. 2012). Four hundred six unigenes for *C. longa* and 413 unigenes for *C. aromatica* showed similarity to R2R3-MYB, a subfamily of MYB whose members were reported to regulate the upstream genes of phenylpropanoid biosynthesis (Omer et al. 2013; Huang et al. 2012b), suggesting the possibility of a similar regulatory mechanism in curcumin biosynthesis also. The transcription factors play a significant role in regulation of secondary metabolite biosynthesis by controlling gene expression (Broun. 2004). It seems likely that one of the processes causing alteration in curcumin content is also due to transcriptional alterations in the curcuminoid pathway regulation. This action may be responsible for the variation in curcumin content in different accessions of *C. longa* and related species of *Curcuma*. This is the first study depicting the entire repertoire of regulatory factors of curcumin biosynthetic pathway in *Curcuma*. However, the specific function of the particular MYB member in curcuminoid metabolism of turmeric needs to be further verified with functional genomics approach.

Gene Validation and Expression Analysis

To check the quality of the assembly and annotation data from the Illumina sequencing, full-length cDNA sequences corresponding to one of the most important downstream enzymes of the curcuminoid biosynthetic pathway viz., curcumin synthase 3 were isolated, sequenced and compared with the assembled sequences of both *C. longa* and *C. aromatica*. Overall, the assembled unigenes covered more than 99 % of the corresponding full-length genes and the complete ORF was predicted to be present. Additionally, the sequence variation was minimal (>99 % pair wise identity), which validated the NGS-based RNA-seq procedures as reliable. To evaluate the dynamic expression patterns, 14 unigenes were chosen for

Fig. 5 List of top 20 up-regulated transcripts in **a** *C. longa* and **b** *C. aromatica*

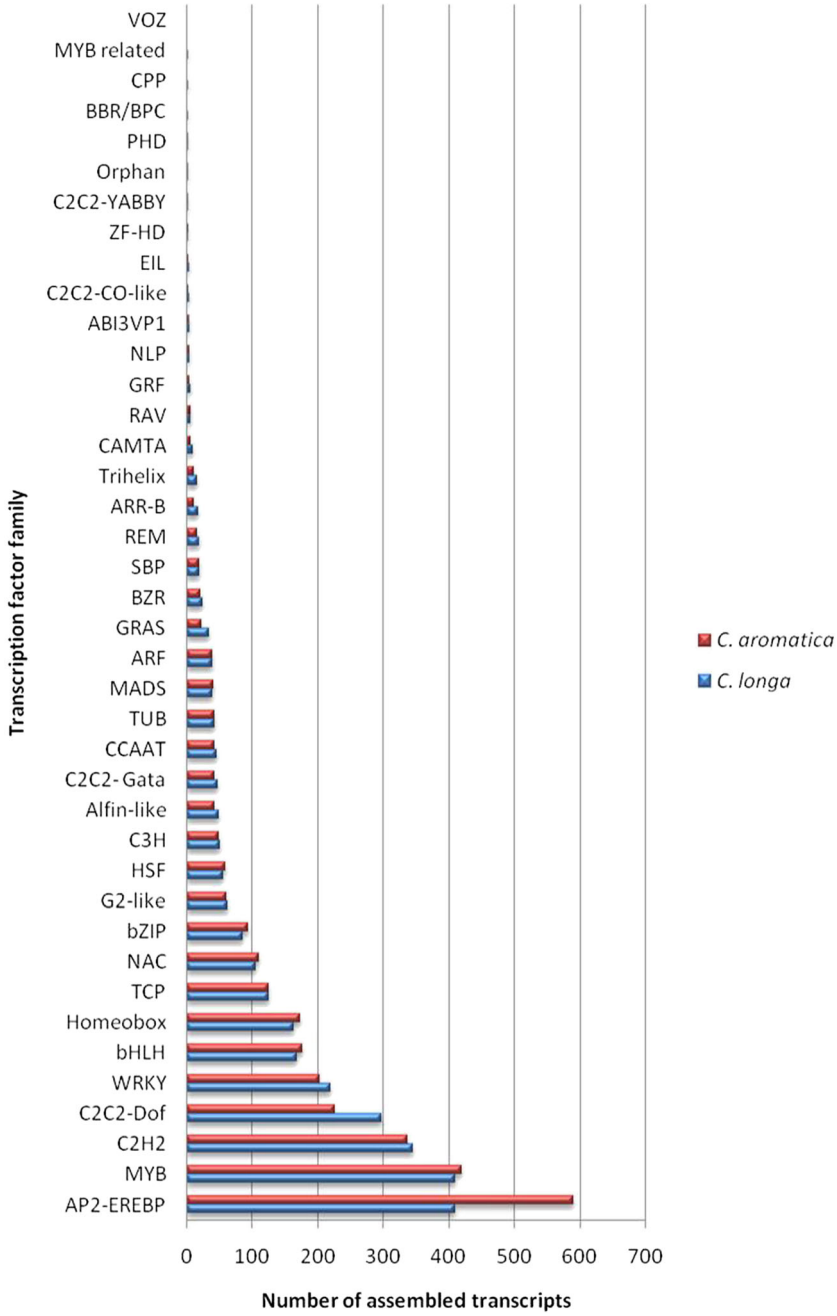

Fig. 6 Distribution of different transcription factors in the assembled transcripts of *C. longa* and *C. aromatica*

qRT-PCR analysis. Although there is change in the magnitude of fold expression levels, all the genes were upregulated in *C. longa*. These upregulated genes possess potential roles in curcuminoid biosynthetic pathway namely *pal*, *c4h*, *4 cl*, *hct*, *c3h*, *comt*, *dcs*, *curs1*, *curs2* and *curs3*; one transcription factor (*wrky*); and two novel polyketide synthases (*clpks1* and *clpks2*) with the exception of a transcription factor, *myb4*. It has been reported that *myb4* acts as a negative regulator of secondary metabolite biosynthesis in many crops (Jin et al. 2000; Wenping et al. 2011) (Fig. 7).

C. aromatica is characterized by pale-yellow-coloured rhizome and camphoraceous aroma whereas *C. longa* possesses yellow to deep-reddish-yellow-coloured rhizomes and typical turmeric-like aroma. Apart from the curcumin biosynthesis-related genes, differential gene expression analysis of these two species revealed that the expression level of beta-eudesmol synthase was higher in *C. aromatica* whereas beta-bisabolene synthase was higher in *C. longa*. Hydrocarbon sesquiterpenes are often responsible for the pungent or aroma flavours specific to plant tissues (Abel et al. 2009). Also, in some plants, volatile sesquiterpenes are emitted by wounding and are considered to function as defence against herbivore attack (Yu et al. 2008). A previous study showed that *Eucalyptus camphora* oil was found to be rich in

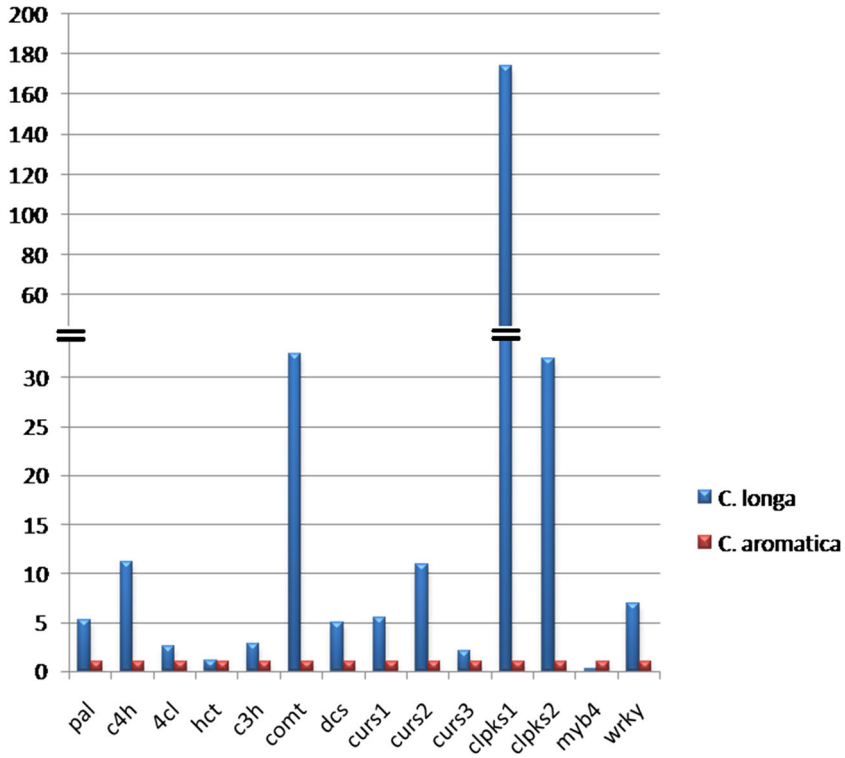

Fig. 7 Expression of putative curcuminoid biosynthetic genes and transcription factors in *C. longa* and *C. aromatica*

eudesmol which is isomeric with ordinary camphor (Smith 1899) whereas beta-bisabolene has a balsamic odour. Further characterization of these genes may help to elucidate the functions of these terpenes in *Curcuma*. The rhizome of *C. longa* is reported to be low in lignin (Nguyen et al. 2014). The transcriptome data showed that laccase, the last key enzyme involved in lignin biosynthesis, had a low level of expression in *C. longa*. Similar trend was found in duckweed (*Landoltia punctata*) which is having low lignin percentage due to the low expression of laccase gene (Tao et al. 2013). The rhizomes of both the species are rich in starch content, which is also inferred from the transcriptome, as all the genes involved in starch biosynthesis remain conserved among these two species. Thus, although a vast set of genes was found to be conserved in both the species, several hundreds of genes were found to be species specific. Thus, the present study generated several testable hypotheses which might shed light on the candidate and regulatory genes involved in the biosynthesis of species-specific compounds in the absence of a genome sequence.

Identification of SSRs and SNPs in Genes for Curcumin Biosynthesis

SSRs and SNPs are abundant markers that are suitable for a species with low genetic diversity such as *C. longa*. A total of 5488 and 5620 SSRs were identified for *C. longa* and *C. aromatica*, respectively. More than one SSR was found to be in 290 transcripts in *C. longa* and 283 transcripts in *C. aromatica*. Compound SSRs were found to be 134 and 141 for *C. longa* and *C. aromatica*, respectively. Trinucleotide SSRs were the most abundant, followed by dinucleotides and tetranucleotides in both the species (Table 6), which is consistent with findings from turmeric (Annadurai et al. 2013) and other monocot crops (La Rota et al. 2005). Four SSRs in *C. longa* and five SSRs in *C. aromatica* were detected in the curcuminoid biosynthesis pathway genes. In case of SNPs, only contigs of at least 150 bp long were considered in order to ensure sufficient flanking region for genoptying purposes. A total of 190 and 108 SNPs were identified in the candidate genes of curcuminoid biosynthetic pathway for *C. longa* and *C. aromatica*, respectively. These markers may have potential for mapping and association studies in *C. longa* based on curcumin content. They will also serve as a valuable source for genetic diversity analysis and extracting core collections from our vast germplasm repository.

Table 6 Summary of SSRs

Unit size	Number of SSRs	
	C.longa	*C.aromatica*
2	1057	1036
3	2317	2513
4	81	75
5	7	8

Identification of Target Site for miRNAs

High-throughput Illumina sequencing of small RNAs from turmeric rhizome had helped us to identify conserved and novel miRNAs (Santhi and Sheeja. 2013). Scanning of these miRNAs against identified transcripts has revealed that 68 and 64 transcripts from *C. longa* and *C. aromatica* have binding sites for some of the identified miRNAs. Our target prediction methods and criteria were very stringent (Allen et al. 2005; Schwab et al. 2005) but still allowed us to capture miRNA targets that are conserved across several plant species. The major miRNA targets identified in the present study are transcription factors including auxin response factor (ARF) and squamosa promoter binding protein-like (SPL) genes, targeted by miR160 and miR156, respectively. It has already been reported that in plants, miR156 plays an important role by regulating SPL (Wu et al. 2009) which is involved in leaf development and their aberrant expressions affect normal leaf and flower development (Ong and Wickneswari 2011). ARF genes are reported to play significant roles in plant development as activators or repressors of auxin-responsive transcription (Tiwari et al. 2003) and serve as targets of miR160 (Liu et al. 2007).

Conclusion

This is the first study to characterize the de novo rhizome transcriptome in *Curcuma*, aimed at identification of all the candidate genes of curcuminoid biosynthetic pathway. The presence of these genes was evident in both *C. longa* and *C. aromatica* without much variation in their expression levels. However, in our study, two novel polyketide synthase genes (*clpks1* and *clpks2*) as compared to *C. aromatica* showed enhanced expression in *C. longa* as evident from the differential gene expression and qRT-PCR analyses, suggesting that they may have key roles in curcuminoid biosynthesis. The present study involving two contrasting *Curcuma* species using NGS and real-time PCR analysis suggests that there may be novel genes involved in the biosynthesis of curcumin. Transcription factors with putative regulatory roles on phenylpropanoid biosynthesis like myb4 were also identified from both the species. The role of transcription factors, miRNAs, etc. in the expression of biosynthetic pathway genes needs to be further elucidated. The present study thus provides useful information for manipulating the curcumin biosynthetic pathway in turmeric as well as its related species. The identified SSRs, SNPs and miRNA targets will be a valuable resource for future studies on genetic mapping and the analysis of useful traits in turmeric.

Acknowledgments This work is partly supported by the Department of Biotechnology (DBT), New Delhi. The authors are grateful to Director, Indian Institute of Spices Research, for providing the facilities. The authors also thank Distributed Information Sub-Centre (DISC), Indian Institute of Spices Research, Kozhikode, for providing the support in sequence analysis and sequence deposition in NCBI.

References

Abel C, Clauss M, Schaub A et al (2009) Floral and insect-induced volatile formation in *Arabidopsis lyrata* ssp. petraea, a perennial, outcrossing relative of *A. thaliana*. Planta 230:1–11. doi:10.1007/s00425-009-0921-7

Allen E, Xie Z, Gustafson AM, Carrington JC (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. Cell 121(2):207–221

Anandaraj M, Prasath D, Kandiannan K et al (2014) Genotype by environment interaction effects on yield and curcumin in turmeric (*Curcuma longa* L.). Ind Crop Prod 53:358–364. doi:10.1016/j.indcrop.2014.01.005

Annadurai RS, Neethiraj R, Jayakumar V et al (2013) De Novo transcriptome assembly (NGS) of *Curcuma longa* L. rhizome reveals novel transcripts related to anticancer and antimalarial terpenoids. PLoS ONE 8:e56217. doi:10.1371/journal.pone.0056217

Broun P (2004) Transcription factors as tools for metabolic engineering in plants. Curr Opin Plant Biol 7:202–209. doi:10.1016/j.pbi.2004.01.013

Deepa K, Sheeja TE, Santhi R, Sasikumar B, Cyriac A, Deepesh PV, Prasath D (2014) A simple and efficient protocol for isolation of high quality functional RNA from different tissues of turmeric (*Curcuma longa* L.). Physiol Mol Biol Plants 20:263–271

del Ramirez-Ahumada MC, Timmermann BN, Gang DR (2006) Biosynthesis of curcuminoids and gingerols in turmeric (*Curcuma longa*) and ginger (*Zingiber officinale*): identification of curcuminoid synthase and hydroxycinnamoyl-CoA thioesterases. Phytochemistry 67:2017–2029. doi:10.1016/j.phytochem.2006.06.028

Elizabeth T, Zachariah JT, Syamkumar S, Sasikumar B (2011) Curcuminoid profiling of Indian turmeric. Int J Med Arom Plants 33(1):36–40, http://220.227.138.214:8080/dspace/handle/123456789/676

Ferreira de Carvalho J, Poulain J, Da Silva C et al (2013) Transcriptome de novo assembly from next-generation sequencing and comparative analyses in the hexaploid salt marsh species *Spartina maritima* and *Spartina alterniflora* (Poaceae). Heredity (Edinb) 110:181–193. doi:10.1038/hdy.2012.76

Garg R, Patel RK, Jhanwar S et al (2011) Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. Plant Physiol 156:1661–1678. doi:10.1104/pp. 111.178616

Ghawana S, Paul A, Kumar H et al (2011) An RNA isolation system for plant tissues rich in secondary metabolites. BMC Res Notes 4:85. doi:10.1186/1756-0500-4-85

Huang H-H, Xu L-L, Tong Z-K et al (2012a) De novo characterization of the Chinese fir (*Cunninghamia lanceolata*) transcriptome and analysis of candidate genes involved in cellulose and lignin biosynthesis. BMC Genomics 13:648. doi:10.1186/1471-2164-13-648

Huang W, Sun W, Lv H et al (2012b) Isolation and molecular characterization of thirteen R2R3-MYB transcription factors from *Epimedium sagittatum*. Int J Mol Sci 14:594–610. doi:10.3390/ijms14010594

Hyun TK, Rim Y, Jang H-J et al (2012) De novo transcriptome sequencing of *Momordica cochinchinensis* to identify genes involved in the carotenoid biosynthesis. Plant Mol Biol 79:413–427. doi:10.1007/s11103-012-9919-9

Iorizzo M, Senalik DA, Grzebelus D et al (2011) De novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. BMC Genomics 12:389. doi:10.1186/1471-2164-12-389

Jin H, Cominelli E, Bailey P et al (2000) Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in *Arabidopsis*. EMBO J 19:6150–6161. doi:10.1093/emboj/19.22.6150

Jitsaeng K (2009) Phytochemical and molecular studies on Musa plants and related species. Dissertation, The Friedrich Schiller University Jena. (http://www.clib-jena.mpg.de/theses/ice/ICE09004.pdf)

Kalra S, Puniya BL, Kulshreshtha D et al (2013) De novo transcriptome sequencing reveals important molecular networks and metabolic pathways of the plant, *Chlorophytum borivilianum*. PLoS ONE 8: e83336. doi:10.1371/journal.pone.0083336

Katsuyama Y, Matsuzawa M, Funa N, Horinouchi S (2007) In vitro synthesis of curcuminoids by type III polyketide synthase from *Oryza sativa*. J Biol Chem 282:37702–37709. doi:10.1074/jbc.M707569200

Katsuyama Y, Kita T, Funa N, Horinouchi S (2009a) Curcuminoid biosynthesis by two type III polyketide synthases in the herb *Curcuma longa*. J Biol Chem 284:11160–11170. doi:10.1074/jbc.M900070200

Katsuyama Y, Kita T, Horinouchi S (2009b) Identification and characterization of multiple curcumin synthases from the herb *Curcuma longa*. FEBS Lett 583:2799–2803. doi:10.1016/j.febslet.2009.07.029

Kita T, Imai S, Sawada H, Kumagai H, Seto H (2008) The biosynthetic pathway of curcuminoid in turmeric (*Curcuma longa*) as revealed by 13C-labeled precursors. Biosci Biotechnol Biochem 72(7):1789–1798

Koo HJ, McDowell ET, Ma X, Greer KA, Kapteyn J, Xie Z, Descour A, Kim H, Yu Y, Kudrna D, Wing RA, Soderlund CA, Gang DR (2013) Ginger and turmeric expressed sequence tags identify signature genes for rhizome identity and development and the biosynthesis of curcuminoids, gingerols and terpenoids. BMC Plant Biol 13:27

Krup V, Prakash HL, Harini A (2013) Pharmacological activities of turmeric (*Curcuma longa* linn): a review. J Homeop Ayurv Med 2:133. doi:10.4172/2167-1206.1000133, http://omicsgroup.org/journals/pharmacological-activities-of-turmeric-curcuma-longa-linn-a-review-2167-1206.1000133.pdf

La Rota M, Kantety RV, Yu J-K, Sorrells ME (2005) Non random distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. BMC Genomics 6:23

Liu PP, Montgomery TA, Fahlgren N, Kasschau KD, Nonogaki H, Carrington JC (2007) Repression of AUXIN RESPONSE FACTOR10 by microRNA160 is critical for seed germination and post-germination stages. Plant J 52(1):133–146. doi:10.1111/j.1365-313X.2007.03218.x/abstract

Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods 25:402–408. doi:10.1006/meth.2001.1262

Logacheva MD, Kasianov AS, Vinogradov DV et al (2011) De novo sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*). BMC Genomics 12:30. doi:10.1186/1471-2164-12-30

Nguyen CM, Nguyen TN, Choi GJ et al (2014) Acid hydrolysis of *Curcuma longa* residue for ethanol and lactic acid fermentation. Bioresour Technol 151:227–235. doi:10.1016/j.biortech.2013.10.039

Omer S, Kumar S, Khan BM (2013) Over-expression of a subgroup 4 R2R3 type MYB transcription factor gene from *Leucaena leucocephala* reduces lignin content in transgenic tobacco. Plant Cell Rep 32:161–171. doi:10.1007/s00299-012-1350-9

Ong SS, Wickneswari R (2011) Expression profile of small RNAs in *Acacia mangium* secondary xylem tissue with contrasting lignin content—potential regulatory sequences in monolignol biosynthetic pathway. BMC Genomics 12(Suppl 3):S13

Prasad S, Aggarwal BB (2011) Turmeric, the golden spice: from traditional medicine to modern medicine. In: Benzie IFF, Wachtel-Galor S (eds) Herbal medicine: biomolecular and clinical aspects, 2nd edn. CRC Press, USA, http://www.ncbi.nlm.nih.gov/books/NBK92752/

Sajitha PK, Prasath D, Sasikumar B (2014) Phenological variation in two species of *Curcuma*. J Plant Crop 42(2):252–255, http://220.227.138.214:8080/dspace/handle/123456789/1196

Santhi R, Sheeja TE (2013) Deep sequencing identifies candidate miRNAs from turmeric with possible regulatory roles on plant and human genes. In: Sasikumar B, Dinesh R, Prasath D, Biju CN, Srinivasan V (eds) Proceedings of the National Symposium on Spices and Aromatic Crops (SYMSAC VII): post-harvest processing of spices and fruit crops. Indian Society for spices, Indian Institute of Spices Research, Kozhikode, p 210, http://www.indianspicesociety.in/iss/pdf/SYMSAC%20VII%20-%20Souvenir.pdf

Sasikumar B (2005) Genetic resources of *Curcuma*: diversity, characterization and utilization. Plant Genet Resour Charact Util 3:230–251. doi:10.1079/PGR200574

Schröder J (1997) A family of plant-specific polyketide synthases: facts and predictions. Trends Plant Sci 2:373–378. doi:10.1016/S1360-1385(97)87121-X

Schwab R, Palatnik JF, Riester M, Schommer C, Schmid M, Weigel D (2005) Specific effects of microRNAs on the plant transcriptome. Dev Cell 8(4):517–527

Singh S, Joshi RK, Nayak S (2013) Identification of elite genotypes of turmeric through agroclimatic zone based evaluation of important drug yielding traits. Ind Crop Prod 43:165–171

Smith HG (1899) On the crystalline camphor of eucalyptus oil (eudesmol), and the natural formation of eucalyptol. J Proc R Soc NSW 33:86–107, http://www.forgottenbooks.com/readbook_text/Journal_and_Proceedings_of_the_Royal_Society_of_New_South_Wales_1000770596/457

Tao X, Fang Y, Xiao Y et al (2013) Comparative transcriptome analysis to investigate the high starch accumulation of duckweed (*Landoltia punctata*) under nutrient starvation. Biotechnol Biofuels 6:72. doi:10.1186/1754-6834-6-72

Tiwari SB, Hagen G, Guilfoyle T (2003) The roles of auxin response factor domains in auxin-responsive transcription. Plant Cell 15(2):533–543

Vaidya K, Ghosh A, Kumar V et al (2013) De novo transcriptome sequencing in *Trigonella foenumgraceum* L. to identify genes involved in the biosynthesis of diosgenin. Plant Genome. doi:10.3835/plantgenome2012.08.0021, https://www.agronomy.org/files/publications/tpg/tpg12-08-0021.pdf

Wenping H, Yuan Z, Jie S et al (2011) De novo transcriptome sequencing in *Salvia miltiorrhiza* to identify genes involved in the biosynthesis of active ingredients. Genomics 98:272–279. doi:10.1016/j.ygeno.2011.03.012

Wu G, Park MY, Conway SR, Wang JW, Weigel D, Poethig RS (2009) The sequential action of miR156 and miR172 regulates developmental timing in *Arabidopsis*. Cell 138(4):750–759

Yang C-Q, Fang X, Wu X-M et al (2012) Transcriptional regulation of plant secondary metabolism. J Integr Plant Biol 54:703–712. doi:10.1111/j.1744-7909.2012.01161.x

Yu F, Harada H, Yamasaki K et al (2008) Isolation and functional characterization of a β-eudesmol synthase, a new sesquiterpene synthase from *Zingiber zerumbet* Smith. FEBS Lett 582:565–572. doi:10.1016/j.febslet.2008.01.020

Zhou Y, Gao F, Liu R et al (2012) De novo sequencing and analysis of root transcriptome using 454 pyrosequencing to discover putative genes associated with drought tolerance in *Ammopiptanthus mongolicus*. BMC Genomics 13:266. doi:10.1186/1471-2164-13-266