Published in final edited form as:
Nat Genet. 2015 December ; 47(12): 1435–1442. doi:10.1038/ng.3435.

The pineapple genome and the evolution of CAM photosynthesis

A full list of authors and affiliations appears at the end of the article.

Abstract

Pineapple (*Ananas comosus* (L.) Merr.) is the most economically valuable crop possessing crassulacean acid metabolism (CAM), a photosynthetic carbon assimilation pathway with high water use efficiency, and the second most important tropical fruit after banana in terms of international trade. We sequenced the genomes of pineapple varieties 'F153' and 'MD2', and a wild pineapple relative *A. bracteatus* accession CB5. The pineapple genome has one fewer ancient whole genome duplications than sequenced grass genomes and, therefore, provides an important reference for elucidating gene content and structure in the last common ancestor of extant members of the grass family (Poaceae). Pineapple has a conserved karyotype with seven pre rho duplication chromosomes that are ancestral to extant grass karyotypes. The pineapple lineage has transitioned from C_3 photosynthesis to CAM with CAM-related genes exhibiting a diel expression pattern in photosynthetic tissues using beta-carbonic anhydrase (βCA) for initial capture of CO_2. Promoter regions of all three β*CA* genes contain a CCA1 binding site that can bind circadian core oscillators. CAM pathway genes were enriched with *cis*-regulatory elements including the morning (CCACAC) and evening (AAAATATC) elements associated with regulation of circadian-clock genes, providing the first link between CAM and the circadian clock regulation. Gene-interaction network analysis revealed both activation and repression of regulatory elements that control key enzymes in CAM photosynthesis, indicating that CAM evolved by reconfiguration of pathways preexisting in C_3 plants. Pineapple CAM photosynthesis is the result of regulatory neofunctionalization of preexisting gene copies and not acquisition of neofunctionalized genes via whole genome or tandem gene duplication.

Christopher Columbus arrived in Guadeloupe in the West Indies on November 4, 1493 during his second voyage to the New World. At a Carib village, he and his sailors encountered pineapple plants and fruit with its astonishing flavor and fragrance that delighted them then and us today. At that time, pineapple was already cultivated on a continental-wide scale following its initial domestication in northern South America, possibly more than 6000 BP[1]. By the end of the 16th century pineapple had become pantropical. Due to the success of industrial production in Hawaii in the last century, pineapple is now not only a routine part of our diet, but has also captured public imagination and become part of pop culture[2,3]. Today, pineapple is cultivated on 1.02 million hectares of land in over 80 countries worldwide, producing 24.8 million metric tonnes of fruit annually with a gross production value approaching 9 billion US dollars[4]. Pineapple has outstanding nutritional and medicinal properties[2] and is a model for studying the evolution of

[†] To whom correspondence should be addressed (rming@life.uiuc.edu; paull@hawaii.edu; qyu@ag.tamu.edu).
[*] These authors contributed equally to this work.

crassulacean acid metabolism (CAM) photosynthesis, which has arisen convergently in many arid regions[5]. Cultivated pineapple, *Ananas comosus* (L.) Merr., is self-incompatible[6], but wild species are self-compatible, providing an opportunity to dissect the molecular basis of self-incompatibility in monocots. As a member of the Bromeliaceae, the pineapple lineage diverged from the lineage leading to grasses (Poaceae) early in the history of the Poales about 100 million years ago[7,8], offering an outgroup and evolutionary reference for investigating cereal genome evolution.

The genome of pineapple variety 'F153', cultivated by Del Monte for 80 years, was sequenced and assembled using data from several sequencing technologies, including 400× coverage of Illumina, 2× coverage of Moleculo synthetic long reads, 1× coverage using 454 sequencing, 5× coverage of PacBio single molecule long reads, and 9,400 bacterial artificial chromosomes (BACs) (see Methods). Due to self-incompatibility, pineapple is cultivated through clonal propagation and, like grape and apple, contains high levels of heterozygosity. To overcome the difficulties of assembling this highly heterozygous genome, we applied a genetic approach to reduce the complexity of the genome utilizing a cross between 'F153' and *Ananas bracteatus* (Lindl.) Schult. & Schult.f. CB5 from Brazil, generating 100× CB5 and 120× F1 genome sequences. Because the F1 contains a haploid genome of both 'F153' and CB5, its sequences were used for haplotype phasing to improve the assembly (see Methods, Supplementary Table 1). The final assembly using this approach substantially improved over the initial Illumina-only assembly, and spans 382 Mb, 72.6% of the estimated 526 Mb pineapple genome[9]. The contig N50 is 126.5 kb and scaffold N50 is 11.8 Mb (Supplementary Table 2). Transposable elements (TEs) accounted for 44% of the assembled genome and 69% of raw reads, indicating 25% of the unassembled genome consists of TEs. The remaining 2.4% are centromeres, telomeres, rDNAs, and other highly repetitive sequences. GC content is 38.3% genome-wide and 51.4% in coding sequences. Endophytic bacterial sequences were identified from raw reads but not in the assembled pineapple genome.

We sequenced 93 F1 individuals from the cross between *A. comosus* 'F153' and *A. bracteatus* CB5 at 10x genome equivalents each, and identified single nucleotide polymorphisms (SNPs) using the 'F153' genome as a reference. Only SNPs that were homozygous for the reference genotype in one parent and heterozygous in the other parent were used, yielding 296,896- segregating SNPs from 'F153'. A genetic map was constructed for 'F153', spanning 3208.6cM at an average of 98.4kb/cM, resulting in 25 linkage groups corresponding to the haploid chromosome number. A total of 564 scaffolds were anchored to the genetic map, covering 316 Mb or 82.7% of the assembled genome (Supplementary Table 3). Scaffolds that mapped to multiple linkage groups were re-assembled with the break points approximated using the information from individual SNPs (2), correcting 119 chimeric scaffolds. Among 18telomeric tracks found, 16 were at the ends of linkage groups (Supplementary Table 4).

We used MAKER to generate a first-pass gene annotation[10]. Each *ab initio* gene model was evaluated against matching transcript and protein evidence to select the most consistent model based on the AED metric. For the final gene set, a MAKER run without repeat masking was selected, followed by extensive filtering of TE-related genes. The original

MAKER run produced 31,893 genes, from which we removed 4,850 TE-related genes, and 19 that were broken during linkage group construction. Among the 27,024 remaining genes, we obtained 24,063 (89.0%) complete gene models, with 11% categorized as partial (Supplementary Table 5). Analysis of transcriptome sequences revealed 10,151 alternative splicing events with intron retentions accounting for 62.8% (Supplementary Table 6). Sequencing small RNA libraries from leaves, flowers and fruits and their analyses revealed 32 miRNA families, including 21 conserved and 11 pineapple specific (Supplementary Table 7).

Transposable elements and expression patterns of LTR retrotransposons

The pineapple genome assembly was searched for TEs that exhibit homology (>80% identity threshold) to currently known TEs. Long terminal repeat (LTR) retrotransposons were identified using structural criteria[11,12]. About 44% of the assembly was accounted for by TEs (Supplementary Table 8). As in other angiosperms, LTR retrotransposons were the most abundant type of TE, representing 33% of the assembly. However, repetitive sequences are under-represented in most shotgun assemblies because identical copies of the same TE are often collapsed into a single sequence and/or masked during the assembly process. We compared the abundance of LTR retrotransposons in the assembly and in the raw reads. The most abundant elements were under-represented in the assembly because of an obligate masking step (Supplementary Table 9). In the most dramatic difference, the Pusofa family made up 28% of all LTR retrotransposon-related sequences in raw reads, but only accounted for 0.5% of all LTR retrotransposon-related sequences in the assembly. In contrast, Wufer, the most abundant family in the assembly (7% of LTR retrotransposons), accounted for ~1.7% of LTR retrotransposons in raw reads. Screening of the raw sequence reads revealed that at least 52% of the nuclear genome is derived from LTR retrotransposons, indicating a total TE content of 69% in the pineapple genome. The abundance of Pusofa, accounting for 28% LTRs and 15% of the pineapple genome, is particularly interesting, because this level of dominance by a single transposable element family is not generally observed. In addition, we identified 20 separate cases in which an LTR retrotransposon had incorporated fragments from one or two genes into the interior of the TE. Interestingly, a recent wave of LTR retrotransposon insertion appears to have occurred in the pineapple lineage about 1.5–2 million years ago (Fig. 1).

About 0.26% of RNA-Seq reads from nine tissues originated from LTR retrotransposons, ranging from 0.16% to 0.52% per tissue (Supplementary Table 10). High LTR expression levels correlates with relatively low copy number (Supplementary Fig. 1). In reads that were mapped to intact elements (0.05% of RNA-Seq reads), the most abundantly expressed family was Sira, a *Copia* element expressed in all nine tissues and accounting for 13% of all LTR retrotransposons expressed, but only 0.2% of LTR retrotransposons in raw reads. An inverse correlation between expression level and LTR retrotransposon abundance has been noted[13], and is indicated here (Supplementary Fig. 1). Different element families exhibited different expression biases as Sira was most highly expressed in flower, Beka in mature fruit, and Ovalut in young fruit (Supplementary Table 9, Supplementary Fig. 2). Individual elements within a family contributed differentially to total family RNA reads. For instance, of the 4 subfamilies of Sira, subfamily sira_1 contributed 96% of RNA-Seq reads mapped to this family. The tissue specificities appeared to be largely the same for each subfamily of any given family (Supplementary Fig. 3). In plants and animals, expression of retrotransposons is dynamic across tissue types, developmental stages and under various stresses and the differentially expressed retroelements discussed here may influence pineapple development.

Heterozygosity in ‘F153’, ‘MD2’, and CB5

Pineapple is cultivated through clonal propagation and is expected to have high levels of residual within genome heterozygosity like other clonal crops such as grape and apple. Breeding efforts have been minimal since the pineapple research institute was dissolved in 1975 and the global pineapple industry is dominated by a small handful of cultivars with limited genetic diversity. ‘MD2’ has been the dominant pineapple variety for the global fresh fruit market for the last 30 years and is a hybrid from the Pineapple Research Institute in Hawaii with a complex pedigree through 5 generations of hybridization. We sequenced the genomes of ‘MD2’ and a wild accession of *A. bracteatus* CB5 at 100× coverage using Illumina paired end reads with different insert size libraries. *De novo* assembly of these two genomes yielded short contigs due to heterozygosity within each coupled with their complex genome structures, demonstrating the effectiveness of using F1 sequences and longer sequence reads for assembling a heterozygous genome. The ‘F153’ genome was used as a reference for assembling these two genomes and assessment of within genome heterozygosity. ‘F153’ has a combined heterozygosity of 1.89% with 1.54% SNPs and 0.35% indels which is similar to ‘MD2’ which has 1.98% heterozygosity with 1.71% SNPs and 0.27% indels. The wild *A. bracteatus* CB5 has higher heterozygosity at 2.93% with 2.53% SNPs and 0.40% indels (Supplementary Table 11). Two homologous pairs of ‘F153’ BACs were identified by probes designed from coding genes and sequenced by Sanger methods to verify heterozygosity rates, which were 2.13% with 1.21% SNPs and 0.92% indels, indicating an underestimation of indels in the three genomes due to the use of a single reference sequence and the technical limitations of aligning reads at such high rates of heterozygosity. The vast majority of heterozygous sites are intergenic but ‘F153’ and ‘MD2’ have 100,743 and 91,876 synonymous and 195,488 and 323,836 non-synonymous sites respectively (Supplementary Table 11). CB5 has 186,520 synonymous and 351,908 non-synonymous sites.

Pineapple karyotype evolution

Intra-genomic syntenic analyses of pineapple show clear evidence of at least two ancient whole genome duplication events (WGDs). Structural comparison of pineapple *vs.* itself revealed 388 intra-genomic blocks including 4,891 pineapple gene pairs derived from WGDs (Supplementary Fig. 4 and 5). Collectively, these collinear blocks span 64% of the annotated gene space and involve each of the 25 pineapple linkage groups, providing strong support for the presence of WGDs. Syntenic depth analyses [14,15] indicated that 35% of the pineapple genome has more than one duplicated segment, as expected if more than one WGD occurred in the pineapple lineage.

The chromosomal organization of pineapple reflects its evolutionary trajectory following the σ and τ whole genome duplications [14,15], starting from a 7-chromosome ancestral monocot genome. We organized the 25 extant chromosomes into major groups corresponding to regions most clearly identifiable as originating from one of the 7 pre-τ chromosomes, Anc1 to Anc7 (Fig. 2). After τ WGD, we inferred 14 chromosomes, which we call $Anc1_1$, $Anc1_2$, $Anc2_1$, $Anc2_2$, $Anc3_1$, $Anc3_2$, $Anc4_1$, $Anc4_2$, $Anc5_1$, $Anc5_2$, $Anc6_1$, $Anc6_2$, $Anc7_1$ and $Anc7_2$. Disrupting this general one-to-one pairing, a translocation of $Anc5_1$ into $Anc3_1$ can be inferred, as well as translocations of $Anc5_2$ into $Anc4_2$ and part of $Anc4_2$ into $Anc3_2$. These events reduced the karyotype to 12 pre-σ chromosomes.

Immediately following the σ event, there were 24 chromosomes, which merged into the 16 extant chromosomes – 3, 4, 8, 10, 11, 12, 13, 14, 16, 17, 18, 19, 21, 22, 23 and 25. One copy of $Anc2_2$ appears to have inserted into one $Anc1_1$ copy to produce extant chromosome 5 while the other $Anc2_2$ copy appears to have fused with one $Anc3_2$ copy to produce chromosome 1. The simplest model suggests that two Anc1 chromosome fissions and one Anc7 chromosome fission produced chromosomes 12, 20 and 24 (Fig. 2).

The high level of retention of most chromosomal identities from the two ancestral monocot WGD events makes pineapple a conservative reference genome for monocots, at least at the level of gene order. Pineapple has few chromosomal rearrangements, and has kept 25 of 28 potential chromosomes as expected from two doublings starting from 7 ancestral chromosomes (7×2×2=28). Similarly, the grapevine genome has played a crucial role in clarifying eudicot genome evolution [16] with 17 of 21 intact chromosomes predicted from the whole genome triplication γ event giving rise to much of the eudicot clade, also from 7 ancestral chromosomes (7×3=21) [17]. The pineapple genome could serve the same comparative role for the monocots because it has conserved most of its karyotype structure during its genome evolution.

Whole genome duplications in pineapple and revised dating of key monocot WGD events

Syntenic analysis of the pineapple genome clarified the genome duplication history of the monocot lineage. We validated and refined phylogenetic dating of three whole genome duplications (WGDs) inferred by previous studies [14,15,17] (Fig. 3A). While the pan-cereal genome duplication event (ρ) is relatively well studied [15], the exact timing of more ancient WGDs (σ and τ) remained controversial because of the high level of degeneration of phylogenetic signals and lack of proper outgroups for each duplication event [14,18]. Because of the pivotal phylogenetic position of pineapple at the base of Poales, we circumscribed the placement of these ancient events based on an integrated syntenic and phylogenetic approach [17,19,20].

Up to four pineapple regions can be aligned to each genomic region in the basal angiosperm *Amborella* (Fig. 3B), that has not experienced WGD since its lineage last shared a common ancestor with all other angiosperms [20]. Both the *Amborella vs.* pineapple comparison and the pineapple self-comparison support two genome doublings in pineapple since its divergence from a shared ancestor with *Amborella*. Microsynteny comparisons to *Amborella* show typical patterns of independent fractionations within four pineapple duplicated regions, as expected from the two WGDs (Fig. 3C; more examples are presented in Supplementary Fig. 6).

An extensive level of synteny conservation is found between pineapple and grass genomes with some large blocks containing over 300 gene pairs (Supplementary Table 12). Rice *vs.* pineapple genome alignments show predominantly 4:2 patterns of syntenic depth (Supplementary Fig. 4), leading to an initial explanation that rice had two WGDs while pineapple had one since diverging from their common ancestor. However, further in-depth microsynteny analyses (Fig. 3C; more examples in Supplementary Fig. 6) show that each pineapple region has up to two highly syntenic rice regions, suggesting that the 4:2 pattern in the rice *vs.* pineapple comparison is best explained by a shared duplication σ, followed by one independent WGD (ρ) in rice, thus reducing the 4:2 syntenic depth ratio to a simpler 2:1 ratio. Higher degrees of microsynteny were observed between rice-pineapple orthologs than rice-pineapple out-paralogs (Supplementary Fig. 5). In addition, the 2:1 syntenic comparisons matched the expected patterns of fractionated gene content in rice following an independent WGD in its lineage [21]. Similar conclusions were found when pineapple was compared to other grass genomes such as sorghum. In addition, retained duplicate genes identified in syntenic blocks within the pineapple genome were sorted into gene families and the timing of duplication events relative to speciation events were inferred through analyses of gene family phylogenies (Supplementary Fig. 8). Taken together, the gene trees and all grass-pineapple syntenic block relationships suggest that the most recent WGD evident in the pineapple genome is σ, an event shared with all members of Poales including the grasses (Fig. 3).

The grass–pineapple genome comparisons have refined previously published time brackets for both the pan–cereal ρ event and the shared σ event [14,17]. The ρ duplication is inferred to have occurred before radiation of lineages leading to rice, wheat and maize, but after the divergence of lineages leading to the grasses and pineapple within the Poales 95–115 MYA[7,8]. The earlier WGD, σ, occurred after the lineage leading to Poales diverged from lineages leading to banana and the palms 100–120 MYA [8,19]. Pineapple represents the closest sequenced lineage to the grasses lacking the pan-grass WGD event ρ, which makes it an excellent outgroup for comparative grass genomic studies (Fig. 3).

Pineapple as a reference genome for monocot comparative genomics

Genome comparisons of pineapple with other non-cereal monocot clades unambiguously identify previously elusive lineage-specific WGD events. Synteny and phylogenomic analyses of banana, palm and grass genomes had indicated the existence of shared and lineage-specific WGD events [8,17,19]. However, precision in dating these events has been limited by sparse sampling of non-cereal monocot genomes.

Genome comparisons to non-cereal genomes using pineapple have much simpler synteny patterns than those using cereals, facilitating easier interpretation. Oil palm had one round of independent WGD, giving rise to mostly 2:2 syntenic depth in comparison with pineapple. While banana had three independent WGDs in its lineage, giving rise to intricate patterns of mostly 8:2 syntenic depth patterns compared to pineapple (Supplementary Fig. 8), our reconstructions of Zingiberales events were considerably less complicated than previous grass-banana comparisons [14,19]. Comparisons of pineapple to orchid in the Asparagales lineage were less definitive, perhaps due to the relatively limited contiguity in the current orchid genome assembly [22]. However, our phylogenomic analyses including genes from the orchid, *Phalaenopsis equestris*, and gene sequences from transcriptome data for agave and garden asparagus, also Asparagales, indicate that an earlier WGD event, τ, occurred in a common Asparagales-commelinids ancestor, the latter including the Poales, Arecales and Zingiberales (Fig. 2A).

Synteny between duckweed (*Spirodela polyrhiza*) and pineapple together with phylogenomic analyses narrowed estimates of the timing of the τ WGD. The duckweed genome in the Alismatales represents one of the earliest diverging monocots [18]. Duckweed-pineapple comparison showed 4:4 syntenic depth, consistent with two known Alismatales-specific WGDs [18], while also confirming independence of the two pineapple WGDs (σ and τ: Fig. 2). This inference was further supported in gene tree analyses (Supplementary Fig. 10). Consequently, we placed τ after the Alismatales-commelinids divergence but before the Asparagales-commelinid divergence (Fig. 2), implying a date between 135-110MYA [8]).

The pineapple genome enables the study of lineage-specific gene family mobility in grasses

Arabidopsis genes have moved around the genome over recent evolutionary time[23], inserting into new places probably by some form of translocation or recombination[24]. To distinguish between gene *insertion* in a query genome versus gene *deletion* in an outgroup, at least two outgroups are required for a confident inference[24]. While Brassicales gene movements have been studied[25], the analysis of mobile genes in grasses has lacked closely-related non-grass genomes, a need now fulfilled by pineapple.

Using pineapple and rice as outgroups, we tested whether the same gene families inferred to be mobile in *Arabidopsis thaliana* (At) (using a papaya outgroup) were also mobile in *Sorghum bicolor* (Sb; using a pineapple outgroup). The most mobile, larger gene families in *Arabidopsis* are F-box genes, MADS-box genes, defensins, and NBS-LRR genes[25]. We queried the *Arabidopsis thaliana* genome using *Arabidopsis lyrata*, peach, and grape as outgroups to determine mobility of genes in *A. thaliana*. We used the same methods to query sorghum against rice and pineapple to determine gene mobility. Our test was whether the number of mobile genes in a family was significantly higher than the number of nonmobile, i.e. syntenic, genes; if so, a gene family was determined to be mobile. We found that the gene families that tend to be mobile in *Arabidopsis* also tend to be mobile in sorghum (Supplementary Table 13), with a few exceptions. The MADS-box genes, while mobile in the *Arabidopsis* lineage, were not mobile in *Sorghum* lineage.

Evolutionarily, plant MADS-box genes are divided into type I and type II based on their specified protein sequences. In general, type II proteins are composed of the most conserved MADS domain for DNA binding, the keratin domain for protein-protein interaction, the intervening domain located between the M and K domains, and the C-terminal domain that is mainly responsible for transcription activation [26]. Unlike type II MADS-box proteins, the structure of type I proteins is simpler because it lacks the K domain. In plants, type I MADS-box genes experienced a faster pace of birth-and-death than type II genes due in part to a higher frequency of gene duplications [27]. Careful examination determined that the Type II MADS-box genes tend to be syntenous in both *Arabidopsis* and sorghum when compared to their respective outgroups (Supplementary Table 13). The more rapidly evolving Type I MADS-box genes tend to be mobile, but there are fewer of these in sorghum, suggesting either loss in the grasses or expansion in *Arabidopsis*. Recent studies indicate that the latter scenario may be the case, because MADS-box genes in the *Arabidopsis* ancestral lineage underwent a burst of mobility ~10 million years ago [25].

Conversely, the GDSL-like lipase/acylhydrolase gene family was not mobile in the Brassicales (*Arabidopsis* lineage), but was mobile in the Poales (*Sorghum* lineage) (Supplementary Table 13). The GDSL esterases/lipases are mainly involved in the regulation of plant development, morphogenesis, synthesis of secondary metabolites, and defense response. This gene family has expanded in the monocot lineage in comparison to eudicots [28]. Our data suggest that much of the GDSL expansion was via gene mobility, and likely has a role specific to grasses. These results demonstrated that pineapple is a useful and, at present, unique outgroup to the grass genomes for evolutionary inference.

Evolution of CAM photosynthesis

Drought is responsible for the majority of global crop loss, so understanding the mechanisms that plants have evolved to survive water stress is vital for engineering drought tolerance in crop species. Plants such as pineapple that use CAM thrive in water-limited environments, potentially achieving greater net CO_2 uptake than their C_3 and C_4 counterparts [29]. By using an alternate carbon assimilation pathway that allows CO_2 to be fixed nocturnally by PEPC and stored transiently as malic acid in the vacuole (Fig. 4B), CAM plants can keep their stomata closed during the daytime while the stored malic acid is decarboxylated and the released CO_2 is refixed through the Calvin-Benson cycle, greatly reducing water loss in evapotranspiration[30]. High water use efficiency and drought tolerance thus make CAM an attractive pathway for engineering crop plants for climate change [31]. The core CAM enzymic steps are well characterized and share similarities with C_4 plants [32], but the regulatory elements of CAM and connections to the circadian clock are largely unknown [33]. CAM photosynthesis is a recurrent adaptation with numerous independent origins across 35 diverse families of vascular plants[34].

We identified genes in the CAM pathway based on homology to C_3/C_4 orthologs in maize, sorghum, and rice. The pineapple genome contains 38 putative genes involved in the carbon fixation module of CAM including the key enzymes carbonic anhydrase (*CA*), phospho*enol*pyruvate carboxylase (*PEPC*), phospho*enol*pyruvate carboxylase kinase (*PPCK*), NAD- and NADP-linked malic enzymes (*ME*), malate dehydrogenase (*MDH*), phospho*enol*pyruvate carboxykinase (*PEPCK*), and pyruvate, orthophosphate dikinase (*PPDK*) (Supplementary Tables 14 and 15). As well as using PEPCK (rather than ME) as its principal decarboxylating enzyme during the daytime [35], pineapple is also distinctive among CAM plants in showing high activities of the alternative glycolytic enzyme PP_i-dependent phosphofructokinase (pyrophosphate:fructose-6-phosphate 1-phosphotransferase) [36,37] and in possessing vacuolar transporters for soluble sugars [38,39], which form the main pool of transitory carbohydrate supplying PEP for nocturnal CO_2 fixation and malic acid synthesis [40,41] (Fig. 4B). Notably, in terms of gene number, pineapple contains fewer of these core metabolic genes compared with other monocots.

To investigate the diel expression patterns of CAM, we collected RNA-seq samples at 2-hour intervals over a 24-hour period from photosynthetic (green tip) and non-photosynthetic (white base) leaf tissue of field grown pineapple (Fig. 4A). Based on contrasting expression patterns between the two tissues, we were able to distinguish the gene family members involved in carbon fixation from the non-CAM related members involved in other processes. Nine genes (*PEPC*, *PPCK*, *PEPCK*, *PPDK*, three copies of *CA* and two *MDH*) have a diurnal expression pattern in the green tissue with low or no expression in the white leaf tissue (Fig. 4C). CAM photosynthesis is divided into four temporal phases that should be largely controlled by the circadian clock. Genes under circadian-clock control were enriched with *cis*-regulatory elements including the morning (CCACAC) and evening element (AAAATATC) [42]. The diurnal expressed photosynthetic genes were enriched ($p = 0.002$) with known circadian clock *cis*-elements compared to the non-photosynthetic gene copies (Fig. 4C), suggesting that the carbon fixation pathway in pineapple is regulated by circadian-clock genes through *cis*- regulatory elements.

Carbonic anhydrase (CA), by catalyzing the conversion of CO_2 into bicarbonate, is responsible for the first step in CO_2 fixation in C_4 and CAM photosynthesis. Of the three carbonic anhydrase families (α, β, and γ) in pineapple, only β*CA* showed a nighttime and early morning expression profile in green tissue. This suggests pineapple uses βCA as the major protein for carbon fixation, which is consistent with the finding in C_4 species in the genus *Flaveria* [43]. Promoter regions of all three β*CA* genes contain a CCA1 binding site that can bind both circadian core oscillators, *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*) and *LATE ELONGATED HYPOCOTYL* (*LHY*) products. Among all β*CA* genes in orchid, rice, maize and sorghum, only one β*CA* gene (Sobic.003G234500) in sorghum contains a CCA1 binding site (Supplementary Table 16) at its promoter and this gene has no known photosynthetic function [44], indicating that β*CA* in pineapple is temporally regulated by the circadian clock to synchronize the expression of its gene product with stomatal opening at night for maximum CO_2 fixation in pineapple.

Although the core CAM pathway genes are well-characterized, little is known about the regulatory networks controlling the temporal phases of CAM. We constructed gene interaction networks comparing the diurnal expression patterns in green and white leaf cells to discriminate CAM-related genes from genes with a general circadian oscillation (see Methods). Two clusters in the networks (clusters 1 and 16) have an enrichment of CAM-related genes including *CA*, *PEPC*, *PPCK*, *NAD-ME*, *MDH* and *PPDK* (Supplementary Fig. 9). A metabolic pathway enrichment analysis of these two clusters suggests they have different biological functions. Cluster 1 is enriched in cellular development pathways such as amino sugar and nucleotide sugar metabolism, ascorbate and aldarate metabolism, and glycerophospholipid metabolism. Cluster 16 is enriched in genes involved in downstream processes associated with carbon fixation, including the citric acid cycle, oxidative phosphorylation, and starch and sucrose metabolism (Supplementary Table 17).

Interestingly, Cluster 1 also contains a significant number of core circadian-clock genes, including CCA1/LHY, GIGANTEA, PSEUDO- RESPONSES REGULATOR 7, and PSEUDO-REPRESONSES REGULATOR 9 (Supplementary Table 18). Furthermore, a promoter enrichment analysis showed that Cluster 1 genes are enriched with circadian related *cis*-acting elements including the G-box and evening motifs, and CCA1 binding sites (Supplementary Table 19). Our network analyses showed that one of the CAM co-expression modules is closely interacting with the circadian-clock pathway, providing empirical evidence connecting CAM with the circadian clock.

We identified putative regulators of CAM by surveying gene-interaction networks. CAM genes are highly connected in the gene interaction network (Figs. 3D and Supplementary Fig. 10). CAM genes have dramatic differences in their regulatory patterns based on gene interactions (Supplementary Table 19). From the network, the increase in expression of β*CA* in the green cells is mainly contributed by the appearance of about 243 potential activators and also disappearance of 2 potential repressors. *PPCK* showed similar regulatory patterns although the number of repression controllers identified was much higher than for βCA. In contrast, the increased expression of *PEPC* was mainly related to the release of repression from potential repression-controllers (35) and relatively less by appearance of potential activators (1). Three isoforms of *MDH* (Aco006122.1, Aco010232.1, and Aco004996.1) showed similar regulatory patterns. Among the identified CAM-related genes, the expression of *NAD-ME2*, *NAD-ME4*, *NAD-MDH*, *PPCRK1*, and *PPCRK 3* showed decreased expression in green tissues compared to that in white tissues. The decreased expression was mostly due to the disappearance of the activation controller together with the appearance of repressors. In summary, different enzymes involved in CAM photosynthesis used different regulatory mechanisms, as reflected in both the interaction partners and also their regulatory patterns, to achieve the position-specific expression patterns (Supplementary Table 20). This result provides strong molecular evidence as to how those regulatory mechanisms controlling the expression of CAM-related genes could have evolved “independently” so often: the capacity was always present, but repressed at the *trans*-acting, cell-specific, and individual gene level. This finding is consistent with the notion that the CAM and/or C_4 photosynthesis evolved as a result of a re-organization of an ancestral metabolic pathway 45. These different features later were assembled to form the functional CAM photosynthesis. The identified candidate genes provide initial targets for detailed functional studies of how the CAM genes have evolved the regulation necessary to gain the observed spatial and temporal expression patterns, but loss of repressors is certainly involved.

We identified CAM-specific genes by comparing genomes of pineapple and the CAM orchid *Phalaenopsis equestris* against genomes of the C_4 grasses sorghum 46 and Setaria [47], and the C_3 grasses Brachypodium [48] and rice [49]. The 198,446 genes in the six genomes were clustered into 23,964 ortholog groups, of which 409 groups (1,295 genes) are shared by the two CAM species, but are absent in C_3 and C_4 species (Supplementary Fig. 11), and are considered to be CAM-specific in this study. Based on a pairwise t-test ($p < 0.05$), 109 orthologous groups were expanded in CAM species relative to C_3 and C_4 species; and five orthologous groups were expanded in both CAM and C_4 species relative to C_3 species. The orthologous groups expanded in CAM species contain 236 pineapple genes. The orthologous groups expanded in both CAM and C_4 species contain 10 pineapple but no

Phalaenopsis genes. There are 568 CAM-specific pineapple genes, among which 306 genes were supported by the time-course RNA-Seq data obtained in the green tip of mature leaves with FPKM value ≥ 5 in at least one of the 13 time-points. A majority of these genes were found to be either transcription regulators such as the Pentatricopeptide repeat (PPR) and tetratricopeptide repeat (TRP) family, F-box and U-box family proteins, or post-transcription regulators such as kinases, ATP/GTP-binding proteins, oxydoreductased, and heat-shock proteins. Some of them are involved in ligand or metal transfer (Supplementary Table 21). Seven of the 27 C_4 CAM-shared pineapple genes (after removing hypothetical proteins and including only those supported by RNA-seq) are ripening-related proteins, together with transcription or post-transcription regulators (Supplementary Table 22). There are 22 C_3 CAM-shared pineapple genes comprising the cytochrome P450 proteins, lipid-transfer proteins and other proteins that may be involved in signaling processes (Supplementary Table 23).

Based on the diel expression pattern, four important gene categories were identified in these CAM-specific genes in pineapple: (1) night-peaking genes showing relatively higher expression during nighttime (Fig. 4E, 75 genes); (2) day-peaking genes showing relatively higher expression during daytime (Fig. 3F, 177 genes); (3) morning-peaking genes that peak in expression near dawn (Fig. 4G, 43 genes); and (4) evening-peaking genes that peak in expression near dusk (Fig. 4H, 11 genes). In addition, 16 orthologous groups are shared by the CAM and C_4 species, but are absent in the C_3 lineages, and are considered to be CAM- and C_4-specific in this study. These CAM- and C_4-specific groups contain 29 pineapple genes, of which 10 pineapple genes were supported by the time-course RNA-Seq data.

Discussion

Pineapple is self-incompatible and all pre-Columbian and most post-Columbian varieties were selected from somatic mutations compared to the extensive breeding history of most crops. Sequencing the genomes of two leading commercial varieties 'F153' and 'MD2' revealed heterozygosity within each genome at about 2%, much higher than seed propagated crops but similar to clonally propagated crops. Self-incompatibility combined with clonal propagation contributes to and maintains the high level of heterozygosity in pineapple. Inbreeding depression from a self-compatible pineapple mutant was so severe that most seedlings died after two generations of self-pollination [50]. The high frequency of non-synonymous SNPs in 'F153' and 'MD2', respectively, may be the cause of such unusually severe inbreeding depression (Supplementary Table 11). The abundance of retrotransposons, such as the Pusofa family (28% of LTR retrotransposons and 15% of the pineapple genome) might have contributed to genome instability in pineapple. Any search for somatic mutations caused by LTR retrotransposons, including those potentially associated with pineapple cultivar improvement, would be best focused on those families that are most highly expressed.

The modified carbon assimilation pathways of CAM and C_4 photosynthesis result in higher water use efficiency (WUE), a highly desirable trait given the need to double food production by 2050 under a changing climate. CAM and C_4 photosynthesis use many of the same enzymes for concentrating CO_2 but differ in spatial (C_4) versus temporal separation of carbon fixation [35]. Understanding the evolution of CAM and C_4 photosynthesis may expedite projects to convert C_3 rice to C_4 [51] and C_3 poplar to CAM[31]. CAM plants have higher WUE than C_3 and C_4 plants and may be better suited for engineering crop drought tolerance. All plants contain the necessary genes for CAM photosynthesis, and the evolution of CAM simply requires rerouting of preexisting pathways. Pineapple has a lower number of CAM related genes compared to other monocots but detailed tissue-specific and diel gene expression profiles identified the candidate gene family members recruited for CAM. CAM pathway genes are enriched with circadian clock associated *cis*-regulatory elements, providing the first link between CAM and the circadian clock. Consistent with this, β*CA* genes in pineapple contain a CCA1 binding site which is absent in C_3 and C_4 monocots. Regulation of CAM is complex and CAM related enzymes use different regulatory mechanisms explaining how CAM evolved independently many times during evolution: the gene content encoding the enzymatic machinery is present, but diel expression patterns are likely silenced or not activated sufficiently at the *cis*-acting, cell-specific, individual gene level. This work provides the first detailed analysis of the expression and regulation patterns associate with CAM which could ultimately be used for engineering better WUE and drought tolerance in crop plants.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Ray Ming[1,2,†,*], Robert VanBuren[1,2,3,*], Ching Man Wai[1,2,*], Haibao Tang[1,4,*], Michael C. Schatz[5], John E. Bowers[6], Eric Lyons[4], Ming-Li Wang[7], Jung Chen[8], Eric Biggers[5], Jisen Zhang[1], Lixian Huang[1], Lingmao Zhang[1], Wenjing Miao[1], Jian Zhang[1], Zhangyao Ye[1], Chenyong Miao[1], Zhicong Lin[1], Hao Wang[6], Hongye Zhou[6], Won C. Yim[9], Henry D. Priest[3], Chunfang Zheng[10], Margaret Woodhouse[11], Patrick P. Edger[11], Romain Guyot[12], Hao-Bo Guo[13], Hong Guo[13], Guangyong Zheng[14], Ratnesh Singh[15], Anupma Sharma[15], Xiangjia Min[16], Yun Zheng[17], Hayan Lee[5], James Gurtowski[5], Fritz J. Sedlazeck[5], Alex Harkess[6], Michael R. McKain[3], Zhenyang Liao[1], Jingping Fang[1], Juan Liu[1], Xiaodan Zhang[1], Qing Zhang[1], Weichang Hu[1], Yuan Qin[1], Kai Wang[1], Li-Yu Chen[1], Neil Shirley[18], Yann-Rong Lin[19], Li-Yu Liu[19], Alvaro G. Hernandez[20], Chris L. Wright[20], Vincent Bulone[18], Gerald A. Tuskan[21], Katy Heath[2], Francis Zee[22], Paul H. Moore[7], Ramanjulu Sunkar[23], James H. Leebens-Mack[6], Todd Mockler[3], Jeffrey L. Bennetzen[6], Michael Freeling[11], David Sankoff[10], Andrew H. Paterson[24], Xinguang Zhu[14], Xiaohan Yang[21], J. Andrew C. Smith[25], John C. Cushman[9], Robert E. Paull[8,†], and Qingyi Yu[15,†]

Affiliations

[1]FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou, Fujian, 350002, China

[2]Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[3]Donald Danforth Plant Science Center, St. Louis, MO, USA

[4]iPlant Collaborative/University of Arizona, Tuscon, AZ 85719, USA

[5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY11724, USA

[6]Department of Plant Biology, University of Georgia, Athens, GA 30602, USA

[7]Hawaii Agriculture Research Center, Kunia, HI 96759, USA

[8]Department of Tropical Plant and Soil Sciences, University of Hawaii, Honolulu, HI 96822, USA

[9]Department of Biochemistry and Molecular Biology, MS330, University of Nevada, Reno, NV 89557-0330, USA

[10]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada K1N 6N5

[11]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA

[12]IRD, UMR DIADE, EVODYN, BP 64501, 34394 Montpellier Cedex 5, France

[13]Department of Biochemistry & Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996, USA

[14]Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

[15]Texas A&M AgriLife Research, Department of Plant Pathology & Microbiology, Texas A&M University System, Dallas, TX 75252, USA

[16]Department of Biological Sciences, Youngstown State University, Youngstown, OH 44555, USA

[17]Faculty of Life Science and Technology, Kunming University of Science and Technology, Kunming, Yunnan 650500, China

[18]ARC Centre of Excellence in Plant Cell Walls, School of Agriculture, Food and Wine, University of Adelaide, Waite Campus Urrbrae, South Australia 5064, Australia

[19]Department of Agronomy, National Taiwan University, Taipei 10617, Taiwan

[20]W.M. Keck Center, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[21]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

[22]USDA-ARS, Pacific Basin Agricultural Research Center, Hilo, HI 96720, USA

[23]Department of Biochemistry and Molecular Biology, 246 Noble Research Center, Oklahoma State University, Stillwater, OK 74078, USA

[24]Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602, USA

[25]Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK

Acknowledgments

We thank Russell Kai and Carol Mayo Riley for maintaining the pineapple plants and collection of leaf tissues; Mr. Mike Conway at Dole Plantation in for assistance on time course leaf sample collection; Garth Sanewski for providing 'MD2' pedigree. This project is supported by the Fujian Agriculture and Forestry University to RM; a USDA T- START grant thought The University of Hawaii to QY, RM, PHM, and REP; and the University of Illinois at Urbana-Champaign to RM. HT is supported by the "100 Talent Plan" award by the Fujian provincial government. Analyses of the pineapple genome are supported by the following sources: National Science Foundation (NSF) Plant Genome Program Grant # 0922545 to RM, PM, QY; NSF IOS 1444567 to JHL; The US National Institutes of Health (award R01-HG006677) and the US National Science Foundation (awards DBI-1350041 and DBI-1265383) to MCS. WCY, H-BG, HG, GAT, XY, and JCC acknowledge support from the Department of Energy, Office of Science, Genomic Science Program under Award Number DE-SC0008834.

References

1. Clement CR, de Cristo-Araújo M, Coppens d'Eeckenbrugge G, Pereira AA, Picanço-Rodrigues D. Origin and domestication of native Amazonian crops. Diversity. 2010; 2:72–106.
2. Bartholomew, DP.; Paull, RE.; Rohrbach, KG., editors. The Pineapple: Botany, Production and Uses. CAB International; 2003.
3. Beauman, F. The Pineapple: King of Fruits. Chatto & Windus; 2005.
4. FAOSTAT. Food and Agriculture Organization of the United Nations, Statistics Division. FAO; 2015.
5. Yang X, et al. A roadmap for research on crassulacean acid metabolism (CAM) to enhance sustainable food and energy production in a hotter, drier world. New Phytol. 2015; 207:491–504. [PubMed: 26153373]
6. Brewbaker JL, Gorrez DD. Genetics of self-incompatibility in the monocot genera, *Ananas* (pineapple) and *Gasteria*. Am J Bot. 1967:611–616.
7. Givnish TJ, et al. Adaptive radiation, correlated and contingent evolution, and net species diversification in Bromeliaceae. Mol Phylogenet Evol. 2014; 71:55–78. [PubMed: 24513576]
8. Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol. 2015; 207:437–453. [PubMed: 25615647]
9. Arumuganathan K, Earle E. Nuclear DNA content of some important plant species. Plant Mol Biol Rep. 1991; 9:208–218.
10. Cantarel BL, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008; 18:188–196. [PubMed: 18025269]
11. McCarthy EM, McDonald JF. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics. 2003; 19:362–367. [PubMed: 12584121]
12. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007; 35:W265–W268. [PubMed: 17485477]
13. Meyers BC, Tingey SV, Morgante M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. 2001; 11:1660–1676. [PubMed: 11591643]
14. Tang H, Bowers JE, Wang X, Paterson AH. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci USA. 2010; 107:472–477. [PubMed: 19966307]

15. Paterson AH, Bowers JE, Chapman BA. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA. 2004; 101:9903–9908. [PubMed: 15161969]

16. Jaillon O, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007; 449:463–467. [PubMed: 17721507]

17. Jiao Y, Li J, Tang H, Paterson AH. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. Plant Cell. 2014; 26:2792–2802. [PubMed: 25082857]

18. Wang W, et al. The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. Nat Commun. 2014; 5:3311. [PubMed: 24548928]

19. D'Hont A, et al. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature. 2012; 488:213–217. [PubMed: 22801500]

20. *Amborella* Genome Project The *Amborella* genome and the evolution of flowering plants. Science. 2013; 342:1241089. [PubMed: 24357323]

21. Lyons E, et al. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. Plant Physiol. 2008; 148:1772–1781. [PubMed: 18952863]

22. Cai J, et al. The genome sequence of the orchid *Phalaenopsis equestris*. Nat Genet. 2015; 47:65–72. [PubMed: 25420146]

23. Freeling M, et al. Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. Genome Res. 2008; 18:1924–1937. [PubMed: 18836034]

24. Woodhouse MR, Pedersen B, Freeling M. Transposed genes in Arabidopsis are often associated with flanking repeats. PLoS Genet. 2010; 6:e1000949. [PubMed: 20485521]

25. Woodhouse MR, Tang H, Freeling M. Different gene families in *Arabidopsis thaliana* transposed in different epochs and at different frequencies throughout the rosids. Plant Cell. 2011; 23:4241–4253. [PubMed: 22180627]

26. Kramer EM, Dorit RL, Irish VF. Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-box gene lineages. Genetics. 1998; 149:765–783. [PubMed: 9611190]

27. Nam J, et al. Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc Natl Acad Sci USA. 2004; 101:1910–1915. [PubMed: 14764899]

28. Chepyshko H, Lai CP, Huang LM, Liu JH, Shaw JF. Multifunctionality and diversity of GDSL esterase/lipase gene family in rice (*Oryza sativa* L. japonica) genome: new insights from bioinformatics analysis. BMC Genomics. 2012; 13:309. [PubMed: 22793791]

29. Nobel PS. Achievable productivities of certain CAM plants: basis for high values compared with C_3 and C_4 plants. New Phytol. 1991; 119:183–205.

30. Osmond CB. Crassulacean acid metabolism: a curiosity in context. Annu Rev Plant Physiol. 1978; 29:379–414.

31. Borland AM, et al. Engineering crassulacean acid metabolism to improve water-use efficiency. Trends in Plant Sci. 2014; 19:327–338. [PubMed: 24559590]

32. Christin PA, et al. Shared origins of a key enzyme during the evolution of C_4 and CAM metabolism. J Exp Bot. 2014; 65:3609–3621. [PubMed: 24638902]

33. Edwards EJ, Ogburn RM. Angiosperm responses to a low-CO_2 world: CAM and C_4 photosynthesis as parallel evolutionary trajectories. Int J Plant Sci. 2012; 173:724–733.

34. Silvera K, et al. Evolution along the crassulacean acid metabolism continuum. Funct Plant Biol. 2010; 37:995–1010.

35. Dittrich P, Campbell WH, Black CC Jr. Phosphoenolpyruvate carboxykinase in plants exhibiting crassulacean acid metabolism. Plant Physiol. 1973; 52:357–361. [PubMed: 16658562]

36. Carnal NW, Black CC. Pyrophosphate-dependent 6-phosphofructokinase, a new glycolytic enzyme in pineapple leaves. Biochem Biophys Res Comm. 1979; 86:20–26. [PubMed: 219853]

37. Carnal NW, Black CC. Soluble sugars as the carbohydrate reserve for CAM in pineapple leaves. Implications for the role of pyrophosphate:6-phosphofructokinase in glycolysis. Plant Physiol. 1989; 90:91–100. [PubMed: 16666775]

38. McRae SR, Christopher JT, Smith JAC, Holtum JAM. Sucrose transport across the vacuolar membrane of *Ananas comosus*. Funct Plant Biol. 2002; 29:717–724.

39. Antony E, Taybi T, Courbot M, Mugford ST, Smith JAC, Borland AM. Cloning, localization and expression analysis of vacuolar sugar transporters in the CAM plant *Ananas comosus* (pineapple). J Exp Bot. 2008; 59:1895–1908. [PubMed: 18408220]

40. Kenyon WH, Severson RF, Black CC Jr. Maintenance carbon cycle in Crassulacean acid metabolism plant leaves Source and compartmentation of carbon for nocturnal malate synthesis. Plant Physiol. 1985; 77:183–189. [PubMed: 16664005]

41. Holtum JAM, Smith JAC, Neuhaus NE. Intracellular carbon transport and pathways of carbon flow in plants with crassulacean acid metabolism. Funct Plant Biol. 2005; 32:429–449.

42. Michael TP, et al. Network discovery pipeline elucidates conserved time-of-day–specific cis-regulatory modules. PLoS Genet. 2008; 4:e14. [PubMed: 18248097]

43. Ludwig M. Carbonic anhydrase and the molecular evolution of C_4 photosynthesis. Plant Cell Environ. 2012; 35:22–37. [PubMed: 21631531]

44. Wang X, et al. Comparative genomic analysis of C_4 photosynthetic pathway evolution in grasses. Genome Biol. 2009; 10:R68. [PubMed: 19549309]

45. West-Eberhard MJ, Smith JAC, Winter K. Photosynthesis, reorganized. Science. 2011; 332:311–312. [PubMed: 21493847]

46. Paterson AH, et al. The *Sorghum bicolor* genome and the diversification of grasses. Nature. 2009; 457:551–556. [PubMed: 19189423]

47. Bennetzen JL, et al. Reference genome sequence of the model plant Setaria. Nature Biotechnol. 2012; 30:555–561. [PubMed: 22580951]

48. Vogel JP, et al. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature. 2010; 463:763–768. [PubMed: 20148030]

49. Project IRGS. The map-based sequence of the rice genome. Nature. 2005; 436:793–800. [PubMed: 16100779]

50. Collins, JL. The pineapple. Leonard Hill; 1960.

51. von Caemmerer S, Quick WP, Furbank RT. The development of C_4 rice: current progress and future challenges. Science. 2012; 336:1671–1672. [PubMed: 22745421]

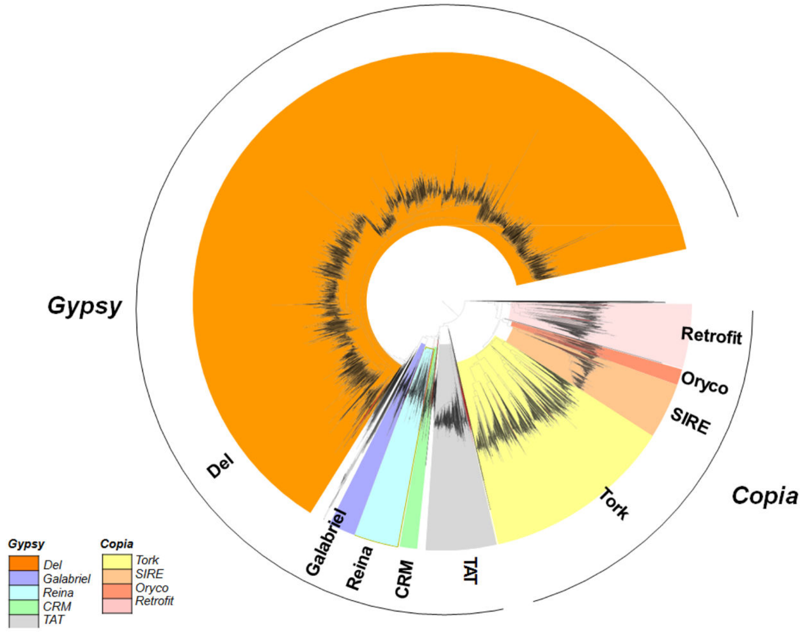

Fig. 1.

Phylogenetic analysis of Pineapple Reverse Transcriptase domains from LTR retrotransposons. Unrooted phylogenetic tree of *Gypsy* and *Copia* elements was constructed based on 6,379 aligned reverse transcriptase.

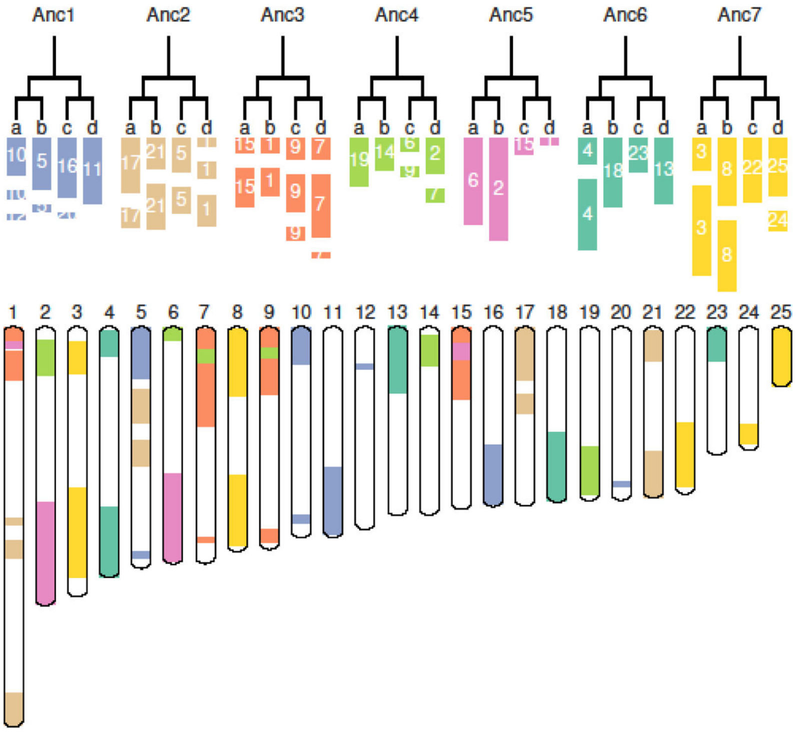

Fig. 2.

Twenty-five pineapple chromosomes organized into pairs of pairs following two whole genome duplication (WGD) events. Each color reflects one of the seven ancestral chromosomes. The left and right pairs represent the two subgenomes produced by WGD τ, and within each pair are the two subgenomes produced by WGD σ.

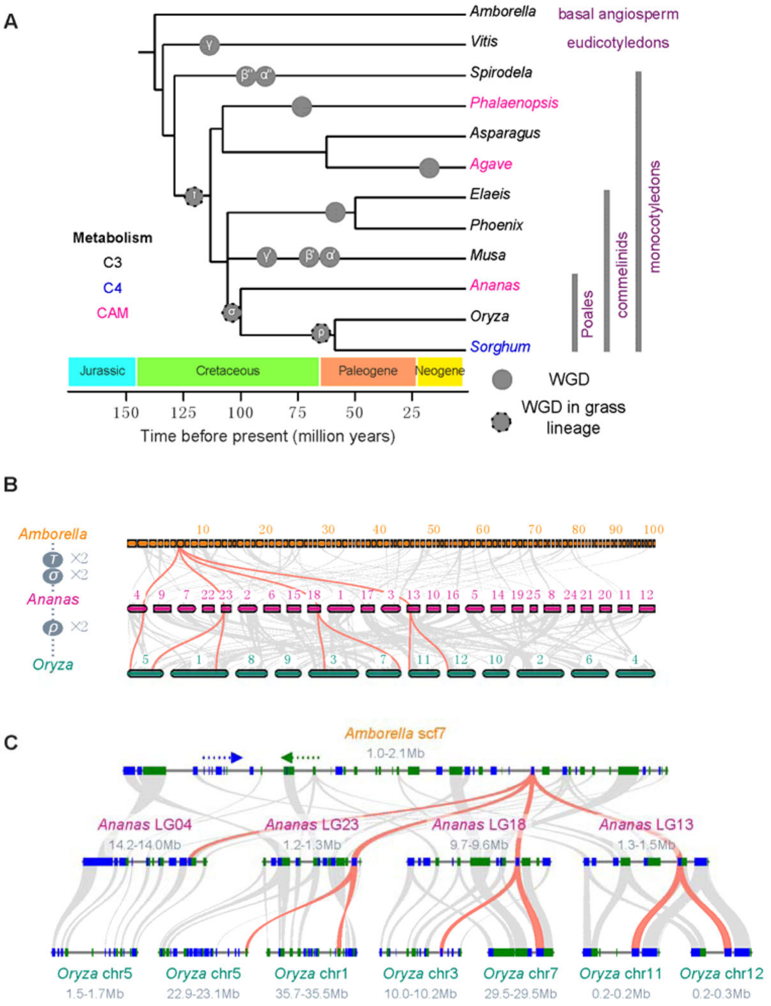

Fig. 3.

Genome evolution in pineapple. (A). Dating of whole genome duplication (WGD) events on the monocot tree of life. Circles represent known WGDs identified previously. The pineapple genome sequence clarified the dating of the three WGDs in the grass lineage – ρ, σ and τ. Taxon labels are colored according to their photosynthetic metabolism – C_3, C_4 or CAM. (B). Genomic alignments between *Amborella trichopoda*, *Ananas comosus* (pineapple) and *Oryza sativa* (rice), tracking gene positions through multiple species and copy numbers arising from multiple genome duplication events. Macro-synteny patterns show that a typical ancestral region in the basal angiosperm *Amborella* can be tracked to up to 4 regions in *Ananas comosus* due to the two genome duplication events σ and τ, and up to 8 regions in *Oryza sativa*. Grey wedges in the background highlight major synteny blocks spanning more than 30 genes between the genomes (illustrated by one syntenic set colored red). (C) Micro-colinearity patterns between genomic regions from *Amborella trichopoda*, *Ananas comosus* and *Oryza sativa*. Rectangles show predicted gene models with blue/green colors showing relative gene orientations. Grey wedges connect matching gene pairs, with one set highlighted in red.

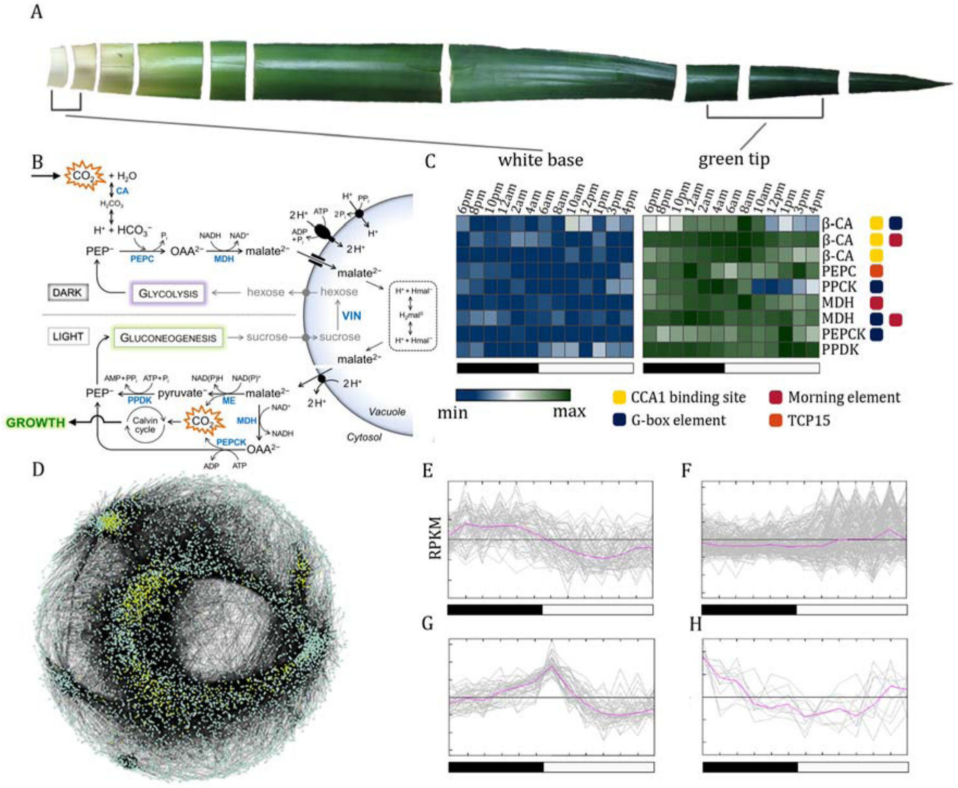

Fig. 4.

Evolution of the CAM pathway in pineapple. (A) Pineapple leaf tissue used to survey the diurnal expression patterns of CAM genes. The fully expanded D leaf of field grown pineapple is shown. Green (photosynthetic) tissue at the leaf tip and white (non-photosynthetic) tissue at the leaf base were collected to distinguish CAM-related gene expression from non-CAM-related circadian oscillation. (B) Overview of the carboxylation (top) and decarboxylation (bottom) pathways of CAM. CAM enzymes are shown in blue. (C) Expression pattern and *cis*-elements of pineapple carbon fixation genes across the diurnal expression data. Log2 transformed RPKM expression profiles are shown. Four known circadian-related binding motif sequences were searched in the 1kb upstream region for each gene. (D) Gene regulatory network of the green leaf tissue. Only the largest module of the network was kept. Genes related to CAM and their interaction partners were highlighted in yellow. (E–H) Co-expression clusters of the CAM-specific genes in pineapple. Cluster of genes with relative higher levels of nocturnal expression (E), diurnal expression (F), expression at dawn (6 am) (G), and expression at dusk (6 pm) (H).