

<b>Manuscript Number:</b>	GIGA-D-18-00275R2	
<b>Full Title:</b>	The draft genomes of five agriculturally important African orphan crops	
<b>Article Type:</b>	Data Note	
<b>Funding Information:</b>	Shenzhen Municipal Government of China (JCYJ20150831201643396)	Dr. Yue Chang
	Shenzhen Municipal Government of China (JCYJ20150529150409546)	Dr. Shifeng Cheng
	Guangdong Provincial Key Laboratory of Genome Read and Write (2017B030301011)	Mr. Haorong Lu
<b>Abstract:</b>	<p><b>Background:</b> Continuous growth of the world population is expected to double the worldwide demand for food by 2050. Eighty-eight percent of countries current face a serious burden of malnutrition, especially in Africa and South and South-East Asia. About 95% of the food energy needs of humans are fulfilled by just 30 species, of which wheat, maize and rice provide the majority of calories. Therefore, to diversify and stabilize global food supply, enhance agricultural productivity and tackle malnutrition, greater use of neglected or underutilized local plants (so-called 'orphan crops', but also including a few plants of special significance to agriculture, agroforestry and nutrition) could be a partial solution.</p> <p><b>Results:</b> Here, we present draft genome information from five agriculturally, biologically, medicinally and economically important underutilized plants native to Africa; <i>Vigna subterranea</i>, <i>Lablab purpureus</i>, <i>Faidherbia albida</i>, <i>Sclerocarya birrea</i>, and <i>Moringa oleifera</i>. Assembled genomes range in size from 217 to 654 Mb. In <i>V. subterranea</i>, <i>L. purpureus</i>, <i>F. albida</i>, <i>S. birrea</i> and <i>M. oleifera</i> we have predicted 31707, 20946, 28979, 18937, 18451 protein-coding genes, respectively. By further analysing the expansion and contraction of selected gene families, we have characterized root nodule symbiosis genes, transcription factors and starch biosynthesis-related genes in these genomes.</p> <p><b>Conclusions:</b> These genome data will be useful to identify and characterize agronomically important genes and understand their modes of action, enabling genomics-based, evolutionary studies, and breeding strategies to design faster, more focused and predictable crop improvement programs.</p>	
<b>Corresponding Author:</b>	Xin Liu, Ph.D. BGI CHINA	
<b>Corresponding Author Secondary Information:</b>		
<b>Corresponding Author's Institution:</b>	BGI	
<b>Corresponding Author's Secondary Institution:</b>		
<b>First Author:</b>	Yue Chang	
<b>First Author Secondary Information:</b>		
<b>Order of Authors:</b>	Yue Chang	
	Huan Liu	
	Min Liu	
	Xuezhu Liao	
	Sunil Kumar Sahu	

	Yuan Fu
	Bo Song
	Shifeng Cheng
	Robert Kariba
	Samuel Muthemba
	Prasad S. Hendre
	Sean Mayes
	Wai Kuan Ho
	Presidor Kendabie
	Sibo Wang
	Linzhou Li
	Alice Muchugi
	Ramni Jamnadass
	Haorong Lu
	Shufeng Peng
	Allen Van Deynze
	Anthony Simons
	Howard Yana-Shapiro
	Xun Xu
	Huanming Yang
	Jian Wang
	Xin Liu, Ph.D.
<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	<p>Dear Dr. Scott,</p> <p>We are glad to resubmit the thoroughly revised and cleaned version of our manuscript entitled "The draft genomes of five agriculturally important African orphan crops", for possible publication in GigaScience as "Data Note".</p> <p>The corrections made and suggested by Dr. Lisa Martin were highly useful to significantly improve the quality of our manuscript. We have carefully implemented all the corrections suggested by her. Now we strongly believe that the revised manuscript is now ready for publication in GigaScience.</p> <p>We look forward to hearing from you at your earliest convenience.</p> <p>Yours sincerely, Xin Liu</p> <p>Reviewer comments:  Reviewer #1: The revisions in the last round satisfactorily addressed my concerns.  Reviewer #2: The revised version is well written and can be accepted for publication.  Response: We thank all the reviewers for their valuable suggestions in the earlier version of the manuscript, and for giving their kind acceptance to our manuscript.</p>
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No

<p><b>Experimental design and statistics</b></p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	<p>Yes</p>
<p><b>Resources</b></p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	<p>Yes</p>
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	<p>Yes</p>

[Click here to view linked References](#)

Chang et al.

Orphan crop genomes

# **The draft genomes of five agriculturally important African orphan crops**

Yue Chang<sup>1,2#</sup>, Huan Liu<sup>1,2,8#</sup>, Min Liu<sup>1,2,8#</sup>, Xuezhu Liao<sup>1,2,8</sup>, Sunil Kumar Sahu<sup>1,2,8</sup>, Yuan Fu<sup>1,2</sup>, Bo Song<sup>1,2</sup>, Shifeng Cheng<sup>1,2</sup>, Robert Kariba<sup>3</sup>, Samuel Muthemba<sup>3</sup>, Prasad S. Hendre<sup>3</sup>, Sean Mayes<sup>5,6,7</sup>, Wai Kuan Ho<sup>6,7</sup>, Presidor Kendabie<sup>5</sup>, Sibongile Wang<sup>1,2</sup>, Linzhou Li<sup>1,2</sup>, Alice Muchugi<sup>3</sup>, Ramni Jamnadas<sup>3</sup>, Haorong Lu<sup>1,2</sup>, Shufeng Peng<sup>1,2</sup>, Allen Van Deynze<sup>3,4</sup>, Anthony Simons<sup>3</sup>, Howard Yana-Shapiro<sup>3,4</sup>, Xun Xu<sup>1,2</sup>, Huanming Yang<sup>1,2</sup>, Jian Wang<sup>1,2</sup> and Xin Liu<sup>1,2,8,9\*</sup>

<sup>1</sup>BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China;

<sup>2</sup>China National GeneBank, BGI-Shenzhen, Jinsha Road, Shenzhen 518120, China;

<sup>3</sup>African Orphan Crops Consortium, World Agroforestry Centre (ICRAF), Nairobi, Kenya;

<sup>4</sup>University of California, 1 Shields Ave, Davis, CA 95616, USA;

<sup>5</sup>Plant and Crop Sciences, Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire LE12 5RD, UK;

<sup>6</sup>Biosciences, University of Nottingham Malaysia Campus, Jalan Broga 43500 Semenyih, Selangor, Malaysia;

<sup>7</sup>Crops For the Future, Jalan Broga, 43500 Semenyih, Selangor, Malaysia;

<sup>8</sup>State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China;

<sup>9</sup>BGI-Fuyang, BGI-Shenzhen, Fuyang 236009, China

22 \*Correspondence address. Xin Liu, BGI-Shenzhen, Beishan Industrial Zone, Yantian

23 District, Shenzhen 518083, China;

24 Tel: +86 18025460332; E-mail: [liuxin@genomics.cn](mailto:liuxin@genomics.cn)

25 ORCID:

26 Xin Liu: 0000-0003-3256-2940;

27 Yue Chang: 0000-0003-3909-0931;

28 Huan Liu: 0000-0002-6902-9931;

29 Liu Min: 0000-0002-8876-7534;

30 Sunil Kumar Sahu: 0000-0002-4742-9870

31 #Equal contribution

## 33 ABSTRACT

34 **Background:** Continuous growth of the world population is expected to double the  
35 worldwide demand for food by 2050. Eighty-eight percent of countries current face a  
36 serious burden of malnutrition, especially in Africa and South and South-East Asia.  
37 About 95% of the food energy needs of humans are fulfilled by just 30 species, of which  
38 wheat, maize and rice provide the majority of calories. Therefore, to diversify and  
39 stabilize global food supply, enhance agricultural productivity and tackle malnutrition,  
40 greater use of neglected or underutilized local plants (so-called ‘orphan crops’, but also  
41 including a few plants of special significance to agriculture, agroforestry and nutrition)

1 42 could be a partial solution.  
2

3 43 **Results:** Here, we present draft genome information from five agriculturally,  
4  
5  
6 44 biologically, medicinally and economically important underutilized plants native to  
7  
8  
9 45 Africa; *Vigna subterranea*, *Lablab purpureus*, *Faidherbia albida*, *Sclerocarya birrea*,  
10  
11 46 and *Moringa oleifera*. Assembled genomes range in size from 217 to 654 Mb. In *V.*  
12  
13 47 *subterranea*, *L. purpureus*, *F. albida*, *S. birrea* and *M. oleifera* we have predicted 31707,  
14  
15 48 20946, 28979, 18937, 18451 protein-coding genes, respectively. By further analysing  
16  
17 49 the expansion and contraction of selected gene families, we have characterized root  
18  
19  
20 50 nodule symbiosis genes, transcription factors and starch biosynthesis-related genes in  
21  
22  
23 51 these genomes.  
24  
25  
26

27 52 **Conclusions:** These genome data will be useful to identify and characterize  
28  
29  
30 53 agronomically important genes and understand their modes of action, enabling  
31  
32  
33 54 genomics-based, evolutionary studies, and breeding strategies to design faster, more  
34  
35  
36 55 focused and predictable crop improvement programs.  
37

38 56  
39  
40 57 **Keywords:** Orphan crops, food security, whole-genome sequencing, transcriptome, root  
41  
42  
43 58 nodule symbiosis, transcription factor.  
44  
45

## 46 59 47 48 60 **Background**

49  
50  
51 61 The world's population is expected to reach 9.8 billion people by 2050. Ensuring a  
52  
53  
54 62 sustainable food supply to meet the energy and nutritional needs of the expanding  
55  
56  
57 63 population is one of the greatest global challenges [1]. Approximately 88% of countries  
58  
59  
60 64 currently face a serious burden of malnutrition [2]. To overcome this burgeoning food  
61  
62  
63  
64  
65

1 65 and nutritional challenge, the use of potential crop plants (both model and non-model)  
2  
3 66 appears to be a better choice. Throughout history, humans have relied on an astonishing  
4  
5  
6 67 variety of plants for energy and nutrition: from 390,000 known plant species, around  
7  
8  
9 68 5,000–7,000 plant species have been cultivated or collected for food [1, 2]. However,  
10  
11  
12 69 in the present century, fewer than 150 species are commercially cultivated for food  
13  
14  
15 70 purposes, and just 30 species provide 95% of human food energy needs. More than half  
16  
17  
18 71 of the protein and calories we obtain from plants are acquired from just three  
19  
20  
21 72 ‘megacrops’: rice, wheat and maize [3]. This narrow range of dietary diversity is partly  
22  
23  
24 73 a result of decades of intensive research, focused on just a few species, which has  
25  
26  
27 74 successfully led to the production of high-yielding varieties of these major crops,  
28  
29  
30 75 usually cultivated under high-input agricultural systems. However, in some regions, we  
31  
32  
33 76 are now witnessing a drastic decrease in their yields and the question has been raised  
34  
35  
36 77 as to whether rice and wheat (in particular) are currently making enough breeding  
37  
38  
39 78 progress to meet the challenge. All three megacrops are high-energy carbohydrate  
40  
41  
42 79 sources, but are limited in protein content. Even if these crops can meet the energy  
43  
44  
45 80 requirement of the increasing world population, they cannot meet the nutritional  
46  
47  
48 81 requirement for active health by themselves [2].

47 82 To diversify the global food supply, enhance agricultural productivity and tackle  
48  
49  
50 83 malnutrition, it is necessary to diversify and focus more on crop plants that are utilized  
51  
52  
53 84 in rural societies as a local source of nutrition and sustenance, but have so far received  
54  
55  
56 85 little attention for crop improvement. These landraces tend to be locally adapted, and  
57  
58  
59 86 can often provide a rich source of nutrition, yet they have largely been ignored by  
60  
61  
62  
63  
64  
65

1 87 modern interventions. The goal of the African Orphan Crops Consortium [4] (AOCC),  
2  
3 88 an international public–private partnership, is to sequence, assemble and annotate the  
4  
5  
6 89 genomes of 101 plants that contribute to traditional African food supplies by 2020.  
7  
8  
9 90 These neglected or orphan plants have been seldom studied by scientists, but are of  
10  
11  
12 91 major importance in many African countries. They are usually grown by smallholder  
13  
14  
15 92 farmers, either for consumption or local sale, and are a major food source for  
16  
17  
18 93 600 million rural Africans [5, 6]. In this study, we sequenced and assembled draft  
19  
20  
21 94 genomes of five African orphan plant species (Figure 1), which are highly important to  
22  
23 95 augment food and nutritional security in Africa.

24  
25 96 *Vigna subterranea* (Bambara groundnut; NCBI: txid115715) belonging to the  
26  
27  
28 97 Fabaceae family, is a leguminaceous plant species that originated in West Africa, and  
29  
30  
31 98 is cultivated in sub-Saharan areas, particularly Nigeria [7, 8]. With good nitrogen-fixing  
32  
33  
34 99 ability and drought tolerance, on average the seeds contain 63% carbohydrate, 19%  
35  
36  
37 100 protein and 6.5% fat, thereby making bambara groundnut a complete food.  
38  
39  
40 101 Approximately 165,000 tons of this species is produced in Africa each year, but yields  
41  
42  
43 102 are low because efforts to improve Bambara have been neglected for many years [9].  
44  
45  
46 103 The genomes of mung bean and adzuki bean, which also belong to the *Vigna* genus,  
47  
48 104 have been published [10, 11].

49  
50 105 *Moringa oleifera* (Moringa; NCBI: txid3735) is a highly nutritious, fast growing  
51  
52  
53 106 and drought-tolerant tree, which is indigenous to northern India, Pakistan and Nepal  
54  
55  
56 107 [12]. Presently, this species is ubiquitously distributed throughout tropical and  
57  
58  
59 108 subtropical countries, and in particular covers the major agro-ecological region in



1 109 Nigeria. The leaves are rich in protein, minerals, beta-carotene and antioxidant  
2  
3 110 compounds, which are generally used as nutrition supplements and in traditional  
4  
5 111 medicine. The seeds are used to extract oil, and seed powder can be used for water  
6  
7  
8 112 purification [13, 14]. There are varying reports of *Moringa* production: India is the  
9  
10 113 largest producer of *Moringa* with an annual production of 1.1–1.3 million tonnes of  
11  
12 114 tender fruits from an area of 38,000 ha. In Limpompo province, *Moringa* is cultivated  
13  
14 115 in relatively small areas (0.25–1 ha), with seed yields of 50–100 kg/ha<sup>-1</sup> [15]. Prior to  
15  
16  
17 116 this study, a draft genome of *Moringa oleifera* from Yunnan (China) was reported [16],  
18  
19  
20 117 which estimated a similar genome assembly size and gene numbers to our version.  
21  
22

23  
24  
25 118 *Lablab purpureus* (Dolichos bean or hyacinth bean; NCBI: txid35936), a member  
26  
27 119 of the Fabaceae family, is one of the most ancient (>3500 years) domesticated and  
28  
29 120 multipurpose legume species, which is used as an intercrop in livestock systems.  
30  
31 121 Although it has large agromorphological diversity in South Asia, its origin appears to  
32  
33 122 be African [17]. It is rich in protein, has good nitrogen-fixing ability, and is highly  
34  
35 123 adaptable to diverse environmental conditions [18]. Limited production data are  
36  
37 124 available, suggesting that yields are low. In south-western parts of Bangladesh, *Lablab*  
38  
39 125 is reported to have a total production area of approximately 48,000 ha [17]. In other  
40  
41 126 areas, it has a similarly relatively low production area; for example, Kenya, approx.  
42  
43 127 10,000 ha [19], and Karnataka, India, 79,000 ha [20].  
44  
45  
46  
47  
48  
49  
50  
51  
52

53 128 *Faidherbia albida* (apple-ring acacia; NCBI: txid138055) is the only tree species  
54  
55 129 in the *Faidherbia* genus (Fabaceae). Its distinctive key features, such as reverse  
56  
57 130 phenology (leaves grow in the long dry season and shed during the rainy season) and  
58  
59  
60  
61  
62  
63  
64  
65

1 131 nitrogen-fixing ability, mean that *F. albida* has been planted as a key agroforestry  
2  
3 132 species in traditional African farming systems for hundreds of years [21]. It originated  
4  
5  
6 133 in the Sahara or eastern and southern Africa, then spread across semi-arid tropical  
7  
8  
9 134 Africa, and later to the Middle East and Arabia. Estimates suggest that, during the last  
10  
11 135 decade, the tree was cultivated over an area of 300,000 ha [22]. Average pod production  
12  
13 136 ranges from 6–135 kg per tree per year in the Sudanian zone. In Mana Pools, Zimbabwe,  
14  
15  
16  
17 137 two trees averaged 161 kg per tree in one year [23]. This yield per unit area is about  
18  
19  
20 138 2,000–3,000kg/ha, assuming a density of ~20 mature trees per hectare [24].  
21

22 139 *Sclerocarya birrea* (Marula; NCBI: txid289766) belongs to the Anacardiaceae  
23  
24  
25 140 family, and is a traditional fruit tree found in southern Africa – mostly south of the  
26  
27  
28 141 Zambezi river [25]. Fruits are eaten fresh, or are used to produce juices and wine, which  
29  
30  
31 142 has substantial socioeconomic and commercialization importance. The seeds of the  
32  
33 143 fruits are rich in nutrition and oil content (56%), and are often consumed raw. It is  
34  
35  
36 144 estimated that the total value of the commercial marula trade is worth USD \$160,000  
37  
38  
39 145 per year to rural communities [26], with values per tree ranging from 315 kg (17,500  
40  
41  
42 146 fruits) to 1,643 kg (91,300 fruits) [26, 27]. A survey in north-central Namibia showed  
43  
44  
45 147 that, on average, there are 5.33 farms per household, with a total of 13,278 fruiting trees.  
46

47 148 Considering the limited systematic efforts to improve the breeding of these  
48  
49  
50 149 understudied tropical crops so far, making their genomic data available will provide  
51  
52  
53 150 much-needed impetus to conduct basic and applied translational research to improve  
54  
55  
56 151 and develop them as important, sustainably cultivated food crops. These efforts will be  
57  
58  
59 152 vital for directly or indirectly improving nutrition for the increasing urban populations  
60  
61  
62  
63  
64  
65

153 in the regions where these crops are grown.

154

## 155 **Data description**

### 156 **Sample collection, library construction, and sequencing**

157 Genomic DNA was extracted either from a tree (*F. albida*, *M. oleifera*) or from nursery  
158 plantlets (*V. subtarranea*, *L. purpureus*, *S. birrea*) grown at the World AgroForestry  
159 Center campus in Kenya using a modified CTAB method [28].

160       Extracted DNA was used to construct paired-end libraries (insert size ranging from  
161 170–800 bp) and mate-pair libraries (insert size >2 kb) following Illumina (San Diego,  
162 USA) protocols. Subsequently, sequencing was performed on a HiSeq 2000 platform  
163 (Illumina, San Diego, CA, USA) using a shotgun sequencing strategy to generate more  
164 than 100 Gb raw data for each species (see Additional file 1: Table S1). Data were  
165 filtered using SOAPfilter (v2.2) [29] as follows: (1) small insert size reads were  
166 discarded; (2) PCR duplicates and adapter contamination were discarded; (3) reads with  
167  $\geq 30\%$  low quality bases (quality score  $\leq 15$ ) were removed; (4) bases with low quality  
168 were trimmed from each end of the reads; (5) reads with  $\geq 10\%$  uncalled (“N”) bases  
169 were removed. At the end, more than 100× high-quality reads were obtained for each  
170 species, according to their estimated genome size (see Additional file 1: Table S1).

171       RNA for transcriptome sequencing was extracted from different tissues of *V.*  
172 *subterranea*, *L. purpureus*, *F. albida*, and *M. oleifera*. The RNA was extracted using the

1 173 PureLink RNA Mini Kit (Thermo Fisher Scientific, Carlsbad, CA, USA) according to  
2  
3 174 the manufacturer's instructions. For each sample, RNA libraries were constructed by  
4  
5  
6 175 following the TruSeq RNA Sample Preparation Kit (Illumina, San Diego, CA, USA)  
7  
8  
9 176 manual, and were then sequenced on the Illumina HiSeq 2500 platform (paired-end,  
10  
11 177 100-bp reads), generating ~36 Gb of sequence data for each species. Data were then  
12  
13  
14 178 filtered using a similar method to that used in DNA filtration, with a slight modification:  
15  
16  
17 179 (1) reads with  $\geq 10\%$  low quality bases (quality score  $\leq 15$ ) were removed; and (2) reads  
18  
19  
20 180 with  $\geq 5\%$  uncalled ("N") bases were removed (see Additional file 1: Table S2). All the  
21  
22  
23 181 transcriptome data from different tissues were compiled, and the combined version was  
24  
25  
26 182 used to check the completeness of the whole genome sequence assembly.

27 183

28 184 Evaluation of genome size

29  
30  
31 185 Clean reads of the paired-end libraries were used to estimate genome sizes (insert size  
32  
33  
34 186 250 bp and 500 bp). k-mer frequency distribution analysis was performed using the  
35  
36  
37 187 following formula:  
38  
39

$$40 \quad 41 \quad 42 \quad 43 \quad 44 \quad 45 \quad 46 \quad 47 \quad 48 \quad 49 \quad 50 \quad 51 \quad 52 \quad 53 \quad 54 \quad 55 \quad 56 \quad 57 \quad 58 \quad 59 \quad 60 \quad 61 \quad 62 \quad 63 \quad 64 \quad 65$$
$$188 \quad Gen = Num * (Len - 17 + 1) / K\_Dep$$

189 Where: *Num* represents the read number of reads used. *Len* represents the read  
190 length, *K* represents the k-mer length, and *K\_Dep* refers to where the main peak is  
191 located in the distribution curve [30].

192 k-mer distributions of *F. albida*, *S. birrea*, and *M. oleifera* showed two distinct  
193 peaks (see Additional file 1: Figure S1), where the second peak was confirmed as the  
194 main one for each of the species. The genome sizes of *V. subterranea*, *L. purpureus*, *F.*

195 *albida*, *S. birrea* and *M. oleifera* were predicted as 550, 423, 661, 356 and 278 Mb,  
196 respectively (see Additional file 1: Table S3).

197

#### 198 *De novo* genome assembly

199 For *de novo* genome assembly, SOAPdenovo2 (SOAPdenovo2, RRID:SCR\_014986)  
200 [29] was used for constructing contigs, followed by scaffolding, and finally gap filling.  
201 To build contigs, libraries ranging from 170–800 bp were used to construct de Bruijn  
202 graphs with the parameters “pregraph -d 2 -K 55”, and contigs were subsequently  
203 formed with the parameters “contig -g -D 1” to delete links with low coverage. In the  
204 scaffolding step, paired-end and mate-pair information was used to order the contigs  
205 with parameters “scaff -g -F” and “map -g -k 55”. Finally, to fill the gaps within  
206 scaffolds, GapCloser version 1.12 (GapCloser, RRID:SCR\_015026) [29] was used with  
207 the parameters “-l 150 -t 32” using the pair-end libraries. Finally, total assembled  
208 lengths of 535.05, 395.47, 653.73, 330.98, and 216.76 Mb were obtained for *V.*  
209 *subterranea*, *L. purpureus*, *F. albida*, *S. birrea* and *M. oleifera* genomes, respectively  
210 (Table 1). This accounted for approximately 97.3%, 93.5%, 98.9%, 92.9% and 77.9%  
211 of their respective estimated genome sizes.

212

#### 213 Genome evaluation

214 Genome assembly completeness was assessed with BUSCO (Benchmarking Universal  
215 Single-Copy Orthologs) version 3.0.1, (BUSCO, RRID:SCR\_015008) [31]. From the  
216 1,440 core embryophyta genes, 1,326 (92.1%), 1,341 (93.2%), 1,315 (91.3%), 1,384

1 217 (96.1%) and 1,297 (90.1%) were identified in the *V. subterranea*, *L. purpureus*, *F.*  
2  
3 218 *albida*, *S. birrea* and *M. oleifera* assemblies, respectively, with 1,244 (86.4%), 1,258  
4  
5  
6 219 (87.4%), 1,231 (85.5%), 1,352 (93.9%) and 1,278 (88.8%) genes, respectively, being  
7  
8  
9 220 complete (Table 2).

10  
11 221 To evaluate the completeness of genes in the assemblies, unigenes were generated  
12  
13 222 from the transcript data of each species using Bridger software with the parameters “–  
14  
15 223 kmer\_length 25 –min\_kmer\_coverage 2” [32], and then aligned to the corresponding  
16  
17 224 assembly using BLAT (BLAT, RRID:SCR\_011919) [33]. The results indicated that  
18  
19  
20 225 each of the assemblies covered about 90% of the expressed unigenes, suggesting that  
21  
22  
23 226 the assembled genomes contained a high percentage of expressed genes (Table 3).

24  
25  
26  
27 227 To confirm the accuracy of the assemblies, some of the paired-end libraries were  
28  
29  
30 228 mapped to the genome assemblies, and the sequencing coverage was calculated using  
31  
32  
33 229 SOAPaligner, version 2.21 (SOAPaligner/soap2 , RRID:SCR\_005503) [34].  
34  
35  
36 230 Sequencing coverage showed that >99% of the bases had a sequencing depth of more  
37  
38  
39 231 than 10×, and confirmed the accuracy at the base level (see Additional file 1: Figure  
40  
41  
42 232 S2). GC content and average depth were also calculated with 10 kb non-overlapping  
43  
44  
45 233 windows. The distribution of GC content indicated a relatively pure single genome  
46  
47  
48 234 without contamination or GC bias (see Additional file 1: Figure S3). The GC content  
49  
50  
51 235 of each sequenced genome was also compared with that of a related species. As  
52  
53 236 expected, close peak positions showed that the related species were similar in GC  
54  
55  
56 237 content (see Additional file 1: Figure S4).

57  
58 238  
59  
60  
61  
62  
63  
64  
65

1 239 Repeat annotation  
2  
3 240 Repetitive sequences were identified using RepeatMasker (version 4-0-5) [35], with a  
4  
5  
6 241 combined Rebase and a custom library obtained through careful self-training. The  
7  
8  
9 242 custom library comprised three parts: MITEs (miniature inverted repeat transposable  
10  
11 243 elements), LTRs (long terminal repeats), and an extensive library that was constructed  
12  
13  
14 244 as follows. First, the annotated MITE library was created using MITE-hunter [36] with  
15  
16  
17 245 default parameters. Then, a library of LTR elements with lengths of 1.5–25 kb, and two  
18  
19  
20 246 libraries of terminal repeats ranging from 100–6000 bp with  $\geq 85\%$  similarity were  
21  
22  
23 247 constructed using LTRharvest [37] integrated in Genometools (version 1.5.8) [38] with  
24  
25 248 parameters “–minlenltr 100, –maxlenltr 6000, –mindistltr 1500, –maxdistltr 25000, –  
26  
27  
28 249 mintsd 5, –maxtsd 5, –similar 90, –vic 10”. Subsequently, we used several strategies to  
29  
30  
31 250 filter the candidates, i.e. 1) presence of intact poly purine tracts or primer binding sites  
32  
33  
34 251 [39] using the eukaryotic tRNA library [40]; 2) removal of contamination from local  
35  
36  
37 252 gene clusters and tandem local repeats by inspecting 50 bases of the upstream and  
38  
39  
40 253 downstream LTR flanks using MUSCLE (MUSCLE, RRID:SCR\_011812) [41] for a  
41  
42  
43 254 minimum of 60% identity; and 3) removal of nested LTR candidates from other types  
44  
45  
46 255 of the elements. Exemplars for the LTR library were extracted from the filtered  
47  
48  
49 256 candidates using a cutoff of 80% identity in 90% of the sequence. Regions of the  
50  
51  
52 257 genome annotated as LTRs and MITEs were masked, and then put into RepeatModeler  
53  
54  
55 258 (version 1-0-8; RepeatModeler, RRID:SCR\_015027) to predict other repetitive  
56  
57  
58 259 sequences for the extensive library. Finally, the MITE, LTR and extensive libraries were  
59  
60  
61 260 integrated into the custom library, which was combined with the Rebase library and

261 taken as an input for RepeatMasker to identify and classify genome-wide repetitive  
262 elements. The pipeline identified 205,189,285 (38.35% of the genome length),  
263 147,050,327 (37.18%), 358,653,534 (54.86%), 149,551,125 (45.18%), and 87,944,150  
264 (40.57%) bases of non-redundant repetitive sequences in *V. subterranea*, *L. purpureus*,  
265 *F. albida*, *S. birrea* and *M. oleifera*, respectively. LTR elements were predominant,  
266 taking up 19.8%, 23.8%, 44.6%, 38.8%, 22.7% of each genome, respectively (Table 4).

267

268 Gene prediction

269 Repetitive regions of the genome were masked before gene prediction. Structures of  
270 protein-coding genes were predicted using the MAKER-P pipeline (version 2.31) [42]  
271 based on RNA, homologous and *de novo* prediction evidence. For RNA evidence, the  
272 clean transcriptome reads were assembled into inchworms using Trinity (version 2.0.6)  
273 [43], and then provided to MAKER-P as expressed sequence tag evidence. For  
274 homologous comparison, protein sequences from the model plant *Arabidopsis thaliana*,  
275 and related species of each sequenced species, were downloaded and provided as  
276 protein evidence. Related species used for homologous evidence were *Arachis*  
277 *duranensis*, *A. ipaensis*, *Glycine max*, *Lotus japonicus*, *Medicago truncatula*, and *Vigna*  
278 *angularis* for *V. subterranea*; *A. duranensis*, *Cajanus cajan*, *G. max*, *M. truncatula*,  
279 *Phaseolus vulgaris*, and *V. angularis* for *L. purpureus*; *C. cajan*, *V. angularis*, *L.*  
280 *japonicus*, *P. vulgaris*, *M. truncatula*, and *G. max* for *F. albida*; *Actinidia chinensis*, and  
281 *Musa acuminata* for *S. birrea*; and *G. max*, *Oryza sativa*, *Populus trichocarpa*, and  
282 *Sorghum bicolor* for *M. oleifera*.



1 283 For *de novo* prediction evidence, a series of training sets was made to optimize  
2  
3 284 different *ab initio* gene predictors. Initially, a set of transcripts was generated by a  
4  
5  
6 285 genome-guided approach using Trinity with the parameters “--full\_cleanup, --  
7  
8  
9 286 jaccard\_clip, --genome\_guided\_max\_intron 10000, --min\_contig\_length 200”. The  
10  
11 287 transcripts were then mapped back to the genome using PASA (version 2.0.2) [44] and  
12  
13 288 a set of gene models with real gene characteristics (e.g., size and number of  
14  
15 289 exons/introns per gene, features of splicing sites) was generated. Complete gene models  
16  
17 290 were picked for training Augustus [45]. Genemark-ES (version 4.21) [46] was self-  
18  
19 291 trained with default parameters. The first round of MAKER-P was run based on the  
20  
21 292 evidence as above, with default parameters except “est2genome” and “protein2genome”  
22  
23 293 being set to “1”, yielding only RNA and protein-supported gene models. SNAP [47]  
24  
25 294 was then trained with these gene models. Default parameters were used to run the  
26  
27 295 second and final rounds of MAKER-P, producing the final gene models.

28 296 The number of protein-coding genes identified in each species was 31,707 in *V.*  
29  
30 297 *subterranea*, 20,946 in *L. purpureus*, 28,979 in *F. albida*, 18,937 in *S. birrea*, and  
31  
32 298 18,451 in *M. oleifera*. Compared to the other sequenced species in the same genus [10,  
33  
34 299 11], *V. subterranea* has a more genes than mung bean (22,427) but less than adzuki bean  
35  
36 300 (34,183). Various gene structure parameters were compared to the related species of  
37  
38 301 each sequenced genome, as summarized in Table 5 and Additional file 1: Figure S5.  
39  
40 302 BUSCO evaluation showed that at least 85% of 1,440 core genes could be identified  
41  
42 303 across all the species, suggesting an acceptable quality of gene annotation for the five  
43  
44 304 sequenced genomes (see Additional file 1: Table S4).  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 305 Non-coding RNA genes in the sequenced genomes were also annotated. Using  
2  
3 306 BLAST, ribosomal RNA (rRNA) genes were searched against the *A. thaliana* rRNA  
4  
5  
6 307 database, or by searching for microRNAs (miRNA) and small nuclear RNA (snRNA)  
7  
8  
9 308 against the Rfam database (Rfam, RRID:SCR\_004276; release 12.0) [48]. tRNAscan-  
10  
11  
12 309 SE (tRNAscan-SE, RRID:SCR\_010835) was also used to scan for tRNAs [49]. The  
13  
14 310 results are summarized in Table 6.

15  
16  
17 31118  
19  
20 312 Functional annotation of protein-coding genes21  
22 313 Functional annotation of protein-coding genes was based on sequence similarity and  
23  
24  
25 314 domain conservation by aligning predicted amino acid sequences to public databases.26  
27  
28 315 Protein-coding genes were first searched against protein sequence databases for best29  
30  
31 316 matches, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG,  
32  
33  
34 317 RRID:SCR\_012773) [50], the National Center for Biotechnology Information (NCBI)35  
36 318 non-redundant (NR) and COG databases [51], SwissProt and TrEMBL [52] using37  
38  
39 319 BLASTP with an E-value cut-off of 1e-5. Then, InterProScan 55.0 (InterProScan,40  
41  
42 320 RRID:SCR\_005829) [53] was used to identify domains and motifs based on Pfam43  
44  
45 321 (Pfam, RRID:SCR\_004726) [54], SMART (SMART, RRID:SCR\_005026) [55],46  
47  
48 322 PANTHER (PANTHER, RRID:SCR\_004869) [56], PRINTS (PRINTS,49  
50  
51 323 RRID:SCR\_003412) [57], and ProDom (ProDom, RRID:SCR\_006969) [58]. In total,52  
53 324 98.0%, 98.2%, 93.6%, 98.1% and 98.8% of genes in *V. subterranea*, *L. purpureus*, *F.*54  
55  
56 325 *albida*, *S. birrea*, and *M. oleifera*, respectively, were functionally annotated. Of the57  
58  
59 326 unannotated genes, 400, 305, 1,514, 293 and 172 were specific to *V. subterranea*, *L.*

327 *purpureus*, *F. albida*, *S. birrea*, and *M. oleifera*, respectively (Table 7).

328

329 Gene family construction

330 Protein and nucleotide sequences from the five sequenced species and nine other

331 species (*A. thaliana*, *Carica papaya*, *Citrus sinensis*, *G. max*, *M. truncatula*, *O. sativa*,

332 *P. vulgaris*, *S. bicolor*, and *Theobroma cacao*) were retrieved to construct gene families

333 using OrthoMCL software [59] based on an all-versus-all BLASTP alignments with an

334 E-value cutoff of 1e-5. A total of 609, 104, 499, 205 and 150 gene families were found

335 specific to *V. subterranea*, *L. purpureus*, *F. albida*, *S. birrea*, and *M. oleifera*,

336 respectively (see Additional file 1: Table S5).

337 Furthermore, the 10,103 gene families of *V. subterranea*, *L. purpureus*, *F. albida*,

338 *M. truncatula*, and *G. max* were clustered (Figure 2A). There were 1,105 orthologous

339 families shared by the four Papilionoideae species, while 808 gene families containing

340 1,966 genes were specific to *F. albida*, 281 gene families containing 538 genes were

341 specific to *L. purpureus*, and 789 gene families containing 3,118 genes were specific to

342 *V. subterranea*.

343 Moreover, 8,184 gene families of *S. birrea*, *M. oleifera*, *C. papaya*, *C. sinensis* and

344 *T. cacao* were clustered (Figure 2B), of which 365 gene families containing 798 genes

345 were specific to *M. oleifera*, and 362 gene families containing 796 genes were specific

346 to *S. birrea*. KEGG pathway enrichment analysis of paralog genes was also conducted

347 (Additional file 1: Table S6, S7). Functional annotation revealed that, in *V. subterranea*,

348 these paralogs corresponded mainly with carbon fixation, zeatin biosynthesis, and

1 349 glyoxylate and dicarboxylate metabolism. However, for *L. purpureus*, the fatty acid  
2  
3 350 elongation pathway was enriched, while in *F. albida*, pathways corresponding to plant–  
4  
5  
6 351 pathogen interactions and cyanoamino acid metabolism were enriched. In *S. birrea*,  
7  
8  
9 352 enrichment occurred in plant–pathogen interaction, starch and sucrose metabolism, and  
10  
11  
12 353 fatty acid biosynthesis pathways. In *M. oleifera*, pathways related to fatty acid and  
13  
14  
15 354 diterpenoid biosynthesis, and cyanoamino acid metabolism were enriched. Using Gene  
16  
17  
18 355 Ontology (GO) analysis, paralog genes in *V. subterranea*, *L. purpureus*, *F. albida*, *M.*  
19  
20 356 *oleifera*, and *S. birrea* were enriched in ion binding, metabolic processes, disease  
21  
22  
23 357 resistance, cell components, and biological processes, respectively.  
24

25 358

26  
27  
28 359 Phylogenetic analysis and estimation of divergence time

29  
30  
31 360 We identified 141 single-copy genes in the 14 species used for the above analysis, and  
32  
33  
34 361 subsequently used them to build a phylogenetic tree. Coding DNA sequence alignments  
35  
36  
37 362 of each single-copy family were generated following protein sequence alignment with  
38  
39  
40 363 MUSCLE (MUSCLE, RRID:SCR\_011812) [41]. The aligned coding DNA sequences  
41  
42  
43 364 of each species were then concatenated to a supergene sequence. The phylogenetic tree  
44  
45  
46 365 was constructed with PhyML-3.0 (PhyML, RRID:SCR\_014629) [60], with the  
47  
48  
49 366 HKY85+gamma substitution model on extracted four-fold degenerate sites. Divergence  
50  
51  
52 367 time was calculated using the Bayesian relaxed molecular clock method with  
53  
54  
55 368 MCMCTREE in PAML (PAML, RRID:SCR\_014932) [61], based on published  
56  
57  
58 369 calibration times (39–59 Mya between *M. truncatula* and the main branch of legumes,  
59  
60  
61 370 15–30 Mya between *G. max* and *P. vulgaris*, and 83–90 Mya between *T. cacao* and *A.*

1 371 *thaliana*) [11, 62].

2  
3 372 Based on the tree constructed using single-copy-family genes, the divergence time  
4  
5  
6 373 between *F. albida* and Papilionoideae was predicted to be 79.1 (70.0–87.0) Mya. This  
7  
8  
9 374 is a little different from a previous prediction of the origin of legumes based on two  
10  
11 375 gene markers (matk and rbcL) [63]. The divergence time between *M. oleifera* and *C.*  
12  
13 376 *papaya* was predicted to be 65.4 (59.2–71.1) Mya, and 67.9 (53.6–77.3) Mya between  
14  
15  
16  
17 377 *S. birrea* and *C. sinensis* (Figure 1).

18  
19  
20 378 Subsequently, to evaluate gene gain and loss, CAFE (CAFE, RRID:SCR\_005983)  
21  
22 379 [64] was employed to estimate the universal gene birth and death rate,  $\lambda$ , under a  
23  
24  
25 380 random birth and death model using the maximum likelihood method. Results for each  
26  
27  
28 381 branch of the phylogenetic tree were estimated and represented in Figure 1.

29  
30  
31 382 GO enrichment analysis was also conducted on gene pathways in expanded  
32  
33 383 families in the lineage of each sequenced species (Additional file 1: Table S8, S9).  
34  
35  
36 384 Terms related to energy and nutrient metabolism were commonly distributed in the  
37  
38  
39 385 enrichment output of *V. subterranea*, *L. purpureus*, *M. oleifera* and *S. birrea*; for  
40  
41  
42 386 example, proton-transporting two-sector ATPase complex, cyclase activity, nutrient  
43  
44  
45 387 reservoir activity and carbohydrate derivative binding.

46  
47 388 In *F. albida*, expanded gene families were related to signal transfer or regulation;  
48  
49  
50 389 e.g., signaling receptor activity, phosphatase regulator activity, and regulation of  
51  
52  
53 390 response to stimulus. Furthermore, the regulatory factors *GLABRA3*, *ENHANCER OF*  
54  
55  
56 391 *GLABRA 3*, *AUX1*, *LAX2*, and *LAX3* [65–67], which are related to the formation of root  
57  
58  
59 392 hairs and lateral roots, were identified in these families. As a traditional agroforestry

1 393 tree in Africa, *F. albida* was previously reported to have a root system architecture that  
2  
3 394 displays wide variation under different environmental factors (soil depth, nutrient  
4  
5  
6 395 amount, or water reservoirs) [68]. This suggests its adaptability to the complex  
7  
8  
9 396 environment, which requires signal transferring and regulation. The results obtained  
10  
11  
12 397 from the GO enrichment analysis were consistent with the biological characteristics of  
13  
14 398 *F. albida*.

15  
16  
17 399

#### 20 400 Mining of transcription factors

21  
22 401 Transcription factors (TFs) in the sequenced species were identified using protein  
23  
24  
25 402 sequences of plant TFs from the plant transcription factor database [69] by BLASTP  
26  
27  
28 403 search with an e-value cutoff of  $10E-10$ , a minimum identity of 40% and a minimum  
29  
30  
31 404 query coverage of 50%. About 59 TF families were revealed across the genes in *M.*  
32  
33  
34 405 *truncatula*, *G. max*, *P. vulgaris*, *C. papaya*, *C. sinensis*, and the five sequenced species  
35  
36 406 (see Additional file 2: Table S14). Among these TFs, bHLH, NAC, ERF, MYB-related,  
37  
38  
39 407 C2H2, MYB, WRKY, bZIP, FAR1, C3H, B3, G2-like, Trihelix, LBD, GRAS, M-type  
40  
41  
42 408 MADS, HD-ZIP, MIKC\_MADS, HSF, GATA were found in abundance (Figure 4).

43  
44  
45 409

#### 47 410 Identification of protein, starch, and fatty acid biosynthesis-related genes

48  
49  
50 411 Using the amino acid, starch and fatty acid synthesis genes in soybean [11, 70] as bait,  
51  
52  
53 412 we performed an ortholog search in *V. subterranea*, *L. purpureus*, *F. albida*, *S. birrea*,  
54  
55  
56 413 *M. oleifera*, *G. max*, *Triticum aestivum*, *Zea mays*, and *O. sativa* (Additional file 1:  
57  
58  
59 414 Tables S10–13). *V. subterranea* is a good source of resistant starch (RS) [71], which has

1 415 the potential to protect against diabetes and reduce the incidence of diarrhea and other  
2  
3 416 inflammatory bowel diseases [72]. High amylose levels can contribute to RS.  
4  
5  
6 417 Previously, studies have shown that deficiency in *SSIIIa* (soluble starch synthase gene)  
7  
8  
9 418 decreases amylopectin biosynthesis and increases amylose biosynthesis by a granule-  
10  
11  
12 419 bound starch synthase (GBSSI) encoded by the *Wx* gene in *O. sativa indica* [73]. Down-  
13  
14  
15 420 regulation of the soluble starch synthase *SSII*, and of *SBE*, leads to higher levels of RS  
16  
17  
18 421 in barley [74]. Interestingly, in *V. subterranea*, two out of four GBSSs underwent  
19  
20  
21 422 expansion, suggesting their vital role in controlling starch synthesis (Figure 5) at the  
22  
23  
24 423 transcriptional and post-transcriptional level. No expansion in GBSS was observed in  
25  
26  
27 424 the genomes of *L. purpureus*, *F. albida*, *S. birrea* or *M. oleifera*, and in *V. subterranea*,  
28  
29  
30 425 soluble starch synthase was not expanded. Therefore, we speculate that the expansion  
31  
32  
33 426 of GBSS might be why *V. subterranea* is rich in RS.

34 427 Similarly, differences in the copy numbers of choline kinase, a key factor in fatty  
35  
36  
37 428 acid synthesis and storage, were found between the four legumes (*V. subterranea*, 7; *F.*  
38  
39  
40 429 *albida*, 4; *L. purpureus*, 2; and *G. max*, 5) and between two orphan species (*S. birrea*,  
41  
42  
43 430 1, and *M. oleifera*, 3). Choline kinase is the first enzyme in the cytidine diphosphate-  
44  
45  
46 431 choline pathway, which is involved in lecithin biosynthesis [75, 76]. Based on these  
47  
48  
49 432 observations, we inferred that all the factors required to synthesize lecithin are present  
50  
51  
52 433 in *V. subterranea*. However, gene expression data remains lacking in terms of the GBSS  
53  
54  
55 434 and choline kinase genes in these the five species. More transcriptomic analysis and  
56  
57  
58 435 chemical tests are required to uncover the mechanisms of their nutrition metabolism.

59 436

## 437 Identification of the root nodule symbiosis pathway

438 Legumes (Fabaceae) are well known for their ability to fix nitrogen; an important trait  
439 to replenish nitrogen supplies in soil and agricultural systems. Being part of the human  
440 food production chain, legumes have a major impact on the global nitrogen cycle.  
441 Nitrogen-fixing plants can fix nitrogen through root nodule symbiosis (RNS) using  
442 symbiotic nitrogen-fixing bacteria. In a previous report, RNS was revealed to be  
443 restricted to Fabales, Fagales, Cucurbitales, and Rosales, which together form the  
444 monophyletic nitrogen-fixing clade. This suggests a predispositional event in their  
445 common ancestor, which enabled their subsequent evolution [77]. Despite this genetic  
446 predisposition, many leguminous members of the nitrogen-fixing clade are non-fixers  
447 [78]. This has raised the question as to whether the nodulation trait evolved  
448 independently in a convergent manner, or originated from a single evolutionary event  
449 followed by multiple losses. The answer to this question cannot be explained with  
450 current genomic approaches, because available genomic information of nodulating  
451 species is, at present, limited to a single subfamily, the Papilionoideae, in the Fabaceae.  
452 Although the Mimosoideae subfamily within the Fabaceae also contains nitrogen-  
453 fixing species, none of its members have been genome-sequenced.

454 In this analysis, we identified 16 root nodulation symbiosis signal (Sym) pathway  
455 genes in three legumes (*V. subterranea*, *L. purpureus*, and *F. albida*) and two non-  
456 legumes (*S. birrea* and *M. oleifera*). First, we collected the protein sequences of  
457 previously reported genes in the Sym pathways of *L. japonicus* and *M. truncatula* [79]  
458 (Figure 3). Using these sequences as bait, we predicted the Sym genes in *V. subterranea*,



1 459 *L. purpureus*, *F. albida*, *S. birrea*, and *M. oleifera* through reciprocal best hits generated  
2  
3 460 by a BLASTP search with an E-value of 1e-5 (Table 8). To verify this prediction with  
4  
5  
6 461 syntenic analysis, ‘all versus all’ BLASTP results were subjected to MCSCANX [80]  
7  
8  
9 462 with default parameters to generate syntenic blocks. The result showed that, among the  
10  
11  
12 463 legumes, all of the components in the pathway were conserved except for  
13  
14 464 *MtNFP/LjNFR5*, *LjCASTOR*, *CCaMK*, *MtCRE1/LjLHK1*, and *NF-YA2*, while many  
15  
16  
17 465 components were missing in the non-legumes. Among the three legumes, the  
18  
19  
20 466 orthologous genes *MtNFP/LjNFR5*, *LjCASTOR* and *MtIPD3/LjCYCLOPS* were absent  
21  
22  
23 467 in *F. albida*. As previously reported, the expression of *NIN* is lower in the *ipd3*-mutant  
24  
25  
26 468 line [81]; analysis of the *M. truncatula* mutant C31 showed that the Nod Factor  
27  
28  
29 469 Perception gene is essential in Nod factor perception at early stages of the symbiotic  
30  
31  
32 470 interaction [82]. Meanwhile, the function of *IPD3* was proved to be partly redundant,  
33  
34  
35 471 which means it is likely that other proteins phosphorylated by CCaMK can partially  
36  
37  
38 472 fulfill this role when *IPD3* is absent [81]. Differences in the components of the RNS  
39  
40  
41 473 pathway (Table 8), together with the relatively weak nitrogen-fixing ability [83] of *F.*  
42  
43  
44 474 *albida*, is thus a good reference for RNS diversification research.

45 475

## 47 476 **Conclusion**

49  
50 477 This comprehensive study reports the sequencing, assembly, and annotation of five  
51  
52  
53 478 genomes of underutilized plants in Africa, along with details of their key evolutionary  
54  
55  
56 479 features. The draft genomes of these species will serve as an important complementary  
57  
58  
59 480 resource for non-model food crops, especially the leguminous plants, and will be

1 481 valuable for both agroforestry and evolutionary research. Improving these underutilized  
2  
3 482 plants using genomics-assisted tools and methods could help to bring food security to  
4  
5  
6 483 millions of people.  
7  
8  
9 484

10  
11 485 **Additional files**

12  
13  
14 486 **Figure S1:** K-mer (K = 17) analysis of five genomes.  
15

16  
17 487 **Figure S2:** Distribution of sequencing depths of the assembly data.  
18

19  
20 488 **Figure S3:** GC content.  
21

22  
23 489 **Figure S4:** Comparison of GC content between closely related species.  
24

25  
26 490 **Figure S5:** Statistical analysis of gene models in *Vigna subterranea*, *Lablab purpureus*,  
27

28 491 *Faidherbia albida*, *Moringa oleifera* and *Sclerocarya birrea*.  
29

30  
31 492 **Figure S6:** Expansion and contraction of gene families.  
32

33  
34 493 **Table S1:** Statistical analysis of raw and clean DNA sequencing data  
35

36  
37 494 **Table S2:** Summary statistics of the transcriptome data of four species  
38

39  
40 495 **Table S3:** Estimation of genome size based on k-mer analysis of five species  
41

42  
43 496 **Table S4:** BUSCO evaluation of the annotated protein-coding genes of five species  
44

45  
46 497 **Table S5:** Analysis of gene families of different species  
47

48  
49 498 **Table S6:** Enriched pathways of unique paralog genes in families  
50

51  
52 499 **Table S7:** Enriched GO terms (level 3) of unique paralog genes in families  
53

54  
55 500 **Table S8:** Enriched GO terms (level 3) of genes in families with expansion  
56

57  
58 501 **Table S9:** Enriched pathways of genes in families with expansion  
59

60  
61 502 **Table S10:** Copy numbers of protein biosynthesis-related genes in each species  
62  
63  
64  
65

1 503 **Table S11:** Copy numbers of starch biosynthesis-related genes in each species

2  
3 504 **Table S12:** Copy numbers of fatty acid synthesis and storage-related genes in each  
4  
5  
6 505 species

7  
8  
9 506 **Table S13:** Copy numbers of fatty acid degradation-related genes in each species

10  
11 507 **Table S14:** Numbers of transcription factors in the studied species

12  
13  
14 508

15  
16  
17 509 **Abbreviations**

18  
19  
20 510 BUSCO: Benchmarking Universal Single-Copy Orthologs; LTR: long terminal repeat;

21  
22 511 TF: transcription factors; MITE: miniature inverted repeat transposable elements;

23  
24  
25 512 NCBI: National Center for Biotechnology Information; RNS: root nodule symbiosis;

26  
27  
28 513 RS.

29  
30  
31 514

32  
33 515 **Funding**

34  
35  
36 516 This work was supported by funding from the Shenzhen Municipal Government of

37  
38  
39 517 China (grant numbers JCYJ20150831201643396 and JCYJ20150529150409546), and

40  
41  
42 518 the Guangdong Provincial Key Laboratory of Genome Read and Write (grant number

43  
44  
45 519 2017B030301011). This work is part of the 10KP project led by BGI-Shenzhen and

46  
47 520 China National GeneBank.

48  
49  
50 521

51  
52 522 **Availability of supporting data**

53  
54  
55 523 The raw data from our genome project was deposited in the NCBI Sequence Read

56  
57  
58 524 Archive database with Bioproject IDs PRJNA453822 and PRJNA474418. Assembly

1 525 and annotation of the five genomes and other supporting data, including BUSCO results,  
2  
3 526 are available in the GigaDB repository [84], and the data reported in this study are also  
4  
5  
6 527 available in the CNGB Nucleotide Sequence Archive (CNSA: <https://db.cngb.org/cnsa> ;  
7  
8  
9 528 accession number CNP0000096).

10  
11 529

### 12 13 530 **Competing interests**

14  
15  
16  
17 531 The authors declare that they have no competing interests.  
18  
19

20 532

### 21 22 533 **Author contributions**

23  
24  
25 534 X.L., X.X., H.Y., J.W., P.S.H., R.J., A.V. and Y.C. conceived the project and supervised  
26  
27  
28 535 the respective components: DNA extraction, sample logistics and collection conducted  
29  
30  
31 536 by the African Orphan Crops Consortium of the World Agroforestry Centre; and data  
32  
33  
34 537 generation and analyses conducted by BGI. Y.C. supervised the analyses. R.K. and S.M.  
35  
36  
37 538 collected and extracted the DNA and RNA. S.B. and F.Y. performed the genome  
38  
39  
40 539 assembly. M.L., X.Z.L., S.B.W. and L.Z.L. performed the genome annotation, gene  
41  
42  
43 540 family analysis and identification of genes related to root growth and root nodule  
44  
45  
46 541 symbiosis. Y.C., M.L, X.Z.L. performed the phylogenetic analysis. Y.C., H.L., S.K.S.,  
47  
48  
49 542 P.S.H. and A.V. wrote the manuscript. H.R.L. and S.F.P. sequenced the samples. S.M.,  
50  
51  
52 543 W.K.H., A.M., P.S.H., J.W., H.M.Y. revised the manuscript. All authors read, edited and  
53  
54  
55 544 approved the final version of the manuscript.

56 545

### 57 58 546 **References**

- 1 547 1. United Nations, Department of Economic and Social Affairs, Population  
2  
3 548 Division. World population prospects: the 2017 revision, Key Findings and  
4  
5  
6 549 Advance Tables. 2017. Working Paper No. ESA/P/WP/248.  
7  
8  
9 550 2. Development Initiatives. Global nutrition report 2017: nourishing the SDGs.  
10  
11 551 Bristol, UK: Development Initiatives. 2017.  
12  
13  
14 552 3. Mouillé, B., Charrondière, U. R., & Burlingame. The contribution of plant  
15  
16  
17 553 genetic resources to health and dietary diversity. Thematic Background Study.  
18  
19  
20 554 2010.  
21  
22  
23 555 4. African Orphan Crops Consortium: <http://www.africanorphancrops.org> (2018).  
24  
25 556 Accessed 20 Nov 2018.  
26  
27  
28 557 5. Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, et al. Draft  
29  
30  
31 558 genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of  
32  
33  
34 559 resource-poor farmers. Nat Biotechnol. 2011;30:83-89. doi:10.1038/nbt.2022.  
35  
36  
37 560 6. Foyer CH, Lam H-M, Nguyen HT, Siddique KHM, Varshney RK, Colmer TD,  
38  
39 561 et al. Neglecting legumes has compromised human health and sustainable food  
40  
41  
42 562 production. Nat Plants. 2016;2:16112. doi:10.1038/nplants.2016.112.  
43  
44  
45 563 7. Borget M. Food legumes. In: The Tropical Agriculturalist, CTA Macmillan.  
46  
47 564 1992.  
48  
49  
50 565 8. Linnemann A.R, Azam–Ali S.N. Bambara groundnut (*Vigna subterranea*)  
51  
52  
53 566 literature review: A revised and updated bibliography. Tropical Crops  
54  
55  
56 567 Communication No. 7. 1993.  
57  
58  
59 568 9. Gbaguidi AA, Dansi A, Dossou-Aminon I, Gbemavo DSJC, Orobiyi A,  
60  
61  
62  
63  
64  
65

- 1 569 Sanoussi F, et al. Agromorphological diversity of local Bambara groundnut  
2  
3 570 (*Vigna subterranea* (L.) Verdc.) collected in Benin. Genet Resour Crop Evol.  
4  
5  
6 571 2018;65(4):1159-1171. doi:10.1007/s10722-017-0603-4.  
7  
8  
9 572 10. Kang YJ, Kim SK, Kim MY, Lestari P, Kim KH, Ha B-K, et al. Genome  
10  
11 573 sequence of mungbean and insights into evolution within *Vigna* species. Nat  
12  
13 574 Commun. 2014;5:5443. doi:10.1038/ncomms6443.  
14  
15  
16  
17 575 11. Yang K, Tian Z, Chen C, Luo L, Zhao B, Wang Z, et al. Genome sequencing of  
18  
19 576 adzuki bean (*Vigna angularis*) provides insight into high starch and low fat  
20  
21 577 accumulation and domestication. Proc Natl Acad Sci U S A.  
22  
23 578 2015;112(43):13213-13218. doi:10.1073/pnas.1420949112.  
24  
25  
26  
27  
28 579 12. Jung IL. Soluble extract from *Moringa oleifera* leaves with a new anticancer  
29  
30 580 activity. PLoS One. 2014;9(4):e95492. doi:10.1371/journal.pone.0095492.  
31  
32  
33  
34 581 13. Leone A, Spada A, Battezzati A, Schiraldi A, Aristil J and Bertoli S. Cultivation,  
35  
36 582 genetic, ethnopharmacology, phytochemistry and pharmacology of *Moringa*  
37  
38 583 *oleifera* Leaves: An Overview. Int J Mol Sci. 2015;16(6):12791-12835.  
39  
40 584 doi:10.3390/ijms160612791.  
41  
42  
43  
44 585 14. Lea M. Bioremediation of turbid surface water using seed extract from *Moringa*  
45  
46 586 *oleifera* Lam. (drumstick) tree. Curr Protoc Microbiol. 2014;33:1G.2.1-G.2.8.  
47  
48 587 doi:10.1002/9780471729259.mc01g02s16.  
49  
50  
51  
52  
53 588 15. Mabapa MP, Ayisi KK, Mariga IK, Mohlabi RC and Chuene RS. Production  
54  
55 589 and utilization of moringa by farmers in Limpopo Province, South Africa.  
56  
57 590 International Journal of Agricultural Research. 1962;12(4):160-171.  
58  
59  
60  
61  
62  
63  
64  
65

- 1 591 doi:10.3923/ijar.2017.160.171.  
2  
3  
4 592 16. Tian Y, Zeng Y, Zhang J, Yang CG, Yan L, Wang XJ, et al. High quality  
5  
6 593 reference genome of drumstick tree (*Moringa oleifera* Lam.), a potential  
7  
8  
9 594 perennial crop. *Science China Life Sciences*. 2015;58(7):627-638.  
10  
11 595 doi:10.1007/s11427-015-4872-x.  
12  
13  
14 596 17. Maass BL, Knox MR, Venkatesha SC, Angessa TT, Ramme S and Pengelly BC.  
15  
16  
17 597 *Lablab purpureus*-a crop lost for Africa? *Trop Plant Biol*. 2010;3(3):123-135.  
18  
19  
20 598 doi:10.1007/s12042-010-9046-1.  
21  
22  
23 599 18. Robotham O and Chapman M. Population genetic analysis of hyacinth bean  
24  
25 600 (*Lablab purpureus* (L.) Sweet, Leguminosae) indicates an East African origin  
26  
27  
28 601 and variation in drought tolerance. *Genet Resour Crop Evol*. 2017;64(1):139-  
29  
30  
31 602 148. doi:10.1007/s10722-015-0339-y.  
32  
33  
34 603 19. Kamotho GN. Evaluation of adaptability potential and genetic diversity of  
35  
36 604 Kenyan Dolichos bean germplasm. PhD thesis. 2015.  
37  
38  
39 605 20. Vankatesha S.C. Molecular characterization and development of mapping  
40  
41  
42 606 populatuions for construction of genetic map in dolichos bean. PhD thesis. 2012.  
43  
44  
45 607 21. Mokgolodi NC, Setshogo MP, Shi L-l, Liu Y-j and Ma C. Achieving food and  
46  
47 608 nutritional security through agroforestry: a case of *Faidherbia albida* in sub-  
48  
49  
50 609 Saharan Africa. *For. Stud. China*. 2011;13(2):123-131. doi:10.1007/s11632-  
51  
52  
53 610 011-0202-y.  
54  
55  
56 611 22. Garrity DP, Akinnifesi FK, Ajayi OC, Weldesemayat SG, Mowo JG,  
57  
58 612 Kalinganire A, et al. Evergreen agriculture: a robust approach to sustainable

- 1 613 food security in Africa. *Food Sec.* 2010;2(3):197-214. doi:10.1007/s12571-010-  
2  
3  
4 614 0070-7.  
5  
6 615 23. DUNHAM KM. Biomass dynamics of herbaceous vegetation in Zambezi  
7  
8  
9 616 riverine woodlands. *African Journal of Ecology.* 1990;28(3):200-212.  
10  
11 617 doi:10.1111/j.1365-2028.1990.tb01153.x.  
12  
13  
14 618 24. Barnes RD and Fagg CW. *Faidherbia albida* monograph and annotated  
15  
16 619 bibliography. Oxford Forestry Inst. 2003;41-267  
17  
18  
19 620 25. Nerd A, Mizrahi Y, Janick J and Simon JE. Domestication and introduction of  
21  
22 621 marula (*Sclerocarya birrea* subsp. *caffra*) as a new crop for the Negev Desert  
23  
24 622 of Israel. *New crops.* 1993;496-499.  
25  
26  
27 623 26. Mng'Omba SA, Sileshi GW, Jamnadass R, Akinnifesi FK and Mhango J. Scion  
28  
29 624 and stock diameter size effect on growth and fruit production of *Sclerocarya*  
30  
31 625 *birrea* (Marula) trees. *J Hortic For.* 2012;4(9):153-60.  
32  
33  
34 626 27. Gouwakinnou GN, Lykke AM, Assogbadjo AE and Sinsin B. Local knowledge,  
35  
36 627 pattern and diversity of use of *Sclerocarya birrea*. *J Ethnobiol Ethnomed.*  
37  
38 628 2011;7 (1):1-9. doi:10.1186/1746-4269-7-8.  
39  
40  
41 629 28. Yang T and Wu C. DNA Extraction for plant samples by CTAB. protocols.io.  
42  
43 630 2018; dx.doi.org/10.17504/protocols.io.pzqdp5w  
44  
45  
46 631 29. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an  
47  
48 632 empirically improved memory-efficient short-read de novo assembler.  
49  
50 633 *GigaScience.* 2012;1(1):1-6. doi:10.1186/2047-217X-1-18.  
51  
52  
53 634 30. Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V, et al. The draft  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65



- 1 635 genome of tropical fruit durian (*Durio zibethinus*). Nat Genet. 2017;49:1633-  
2  
3  
4 636 1641. doi:10.1038/ng.3972.  
5  
6 637 31. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM.  
7  
8  
9 638 BUSCO: assessing genome assembly and annotation completeness with single-  
10  
11  
12 639 copy orthologs. Bioinformatics. 2015;31(19):3210-3212.  
13  
14 640 doi:10.1093/bioinformatics/btv351.  
15  
16  
17 641 32. Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, et al. Bridger: a new framework  
18  
19  
20 642 for de novo transcriptome assembly using RNA-seq data. Genome Biol.  
21  
22 643 2015;16:30. doi:10.1186/s13059-015-0596-2.  
23  
24  
25 644 33. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res.  
26  
27  
28 645 2002;12(4):656-664. doi:10.1101/gr.229202.  
29  
30  
31 646 34. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: an improved  
32  
33  
34 647 ultrafast tool for short read alignment. Bioinformatics. 2009;25(15):1966-1967.  
35  
36 648 doi:10.1093/bioinformatics/btp336.  
37  
38  
39 649 35. Tarailo-Graovac M and Chen N. Using RepeatMasker to identify repetitive  
40  
41  
42 650 elements in genomic sequences. Curr Protoc Bioinformatics. 2009;25(1) 4.10.1-  
43  
44 651 4.10.14. doi:10.1002/0471250953.bi0410s25.  
45  
46  
47 652 36. Han Y and Wessler SR. MITE-Hunter: a program for discovering miniature  
48  
49  
50 653 inverted-repeat transposable elements from genomic sequences. Nucleic Acids  
51  
52 654 Res. 2010;38(22):e199-e199. doi:10.1093/nar/gkq862.  
53  
54  
55 655 37. Ellinghaus D, Kurtz S and Willhoeft U. LTRharvest, an efficient and flexible  
56  
57  
58 656 software for de novo detection of LTR retrotransposons. BMC Bioinformatics.  
59  
60  
61  
62  
63  
64  
65

- 1 657 2008;9:18. doi:10.1186/1471-2105-9-18.  
2  
3 658 38. Gremme G, Steinbiss S and Kurtz S. GenomeTools: a comprehensive software  
4  
5  
6 659 library for efficient processing of structured genome annotations. IEEE/ACM  
7  
8  
9 660 Trans Comput Biol Bioinform. 2013;10(3):645-656. doi:10.1109/tcbb.2013.68.  
10  
11 661 39. Steinbiss S, Willhoeft U, Gremme G and Kurtz S. Fine-grained annotation and  
12  
13  
14 662 classification of de novo predicted LTR retrotransposons. Nucleic Acids Res.  
15  
16  
17 663 2009;37(21):7002-7013. doi:10.1093/nar/gkp759.  
18  
19  
20 664 40. Chan PP and Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA  
21  
22  
23 665 genes identified in complete and draft genomes. Nucleic Acids Res. 2016;44  
24  
25  
26 666 (D1):D184-D9. doi:10.1093/nar/gkv1309.  
27  
28 667 41. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high  
29  
30  
31 668 throughput. Nucleic Acids Res. 2004;32(5):1792-1797. doi:10.1093/nar/gkh340.  
32  
33  
34 669 42. Campbell MS, Holt C, Moore B and Yandell M. Genome annotation and  
35  
36  
37 670 curation using MAKER and MAKER-P. Curr Protoc Bioinformatics.  
38  
39  
40 671 2014;48(1): 4.11.1-4.11.39. doi:10.1002/0471250953.bi0411s48.  
41  
42 672 43. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al.  
43  
44  
45 673 De novo transcript sequence reconstruction from RNA-seq using the Trinity  
46  
47  
48 674 platform for reference generation and analysis. Nat Protoc. 2013;8:1494–1512.  
49  
50  
51 675 doi:10.1038/nprot.2013.084.  
52  
53 676 44. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated  
54  
55  
56 677 eukaryotic gene structure annotation using EVidenceModeler and the program  
57  
58  
59 678 to assemble spliced alignments. Genome Biol. 2008;9(1):R7. doi:10.1186/gb-

- 1 679 2008-9-1-r7.  
2  
3  
4 680 45. Stanke M, Schoffmann O, Morgenstern B and Waack S. Gene prediction in  
5  
6 681 eukaryotes with a generalized hidden Markov model that uses hints from  
7  
8  
9 682 external sources. *BMC Bioinformatics*. 2006;7:62. doi:10.1186/1471-2105-7-  
10  
11 683 62.  
12  
13  
14 684 46. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO and Borodovsky M. Gene  
15  
16  
17 685 identification in novel eukaryotic genomes by self-training algorithm. *Nucleic*  
18  
19  
20 686 *Acids Res*. 2005;33(20):6494-6506. doi:10.1093/nar/gki937.  
21  
22  
23 687 47. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 2004;5:59.  
24  
25 688 doi:10.1186/1471-2105-5-59.  
26  
27  
28 689 48. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al.  
29  
30  
31 690 Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*.  
32  
33  
34 691 2015;43(D1):D130-D137. doi:10.1093/nar/gku1063.  
35  
36  
37 692 49. Lowe TM and Chan PP. tRNAscan-SE On-line: integrating search and context  
38  
39 693 for analysis of transfer RNA genes. *Nucleic Acids Res*. 2016;44(W1):W54-  
40  
41  
42 694 W57. doi:10.1093/nar/gkw413.  
43  
44  
45 695 50. Tanabe M and Kanehisa M. Using the KEGG database resource. *Curr Protoc*  
46  
47 696 *Bioinformatics*. 2012; 38(1):1.12.1-1.12.43.  
48  
49  
50 697 doi:10.1002/0471250953.bi0112s38.  
51  
52  
53 698 51. Tatusov RL, Koonin EV and Lipman DJ. A genomic perspective on protein  
54  
55  
56 699 families. *Science*. 1997;278(5338):631-637.  
57  
58  
59 700 52. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E,

- 1 701 et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in  
2  
3 702 2003. *Nucleic Acids Res.* 2003;31(1):365-370.  
4  
5  
6 703 53. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan  
7  
8  
9 704 5: genome-scale protein function classification. *Bioinformatics.*  
10  
11 705 2014;30(9):1236-1240. doi:10.1093/bioinformatics/btu031.  
12  
13  
14 706 54. Finn RD, Mistry J, Tate J, Coghill P, Heger A, Pollington JE, et al. The Pfam  
15  
16  
17 707 protein families database. *Nucleic Acids Res.* 2010;38 suppl 1:D211-D222.  
18  
19  
20 708 doi:10.1093/nar/gkp985.  
21  
22  
23 709 55. Letunic I, Doerks T and Bork P. SMART 6: recent updates and new  
24  
25 710 developments. *Nucleic Acids Res.* 2009;37 suppl 1:D229-D232.  
26  
27  
28 711 doi:10.1093/nar/gkn808.  
29  
30  
31 712 56. Mi H, Muruganujan A, Casagrande JT and Thomas PD. Large-scale gene  
32  
33 713 function analysis with the PANTHER classification system. *Nat Protoc.*  
34  
35  
36 714 2013;8:1551-1566. doi:10.1038/nprot.2013.092  
37  
38  
39 715 57. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, et  
40  
41  
42 716 al. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*  
43  
44  
45 717 2003;31(1):400-402.  
46  
47  
48 718 58. Corpet F, Servant F, Gouzy J and Kahn D. ProDom and ProDom-CG: tools for  
49  
50 719 protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*  
51  
52  
53 720 2000;28(1):267-269.  
54  
55  
56 721 59. Stichting C, Centrum M and Dongen SV. A Cluster Algorithm for Graphs.  
57  
58 722 *Information Systems [INS].* 2000:1-40.  
59  
60  
61  
62  
63  
64  
65

- 1 723 60. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W and Gascuel O.  
2  
3 724 New algorithms and methods to estimate maximum-likelihood phylogenies:  
4  
5  
6 725 assessing the performance of PhyML 3.0. *Syst Biol.* 2010;59(3):307-321.  
7  
8  
9 726 doi:10.1093/sysbio/syq010.  
10  
11 727 61. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol.*  
12  
13  
14 728 2007;24(8):1586-1591. doi:10.1093/molbev/msm088.  
15  
16  
17 729 62. He N, Zhang C, Qi X, Zhao S, Tao Y, Yang G, et al. Draft genome sequence of  
18  
19  
20 730 the mulberry tree *Morus notabilis*. *Nat Commun.* 2013;4:2445.  
21  
22  
23 731 doi:10.1038/ncomms3445.  
24  
25 732 63. Lavin M, Herendeen PS, Wojciechowski MF and Linder P. Evolutionary rates  
26  
27  
28 733 analysis of leguminosae implicates a rapid diversification of lineages during the  
29  
30  
31 734 Tertiary. *Syst Biol.* 2005;54(4):575-594. doi:10.1080/10635150590947131.  
32  
33  
34 735 64. De Bie T, Cristianini N, Demuth JP and Hahn MW. CAFE: a computational tool  
35  
36  
37 736 for the study of gene family evolution. *Bioinformatics.* 2006;22(10):1269-1271.  
38  
39  
40 737 doi:10.1093/bioinformatics/btl097.  
41  
42 738 65. Bernhardt C, Lee MM, Gonzalez A, Zhang F, Lloyd A and Schiefelbein J. The  
43  
44  
45 739 bHLH genes *GLABRA3* (GL3) and *ENHANCER OF GLABRA3* (EGL3)  
46  
47  
48 740 specify epidermal cell fate in the *Arabidopsis* root. *Development.*  
49  
50  
51 741 2003;130(26):6431-6439. doi:10.1242/dev.00880.  
52  
53 742 66. Paponov IA, Paponov M, Teale W, Menges M, Chakrabortee S, Murray JA, et  
54  
55  
56 743 al. Comprehensive transcriptome analysis of auxin responses in *Arabidopsis*.  
57  
58  
59 744 *Mol Plant.* 2008;1(2):321-337. doi:10.1093/mp/ssm021.  
60  
61  
62  
63  
64  
65

- 1 745 67. Vanneste S, Rybel BD, Beemster GTS, Ljung K, Smet ID, Isterdael GV, et al.  
2  
3 746 Cell cycle progression in the pericycle is not sufficient for SOLITARY  
4  
5  
6 747 ROOT/IAA14-mediated lateral root initiation in *Arabidopsis thaliana*. Plant  
7  
8  
9 748 Cell. 2005;17(11):3035-3050. doi:10.1105/tpc.105.035493.  
10  
11 749 68. Vandenbeldt RJ. *Faidherbia albida* in the West African semi-arid tropics.  
12  
13 750 ICRISAT. 1992. p. 107-110.  
14  
15  
16 751 69. Jin J, Tian F, Yang D-C, Meng Y-Q, Kong L, Luo J, et al. PlantTFDB 4.0: toward  
17  
18 752 a central hub for transcription factors and regulatory interactions in plants.  
19  
20  
21 753 Nucleic Acids Res. 2017;45(D1):D1040-D5. doi:10.1093/nar/gkw982.  
22  
23  
24 754 70. Jang YE, Kim MY, Shim S, Lee J and Lee S-H. Gene expression profiling for  
25  
26 755 seed protein and oil synthesis during early seed development in soybean. Genes  
27  
28  
29 756 Genom. 2015;37(4):409-418. doi:10.1007/s13258-015-0269-2.  
30  
31  
32 757 71. Bamshaiye OM, Adegbola JA and Bamishaiye EI. Bambara groundnut : an  
33  
34 758 under-utilized nut in Africa. Adv Agric Biotechnol. 2011;1:60-72.  
35  
36  
37 759 72. Raigond P, Ezekiel R and Raigond B. Resistant starch in food: a review. J Sci  
38  
39  
40 760 Food Agric. 2015;95(10):1968-1978.  
41  
42  
43 761 73. Zhou H, Wang L, Liu G, Meng X, Jing Y, Shu X, et al. Critical roles of soluble  
44  
45 762 starch synthase SSIIIa and granule-bound starch synthase Waxy in synthesizing  
46  
47  
48 763 resistant starch in rice. Proc Natl Acad Sci U S A. 2016;113(45):12844-12849.  
49  
50  
51 764 doi:10.1073/pnas.1615104113.  
52  
53  
54 765 74. Bird AR, Flory C, Davies DA, Usher S and Topping DL. A novel barley cultivar  
55  
56  
57 766 (*Himalaya 292*) with a specific gene mutation in starch synthase IIa raises large  
58  
59  
60  
61  
62  
63  
64  
65

- 1 767            bowel starch and short-chain fatty acids in rats. *J Nutr.* 2004;134(4):831-835.  
2  
3 768            doi:10.1093/jn/134.4.831.  
4  
5  
6 769    75.    Morre DJ, Nyquist S and Rivera E. Lecithin biosynthetic enzymes of onion stem  
7  
8 770            and the distribution of phosphorylcholine-cytidyl transferase among cell  
9  
10 771            fractions. *Plant Physiol.* 1970;45(6):800-804.  
11  
12 772    76.    Johnson KD and Kende H. Hormonal control of lecithin synthesis in barley  
13  
14 773            aleurone cells: regulation of the CDP-choline pathway by gibberellin. *Proc Natl*  
15  
16 774            *Acad Sci U S A.* 1971;68(11):2674-2677.  
17  
18 775    77.    Soltis DE, Soltis PS, Morgan DR, Swensen SM, Mullin BC, Dowd JM, et al.  
19  
20 776            Chloroplast gene sequence data suggest a single origin of the predisposition for  
21  
22 777            symbiotic nitrogen fixation in angiosperms. *Proc Natl Acad Sci U S A.*  
23  
24 778            1995;92(7):2647-2651.  
25  
26 779    78.    Doyle JJ. Phylogenetic perspectives on the origins of nodulation. *Mol Plant*  
27  
28 780            *Microbe Interact.* 2011;24(11):1289-1295. doi:10.1094/MPMI-05-11-0114.  
29  
30 781    79.    Geurts R, Xiao TT and Reinhold-Hurek B. What does it take to evolve a  
31  
32 782            nitrogen-fixing endosymbiosis? *Trends Plant Sci.* 2016;21 (3):199–208.  
33  
34 783            doi:10.1016/j.tplants.2016.01.012.  
35  
36 784    80.    Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCSScanX: A toolkit  
37  
38 785            for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic*  
39  
40 786            *Acids Res.* 2012;40(7):e49. doi:10.1093/nar/gkr1293.  
41  
42 787    81.    Horváth B, Li HY, Domonkos Á, Halász G, Gobbato E, Ayaydin F, et al.  
43  
44 788            *Medicago truncatula* IPD3 is a member of the common symbiotic signaling  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- 1 789 pathway required for rhizobial and mycorrhizal symbioses. *Mol Plant Microbe*  
2  
3 790 *Interact.* 2011;24(11):1345-1358. doi:10.1094/MPMI-01-11-0015.  
4  
5  
6 791 82. Amor BB, Shaw SL, Oldroyd GED, Maillet F, Penmetsa RV, Cook D, et al. The  
7  
8 792 NFP locus of *Medicago truncatula* controls an early step of Nod factor signal  
9  
10 transduction upstream of a rapid calcium flux and root hair deformation. *Plant*  
11 793 *J.* 2003;34(4):495-506.  
12  
13 794  
14  
15  
16  
17 795 83. Ndoye I, Gueye M, Danso SKA and Dreyfus B. Nitrogen fixation in *Faidherbia*  
18  
19 796 *albida*, *Acacia raddiana*, *Acacia senegal* and *Acacia seyal* estimated using the  
20  
21 <sup>15</sup>N isotope dilution technique. *Plant Soil.* 1995;172(2):175-180.  
22 797  
23  
24  
25 798 doi:10.1007/BF00011319.  
26  
27  
28 799 84. Chang Y, Liu H, Liu M, Liao X, Sahu SK, Fu Y, Song B; Cheng S, Kariba R,  
29  
30 800 Muthemba S, Hendre PS, Mayes S, Ho WK, Kendabie P, Wang S, Li L,  
31  
32 801 Muchugi A, Jamnadass R, Lu H, Peng S, Deynze AV, Simons A, Yana-Shapiro  
33  
34 802 H, Xu X, Yang H, Wang J, Liu X. Supporting data for "The draft genomes of  
35  
36 803 five agriculturally important African orphan crops". *GigaScience Database*  
37  
38 804 2018. <http://dx.doi.org/10.5524/100504>.  
39  
40  
41  
42  
43 805 85. Bioinformatics & Evolutionary Genomics:  
44  
45  
46 806 <http://bioinformatics.psb.ugent.be/webtools/Venn/>. Accessed 20 Nov 2018.  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65



**Table 1:** Statistical analysis of the final *de novo* genome assembly of *Vigna subterranea*, *Lablab purpureus*, *Faidherbia albida*, *Sclerocarya birrea* and *Moringa oleifera*

Parameters	<i>V. subterranea</i>		<i>L. purpureus</i>		<i>F. albida</i>		<i>S. birrea</i>		<i>M. oleifera</i>	
	Contig	Scaffold	Contig	Scaffold	Contig	Scaffold	Contig	Scaffold	Contig	Scaffold
N90	3,804	75,271	785	860	8,254	95,167	3,661	21,833	6,676	57,837
N80	7,872	197,296	8,009	61,348	16,321	251,730	7,649	82,385	16,503	241,828
N70	11,464	325,826	16,144	205,392	24,165	380,587	11,885	155,416	25,754	441,152
N60	15,122	474,616	24,010	359,168	32,440	534,880	16,393	243,236	35,081	644,014
N50	19,154	640,666	32,223	621,373	42,029	692,039	21,349	335,449	45,268	957,246
N40	23,828	865,081	42,690	950,808	53,479	881,230	26,914	485,585	58,406	1,446,587
N30	29,382	1,133,817	54,401	1,489,002	69,167	1,197,388	33,914	705,409	74,710	1,878,891
N20	36,928	1,503,436	70,790	1,971,744	92,147	1,501,241	43,984	1,098,843	96,626	2,565,629
N10	49,695	2,049,645	95,643	2,606,483	139,388	1,925,526	62,875	2,089,533	136,952	3,296,678

16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Chang et al.

Orphan crop genomes

	N90	29,245	1,087	26,272	9,409	16,834	1,132	17,585	1,537	5,524	366
	N80	20,188	664	9,869	715	11,420	727	11,678	787	3,574	191
	N70	14,829	453	6,576	366	8,198	514	8,313	499	2,542	125
	N60	10,943	315	4,630	222	5,898	370	6,001	332	1,833	84
<b>Number</b>	N50	7,932	220	3,244	138	4,151	263	4,277	214	1,295	56
	N40	5,532	147	2,204	86	2,791	179	2,929	131	876	37
	N30	3,590	93	1,403	52	1,728	114	1,857	74	553	24
	N20	2,024	52	776	29	912	64	1,012	36	300	13
	N10	806	22	306	12	326	26	387	12	112	6
	Maximum length	148,612	3,684,321	240,194	5,699,750	529,842	4,746,824	227,874	5,850,796	449,426	4,637,711
	Total length	512,516,846	535,052,523	385,303,786	395,472,305	644,456,383	653,726,905	322,977,033	330,983,508	213,739,255	216,759,177
	Total number $\geq$										
100 bp		104,575	65,586	135,039	118,976	75,572	51,470	64,158	40,280	29,972	22,329

15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Chang et al.

Orphan crop genomes

---

Total number	$\geq$											
2000 bp	35,465	2,920	15,984	4,265	26,459	5,758	22,172	4,852	8,300	2,166		
N content (%)	4.21		2.57		1.42		2.42		1.39			

---

809

810 **Table 2:** Completeness evaluation of genome assembly using BUSCO database in five  
 811 species

BUSCOs	<i>Vigna</i>		<i>Lablab</i>		<i>Faidherbia</i>		<i>Sclerocarya</i>		<i>Moringa</i>	
	<i>subterranea</i>		<i>purpureus</i>		<i>albida</i>		<i>birrea</i>		<i>oleifera</i>	
	N	%	N	%	N	%	N	%	N	%
Complete single copy	1,244	86.39	1,258	87.40	1,231	85.50	1352	93.90	1,278	88.80
Complete duplicated	82	5.69	83	5.80	84	5.80	32	2.20	19	1.30
Fragmented	28	1.94	20	1.40	34	2.40	21	1.50	23	1.60
Missing	86	5.97	79	5.40	91	6.30	35	2.40	120	8.30
Total	1440	/	1440	/	1440	/	1440	/	1440	/

812 Abbreviation: BUSCO, Benchmarking universal single-copy orthologs; N, number

813

814 **Table 3:** Gene coverage of candidate species based on transcriptome data

Species	Dataset	Number	Total length (bp)	Base coverage by assembly (%)	Sequence coverage by assembly (%)
	All	116,223	161,077,155	89.61	98.21
<i>Vigna</i>	>200 bp	116,223	161,077,155	89.61	98.21
<i>subterranea</i>	>500 bp	72,139	147,068,299	89.03	98.00
	>1000 bp	47,952	129,884,929	88.33	97.52
	All	86,867	80,837,182	93.59	99.25
<i>Lablab</i>	>200 bp	86,867	80,837,182	93.59	99.25
<i>purpureus</i>	>500 bp	41,252	66,764,786	92.94	99.18
	>1000 bp	24,627	55,074,989	92.32	99.02
	All	50,294	46,650,067	93.62	98.85
<i>Faidherbia</i>	>200 bp	50,294	46,650,067	93.62	98.85
<i>albida</i>	>500 bp	26,352	39,282,694	93.32	99.05
	>1000 bp	15,569	31,560,858	92.78	98.95
	All	60,964	57,114,636	88.98	92.16
<i>Moringa</i>	>200 bp	60,964	57,114,636	88.98	92.16
<i>oleifera</i>	>500 bp	29,581	47,523,018	88.85	92.69
	>1000 bp	18,322	39,528,310	88.70	92.99

815

816 **Table 4:** Proportion of different classes of repeats (%) in five species

Repeat type	<i>Vigna subterranea</i>		<i>Lablab purpureus</i>		<i>Faidherbia albida</i>		<i>Sclerocarya birrea</i>		<i>Moringa oleifera</i>	
	% in	Length	% in	Length	% in	Length	% in	Length (bp)	%in	Length
	genome	(bp)	genome	(bp)	genome	(bp)	genome		genome	(bp)
SINE	0	313	0.005	19,444	< 0.01	1,966	0.02	69,836	0.11	248,569
LINE	0.25	1,387,567	0.45	1,784,785	0.91	6,003,271	0.19	647,579	1.83	3,970,802
LTR	19.77	105,828,735	23.78	94,062,428	44.65	291,901,514	38.78	128,362,381	22.69	49,200,625
DNA	7.15	38,294,871	4.76	18,851,402	4	26,164,519	1.76	5,829,982	5.81	12,599,607
Satellite	0.01	71,679	0.02	107,451	0.01	110,749	0	18,597	0.74	1,623,399
Simple repeat	0.35	1,922,719	0.2	821,773	0.04	308,481	0.04	153,135	0.29	630,662
Others	11.94	63,926,350	8.95	35,400,400	6.48	42,426,306	5.11	16,918,179	10.35	22,439,026
<b>Total</b>	<b>38.35</b>	<b>205,189,285</b>	<b>37.18</b>	<b>147,050,327</b>	<b>54.86</b>	<b>358,653,534</b>	<b>45.18</b>	<b>149,551,125</b>	<b>40.57</b>	<b>87,944,150</b>

817 Abbreviations: bp, base pairs; DNA, deoxyribonucleic acid; LINE, long interspersed nuclear element; LTR, long terminal repeats; SINE, short interspersed nuclear element.

818 **Table 5:** Gene structure parameters of *Vigna subterranea*, *Lablab purpureus*,  
 819 *Faidherbia albida*, *Medicago truncatula*, *Glycine max*, *Sclerocarya birrea*, *Moringa*  
 820 *oleifera*, *Carica papaya*, *Theobroma cacao* and *Citrus sinensis*

Species	Protein-coding gene number	Mean gene length (bp)	Mean coding sequence length (bp)	Mean exons per gene	Mean exon length (bp)	Mean intron length (bp)
<i>V. subterranea</i>	31,707	3,287	1,163	5	222	501
<i>L. purpureus</i>	20,946	3,696	1,276	5	239	557
<i>F. albida</i>	28,979	3,396	1,207	5	226	504
<i>M. truncatula</i>	50,358	2,334	986	4	243	440
<i>G. max</i>	55,137	3,144	1,169	5	232	488
<i>S. birrea</i>	18,937	3,561	1,343	6	239	479
<i>M. oleifera</i>	18,451	3,308	1,238	5	232	478
<i>C. papaya</i>	24,107	2,531	962	4	223	473
<i>T. cacao</i>	41,951	3,684	1,323	6	223	479
<i>C. sinensis</i>	35,182	3,797	1,424	6	237	475

821 **Table 6:** Annotation of non-coding RNA genes in the genomes of *Vigna subterranea*,  
 822 *Lablab purpureus*, *Faidherbia albida*, *Sclerocarya birrea* and *Moringa oleifera*

Species	Type	Copy	Average length (bp)	Total length (bp)	% of genome	
<i>V. subterranea</i>	miRNA	102	122	12,466	0.002330	
	tRNA	756	75	56,639	0.010586	
	rRNA	rRNA	1,080	124	134,185	0.025079
	18S	55	560	30,798	0.005756	

1		28S	62	126	7,793	0.001456
2						
3		5.8S	17	124	2,110	0.000394
4						
5		5S	946	99	93,484	0.017472
6						
7	snRNA	snRNA	523	117	61,006	0.011402
8						
9		CD-box	327	100	32,643	0.006101
10						
11		HACA-box	47	133	6,236	0.001165
12						
13		splicing	149	149	22,127	0.004135
14						
15						
16		miRNA	109	123	13,398	0.003388
17						
18		tRNA	611	75	45,748	0.011568
19						
20	rRNA	rRNA	633	227	143,466	0.036277
21						
22		18S	213	446	95,074	0.024041
23						
24		28S	283	121	34,186	0.008644
25						
26						
27	<i>L. purpureus</i>	5.8S	53	135	7,177	0.001815
28						
29		5S	84	84	7,029	0.001777
30						
31	snRNA	snRNA	457	118	54,029	0.013662
32						
33		CD-box	278	97	26,915	0.006806
34						
35		HACA-box	48	133	6,371	0.001611
36						
37		splicing	131	158	20,743	0.005245
38						
39						
40		miRNA	126	122	15,364	0.002350
41						
42		tRNA	458	75	34,388	0.005260
43						
44	rRNA	rRNA	1,008	107	107,518	0.016447
45						
46		18S	25	321	8,034	0.001229
47						
48		28S	26	118	3,063	0.000469
49	<i>F. albida</i>					
50		5.8S	6	118	710	0.000109
51						
52		5S	951	101	95,711	0.014641
53						
54	snRNA	snRNA	1,996	108	216,482	0.033115
55						
56		CD-box	1,836	106	194,676	0.029779
57						
58						
59						
60						
61						
62						
63						
64						
65						



		HACA-box	42	132	5,548	0.000849
		splicing	118	138	16,258	0.002487
		miRNA	106	122	12,899	0.003897
		tRNA	564	75	42,181	0.012744
	rRNA	rRNA	313	142	44,378	0.013408
		18S	80	240	19,239	0.005813
		28S	57	113	6,460	0.001952
	<i>S. birrea</i>	5.8S	16	103	1,644	0.000497
		5S	160	106	17,035	0.005147
	snRNA	snRNA	841	115	96,517	0.029161
		CD-box	638	105	67,216	0.020308
		HACA-box	34	124	4,217	0.001274
		splicing	169	148	25,084	0.007579
		miRNA	111	119	13,161	0.006072
		tRNA	1,241	75	93,620	0.043191
	rRNA	rRNA	8,406	309	2,598,079	1.198602
		18S	3,256	608	1,979,080	0.913032
		28S	3,808	113	430,280	0.198506
	<i>M. oleifera</i>	5.8S	1,182	150	177,612	0.08194
		5S	160	69	11,107	0.005124
	snRNA	snRNA	229	119	27,158	0.012529
		CD-box	119	97	11,578	0.005341
		HACA-box	38	132	4,999	0.002306
		splicing	72	147	10,581	0.004881

**Table 7:** Statistical analysis of the functional annotations of protein-coding genes in the genomes of *Vigna subterranea*, *Lablab purpureus*, *Faidherbia albida*, *Sclerocarya birrea* and *Moringa oleifera*

Database	<i>V. subterranea</i>		<i>L. purpureus</i>		<i>F. albida</i>		<i>S. birrea</i>		<i>M. oleifera</i>	
	N	%	N	%	N	%	N	%	N	%
Nr	31,013	97.81	20,540	98.06	27,021	93.24	18,547	97.94	18,203	98.65
Swissprot	22,496	70.95	15,905	75.93	21,247	73.32	15,513	81.92	15,109	81.88
KEGG	22,141	69.83	14,699	70.18	20,184	69.65	14,623	77.22	14,044	76.11
COG	10,814	34.11	7,854	37.50	10,526	36.32	7,715	40.74	7,662	41.52
TrEMBL	30,964	97.66	20,489	97.82	26,828	92.58	18,477	97.57	18,193	98.60
Interpro	22,744	71.73	18,911	90.28	25,401	87.65	15,537	82.05	15,134	82.02
GO	18,894	59.59	13,811	65.94	15,182	52.39	11,505	60.75	11,877	64.37
Overall	31,074	98.00	20,574	98.22	27,118	93.58	18,573	98.08	18,236	98.83

16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

---

Unannotated	633	2.00	372	1.78	1,861	6.86	364	1.92	216	1.17
-------------	-----	------	-----	------	-------	------	-----	------	-----	------

---

826

827 **Table 8:** Orthologs of nitrogen fixation genes in *Vigna subterranea*, *Lablab purpureus*, *Faidherbia albida*, *Moringa oleifera* and *Sclerocarya*

828 *birrea*

---

Gene	<i>V. subterranea</i>	<i>L. purpureus</i>	<i>F. albida</i>	<i>M. oleifera</i>	<i>S. birrea</i>
MtLYK3/LjNFR1	Vigsu176S22567_VIGSU	Labpu216S12485_LABPU	Faial2789S13350_FAIAL	—	—
MtNFP/LjNFR5	Vigsu189S04417_VIGSU	Labpu54S03611_LABPU	—	—	Scibi409S02347_SCLBI
MtDMI2/LjSYMVK	Vigsu107959S16599_VIGSU	Labpu4785S15752_LABPU	Faial1833S08172_FAIAL	Morol36160S02362_MOROL	Scibi5995S15146_SCLBI
LjCASTOR	Vigsu108012S17109_VIGSU	Labpu27S13484_LABPU	—	—	—
MtHMGR1	—	—	—	—	—
MtDMI1/LjPOLLUX	Vigsu108496S19983_VIGSU	Labpu4332S15101_LABPU	Faial363S16033_FAIAL	Morol36085S07630_MOROL	—
NSP1	Vigsu2922S08781_VIGSU	Labpu723S04373_LABPU	Faial1104S01086_FAIAL	Morol36102S01150_MOROL	Scibi5005S02593_SCLBI

15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Chang et al.

Orphan crop genomes

---

NSP2	Vigsu107793S01507_VIGSU	Labpu887S08157_LABPU	Faial757S23006_FAIAL	Morol36224S03158_MOROL	Sclbi2944S01716_SCLBI
CCaMK	Vigsu91S05737_VIGSU	—	Faial752S22546_FAIAL	—	—
MtIPD3/LjCYCLOPS	Vigsu104856S09608_VIGSU	Labpu701S17462_LABPU	—	—	Sclbi2578S10386_SCLBI
NIN	Vigsu273S23676_VIGSU	Labpu165S10337_LABPU	Faial788S23538_FAIAL	Morol36195S02810_MOROL	Sclbi2838S04948_SCLBI
MtCRE1/LjLHK1	—	Labpu2293S02028_LABPU	Faial1226S02883_FAIAL	—	—
NF-YA1	Vigsu107799S13964_VIGSU	Labpu193775S11413_LABPU	Faial246S12019_FAIAL	Morol36154S02289_MOROL	Sclbi406S12278_SCLBI
NF-YA2	—	—	Faial858S26716_FAIAL	—	—
MtERN1	Vigsu107612S00570_VIGSU	Labpu210S01798_LABPU	Faial719S21851_FAIAL	Morol36040S00658_MOROL	Sclbi1920S01196_SCLBI
MtERN2	Vigsu108137S07511_VIGSU	Labpu448S03276_LABPU	Faial4604S17896_FAIAL	—	—

---

829

**830 Figure legends****831 Figure 1: Phylogenetic and evolutionary analysis.**

832 Scale bar = 10 million years. Values at branch points indicate estimates of divergence  
833 time (million years ago, Mya); blue numbers show divergence time (Mya); red nodes  
834 indicate previously published calibration times. *V. sub* shows seeds of *Vigna*  
835 *subterranean*; *L. pur*, flowers of *Lablab purpureus*; *F. alb*, seed pods of *Faidherbia*  
836 *albida*; *S. bir*, fruit of *Sclerocarya birrea*; *M. ole*, flowers of *Moringa oleifera*.

**838 Figure 2: The groups of orthologs shared by the orphan crops.**

839 (A) Groups of orthologs shared between *Lablab purpureus* (*L. pur*), *Faidherbia albida*  
840 (*F. alb*), *Glycine max* (*G. max*), *Medicago truncatula* (*M. tru*) and *Vigna subterranea*  
841 (*V.sub*). (B) Groups of orthologs shared between *Sclerocarya birrea* (*S. bir*), *Moringa*  
842 *oleifera* (*M. ole*), *Carica papaya* (*C. pap*), *Citrus sinensis* (*C. sin*) and *Theobroma*  
843 *cacao* (*T. cac*). Venn diagram generated using [85].

**845 Figure 3: The common symbiosis signaling pathway among the orphan crops.**

846 Sixteen root nodulation symbiosis signal (Sym) pathway genes were identified in three  
847 legumes (*Vigna subterranea*, *Lablab purpureus* and *Faidherbia albida*) and two non-  
848 legumes (*Sclerocarya birrea* and *Moringa oleifera*). Lj, *Lotus japonicas*; Mt, *Medicago*  
849 *truncatula*; LCOs, Lipochitooligosaccharides.

**851 Figure 4: Percentages of transcription factors in five orphan species.**

852 Blastp was used to search against 58 plant transcription factor families obtained from  
853 PlantTFDB [69] (see Additional file 2: Table S14). In this figure, MADS includes M-  
854 type\_MADS and MIKC\_MADS. MYB includes MYB and MYB\_related. NF-YA/B/C

1 855 includes NF-YA, NF-YB and NT-YC. “Others” comprises 31 types of transcription  
2 856 factors (E2F/DP, Nin-like, TALE, YABBY, GeBP, BES1, DBB, CO-like, CPP, SBP,  
3  
4 857 STAT, WOX, BBR-BPC, CAMTA, AP2, ZF-HD, S1Fa-like, ARR-B, SRS, GRF, LSD,  
5  
6  
7 858 NF-X1, EIL, RAV, HRT-like, HB-PHD, VOZ, Whirly, SAP, LFY and NZZ/SPL) whose  
8  
9 859 percentage was less than 1%.

10  
11  
12 860

13  
14 861 **Figure 5: Identification of genes involved in the starch biosynthesis pathway.**

15  
16  
17 862 Genes identified as being involved in starch synthesis are shown in red. Numbers of  
18  
19 863 homolog genes are presented in Additional file 2: Table S11. AGP, ADP-glucose  
20  
21 864 pyrophosphorylase; AGPL, AGP large subunit; AGPS, AGP small subunit; PHOH,  
22  
23 865 starch phosphorylase H (cytosolic type); GBSS, granule-bound starch synthase; SS,  
24  
25  
26 866 soluble starch synthase; BE, starch branching enzyme; ISA, isoamylase; DPE, starch  
27  
28  
29 867 debranching enzyme.  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

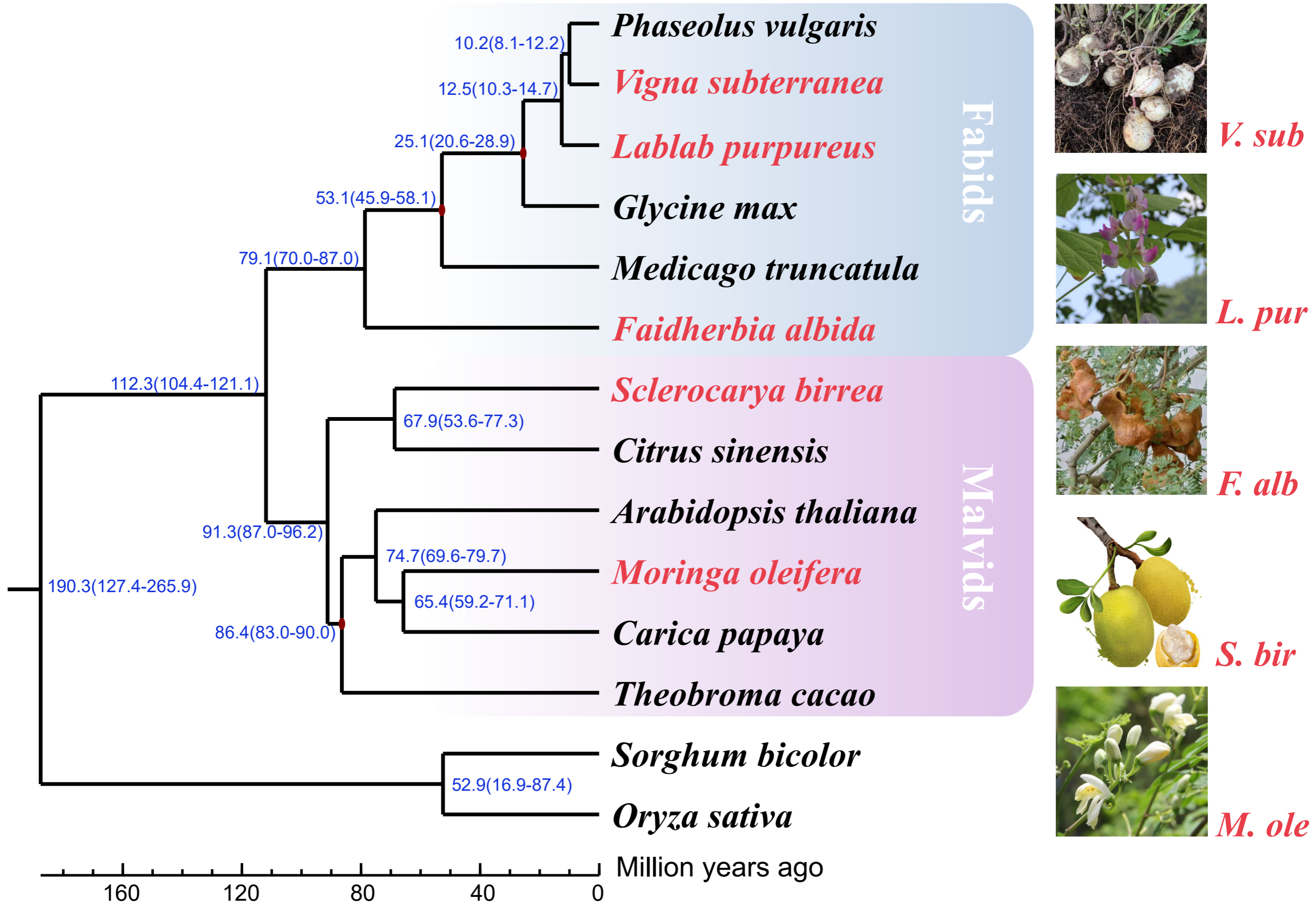
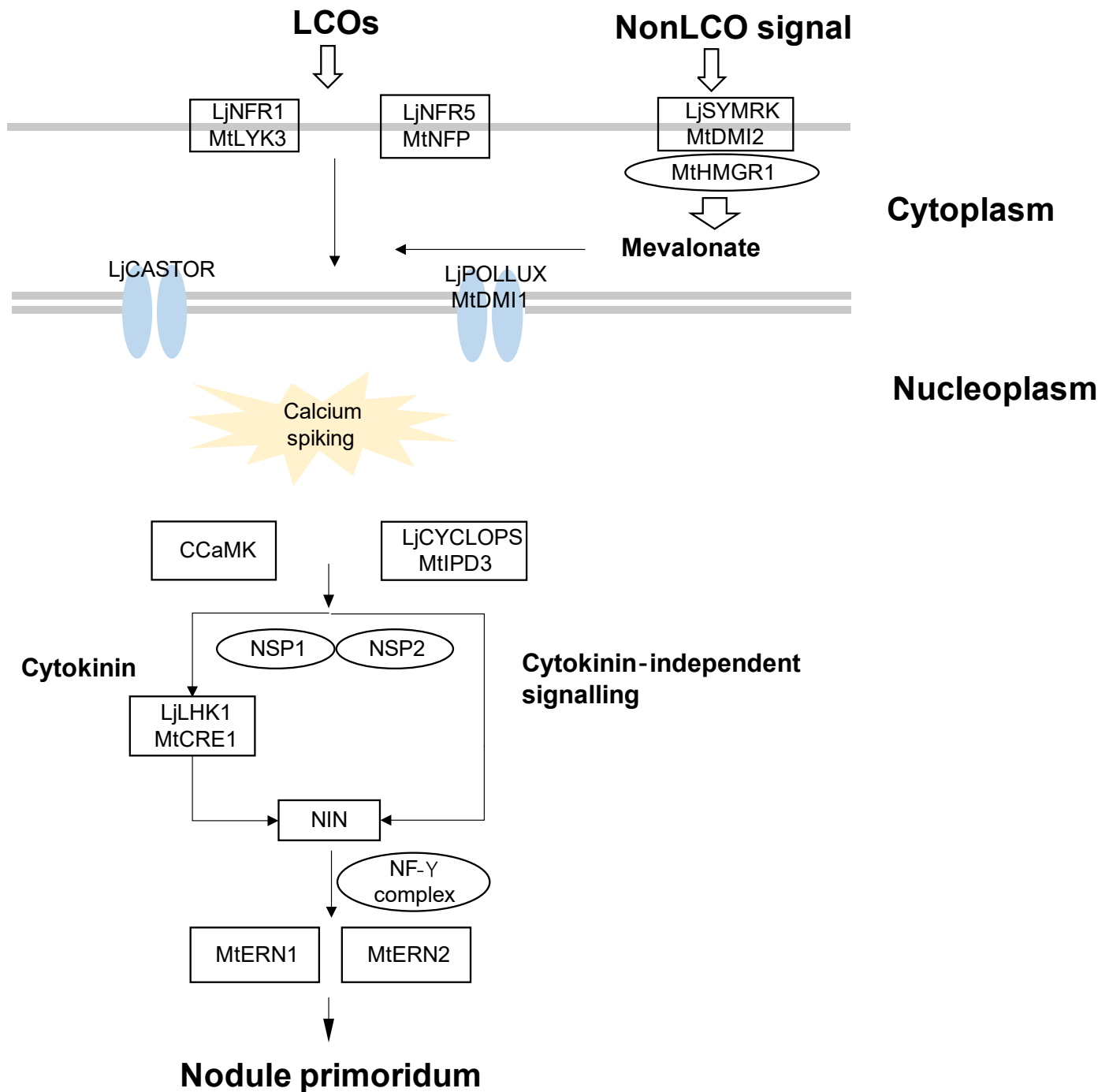
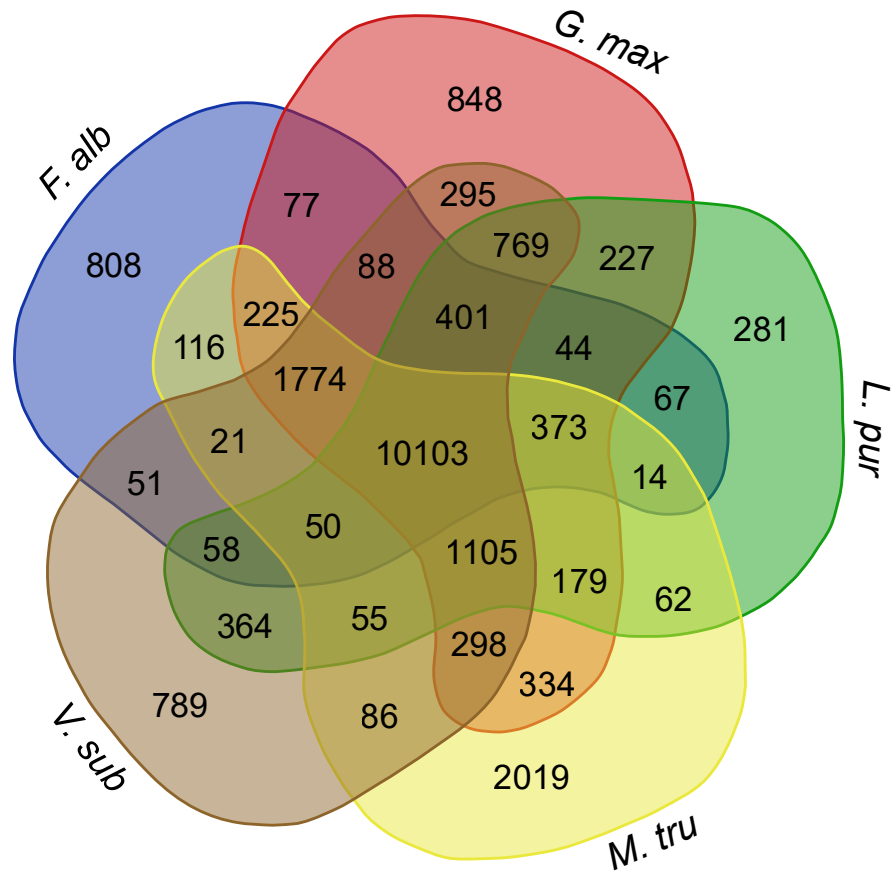


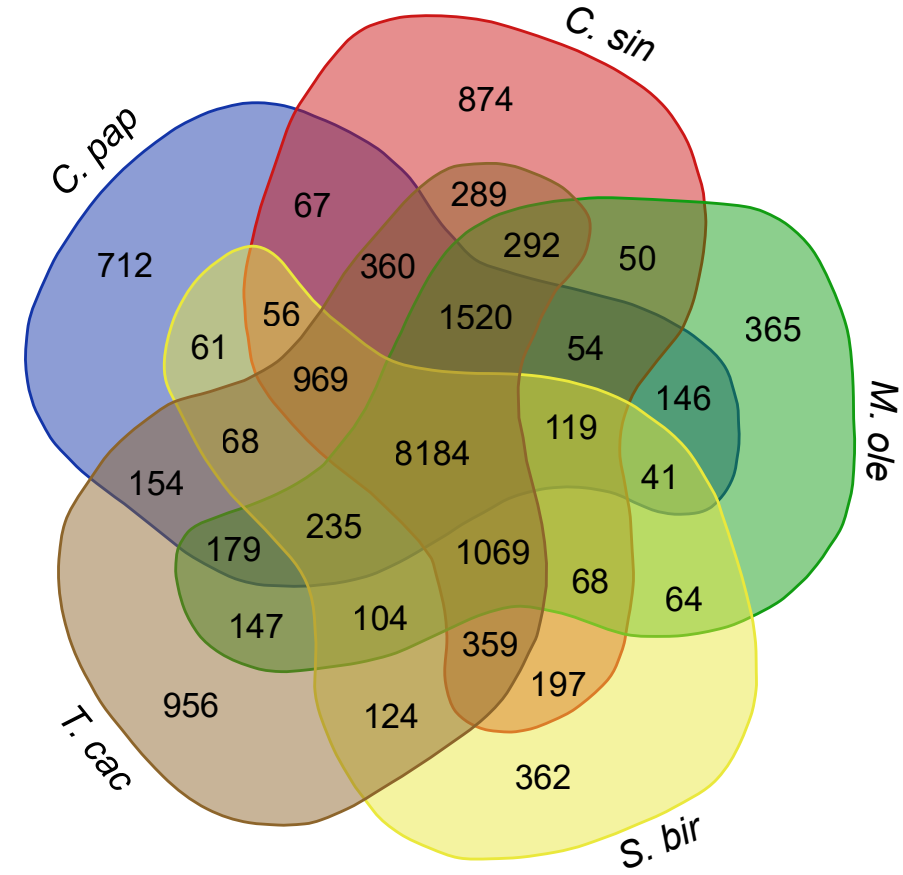
Figure 3



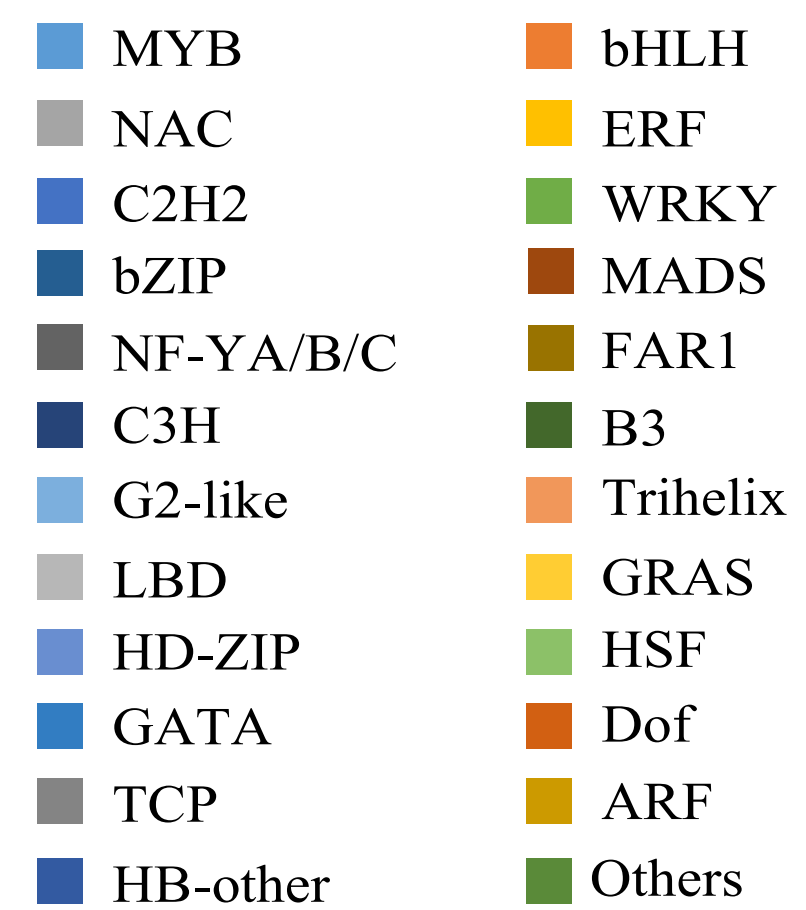
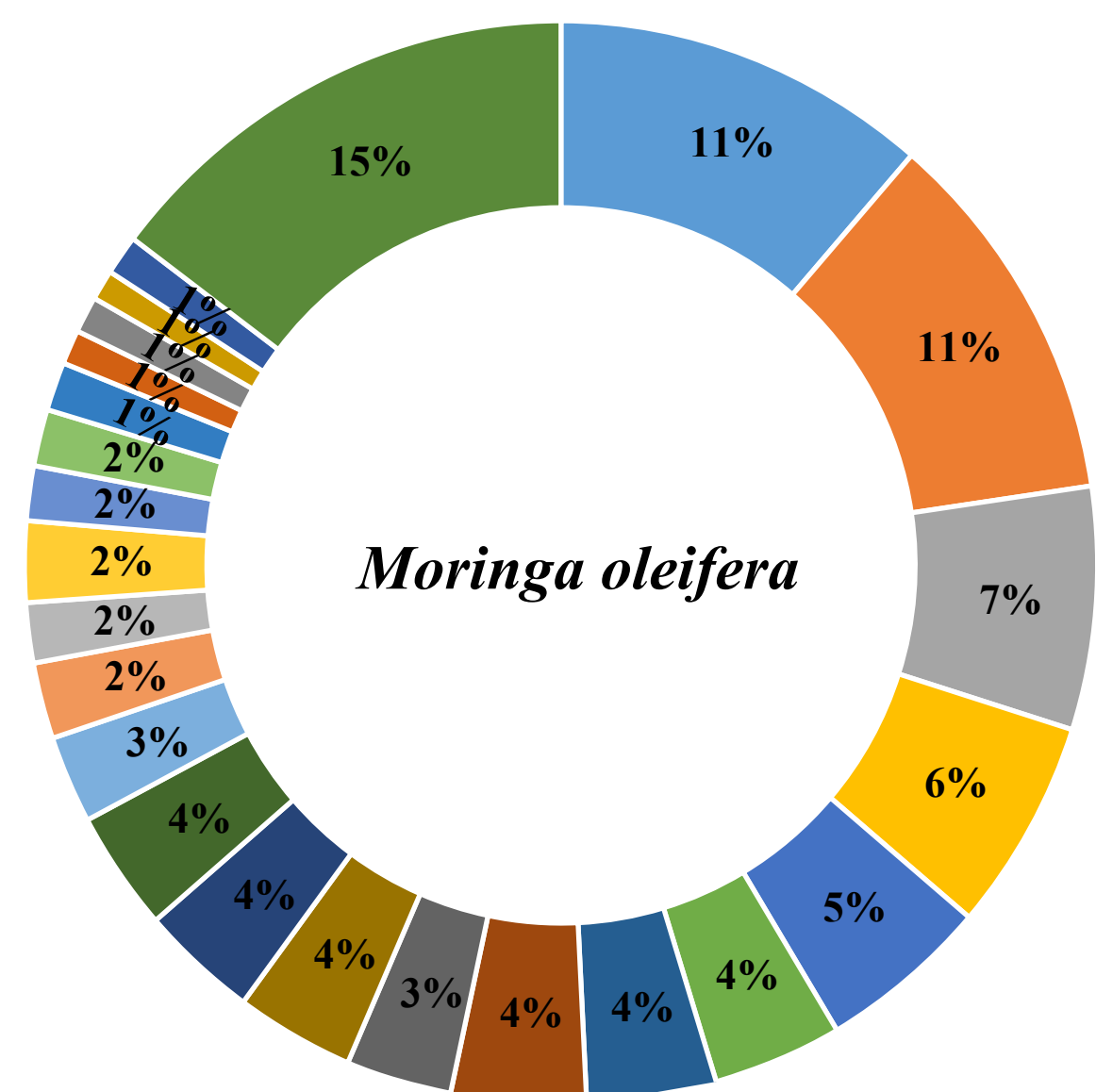
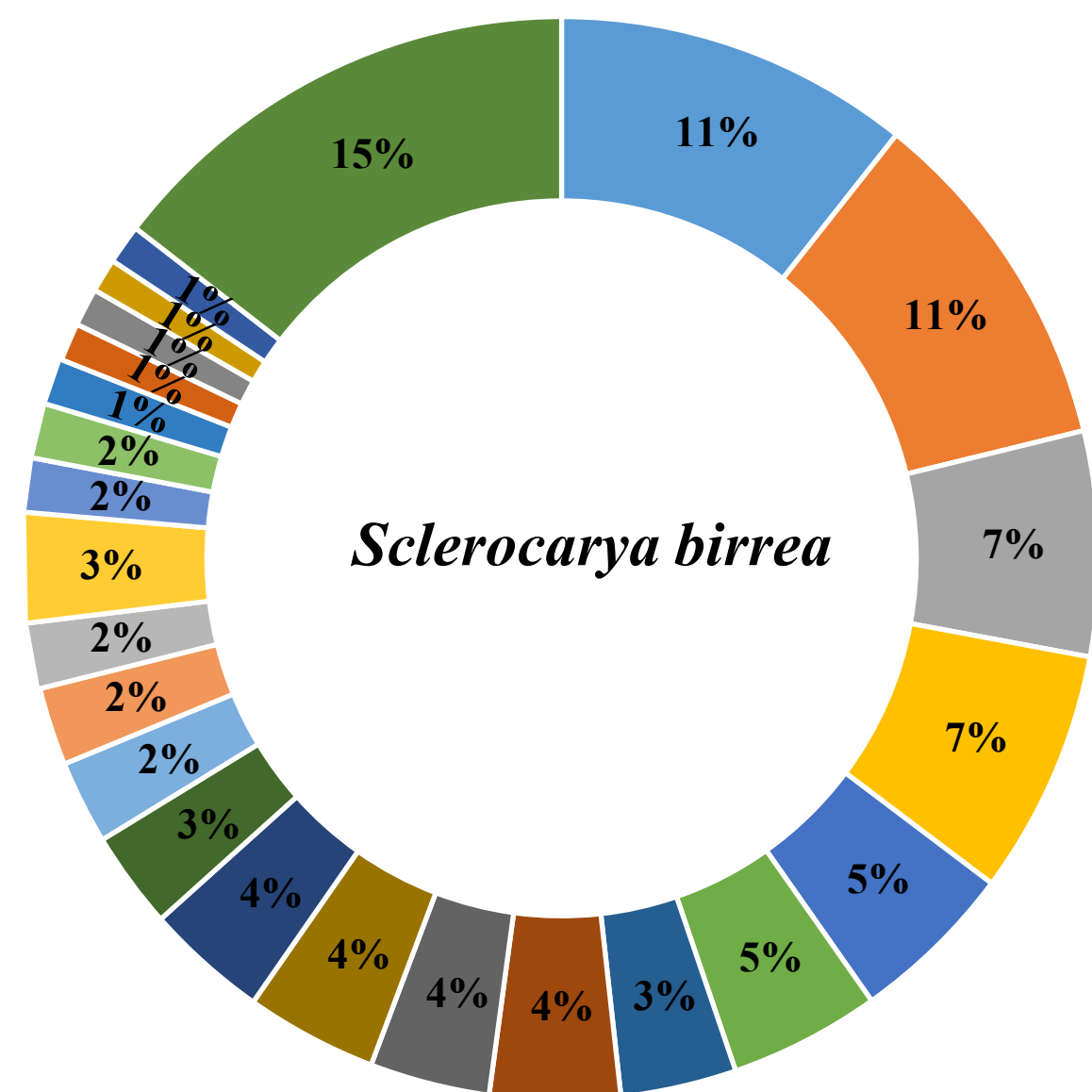
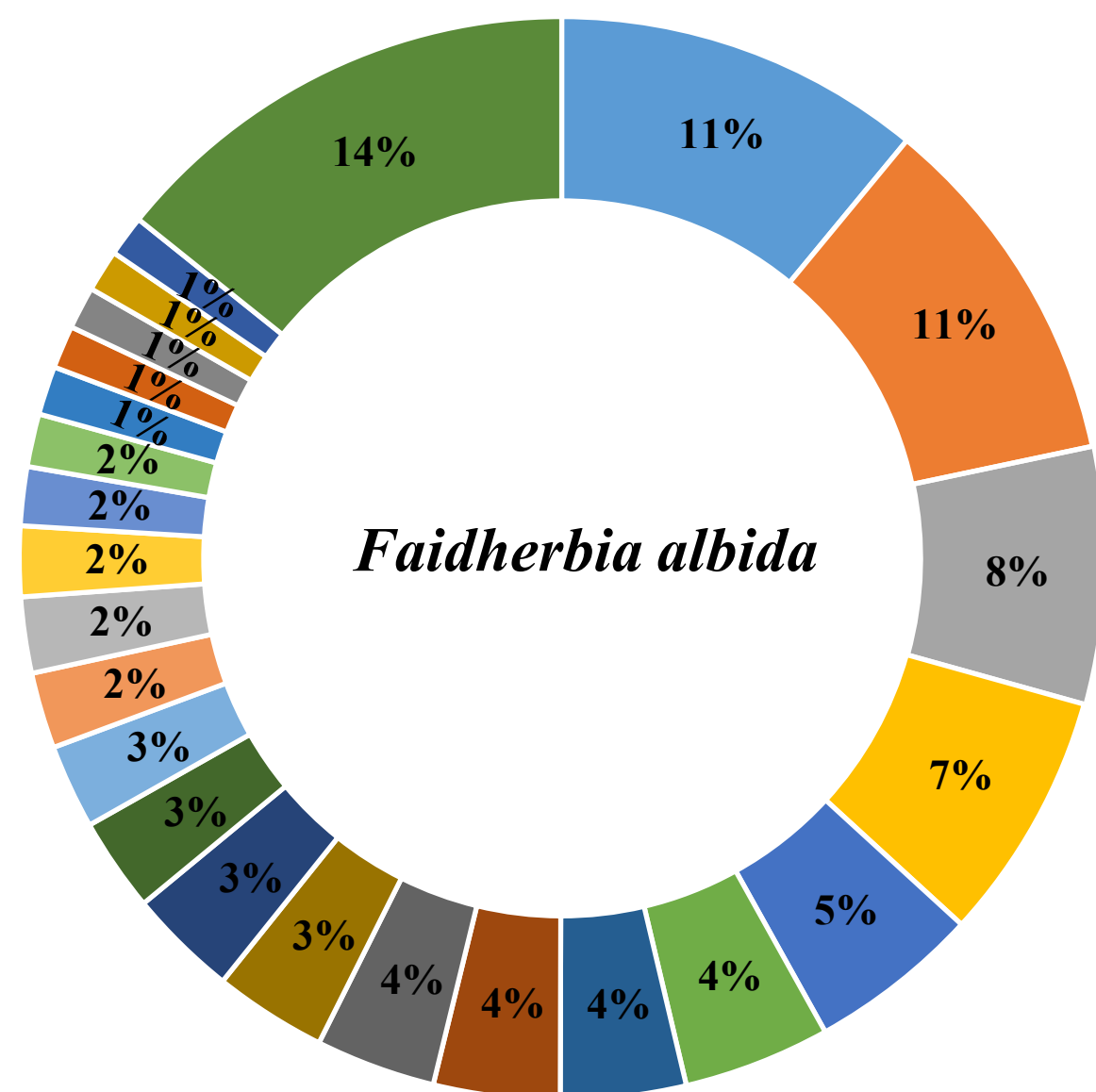
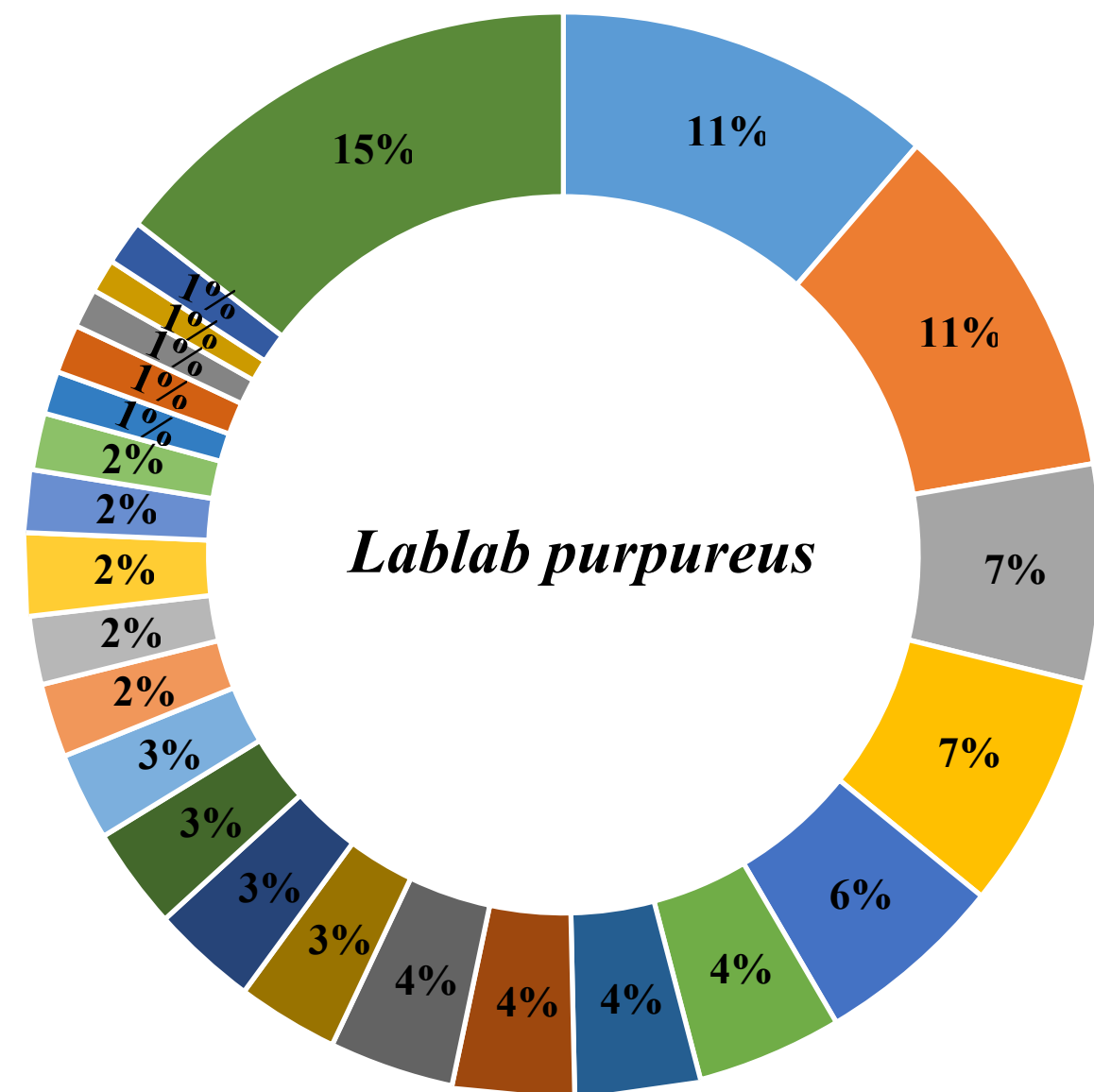
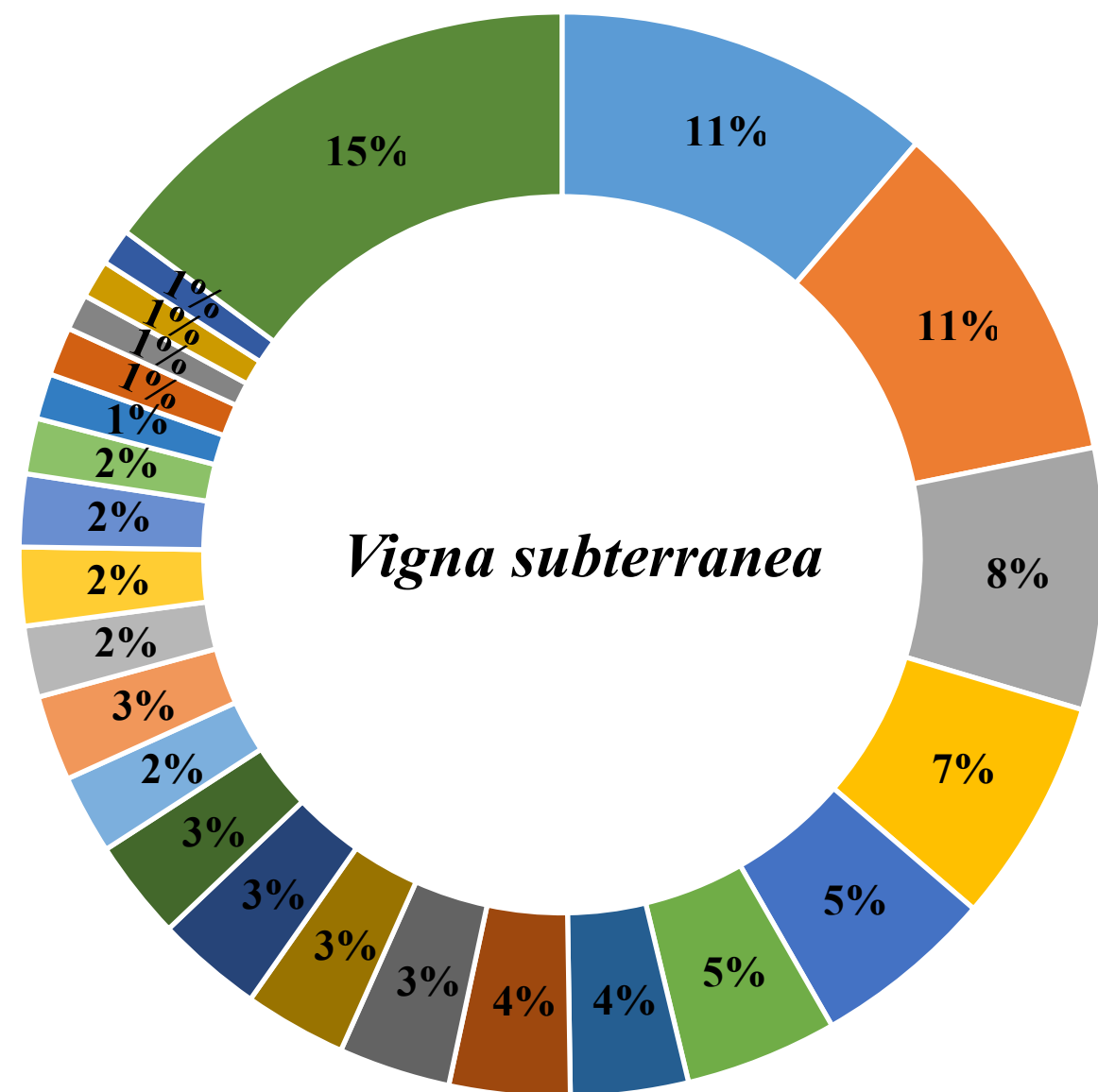


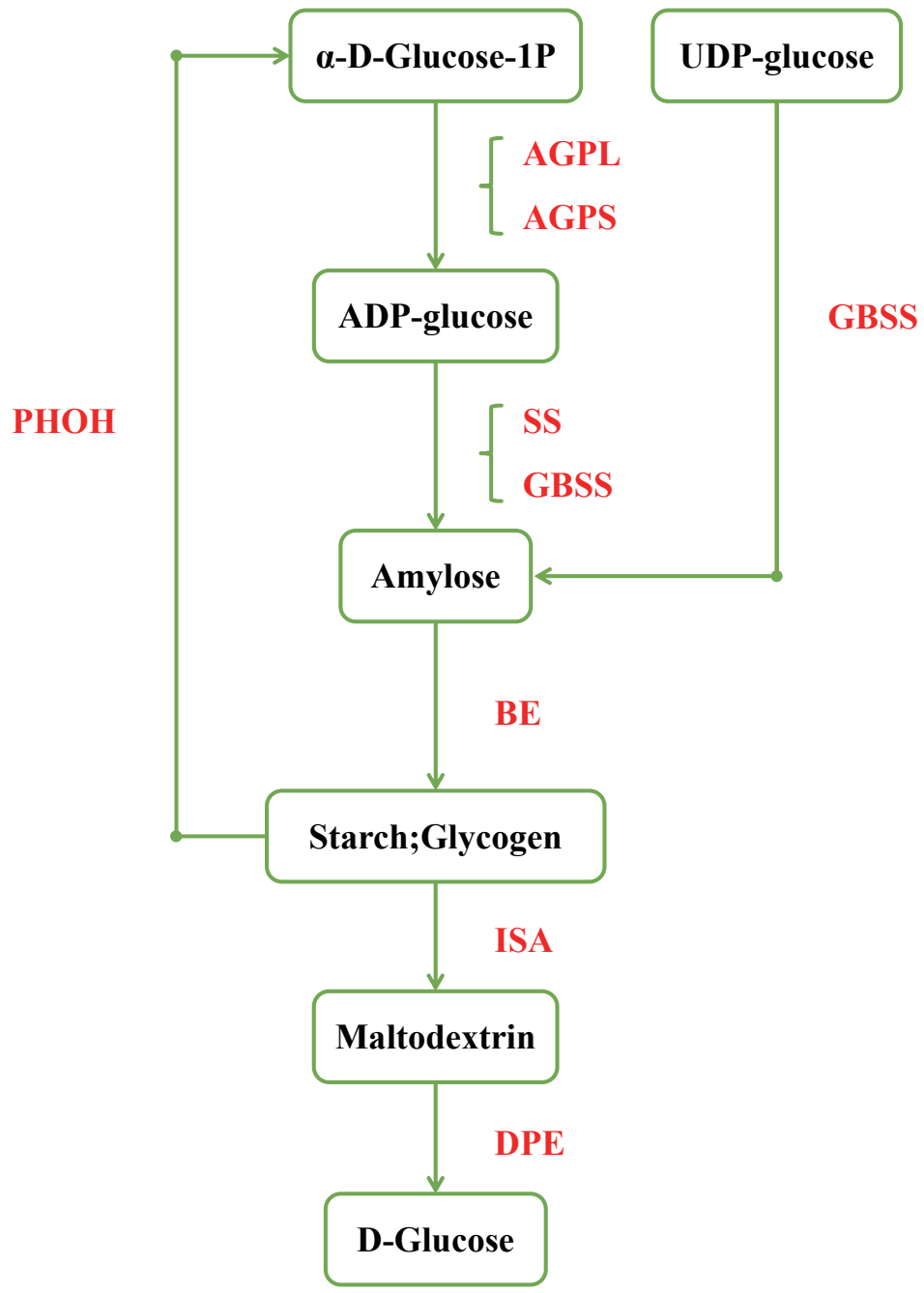


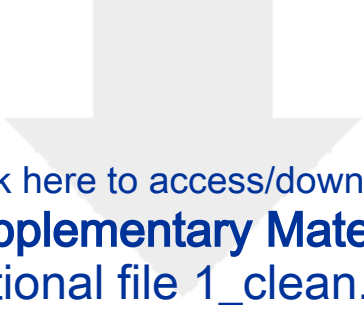
(A)



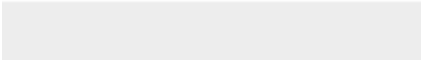

(B)

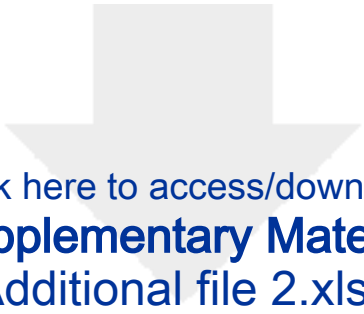







Click here to access/download  
**Supplementary Material**  
Additional file 1\_clean.docx





Click here to access/download  
**Supplementary Material**  
Additional file 2.xlsx



Dear Dr. Scott,

**Sub: Submission of the re-revised manuscript GIGA-D-18-00275R1**

We are glad to resubmit the thoroughly revised and cleaned version of our manuscript entitled “The draft genomes of five agriculturally important African orphan crops”, for possible publication in GigaScience as “Data Note”.

The corrections made and suggested by Dr. Lisa Martin were highly useful to significantly improve the quality of our manuscript. We have carefully implemented all the corrections suggested by her. Now we strongly believe that the revised manuscript is now ready for publication in GigaScience.

We look forward to hearing from you at your earliest convenience.

Yours sincerely,

Xin Liu

Reviewer reports:

Reviewer #1: The revisions in the last round satisfactorily addressed my concerns.

Reviewer #2: The revised version is well written and can be accepted for publication.

[Response: We thank all the reviewers for their valuable suggestions in the earlier version of the manuscript, and for giving their kind acceptance to our manuscript.](#)