**Project title:**

Phylogenomic discovery and engineering of nitrogen fixation into the bioenergy woody crop poplar

**Applicant/Institution:**

University of Florida

**Street Address/City/State/Zip:**

G022 McCarty Hall-D, PO Box 110110

Gainesville, FL 32611-0110

**Postal Address:**

118 Newins-Ziegler Hall, PO Box 110410

Gainesville, FL 32611-0410

**Lead PI name, telephone number, email:**

Matias Kirst, 352 846 0900, mkirst@ufl.edu

**Administrative Point of Contact name, telephone number, email:**

Nate Rushing, 352 846 0100, sfrc-grant@ifas.ufl.edu

**Funding Opportunity FOA Number:** DE-FOA-0001650

**DOE/Office of Science Program Office:** Biological and Environmental Research

**DOE/Office of Science Program Office Technical Contact:** Pablo Rabinowicz

**PAMS Preproposal tracking number:** PRE-0000011714

**Research area or areas as identified in Section I of this FOA:**

Plant systems design for bioenergy

**COLLABORATIONS**

*Collaborating institutions, institution’s principal investigator and point of contact:*

University of Florida – Matias Kirst

University of Wisconsin-Madison – Jean-Michel Ané

*Leadership structure and responsibilities:*

The PI of the project, Dr. Matias Kirst, will be the overall project director. The project will fund a Research Manager who will coordinate activities among the groups at UF and UW-Madison, ensuring proper and timely transfer of information and results among them. The Research Manager is expected to spend ½ time for the duration of the project (5 years) on this activity. Specific responsibilities are described below:

* *Matias Kirst* (University of Florida), PI. The PI Kirst will oversee the overall project progress and directly coordinate the activities of *Aim III*.
* *Jean-Michel Ané* (University of Wisconsin - Madison), co-PI. The co-PI Ané will be responsible for research activities of *Aim II*, the evaluation of genes uncovered by the comparative phylogenomic approach.
* *Douglas Soltis* and *Pamela Soltis* (University of Florida), co-PIs. The co-PIs Soltis will be responsible for research activities of *Aim I*, specifically the acquisition of plant resources (samples and data), the phylogenomic analysis, comparative transcriptomics, and prioritization of genes for *Aims II* and *III*.
* *Robert Guralnick* (University of Florida), co-PI; *Ryan Folk* (University of Florida), Sr. Pers. Guralnick and Folk will be responsible for data acquisition and management for the phylogenetic and transcriptome analyses.
* *Sushmita Roy* (University of Wisconsin - Madison), co-PI. Roy is an expert in transcriptome analysis and will provide support in *Experiment I–3*.
* *Brian O’Meara* (University of Tennessee - Knoxville), collaborator. O’Meara will provide expertise in the development of improved HMM approaches and other comparative phylogenetic tasks for the phylogenomic analysis of *Aim I*.

External Advisory Board (EAB): The function of the advisory board will be primarily to provide support in the selection of suitable candidate genes (*Aim I*) to be moved to functional characterization in *Aim II* and *Aim III*, in addition to advising on research strategies. The selected members have an extensive and exceptional record in research on the discovery of genes involved in establishing nodulation and N fixation, and phylogenomic analysis of that process: Dr. Giles Oldroyd (John Innes Center, UK), Gijsbert D. A. Werner (University of Oxford, UK), and Michael Udvardi (Samuel Roberts Noble Foundation, US). The EAB will meet with the project investigators bi-annually (once remotely and once in person).

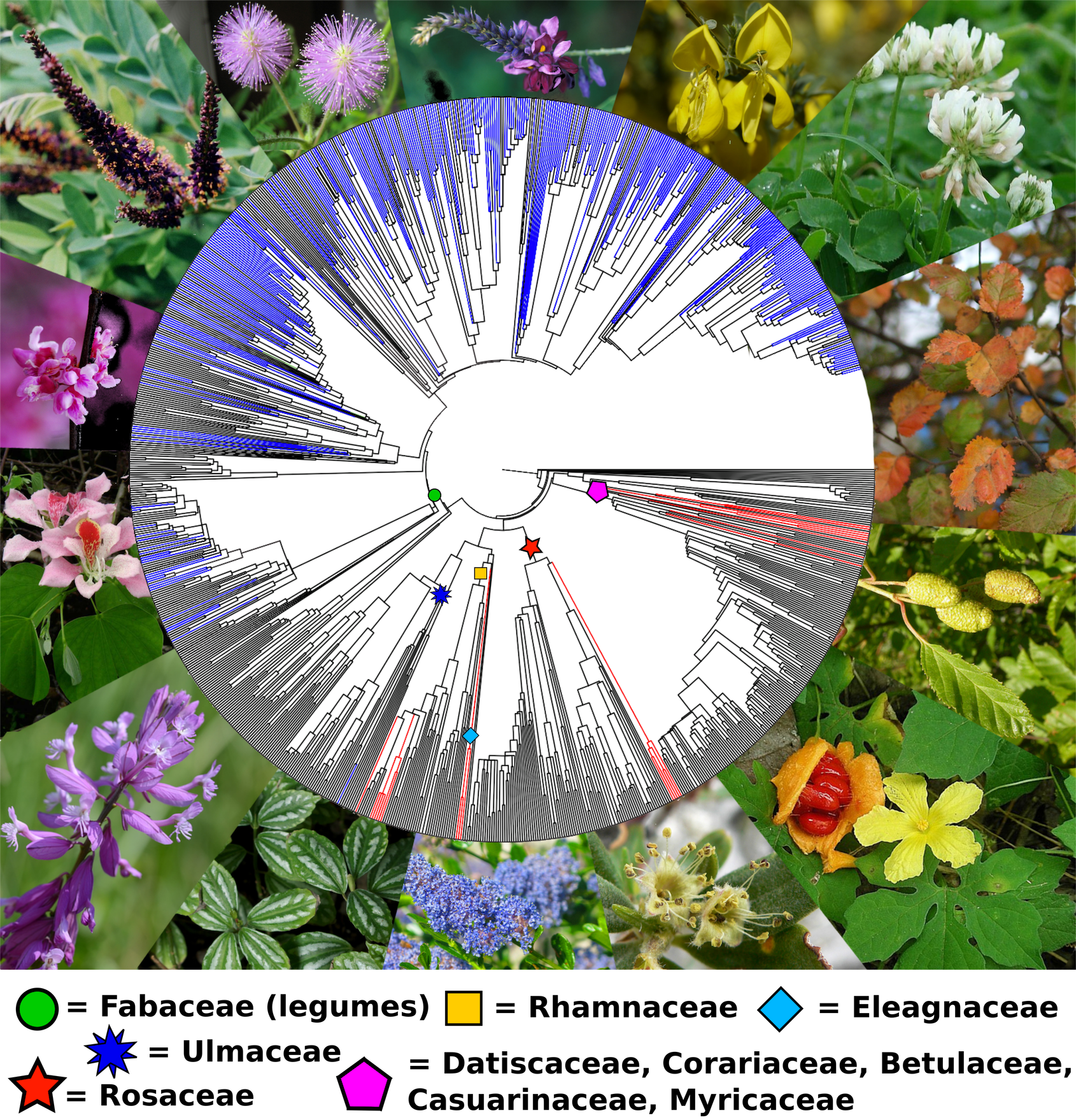
*Collaborative application budget information (in thousands):*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Names** | **Institution** | **Yr1** | **Yr2** | **Yr3** | **Yr4** | **Yr5** | **Total** |
| PI  Co-PI | M. Kirst  R. Guralnick,  D. Soltis,  P. Soltis | U Florida | $1,244 | $1,667 | $1,219 | $657 | $658 | $5,444 |
| Co-PI | J.-M. Ané,  S. Roy | U Wisconsin-Madison | $782 | $599 | $608 | $611 | $615 | $3,214 |

**BACKGROUND/INTRODUCTION**

***Nitrogen (N) availability is critical for high biomass productivity, particularly in marginal lands, yet its application is costly, environmentally damaging, and potentially hazardous to human health.*** Nitrogen is the most common chemical component of Earth’s atmosphere and the mineral nutrient required in the greatest amount by plants because of its role as the primary building block of DNA, RNA, and amino acids. Despite its abundance and critical importance for growth and development, plants cannot access N2 from the atmosphere directly. Instead, plants must absorb available N in the soil as nitrate, ammonium, or amino acids. Intensive fertilization with reactive forms of N is used to compensate for its low availability in agricultural lands. Over 118 million metric tons of N are used annually, produced from natural gas by the Haber-Bosh process that releases ~3% of all global carbon emissions and represents up to 50% of agriculture’s operational costs (Ladha *et al.*, 2016; Jez *et al.*, 2016). The dependence of fertilizer production on a fossil fuel is worrying for the long-term sustainability of modern agriculture. Additionally, of the N applied to agricultural lands, 50-75% is not captured by plants and is instead leached into waterways or released to the atmosphere as N gases. Leached N increases environmental degradation and leads to indirect adverse effects after being naturally converted to different chemical forms, in addition to negatively impacting human health (Gutierrez, 2012). *Clearly, more efficient and cost-effective approaches are needed to enable bioenergy and agricultural crops in general to acquire the N required to maximize growth while minimizing inputs and environmental impact.*

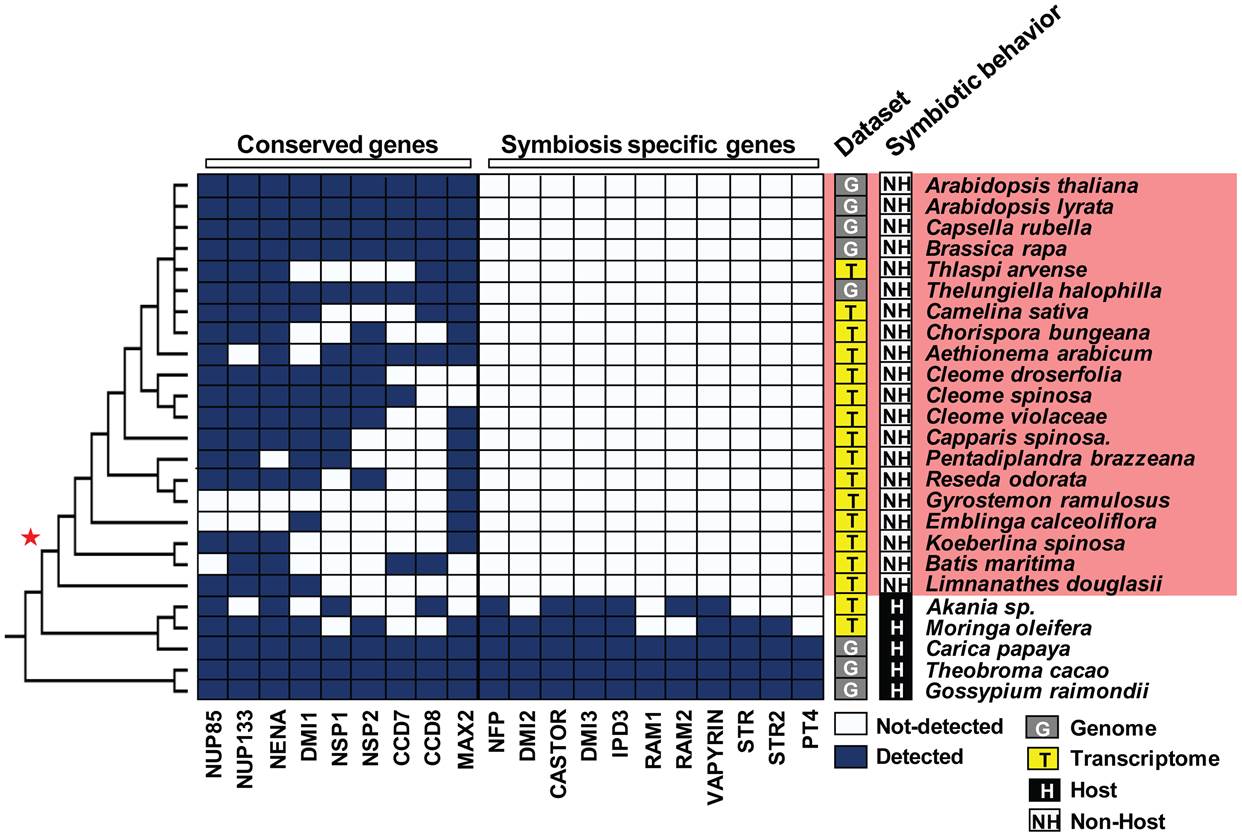
***Some plants acquired the capability to obtain N through a mutualistic relationship with bacteria and archaea.*** Nitrogen from the atmosphere can be converted from N2 to NH3 by certain bacteria and archaea. These prokaryotes capable of N-fixation use a nitrogenase enzyme complex, which catalyzes the conversion of N2 gas to the usable form NH3. This uniquely microbial process has been co-opted by a small number of plant species through a mutualistic symbiotic relationship in root nodules. In this symbiosis, the host plant releases signals that attract the symbiotic bacteria. After the bacteria enter the host, the plant develops new organs, the root nodules that provide a suitable environment for the function of the nitrogenase complex, while providing the necessary resources for the bacteria to thrive. *Unfortunately, the capability to host N-fixing bacteria is limited in the plant kingdom and absent from most bioenergy crops.*



**Figure 1.** Phylogeny of the nodulating clade showing rhizobial lineages (blue) and actinorhizal nodulating lineages (red). Gray and black represent unknown or no N-fixing symbiosis. Major representative lineages are represented by symbols and marginal photographs.

***Two decades ago, we discovered that all flowering plant lineages known to undergo root nodule N-fixation with bacterial symbionts occur within a single clade of angiosperms, with a single underlying predisposition that has yet to be genetically characterized*** (Soltis *et al.*, 1995; Werner *et al*. 2014; Li *et al*. 2015). This ‘N-fixing clade’ includes ~8% of all flowering plants or ~31,200 species. While plant-microbe symbioses exist in other parts of the plant tree of life, no other clade has evolved nodules, which facilitate and greatly increase the efficiency of N-fixation (Bryan *et al.*, 1996; Stokstad, 2016). In other words, one of the most economically and ecologically pivotal symbioses on the planet is tightly packed in a single branch of the tree of life, which includes legumes and nine other families of angiosperms (**Fig. 1**). The best-known and most species-rich family with root nodule symbioses is the legumes (Fabaceae), which associate with rhizobia (alpha or beta-proteobacteria), although many species of legumes do not have this symbiotic relationship. The non-legume families in this clade have established relationships using a diversity of root nodule structures; with one exception, these lineages instead use actinorhizal bacteria (genus *Frankia*). At least nine origins of actinorhizal symbioses have been suggested, as well as numerous origins of rhizobial symbioses. *These parallel origins make the clade as a whole ideal for the discovery of the genomic innovations required for N-fixation.*

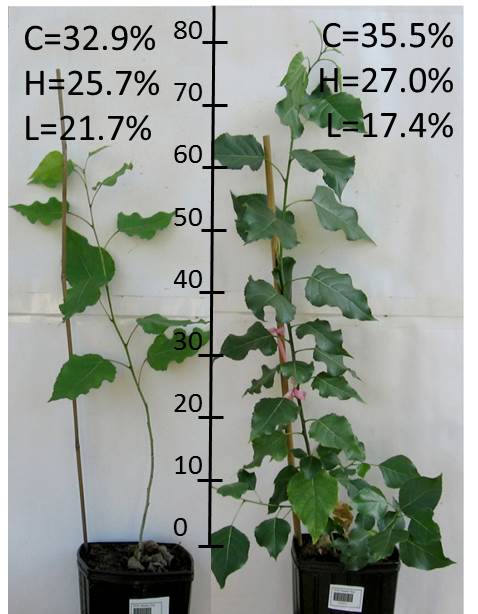
***Phylogenetic approaches and genomic resources allow the discovery of clade innovations and uncover gene gains or losses linked to the evolution of N-fixing symbioses by comparing close relatives that have maintained or lost this ability* (Delaux *et al.*, 2015)*. We demonstrated the utility of this approach by identifying genes potentially required for arbuscular mycorrhizal (AM) symbiosis* (Delaux *et al.*, 2014)*.*** By analyzing the correlation between loss of AM and absence of symbiotic genes in Brassicales, we were able to identify 174 genes that were consistently lost in non-host species, many of which have been related to symbiosis (**Fig. 2**). These results were corroborated by an independent study that evaluated a different set of species to address a similar question and uncovered a comparable set of genes (Favre *et al.*, 2014).While these approaches require large-scale data, next-generation sequencing methods offer unique opportunities to deeply characterize genomes and transcriptomes and create the necessary resources for comparing closely related species to identify critical nodulation genes. *We aim to apply a similar approach in the proposed research to identify specific gene gains or losses linked to the evolution of N-fixing symbioses. Once identified, these genes can be engineered into bioenergy crops to produce the capability to fix N and thus support their cultivation in marginal lands with low inputs.*



**Figure 2.** Loss of the ‘symbiosis-specific’ genes in the Brassicales. Conserved genes are present in both host and non-host Brassicales species. In contrast, ‘symbiosis-specific’ ones are not detected in the genomes and transcriptomes of species having diverged after the loss of the AM symbiosis (red star). Modified from (Delaux *et al.*, 2014).

***The genus* Populus *(poplars) includes the main short-rotation woody bioenergy species in the U.S., many of which are models to develop specialty designed crops for biofuel production. However, the commercial cultivation of the most productive poplar species requires intensive and costly inputs to generate the required productivity, making it an ideal candidate species in which to introduce nodulation capabilities.*** *Populus* species are fast-growing, woody plants that are of worldwide importance, cultivated extensively in Europe and North America. They are considered the principal short rotation woody crop (SRWC) species in the U.S., suitable for planting on more than 40 million acres of surplus or idle agricultural land. *Populus* contains >30 inter-fertile species of woody perennials that provide excellent genetic resources, with single species adapted to large ranges of soils, latitudes, and environments (Bradshaw *et al.*, 2000; Brunner *et al*., 2004). However, most poplar species of significant potential for bioenergy production are adapted to water- and nutrient-rich riparian environments. These species require intensive fertilization to be productive on marginal lands where N availability is limited; N is the most limiting nutrient for tree growth (Oren *et al.*, 2001; Finzi *et al.*, 2007). We have also shown that high N availability has very positive impacts on biomass productivity, and also for conversion of biomass to biofuels, including an increase in cellulose and hemicellulose content, and a decrease in lignin (**Fig. 3**; Novaes *et al.*, 2009). *Genetic engineering of N-fixation capabilities in poplar can address the existing limitation to poplar cultivation in marginal lands, improving productivity and reducing both the cost of biomass production and its environmental impact.*

***Poplar is an ideal plant system in which to develop the approach to introduce N-fixation into bioenergy crops because of the extensive genomic resources available and ease of transformation, and because we have developed innovative systems to rapidly test large numbers of genes and their phenotypic effect on root development/nodulation.*** The molecular and genomic resources available for *Populus* are unmatched by any other woody species (Taylor, 2002; Brunner *et al.*, 2004; Tuskan *et al.*, 2006).Furthermore, efficient methods for genetic transformation of select species and hybrids are well established, and standard protocols are widely used worldwide (Leple *et al.*, 1992). Finally, facile clonal propagation methods that permit the evaluation of genetically identical individuals over different environments and treatment conditions are readily available. To complement these capabilities, we have adopted and improved a rapid root expression system to test the function of target genes (adapted from Yoshida *et al.*, 2015). The induction of transgenic roots by *Agrobacterium rhizogenes* offers a method of fast and efficient transformation and can be immediately applied to test genes discovered using a phylogenetic approach. *In summary, the availability of methods for rapid, high-throughput evaluation of the contribution of newly introduced genes to root nodulation is a perfect complementation to the phylogenetic approaches proposed.*



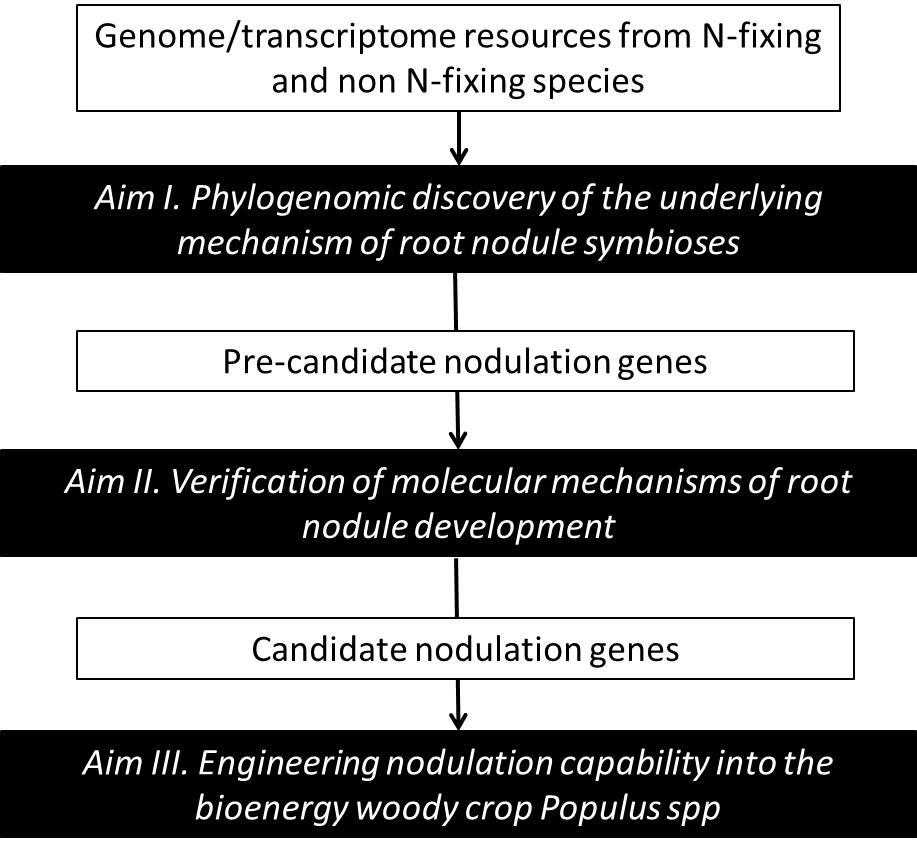
**Figure 3.** Effect of N fertilization on biomass in a *Populus* hybrid grown under N deficiency (left, 0 mM of NH4NO3) and 25 mM NH4NO3 (right). Height (cm) is depicted in the center; biomass composition is C=cellulose, H=hemicellulose, and L= lignin = L; *p* <0.001) (Novaes *et al.* 2009).

*Identifying the genomic innovations that led to nodule development and N-fixation is an obvious prerequisite to genetically engineering bioenergy crops with the capability to fix N and support their cultivation on marginal lands.**The* ***overall goal*** *of this proposal is to discover the underlying genomic novelties that enabled the mutualistic symbiotic relationship between plants and N-fixing bacteria, using a comparative phylogenomic approach across the nodulating clade, to support the genetic engineering of this capability into bioenergy crops. Our* ***overall hypothesis*** *is that a common minimal genomic toolkit, derived from the ancestral genetic predisposition, is shared among plants that carry out root nodule symbioses and that this mechanism can be co-opted into bioenergy crops such as poplar. To evaluate this hypothesis, we will carry out the three aims described below (****Fig. 4****).*

*Aim I. Phylogenomic discovery of the underlying genetic toolkit of root nodule symbioses*

One single clade contains all flowering plant lineages known to undergo nodule development with bacterial symbionts. The *objective* of the first aim is to uncover the genomic novelties that were required for the evolution of these root nodule symbioses. To achieve this goal, we will massively improve knowledge of relationships in the N-fixing clade (**Fig. 1**) and provide a robust, revised understanding of the exact ancestral origin of nodulation and the evolution of the predisposition to nodulate. In contrast to previous studies that have focused largely on legumes, we will extend this phylogenomic approach to comparisons across many of the origins of the trait, including both actinorhizal and rhizobial origins (red and blue, respectively, **Fig. 1**). This phylogenomic framework will inform a series of experiments comparing close relatives that nodulate or do not nodulate to identify the genes underlying nodulation. Our *hypothesis* is that the “predisposition” to nodulation is related to a conserved minimal genetic toolkit that enabled bacterial symbioses. Testing this requires a robust phylogenetic framework, a large-scale nodulation database, querying of genomes for genes related to nodulation in related but nodulation-contrasting species, and further surveying those genes for involvement in symbiosis through transcriptome and evolutionary analysis. The key outcomes of *Aim I* will be a robust inference of ancestral nodulation states and prioritized gene candidates for verification and transformation in *Aim II*.

*Aim II. Verification of molecular mechanisms of root nodule development*

**Figure 4.** General project workflow. Black boxes identify the primary aims of the proposed project. White boxes indicate input/output data and information for and from each aim.

Genetic perturbation of the molecular mechanisms that regulate root nodule development is likely to result in the elimination of this capability in species that acquired it. In contrast, nodule development may be genetically engineered into species that lack this capability by introducing the gene(s) that control the trait through genetic transformation. The *objective* of the second aim is to verify the function of targets discovered in *Aim I* for their effect on root nodule development in *Medicago truncatula* (nodulating) and poplar root organ cultures (non-nodulating). Our *hypothesis* is that genetic perturbation of these targets will disable nodule development and, consequently, nodulation in plants with this capability. In contrast, nodule structures are expected to develop in poplar roots. Combinatorial modifications of candidate genes will be constructed and rapidly evaluated in these model systems before their detailed evaluation in the bioenergy woody crop poplar in *Aim III*.

*Aim III. Engineering nodulation capability into the bioenergy woody crop* Populus *spp.*

Most poplar species are well-adapted to fertile sites in riparian zones and require substantive N-fertilization to reach suitable productivity levels. In *Aims I* and *II*, we anticipate discovering key elements of the ancestral gene set that underlies the predisposition to nodulation and subsequent gains of the symbiosis, leading to the development of nodule structures to support N-fixation. The last objective of this proposal is to engineer nodule development in entire poplar plants and test the impact of these structures on N-fixation. To achieve this aim, we will test targets integrated from previous genetic work in legumes and from *Aims I* and *II* for their effect in nodule development. The ability of these new organs on poplar roots to support N-fixation will be evaluated with a wide range of N-fixing bacteria. Our *hypothesis* is that introduction of these genetic changes in poplar will result in the development of nodule-like structures that will improve the association between poplar and N-fixing bacteria. Moreover, this aim will serve as a platform for future engineering of intracellular accommodation (infection) processes.

This proposal is well aligned with the research areas of the *Plant Systems Design for Bioenergy* *Program*, by aiming to engineer the unique capabilities of N-fixing symbioses into the bioenergy crop poplar to support high biomass productivity on marginal lands, with a reduction in both cost and environmental impact. Three critical knowledge areas are required for completion of the proposed research goals: phylogenomic analysis, the biology of N-fixing symbioses, and bioenergy poplar. The *phylogenomics* effort will be directed by D.E. Soltis and P.S. Soltis (co-PIs, University of Florida) who have led some of the most extensive phylogenetic studies in plants, including first hypothesizing the nodulating predisposition in the N-fixing clade. Research related to the *biology of N-fixing symbioses* will be led by J.M. Ané (co-PI, University of Wisconsin-Madison), an expert in the molecular and genetic analysis of nodule development and N-fixation. The incorporation of the molecular mechanisms into poplar will be coordinated by M. Kirst, a tree genomicist working on the development of *bioenergy poplar* cultivars that are more productive and suitable for conversion to biofuels. Data integration across the project will be accomplished through an informatics platform, along with management and analysis of trait information, developed by R.P. Guralnick (co-PI) and R. Folk (Sr. Pers.) at the University of Florida, and S. Roy (co-PI) at the University of Wisconsin-Madison. Collaborators will provide additional support in novel data analysis methods (B. O’Meara, University of Tennessee). Finally, an External Advisory Board will be established to provide critical expertise in phylogenomic analysis and the discovery of genes involved in establishing nodulation and N-fixation (G. Oldroyd, John Innes Center, UK; G.D.A. Werner, University of Oxford, UK; M. Udvardi, Samuel Roberts Noble Foundation, USA).

**PROPOSED RESEARCH AND METHODS**

**AIM I. Phylogenomic discovery of the underlying genetic toolkit of root nodule symbioses**

***Introduction***—A phylogenomic approach will be used to reconstruct gains and losses of N-fixing symbioses and identify candidate genes involved in the predisposition of this crucial trait (Delaux *et al.*, 2014, 2015). The research proposed in *Aim I* will provide a vastly improved comparative framework, comprising a phylogenetic tree of unparalleled robustness and species sampling that will inform a subsequent series of ambitious transcriptome experiments, genomic screens for nodulation genes, and downstream analyses based on evolutionarily optimized nodulating and non-nodulating species pairs from our ancestral nodulation reconstructions. **We will integrate the comparative phylogenetic, comparative transcriptomic, and genomic products of *Aim I* to discover gene candidates that potentially underlie nodule development and represent components of the ancestral gene toolkit.** These candidate genes will then be tested for function in nodulating and non-nodulating model systems in *Aim II* and *Aim III*. For discovery of genes that determine nodule development, a clear understanding of the evolutionary lability and thus likelihood of successful transferability of N-fixing symbioses among lineages of angiosperms is necessary and requires: (1) robust phylogenetic inferences on the origin of the predisposition to nodulation, as well as subsequent gains and losses of nodulation itself, and (2) comparative transcriptomic analyses, where cumulative homology among transcriptomes is maximized via sister lineage comparisons and integrated with genomic screens of gene candidates.

Multiple phylogenetic analyses have been conducted on the N-fixing clade with the aim of elucidating the origins of N-fixing symbioses, among which the most comprehensive and widely used is that of Werner *et al.* (2014). This study used an angiosperm-wide phylogeny, in combination with a comprehensive nodulation dataset and a Hidden Markov Model (HMM) approach, to estimate both the ancestral nodulation states and the evolutionary rates of nodulation across the tree. This powerful approach allowed the identification of ancestral species without the nodulation state but with an elevated rate of nodule evolution (= predisposed lineages, the predisposition of which we hypothesize to be related to a minimal ancestral symbiosis-enabling genetic toolkit). This work demonstrated not only multiple origins of nodulation, but, critically, statistical support for a single origin of the predisposition for N-fixation at the origin of the clade (a corroboration of hypotheses in Soltis *et al.*, 1995). However, the Werner *et al.* (2014) phylogeny was derived from an angiosperm-wide tree developed over five years ago from sequence data for only a few genes available in GenBank, rather than a tree explicitly developed from genome-scale data to address the question of the evolutionary lability of nodulation. Thus, the Werner *et al.* (2014) tree includes only 3,467 species, of which approximately only half are members of the N-fixing clade. Hence, the N-fixing clade, which comprises 31,000 species in total, was grossly under-sampled. Moreover, because of limited sampling and data available at the time of tree construction, the tree contains known topological errors (Doyle 2017) that impede efforts to understand the lability of nodulation symbioses and develop models for transferability of this trait across evolutionary lineages. Likewise, improved HMMs developed recently can enable more accurate evolutionary interpretations of gains, losses, and rates of nodule evolution, particularly in combination with updated databases on nodule-related traits. Hence, a revised phylogeny—based on deliberate and extensive sampling, phylogenomic data, and rigorous statistical analysis—is needed to infer more accurately the precursors of N-fixing symbioses, gain and loss events, and potential transferability. This tree-building step is a critical antecedent to identifying phylogenetically optimal sister pairs for further comparison.

*Aim I* has three components, which collectively will yield a set of candidate nodulation genes for *Aims II* and *III* (**Fig. 5**). The primary objective of *Experiment I-1* is to develop a robust comparative framework to confidently infer the number and distribution of independent origins and losses of N-fixing symbioses, the phylogenetic location of the underlying predisposition, and optimal sister clade comparisons for transcriptomic experimentation and genomic candidate screening. We will build a vastly improved phylogenetic estimate compared to Werner *et al*. with unmatched statistical robustness and taxonomic sampling. Because independent origins and losses of the N-fixing symbiosis represent evolutionary replication, the integration of comparative phylogenetics with transcriptomic and genomic analyses in this *Aim* will provide an explicit test of the hypothesis that the single predisposition to nodulation in the N-fixing clade represents a single ancestral minimal genetic toolkit enabling the symbiosis. The robust tree we obtain not only will be foundational for our investigation, but also represents a central community resource for researchers in this intensely investigated group. In collaboration with B. O’Meara, a leader in statistical comparative phylogenetics (e.g., Beaulieu *et al.* 2012, 2014), we will repeat the HMM model of nodule evolution proposed by Werner *et al.* (2014), but now including an improved nodulation dataset, a superior phylogenetic estimate, and methodological extensions that provide the best estimate to date of the pattern of origination and losses of nodulating N-fixing symbioses. Using this refined phylogeny, we will identify (*Experiment I-2*) evolutionarily optimized pairs of closely related nodulating and non-nodulating species for comparative analyses in *Experiment I-3*. In *Experiment I-3*, we will execute a series of transcriptomic comparisons and genomic screens of these sister pairs to identify candidate genes underlying nodulation and its predisposition. The overall goal of *Aim I*, and the key aspect of *Experiment I-3*, is to prioritize the selection of plant nodulation gene candidates through a series of criteria based on analyses of gene expression, evolution, and presence/absence patterns that are predictive of nodulation phenotypes. We hypothesize that our approach will yield *a robust, prioritized set of candidate genes* representative of the ancestral genetic toolkit and subsequent origins of nodulation, including the putative critical regulator(s) of nodule development.

Grant/DOE%20nitrogen%20fixation/New_aim_one_pipeline.pdf

**Figure 5.** *Aim I* experimental workflow. The three experiments will be used to select candidate nodulation genes to be tested in *Aim II*.

***Responsibilities, timeline, and deliverables***—Co-PIs D. Soltis, P. Soltis, and R. Guralnick will be responsible for sequencing and phylogenomic activities of *Aim I*. The co-PI Roy will be the primary coordinator of the transcriptome analysis. These activities will begin in the first year of the project, with the immediate construction of a phylogenetic framework of the N-fixing clade, critically improving largely unknown relationships that particularly impact comparisons in the legumes and among certain actinorhizal groups. Screening for existing candidate genes for closely related nodulating and non-nodulating species will also start in the first year, focusing on an initial small set of contrasts for which the phylogeny is currently well understood (see below). As the phylogenetic framework is improved in years one and two, additional sister groups will be selected for transcriptomic analyses and genomic screens. Finally, candidate genes derived from these comparisons will be filtered through a series of methods for use in *Aims II* and *III*. From *Aim I*, we anticipate the yearly delivery of a rank of candidate nodulation genes, where the 50 most highly ranked will be evaluated annually in *Aim II*. For the duration of the project, up to 200 candidate genes will be delivered, starting at the end of the first year and continuing until the end of the fourth year.

***Experimental Design***—The overall workflow of *Aim I* is described in **Fig. 5**. The first two experiments of *Aim I*, involving the construction of a phylogenetic framework of the N-fixing clade (*Experiment I–1*) and the selection of phylogenetically informed species pairs (*Experiment I–2*), will initiate simultaneously. The phylogeny of the N-fixing clade will continue to be refined for the first three years of the project, to improve the selection of species for comparative transcriptomic analyses of nodule development and genomic queries of nodulation genes to identify a suite of candidates involved in nodule development (*Experiment I–3*).

Experiment I–1. Construction of a phylogenetic framework of the N-fixing clade

*Species sampling*—Our approach aims to correct for uneven sampling in available data, guided by current phylogenetic knowledge (**Fig. 1**), taxonomy (for non-sequenced species), and the inclusion of a comprehensive representation of known diversity of N-fixating species. This balanced sampling approach is critical for probabilistic comparative methods that will be deployed, which rely on topology and branch length to infer ancestral nodulation patterns. We will link newly collected data with existing transcriptomic resources in the Sequence Read Archive ([www.ncbi.nlm.nih.gov/sra](http://www.ncbi.nlm.nih.gov/sra)) and the 1KP project ([www.onekp.com)](http://www.onekp.com)), currently comprising 311 species, for extensive outgroup sampling. To enable comprehensive sampling on an ambitious scale, DNA materials will be derived primarily from historical DNAs, leveraging strategic museum collections with geographic strengths that cover diversity spots for the N-fixing clade (Royal Botanic Garden Edinburgh, Royal Botanic Garden Kew, New York and Missouri Botanical Gardens, Smithsonian, National Herbarium of the Netherlands, Jepson Herbarium, National Herbarium of Mexico, Australian National University). Investigators at the Kunming Institute of Botany (China) have also agreed to commit extensive silica-dried East Asian collections. The co-PIs of *Aim I* have extensive experience with phylogenomic studies using historical DNAs from large collections of species. We will sample 15,000 species in the phylogenomic dataset and include ~100 outgroups, yielding a robust phylogenomic backbone covering approximately 50% of the clade. As a final comprehensive approach incorporating all current phylogenetic knowledge, we will integrate all available species with GenBank data, inferring a phylogeny constrained by the topology of the optimal phylogenomic tree. This tree will include over 65% of the 31,000 species that belong to the N-fixing clade.

*Phylogenomic sequencing and analysis*—R. Folk (Sr. Pers.) has already developed a targeted genomic sequencing tool based on sequence capture (100 genes, mean length 700 bp) that applies to all rosids, which includes the N-fixing clade, to support comprehensive sampling of outgroups. This resource was developed using a program we developed (MarkerMiner, Chamala *et al.*, 2015) with transcriptomic resources from the 1KP project (onekp.com/; Matasci *et al.*, 2014; representatives included for 78 families). We will also obtain near-complete chloroplast genomes and rDNA, resulting in ~170,000 coding bp sequenced per individual. To overcome millions of years of nucleotide divergence, ortholog variants have been incorporated into the sequence capture probe design (cf. Mandel *et al.*, 2014). For data combination strategies, we will link with ongoing efforts in our labs to harvest genomic data from public infrastructure, enabling meta-analyses of published nucleotide and genomic data at scales that are currently not feasible. Sequence assembly and analysis will follow aTRAM-based pipelines that were previously developed by the Senior Personnel Folk (Folk and Allen, unpublished; Allen *et al*., 2015; Allen *et al.*,2017).

*Testing the origin of the predisposition*—The phylogenetic comparative analysis of Werner *et al.* (2014) has greatly impacted subsequent nodulation research by explicitly identifying lineages with the predisposition for the symbiosis using a HMM approach. This is a critical development because these lineages could represent some of the most likely candidates for assessing the origin of N-fixation and the feasibility of transforming non-nodulators with symbiotic competence. However, as Doyle (2017) noted, the underlying phylogenetic data used by Werner *et al*. (2014) (developed before the widespread availability of phylogenomic data for thousands of flowering plants) are limited, and the tree in several respects does not agree with our current understanding of higher-level relationships in legumes (LPWG, 2017). This means that, while the conclusion that a single predisposition arose in the most recent common ancestor of the N-fixing clade is likely to remain robust to future analyses, finer-scale determinations of the presence of symbiotic tendencies are sensitive to alternative resolutions of topology and branch length in the underlying phylogeny. That is, finer-scale inferences in Werner *et al.* (2014) that are crucial to our proposed work are likely erroneous. Identification of nodulation in the basal legumes are particularly suspect, indicating that the improved resolution that will result from our proposed phylogenomic analysis will be a major contribution. Phylogenomic estimates will improve the comparative framework for inferring the genes underlying nodulation, through more robust clade resolution, especially in the recalcitrant legumes, and more robust estimates of divergence times. We will integrate our more robust phylogenomic hypothesis with an improved version of the nodulation database described below and maintained by nodulation expert J. Sprent, and extend the HMM approach applied previously (Werner *et al.*, 2014). We have been in contact with J. Sprent, who has offered her expertise in nodule morphology. However, due to current health issues, J. Sprent is currently unable to serve as a formal collaborator, but will assist in all possible within her capabilities.

As part of our extension of the nodulation database, we will incorporate a previously described classification system ([www.ildon.org/ildon.html](http://www.ildon.org/ildon.html)) for nodular morphology, asking whether differing and possibly non-homologous tissue structures have an impact on the conclusions of the model. Our work will also involve extending the R-based methods to integrate over phylogenetic and trait model uncertainty utilizing a bootstrapping approach, avoiding the potential for positively misleading predisposition identifications. This work will be carried out with the support of collaborator B. O’Meara (University of Tennessee), a leader in statistical phylogenetic comparative methods.

Experiment I–2. Selection of phylogenetically informed comparisons of nodulating and non-nodulating species

We propose a suite of comparisons *between nodulating and non-nodulating close relatives* (i.e., sister lineages; rather than a random assortment of nodulating and non-nodulating species) to maximize cumulative sequence and regulatory homology among genomes and transcriptomes and therefore our ability to elucidate the underlying genetic syndrome involved in N-fixing symbioses. In *Experiment I-2*, we will use the phylogeny we provide above, as well as published literature, to select optimal sister pairs for transcriptomic analyses and genomic screens to be conducted in *Experiment I-3*. Most previous comparisons (as well as ongoing comparisons of nodulating and non-nodulating species) have involved legumes that associate with rhizobia. In contrast, actinorhizal comparisons have been largely overlooked (although several recent/unpublished comparisons are underway). These lineages are an ideal focal point for multiple comparisons because actinorhizal nodulation has evolved multiple times, and some of these relationships are clear. *Experiment I-3* will begin immediately on select and unambiguous comparisons of both actinorhizal and rhizobial species pairs. The phylogenomic analyses to be carried out in *Experiment I–1* will critically inform appropriate transcriptomic and genomic comparisons continuously through this project by reconstructing the gains and losses of nodulation, predisposition of nodulation, and the timing of the gain-loss events. Below, we present 16 possible comparisons, of which a subset of the phylogenetically best-resolved species pairs will be selected for *Experiment I-3*; the remaining examples illustrate considerations underway and will be further evaluated through our phylogenetic analysis. We stress that many of these comparisons are preliminary suggestions, and while representative of the scope of our aims, will be continually informed by phylogenomic efforts.

Actinorhizal Comparisons

We propose to carry out the comparisons of lineages described in Table 1.

**Table 1.** Actinorhizal lineages of separate origin and non-N-fixing sister group for comparison. Numbers in the first column identify the specific comparisons that are described in the following paragraphs.

|  |  |  |
| --- | --- | --- |
| ***Actinorhizal N-fixing lineages*** | ***Sister group (no- N-fixing lineages)*** | ***Data source and species (G=genome, T=transcriptome)*** |
| **1**—*Myrica* + *Comptonia* | *Canacomyrica* | Novel comparison |
| **2**—Casuarinaceae | *Canacomyrica* and Betulaceae (other than *Alnus)* | H. Gherbi, V. Hocher & S. Svistoonoff (INRA/CIRAD/Université Montpellier, France) - *Casuarina glauca* (G) |
| **3**—*Alnus* | Other Betulaceae | P. Normand (INRA, France)—*Alnus glutinosa* (G), *Betula pendula* (T), and *Fagus sylvatica* (G) |
| **4**—*Coriaria* | *Corynocarpus* | Novel comparison |
| **5**—*Datisca* | Tetramelaceae | K. Pawlowski (Stockholm U., Sweden)—*Datisca glomerata* (G) and *Begonia fuchsioides* (G) |
| **6**—Dryadoideae (Rosaceae)  variation within *Dryas* | Various Rosoideae (*Rubus*, *Fragaria*) | A.M. Berry (UC Davis)—*Purshia tridentata* (G)  Martin Parniske (Ludwig-Maximilians U, Germany)—*Dryas drummondii* (G) and *Dryas octopetala* (G) |
| **7**—Elaeagnaceae | *Barbeya* (and *Dirachma*) | Novel comparison |
| **8**—Colletieae (Rhamnaceae)\* | *Granitites* + *Alphitonia\** | L.G. Wall (National U. of Quilmes, Argentina)—*Discaria trinervis* (G)  *Granitites* and *Alphitonia*—novel comparison |
| **9**—*Ceanothus*\* (Rhamnaceae) | Pomaderreae (*Spyridium*)\* | A.M. Berry (UC Davis)—*Ceanothus thyrsiflorus* (T) |

Novel actinorhizal comparisons not previously addressed (number identifies the comparison in Table 1):

**1. Myricaceae**—This family comprises three genera; *Myrica* (55 spp.) and *Comptonia* (1 sp.) have N-fixing symbioses and are collectively sister to *Canacomyrica* (1 sp. endemic to New Caledonia) which does not. *Canacomyrica* germplasm will be provided by a colleague (Gildas Gâteblé—Institut Agronomique Néo-Calédonien); *Myrica* is a US native and easily obtained.

**4. Coriariaceae**—This family has one genus (*Coriaria*, 14 spp. in China, New Zealand, South America); species of *Coriaria* have N-fixing symbioses. However, their sister group, *Corynocarpus* (5 species; New Zealand and New Caledonia, New Guinea), the only genus in Corynocarpaceae, is not symbiotic. Germplasm of both genera is readily obtainable. Initial phylogenetic results indicating an apparent gain in *Coriaria* can also be leveraged with ongoing parallel comparative work focusing on a closely related gain in *Datisca* (compared to non-symbiotic Begoniaceae).

**7. Elaeagnaceae**—The family consists of 60 species in three genera; all three genera and all component species (*Elaeagnus* ~50 spp.; *Hippophae* 5–7 spp.; *Shepherdia* 3 spp.) have N-fixing symbioses, and germplasm is easily obtained. The family represents another clear gain of the symbiosis, yet relationships remain uncertain. The sister group in most analyses is Barbeyaceae (*Barbeya*, one sp.; Arabian peninsula), although Dirachmaceae (*Dirachma*, two spp.; Somalia) may be sister to *Barbeya* (**Fig. 6**). Large-scale phylogenomic analyses will clarify these poorly supported relationships.

Actinorhizal lineages with enhanced comparisons

**8 and 9. Rhamnaceae**—Ongoing work on N-fixing members of Rhamnaceae has considered the two symbiotic Rhamnaceae (tribe Colletieae and *Ceanothus*) as a single origin of nodulation in comparative experiments with two other non-symbiotic members of Rhamnaceae (*Emmenosperma* and *Rhamnus*). However, *Ceanothus* and *Discaria* (the genus of Colletieae under investigation) are in separate clades and clearly represent either two separate origins of the symbiosis or one ancient gain followed by losses. The relationships of *Ceanothus* are still unclear (though analyses to date suggest *Ceanothus* should be compared to other Pomaderreae, e.g., *Spyridium*), as are relationships in the Ziziphoid clade of Rhamnaceae; clarifying these relationships would critically improve experimental interpretations in this clade performed to date.

../../../../../Desktop/Nitrogen-fixer-phylogeny-figures/Eleagnaceae.p

**Figure 6.** Example of an evolutionarily optimized experiment, using exemplars of Elaeagnaceae, for which the optimal comparison appears to be *Barbeya*.

Rhizobial Comparisons

We propose to carry out the comparisons of lineages described in Tables 2 and 3.

**Table 2.** Rhizobial lineages (non-legumes) of separate origin and non-N-fixing sister group for comparison. Numbers in the first column identify the specific comparisons that are described below the table in detail.

|  |  |  |
| --- | --- | --- |
| ***Rhizobial N-fixing lineage, non-legume*** | ***Sister group (non N-fixing lineages)*** | ***Data source and species (G=genome, T=transcriptome)*** |
| 1–*Parasponia* (Cannabaceae) | *Trema* (Cannabaceae) | R. Geurts (Wageningen University, Netherlands)—*Parasponia* and *Trema* (G) |

Rhizobial lineages with enhanced comparisons

Rhizobial N-fixation has evolved multiple times: once in *Parasponia* or the ancestor of this lineage (Table 2; **Fig. 1**), and a large but indeterminate number of times in the legumes, in which gain has been followed by multiple losses. Due to uncertainty regarding the phylogeny of legumes, the number of gains and losses and where these occurred in the history of this group remain poorly understood; this problem is most acute for basal lineages and therefore for our understanding of evolutionary dynamics of leguminous nodulation at its earliest stages. In fact, despite recent progress (LPWG, 2017), the poor state of our knowledge of higher-level legume relationships remains one of the biggest challenges in the study of N-fixation in rhizobia-associated lineages, precluding selection of optimal species pairs for comparative genomics, physiology, and other fields (see below). Hence, one of our initial goals (*Experiment I–1*) will be to resolve relationships among basal legumes, where the greatest uncertainty in relationships exists and where it is most critical to understanding the early evolution of rhizobial N-fixing symbiosis.

**1. *Parasponia***—This nodulating lineage is sister to *Trema* (non-nodulating). The origin of nodulation in *Parasponia* involves *Rhizobium,* and this is the only instance of a rhizobial interaction outside of the legumes and, therefore, an important comparison. Rene Geurts (Wageningen University, The Netherlands) has sequenced *Trema tomentosa* along with *Parasponia andersonii* (manuscript in prep.) and will provide material for use in this study (see *Letter of collaboration - Rene Geurts*).

***Legumes***—A major outcome of our comprehensive sampling effort and phylogenomic analysis to be carried out in *Experiment I–1* will be in the clarification of the evolutionary pattern of symbioses in the legumes (Fabaceae). We will focus on elucidating relationships among the basal lineages, the region of the phylogeny that shows the greatest historical dynamism and that is critical to determining the origins, gains, and losses of N-fixing symbioses. The distribution of nodulation suggests multiple gains and losses of this association, but the direction of change and the number of evolutionary transitions remain unclear due to the lack of a firm phylogenetic underpinning. Numerous comparisons of nodulating and non-nodulating taxa could be made within legumes, providing robust evolutionary replication. On the basis of current phylogenetic knowledge, we propose several possible comparisons (Table 3), representing the general sampling scheme but highly contingent on phylogenomic results.

**Table 3.** Rhizobial lineages (legumes) of separate origin and non-N-fixing sister group for comparison. Numbers in the first column identify the specific comparisons that are described below the table in detail.

|  |  |  |
| --- | --- | --- |
| ***Rhizobial N-fixing lineage, legume*** | ***Sister group (non N-fixing lineages)*** | ***Data source and species (G=genome, T=transcriptome)*** |
| 1–*Chamaecrista* (Caesalpinioid) | *Senna* (Caesalpinioid) | Co-PI J.-M. Ané (U. Wisconsin)—*Chamaecrista fasciculata* (G) and *Senna alexandrina* (T) |
| 2– *Mimosa* and *Anadenanthera* (Mimosoid)\* | *Parkia* and *Adenopodia* (Mimosoid) | Co-PI J.-M. Ané (U. Wisconsin)—*Mimosa pudica* (G) and novel (*Parkia* or *Anadenanthera)* |
| 3- *Dimorphandra* Group + mimosoids. Sampling the nodulating and non-nodulating clades of the [*Dimorphandra* Group B + mimosoids] clade might give us data for several independent gains and/or losses of nodulation. This would be tied to 2, above. | | |
| 4a–*Lotus* (Pap.-Robinioid.) | *Peteria* or *Genistidium* | Novel comparison |
| 4b*–Medicago* (Pap.-Galegoid.) | *Afgekia* |
| 5. Multiple gains and loss events within the 50-kb inversion clade  *Arachis/Zornia* (Pap.-Dalbergioid.) | *Nissolia* / *Chaetocalyx* (Pap.-Dalbergioid) | Novel comparison  P.-M. Delaux (CNRS, U. Toulouse, France)—*Nissolia schottii* (G)  Novel comparison by adding *Zornia* |
| 6. Losses within *Trifolium* |  | Novel comparison |

**1. Comparison of *Chamaecrista* and *Senna* (caesalpinioids)**—As with other basal legumes (particularly those in caesalpinoids, which are not monophyletic), it is difficult to ascertain the evolutionary pathway of gain vs loss of nodulation without a better-supported topology.

**2. *Mimosa***—Studies are ongoing on *Mimosa* (mimosoid), but comparisons to close relatives are not yet being pursued. *Parkia* represents loss of nodulation in the mimosoids and could be compared to *Mimosa*. Also, its sister, *Anadenanthera* (which nodulates)*,* provides another valuable comparison. Another important comparison involves *Adenopodia* (4 species in Africa, 3 in Central America), which is not a nodulator, and the apparent sister to *Mimosa*. Both gains and losses of nodulation may have occurred in mimosoid legumes, but the patterns are complex. Because the phylogenetic relationships in the mimosoid part of the legume tree are poorly supported, the improved phylogenetic framework to be constructed in *Experiment I-1* will be a critical support tool.

**3. *Dimorphandra* Group + mimosoids**—Within a clade of some caesalpinioids (non-monophyletic; see Li *et al*. 2015; LPGW, 2017) and mimosoids, there are apparently several independent gains and losses of nodulation. This area of the phylogenetic tree needs additional resolution to interpret gain-loss events, to be generated in *Experiment I-1*. Sampling nodulating and non-nodulating close relatives of the *Dimorphandra* Group + mimosoids clade would provide information on several independent gains and/or losses of nodulation. For example, *Dimorphandra* could be compared to a member of its putative sister group (*Mora*). *Mora* represents a unique inclusion because it is one of several non-nodulating legume species that form nodule-like outgrowths.

**4a. *Lotus* and *Medicago***—These lineages represent a unique origin of nodulation in papilionoid legumes, most of which nodulate. Current studies are comparing the nodulators *Lotus* and *Medicago* to the non-nodulators *Nissolia* and *Chaetocalyx*. However, the latter two genera are only distantly related to *Lotus* and *Medicago*, despite both being papilionoid legumes. Thus, we propose a more suitable approach, where *Lotus* would be compared to a close relative that is non-nodulating (e.g., *Peteria* or *Genistidium,* contingent on phylogenomic results). Both genera appear to represent losses of nodulation—thus, this would provide a novel comparison of papilionoid legumes.

**4b. *Medicago* and Baphieae**—*Medicago* has been investigated extensively, with a genome available for the model legume *Medicago truncatula* (Tang *et al*., 2014). We propose a novel comparison with *Leucomphalos* or *Baphiopsis* (non-nodulating). These two genera will also be part of a comparison within the ‘50-kb inversion clade’ (see below). Furthermore, comparisons can also be made of *Medicago* with other members of this clade. Comparison of *Medicago* to *Leucomphalos* or *Baphiopsis* (apparent loss) will permit another gain-loss comparison in another legume subclade (the IRLC clade).

**5. 50-kb inversion clade**—This is another part of the legume phylogeny with multiple apparent gain-loss events, ripe for investigation of the evolution of nodulation. Again, as in other parts of the proposed study, increased resolution and support of relationships are central to moving forward with investigations of the symbiosis. We propose several possible comparisons (contingent on the phylogenetic results obtained) that represent different portions of the 50-kb clade: 1) *Cladastris* clade–Swartzieae: *Cladrastis* (nodulating) vs. *Swartzia* (non-nodulating); 2) Baphieae: *Baphia* (nodulating) vs. *Leucomphalos* or *Baphiopsis* (non-nodulating; these genera will also be part of a comparison with *Medicago*, a more distant relative (above); and 3) Dalbergioids: *Zornia* (nodulating) vs. *Nissolia* (non-nodulating); *Nissolia schottii* is already under investigation by Dr. P.-M. Delaux at the Laboratoire de Recherche en Sciences Végétales, France, using genome sequencing and comparisons with *Arachis ipaensis*. This comparison is phylogenetically suboptimal. We will add a novel comparison by adding *Zornia,* a closer relative of *Nissolia*. We will collaborate with these investigators to obtain transcriptomes; these data, combined with a genome for *Nissolia,* will elucidate the putative loss of symbiosis in *Nissolia.* Representative germplasm of most of the lineages noted above are either widely obtainable, or in the case of *Errazurizia*, available from the U.S. National Germplasm System. The exceptions are members of Baphieae, which are infrequently cultivated and more challenging to obtain; but will likely be obtainable through collaborators, if necessary.

**6. Recent putative losses**—There are a few instances of losses within a genus (*Senegalia*, *Pentaclethra*, *Pterocarpus*, and *Trifolium*); remarkably, these types of investigations have never been conducted at such shallow phylogenetic scales. In such recent losses of the symbiosis, the molecular homology of the genes needed for nodulation is expected to be stronger than for more temporally remote events (e.g., clades deeper in the tree), possibly simplifying expression and gene network comparisons. For example, *Trifolium polyphylla* is reported not to nodulate (Makarov *et al.*, 2011; Soudzilovskaia *et al.*, 2012). Its tentative sister, *Trifolium alpinum* (Werner *et al.*, 2014), nodulates (Sprent, 2009). We can compare the non-nodulating *T. polyphylla* to *T. alpinum* and/or to other species of *Trifolium*, which typically nodulate; these are all field-accessible and partially represented in germplasm resources.

Experiment I-3. Comparative analyses of nodulation and detection of the genetic nodulation toolkit

*A. Transcriptomic Experiment*

*Goals*—Using the species pairs identified through *Experiments I-1* and *I-2*, we will conduct comparative transcriptomic analyses of nodule development to identify a suite of candidate genes involved in nodulation. Although previous studies have addressed this goal in some species, our experimental design will yield novel results for the following key reasons: (i) we have developed a phylogenetically informed sampling strategy that will be based on improved inferences of losses and gains of nodulation, (ii) our sampling spans a range of species with ‘independent’ gains of nodulation, thus providing *evolutionary* replication increasing our power to find candidate genes showing consistent associations with the nodulating phenotype, and (iii) we include transcriptomic comparisons of roots with and without nodules in nodulating species and under low- and high-N conditions.

*Experimental design*—We propose 16 pairwise comparisons of nodulating and non-nodulating species (above), the selection of which will be evolutionarily optimized via the output of *Experiments I-1* and *I-2*, although as noted we anticipate several potential options described in *I-2* above that are likely to remain stable and are immediately actionable. Our choice for non-nodulating species in each pair will be based on our phylogenetic and comparative analyses (*Experiment I-1*), including at least one instance where a loss of the predisposition is inferred. Some of our comparisons involve a gain of nodulation from a non-nodulating ancestor, whereas in other cases our proposed comparison involves the loss of nodulation compared to a non-nodulating relative. Some of our proposed comparisons are in deep evolutionary time (sister families) and others are shallow, involving closely related genera or even species. Together, this suite of comparisons will afford unique opportunities to elucidate the genetic underpinning of nodulation—a broad phylogenetically informed approach that has not been employed previously. We will also include poplar (last aim), given that poplar is not part of the N-fixing clade and thus has never had a predisposition to nodulate. We will perform a factorial experiment where the following factors will be tested in each inter-species comparison: (i) bacterial-supplemented vs. unsupplemented (rhizobia/*Frankia* cultures currently available; see *Aim II*); (ii) N-supplemented vs. unsupplemented (to suppress/promote nodulation); and (iii) young vs. mature roots (i.e., mature plants with nodulating roots vs. young plants at pre-nodulation stage). In total, this comprises 8 (2^3) factor combinations for each pair (16 total species-treatments); each will have three replicates for 48 transcriptomes per species-pair comparison (768 total).

*Experimental methods*—We will establish multiple individuals of each species and will include three biological replicates for all comparisons. Also, roots of non-nodulating species will be sampled at analogous periods and stages of development. For all comparisons, plants will be grown in a 1:1 mix of sand:Turface® and watered with nutrient solutions that contain added nitrate at concentrations that do (15 mM) or do not (0.5 mM) inhibit root nodulation. RNAs will be extracted following a range of methods used in the Soltis lab, developed for successful RNA extraction across all land plants (Jordon-Thaden *et al.*, 2015). Because some of the plants are woody and non-model species, establishment and development of roots in growth facilities may be slow, and we may need to obtain transcriptomes for nodulated and non-nodulated root segments regardless of whether the non-nodulated segments are from the same or different roots. We will sequence 6 Gb of transcriptome data from each sample (a standard depth-of-coverage value assuming a standard angiosperm genic content) using the Illumina HiSeq 3000 platform. Gene expression level will be quantified using the RSEM (Li & Dewey, 2011) program that outputs Transcripts Per Million (TPM) that will be log transformed, normalized across samples, and used as inputs for downstream network and module analysis.

*B. Genome screening for nodulation genes*

*Goals*— Gene gain/loss cannot be reliably inferred from transcriptomic data. We will therefore test whether nodulation-specific genes that were not recovered in transcriptomes are in fact absent in non-host species, based on genomic queries, following the logic of Delaux *et al.* (2015). Patterns of gene family expansion/contraction will also be important for interpreting those gene copies that are expressed and potentially associated with gains of nodulation. Because genome sequencing for even the species pairs identified in *Experiment I-2* is beyond the scope of this project, we will query the genomes of members of the N-fixing clade using target capture, a highly economical and sensitive genome reduction method.

*Lab methods*—To obtain genomic data in a cost-efficient manner, we propose a two-pronged genomic screening strategy. First, we propose a taxonomically ambitious, large-scale genomic screen of the 15,000 species included in our phylogenomic work, using candidates known from previous studies to play an important role in nodule development and symbiotic infection, comprising approximately 70 candidates such as *CCaMK/DMI3, CYCLOPS/IPD3, NIN*, or *RPG* (Delaux *et al.*, 2013, 2014; Venkateshwaran *et al*., 2013; ongoing work in our labs). Second, we propose a small-scale assay, using gene candidates from identified through transcriptomic and evolutionary analyses (see below); the size of this second capture set will be determined from these analyses, but can accommodate up to 300 genes. This assay will be performed on all 16 species pairs of germplasm accessions involved in the transcriptome experiment to provide matched transcriptome- and genome-level data for optimal data interpretation. Capture products will be sequenced on an Illumina HiSeq 3000, largely following the methods developed for phylogenomic enrichment of nuclear loci.

*Analytical methods*—We will use aTRAM (Allen *et al.,* 2015), an assembly pipeline well-suited to target-capture data, to assemble all gene copies per accession. We will build individual gene trees to infer orthologs and assess gain/loss events, scoring these as a binary matrix for correlation calculations.

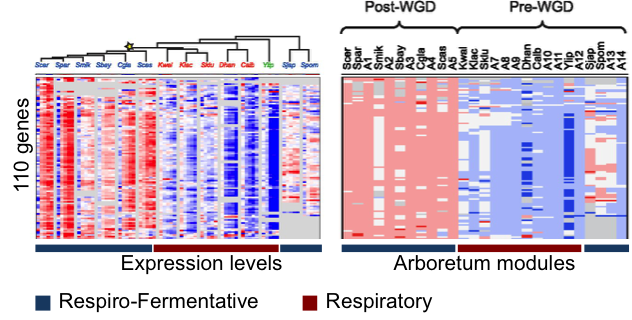
*C. Selection and prioritization of genes underlying nodulation*

To build a predictive genotype-phenotype framework, we will correlate transcriptomic and genomic data from Sections A and B above with nodulation states in our experimental species. For the genotypic component, three quantitative approaches are proposed to maximize our power to detect genotype-phenotype associations: (a) individual gene and gene expression regulatory module expression patterns; (b) gene presence/absence, and (c) rates and patterns of gene evolution. The expectation is that each analytical approach proposed will provide an independent and complementary method to assess whether or not a gene is predictive of nodulating phenotypes. Genetic data will be correlated with nodulation phenotypes described for each species by leveraging the comprehensive trait matrix assembled for *Aim I-1*. To control for similarity in observed patterns due only to close phylogenetic relationship and not due to phenotype, we will evaluate the relationship between variables (gene gain/loss, nodulation phenotypes) using phylogenetically adjusted ANOVA and MANOVA methods, and other phylogeny-aware genotype-phenotype models. We anticipate identifying up to 200 candidate genes for further screening in *Aim II*, and we will share at least 50 such candidates to *Aim II* annually, with candidates suggested later in the project having undergone greater vetting prior to *Aim II* validation than those early on.

Gene regulatory modules

*Rationale and approach*—Ideally, differences in gene expression patterns associated with different treatments and nodulation capacity (as incorporated in our approach) will yield a set of strong candidate genes with roles in nodulation. However, transcriptomes may not provide clear results for many reasons, including variation in expression levels, weak expression of single genes, altered patterns of expression among species, etc. Divergence in gene regulatory networks has repeatedly been shown to contribute to phenotypic divergence (King and Wilson, 1975; Romero *et al*., 2012). To gain insight into transcriptional regulatory networks associated with nodule development, we will systematically compare the transcriptomes of the 16 pairs of nodulating and non-nodulating species to identify gene expression modules (defined as groups of genes that are co-expressed and could be co-regulated) in each species and study their evolutionary dynamics using the Arboretum algorithm and its associated downstream analysis (Roy *et al.*, 2013; Thomson & Roy, 2013). The rationale for using Arboretum is that transcriptional gene regulatory networks are organized into regulatory modules of co-expressed genes; accounting for gene modules and networks is expected to increase our statistical power to detect nodulation-significant candidates over naïve individual gene-level expression analysis. The Arboretum algorithm aims to identify modules of co-expressed genes in multiple species related by a phylogeny. Arboretum is based on a probabilistic model of module membership evolution along a path from an ancestral to a descendent species. Two unique aspects of Arboretum are (1) the ability to reconstruct the hidden ancestral module membership, and (2) explicitly model gene trees that are not necessarily identical to the species tree because of gene duplication and loss events. The ability to handle gene duplication events, a major mechanism by which networks can rewire (Hittinger & Carroll, 2007), allows us to work with large phylogenies with complex many-to-many relationships. Using Arboretum, we have reconstructed evolutionarily conserved and diverged modules of co-expressed genes in 15 yeast species responding to glucose starvation (Thompson, Roy *et al*., 2013) and 8 species responding to multiple transient stresses (Roy *et al.*, 2013). We found significant conservation in modules across species. Genes that diverge in expression do so in lifestyle and clade-specific ways. We found that gene duplication is a significant contributor to the divergence of module memberships of genes across species.

Arboretum takes as input transcriptome profiles across multiple conditions for each species in the phylogeny, gene trees to capture complex orthology and paralogy relationships, the species tree, and the number of modules. Arboretum then identifies gene modules in each species exploiting the gene tree structure to handle complex orthology relationships. To identify gene trees, we will apply the OrthoMCL algorithm to define putatively orthologous gene clusters. Next, we will apply MUSCLE and RAxML to construct gene trees and finally use TreeFix to obtain reconciled trees. To determine the number of modules, we will use a penalized log-likelihood based criterion as described in Roy *et al*. (2013), as well as additional metrics of module coherence (e.g. the Silhouette index, module enrichment, and the number of distinct expression patterns captured by the modules).

   
**Figure 7.** Divergence of modules in a life-style specific manner. Expression (left) and module assignments (right) of genes induced (red) in respiro-fermentative yeasts but not respiratory. Red corresponds to induced expression and blue corresponds to repressed expression. Arboretum module assignments are shown in different colors (adapted from Thompson & Roy *et al.*, 2013).

*Transcriptome modules to gene candidates*—To identify important genes associated with nodulation, we will examine the Arboretum modules for the enrichment of *cis*-regulatory elements using sequence-specific motifs available in the Cis-BP (Weirauch *et al*., 2014) and JASPAR (Matheleir *et al*., 2014) databases, and Gene Ontology processes. This will enable us to identify which biological processes, pathways, and regulators are associated with each module. Arboretum infers modules for both the extant and ancestral species. We will mine these inferred module trajectories to identify individual genes as well as sets of genes exhibiting different patterns of phylogenetic conservation of module membership. We have used this approach in our 15-species yeast study (Thompson & Roy, 2013), which comprised a combination of respiratory and respire-fermentative species, and found groups of genes that were associated with one module in the respiratory species and another module in the respiro-fermentative species **(Fig. 7**). One immediate example is to look for genes that switch from a module enriched with nodule development-related processes to a module that is not enriched for nodule-specific functionality. For any given pattern with a sufficient number of genes to allow for enrichment analysis, we will again perform *cis*-regulatory enrichment to identify candidate regulatory proteins. Individual genes, gene sets with interesting patterns of module divergence (e.g. associated with clades of species forming nodules versus not forming nodules), and regulators associated with these gene sets will provide useful candidates for downstream functional analysis and validation.

Rates and patterns of gene evolution

Genes that are essential for nodulation are expected to differ in sequence evolution dynamics and may carry signatures of selection that distinguish them from homologs in non-nodulating species. We will test these hypotheses for genes that are detected in both nodulating and non-nodulating species, using diverse methods including recent approaches (e.g. Zhang and Yang, 2015). For example, dN/dS, the ratio of non-synonymous to synonymous amino acid substitutions, will be estimated from transcript assemblies between pairwise nodulating and non-nodulating species; values of <1 signify purifying selection, 0 = neutral patterns of substitution, and >1 is positive selection. We expect genes involved in nodulation to be under strong selection, and we can use dN/dS to test for selection in all candidate genes. The functional importance of a gene can also impact its evolutionary rate. In addition to rigorous comparisons of rates of gene evolution, measures of directional selection such as Tajima’s D will be estimated to compare the degree of species differentiation. Using these and other tests of selection and evolutionary rates, we will search for patterns that are consistent with genes that may be under selection or have outlier sequence evolution parameters, as we would predict for nodulation genes.

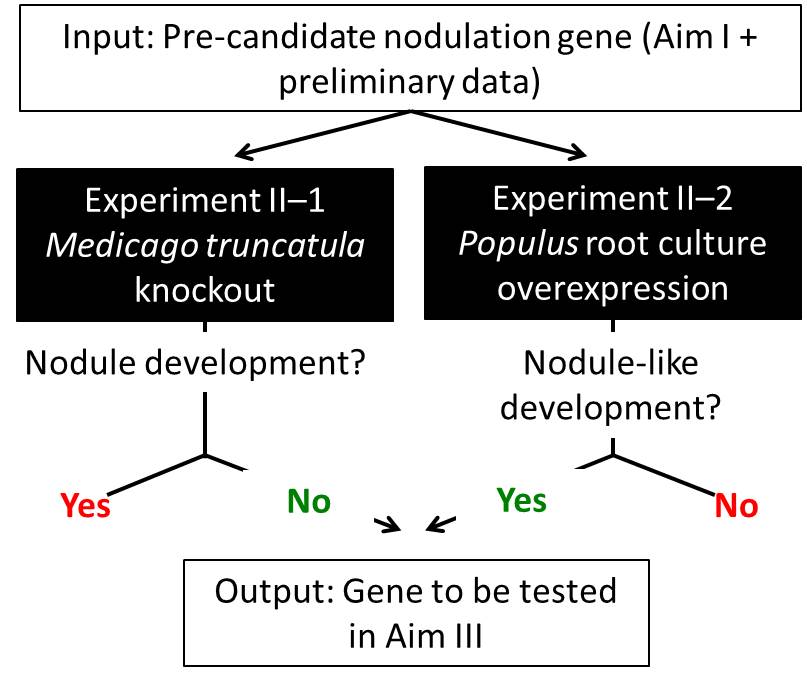
***Potential pitfalls and proposed solutions***—We are confident the phylogenomic portion of this project will be successful given our (R. Folk, D. Soltis, P. Soltis, R. Gusralnick) background in large-scale phylogenomics and the power of phylogenomic data to resolve deep and shallow relationships in the tree of life. Likewise, in collaboration with B. O’Meara, we are poised to produce cutting-edge comparative phylogenetic work with novel results for legume and non-legume predisposition. One potential pitfall is the possibility that we will find little overlap in gene expression patterns among taxa, limiting the search for an underlying toolkit that enabled the symbiosis. Our response to this challenge involves (1) the most ambitious and thorough taxonomic representation to date in this type of experiment, (2) explicitly accounting for gene expression modules and gene tree-species tree discordance due to gene gain/loss (through genomic queries of known and hypothesized nodulation genes in all species), and (3) the use of multiple complementary approaches to querying the experimental taxa for genotype-phenotype correlation (gene gain/loss, gene module/network expression, and gene evolutionary rates).

**AIM II. Verification of molecular mechanisms of root nodule development**

***Introduction—***The first aim of the proposed research will seek to uncover the genomic novelties required for the evolution of root nodule symbioses, using a comparative phylogenetic framework that contrasts related species that possess or lack the nodulation ability within subclades. *In the current aim, we will test the contribution of these novelties to root development in nodulating and non-nodulating model systems, to identify those that are critical for this developmental process.* In the last aim of the proposal, those targets that display promising phenotypes when modified in model systems will be genetically engineered into poplar, and their impact on nodule development, N-fixation, and whole plant development will be assessed.

Genetic perturbation of the molecular mechanisms that regulate root nodule development is expected to result in the disruption of this capability in nodulating species. In contrast, nodule development may be genetically engineered into species that lack this capability by introducing the gene(s) that control the nodule developmental pathway. This rationale defines the *objective* of the second aim, which will be to verify the function of target gene candidates discovered in *Aim I* for their effect on root nodule development in *M. truncatula* (nodulating) and poplar root organ cultures (non-nodulating). Our *hypothesis* is that genetic perturbation of these targets will disable nodule development and, consequently, nodulation in *M. truncatula*. In contrast, nodule-like structures are expected to develop in poplar root cultures. Modifications of candidate genes will be rapidly evaluated in these model systems before their detailed evaluation in the bioenergy woody crop poplar in the last aim of the project.

***Responsibilities, timeline, and deliverables***—Co-PI Ané will be *responsible* for coordination and execution of all research activities of *Aim II*. Because the phylogenetic analysis proposed in *Aim I* is not expected to deliver targets until the end of the first year, the activities of Aim II will focus during the first 12 months on genes for which there has been reported evidence of a significant role in nodule development. As the phylogenetic analysis performed in *Aim I* delivers novel targets; these will be gradually integrated into the pipeline of verification of function of candidates and evaluated in years 2-5. Deliverables of *Aim II* will occur in every year of the project, starting with the outcome of the analysis of candidate genes previously reported in the literature (year 1), which will be introduced or (if putative orthologs are available) over-expressed in polar root culture. In years 2-5 the deliverables will expand to the analysis of 50 candidate genes discovered in *Aim I* each year, to be introduced in polar transgenic root cultures and knocked-out in *M. truncatula*. Overall we expect that, at the conclusion of the project, *Aim II* will have delivered the evaluation of 200+ candidate genes for nodule development in *M. truncatula* and poplar root culture model systems.



**Figure 8.** *Aim II* experimental workflow. Candidate nodulation genes derived from *Aim I* (input) will be tested for function in nodule development in nodule-forming *Medicago* (*Experiment II–1*) and in the non-nodulating poplar root culture. Gene knockout that results in loss of nodule development in *Medicago*, and/or acquisition of nodule-like structures in *Populus*, provide genes to be tested (output) in whole-plants in *Aim III*.

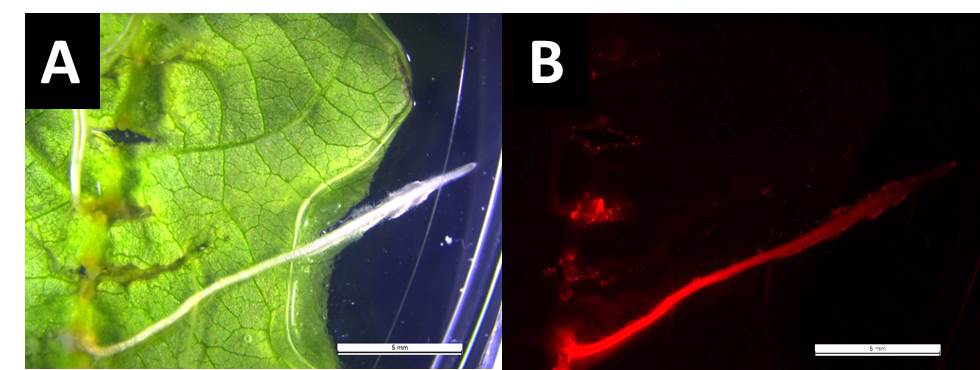
***Experimental Design***—Genes previously known to be involved in nodulation, as well as candidate genes identified in the first aim will be continuously evaluated by the co-PI Ané for their potential function in nodule development (**Fig. 8**). In the first year, known nodulation genes will be introduced in transgenic root organ cultures of poplar (non-nodulating) using *A. rhizogenes* to assess developmental changes that indicate the development of nodules (*Experiment II–1*). This effort will continue in the following years with candidate genes uncovered in *Aim I*. These novel candidates will, in parallel, be disrupted in the model *M. truncatula* (nodulating) to verify their developmental effect on the development of this nodulating species (*Experiment II–2*). Use of both plant model systems will allow rapid evaluation of large numbers of candidate genes, and their combinatorial effect using the GoldenGate assembly system, before the whole-plant evaluation of the most promising candidate genes in *Populus* in *Aim III*.

Experiment II–1. Gain-of-function in poplar transgenic poplar root culture

Although the primary candidate genes for gain-of-function studies in poplar will be those coming from *Aim I*, these candidate genes may not be available until the end of year 1. As such, the candidate genes below will be initially evaluated in poplar root culture.

*Candidate genes*—We will initiate the gain-of-function approach using candidate genes identified from the literature, including genes that encode histidine kinase cytokinin receptors (LHK/CRE1) and Nodule INception (NIN) or NIN-like proteins (NLPs). A constitutively active mutant of an LHK/CRE histidine kinase cytokinin receptor in *Lotus japonicus* is sufficient to trigger spontaneous nodule development in the absence of rhizobia (Tirichine *et al.*, 2007). Several LHK/CRE histidine kinase cytokinin receptors are present in poplar, and five LHK/CRE1 homologs (*PtCRE1a*, *PtCRE1b*, *PtHK2*, *PtHK3a* and *PtHK3b*) have been identified in the *Populus trichocarpa* genome(Nieminen *et al.*, 2008). Per the poplar expression atlas PopGenIE, *PtCRE1a*, *PtCRE1b*, *PtHK2* are expressed in roots while *PtHK3a* and *PtHK3b* are not. Similarly, overexpression of *NIN*, an essential transcription factor, induced root nodule primordium-like structures in *Lotus japonicus* that originated from cortical cells and in the absence of bacterial symbionts. These results indicate that *NIN* is not only necessary but also sufficient for initiating nodule development (Soyano *et al.*, 2013). Interestingly, *NIN* homologs have also been identified in non-legumes (Schauser *et al.*, 2005). We identified one strong homolog(*PtNIN*), and a weaker one (*PtNLP*) in the poplar genome. As with the LHK/CRE histidine kinase cytokinin receptors, *PtNIN* and *PtNLP* are expressed in roots according to PopGenIE.

*Genetic transformation of poplar root culture*—We will use GoldenGate technology to assemble multi-gene constructs and express candidate genes (Patron *et al.*, 2015). This technology is used routinely in the Co-PI Ané lab, and most of the parts (vectors, promoters, terminators, and adaptators) necessary for this project are readily available. In Year 1, we will express constitutive active mutants of *PtCRE1a*, *PtCRE1b*, *PtHK2* as well as wild-type *PtNIN* and *PtNLP* both individually and in combinations. Each gene will be driven by a different strong and constitutive eudicot promoter. We will use β-estradiol-inducible promoters for constructs that may affect development significantly and prevent root development (*i.e.* the constitutively active cytokinin receptors) (Schlücking *et al.*, 2013). These constructs will be transformed into poplar roots using *Agrobacterium rhizogenes*-mediated transformation. The CoPI Ané’s lab has optimized a fast procedure for poplar transformation using *A. rhizogenes* (**Fig. 9 A, B**). The transgenic roots can be sub-cultured to give rise to root organ cultures. These transgenic roots and root organ cultures develop lateral roots normally. Therefore, we anticipate being able to detect without any difficulty the presence of nodule-like structures. When such structures are observed, thin sections will be used to test the presence of a vasculature system, a meristem and more generally to describe the developmental process in detail as described previously (Xiao *et al.*, 2014).

****

**Figure 9.** Poplar leaves transformed with *Agrobacterium rhizogenes* ARqua1 develop transgenic roots 2-3 weeks post inoculation (A), and transgenic roots can be identified using DsRed1 fluorescence provided by the transgene (B). (A, bright field; B, DsRed1 fluorescence). Scale bars = 5 mm.

Experiment II–2. Loss-of-function approach in the model legume, *Medicago truncatula*

*Genetic transformation of M. truncatula*—Candidate genes from *Aim I* will be tested using a rapid knock-out screen with the CRISPR/Cas9 genome editing technique in transgenic roots generated with *Agrobacterium rhizogenes* as described previously (Cai *et al.*, 2015). CRISPR/Cas9 constructs and four small guide RNAs targeting the 3’ and 5’ regions of the gene (two sgRNAs per end) will be assembled into binary vectors using the GoldenGate technology. These vectors will also include a fluorescent marker (DsRed1) for rapid detection of transgenic roots. The assembled constructs will be transformed into *M. truncatula* Jemalong A17 using *A. rhizogenes* strain MSU440 as described previously (Boisson-Dernier *et al.*, 2001; Kevei *et al.*, 2007; Venkateshwaran *et al.*, 2012)*.* Two weeks post inoculation, transgenic roots growing axenically on Fahräeus medium will be identified based on DsRed1 fluorescence under a binocular scope. Transgenic roots expressing just DsRed1 will be used as positive controls. Plants with transgenic roots will be inoculated with *Sinorhizobium meliloti* strain 2011, and nodulation phenotypes will be tested as described in (Horváth *et al.*, 2011; Sun *et al.*, 2015; Venkateshwaran *et al.*, 2015). Briefly, transgenic roots will be inoculated with *S. meliloti* 2011 strains expressing the β-galactosidase or the green fluorescent protein (GFP) constitutively. At various time points after inoculation (3, 7, 14 and 21 days), we will analyze root hair deformations, infection thread formation, and cortical cell divisions. If loss of nodulation occurs or nodule number is reduced compared to the control, we will analyze the nodule development defects in detail as in (Xiao *et al.*, 2014) as well as the responses to purified signals (Nod factors) from *S. meliloti* such as nodulin gene expression and calcium spiking assays. The expression of standard nodulin genes (*ENOD11*, *ERN1*, *ENOD40* and *NIN*) will be monitored using quantitative RT-PCR and promoter-GUS fusions. Briefly, calcium spiking will be monitored using a Green-GECO calcium sensor (Uslu & Grossmann, 2016). This sensor will be expressed in transgenic roots along with the CRISPR/Cas construct. Imaging will be performed at the UW-Madison Newcomb Imaging Center. For genes showing interesting nodulation phenotypes in transgenic roots, *Tnt1*-insertion mutants in will be obtained through the Samuel Roberts Noble Foundation. If necessary, stable CRISPR/Cas9 knockouts generated at the Wisconsin Crop Improvement Center (WCIC) transformation facility. Nodulation phenotypes and responses to purified Nod factors will be performed on mutants as described previously.

***Potential pitfalls and proposed solutions***—We do not anticipate any technical pitfall for this aim since the approaches proposed are already used in the CoPI Ané’s lab. It seems very likely that loss-of-function studies in *M. truncatula* will lead to interesting nodulation phenotypes. We acknowledge that obtaining gain-of-function phenotypes is an ambitious goal, but this is also a high-reward one. It is possible that only few candidate genes from *Aim I* will lead to gain-of-function (nodule development) in poplar root culture. However, the candidate genes mentioned previously (P*tCRE1a, PtCRE1b, PtHK2, PtNIN* and *PtNLP*) offer an excellent starting hypothesis to trigger the development of nodule-like structures. While the existing phylogenetic analysis of nodulation has suggested a relatively simple mechanism of predisposition, controlled by one or few genes, it is also possible that combinations of targets may be required to trigger nodule development. To address this potential limitation, we have the ability to engineer multiple genes simultaneously (using GoldenGate) and can use information derived from the transcriptome studies derived from *Aim I* to identify and evaluate combinations of genes.

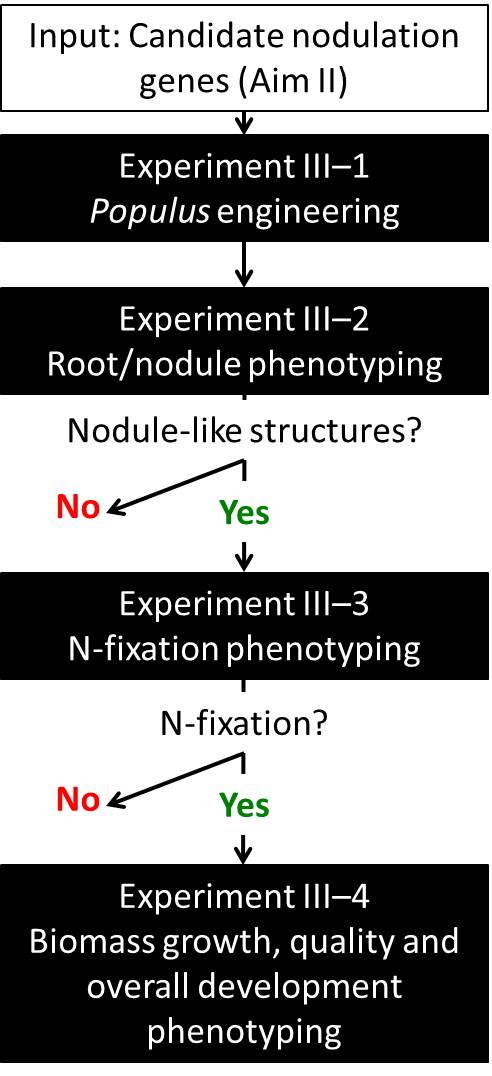
**AIM III. Engineering nodulation capability into the bioenergy woody crop Populus**

***Introduction***—The first aim of the proposed project will define genome innovations associated with nodulation. These will be tested in the second aim for loss-of-function in nodule-forming plants (*M. truncatula*) and gain-of-function in plants that lack that capability (poplar root organ cultures). While the assessment of gain-of-function in poplar root culture will identify genes that contribute to the development of nodule-like structures, it will not provide a rigorous quantification of their effect in the whole-plant phenotype in the presence of N-fixing bacteria, including N assimilation, plant development, and biomass productivity and quality. The *objective* of the last aim of this project is to address this limitation. More specifically, we will engineer nodule development genes selected in *Aim II* into poplar, evaluate the development of nodules in poplar roots, inoculate with N-fixers, and test their impact on N-fixation and whole-plant properties. The ability of nodule-like structures to support N-fixation will be evaluated with a wide range of bacteria. We previously observed that N fertilization significantly increases wood cellulose content and decreases lignin content and extractability (measured as S/G ratio). Thus, phenotyping will extend beyond N assimilation, root phenotypes, and biomass productivity to also include the quantification of changes in biomass composition.

The observation that species that are phylogenetically closely related have maintained or lost the capability of N-fixing symbioses suggests that specific, individual genes may be uniquely necessary to engineer this capability into a non-N-fixing plant. Studies have shown that the manipulation of single genes is necessary and sufficient to trigger nodule development in model legumes (e.g. Tirichine *et al.*, 2006). We *hypothesize* that, similarly, the introduction of one or few regulatory genes will be sufficient to trigger an endogenous machinery of poplar lateral root development that leads to the development of nodule-like structures. We also *hypothesize* that nodule-like structures will improve the association with N-fixing bacteria. Our expectation is that the discovery of genes that trigger the development of nodule-like structures will serve as a platform for future engineering of intracellular accommodation (infection) processes not only in poplar, but may also be transferable to other bioenergy crops.

***Responsibility, timeline, and deliverables*—**The PI Kirst will be responsible for the research activities of *Aim III*. Some of the proposed activities will require measuring N fixation; these will be supported by the co-PI Ané, who has the necessary expertise, equipment, and facilities to carry out these experiments. The effort to genetically engineer poplar for development of nodule-like structures in poplar will initiate in the first year, targeting histidine kinase cytokinin receptors (LHK/CRE) and Nodule INception (NIN) and NIN-like proteins (NLP) previously identified as critical in the process of nodule development (Tirichine *et al.*, 2006; Soyano & Hayashi, 2014). These genes will be evaluated in poplar root culture in *Aim II*, during the first six months of the project. The effort will expand in years 2-5 to include genes identified by the phylogenetic approach (*Aim I*) and verified for function in *M. truncatula* and poplarroot culture (*Aim II*). Deliverables from *Aim III* will include the engineering of up to 10 candidate genes yearly (years 2-5), and the detailed greenhouse phenotyping of transgenic lines for nodule development, N fixation, and whole-plant development.

***Experimental design*—** To engineer nodule development and test the impact of these structures on N fixation and whole-plant properties in poplar, putative nodulating genes will be introduced in the *Populus tremula × P. alba* genotype INRA 717-1-B4 and phenotyped (**Fig. 10**). This genotype will be used previously for testing candidate genes in the root culture system (*Aim II*). INRA 717-1-B4 is regularly used by the Kirst lab and in tree research in general for functional genomic studies of woody perennial species. Genes engineered into poplar (*Experiment III–1*) will be evaluated for root development to quantify the development of nodule-like structure in greenhouse conditions, inoculated with N-fixing bacteria (*Experiment III–2*). N fixation will be evaluated by Acetylene Reduction Assays (ARA), 15N dilution, and 15N gas enrichment techniques (*Experiment III–3*). Finally, phenotyping will extend to the evaluation of biomass growth and chemical/physical properties, as well as overall plant development in greenhouse conditions, to evaluate other phenotypic consequences that may impact the suitability for bioenergy (*Experiment III–4*).

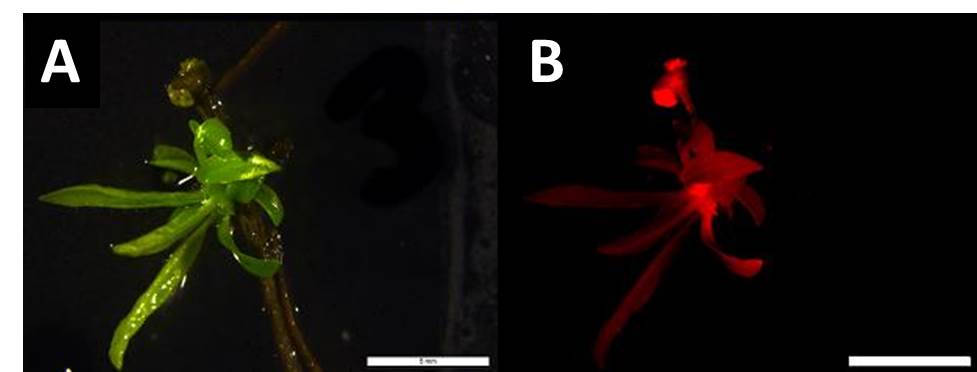


**Figure 10.** *Aim III* experimental workflow. Function of candidate nodulation genes derived from *Aim II* (input) will be verified by introducing them in poplar (*Experiment III–1*) and verifying development of nodule-like structures (*Experiment III–2*). If the phenotype is confirmed, engineered plants will be phenotyped for N-fixation capability (*Experiment III–3*) and plant development, growth and biomass composition (*Experiment III–4*).

Experiment III–1. Genetic engineering of *Populus tremula × P. alba* (clone INRA 717-1-B4)

Candidate genes will be introduced into poplar using two alternative strategies: (a) *A. rhizogenes*-mediated transformation, extending the use of transgenic lines derived from *Aim II*, or (b) classical transformation using *A. tumefaciens*.

*Agrobacterium rhizogenes*-mediated transformation—whole-plant regeneration: As described previously, the Co-PI Ané’s lab optimized poplar transformation using *A. rhizogenes* (*Experiment II–2*, **Fig. 9 A, B**) to generate transgenic roots. Furthermore, placement of these transgenic poplar roots on medium containing trans-zeatin and incubation under low-light conditions allows the development of transgenic shoots and the rapid generation of fully transformed plants in less than six months (**Fig. 11 A, B**). Because the *A. rhizogenes* transgenic plants will be generated as part of the previous aim, for every gene selected for whole-plant analysis, we will initially evaluate the possibility of generating fully transformed plants using this approach. In the case of failure, we will proceed to utilize the standard transformation method that uses *A. tumefaciens*.



**Figure 11.** A,B: Transgenic shoots generated from transgenic roots using hormones (A, bright field; B, DsRed1 fluorescence). Scale bars = 5 mm.

Agrobacterium tumefaciens*-mediated transformation*—We will develop transgenic lines using ectopic expression driven by the 35S promoter, to generate gain-of-function variants. Alternatively, expression driven by a root-specific promoter will be utilized; see *Potential pitfalls and proposed solutions*. The coding sequence of genes uncovered phylogenetically (*Aim I*) and tested by *A. rhizogenes* transformation of poplar leaves (*Aim II*) will be cloned into the pCAPO vector for overexpression. The *Agrobacterium*-mediated transformation will be performed using the strain GUV3101 in the *Populus tremula* × *Populus alba* 717-1B4 genotype. For each transgene construct, we will obtain ten independent transgenic lines, from which we will identify using qRT-PCR three lines with highest over-expressed transcript levels of the candidate genes. Rooted plants will be acclimated before being transferred 17.8 cm deep nursery pots (Classic 400, Nursery Supplies Inc., Kissimmee, FL, USA) with Fafard 4MIX soil (Canadian Sphagnum Peat 40%, processed pine bark, and vermiculite; Sun Gro Horticulture, Agawam, MA, USA). These plants will be maintained in the greenhouse as a source of vegetative clonal replicates for the experiment that will follow.

*Plant vegetative propagation*—After generation of transgenic poplar and acclimation to greenhouse conditions, plants will be propagated vegetatively to confirm root nodulation traits observed in the previous aim, and establish the resources for detailed whole-plant phenotypes to be carried out in the remaining experiments. For plant propagation, apical meristem cuttings will be planted in peat pellets and placed on benches with intermittent mist for approximately two weeks, when rooted cuttings will be hardened off for three days. Under our standard greenhouse growth conditions, the genotype INRA 717-1-B4 does not enter dormancy, resulting in the ability to generate and collect vegetative cuttings throughout the year. Following the development of rooted cuttings, plants will be transferred to distinct growth/experimental conditions to phenotype each specific trait.

Experiment III–2. Root development to quantify the development of nodule-like structures

Transgenic poplar plants will be initially grown in a greenhouse, to confirm root nodule phenotypes observed in *Aim II*.

*Greenhouse growth conditions*—Rooted cuttings will be transferred to 41-cm-deep pots to allow unrestricted root growth. Pots will contain sterile 1:1 sand:Turface®, and plants will be fertilized with a low-N Fahräeus medium and placed on ebb-n-flow benches. Turface® is a widely used plant growth substrate in studies of nodulation. Benches will be flooded twice daily with low-N Fahräeus medium.

*Inoculation with N-fixing bacteria*—As described previously (Experiment II–2; *Aim I*), plants will be inoculated with a mixture of N-fixing bacteria able to colonize a wide range of plants within the nodulating clade: *Azorhizobium caulinodans* ORS571, *Sinorhizobium* sp. BR816, *Sinorhizobium* sp. NGR234, *Bradyrhizobium* sp. ANU289, *Bradyrhizobium* sp. NC9, *Bradyrhizobium* sp. DOA9, *Rhizobium freirei* PRF 81, *Rhizobium* sp. GRH2, *Rhizobium* sp. LPU83, *Rhizobium tropici* CIAT 89, *Mesorhizobium plurifarium* ORS 1032, *Burkholderia phymatum* STM815, *Cupriavidus taiwanensis* LMG 19424, *Frankia alni* ACN14a, *Frankia* sp. BMG5.1, *Frankia* sp. CcI3, *Frankia* sp. EAN1pec, *Frankia* sp. Eul1c.

*Detection of nodule-like structures and infection threads on transgenic roots*—Nodule structures will be detected as described previously in *Experiment II–2*. Briefly, at regular (bi-weekly) intervals after inoculation, we will visually inspect roots for root hair deformations, infection thread formation, and cortical cell divisions. If nodules are detected, those will be analyzed in detail for morphology and expression of standard nodulin genes (*ENOD11*, *ERN1*, *ENOD40* and *NIN*), using quantitative RT-PCR

Experiment III–3. N-fixation transgenic in INRA 717-1-B4

In the third experiment, we will quantify N fixation in stable transgenic lines that present altered root phenotypes that indicate the presence of nodule-like structures. Alternative methods are currently adopted to quantify N fixation in plants, all of which have advantages and limitations. The simplest technical approach evaluates nitrogenase activity by acetylene reduction assays (ARA). While ARA provides a simple and rapid screen for N fixation by symbiotic bacteria, it does not quantify the assimilation of the N in the host-plant. Methods such 15N gas enrichment and 15N dilution address this limitation, but are more technically difficult and of lower throughput. In this experiment, stable lines expressing selected candidates will be initially screened for nitrogenase activity by ARA at the University of Florida. Lines that display significant nitrogenase activity will be evaluated using the 15N dilution and 15N gas enrichment (Ruppel and Merbach, 1997) techniques, providing the link between N fixation and the assimilation of fixed N by the plant. Due to their higher technical complexity, 15N dilution and 15N gas enrichment will be carried out by the co-PI Ané at the University of Wisconsin - Madison, where necessary facilities and equipment are readily available.

*Evaluation of nitrogenase activity by Acetylene Reduction Assay*—Stable lines developed in *Experiment III–1* will be grown in Magenta bottles in a growth chamber. The plants will be inoculated one month after planting with individual diazotrophs, various mixtures of these diazotrophs or with sterile water (control). The tested diazotrophs will include 18 well-characterized N-fixing bacteria able to colonize a wide range of plants within the nodulating clade, as described previously*.* Root nodules will be harvested at 5, 7, 14, 21, and 28 days post-inoculation and used for ARA as described previously (Francisco and Akao, 1993). Non-inoculated roots of the transgenic lines with and without nodules will be used as negative controls to evaluate the basal level of ethylene production of the plant. However, some of these diazotrophs (*Azorhizobium caulinodans* ORS571 and *Burkholderia phymatum* STM815) can fix N in free-living conditions, so roots that do not carry nodules and inoculated with these bacteria will also be used as negative controls. Soybean roots inoculated with *Bradyrhizobium japonicum* USDA110 will be used as positive controls. Increased ARA activity in nodulated roots will indicate that these nodules can provide a suitable environment for nitrogenase activity.

*Evaluation of N fixation and assimilation using 15N gas enrichment*—Engineered poplar plants that display evidence of nitrogenase activity from the ARA experiments will be evaluated using the 15N gas enrichment technique (Ruppel & Merbach, 1997). This experiment will provide the link between N fixation and assimilation by the plant. Plants grown in Magenta bottles in a growth chamber will be inoculated one month after planting with individual diazotrophs, mixtures of them or with sterile water (control), as described previously. One month after inoculation, 15N2 dinitrogen gas (15% (v/v)) will be injected into the Magenta bottles and incubated for one week. After incubation, nodules will be harvested from the samples and subjected to IRMS analysis at the Soil Science Analytical Chemistry and Stable Isotope Facility (UW–Madison). To further confirm that 15N enrichment in nodules is due to the transfer of fixed N from the diazotrophs to the plant (and not merely detection of the diazotrophs present in the nodules), pheophytin will be extracted from transgenic plants both in the inoculated and control samples (following the protocol from Kahn *et al*., 2002) and will be subjected toIRMS analysis.

*Evaluation of N fixation and assimilation by 15N dilution*—This experiment will be performed in pots at the Biotron Greenhouse facility, UW–Madison (<https://biotron.wisc.edu/>). Promising stable lines by ARA and 15N gas enrichment will be grown in Magenta bottles in a growth chamber for one month. These lines will be transferred to pots containing 7 kg/pot of Metro-Mix® and evaluated in four replicates. After one month in pots, 500 mg of 15N-labeled ammonium sulfate (10 atom % excess) will be added per pot. These plants will be inoculated with individual diazotrophs, mixtures of them, or sterile water (control). Soil samples, young leaf samples, and whole plant samples will be collected from these lines at 0, 7, 14, 28, and 56 days post-inoculation. Soil N content, total plant N, soil and plant 15N (using IRMS) will be analyzed for each time point to determine N fixation by the poplar transgenic lines.

Experiment III–4. Biomass and whole-plant properties of transgenic INRA 717-1-B4

To quantify biomass growth and properties in transgenic plants, as well as whole-plant developmental properties, six biological replicates of each transgenic line will be grown in a greenhouse and inoculated with N-fixing bacteria as described in *Experiment III–2*. Plants will be grown for at least eight weeks, in ebb-and-flow benches, alongside negative controls (wild-type and empty vector-transformed INRA 717-1-B4). Because we anticipate that up to 180 transgenic lines will need to be tested in parallel (10 candidate genes × three transgenic lines × six replicates), in addition to controls, plants will be accommodated in three ebb-and-flow benches, in a complete block design with two biological replicates per block. Each plant pot will be identified by a barcode for tracking of samples and physical location at the time of tissue collection.

*Tissue harvest and biomass growth and allocation measurement*—After eight weeks of growth in the greenhouse, leaf, stem, and roots will be collected and stored in a drying room for four weeks. Biomass of root, stem, and leaves will be determined separately for each plant, and ratios of biomass allocated to leaves, stems, and roots will be estimated individually. At the onset of tissue collection, the plant barcode will be scanned to record genotype, tissue, physical location and time of sampling.

*Root architecture phenotyping*—before drying the root component, they will be scanned on flatbed scanners for root architecture analysis. Image analysis software developed for the analysis of root systems (WinRhizo®, Regent Instruments, Quebec City, Canada) will be used to derive architectural traits from the digitized image of each root system. Parameters calculated will include total and frequency distributions of root length, the number of tips, projected area, surface area, and volume.

*Biomass composition (lignin analysis)—*The dried stems will be pulverized with a Wiley mill to a fine powder for measurement of cell wall chemical compositions. Pyrolysis molecular beam mass spectrometry (pyMBMS) is well suited for rapid quantification of lignin, and analytical methods have been developed and applied previously by us to the analysis of poplar. A spectrum ranging from 30 to 450 mass-to-charge (m/z) ratio will be generated for each sample analyzed (two technical replicates for each of six biological replicates) as a service by the Complex Carbohydrate Research Center (CCRC, University of Georgia). Peak intensity will be mean normalized and technical replicates will be averaged for each genotype. Peaks associated with lignin content and monomer composition will be summed as we described previously (Sykes *et al.*, 2009; Novaes *et al.*, 2009), dividing the syringyl (S) lignin peak sum by the guaiacyl (G) lignin peak sum to obtain an estimate of lignin S:G ratio for each genotype. While pyMBMS also provides estimates of carbohydrate (lignin and cellulose) content, those estimates are less reliable than those generated for lignin. Thus, for samples that display a significant departure in lignin content compared to controls, we will proceed to carry out a detailed carbohydrate analysis at CCRC/UGA (<https://ast.uga.edu/home/cellulose-hemicellulose-analysis/>).

***Potential pitfalls and proposed solutions***—The methods proposed in the current aim are all in use in the PI’s or co-PI’s laboratories. Phenotyping experiments of a larger scale than those proposed have been carried out in the past (Novaes *et al.*, 2009; Fahrenkrog *et al.*, 2017). It is possible that unexpected, deleterious phenotypic outcomes may result for the non-specific (35S driven) over-expression of candidate genes—some may result in negative consequences to overall plant development. As a possible solution to this potential limitation, we are currently developing a root-specific promoter that can replace 35S in circumstances where the original construct causes deleterious developmental effects. To identify a suitable promoter, we have identified root-specific genes from a gene transcriptome atlas that we previously generated (Quesada *et al.*, 2008), as well as new transcriptome resources developed recently (i.e. PopGenIE). Constructs that incorporate those putative root promoters driving reporter genes will be evaluated in the first six months of the project, as a possible replacement for the proposed approach.

**Research Project Management and Timeline**

Management team organization: The PI of the project, Dr. Matias Kirst, will be the overall project director. The project will fund a Research Manager who will coordinate activities among the groups at UF and UW-Madison, ensuring proper and timely transfer of information and results among them. The Research Manager is expected to spend ½ time for the duration of the project (5 years) on this activity.

External Advisory Board (EAB): The function of the advisory board will be primarily to provide support in the selection of suitable candidate genes (*Aim I*) to be moved to functional characterization in *Aim II* and *Aim III*, in addition to advising on research strategies. The selected members have an extensive and exceptional record in research on the discovery of genes involved in establishing nodulation and N fixation, and phylogenomic analysis of that process: Dr. Giles Oldroyd (John Innes Center, UK), Gijsbert D. A. Werner (University of Oxford) and Michael Udvardi (Samuel Roberts Noble Foundation, US). The EAB will meet with the project investigators bi-annually (once remotely and once in person).

PI, co-PI and collaborator responsibilities:

* *Matias Kirst* (University of Florida), PI. The PI Kirst will oversee the overall project progress, and directly coordinate the activities of *Aim III*. These activities will involve testing genes selected for further analysis in *Aim II*, by genetically engineering them into poplar and phenotyping transgenic plants for nodulation, N fixation, and overall development. Kirst will supervise the Research Manager.
* *Jean-Michel Ané* (University of Wisconsin - Madison), co-PI. The co-PI Ané will be responsible for research activities of Aim II, the evaluation of genes uncovered by the comparative phylogenomic approach. He will also play a critical role in the selection of which genes to pursue, considering his extensive background in the molecular process of symbiosis and N fixation. Finally, Ané will also support the research activities involved in evaluating N-fixation in whole-plant (*Aim III*).
* *Douglas Soltis and Pamela Soltis* (University of Florida), co-PIs. The co-PIs Soltis will be responsible for research activities of *Aim I*, specifically the acquisition of plant resources (samples and data) and the follow-up phylogenomic comparative analysis.
* *Robert Guralnick* (University of Florida), co-PI; *Ryan Folk* (University of Florida), Sr. Pers. Guralnick and Folk will be responsible for data acquisition and management for the phylogenetic analysis, as well as the following steps required for selection of genes to be tested in *Aim II*.
* *Sushmita Roy* (University of Wisconsin), co-PI. Roy is an expert in transcriptome analysis and will coordinate all the activities involved in the transcriptome analysis of nodulation, that are part of *Experiment I–3*.
* *Brian O’Meara* (University of Tennessee Knoxville), collaborator. O’Meara will provide expertise in the development of improved HMM approaches and other comparative phylogenetic tasks for the phylogenomic analysis of *Aim I*.

Communication:

* *In person meetings:* Investigators (PI and co-PIs) will meet in person at least once annually for a project meeting. Annual meetings will be followed by a joint meeting with the EAB, to discuss annual progress and research strategy for the following year. Also, we anticipate that investigators and postdocs will be involved in additional in-person meetings aimed at developing and exchanging technical capabilities.
* *Teleconferences:* Teleconferences will be held quarterly among the PI and co-PIs to coordinate data sharing among participating groups.
* *Website:* A project website will be developed to communicate project progress and make project outcomes easily available to the public.

The timeline below describes the timeframe expected for completion of milestones of the project.



**APPENDIX 1: BIOGRAPHICAL SKETCH**

**JEAN-MICHEL ANÉ**

Departments of Bacteriology and Agronomy

University of Wisconsin

jeanmichel.ane@wisc.edu

**a. Professional preparation**

University of Lyon (France) B.S. Molecular and Cellular Biology 1995

University of Toulouse (France) M.S. Plant Cellular and Molecular Biology 1997

University of Toulouse (France) Ph.D. Plant Cellular and Molecular Biology 2002

University of California Davis Post-doc. Plant Pathology 2002-2004

**b. Appointments**

2014 - present Professor University of Wisconsin - Madison

2010 – 2014 Associate Professor University of Wisconsin - Madison

2004 - 2010 Assistant Professor University of Wisconsin - Madison

2002-2004 Postdoctoral Researcher University of California - Davis

**c. Publications (10 most related to proposal)**

Kamel L, Keller-Pearson M, Roux C, Ané JM. Biology and evolution of arbuscular mycorrhizal symbiosis in the light of genomics. **New Phytologist** (in press).

Mus F, Crook M, Garcia K, Garcia-Costas A, Geddes B, Kouri ED, Paramasivan P, Ryu MH, Oldroyd GED, Poole P, Udvardi M, Voigt CA, Ané JM, Peters J (2016). Symbiotic nitrogen fixation and challenges to extending it to non-legumes. **Applied and Env. Microbiology** 82(13): 3698–3710.

Marx H, Minogue CE, Jayaraman D, Richards AL, Kwiecien NW, Sihapirani AF, Rajasekar S, Maeda J, Garcia K, Del Valle-Echevarria AR, Volkening J, Westphall MS, Roy S, Sussman MR, Ané JM, Coon JJ (2016). A Proteomic Atlas of the Legume, *Medicago truncatula*, and its Nitrogen-Fixing Endosymbiont, *Sinorhizobium meliloti*. **Nature Biotechnology** 34, 1198–1205.

Poinsot V, Crook M, Erdn S, Maillet F, Bascaulesa A, Ané JM (2016). New insights into Nod factor biosynthesis: Analyses of chitooligomers and lipo-chitooligomers of *Rhizobium* sp. IRBG74 mutants. **Carbohydrate Research** 434, 83–93.

Delaux PM, Radhakrishnan GV, Jayaraman D, Cheemah J, Malbreil M, Volkening JD, Sekimoto H, Nishiyama T, Melkonian M, Pokorny L, Rothfels CJ, Winter-Sederoff H, Stevenson DW, Surek B, Zhang Y, Sussman MR, Dunand C, Morris RJ, Roux C, Wong GKS, Oldroyd GED, Ané JM (2015). The algal ancestor of land plants was pre-adapted for symbiosis. **Proceedings of the National Academy of Sciences USA** 112 (43) 13390–13395.

Garcia K, Delaux PM, Cope K, Ané JM (2015). Molecular signals required for the establishment and maintenance of ectomycorrhizal symbioses. **New Phytologist** 208, 79-87.

Venkateshwaran M, Jayaraman D, Chabaud M, Genre A, Balloon AJ, Maeda J, Forshey K, Den Os D, Kwiecien NW, Coon JJ, Barker DG, Ané JM (2015). A role for the mevalonate pathway in early plant symbiotic signaling. **Proceedings of the National Academy of Sciences USA** 112 (31), 9781-9786.

Delaux PM, Varala K, Edger PP, Pires JC, Ané JM (2014). Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. **PLoS Genet.,** 10(7): e1004487.

Jayaraman D, Gilroy S, Ané JM (2014) Staying in touch: mechanical signals in plant-microbe interactions. **Current Opinion in Plant Biology**, 20:104–109.

Delaux PM, Séjalon-Delmas N, Bécard G, Ané JM (2013) Evolution of the plant – microbe symbiotic “toolkit”. **Trends in Plants Sciences**, 18(6), 298-304.

**d. Synergistic Activities**

* Participation to several Summer Research Programs (SRPs) at the University of Wisconsin - Madison offering training and research experience to undergraduate students of underrepresented groups, and to the PEOPLE program offering similar research experiences to high school students of underrepresented groups.
* Outreach activities for the Wisconsin Soybean Marketing Board, the Wisconsin Crop Improvement Association, the Family Horticulture day, the Science Expeditions, Darwin day, Martin Luther King Day of Service, the French-American Science Festival (Chicago), Middle School Science Symposium, and the Madison Children’s Museum.

**e. Collaborators and Other Affiliations**

***(i) Collaborators/co-authors***

Balloon A (UW Madison); Barker D (CNRS, France); Barrett-Wilt G (UW Madison); Barthelemy-Delaux C (UW Madison); Bascaules (Paul Sabatier U, France); Bécard G (U Toulouse); Blackwell H (UW Madison); Chabaud M (CNRS, France); Cheema J (JIC, UK); Chen R (UW Madison); Conley S (UW Madison); Cook D (UC Davis); Coon J (UW Madison); Coruzzi G (NYU); Cottaz S (CNRS, France); Das S (UW Madison); den Os D (Hanze U of Applied Sciences, Netherlands); Edger P (U Missouri); Erdn S (CNRS, France); Esker P (U Costa Rica); Farré G (JIC, UK); Fort S (CNRS, France); Garcia Costas A (Montana SU); Geddes B (U Oxford); Gemperline E (UW Madison); Genre A (U Torino); Gilroy S (UW Madison); Gobbato E (Agricultural Biotech C, Hungary); Graham D (UW Madison); Graham J (UW Madison); Graham L (UW Madison); Granqvist E (JIC, UK); Gyaneshwar P (UW Milwaukee); Haseloff J (U Cambridge); Howes-Podoll M (UW Madison); James E (James Hutton Inst, UK); Kamel L (U Toulouse, France); Kaspar C (UW Madison); Keller C (UW Madison); Knack J (UW Madison); Kouri ED (Noble Foundation); Kwiecien (UW Madison); Lauer J (UW Madison); Lazar S (Florida Inst Tech); Li L (UW Madison); Maeda J (UW Madison); Maillet F (INRA, France); Marx H (UW Madison); Minogue C (UW Maidson); Mitra S (UW Milwaukee); Mus F (Montana SU); Nguyen T (UW Madison); Oldroyd G (JIC, UK); Owen H (UW Madison); Palmer A (Florida Inst Tech); Paramasivan P (JIC, UK); Patron N (Sainsbury Lab, UK); Peters J (Montana SU); Piotrowski M (UW Madison); Pires C (U Missouri); Poinsot V (Paul Sabatier U, France); Poole P (JIC, UK); Rajasekar S (UW Madison); Reddy P (Energy Resources Inst, India); Richards A (UW Madison); Rose CM (UW Madison); Roux C (U Toulouse); Roy S (UW Madison); Ryu M (MIT); Sadowsky M (U Minnesota); Samain E (CNRS, France); Séjalon-Delmas N (ENSAT, France); Senechal A (UW Madison); Siahpirani A (UW Madison); Stacy D (UW Madison); Sun J (JIC, UK); Sussman M (UW Madison); Udvardi M (Noble Foundation); Van Ness L (UW Madison); Varala K (NYU); Voigt C (MIT); Volkening J (UW Madison); Westphall M (UW Madison); Wilcox L (UW Madison); Ye H (UW Madison).

***(ii) Graduate (2) and Postdoctoral (1) Advisors***

Jean Dénarié and Charles Rosenberg (Graduate Advisors, INRA Toulouse, France)

Douglas R. Cook (Postdoctoral Advisor, U. of California Davis)

***(iii) Thesis Advisor and Postgraduate Scholar Sponsor***

Li Huey Yeun (Ph.D., May 2011, UW Madison); Kari Forshey (M.S., 2012, UW Madison); Audrey K. Wiley (Ph.D., 2015, UW Madison); Arijit Mukherjee (Postgraduate Scholar, January 2008 – August 2012); David Marburger (Ph.D., 2016, UW Madison); Dhileepkumar Jayaraman (Postgraduate Scholar, June 2011 – December 2012, September 2013 - present); Muthusubramanian Venkateshwaran (Postgraduate Scholar, May 2010 – August 2013); Pierre-Marc Delaux, (Postgraduate Scholar, November 2011 – September 2014); Oswaldo Valdés-López (Postgraduate Scholar, October 2012 – August 2013); Matthew Crook (Postgraduate Scholar, January 2013 – present); Kevin Cope (Ph.D. expected in 2017, UW Madison); Taylor Wahlig (Ph.D. expected in 2017, UW Madison); Kevin Garcia (Postgraduate Scholar, February 2014 – present), Angel R.Valle-Echevarria (Postgraduate Scholar, January 2015 – January 2016); Marian Lund (Ph.D. expected in 2018, UW Madison); Michelle Keller-Pearson (Ph.D. expected in 2018, UW Madison); Shane Bernard (Ph.D. expected in 2018, UW Madison); Anthony Bortolazzo (Ph.D. expected in 2018, UW Madison); Zachary Keyser (Ph.D. expected in 2020, UW Madison); Ray Collier (Postgraduate Scholar, July 2016 – present)

**Robert P. Guralnick**

Department of Natural History and the [Florida Museum of Natural History](http://www.flmnh.ufl.edu/)

University of Florida

rguralnick@flmnh.ufl.edu

**a. Professional preparation**

U.C. Berkeley Berkeley, CA Psychology BA with high honors 1992

U.C. Berkeley Berkeley, CA Integrative Biology Doctor of Philosophy 1999

U.C. Berkeley Berkeley, CA Postdoctoral Fellow 1999

**b. Appointments**

Associate Curator 2014-present Dept. of Nat. Hist. University of Florida

Associate Professor 2007-2014 Ecol. & Evol. Biol. University of Colorado

Assistant Professor 2000-2007 Ecol. & Evol. Biol. University of Colorado

Curator of Zoology 2000-2014 CU Museum of University of Colorado

Natural History

**c. Publications (10 most related to proposal)**

Meyer, C, H. Kreft, R. Guralnick and W. Jetz. 2015. Global priorities for an effective knowledge base of biodiversity distributions. *Nature Communications* 6:8221[DOI:10.1038/ncomms9221]

Metcalf, J., S. Prost, D. Nogues-Bravo, E. DeChaine, C. Anderson, M. Araujo, A. Cooper and R. P Guralnick\*. 2014. Integrating multiple lines of evidence into historical biogeography hypothesis testing: A *Bison bison* case study. Proceedings of the Royal Society B 281: 20132782. [corresponding author]

Prost, S, R. Guralnick, V. B. Fedorov, E. Kuzmina, N. Smirnov, T. van Kolfschoten, E. Waltari, M. Hofreiter, and K. Vrieling. 2013. Losing ground: Past history and future fate of Arctic small mammals in a changing climate. *Global Change Biology* 19(6):1854-1864. [DOI: 10.1111/gcb.12157]

Parr, C. S., R. P. Guralnick\*, N. Cellinese, and R. Page. 2012. Evolutionary informatics: Unifying knowledge About life’s diversity. *Trends in Ecology and Evolution* 27(2):94–103

Jetz, W., J. MacPherson and R. P. Guralnick. 2012. Integrating biodiversity distribution knowledge: Toward a global map Of life. *Trends in Ecology and Evolution* 27(3):151-159 [doi:10.1016/j.tree.2011.09.007].

Yilmaz, P., R. Kottman, D. Field, R. Knight, J. R. Cole, and 63 other co-authors (31st of 63). 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology* 29: 415–420*.*

Constable, H., R. P. Guralnick, J. Wieczorek, C. Spencer, A. T. Peterson and the VertNet Steering Committee. 2010. VertNet: A new model for biodiversity data sharing*. PLoS Biology* 8(2): e1000309

Hill, A. W., R. P. Guralnick, M. J. C. Wilson, F. Habib and D. Janies. 2009. Evolution of drug Resistance in multiple distinct lineages of H5N1 Avian Influenza. *Infection, Genetics and Evolution* 9:163-172

Map of Life ([http://mappinglife.org](http://mappinglife.org/)) highly accessed and awarded website mapping application and mobile app for examining multiple different distributional products for hundreds of thousands of species.

Flemons, P., R. Guralnick, J. Krieger, A. Ranipeta, and D. Neufeld. 2007. A Web based GIS Tool for Exploring the World's Biodiversity: The global biodiversity information facility mapping and analysis portal application (GBIF MAPA). *Ecological Informatics* 2:49-60.

**d. Synergistic Activities**

* Service to the scientific community: I serve as an Associated Editor for the journal Biodiversity Informatics and PLoS ONE. Both are open access and fit an ethical stance about the importance of open access and open data. I have coordinated multiple community-oriented projects, including leading working groups or catalysis meetings at EOL-BioSync, NESCent, NCEAS and via ESA in the past 5 years. I have been the Steering Committee Chair for a data sharing project called VertNet (vertnet.org) and have served as Steering Committee member or Co-Chair for iEvoBio, a satellite conference with Evolution (2010-2012). I served on a President’s Council of Advisors on Science and Technology working group on biodiversity and ecosystems in 2010 and on multiple advisory boards. I have organized multiple workshops and hackathons (two on citizen science and one as part of an International Biogeographic Society workshop series); these focus in particular on graduate training courses outside traditional class programs. More recently, I have co-moderated a Powell Center workshop on formalizing phenology data integration approaches, and am a member of the NEON Science, Technology, and Education Advisory Committee.
* Service to the broader community: I have served as a member of Board of Directors and am the current Vice President of the Board (after two years as Board President) of the JRS Biodiversity Foundation. This is a primarily grant-making foundation that works in Africa and Latin America. Executive roles on the Board requires significant time during the year plus two in-person board meetings a year. I also serve on an oversight board for CONABIO in Mexico. I have been active in research review and evaluation committees at CU Boulder and continue service at UF.
* Broadening participation: My lab group has always remained diverse along gender and cultural lines. I have endeavored to bring undergraduates and volunteers both into my lab and into Museum collections. I have mentored 4 female honor’s students, two of whom are now finishing Ph.D. degrees and two who have remained involved in research activities and field station work. I have focused my training more and more towards international students at the graduate and postdoctoral level (1 Spanish and 3 Indian postdoctoral students or graduate students) and am actively bringing on board more underrepresented students (6 so far).

**e. Collaborators and Other Affiliations**

***(i) Collaborators/co-authors***

Nico Cellinese (Univ. of Florida), Carla Cicero (Univ. of California, Berkeley), Joel Cracraft (American Museum of Natural History), John Deck (Univ. of California at Berkeley), Kitty Emery (Univ. Florida), Clint Francis (Cal Poly), Walter Jetz (Yale), Akito Kawahara (Univ. Florida), David Lohman (CUNY), Austin Mast (Florida State University), Carsten Meyer (iDiv); Lucinda McDade (Rancho Santa Ana Botanic Gardens), Bruce MacFadden (UF), Doug Soltis (UF), Pam Soltis (UF), Chris Ray (CU Boulder), Leslie Reis (Georgetown), Dave Vieglais (University of Kansas), Ramona Walls (iPlant/UA), John Wieczorek (University of California at Berkeley).

***(ii) Advisors***

Dr. David Lindberg, University of California at Berkeley

***(iii) Thesis Advisor and Postgraduate-Scholar Sponsor***

**CU EBIO Masters and Ph.D. program (current)**: Aidan Beers; EBIO (M), Brian Stuckey (Ph.D., Co-advise w. Deane Bowers, now also postdoc @ UF), Gaurav Vaidya (Ph.D.), Nathan Kleist (Ph.D.). CU **EBIO Graduate Program (graduated)**: Kathleen Weaver (University of LaVerne, tenured), Jonathan Krieger (Royal Botanic Gardens, Kew), David Daitch (Natural Resources Lead at SWCA Environmental Consultants), Heidi Schutz (Pacific Luthern University, tenure-track), Andrew Hill (Chief Scientist, Vizzuality), Peter Erb (Biological Sciences Initiative), Liesl Peterson (Warren Wilson College), Robert Jadin (Northwestern University), Natalie Robinson (NEON). **Postdoctoral students**: Jess Oswald (UF, 2016-present), Julie Allen (UF, 2016-present), Vijay Barve (UF, 2015-), Narayani Barve (UF, 2015-), Hannah Owens (UF, 2015-), Ryan Folk (UF, 2015-), Chris Ray (CU, 2014); Javier Otegui (CU, 2012-2014), Stephen Mayor (CU, 2013-present), Tamara Anderson (2005-2006), Eric Waltari (2005-2007).

**MATIAS KIRST**

Genetics Institute and School of Forest Resources and Conservation

University of Florida

[mkirst@ufl.edu](mailto:mkirst@ufl.edu)

**a. Professional preparation**

Federal University of Santa Maria (Brazil) Forestry Engineering B.S. 1996

Federal University of Viçosa (Brazil) Genetics and Improvement M.Sc. 1999

NC State University Genetics and

Functional Genomics (co-major) Ph.D. 2003

Cornell University (Buckler Lab) Genomics of maize diversity Postdoc 2004

**b. Appointments**

2016 Director, Plant Molecular and Cellular Biology Graduate Program, University of Florida

2015 Professor, School of Forest Resources and Conservation, University of Florida

2011 Founder RAPiD Genomics LLC, CEO (2011-2012), Scientific Consultant (2012-present)

2010 Associate Professor, School of Forest Resources and Conservation, University of Florida

2010 Co-Director Cooperative Forest Genetics Research Program

2005 Member Genetics Institute, University of Florida

2004 Assistant Professor, School of Forest Resources and Conservation, University of Florida

**c. Publications (10 most related to proposal)**

Fahrenkrog AM, Neves LG, Resende MF Jr, Vazquez AI, de Los Campos G, Dervinis C, Sykes R, Davis M, Davenport R, Barbazuk WB, Kirst M. Genome-wide association study reveals putative regulators of bioenergy traits in Populus deltoides. New Phytol. 2017 Jan;213(2):799-811.

Ribeiro CL, Silva CM, Drost DR, Novaes E, Novaes CR, Dervinis C, Kirst M. Integration of genetic, genomic and transcriptomic information identifies putative regulators of adventitious root formation in Populus. BMC Plant Biol.2016 Mar 16;16:66.

Drost DR, Puranik S, Novaes E, Novaes CR, Dervinis C, Gailing O, Kirst M. Genetical genomics of Populus leaf shape variation. BMC Plant Biol. 2015 Jun 30;15:166.

Harfouche A, Meilan R, Kirst M, Morgante M, Boerjan W, Sabatti M, Scarascia Mugnozza G. Accelerating the domestication of forest trees in a changing world. Trends Plant Sci. 2012 Feb;17(2):64-72.

Novaes E, Kirst M, Chiang V, Winter-Sederoff H, Sederoff R. Lignin and biomass: a negative correlation for wood formation and lignin content in trees. Plant Physiol. 2010 Oct;154(2):555-61.

Drost DR, Benedict CI, Berg A, Novaes E, Novaes CR, Yu Q, Dervinis C, Maia JM, Yap J, Miles B, Kirst M. Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of Populus. Proc Natl Acad Sci U S A. 2010 May 4;107(18):8492-7.

Sykes R, Yung M, Novaes E, Kirst M, Peter G, Davis M. High-throughput screening of plant cell-wall composition using pyrolysis molecular beam mass spectroscopy. Methods Mol Biol. 2009;581:169-83.

Novaes E, Osorio L, Drost DR, Miles BL, Boaventura-Novaes CR, Benedict C, Dervinis C, Yu Q, Sykes R, Davis M, Martin TA, Peter GF, Kirst M. Quantitative genetic analysis of biomass and wood chemistry of Populus under different nitrogen levels. New Phytol. 2009 Jun;182(4):878-90.

Drost DR, Novaes E, Boaventura-Novaes C, Benedict CI, Brown RS, Yin T, Tuskan GA, Kirst M. A microarray-based genotyping and genetic mapping approach for highly heterozygous outcrossing species enables localization of a large fraction of the unassembled Populus trichocarpa genome sequence. Plant J. 2009 Jun;58(6):1054-67.

Tuskan et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 2006 Sep 15;313(5793):1596-604.

**d. Synergistic Activities**

* *Industry/University research collaboration –*Kirst is the co-Director of the Cooperative Forest Genetics Research Program. Kirst also co-founded RAPiD Genomics LLC ([www.rapid-genomics.com](http://www.rapid-genomics.com)), a start-up company that provides genotyping platforms to academic scientists and industry. (2) *Professional/graduate training –* Kirst leads the development and organization of several training workshops on genomic prediction, including the "Phenotype Prediction Using Genomic Data Workshop"; (3) *Mentoring of high-school minority students and teachers* **–** Kirst hosts two or more high-school students yearly, and host several high-school teachers in his lab; (4) *Graduate and high-school curriculum development –* Kirst developed the curriculum of the main course in advanced genomics at the University of Florida, and is currently the lead instructor; (5) *Scientific or external advisory board member*for four projects funded by Genome Canada and the European Commission (2011-present).

**e. Collaborators and Other Affiliations**

**Collaborators and co-editors:** Albert, Patrice S. (U. of Missouri); Barbazuk, William “Brad” (UF); Berg, Arthur (Penn State U.); Birchler, James A. (U. of Missouri); Bradbury, Peter (USDA & Cornell); Chiang, Vincent L. (NCSU); Coluccio, Alison E. (USDA); Danilova, Tatiana V. (U. of Missouri); Davis John M. (UF); Davis, Mark (NREL, DOE); Dervinis, Christopher (UF); DiFazio, Steve (UWV); Eckert, Andrew (U. of Virginia); Ferl, Robert (UF); Fernando Rohan L. (Iowa State U. ); Garrick, Dorian J. (Iowa State U. ); Gezan, Salvador (UF); Guy, Charles (UF); Holliday, Jason (Virginia Tech); Huber, Dudley A. (UF); Isik Fikret (NCSU); Jokela, Eric J. (UF); Kochian, Leon V. (USDA); Krill, Allison M. (USDA); Ma, Jian F. (UF); Jessica M. (Duke U. ); Maron, Lyza G. (Cornell); Martin, Tim A. (UF); Meilan, Richard (Purdue U.); Miles, Brianna (US Forest Service); Mitani, Namiki (UF); Muñoz, Patricio (UF); Neale, David B (UC Davis); Noirot, Céline (INRA, France); Novaes, Evandro (Universidade Federal de Goias, Brazil); Pappas Jr., Georgios (Universidade de Brasilia, Brazil); Paul, Anna-Lisa (UF); Peter, Gary F. (UF); Pineros, Miguel A. (USDA); Plomion, Christophe (INRA, France); Prosdocimi, Francisco (EMBRAPA, Brazil); Quesada, Tania (UF); Ralph, Steve (UND); Rathinasabapathi, Bala (UF); Riveros, Alejandro (UF); Rongling, Wu (Penn State U. ); Schatz, Michael C. (Cold Spring Harbor Laboratory); Schuerger, Andrew C. (UF); Soltis, Doug (UF); Soltis, Pamela (UF); Sykes, Robert (NREL, DOE); Tieman, Denise M. (UF); Tuskan, Gerald A. (ORNL, DOE); Visscher, Anne M. (UF); Wegrzyn, Jill L. (UConn); Westbrook, Jared (UF); Whetten, Ross (NCSU); Wing, Rod A. (U. of Arizona); Winter-Sederoff, Heike (NCSU).

**Graduate and postdoctoral advisors:** Buckler, Edward (USDA & Cornell); Grattapaglia, Dario (EMBRAPA, Brazil); Mackay, Trudy (NCSU); Sederoff, Ronald R. (NCSU)

**Thesis Advisor and Postgraduate-Scholar Sponsor:** Acosta, Juan J. (Research Assistant Professor, NCSU); Drost, Derek (Seminis, USA); Fahrenkrog, Annette (Ph.D. student, UF); Neves, Leandro (CSO, RAPiD Genomics LLC); Novaes, Evandro (Assistant Professor, Universidade Federal de Goias, Brazil); Leite Ribeiro, Cíntia (Ph.D. student, UF); Munoz, Patricio (Assistant Professor, UF); Resende, Márcio (Research Associate, UF); Resztak, Justyna (Ph.D. student, UF).

**SUSHMITA ROY**

Biostatistics and Medical Informatics

University of Wisconsin – Madison

sroy@biostat.wisc.edu

**a. Professional Preparation**

University of Pune B.E. Computer Engineering 2000

University of New Mexico PhD Computer Science 2009

Broad Institute of MIT & Harvard Postdoc Computational Biology 2009-2011

**b. Appointments**

Assistant Professor, Biostatistics and Medical Informatics, University of Wisconsin, Madison. 2012-present.

**c. Publications**

K. Garcia, D. Chasman, S. Roy, and J.-M. Ane, "Physiological responses and gene co-expression network of mycorrhizal roots under k+ deprivation," *Plant Physiology*, pp. pp.01 959.2016+, Feb. 2017. [Online]. Available: http://dx.doi.org/10.1104/pp.16.01959

H. Marx, C. E. Minogue, D. Jayaraman, A. L. Richards, N. W. Kwiecien, A. F. Sihapirani, S. Rajasekar, J. Maeda, K. Garcia, A. R. Del Valle-Echevarria, J. D. Volkening, M. S. Westphall, S. Roy, M. R. Sussman, J. M. Ane, and J. J. Coon, "A proteomic atlas of the legume medicago truncatula and its nitrogen-fixing endosymbiont sinorhizobium meliloti," *Nature Biotechnology*, vol. advance online publication, Oct. 2016. [Online]. Available: <http://dx.doi.org/10.1038/nbt.3681>

D. Chasman, K. B. Walters, T. J. S. Lopes, A. J. Eisfeld, Y. Kawaoka, and S. Roy, "Integrating transcriptomic and proteomic data using predictive regulatory network models of host response to pathogens," *PLoS Comput Biol*, vol. 12, no. 7, pp. e1 005 013+, Jul. 2016.

Z. Niu, D. Chasman, A. J. Eisfeld, Y. Kawaoka, and **S. Roy**, "Multi-task consensus clustering of genome-wide transcriptomes from related biological conditions." *Bioinformatics (Oxford, England)*, Jan. 2016. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/26801959

E. Larrainzar, B. Riely, S. C. Kim, N. Carrasquilla-Garcia, H. J. Yu, H.-J. Hwang, M. Oh, G. B. Kim, A. Surendrarao, D. Chasman, A. F. Siahpirani, R. V. Penmetsa, G.-S. Lee, N. Kim, **S. Roy**, J.-H. Mun, and D. R. Cook, "Deep sequencing of the medicago truncatula root transcriptome reveals a massive and early interaction between nod factor and ethylene signals," *Plant Physiology*, pp. pp.00 350.2015+, Jul. 2015. [Online]. Available: http://dx.doi.org/10.1104/pp.15.00350

K. A. Tran, S. A. Jackson, Z. P. Olufs, N. Z. Z. Zaidan, N. Leng, C. Kendziorski, **S. Roy**, and R. Sridharan. (2015) "Collaborative rewiring of the pluripotency network by chromatin and signalling modulating pathways." *Nature communications*, vol. 6, 2015. [Online]. Available: <http://view.ncbi.nlm.nih.gov/pubmed/25650115>

**S. Roy**, S. Lagree, Z. Hou, J. A. Thomson, R. Stewart, A. P. Gasch.(2013) Integrated module and gene-specific regulatory inference implicates upstream signaling networks. ," ***PLoS Comput Biol***, vol. 9, no. 10, pp. e1 003 252+, Oct. 2013. [Online]. Available: <http://dx.doi.org/10.1371/journal.pcbi.1003252>.

D. Thompson+, **S. Roy**+, Michelle Chan, Mark Styczynski, Jenna Pfiffner, Courtney French, Amanda Socha, Anne Thielke, Sara Napolitano, Paul Muller Jr., Manolis Kellis, Jay Konieczka, Ilan Wapinski1,5 and Aviv Regev (2013). Evolutionary principles of modular gene regulation in yeasts. (+Equal contribution). *eLife*, vol. 2, Jun. 2013. [Online]. Available: http://dx.doi.org/10.7554/elife.00603

**S.Roy**, I.Wapinski, J.Pfiffner, C.French, A.Socha, J.Konieczka, N.Habib, M.Kellis, D.Thompson, and A. Regev (2013). Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. Genome Research. vol. 23, no. 6, pp. 1039-1050, Jun. 2013. [Online]. Available: http://dx.doi.org/10.1101/gr.146233.112

N. Rhind, Z. Chen, M. Yassour+, D. A. Thompson+, B. J. Haas+, N. Habib+, I. Wapinski+, **S. Roy**+, M. F. Lin, D. I. Heiman, et al**.,** Comparative Functional Genomics of the Fission Yeasts**.** *Science*, vol. 332, no. 6032, pp. 930-936, May 2011. [Online]. Available: http://dx.doi.org/10.1126/science.1203357. (+Equal contributions).

**d. Synergistic Activities**

1. Reviewer for Bioinformatics, Intelligent Systems in Molecular Biology, International Journal of Artificial Intelligence research, International Journal of Approximate Reasoning, International Conference of Machine Learning, IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics, Proteins, BMC Bioinformatics, Association of Artificial Intelligence, Pacific Symposium of Biocomputing, IEEE Transactions in Neural Networks, PLos One, Genome Biology, Plos compbio, ACM-BCB 2013 Program Committee.

2. NIH Early Career Reviewer for Genetic Variation and Evolution, NSF IIS panel member, NSF Ad hoc reviewer.

3. Steering committee of 2012 Math-Bio Symposium, University of Wisconsin, Madison

4. Mentor for UW Madison Computational Biology and Biostatistics program 2012,2015

5. Mentor for Broad institute Summer Research Program in Genomics (SRPG) 2010,2011.

**e. Collaborators & Other Affiliations:** Ferhat Ay, U. of Washington; Aviv Regev, Broad institute/MIT; Manolis Kellis, Broad institute/MIT; Margaret Werner-Washburne, U. of New Mexico; Terran Lane, U. of New Mexico; Rupa Sridharan, U. of Wisconsin; Randolph Ashton, U. of Wisconsin; William Murphy, U. of Wisconsin; Joshua Coon, U. of Wisconsin; Michael Sussman, U. of Wisconsin; Jean-Michel Ane, U. of Wisconsin; Audrey Gasch, U. of Wisconsin; Ron Stewart, U. of Wisconsin; James Thomson, U. of Wisconsin; Dawn Anne Thompson, U. of Wisconsin; Chris Hittinger, U. of Wisconsin; Dave Pagliarini, U. of Wisconsin; John Denu, U. of Wisconsin; Douglas Cook, University of California, Davis; Michael Wilson, University of Toronto; Federica DiPalma, The Genome Analysis Center (UK); Jay Konieczka, Enevolve; Daniel Marbach, University of Lausanne; Tomas A. Prolla, U. of Wisconsin; Chrisitina Kendzrioski, U. of Wisconsin; Mark Styczynszki, Georgia Tech; Ilan Wapinski, Enevolve; George Davidson, Sandia National Labs; Larry Sklar, U. of New Mexico; Osorio Meirelles, U. of New Mexico; Fariba Assadi Porter, Texas Tech University; Michael Westphall, U. of Wisconsin; Mark Carter, U. of New Mexico; Susan Young, U. of New Mexico; Bruce Edwards, U. of New Mexico; Christopher Bristow, MD Anderson; Tamer Kahveci, U of Florida; Patrick Meyer, Université de Liège; Vince Calhoun, U. of New Mexico; Sergey Plis, U. of New Mexico.

**Graduate and Postdoc Advisors:** Terran Lane (Univ. of New Mexico), Margaret Werner-Washburne (Univ. of New Mexico), Aviv Regev (Broad institute, MIT), Manolis Kellis (MIT, Broad Institute)

**Thesis/Postdoc Advisor:** Steve Lagree (UW Madison), Rob Atlas (UW Madison), Alireza Fotuhi Sihapirani (UW Madison), Dongyoung Cho (UW Madison), Zhen Niu (UW Madison), Christopher Koch (UW Madison), Chang Wang (UW Madison), Sara Knaack (PhD Physics), Deborah Chasman (PhD Computer Science).

**DOUGLAS E. SOLTIS**

Florida Museum of Natural History, Department of Biology, & the Genetics Institute,

University of Florida

dsoltis@ufl.edu

**a. Professional preparation**

College of William and Mary Biology B.S. 1975

Indiana University Biology M.A. 1977

Indiana University Biology Ph.D. 1980

University of British Columbia Postdoctoral 1981

**b. Appointments**

Co-Director Grad Program, Genetics and Genomics, University of Florida, 2016-present

Distinguished Professor, University of Florida, 2008-present

Chair, Department of Botany, University of Florida, 2006-2008

Professor, University of Florida, 2000-present

Professor, Washington State University, 1990-2000

Acting Director of the Ownbey Herbarium, 1990-91

Associate Professor, Washington State University, 1986-1990

Assistant Professor, Washington State University, 1983-1986

Assistant Professor, The University of North Carolina at Greensboro, 1980-1983

Postdoctoral experience, University of British Columbia, Summer 1981

William R. Ogg Fellowship, Indiana University 1979-1980

**c. Publications (10 most related to proposal)**

Soltis, D. E., P. S. Soltis, D. R. Morgan, S. M. Swensen, B. C. Mullin, J. M. Dowd, and P. G. Martin. 1995. Chloroplast gene sequence data suggest a single origin of symbiotic nitrogen fixation in angiosperms. *Proceedings of the National Academy of Sciences, USA* 92: 2647-2651.

Wang, H., M. J. Moore, P. S. Soltis, C. D. Bell, S. Brockington, R. Alexandre, C. C. Davis, M. Latvis, S. R. Manchester, and D. E. Soltis. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc. Nat. Acad. Sciences USA* 106: 3853-3858.

Li, H.L., W. Wang, P. E. Mortimer, R-Q Li, D-Z Li, K. D. Hyde, J-C Xu, D. E. Soltis and Z-D Chen. 2015. Large-scale phylogenetic analyses reveal multiple gains of symbiotic actinorhizal nitrogen fixation in angiosperms associated with climate changes. *Scientific Reports* 5:14023.

Ruhfel, B. R., M. A. Gitzendanner, P. S. Soltis, D. E. Soltis, J. G.Burleigh. 2014. From Algae to Angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biology* 4:23.

Wang, H., M. J. Moore, P. S. Soltis, C. D. Bell, S. Brockington, R. Alexandre, C. C. Davis, M. Latvis, S. R. Manchester, and D. E. Soltis. 2009. Rosids radiation and the rapid rise of angiosperm-dominated forests. *Proceedings of the National Academy of Sciences* 106: 3853-3858*.*

Hinchliff, C.E., S. A. Smith, J. F. Allman, J. G. Burleigh, R. Chaudhary, L. M. Coghill, K. A. Crandall, J. Deng, B. T. Drew, R. Gazisg, K. Gude, D. S. Hibbettg, L. A. Katz, H. D. Laughinghouse, E. J. McTavish, P. E. Midford, C. L. Owen, R. H. Reed, J. A. Rees, D. E. Soltis, T. Williams, and K. A. Cranston. 2015. Synthesis of phylogeny and taxonomy into a Comprehensive tree of life. *Proc. National Acad. Sciences* 112: 12764–12769.

Zanne, A. E. D. C. Tank, W. K. Cornwell, J. M. Eastman, S. A. Smith, R. G. FitzJohn, D. J. McGlinn, B. C. O'Meara, A. T. Moles, P. B. Reich, D. L. Royer, D. E. Soltis, P. F. Stevens, M. Westoby, I. J. Wright, L. Aarssen, R. I. Bertin, A. Calaminus, R. Govaerts, F. Hemmings, M. R. Leishman, J. Oleksyn, P. S. Soltis, N. G. Swenson, L. Warman, and J. M. Beaulieu. 2013. Three keys to the radiation of angiosperms into freezing environments. *Nature* doi:10.1038/nature12872.

AMBORELLA GENOME PROJECT (citation) authors: Chamala, S., A. Chanderbali, J. Der, T. Lan, B. Walts, V. Albert, C. dePamphilis, J. Leebens-Mack, S. Rounsley, S. Schuster, R. Wing, N. Xiao, R. Moore, P. S. Soltis, D. E. Soltis, W. B. Barbazuk, et al.). 2013. The complete nuclear genome of *Amborella trichopoda*: an evolutionary reference genome for the angiosperms. *Science* 342: no 6165.

Soltis, D. E., P. S. Soltis, P.K. Endress, M. W. Chase, W. S. Judd, L. Majure, E. Mavrodiev, S. M. Manchester. 2017. *Phylogeny and evolution of the angiosperms*. New Edition, University of Chicago Press, in press.

**d. Synergistic Activities**

Co-Organizer of PAG polyploidy symposia, 2005-2009

Associate editor for *American Journal of Botany*, *Webbia*, *Phytokeys*, *PLoS Currents Tree of Life*

Development and implementation of GISH/FISH Cytogenetics workshop, Botany 2012, Columbus

Advisory Committee Hardwood Genomics—Plant Genome Grant, 2011-present

Chinese Academy of Sciences—Visiting Professorship for Senior Scientists 2011-2013

**e. Collaborators and other affiliations:** D. Albach, U. Mainz; V. Albert, U. Buffalo; B. Bremer, Royal Swedish Acad. Sci.; K. Bremer, Stockholm U.; M. W. Chase, Royal Botanic Gardens, Kew; J. Chen, U. Texas; P. Crane, Yale; J. I. Davis, Cornell U.; C. dePamphilis, Penn State U.; M. J. Donoghue, Yale; J. J. Doyle, Cornell U.; C. Freeman, U. Kansas; M. Fay, Kew; T. Givnish, U. Wisconsin; K. Hilu, Virginia Polytechnic Inst.; J. Leebens-Mack, U. Georgia; A. Leitch, U. London; I. Leitch, Kew; H. Ma, Penn State U.; P. Manos, Duke U.; S. Mathews, Harvard; D. Nickrent, Southern Illinois U.; R. G. Olmstead, U. Washington; A. Paterson, U. Georgia; Y-L. Qiu, U. Michigan; J. L. Reveal, U. Maryland; P. Rudall, Kew; V. Savolainen, Imperial College, Silwood; P. Schnable, Iowa State U.; P. F. Stevens, U. Missouri-St. Louis; D. W. Stevenson, New York Bot. Garden; K. Sytsma, U. Wisconsin; D. Tank, U. Idaho; J. Wendel, Iowa State U.: K. Wurdack, Smithsonian; L. Zahn, *Science*

**Graduate and Postdoc Advisors:** M.A./Ph.D. Advisor: G. J. Gastony (retired); Postdoctoral Advisor:B. A. Bohm (retired)

**Thesis/Postdoc Advisor:** L. Rieseberg (1990; Univ. British Columbia); Steve Brunsfield (1994; Univ. Idaho); Paul Wolf (1996; Utah State); Mark Mort (1999; Univ. Kansas; Greg Plunkett, (1994; New York Botanical Garden); Q. Xiang (1995; North Carolina State University); Linda Vook (1998; Washington State Univ); Christine Edwards (2005; Missouri Botanical Garden); Brockington (2009; Cambridge); J. Clayton (2009; U.K.); J. Koh (2010; U. Florida); L. Majure (2012); A. Payton (2012); M. Latvis (2013); R. Hodel (2011-); X. Liu (2011-); C. Visger (2017; Sacramento State); H.-R. Kates (2011-); I. Molgo (2011-); B. Marchant (2013-). **Postdocs:** D. Morgan (1987-1990; W. Washington); E. Conti (1990-1992; Univ. Zurich); J. Tate (2000-2005; Massey Univ, NZ); V. Symmonds (2002-2005, Massey Univ, NZ); E. Mavrodiev (2002-2011; U. Florida); M. Moore (2003-2006; Oberlin); R. Buggs (2007-2010; (QM U. London); L. Zhang (2008-2009; Shanxi Normal U.); Q. Xu (2009-2010; Chinese Acad. Sci.); M. Chester (2010-14; Kew Gardens); I. Jordon-Thaden (2010-2012; Bucknell U.); B. Ruhfel (2011-2012; E. Kentucky U.); S. Pissis, Oxford (2011-2013); B. Drew (2012-14; U. Nebraska-Kearney); A. Chanderbali (2003-; Univ. Florida); J. Koh (2011-; University of Florida).

**PAMELA S. SOLTIS**

Florida Museum of Natural History and the Genetics Institute

University of Florida

psoltis@flmnh.ufl.edu

**a. Professional preparation**

B.A. (Biology), Central College, Pella, IA, 1980, Summa Cum Laude

M. Phil. (Botany), University of Kansas, 1984, with Honors

Ph.D. (Botany), University of Kansas, 1986

**b. Appointments**

Curator, Florida Museum of Natural History, and Professor, Genetics Institute, University of Florida, October 2000 – present

Distinguished Professor, University of Florida, 2007 – present

Director, UF Biodiversity Institute, 2016 – present

Fulbright Distinguished Scholar, Royal Botanic Gardens, Kew, England, and Imperial College, Silwood Park, England, 2000 – 2001

Assistant, Associate, and Full Professor, Department of Botany, Washington State University, 1986-2000

Mellon Senior Fellow, Smithsonian Institution, 1994-95

**c. Publications (10 most related to proposal)**

Sun, M., D. E. Soltis, P. S. Soltis, G. J. Burleigh, and Z. Chen. 2015. Deep phylogenetic incongruence in the angiosperm clade Rosidae. *Molecular Phylogenetics and Evolution* 83: 156–166.

Tank, D. C., J. M. Eastman, M. W. Pennell, P. S. Soltis, D. E. Soltis, C E. Hinchliff, J. W. Brown, and L. J. Harmon. 2015. Progressive radiations and the pulse of angiosperm diversification. *New Phytologist* 207: 454-467.

Zanne, A. E. D. C. Tank, W. K. Cornwell, J. M. Eastman, S. A. Smith, R. G. FitzJohn, D. J. McGlinn, B. C. O'Meara, A. T. Moles, P. B. Reich, D. L. Royer, D. E. Soltis, P. F. Stevens, M. Westoby, I. J. Wright, L. Aarssen, R. I. Bertin, A. Calaminus, R. Govaerts, F. Hemmings, M. R. Leishman, J. Oleksyn, P. S. Soltis, N. G. Swenson, L. Warman, and J. M. Beaulieu. 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature* doi:10.1038/nature12872.

Soltis,P. S. and D. E. Soltis. 2013. Angiosperm phylogeny: A framework for studies of genome evolution. In I. Leitch, J. Dolezel, J. Greilhuber (eds.), *Diversity of Plant Genomes*. Springer, Vienna.

Stull, G. W., M. J. Moore, V. S. Mandala, N. A. Douglas, H.-R. Kates, X. Qi, S. F. Brockington, P. S. Soltis, D. E. Soltis, and M. A. Gitzendanner. 2013. A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Scienc*es 1: 1200497.

Soltis, P. S. and D. E. Soltis (eds.). 2012. *Polyploidy and Genome Evolution*. Springer: Heidelberg.

Soltis, D. E., S. A. Smith, N. Cellinese, K. J. Wurdack, D. C. Tank, S. F. Brockington, N. F. Refulio-Rodriguez, J. B. Walker, M. J. Moore, B. S. Carlsward, C. D. Bell, M. Latvis, S. Crawley, C. Black, D. Diouf, Z. Xi, C. A. Rushworth, M. A. Gitzendanner, K. J. Sytsma, Y.-L. Qiu, K. W. Hilu, C. C. Davis, M. J. Sanderson, R. S. Beaman, R. G. Olmstead, W. S. Judd, M. J. Donoghue, and P. S. Soltis. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98: 704-730.

Moore, M. J., P. S. Soltis, C. D. Bell, J. G. Burleigh, and D. E. Soltis. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences, USA* 107: 4623-4628.

Wang, H., M. J. Moore, P. S. Soltis, C. D. Bell, S. Brockington, R. Alexandre, C. C. Davis, M. Latvis, S. R. Manchester, and D. E. Soltis. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings National Academy of Sciences, USA* 106: 3853-3858.

Soltis, D. E., P. S. Soltis, D. R. Morgan, S. M. Swensen, B. C. Mullin, J. M. Dowd, and P. G. Martin. 1995. Chloroplast gene sequence data suggest a single origin of symbiotic nitrogen fixation in angiosperms. *Proceedings of the National Academy of Sciences, USA* 92: 2647-2651.

**d. Synergistic Activities**

*iDigBio, Director for Research:* Developed scientific requirements for iDigBio portal

Leading development of US Virtual Herbarium portal

Developing use cases for specimen data in research & education

Organizer/Co-organizer of research-related workshops and symposia, including Phenology Workshop, March 2016

Organizer/Co-organizer of training workshops on data analysis, 2014-present

*Society Editorial and Service:* Strategic Planning Committee, Botanical Society of America, 2014

Merit Awards Committee, Botanical Society of America, 2013-15

Nominating Committee, Society of Systematic Biologists, 2009-11

Consulting Editor, *The Plant Cell*, 2015 -

Associate Editor, *Applications in Plant Sciences*, 2013-

Associate Editor, *PhytoKeys*, 2010-

Editor-in-Chief, *PLoS Currents Tree of Life*, 2010-

*NSF Workshops:* The Future of Evolution, Norman, OK, 2011

DNA Banking, Missouri Botanical Garden, St. Louis, 2013

CollectionsWeb, Washington, DC, 2013

Representative from iDigBio to Assembling, Visualizing, and Analyzing the Tree of Life PIs Meeting, Arlington, VA, 2013

US-China Biodiversity Meeting, Raleigh, NC, 2015; Hangzhou, 2016

Dimensions of Biodiversity PIs Meeting, Arlington, VA, 2016

**e. Collaborators and other affiliations:** V. Albert (U. Oslo), N. Altman (Penn State U.), T. Borsch (U. Bonn), P. Cantino (Ohio U.), J. Carlson (Penn State U.), M. Chase (Royal Bot. Gard., Kew), J. Chen (U. Texas), C. Davis (Harvard), C. dePamphilis (Penn State U.), M. Donoghue (Yale), J. Doyle (Cornell), S. Farris (Stockholm U.), M. Frohlich (Natural History Museum, London), K. Hilu (Virginia Tech U.), R. Huck (U. Florida), L. Hufford (Washington State U.), H. Kong (Chinese Acad. Sci.), A. Kovarik (Czech Acad. Sci.), J. Leebens-Mack (U. Georgia), A. Leitch (U. London), I. Leitch (Royal Bot. Gard., Kew), H. Ma (Penn State U.), R. Matyasek (Czech Acad. Sci.), R. Olmstead (U. Washington), Y.-L. Qiu (U. Michigan), M. Sanderson (U. Arizona), V. Savolainen (Imperial College, Silwood), S. Schlarbaum (U. of Tennessee), K. Sytsma (U. Wisconsin), G. Theissen (U. Jena), L. Zahn (Science), G. Riccardi (Florida St. U.)

**Graduate and Postdoc Advisors:** W. L. Bloom (retired)

**Thesis/Postdoc Advisor:** Students (24 PhD, 4 MS): S. Brockington (PhD 2009; Cambridge U.); R. Vergara (PhD 2011); C. Germain-Aubrey (PhD 2012; U. Florida); N. Miles (PhD 2013); C. Segovia (PhD 2014); G. Godden (PhD 2014); M. Heaney (2008-); N. Garcia (2009-); G. Stull (2010-); J. Landis (2011-); I. Molgo (2011-); H.R. Kates (2011-); R. Moraski (2012-); B. Marchant (2013-); L. Gonzalez (2014-). Post-docs (24): M. Moore (Oberlin); S. Jian (Chinese Acad. Sci.); H. Wang (Chinese Acad. Sci.); A. Doust (Oklahoma State U.); L. Zhang (Shanxi Normal U.); R. Buggs (QM U. London); E. Mavrodiev (U. Florida); I. Jordon-Thaden (Bucknell U.); A. Veruska (Embrapa, Brazil); B. Ruhfel (Eastern Kentucky U.); S. Pissis (Kings, U. London); M. Chester (Kew); A. Chanderbali; C. Germain-Aubrey

**RYAN ANDREW FOLK (SENIOR PERSONNEL)**

Florida Museum of Natural History

University of Florida

ryanfolk@ufl.edu

**a. Professional preparation**

University of Akron Biology, minor in Chemistry B.S. *summa cum laude* 2010

The Ohio State University Evolution, Ecology, and Organismal Biology Ph.D. 2015

University of Florida/FLMNH NSF PRFB postdoc (collections) 2015-2017

**b. Appointments**

NSF Postdoctoral Fellow [*postdoctoral fellowship focused on transformative uses of biological museum collections, NSF PRFB program, Collections competitive area*], Florida Museum of Natural History, University of Florida, September 2015 – present.

Graduate Research Assistant [*specimen informatics*], Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, May 2015 – August 2015.

Graduate Candidate, Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, December 2012 – August 2015.

Graduate Teaching Assistant, Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, October 2012 – April 2014.

Graduate Fellow [*University-level graduate fellowship*], Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, September 2010 – August 2012; May 2014 – April 2015.

**c. Publications**

Folk, R.A., J.R. Mandel, and J.V. Freudenstein. 2016. Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. *Systematic Biology* (advance access).

Freudenstein, J.V, M.B. Broe, R.A. Folk, B.T. Sinn. 2016. Biodiversity and the species concept – Lineages are not enough. *Systematic Biology* (advance access)*.*

Folk, R.A., J.R. Mandel, and J.V. Freudenstein. 2015. A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: A phylogenomic example from *Heucher*a (Saxifragaceae). *Applications in Plant Sciences* 3(8): 1500039.

Folk, R.A. and J.V. Freudenstein. (2015) "Sky islands" in the eastern U.S.A.? – Strong phylogenetic structure in the *Heuchera parviflora* group (Saxifragaceae). *Taxon* 64(2): 254–271.

Folk R.A. and P.J. Alexander. (2015) Two new species, *Heuchera soltisii* and *H. adulterina*, with further taxonomic and phylogenetic notes for the western group of *Heuchera* section *Heuchera* (Saxifragaceae). *Systematic Botany* 40(2):489-500.

Folk, R.A. and J.V. Freudenstein. 2014. Revision of *Heuchera* section *Rhodoheuchera* subsections *Hemsleyanae* and *Rosendahliae* subsectio nova (Saxifragaceae). *Systematic Botany* 39(3): 850-874.

Folk, R.A. and J.V. Freudenstein. 2014. Phylogenetic relationships in the genus *Heuchera* L. (Saxifragaceae) on the basis of nuclear loci. *American Journal of Botany* 101(9): 1532-1550.

Folk, R.A. (2013) *Heuchera lakelae* (Saxifragaceae), a new species from the Sierra La Marta and Sierra Coahuilón, Coahuila and Nuevo León, Mexico. *Phytotaxa* 124(1): 37-42.

**d. Synergistic Activities**

*Training -* Mentoring of seven undergraduate researchers, including a summer REU; mentoring of a high school student under UF’s summer SSTP program

*Workshops –* Organizing volunteer and presenter, iDigBio workshop at the 2016 Botanical Society of America meeting

*Reviewing –* Peer reviews for *Molecular Ecology*, *Evolution, American Journal of Botany, Taxon, PLoS ONE, Botanical Letters, Phytotaxa*; external reviewer for NSF OPUS

*Outreach –* Organizing volunteer for the OSU Museum of Biological Diversity Open House (an annual public outreach event), 2011-2015, including a hallway display of my botanical illustration work; public outreach lecture on my dissertation research on *Heuchera* (2014, Ohio Botany Research Symposium); several pop-up museum public events with FLMNH

**e. Collaborators and other affiliations:** P.J. Alexander (Bureau of Land Management), J.M. Allen (University of Florida), M. Broe (The Ohio State University), N. Cellinese (University of Florida), J. de Vos (Kew Gardens), W. Eiserhardt (Kew Gardens), A. Floden (University of Tennessee Knoxville), N. García (Universidad de Chile), T. Heath (Iowa State University), S. Manchester (University of Florida), J.R. Mandel (University of Memphis), A. Moerland (Kew Gardens), M.E. Mort (University of Kansas), J. Oliverio (University of Florida), B.T. Sinn (The Ohio State University), R.L. Stubbs (University of Florida), B. Stucky (University of Florida), M. Sun (University of Florida), J. Thompson (University of California Santa Cruz), C.J. Visger (University of Florida), C. Xiang (Kunming Institute of Botany).

**Graduate and Postdoc Advisors:** J.V. Freudenstein (The Ohio State University), D.E. Soltis (University of Florida), P.S. Soltis (University of Florida), R.P. Guralnick (University of Florida).

**Thesis/Postdoc Advisor:** N/A

**APPENDIX 2: CURRENT AND PENDING SUPPORT**

|  |
| --- |
| **CURRENT & PENDING SUPPORT** |
|  |
| **Name: JEAN-MICHEL M ANÉ** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **NAME**  **(List/PD #1 first)** | **SUPPORTING AGENCY AND AGENCY ACTIVE AWARD/PENDING PROPOSAL NUMBER** | **TOTAL $ AMOUNT** | **EFFECTIVE AND EXPIRATION DATES** | **% OF TIME COMMITTED** | **TITLE OF PROJECT** |
| Sussman, Michel; Ané, Jean-Michel M  Ané, Jean-Michel M  Peters, John; Ané, Jean-Michel M  Ané, Jean-Michel M  Coruzzi, Gloria; Ané, Jean-Michel M  Ané, Jean-Michel M  Prasad, Gvaneshwar; Ané, Jean-Michel M | Active:  NSF - PGRP  USDA  Hatch #WIS01695  NSF - #1331098  USDA  Department of Energy – DE-SC0014377  USDA-AFRI #2015-670-22899  NSF IOS #1256664 | $3,868,166  $180,681  $3,953,828  $1,986,681  $1,299,228  $499,974  $599,615 | 07/01/2016-06/30/2019  10/01/2013-09/30/2018  09/01/2013-08/31/2018  10/01/2013-09/30/2018  09/01/2015- 08/31/2020  01/01/2015- 12/31/2017  05/01/13- 04/30/2017 | 8.33%  4.16 %  8.33%  4.16%  4.16%  8.33%  8.33% | RESEARCH PGR: An interdisciplinary approach to deciphering molecular  signaling pathways controlling plant-symbiont associations in legumes and cereals  Poplar as a model to dissect endo- and ectomycorrhizal symbioses in woody perennials  Engineering Synthetic Symbioses Between Plant and Bacteria to Deliver Nitrogen to Crops  Colonization of legume sprouts by *Salmonella enterica*  EVONET: A Phylogenomic and Systems Biology Approach to Identify Genes Underlying Plant Survival in Marginal, Low-N Soils  Regulation of nuclear cation channels controlling the establishment of plant- microbe symbioses  Understanding and utilizing a unique association between rice and Rhizobium |
| Kirst, Matias; Ané, Jean-Michel; Guralnic, Robert; Roy, Sushmita; Soltis, Douglas; Soltis, Pamela  Prasad, Gvaneshwar; Ané, Jean-Michel M | Pending:  DOE  USDA | $8,657,898  $129,259 | 9/20/17-9/20/22  10/01/2017-09/30/2021 | 8.33%  8.33% | Phylogenomic discovery and engineering of nitrogen fixation into the bioenergy woody crop poplar (**THIS PROPOSAL**)  Collaborative proposal: Role of HMG-CoA Reductases in Microbial Symbioses of Legumes and Cereals |

|  |
| --- |
| **CURRENT & PENDING SUPPORT** |
|  |
| **Name: ROBERT GURALNICK** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **NAME**  **(List/PD #1 first)** | **SUPPORTING AGENCY AND AGENCY ACTIVE AWARD/PENDING PROPOSAL NUMBER** | **TOTAL $ AMOUNT** | **EFFECTIVE AND EXPIRATION DATES** | **% OF TIME COMMITTED** | **TITLE OF PROJECT** |
| Guralnick, Robert  Guralnick, Robert  Guralnick, Robert  Guralnick, Robert  Guralnick, Robert  Guralnick, Robert  Guralnick, Robert  Guralnick, Robert | Active:  USGS  Bureau of Land Management  NSF GoLife  NSF ABI  Advancing Digitization of Biodiversity Collections/NSF  NSF GoLife  NSF ABI  NSF Dimensions of Biodiversity | $30,000  $30,000  $2,500,000  $594,428  $65,000  $2,500,000  $909,541  $211,066 | 08/01/2016-07/31/2017  0701/2016-06/30/2017  08/01/2015-07/31/2019  03/01/2015-02/29/2018  06/01/2014- 05/31/2018  10/01/2014- 09/30/2018  08/01/2013- 07/31/2017  09/01/2012- 08/31/2017 | 0%  0 %  0%  4.16%  0%  3.33%  5%  1.25% (2011-2016) | Developing the Plant Phenology Ontology  Monadenia terrestrial snail genetics and taxonomy  Collaborative Research: ButterflyNet — An integrative framework for comparative biology  Collaborative Research: ABI DEVELOPMENT: Notes from Nature: Advancing a Next Generation Citizen Science Platform For Biocollection Transcription  ADBC Proposal: Digitization TCN: Collaborative Research: The Key to the Cabinets: Building and sustaining a research database for a global biodiversity hotspot  Collaborative Research: VertLife Terrestrial: A complete, global assembly of phylogenetic, trait, spatial and environment characteristics for a model clade  Advancing Map of Life's Impact and Capacity for Sharing, Integrating, and Using Global Spatial Biodiversity Knowledge  Dimensions: Collaborative Research: Assembly and evolution of the  Amazonian biota and its environment: an integrated approach |
| Kirst, Matias; Ané, Jean-Michel; Guralnick, Robert; Roy, Sushmita; Soltis, Douglas; Soltis, Pamela  Guralnick,  Robert  Guralnick,  Robert | Pending:  DOE  NASA GEO  NASA AIST | $8,657,898  $59,948  $56,672 | 09/1/2017-08/31/2022  09/01/2017- 08/31/2019  07/01/2017- 06/30/2019 | 4.17%  5%  5% | Phylogenomic discovery and engineering of nitrogen fixation into the bioenergy woody crop poplar (**THIS PROPOSAL**)  Activities to advance, build, and deliver remote-sensing supported species distribution and species abundance EBVs  Software workflows and tools for integrating remote sensing and organismal occurrence data streams to assess and monitor biodiversity change |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Guralnick,  Robert  Guralnick,  Robert  Guralnick,  Robert  Guralnick,  Robert  Guralnick,  Robert  Guralnick,  Robert | Pending:  NSF National Research Traineeship program  NSF Macrosystems  NSF Macrosystems  NSF ADBC  NSF ABI  NSF | $3,000,000  $291,282  $2,500,000  $331,086  $428,462  $158,783 | 07/01/2017-06/30/2022  05/01/2017- 04/30/20121  04/15/2017- 04/14/2021  07/01/2017- 06/30/2021  05/01/2017- 04/30/2020  05/01/2017-04/30/2020 | 7%  2.5%  1%  1%  1.25%  0% | NRT: Integrated Research and Data Sciences for Biodiversity (IREADBIO)  Collaborative Proposal: MSB-ENSA: Enhancing NEON with novel biodiversity change models, products and tools  Collaborative Proposal: MSB-FRA: Causes, consequences, and cross-scale linkages of climate-driven phenological mismatch across three trophic levels  Collaborative Research: Digitization TCN: Carabid Beetle Collective: digitizing a dominant insect group to enable research on diversification and ecological dynamics.  Collaborative Research: ABI Development: MOL-SDM - Infrastructure for next-generation model-based development of species distribution knowledge  Collaborative Research: Documenting Turkey Husbandry and Domestication in Ancient Mesoamerica |

|  |
| --- |
| **CURRENT & PENDING SUPPORT** |
|  |
| **Name: MATIAS KIRST** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **NAME**  **(List/PD #1 first)** | **SUPPORTING AGENCY AND AGENCY ACTIVE AWARD/PENDING PROPOSAL NUMBER** | **TOTAL $ AMOUNT** | **EFFECTIVE AND EXPIRATION DATES** | **% OF TIME COMMITTED** | **TITLE OF PROJECT** |
| Kirst, Matias; Munoz, Patricio  Kirst, Matias; Munoz, Patricio; Peter, Gary  Kirst, Matias; Resende, Marcio; de los Campos, Gustavo; Barbazuk, William | Active:  USDA  (2013-67013-21159)  USDA  (2013-67009-21200)  National Science Foundation (IOS 1444543) | $450,000  $1,000,000  $1,956,424 | 08/01/2013-07/31/2017  09/01/2013-08/30/2017  03/15/2015-02/28/2019 | 8.33%  8.33%  8.33% (yr 1-2), 16.66% (yr 3-4) | Accelerated breeding by improved accuracy and mate-allocation using genome-wide selection  Accelerated development of optimal feedstock for bioenergy using genome-wide selection  Genome and transcriptome based prediction, and regulator inference, of molecular and whole-plant phenotypes |
| Kirst, Matias; Ané, Jean-Michel; Guralnick, Robert; Roy, Sushmita; Soltis, Douglas; Soltis, Pamela  Kirst, Matias; Resende, Marcio; de los Campos, Gustavo; Kirst, Mariana  Kirst, Matias (co-PI with 50+ co-PIs)  Kirst, Matias; Meilan, Richard; Smith, Jason  Brunner, Amy; Tsai, C.-J.; Holiday, Jason; Schmitz, Robert; Meilan, Richard; Kirst, Matias  Assmann, Sarah; Chen, Sixue; Kirst, Matias | Pending:  DOE  National Science Foundation  USDA  DOE/USDA  DOE  DOE | $8,657,898  $851,391  $15,000,000  $1,200,000  ~$5,000,000  ˜$5,000,000 | 9/20/17-9/20/22  1/1/2017-12/31/2019  1/1/17-12/31/21  9/1/17-8/31/20  9/01/17-8/31/22  9/01/17-8/31/22 | 16.66%  8.33%  8.33%  8.33%  8.33%  8.33% | Phylogenomic discovery and engineering of nitrogen fixation into the bioenergy woody crop poplar (**THIS PROPOSAL**)  The genetic regulation of the endophytic microbiome and its contribution to plant productivity  PopuluSolv: A Sustainable Supply Chain for the Southeast’s Bioeconomy  Genome-editing of elite *Populus deltoides* (eastern cottonwood) for disease resistance  POLyGENE: POpuLus Growth and Environmental Network Engineering  Engineering C3 to CAM Stomatal Transition into Poplar for Enhanced Biofuel Production |

|  |
| --- |
| **CURRENT & PENDING SUPPORT** |
|  |
| **Name: SUSHMITA ROY** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **NAME**  **(List/PD #1 first)** | **SUPPORTING AGENCY AND AGENCY ACTIVE AWARD/PENDING PROPOSAL NUMBER** | **TOTAL $ AMOUNT** | **EFFECTIVE AND EXPIRATION DATES** | **% OF TIME COMMITTED** | **TITLE OF PROJECT** |
| Roy, S.  Summan, M.; Roy, S.  Crave, R; Roy, S.  Murphy, W.; Roy, S.  Sussman, M.; Roy, S.  Roy, S.  Roy, S. | Active:  NSF-DBI  NSF-MCB  NIH  EPA  NSF  USDA  James S. McDonnel Foundation  NIH | $323,415  $40,758  $337,500  $194,272  $190,632  $351,590  $1,010,000 | 08/01/2014-07/31/2019  07/01/2014-06/30/2017  09/29/2014-04/30/2019  12/01/2014-11/30/2018  09/01/2015- 08/31/2020  08/01/2016- 07/31/2022  09/12/2016- 06/30/2021 | 8.33%  4.16%  16.66%  8.33%  8.33%  4.16%  30% | Career: Comparative Network Biology to Study the Evolution of Regulatory Networks  An Isotope-Assisted Quantitative Phosphoproteomic Analysis of Signaling Pathways Initiated at the Plasma Membrane of Arabidopsis thaliana  The Center for Predictive Computational Phenotyping  Human Models for Analysis of Pathways (H-MAPs)  An interdisciplinary approach to deciphering molecular signaling pathways controlling plant-symbiont associations in legumes and cereals  Learning network-based predictive models of complex phenotypes  Computational inference of cell-type specific predictive regulatory network models |
| Kirst, Matias, Ané, Jean-Michel; Guralnic, Robert; Roy, Sushmita; Soltis, Douglas; Soltis, Pamela | Pending:  DOE | $8,657,898 | 9/20/17-9/20/22 | 8.33% | Phylogenomic discovery and engineering of nitrogen fixation into the bioenergy woody crop poplar (**THIS PROPOSAL**) |

|  |
| --- |
| **CURRENT & PENDING SUPPORT** |
|  |
| **Name: DOUGLAS SOLTIS** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **NAME**  **(List/PD #1 first)** | **SUPPORTING AGENCY AND AGENCY ACTIVE AWARD/PENDING PROPOSAL NUMBER** | **TOTAL $ AMOUNT** | **EFFECTIVE AND EXPIRATION DATES** | **% OF TIME COMMITTED** | **TITLE OF PROJECT** |
| Soltis, D., Visger, C.  Soltis, D.,  Hodel, R.  Buell, R; Soltis, D, Soltis P., O’Conner, S., N. Dudareva  Soltis, D., Beach, J., Smith, S., Soltis, P.  Soltis, P., Soltis, D., Lichtstein, J., Mack, M., Triplett, E., Gilbert, J., Bohlman, S., Xoiang, J., Shi, W | Active:  NSF  DDIG  NSF  DDIG  National Science Foundation  Plant Genome  National Science Foundation  ABI  National Science Foundation  Dimenions | $19,353  $19,455  $865,668  $834,828  $1,199,043 | 6/1/15 - 5/31/17  06/01/15 - 05/31/17  1/1/15-12/31/19  1/1/15-12/31/19  1/1/15-12/31/19 | 2%  2%  8%  10%  8% | Dissertation Research: The evolutionary significance of autopolyploidy in *Tolmiea* (Saxifragaceae)  Dissertation Research: Comparative phylogeography of three co-distributed Neotropical mangrove species  PGRP: Evolution of specialized metabolite biosynthetic pathways in the Lamiaceae: Sources of chemical diversity for molecules essential for human use and plant defense  Collaborative Research: ABI Innovation: Connecting resources to enable large-scale biodiversity analyses  Dimensions US-China: Collaborative Research: How historical constraints, local adaptation, and species interactions shape biodiversity across an ancient floristic disjunction |
| Kirst, Matias, Ané, Jean-Michel; Guralnick, Robert; Roy Susmnita; Soltis, Douglas; Soltis, Pamela  Soltis, Douglas;  Liu, Xiaoxian | Pending:  DOE  National Science Foundation  DDIG | $8,657,898  $19,825 | 9/1/17-8/31/22  05/01/2017-04/31/2019 | 8.33%  2% | Phylogenomic discovery and engineering of nitrogen fixation into the bioenergy woody crop poplar (**THIS PROPOSAL**)  Doctoral Dissertation Research: Evolutionary impact of genome duplication on alternative splicing: Genome-wide assessment in a polyploid plant (*Tragopogon*) |

|  |
| --- |
| **CURRENT & PENDING SUPPORT** |
|  |
| **Name: PAMELA SOLTIS** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **NAME**  **(List/PD #1 first)** | **SUPPORTING AGENCY AND AGENCY ACTIVE AWARD/PENDING PROPOSAL NUMBER** | **TOTAL $ AMOUNT** | **EFFECTIVE AND EXPIRATION DATES** | **% OF TIME COMMITTED** | **TITLE OF PROJECT** |
| Page, Larry;  Soltis, Pamela;  MacFadden, Bruce;  Fortes, Jose;  Riccardi, Greg  Soltis, Douglas;  Soltis, Pamela;  Fortes, Jose  Buell, Robin (MSU);  Soltis, Douglas (PI @ UF);  Soltis, Pamela  Soltis, Pamela (lead PI, multi-institutional);  Soltis, Douglas;  Lichstein, Jeremy;  Mack, Michelle  Kawahara, Akito;  Willmott, Keith;  Miller, Jacqueline;  Soltis, Pamela  dePamphilis, Claude (PSU);  Soltis, Pamela (PI @ UF);  Soltis, Douglas;  Barbazuk, Brad  Page, Larry;  Soltis, Pamela;  MacFadden, Bruce;  Fortes, Jose;  Riccardi, Greg | Active:  NSF  NSF  NSF  NSF  NSF  NSF  NSF | $15,486,747  $834,828  $810,477 to UF;  ~$4M total  $1,199,043  $497,364  $1,704,138 to UF; ~$7.5M total  $11,846,048 | 09/01/2016-08/31/2021  05/1/15-04/30/18  04/1/15-03/31/20  01/1/15-12/31/19  07/1/14-06/30/17  07/1/10-6/30/18  07/1/11-06/30/17 | 8.33%  0% paid time  4.16%  0% paid time  0% paid time  4.16%  4.16% | Digitization: iDigBio: Integrated Digitized Biocollections Phase 2  Collaborative Research: ABI Innovation: Connecting resources to enable large-scale biodiversity analyses  PGRP: Evolution of specialized metabolite biosynthetic pathways in the Lamiaceae: Sources of chemical diversity for molecules essential for human use and plant defense  Dimensions US-China: Collaborative Research: How historical constraints, local adaptation, and species interactions shape biodiversity across an ancient floristic disjunction  CSBR: Natural History Collections: Building a central database and curation improvements for The McGuire Center for Lepidoptera at the Florida Museum of Natural History  TRPGR: The *Amborella* Genome: An Evolutionary Reference for Plant Biology  Digitization HUB: A Collections Digitization Framework for the 21st Century |
| Kirst, Matias, Ané, Jean-Michel; Guralnick, Robert; Roy, Sushmita; Soltis, Douglas; Soltis, Pamela  Guralnick, Robert;  Soltis, Pamela  Soltis, Douglas;  Liu, Xiaoxian;  Soltis, Pamela  Fortes, Jose;  Soltis, Pamela;  Barbazuk, Brad  Michailidis, George;  Soltis, Pamela;  Guralnick, Robert | Pending:  DOE  NSF  NSF  NSF  NSF | $8,657,898  $463,241  $19,826  $1,200,000  $3,000,000 | 9/1/17-8/31/22  05/01/17 - 04/30/18  09/1/2017-08/31/2020  09/1/2017-08/31/2022 | 4.17%  0% paid time  0% paid time  0% paid time  4.16% | Phylogenomic discovery and engineering of nitrogen fixation into the bioenergy woody crop poplar (**THIS PROPOSAL**)  Collaborative Proposal: ABI Development: CreatureFeatures: A semantic toolkit for biodiversity trait data  DISSERTATION RESEARCH: Evolutionary impact of genome duplication on alternative splicing: Genome-wide assessment in a polyploid plant (*Tragopogon*)  Acquisition of Software-defined Instrument for In-Silico Observation and Analysis of Biodiversity  Data Sciences and Informatics Training for Integrative Biodiversity Research |

|  |
| --- |
| **CURRENT & PENDING SUPPORT** |
|  |
| **Name: RYAN FOLK (SENIOR PERSONNEL)** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **NAME**  **(List/PD #1 first)** | **SUPPORTING AGENCY AND AGENCY ACTIVE AWARD/PENDING PROPOSAL NUMBER** | **TOTAL $ AMOUNT** | **EFFECTIVE AND EXPIRATION DATES** | **% OF TIME COMMITTED** | **TITLE OF PROJECT** |
| Folk, R. | Active:  NSF | $138,000 | 09/01/15-8/31/17 | 100% | NSF Postdoctoral Fellowship in Biology |

**APPENDIX 3: BIBLIOGRAPHY & REFERENCES CITED**

**Allen JM, Huang DI, Cronk QC, Johnson KP**. **2015**. aTRAM - automated target restricted assembly method: a fast method for assembling loci across divergent taxa from next-generation sequencing data. *BMC Bioinformatics* **16**: 98.

**Allen JM, Boyd B, Nguyen N-P, Vachaspati P, Warnow T, Huang DI, Grady PGS, Bell KC, Cronk QCB, Mugisha L, Pittendrigh BR, Leonardi MS, Reed DL, Johnson KP.** 2017. Phylogenomics from whole genome sequences using aTRAM. *Systematic Biology* syw105. http://doi.org/10.1093/sysbio/syw105

**Beaulieu JM, Jhwueng D-C, Boettiger C, O’Meara BC**. **2012**. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution; international journal of organic evolution* **66**: 2369–83.

**Beaulieu JM, O’Meara BC, Donoghue MJ**. **2013**. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. *Systematic Biology* **62**: 725–37.

**Boisson-Dernier A, Chabaud M, Garcia F, Bécard G, Rosenberg C, Barker DG**. **2001**. Agrobacterium rhizogenes-transformed roots of *Medicago truncatula* for the study of nitrogen-fixing and endomycorrhizal symbiotic associations. *Molecular plant-microbe interactions : MPMI* **14**: 695–700.

**Bradshaw HD, Ceulemans R, Davis J, Stettler R**. **2000**. Emerging model systems in plant piology: poplar (*Populus*) as a model forest tree. *Journal of PlantGrowth Regulation* **19**: 306–313.

**Brunner AM, Busov VB, Strauss SH**. **2004**. Poplar genome sequence: functional genomics in an ecologically dominant plant species. *Trends in plant science* **9**: 49–56.

**Bryan JA, Berlyn GP, Gordon JC**. **1996**. Toward a new concept of the evolution of symbiotic nitrogen fixation in the Leguminosae. *Plant and Soil* **186**: 151–159.

**Cai Y, Chen L, Liu X, Sun S, Wu C, Jiang B, Han T, Hou W**. **2015**. CRISPR/Cas9-mediated genome editing in soybean hairy roots. *PloS one* **10**: e0136064.

**Chamala S, García N, Godden GT, Krishnakumar V, Jordon-Thaden IE, Smet R De, Barbazuk WB, Soltis DE, Soltis PS**. **2015**. MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences* **3**: 1400115.

**Delaux P-M, Radhakrishnan G, Oldroyd G**. **2015**. Tracing the evolutionary path to nitrogen-fixing crops. *Current Opinion in Plant Biology* **26**: 95–99.

**Delaux P-M, Séjalon-Delmas N, Bécard G, Ané J-M**. **2013**. Evolution of the plant–microbe symbiotic ‘toolkit’. *Trends in Plant Science* **18**: 298–304.

**Delaux P-M, Varala K, Edger PP, Coruzzi GM, Pires JC, Ané J-M**. **2014**. Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. (JM McDowell, Ed.). *PLoS Genetics* **10**: e1004487.

**Doyle JJ**. **2016**. Chasing unicorns: Nodulation origins and the paradox of novelty. *American Journal of Botany* **103**: 1865–1868.

**Fahrenkrog AM, Neves LG, Resende MFR, Vazquez AI, de los Campos G, Dervinis C, Sykes R, Davis M, Davenport R, Barbazuk WB, *et al.*** **2017**. Genome-wide association study reveals putative regulators of bioenergy traits in *Populus deltoides*. *New Phytologist* **213**: 799–811.

**Favre P, Bapaume L, Bossolini E, Delorenzi M, Falquet L, Reinhardt D**. **2014**. A novel bioinformatics pipeline to discover genes related to arbuscular mycorrhizal symbiosis based on their evolutionary conservation pattern among higher plants. *BMC plant biology* **14**: 333.

**Finzi AC, Norby RJ, Calfapietra C, Gallet-Budynek A, Gielen B, Holmes WE, Hoosbeek MR, Iversen CM, Jackson RB, Kubiske ME, *et al.*** **2007**. Increases in nitrogen uptake rather than nitrogen-use efficiency support higher rates of temperate forest productivity under elevated CO2. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 14014–9.

**Francisco PB, Akao S**. **1993**. Autoregulation and nitrate inhibition of nodule formation in soybean cv. Enrei and its nodulation mutants. *Journal of Experimental Botany* **44**: 547–553.

**Gutierrez RA**. **2012**. Systems biology for enhanced plant nitrogen nutrition. *Science* **336**: 1673–1675.

**Hittinger CT, Carroll SB**. **2007**. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**: 677–681.

**Horváth B, Yeun LH, Domonkos A, Halász G, Gobbato E, Ayaydin F, Miró K, Hirsch S, Sun J, Tadege M, *et al.*** **2011**. *Medicago truncatula* IPD3 is a member of the common symbiotic signaling pathway required for rhizobial and mycorrhizal symbioses. *Molecular plant-microbe interactions : MPMI* **24**: 1345–58.

**Jez JM, Lee SG, Sherp AM**. **2016**. The next green movement: plant biology for the environment and sustainability. *Science* **353**: 1241–1244.

**Jordon-Thaden IE, Chanderbali AS, Gitzendanner MA, Soltis DE**. **2015**. Modified CTAB and TRIzol protocols improve RNA extraction from chemically complex Embryophyta. *Applications in Plant Sciences* **3**: 1400105.

**Kevei Z, Lougnon G, Mergaert P, Horváth G V, Kereszt A, Jayaraman D, Zaman N, Marcel F, Regulski K, Kiss GB, *et al.*** **2007**. 3-hydroxy-3-methylglutaryl coenzyme a reductase 1 interacts with NORK and is crucial for nodulation in *Medicago truncatula*. *The Plant cell* **19**: 3974–89.

**King MC, Wilson AC**. **1975**. Evolution at two levels in humans and chimpanzees. *Science (New York, N.Y.)* **188**: 107–16.

**Ladha JK, Tirol-Padre A, Reddy CK, Cassman KG, Verma S, Powlson DS, van Kessel C, de B. Richter D, Chakraborty D, Pathak H**. **2016**. Global nitrogen budgets in cereals: A 50-year assessment for maize, rice, and wheat production systems. *Scientific Reports* **6**: 19355.

**Leple J, Brasileiro A, Michel M, Delmotte F, Jouanin L**. **1992**. Transgenic poplars: expression of chimeric genes using four different constructs. *Plant Cell Reports* **11**: 137–141.

**Li B, Dewey CN**. **2011**. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

**Li H-L, Wang W, Mortimer PE, Li R-Q, Li D-Z, Hyde K-D, Xu J-C, Soltis DE, Chen Z-D.** 2015. Large-scale phylogenetic analyses reveal multiple gains of actinorhizal nitrogen-fixing symbioses in angiosperms associated with climate change. *Scientific Reports.* **5:** 14023.

**Makarov MI, Malysheva TI, Ermak AA, Onipchenko VG, Stepanov AL, Menyailo O V.** **2011**. Symbiotic nitrogen fixation in the alpine community of a lichen heath of the Northwestern Caucasus Region (the Teberda Reserve). *Eurasian Soil Science* **44**: 1381–1388.

**Mandel JR, Dikow RB, Funk VA, Masalia RR, Staton SE, Kozik A, Michelmore RW, Rieseberg LH, Burke JM**. **2014**. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Applications in Plant Sciences* **2**: 1300085.

**Matasci N, Hung L-H, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M, *et al.*** **2014**. Data access for the 1,000 Plants (1KP) project. *GigaScience* **3**: 17.

**Mathelier A, Fornes O, Arenillas DJ, Chen C, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R, *et al.*** **2016**. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **44**: D110–D115.

**Nieminen K, Immanen J, Laxell M, Kauppinen L, Tarkowski P, Dolezal K, Tahtiharju S, Elo A, Decourteix M, Ljung K, *et al.*** **2008**. Cytokinin signaling regulates cambial development in poplar. *Proceedings of the National Academy of Sciences, USA* **105**: 20032–20037.

**Novaes E, Osorio L, Drost DR, Miles BL, Boaventura-Novaes CRD, Benedict C, Dervinis C, Yu Q, Sykes R, Davis M, *et al.*** **2009**. Quantitative genetic analysis of biomass and wood chemistry of *Populus* under different nitrogen levels. *The New Phytologist* **182**: 878–90.

**Oren R, Ellsworth DS, Johnsen KH, Phillips N, Ewers BE, Maier C, Schäfer KVR, McCarthy H, Hendrey G, McNulty SG, *et al.*** **2001**. Soil fertility limits carbon sequestration by forest ecosystems in a CO2-enriched atmosphere. *Nature* **411**: 469–472.

**Patron NJ, Orzaez D, Marillonnet S, Warzecha H, Matthewman C, Youles M, Raitskin O, Leveau A, Farré G, Rogers C, *et al.*** **2015**. Standards for plant synthetic biology: a common syntax for exchange of DNA parts. *The New Phytologist* **208**: 13–9.

**Quesada T, Li Z, Dervinis C, Li Y, Bocock PN, Tuskan GA, Casella G, Davis JM, Kirst M**. **2008**. Comparative analysis of the transcriptomes of *Populus trichocarpa* and *Arabidopsis thaliana* suggests extensive evolution of gene expression regulation in angiosperms. *The New Phytologist* **180**: 408–20.

**Roy S, Wapinski I, Pfiffner J, French C, Socha A, Konieczka J, Habib N, Kellis M, Thompson D, Regev A**. **2013**. Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Research* **23**: 1039–50.

**Ruppel S, Merbach W**. **1997**. Effect of ammonium and nitrate on 15N2-fixation of *Azospirillum* spp. and *Pantoea agglomerans* in association with wheat plants. *Microbiological Research* **152**: 377–383.

**Schauser L, Wieloch W, Stougaard J**. **2005**. Evolution of NIN-like proteins in Arabidopsis, rice, and Lotus japonicus. *Journal of molecular evolution* **60**: 229–37.

**Schlücking K, Edel KH, Köster P, Drerup MM, Eckert C, Steinhorst L, Waadt R, Batistic O, Kudla J**. **2013**. A new β-estradiol-inducible vector set that facilitates easy construction and efficient expression of transgenes reveals CBL3-dependent cytoplasm to tonoplast translocation of CIPK5. *Molecular plant* **6**: 1814–29.

**Soltis DE, Soltis PS, Morgan DR, Swensen SM, Mullin BC, Dowd JM, Martin PG**. **1995**. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. *Proceedings of the National Academy of Sciences of the United States of America* **92**: 2647–51.

**Soudzilovskaia NA, Aksenova AA, Makarov MI, Onipchenko VG, Logvinenko OA, ter Braak CJF, Cornelissen JHC**. **2012**. Legumes affect alpine tundra community composition via multiple biotic interactions. *Ecosphere* **3**: art33.

**Soyano T, Hayashi M**. **2014**. Transcriptional networks leading to symbiotic nodule organogenesis. *Current Opinion in Plant Biology* **20**: 146–154.

**Soyano T, Kouchi H, Hirota A, Hayashi M**. **2013**. Nodule inception directly targets NF-Y subunit genes to regulate essential processes of root nodule development in Lotus japonicus. (JM McDowell, Ed.). *PLoS Genetics* **9**: e1003352.

**Sprent J. 2009.** *Legume Nodulation: A Global Perspective.* New York: Wiley. 183 pp.

**Stokstad E**. **2016**. The nitrogen fix. *Science (New York, N.Y.)* **353**: 1225–7.

**Sun J, Miller JB, Granqvist E, Wiley-Kalil A, Gobbato E, Maillet F, Cottaz S, Samain E, Venkateshwaran M, Fort S, *et al.*** **2015**. Activation of symbiosis signaling by arbuscular mycorrhizal fungi in legumes and rice. *The Plant Cell* **27**: 823–38.

**Sykes R, Yung M, Novaes E, Kirst M, Peter G, Davis M**. **2009**. High-throughput screening of plant cell-wall composition using pyrolysis molecular beam mass spectroscopy. *Methods in molecular biology (Clifton, N.J.)* **581**: 169–83.

**Taylor G**. **2002**. Populus: Arabidopsis for forestry. Do we need a model tree? *Annals of Botany* **90**: 681–9.

**Thompson DA, Roy S, Chan M, Styczynsky MP, Pfiffner J, French C, Socha A, Thielke A, Napolitano S, Muller P, *et al.*** **2013**. Evolutionary principles of modular gene regulation in yeasts. *eLife* **2**: e00603.

**Tirichine L, Sandal N, Madsen LH, Radutoiu S, Albrektsen AS, Sato S, Asamizu E, Tabata S, Stougaard J**. **2006**. A gain-of-function mutation in a cytokinin receptor triggers spontaneous root nodule organogenesis. *Plant Cell Mol. Plant Microbe Interact. D. P. Lohar et al. Plant J* **18**: 2680–203.

**Tirichine L, Sandal N, Madsen LH, Radutoiu S, Albrektsen AS, Sato S, Asamizu E, Tabata S, Stougaard J**. **2007**. A gain-of-function mutation in a cytokinin receptor triggers spontaneous root nodule organogenesis. *Science (New York, N.Y.)* **315**: 104–7.

**Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, *et al.*** **2006**. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science (New York, N.Y.)* **313**: 1596–604.

**Uslu V V, Grossmann G**. **2016**. The biosensor toolbox for plant developmental biology. *Current opinion in plant biology* **29**: 138–47.

**Venkateshwaran M, Cosme A, Han L, Banba M, Satyshur KA, Schleiff E, Parniske M, Imaizumi-Anraku H, Ané J-M**. **2012**. The recent evolution of a symbiotic ion channel in the legume family altered ion conductance and improved functionality in calcium signaling. *The Plant Cell* **24**: 2528–45.

**Venkateshwaran M, Jayaraman D, Chabaud M, Genre A, Balloon AJ, Maeda J, Forshey K, den Os D, Kwiecien NW, Coon JJ, *et al.*** **2015**. A role for the mevalonate pathway in early plant symbiotic signaling. *Proceedings of the National Academy of Sciences* **112**: 9781–9786.

**Venkateshwaran M, Volkening JD, Sussman MR, Ané J-M**. **2013**. Symbiosis and the social network of higher plants. *Current Opinion in Plant Biology* **16**: 118–127.

**Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, *et al.*** **2014**. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**: 1431–1443.

**Werner GDA, Cornwell WK, Sprent JI, Kattge J, Kiers ET**. **2014**. A single evolutionary innovation drives the deep evolution of symbiotic N2-fixation in angiosperms. *Nature Communications* **5**: 217–248.

**Xiao TT, Schilderink S, Moling S, Deinum EE, Kondorosi E, Franssen H, Kulikova O, Niebel A, Bisseling T**. **2014**. Fate map of *Medicago truncatula* root nodules. *Development (Cambridge, England)* **141**: 3517–28.

**Yoshida K, Ma D, Constabel CP**. **2015**. The MYB182 protein down-regulates proanthocyanidin and anthocyanin biosynthesis in poplar by repressing both structural and regulatory flavonoid genes. *Plant Physiology* **167**: 693–710.

**Zhang J, Yang J-R**. **2015**. Determinants of the rate of protein sequence evolution. *Nature reviews. Genetics* **16**: 409–20.

**APPENDIX 4 - FACILITIES AND OTHER RESOURCES**

**UNIVERSITY OF WISCONSIN – MADISON (Ané and Roy)**

The co-PI Ané laboratory is comprised of approximately 3,300 square feet of wet lab plus 180 square feet of office space for lab staff. An excellent support staff is available to assist faculty and other research personnel with ordering supplies, accounting and budget management, audiovisual assistance, computer support and troubleshooting, travel arrangements, web pages, and instructional materials.

Most of the microscopy studies for this project will be conducted at the Plant Imaging Center (PIC) (<http://www.botany.wisc.edu/pic/>) which provides easy access to a Zeiss 510 Meta confocal laser scanning microscope as well as an Olympus BX60 Epifluorescence Microscope. Microscopy training is also provided through the PIC. Many other facilities are also available on UW-Madison campus giving easy access to state-of-the-art microscopy and computer equipment.

Other facilities include (1) the Biotechnology Center which provides fee-for-service DNA sequencing, oligonucleotide synthesis, electrospray ionization mass spectroscopy, peptide synthesis, and amino acid analysis; (2) the Gene Expression Center which provides a core facility with equipment and expert staff to perform and analyze RNAseq experiments; (3) the Soil Testing Laboratory that conducts elemental analyses using an inductively coupled plasma-mass spectrometer (ICP-MS); (4) the Statistical Support and Analysis Laboratory; and (5) the Biology New Media Center that provides technical and support services to generate and fully render digital video for web streaming.

Also available is the Biotron that is a controlled environment facility for biological research at the University of Wisconsin - Madison. The Biotron provides researchers with the ability to manipulate parameters of the physical environment to simulate natural environments or changes in the environment resulting from human activities. Several greenhouses facilities are also available on campus. Greenhouses are located at Walnut Street and have been completely renovated in 2005.

**UNIVERSITY OF FLORIDA (R. Guralnick, M. Kirst, P. Soltis, D. Soltis)**

**UF Genetics Institute**

The PI Kirst laboratory is located in the University of Florida Genetics Institute, with 2,400 sq ft of laboratory and office space. The UF Genetics Institute has several small conference rooms (4-6 people) as well as larger rooms (20-30) persons and an auditorium (150 persons). The UFGI also has a copy, fax, and all other necessary office support.

Available facilities for plant propagation, growth, and maintenance in the University of Florida Genetics Institute include four walk-in growth chambers and over 2000 sq. ft. of climate-controlled greenhouse space equipped with 30 ebb-and-flow benches that can support up to 140 plants/each. Also, the PIs have access to greenhouse facilities and walk-in growth chambers at the University of Florida’s Genetics Institute and Institute of Food and Agricultural Sciences (IFAS). The Genetics Institute Plant Growth Facility is composed of two parts: 3 greenhouses and 10 Percival growth chambers and growth rooms. The greenhouses occupy 5,000 square feet of space and are located on the roof of the building. The PIs have full access to these growth facilities. The IFAS facilities are available at minimal cost.

Field trial facilities are available at the Institute of Food and Agricultural Sciences (IFAS) Research and Education Center (REC) near Citra, FL, approximately 30 miles south of the Gainesville Campus. The REC has dedicated full-time staff, farm equipment, and irrigation systems to manage the plots.

**Florida Museum of Natural History & Dept. Biol.**

The co-PIs P. Soltis and D. Soltis share two laboratories, in the Florida Museum of Natural History (Dickinson Hall) and the Department of Biology (Bartram Hall). The Department of Biology maintains excellent greenhouse facilities. Two rooms of one greenhouse (over 1000 sq feet) are dedicated for use by the PIs.

**Historical DNA isolation facility**

Co-PI Guralnick maintains a dedicated facility at the Florida Museum of Natural History for isolating historical materials, such as herbarium/museum specimens, for phylogenomics projects.

**Genetic Resources Repository**

Established in 2006, the Florida Museum of Natural History's Genetic Resources Repository (GRR) archives nearly 60,000 tissue samples and DNA and RNA preparations from physical specimens in the Museum. Through its ongoing research and conservation projects, the Museum estimates it will add more than 100,000 specimens to the collection in the next 15 to 20 years. The Repository's mission is to guarantee the proper preservation of the Museum's genetic resources and facilitate their use by the global scientific community. The integrity of the specimens is ensured by their cryogenic storage in a nitrogen-cooled freezer with a temperature of -300° Fahrenheit. The information linking the holdings to their sources and other data relevant to their molecular value is managed in a searchable online database (although, at the time of this writing, the database is undergoing an upgrade and migration to Specify and is not currently online). Staffing of the GRR includes a 40%-time Collections Manager, Terry Lott. Samples from this project will be archived in the GRR.

**University of Florida Herbarium (FLAS)**

The Herbarium (N. Williams, Keeper), established in 1891 as the Herbarium of Florida Agricultural College, contain approximately 1/2 million specimens. There are currently approximately 230,000 accessioned sheets of vascular plants with an excellent representation of the vascular flora, including many rosids, of Florida, the southeastern United States coastal plain, and Haiti. The collection allows destructive sampling for approved projects, and it has provided extensive materials for DNA analysis. The collection is rich in historical material dating back to the mid-19th century, providing a valuable resource for identification and taxonomy. The University of Florida Bryophyte and Lichen Collection contains approximately 160,000 bryophyte and lichen specimens. The collection is worldwide in scope with an excellent representation of species from Florida and tropical areas such as Costa Rica, Venezuela, and Brazil. The UF Herbarium’s collection databases and image galleries provide interactive virtual access. Searchable databases contain ca. 122,000 of the 450,000 specimens in the vascular plant, bryophyte, lichen and algal collections.

The Herbarium houses two high-resolution digital copy stands. One with a 33 Megapixel Sinar Evolution H75 digital back and one with a 36.3 Megapixel Nikon D810 DSLR deployed with a MK Photo- eBoxTM system. These units can generate high-resolution (ca. 400 ppi) images and may be used to produce whole specimen images of vouchers for this project. Image galleries are being developed with a thematic focus. Digital image sets include potentially invasive, insectivorous, poisonous, and endangered plants of Florida, type specimens, orchids, Melastomataceae, and vouchers for a variety of floristic projects. All southeastern U.S. specimens are currently being digitized as part of the NSF-sponsored TCN project, Collaborative Research: The Key to the Cabinets: Building and Sustaining a Research Database for a Global Biodiversity Hotspot. The specimen-based images are associated with label information in the collection catalog, including data on habitat, flowering and fruiting period, frequency, and distribution. A common name search tool provides a walkway to the scientific names used in the catalogs. Support for the herbarium’s digitization effort is provided by the Florida Museum of Natural History, UF / Institute of Food and Agricultural Sciences, National Science Foundation, United States Department of Agriculture, Andrew W. Mellon Foundation, UF Libraries Digital Library Center, Florida Center for Library Automation, and the Florida Museum Associates.

**Interdisciplinary Center for Biotechnology Research**

UF supports the Interdisciplinary Center for Biotechnology Research (ICBR), which provides services and training in 7 Core Facilities. ICBR is vital to campus-wide training in technology-enabled science and bioinformatics through sponsored seminars, symposia, industry visits to campus, and training programs. ICBR also offers and encourages consultation with UF faculty and students to improve approaches to research problems. Although data generated through all of the Cores may be relevant to biodiversity science, we provide here summaries of only the Bioinformatics, Gene Expression and Genotyping, and NextGen DNA Sequencing Cores, as these are likely to be most relevant to our project. The Bioinformatics Core offers bioinformatics and biostatistics consulting and data analysis services to help researchers toward an in-depth understanding of large-scale data sets acquired from next-generation DNA sequencing, gene expression, and mass-spectrometry analytical technologies. Bioinformatics staff members have extensive experience in bioinformatics, genomics, transcriptomics, medical informatics, and translational research, and their expertise includes software and database development, big data analysis and statistics, and high-performance computing. The Gene Expression & Genotyping Core provides a variety of services related to Affymetrix and Agilent array processing, including genome-wide gene expression arrays, miRNA arrays, DNA arrays, and Gene Regulation Analysis arrays. This core also offers real-time PCR services, RNA-Seq, SAGE, small RNA and Sequence Capture library construction, and cDNA normalization for NextGen Platforms. The core provides basic molecular biology services such as DNA isolations, PCR reactions, and PCR purifications for sequencing, as well as fragment genotyping analysis and DNA and RNA quality assessment services using an Agilent Bioanalyzer. The NextGen DNA Sequencing Core provides quality, massively parallel, high-throughput sequence data using the most current instrumentation: i.e., multiple Illumina platforms and the Pacific Biosciences SMRT system, including the new Sequel platform. Free consultation services help researchers navigate through the complex matrix of experimental options represented by available sequencing technologies for an ever-broadening range of applications.

**APPENDIX 5 - EQUIPMENT**

**UNIVERSITY OF WISCONSIN – MADISON (Ané and Roy)**

The office space is fully equipped with the necessary equipment (including two network laser printers and a scanner) and supplies. Eight workstations are fully equipped with recent and relevant software for the proposed project (Sequencher, Geneious, etc.) and are shared in the PI’s laboratory.

The PI’s laboratory is fully equipped for research in molecular biology and protein biochemistry. Our equipment includes a BioRad CFX96 Real-TimePCR Detection System, a fully automated LC-2010A HPLC with autosampler for Nod factor purification, a Tecan GENios microplate fluorescence, absorbance and luminescence reader, three Eppendorf Mastercycler PCR machines, BioRad MyCycler and T100 PCR machines, a FOTODYNE Analyst image acquisition system, five stereoscopes including a LEICA M165FC fluorescence stereoscope equipped with a LEICA DFC310FX camera and a LEICA DMi1 stereoscope, five microscopes including a LEICA DMi8 fluorescence microscope equipped with a LEICA DFC365FX camera, a Vibratome 1000Plus, a Leica RM2245 microtome, four ultra-low temperature freezers, a Millipore water purification system, a Getinge 533LS autoclave, three laminar flow hoods, four shaking and four non-shaking incubators including one for tissue culture, a Fisher Isotemp CO2 incubator, a BioRad GenePulserII electroporation system, a Retsch MM300 mixer mill for high throughput plant DNA and RNA extraction, various Owl electrophoresis systems, BioRad protein electrophoresis and blotting systems, a GE ImageQuant LAS500 chemiluminescence detection system, an Interscience easySpiral automatic spiral plater, precision balances, a Sorvall RC-6+ Refrigerated Centrifuge, two Fisher accuSpin Micro microcentrifuges, a Eppendorf 5430R refrigerated microcentrifuge, two Fisher accuSpin 3R centrifuges equipped for microtiter plate applications, a Fisher Sonic Dismembrator, a BioRad spectrophotometer, two Cannon digital cameras, a NanoDrop 1000 for nucleic acids and protein quantification, and a Labonco dishwasher. For nitrogen fixation assays, the laboratory is equipped with a Qubit System and a Shimazu GC-2010 gas chromatograph with an autosampler for Acetylene Reduction Assays (ARA).

The PI’s laboratory has five plant growth chambers located in the basement of the Plant Sciences building and three in the Microbial Sciences Building. Two of them are equipped with aeroponic systems allowing an efficient production of about large amounts of root tissue.

**UNIVERSITY OF FLORIDA (R. Guralnick, M. Kirst, P. Soltis, D. Soltis)**

The laboratory of the PI Kirst is well equipped for the analysis proposed in this project. The laboratory has all equipment required for high-throughput extraction, analysis and storage of RNA and DNA, and transcript level quantitations by Q-PCR. The equipment available includes tabletop, high-speed centrifuges, -80oC freezers, QuantStudio 12k Flex Real-time PCR System, and a GenoGrinder2000 (Spex Certiprep, Inc.) for simultaneous extraction of total RNA or DNA from 96 tissue samples. For biomass density and 3D structure analysis, we have an X-ray microCT, a uCT40 instrument from Scanco Medical, equipped with a sample autoloader and dedicated workstation. For analysis of biomass composition, we have a Perkin-Elmer Spectrum 400 FTIR/NIRA scanning instrument with 96 well autosampler and UATR probe.

The co-PIs P. Soltis and D. Soltis share two laboratories. The laboratory in the Florida Museum of Natural History (Dickinson Hall) is set up primarily for DNA and RNA extraction, PCR, library construction, and phylogenetic and population genetic analysis. Items relevant to the proposal include two refrigerators, five freezers, two ultra-cold freezers, an ice machine, tabletop centrifuges, microfuges, Speed-Vac, vacuum pump, five thermal cyclers, several mini-gel rigs and power supplies, UV light box and gel imaging system, and Apple and Dell computers. The laboratory in the Department of Biology (Bartram Hall) is used for developmental morphology, RNA isolation, cDNA library construction, screening libraries, *in situ* hybridizations of gene expression, and molecular cytogenetics. This lab has an Olympus SZH-10 stereomicroscope and Zeiss compound microscope, plus the equipment for RNA research.

Co-PI Guralnick maintains a dedicated facility at the Florida Museum of Natural History that contains a laminar flow hood with UV, dedicated pipettes and other equipment that never come into contact with modern DNAs, and a set of equipment for working with plant and animal historical and paleo-materials (bone drills, etc.).

The co-PIs have direct access to UF Research Computing, which operates the HiPerGator supercomputer, a cluster-based system with a combined capacity of about 21,000 cores in multi-core servers. In November 2015, this capacity was expanded by adding 30,000 new Intel cores, bringing the total to 51,000 cores. The servers are part of an integrated InfiniBand fabric. The clusters share over 5 PetaBytes of distributed storage via the Lustre parallel file system. Also, Research Computing houses about 2 PB of storage for the High Energy Physics collaboration of the Compact Muon Solenoid (CMS) experiment. The system includes over 100 NVIDIA GPU accelerators and 24 Intel Xeon Phi accelerators, available for experimental and production research, as well as for training and teaching. If additional resources are needed, UF is directly connected to the Florida and National LambdaRails, allowing rapid data transfer and sharing, facilitating the use of additional national research facilities such as the XSEDE resources. Research Computing provides advanced support and training to the user community. Many training materials are now available online. User feedback meetings are held as well as periodic training workshops called *Research Computing Day* are organized every semester. Several graduate courses use HiPerGator and train and prepare graduate students to use the clusters and the software for their thesis research. The coPIs have a substantial allocation (Guralnick – 50 cores). Startup resources support storage and computing resources on these boxes.

In addition to university-wide high-performance computing solutions, co-PI Guralnick oversees a number of hardware and software assets for managing computing and data resources for this project. In particular, he and his staff will be maintaining two PowerEdge servers (40 and 40-core, total 50 TB replicated disk space with increases to 74TB planned shortly) that host multiple informatics projects developed at the University of Florida. The servers will be housed in his laboratory. The dual socket servers have vanilla Linux and Xen installed to provide direct access to machines and easy ability to virtualize. These run core software (R, Python) and are well-used for medium scale parallelizable computing tasks and particularly interactive development. Other funding sources also include database and software development related to management of large-scale biodiversity data.

Guralnick also has access to 176 computational nodes through the NSF XSEDE program (using the Jetstream cluster), and 4 Tb of storage; other startup allocations on Stampede and similar systems collectively comprise 100,000 processor-hours and 1 Tb of storage; we plan to expand these resources through a full XSEDE proposal to be submitted this summer.

The Florida Museum of Natural History also provides us file services, hosting about 1.5 Tb of Soltis Lab data on a redundant, clustered file system with daily off-site backup. The Guralnick group works with the Office of Museum Technology (OMT) to assure secure access to our database resources remotely and that any services we develop for external communities (e.g. web content, APIs, etc) meet University security standards.

Finally, the investigators also have access to all equipment available at the Interdisciplinary Center for Biotechnology Research (ICBR, www.biotech.ufl.edu), also located at the Genetics Institute. This is a state-of-the-art facility that disposes of the latest technology in DNA sequencing, microarray analysis, proteomics, and metabolomics. The technology available includes an Illumina NextSeq500, MiSeq and HiSeq3000 DNA sequencers, robotic workstations, and most other equipment that may be necessary for preparation of genomic DNA and RNA libraries, and next generation sequencing.

**APPENDIX 6: DATA MANAGEMENT PLAN**

**Overview and Roles/Responsibilities**

All investigators will follow the required policies dictated DOE, on the dissemination and sharing of research results. The PI Kirst will have fiduciary responsible and be accountable for execution of this data management plan. He will coordinate the data resources, in consultation with the rest of the team, on all data aspects related to sequencing efforts, and tree assembly. The most extensive data generated in this project will involve DNA/genome sequence, phylogenetic information, and nodulation trait data – several investigators in the project have extensive experience in managing this type of data and making in available. The co-PI Guralnick will be the technical lead and assure that this effort is managed according to the plan below. We will work to ensure proper licensing, of products and that ability for providers to will, accrue professional credit for their contributions.

**Raw Research Data Products**

*Raw materials and DNA/RNA extracts:* All sampled museum specimens (DNA) will be imaged with captured minimal data (original species ID, determiner, country, museum barcode) and linked to a database that will be managed by Co-PI Guralnick. The large number of DNA/RNA extracts will be linked by miniature barcodes and barcode scanners; ultimately they will be accessioned in the FLMNH Genetics Resources Repository and (for genomic data) several partner herbaria (e.g., Kew, RBGE) in accordance with their destructive sampling policies (part of routine biological collections activities).

*Sequencing and assemblies:* Raw sequence reads, assemblies, and assembly consensus calls will be identified by unique object identifiers to be linked with taxonomy and original herbarium specimens/germplasm accessions.

*Nodulation database:*Nodulation status for species will be linked to all other resources via taxonomy tables and stored in a local database collated and curated from the ILDON and TRY databases. This database approach critically reconciles taxonomy between phylogenomics, transcriptomics, and the literature in an expandable, open-science approach that will serve as a community resource.

*Vouchering:* Phylogenomics will be primarily based on existing vouchered collections, which will be digitized with minimal data and disseminated online following community best practices. Vouchers will be prepared for all living germplasm collections to reproducibly substantiate identification.

**Data Storage and Preservation**

All data assets, from specimen provenance data to nodulation data, phylogenetic trees, and analytical products such as assemblies, will be regularly archived. This will be accomplished utilizing existing databases for managing specimens to genomic sequences already maintained by the co-PI Guralnick. We will locally archive all raw data products and subsequently deposit them in public repositories (see Data Sharing below). Data resources that are “live” in the system (e.g. phylogenetic trees, nodulation data, specimen data, and taxonomy) will have APIs that ultimately allow us to assure centralization of resources and ability to provide web-based curation. Guralnick is actively working on long-term preservation solutions and will set up pipelines for moving data resources to Cyverse and DataONE, which provide data replication services along with assignment of GUIDs (e.g. DOIs); both follow Lots of Copies Keep Stuff Safe (LOCKSS) best practices.

**Data Formats and Metadata**

We will publish the database model linking all portions of the project. This project will leverage existing infrastructure and standards where data assets include raw schematized data (e.g., phylogenetic trees in typical formats, trait information).

*Sequence and Tree Data*: Data formats for DNA data will follow current standards – FASTQ for raw reads, SAM for assemblies, FASTA or FASTQ for assembly consensus. For tree data, we will build metadata descriptions utilizing an application profile of MIAPA (<https://github.com/miapa>) for all trees generated and stored here, and conforming to Open Tree standards when pushed to that platform.

*Nodulation database:*Our nodulation database will follow the ontology of <http://www.ildon.org/ildon.html>, including classification of the structure.

*Serializations:* Output formats will have a strong focus on machine-readable products, especially CSVs using community-based semantics. APIs will represent data assets using JSON-LD, KML, GeoJSON, and CSV.

**Data Sharing and Public Access Policies**

*Raw materials and DNA/RNA extracts:* A public, searchable list of sampled taxa and source museum specimens will be released online, following the template of the 1KP project list (<http://www.onekp.com/samples/list.php>). Raw, unprocessed DNA data will be uploaded to the sequence read archive (SRA; <http://www.ncbi.nlm.nih.gov/sra>) and publicly released at publication, following community standards if needs arise to change repositories.

*Sequencing and assemblies:* Gene assemblies and consensus calls, alignments, and further downstream products (multilocus coalescent and concatenated trees, individual gene trees) will be uploaded to Dryad, and an optimal total-evidence tree will be submitted to the Open Tree of Life project (<http://opentreeoflife.org/>). Consensus calls on gene assemblies will additionally be annotated and uploaded to GenBank.Provisioning the raw and output data resources to open repositories assures repeatability.

*Software development:* While the project relies primarily on database architectures already in place, we expect further development of informatics pipelines (particularly *Aim I*); software tools developed during the project will be publicly available on GitHub (GNU GPLv3), including public version control of the source code.

*Provenance and ownership:* All quality assurance, provenance and process metadata, along with data licensing, confidentiality, etc. will be properly described at the level of the whole project and its goals as well as at the level of individual resources. Assigning licenses to data products is already in place in MOL. We will not long-store or expose any copyrighted material as part of this project but do plan to mine factual content under fair-use and provision those data via our infrastructure.

**Publication of machine readable content linked to publications**

We plan to deposit input and output data analyses to Dryad (http://dryad.org), as an archival source of data files used in analyses, and to assure replication. We also will set up, as is feasible (particularly for post-NGS-assembly analysis tasks such as visualizing expression), Jupyter and R notebooks. These provide means for others to also not only examine the code for analysis, but run those analyses directly, and to have annotated and documented code as well. In this way, publication results, including even figures and tables, can be replicated more efficiently.

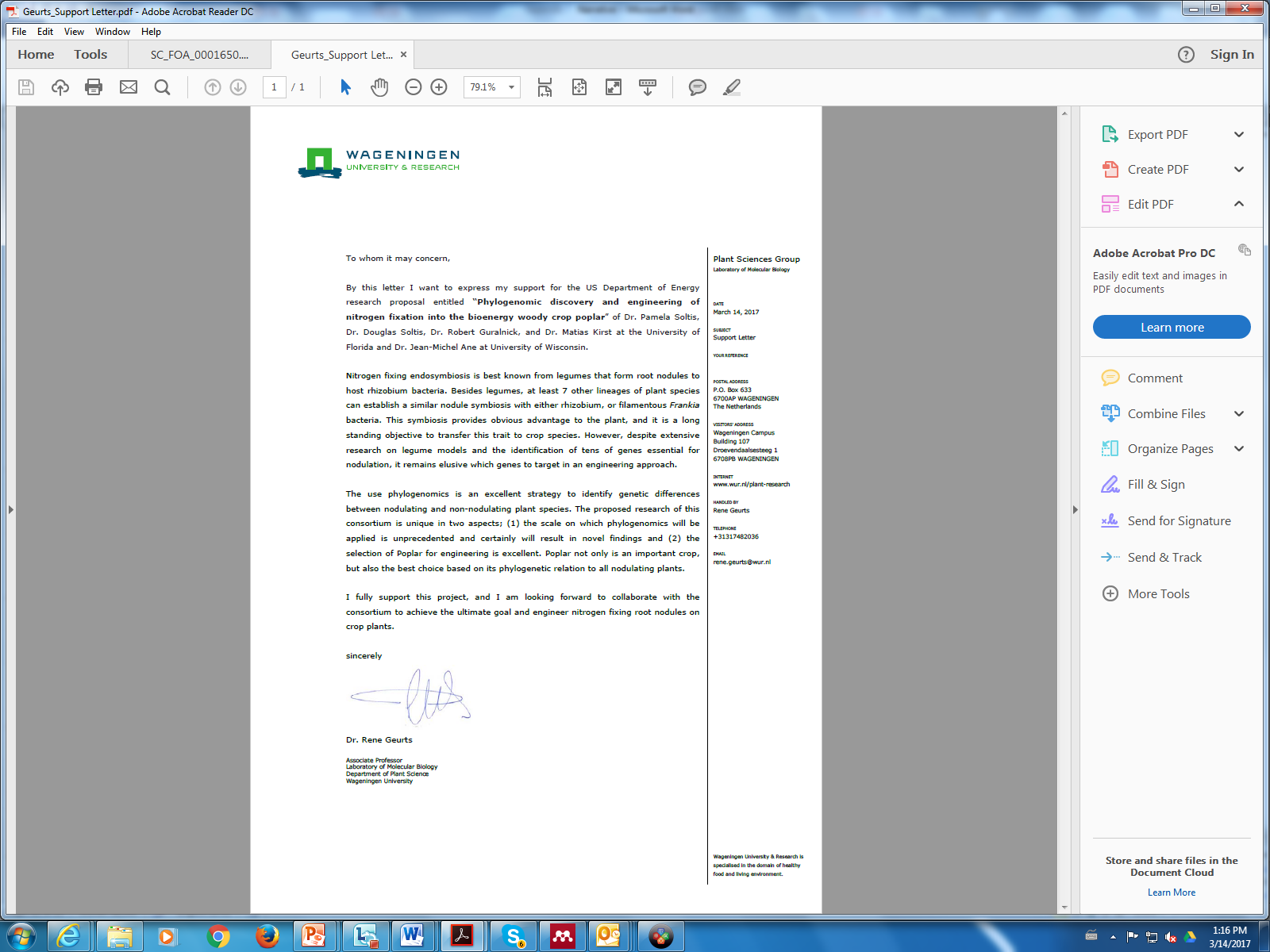
**Embargo policies**

We will only embargo phylogenomic and transcriptomic data until the publication of results from those data, at which point we will make data available via typical resources, such as Genbank/SRA. Other datasets, including experimental outcomes of experiments to induce nitrogen fixing, may require longer time frames on data embargoes.

**Patent policies and data licensing**

We do see the potential for patentable products derived from this work, but gene/genomic and transcriptomic data are not patentable, and we will make those data available without restriction. All data products, unless specifically required by a partnering repository, or considered to be of potential commercial use, will be licensed under creative commons attribution licenses.

**APPENDIX 7 – LETTERS OF COLLABORATION**



From: Brian O'Meara.

By transmitting electronically, I acknowledge that I am listed as a collaborator on the proposal entitled " Phylogenomic discovery and engineering of nitrogen fixation into the bioenergy woody crop poplar with Matias Kirst as the Principal Investigator. I agree to undertake the tasks associated with me as described in the project description of this proposal, and commit to provide or make available the resources designated in the proposal.

Thanks,

Brian