SNP discovery of baru tree (*Dipteryx alata* Vogel) accessions enriches its genomic toolkit for species conservation, domestication, and breeding

Pessoa Filho, Marco (1); de Campos Telles, Mariana P. (2)(3); Coelho, Alexandre S. G. (2); Chaves, Lázaro J. (2); Soares, Thannya N. (2) André, Thiago (2).
(1) Embrapa Recursos Genéticos e Biotecnologia; (2) Universidade Federal de Goiás; (3) Pontifícia Universidade Católica de Goiás; marco.pessoa@embrapa.br

Baru seeds are protein-rich sources of nutrients with a supply chain primarily based on extractivism. The baru tree (*Dipteryx alata* Vogel, Fabaceae) is a neotropical species native to Latin American savannas, included as vulnerable in the IUCN Red List of Threatened Species. Conservation, domestication, and breeding efforts would benefit from a rich set of genomic resources, which already includes a draft genome assembly. We used whole-genome sequencing (WGS) of individual baru trees to discover and genotype single-nucleotide polymorphisms (SNPs) on a genome-wide scale. Trees belonging to 24 accessions of the baru germplasm collection at Universidade Federal de Goias were selected for WGS. Young leaf tissue was collected for DNA extraction, and an Illumina DNA library was prepared with Nextera DNA Flex Indexes. The library was paired-end sequenced (2 x 300 bp) on a NextSeq 1000. A customized pipeline with successive iterations of analyses on the GATK and FreeBayes was used to build a high-quality SNP database. Adapter-marked reads were mapped to the draft assembly with BWA, followed by the marking of duplicates. The HaplotypeCaller and GenotypeGVCFs tools of the GATK were used in a first round of genotyping and hard-filtering, generating a dataset for base-quality score recalibration (BQSR). Recalibrated BAMs were used to call SNPs with FreeBayes, followed by hard filtering based on QUAL and DP annotations. Intersection of common variants between the GATK and Freebayes was performed, followed by selecting biallelic variants with MAF > 0.05 and pruning for one variant every 150 bp. This variant resource was used for a first round of Variant Quality Score Recalibration to obtain a sensitive variant dataset with a truth sensitivity threshold of 95.0. This dataset was used on a second round of BQSR, after which the pipeline was rerun with the re-recalibrated BAMs, including a second round of VQSR and selection of variants in the truth sensitivity tranche 90.0. The final dataset includes 14,428,145 variants, of which 8,509,545 are biallelic with MAF > 0.05, representing one variant every 94 bp of the baru tree genome. Genome-wide linkage disequilibrium (LD) estimates from a pruned dataset of 2 million SNPs showed a mean value of 0.30 for r^2 and an LD decay of 5.4 kbp. Ongoing efforts of chromosome-scale scaffolding of the draft assembly and prediction and annotation of gene models will allow further selection of SNPs for germplasm characterization and breeding.

Apoio: CNPq, Embrapa.

Palavras-chave: Plant genetic resources. Plant genomics.