Transcriptome-Derived Microsatellite Markers for Population Diversity Analysis in *Archidendron Clypearia* (Jack) I.C. Nielsen

Dandan Li
Research Institute of Tropical Forestry, Chinese Academy of Forestry
Mei Li
Research Institute of Tropical Forestry, Chinese Academy of Forestry
Fagen Li
Research Institute of Tropical Forestry, Chinese Academy of Forestry
Qijie Weng
Research Institute of Tropical Forestry Chinese Academy of Forestry
Changpin Zhou
Research Institute of Tropical Forestry, Chinese Academy of Forestry
Shineng Huang
Research Institute of Tropical Root and Tuber Crops: Instituto de Investigaciones de Viandas Tropicales
Siming Gan (siminggan@caf.ac.cn)
Research Institute of Tropical Forestry, Chinese Academy of Forestry https://orcid.org/0000-0001-6677-9860

Short Report

Keywords: Archidendron clypearia (Jack) I.C. Nielsen, Medicinal tree, Transcriptome, Genic microsatellite marker, Genetic diversity

Posted Date: June 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-580743/v1

Version of Record: A version of this preprint was published at Molecular Biology Reports on October 16th, 2021. See the published version at https://doi.org/10.1007/s11033-021-06773-4.

Abstract

Genomic resources including transcriptomic sequences and molecular markers remain scarce in the medicinally important woody legume genus *Archidendron* F. Mueller. Here we conducted transcriptome sequencing, genic microsatellite marker development, and population diversity analysis in *Archidendron clypearia* (Jack) I.C. Nielsen. Flower and flower bud transcriptomes were *de novo* assembled into 173,172 transcripts, with an average transcript length of 1597.3 bp and an N50 length of 2427 bp. A total of 34,701 microsatellite loci were identified from 26,716 (15.4%) transcripts. Primer pairs were designed for 718 microsatellite loci, of which 456 (63.5%) were polymorphic. Of the 456 polymorphic markers, 391 (85.7%) and 402 (88.1%) were transferable to *A. lucidum* (Benth.) I.C. Nielsen and *A. multifoliolatum* (H.Q. Wen) T.L. Wu, respectively. Using a subset of 15 microsatellite markers, relatively high genetic diversity was detected over two *A. clypearia* populations, with overall mean expected heterozygosity (H_e) being 0.707 and demonstrating the necessity of conservation. Relatively low differentiation between the two populations was revealed despite the distant separation (about 700 km), with overall inbreeding coefficient of sub-population to the total population (F_{st}) being 8.7%. This suggests that *A. clypearia* has mainly an outcrossing mating system and weak genetic structure. These results will offer valuable resources and information for further genetic studies and practical applications in *Archidendron* and the related taxa.

Introduction

Archidendron F. Mueller is a recently renewed woody genus (tribe Ingeae, subfamily Caesalpinioideae DC., family Leguminosae Juss. or Fabaceae Lindl.) of approximately 100 species, which takes over the two earlier genera *Cylindrokelupha* Kosterm. and *Paralbizzia* Kosterm. and most species of *Pithecellobium* Martius [1, 2]. *Archidendron* species occur widely in tropical Asia [1], ranging from India and southern China to New Guinea (http://www.asianplant.net/). Their leaves, twigs, and branches contain diverse structures such as polyphenols, flavonoids, lignans, and terpenoids that have anti-virus, anti-bacterium, anti-allergy, and/or anti-oxidation functions, and have been long used as herbal resources for curing upper respiratory tract infection, acute pharyngitis, tonsillitis, and gastroenteritis [3]. Of the genus, *Archidendron clypearia* (Jack) I.C. Nielsen (syn. *Pithecellobium clypearia* Benth) represents the most important species for medicinal and industrial applications [4].

Molecular marker technology provides powerful tools for a range of applications, such as population diversity investigation and variety fingerprinting. However, so far to our knowledge, no molecular marker has been developed for the genus *Archidendron*. Though eight microsatellite (or simple sequence repeats, SSR) markers were reported in the closely related genus *Pithecellobium* [5, 6], their transferability to *Archidendron* has not been investigated yet. Moreover, *A. clypearia* has become a species with extremely small populations due to human activities [7], and its genetic diversity remains to be investigated, especially using molecular markers.

Here we present the first report on transcriptome-derived SSR markers in the genus *Archidendron*. Transcriptome sequencing (or RNA sequencing, RNA-seq) has emerged as an innovative tool for generating large expressed sequence tag (EST) data, comprehensive transcriptome profiling, and molecular marker development in many organisms [8]. SSR markers are sound for many applications due to such characteristics as co-dominance, multi-allelism, high reproducibility, and abundance within the genome [9, 10]. Further,

transcriptome-derived SSR markers represent functionally transcribed genomic loci and are most likely transferable across related species [10]. Thus, the objectives of this study were to develop polymorphic EST-SSR markers in *A. clypearia*, test their cross-species transferability, and investigate the genetic diversity of and differentiation between two *A. clypearia* populations.

Materials And Methods

Plant material, RNA, and DNA isolation

A mature tree of *A. clypearia* growing in Huolushan Forest Park (23°10'51" N, 113°23'26" E), Guangdong, China was used for sampling flower and flower bud for RNA isolation and sequencing. The tree and other two *A. clypearia* mature trees from Erlongshan Ecological Park (EEP population, 23°21'08" N, 113°44'08" E), Guangdong, China were leaf sampled for screening primer pairs of effective polymerase chain reaction (PCR). Also, 16 additional mature trees were leaf collected from EEP (totaling at 18 presumably unrelated trees with at least 50 m distance apart) and included for estimation of marker polymorphism and population diversity. For cross-species transferability investigation, three trees of each of *A. lucidum* (Benth.) I.C. Nielsen and *A. multifoliolatum* (H.Q. Wen) T.L. Wu were leaf sampled from Nongla Ecological Tourism Area (23°39'24" N, 108°19'56" E), Guangxi, China and Zengcheng Forest Farm (23°18'05" N, 113°48'34" E), Guangdong, China, respectively.

Besides EEP population, 16 unrelated trees (at least 50 m apart) were leaf sampled from Jianfengling National Forest Park (JNFP population, 18°44'39" N, 108°49'55" E), Hainan, China for genetic diversity and differentiation analyses. As *A. clypearia* is characteristic of extremely small populations [7], these two populations are the largest we have ever found in China.

Total RNA was isolated from flower and flower bud samples using the EASYspin Plus Plant RNA Kit (Aidlab Biotechnologies, Beijing, China). Genomic DNA was extracted from leaf samples using a modified cetyltrimethyl ammonium bromide method [11].

RNA sequencing, De novo **assembly of transcriptomes, and SSR identification**

A cDNA library was constructed per sample using a VAHTS® Stranded mRNA-seq Library Prep Kit for Illumina (Vazyme Biotech Co., Nanjing, Jiangsu Province, China) and sequenced on a NovaSeq 6000 system (Illumina Inc., San Diego, CA, USA) using paired-end 150 bp read chemistry. Raw reads were filtered out for adaptors and low quality reads (≥ 5.0% uncertain nucleotides, ≥ 20.0% low-quality bases with Q ≤ 20, and final read length ≤ 85 bp) using Trimmomatic 0.39 [12]. The clean reads of the two transcriptomes were *de novo* assembled into contigs using Trinity 1.6 [13] under default parameters. To reduce the assembly redundancy, identical or nearly identical contigs were clustered to define the final transcript using Cd-hit-est [14] with an identity threshold of 95.0%.

The transcripts were subjected to SSR identification using MSATCOMMANDER 0.8.28 [15], in which di-, tri-, tetra-, penta-, and hexa-nucleotide repeat microsatellite was determined at a minimum of 12, 12, 16, 20, and 24 bases, respectively. A total of 718 primer pairs were then designed for di-, tri-, and tetra-nucleotide microsatellites using GMATA [16].

Marker Amplification And Genotyping

All primer pairs were screened for effective PCR against the equimolar DNA mixture of the three mature trees. Routine PCR (10 μL) was performed as described earlier [11] with specific melting temperature (T_m; 58, 56, or 60 °C).

The effective primer pairs each with an amplicon less than 500 bp (the maximal size of internal standard in SSR detection) were included in polymorphism estimation and cross-species transferability investigation. Fifteen EST-SSR markers (Table 1) were selected to genotype the 16 individuals of JNFP population. PCR and SSR detection were performed following a fluorescent-dUTP-based SSR genotyping protocol [11]. SSR markers were named with the prefix of ARCeSSR (*Archidendon* EST-SSR) and the suffix of sequential number (three numerals).

Table 1
The diversity parameters of 15 expressed sequence tag (EST) derived simple sequence repeats (EST-SSR) markers over the two populations of *Archidendron clypearia* (Jack) I.C. Nielsen

Serial no.	EST-SSR	N_a	N_e	H_o	H_e	N_{pa}	N_{da}	F_{is}	F_{it}	F_{st}	HWE
1	ARCeSSR141	9	4.3	0.882	0.769	3	0	−0.181	−0.140	0.035	
2	ARCeSSR266	8	4.6	0.853	0.782	5	2	−0.330	−0.098	0.174	
3	ARCeSSR277	11	5.8	0.824	0.827	6	1	−0.046	0.005	0.049	
4	ARCeSSR304	16	9.1	1.000	0.890	8	1	−0.166	−0.121	0.039	
5	ARCeSSR665	14	9.8	0.971	0.898	7	3	−0.133	−0.081	0.046	
6	ARCeSSR425	10	5.2	0.882	0.807	4	0	−0.170	−0.092	0.067	
7	ARCeSSR464	7	4.1	0.618	0.756	2	0	0.073	0.180	0.116	
8	ARCeSSR006	8	2.9	0.735	0.654	6	2	−0.236	−0.139	0.079	***
9	ARCeSSR095	16	10.2	1.000	0.902	8	1	−0.157	−0.109	0.041	
10	ARCeSSR075	7	3.8	0.853	0.740	5	2	−0.387	−0.165	0.160	
11	ARCeSSR649	10	5.5	0.882	0.820	6	2	−0.270	−0.072	0.155	
12	ARCeSSR288	6	2.9	0.647	0.652	4	2	−0.145	0.013	0.138	
13	ARCeSSR366	9	4.8	0.853	0.791	2	0	−0.097	−0.072	0.022	
14	ARCeSSR448	7	2.0	0.500	0.494	5	1	−0.094	0.003	0.088	
15	ARCeSSR474	8	4.9	0.882	0.795	2	0	−0.220	−0.107	0.092	***
Total		146	79.9	–	–	73	17	–	–	–	–
Mean		9.7	5.33	0.825	0.772	4.9	1.1	−0.171	−0.066	0.087	–

N_a number of alleles, N_e number of effective alleles, H_o observed heterozygosity, H_e expected heterozygosity, N_{pa} number of private alleles being present only in one population, N_{da} number of diagnostic alleles being present in only one population at a frequency greater than 10.0%, F_{is} inbreeding coefficient of individuals relative to the subpopulation, F_{it} inbreeding coefficient of individuals relative to the total population/species, F_{st} inbreeding coefficient of sub-population relative to the total population/species, HWE Hardy-Weinberg equilibrium, *** departure at 0.001 significance level with Bonferroni correction

Data analysis

For each EST-SSR marker, the number of alleles (N_a), number of effective alleles (N_e), allele size range (*ASR*), and fixation index (*F*) were calculated over the 18 trees of EEP population using GenAlEx 6.5 [17]. The polymorphism parameters including observed heterozygosity (H_o), expected heterozygosity (H_e), and polymorphic information content (*PIC*) were also estimated across EEP population using GenAlEx 6.5 [17].

For the two populations EEP and JNFP, Hardy-Weinberg equilibrium (HWE) was tested using GenAlEx 6.5 [17], with manual Bonferroni correction. Number of private alleles (N_{pa}) being present in only one population, number of diagnostic alleles (N_{da}) being present in only one population at a frequency greater than 10.0%, and inbreeding coefficients of individuals relative to the sub-population (F_{is}) and to the total population/species (F_{it}) and of sub-population to the total population/species (F_{st}) were also calculated using GenAlEx 6.5 [17]. Evidence for a recent bottleneck was investigated by sign and Wilcoxon sign-rank tests using two-phase model (TPM) and stepwise mutation model (SMM) in BOTTLENECK 1.2.02 [18]. In addition, allele frequencies were assessed for deviation from a normal L-shaped distribution, being indicative of possible bottleneck through loss of rare alleles.

Results And Discussion

The statistics data of *Archidendron clypearia* transcriptome sequencing are listed in Supplementary Table S1. After assembly redundancy reduction, a final number of 173,172 transcripts were retained with an average transcript length of 1597.3 bp and an N50 length of 2427 bp. These length values were remarkably higher than those of other medicinal leguminous plants, e.g. an average transcript length of 626 bp and an N50 length of 987 bp in *Mucuna pruriens* (L.) DC. [19].

A total of 34,701 SSR loci were identified from 26,716 (15.4%) transcripts. Of the di- to penta-nucleotide repeat motifs, the most frequent was tri-nucleotide (17,751, 51.2% of the 26,716 transcripts), followed by di- (14,760, 42.5%), tetra- (1557, 4.5%), hexa- (378, 1.1%), and penta-nucleotide (254, 0.7%; Supplementary Table S2). The three most abundant repeat motifs were AG/CT (6908, 19.9%), AAG/CTT (5614, 16.2%), and AC/GT (3933, 11.3%; Supplementary Table S2). The prevalence of AG/CT motif is consistent with most plant species [20].

Out of the 718 primer pairs designed, 644 (89.7%) were of effective amplification. After excluding 70 effective primer pairs each with an amplicon larger than 500 bp, 456 polymorphic (Supplementary Table S3), 99 monomorphic, and 19 no-fluorescent-signal SSRs were identified. For all the polymorphic markers, N_a, H_o, H_e, and *PIC* ranged between 2–16 (mean 3.0), 0.0–1.0 (mean 0.776), 0.056–0.944 (mean 0.511), and 0.053–0.912 (mean 0.409), respectively (Supplementary Table S4). The polymorphism estimates are generally higher than those reported in other legumes, e.g. *PIC* and H_e being 0.24 and 0.41 in *M. pruriens* [19] and 0.1956 and 0.1081 in *Cyamopsis tetragonoloba* (L.) Taub. [21].

Of the 456 polymorphic markers, 391 (85.7%) and 402 (88.1%) were transferable to *A. lucidum* and *A. multifoliolatum*, respectively, with 376 shared by both species (Supplementary Table S4). The cross-species transferability levels are higher than those reported for 89 EST-SSR markers from *C. tetragonoloba* in two *Cyamopsis* species (82% and 69%) [21].

The 15 polymorphic EST-SSR markers resulted in a total of 146 alleles over the two populations, with 6–16 alleles per marker (mean 9.7; Table 1). Only two markers showed significant deviation from HWE expectation after Bonferroni correction (Table 1). Locus H_o and H_e ranged from 0.500 to 1.000 (mean 0.825) and from 0.494 to 0.902 (mean 0.772), respectively (Table 1). Population H_o and H_e were 0.815 and 0.696 for EEP and 0.838 to 0.719 for JNFP (mean 0.826 and 0.707), respectively (Table 2). The genetic diversity is relatively high as compared to other legumes, e.g., H_e being 0.41 in *M. pruriens* [19] and 0.1081 in *C. tetragonoloba* [21]. This

result demonstrates that the two populations are clearly needed to conserve the genetic diversity. Additionally, the higher H_o than H_e estimates indicate certain magnitude of heterozygote excess.

Table 2
Genetic diversity estimates for two *A. clypearia* populations

Population	*N*	N_a	N_e	H_o	H_e	*F*
EEP	18	113	57.4	0.815	0.696	−0.171
JNFP	16	106	69.2	0.838	0.719	−0.164
Total	34	146	79.9	–	–	–
Mean	–	–	–	0.826	0.707	−0.168

N_a, number of alleles; N_e, number of effective alleles; H_o, observed heterozygosity; H_e, expected heterozygosity; *F*, Wright's fixation (inbreeding) coefficient; EEP, Erlongshan Ecological Park; JNFP, Jianfengling National Forest Park

The indices *F* and F_{is} suggested certain degree of inbreeding, with general mean of *F* and F_{is} being − 0.168 (Table 2) and − 0.171 (Table 1). On the other hand, there were 73 (50.0%) private and 17 (11.6%) diagnostic alleles over EEP and JNFP populations (Table 1), indicating some difference between the two populations. An overall F_{st} of 8.7% (0.087; Table 1) was detected, suggesting relatively low level of population differentiation despite the long distance (about 700 km) separation of the two populations. This also indicates that *A. clypearia* has mainly an outcrossing mating system as population differentiation coefficient is usually less than 19% for outcrossing species [22]. In addition, significant bottleneck effect and allelic frequency distribution mode shift were not detected for EEP and JNFP populations under TPM and SMM models (Supplementary Table S5), indicating no recent bottleneck experienced by both populations in spite of certain magnitude of heterozygote excess.

Conclusions

This study represents the first attempt to conduct transcriptome sequencing, SSR marker development, and population genetics analysis in the medicinally important genus *Archidendron*. High quality of flower and flower bud transcriptomes (52,895,540 raw reads, 15.87 Gb in total) were generated and *de novo* assembled in *A. clypearia*. A relatively large number (456) of EST-SSR markers were developed in *A. clypearia*, and high cross-species transferability was revealed in *A. lucidum* (85.7%) and *A. multifoliolatum* (88.1%). The two *A. clypearia* populations analyzed need to be conserved considering their relatively high genetic diversity. Relatively low population differentiation in *A. clypearia* suggests that the species has mainly an outcrossing mating system and weak genetic structure. These results will offer valuable resources and information for further population genetics analysis and breeding applications in *Archidendron* and the related taxa.

Declarations

Acknowledgments The authors would like to thank Yaqin Wang and Jiabin Lv for valuable assistance in microsatellite marker experiments and data analyses. We are also grateful to Jingjing Yan for kind help with

collection of JNFP samples.

CRediT authorship contribution **D. Li:** Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization; **M. Li:** Conceptualization, Methodology, Resources, Writing - review & editing, Project administration, Funding acquisition; **F. Li:** Formal analysis, Investigation, Data curation; **Q. Weng:** Investigation, Resources; **C. Zhou:** Investigation, Resources, Data curation; **S. Huang:** Investigation, Resources; **S. Gan:** Conceptualization, Methodology, Formal analysis, Resources, Data curation, Writing - original draft, Writing - review & editing, Supervision.

Funding This work was financially supported by the Fundamental Research Funds of Chinese Academy of Forestry (CAFYBB2018SY019) and the Public Welfare Research and Capacity Building Funds for Social Development from the Department of Science and Technology of Guangdong Province, China (2017A020213002).

Data availability The raw reads of transcriptome sequencing were deposited in the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra/) under BioProjects PRJNA675372 and PRJNA675437 for flower and flower bud, respectively. SSR primer sequences and other data were available from the supplementary tables.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Informed consent All authors consent to publication of these data.

References

1. Wu DL, Nielsen IC (2010) Tribe Ingeae. In: Xu L, Chen D, Zhu X et al (eds) Fabaceae (Leguminosae), flora of China, vol 10. Science Press, Beijing, China & Missouri Botanical Garden, St. Louis, pp 60–71
2. Jiang K, Pan B, Tian B (2019) Recent taxonomic changes for Fabaceae (Leguminosae) genera in China. Biodiv Sci 27:689–697. http://doi.org/10.17520/biods.2019032
3. Liu L-Y, Kang J, Chen R-Y (2013) Research progress in chemical constituents and pharmacological activities of plants in *Pithecellobium* Mart. Chin Trad Herb Drug 44:2623–2629
4. Li M, Huang S, Chen Z et al (2018) The research status and utilization prospect of medicinal tree species of *Archidendron clypearia*. Sci Silvae Sin 54:142–154. http://doi.org/10.11707/j.1001-7488.20180417
5. Chase M, Kesseli R, Bawa K (1996) Microsatellite markers for population and conservation genetics of tropical trees. Am J Bot 83:51–57. https://doi.org/10.1002/j.1537-2197.1996.tb13873.x
6. Chase MR, Moller C, Kesseli R et al (1996) Distant gene flow in tropical trees. Nature 383:398–399. https://doi.org/10.1038/383398a0
7. Ma XY, Li M, Jin WY et al (2017) Natural regeneration characteristics of *Archidendron clypearia*. Bull Bot Res 37:761–767. https://doi.org/10.7525/j.issn.1673-5102.2017.05.017
8. Wolf JBW (2013) Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Mol Ecol Resour 13:559–572. https://doi.org/10.1111/1755-0998.12109

9. Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends Plant Sci 1:215–222. https://doi.org/10.1016/1360-1385(96)86898-1
10. Varshney R, Granner A, Sorrells M (2005) Genic microsatellite markers in plants: features and applications. Trends Biotech 23:48–55. https://doi.org/10.1016/j.tibtech.2004.11.005
11. Li F, Gan S (2011) An optimised protocol for fluorescent-dUTP based SSR genotyping and its application to genetic mapping in *Eucalyptus*. Silvae Genet 60:18–25. https://doi.org/10.1515/sg-2011-0003
12. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170
13. Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652. https://doi.org/10.1038/nbt.1883
14. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. https://doi.org/10.1093/bioinformatics/btl158
15. Faircloth BC (2008) MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. Mol Ecol Resour 8:92–94. https://doi.org/10.1111/j.1471-8286.2007.01884.x
16. Wang X, Wang L (2016) GMATA: an integrated software package for genome-scale SSR mining marker development and viewing. Front Plant Sci 7:1350. https://doi.org/10.3389/fpls.2016.01350
17. Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–An update. Bioinformatics 28:2537–2539. https://doi.org/10.1093/bioinformatics/bts460
18. Piry S, Luikart G, Cornuet J-M (1999) BOTTLENECK: a computer program for detecting recent reductions in the effective population size using allele frequency data. J Hered 90:502–503. https://doi.org/10.1093/jhered/90.4.502
19. Sathyanarayana N, Pittala R, Tripathi P et al (2017) Transcriptomic resources for the medicinal legume *Mucuna pruriens*: *de novo* transcriptome assembly, annotation, identification and validation of EST-SSR markers. BMC Genom 18:409. https://doi.org/10.1186/s12864-017-3780-9
20. Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200. https://doi.org/10.1038/ng822
21. Tribhuvan KU, Mithra SVA, Sharma P et al (2019) Identification of genomic SSRs in cluster bean (*Cyamopsis tetragonoloba*) and demonstration of their utility in genetic diversity analysis. Ind Crop Prod 133:221–231. https://doi.org/10.1016/j.indcrop.2019.03.028
22. Hamrick JL, Godt MJW (1996) Effects of life history traits on genetic diversity in plant species. Philos Trans R Soc Lond B Biol Sci 351:1291–1298. https://doi.org/10.1098/rstb.1996.0112

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESMLiDetalTablesS1S5.pdf