Increasing the genomics knowledge in Pennycress: a promissory crop for biofuel production.

Tatiana García[1], Cintia Arias[2], Eric Mukundi[1], Ana Paula Alonso[2] and Erich Grotewold[1]

[1]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, 48824, USA.

[2]BioDiscovery Institute and Department of Biological Sciences, University of North Texas, Denton, TX, 76203, USA.

Corresponding author e-mail address:garci496@msu.edu, grotewol@msu.edu

Thlaspi arvense (Pennycress) belongs to the Brassicaceae family and is native to the Mediterranean region. Pennycress survives temperatures below −30°C, and it can grow from fall to spring with minimal agricultural inputs. In the last several years, research interests in this plant have been steadily increasing because of its seed fatty acid composition and potential use as a biofuel crop. It has been reported that oil content/seed (% weight) in pennycress ranges between 20-36%, being highly competitive when compared to soybean with 18-22% and Camelina with 36-47%. From a genomic perspective, pennycress has a diploid genome (2n = 14) of approximately 539 Mb, with 27,390 gene models. The genomes of the annual winter line MN106 and inbred line Spring 32-10 are currently available. Besides, Pennycress anatomy is similar to Arabidopsis, and transformation using *Agrobacterium*-mediated floral dip is possible.

In this study, we improved the genome annotation by combining short and long-reads of RNA-Seq embryonic data of 22 pennycress accessions and the integration of public data. We identified 27,213 protein-coding genes, including 85% of the core genes for plant species. A total of 22,045 gene ontology terms were assigned. Besides, we determined new alternative splicing isoforms for 2,842 genes. We evaluated the natural variation of pennycress with potential use in pennycress breeding programs. As a result, we identified 6,188 SNP markers through NGSEP V4.0. Also, we estimated the transition/transversion rate (Ts:Tv), which was found to be 1.5. On the other hand, we used the population genomic analysis and the SNPs matrix to generate a phylogenetic tree, principal component analysis (PCA),

discriminant analysis of principal components (DAPC), and STRUCTURE using the admixture model. In all analyses, we identified a high differentiation between the accession from Armenia (Ames 3287) compared to other accessions whose samples were mainly clustered according to the geographical origin.

In addition to the Pennycress annotation improvement, we evaluated the 200 most variable genes in each stage (10 and 16 days after pollination). Interestingly, we found 92 genes with high variation in mRNA accumulation at both stages. Using the KEGG pathways database of those 200 most variable genes, we observed that the most significantly enriched process was “Protein processing in endoplasmic reticulum” for both developmental stages. We anticipate that this work will contribute to pennycress's basic genomic knowledge and provide resources for molecular breeding programs.