Genome sequencing of turmeric provides evolutionary insights into its medicinal properties

Abhisek Chakraborty1, Shruti Mahajan1, Shubham K. Jaiswal1 & Vineet K. Sharma1

Curcuma longa, or turmeric, is traditionally known for its immense medicinal properties and has diverse therapeutic applications. However, the absence of a reference genome sequence is a limiting factor in understanding the genomic basis of the origin of its medicinal properties. In this study, we present the draft genome sequence of C. longa, belonging to Zingiberaceae plant family, constructed using 10x Genomics linked reads and Oxford Nanopore long reads. For comprehensive gene set prediction and for insights into its gene expression, transcriptome sequencing of leaf tissue was also performed. The draft genome assembly had a size of 1.02 Gbp with ~70% repetitive sequences, and contained 50,401 coding gene sequences. The phylogenetic position of C. longa was resolved through a comprehensive genome-wide analysis including 16 other plant species. Using 5,388 orthogroups, the comparative evolutionary analysis performed across 17 species including C. longa revealed evolution in genes associated with secondary metabolism, plant phytohormones signaling, and various biotic and abiotic stress tolerance responses. These mechanisms are crucial for perennial and rhizomatous plants such as C. longa for defense and environmental stress tolerance via production of secondary metabolites, which are associated with the wide range of medicinal properties in C. longa.
Turmeric, a common name for *C. longa*, has been traditionally used as a herb and spice for 4000 years in Southern Asia. It has a long history of usage in medicinal applications, as an edible dye, as a preservative in many food materials, in religious ceremonies, and is now widely used in cosmetics throughout the world. It is a perennial rhizomatous monocot herbal plant of the *Curcuma* genus comprising of more than 130 known species affiliated with the family of Zingiberaceae comprising more than 1300 species, which are widely distributed in tropical Africa, Asia, and America. The Zingiberaceae family is enriched in rhizomatous and aromatic plants that produce a variety of bioactive compounds such as curcumin. The association of endophytes with Zingiberaceae family plants helps in enhanced production of various secondary metabolites that further confer medicinal properties to species such as *C. longa*.

Secondary metabolism is one of the key adaptations in plants to cope with the environmental conditions through production of a wide range of common or plant-specific secondary metabolites. These secondary metabolites also play a key role in plant defense mechanisms, and several of these metabolites have numerous pharmacological applications in both medicinal and phytotherapy. The pathways for biosynthesis of various secondary metabolites including phenylpropanoids, flavonoids (such as curcuminoids), terpenoids, and alkaloids are found in *C. longa*. Other constituents such as volatile oils, proteins, resins, and sugars are also present in *C. longa*. Flavonoids are known to have anti-inflammatory, antioxidatant, and anti-cancer activities. Phenylpropanoids are also of great importance because of their antioxidant, anti-cancer, anti-microbial, anti-inflammatory, and wound-healing activities. The other class of secondary metabolites, terpenoids, are also known to possess anti-cancer and anti-malarial properties.

The three curcuminoids namely curcumin, demethoxy-curcumin, and bisdemethoxycurcumin, are responsible for the yellow color of turmeric. Among these, the primary bioactive component of turmeric is curcumin, which is a polyphenol-derived flavonoid compound and also known as diferuloylmethane. Curcumin shows broad-spectrum antimicrobial properties against bacteria, fungi and viruses, and also possesses anti-diabetic, anti-inflammatory, anti-fertility, anti-coagulant, hepatoprotective, and hypertension protective properties. Being an excellent scavenger of reactive oxygen species and reactive nitrogen species, its antioxidatant activity also controls DNA damage by lipid peroxidation mediated by free-radicals, and thus provides it with anti-carcinogenic properties. Due to these medicinal properties, turmeric has been of interest for scientists from many decades. Notably, the major bioactive compound of turmeric i.e., curcumin, is being recognized as “Pan-assay interference compounds” (PAINS), and “Invalid metabolic panaceas” (IMPS) candidate. As a PAINS candidate, curcumin can result in false-positive assay readouts, which are not the actual results of the interactions with other compounds in the assay but artefacts. Similarly, another report considered curcumin as a poor drug lead and IMPS candidate because of its promiscuous bioactivity and metabolic instability.

However, recently, the efficiency of curcumin as a drug lead was improved through prodrug-based curcumin nanoparticles generation that increased its chemical stability, and reduced its aggregation.

Several studies have been carried out to study the secondary metabolites and medicinal properties of this plant. Recently, the transcriptome profiling and analysis of *C. longa* using rhizome samples have been carried out to identify the secondary metabolite pathways and associated transcripts. However, its reference genome sequence is not yet available, which is much needed to understand the genomic and molecular basis of evolution of the unique characteristics of *C. longa*. According to the Plant DNA c-values database, *C. longa* genome has an estimated size of 1.33 Gbp with 2n = 63 chromosomes, but a wide range of genome size variation (4C values ranging from 4.30 to 8.84 pg) and chromosome number variation (2n = 48 to 2n = 64) in *C. longa* was suggested. Recent studies showed evidence for a ploidy level of 3X (2n = 63 chromosomes, basic chromosome number X = 21).

Therefore, we performed the draft genome sequencing and assembly of *C. longa* using Oxford Nanopore long reads and 10x Genomics linked reads generated on Illumina platform. The transcriptome of rhizome tissue for this plant has been known from several previous studies, and one study also reported the transcriptome of leaf tissue. Here, we carried out an extensive transcriptome sequencing of leaf tissue followed by a comprehensive transcriptome analysis, which also helped in the gene set construction. We also constructed a genome-wide phylogeny of *C. longa* with other available monocot genomes. The comparative analysis of *C. longa* with other monocot genomes revealed adaptive evolution in genes associated with plant defense and secondary metabolism, and provided genomic insights into the medicinal properties of this species.

**Results**

**Sequencing of genome and transcriptome.** A total of 94.8 Gbp of 10x Genomics linked read data, 47.2 Gbp Oxford Nanopore long-read data, and 32.4 Gbp of RNA-Seq data was generated from leaf tissue (Supplementary Tables 1–2). The total genomic data corresponded to ~82.4X coverage of 10x Genomics linked read data, and ~41X coverage of Nanopore long read data based on the estimated genome size of 1.15 Gbp using SGA-preqc. To carry out the genome annotation, de novo transcriptome assembly was performed using RNA-Seq data from this study. All the paired-end RNA-Seq reads were trimmed and quality filtered using Trimomatic v0.38 and used for de novo transcriptome assembly. The detailed workflow for genome and transcriptome analysis is shown in Supplementary Fig. 1.

**Assembly of the *C. longa* genome and transcriptome sequence.** The genome size of *C. longa* was estimated to be 1.15 Gbp using SGA-preqc with barcode-filtered 10x Genomics linked reads, which is close to the previously estimated genome size of 1.33 Gbp. *C. longa* genome was estimated to contain 4.83% heterozygosity. The *C. longa* genome sequence, assembled using Supernova v2.1.1 and Flye v2.4.3 had the N50 values of 15.8 Kbp and 60.9 Kbp, respectively. After correction of mis-assemblies, scaffolding, gap-closing and polishing, the final draft genome assembly (contigs with length of ≥3000 bp after scaffolding) of *C. longa* had the total size of 1.02 Gbp that comprised of 22,470 contigs with N50 value of 100.6 Kbp and GC-content of 38.75% (Supplementary Table 3). BUSCO (Benchmarking Universal Single-Copy Orthologs) completeness for Supernova assembled genome was 75.9% (Complete BUSCOs) and Flye assembled genome was 71.5% (Complete BUSCOs), which was improved to 92.4% (Complete BUSCOs) in the final polished draft *C. longa* genome assembly (Supplementary Table 4). Further, 98.9% 10x Genomics linked reads, 92.4% Nanopore long-reads, and 92.9% RNA-Seq reads were mapped on this final genome assembly. LAI value of *C. longa* genome (≥35 Kbp) that covered 72.2% of the estimated genome size was calculated at 10.26 (Supplementary Table 5) (see Methods). The genome of *C. longa* was predicted as triploid since at the variable sites, both the distributions of base frequencies showed the smallest Δlog-likelihood value for the triploid fixed model, before and after denoising (see “Methods”) (Supplementary Fig. 2a). Also,
heterozygous k-mer pair coverage pattern distribution using Smudgeplot\(^4\) (see “Methods”) showed that 83% of the k-mer pairs corroborated to total coverage of k-mer pair = 3n and normalized minor k-mer coverage = 1/3, and thus the genome was inferred as triploid (Supplementary Fig. 2b).

The de novo transcriptome assembly of \textit{C. longa} (from this study) using Trinity v2.9.1\(^{35}\) had a total size of 86,158,697 bp, with a total of 84,520 predicted transcripts corresponding to 36,510 genes. The complete assembly had an N50 value of 1086 bp, an average transcript length of 1019 bp and GC-content of 45.45% (Supplementary Table 6). A total of 30,552 unigenes were identified after clustering using CD-HIT-EST v4.8.1\(^{36}\) to remove the redundant gene sequences. The coding sequence (CDS) prediction from these unigenes using TransDecoder v5.5.0 resulted in 23,943 coding genes.

**Genome annotation and gene set construction.** For repeat identification, a de novo custom repeat library was constructed using the final polished \textit{C. longa} genome by RepeatModeler v2.0.1\(^{37}\), which resulted in a total of 2430 repeat families. The repeat families were clustered into 1977 representative sequences. These were used to soft-mask the genome assembly using RepeatMasker v4.1.0, which predicted 64.16% of \textit{C. longa} genome as repetitive sequences, of which 62.37% was identified as interspersed repeats (31.61% unclassified, 28.50% retroelements and 2.26% DNA transposons). Retroelements consisted of 27.37% LTR (long terminal repeat) elements (17.19% Ty1/Copia and 2.26% DNA transposons). Retroelements were clustered using CD-HIT-EST v4.8.1\(^{36}\) to remove the redundant gene sequences. The coding sequence (CDS) prediction from these unigenes using TransDecoder v5.5.0 resulted in 23,943 coding genes.

**Genes with signatures of adaptive evolution.** Genes with site-specific signatures of adaptive evolution were identified in \textit{C. longa}. 3230 genes showed unique amino acid substitution with respect to the other selected species. Among these 3230 genes, 2429 genes were identified to have functional impacts using sorting intolerant from tolerant (SIFT), and were considered further. Further, 569 genes were found to contain positively selected codon sites with greater than 95% probability. In addition to these site-specific signatures of evolution, 63 genes showed higher rate of nucleotide divergence, and 306 genes showed positive selection in \textit{C. longa} with FDR (false discovery rate)-corrected \(p\)-values < 0.05. These positively selected genes had positively selected codon sites with greater than 95% probability. A total of 188 genes were identified containing more than one of the signatures of adaptive evolution namely positive selection, unique amino acid substitution with functional impact and higher rate of nucleotide divergence.

The positively selected genes, genes with higher nucleotide divergence, genes showing site-specific evolutionary signatures, and MSA (multiple signs of adaptive evolution) genes were mapped on KEGG (Kyoto encyclopedia of genes and genomes) pathways, and classified in eggNOG COG (clusters of

**Fig. 1 Phylogenetic position of \textit{C. longa}.** Phylogenetic tree of \textit{C. longa} with 15 other selected species and \textit{Arabidopsis thaliana} as an outgroup species. The values mentioned at the nodes correspond to the bootstrap values.
Adaptive evolution of plant defense associated genes. Several genes known to provide plant immunity against pathogen infection or disease were found in MSA genes. Since plants lack adaptive immune response, the innate immune response in plants is provided by PAMP-triggered (PTI) and effector-triggered (ETI) immunity with the help of two plant stress hormones, salicylic acid (SA) and jasmonic acid (JA) signaling pathway. Among these MSA genes, JAR1 (Jasmonate resistant 1) is required for conversion of jasmonic acid to its bioactive form jasmonoyl-L-isoleucine (JA-Ile), COI1 (Coronatine- insensitive 1) is a receptor of JA-Ile and thus regulates downstream JA-signaling processes, MPK9 (Map kinase 9) gene expression is regulated by JA as well as SA treatments and is involved in PAMP-triggered immunity, BSK1 (Brassinosteroid-signaling kinase 1) plays an important role in brassinosteroid signaling pathway that is involved in plant innate immunity and also has an antagonistic relationship with JA signaling effects. Also, WRKY transcription factor is induced by pathogen attack and is involved in SA-signaling pathway mediated plant immunity, MYB48 (Myeloblastosis 8) transcription factor that is involved in salicylic acid-mediated response negatively regulates effector-triggered immunity, and EIN3 (ethylene-insensitive 3) negatively regulates SA levels and PAMP-triggered plant innate immunity. Jasmonic acid and salicylic acid elicit the production and accumulation of secondary metabolites such as phenolics, terpenoids, alkaloids, and glycosides, in medicinal plants.

Three O-fucosyltransferase family proteins—AT1G22460, AT3G11540, and AT4G08810 were found in MSA genes category, and are involved in plant immunity. Previously it has been shown that the lack of fucosylation of genes led to increased disease susceptibility in Arabidopsis sp. by affecting PTI, ETI as well as stomatal and apoplastic defense. Three ubiquitin-conjugating enzymes—UBE2E (ubiquitin-conjugating enzyme E2 E), UBE2D (ubiquitin-conjugating enzyme E2 D), AT2G16920 and six E3 ubiquitin-protein ligase genes—COPI (constitutive photomorphogenic protein 1), AT1G55250, AT4G27880, AT4G28370, AT3G26730, and AT5G45360 also showed multiple signs of adaptive evolution. Ubiquitination-related proteins affect hypersensitive-response (HR) and phytohormone signaling mediated pathogen defense by targeting proteins for proteasomal degradation. These ubiquitin ligase proteins also regulate various abiotic stress responses such as drought, salinity, temperature. Among the MSA genes, three ribosomal subunit proteins are regulated by signaling molecules such as methyl jasmonate, salicylic acid and environmental stress effects. These two cell cycle related MSA genes—cyclin-A and APC10 (anaphase promoting complex subunit 10) are involved in plant immunity, disease resistance as well as abiotic stress responses. Cell cycle mediated pathogen defense by targeting proteins for proteasomal degradation and泛素化相关蛋白影响对植物活性的影响。
feruloyl-CoA showed unique amino acid substitution with functional impact (Fig. 3a).

Further, coumaroyl-CoA and feruloyl-CoA are used for the production of curcumin, demethoxycurcumin, and bisdemethoxycurcumin via coumaroyl-diketide-CoA and feruloyl-diketide-CoA, catalyzed by four enzymes—CURS1 (curcumin synthase 1), CURS2 (curcumin synthase 2), CURS3 (curcumin synthase 3), and DCS (diketide-CoA synthase)75. To identify these enzymes in the genome and transcriptome assemblies constructed in this study, we mapped the coding gene sequences of CURS1, CURS2, CURS3, and DCS on these assemblies. All four enzymes were found to be present in the de novo genome assembly, in the gene set derived from de novo transcriptome assembly of C. longa. Using Exonerate, we further constructed the gene trees for these four major curcuminoid biosynthesis genes (Fig. 3b–e). Each of the CURS1, CURS2 and CURS3 genes consisted of two exons and one intron. DCS gene consisted of three exons and two introns. DCS and CURS genes are members of chalcone synthase (CHS) family77, and the genes from CHS family generally consist of two exons and one intron78, which is consistent with previous studies and also further supported by our findings.

Coumaroyl-CoA is also a precursor for biosynthesis of anthocyanins, flavonols and catechins76. A key enzyme CHS, responsible for conversion of coumaroyl-CoA to chalcone, showed unique substitution with functional impact (Fig. 3a). The FLS (flavonol synthase) gene, required for flavonols synthesis76, possessed unique amino acid substitution. Another intermediate of curcuminoid biosynthesis pathway, feruloyl-CoA, is a precursor for lignin biosynthesis79. Enzymes involved in production of lignins from feruloyl-CoA also showed signatures of evolution. CAD (cinnamyl alcohol dehydrogenase) was positively selected, PRX (peroxidase) and CCR (cinnamyl-CoA reductase) exhibited unique substitution with functional impact, and LAC (laccase) showed both unique amino acid substitution and positively selected codon site.

Terpenoid biosynthesis pathway also showed distinct evolutionary signatures in C. longa. 7 out of 10 enzymes in the mevalonate pathway of terpenoid backbone biosynthesis80 were found to be evolved in comparison to the other selected species (Fig. 3f). Among these, AACT (acetoacetyl-CoA thiolase), HMGS (HMG-CoA synthase), HMGR (HMG-CoA reductase), MK (mevalonate kinase), PMK (phosphomevalonate kinase), MDD (mevalonate-5-diphosphate decarboxylase), IDI (isopentenyl diphosphate isomerase), GPPS (geranyl diphosphate synthase), FPPS (farnesyl diphosphate synthase), GGPPS (geranylgeranyl diphosphate synthase).
had ancestor genes from bacterial origin, and fungal orthologs were also observed in the case of six of these enzymes. Identified orthologs of PAL, 4CL, and OMT (O-methyltransferase) enzymes (Supplementary Fig. 3) show ammonia-lyase activity, catalytic activity, and methyltransferase activity, respectively. The four key enzymes unique to curcuminoid biosynthesis pathway i.e., DCS, CURS1, CURS2, and CURS3 (Type III polyketide synthases) had similar bacterial and fungal orthologs (Fig. 4a–d) that were annotated as 3-oxoacyl-ACP synthase, and chalcone and stilbene synthase, respectively. 3-oxoacyl-ACP synthase plays a role in fatty acid biosynthesis in bacteria82, and chalcone and stilbene synthase that was identified as a fungal ortholog is a polyketide synthase and a key enzyme involved in secondary metabolite biosynthesis pathways83.

Gene family evolution analysis (using CAFÉ v4.2.1)84 for these ten enzymes showed that gene families of 8 out of 10 enzymes were expanded in C. longa compared to its immediate ascending node, and gene families of HCT and OMT genes were contracted (Supplementary Table 22). Among these, 4CL gene family showed comparatively more expansion in terms of gene numbers. DCS, CURS1, CURS2, and CURS3 genes were identified as members of the same gene family, which underwent expansion in the CAFÉ analysis.

Discussion

C. longa is a monocot species from Zingiberaceae plant family and is widely known for its medicinal properties and therapeutic applications15. In this study, we carried out the whole-genome sequencing and reported the draft genome sequence of C. longa. This is the first whole-genome sequenced and analyzed from Zingiberaceae plant family to the best of our knowledge, which comprises of more than 1300 species, and thus will act as a valuable reference for studying the members of this family including those of Curcuma genus. Genomic polyploidy in members of Curcuma genus is well known from previously reported experimental studies27,85,86. In this study, we estimated the ploidy level of C. longa genome using next-generation sequencing (NGS) reads, and showed the triploid nature of C. longa genome, which is also supported by the previous experimental studies27,28,85,86. The application of Oxford Nanopore long-reads and 10x Genomics linked read sequencing that has the potential to resolve complex polyploid genomes87, helped in successfully constructing the C. longa draft genome of 1.02 Gbp with a decent N50 of 100.6 Kbp. After assembly correction, scaffolding, gap-closing and polishing, the BUSCO completeness of final C. longa genome improved to 92.4%, which is similar to other plant genomes, thus indicating the usefulness of post-assembly processing88. It is noteworthy that the LAI value of C. longa genome (≥35 Kbp) was estimated at 10.26, which also corresponds to a reference quality genome assembly89.

Since the construction of a comprehensive gene set was essential to explore the genetic basis of its medicinal properties, both genome and transcriptome assemblies, and an integrated approach using de novo and homology-based methods were used, which resulted in the final set of 50,401 genes. The identification of all Type III polyketide synthase genes CURS1, CURS2, CURS3, and DCS, involved in the biosynthesis of the three most important secondary metabolites (curcuminoids)—curcumin, demethoxycurcumin and bisdemethoxycurcumin, in both genome and transcriptome assemblies also attests to the quality and comprehensiveness of our genome and transcriptome assembly. Further, the revelation of complete gene structures of the above four biosynthesis genes of curcuminoid pathway from the draft genome of this plant is likely to help further studies and improve commercial exploitation of these curcuminoids that find wide applications as coloring agents, food additives and possess antioxidant, anti-inflammatory, anti-microbial, neuroprotective, anticancer, and many other medicinal properties90.

Repetitive sequence prediction revealed that ~70% of the genome consisted of repeat elements, which is similar to other
plant genome, such as *Triticum urartu*⁴¹. Notably among the LTR repeat elements, Ty1/Copia elements (17.19%) were more abundant than Gypsy/DIRS1 elements (9.42%), which corroborates with the observations made in the case of *Musa acuminate* species from the same Zingiberaceae plant order, and thus appears to be a specific signature of repeat elements in Zingiberaceae order⁴².

The genome-wide phylogenetic analysis of *C. longa* with 15 other representative monocot species available on Ensembl plants revealed the relative position of *C. longa*, which was supported by previously reported phylogenies using 1685 gene partitions, and using phytocytstat gene CypP1⁴²,⁴³. Ren et al. also showed similar phylogenetic position of *C. longa* with other selected monocots – *Dioscorea sp.*, *Musa acuminata*, *Brachypodium distachyon*, *Oryzia sativa*, *Panicum sp.*, *Setaria italica*, *Zea mays*, *Sorghum bicolor* using genome and transcriptome data of 105 angiosperms⁴⁴. Also, an updated megaphylogeny for vascular plants showed similar relative phylogenetic position of *C. longa* with respect to the selected monocots⁴⁵. Further, selected species from Poales order also showed similar relative positions with respect to each other⁴⁶. Absence of any polyphony in the phylogenetic tree is because of large number of genomic loci and a high bootstrap value, or no multiple speciation events took place at the same time. Taken together, the genome-wide phylogenetic analysis of *C. longa* confirmed its phylogenetic position and will be a useful reference for further studies.

Analysis of genes with signatures of adaptive evolution using 5388 orthodox gene sets revealed that a large proportion (~91%) of the genes with multiple signs of adaptive evolution (MSA) were associated with plant defense mechanisms against biotic and abiotic stress responses, and secondary metabolism. Notable ones among these are the genes associated with Jasmonic acid and salicylic acid signaling pathways. These two pathways are important components of plant innate immune response⁴⁶, and also affect plant secondary metabolism by regulating the production of secondary metabolites⁴³. Thus play a crucial role in plant-pathogen interaction. Jasmonic acid is also reported to have a role in induction and growth of rhizome in vitro through its interaction with ethylene, which is important for a rhizomatous plant like *C. longa*⁴⁷. Further, one of the genes (PAL) for the enzymes involved in curcuminoid biosynthesis pathway was also found to have signatures of adaptive evolution, which is an important observation because curcuminoid is the most important secondary metabolite of *C. longa*. The genes for the four key enzymes (*CURS1, CURS2, CURS3, DCS*) of this pathway could not be found in the list of MSA genes since these genes are unique to *Curcuma* genus and were absent in the other species considered for the evolutionary analysis.

The gene family evolution analysis showed that gene families of all ten enzymes of curcuminoid biosynthesis pathway underwent expansion/contraction compared to the immediate ascending node of *C. longa* in the species phylogenetic tree, which suggests evolution of this pathway in *C. longa*. Further, evolutionary origin of these ten enzymes revealed that homologs for the enzymes exist in bacterial and fungal species indicating ancestral origin of these genes. Interestingly, in case of the four key enzymes (*CURS1, CURS2, CURS3, DCS*), the bacterial ancestor genes were involved in primary metabolism, and fungal ancestor genes (member of polyketide synthase family) were involved in secondary metabolism. Taken together, these observations suggest an evolution of these genes in *C. longa* to play key roles in curcuminoid biosynthesis pathway. Furthermore, several major secondary metabolism pathways were also found to be evolved in *C. longa* compared to the other selected plant species in this study. Also, the key enzymes involved in the biosynthesis pathways of terpenoid backbone and important compounds in phe-nolic group of secondary metabolites (e.g., curcuminoids, anthocyanins, lignins, phenylpropanoid) showed signatures of adaptive evolution, which is an important observation since these pathways are associated with the wide range of medicinal properties of *C. longa*.

It is important to mention here that the biosynthesis of secondary metabolites such as polyketides (curcuminoids), which are crucial for plants survival under environmental challenges, are regulated by biotic and abiotic stress responses⁴⁹,⁵⁰. Also, it is known that secondary metabolites in plants are primarily produced in response to environmental stress and for plant defense, which help in better survival under various environmental conditions⁵⁰, and several of these secondary metabolites also possess medicinal values. This also seems to be the case with *C. longa* where the observed abundance of adaptively evolved genes associated with plant defense mechanisms and secondary metabolism makes it tempting to speculate that these genes gradually evolved for environmental adaptation and to confer resistance to a perennial rhizomatous plant like *C. longa*. Several of the metabolites produced in the above processes possess diverse medicinal properties, and thus provide *C. longa* with its medicinal characteristics and traditional significance.

**Methods**

**Sample collection, library preparation, and sequencing.** The plant sample was collected from an agricultural farm (23°30’26”N 77°30’89”E) located in Bhopal, Madhya Pradesh, India. The leaves were homogenized in liquid nitrogen for DNA extraction using Carlson lysis buffer. Species identification was performed by PCR (polymerase chain Reaction) amplification of a nuclear gene (internal transcribed spacer ITS) and a chloroplast gene (Maturase K), followed by Sanger sequencing at the in-house facility. The linked read library construction from the extracted DNA was done with the help of Chromium Controller instrument (10x Genomics) using Chromium™ Genome Library & Gel Bead Kit v2 by following the manufacturer’s instructions. The Nanopore library was prepared using SQK-LSK109 kit and sequenced on MiniION platform using FLO-MIN106 flow cell. The raw reads from the plant sample, DNA extraction was carried out using DNA powder leaves using TriZol reagent (Invitrogen, USA). The transcriptomic library was prepared with TruSeq Stranded Total RNA Library Preparation kit by following the manufacturer’s protocol with Rib-o-Zero Workflow (Illumina, Inc., USA). The quality of 10x Genomics and transcriptomic libraries was evaluated on Agilent 2200 TapeStation using High Sensitivity D1000 ScreenTape (Agilent, Santa Clara, CA) prior to sequencing. The prepared genomic (10x Genomics) and transcriptomic libraries were sequenced on NovaSeq 6000 (Illumina, Inc., USA) generating 150 bp paired-end reads. The detailed DNA and RNA extraction steps and other methodologies are mentioned in Supplementary Notes 1.

**Genomic data processing and assembly.** The barcode sequences were trimmed from raw 10x Genomics linked reads using a set of python scripts (https://github.com/ucdavis-bioinformatics/procedures/proc10xG). The genome size of *C. longa* was estimated using a k-mer count distribution method implemented in SGA-preg⁵⁰ (Supplementary Notes 2). A total of 631.11 million raw 10x Genomics linked reads corresponded to ~82.4X coverage were used for generating a de novo assembly using Supernova assembler v2.1.1 with maxreads = all option and other defaults settings⁵¹. The haplotype-phased assembled genome was generated using Supernova mkoutput in ‘pseudohap’ style. The 10x Genomics linked reads were run through Longranger basic v2.2.2 (https://support.10xgenomics.com/gene-expression/software/pipelines/latest/installation) for barcode processing and were used to detect and correct mis-matched base calls of Barcode genome using Trimgen v1.1.27. The first round of scaffolding was carried out using ARCS v1.1.1 (default parameters) to generate more contiguous assembly using 10x Genomics linked reads⁵⁶. Further scaffolding was performed to improve the contiguity using AGOUTI v0.3.3 with the quality filtered paired-end RNA-Seq reads from our study, which was also used in de novo transcriptome assembly⁵⁷. Adapter-processed Nanopore long-reads (~20 Kb) were also used for scaffolding of the genome assembly using LINKS v1.8.6 with default parameters⁵⁸. Oxford Nanopore long-read data was base-called using Guppy v4.4.0 (Oxford Nanopore Technologies), and adapter sequences were removed using Porechop v0.2.4 (Oxford Nanopore Technologies). The adapter-processed Nanopore reads were used to perform long reads-based de novo assembly of *C. longa* genome by Flye⁵⁹ with default parameters using the version v2.4.2 that provides better assembly coverage and contiguity. This assembly was polished using barcode-processed 10x Genomics linked reads using Pilon⁶⁰, v1.23 in three iterations to fix local mis-assemblies, small indels, or individual base errors that could be introduced from long, error-prone Nanopore reads. Scaffolding of this polished assembly was performed with barcode-processed 10X Genomics linked reads,
quality-filtered RNA-Seq reads from our study, and Nanopore long reads (≥20 Kb) using ARCs v1.1.130, AGOUTI v0.3.379, and LINKS v1.8.6100, respectively. The genome assembly of C. longa generated from 10x genomics linked reads and Nanopore long reads were merged together using Quickmerge v0.3 in order to achieve a more contiguous assembly102. Gap-closing of this scaffolded assembly was performed with barcode-process linked reads using Sealer v2.1.5 with k-mer value from 30 to 70 with an interval of 10 bp using a Bloom filter-based approach101 and L.R. 2050 v2.0103,104 with Nanopore long reads. Finally, the assembly quality was improved by Pilon v1.23 using barcode-process linked reads to fix small indels, individual base errors, or local mis-assemblies that could be introduced by the previous scaffolding steps105. The other details about the genome assembly polishing parameters are mentioned in Supplementary Note 2. Further, in order to validate the final genome assembly of C. longa, barcode-removed 10x Genomics linked reads, adapter-processed Nanopore long reads, and quality-filtered RNA-Seq reads from this study were individually mapped to the assembly using BWA-MEM v0.7.17105 with default parameters, and samtools v1.9 was used to generate unfiltered indexed alignments. These alignments were processed using nQuery with default parameters to extract the variable sites with the free model and the fixed models. Distribution of these base frequencies was denoted and both the distributions, before and after denosing, were used to estimate Alog-likelihood values for the fixed models. Also, Smudgeplot v0.4.0.2 was used to estimate k-mer counting, and k-mer frequency-based histogram generation using KMC112 v3.1.1 with the parameters: k-mer length of 21, excluding k-mers occurring less than 1 time, maximum value of a counter of 10,000, and depth ≥ 1000 bp. Accordingly we restricted this analysis by only considering the monocot species, and NCBI non-redundant (nr) database using BLASTP (e-value cut-off 10−5), and Pams2 (v3.2.0) database using HMMER v3.3.2 with an e-value cut-off 10−5. Further, sequence variation between different alleles in the coding gene regions were analyzed in the final coding gene set using quality-filtered paired-end RNA-Seq data from this study. First, filtered paired-end reads were mapped from the quality-filtered RNA-Seq reads using FastUniq v1.1.126. The resultant reads were mapped to the coding genes using BWA-MEM v0.7.17105, and SAMTools v1.398 was used to generate the alignment in BAM format. Using this alignment, BCFTools127 (v1.9) "mpileup" was used for variant calling in the coding genes, and further filtering of false-positive variants based on the following parameters128—mapping quality ≥50, variant sites with quality ≥30, sequencing depth ≥30.

Orthologs identification. Representative species from all 15 monocot genus available in Ensembl plants release 47, and model organism Arabidopsis thaliana as an outgroup species were selected for orthologs identification19. To construct the orthologs, the protein sequences of C. longa obtained from TransDecoder and proteome files for other selected species i.e., Agrostis tanschii, Ananas comosus, Brachiadipodium distachyon, Dioscorea rotundata, Eravgrisit tef, Hordeum vulgare, Leersia oryza, Musa acuminata, Oryza sativa, Panicum hallii, Setaria italica, Sorghum bicolor, Triticum aestivum, Zea mays, and Arabidopsis thaliana obtained from Ensembl release 47, were used. The longest isoforms for all proteins were extracted for all selected species to construct the orthologs using OrthoFinder v2.3.9140.

Construction of orthologous gene set and phylogenetic tree. Only those orthologs that contained genes from all 17 species were extracted from all the identified orthologs. The fuzzy one-to-one orthologous containing genes from all 17 species were identified from these orthologs, and extracted using KinFin v1.051. For cases where the orthologous gene sets comprised of multiple genes for a species, the longest gene was extracted. The fuzzy one-to-one orthologous were further aligned individually using MAFFT v7.467130 for reliable results. The amino acid alignment of orthogroups was performed using PROTGAMMAAUTO161. Further, sequence variation between different alleles in the coding gene regions were analyzed in the final coding gene set using quality-filtered paired-end RNA-Seq data from this study. First, filtered paired-end reads were mapped from the quality-filtered RNA-Seq reads using FastUniq v1.1.126. The resultant reads were mapped to the coding genes using BWA-MEM v0.7.17105, and SAMTools v1.398 was used to generate the alignment in BAM format. Using this alignment, BCFTools127 (v1.9) "mpileup" was used for variant calling in the coding genes, and further filtering of false-positive variants based on the following parameters128—mapping quality ≥50, variant sites with quality ≥30, sequencing depth ≥30.

Identification of genes with higher nucleotide divergence. Protein sequences of all the orthologs across 17 species were aligned individually using MAFFT v7.467 with individual-based parameter set. Multiple sequence alignments of all fuzzy one-to-one orthologous gene sets across 17 species. This concatenated alignment was used by the rapid hill climbing algorithm-based RAXML v8.2.12 for construction of maximum likelihood species phylogenetic tree with "PROTGAMMAAUTO" amino acid substitution model using 100 bootstrap values131.

Identification of genes with unique substitution having functional impact. The unique substitutions in genes that have impact on protein function can identify species-specific amino acid substitutions and are considered as a site-specific evolutionary signature. However, the inclusion of phylogenetically distant species in this analysis may erroneously increase the number of uniquely substituted genes; therefore, we restricted this analysis only considering the monocot species (available on the Ensembl plant release 47) for reliable results. The amino acid positions that were identical across the other 16 species in the individual multiple
sequence alignments of all orthogroups but different in C. longa were considered as the uniquely substituted amino acid positions.

For the identification of uniquely substituted sites, an in-house python script was used. Any gap and ten amino acid sites around any gap in the alignments were not considered in this analysis. The impact of these uniquely amino acid substitutions on protein function was identified using sorting intolerant from tolerant (SIFT), by utilizing UniProt as the reference database133,134.

Identification of positively selected genes. The nucleotide sequences of all the orthogroups across 17 species were aligned using MAFFT v7.670, 'codeml' from PAML package v4.9a that uses a branch-site model was used to identify positively selected genes using nucleotide alignments of all the orthologs in phylip format and the species phylogenetic tree generated in the previous steps135. Log-likelihood values were used to perform likelihood ratio tests and chi-square analysis-based p-values were calculated. The genes that qualified against the null model (fixed omega) (FDR-corrected p-values < 0.05) were identified as positively selected genes. All codon sites showing greater than 95% probability for foreground lineage based on Bayes Empirical Bayes (BEB) analysis were termed as positively selected sites.

Genes with multiple signs of adaptive evolution (MSA). Among the three signs of adaptive evolution—higher nucleotide divergence, unique substitution having functional impact and positive selection, the C. longa genes that showed at least two of these two signs were termed as genes with multiple signs of adaptive evolution or MSA genes136. MSA (multiple signs of adaptive evolution) genes are obtained by taking the intersection of the genes showing different evolutionary signatures, and because of the presence of more than one evolutionary signature, these genes can be considered as the highly evolved genes. Thus, these genes are useful to decipher and to strongly support the mechanisms or pathways responsible for adaptive evolution of the species.

Functional annotation. KAAS genome annotation server v2.1 was used to assign KEGG Orthology (KO) identifiers and KEGG pathways to the genes137. eggNOG-mapper v2 was used for functional annotation of genes using precomputed orthologous groups from eggNOG clusters138. WebGestalt web server was used for GO enrichment analysis, and only the GO categories showing p-values < 0.05 in over-representation enrichment analysis were considered further139. The assign-ment of genes into functional categories was manually curated.

Curcuminoid biosynthesis pathway. Coding sequences of four key genes involved in curcuminoid biosynthesis pathway, namely, curcumin synthase 1 (CURS1, NCBI accession number BA562262), curcumin synthase 2 (CURS2, NCBI accession number AB506762), curcumin synthase 3 (CURS3, NCBI accession number AB506763), and diketide-CoA synthase (DCS, NCBI accession number BA562253) were retrieved139. The sequences of these four genes were mapped to the gene set derived from de novo transcriptome assembly generated in this study, and the gene set derived from MAKER annotation pipeline using BLASTN with query coverage ≥50% and e-value 10−9. These sequences were also aligned to de novo genome assembly of C. longa constructed in this study, using Exonerate v2.2.0 (https://github.com/nathanwee/exonerate) with 95% of maximal alignment score and 95% quality threshold, and the best hits were selected to construct the gene structures.

Further, the ten enzymes involved in curcuminoid biosynthesis pathway (Fig. 3a) were searched, and identified in the prototype sequences of C. longa using CAPS protocol141. EC numbers or NCBI accession numbers of these ten enzymes (Supplementary Table 22) were used for homolog identification for each enzyme in UniProt database134, and the top hits that were found in UniProt database were retained. These homolog sequences were then aligned using Clustal Omega v1.2.4 (https://www.ebi.ac.uk/Tools/msa/clustalo/) (default parameters). In order to identify the true homologs, the functionally important residues (FIR)-binding site and active site amino acid residues for each enzyme (Supplementary Table 23) were detected from UniProt134, database, and sequences that did not contain those residues were removed from the alignments. These filtered alignments were queried against the prototype sequences of C. longa using PSI-BLAST24 with e-value of 10−3, inclusion threshold of 10−3, query coverage ≥70%, sequence identity ≥40%, and 2 iterations, as used in CAPS protocol85. The PSI-BLAST hits were again searched for the presence of FIRs, and the best identical hits were retained.

Evolution and phylogenetic analysis of curcuminoid biosynthesis pathway in C. longa. In order to elucidate the origin of the candidate enzymes involved in curcuminoid biosynthesis pathway, the ten genes identified in the previous step were used for phylogenetic analysis of these enzymes. The amino acid sequences of the identified genes were mapped against UniRef30 database235 using HHblits142 web server (default parameters). The top 20 hits were searched to extract one gene for each unique genus, and the target sequences were mapped for sequence domains using Pfam-A (v23.0) database, and only those sequences with the identified domains for each enzyme were selected as candidate homologs of the corresponding enzymes. The selected homologs were aligned using MAFFT v7.670, the empty sites were removed from the multiple sequence alignments using BeforePhylm v0.9.0, and the filtered alignments were used for construction of maximum likelihood-based gene phylogenetic tree for individual genes using RAxML v8.212 with bootstrap values of 1000 and “PROTGRAMMAUTO” amino acid substitution model.

CAFE v4.2.134 was used to analyse the evolution of the gene families that included the genes involved in curcuminoid biosynthesis pathway. The protein sequences of the selected 17 plant species (including Arabidopsis thaliana as an outgroup) were used for all-versus-all iBLASTp homology search, and subsequent clustering using MCL135. v14.137. After clustering, the gene sets contained ≥100 gene copies for at least one species were removed. The filtered gene families and the ultrametric species phylogenetetic tree were used for gene family expansion and contraction analysis using two- lambda (λ) model. In this two- lambda (λ) model, the clade formed by C. longa and Musa acuminata was assigned separate λ-value compared to the rest of the species (Supplementary Fig. 4).

Statistics and reproducibility. Computational data analyses were performed using Linux, Perl, and Python custom scripts. Statistical tests (chi-square, Bayes Empirical Bayes) was used in positive selection analysis were performed using PAML v4.9a135. Statistical significance levels are mentioned as p < 0.05. Statistically significant GO enriched categories were analyzed using WebGestalt web server138. Branch length distance values were calculated for higher nucleotide divergence analysis using ‘ade phylo’ package in R v3.6.0. For DNA–RNA extraction and sequencing, a single plant individual (n = 1) was used.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw genome and transcriptome reads of C. longa have been deposited in National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject accession—PRJNA669066, BioSample accession—SAMN1594602, SRA accessions—SRR12560783, SRR12560784, SRR12560785, SRR15204660, SRR15204661. Detailed information related to the MSA genes of C. longa have been provided in Supplementary Data 1.

Received: 29 September 2020; Accepted: 13 August 2021; Published online: 15 October 2021

References

1. Prasad, S. & Aggarwal, B. Turmeric, the Golden Spice. in Herbal Medicine: Biomolecular and Clinical Aspects 2nd edn, https://doi.org/10.1201/b10787-14 (2011).
2. Al-baititi, N. H. A study of preservative effects of sesame oil (Sesamum indicum L.) on mashed potatoes. Int. J. Sci. Res. Innov. Technol. 2, 6–10 (2015).
3. Chakraborty, A., Kundu, S., Mukherjee, S. & Ghosh, B. Endophytism in Zingiberales: Elucidation of Beneficial Impact. in Endophytes and Secondary Metabolites https://doi.org/10.1007/978-3-319-90484-9_31 (2019).
4. Kreymann, J. Natural diversity and adaptation of secondary metabolism.Curr. Opin. Plant Biol, https://doi.org/10.1016/j.pbi.2011.03.021 (2011).
5. Berini, J. et al. Combinations of abiotic factors differentially alter production of plant secondary metabolites in five woody plant species in the boreal-temperate transition zone. Front. Plant Sci, https://doi.org/10.3389/fpls.2018.01257 (2018).
6. Wink, M. Modes of action of herbal medicines and plant secondary metabolites. Medicines https://doi.org/10.3390/medicines2030251 (2015).
7. Koo, H. J. & Gang, D. R. Suites of terpene synthases explain differential terpenoid production in ginger and turmeric tissues. PLoS ONE https://doi.org/10.1371/journal.pone.0051481 (2012).
8. Sheeba, J. E., Deepa, K., Santhi, R. & Saiakumar, B. Comparative transcriptional analysis of two species of curcuma contrasting in a high-value compound: curcumin: insights into genetic basis and regulation of biosynthesis. Plant Mol. Biol. Report, https://doi.org/10.1105/0115S-05-0878-6 (2015).
9. Singh, N. & Sharma, A. Curcuminic (Curcuma longa): miRNAs and their regulating targets are involved in development and secondary metabolite production. C.R. Biol. https://doi.org/10.1016/j.crvi.2017.09.009 (2017).
10. Iurevna, J. S. Anti-inflammatory properties of curcumin, a major constituent of Curcuma longa: a review of preclinical and clinical research. Altern. Med. Rev. 14, 141–153 (2009).
11. Gupta, A. et al. Association of flavonoid-degrading plasmin, a flavonoid-degrading bacterium, with the gut microbiome of colorectal cancer patients in India. BioSystems https://doi.org/10.1155/2019/0438-19 (2019).
12. Korkina, L. G. Phenylpropanoids as naturally occurring antioxidants: from plant defense to human health. Cell. Mol. Biol. https://doi.org/10.1170/T772 (2007).
13. Annadurai, R. S. et al. De novo transcriptome assembly (NGS) of Curcuma longa L. Rhizome reveals novel transcripts related to anticancer and antimarial terpenoids. PLoS ONE https://doi.org/10.1371/journal.pone.0056217 (2013).

14. Zoroftichian Moghadamtousi, S. et al. A review on antibacterial, antiviral, and antifungal activity of curcumin. BioMed Res. Int. https://doi.org/10.1155/2014/186864 (2014).

15. Chattopadhyay, L., Biswas, K., Bandyopadhyay, U. & Aldebs, Y. Role of curcumin in disease prevention and treatment. Adv. Biomed. Res. https://doi.org/10.4103/abr.abr.147.16 (2018).

16. Nelson, K. M., et al. The essential medicinal chemistry of curcumin. J. Med. Chem. https://doi.org/10.1021/acsmedchemlett.6b00975 (2017).

17. Bhardwaj, R. S., Bhardwaj, K. S., Ranjeet, D. & Ganesh, N. Curcuma longa rhizome reveals novel transcripts related to anticancer and plant immunity by amino acid metabolic pathways. Plant Cell Environ. https://doi.org/10.1111/pce.12122 (2013).

18. Ruan, J. et al. Jasmonic acid signaling pathway in plants. Int. J. Mol. Sci. https://doi.org/10.3390/ijms10022479 (2009).

19. Jagodzik, P., Tajdel-Zielinska, M., Ciesla, A., Marczak, M. & Ludwikow, A. Mitogen-activated protein kinase cascades in plant hormone signaling. Front. Plant Sci. https://doi.org/10.3389/fpls.2018.01587 (2018).

20. Yu, M. H., Zhao, Z. Z. & He, J. X. Brassinosteroid signaling in plant–microbe interactions. Int. J. Mol. Sci. https://doi.org/10.3390/ijms19124091 (2018).

21. Zhou, X., Jiang, Y. & Yu, D. WRKY22 transcription factor mediates dark-induced leaf senescence in Arabidopsis. Mol. Cells https://doi.org/10.1007/s10059-011-0547-1 (2011).

22. Imran, Q. M. et al. Transcriptome profile of NO-induced Arabidopsis transcription factor genes suggests their putative regulatory role in multiple biological processes. Sci. Rep. https://doi.org/10.1038/s41598-017-18850-5 (2018).

23. Chen, H. et al. Ethylene insensitive3 and ethylene insensitive3-like repress salicylic acid induction deficient2 expression to negatively regulate plant innate immunity in Arabidopsis. Plant Cell https://doi.org/10.1105/tpc.111.097772 (2009).

24. Kojo, K. et al. Regulatory mechanisms of RIO1 generation are affected by rice spl mutations. Plant Cell Physiol. https://doi.org/10.1093/pcp/pcs074 (2006).

25. Vee, D. & Goring, D. R. The diversity of plant U-box E3 ubiquitin ligases: from upstream activators to downstream target substrates. J. Exp. Bot. https://doi.org/10.1093/jxb/erm369 (2009).

26. Moin, M. et al. Rice ribosomal protein large subunit genes and their spatio-temporal and stress regulation. Front. Plant Sci. https://doi.org/10.3389/fpls.2016.01284 (2016).

27. Nagaraj, S., Senthil-Kumar, M., Ramu, V. S., Wang, K. & Mysore, K. S. Plant ribosomal proteins, RPL12 and RPL19, play a role in nonhost disease resistance against bacterial pathogens. Front. Plant Sci. https://doi.org/10.3389/fpls.2015.01192 (2016).

28. Qi, F. & Zhang, F. Cell cycle regulation in the plant response to stress. Front. Plant Sci. https://doi.org/10.3389/fpls.2019.01765 (2020).

29. Bao, Z. & Hua, J. Interaction of CPR5 with cell cycle regulators UV14 and OSD1 in Arabidopsis. PLoS ONE (2014), https://doi.org/10.1371/journal.pone.0103347 (2014).

30. Miller, J. C., Chezem, W. R. & Clay, N. K. Terracyclic W40 repeat-containing protein complex: evolution, composition and roles in plant immunity. Front. Plant Sci. https://doi.org/10.3389/fpls.2015.01108 (2016).

31. Sharma, M. & Pandey, G. K. Expansion and function of repeat domain proteins during stress and development in plants. Front. Plant Sci. https://doi.org/10.3389/fpls.2015.01218 (2016).

32. Choe, S. et al. The DWF4 gene of Arabidopsis encodes a cytochrome P450 that mediates multiple 22a-hydroxylation steps in brassinosteroid biosynthesis. Plant Cell Physiol. https://doi.org/10.1093/pcp/25.10.231 (1998).

33. Sun, W. et al. Chalcone isomerase a key enzyme for anthocyanin biosynthesis in ophiirita japonica. Front. Plant Sci. https://doi.org/10.3389/fpls.2019.00865 (2019).

34. Corea, O. R. A., Bedgar, D. L., Davin, L. B. & Lewis, N. G. The arogenate dehydratase gene family: towards understanding differential regulation of carbon flux through phenylalanine into primary versus secondary metabolic pathways.
