Molecular characterization and genetic structure of American oil palm (*Elaeis oleifera*) based on genome-wide SNP markers

Valquiria M. Pereira[12]; Jaire A. Ferreira Filho[12]; André P. Leão[1]; Eduardo F. Formighieri[1]; Manoel T. Souza Junior[12], Sara A. Rios[3] and Alexandre A. Alves[1]*

Background

African oil palm (*Elaeis guineensis*, Jacq.) is by far the most productive oil crop and alone is capable to fulfill the large and growing world demand for vegetable oils that is estimated to reach 240 million tons by 2050 (BARCELOS et al. 2015) . Per hectare of cropland, oil palm plantations give 3–8 times more oil than any other temperate or tropical oil crop. The species is, however, susceptible to a wide range of diseases including bud rot (or fatal yellowing) and basal stem rot caused by *Ganoderma* sp. In African germplasm there are no good sources of natural genetic disease resistance. Also African oil palm trees can reach 15–18 meters in height, making its exploration challenging. Because of that, *Elaeis oleifera*, also known as, American Oil Palm is currently being used as source of variability for disease resistance, plant architecture and oil quality, in the genetic improvement programs of Oil palm (*E. guineensis*). These two species hybridize well producing fertile offspring. Therefore, genetic improvement programs are focusing its effort in generating and evaluating interspecific hybrids. Despite the fact that relatively large germplasm collections have been established (MORETZSOHN et al. 2002) and evaluated for both species, little is known about its population dynamics and patterns of genetic diversity in its natural range of occurrence. Here, we evaluated the genetic diversity patterns and structure of subpopulations of *E. oleifera* in the Amazon River basin.

Methods

To access the natural genetic diversity *Elaeis oleifera* we sampled 552 palms that represent collections made in 19 different sites (presumably populations) within 6 major regions. The number of plants analyzed per population varied between 2 and 140. Total genomic DNA was extracted from freeze-dried adult leaf tissue using a modified CTAB protocol and quantified using NanoDrop®. Samples were then genotyped using the DArTSeq

1 Embrapa Agroenergy, PqEB, Av W3 Norte (Final), Brasilia- DF, Brazil, 70770-901
2 Plant Biotechnology Graduate Program, Federal University of Lavras, Caixa Postal 3037, Lavras- MG, Brazil, 37200-000
3 Embrapa Western Amazon, Rodovia AM 10- s/n Km 28, Manaus- AM, Brazil. 69090-000.
valquiria.pereira@coloborador.embrapa.br, alexandre.alonso@embrapa.br

platform. Briefly, genomic DNA of the 5552 plants were sent to Australia in dry ice, and after checks of quantity and quality, it was treated with a combination of restriction enzymes (Res) (a common cut, e.g. *BstNI* and rare cutting, *PstI*) to reduce the complexity of the DNA being sequenced. The reduced complexity of DNA samples was then individually labeled by binding PstI adapters containing barcodes specifically designed for each of the 552 plants. The resulting products were amplified by PCR, and subsequently sets of 96 samples were mixed and sequenced together in a lane of Illumina HiSeq 2000 platform (SE 100bp). The *PstI* adapter includes a sequencing primer so that all the generated tags are always read from the *PstI* restriction site. The resulting sequences were filtered and allocated to their respective genotypes based on the barcodes. The sequences were then trimmed at 69 bp (5 bp of the restriction site plus 64 bases with a min Q score of 20). A proprietary pipeline of DArT Pty was be used to carry out the score of the presence absence variants (PAV markers). Cases in which the presence or absence of reads of the fragment resulting from restriction enzyme cleavage were not clear (i.e. less than 6x) were recorded as missing data. Reads almost identical (i.e. less than 3 polymorphisms) were combined so that one or more SNPs in the read would not confuse the analysis. In parallel the total sequences generated were used to generate a low-coverage consensus sequence that was used as reference, since by the time the genotyping was carried out no oil palm reference genome was available. The SNP and genotype calling was then be performed by aligning reads of 69 bp to the reference consensus sequence, using the bowtie v0.12 software. To characterize the polymorphic sites, we aligned the sequences on which SNPs were discovered, to the public available reference genome (RG) (SINGH et al. 2013) (portioned in 130 5M intervals) using BLASTN. This allowed us to check for genome-wide distribution of the markers, and to assign the SNPs to one of the 16 chromosomes. The molecular data were then used to estimate the allelic frequencies, the heterozygosity, the F statistics, population effective sizes (Ne) and the genetic distance among the subpopulations and among individuals using the PowerMarker v3.25 software. The genetic distance between the individuals was estimated according to the proportion of common alleles and grouping performed using the neighbor-joining method. Population structuration was then evaluated through a bayesian clustering analysis (Structure) and AMOVA. The most probable number of clusters in the Structure analysis was determined by ΔK value.

Results and Conclusions

The genotyping platform provided over 1.500 high-quality SNPs (Call Rate> 0.90; MAF> 0.05). Of these 812 were effectively mapped against the RG with an average of 6.25 SNPs per bin, providing relatively good coverage of all chromosomes. Based on this genome-wide set of SNP markers, we found only moderate genetic diversity (0.301) among the subpopulations. The neighbor joining analysis revealed the existence of 06 (six) major

groups, being the grouping pattern mainly related to the geographical origin of the samples. In general, samples from the same river basin tended to be grouped together, corroborating previous findings of Moretzhsohn et al. (2002). This same pattern was also evidenced by PCA analysis. The F index, which measures the inbreeding coefficient, was relatively low for all subpopulations. Conversely, the observed heterozygosity (Ho) was moderate (0.185 on average) which may be indicative that the subpopulations reproduce mainly by crosses between unrelated individuals. The F_{ST} index was, however, high (0.315 on average), demonstrating the existence of structuration, or genetic differentiation among the *E. oleifera* subpopulations. Structure analysis indeed revealed the existence of 02 to 03 major clusters and the AMOVA indicated that only 54% of the molecular variance occurs within populations. This results contrasts with previous findings that were indicative of major contribution of populations to the overall genetic variance. Based on the results reported in here, for *ex situ* conservation, new collections should include as many sites as possible. For breeding purposes, testing of different subpopulations in hybrids crosses should prove to be more effective based on the distribution of SNP variation within and among subpopulations.

Financial Support

This work was supported by PRODENDE and DENDEPALM grants provided by FINEP/MCTI to Embrapa and its partners. VMP and JAF were supported, respectively, with doctoral and masters fellowships provided by CNPq.

References

BARCELOS, E. et al. Oil palm natural diversity and the potential for yield improvement. **Frontiers in Plant Science**, Lausanne, v. 6, n. 1, p. 1-16, 2015.

MORETZSOHN, M. C. et al. Genetic diversity of Brazilian oil palm (*Elaeis oleifera* H.B.K.) germplasm collected in the Amazon Forest. **Euphytica**, Wageningen, v. 124, n.1, p. 35-45, 2002.

SING, R., et al. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. **Nature**, London, v. 500, n. 7462, p. 335-339, 2013.