


## Genome sequencing of turmeric provides evolutionary insights into its medicinal properties

Abhisek Chakraborty<sup>1</sup>, Shruti Mahajan<sup>1</sup>, Shubham K. Jaiswal<sup>1</sup> & Vineet K. Sharma<sup>1</sup>  <sup>✉</sup>

*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*.

<sup>1</sup>MetaBioSys Group, Department of Biological Sciences, Indian Institute of Science Education and Research Bhopal, Bhopal, India.

<sup>✉</sup>email: [vineetks@iiserb.ac.in](mailto:vineetks@iiserb.ac.in)

**T**urmeric, a common name for *C. longa*, has been traditionally used as a herb and spice for 4000 years in Southern Asia<sup>1</sup>. 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<sup>1,2</sup>. 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<sup>1,3</sup>. 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*<sup>3</sup>.

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<sup>4,5</sup>. These secondary metabolites also play a key role in plant defense mechanisms, and several of these metabolites have numerous pharmacological applications in herbal medicines and phytotherapy<sup>6</sup>. The pathways for biosynthesis of various secondary metabolites including phenylpropanoids, flavonoids (such as curcuminoids), terpenoids, and alkaloids are found in *C. longa*<sup>7–10</sup>. Other constituents such as volatile oils, proteins, resins, and sugars are also present in *C. longa*<sup>10</sup>. Flavonoids are known to have anti-inflammatory, antioxidant, and anti-cancer activities<sup>11</sup>. Phenylpropanoids are also of great importance because of their antioxidant, anti-cancer, anti-microbial, anti-inflammatory, and wound-healing activities<sup>12</sup>. The other class of secondary metabolites, terpenoids, are also known to possess anti-cancer and anti-malarial properties<sup>13</sup>.

The three curcuminoids namely curcumin, demethoxycurcumin, and bisdemethoxycurcumin, are responsible for the yellow color of turmeric<sup>10</sup>. Among these, the primary bioactive component of turmeric is curcumin, which is a polyphenol-derived flavonoid compound and also known as diferuloylmethane<sup>10</sup>. Curcumin shows broad-spectrum antimicrobial properties against bacteria, fungi and viruses<sup>14</sup>, and also possesses anti-diabetic, anti-inflammatory, antifertility, anti-coagulant, hepatoprotective, and hypertension protective properties<sup>10,15</sup>. Being an excellent scavenger of reactive oxygen species (ROS) and reactive nitrogen species, its antioxidant activity also controls DNA damage by lipid peroxidation mediated by free-radicals, and thus provides it with anti-carcinogenic properties<sup>16</sup>. 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<sup>17–19</sup>. 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<sup>17</sup>. 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<sup>20</sup>.

Several studies have been carried out to study the secondary metabolites and medicinal properties of this plant<sup>21,22</sup>. 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<sup>8,13,23,24</sup>. 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<sup>25</sup>, *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<sup>26</sup>. Recent studies showed evidence for a ploidy level of  $3X$  ( $2n = 63$  chromosomes, basic chromosome number  $X = 21$ )<sup>27,28</sup>.

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<sup>8,13,23,24</sup>, and one study also reported the transcriptome of leaf tissue<sup>29</sup>. 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<sup>30</sup>. 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 Trimmomatic 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<sup>30</sup> with barcode-filtered 10x Genomics linked reads, which is close to the previously estimated genome size of 1.33 Gbp<sup>25</sup>. *C. longa* genome was estimated to contain 4.83% heterozygosity. The *C. longa* genome sequence, assembled using Supernova v2.1.1<sup>31</sup> and Flye v2.4.2<sup>32</sup> 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  $\geq 3000$  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 ( $\geq 35$  Kbp) that covered 72.2% of the estimated genome size was calculated as 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  $\Delta\log$ -likelihood value for the triploid fixed model, before and after denoising (see “Methods”)<sup>33</sup> (Supplementary Fig. 2a). Also,

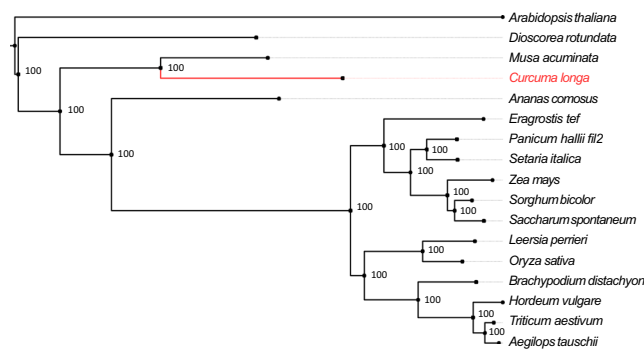
heterozygous k-mer pair coverage pattern distribution using Smudgeplot<sup>34</sup> (see “Methods”) showed that 83% of the k-mer pairs corroborated to total coverage of k-mer pair = 3n and normalized minor k-mer coverage = 1/3, and thus the genome was inferred as triploid (Supplementary Fig. 2b).

The de novo transcriptome assembly of *C. longa* (from this study) using Trinity v2.9.1<sup>35</sup> had a total size of 86,158,097 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<sup>36</sup> 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 *C. longa* genome by RepeatModeler v2.0.1<sup>37</sup>, 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 *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 9.42% Gypsy/DIRS1 elements) (Supplementary Table 7). Additionally, 7.64% of *C. longa* genome was identified as simple repeats using TRF v4.09<sup>38</sup>. Thus, ~70% of the genome was predicted to be constituted of simple and interspersed repetitive sequences. Among the non-coding RNAs, 1826 standard amino acid specific tRNAs (transfer RNAs) and 335 hairpin miRNAs (micro RNAs) were predicted in *C. longa* genome.

MAKER genome annotation pipeline<sup>39</sup> and TransDecoder v5.5.0 predicted a total of 57,486 and 23,943 coding sequences from genome and transcriptome assemblies, respectively. Length-based filtering criteria ( $\geq 150$  bp) of the above two coding gene sets resulted in 56,043 and 23,943 coding sequences from genome and transcriptome assemblies, respectively. MAKER derived filtered coding sequences were further clustered at 95% sequence identity, which resulted in 45,307 non-redundant coding sequences. These two coding gene sets were merged using BLASTN, resulting in the final *C. longa* gene set comprising of 50,401 coding gene sequences. BUSCO analysis revealed presence of 85.8% BUSCO genes (complete + fragmented) in this coding gene set of *C. longa*. Among these 50,401 genes, a total of 45,479 genes, 36,724 genes and 33,338 genes were mapped to NCBI (nr), Swiss-Prot, and Pfam-A (v32.0) databases, respectively. A total of 45,576 genes (90.43%) could be annotated against any of these three databases, and 4825 (9.57%) genes remained unannotated. Further investigations of these coding gene sequences revealed 56,936 variable nucleotide sites out of total 53,278,136 bases (0.11%) in the coding gene set. A total of 9951 out of 50,401 coding genes showed sequence variation.

**Resolving the phylogenetic position of *C. longa*.** From the selected 17 plant species, a total of 104,746 orthogroups were identified using protein sequences by OrthoFinder v2.3.9<sup>40</sup>. Among these orthogroups, 5388 contained protein sequences from all 17 species, and were used for evolutionary analysis. Further, KinFin v1.0<sup>41</sup> predicted a total of 1053 fuzzy one-to-one orthogroups containing protein sequences from all 17 selected species, which were used to construct the maximum likelihood-



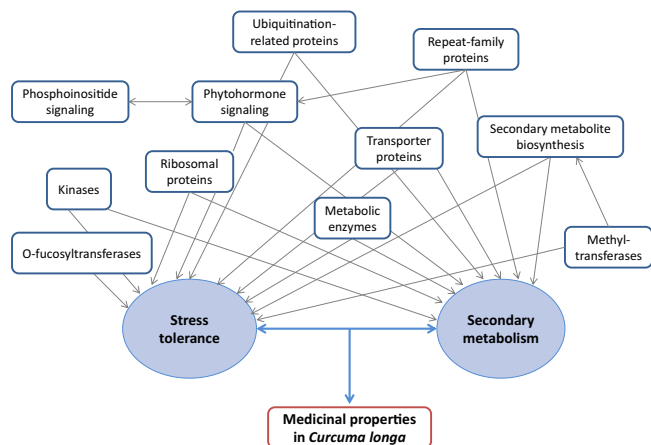
**Fig. 1 Phylogenetic position of *C. longa*.** Phylogenetic tree of *C. longa* with 15 other selected species and *Arabidopsis thaliana* as an outgroup species. The values mentioned at the nodes correspond to the bootstrap values.

based phylogenetic tree of *C. longa* with 15 other monocot species and *Arabidopsis thaliana* as an outgroup.

All 1053 fuzzy one-to-one orthogroups were aligned, concatenated and filtered for undetermined or missing values. This filtered alignment data consisting of 851,765 alignment positions was used to construct the maximum likelihood-based species phylogenetic tree of *C. longa* with 15 other monocot species available on Ensembl plants release 47 and *Arabidopsis thaliana* as the outgroup (Fig. 1). Position of *C. longa* and the other selected monocots in this genome-wide phylogeny were supported by previously reported phylogenies<sup>42–45</sup>. In our phylogeny, *Ananas comosus* showed an earlier divergence among the monocots from Poales order, which was also supported by previously reported studies<sup>42,45</sup>. Among all selected monocot species in our study, species from Dioscoreales order showed the earliest divergence, supported by the previously reported studies<sup>44,45</sup>. From our genome-wide phylogeny constructed using the selected monocots, it is apparent that *Musa acuminata* was comparatively closer to *C. longa*. Belonging to the same phylogenetic order Zingiberales, *C. longa* and *Musa acuminata* shared the same clade in the species phylogenetic tree. Species from the Zingiberales order are closer related to species from the Poales order in comparison to the species from the Dioscoreales order.

**Genes with signatures of adaptive evolution.** Genes with site-specific signatures of adaptive evolution were identified in *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 *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. 2 Functional categories of MSA genes.** Functional association of MSA genes with medicinal properties of *C. longa*.

orthologous groups) categories and GO (Gene ontology) enrichment categories (Supplementary Tables 8–20). A total of 172 out of 188 MSA genes (Supplementary Data 1) were associated with categories essential for plant secondary metabolism and defense responses against environmental stress conditions i.e., biotic stress and abiotic stress (Fig. 2).

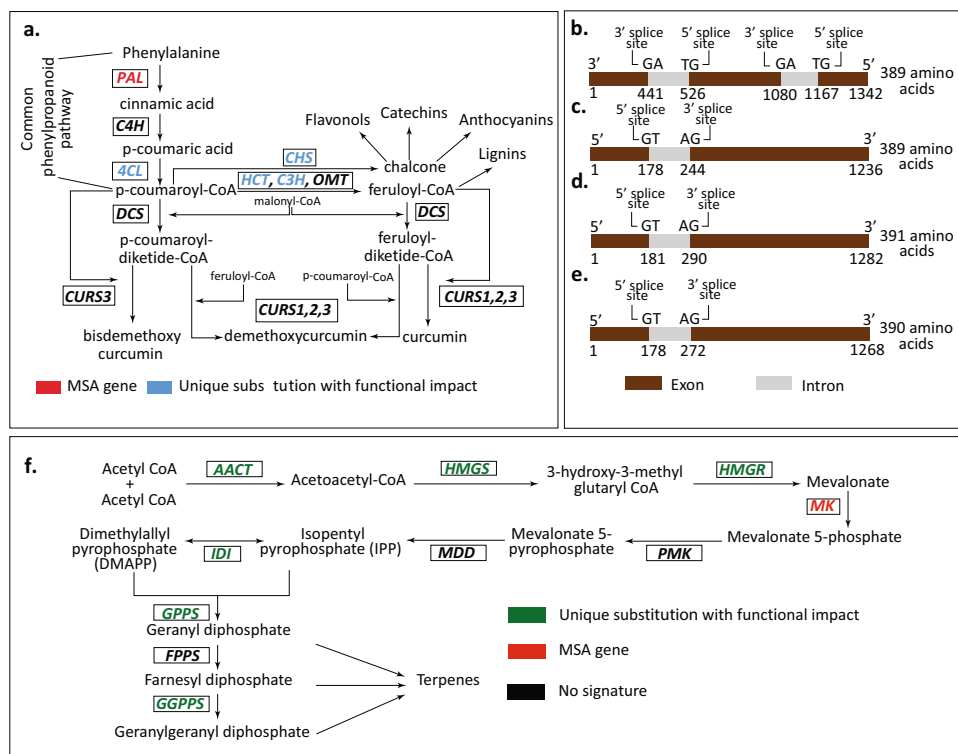
**Adaptive evolution of plant defense associated genes.** Several genes known to provide plant immunity against pathogen infection or disease were found in MSA genes. Since plants lack adaptive immune response, the innate immune response in plants is provided by PAMP-triggered (PTI) and effector-triggered (ETI) immunity with the help of two plant stress hormones, salicylic acid (SA) and jasmonic acid (JA) signaling pathway<sup>46</sup>. 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<sup>47–49</sup>. Also, *WRKY* transcription factor is induced by pathogen attack and is involved in SA-signaling pathway mediated plant immunity, *MYB48* (Myeloblastosis 48) 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<sup>50–52</sup>. Jasmonic acid and salicylic acid elicit the production and accumulation of secondary metabolites such as phenolics, terpenoids, alkaloids, and glycosides, in medicinal plants<sup>53</sup>.

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<sup>54</sup>. 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<sup>55,56</sup>. These ubiquitin ligase proteins also regulate various abiotic stress responses e.g., drought, salinity, temperature<sup>56</sup>. Among the MSA genes, three ribosomal subunit proteins are regulated by signaling molecules such as methyl jasmonate, salicylic acid and environmental stress e.g., cold, heat, ultraviolet (UV), drought, salinity<sup>57,58</sup>. 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<sup>59,60</sup>. 19 different repeat family genes belonging to WD-family repeats, leucine-rich repeats (LRR), pentatricopeptide repeats (PPR), tetratricopeptide repeat (TPR)-like superfamily repeats, armadillo (ARM) superfamily repeats and ankyrin repeats were identified as MSA genes. These repeat family proteins are involved in plant innate immunity as well as abiotic stress tolerance<sup>61,62</sup>.

**Abundance of secondary metabolism pathways in *C. longa* genome.** Genes associated with secondary metabolite biosynthesis and secondary metabolism were found to be abundant (104 out of 188 genes) among the MSA genes in *C. longa*. Of these, *DWF4* (Dwarf 4) is involved in brassinosteroid biosynthesis, *CHI* (chalcone isomerase) plays a key role in anthocyanin biosynthesis, *ADT* (arogenate dehydratase) has a role in lignin biosynthesis, *GST* (glutathione S-transferase) aids in glucosinolate biosynthesis<sup>63–66</sup>. Shikimate kinase, which is a central enzyme involved in shikimic acid pathway required for production of a wide range of phenolic group of secondary metabolites<sup>67</sup>, was identified as an MSA gene. Methyltransferases e.g., *AT1G04430*, serine hydroxymethyltransferase, lysine methyltransferase-like, Protein arginine methyltransferase are also involved in secondary metabolites biosynthesis and responsible for methylation of secondary metabolites, which in turn assist in disease resistance and stress tolerance in plants<sup>68,69</sup>. These secondary metabolites possess various medicinal applications such as anti-inflammatory, antimicrobial, antioxidant, and anti-carcinogenic properties<sup>6,12,70</sup>. The functional role of selected MSA genes involved in secondary metabolite biosynthesis and medicinal properties of the resultant metabolites are mentioned in Supplementary Table 21. Transporter genes e.g., three ATP-binding cassette proteins, *KT1* (K<sup>+</sup> Transporter 1) and *KEA3* (K<sup>+</sup> Efflux Antiporter 3) potassium transporters, *NRT* (nitrate transporter), *SULTR2;1* (sulfate transporter 2;1) also showed multiple signs of adaptation. The evolution of these transporter genes appears important since the alteration in concentration of metal ions regulate accumulation and translocation of secondary metabolites<sup>71–73</sup>.

Polyketides, such as curcuminoids, represent a diverse class of secondary metabolites, which are crucial for plants survival under environmental challenges<sup>74</sup>. Curcuminoid biosynthesis pathway is the most important secondary metabolism pathway in *C. longa* where phenylalanine is converted to coumaroyl-CoA via cinnamic acid and coumaric acid<sup>75</sup>. Conversion of phenylalanine to coumaroyl-CoA is also part of phenylpropanoid biosynthesis pathway<sup>76</sup>. The two key enzymes *PAL* (phenylalanine ammonia lyase) and *4CL* (4-coumarate-CoA ligase), in this step showed evolutionary signatures. *4CL* gene possessed unique amino acid substitution with functional impact. Notably, the *PAL* gene, which is involved in conversion of phenylalanine to cinnamic acid in the curcuminoid biosynthesis pathway<sup>75</sup>, showed positive selection and unique amino acid substitution in this study, and thus was identified as an MSA gene. Also, the *C3H* (cinnamate-3-hydroxylase) and *HCT* (hydroxycinnamoyl transferase) genes that are involved in conversion of coumaroyl-CoA to



**Fig. 3** Pathways and genes involved in secondary metabolism of *C. longa*. **a** Curcuminoid biosynthesis pathway in *C. longa*<sup>75,76,79</sup>. PAL = phenylalanine ammonia lyase, C4H = cinnamate-4-hydroxylase, 4CL = 4-coumarate-CoA ligase, HCT = hydroxycinnamoyl transferase, C3H = cinnamate-3-hydroxylase, OMT = O-methyltransferase, CHS = chalcone synthase, DCS = diketide-CoA synthase, CURS1 = curcumin synthase 1, CURS2 = curcumin synthase 2, CURS3 = curcumin synthase 3. **b** Structure of *DCS* gene. **c** Structure of *CURS1* gene. **d** Structure of *CURS2* gene. **e** Structure of *CURS3* gene. **f** Evolutionary signatures in mevalonate pathway of terpenoid biosynthesis<sup>80</sup>. 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.

feruloyl-CoA showed unique amino acid substitution with functional impact (Fig. 3a).

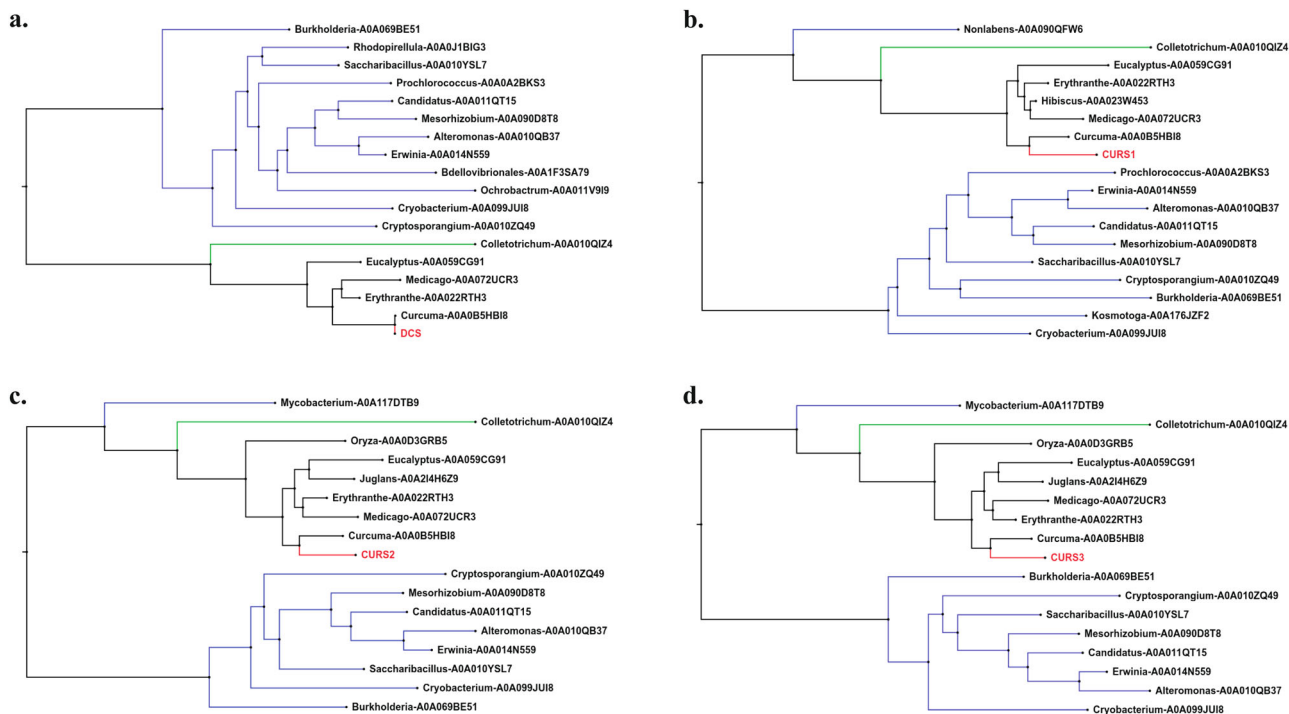
Further, coumaroyl-CoA and feruloyl-CoA are used for the production of curcumin, demethoxycurcumin, and bisdemethoxycurcumin via coumaroyl-diketide-CoA and feruloyl-diketide-CoA, catalyzed by four enzymes—*CURS1* (curcumin synthase 1), *CURS2* (curcumin synthase 2), *CURS3* (curcumin synthase 3), and *DCS* (diketide-CoA synthase)<sup>75</sup>. 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* genes on these assemblies. All four enzymes were found to be present in the de novo genome assembly, in the gene set derived from MAKER pipeline, and in the gene set derived from de novo transcriptome assembly of *C. longa*. Using Exonerate, we further constructed the gene structures 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*) family<sup>77</sup>, and the genes from *CHS* family generally consist of two exons and one intron<sup>78</sup>, 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 catechins<sup>76</sup>. 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 synthesis<sup>76</sup>, possessed unique amino acid substitution. Another intermediate of curcuminoid biosynthesis pathway, feruloyl-CoA, is a precursor for lignin biosynthesis<sup>79</sup>. 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* (cinnamoyl-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 biosynthesis<sup>80</sup> 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), *IDI* (isopentenyl diphosphate isomerase), *GPPS* (geranyl diphosphate synthase), *GGPPS* (geranylgeranyl diphosphate synthase) showed unique amino acid substitutions with functional impact. *MK* (mevalonate kinase) exhibited both unique substitution and positive selection, and thus was found among the MSA genes.

**Evolution of curcuminoid biosynthesis pathway.** The ten enzymes involved in curcuminoid biosynthesis pathway were analyzed to elucidate their origin by remote homology finding and to examine their phylogenetic position with respect to the genes from other plant species (see “Methods”). These enzymes were identified in the coding gene set of *C. longa* using CAPS\_protocol<sup>81</sup>. Among the ten enzymes, *HCT* had all the phylogenetically closer orthologs (transferases) from angiosperm plant division. However, *C4H* (cinnamate-4-hydroxylase) had a Bryophyte ortholog and *C3H* had a Gymnosperm ortholog, along with other angiosperm orthologs (monooxygenase and oxidoreductase) (Supplementary Fig. 3). The remaining seven enzymes



**Fig. 4** Phylogenetic relationships of candidate enzymes of lower curcuminoid biosynthesis pathway in *C. longa* with their distant orthologous genes. **a** *DCS* gene. **b** *CURS1* gene. **c** *CURS2* gene. **d** *CURS3* gene. Blue colored line denotes Bacterial orthologs, dark green colored line denotes fungal orthologs, black colored line denotes angiosperm orthologs, and red colored line denotes the genes of interest in *C. longa*.

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 bacteria<sup>82</sup>, 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 pathways<sup>83</sup>.

Gene family evolution analysis (using CAFÉ v4.2.1)<sup>84</sup> 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 applications<sup>15</sup>. 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 studies<sup>27,85,86</sup>. 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 studies<sup>27,28,85,86</sup>. The application of Oxford Nanopore long-reads and 10x Genomics linked read sequencing that has the potential to resolve complex polyploid genomes<sup>87</sup>, 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 processing<sup>88</sup>. It is noteworthy that the LAI value of *C. longa* genome ( $\geq 35$  Kbp) was estimated at 10.26, which also corresponds to a reference quality genome assembly<sup>89</sup>.

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, anti-cancer, and many other medicinal properties<sup>90</sup>.

Repetitive sequence prediction revealed that ~70% of the genome consisted of repeat elements, which is similar to other

plant genome, such as *Triticum urartu*<sup>91</sup>. 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 acuminata* species from the same Zingiberales plant order, and thus appears to be a specific signature of repeat elements in Zingiberales order<sup>92</sup>.

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 phytocystatin gene *CypC1*<sup>42,43</sup>. Ren et al. also showed similar phylogenetic position of *C. longa* with other selected monocots – *Dioscorea sp.*, *Musa acuminata*, *Brachypodium distachyon*, *Oryza sativa*, *Panicum sp.*, *Setaria italica*, *Zea mays*, *Sorghum bicolor* using genome and transcriptome data of 105 angiosperms<sup>44</sup>. Also, an updated megaphylogeny for vascular plants showed similar relative phylogenetic position of *C. longa* with respect to the selected monocots<sup>45</sup>. Further, selected species from Poales order also showed similar relative positions with respect to each other<sup>45</sup>. Absence of any polytomy 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 orthologous 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 response<sup>46</sup>, and also affect plant secondary metabolism by regulating the production of secondary metabolites<sup>53</sup>, 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*<sup>93</sup>. 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 other 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 phenolic group of secondary metabolites (e.g., curcuminoids,

anthocyanins, lignins, phenylpropanoids) 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 responses<sup>94,95</sup>. 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 conditions<sup>96</sup>, 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 (23.2280252°N 77.2088987°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 MinION platform using FLO-MIN106 flow cell. Using the same plant sample, RNA extraction was carried out from the powdered 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 Ribo-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/proc10xG>). The genome size of *C. longa* was estimated using a k-mer count distribution method implemented in SGA-preqc<sup>30</sup> (Supplementary Notes 2). A total of 631.11 million raw 10x Genomics linked reads corresponding 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 settings<sup>31</sup>. 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/genome-exome/software/pipelines/latest/installation>) for barcode processing and were used to detect and correct mis-assemblies in Supernova assembled genome using Tigmint v1.1.2<sup>97</sup>. The first round of scaffolding was carried out using ARCS v1.1.1 (default parameters) to generate more contiguous assembly using 10x Genomics linked reads<sup>98</sup>. 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 assembly<sup>99</sup>. Adapter-processed Nanopore long-reads (>20 Kb) were also used for scaffolding of the genome assembly using LINKS v1.8.6 with default parameters<sup>100</sup>.

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 Flye<sup>32</sup> 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 Pilon<sup>101</sup> 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 v1.1.1<sup>98</sup>, AGOUTI v0.3.3<sup>99</sup>, and LINKS v1.8.6<sup>100</sup>, 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 assembly<sup>102</sup>. Gap-closing of this scaffolded assembly was performed with barcode-processed linked reads using Sealer v2.1.5 with k-mer value from 30 to 120 with an interval of 10 bp using a Bloom filter-based approach<sup>103</sup>, and LR\_Gapcloser<sup>104</sup> with Nanopore long-reads. Finally, the assembly quality was improved by Pilon v1.23 using barcode-processed linked reads to fix small indels, individual base errors, or local mis-assemblies that could be introduced by the previous scaffolding steps<sup>101</sup>. The other details about the genome assembly post-processing are mentioned in Supplementary Notes 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<sup>105</sup> (v0.7.17), Minimap2<sup>106</sup> (v2.17), and HISAT2<sup>107</sup> (v2.2.1) respectively, and the mapping statistics was calculated using samtools<sup>108</sup> (v1.9) “flagstat”. BUSCO v5.2.1 was used to assess the genome assembly completeness using embryophyta\_odb10 database<sup>109</sup>. Also, LTR Assembly Index (LAI) was used to assess the assembly continuity using LTR retrotransposons (LTR-RTs) that are highly abundant and difficult to assemble in repeat rich plant genomes<sup>89</sup>. GenomeTools<sup>110</sup> v1.6.1, and LTR\_retriever<sup>111</sup> v2.9.0 (default parameters) was used to calculate LAI score in the final *C. longa* genome assembly.

The genome ploidy was estimated with a statistical approach using Gaussian Mixture Model (GMM), implemented in nQuire<sup>33</sup>. After removing the bacterial sequences, 10x Genomics linked reads were mapped to the final draft genome of *C. longa* using BWA-MEM v0.7.17<sup>105</sup> with default parameters, and samtools v1.9<sup>108</sup> was used to generate the sorted and indexed alignments. These alignments were processed using nQuire with default parameters to extract the variable sites with the free model and the fixed models. Distribution of these base frequencies was denoised and both the distributions, before and after denoising, were used to estimate Alog-likelihood values for the fixed models. Also, Smudgeplot<sup>34</sup> v0.2.2 was used to infer the genomic ploidy of *C. longa*. First, barcode-filtered 10x Genomics linked reads were used for k-mer counting, and k-mer frequency-based histogram generation using KMC<sup>112</sup> 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 excluding k-mers occurring more than 10,000 times. After extracting the k-mers between an upper and lower coverage range, Smudgeplot was used for computing the k-mer pairs set, and ploidy estimation.

The heterozygosity of *C. longa* genome was analyzed in order to assess the genomic complexity, and ease of genome assembly. The k-mer frequency-based histogram that was generated using KMC v3.1.1 in the previous step was used to estimate the heterozygosity content using GenomeScope v2.0<sup>113</sup> with triploid model.

**Transcriptome assembly.** The RNA-Seq data from our study and from previously available transcriptome studies of *C. longa* were used for de novo transcriptome assembly<sup>8,13,23,24,29</sup>. Trimmomatic v0.38 was used for adapter removal and quality filtration of raw Illumina sequence data<sup>114</sup> (Supplementary Notes 2). Finally, Trinity v2.9.1 was used with default parameters to perform de novo transcriptome assembly of quality-filtered paired-end and single-end reads<sup>35</sup>.

Further, the quality-filtered paired-end and single-end RNA-Seq reads obtained from our study were separately used for de novo transcriptome assembly using Trinity v2.9.1<sup>35</sup> with strand-specific parameter and other default settings. The strand-specificity parameter “--SS\_lib\_type” in Trinity v2.9.1 was used since stranded RNA-Seq data aids in more accurate transcriptome assembly and gene prediction by retaining the information of the overlapping genes from which the transcript arises<sup>115</sup>. Only the transcripts with length  $\geq 500$  bp were retained, and assembly statistics was calculated using a Perl script available in Trinity software package. Since this transcriptome assembly was obtained for the same individual plant that was used for genomic DNA extraction as well as genome assembly, it was used for downstream coding gene prediction steps to avoid any ambiguity. The longest isoforms for all genes were extracted from the assembled transcripts. These sequences were clustered to identify the unigenes using CD-HIT-EST v4.8.1 (90% sequence identity, 8 bp seed size)<sup>36</sup>.

**Genome annotation.** The final polished assembly (contigs with length of  $\geq 1000$  bp after scaffolding) was used for genome annotation. RepeatModeler v2.0.1 was used to generate a de novo repeat library for this genome<sup>37</sup>. The resultant repeat sequences were further clustered using CD-HIT-EST v4.8.1 (90% sequence identity, seed size = 8 bp)<sup>36</sup>. This repeat library was used to soft-mask *C. longa* genome using RepeatMasker v4.1.0 (<http://www.repeatmasker.org>) and was further used for gene set construction. Genome annotation was carried out using MAKER pipeline to predict the final gene models using ab initio gene prediction programs and evidence-based approaches<sup>39</sup>. The transcriptome assembly of *C. longa* using previously available data along with data from our study, and protein sequences of *C. longa* along with its closest species *Musa acuminata* were used as empirical evidence in MAKER pipeline. AUGUSTUS v3.2.3, BLAST and Exonerate v2.2.0 were used in MAKER pipeline for ab initio gene prediction, evidence alignments and alignments polishing, respectively<sup>116,117</sup>. Tandem repeat finder (TRF) v4.09 was

used to detect the tandem repeats present in *C. longa* genome<sup>38</sup>. Additionally, miRBase database and tRNAscan-SE v2.0.5 were used for homology-based identification of miRNAs and de novo prediction of tRNAs, respectively<sup>118,119</sup>. The other details about genome annotation are mentioned in Supplementary Notes 3.

**Construction of gene set.** The coding genes were predicted by TransDecoder v5.5.0 from the unigenes identified in transcriptome assembly obtained from our sequencing data, using Uniref90 and Pfam (v32.0) databases for homology-based searching as ORF (open reading frame) retention criteria (<https://github.com/TransDecoder/TransDecoder>)<sup>120,121</sup>. Also, the MAKER pipeline-based gene models were clustered using CD-HIT-EST v4.8.1 (95% sequence identity, 8 bp seed size)<sup>36</sup>. The TransDecoder pipeline derived genes ( $\geq 150$  bp) were aligned against the MAKER derived gene set ( $\geq 150$  bp) using BLASTN, and the genes that did not match with identity  $\geq 50\%$ , query coverage  $\geq 50\%$  and e-value  $10^{-9}$  were added to the MAKER gene set to prepare the final gene set of *C. longa*<sup>122,123</sup>. The two coding gene sets were merged to consider any unique coding gene that was not predicted from de novo genome assembly using MAKER pipeline<sup>122</sup>.

Further, BUSCO v5.2.1 analysis was performed using embryophyta\_odb10 dataset<sup>109</sup>, to assess the quality, and completeness of the final coding gene set. For functional annotation, this gene set was mapped against the Swiss-Prot<sup>124</sup>, NCBI non-redundant (nr) database using BLASTP (e-value cut-off  $10^{-5}$ ), and Pfam<sup>121</sup> (v32.0) database using HMMER<sup>125</sup> 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, duplicate reads were identified and removed from the quality-filtered RNA-Seq reads using FastUniq v1.1<sup>126</sup>. The resultant reads were mapped to the coding genes using BWA-MEM v0.7.17<sup>105</sup>, and SAMtools v1.9<sup>108</sup> was used to generate the alignment in BAM format. Using this alignment, BCFtools<sup>127</sup> (v1.9) “mpileup” was used for variant calling in the coding genes, and for further filtering of false-positive variants based on the following parameters<sup>128</sup>—mapping quality  $\geq 50$ , variant sites with quality  $\geq 30$ , sequencing depth  $\geq 30$ .

**Orthogroups 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 orthogroups identification<sup>129</sup>. To construct the orthogroups, the protein sequences of *C. longa* obtained from TransDecoder and proteome files for other 16 selected species i.e., *Aegilops tauschii*, *Ananas comosus*, *Brachypodium distachyon*, *Dioscorea rotundata*, *Eragrostis tef*, *Hordeum vulgare*, *Leersia perrieri*, *Musa acuminata*, *Oryza sativa*, *Panicum hallii fil2*, *Saccharum spontaneum*, *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 orthogroups using OrthoFinder v2.3.9<sup>40</sup>.

**Construction of orthologous gene set and phylogenetic tree.** Only those orthogroups that contained genes from all 17 species were extracted from all the identified orthogroups. The fuzzy one-to-one orthogroups containing genes from all 17 species were identified from these orthogroups, and extracted using KinFin v1.0<sup>41</sup>. For cases where the orthologous gene sets comprised of multiple genes for a species, the longest gene was extracted. The fuzzy one-to-one orthogroups were further aligned individually using MAFFT v7.467 for species phylogenetic tree construction<sup>130</sup>. BeforePhylo v0.9.0 (<https://github.com/qiyunzhu/BeforePhylo>) was used to trim the multiple sequence alignments to remove empty sites and to concatenate the 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 values<sup>131</sup>.

**Identification of genes with higher nucleotide divergence.** Protein sequences of all the orthogroups across 17 species were aligned individually using MAFFT v7.467<sup>130</sup>, and individual maximum likelihood-based phylogenetic trees were built using these alignments by RAXML v8.2.12 (‘PROTGAMMAAUTO’ amino acid substitution model, 100 bootstrap values)<sup>131</sup>. The ‘adephylo’ package in R was used to calculate root-to-tip branch length distance values for each species<sup>132</sup>. *C. longa* genes that showed comparatively higher root-to-tip branch length with respect to the other selected species were extracted, and were considered as the genes with higher nucleotide divergence or higher rate of evolution.

**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 by 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 unique amino acid substitutions on protein function was identified using sorting intolerant from tolerant (SIFT), by utilizing UniProt as the reference database<sup>133,134</sup>.

**Identification of positively selected genes.** The nucleotide sequences of all the orthogroups across 17 species were aligned using MAFFT v7.467<sup>130</sup>. '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 steps<sup>135</sup>. 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 signs were termed as genes with multiple signs of adaptive evolution or MSA genes<sup>136</sup>. 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 genes<sup>137</sup>. eggNOG-mapper v2 was used for functional annotation of genes using precomputed orthologous groups from eggNOG clusters<sup>138</sup>. 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 further<sup>139</sup>. The assignment 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 BAH56226), curcumin synthase 2 (*CURS2*, NCBI accession number AB506762), curcumin synthase 3 (*CURS3*, NCBI accession number AB506763), and diketide-CoA synthase (*DCS*, NCBI accession number BAH56225) were retrieved<sup>140</sup>. 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<sup>117</sup> with query coverage  $\geq 50\%$  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/nathanweeks/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 proteome sequences of *C. longa* using CAPS\_protocol<sup>81</sup>. EC numbers or NCBI accession numbers of these ten enzymes (Supplementary Table 22) were used for homolog identification for each enzyme in UniProt database<sup>134</sup>, 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 UniProt<sup>134</sup> database, and sequences that did not contain those residues were removed from the alignments. These filtered alignments were queried against the proteome sequences of *C. longa* using PSI-BLAST<sup>141</sup> with  $e$ -value of  $10^{-5}$ , inclusion threshold of  $10^{-5}$ , query coverage  $\geq 70\%$ , sequence identity  $\geq 40\%$ , and 2 iterations, as used in CAPS\_protocol<sup>81</sup>. 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 database<sup>120</sup> using HHblits<sup>142</sup> 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 (v32.0) 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 v7.467, the empty sites were removed from the multiple sequence alignments using BeforePhylo v0.9.0, and the filtered alignments were used for construction of

maximum likelihood-based gene phylogenetic tree for individual genes using RAxML v8.2.12 with bootstrap values of 1000 and 'PROTGAMMAAUTO' amino acid substitution model.

CAFÉ v4.2.184 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 BLASTP homology search, and subsequent clustering using MCL<sup>143</sup> v14.137. After clustering, the gene families that contained  $\geq 100$  gene copies for at least one species were removed. The filtered gene families and the ultrametric species phylogenetic tree were used for gene family expansion and contraction analysis using two-lambda ( $\lambda$ ) model. In this two-lambda ( $\lambda$ ) model, the clade formed by *C. longa* and *Musa acuminata* was assigned separate  $\lambda$ -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) used in positive selection analysis were performed using PAML v4.9a<sup>135</sup>. Statistical significance levels are mentioned as  $p < 0.05$ . Statistically significant GO enriched categories were analyzed using WebGeStalt web server<sup>139</sup>. Branch length distance values were calculated for higher nucleotide divergence analysis using 'adephylo'<sup>132</sup> 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—PRJNA660606, BioSample accession—SAMN15954062, 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

- 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).
- Al-bahititi, 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).
- Chakraborty, A., Kundu, S., Mukherjee, S. & Ghosh, B. Endophytism in Zingiberaceae: Elucidation of Beneficial Impact. in *Endophytes and Secondary Metabolites* [https://doi.org/10.1007/978-3-319-90484-9\\_31](https://doi.org/10.1007/978-3-319-90484-9_31) (2019).
- Kroymann, J. Natural diversity and adaptation in plant secondary metabolism. *Curr. Opin. Plant Biol.* <https://doi.org/10.1016/j.pbi.2011.03.021> (2011).
- Berini, J. L. 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).
- Wink, M. Modes of action of herbal medicines and plant secondary metabolites. *Medicines* <https://doi.org/10.3390/medicines2030251> (2015).
- 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).
- Sheeja, T. E., Deepa, K., Santhi, R. & Sasikumar, B. Comparative transcriptome 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/10.1007/s11105-015-0878-6> (2015).
- Singh, N. & Sharma, A. Turmeric (*Curcuma longa*): miRNAs and their regulating targets are involved in development and secondary metabolite pathways. *C R Biol.* <https://doi.org/10.1016/j.crvi.2017.09.009> (2017).
- Jurenka, 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).
- Gupta, A. et al. Association of flavonifractor plautii, a flavonoid-degrading bacterium, with the gut microbiome of colorectal cancer patients in India. *mSystems* <https://doi.org/10.1128/mSystems.00438-19> (2019).
- 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 antimalarial terpenoids. *PLoS ONE* <https://doi.org/10.1371/journal.pone.0056217> (2013).
14. Zorofchian 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, I., Biswas, K., Bandyopadhyay, U. & Banerjee, R. K. Turmeric and curcumin: biological actions and medicinal applications. *Curr. Sci.* **87**, 44–53 (2004).
16. Rahmani, A., Alsahli, M., Aly, S., Khan, M. & Aldebasi, Y. Role of curcumin in disease prevention and treatment. *Adv. Biomed. Res.* [https://doi.org/10.4103/abr.abr\\_147\\_16](https://doi.org/10.4103/abr.abr_147_16) (2018).
17. Nelson, K. M. et al. The essential medicinal chemistry of curcumin. *J. Med. Chem.* <https://doi.org/10.1021/acs.jmedchem.6b00975> (2017).
18. Baker, M. Deceptive curcumin offers cautionary tale for chemists. *Nature* <https://doi.org/10.1038/541144a> (2017).
19. Baell, J. & Walters, M. A. Chemistry: Chemical con artists foil drug discovery. *Nature* **513**, 481–483 (2014).
20. Wang, J. et al. Enzymatic formation of curcumin in vitro and in vivo. *Nano Res.* <https://doi.org/10.1007/s12274-018-1994-z> (2018).
21. Bhardwaj, R. S., Bhardwaj, K. S., Ranjeet, D. & Ganesh, N. Curcuma longa leaves exhibits a potential antioxidant, antibacterial and immunomodulating properties. *Int. J. Phytomedicine* **3**, 270 (2011).
22. Dutta, B. Study of secondary metabolite constituents and curcumin contents of six different species of genus *Curcuma*. *J. Med. Plants Stud.* **3**, 116–119 (2015).
23. Sahoo, A., Jena, S., Sahoo, S., Nayak, S. & Kar, B. Resequencing of *Curcuma longa* L. cv. Kedaram through transcriptome profiling reveals various novel transcripts. *Genomics Data* <https://doi.org/10.1016/j.gdata.2016.08.010> (2016).
24. Sahoo, A., Kar, B., Sahoo, S., Ray, A. & Nayak, S. Transcriptome profiling of *Curcuma longa* L. cv. Suvarna. *Genomics Data* <https://doi.org/10.1016/j.gdata.2016.09.001> (2016).
25. Pellicer, J. & Leitch, I. J. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol.* <https://doi.org/10.1111/nph.16261> (2020).
26. Leong-Skornickova, J. et al. Chromosome numbers and genome size variation in Indian species of *Curcuma* (Zingiberaceae). *Ann. Bot.* <https://doi.org/10.1093/aob/mcm144> (2007).
27. Chen, J., Xia, N., Zhao, J., Chen, J. & Henny, R. J. Chromosome numbers and ploidy levels of Chinese *Curcuma* species. *HortScience* <https://doi.org/10.21273/hortsci.48.5.525> (2013).
28. Anamthawat-Jónsson, K. & Umpunjun, P. Polyploidy in the ginger family from Thailand. in *Chromosomal Abnormalities* <https://doi.org/10.5772/intechopen.92859> (2020).
29. Matasi, N. et al. Data access for the 1,000 Plants (1KP) project. *GigaScience* <https://doi.org/10.1186/2047-217X-3-17> (2014).
30. Simpson, J. T. & Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* <https://doi.org/10.1101/gr.126953.111> (2012).
31. Weisenfeld, N. L., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* <https://doi.org/10.1101/gr.214874.116> (2017).
32. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-019-0072-8> (2019).
33. Weib, C. L., Pais, M., Cano, L. M., Kamoun, S. & Burbano, H. A. nQuire: A statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics* <https://doi.org/10.1186/s12859-018-2128-z> (2018).
34. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* <https://doi.org/10.1038/s41467-020-14998-3> (2020).
35. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* <https://doi.org/10.1038/nprot.2013.084> (2013).
36. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/bts565> (2012).
37. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* <https://doi.org/10.1073/pnas.1921046117> (2020).
38. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/27.2.573> (1999).
39. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinforma.* (2014), <https://doi.org/10.1002/0471250953.bi0411s48> (2014).
40. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* (2019), <https://doi.org/10.1186/s13059-019-1832-y>.
41. Laetsch, D. R. & Blaxter, M. L. KinFin: Software for taxon-aware analysis of clustered protein sequences. *G3 Genes, Genomes, Genet.* <https://doi.org/10.1534/g3.117.300233> (2017).
42. Singh, R. et al. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature* <https://doi.org/10.1038/nature12309> (2013).
43. Chan, S. N., Abu Bakar, N., Mahmood, M., Ho, C. L. & Shaharuddin, N. A. Molecular cloning and characterization of novel phytoalexin gene from turmeric, *Curcuma longa*. *Biomed Res. Int.* <https://doi.org/10.1155/2014/973790> (2014).
44. Ren, R. et al. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Mol. Plant* <https://doi.org/10.1016/j.molp.2018.01.002> (2018).
45. Qian, H. & Jin, Y. An updated megaphylogeny of plants, a tool for generating plant phylogenies and an analysis of phylogenetic community structure. *J. Plant Ecol.* <https://doi.org/10.1093/jpe/rtv047> (2016).
46. Zeier, J. New insights into the regulation of plant immunity by amino acid metabolic pathways. *Plant Cell Environ.* <https://doi.org/10.1111/pce.12122> (2013).
47. Ruan, J. et al. Jasmonic acid signaling pathway in plants. *Int. J. Mol. Sci.* <https://doi.org/10.3390/ijms20102479> (2019).
48. Jagodzki, 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.2018.01387> (2018).
49. 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).
50. Zhou, X., Jiang, Y. & Yu, D. WRKY22 transcription factor mediates dark-induced leaf senescence in Arabidopsis. *Mol. Cells* <https://doi.org/10.1007/s10059-011-0047-1> (2011).
51. 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).
52. Chen, H. et al. Ethylene insensitive3 and ethylene insensitive3-like1 repress salicylic acid induction deficient2 expression to negatively regulate plant innate immunity in Arabidopsis. *Plant Cell* <https://doi.org/10.1105/tpc.108.065193> (2009).
53. Singh, A., Dwivedi, P. & Padmanabh Dwivedi, C. Methyl-jasmonate and salicylic acid as potent elicitors for secondary metabolite production in medicinal plants: a review. *J. Pharmacogn. Phytochem.* **7**, 750–757 (2018).
54. Zhang, L., Paasch, B. C., Chen, J., Day, B. & He, S. Y. An important role of l-fucose biosynthesis and protein fucosylation genes in Arabidopsis immunity. *New Phytol.* <https://doi.org/10.1111/nph.15639> (2019).
55. Kojko, K. et al. Regulatory mechanisms of ROI generation are affected by rice spl mutations. *Plant Cell Physiol.* <https://doi.org/10.1093/pcp/pcj074> (2006).
56. Yee, 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/ern369> (2009).
57. 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).
58. 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).
59. 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).
60. Bao, Z. & Hua, J. Interaction of CPR5 with cell cycle regulators UVI4 and OSD1 in Arabidopsis. *PLoS ONE* (2014), <https://doi.org/10.1371/journal.pone.0100347> (2014).
61. Miller, J. C., Chezem, W. R. & Clay, N. K. Ternary WD40 repeat-containing protein complexes: evolution, composition and roles in plant immunity. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2015.01108> (2016).
62. 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).
63. Choe, S. et al. The DWF4 gene of Arabidopsis encodes a cytochrome P450 that mediates multiple 22 $\alpha$ -hydroxylation steps in brassinosteroid biosynthesis. *Plant Cell* <https://doi.org/10.1105/tpc.10.2.231> (1998).
64. Sun, W. et al. Chalcone isomerase a key enzyme for anthocyanin biosynthesis in ophiorrhiza japonica. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2019.00865> (2019).
65. 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. *Phytochemistry* <https://doi.org/10.1016/j.phytochem.2012.05.026> (2012).
66. Dixon, D. P., Skipsey, M. & Edwards, R. Roles for glutathione transferases in plant secondary metabolism. *Phytochemistry* <https://doi.org/10.1016/j.phytochem.2009.12.012> (2010).
  67. Francenia Santos-Sánchez, N., Salas-Coronado, R., Hernández-Carlos, B. & Villanueva-Cañongo, C. Shikimic acid pathway in biosynthesis of phenolic compounds. *Plant Physiological Asp. Phenolic Compd.* <https://doi.org/10.5772/intechopen.83815> (2019).
  68. Moffatt, B. A. & Weretilnyk, E. A. Sustaining S-adenosyl-L-methionine-dependent methyltransferase activity in plant cells. *Physiologia Plantarum* <https://doi.org/10.1034/j.1399-3054.2001.1130401.x> (2001).
  69. Bureau, T., Lam, K. C., Ibrahim, R. K., Behdad, B. & Dayanandan, S. Structure, function, and evolution of plant O-methyltransferases. *Genome* <https://doi.org/10.1139/G07-077> (2007).
  70. Kohli, S. K. et al. Therapeutic potential of brassinosteroids in biomedical and clinical research. *Biomolecules* <https://doi.org/10.3390/biom10040572> (2020).
  71. Shitan, N. Secondary metabolites in plants: transport and self-tolerance mechanisms. *Biosci. Biotechnol. Biochem.* <https://doi.org/10.1080/09168451.2016.1151344> (2016).
  72. Gigolashvili, T. & Kopriva, S. Transporters in plant sulfur metabolism. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2014.00442> (2014).
  73. Xie, Q. et al. Multiple high-affinity K<sup>+</sup> transporters and ABC transporters involved in K<sup>+</sup> uptake/transport in the potassium-hyperaccumulator plant *Phytolacca acinosa* Roxb. *Plants* <https://doi.org/10.3390/plants9040470> (2020).
  74. Wang, X. et al. Identification and functional characterization of three type III polyketide synthases from *Aquilaria sinensis* calli. *Biochem. Biophys. Res. Commun.* <https://doi.org/10.1016/j.bbrc.2017.03.159> (2017).
  75. Rodrigues, J. L., Prather, K. L. J., Kluskens, L. D. & Rodrigues, L. R. Heterologous production of curcuminoids. *Microbiol. Mol. Biol. Rev.* <https://doi.org/10.1128/mmb.00031-14> (2015).
  76. Zhu, Q. et al. Ectopic expression of the *coless* MYB-type proanthocyanidin regulator gene *SsMYB3* alters the flower color in transgenic tobacco. *PLoS ONE* <https://doi.org/10.1371/journal.pone.0139392> (2015).
  77. Wannapinpong, S., Srikulnath, K., Thongpan, A., Choowongkamon, K. & Peyachoknagul, S. Molecular cloning and characterization of the CHS gene family in turmeric (*Curcuma longa* Linn.). *J. Plant Biochem. Biotechnol.* (2013), <https://doi.org/10.1007/s13562-013-0232-8> (2013).
  78. Pang, Y. et al. Characterization and expression of chalcone synthase gene from *Ginkgo biloba*. *Plant Sci.* <https://doi.org/10.1016/j.plantsci.2005.02.003> (2005).
  79. Xie, M. et al. Regulation of lignin biosynthesis and its role in growth-defense tradeoffs. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2018.01427> (2018).
  80. Abdallah, I. I. & Quax, W. J. A Glimpse into the biosynthesis of terpenoids. *KnE Life Sci.* <https://doi.org/10.18502/kl.v3i5.981> (2017).
  81. Joshi, A. G. et al. A knowledge-driven protocol for prediction of proteins of interest with an emphasis on biosynthetic pathways. *MethodsX* <https://doi.org/10.1016/j.mex.2020.101053> (2020).
  82. Guo, Q. Q. et al. Characterization of 3-oxacyl-acyl carrier protein reductase homolog genes in *Pseudomonas aeruginosa* PAO1. *Front. Microbiol.* (2019), <https://doi.org/10.3389/fmicb.2019.01028>.
  83. Dao, T. T. H., Linthorst, H. J. M. & Verpoorte, R. Chalcone synthase and its functions in plant resistance. *Phytochem. Rev.* <https://doi.org/10.1007/s11101-011-9211-7> (2011).
  84. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btl097> (2006).
  85. Lamo, J. M. & Rao, S. R. Chromosome counts two species *Curcuma* Linnaeus (Zingiberaceae) North-East India. *Pleione* **8**, 435–438 (2014).
  86. Nair, R. R. & Sasikumar, B. Chromosome number variation among germplasm collections and seedling progenies in turmeric, *Curcuma longa* L. *Cytologia (Tokyo)*. <https://doi.org/10.1508/cytologia.74.153> (2009).
  87. Ott, A. et al. Linked read technology for assembling large complex and polyploid genomes. *BMC Genomics* <https://doi.org/10.1186/s12864-018-5040-z> (2018).
  88. Xu, C. Q. et al. Genome sequence of *Malaria oleifera*, a tree with great value for nervonic acid production. *Gigascience* <https://doi.org/10.1093/gigascience/giy164> (2019).
  89. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gky730> (2018).
  90. Amalraj, A., Pius, A., Gopi, S. & Gopi, S. Biological activities of curcuminoids, other biomolecules from turmeric and their derivatives—a review. *J. Tradit. Complement. Med.* <https://doi.org/10.1016/j.jtcm.2016.05.005> (2017).
  91. Ling, H. Q. et al. Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. *Nature* <https://doi.org/10.1038/s41586-018-0108-0> (2018).
  92. D'hont, A. et al. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* <https://doi.org/10.1038/nature11241> (2012).
  93. Rayirath, U. P., Lada, R. R., Caldwell, C. D., Asiedu, S. K. & Sibley, K. J. Role of ethylene and jasmonic acid on rhizome induction and growth in rhubarb (*Rheum rhabarbarum* L.). *Plant Cell. Tissue Organ Cult.* <https://doi.org/10.1007/s11240-010-9861-y> (2011).
  94. Pandith, S. A. et al. Functional promiscuity of two divergent paralogs of type III plant polyketide synthases. *Plant Physiol.* <https://doi.org/10.1104/pp.16.00003> (2016).
  95. Ramakrishna, A. & Ravishankar, G. A. Influence of abiotic stress signals on secondary metabolites in plants. *Plant Signal. Behav.* <https://doi.org/10.4161/psb.6.11.17613> (2011).
  96. Isah, T. Stress and defense responses in plant secondary metabolites production. *Biological Res.* <https://doi.org/10.1186/s40659-019-0246-3> (2019).
  97. Jackman, S. D. et al. Tigmint: Correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics* <https://doi.org/10.1186/s12859-018-2425-6> (2018).
  98. Yeo, S., Coombe, L., Warren, R. L., Chu, J. & Birol, I. ARCS: Scaffolding genome drafts with linked reads. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btx675> (2018).
  99. Zhang, S. V., Zhuo, L. & Hahn, M. W. AGOUTI: Improving genome assembly and annotation using transcriptome data. *Gigascience* <https://doi.org/10.1186/s13742-016-0136-3> (2016).
  100. Warren, R. L. et al. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience* <https://doi.org/10.1186/s13742-015-0076-3> (2015).
  101. Walker, B. J. et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* (2014), <https://doi.org/10.1371/journal.pone.0112963>.
  102. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. J. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkw654> (2016).
  103. Paulino, D. et al. Sealer: A scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics* (2015), <https://doi.org/10.1186/s12859-015-0663-4> (2015).
  104. Xu, G. C. et al. LR-GapCloser: A tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* <https://doi.org/10.1093/gigascience/giy157> (2018).
  105. Li, H. Aligning sequence reads, clone sequences Assem. contigs BWA-MEM. *Arxiv*. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
  106. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/bty191> (2018).
  107. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* <https://doi.org/10.1038/nmeth.3317> (2015).
  108. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btp352> (2009).
  109. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btv351> (2015).
  110. Gremme, G., Steinbiss, S. & Kurtz, S. Genome tools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* <https://doi.org/10.1109/TCBB.2013.68> (2013).
  111. Ou, S. & Jiang, N. LTR\_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* <https://doi.org/10.1104/pp.17.01310> (2018).
  112. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btx304> (2017).
  113. Vurture, G. W. et al. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* (2017), <https://doi.org/10.1093/bioinformatics/btx153> (2017).
  114. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btu170> (2014).
  115. Zhao, S. et al. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics* <https://doi.org/10.1186/s12864-015-1876-7> (2015).
  116. Stanke, M. et al. AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkl200> (2006).
  117. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) (1990).
  118. Griffiths-Jones, S., Saini, H. K., Van Dongen, S. & Enright, A. J. miRBase: Tools for microRNA genomics. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkm952> (2008).

119. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* [https://doi.org/10.1007/978-1-4939-9173-0\\_1](https://doi.org/10.1007/978-1-4939-9173-0_1) (2019).
120. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btu739> (2015).
121. Bateman, A. The Pfam protein families database. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkh121> (2004).
122. Jaiswal, S. K., Mahajan, S., Chakraborty, A., Kumar, S. & Sharma, V. K. The genome sequence of Aloe vera reveals adaptive evolution of drought tolerance mechanisms. *iScience* <https://doi.org/10.1016/j.isci.2021.102079> (2021).
123. Jaiswal, S. K. et al. Genome sequence of peacock reveals the peculiar case of a glittering bird. *Front. Genet.* <https://doi.org/10.3389/fgene.2018.00392> (2018).
124. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/28.1.45> (2000).
125. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkr367> (2011).
126. Xu, H. et al. FastUniq: A fast de novo duplicates removal tool for paired short reads. *PLoS ONE* <https://doi.org/10.1371/journal.pone.0052249>. (2012)
127. Narasimhan, V. et al. BCFTools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btw044> (2016).
128. Zhang, L. et al. RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. *Nat. Commun.* <https://doi.org/10.1038/s41467-017-02445-9> (2017).
129. Bolser, D., Staines, D. M., Pritchard, E. & Kersey, P. Ensembl plants: Integrating tools for visualizing, mining, and analyzing plant genomics data. *Methods Mol. Biol.* [https://doi.org/10.1007/978-1-4939-3167-5\\_6](https://doi.org/10.1007/978-1-4939-3167-5_6) (2016).
130. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* <https://doi.org/10.1093/molbev/mst010> (2013).
131. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btu033> (2014).
132. Jombart, T. & Dray, S. Adephylo: exploratory analyses for the phylogenetic comparative method. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btq292> (2010).
133. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkg509> (2003).
134. Bateman, A. et al. UniProt: A hub for protein information. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gku989> (2015).
135. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* <https://doi.org/10.1093/molbev/msm088> (2007).
136. Mittal, P., Jaiswal, S. K., Vijay, N., Saxena, R. & Sharma, V. K. Comparative analysis of corrected tiger genome provides clues to its neuronal evolution. *Sci. Rep.* <https://doi.org/10.1038/s41598-019-54838-z> (2019).
137. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAA5: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkm321> (2007).
138. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* <https://doi.org/10.1093/molbev/msx148> (2017).
139. Liao, Y., Wang, J., Jaehnic, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkz401> (2019).
140. Katsuyama, Y., Kita, T. & Horinouchi, S. Identification and characterization of multiple curcumin synthases from the herb *Curcuma longa*. *FEBS Lett.* <https://doi.org/10.1016/j.febslet.2009.07.029> (2009).
141. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/25.17.3389> (1997).
142. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* <https://doi.org/10.1038/nmeth.1818> (2012).
143. Van Dongen, S. & Abreu-Goodger, C. Using MCL to extract clusters from networks. *Methods Mol. Biol.* [https://doi.org/10.1007/978-1-61779-361-5\\_15](https://doi.org/10.1007/978-1-61779-361-5_15) (2012).

## Acknowledgements

A.C. and S.M. thank Council of Scientific and Industrial Research (CSIR) for fellowship. S.K.J. thanks Department of Science and Technology for the DST-INSPIRE fellowship. The authors thank the intramural research funds provided by IISER Bhopal and the NGS facility at IISER Bhopal.

## Author contributions

VKS conceived and coordinated the project. SM prepared the DNA and RNA samples, prepared the samples for sequencing, performed the Nanopore sequencing, and the species identification assays. AC and VKS designed the computational framework of the study. AC performed all the computational analysis presented in the study. AC, SM, and SKJ performed the functional annotation of gene sets. AC, SM, SKJ and VKS analyzed the data and interpreted the results. AC constructed the figures. AC, SM, SKJ and VKS wrote the manuscript. All the authors have read and approved the final version of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42003-021-02720-y>.

**Correspondence** and requests for materials should be addressed to Vineet K. Sharma.

**Peer review information** *Communications Biology* thanks Atul Upadhyay, Daniela Holtgräwe and the other, anonymous, reviewer for their contribution to the peer review of this work. Primary Handling Editor: Caitlin Karniski.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021