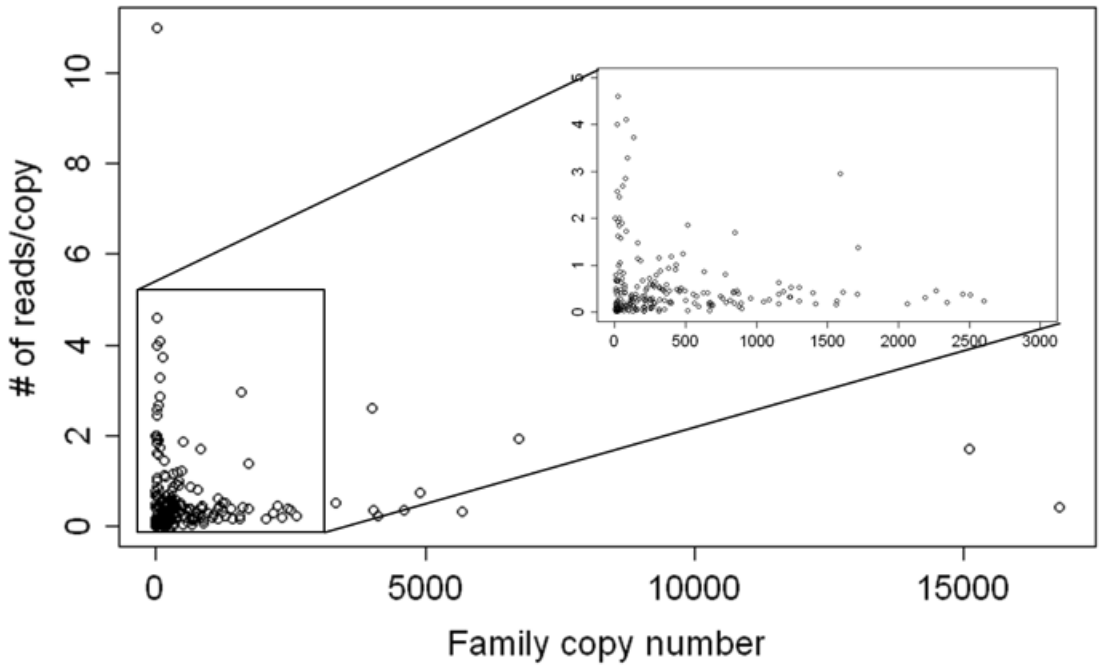
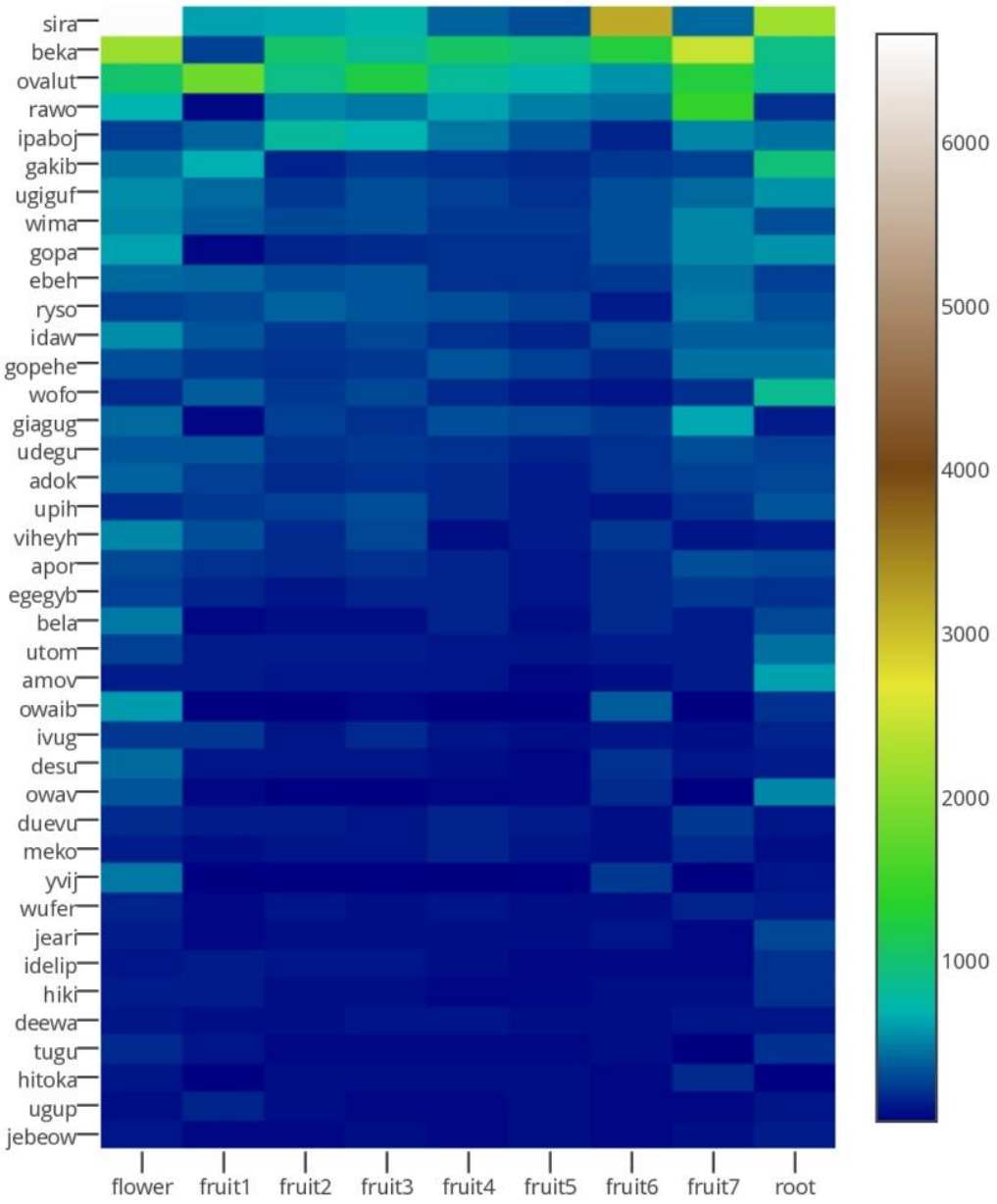


Supplementary Figure 1

Correlation between family copy number and expression level of LTR elements.

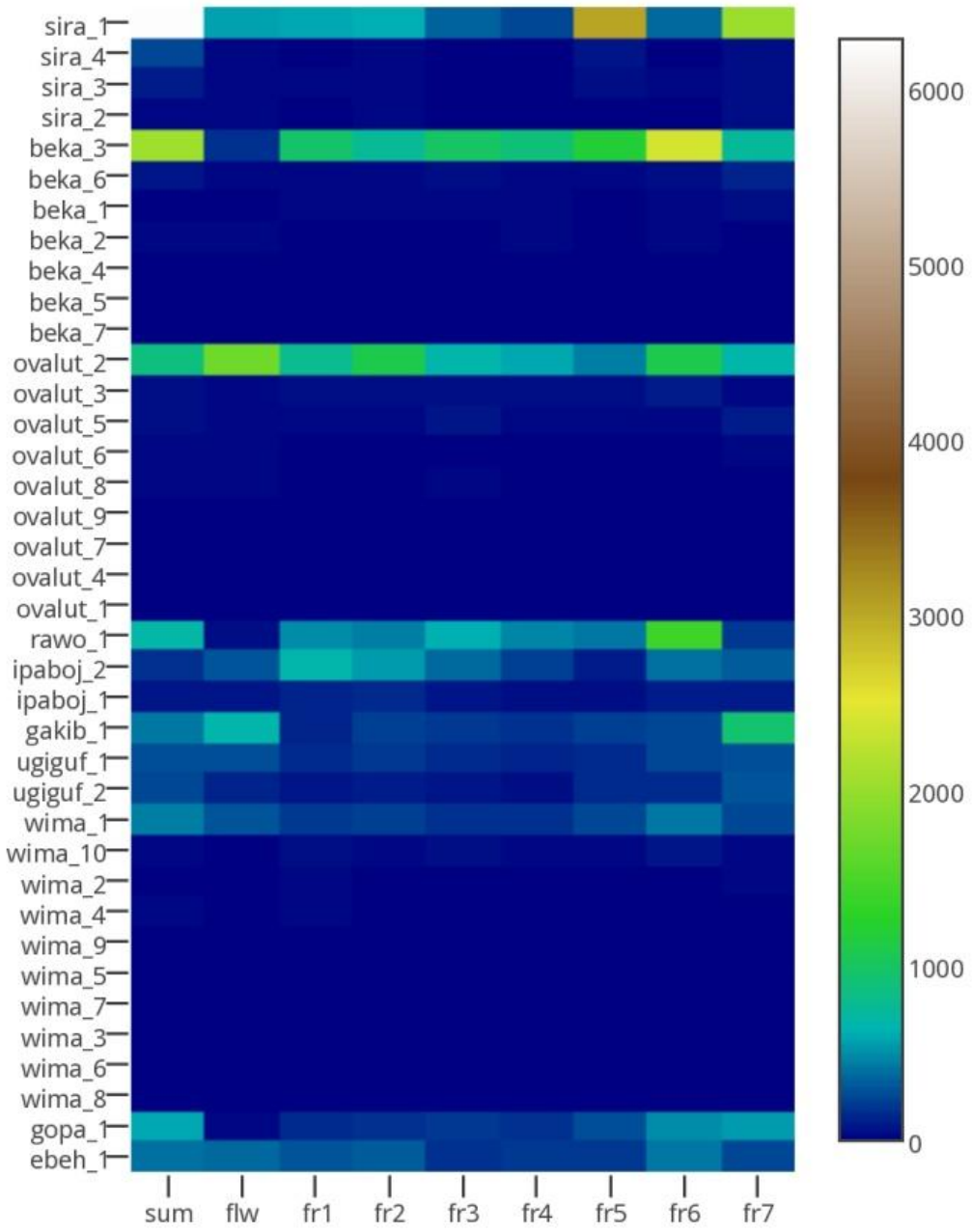
The data indicate that high expression levels of LTR elements are correlated with a relatively low copy number of their family.



Supplementary Figure 2

Expression of intact LTR retrotransposons in nine pineapple tissue samples.

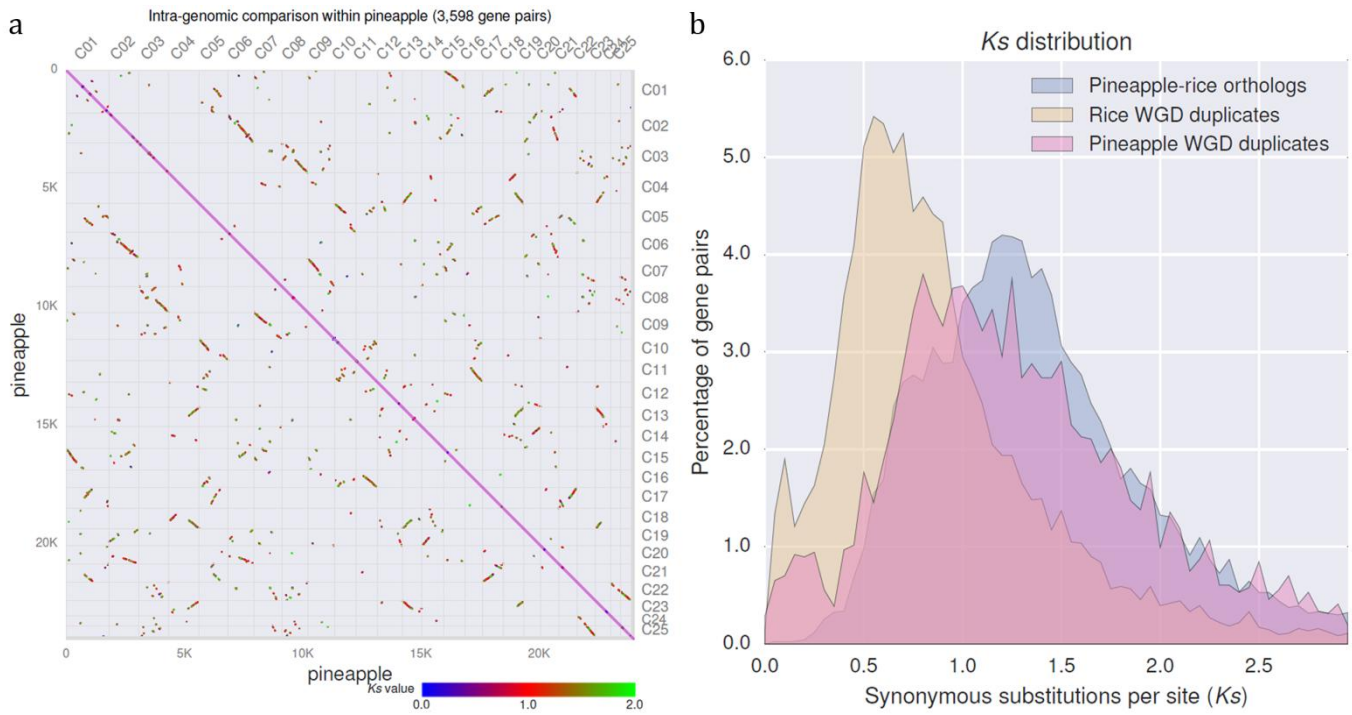
This heat map shows the number of RNA-seq reads mapped to the top 40 most highly expressed LTR retrotransposon families. Family names are shown as row labels, and tissue names are given as column labels. From top to bottom, the rows are sorted by total counts of mapped reads in families.



Supplementary Figure 3

Expression of subfamilies of LTR retrotransposons in nine pineapple tissue samples.

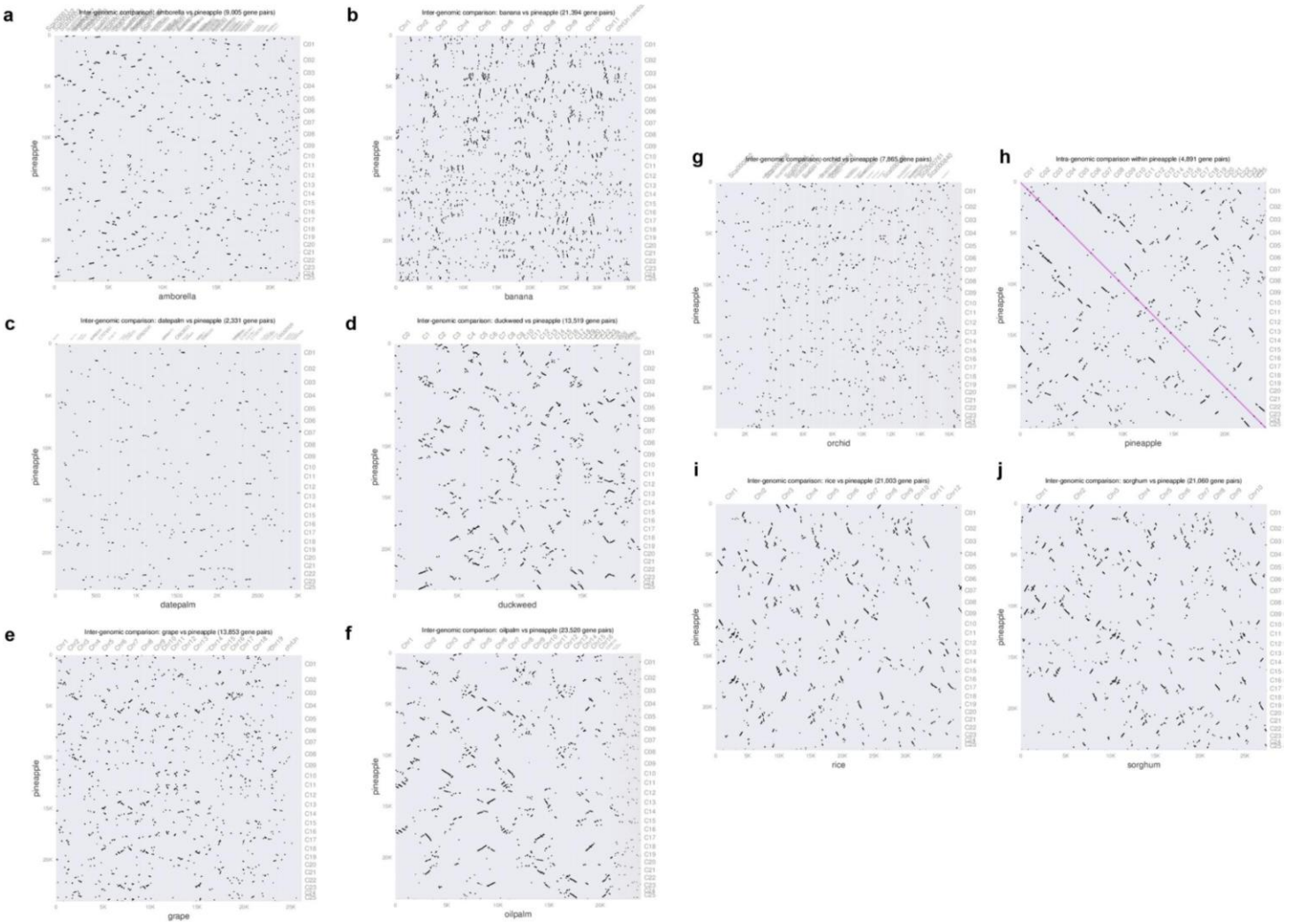
The heat map shows the number of RNA-seq reads mapped to the top ten most highly expressed LTR retrotransposon families. Each row represents a subfamily, and each column represents a tissue. The numbers following family names give subfamily IDs. From top to bottom, the rows are sorted by total counts of mapped reads in families. Within each family, the rows are further sorted by total counts of mapped reads in subfamilies.



Supplementary Figure 4

Synonymous substitutions per site (*Ks*) values between inferred whole-genome duplicates in pineapple.

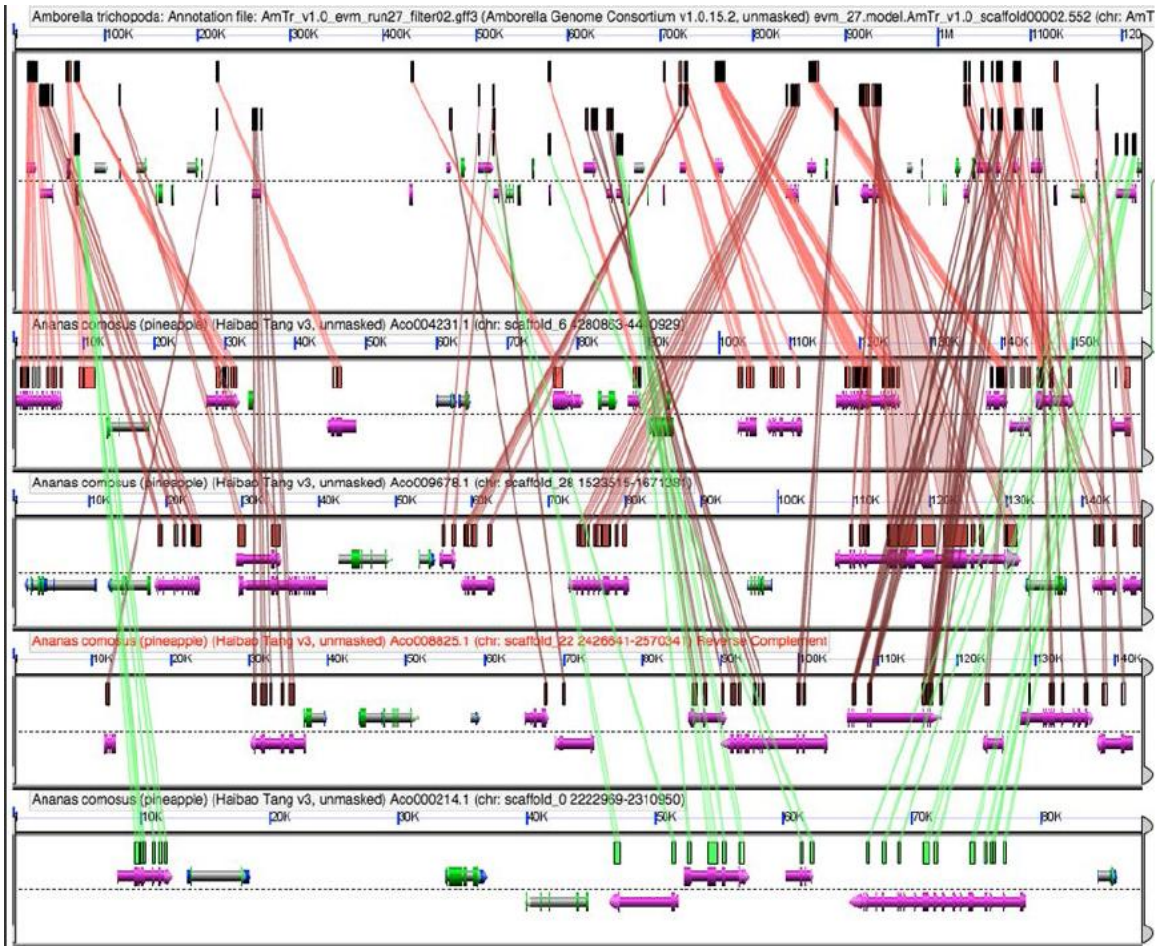
(**a**) Syntenic dot plot in pineapple versus pineapple comparison, with *Ks* values color coded; only the gene pairs with a *Ks* value between 0 and 2 are plotted. (**b**) Histogram of *Ks* values for pineapple-rice orthologs, rice whole-genome duplicates and pineapple whole-genome duplicates.



Supplementary Figure 5

Pairwise genome comparisons between pineapple and ten related plant species.

Pairwise comparisons (dot plots) between pineapple (*y* axis) and a total of ten related plant genomes (*x* axis), including (**a**–**j**) *Amborella*, banana, date palm, duckweed, grape, oil palm, orchid, pineapple (i.e., self-comparison), rice and sorghum. For clarity, only gene pairs within synteny blocks of at least size 4 are shown.



Supplementary Figure 6

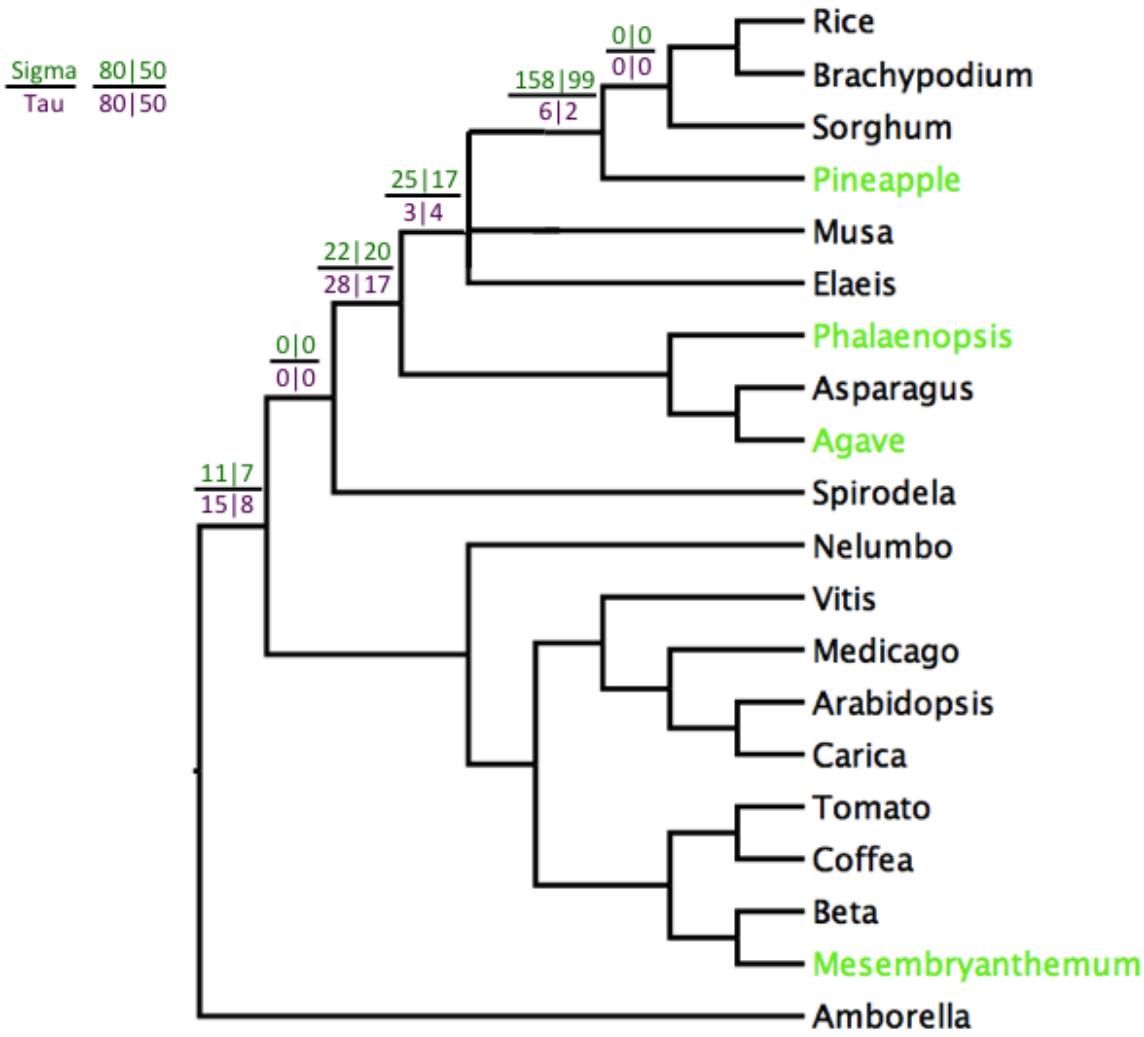
Microsynteny fractionation for 4:1 pineapple to *Amborella*, providing evidence that pineapple has undergone two WGDs in its lineage since their divergence.

Five exemplar regions are shown. Each panel contains multiple parallel tracks representing syntenic regions in rice and pineapple. Connecting lines show sequence similarities between the regions. CoGe, https://genomevolution.org/r/e426, https://genomevolution.org/r/e428, https://genomevolution.org/r/e427, https://genomevolution.org/r/e448 and https://genomevolution.org/r/e446.

Supplementary Figure 7

Microsynteny fractionation for 1:2 pineapple to rice, providing evidence that rice has undergone one WGD in its lineage (ρ) since its divergence from pineapple.

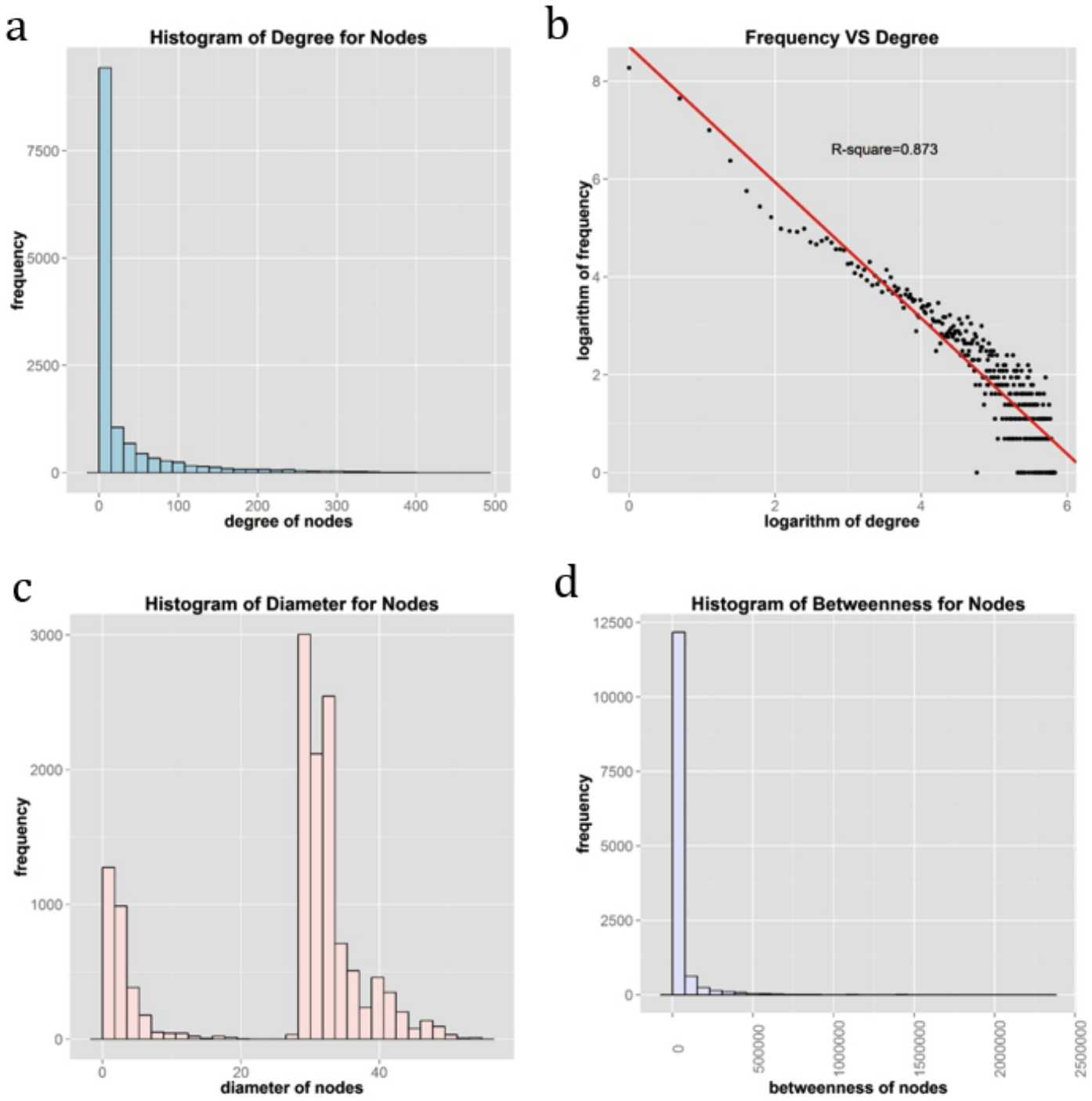
Three exemplar regions are shown. Each panel contains multiple parallel tracks representing syntenic regions in *Amborella* and pineapple. Connecting lines show sequence similarities between the regions. CoGe, https://genomevolution.org/r/e3kg, https://genomevolution.org/r/e3kw, https://genomevolution.org/r/e3k4.



Supplementary Figure 8

Dating of whole-genome duplication (WGD) events on the flowering plant tree.

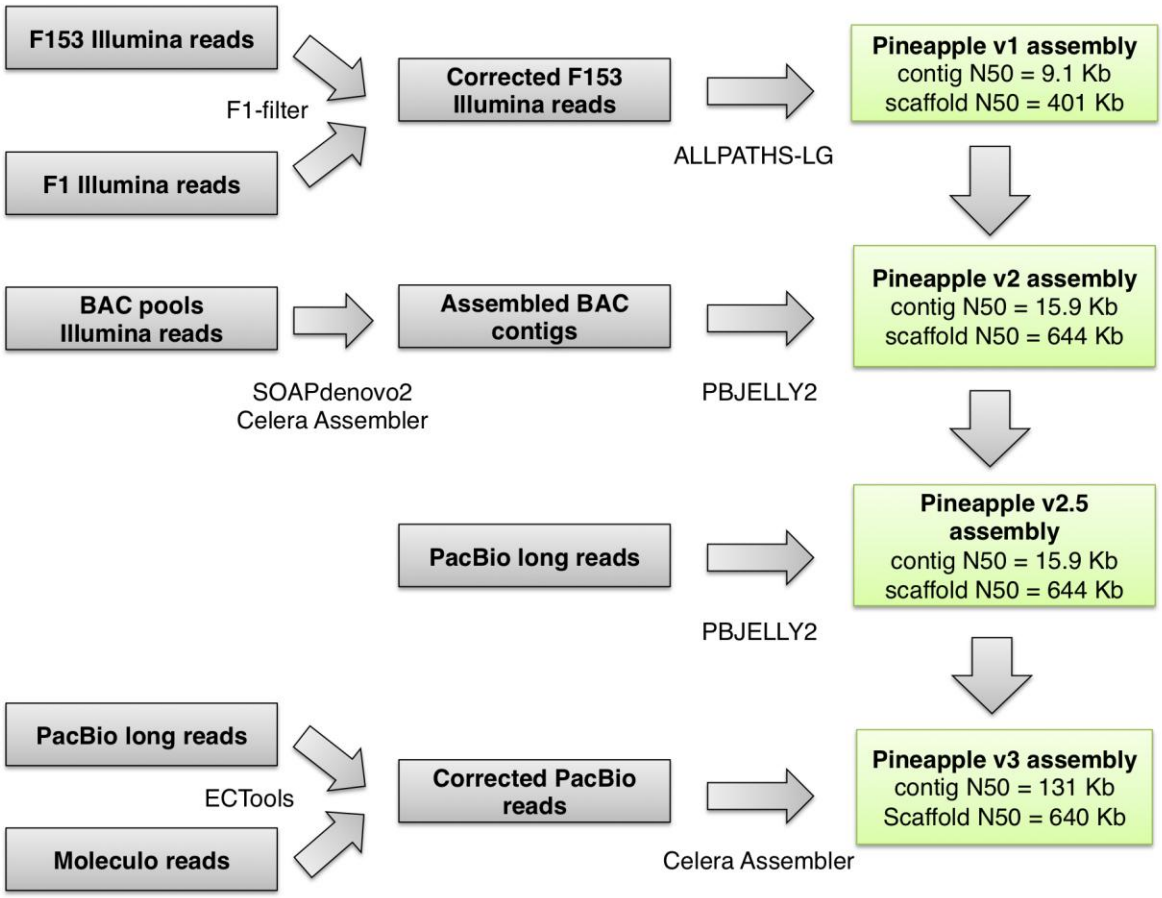
Letters represent previously identified WGDs. Estimated gene family phylogenies including genes on syntenic blocks corresponding to the σ and τ WGDs were queried to identify the timing of implied gene duplications relative to speciation events. The numbers below each lineage in the monocot clade represent gene duplication events corresponding to the σ (green) and τ (purple) synteny blocks. Trees with inferred duplication events supported by greater than 80% (left) and between 80% and 50% (right) bootstrap support values are shown for each node. Taxon names are color coded as in **Figure 2**.



Supplementary Figure 9

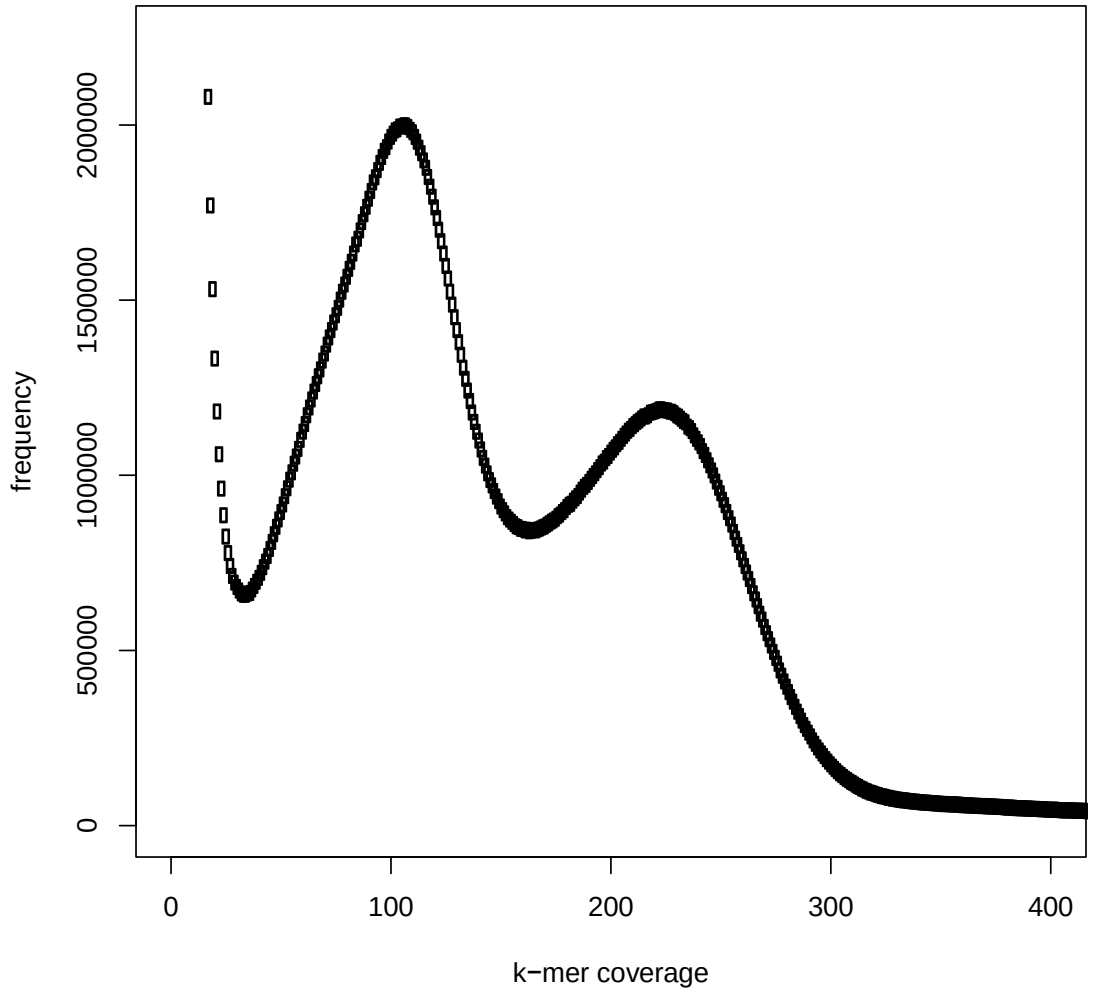
Property of leaf green tip gene interaction network.

(**a**,**c**,**d**) Distributions of the node degree, diameter and betweenness attribute. (**b**) Relationship between node degree and frequency in logarithmic coordinates.



Supplementary Figure 10

Schematic workflow of the pineapple genome assembly and improvement.



Supplementary Figure 11

***k*-mer coverage of the F153 fragment library (*k* = 23).**

Supplementary Note

Detailed summary of assembly protocol

ALLPATHS-LG assembly (v1 assembly). The first assembly of the genome used the ALLPATHS-LG software using 85× coverage of fragment library (2 × 100bp, 180bp insert length), 60× coverage of 1.5 kb jumping mates, 20× coverage of 3 kb jumping mates, and 11× coverage of 8kb jumping mates. We selected ALLPATHS-LG for the assembly based on our own prior successful results with plant genomes, as well as those of several independent evaluations. However, this resulted in a rather poor assembly with a contig N50 size of only 2kbp and a scaffold N50 of only 13kb.

On further inspection, we observed a high rate of heterozygosity in the genome (1% to 2%) that was the probable cause of the poor assembly. Notably, the histogram of k-mer coverage (k=23) frequencies of the 'F153' libraries is clearly bimodal, with homozygous k-mers at ~110x coverage and a second peak of heterozygous k-mers at ~220x coverage. Note that the term *coverage* refers to *k-mer coverage* rather than base coverage, so it is 23% below base coverage. We made several attempts to overcome the heterozygosity, including using the "HAPLOIDIFY" option in ALLPATHS-LG, and a similar algorithm of our own implementation. These approaches search for pairs of k-mers in the reads that differ by a single base, representing the two heterozygous alleles, and then systematically replace one of the k-mers with the other. This approach modestly improved the assembly to a 4.5 kb contig N50 size and a 32kb scaffold N50.

Ultimately, though, we had the best results using a novel assembly strategy leveraging our experimental design in which we had sequenced 'F153' as well as an F1 cross of 'F153' with the CB5 variety. The basis for this approach was that for any given region of the F1 genome, the F1 would inherit just one of the two chromosomes from 'F153' together with one chromosome from CB5. Thus any reads containing k-mers from 'F153' not present in the F1 must have originated from the second allele of 'F153'. We discarded those reads and their mate-pairs forming a pseudo-haploid representation of the 'F153' genome.

Specifically, we used the jellyfish algorithm to count k-mer frequencies in the 'F153' libraries and in the F1 libraries. We then discarded reads from the 'F153' libraries using the program *f1-filter* available in the AMOS package if it contained at least a single k-mer that occurred at least 40 times in the 'F153' dataset but no more than eight times in the F1 dataset. In theory any occurrences in the F1 would indicate it was inherited from 'F153', but we allowed up to eight of these k-mers in the F1 to account for sequencing errors that can give rise to artificial k-mers, especially as heterozygous k-mers often differ by just a single base. Similarly, we required the k-mer to occur at least 40 times in the 'F153' dataset to ensure it is reliable.

This filtering approach will not successfully filter reads if the CB5variety happens to pass along k-mers from the second allele of 'F153' or if there other biases in the sequencing. Nevertheless this approach filtered out from 15% to 27% of the reads of each library from the 'F153' dataset. The assembly of the resulting F1-filtered dataset had greatly improved contiguity statistics: the contig N50 size improved to 9.1kbp and the scaffold N50 size jumped to 401 kb because of the reduced heterozygosity in the reads.

The assembly also incorporated 1.7M 20 kb mates sequenced with 454, using the approach we previously used to include them within the ALLPATHS-LG analysis, as it does not natively support 454 sequencing.

This v1 intermediate assembly is available online here: *http://schatzlab.cshl.edu/data/pineapple/f1_filter.frag_ec_2.fasta.gz*

Assembly of BAC pools. Bacterial artificial chromosomes (BACs) partition the genome into smaller segments (~200kbp in length) that contain a single haplotype, thus presenting much less of a challenge for assembling a large heterozygous genome like pineapple. A total of 219 pools of pineapple BACs were sequenced, with each BAC pool containing 48, 64, 96 or 384 clones, from different batches of sequencing. The majority of the pools contained 48 clones. The aggregate coverage of the BACs is approximately 2× the length of the pineapple genome. Reads from separate BAC pools were assembled using SOAPdenovo2 [1] to generate contigs. Contigs were pooled together and assembled in Celera Assembler [2] to resolve overlapping BACs. Based on gene coverage analyses, only 63.9% of the TRINITY transcripts are considered mapped to the BAC contigs. This suggests that although ~2× genome depth of the BACs should provide 86% theoretical coverage based on Lander-Waterman model, we still missed a substantial amount of the genome using BAC method alone. Possible causes of loss of coverage might be non-random shearing of BACs, or contaminants and uneven growth among BACs within a pool.

Incorporation of BAC sequences (v2 assembly). We used PBJELLY to patch in the BAC contigs into the v1 assembly (with BLASR option: *"-minMatch 20 -minPctIdentity 96 -maxScore -500"*) [3]. Following PBJELLY, we used SSPACE with the 3 kb, 8 kb mate pair libraries to perform additional scaffolding with default settings [4]. This processing improved the assembly statistics to a 15.8 kb contig N50 size, and a 643kb scaffold N50 size.

Incorporation of PacBio and Moleculo sequences (v3 assembly). We also sequenced approximately 15× coverage of the genome using long PacBio reads (mean length: 6,232 bp, max: 35,290 bp), and used these reads with the PBJelly algorithm to close gaps in the v2 assembly. For this we used the parameters recommended for BLASR when aligning raw PacBio reads: "-minMatch 8 -minPctIdentity 70 -bestn 5 -nCandidates 20 -maxScore -500 -nproc 20 –noSplitSubreads". The result of this analysis had a marginal effect on scaffold N50 size, improving it from 643 kb to 653 kb, but significantly improved the contig N50 size, from 15.8 kb to 131 kb as it was able to close virtually all of the small scaffolding gaps in the v2 assembly.

The final assembly step was to include the long Moleculo reads that we had sequenced (mean length: 3,248bp, max: 16,672bp). These reads have the advantage of being much longer than standard Illumina sequencing reads and having a very low error rate (<1%), but we were only able to sequence the genome at ~2.3× coverage. Nevertheless, we attempted to include the Moleculo reads by using them to error-correct the PacBio reads using ECTools, our new pipeline for error-correcting PacBio reads. Briefly, ECTools uses the nucmer sequence alignment algorithm to align the Moleculo reads to the PacBio reads. It then uses a dynamic programming algorithm based on the

length and identity of the alignments to select the best set of alignments that spans each PacBio read. Those alignments are then used to error correct the raw PacBio reads with the nearly perfect Moleculo sequences. After error correction, we assembled the PacBio reads de novo using the Celera Assembler (v8), leading to an assembly with a contig N50 size of 36.8 kb (no scaffolds were generated because there were no mate pairs used in this assembly)

We evaluated the two PacBio based assemblies, one assembled using PBJelly with the BAC-sequences and one de novo assembly, based on CEGMA (eukaryotic conserved genes) and transcript coverage. We found that the PBJELLY assembly was much more complete than the *de novo* assembly, as expected due to our relatively low PacBio coverage. The only notable exceptions were six novel KOGs in the CEGMA test and 484 novel transcripts in the transcript coverage test that were only found in the *de novo* assembly on 244 contigs, with a total length of 7.8 Mb. To maximize the gene content of our analysis, the 7.8 Mb novel sequences were added to the PBJELLY assembly to create the final v3 assembly. These sequences did not change the overall contig or scaffold N50 sizes.

Quality assessment and improvement. The final pineapple v3 assembly is estimated to be 93.4% complete based on the mapping of TRINITY transcripts (requiring identity $\geq$ 98%, and coverage $\geq$ 50% of each transcript) that were assembled from diverse RNA-seq libraries. We also assessed the completeness of the assembly through coverage of 248 ultra-conserved CEGs using CEGMA [5]. A total of 220 (88.7%) CEGs can be found in full length while 243 (98.0%) can be found in partial or full length, indicating that most genic sequences were present in the current assembly. The combination of transcript coverage and CEG analyses supported a relatively complete genome assembly.

Genomic scaffolds were also compared against Sanger-sequenced pineapple BACs using NUCMER followed by MUMMERPLOT to visualize the alignments [6]. The set of pineapple BAC references includes seven BACs with a total sequence size of 582 Kb. Our final assembly covered ~85.7% of the BAC sequences. Quality assessment was performed during each round of assembly upgrade to confirm the level of improvements between the releases (Supplementary Table 17).

References

1 Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18, doi:10.1186/2047-217X-1-18 (2012).

2 Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science* **287**, 2196-2204 (2000).

3 English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**, e47768, doi:10.1371/journal.pone.0047768 (2012).

4 Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579, doi:10.1093/bioinformatics/btq683 (2011).

5 Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067, doi:10.1093/bioinformatics/btm071 (2007).
6 Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12, doi:10.1186/gb-2004-5-2-r12 (2004).
7 Paterson, A. H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-556, doi:10.1038/nature07723 (2009).

Supplementary Tables

Supplementary Table 1. Integrated assembly of the pineapple genome. "Smoothing" refers to the k-mer based preprocessing to alleviate the impact of heterozygosity.

	Illumina reads	**After smoothing Illumina**	**Smoothing + 20Kb + 454 + matepairs**	**BAC only assembly**	**Illumina+ 454+BACs**	**PACbio + Illumina + BACs + Moleculo**
Contig N50	2kb	6.5kb	9.5kb	3.5kb	15.9 kb	116 kb
Scaffold N50	7kb	91kb	408kb	23kb	644 kb	640 kb
Total Scaffold Length	427Mb	324kb	326Mb	231Mb	362 Mb	382 Mb

Supplementary Table 2. Summary of genome assembly and annotation of pineapple variety 'F153'.

(a) Assembly	Status	Number	N50 (kb)	Longest (kb)	size (Mb)	% assembly
Contigs	All	8986	126.5	1589.4	375.1	71.3
Scaffold	All	3133	11759.3	24880.7	381.9	72.6

(b) Annotation	number	Average size (bp)	Median size (bp)	Total length (Mb)	% of genome	% GC
Gene	27,024	4,893	3,428	132	35	39
Exons	157,953	252	138	40	10	49
Introns	130,929	705	294	92	24	35
miRNA	158	169	176	0.0267	3.32E–05	45.8
tRNA	502	75	73	0.0377	4.69E–05	55.3

Supplementary Table 3. Summary of 'F153' ultra high-density linkage map statistics and anchored scaffolds.

Linkage group	Physical size (bp)	Genetic distance (cM)	No. of SNPs	No. of bins	No. of scaffolds anchored
LG01	24,880,688	251.6	22,688	278	64
LG02	17,334,668	158.1	14,753	176	24
LG03	16,781,886	187.1	18,580	216	31
LG04	15,584,765	162.4	16,769	190	28
LG05	15,024,279	261.3	16,748	187	25
LG06	14,747,297	163.4	13,055	154	17
LG07	14,728,688	125.8	12,840	131	23
LG08	13,970,067	130.1	14,176	167	28
LG09	13,841,557	121.5	15,896	165	21
LG10	13,112,115	145.2	7,758	90	17
LG11	13,096,665	136.6	10,474	127	23
LG12	12,612,916	151.6	13,377	125	15
LG13	11,759,267	109.7	9,976	127	26
LG14	11,705,491	93.5	12,472	140	32
LG15	11,365,478	115.1	12,206	141	19
LG16	11,174,208	95.7	10,141	114	23
LG17	11,156,084	88.2	8,730	101	14
LG18	10,911,977	114	10,640	120	20
LG19	10,815,949	89.2	5,664	82	24
LG20	10,792,357	93.5	12,774	141	30
LG21	10,645,290	146.2	6,123	77	17
LG22	10,398,566	90.3	10,699	110	19
LG23	7,979,763	71	8,206	78	3
LG24	7,684,463	33.3	6,121	72	19
LG25	3,739,203	74.2	6,030	52	2
Total	315,843,687	3,208.6	296,896	3,361	564

Supplementary Table 4. Scaffolds containing telomere tracks are at the end of linkage groups with two exceptions.

Telomere position	'F153' map linkage group bin
scaffold 1	f270,f271
scaffold 101	u002,u003
scaffold 114	j273,j274
scaffold 14	n112*
scaffold 146	i001
scaffold 15	a001
scaffold 160	s237
scaffold 18	g156
scaffold 238	a193
scaffold 25	q180
scaffold 318	w003*
scaffold 40	h165,h166
scaffold 64	g001
scaffold 643	p256,p257
scaffold 78	h001,h002
scaffold 8	l175,l176
scaffold 86	d004,d005
scaffold 96	m001

*: bin not at the end of linkage group

Nature Genetics: doi:10.1038/ng.3435

Supplementary Table 5. Summary of gene model annotations.

(in a separate Excel file).

Supplementary Table 6. Alternative splicing events in annotated pineapple gene models.

AS event type	No. of events	Percentage of events (%)
exon skipping	329	3.2
alternative donor sites	680	6.7
alternative acceptor sites	1, 146	11.3
intron retention	6, 375	62.8
others (complex events)	1, 621	16
Total	**10, 151**	

Nature Genetics: doi:10.1038/ng.3435

Supplementary Table 7. Frequencies of miRNA families identified in leaves, flowers and fruits of pineapple.

miRNA family	Number of reads in flowers	Normalized frequency (RPTM) in flowers	Number of reads in fruits	Normalized frequency (RPTM) in fruits	Number of reads in leaves	Normalized frequency (RPTM) in leaves	Note
miR156	1,649	4,876	53	2,633	850	2,800	Conserved
miR159	1,653	4,888	307	15,249	872	2,873	Conserved
miR160	57	169	2	99	57	188	Conserved
miR162	1,629	4,817	211	10,481	1,242	4,092	Conserved
miR164	700	2,070	40	1,987	885	2,916	Conserved
miR165/166	340,019	1,005,516	3,710	184,280	55,621	183,236	Conserved
miR167	1,488	4,400	21	1,043	853	2,810	Conserved
miR168	1,107	3,274	228	11,325	419	1,380	Conserved
miR169	119	352	0	0	31	102	Conserved
miR170/171	1,414	4,182	1	50	800	2,635	Conserved
miR172	54	160	0	0	8	26	Conserved
miR319	0	0	2	99	1	3	Conserved
miR390	22	65	0	0	9	30	Conserved
miR393	175	518	4	199	25	82	Conserved
miR394	224	662	1	50	56	184	Conserved
miR395	2	6	0	0	23	76	Conserved
miR396	51,224	151,481	1,241	61,642	8,766	28,878	Conserved
miR397	15	44	0	0	97	320	Conserved
miR398	0	0	1	50	0	0	Conserved
miR399	6	18	0	0	5	16	Conserved
miR408	52	154	1	50	976	3,215	Conserved
miR477	0	0	1	50	0	0	Conserved
miR479	2	6	0	0	2	7	Conserved

miR529	116	343	16	795	141	465	Conserved
miR530	14	41	10	497	5	16	Conserved
miR535	726	2,147	14	695	185	609	Conserved
miR827	443	1,310	0	0	499	1,644	Conserved
miR845	9	27	0	0	1	3	Conserved
miR858	1	3	0	0	0	0	Conserved
miR444	3,166	9,.398	233	11,573	609	2,006	Enriched in monocots
miR528	33	98	1	50	2,373	7,818	Enriched in monocots
miR9677	3	9	0	0	1	3	Enriched in monocots
miR7782	57	169	12	596	71	234	Enriched in monocots

Supplementary Table 8. Summary of TEs and other repeats in the assembly.

TE category	DNA masked (bp)	Percentage
Class1/DIRS	77792	0.02%
Class1/LINE	4053195	1.06%
Class1/LTR	121024239	31.68%
Class1/SINE	1760427	0.46%
Class1/TRIM-LARD	27147431	7.11%
Class2/*Helitron*	1593335	0.42%
Class2/Maverick	236650	0.06%
Class2/MITE	9900150	2.59%
Class2/Transposon	4062169	1.06%
Low_complexity	2516223	0.66%
Simple_repeat	8825293	2.31%
Unclassified repeats	16862716	4.41%
Total	198059620	51.84%

Supplementary Table 9. LTR retrotransposons in the assembly and in raw reads.

Family name	Percentage in raw reads	Percentage in assembly
pusofa	28.156	0.498
ovalut	5.2	5.788
wima	3.669	3.415
gopehe	2.366	2.129
amov	2.362	2.069
afuka	2.361	1.536
jafuka	2.2	2.747
duevu	1.999	2.607
wufer	1.676	7.061
paikir	1.648	1.616
utom	1.558	1.918
mawe	1.364	0.848
duovy	1.336	0.922
jeari	1.324	1.538
beka	1.268	1.522
ijid	1.261	0.753
adok	1.216	1.018

Only families with >1% reads mapped are listed.

Supplementary Table 10. Abundance of RNA-Seq reads derived from LTR retrotransposons in nine pineapple tissue samples.

Tissue	flower	fruit 1	fruit 2	fruit 3	fruit 4	fruit 5	fruit 6	fruit 7	root	Total
Percentage	0.36	0.52	0.2	0.22	0.18	0.19	0.32	0.16	0.4	0.26

Nature Genetics: doi:10.1038/ng.3435

Supplementary Table 11. Summary of within-genome heterozygosity of 'F153', 'MD2', and CB5.

Accession	SNP/Indel	Total	Intergenic	5' UTR	CDS	Intron	3' UTR	Synony-mous[1]	Non-synonymous	Heterozygosity rate[2]
'F153'	SNP	5,371,423	3,818,616	32,394	298,600	1,168,573	53,240	100,743	195,488	1.54%
	Indel	1,212,898	841,137	17,859	39,397	296,723	17,782			0.35%
'MD2'	SNP	5,986,729	4,163,977	33,263	418,669	1,306,907	63,913	91,876	323,836	1.71%
	Indel	928,884	615,432	8,777	40,904	246,933	16,838			0.27%
CB5	SNP	8,825,470	5,878,978	54,961	543,348	2,218,761	129,422	186,520	351,908	2.53%
	Indel	1,398,306	901,913	13,808	32,388	416,602	33,595			0.40%

[1] SNPs located in overlapping regions of different transcripts were annotated separately. Some SNPs are synonymous in one transcript, but appear as non-synonymous SNPs in another overlapping transcript. Additionally, we excluded trimorphic SNPs. Consequently, the sum of synonymous and non-synonymous SNPs is not equal to the number of SNPs in the CDS regions.

[2] The masked genome size of pineapple is 349,178,920 bp, which was used as the denominator for calculation of heterozygosity rate.

Supplementary Table 12. Summary of synteny blocks with at least four gene pairs (' syntenic anchors') between pineapple and the selected genomes. Visualization of these syntenic blocks are in Supplementary Figure 5.

Comparison of pineapple	No. of blocks	No. of syntenic anchors	Max block size	Inferred syntenic depth ratio
vs. Amborella	804	9,005	61	4:1
vs. banana	2,009	21,394	86	2:8
vs. date palm	325	2,331	29	2:2
vs. duckweed	985	13,519	147	4:4
vs. grape	1,112	13,853	104	4:3
vs. oil palm	1,022	23,520	394	2:2
vs. orchid	796	7,865	51	2:2
vs. rice	992	21,003	313	1:2
vs. sorghum	912	21,060	312	1:2
vs. self	388	4,891	102	4:4 including diagonal

Supplementary Table 13. Comparison of mobile and syntenic gene families between *Arabidopsis* and sorghum

	At genes								**Sb genes**							
	all genes in study	**M**	**S**	**total M+S**	**M/(M+S)**	**S/(M+S)**	**ratio of M:S for X2 calculation**		**all genes in study**	**M**	**S**	**total M+S**	**M/(M+S)**	**S/(M+S)**	**ratio of M:S for X2 calculation**	
mobile and syntenous genes in background	20,732	3,634	4,431	8,065	0.45	0.55	45/55		31,114	8,675	7,611	16,286	0.53	0.47	53/47	
gene description	**genes families with >25 copies**	**M**	**S**	**total M+S**	**exp M ((M+S)*.45)**	**exp S ((M+S)*.55)**	***p* value if 45:55 M:S**	**mobile family?**	**genes families with >25 copies**	**M**	**S**	**total M+S**	**exp M ((M+S)*.53)**	**exp S ((M+S)*.47)**	**p value if 53:46 M:S**	**mobile family?**
F-box	967	352	23	375	168.75	206.25	1E–80	YES	619	321	67	388	205.64	182.36	8.5E–32	YES
CC/TIR-NBS-LRR	119	58	2	60	27	33	9E–16	YES	57	34	1	35	18.55	16.45	1.7E–07	YES
defensins	155	45	0	45	20.25	24.75	1E–13	YES	36	14	0	14	7.42	6.58	4.3E–04	YES
TRAF	55	17	1	18	8.1	9.9	2E–05	YES	86	57	0	57	30.21	26.79	1.2E–12	YES
B3	29	13	1	14	6.3	7.7	3E–04	YES	58	27	9	36	19.08	16.92	8.2E–03	YES
LRR	34	17	3	20	9	11	3E–04	YES	50	29	3	32	16.96	15.04	2.0E–05	YES
beta glucosidase	38	15	2	17	7.65	9.35	3E–04	YES	42	24	7	31	16.43	14.57	6.4E–03	YES
TERF	29	10	4	14	6.3	7.7	5E–02	YES	30	16	3	19	10.07	8.93	6.4E–03	YES
thionin	50	34	0	34	15.3	18.7	1E–10	YES	13	6	1	7	3.71	3.29	0.083	weak
MADS [7]	85	31	11	42	18.9	23.1	2E–04	YES	68	16	21	37	19.61	17.39	0.23	no
GDSL-like lipase/acylhydrolase	55	8	17	25	11.25	13.75	0.19	no	137	57	23	80	42.4	37.6	0.00	YES

Supplementary Table 14. List of putative pineapple CAM-related carbon fixation genes.

Gene ID	Enzyme description	Gene symbol
Aco007803.1	alpha carbonic anhydrase 1	*alpha-CA*
Aco016727.1	alpha carbonic anhydrase 1	*alpha-CA*
Aco001338.1	alpha carbonic anhydrase 7	*alpha-CA*
Aco002732.1	beta carbonic anhydrase 5	*beta-CA*
Aco006181.1	beta carbonic anhydrase 2	*beta-CA*
Aco005402.1	beta carbonic anhydrase 2	*beta-CA*
Aco014975.1	Gamma carbonic anhydrase 1	*gamma-CA*
Aco023760.1	Gamma carbonic anhydrase 1	*gamma-CA*
Aco019038.1	Gamma carbonic anhydrase-like 2	*gamma-CA*
Aco010025.1	phosphoenolpyruvate carboxylase 3	*PEPC*
Aco018093.1	phosphoenolpyruvate carboxylase 3	*PEPC*
Aco016429.1	phosphoenolpyruvate carboxylase 4	*PEPC*
Aco010095.1	phosphoenolpyruvate carboxylase kinase 1	*PPCK*
Aco013938.1	phosphoenolpyruvate carboxylase kinase 1	*PPCK*
Aco022525.1	phosphoenolpyruvate carboxylase-related kinase 1	*PEPC-related kinase*
Aco001261.1	phosphoenolpyruvate carboxylase-related kinase 2	*PEPC-related kinase*
Aco006122.1	malate dehydrogenase	*MDH*
Aco007734.1	malate dehydrogenase	*MDH*
Aco013935.1	malate dehydrogenase	*MDH*
Aco002885.1	malate dehydrogenase	*MDH*
Aco004349.1	malate dehydrogenase	*MDH*
Aco014690.1	malate dehydrogenase	*MDH*
Aco017525.1	malate dehydrogenase	*MDH*
Aco017526.1	malate dehydrogenase	*MDH*
Aco017527.1	malate dehydrogenase	*MDH*
Aco017528.1	malate dehydrogenase	*MDH*
Aco019631.1	malate dehydrogenase	*MDH*
Aco010232.1	malate dehydrogenase	*MDH*
Aco004996.1	malate dehydrogenase	*MDH*
Aco008626.1	malate dehydrogenase	*MDH*
Aco017762.1	phosphoenolpyruvate carboxykinase 1	*PEPCK*
Aco009967.1	NADP-malic enzyme 1	*NADP-ME*
Aco005631.1	NADP-malic enzyme 4	*NADP-ME*
Aco005989.1	NADP-malic enzyme 3	*NADP-ME*
Aco016569.1	NAD-dependent malic enzyme	*NAD-ME*

Aco007622.1	NAD-dependent malic enzyme 2	*NAD-ME*
Aco024818.1	Pyruvate, orthophosphate dikinase	*PPDK*
Aco014488.1	PPDK regulatory protein	*PPDK* regulatory protein

Supplementary Table 15. Potential binding motifs located on carbonic anhydrase promoter sequences of C_3, C_4, and CAM species.

	Pineapple		Orchid		Rice		Maize		Sorghum	
Gene	Gene ID	Binding motif	Gene ID	Binding motif	Gene ID	Binding motif	Gene ID	Binding motif	Gene ID	Binding motif
alpha-CA	Aco007803.1		PEQU_42036	Gbox	LOC_Os02g33030	ME	GRMZM2G024495	ME	Sobic.006G069300	ME
	Aco016727.1		PEQU_21377		LOC_Os11g05520	ME	GRMZM2G164182	ME	Sobic.004G166000	ME
	Aco001338.1		PEQU_10155	ME	LOC_Os08g36630		GRMZM5G807267	Gbox	Sobic.005G039000	
			PEQU_35651	ME	LOC_Os04g33660	ME	GRMZM2G087259		Sobic.007G155000	ME
			PEQU_35653	ME	LOC_Os12g05730		GRMZM2G113191		Sobic.007G155100	
			PEQU_41624		LOC_Os08g32750	EE	GRMZM2G113165		Sobic.007G155200	
			PEQU_33524		LOC_Os08g32780	EE	GRMZM2G009633	TCP15	Sobic.007G154800	CCA1,ME,TCP1
			PEQU_33529		LOC_Os08g32840		GRMZM2G088208		Sobic.002G224000	Gbox
			PEQU_38002	CCA1	LOC_Os09g28150	ME, Gbox				
			PEQU_40137	EE	LOC_Os08g36680	Gbox				
			PEQU_41718	EE						
beta-CA	Aco002732.1	CCA1, Gbox	PEQU_27755	ME	LOC_Os01g45274	ME, TCP15	GRMZM2G121878		Sobic.002G230100	Gbox, ME
	Aco006181.1	CCA1, ME	PEQU_37822		LOC_Os09g28910		GRMZM2G348512	ME	Sobic.003G234200	Gbox
	Aco005402.1	CCA1, CCA1	PEQU_31387				GRMZM2G094165	ME	Sobic.003G234400	
							GRMZM2G414528		Sobic.003G234500	CCA1
							GRMZM2G145101		Sobic.003G234600	
gamma-CA	Aco014975.1	ME	PEQU_08854	N/A	LOC_Os12g07220	CCA1	GRMZM2G046924	CCA1	Sobic.002G39500	
	Aco023760.1				LOC_Os01g18070		GRMZM2G140885		Sobic.003g135600	
	Aco019038.1				LOC_Os07g44840	CCA1,EE,Gbox	GRMZM2G037177	ME	Sobic.004g155100	Gbox
					LOC_Os02g30460	ME				

Supplementary Table 16. Summary of interaction partners for pineapple CAM genes.

Gene_ID	Gene name	Expression	Activator	Silencer	Repression controller	Activation controller
Aco005402.1	βCA-2	increase	243	NA	2	NA
Aco010025.1	PEPC	increase	1	NA	35	NA
Aco013938.1	PPCK	increase	161	NA	30	NA
Aco022525.1	PPCrk1	decrease	NA	3	NA	32
Aco001261.1	PPCrk3	decrease	NA	81	NA	1
Aco006122.1	MDH	increase	7	NA	32	NA
Aco007734.1	MDH	decrease	NA	5	NA	105
Aco010232.1	MDH	increase	1	NA	9	NA
Aco009967.1	NAD-ME4	decrease	NA	0	NA	254
Aco007622.1	NAD-ME2	decrease	NA	1	NA	55
Aco004996.1	pNAD-MDH	increase	24	NA	87	NA

Supplementary Table 17. Quality assessment of incrementally improved pineapple assemblies.

	v1 assembly	v2 assembly	v3 assembly
Complete CEGMA	86.30%	87.90%	88.70%
TRINITY transcript	88.90%	92.30%	93.40%
Contig N50	9.1 Kb	15.9 Kb	116 Kb
Contig Length	277 Mb	343 Mb	375 Mb
Scaffold N50	401 Kb	644 Kb	640 Kb
Scaffold Length	329 Mb	362 Mb	382 Mb
Coverage of Sanger-sequenced BACs	71.70%	86.20%	85.70%