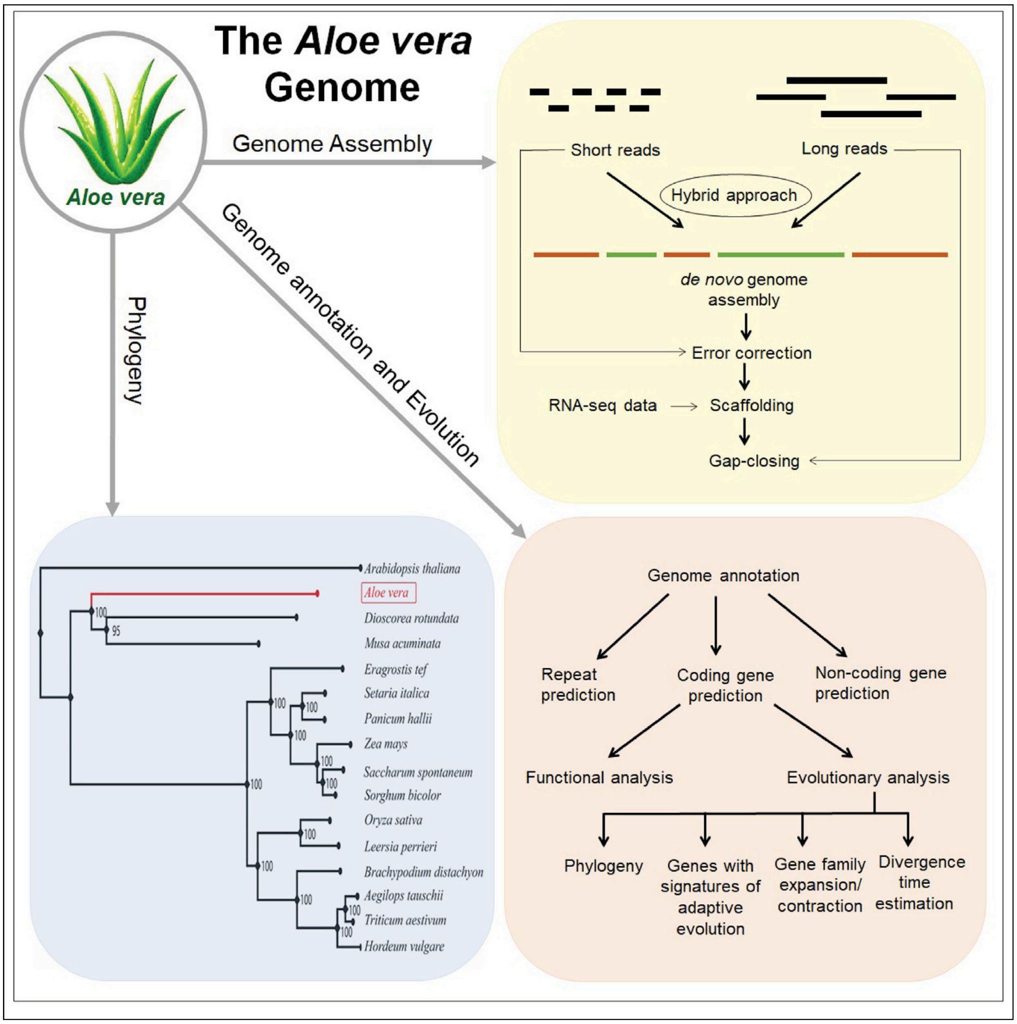
iScience

CellPress
OPEN ACCESS

Article

The genome sequence of *Aloe vera* reveals adaptive evolution of drought tolerance mechanisms



Shubham K. Jaiswal, Shruti Mahajan, Abhisek Chakraborty, Sudhir Kumar, Vineet K. Sharma

vineetks@iiserb.ac.in

HIGHLIGHTS

First *Aloe vera* genome of 12.93 Gbp harboring 86,177 protein-coding genes

First genome from Asphodelaceae family and phylogeny of monocots with *Aloe vera*

Adaptive evolution in drought stress response, CAM pathway, circadian rhythm genes

Clues on the genetic basis of evolution of drought stress tolerance capabilities

Jaiswal et al., iScience *24*, 102079
February 19, 2021 © 2021
https://doi.org/10.1016/j.isci.2021.102079

iScience

CellPress
OPEN ACCESS

Article

The genome sequence of *Aloe vera* reveals adaptive evolution of drought tolerance mechanisms

Shubham K. Jaiswal,[1,2] Shruti Mahajan,[1,2] Abhisek Chakraborty,[1,2] Sudhir Kumar,[1] and Vineet K. Sharma[1,3,*]

SUMMARY

***Aloe vera* is a species from Asphodelaceae family having characteristics like drought resistance and numerous medicinal properties. However, the genetic basis of these phenotypes is yet unknown primarily due to unavailability of its genome sequence. Thus, we report the first *Aloe vera* genome sequence comprising of 12.93 Gbp and harboring 86,177 protein-coding genes. It is the first genome from Asphodelaceae family and the largest angiosperm genome sequenced and assembled till date. We also report the first genome-wide phylogeny of monocots including *Aloe vera* to resolve its phylogenetic position. The comprehensive comparative analysis of *Aloe vera* with other available high-quality monocot genomes revealed adaptive evolution in several genes of drought stress response, CAM pathway, and circadian rhythm and positive selection in DNA damage response genes in *Aloe vera*. This study provides clues on the genetic basis of evolution of drought stress tolerance capabilities of *Aloe vera*.**

INTRODUCTION

Aloe vera is a succulent and drought-resistant plant belonging to the genus Aloe of family Asphodelaceae (Silva et al., 2010). More than 400 species are known in genus *Aloe*, of which four have medicinal properties, with *Aloe vera* being the most potent species (Grace et al., 2015). *Aloe vera* is a perennial tropical plant with succulent and elongated leaves having a transparent mucilaginous tissue consisting of parenchyma cells in the center referred to as *Aloe vera* gel (Reynolds and Dweck, 1999). The plant is extensively used as a herb in traditional practices in several countries and in cosmetics and skin care products due to its pharmacological properties including anti-inflammatory, anti-tumor, anti-viral, anti-ulcers, fungicidal, etc. (Gupta and Malhotra, 2012; Raksha et al., 2014). These medicinal properties emanate from the presence of numerous chemical constituents such as anthraquinones, vitamins, minerals, enzymes, sterols, amino acids, salicylic acids, and carbohydrates (Choudhri et al., 2018; Hamman, 2008; Joseph and Raj, 2010). These properties make it commercially important, with a global market worth 1.6 billion (Choudhri et al., 2018).

One of the key characteristics of this succulent plant is drought resistance that enables it to survive in adverse hot and dry climates (Silva et al., 2010). The plant has thick leaves arranged in an attractive rosette pattern to the stem. As an adaptation to the hot climate, the plant is able to perform a photosynthetic pathway known as crassulacean acid metabolism (CAM) that helps in limiting the water loss by transpiration (Nobel and Jordan, 1983). Moreover, the leaves have the capacity to store a large volume of water in their tissues (Jin et al., 2007). It is also known to synthesize more of soluble carbohydrates to make the osmotic adjustments under the limited water conditions, thus improving the water use efficiency (Delatorre-Herrera et al., 2010). The transcriptome and chloroplast genome of *Aloe vera* are available from previous studies, and a few studies have also highlighted the drought stress tolerance and potential benefits of *Aloe vera* (Choudhri et al., 2018; Ren et al., 2020). However, the unavailability of its reference genome sequence has been a deterrent in understanding the genetic basis and molecular mechanisms of the unique characteristics of this medicinal plant.

In addition to the functional analysis, the resolution of the phylogenetic position has the potential to reveal the evolutionary history and understand the correlations between phylogenetic diversity and important traits of interest. Multiple attempts have been made to resolve the phylogenetic position of *Aloe* genus and *Aloe vera*; however, these efforts only used a few conserved loci such as rbcL, psbA, matK, and ribosomal genes and were not performed at the genome-wide level due to the unavailability of the genomic sequence (Adams et al., 2000; Grace et al., 2015; Treutlein et al., 2003). The previous phylogenies have

[1] MetaBioSys Group, Department of Biological Sciences, Indian Institute of Science Education and Research Bhopal, Bhopal, Madhya Pradesh, India

[2] These authors contributed equally

[3] Lead contact

*Correspondence: vineetks@iiserb.ac.in

https://doi.org/10.1016/j.isci.2021.102079

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

reported that *Aloe vera* shared the most common recent ancestor with the species of Poales and Zingiberales order, also within the Asparagales order; it was closest to the other succulent genera such as *Haworthia*, *Gasteria*, and *Astroloba* (Chase et al., 2016; Qian and Jin, 2016).

Aloe vera has an estimated genome size of 16.04 Gbp with a diploid ploidy level containing 14 (2n) chromosomes according to the Plant DNA c-value database (Zonneveld, 2002). The unavailability of the genome sequence of *Aloe vera* is noteworthy given the fact that the representative genomes of species from almost all the plant families including Brassicaceae, Cannabaceae, Cucurbitaceae, Euphorbiaceae, Fabaceae, Malvaceae, Rosaceae, Solanaceae, Poaceae, Orchidaceae, and Betulaceae have been sequenced and studied. However, no species from the Asphodelaceae plant family has been sequenced till date. Thus, the availability of *Aloe vera* genome sequence will help to reveal the genomic signatures of Asphodelaceae family and will also be useful in understanding the genetic basis of the important phenotypes such as medicinal properties and drought resistance in *Aloe vera*.

Therefore, in this study, we report the first draft genome sequence of *Aloe vera* using a hybrid sequencing and assembly approach by combining the Illumina short-read and Oxford Nanopore long-read sequences to construct the genome sequence. The transcriptome sequencing and analysis of two tissues, root and leaf, was carried out to gain deep insights into the gene expression and to precisely determine its gene set. The genome-wide phylogeny of *Aloe vera* with other available monocot genomes was also constructed to resolve its phylogenetic position. The comparative analysis of *Aloe vera* with other monocot genomes revealed adaptive evolution in its genes and provided insights on the stress tolerance capabilities of this species.

RESULTS

Sequencing of *Aloe vera* genome and transcriptome

To comprehensively cover the large genome of *Aloe vera* species, a total of 506.4 Gbp (37.15X) of short-read and 146.8 Gbp (~10.77X) of long-read data were generated using Illumina and nanopore platforms, respectively (Tables S1 and S2) (Dolezel et al., 2003; Zonneveld, 2002). For transcriptome, a total of 6.6 Gbp and 7.3 Gbp of RNA-seq data were generated from leaf and root, respectively. The transcriptome data from this study and the publicly available RNA-seq data from previous studies (Choudhri et al., 2018; Wickett et al., 2014) were combined together, resulting in a total of 37.1 Gbp of RNA-seq data for *Aloe vera*, which was used for the analysis (Table S3). All the short-read genomic and RNA-seq data were trimmed and filtered using Trimmomatic, and only the high-quality read data were used to construct the final genome and transcriptome assemblies. The complete workflow of the sequence analysis is shown in Figure S1.

Assembly of *Aloe vera* genome

The percent heterozygosity was estimated to be 11.3% for *Aloe vera* species. The final draft genome assembly of *Aloe vera* had the size of 12.93 Gbp and N50 of 14.6 kbp, of which 11.12 Gbp had length >500 bp and N50 of 20.4 kbp (Table S4). The genomic coverage and N50 attained in case of the first draft assembly of *Aloe vera* genome appears reasonable for such a challenging and a gigantic plant genome and is also comparable to the other large plant genomes assembled till date (Birol et al., 2013; Neale et al., 2014; Nystedt et al., 2013; Stevens et al., 2016). This was achieved by the hybrid assembly of short-read and long-read data, which was further polished by correction using SeqBug, RNA-seq-data-based scaffolding using Rascaf, and long-read-based gap-closing using LR-gap closer. The k-mer-count-distribution-based method using only the short Illumina reads estimated a genome size of 13.63 Gb, which was smaller than the c-value-based genome size estimation of 16.04 Gbp, conceivably due to the usage of only short-read data for the genome size estimation (Figure S2). The percent GC for the final assembly was 41.98%.

The analysis of repetitive sequences revealed 557,638,058 bp of tandem repeats corresponding to 3.41% of the complete genome. For interspersed repeat identification, a total of 1,820 repeat families identified using RepeatModeler were used as custom repeat library for repeat identification in *Aloe vera* genome. Out of the 1,820 repeat families, 1,550 families could not be annotated by RepeatModeler. Identification of repeats was carried out on short-read assembly using RepeatMasker. It revealed that 82.66% of *Aloe vera* genome is constituted by interspersed repeats, of which 55.57% was unclassified and 26.96% was identified as retroelements. Among the retroelements, 26.71% was LTR repeats (7.34% Ty1/Copia and 19.37% Gypsy/

DIRS1 elements) and 0.13% was DNA transposons. Similarly, the identification of repeats in hybrid assembly revealed that interspersed repeats constitute 78.70% of *Aloe vera* genome, of which 51.26% was unclassified and 27.28% was identified as retroelements. Among the retroelements, 26.94% was LTR repeats (8.87% Ty1/Copia and 18.07% Gypsy/DIRS1 elements) and 0.16% was DNA transposons.

Transcriptome assembly

The Trinity assembly of transcriptomic reads resulted in a total size of 163,190,792 bp with an N50 value of 1,268 bp and an average contig length of 796 bp (Table S5). The mapping of filtered RNA-seq reads on the Trinity transcripts using hisat2 resulted in the overall percentage mapping of 92.49%. The complete BUSCO score (addition of single copy and duplicates) on the transcripts was 90.5%. A total of 205,029 transcripts were predicted, corresponding to 108,133 genes with the percent GC of 43.69. The clustering of gene sequences using CD-HIT-EST to remove the redundancy resulted in 107,672 unigenes. The coding genes (CDS) from the unigenes were predicted using TransDecoder resulting in 34,269 coding genes.

Genome annotation and gene set construction

A total of 1,978 standard amino-acid-specific tRNAs and 378 hairpin miRNAs were identified in the *Aloe vera* genome (Table S6). The MAKER-pipeline-based gene prediction resulted in a total of 114,971 coding transcripts, of which 63,408 transcripts (≥300 bp) were considered further for clustering at 95% identity resulting in 57,449 unique coding gene transcripts. Application of the same length-based selection criteria (≥300 bp) on trinity-identified 34,269 coding gene transcripts resulted in 33,998 coding gene transcripts. The merging of these two coding gene transcript sets resulted in the final gene set of 86,177 genes for *Aloe vera*, which had the complete BUSCO score of 74.6% and single-copy BUSCO score of 72.4%.

Identification of orthologs across selected plant species

A total of 104,543 orthogroups were identified using OrthoFinder across the selected 16 plant species. Of which, only a total of 5,472 orthogroups had sequences from all the 16 plant species and were used for the identification of orthologs. For these 5,472 orthogroups, in case of presence of more than one gene from a species in an orthogroup, the longest gene representative from that species was selected to construct the final orthologous gene set for any orthogroup. Thus, including one gene from each of the 16 species in an orthogroup, a total of 5,472 orthologs were identified. In addition, the fuzzy one-to-one orthologs finding approach applied using KinFin resulted in a total of 1,440 fuzzy one-to-one orthologs that were used for constructing the maximum likelihood species phylogenetic tree.

Resolving the phylogenetic position of *Aloe vera*

Each of the 1,440 fuzzy one-to-one orthologous gene set was aligned and concatenated, and the resultant concatenated alignment had a total of 1,453,617 alignment positions. The concatenated alignment was filtered for the undetermined values, which were treated as missing values, and a total of 1,157,550 alignment positions were retained. The complete alignment data and the filtered alignment data were both used to construct maximum likelihood species trees using RAxML with the bootstrap value of 100, and both the alignment data resulted in the same phylogeny. Thus, the phylogeny based on the filtered data was considered as the final genome-wide phylogeny of *Aloe vera* including all the representative monocot genomes available on Ensembl plants database with *Arabidopsis thaliana* as an outgroup (Figure 1). This phylogeny also corroborated with the earlier reported phylogenies by Silvera et al. (2014), Dunemann et al. (2014), and Wang and Deng, 2016, which were constructed using a limited number of genetic loci (Dunemann et al., 2014; Silvera et al., 2014; Wang and Deng, 2016). It is apparent from the phylogeny that *Dioscorea rotundata* and *Musa acuminata* are the most closely related to *Aloe vera* and share the same clade (Figure 1). All other selected monocots are distributed in separate clade, with *Triticum aestivum* and *Aegilops tauschii* being the most distantly related to *Aloe vera*.

Recently an updated plant megaphylogeny has been reported for the vascular plants (Qian and Jin, 2016). The species of Poales order showed similar relative positions in our reported phylogeny and this megaphylogeny. In the megaphylogeny, *Musa acuminata* was reported to share the most common recent ancestor with the species of Poales order, but in our phylogeny we observed that *Musa acuminata* shared the most common recent ancestor with *Dioscorea rotundata* from Dioscoreales order (Figures 1 and S3). Also, among the selected monocots, the species of Dioscoreales order was reported to show the earliest divergence. However, in our genome-wide phylogeny, *Aloe vera* showed the earliest divergence.

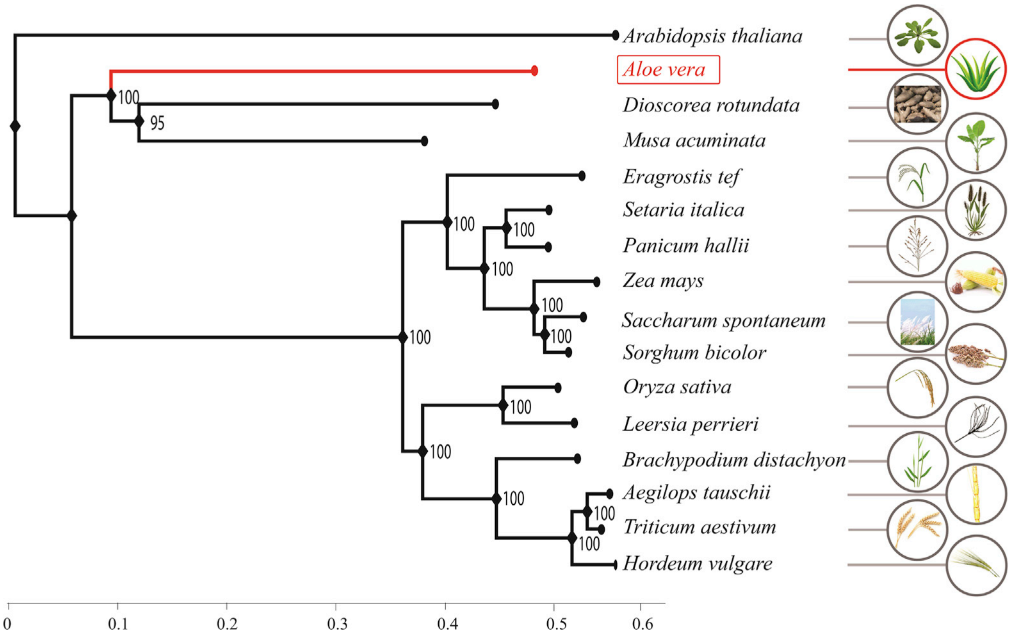


Figure 1. The phylogenetic tree of the selected 14 monocot species, *Aloe vera*, and *Arabidopsis thaliana* as an outgroup

The values mentioned at the nodes are the bootstrap values. The scale mentioned is the nucleotide substitutions per base.

See also Figure S3.

With reference to the reported phylogeny of angiosperms, at the order level the Poales and Zingiberales formed a clade, and their ancestor shared the most recent common ancestor with Asparagales, then all three shared a recent ancestor with Dioscoreales (Chase et al., 2016). In our genome-wide phylogeny, Zingiberales and Dioscoreales shared the most recent common ancestor, and their ancestor shared the most recent common ancestor with Asparagales, and the three shared a recent ancestor with Poales.

The divergence time for *Aloe vera* and clade formed by *Dioscorea rotundata* and *Musa acuminata* was estimated to be 104.97 Mya with equal-tailed confidence interval of 101.32–112.22 Mya, which is in agreement with the previous studies (Bremer, 2000; Fiz-Palacios et al., 2011; Pouget et al., 2016). Further, the divergence time of the clade leading to *Aloe vera*, *Dioscorea rotundata*, and *Musa acuminata* and the clade formed by the other species of Poales order was estimated to be 107.66 Mya with equal-tailed confidence interval of 103.47–116.73 Mya.

Gene families with expansion and contraction in *Aloe vera*

For the identification of gene family expansion and contraction, a total of 52,357 families were obtained by clustering. Of these, 23,016 families having <100 gene copies for every species were used for further analysis. A total of 5,143 families were found to be expanded, and 2,977 families showed contraction in *Aloe vera* in comparison to the recent ancestor (Figure S4). Among the top 20 expanded families in *Aloe vera*, ABC transporter, RNA-mediated transposition, transcription initiation factor TFIIB, ATPases, ribonuclease H, and serine arginine-rich splicing factor were the families with known function. In contrast, the top 20 contracted families in *Aloe vera* included the functions related to MADS-box transcription factor, synthase, RNA polymerase II transcription regulator recruiting activity, peptidase S10, cysteine-rich receptor-like protein kinase, galactoside, WRKY transcription factor, AP2-like ethylene-responsive transcription factor, and reverse transcriptase.

Genes with a higher rate of evolution

A total of 85 genes showed higher rates of evolution in *Aloe vera* in comparison to the other monocot species. These genes belonged to several eggNOG categories and KEGG pathways as mentioned in Tables S7 and S8, with a higher representation of ribosomal genes. The distribution of enriched (p value<0.05) biological process GO terms is mentioned in Table S9. Also, among these 85 genes three molecular function GO terms, rRNA binding, structural constituent of cytoskeleton, and structural constituent of ribosome showed an enrichment (p value<0.05) (Table S10). Five transcription factors—WRKY, MYB, bHLH, CPP, and LBD—showed higher rates of evolution in *Aloe vera*. Among these, WRKY, MYB, and bHLH are known to be involved in drought stress tolerance (Fei et al., 2019; Waseem and Li, 2019; Zhao et al., 2018b). There were six chloroplast-functioning-related genes, namely EMB3127, PnsB3, TL29, IRT3, PDV2, and SIRB, that showed a higher rate of evolution. Notably, the chloroplast-function-related genes have been implicated in different abiotic stress conditions in plants, including drought (Yoo et al., 2019; Zhao et al., 2018a).

Identification of positively selected genes

A total of 199 genes showed positive selection in *Aloe vera* with the FDR q-value threshold of 0.05. The distribution of these genes in eggNOG categories, KEGG pathways, and GO term categories are mentioned in Tables S11–S15. Among the genes with positive selection, several genes were involved in key functions with specific phenotypic consequences (Figure S5). These included flowering-related genes that are important for the reproductive success, calcium-ion-binding and transcription factors/sequence-specific DNA-binding genes involved in signal transduction for response to external stimulus, carbohydrate catabolism genes required for energy production, and genes involved in abiotic stress response (Agarwal and Jha, 2010; Takatsuji, 1998; Tuteja and Mahajan, 2007). Among the abiotic stress response genes, there were four categories of genes: water-related stress response genes, DNA damage response genes involved in reactive oxidative species (ROS) stress response, nuclear pore complex genes involved in plant stress response by regulating the nucleo-cytoplasmic trafficking, and secondary metabolites-biosynthesis-related genes that deal with different types of biotic and abiotic stresses (Naik and Al-Khayri, 2016; Roldán-Arjona and Ariza, 2009; Yang et al., 2017). The robust and efficient DNA damage response mechanism is essential for biotic and abiotic stress tolerance and for the genomic stability (Nisa et al., 2019). Thus, adaptive evolution in this pathway seemingly contributes toward the stress tolerance capabilities and genomic stability in *Aloe vera*.

Another gene G6PD5 that showed positive selection in *Aloe vera* protects plants against different types of stress such as salinity stress by producing nitric oxide (NO) molecule, which leads to the expression of defense response genes (Arasimowicz and Floryszak-Wieczorek, 2007; Liu et al., 2007). Regulation of osmotic potential under drought stress is acquired by different ion channels, transporters, and carrier proteins (Bray, 1993). In this study, K^+ transporter 1(KT1), bidirectional amino acid transporter 1(BAT1), and "BASS6," a sodium/metabolite co-transporter gene, were found to be positively selected in *Aloe vera*.

The abscisic acid (ABA) responsive element binding factor (ABF) gene was found to be positively selected. This gene is differentially expressed under drought and other abiotic stress and alters specific target gene expression by binding to ABRE (abscisic-acid-response element), the characteristic element of ABA-inducible genes (Feng et al., 2019). A previous study showed that mutations in this gene lead to increased sensitivity to drought (Yoshida et al., 2015). ABA also regulates stomatal closure and solute transport and thus have implications in drought tolerance (Yamaguchi-Shinozaki and Shinozaki, 2006). The trehalase 1 (TRE1) gene was also found to be positively selected, and the overexpression of this gene causes better drought tolerance through ABA-guided stomatal closure (Van Houtte et al., 2013).

Genes with site-specific signs of evolution

Two types of site-specific signatures of adaptive evolution, i.e., positively selected codon sites and unique amino acid substitutions with significant functional impact, were identified in *Aloe vera*. The unique substitutions analysis has the potential to reveal the amino acid substitutions, which are specific to the species of interest and can significantly affect the protein function. However, the number of genes with unique substitutions may increase with the increase in genetic distance, and thus the usage of closely related species is desired to make the analysis more reliable. Therefore, for this analysis we have only used the monocot genomes that were available on the Ensembl plant genome database to make the analysis more robust and reliable. The positively selected codon sites analysis identifies the codon sites that are under positive selection in our species of interest; the setup of using monocot genomes along with *Arabidopsis thaliana* as outgroup also helped in effectively identifying the positively selected codon sites with higher accuracy.

A total of 1,848 genes had positively selected codon sites, and a total of 2,669 genes had unique amino acid substitutions with functional impact. The distribution of genes with positively selected codon sites and unique amino acid substitutions with functional impact in eggNOG categories, KEGG pathways, and GO term categories are mentioned in Supplemental Tables S16–S25.

One of the characteristics of succulent plants such as *Aloe vera* is the ability to efficiently assimilate the atmospheric CO_2 and reduce water loss by transpiration through the CAM pathway, a specific mode of photosynthesis. The evolution of CAM is an adaptation to the limited CO_2 and limited water condition, and a significant correlation between higher succulence and increased magnitude of CAM metabolism has been observed (Teeri et al., 1981). In this study, several crucial genes of CAM metabolism showed site-specific signatures of adaptive evolution in *Aloe vera* (Figure 2A). The potassium channel involved in stomatal opening/closure (KAT2), malic enzyme (ME) that converts malic acid to pyruvate, and phosphoenolpyruvate carboxylase (PEPC) that converts phosphoenolpyruvate to oxaloacetate and assimilates the environmental CO_2 showed both the signs of site-specific adaptive evolution. In addition, the other CAM genes including potassium transport 2/3 (KT2/3), pyruvate orthophosphate dikinase (PPDK), phosphoenolpyruvate carboxylase kinase 1 (PPCK1), carbonic anhydrase 1 (CA1), peroxisomal NAD-malate dehydrogenase 2 (PMDH2), tonoplast dicarboxylate transporter (TDT), and aluminum-activated malate transporter family protein (ALMT9) showed unique substitutions with functional impact on *Aloe vera*.

CAM metabolism evolution is known to be a result of modified circadian regulation at the transcription and posttranscriptional levels (Mallona et al., 2011). CAM evolution is a well-characterized physiological rhythm in plants, and it is also a specific example of circadian clock-based specialization (Mallona et al., 2011; Silvera et al., 2010). Several plant circadian rhythm genes showed site-specific signs of adaptive evolution in *Aloe vera* (Figure 2B). Three essential genes of red light response—PHYB, ELF3, and LHY showed both the signs of site-specific adaptive evolution. Also, the FT gene important for flowering and under the control of circadian rhythm showed both the signs of site-specific adaptive evolution. The PHYA gene, which is also a part of the red light response, had unique substitutions with functional impact. Among the blue light response genes, three genes—GI, FKF1, and SPA2 had unique substitutions with functional impact, and two genes—HY5 and CHS had positively selected codon sites. The blue light response regulates the UV protection and photomorphogenesis.

Plant hormone signaling regulates plant growth, development, and response to different types of biotic and abiotic stress (Santner and Estelle, 2009). Multiple genes of auxin, cytokinin, and brassinosteroid hormone signaling involved in cellular growth and elongation having implications in cellular and tissue succulence showed site-specific signatures of adaptive evolution (Figure 2C). The genes of the abscisic acid hormone signaling involved in stomatal opening/closure required for CAM metabolism and different biotic and abiotic stress response (Feng et al., 2019) had positively selected codon sites and unique substitution sites with functional impact (Figure 2C). Also, the genes involved in salicylic acid signaling important for providing disease resistance and help in biotic stress response showed site-specific signatures of adaptive evolution (Figure 2C).

Genes with multiple signs of adaptive evolution

Among the three signatures of adaptive evolution i.e., positive selection, a higher rate of evolution, and unique amino acid substitutions with functional impact, a total of 148 genes showed two or more signs of adaptive evolution and were identified as the genes with multiple signs of adaptive evolution (MSA) (Chakraborty et al., 2020; Jaiswal et al., 2018; Mittal et al., 2019). The distribution of these genes in eggNOG categories, KEGG pathways, and GO categories are mentioned in Tables S26–S29. Another study that performed the proteomic analysis of drought stress response in wild peach also found similar categories to be enriched in the proteins that were differentially expressed under drought conditions (Cao et al., 2017). A total of 90 genes out of the 148 MSA genes in *Aloe vera* were from the specific categories that are associated with drought stress tolerance (Supplementary Data Sheet 1.). The literature references for MSA genes that were considered to be associated with drought stress tolerance mechanisms are mentioned in Supplementary Data Sheet 1. The specific groups of proteins and their relation with the drought stress tolerance are mentioned in Figure 3.

Several ribosomal genes and transcription factors genes were found to be MSA genes in this study, and these were also found to be overexpressed under drought conditions in different proteomic and transcriptomic studies and aid in better drought stress survival (Cao et al., 2017; Janiak et al., 2018; Moin et al., 2016). Mutational studies in these genes have also shown their role in drought resistance (Moin et al., 2016). Many

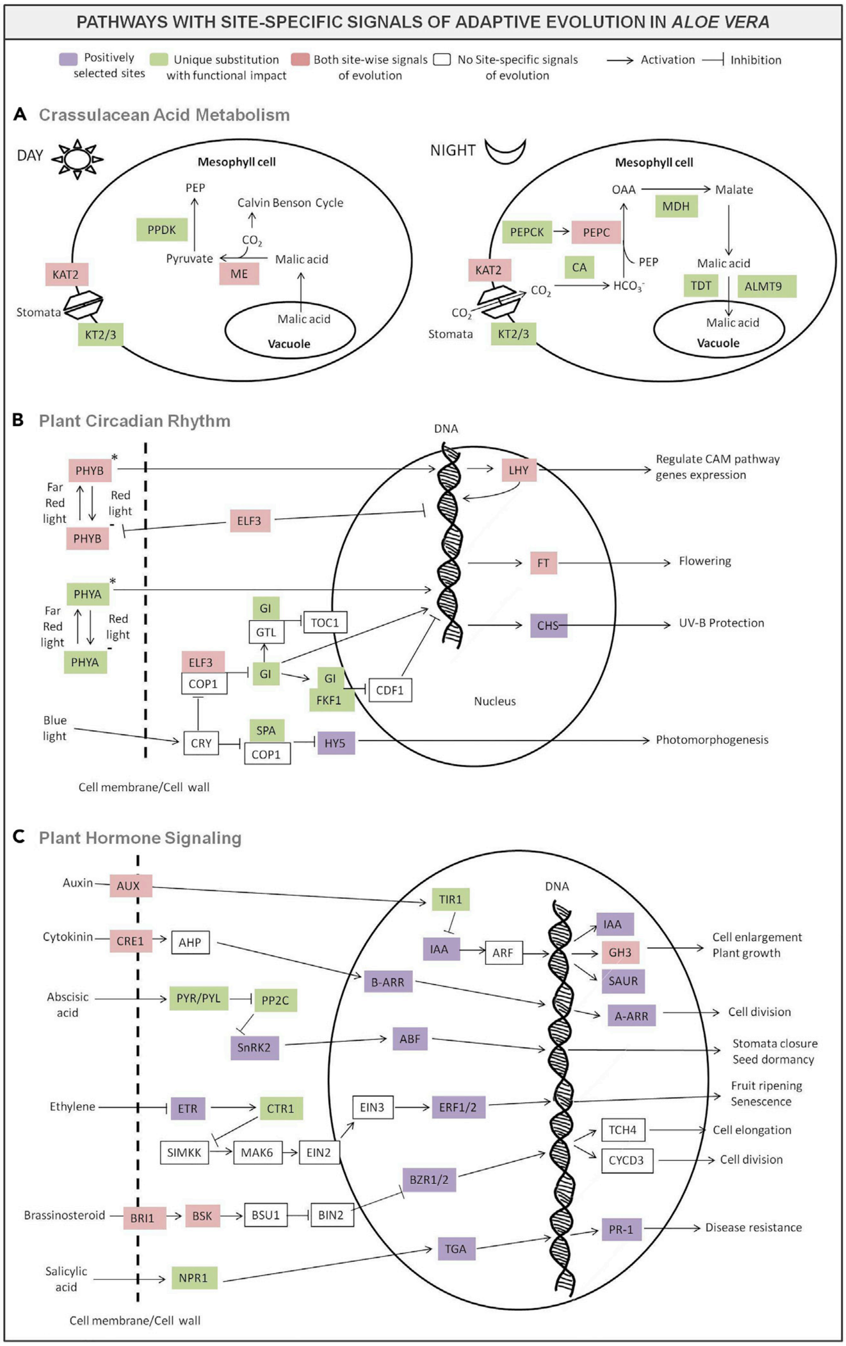


Figure 2. The adaptive evolution of CAM pathway, plant circadian rhythm, and plant hormone signaling in *Aloe vera*

The important genes of the CAM pathway, plant circadian rhythm, and plant hormone signaling are shown with their function: (A) for CAM pathway, (B) for plant circadian rhythm, and (C) for plant hormone signaling. The genes in Lavender had positively selected codon sites, the genes in Green had unique substitutions with function impact, and the genes in Red showed both the signs of site-specific adaptive evolution in *Aloe vera*. There were no CAM pathway genes that had only positively selected codon sites.

See also Tables S16–S25.

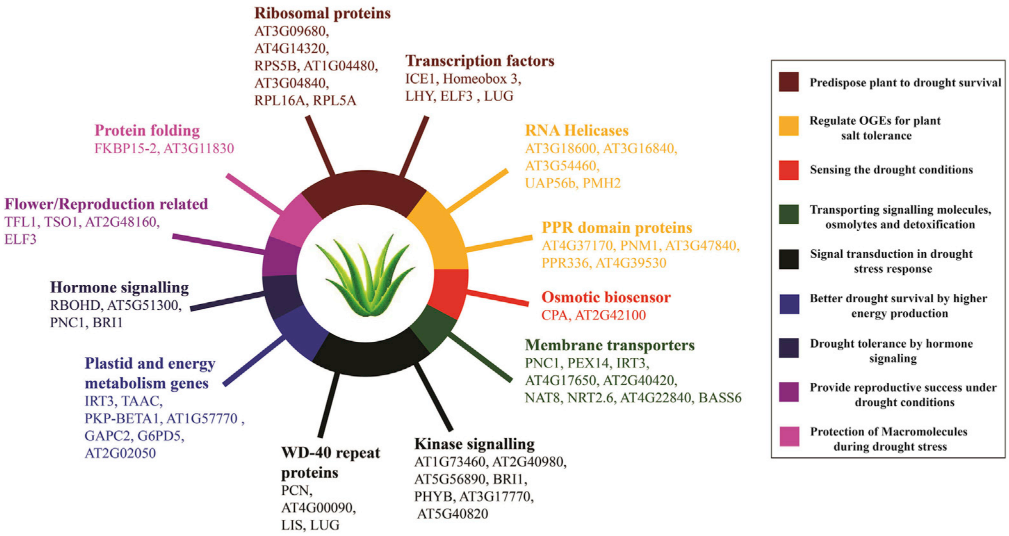


Figure 3. The MSA genes in *Aloe vera* that are involved in drought stress response

The relation of specific categories of genes with drought stress response was determined from the literature. The standard *Arabidopsis thaliana* gene IDs were used in case of genes that did not have a standard gene symbol.
See also Tables S26–S30.

nuclear genes are involved in the functioning of symbiotic organelles chloroplast and mitochondria. Some of these genes are also involved in the organellar gene expression (OGEs) regulation, and their mutants are known to show altered response to different abiotic stress, including high salinity stress (Leister et al., 2017; Robles and Quesada, 2019). Several of these genes belonging to two categories, RNA helicases and PPR domain proteins, were found to be MSA genes. Thus, in the *Aloe vera* species, these genes have been adaptively evolved to provide this species with better salt tolerance.

Two osmotic biosensor genes—"CPA" and "AT2G42100" were found to be among the MSA genes in *Aloe vera*. Different membrane transporters that can transport signaling molecules, osmolytes, and metals were also among the MSA genes (Figure 3). These included two peroxisomal transporters—"PNC1" a nucleotide carrier protein and "PEX14" a transporter for PTS1 and PTS2 domain containing signaling proteins and different heavy metal transporters such as "IRT3," an iron transporter, "AT4G17650," a lipid transporter, "AT2G40420," an amino acid transporter, "ALA1," a phospholipid transporter, "NAT8," a nucleobase-ascorbate transporter, "NRT2.6," a high-affinity nitrate transporter, and "BASS6," a sodium/metabolite co-transporter. These osmotic sensors and transporters provide significant enhancement in function in drought stress condition and help in adjusting to the water scarcity (Iqbal, 2018; Jarzyniak and Jasiński, 2014).

The genes for several kinases and WD-40 repeat proteins were also found to be among the MSA genes in *Aloe vera*. These proteins are involved in signaling and transcription regulation required for the drought stress tolerance (Feyissa et al., 2019; Janiak et al., 2018; Liu et al., 2017; Ranjan and Sawant, 2015). Also, the genes involved in energy generation and are part of the thylakoid membrane showed MSA. The stability of thylakoid membrane proteins has been associated with drought resistance, and these energy-production-related genes are crucial in survival during the drought stress (Janiak et al., 2018; Tian et al., 2013). Two genes that assist in protein folding were found to show MSA (Figure 3), and these proteins are very important in protecting the macromolecules of the cells under drought stress conditions (Shinozaki and Yamaguchi-Shinozaki, 2007). Four genes involved in plant hormone signaling were also among the MSA genes. The plant hormone signaling is central to the signaling pathways required for drought stress tolerance (Tiwari et al., 2017). Four genes involved in flowering and reproduction regulation were also found to be among the MSA genes in *Aloe vera*. The flowering and reproduction-related genes are known to be regulated for better reproductive success under drought stress conditions as part of the drought tolerance strategy used by many plants (Monroe et al., 2018; Song et al., 2017).

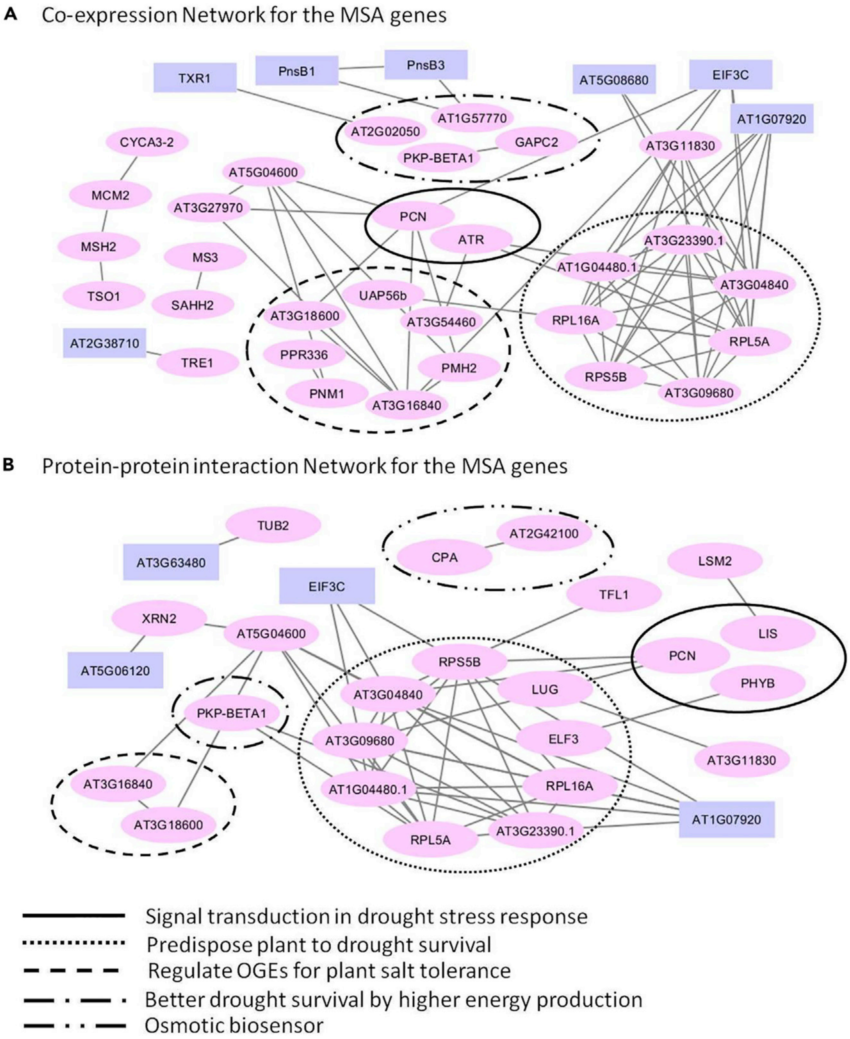


Figure 4. Evaluating the co-expression and physical interaction of MSA genes in *Aloe vera*

(A) The co-expression network of the MSA genes is shown. Only the MSA genes that showed at least one co-expression connection are shown. The nodes represent the genes, and the edges represent the co-expression of the connected nodes.

(B) The protein-protein interaction network of the MSA genes is shown. Only the MSA genes that showed at least one protein-protein interaction are shown. The nodes represent the genes, and the edges represent the protein-protein interaction between the connected nodes.

The genes shown in Pink are the genes involved in drought stress tolerance mechanisms, and genes shown in Lavender had other functions. To check the literature reference for the genes associated with the drought stress tolerance mechanisms, please refer to Table S30. The standard *Arabidopsis thaliana* gene IDs were used in case of genes that did not have a standard gene symbol.

The co-expression of MSA genes was examined using the co-expression data from the STRING database (Szklarczyk et al., 2017), and the MSA genes that co-express with at least one other MSA gene are displayed as a network diagram (Figure 4A). From the network, it is evident that majority of co-expressing MSA genes are associated with drought stress tolerance mechanisms, and the genes forming the dense network are also related to drought stress tolerance. Predominantly, three categories of drought-stress-tolerance-related MSA genes have shown co-expression: genes involved in energy production, genes involved in OGEs regulation, and genes that predispose plants to drought stress tolerance.

Similarly, a network diagram was constructed using the protein-protein interaction data of MSA genes from the STRING database (Szklarczyk et al., 2017). The genes with physical interaction known from the experimental studies are shown as a network diagram in Figure 4B. From the network, it is apparent that among the interacting MSA genes, all of them except four are associated with drought stress tolerance mechanisms. Further, among the MSA genes, primarily the genes that predispose plants to drought stress tolerance showed the physical interaction. In addition, three genes involved in signal transduction in drought stress response, two genes that function as osmotic biosensors, and two OGEs regulation genes also displayed physical interaction.

DISCUSSION

In this work, we have reported the complete draft genome sequence of *Aloe vera*, which is an evolutionarily important, ornamental, and widely used plant species due to its medicinal properties, pharmacological applications, traditional usage, and commercial value. The availability of *Aloe vera* genome sequence is also important because it is the first genome sequenced from the Asphodelaceae plant family and is the largest angiosperm and the fifth largest genome sequenced so far. It is also the largest genome sequenced using the Oxford Nanopore technology till date.

The high level of heterozygosity is one of the key challenges in genome assembly because the *de novo* assembler tries to generate a single haploid output from the allelic differences, which is difficult to achieve with high levels of heterozygosity (Asalone et al., 2020). The estimated heterozygosity for the *Aloe vera* genome was 11.3%, which is even higher than the heterozygosity of wheat genome (10.1%) that required the combined effort of several research institutions across the world and more than a decade for its completion. Thus, the hybrid approach of using short-read (Illumina) and long-read (Nanopore) sequence data emerged as a successful assembly strategy to overcome the challenge of constructing one of the largest plant genomes. The study also reported the gene set of *Aloe vera* constructed using the combination of *de novo* and homology-based gene predictions, and also using the data from the genomic assembly and the transcriptomic assembly from multiple tissues, thus indicating the comprehensiveness of the approach.

This study reported the first genome-wide phylogeny of *Aloe vera* with all other monocot species available on the Ensembl plant database and with *Arabidopsis thaliana* as an outgroup. A few previous studies have also examined the phylogenetic position of *Aloe vera* with respect to other monocots but used a few genomic loci. Thus, this is the first genome-wide phylogeny of monocots that resolves the phylogenetic position of *Aloe vera* with respect to the other monocots by using 1,440 different loci distributed throughout their genomes. The very high bootstrap values for the internal nodes and existence of no polytomy in the phylogeny further attest to the correctness of the phylogeny. This phylogeny is mostly in agreement with the previously known phylogenies and also provided new insights (Dunemann et al., 2014; Grace et al., 2015; Qian and Jin, 2016; Silvera et al., 2014; Wang and Deng, 2016).

An earlier phylogeny constructed using "ppc-aL1a" gene showed that *Sorghum bicolor*, *Zea mays*, *Setaria italica*, *Brachypodium distachyon*, *Hordeum vulgare,* and *Oryza sativa* form a monophyletic group, which was also observed in our phylogeny (Silvera et al., 2014). Similarly, the relative positions of *Hordeum vulgare*, *Saccharum officinarum*, *Zea mays,* and *Oryza sativa* in another phylogeny based on "CENH3" gene were in agreement with our phylogeny (Dunemann et al., 2014). Using the "NORK" gene, another recent study reported the relative phylogenetic position of four monocot species: *Oryza sativa, Zea mays*, *Sorghum bicolor,* and *Setaria italica* (Wang and Deng, 2016). *Zea mays*, *Sorghum bicolor,* and *Setaria italica* were found to share a recent last common ancestor, and *Oryza sativa* had diverged earlier from their common ancestor, which is also supported by the genome-wide phylogeny reported in this study.

Although the genome-wide phylogeny showed the species of Poales order with similar topology as reported in earlier studies, a different topology was observed for the relative position of Musa acuminata, *Dioscorea rotundata,* and *Aloe vera* from the orders Zingiberales, Dioscoreales, and Asparagales, respectively (Chase et al., 2016; Qian and Jin, 2016). The observed differences could be due to the usage of a few genomic loci in the previous phylogenies, whereas the phylogeny reported in this study is a genome-wide phylogeny constructed using 1,440 one-to-one orthologs distributed across the genome. The availability of more complete genomes from monocots and the inclusion of more genomic loci in the phylogenetic analysis will help in explaining the observed differences and confirm the relative positions of these species.

One of the key highlights of the study was the revelation of adaptive evolution of genes involved in drought stress response, which provides a genetic explanation for the drought stress tolerance properties of *Aloe* *vera*. This plant is known to display a number of phenotypes such as perennial succulent leaves and CAM mechanism for carbon fixation that provide it with better drought stress survival (Jin et al., 2007). Several experimental studies have also reported that it can make adjustments such as increased production of sugars and increased expression of heat-shock and ubiquitin proteins for efficient water utilization and osmotic maintenance that eventually provide better drought survival (Delatorre-Herrera et al., 2010; Hazrati et al., 2017; Huerta et al., 2013). In this study, the majority (60.81%) of genes that showed multiple signs of evolution (MSA) were involved in drought-stress-tolerance-related functions. These genes were also found to be co-expressing and physically interacting with each other, which further point toward the adaptive evolution of the drought stress tolerance mechanisms in this species. The adaptive evolution of genes involved in drought stress tolerance provides insights into the genetic basis of drought resistance property of *Aloe vera*.

Several crucial genes of CAM pathway and circadian rhythm have also shown site-specific signs of adaptive evolution in *Aloe vera* in comparison to the other monocot species. The CAM pathway has very high water use efficiency and is known to have evolved convergently in many arid regions for better drought survival (Ming et al., 2015). Also, the CAM pathway is a physiological rhythm with temporal separation of atmospheric CO_2 assimilation and Calvin-Benson cycle and is under the control of plant circadian rhythm (Mallona et al., 2011; Yin et al., 2018). This CAM pathway evolution is known to be a specific type of circadian rhythm specialization (Hartwell, 2018; Silvera et al., 2010). Thus, the observed adaptive evolution of CAM pathway and its controller circadian rhythm in this study point toward its role in providing this species an evolutionary advantage for efficient drought stress survival. Furthermore, it should be noted that among the genes, which showed adaptive evolution in *Aloe vera* such as the ones involved in drought stress response, CAM pathway, and circadian rhythm, only the genes with previously known functional role from experimental studies were used for the interpretation. Therefore, it is likely that the genes with adaptive evolution in *Aloe vera* may have phenotypic consequences.

The evolutionary success of the *Aloe* genus is also known to be due to the succulent leaf Mesophyll tissue (Grace et al., 2015). Particularly, the medicinal use of *Aloe vera* is much associated with the succulent leaf mesophyll tissue, and a loss of this tissue leads to the loss of medicinal properties (Reynolds and Dweck, 1999). The plant species with CAM pathway have large vacuoles in comparison to the non-CAM plants, and therefore, the leaf succulence is also higher in CAM plants. Thus, it is tempting to speculate that the observed evolution of CAM pathway in *Aloe vera* may also be crucial for the higher leaf mesophyll succulence contributing to its medicinal properties. Also previously, it has been proposed that the specific properties of *Aloe vera* such as the high leaf succulence, medicinal properties, and drought resistance are the consequences of evolutionary processes such as selection and speciation rather than due to phylogenetic diversity or isolation (Grace et al., 2015). The signatures of adaptive evolution in drought tolerance and CAM pathway genes in *Aloe vera* further substantiate this notion.

Limitations of the study

Because of the unavailability of any other genome from the *Aloe* genus or from the Asphodelaceae family, the phylogenetic and adaptive evolution analysis was performed on the available species that were distantly related to *Aloe vera*. Thus, the availability of more genomes from closely related plant species will provide more insights into the unique properties of this species and its family. Further, the large genome size and high heterozygosity in plant genomes such as *Aloe vera* are among the key challenges for their complete assembly, and the availability of more data, improved algorithms, and tools are likely to help in generating a more comprehensive and contiguous assembly of this species.

Resource availability

Lead contact

Further information, requests, and inquiries should be directed to the Lead Contact, Vineet K. Sharma (vineetks@iiserb.ac.in).

Materials availability

All the materials and methods used for the generation of data and analysis are mentioned in the manuscript text. The generated data were deposited to the public repositories, and accession numbers are mentioned in ''Data and code availability'' section.

Data and code availability

The accession number for the *Aloe vera* sequence data (DNA and RNA) reported in this paper are NCBI BioProject accession number: PRJNA634897, NCBI Biosample accession number: SAMN15010737, and NCBI accession codes: SRR11842980, SRR11842979, SRR11842978, and SRR11842977.

METHODS

All methods can be found in the accompanying Transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102079.

ACKNOWLEDGMENTS

The author SKJ thanks Department of Science and Technology, India for the DST-INSPIRE fellowship. AC and SM thank Council for Scientific and Industrial Research (CSIR), India for fellowship. SK thanks University Grants Commission (UGC), India for the fellowship. The authors thank the intramural research funds provided by IISER Bhopal, India. We also acknowledge the help of Mr. Vishnu Prasoodanan PK in preparing Figures 1 and 3.

AUTHOR CONTRIBUTIONS

VKS conceived and coordinated the project. SM prepared the DNA and RNA samples, performed sequencing, and the species identification assay. SKJ and VKS designed the computational framework of the study. SKJ and AC performed the genome assembly, transcriptome assembly, genome annotation, gene set construction, orthology analysis, and species phylogenetic tree construction. SKJ performed the root-to-tip branch length, positive selection, unique substitution with functional impact, network, and statistical analysis. SKJ, AC, SK, and SM performed the functional annotation of gene sets. SKJ, AC, and VKS analyzed the data. SKJ, AC, and VKS interpreted the results. SKJ constructed the figures. SKJ, AC, SM, SK, and VKS wrote and revised the manuscript. All the authors have read and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 1, 2020
Revised: December 18, 2020
Accepted: January 14, 2021
Published: February 19, 2021

REFERENCES

Adams, S.P., Leitch, I.J., Bennett, M.D., Chase, M.W., and Leitch, A.R. (2000). Ribosomal DNA evolution and phylogeny in Aloe (Asphodelaceae). Am. J. Bot. *87*, 1578–1583.

Agarwal, P., and Jha, B. (2010). Transcription factors in plants and ABA dependent and independent abiotic stress signalling. Biol. Plant. *54*, 201–212.

Arasimowicz, M., and Floryszak-Wieczorek, J. (2007). Nitric oxide as a bioactive signalling molecule in plant stress responses. Plant Sci. *172*, 876–887.

Asalone, K.C., Ryan, K.M., Yamadi, M., Cohen, A.L., Farmer, W.G., George, D.J., Joppert, C., Kim, K., Mughal, M.F., and Said, R. (2020). Regional sequence expansion or collapse in heterozygous genome assemblies. PLoS Comput. Biol. *16*, e1008104.

Birol, I., Raymond, A., Jackman, S.D., Pleasance, S., Coope, R., Taylor, G.A., Yuen, M.M.S., Keeling, C.I., Brand, D., and Vandervalk, B.P. (2013). Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics *29*, 1492–1497.

Bray, E.A. (1993). Molecular responses to water deficit. Plant Physiol. *103*, 1035.

Bremer, K. (2000). Early Cretaceous lineages of monocot flowering plants. Proc. Natl. Acad. Sci. U S A *97*, 4707–4711.

Cao, Y., Luo, Q., Tian, Y., and Meng, F. (2017). Physiological and proteomic analyses of the drought stress response in Amygdalus Mira (Koehne) Yü et Lu roots. BMC Plant Biol. *17*, 53.

Chakraborty, A., Mahajan, S., Jaiswal, S.K., and Sharma, V.K. (2020). Genome sequencing of turmeric provides evolutionary insights into its medicinal properties. bioRxiv. https://doi.org/10.1101/2020.09.07.286245.

Chase, M.W., Christenhusz, M., Fay, M., Byng, J., Judd, W.S., Soltis, D., Mabberley, D., Sennikov, A., Soltis, P.S., and Stevens, P.F. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot. J. Linn. Soc. *181*, 1–20.

Choudhri, P., Rani, M., Sangwan, R.S., Kumar, R., Kumar, A., and Chhokar, V. (2018). De novo sequencing, assembly and characterisation of *Aloe vera* transcriptome and analysis of expression profiles of genes related to saponin and anthraquinone metabolism. BMC Genomics *19*, 427.

Delatorre-Herrera, J., Delfino, I., Salinas, C., Silva, H., and Cardemil, L. (2010). Irrigation restriction effects on water use efficiency and osmotic adjustment in *Aloe vera* plants (Aloe barbadensis Miller). Agric. Water Manage. *97*, 1564–1570.

Dolezel, J., Bartos, J., Voglmayr, H., and Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. Cytometry A J. Int. Soc. Anal. Cytol. *51*, 127–128, author reply 129.

Dunemann, F., Schrader, O., Budahn, H., and Houben, A. (2014). Characterization of centromeric histone H3 (CENH3) variants in cultivated and wild carrots (Daucus sp.). PLoS One *9*, e98504.

Fei, X., Hou, L., Shi, J., Yang, T., Liu, Y., and Wei, A. (2019). Patterns of drought response of 38 WRKY transcription factors of Zanthoxylum bungeanum Maxim. Int. J. Mol. Sci. *20*, 68.

Feng, R.-J., Ren, M.-Y., Lu, L.-F., Peng, M., Guan, X., Zhou, D.-B., Zhang, M.-Y., Qi, D.-F., Li, K., and Tang, W. (2019). Involvement of abscisic acid-responsive element-binding factors in cassava (Manihot esculenta) dehydration stress response. Scientific Rep. *9*, 1–12.

Feyissa, B.A., Arshad, M., Gruber, M.Y., Kohalmi, S.E., and Hannoufa, A. (2019). The interplay between miR156/SPL13 and DFR/WD40–1 regulate drought tolerance in alfalfa. BMC Plant Biol. *19*, 1–19.

Fiz-Palacios, O., Schneider, H., Heinrichs, J., and Savolainen, V. (2011). Diversification of land plants: insights from a family-level phylogenetic analysis. BMC Evol. Biol. *11*, 341.

Grace, O.M., Buerki, S., Symonds, M.R., Forest, F., van Wyk, A.E., Smith, G.F., Klopper, R.R., Bjorå, C.S., Neale, S., and Demissew, S. (2015). Evolutionary history and leaf succulence as explanations for medicinal use in aloes and the global popularity of *Aloe vera*. BMC Evol. Biol. *15*, 29.

Gupta, V.K., and Malhotra, S. (2012). Pharmacological attribute of *Aloe vera*: Revalidation through experimental and clinical studies. Ayu *33*, 193.

Hamman, J.H. (2008). Composition and applications of *Aloe vera* leaf gel. Molecules *13*, 1599–1616.

Hartwell, J. (2018). The circadian clock in CAM plants. Annu. Plant Rev. Online *21*, 211–236.

Hazrati, S., Tahmasebi-Sarvestani, Z., Mokhtassi-Bidgoli, A., Modarres-Sanavy, S.A.M., Mohammadi, H., and Nicola, S. (2017). Effects of zeolite and water stress on growth, yield and chemical compositions of *Aloe vera* L. Agric. Water Management *181*, 66–72.

Huerta, C., Freire, M., and Cardemil, L. (2013). Expression of hsp70, hsp100 and ubiquitin in Aloe barbadensis Miller under direct heat stress and under temperature acclimation conditions. Plant Cell Rep. *32*, 293–307.

Iqbal, M.J. (2018). Role of osmolytes and antioxidant enzymes for drought tolerance in wheat. Glob. Wheat Prod. *51*, https://doi.org/10.5772/intechopen.75926.

Jaiswal, S.K., Gupta, A., Saxena, R., Prasoodanan, V.P.K., Sharma, A.K., Mittal, P., Roy, A., Shafer, A.B.A., Vijay, N., Sharma, V.K., et al. (2018). Genome sequence of peacock reveals the peculiar case of a glittering bird. Front. Genet. https://doi.org/10.3389/fgene.2018.00392.

Janiak, A., Kwasniewski, M., Sowa, M., Gajek, K., Żmuda, K., Kościelniak, J., and Szarejko, I. (2018). No time to waste: transcriptome study reveals that drought tolerance in barley may be attributed to stressed-like expression patterns that exist before the occurrence of stress. Front. Plant Sci. *8*, 2212.

Jarzyniak, K.M., and Jasiński, M. (2014). Membrane transporters and drought resistance–a complex issue. Front. Plant Sci. *5*, 687.

Jin, Z.M., Wang, C.H., Liu, Z.P., and Gong, W.J. (2007). Physiological and ecological characters studies on *Aloe vera* under soil salinity and seawater irrigation. Process Biochem. *42*, 710–714.

Joseph, B., and Raj, S.J. (2010). Pharmacognostic and phytochemical properties of *Aloe vera* linn an overview. Int. J. Pharm. Sci. Rev. Res. *4*, 106–110.

Leister, D., Wang, L., and Kleine, T. (2017). Organellar gene expression and acclimation of plants to environmental stress. Front. Plant Sci. *8*, 387.

Liu, W.C., Li, Y.H., Yuan, H.M., Zhang, B.L., Zhai, S., and Lu, Y.T. (2017). WD40-REPEAT 5a functions in drought stress tolerance by regulating nitric oxide accumulation in Arabidopsis. Plant Cell Environ. *40*, 543–552.

Liu, Y., Wu, R., Wan, Q., Xie, G., and Bi, Y. (2007). Glucose-6-phosphate dehydrogenase plays a pivotal role in nitric oxide-involved defense against oxidative stress under salt stress in red kidney bean roots. Plant Cell Physiol. *48*, 511–522.

Mallona, I., Egea-Cortines, M., and Weiss, J. (2011). Conserved and divergent rhythms of crassulacean acid metabolism-related and core clock gene expression in the cactus Opuntia ficus-indica. Plant Physiol. *156*, 1978–1989.

Ming, R., VanBuren, R., Wai, C.M., Tang, H., Schatz, M.C., Bowers, J.E., Lyons, E., Wang, M.-L., Chen, J., and Biggers, E. (2015). The pineapple genome and the evolution of CAM photosynthesis. Nat. Genet. *47*, 1435–1442.

Mittal, P., Jaiswal, S.K., Vijay, N., Saxena, R., and Sharma, V.K. (2019). Comparative analysis of corrected tiger genome provides clues to its neuronal evolution. Sci. Rep. https://doi.org/10.1038/s41598-019-54838-z.

Moin, M., Bakshi, A., Saha, A., Udaya Kumar, M., Reddy, A.R., Rao, K., Siddiq, E., and Kirti, P. (2016). Activation tagging in indica rice identifies ribosomal proteins as potential targets for manipulation of water-use efficiency and abiotic stress tolerance in plants. Plant Cell Environ. *39*, 2440–2459.

Monroe, J.G., Powell, T., Price, N., Mullen, J.L., Howard, A., Evans, K., Lovell, J.T., and McKay, J.K. (2018). Drought adaptation in Arabidopsis thaliana by extensive genetic loss-of-function. Elife *7*, e41038.

Naik, P.M., and Al-Khayri, J.M. (2016). Abiotic and biotic elicitors–role in secondary metabolites production through in vitro culture of medicinal plants. In Abiotic and Biotic Stress in Plants: Recent Advances and Future Perspectives, A. Shanker and C. Shanker, eds., pp. 247–277.

Neale, D.B., Wegrzyn, J.L., Stevens, K.A., Zimin, A.V., Puiu, D., Crepeau, M.W., Cardeno, C., Koriabine, M., Holtz-Morris, A.E., and Liechty, J.D. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. *15*, R59.

Nisa, M.-U., Huang, Y., Benhamed, M., and Raynaud, C. (2019). The plant DNA damage response: signaling pathways leading to growth inhibition and putative role in response to stress conditions. Front. Plant Sci. *10*, https://doi.org/10.3389/fpls.2019.00653.

Nobel, P.S., and Jordan, P.W. (1983). Transpiration stream of desert species: resistances and capacitances for a C3, a C4, and a CAM plant. J. Exp. Bot. *34*, 1379–1391.

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., and Alexeyenko, A. (2013). The Norway spruce genome sequence and conifer genome evolution. nature *497*, 579–584.

Pouget, M., Youssef, S., Dumas, P.-J., Baumberger, T., San Roman, A., Torre, F., Affre, L., Médail, F., and Baumel, A. (2016). Spatial mismatches between plant biodiversity facets and evolutionary legacy in the vicinity of a major Mediterranean city. Ecol. Indicators *60*, 736–745.

Qian, H., and Jin, Y. (2016). An updated megaphylogeny of plants, a tool for generating plant phylogenies and an analysis of phylogenetic community structure. J. Plant Ecol. *9*, 233–239.

Raksha, B., Pooja, S., and Babu, S. (2014). Bioactive compounds and medicinal properties of *Aloe vera* L.: an update. J. Plant Sci. *2*, 102–107.

Ranjan, A., and Sawant, S. (2015). Genome-wide transcriptomic comparison of cotton (Gossypium herbaceum) leaf and root under drought stress. 3 Biotech. *5*, 585–596.

Ren, J.-J., Wang, J., Lee, K.-K., Deng, H., Xue, H., Zhang, N., Zhao, J.-C., Cao, T., Cui, C.-L., and Zhang, X.-H. (2020). The complete chloroplast genome of *Aloe vera* from China as a Chinese herb. Mitochondrial DNA B *5*, 1092–1093.

Reynolds, T., and Dweck, A. (1999). *Aloe vera* leaf gel: a review update. J. Ethnopharmacology *68*, 3–37.

Robles, P., and Quesada, V. (2019). Transcriptional and post-transcriptional regulation of organellar gene expression (OGE) and its roles in plant salt tolerance. Int. J. Mol. Sci. *20*, 1056.

Roldán-Arjona, T., and Ariza, R.R. (2009). Repair and tolerance of oxidative DNA damage in plants. Mutat. Res. Rev. Mutat. Res. *681*, 169–179.

Santner, A., and Estelle, M. (2009). Recent advances and emerging trends in plant hormone signalling. Nature *459*, 1071–1078.

Shinozaki, K., and Yamaguchi-Shinozaki, K. (2007). Gene networks involved in drought stress response and tolerance. J. Exp. Bot. *58*, 221–227.

Silva, H., Sagardia, S., Seguel, O., Torres, C., Tapia, C., Franck, N., and Cardemil, L. (2010). Effect of water availability on growth and water use efficiency for biomass and gel production in *Aloe vera* (Aloe barbadensis M.). Ind. Crops Prod. *31*, 20–27.

Silvera, K., Neubig, K.M., Whitten, W.M., Williams, N.H., Winter, K., and Cushman, J.C. (2010). Evolution along the crassulacean acid metabolism continuum. Funct. Plant Biol. *37*, 995–1010.

Silvera, K., Winter, K., Rodriguez, B.L., Albion, R.L., and Cushman, J.C. (2014). Multiple isoforms of phospho enol pyruvate carboxylase in the Orchidaceae (subtribe Oncidiinae): implications for the evolution of crassulacean acid metabolism. J. Exp. Bot. *65*, 3623–3636.

Song, K., Kim, H.C., Shin, S., Kim, K.-H., Moon, J.-C., Kim, J.Y., and Lee, B.-M. (2017). Transcriptome analysis of flowering time genes under drought stress in maize leaves. Front. Plant Sci. *8*, 267.

Stevens, K.A., Wegrzyn, J.L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., Paul, R., Gonzalez-Ibeas, D., Koriabine, M., and Holtz-Morris, A.E. (2016). Sequence of the sugar pine megagenome. Genetics *204*, 1613–1626.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic. Acids Res. *45*, D362–D368.

Takatsuji, H. (1998). Zinc-finger transcription factors in plants. Cell Mol. Life Sci. CMLS *54*, 582–596.

Teeri, J., Tonsor, S., and Turner, M. (1981). Leaf thickness and carbon isotope composition in the Crassulaceae. Oecologia *50*, 367–369.

Tian, F., Gong, J., Zhang, J., Zhang, M., Wang, G., Li, A., and Wang, W. (2013). Enhanced stability of thylakoid membrane proteins and antioxidant competence contribute to drought stress resistance in the tasg1 wheat stay-green mutant. J. Exp. Bot. *64*, 1509–1520.

Tiwari, S., Lata, C., Singh Chauhan, P., Prasad, V., and Prasad, M. (2017). A functional genomic perspective on drought signalling and its crosstalk with phytohormone-mediated signalling pathways in plants. Curr. Genomics *18*, 469–482.

Treutlein, J., Smith, G.F., Van Wyk, B.-E., and Wink, M. (2003). Phylogenetic relationships in Asphodelaceae (subfamily Alooideae) inferred from chloroplast DNA sequences (rbcL, matK) and from genomic fingerprinting (ISSR). Taxon *52*, 193–207.

Tuteja, N., and Mahajan, S. (2007). Calcium signaling network in plants: an overview. Plant Signal. Behav. *2*, 79–85.

Van Houtte, H., Vandesteene, L., López-Galvis, L., Lemmens, L., Kissel, E., Carpentier, S., Feil, R., Avonce, N., Beeckman, T., and Lunn, J.E. (2013). Overexpression of the trehalase gene AtTRE1 leads to increased drought stress tolerance in Arabidopsis and is involved in abscisic acid-induced stomatal closure. Plant Physiol. *161*, 1158–1171.

Wang, L., and Deng, L. (2016). GmACP expression is decreased in GmNORK knockdown transgenic soybean roots. Crop J. *4*, 509–516.

Waseem, M., and Li, Z. (2019). Dissecting the role of a basic helix-loop-helix transcription factor, SlbHLH22, under salt and drought stresses in transgenic Solanum lycopersicum L. Front. Plant Sci. *10*, 734.

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., and Gitzendanner, M.A. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. U S A *111*, E4859–E4868.

Yamaguchi-Shinozaki, K., and Shinozaki, K. (2006). Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. Annu. Rev. Plant Biol. *57*, 781–803.

Yang, Y., Wang, W., Chu, Z., Zhu, J.-K., and Zhang, H. (2017). Roles of nuclear pores and nucleo-cytoplasmic trafficking in plant stress responses. Front. Plant Sci. *8*, 574.

Yin, H., Guo, H.-B., Weston, D.J., Borland, A.M., Ranjan, P., Abraham, P.E., Jawdy, S.S., Wachira, J., Tuskan, G.A., and Tschaplinski, T.J. (2018). Diel rewiring and positive selection of ancient plant proteins enabled evolution of CAM photosynthesis in Agave. BMC Genomics *19*, 588.

Yoo, Y.-H., Hong, W.-J., and Jung, K.-H. (2019). A Systematic view exploring the role of chloroplasts in plant abiotic stress responses. Biomed. Res. Int. *2019*, https://doi.org/10.1155/2019/6534745.

Yoshida, T., Fujita, Y., Maruyama, K., Mogami, J., Todaka, D., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2015). Four A rabidopsis AREB/ABF transcription factors function predominantly in gene expression downstream of SnRK2 kinases in abscisic acid signalling in response to osmotic stress. Plant Cell Environ. *38*, 35–49.

Zhao, C., Haigh, A.M., Holford, P., and Chen, Z.-H. (2018a). Roles of chloroplast retrograde signals and ion transport in plant drought tolerance. Int. J. Mol. Sci. *19*, 963.

Zhao, Y., Cheng, X., Liu, X., Wu, H., Bi, H., and Xu, H. (2018b). The wheat MYB transcription factor TaMYB31 is involved in drought stress responses in Arabidopsis. Front. Plant Sci. *9*, 1426.

Zonneveld, B.J. (2002). Genome size analysis of selected species of Aloe (Aloaceae) reveals the most primitive species and results in some new combinations. Bradleya *2002*, 5–12.

iScience, Volume *24*

Supplemental Information

The genome sequence of *Aloe vera* reveals adaptive evolution of drought tolerance mechanisms

Shubham K. Jaiswal, Shruti Mahajan, Abhisek Chakraborty, Sudhir Kumar, and Vineet K. Sharma

SUPPLEMENTARY TABLES

Supplementary Table S1. Summary of the Illumina sequence data for *Aloe vera* genome (Related to "Figure 1" of main text)

Paired-end Insert Size	Average Read Length	Number of Reads	Total Data	Sequence Coverage
447 bp and 600 bp	150 bp	3,393,209,648	506.4 Gb	37.15X

Supplementary Table S2. Summary of the nanopore sequence data for *Aloe vera* genome (Related to "Figure 1" of main text)

Average Read Length	Number of Reads	Total Data	Sequence Coverage
3,327 bp	48,240,314	146.8 Gb	10.77X

Supplementary Table S3. Summary of the transcriptome data for *Aloe vera* genome (Related to "Figure 1" of main text)

Tissue	Average read length R1 (bp)	Average read length R2 (bp)	Total number of read pairs	Total number of R1 (bp)	Total number of R2 (bp)	Total number of bases (bp)
Leaf[1]	101	101	32,776,695	3,310,446,195	3,310,446,195	6,620,892,390
Root[1]	101	101	36,212,970	3,657,509,970	3,657,509,970	7,315,019,940
Leaf[2]	101	101	29,247,010	2,953,948,010	2,953,948,010	5,907,896,020
Root[2]	145.6	145.4	51,078,070	7,440,880,511	7,427,363,139	14,868,243,650
1K genome project[3]	73	75	16,218,326	1,183,937,798	1,216,374,450	2,400,312,248
Total data			165,533,071	18,546,722,484	18,565,641,764	37,112,364,248

[1] Our study, [2] Choudhri et al., 2018 (Choudhri et al., 2018), [3] 1KP project (Wickett et al., 2014)

Supplementary Table S4. Summary statistics of the final assembly for *Aloe vera* genome (Related to "Figure 1" of main text)

Genome statistics*	Value
Number of scaffold (> 300 bp)	7,545,697
Number of scaffold (≥ 1000 bp)	1,687,410
Number of scaffold (≥ 5000 bp)	433,182
Number of scaffold (≥ 10000 bp)	254,572
Number of scaffold (≥ 25000 bp)	104,910
Number of scaffold (≥ 50000 bp)	32,389
Total length (> 300 bp)	12,934,659,027
Total length (≥1000 bp)	10,404,862,008
Total length (≥ 5000 bp)	8,523,711,385

Total length (≥ 10000 bp)	7,293,656,262
Total length (≥ 25000 bp)	4,912,010,614
Total length (≥ 50000 bp)	2,385,794,989
Largest scaffold	4,941,863
GC% (> 300 bp)	41.90
N50 (> 300 bp)	14,560
Number of N's per 100 kbp (> 300 bp)	175.12

Supplementary Table S5. Summary statistics of the transcriptome assembly for *Aloe vera* (Related to "Figure 1" of main text)

Statistics based on all transcript contigs	
Contig N10	3151
Contig N20	2402
Contig N30	1942
Contig N40	1584
Contig N50	1268
Average contig	795.94
Total assembled bases	163,190,792
Statistics based on only longest isoform per gene	
Contig N10	3,098
Contig N20	2,338
Contig N30	1,838
Contig N40	1,431
Contig N50	1,061
Average contig	652.68
Total assembled bases	70,576,159
Counts of genes and transcripts	
Total trinity 'genes'	108133
Total trinity transcripts	205029

Supplementary Table S6. The tRNAs compared across different monocot species including *Aloe vera* (Related to "Figure 1" of main text)

Species	tRNA
Aegilops tauschii	547
Brachypodium distachyon	289
Dioscorea rotundata	154
Hordeum vulgare	727
Leersia perrieri	196
Musa acuminata	1,254
Oryza sativa	242
Setaria italicaa	346
Sorghum bicolor	324
Triticum aestivum	1,915
Zea mays	2,834
Arabidopsis thaliana	689
Aloe vera	1,978*

*Only the tRNAs specific to standard amino acids are mentioned

The data for the other species was retrieved from the Ensembl plants genome browser, and for the *Aloe vera* species the tRNAs were identified as mentioned in the **Supplementary Text S3**.

Supplementary Table S7. The distribution of genes with higher rate of evolution in different eggNOG categories in *Aloe vera* (Related to "Figure 2" of main text)

eggNOG category	Number of genes
Function unknown	17
Translation, ribosomal structure and biogenesis	9
Posttranslational modification, protein turnover, chaperones	8
Inorganic ion transport and metabolism	6
Intracellular trafficking, secretion, and vesicular transport	5
Transcription	5
Energy production and conversion	5
RNA processing and modification	5
Carbohydrate transport and metabolism	4
Cytoskeleton	4
Signal transduction mechanisms	3
Coenzyme transport and metabolism	2
Amino acid transport and metabolism	2
Replication, recombination and repair	2
Nucleotide transport and metabolism	1
Lipid transport and metabolism	1
Cell cycle control, cell division, chromosome partitioning	1

Supplementary Table S8. The distribution of genes with higher rate of evolution in different KEGG pathways in *Aloe vera* (Only pathways relevant to plants and with more than one genes are mentioned) (Related to "Figure 2" of main text)

KEGG Pathway	Number of genes
Ribosome	8
Oxidative phosphorylation	2
Glutathione metabolism	2
Spliceosome	2
RNA transport	2

Supplementary Table S9. The biological process GO categories that were enriched in the genes with higher rate of evolution in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to "Figure 2" of main text)

GO Term ID	Description	p-value
GO:0051187	cofactor catabolic process	0.0156
GO:0072511	divalent inorganic cation transport	0.0234
GO:0009624	response to nematode	0.0278
GO:0048285	organelle fission	0.0299
GO:0070925	organelle assembly	0.0350
GO:0007017	microtubule-based process	0.0401

Supplementary Table S10. The molecular function GO categories that were found enriched in the genes with higher rate of evolution in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to "Figure 2" of main text)

GO Term ID	Description	p-value
GO:0019843	rRNA binding	0.0014
GO:0005200	structural constituent of cytoskeleton	0.0042
GO:0003735	structural constituent of ribosome	0.0384

Supplementary Table S11. The distribution of genes with positive selection in different eggNOG categories in *Aloe vera* (Related to "Figure 2" of main text)

eggNOG category	Number of genes
Function unknown	47
Transcription	19
Carbohydrate transport and metabolism	15
RNA processing and modification	13
Signal transduction mechanisms	12
Posttranslational modification, protein turnover, chaperones	11
Inorganic ion transport and metabolism	9
Intracellular trafficking, secretion, and vesicular transport	9
Translation, ribosomal structure and biogenesis	8
Replication, recombination and repair	8
Energy production and conversion	6
Amino acid transport and metabolism	5
Lipid transport and metabolism	5
Coenzyme transport and metabolism	5
Cell cycle control, cell division, chromosome partitioning	5
Secondary metabolites biosynthesis, transport and catabolism	4
Chromatin structure and dynamics	3
Cell wall/membrane/envelope biogenesis	3
Defence mechanisms	2
Cytoskeleton	2
Nuclear structure	1

Supplementary Table S12. The distribution of genes with positive selection in different KEGG pathways in *Aloe vera* (Only pathways relevant to plants and with more than one gene are mentioned) (Related to "Figure 2" of main text)

KEGG Pathway	Number of Genes
Cell cycle	5
RNA transport	4
Glycolysis / Gluconeogenesis	3
Starch and sucrose metabolism	3
alpha-Linolenic acid metabolism	3
Cysteine and methionine metabolism	3
Aminoacyl-tRNA biosynthesis	3
Plant hormone signal transduction	3
Cellular senescence	3
Circadian rhythm - plant	3
Plant-pathogen interaction	3

Supplementary Table S13. The biological process GO categories that were enriched in the genes with positive selection in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to "Figure 2" of main text)

GO Term ID	Description	p-value
GO:0051128	regulation of cellular component organization	0.003
GO:0009415	response to water	0.004
GO:0032504	multicellular organism reproduction	0.007
GO:0051172	negative regulation of nitrogen compound metabolic process	0.008
GO:0015748	organophosphate ester transport	0.008
GO:0009890	negative regulation of biosynthetic process	0.010
GO:0040008	regulation of growth	0.011
GO:0016052	carbohydrate catabolic process	0.014
GO:0104004	cellular response to environmental stimulus	0.016
GO:0071496	cellular response to external stimulus	0.024
GO:2000241	regulation of reproductive process	0.024
GO:0015979	photosynthesis	0.025
GO:1905392	plant organ morphogenesis	0.031
GO:0051726	regulation of cell cycle	0.031
GO:0006974	cellular response to DNA damage stimulus	0.031
GO:0009735	response to cytokinin	0.036
GO:0080134	regulation of response to stress	0.040
GO:0009409	response to cold	0.043
GO:0009308	amine metabolic process	0.044
GO:0006325	chromatin organization	0.046
GO:0010038	response to metal ion	0.048
GO:0045165	cell fate commitment	0.049

Supplementary Table S14. The cellular component GO categories that were enriched in the positively selected genes in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to "Figure 2" of main text)

GO Term ID	Description	p-value
GO:0030133	transport vesicle	0.031
GO:0005635	nuclear envelope	0.046

Supplementary Table S15. The molecular function GO categories that were enriched in the positively selected genes in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to "Figure 2" of main text)

GO Term ID	Description	p-value
GO:0005509	calcium ion binding	0.006
GO:0016798	hydrolase activity, acting on glycosyl bonds	0.037

GO:0017056	structural constituent of nuclear pore	0.038

Supplementary Table S16. The distribution of genes containing positively selected codon sites in different eggNOG categories in *Aloe vera* (Related to "Figure 2" of main text)

eggNOG category	Number of genes
Function unknown	443
Signal transduction mechanisms	167
Posttranslational modification, protein turnover, chaperones	157
Transcription	128
Carbohydrate transport and metabolism	123
RNA processing and modification	87
Intracellular trafficking, secretion, and vesicular transport	81
Amino acid transport and metabolism	80
Translation, ribosomal structure and biogenesis	79
Secondary metabolites biosynthesis, transport and catabolism	65
Lipid transport and metabolism	63
Energy production and conversion	59
Inorganic ion transport and metabolism	48
Replication, recombination and repair	42
Cell cycle control, cell division, chromosome partitioning	30
Cell wall/membrane/envelope biogenesis	28
Cytoskeleton	27
Coenzyme transport and metabolism	25
Chromatin structure and dynamics	24
Nucleotide transport and metabolism	23
Defence mechanisms	15
Nuclear structure	1

Supplementary Table S17. The distribution of genes containing positively selected codon sites in different KEGG pathways in *Aloe vera* (Only pathways relevant to plants and with more than ten genes are mentioned) (Related to "Figure 2" of main text)

KEGG Pathway	Number of Genes
Ribosome	25
Protein processing in endoplasmic reticulum	24
Purine metabolism	21
Starch and sucrose metabolism	18
Spliceosome	18
RNA transport	18

Plant hormone signal transduction	18
Cysteine and methionine metabolism	17
Amino sugar and nucleotide sugar metabolism	16
Cell cycle	16
Glycolysis / Gluconeogenesis	15
Glycerophospholipid metabolism	14
Lysosome	13
Ubiquitin mediated proteolysis	12
AMPK signaling pathway	12
Plant-pathogen interaction	12

Supplementary Table S18. The biological process GO categories that were enriched in the genes containing positively selected codon sites in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to "Figure 2" of main text)

GO Term ID	Description	p-value
GO:0010038	response to metal ion	0.002
GO:0044087	regulation of cellular component biogenesis	0.005
GO:0021700	developmental maturation	0.005
GO:0015850	organic hydroxy compound transport	0.009
GO:0072330	monocarboxylic acid biosynthetic process	0.019
GO:0006325	chromatin organization	0.019
GO:0006928	movement of cell or subcellular component	0.019
GO:0006366	transcription by RNA polymerase II	0.022
GO:0071669	plant-type cell wall organization or biogenesis	0.027
GO:0044419	interspecies interaction between organisms	0.029
GO:1905392	plant organ morphogenesis	0.030
GO:0006638	neutral lipid metabolic process	0.042
GO:0019932	second-messenger-mediated signaling	0.045
GO:0006857	oligopeptide transport	0.045
GO:0009812	flavonoid metabolic process	0.049

Supplementary Table S19. The cellular component GO categories that were enriched in the genes containing positively selected codon sites in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to "Figure 2" of main text)

GO Term ID	Description	p-value
GO:0009505	plant-type cell wall	0.006

Supplementary Table S20. The molecular function GO categories that were enriched in the genes containing positively selected codon sites in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to "Figure 2" of main text)

GO Term ID	Description	p-value

GO:0016798	hydrolase activity, acting on glycosyl bonds	0.006
GO:0001098	basal transcription machinery binding	0.021
GO:0005509	calcium ion binding	0.035

Supplementary Table S21. The distribution of genes containing unique substitutions with functional impact in different eggNOG categories in *Aloe vera* (Related to "Figure 2" of main text)

eggNOG category	Number of genes
Function unknown	642
Signal transduction mechanisms	242
Posttranslational modification, protein turnover, chaperones	224
Carbohydrate transport and metabolism	169
Translation, ribosomal structure and biogenesis	143
Transcription	142
RNA processing and modification	131
Intracellular trafficking, secretion, and vesicular transport	122
Amino acid transport and metabolism	111
Lipid transport and metabolism	97
Energy production and conversion	85
Inorganic ion transport and metabolism	80
Secondary metabolites biosynthesis, transport and catabolism	75
Replication, recombination and repair	68
Cell cycle control, cell division, chromosome partitioning	60
Coenzyme transport and metabolism	43
Cytoskeleton	41
Nucleotide transport and metabolism	37
Chromatin structure and dynamics	36
Cell wall/membrane/envelope biogenesis	32
Defence mechanisms	18

Supplementary Table S22. The distribution of genes containing unique substitutions with functional impact in different KEGG pathways in *Aloe vera* (Only pathways relevant to plants and with more than ten genes are mentioned) (Related to "Figure 2" of main text)

KEGG Pathway	Number of Genes
RNA transport	36
Spliceosome	31
Protein processing in endoplasmic reticulum	30
Ribosome	29
Purine metabolism	26
Ribosome biogenesis in eukaryotes	24
Glycolysis / Gluconeogenesis	23
Cysteine and methionine metabolism	23

Starch and sucrose metabolism	22
Ubiquitin mediated proteolysis	22
mRNA surveillance pathway	21
Endocytosis	20
Cell cycle	20
Pyruvate metabolism	19
Amino sugar and nucleotide sugar metabolism	18
Glycerolipid metabolism	18
Aminoacyl-tRNA biosynthesis	18
Lysosome	17
RNA degradation	16
Plant hormone signal transduction	16
Peroxisome	16
Carbon fixation in photosynthetic organisms	14
Glycerophospholipid metabolism	14
Porphyrin and chlorophyll metabolism	14
AMPK signaling pathway	14
Alanine, aspartate and glutamate metabolism	13
Glycine, serine and threonine metabolism	13
Phenylalanine, tyrosine and tryptophan biosynthesis	13
Glutathione metabolism	13
MAPK signaling pathway - plant	13
HIF-1 signaling pathway	13
Plant-pathogen interaction	13
DNA replication	12
Nucleotide excision repair	12
Cellular senescence	12
Glyoxylate and dicarboxylate metabolism	11
Oxidative phosphorylation	11
Valine, leucine and isoleucine degradation	11
Arginine biosynthesis	11
Terpenoid backbone biosynthesis	11
Proteasome	11
Homologous recombination	11
mTOR signaling pathway	11
Circadian rhythm - plant	11
Pentose phosphate pathway	10
Fructose and mannose metabolism	10
Inositol phosphate metabolism	10
Methane metabolism	10
N-Glycan biosynthesis	10
Various types of N-glycan biosynthesis	10
FoxO signaling pathway	10

Supplementary Table S23. The biological process GO categories that were enriched in the genes containing unique substitutions with functional impact in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to "Figure 2" of main text)

GO Term ID	Description	p-value
GO:0009624	response to nematode	0.001
GO:0030001	metal ion transport	0.003
GO:0016051	carbohydrate biosynthetic process	0.004
GO:0015849	organic acid transport	0.006
GO:0051187	cofactor catabolic process	0.007
GO:0009617	response to bacterium	0.008
GO:0051128	regulation of cellular component organization	0.011
GO:0072521	purine-containing compound metabolic process	0.012
GO:0046777	protein autophosphorylation	0.016
GO:0051094	positive regulation of developmental process	0.018
GO:0070085	Glycosylation	0.022
GO:0007166	cell surface receptor signaling pathway	0.026
GO:0042180	cellular ketone metabolic process	0.026
GO:0006820	anion transport	0.026
GO:0048878	chemical homeostasis	0.028
GO:0009308	amine metabolic process	0.028
GO:0005976	polysaccharide metabolic process	0.030
GO:0009100	glycoprotein metabolic process	0.033
GO:0051240	positive regulation of multicellular organismal process	0.033
GO:0006090	pyruvate metabolic process	0.034
GO:0071669	plant-type cell wall organization or biogenesis	0.034
GO:0062012	regulation of small molecule metabolic process	0.034
GO:0044262	cellular carbohydrate metabolic process	0.036
GO:0016049	cell growth	0.038
GO:0010038	response to metal ion	0.040
GO:0006631	fatty acid metabolic process	0.043
GO:0022603	regulation of anatomical structure morphogenesis	0.048

Supplementary Table S24. The cellular component GO categories that were enriched in the genes containing unique substitutions with functional impact in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to "Figure 2" of main text)

GO Term ID	Description	p-value
GO:0005802	trans-Golgi network	0.002
GO:0000325	plant-type vacuole	0.005
GO:0005768	Endosome	0.018
GO:0098552	side of membrane	0.021
GO:0000139	Golgi membrane	0.040
GO:0099023	tethering complex	0.042

Supplementary Table S25. The molecular function GO categories that were enriched in the genes containing unique substitutions with functional impact in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to "Figure 2" of main text)

GO Term ID	Description	p-value
GO:0008194	UDP-glycosyltransferase activity	0.002
GO:0022803	passive transmembrane transporter activity	0.012
GO:0042562	hormone binding	0.017
GO:0043177	organic acid binding	0.018
GO:0052689	carboxylic ester hydrolase activity	0.020
GO:0005102	signaling receptor binding	0.022
GO:0016874	ligase activity	0.022
GO:0030551	cyclic nucleotide binding	0.034
GO:0043178	alcohol binding	0.034
GO:0016758	transferase activity, transferring hexosyl groups	0.035

Supplementary Table S26. The distribution of MSA genes in different eggNOG categories in *Aloe vera* (Related to "Figure 3" and "Figure 4" of main text)

eggNOG category	Number of genes
Function unknown	31
Translation, ribosomal structure and biogenesis	13
Signal transduction mechanisms	11
Carbohydrate transport and metabolism	11
RNA processing and modification	11
Posttranslational modification, protein turnover, chaperones	10
Inorganic ion transport and metabolism	9
Transcription	6
Intracellular trafficking, secretion, and vesicular transport	6
Cytoskeleton	5
Coenzyme transport and metabolism	5
Energy production and conversion	5
Amino acid transport and metabolism	4
Replication, recombination and repair	4
Cell cycle control, cell division, chromosome partitioning	3
Lipid transport and metabolism	3
Cell wall/membrane/envelope biogenesis	3
Secondary metabolites biosynthesis, transport and catabolism	2
Nucleotide transport and metabolism	1
Chromatin structure and dynamics	1

Supplementary Table S27. The distribution of MSA genes in different KEGG pathways in *Aloe vera* (Only pathways relevant to plants and with more than one gene are mentioned) (Related to "Figure 3" and "Figure 4" of main text)

KEGG Pathway	Number of genes
Ribosome	7
RNA transport	5
Cysteine and methionine metabolism	4
Glycolysis / Gluconeogenesis	3
Spliceosome	3
Aminoacyl-tRNA biosynthesis	3
Circadian rhythm - plant	3
Plant-pathogen interaction	3

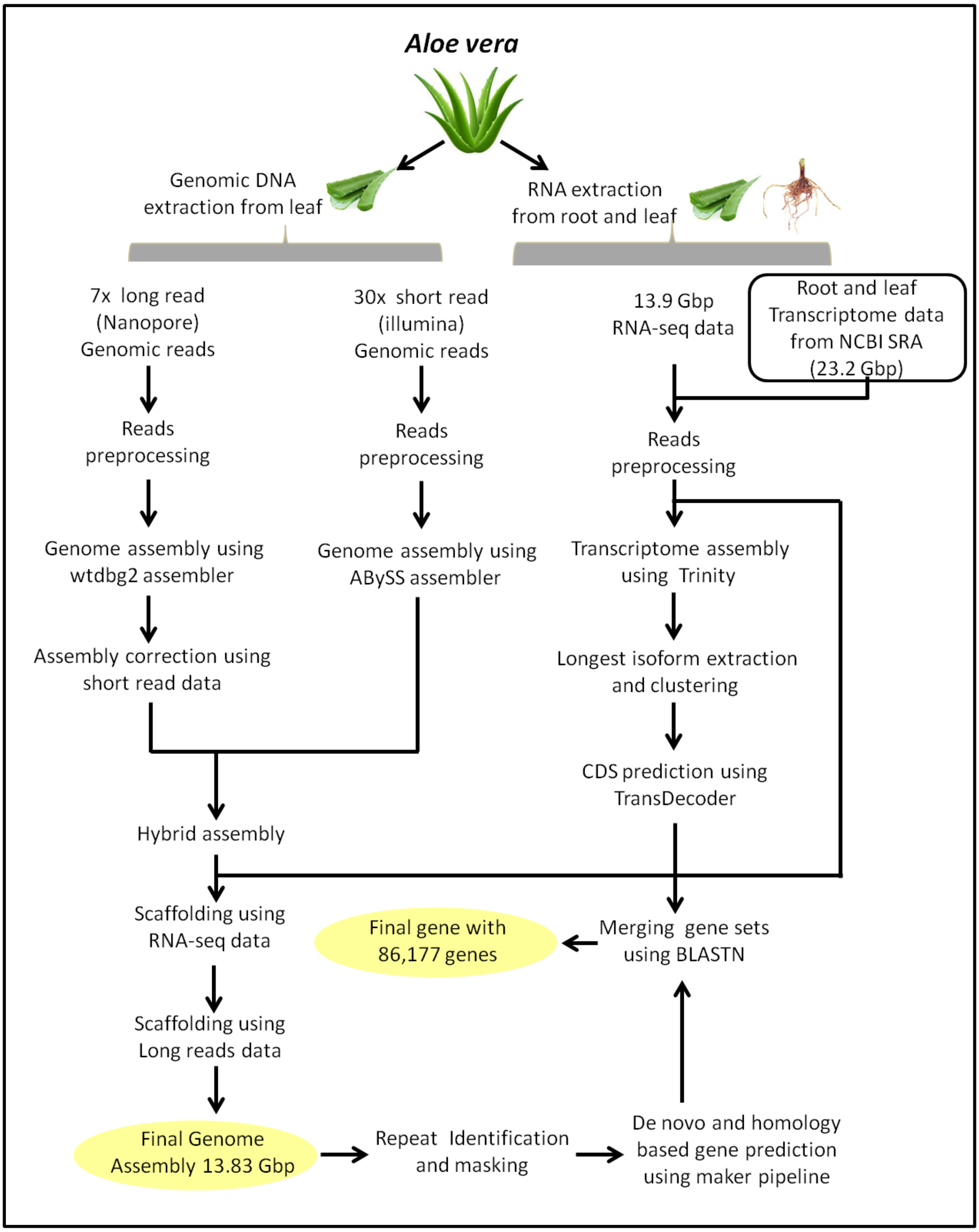
Supplementary Table S28. The biological process GO categories that were enriched in the MSA genes in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to “Figure 3” and “Figure 4” of main text)

GO term ID	Description	p-value
GO:0015748	organophosphate ester transport	0.00263
GO:0010038	response to metal ion	0.00621
GO:0032504	multicellular organism reproduction	0.00868
GO:0046700	heterocycle catabolic process	0.01165
GO:0048589	developmental growth	0.01165
GO:0019439	aromatic compound catabolic process	0.01252
GO:0044270	cellular nitrogen compound catabolic process	0.01252
GO:0016052	carbohydrate catabolic process	0.01535
GO:0016049	cell growth	0.01538
GO:1901361	organic cyclic compound catabolic process	0.01643
GO:0009624	response to nematode	0.01748
GO:0045165	cell fate commitment	0.02348
GO:2000241	regulation of reproductive process	0.02462
GO:0080134	regulation of response to stress	0.02652
GO:0042157	lipoprotein metabolic process	0.02825
GO:0007017	microtubule-based process	0.02968
GO:0008283	cell proliferation	0.02968
GO:0072524	pyridine-containing compound metabolic process	0.02968
GO:1901698	response to nitrogen compound	0.03029
GO:0070482	response to oxygen levels	0.03337
GO:0009735	response to cytokinin	0.03447
GO:0015931	nucleobase-containing compound transport	0.03869
GO:0090351	seedling development	0.03869
GO:0022603	regulation of anatomical structure morphogenesis	0.03883
GO:0021700	developmental maturation	0.04894

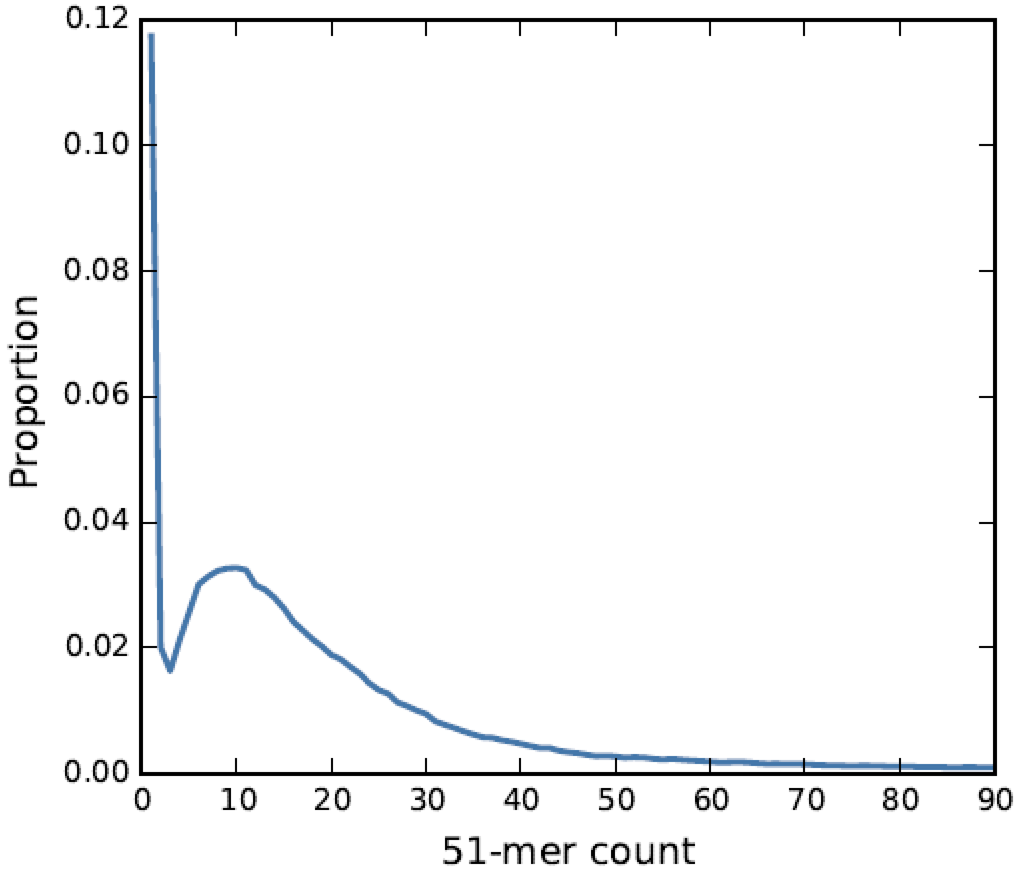
Supplementary Table S29. The molecular function GO categories that were enriched in the MSA genes in *Aloe vera* (Only statistically significant GO terms p<0.05 are mentioned) (Related to "Figure 3" and "Figure 4" of main text)

GO term ID	Description	p-value
GO:0015605	organophosphate ester transmembrane transporter activity	0.01886
GO:0005200	structural constituent of cytoskeleton	0.02383
GO:0015932	nucleobase-containing compound transmembrane transporter activity	0.04144
GO:1901505	carbohydrate derivative transmembrane transporter activity	0.04812

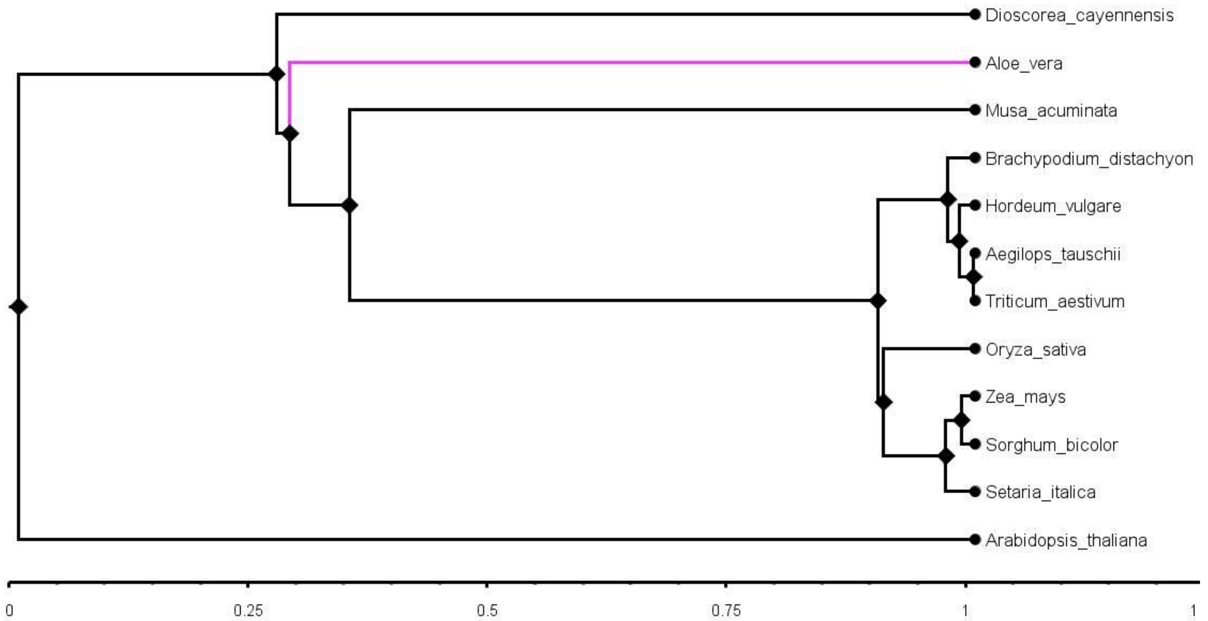
SUPPLEMENTARY FIGURES


Aloe vera
Genomic DNA extraction from leaf
RNA extraction from root and leaf
7x long read (Nanopore) Genomic reads
30x short read (illumina) Genomic reads
13.9 Gbp RNA-seq data
Root and leaf Transcriptome data from NCBI SRA (23.2 Gbp)
Reads preprocessing
Reads preprocessing
Reads preprocessing
Genome assembly using wtdbg2 assembler
Genome assembly using ABySS assembler
Transcriptome assembly using Trinity
Assembly correction using short read data
Longest isoform extraction and clustering
CDS prediction using TransDecoder
Hybrid assembly
Scaffolding using RNA-seq data
Final gene with 86,177 genes
Merging gene sets using BLASTN
Scaffolding using Long reads data
Final Genome Assembly 13.83 Gbp
Repeat Identification and masking
De novo and homology based gene prediction using maker pipeline

Supplementary Figure S1. The complete workflow of the genomic and transcriptomic data analysis for *Aloe vera* (Related to "Figure 1" of main text)



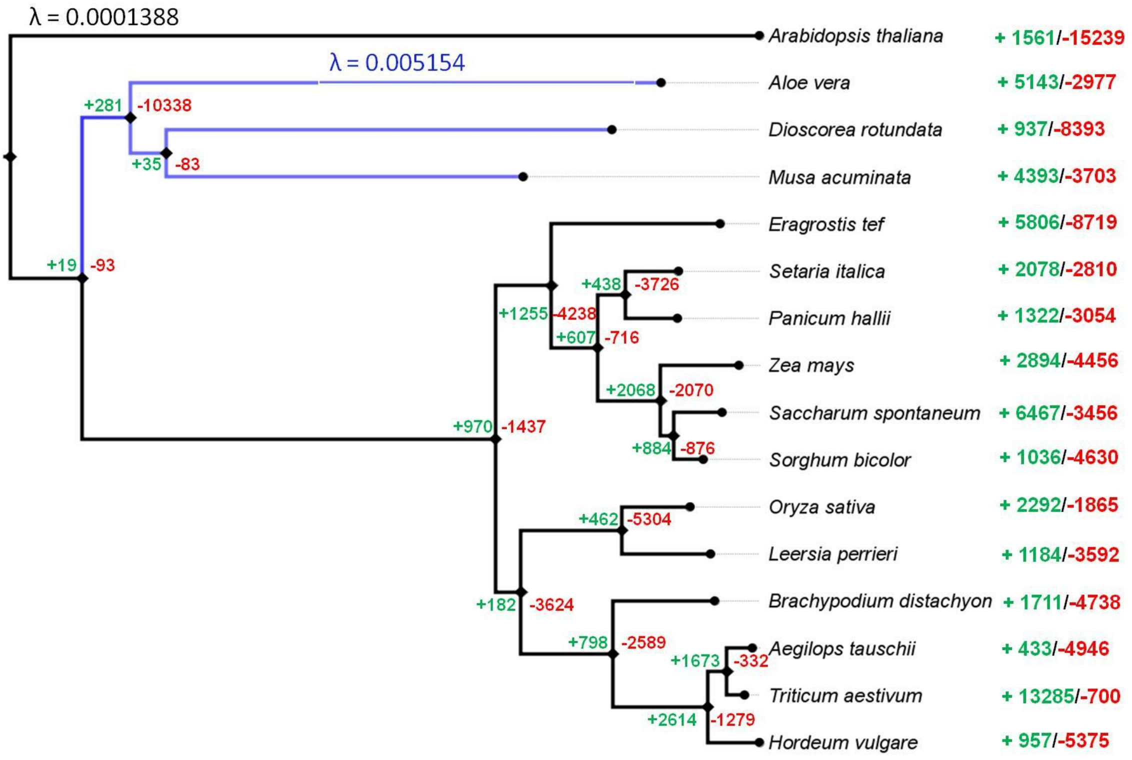
Supplementary Figure S2. k-mer count distribution for the 51-mer. The y-axis is the proportion of 51-mers and x-axis is the count of the 51-mer. (Related to "Figure 1" of main text)



Supplementary Figure S3. The phylogenetic tree of the monocot species that were common in our study and plant megaphylogeny is shown (Qian and Jin, 2016). The *Arabidopsis thaliana* was used as an outgroup.

***Dioscorea rotundata* is a subspecies of *Dioscorea cayennensis* and both are scientific names for white yam**

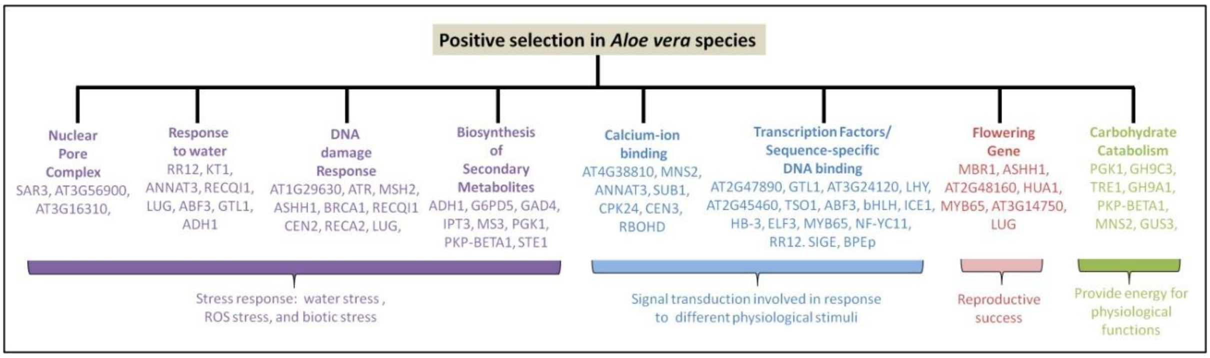
(Related to "Figure 1" of main text)



Supplementary Figure S4. The gene family expansion/contraction for the selected monocot plant species including *Aloe vera* with *Arabidopsis thaliana* as an outgroup.

The number of gene families with expansion and contraction are shown for all the extant species and ancestor nodes. The expansion numbers are shown with "+" symbol in Green colour and the contraction numbers are shown with "-" symbol in Red colour. The clade formed by *Aloe vera*, *Dioscorea rotundata*, and *Musa acuminata* had the λ value of 0.005154 and rest of the phylogeny had the λ value of 0.0001388.

(Related to "Figure 1" of main text)


Positive selection in *Aloe vera* species
Nuclear Pore Complex
SAR3, AT3G56900, AT3G16310,
Response to water
RR12, KT1, ANNAT3, RECQI1, LUG, ABF3, GTL1, ADH1
DNA damage Response
AT1G29630, ATR, MSH2, ASHH1, BRCA1, RECQI1 CEN2, RECA2, LUG,
Biosynthesis of Secondary Metabolites
ADH1, G6PD5, GAD4, IPT3, MS3, PGK1, PKP-BETA1, STE1
Calcium-ion binding
AT4G38810, MNS2, ANNAT3, SUB1, CPK24, CEN3, RBOHD
Transcription Factors/ Sequence-specific DNA binding
AT2G47890, GTL1, AT3G24120, LHY, AT2G45460, TSO1, ABF3, bHLH, ICE1, HB-3, ELF3, MYB65, NF-YC11, RR12. SIGE, BPEp
Flowering Gene
MBR1, ASHH1, AT2G48160, HUA1, MYB65, AT3G14750, LUG
Carbohydrate Catabolism
PGK1, GH9C3, TRE1, GH9A1, PKP-BETA1, MNS2, GUS3,
Stress response: water stress , ROS stress, and biotic stress
Signal transduction involved in response to different physiological stimuli
Reproductive success
Provide energy for physiological functions

Supplementary Figure S5. The functional categories of genes that showed positive selection in *Aloe vera.*

The standard *Arabidopsis thaliana* gene IDs were used in case of genes that did not have a standard gene symbol.

(Related to “Figure 2”, “Figure 3” and “Figure 4” of main text)

TRANSPARENT METHODS

Sample collection and species identification

The *Aloe vera* plant was bought from a plant nursery in Bhopal, India. The pulp or gel from the leaf was scrapped out and the rest was used for the DNA extraction followed by amplification of complete ITS1 and ITS2 (Internal Transcribed Spacer) and Maturase K (MatK) regions for species identification. The DNA was extracted using DNeasy Plant mini kit (Qiagen, United States) and was quantified on Qubit 2.0 fluorometer by using Qubit DNA BR assay kit (Life Technologies, United States). The DNA was used for amplification of complete ITS1 and ITS2 (Internal Transcribed Spacer) and Maturase K (MatK) regions by using following primer sets:

(i)	Complete ITS region:	forward primer 5′-TCCGTAGGTGAACCTGCGG-3′
		reverse primer 5′-TCCTCCGCTTATTGATATGC-3′
(ii)	ITS1 region:	forward primer 5′-TCCGTAGGTGAACCTGCGG-3′
		reverse primer 5′-GCTGCGTTCTTCATCGATGC-3′
(iii)	ITS2 region:	forward primer 5′-GCATCGATGAAGAACGCAGC-3′
		reverse primer 5′-TCCTCCGCTTATTGATATGC-3′
(iv)	Mat K region:	forward primer 5′-CGATCTATTCATTCAATATTTC-3′
		reverse primer 5′-TCTAGCACACGAAAGTCGAAGT-3′

The PCR programme run on Veriti 96 well thermal cycler (Applied Biosystems) for ITS regions was 94 ˚C for 3 mins, 35 cycles of 94 ˚C for 1 min, 55 ˚C for 1 min and 72 ˚C for 2.5 mins and 72 ˚C for 10 mins. Similarly, the programme for MatK was 95 ˚C for 3 mins, 35 cycles of 95 ˚C for 30 sec, 50 ˚C for 3 mins and 72 ˚C for 1:15 min and final extension at 72 ˚C for 7 mins. The amplification products were assessed by running them on 2% agarose gel electrophoresis. The amplified products were purified and sequenced at in-house Sanger sequencing facility. All the sequences were checked for alignment with NCBI database using BLASTN and showed highest identity with *Aloe vera* which confirmed the species as *Aloe vera*.

Genome sequencing

Short read sequencing

DNA extraction: The DNA was initially isolated from DNeasy Plant mini kit (Qiagen, United States). The pulp or gel from the leaf was scrapped out before grinding. While grinding with liq. Nitrogen, 1 ml of AP1 buffer was added to increase the yield. Further steps were followed as given in the kit. The DNA was eluted in 50 µl elution buffer (Qiagen, United States). The DNA was quantified using Qubit DNA BR assay kit on Qubit 2.0 fluorometer (Life Technologies, United States). The library was prepared using NEBNext Ultra II DNA Library preparation Kit for Illumina (New England Biolabs, England) and TruSeq DNA Nano Library preparation kits (Illumina, Inc., United States). The library size was evaluated by Agilent 2100 Bioanalyzer and library was quantified by qPCR. The library was sequenced on Illumina HiSeq X ten platform and NovaSeq 6000 (Illumina, Inc., United States) for 150 bp paired end reads.

Long read sequencing

The DNA extraction for long read sequencing was done by using Carlson lysis buffer [100 mM Tris; 2% CTAB; 1.4 M NaCl; 1% PEG 8000; 20 mM Ethylene read sequencing Diamine Tetra Acetic acid (EDTA)]. To 50 ml of Carlson buffer, β-mercaptoethanol (125 µl) was added and vortexed to mix it properly. The plant part taken for DNA extraction was leaf. The pulp or gel from the leaf was scrapped out before homogenization. The sample was homogenized in liquid nitrogen by using a mortar and pestle (autoclaved and precooled at -20˚C for 30 mins) and transferred to a microcentrifuge tube. Carlson lysis buffer (1 ml) was preheated at 65˚C for 30 mins and added to the sample. After adding 2 µl of RNase A (20 mg/ml) and 25 ul of Proteinase K (20 µL/mL) and vortexing for 5 sec, the sample was incubated at 65˚C for 1 hr. The sample was mixed in between by inverting 10 times. After incubation, the sample was allowed to cool at room temperature for 5 mins. The sample tube was provided with 1 ml of chloroform, vortexed and centrifuged at 5,000 xg for 15 mins at 4˚C. The top aqueous layer was transferred with wide bore tip to a new centrifuge tube. The 0.7X volume of isopropanol was added, mixed by inverting 10 times and incubated at -20˚C for overnight. The tube was centrifuged at 5,000 xg for 45 mins at 4˚C. In conventional method, the supernatant was discarded and pellet was washed with 1 ml of ice-cold 70 % ethanol by centrifuging at 5,000 xg for 10 mins at 4˚C. The supernatant was again discarded and the pellet was air dried to evaporate all of the ethanol residues. The DNA was eluted in 50 µl of nuclease free water.

In kit-based method, Blood and Cell culture kit with Genomic tip 20 (Qiagen, United States) was used. The Pelleted DNA was not washed with 70% ethanol but it was dissolved in G2 buffer by incubating at 50˚C for 30 mins. The dissolved DNA was passed through equilibrated Genomic tip 20. The column was washed thrice with QC buffer (1 ml) and eluted in 1 ml of QF buffer (pre heated at 56˚C). The DNA was allowed to precipitate in 0.7X Isopropanol for overnight at -20˚C. The precipitated DNA was washed and eluted same as in conventional method.

The DNA was quantified with Qubit 2.0 fluorometer by using Qubit DNA BR assay kit (Life Technologies, United States). The DNA quality was checked by running on agarose gel and NanoDrop™ 8000 Spectrophotometer (ThermoFisher Scientific, USA). To reach the required purity of samples, they were purified with Ampure XP beads (Beckman Coulter, USA). The purified samples were used for library preparation by following the protocol Genomic DNA by Ligation using SQK-LSK109 kit (Oxford Nanopore). The library was loaded on FLO-MIN106 Flow cell (R 9.4.1) and sequenced on MinION (Oxford Nanopore, UK) using MinKNOW software (versions 3.4.5 and 3.6.0).

Transcriptome sequencing

The leaf and root part of plant were taken as sample for RNA extraction. The samples were grinded in liquid nitrogen with the help of mortar and pestle. The powdered sample (100 mg) was transferred to centrifuge tube to which 1 ml of TRIzol reagent (Invitrogen, USA) was added and shaken for 5 mins. For complete dissociation of nucleoprotein complexes the tubes were incubated for 5 mins at room temperature. Chloroform (200 µl) was added to the tubes and vortexed for 15 sec and incubated at room temperature for 10 mins. After incubation the tubes were centrifuged at 12,000 xg for 15 mins at 4˚C and upper aqueous phase was transferred to a new centrifuge tube. Isopropanol (500 µl) was added, mixed thoroughly and allowed to precipitate for 5-10 mins at room temperature. The RNA was pellet down by centrifuging at 12,000 xg for 10 mins at 4˚C and supernatant was discarded. The pellet was washed with 1ml of 75% ethanol by centrifuging at 7,500 xg for five mins at 4˚C. The supernatant was discarded and was kept at 37˚C for 30 mins to evaporate the residual ethanol. The RNA pellet was resuspended in 30 ul of nuclease free water, dissolved the pellet by pipette mixing and incubated at 55-60˚C for 10-15 mins (Johnson et al., 2012). The RNA was diluted 10 times and quantified on Qubit 2.0 fluorometer by using Qubit HS Assay kit (Invitrogen, USA). The library was prepared by using TruSeq Stranded mRNA LT Sample Prep kit and following TruSeq Stranded mRNA Sample Preparation Guide (Illumina, Inc., United States) and sequenced on Illumina NovaSeq 6000 platform for 101 basepair paired end reads.

Data preprocessing

The raw Illumina sequence data was processed using the Trimmomatic v0.38 tool (Bolger et al., 2014). The adapters used for the sequencing were trimmed using the parameters: 2 mismatches to be allowed in the seed matching with seed length of 16, palindrome clip threshold of 30, and simple clip threshold of 10. The low quality bases or N's were removed from the leading and trailing ends of the reads with the quality threshold of 15. The reads were scanned with a sliding window of 4 bp and the reads were trimmed when average PHRED quality score per base went below 15. After these steps, all the reads smaller than 60 bp were removed. For nanopore data the raw sequencing reads were obtained in fast5 from the MinKNOW v3.6.0 and basecalling was performed using Guppy v3.2.1. Adapter sequences were then removed based on all known adapters by using Porechop v0.2.3.

Genome characteristics (heterozygosity and genome size estimation)

The percent heterozygosity was estimated using the quality-filtered paired-end short reads to assess the complexity of the genome. The k-mer frequency data was generated using Jellyfish (v2.2.10), and was used for the percent heterozygosity calculations using GenomeScope (v2.0) (Marçais and Kingsford, 2011; Vurture et al., 2017).

SGA-preqc was used to estimate the genome size of *Aloe vera* using a k-mer count distribution method (Simpson and Durbin, 2012). It uses a k-mer count distribution method, where only the k-mers with higher occurrences are considered for genome size estimation. Thus, it reduces the impact of sequencing errors on the estimation which is very useful for higher repeat containing plant genomes. At first the sga preprocess was run (with the option -pe-mode set to 0 to consider all the paired-end and single-end filter reads for the analysis) to preprocess the raw reads. Next, the sga index was run with 'ropebwt' algorithm and --no-reverse option to index the preprocessed reads, and finally, the sga preqc was run with default options for genome size estimation.

Genome assembly

The filtered paired and unpaired Illumina reads were *de novo* assembled using ABySS v2.1.5 with bloom-filter function for a Bloom filter size of 950 GB, the bloom filter hash function and minimum k-mer count threshold for bloom-filter assembly were used as default (Birol et al., 2009). The other parameters were: minimum alignment length of a read of 40 bp, minimum unitig size required for building contigs of 500 bp, and minimum contig size required for building scaffolds of 500 bp. Different assemblies were generated on a sample dataset at different k-mer values: 41, 87, 96, 107, 117, 127. The best assembly resulted on k-mer value of 107 hence, the final assembly on complete data was performed at the k-mer value of 107.

The preprocessed nanopore reads were *de novo* assembled using wtdbg2 v2.0.0 with an estimated genome size of 16 GB, subsampling k-mer value of 1.0, minimum read depth of a valid edge of 2, with keeping the contained reads during alignments (Ruan and Li, 2020). The minimum length of alignment between reads was set to 2,048 bp.

The obtained genome assembly was first corrected for the assembly and sequencing errors using short-read data by SeqBug (Mittal et al., 2019). The scaffolding of short-read assembly was performed by utilizing the long-read assembly using QuickMerge (v0.3) (Chakraborty et al., 2016). All the ABySS contigs were searched in the wtdbg2 contigs and all the matching contigs with the criteria of 50% query coverage, e-value of $<10^{-6}$ and 90% identity were removed, and the unique contigs from ABySS and wtdbg2 assembly were merged to construct a hybrid assembly. The merging of two assemblies using homology search by BLASTN is one of the standard methods that is commonly used by similar studies (Schmidt et al., 2020). To further improve the assembly and to remove the undetermined bases in the assembly, the RNA-seq data based scaffolding and gap closing of the assembly was performed using the long-read data. The RNA-seq data based scaffolding was performed using 'Rascaf' (Song et al., 2016), followed by the long-read based gap-closing performed using LR_Gapcloser to generate the final *Aloe vera* genome assembly.

Genome annotation

The genome annotation was performed on all the contigs of hybrid assembly. The tandem repeats in the genome were identified using the Tandem Repeat Finder (TRF) v4.09 with the parameters: matching weight = 2, mismatching penalty = 7, indel penalty = 7, match probability = 0.8, indel probability = 0.1, minimum alignment score = 50, and maximum period size = 2,000 (Benson, 1999). To identify the interspersed repeats, the *de novo* repeat library was constructed using the nanopore reads (>40 kb) in *Aloe vera* genome using RepeatModeler v2.0.1 (Flynn et al., 2020). The identified repeat families were clustered using CD-HIT-EST v4.8.1 with 90% sequence identity and seed size of 8 bp (Fu et al., 2012). This resultant repeat library was used to identify interspersed repeat elements in *Aloe vera* using RepeatMasker v4.1.0 (RepeatMasker Open-4.0, http://www.repeatmasker.org).

The tRNAs are very large and complex non-coding RNA families and present in all the living organisms. Thus, tRNAs in the *Aloe vera* genome were predicted on the final gap-closed assembly using tRNAscan-SE v2.0.5 on the default parameters (Chan and Lowe, 2019; Lowe and Eddy, 1997). The tRNAscan-SE uses a companion Genomic tRNA Database and UCSC genome browser to identify the tRNAs present in a genome. In this method a total of 3,119 tRNAs were identified, of which standard amino acids related tRNAs were 1,978, possible suppressor tRNAs (CTA, TTA, TCA) were 9, undetermined/unknown isotypes tRNAs were 29, and predicted pseudogenes tRNAs were 1,103. Further, a total of 128 tRNAs with introns were also identified.

A total of 38,589 hairpin miRNAs were retrieved from the miRBase database (Griffiths-Jones et al., 2007). These hairpin miRNAs were clustered separately to remove redundancies using CD-HIT-EST v4.8.1 (Fu et al., 2012). After clustering, the hairpin miRNAs dataset had a total of 22,365 sequences.

The non-redundant sequences were used to identify the hairpin miRNAs in the *Aloe vera* genome using homology-based search by BLASTN alignment tool with the thresholds: identity ≥80% and e-value <1e-03 (Altschul et al., 1990).

Transcriptome assembly

The transcriptome assembly of *Aloe vera* was carried out using the RNA-seq data generated from the root and leaf tissue in this study and previous studies (Choudhri et al., 2018; Wickett et al., 2014). All the quality-filtered paired and unpaired transcriptome sequencing reads were *de novo* assembled using Trinity v2.6.6 software with default parameters to generate the assembled transcripts (Haas et al., 2013). The transcriptome assembly was evaluated by mapping the filtered RNA-seq data on the assembled transcripts using hisat2 v2.1.0 (Kim et al., 2015). The BUSCO score was used to assess the completeness of the transcriptome assembly calculated by BUSCO v4.1.4 software using the standard database specific to the embryophyta clade known as 'embryophyta_odb10' (Simão et al., 2015; Waterhouse et al., 2018).

Gene set construction

The MAKER pipeline was used for gene set construction of the *Aloe vera* genome (Campbell et al., 2014). The soft-masked genome of *Aloe vera* (contigs ≥300 bp) generated using RepeatMasker v4.1.0 with Repbase repeat library (RepeatMasker Open-4.0, http://www.repeatmasker.org) was used for the gene prediction using the MAKER pipeline. Both the *ab initio* and empirical evidence were used for the gene predictions. The *Aloe vera* EST evidence from the RNA-seq assembly of *Aloe vera* species, protein sequences of the closest species *Dioscorea rotundata* and *Musa acuminata*, and *ab initio* gene predictions of the *Aloe vera* genome were used to construct the gene set using the MAKER pipeline. AUGUSTUS v3.2.3 was used for the *ab initio* gene prediction, and the BLAST alignment tool was used for homology-based gene prediction using the EST evidence in the MAKER pipeline (Altschul et al., 1990; Stanke et al., 2006; Stanke et al., 2004). Further, Exonerate v2.2.0 was used to polish and curate the BLAST alignment results (https://github.com/nathanweeks/exonerate). The evidence from *ab initio* and homology-based methods were integrated to perform the final gene predictions.

The genes from predicted transcripts were identified by extracting the longest isoforms. The unigenes were identified by performing the clustering using CD-HIT-EST v4.8.1 program, and coding regions were predicted using TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder) (Bateman et al., 2004; Buchfink et al., 2015; Finn et al., 2011; Fu et al., 2012; Suzek et al., 2015). The gene set constructed using the MAKER pipeline and transcriptome assembly was filtered, and only the genes with ≥300 bp length were considered further. The clustering of remaining MAKER pipeline based genes was performed using CD-HIT-EST v4.8.1 program with 95% identity and a seed size of 8 bp (Fu et al., 2012). The transcriptome gene set was searched in the MAKER gene set using BLASTN. The genes from the transcriptome assembly gene set that matched to the MAKER gene set with the parameters: identity ≥50%, e-value $<10^{-9}$, and query coverage ≥50% were removed. The remaining genes for the transcriptome assembly gene set were directly added to the MAKER gene set to construct the final gene set of *Aloe vera*. This approach of gene set construction is a standard method and has been used for other genomes (Chakraborty et al., 2020; Cho et al., 2013; Jaiswal et al., 2018).

Orthogroups identification

For orthogroups identification, the representative of monocot species from all the genera, for which high-quality genomes were available on Ensembl plants database, were selected along with an outgroup species, the model plant *Arabidopsis thaliana*. The selected monocot species were *Aegilops tauschii*, *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*, and *Zea mays*. The proteome files containing all the protein sequences of the 15 species retrieved from Ensembl plants release 46 (Zerbino et al., 2018), and the protein-coding genes from the transcriptome assembly of *Aloe vera* were used to construct the orthogroups. The longest transcript for each gene was extracted for each species using in-house python scripts. The proteome files with longest transcripts were used for the orthogroups identification using OrthoFinder v2.3.9 (Emms and Kelly, 2019). The OrthoFinder v2.3.9 analysis included a total of 16 species, i.e., 14 monocot species, the model species *Arabidopsis thaliana* as an outgroup, and *Aloe vera* sequenced in this study.

Orthologous gene set construction

From the orthogroups identified by the OrthoFinder analysis, the orthogroups with the taxon count of 16 were extracted, which included genes from each of the 16 species. A total of 5,472 orthogroups were extracted using this criterion. Only the longest gene of each species was retained in each of these orthogroups to construct the orthologous gene set. Thus, a total of 5,472 orthologs were identified across 16 species. From these, 5,472 orthologs one-to-one orthologs were extracted. To include maximum number of genes in the one-to-one orthology, the fuzzy one-to-one orthogroups instead of true one-to-one orthogroups were identified using KinFin v1.0 (Laetsch and Blaxter, 2017). A total of 1,440 one-to-one orthologs were extracted using this method across the selected 16 species.

Phylogenetic tree construction

The phylogenetic species tree was constructed with the fuzzy one-to-one orthologous genes. The individual orthologous sets were aligned using MAFFT v7.455 (Katoh and Standley, 2013). The alignments were trimmed using BeforePhylo v0.9.0 (https://github.com/qiyunzhu/BeforePhylo) to remove the poorly aligned regions. All protein sequence alignments of orthologs across 16 species were concatenated using BeforePhylo v0.9.0, followed by species phylogenetic tree construction using RAxML v8.2.12 (Stamatakis, 2014). The maximum likelihood phylogenetic tree was constructed using the rapid hill climbing algorithm with 100 bootstrap replicates. Since the amino acid sequences were used, the 'PROTGAMMAGTR' substitution model was utilized to construct the species tree.

Divergence Time estimations

The divergence time was estimated using MCMCTree methodology of the PAML (v4.9) package (Yang, 2007). Supermatrix from the concatenated protein sequence alignments of fuzzy one-to-one orthologs across 16 species and the maximum likelihood phylogeny constructed in the previous step were used for the MCMCTree analysis. Two calibration points, one for the divergence of *Aloe vera* and *Dioscorea rotundata*, and the other for the divergence of *Aloe vera* and *Aegilops tauschii* were used from the TimeTree database (Hedges et al., 2015), which is a public database containing information on evolutionary times for the tree of life for more than 50,632 species from the published studies.

Gene family expansion and contraction analysis

The gene family expansion and contraction analysis was performed using CAFE (v4.2.1) with a random birth and death model for estimating the gene gain and loss in *Aloe vera* genome (Han et al., 2013). The species phylogenetic tree constructed in this study using fuzzy one-to-one orthologs was converted into ultrametric tree using the divergence time between *Aloe vera* and *Dioscorea rotundata* of 122 million years as the calibration point. This ultrametric tree was used for the expansion and contraction analysis with *Arabidopsis thaliana* as an outgroup. For each gene, only the longest protein isoform sequence was used for all of the species. An all-versus-all homology search was performed for all the protein sequences from each species using BLASTP, and the output was utilized for performing the clustering using MCL (v14.137) (Van Dongen and Abreu-Goodger, 2012). The gene families with ≥ 100 gene copies in any one or more species were removed and the remaining were used for further analysis. The random birth and death-based two-lambda model was used for the gene expansion and contraction analysis using CAFE methodology. In the two-lambda model, the clade formed by *Aloe vera*, *Dioscorea rotundata*, and *Musa acuminata* were given one lambda value, and the rest of the species were assigned with the other lambda value.

Identification of genes with a higher rate of evolution

The genes that show higher root-to-tip branch length are considered to have a higher rate of nucleotide divergence or mutation, indicating a higher rate of evolution. For this analysis, the individual maximum likelihood phylogenetic trees were constructed using the protein sequences of the 5,472 orthologs identified across the 16 species. The maximum likelihood phylogenetic trees with 100 bootstrap replicates were constructed using the rapid hill climbing algorithm with the 'PROTGAMMAGTR' substitution model by using RAxML v8.2.12 (Stamatakis, 2014). The root-to-tip branch length values were calculated for each of the 16 extant species using the 'adephylo' package in R (Jombart and Dray, 2010; Jombart et al., 2017). All the genes that showed a significantly higher root-to-tip branch length for *Aloe vera* in comparison to rest of the species were extracted using in-house Perl scripts and were considered to be the genes with a higher rate of evolution in *Aloe vera*.

Identification of positively selected genes

The positively selected genes in *Aloe vera* were identified using the branch-site model implemented in the PAML software package v4.9a (Yang, 2007). An iterative program for sequence alignment, SAT'e, was utilized to perform the alignments of the 5,472 ortholog protein sequences. The combination of Prank, MUSCLE, and RaxML was used to perform the SAT'e based alignment to control the false positives and false negatives in the alignment (Liu et al., 2012). The protein-sequence alignment guided codon alignment was performed for the 5,472 ortholog nucleotide sequences using ‘TRANALIGN’ program of EMBOSS v6.5.7 package (Rice et al., 2000). The ‘codeml’ was run on ortholog codon alignments using the species phylogenetic tree constructed in previous steps. The alignments were filtered for the ambiguous codon sites and gaps and only the clean sites were considered for the positive selection analysis. The likelihood ratio tests were performed using the log-likelihood values for the null and alternative models, and the p-values were calculated based on the χ^2-distribution. Further, the FDR corrected p-values or FDR q-values were also calculated. All genes with FDR-corrected p-values <0.05 were considered to be the genes with positive selection in *Aloe vera*. Further, all codon sites with >0.95 probability of being positively selected in the ‘foreground’ branch based on the Bayes Empirical Bayes analysis were considered to be the positively selected codon sites in a gene.

Identification of genes with unique substitutions that have functional impact

The genes with unique amino acid substitutions in *Aloe vera* species in comparison to all the selected species were identified. The protein alignments for the 5,472 orthologs were generated using the MAFFT v7.455 (Katoh and Standley, 2013). The positions that are identical in all the species but different in *Aloe vera* were identified and considered to be the sites with unique amino acid substitutions in *Aloe vera*. In this analysis, the gaps were ignored, and also the sites with gaps present in the 10 amino acids flanking regions on both sides were ignored. This step helped in considering only the sites with proper alignment for the unique substitution analysis. The identification of unique amino acid sites was performed by using the in-house python scripts. The functional impact of the unique amino acid substitutions on the protein function was identified using the Sorting Intolerant From Tolerant (SIFT) tool with UniProt database as reference (Boutet et al., 2007; Ng and Henikoff, 2003).

Identification of genes with multiple signs of adaptive evolution (MSA)

The genes that showed at least two signs of adaptive evolution among the three signs of adaptive evolution tested above (higher rate of evolution, positive selection, and unique substitution with functional impact) were considered as the genes with multiple signs of adaptive evolution or MSA genes in *Aloe vera*.

Functional annotation

The functional annotation of gene sets was performed using multiple methods. The functional annotation and functional categorization of genes into different eggNOG categories was performed using the eggNOG-mapper (Huerta-Cepas et al., 2017). The genes were assigned to different KEGG pathways, and also the KEGG orthology was determined using the most updated KAAS genome annotation server (Moriya et al., 2007). The gene ontology enrichment or GO term enrichment analysis was performed using the WebGestalt web server (Liao et al., 2019). In the over representation analysis, only the GO categories with the p-value <0.05 in the hypergeometric test were considered to be functionally enriched in the gene set. Further, the functional annotation of genes was also manually curated. The assignment of genes to the specific categories and phenotypes was performed by manual annotation. The protein-protein interaction and co-expression data were extracted from the STRING database, and the network analysis was performed using Cytoscape (Shannon et al., 2003; Szklarczyk et al., 2016).

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., and Sonnhammer, E.L. (2004). The Pfam protein families database. Nucleic acids research *32*, D138-D141.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research *27*, 573-580.

Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y., Hirst, M., and Schein, J.E. (2009). De novo transcriptome assembly with ABySS. Bioinformatics *25*, 2872-2877.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). Uniprotkb/swiss-prot. In Plant bioinformatics (Springer), pp. 89-112.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature methods *12*, 59.

Campbell, M.S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. Current protocols in bioinformatics *48*, 4.11. 11-14.11. 39.

Chakraborty, A., Mahajan, S., Jaiswal, S.K., and Sharma, V.K. (2020). Genome sequencing of turmeric provides evolutionary insights into its medicinal properties. bioRxiv.

Chakraborty, M., Baldwin-Brown, J.G., Long, A.D., and Emerson, J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic acids research *44*, e147-e147.

Chan, P.P., and Lowe, T.M. (2019). tRNAscan-SE: searching for tRNA genes in genomic sequences. In Gene Prediction (Springer), pp. 1-14.

Cho, Y.S., Hu, L., Hou, H., Lee, H., Xu, J., Kwon, S., Oh, S., Kim, H.-M., Jho, S., and Kim, S. (2013). The tiger genome and comparative analysis with lion and snow leopard genomes. Nature communications *4*, 2433.

Choudhri, P., Rani, M., Sangwan, R.S., Kumar, R., Kumar, A., and Chhokar, V. (2018). De novo sequencing, assembly and characterisation of Aloe vera transcriptome and analysis of expression profiles of genes related to saponin and anthraquinone metabolism. BMC genomics *19*, 427.

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome biology *20*, 1-14.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic acids research *39*, W29-W37.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and Smit, A.F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences *117*, 9451-9457.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150-3152.

Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. (2007). miRBase: tools for microRNA genomics. Nucleic acids research *36*, D154-D158.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., and Lieber, M. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols *8*, 1494.

Han, M.V., Thomas, G.W., Lugo-Martinez, J., and Hahn, M.W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Molecular biology and evolution *30*, 1987-1997.

Hedges, S.B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. Molecular biology and evolution *32*, 835-845.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., Von Mering, C., and Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Molecular biology and evolution *34*, 2115-2122.

Jaiswal, S.K., Gupta, A., Saxena, R., Prasoodanan, V.P., Sharma, A.K., Mittal, P., Roy, A., Shafer, A., Vijay, N., and Sharma, V.K. (2018). Genome sequence of peacock reveals the peculiar case of a glittering bird. Frontiers in genetics *9*, 392.

Johnson, M.T., Carpenter, E.J., Tian, Z., Bruskiewich, R., Burris, J.N., Carrigan, C.T., Chase, M.W., Clarke, N.D., Covshoff, S., and dePamphilis, C.W. (2012). Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. Plos one *7*.

Jombart, T., and Dray, S. (2010). adephylo: exploratory analyses for the phylogenetic comparative method. Bioinformatics *26*, 1-21.

Jombart, T., Dray, S., and Dray, M.S. (2017). Package 'adephylo'.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution *30*, 772-780.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature methods *12*, 357-360.

Laetsch, D.R., and Blaxter, M.L. (2017). KinFin: software for Taxon-Aware analysis of clustered protein sequences. G3: Genes, Genomes, Genetics *7*, 3349-3357.

Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic acids research *47*, W199-W205.

Liu, K., Warnow, T.J., Holder, M.T., Nelesen, S.M., Yu, J., Stamatakis, A.P., and Linder, C.R. (2012). SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Systematic biology *61*, 90.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res *25*, 955-964.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764-770.

Mittal, P., Jaiswal, S.K., Vijay, N., Saxena, R., and Sharma, V.K. (2019). Comparative analysis of corrected tiger genome provides clues to its neuronal evolution. Scientific reports *9*, 1-11.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic acids research *35*, W182-W185.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research *31*, 3812-3814.

Qian, H., and Jin, Y. (2016). An updated megaphylogeny of plants, a tool for generating plant phylogenies and an analysis of phylogenetic community structure. Journal of Plant Ecology *9*, 233-239.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite (Elsevier current trends).

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nature methods *17*, 155-158.

Schmidt, H., Hellmann, S.L., Waldvogel, A.-M., Feldmeyer, B., Hankeln, T., and Pfenninger, M. (2020). A high-quality genome assembly from short and long reads for the non-biting midge Chironomus riparius (Diptera). G3: Genes, Genomes, Genetics *10*, 1151-1157.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research *13*, 2498-2504.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210-3212.

Simpson, J.T., and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. Genome research *22*, 549-556.

Song, L., Shankar, D.S., and Florea, L. (2016). Rascaf: improving genome assembly with RNA sequencing data. The plant genome *9*.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312-1313.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic acids research *34*, W435-W439.

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic acids research *32*, W309-W312.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., and Consortium, U. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926-932.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., and Bork, P. (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research, gkw937.

Van Dongen, S., and Abreu-Goodger, C. (2012). Using MCL to extract clusters from networks. In Bacterial Molecular Networks (Springer), pp. 281-295.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., and Schatz, M.C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics *33*, 2202-2204.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular biology and evolution *35*, 543-548.

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., and Gitzendanner, M.A. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences *111*, E4859-E4868.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution *24*, 1586-1591.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., and Girón, C.G. (2018). Ensembl 2018. Nucleic acids research *46*, D754-D761.