NitFix: Identification of Conserved Non-Coding Sequences in Nitrogen Fixing Plants

Wendell J. Pereira,[1] Jean-Michel Ané,[2] Pamela Soltis,[3] Douglas Soltis,[3] Robert Guralnick,[3] Ryan Folk,[3] Sushmita Roy,[2] and **Matias Kirst**[1]* (mkirst@ufl.edu)

[1]University of Florida, Gainesville, FL; [2]University of Wisconsin, Madison, WI; [3]University of Florida, Florida Museum of Natural History, Gainesville, FL

http://nitfix.org; Twitter: @nit_fix

Project Goals: The collaborative project "Phylogenomic discovery and engineering of nitrogen fixation into the bioenergy woody crop poplar" is focused on identifying the genomic novelties that enable the symbiotic relationship between nodulating plants and nitrogen-fixing bacteria to support genetically engineering this capability into bioenergy crops. We are using a comparative phylogenetic framework to contrast related species that possess and lack this ability. Genomic novelties are being evaluated for their effect on root nodule development in *Medicago* (nodulating) and poplar (non-nodulating). Genes identified are being engineered into poplar, to test their impact on nitrogen-fixation, whole-plant development and biomass productivity and composition. This abstract focuses on one component of the comparative phylogenetic framework, in which conserved non-coding sequences that may contribute to nodule development are being identified among species of the nitrogen-fixing clade.

Nitrogen (N) is one of the most important nutrients required for plant growth because of its fundamental role as a component of DNA, RNA, and amino acids. However, plants are not able to obtain N directly from the atmosphere, but depend on its availability in the soil, in the form of nitrate, ammonium, or amino acids. Yearly, approximately 120 million metric tons of N fertilizer are produced and used in agriculture. Nitrogen fertilization is responsible for a significant cost of crop production, and an important source of environmental contamination.

Several species of four angiosperm orders (Fabales, Fagales, Rosales and Cucurbitales), developed the capacity of establish a symbiotic relationship with bacteria that are capable of conversion of atmospheric N to ammonium. These bacteria are attracted by signaling molecules secreted by the plant. After the infection is established, root nodules develop to host the bacteria, while the plant benefits by absorbing the produced ammonia. Together, these plant orders are known as the nitrogen-fixing root nodule (NFN) clade.

The restricted phylogenetic occurrence of plant symbiosis with nitrogen-fixing bacteria suggests that a unique genetic event established the foundation for this symbiotic relationship. Previous studies have tried to identify the genetic component of this evolutionary event by searching for conserved genes among species on the NFN clade. However, a complete understanding of the genetic origin of the symbiosis, that encompasses entire genome sequences (including putative regulatory sequences) is lacking. In recent years, the role of conserved, non-coding sequences (CNS) in gene regulation has been extensively demonstrated. In plants, a few studies showed the

presence of many CNS, suggesting that they are under selection and possibly involved in critical biological processes. The main objective of this study was to identify CNS among species that are able to promote nitrogen-fixing root nodule formation. Additionally, by comparing CNS with those reported in species that don't belong to the NFN clade, we sought to detect those that may be involved in the development of NFN symbioses in plants.

The search for conserved sequences across genomes of species capable of nitrogen fixation was carried out by applying a CNS pipeline (Liang et al., 2018) to the analysis of 34 plant genomes. To select the species to be used in this study, all genomes of the orders belonging to the NFN clade and available in the NCBI databases, RefSeq and GenBank, were classified into those capable of N fixation or not. If improved version of these genomes were identified in other databases (e.g. Phytozome), they were used instead. Additionally, nine species with a genome deposited in RefSeq were chosen as an outgroup, representing different orders outside of the NFN clade. All genomes were used in pairwise whole genome alignments using LAST, with the *Medicago truncatula* genome v.5.0 (Pecrix et al., 2018) used as reference. The reference-dependent multiple alignment was produced using ROAST. Next, the identification of conserved element was made according to a scoring system (Liang et al., 2018). The selection of CNS sequences potentially related to the NFN event was done by excluding those detected in all genomes, from those observed in the NFN clade. Finally, the genome context of CNS was inferred from *Medicago truncatula* annotation using bedtools and R scripts. After the exclusion of CNS present in all species, and considering a conservation score of 0.9, 76,989 CNS that are potentially exclusive of the NFN clade were identified. From these, the vast majority (>95%) are within 5Kb of annotated elements (genes, TEs or miRNAs), demonstrating a potential rule in promoting *cis*-regulation of gene expression.

Data derived from this experiment is being continuously improved by the inclusion of new genomes. It will be used now in combination with chromatin accessibility and transcriptome data to further filter for target genes to be evaluated in poplar root cultures.

References
Liang Y et al. (2018) Single-base resolution map of evolutionary constraints and annotation of conserved elements across major grass genomes. Genome Biol. Evol. 10:473–488.

Pecrix Y, et al. (2018) Whole-genome landscape of Medicago truncatula symbiotic genes. Nature Plants, 4:1017–1025.