Genomics & Bioinformatics

LETTER

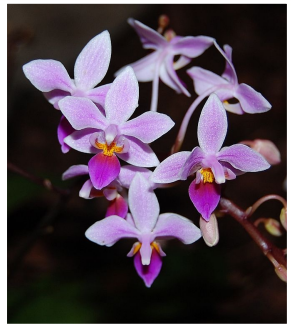
OPEN
doi:10.1038/nature23897

The *Apostasia* genome and the evolution of orchids

Guo-Qiang Zhang[1]*, Ke-Wei Liu[1]*, Zhen Li[2,3]*, Rolf Lohaus[2,3]*, Yu-Yun Hsiao[4,5]*, Shan-Ce Niu[1,6], Jie-Yu Wang[1,7], Yao-Cheng Lin[2,3]†, Qing Xu[1], Li-Jun Chen[1], Kouki Yoshida[8], Sumire Fujiwara[9], Zhi-Wen Wang[10], Yong-Qiang Zhang[1], Nobutaka Mitsuda[9], Meina Wang[1], Guo-Hui Liu[1], Lorenzo Pecoraro[1], Hui-Xia Huang[1], Xin-Ju Xiao[1], Min Lin[1], Xin-Yi Wu[1], Wan-Lin Wu[1,4], You-Yi Chen[4,5], Song-Bin Chang[4,5], Shingo Sakamoto[9], Masaru Ohme-Takagi[9,11], Masafumi Yagi[12], Si-Jin Zeng[1,7], Ching-Yu Shen[13], Chuan-Ming Yeh[11], Yi-Bo Luo[6], Wen-Chieh Tsai[4,5,13], Yves Van de Peer[2,3,14] & Zhong-Jian Liu[1,7,15,16]

nature genetics
OPEN

- **Orchidaceae is the second most species rich plant family.**
- **High economical interest.**
- **Several CAM species.**
- **Plethora of genomic data available.**

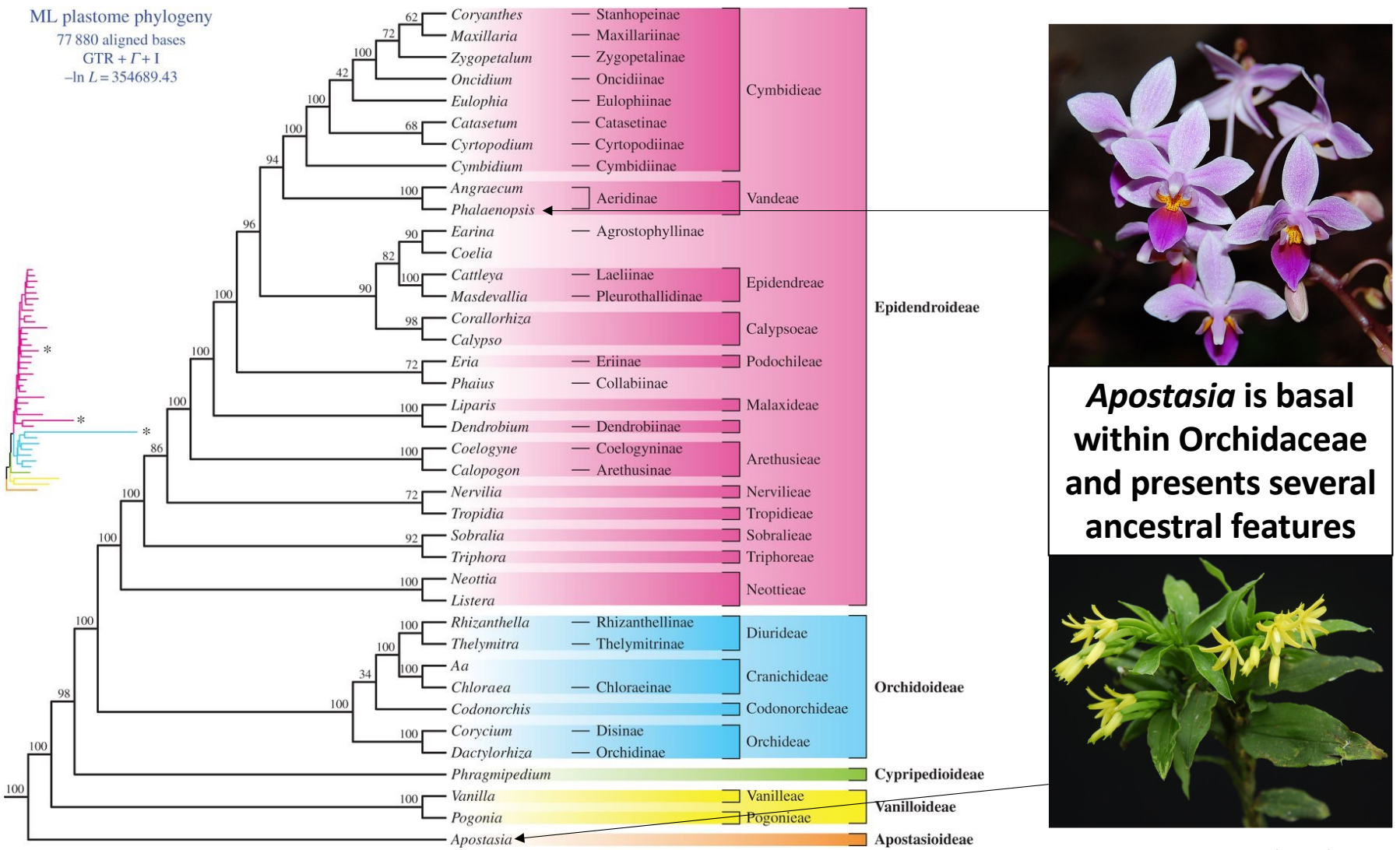


The genome sequence of the orchid *Phalaenopsis equestris*

Jing Cai[1–3,15], Xin Liu[4,15], Kevin Vanneste[5,6,15], Sebastian Proost[5,6,15], Wen-Chieh Tsai[7,15], Ke-Wei Liu[1–3,15], Li-Jun Chen[1], Ying He[5,6], Qing Xu[8], Chao Bian[4], Zhijun Zheng[4], Fengming Sun[4], Weiqing Liu[4], Yu-Yun Hsiao[9], Zhao-Jun Pan[9], Chia-Chi Hsu[9], Ya-Ping Yang[9], Yi-Chin Hsu[9], Yu-Chen Chuang[9], Anne Dievart[10], Jean-Francois Dufayard[10], Xun Xu[4], Jun-Yi Wang[4], Jun Wang[4], Xin-Ju Xiao[1], Xue-Min Zhao[11], Rong Du[11], Guo-Qiang Zhang[1], Meina Wang[1], Yong-Yu Su[12], Gao-Chang Xie[1], Guo-Hui Liu[1], Li-Qiang Li[1], Lai-Qiang Huang[1–3,12], Yi-Bo Luo[8], Hong-Hwa Chen[9,13], Yves Van de Peer[5,6,14] & Zhong-Jian Liu[1,2,12]

Orchidaceae, renowned for its spectacular flowers and other reproductive and ecological adaptations, is one of the most diverse plant families. Here we present the genome sequence of the tropical epiphytic orchid *Phalaenopsis equestris*, a frequently used parent species for orchid breeding. *P. equestris* is the first plant with crassulacean acid metabolism (CAM) for which the genome has been sequenced. Our assembled genome contains 29,431 predicted protein-coding genes. We find that contigs likely to be underassembled, owing to heterozygosity, are enriched for genes that might be involved in self-incompatibility pathways. We find evidence for an orchid-specific paleopolyploidy event that preceded the radiation of most orchid clades, and our results suggest that gene duplication might have contributed to the evolution of CAM photosynthesis in *P. equestris*. Finally, we find expanded and diversified families of MADS-box C/D-class, B-class *AP3* and *AGL6*-class genes, which might contribute to the highly specialized morphology of orchid flowers.

Why producing an annotated draft genome sequence for *Apostasia*?

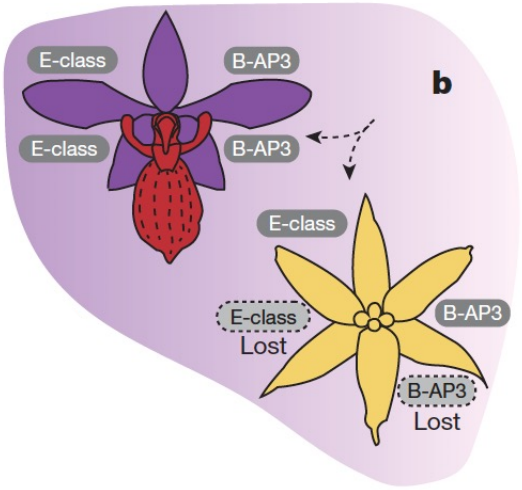


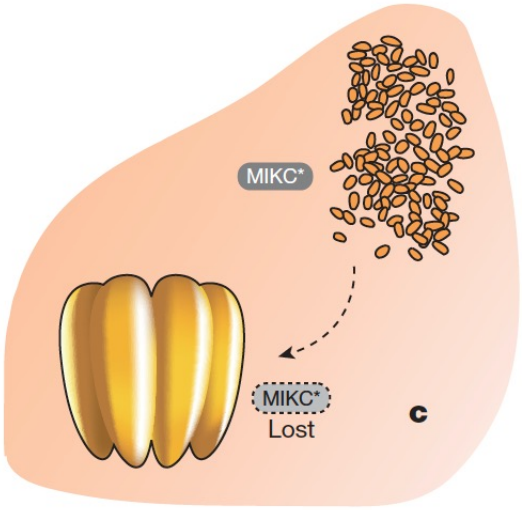
Apostasia is basal within Orchidaceae and presents several ancestral features

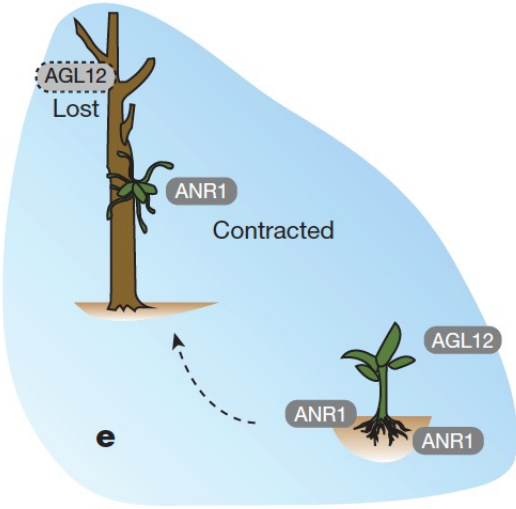
Givnish *et al.* (2015) *Proc. Royal Soc. B*

***Apostasia* presents a number of characters that are plesiomorphic in orchids:**

1. Actinomorphic perianth with an undifferentiated labellum.
2. Gynostemium with partially fused androecium and gynoecium (*Solanum*-like flower).
3. Pollen not aggregated into pollinia.
4. Underground roots for terrestrial growth.





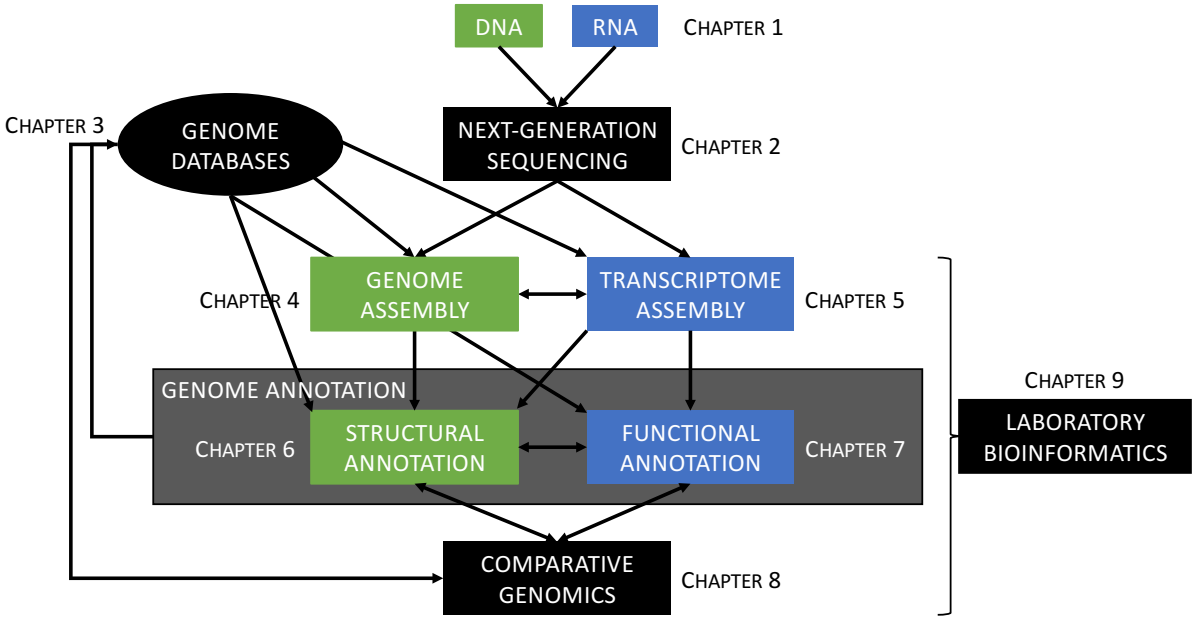


Zhang *et al.* (2017) *Nature*

This study aims at shedding light into the genetic mechanisms underpinning key innovations:

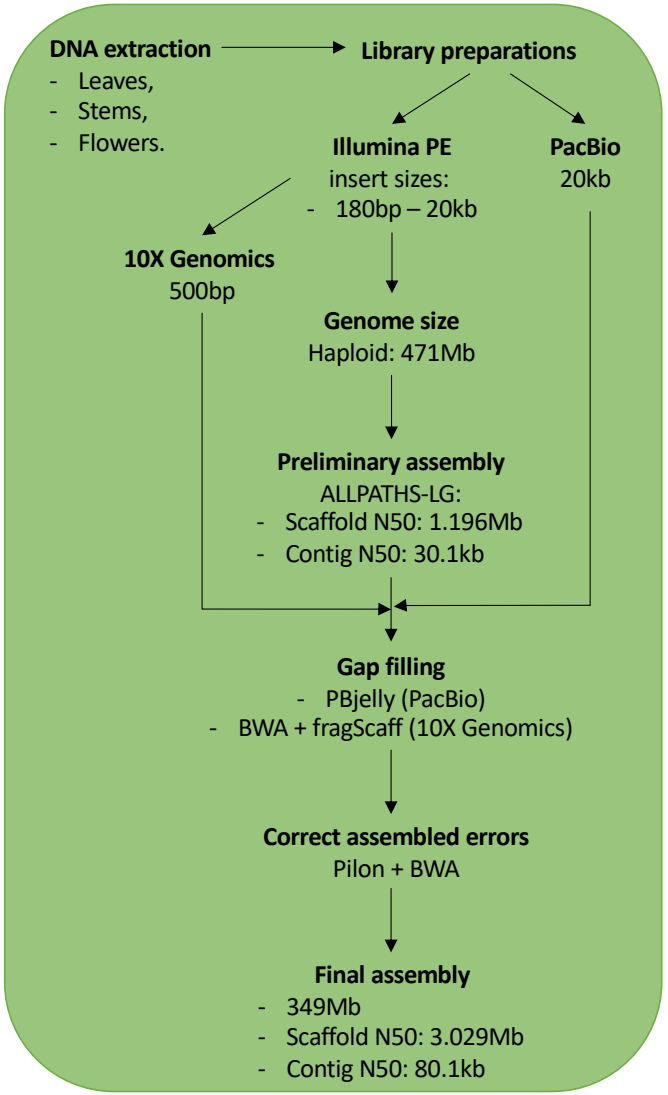
- Flower development (labellum, gynostegium, pollinia and seeds without endosperm).
- Evolution of habitus (especially epiphytism).
- Clarify evolutionary history of Orchidaceae within angiosperms.

How: By producing an annotated draft genome of *Apostasia* and conducting comparative analyses



Material & Methods

Genome assembly

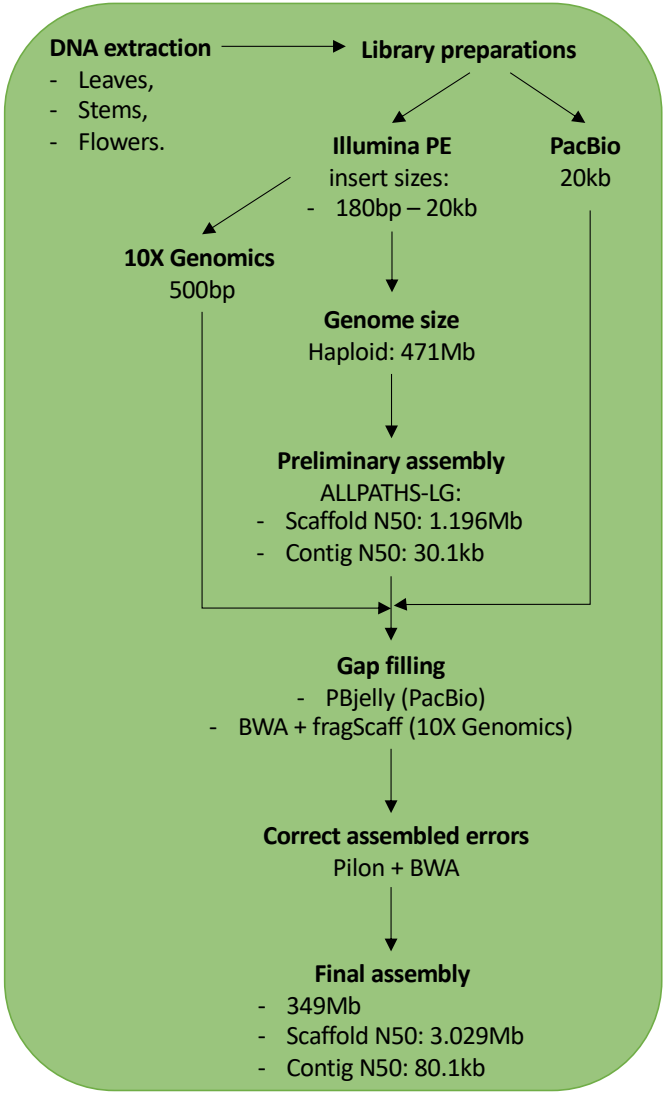


Supplementary Table 1 | Summary of the *A. shenzhenica* genome sequencing data derived from the Illumina technology.

Insert size(bp)	Read length(bp)	Number of reads	Total data(Gb)	Sequence depth(X)
180	90	168,223,606	15.14	34.97
500	100	275,127,442	27.51	63.54
800	90	98,419,616	8.86	20.46
2000	90	91,713,060	8.25	19.06
5000	90	119,967,670	10.80	24.94
10000	90	46,631,366	4.20	9.69
20000	125	58,492,233	5.26	12.23

Supplementary Table 3 | Summary of the 10X genomics Linked-Reads sequencing derived from the Illumina technology.

Species	Read length (bp)	Raw paired reads	Raw bases	Filtered paired reads	Filtered bases
A. shenzhenica	150	369,749,121	110,924,736,300	318,763,894	95,629,168,200

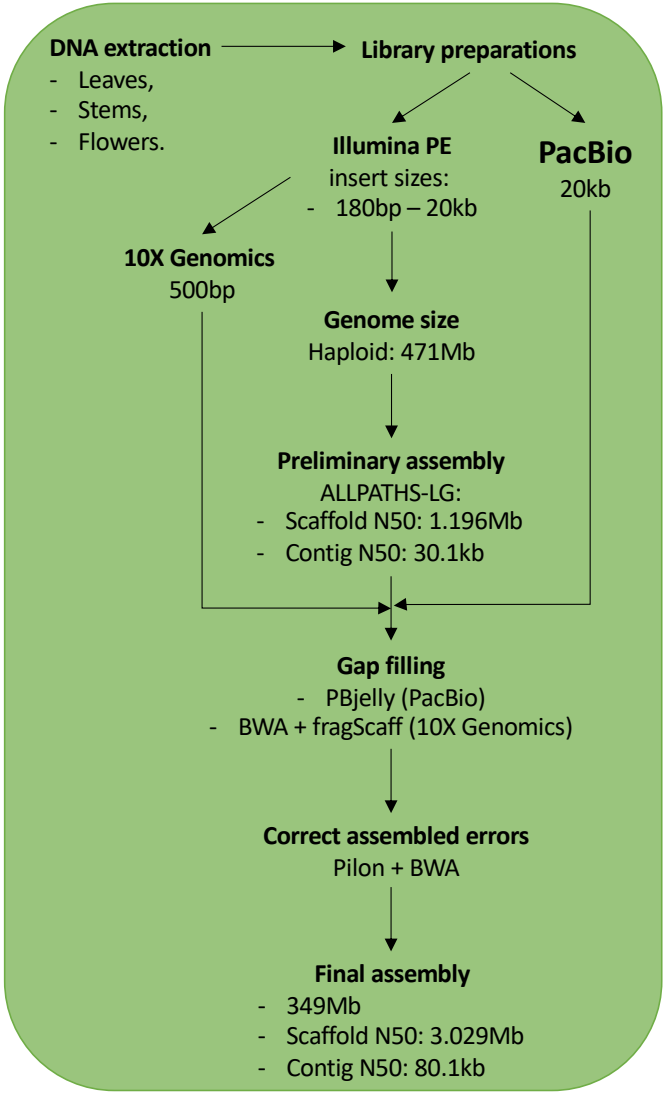


Material & Methods

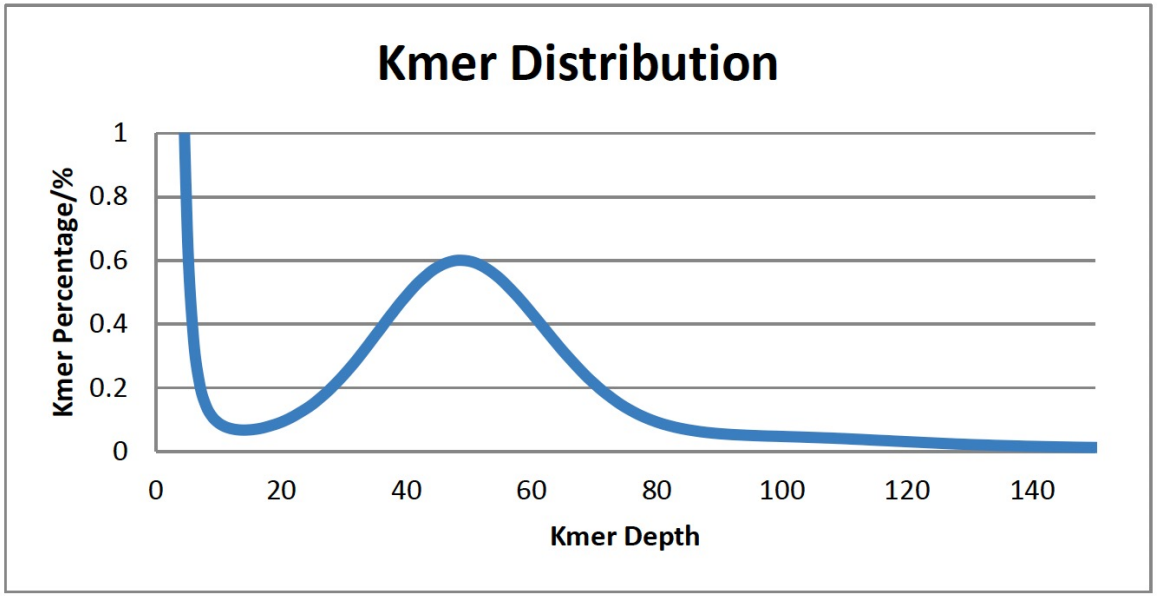
Genome assembly

Supplementary Table 2 | Summary of the 3rd generation sequencing derived from the PacBio RS II.

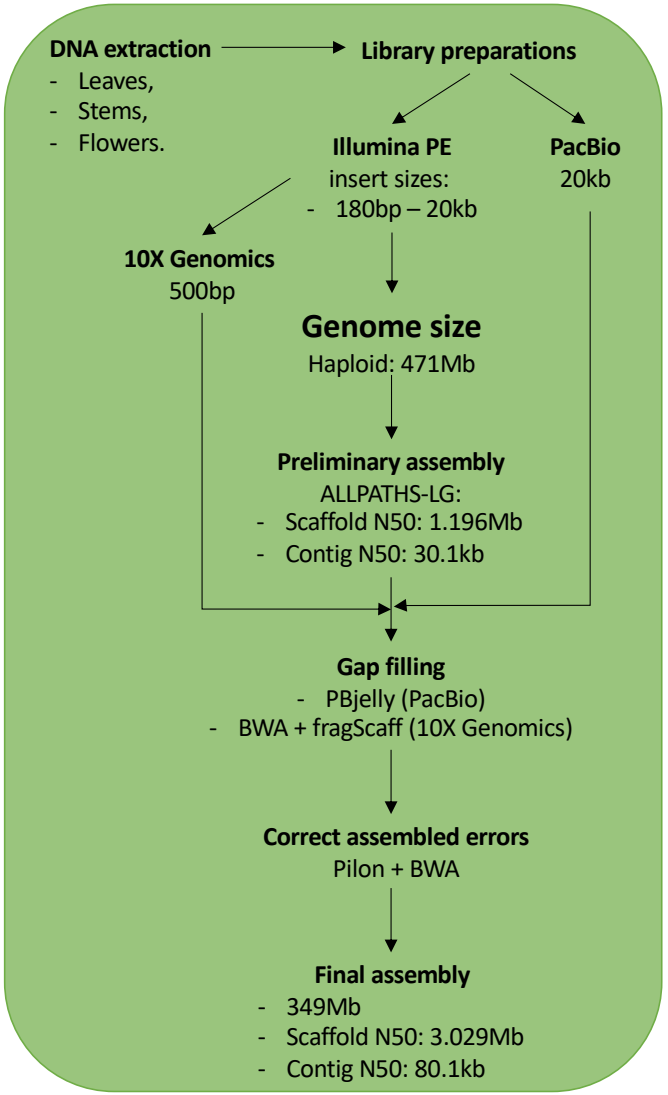
Species	Number of bases	Number of reads	Mean read length (bp)
A. shenzhenica	5,441,238,461	1,352,628	4,023



MATERIAL & METHODS

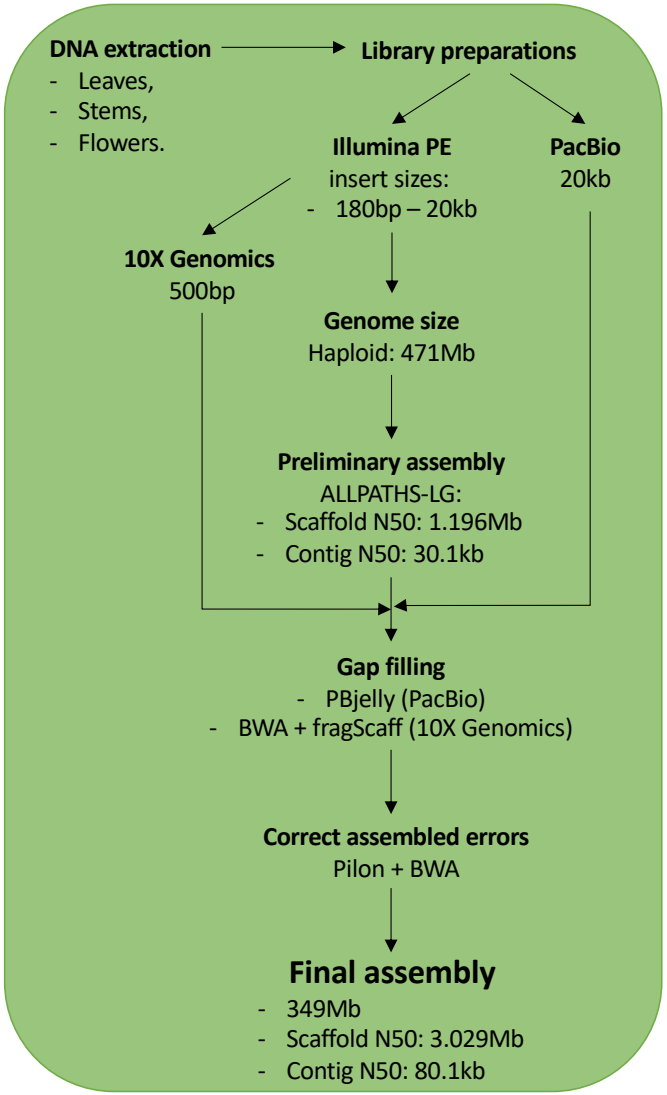


Supplementary Figure 17 | *K-mer* distribution of sequencing reads. According to the distribution, we estimate that the genome size of *A. shenzhenica* is approximately 471 Mb. The analysis is based on the Illumina data.



Supplementary Table 4 | Summary of the *A. shenzhenica* genome assembled by Illumina, PacBio and 10X genomics technologies.

	Scaffold		Contig	
	Length(bp)	Number	Length(bp)	Number
max_len	12,424,053		556,054	
N10	10,110,636	4	223,148	112
N20	6,237,011	8	166,933	283
N30	5,003,307	14	130,671	503
N40	3,457,059	22	103,308	780
N50	3,029,156	32	80,069	1,136
N60	2,413,737	45	63,275	1,590
N70	1,972,814	61	47,252	2,184
N80	1,402,703	82	31,086	3,022
N90	765,391	115	15,048	4,473
Total_length	348,734,287		322,901,144	
GC_rate	31.2%		33.7%	



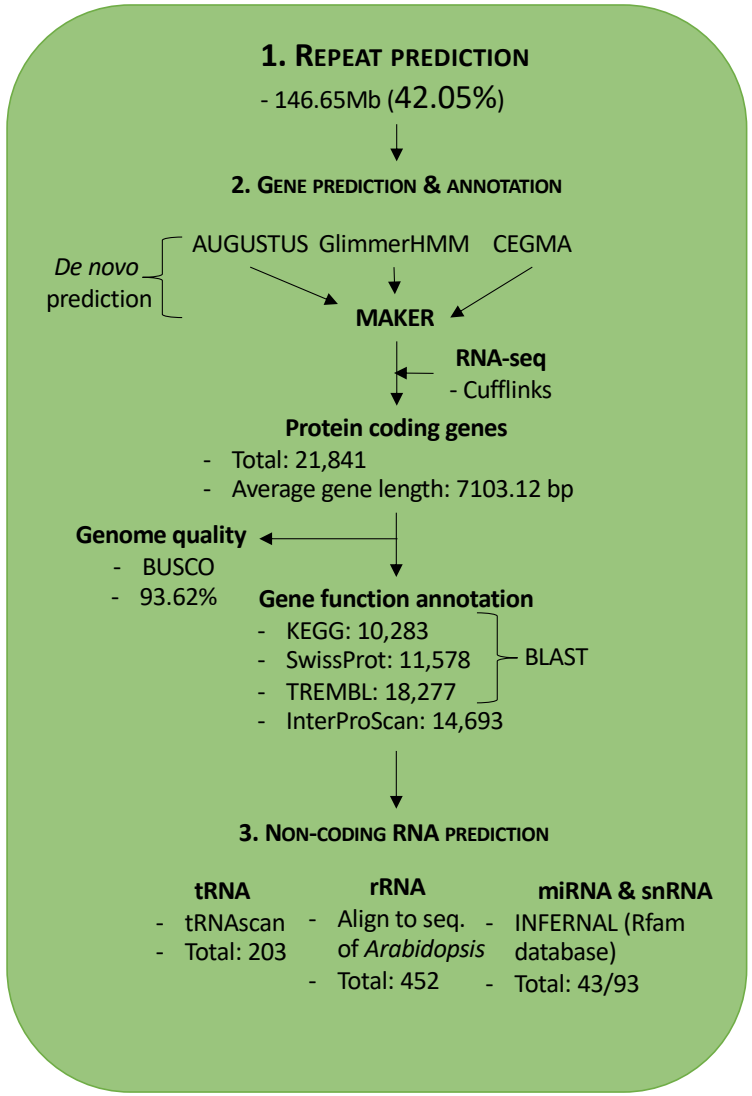
MATERIAL & METHODS

GENOME ANNOTATION

First step is to annotate repeats and then mask those for gene prediction and annotation

Supplementary Table 9 | Summary of repeat annotation of *A. shenzhenica*.

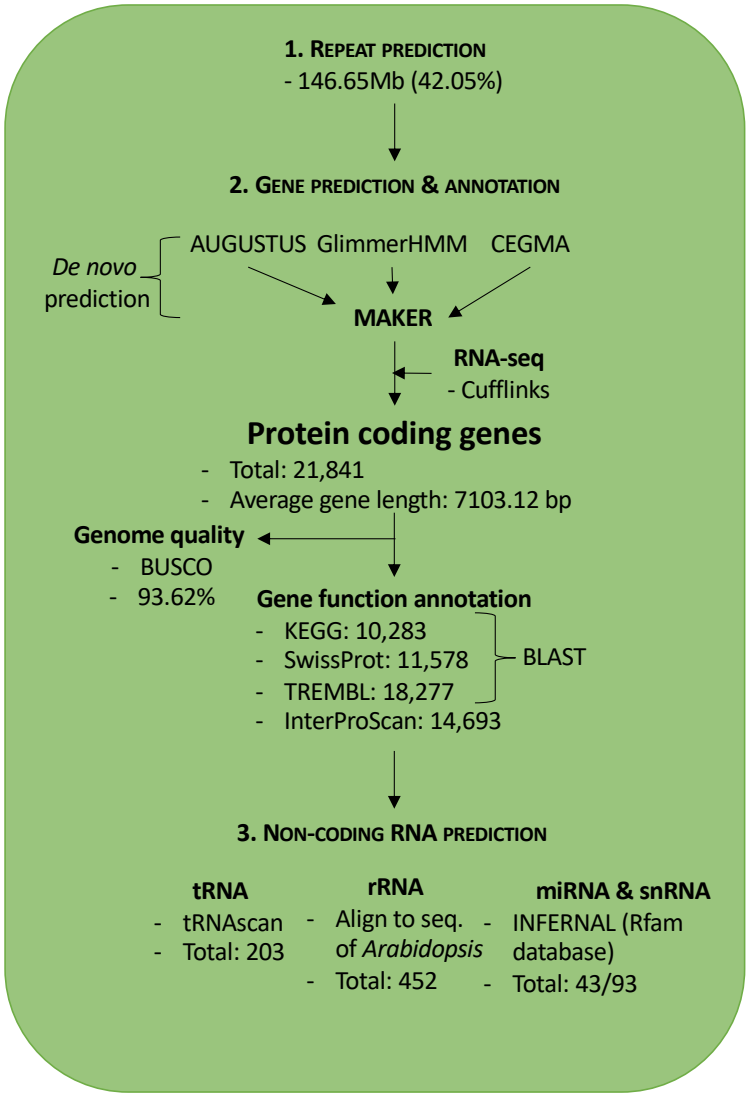
Type of repeats	RepBase TEs		TE Proteins		*De novo*		Combined TEs	
	Length (bp)	%in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome
DNA	3,604,289	1.03	3,106,810	0.89	18,995,301	5.45	22,534,396	6.46
LINE	10,240,458	2.94	9,511,200	2.73	41,316,780	11.85	44,203,442	12.68
SINE	12,411	0.00	0	0.00	149,789	0.04	161,345	0.05
LTR	10,644,451	3.05	15,167,007	4.35	72,767,333	20.87	76,930,066	22.06
Other	5,732	0.00	0	0.00	0	0.00	5,732	0.00
Unknown	38,684	0.01	0	0.00	20,482,533	5.87	20,520,835	5.88
Total	24,555,914	7.04	27,699,462	7.94	137,241,384	39.35	146,653,786	42.05



Material & Methods
Genome annotation

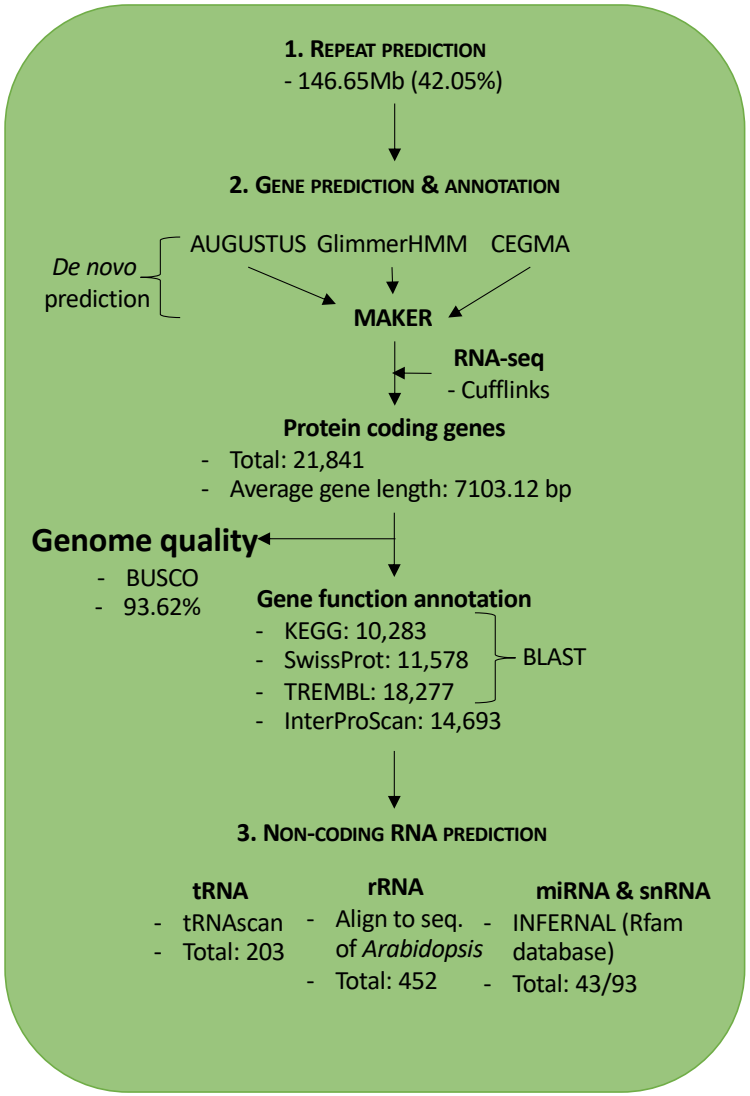
Supplementary Table 5 | Summary of gene annotation of *A. shenzhenica*.

Gene set		Protein coding gene number	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	AUGUSTUS	26,015	8771.89	1128.15	4.68	241.14	2078.02
	GlimmerHMM	36,406	8249.20	701.52	3.39	207.17	3163.09
Homolog	*A. thaliana*	19,532	5099.79	904.39	4.02	225.25	1391.48
(exonerate)	*O. sativa*	21,804	4905.70	870.30	3.81	228.44	1436.23
	P. equestris	28,179	3899.98	775.81	3.35	231.88	1331.82
	S. bicolor	20,483	4829.59	881.21	3.93	223.95	1345.31
	Z. mays	20,929	4620.41	852.71	3.79	224.79	1348.81
RNA-seq (Cufflinks)		20,202	9588.04	1144.15	4.77	239.67	1471.21
CEGMA		448	11532.37	1225.80	8.46	144.82	1380.78
MAKER		23,181	7866.12	994.09	4.08	243.45	1915.22
Final set		21,841	7103.12	1099.99	4.51	244.07	1436.59



Material & Methods
Genome annotation

	A. shenzhenica			
	Assembly		**Gene set**	
	Proteins	**Percentage**	**Proteins**	**Percentage**
Complete Single-Copy BUSCOs	685	71.65%	575	60.15%
Complete Duplicated BUSCOs	210	21.97%	304	31.8%
Fragmented BUSCOs	20	2.09%	38	3.97%
Missing BUSCOs	41	4.29%	39	4.08%
Total BUSCO groups searched	956	100%	956	100%

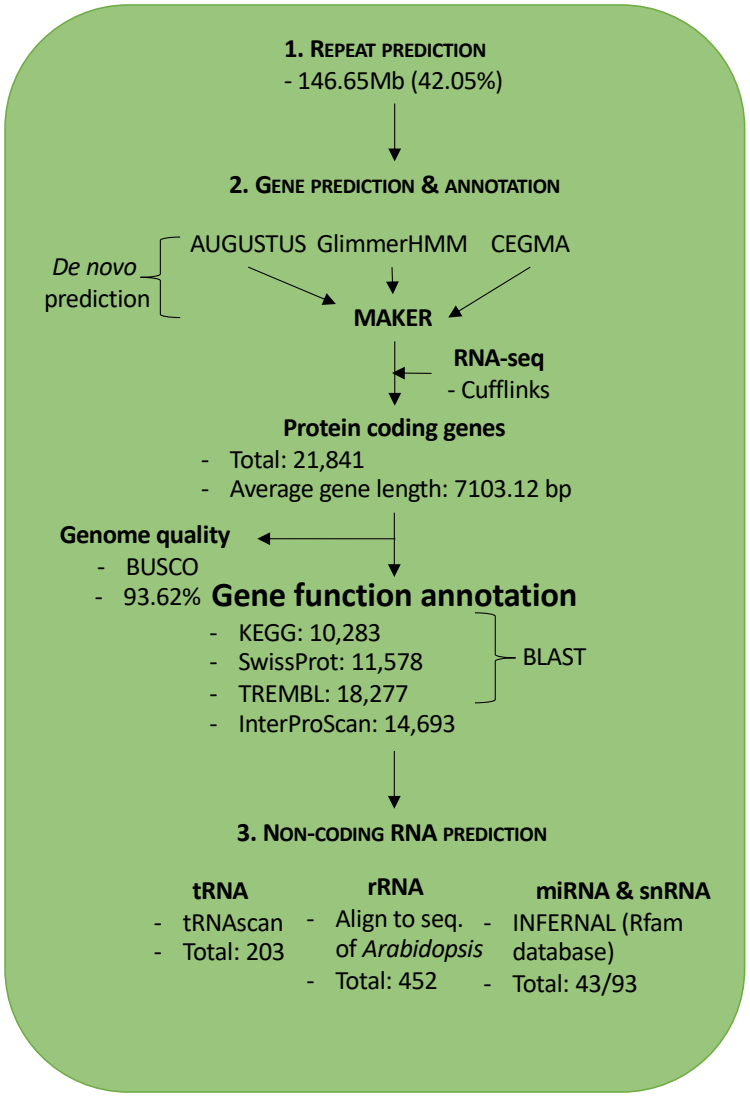


Material & Methods

Genome annotation

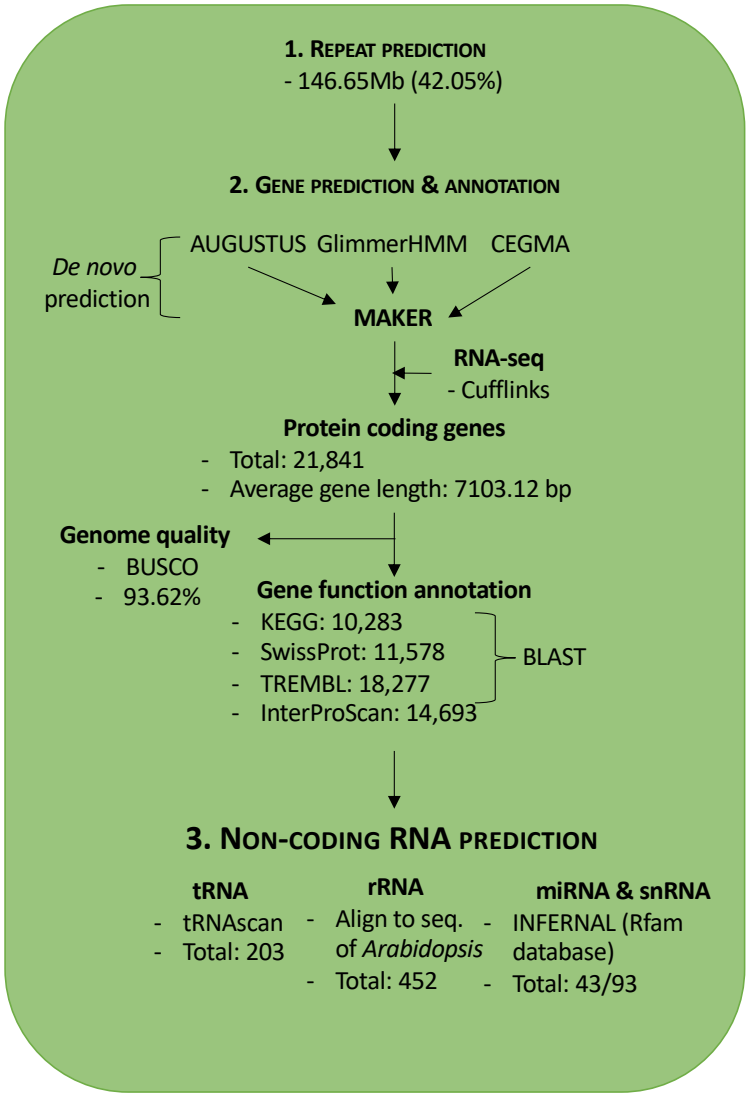
Supplementary Table 11 | Gene function annotation of *A. shenzhenica*.

		Number	Percent (%)
Total		21,841	
Annotated	InterPro	14,693	67.27
	GO	10,499	48.07
	KEGG	10,283	47.08
	SwissProt	11,578	53.01
	TrEMBL	18,277	83.68
	NCBI non-redundant	18,243	83.52
Unannotated		3,449	15.79



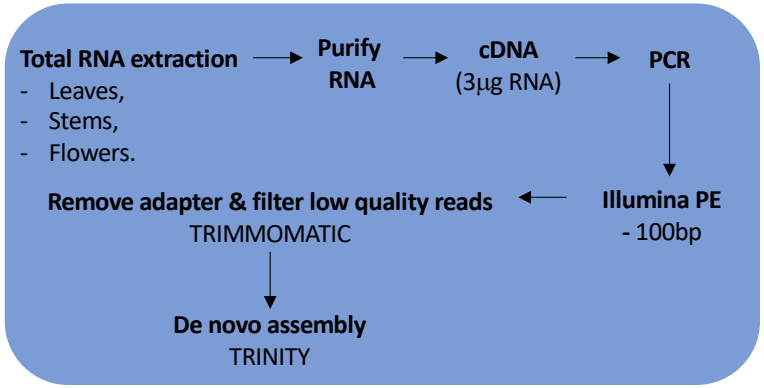
Supplementary Table 12 | Summary of ncRNA annotation of *A. shenzhenica*.

Type		Number	Average length (bp)	Total length (bp)	% of genome
miRNA		43	125.56	5,399	0.00155
tRNA		203	74.75	15,174	0.00435
rRNA	rRNA	452	162.68	73,530	0.02109
	18S	35	771.83	27,014	0.00775
	28S	26	192.96	5,017	0.00144
	5.8S	11	151.73	1,669	0.00048
	5S	380	104.82	39,830	0.01142
snRNA	snRNA	93	103.60	9,635	0.00276
	CD-box	45	98.42	4,429	0.00127
	HACA-box	0	0.00	0	0.00000
	splicing	48	108.46	5,206	0.00149
	scaRNA	0	0.00	0	0.00000



MATERIAL & METHODS

TRANSCRIPTOME ASSEMBLY



Supplementary Table 13 | Information about the transcriptomes used in this study.

Family	Subfamily	Species	Tissues used in genome annotation	Tissues used in expression analysis	Tissues used in WGD and phylogenetic analysis
Orchidaceae	Apostasioideae	*Apostasia shenzhenica*	Flower bud, leaf, root, seed, stem, tuber	flower bud, pollen, stem, root, leaf, seed	
		Apostasia odorata			flower bud
		Neuwiedia malipoensis			flower

Material & Methods

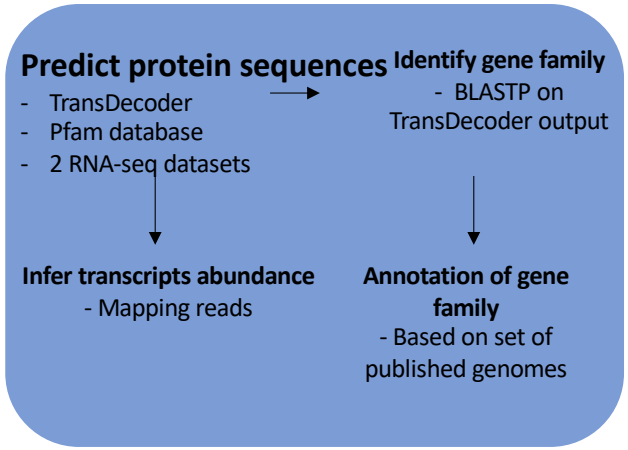
Transcriptome annotation

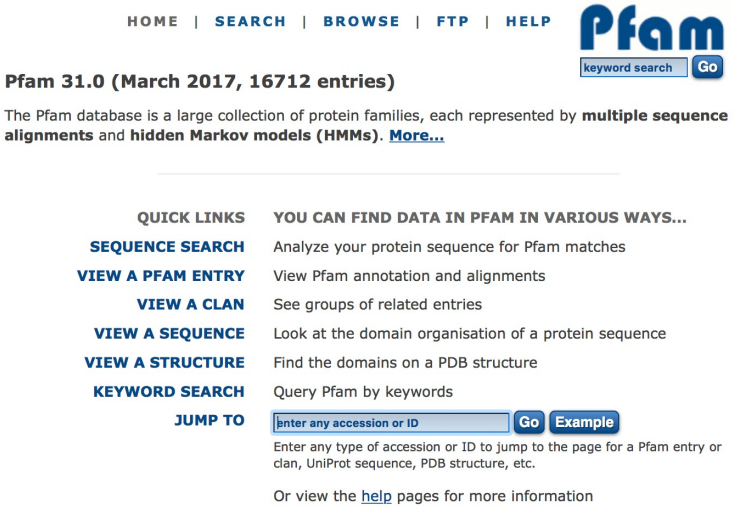
TransDecoder (Find Coding Regions Within Transcripts)

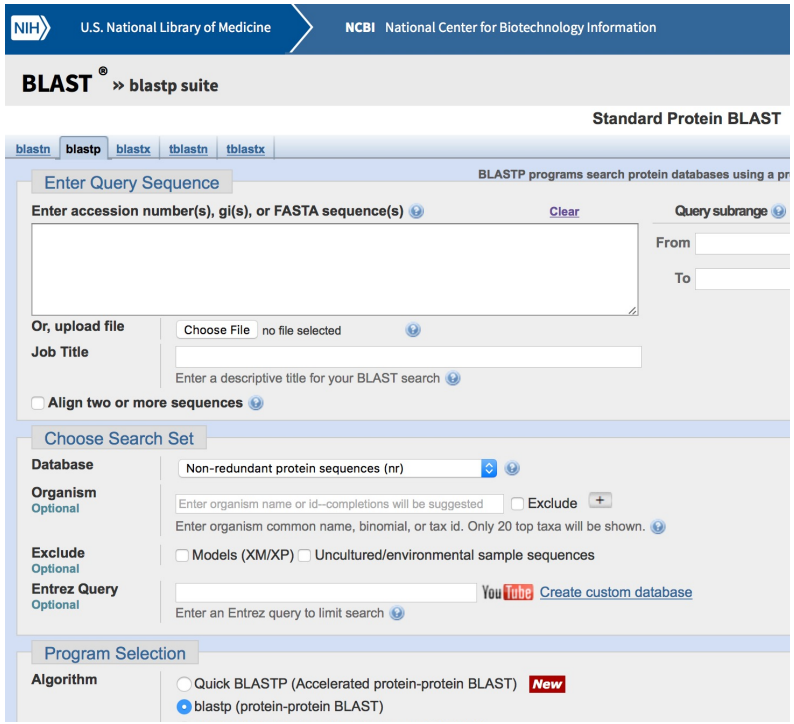
TransDecoder identifies candidate coding regions within transcript sequences, such as those generated by de novo RNA-Seq transcript assembly using Trinity, or constructed based on RNA-Seq alignments to the genome using Tophat and Cufflinks.

TransDecoder identifies likely coding sequences based on the following criteria:

- a minimum length open reading frame (ORF) is found in a transcript sequence
- a log-likelihood score similar to what is computed by the GeneID software is > 0.
- the above coding score is greatest when the ORF is scored in the 1st reading frame as compared to scores in the other 2 forward reading frames.
- if a candidate ORF is found fully encapsulated by the coordinates of another candidate ORF, the longer one is reported. However, a single transcript can report multiple ORFs (allowing for operons, chimeras, etc).
- a PSSM is built/trained/used to refine the start codon prediction.
- **optional** the putative peptide has a match to a Pfam domain above the noise cutoff score.







Predict protein sequences
- TransDecoder
- Pfam database
- 2 RNA-seq datasets

Identify gene family
- BLASTP on TransDecoder output

Infer transcripts abundance
- Mapping reads

Annotation of gene family
- Based on set of published genomes

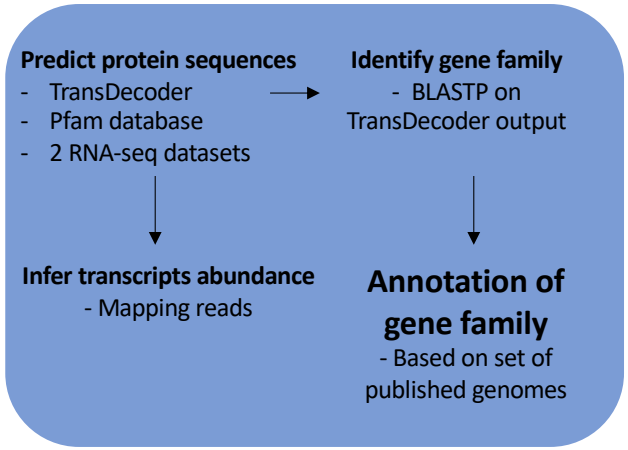
Supplementary Table 13 | Information about the transcriptomes used in this study.

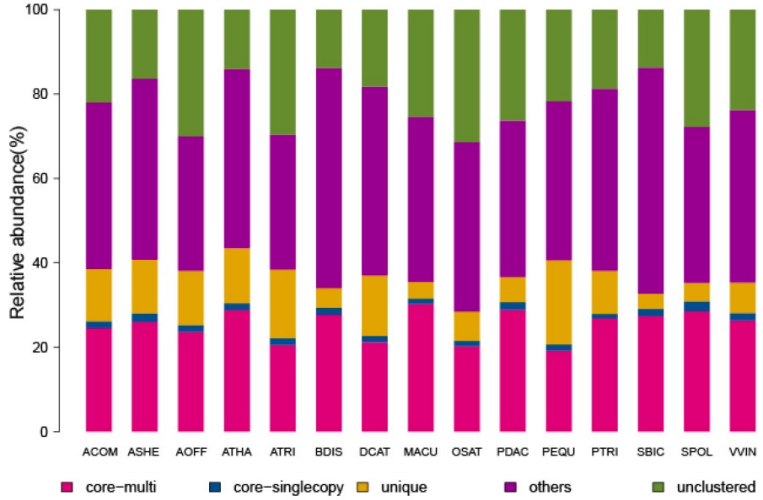
Family	Subfamily	Species	Tissues used in genome annotation	Tissues used in expression analysis	Tissues used in WGD and phylogenetic analysis	Number of predicted proteins	Number of predicted proteins with plant homologs
						TransDecoder	BLASTP
Orchidaceae	Apostasioideae	*Apostasia shenzhenica*	Flower bud, leaf, root, seed, stem, tuber	flower bud, pollen, stem, root, leaf, seed			
		Apostasia odorata			flower bud	23,504	18,030
		Neuwiedia malipoensis			flower	25,211	23,011

Supplementary Table 14 | Summary of orthologous gene families in 15 sequenced plant species.

Species	Genes	Unclustered genes	Clustered genes	Familys	Unique families	Unique families genes	Common families	Common families genes	Single Copy	Average genes per family
A. comosus	27,024	5,950	21,074	13,279	936	3,346	4,120	7,079	439	1.587
A. shenzhenica	21,841	3,573	18,268	11,995	562	2,789	4,120	6,121	439	1.523
A. officinalis	27,375	8,220	19,155	12,014	901	3,521	4,120	6,920	439	1.594
A. thaliana	26,637	3,750	22,887	12,719	859	3,466	4,120	8,108	439	1.799
A. trichopoda	25,933	7,699	18,234	12,200	1,044	4,206	4,120	5,758	439	1.495
B. distachyon	26,415	3,655	22,760	15,344	421	1,240	4,120	7,748	439	1.483
D. catenatum	29,257	5,339	23,918	14,050	1,036	4,183	4,120	6,638	439	1.702
M. acuminata	34,241	8,710	25,531	12,865	538	1,359	4,120	10,792	439	1.985
O. sativa	35,402	11,106	24,296	16,352	958	2,473	4,120	7,604	439	1.486
P. dactylifera	23,890	6,281	17,609	11,011	444	1,431	4,120	7,331	439	1.599
P. equestris	29,545	6,420	23,125	13,752	1,197	5,887	4,120	6,112	439	1.682
P. trichocarpa	40,984	7,683	33,301	14,471	1,362	4,181	4,120	11,440	439	2.301
S. bicolor	27,160	3,723	23,437	15,749	361	984	4,120	7,893	439	1.488
S. polyrrhiza	18,357	5,095	13,262	10,076	264	797	4,120	5,672	439	1.316
V. vinifera	25,328	6,032	19,296	12,808	643	1,833	4,120	7,113	439	1.507

Unique families = families present only in one species





Emergence of key innovations in orchids were unveiled by the analysis of transcriptomes

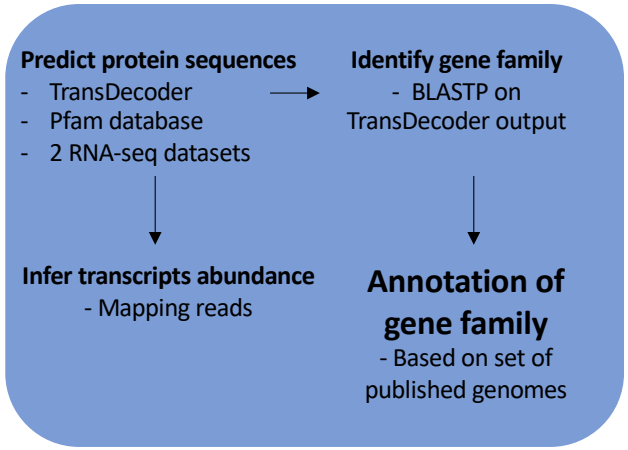


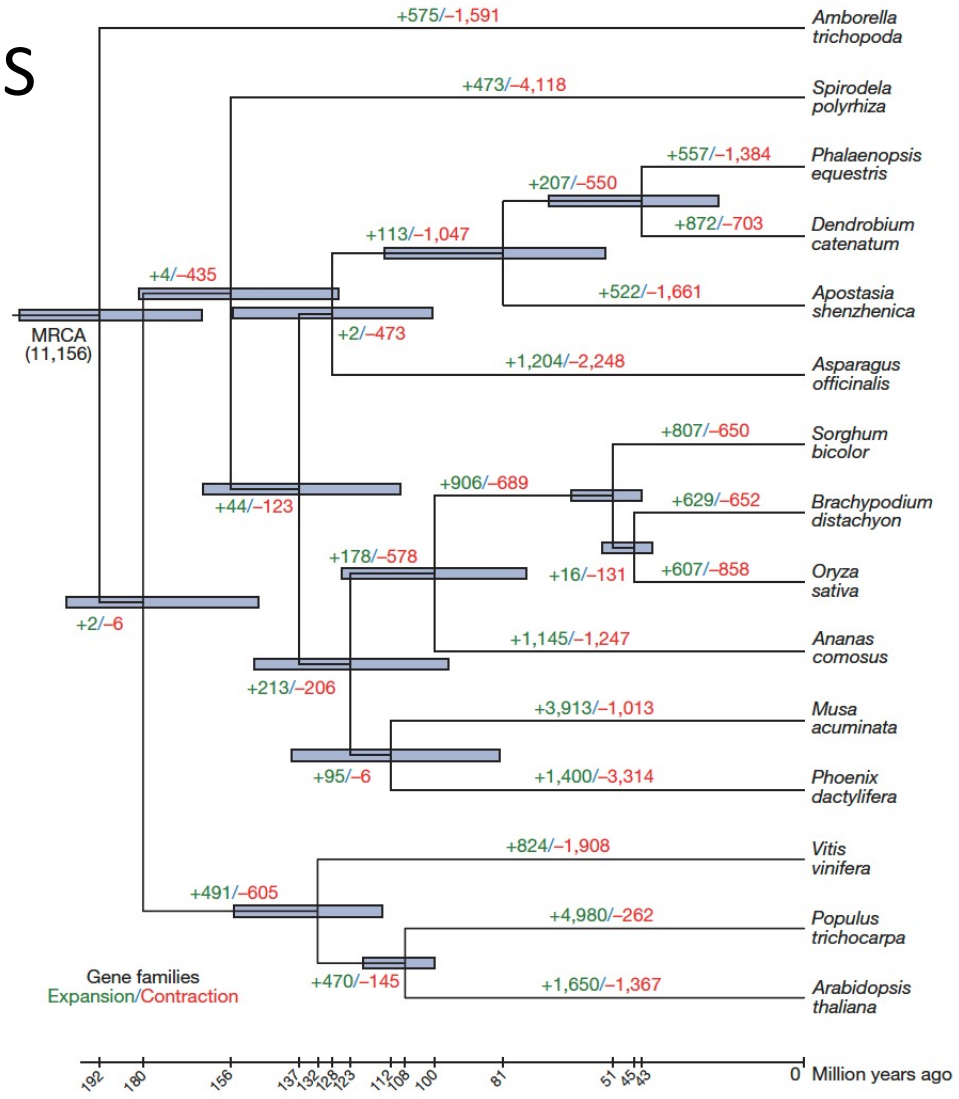
Table 1 | MADS-box genes in the *A. shenzhenica*, *P. equestris*, *D. catenatum*, *P. trichocarpa*, *A. thaliana* and *O. sativa* genomes

Category	*A. shenzhenica*		*P. equestris*		*D. catenatum*		*P. trichocarpa**		*A.thaliana**		*O. sativa**	
	Functional	Pseudo	Functional	Pseudo	Functional	Pseudo	Functional	Pseudo	Functional	Pseudo	Functional	Pseudo
Type II (Total)	27	4	29	1	35	11	64	3	47	5	48	1
MIKC^c	25	3	28	1	32	9	55	2	43	4	47	1
MIKC*	2	1	1	0	3	2	2	0	2	0	1	0
Mδ	0	0	0	0	0	0	7	1	4	1	0	0
Type I (Total)	9	0	22	8	28	1	41	9	62	36	32	6
Mα	5	0	10	6	15	1	23	4	20	23	15	2
Mβ	0	0	0	0	0	0	12	5	17	5	9†	1
Mγ	4	0	12	2	13	0	6	0	21	8	8	3
Total	36	4	51	9	63	12	105	12	107	41	80	7

*Genes with stop codon in MADS-box domain were categorized as pseudogenes[29].
†Nine MADS-box genes belonging to the M_β subgroup were identified[30].

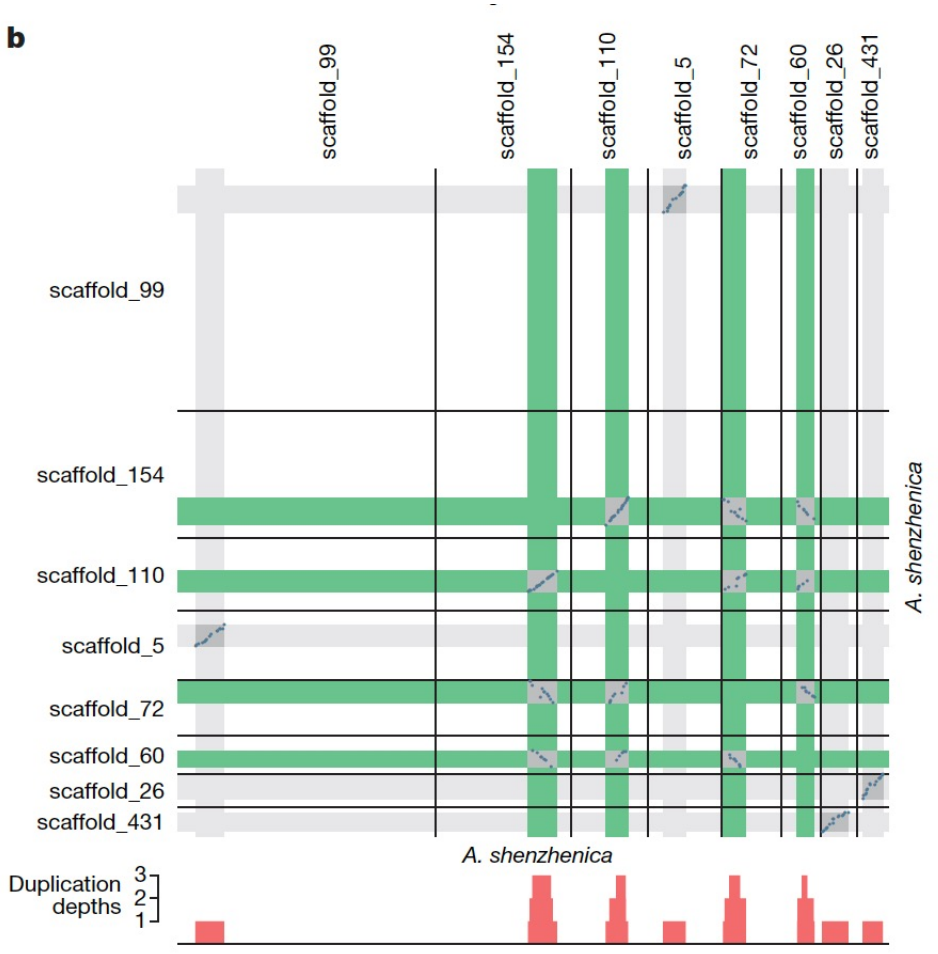
COMPARATIVE GENOMICS

Annotated genome allows inferring expansion and contraction of gene families along phylogenetic tree



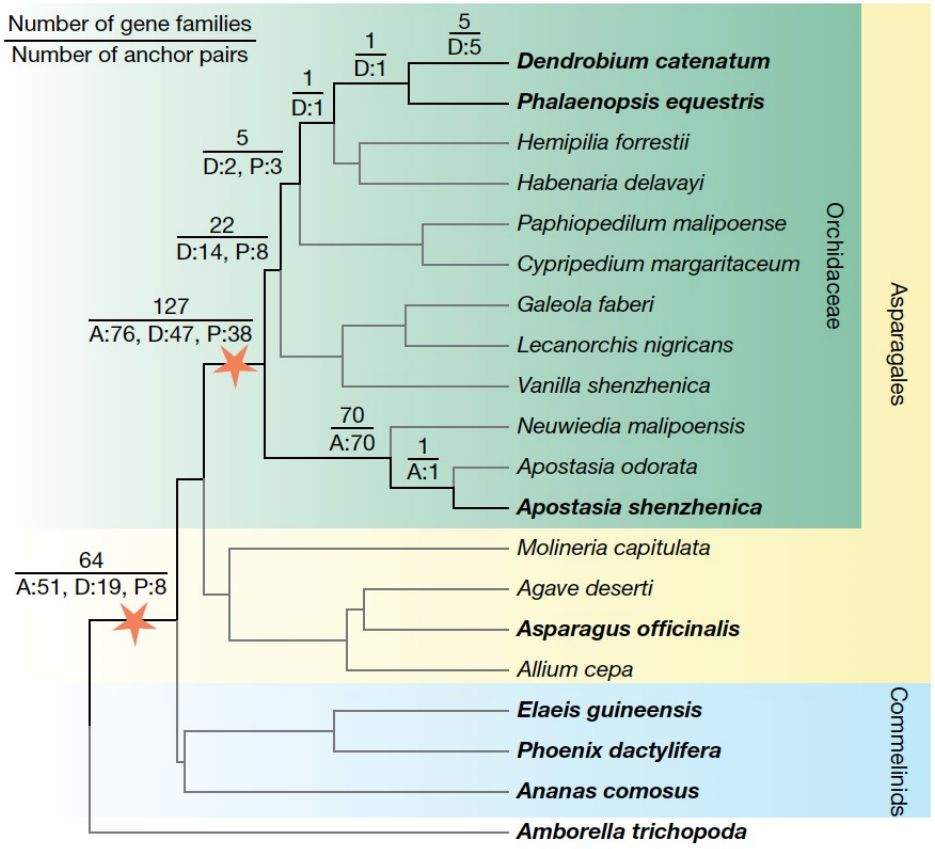
COMPARATIVE GENOMICS

Annotated genome allows inferring duplication events (in this case whole genome doublings) using co-linear comparisons of scaffolds



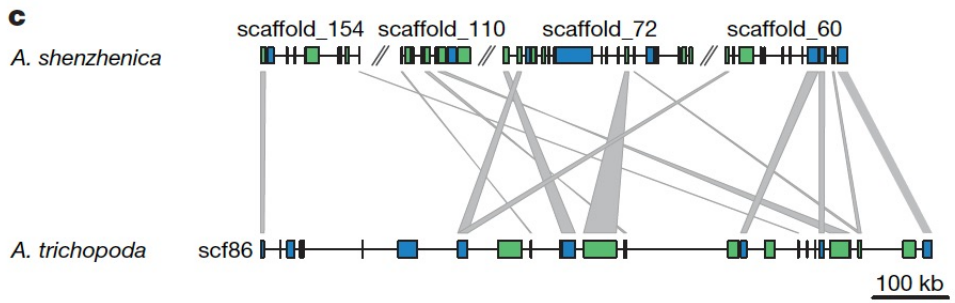
COMPARATIVE GENOMICS

Annotated genome allows inferring whole genome doublings along phylogenetic tree



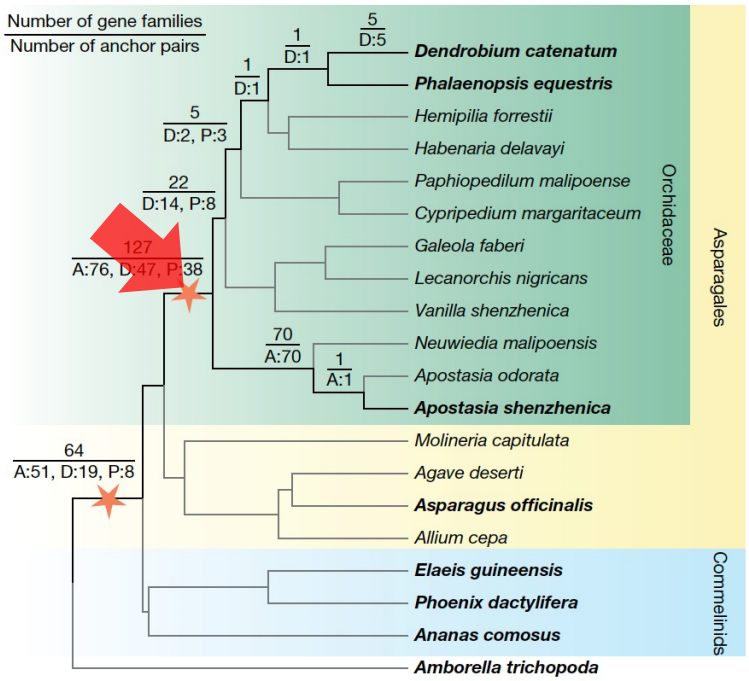
COMPARATIVE GENOMICS

Annotated genome allows comparing genome structures with other species



Synteny: conservation of blocks of genes/DNA between two or more sets of chromosomes belonging to different species.

COMPARATIVE GENOMICS



Annotated genome allows

studying emergence of key innovations