1. **Signficance..**
2. **INNOVATION..**
3. APPROACH

Historically, the primary challenge in developing plant-derived drugs is to identify the active molecules and their biosynthetic processes in the plant. Both stages are extremely cost and labour intensive and a diminishing rate of return has heavily reduced the investment in plant natural product discovery of recent [Koehn and Carter, Nature Reviews Drug Discovery 4, 206-220 (March 2005)]. Thus, while a few important plant-derived therapeutic chemicals (quinine, taxol, vincristine) have been identified and even studied at the genomic level in 14 species [http://medicinalplantgenomics.msu.edu/], a vast and diverse collection of medicinal plants used for centuries by native populations, *have not* been studied or even known to modern medicine.

Our approach offers a radical shift in the traditional paradigm for drug discovery that harnesses the power of evolutionary genomics. The enabling innovation is the development of a phylogenomic pipeline to identify genes that underlie the evolution of therapeutic compound production in species without the need of identification of the active ingredient first.

To implement this approach, we performed a phylogenomic analysis of 22,000 plant orthologs across 150 plant genomes [1]. In a proof-of-principle example, this genome-scale phylogeny revealed that genes responsible for antioxidant production (2 steps in glutathione synthesis) are significantly over-represented in the set of genes that provide positive branch support for three independent clades in the rosids/asterids, plants known for their antioxidant traits [1]. Thus, this phylogenomic approach enabled us to exploit convergent evolution in this clade to derive defined hypotheses for genes and pathways involved in the “nutriceutical” use of Rosid/asteroid plants based on plant genome data.

We now propose to implement this phylogenomic approach to the genomes of plant species based on their medicinal use traits. For this, we will exploit a unique collection of 500 specimens in the “Wild Medicine” collection at NYBG and their database of “medicinal uses” collected by NYBG botantists and anthropologists, working with native healers around the world. This knowledge can be used as “Medicinal” traits that appear in close or distantly related plants, and can therefore be treated as a means to uncover the genomic basis for convergent evolution towards the medicinal trait and to uncover the underlying genes, as detailed below.

In this approach, the relevant genes and pathways identified using genomic data from Aim 1 and phylogenomic analysis in Aim 2, will enable us to derive hypotheses for specific classes of compounds to extract and test in bioassays in Aim 3 reducing the potential pool of bioactive candidates substantially. We will focus on plants with “anti-infectious” properties in a proof-of-principle study, owing to the ease of bioassays, and to fill the urgent current need in the US and world-wide for new antimicrobial, fungal and parasitic agents.

**Aim 1.** “**Wild Medicine Collection @ NYBG”:**

**A. Species selection and sampling for genomic analysis**.

**Rationale:** In this aim, we will focus our genomic analysis initially on 50-100 species from the Wild Medicine collection. Species are selected based on: 1. Specificity for medicinal use, 2. Evolutionary diversity, and 3. Conservation status. Our focus for this proof-of-principle study is on plants with “anti-infectious” properties benefits due to the relative ease of bioassays (in Aim 3), and the potential to fill an urgent need for new antimicrobial, fungal and parasitic agents. In each clade, we will also select one non-medicinal plant (for this trait) as an outgroup to provide a higher phylogenetic resolution. For each of the initial species selected, we will perform RNA-seq analysis for 2 tissues using an efficient and economical “balanced block design”. The goal is to create an effective coverage of the expressed genes (e.g. “Gene Space”) to build a phylogenetic tree (Aim 2A), and to also generate expression datasets for correlation networks (in Aim 2B).

**The “Wild Medicine” Collection at The New York Botanical Garden:** Field botantists/anthroplogists at The New York Botanical Garden (NYBG) have collected over 500 live species of medicinal plants that were recently part of an NYBG exhibit and collection called “Wild Medicine” (May 18-Sept 8, 2013). The Wild Medicine collection represents the largest repository of live samples of medicinal plants in the world. The curators who built this collection have simultaneously collated extensive information about their medicinal usage from local people including shamans, thus synthesizing the largest database of medicinal uses of plants. We aim to utilize this unique botanical collection of live specimens and the database of their uses to broadly expand our knowledge of plant-derived medicinal compounds and more importantly, to specifically identify the underlying metabolic pathways that lead to their synthesis.

**Creation of Medicinal “signatures” for the “Wild Medicine” collection**.

Our ultimate goal is to connect “medicinal traits” with genes via a phylogeny. To accomplish this, we devised a database and binary code to group species according to their “Medicinal Signatures”. We first applied text-mining techniques to the medicinal use annotation of each of the 500 species in the “Wild Medicine” collection. This identified 27 general categories and each species was assigned a binary character (0, 1) for each medicinal property (e.g. GI, ENT, INF) (Table X). This binary matrix was then used to identify groups of species that are used to treat the same conditions (Table 1). Next, these “medicinal trait” signatures for each species (0101010101000…..) were scored for their “specificity” for the trait of interest. As shown in Table I, species in Group A show high specificity for the trait “infectious”, while Group B shows low specificity for this medicinal trait. Despite the unstructured nature of descriptions of medical uses provided for these species, selecting plants with a “high specificity” signature for use to treat “Infections” (289 species) reveals commonalities to the recorded usage (Fig. 1). It is noteworthy that these 289 species grouped under the “anti-infectious” set cover a broad range of plant evolution spread out over hundreds of millions of years. And yet, this set of species has independently evolved to produce bioactive compounds underlying their common use for “anti-infectious” treatments. This congruence of evolution based on medicinal traits can be used to identify pathways that are shared between such species, despite their very large evolutionary divergence and phylogenetic distance, as described below.

**Selection of Species**: Species to be studied are selected on the basis of the following criteria:

1. ***Specificity of bioactivity****:* “High Specificity”. E.g. Species used against infections with little to no known usage for other ailments. Our initial focus on “anti-infectious” in this study, owes to the ease of bioassays, and to address the urgent need for new antimicrobial, fungal and parasitic agents.
2. ***Phylogenetic diversity****:* Species are selected to span diverse clades to increase the chances of detecting signatures of convergent evolution.
3. ***Conservation status****:* Species abundant in nature are prioritized to allow ease of collection and prevent damage to endangered species.

An initial set of 50 species for “anti-infectious” were identified for this initial study using these three criteria and multiple individuals from each of these species are being maintained as live specimens in the NYBG. We anticipate adding an additional 25-50 species (for a total of 100 species) based on increased sampling in evolutionary clades that show promising results either in the classes of compounds they make (Aim 2&3) and/or the Bioassays (Aim 3).

**B. Transcriptome sequencing by RNA-seq analysis**

**RNA-sequencing the species “Gene space”,**

We will sequence the transcriptomes of multiple organs to effectively uncover the “gene space” of expressed genes as a proxy for the genome.For each of the initial 50 species selected above, at least two organs will be collected for RNA extraction. The choice of organs will be such that we sample one therapeutic and another non-therapeutic organ for each species. In specialized cases where there is evidence (either from the literature or chemical information from Aim 3) that a potential source tissue i.e., origin of bioactive compound is different from the therapeutically used tissue, additional sampling of the source tissue will be performed to capture genes in the biosynthetic pathway.

**A Balanced Block Design for maximal coverage at minimal cost**: RNA-seq libraries made from RNA isolated from these tissues will be profiled using the Illumina platform to generate a high depth profile of gene content for each species. To perform economical and reliable RNA comparison across tissues (i.e. leaves, roots, influorescence) and species (50-100 species), we will use a “balanced block” RNA-Seq design with multiple levels of replication [Auer & Doerge Genetics 2010]. This design, successfully implemented in the Coruzzi lab and CSHL, involves RNA-Seq libraries from 2-3 replicates of each tissue, and barcoding each library and sequencing pooled libraries with replication. This strategy will allow us to reliably sample one species per lane of Illumina Hi Seq machine.

**DATA ANALYSIS**

**Representation of Species “Gene Space”**: The “gene space” for each species will be assembled using the state-of-the art transcriptome assembler Trinity [Haas et al. Nat Protoc. 2013 Aug;8(8):1494-512.] Each lane of sequencing will yield roughly 200 million reads which will allow us to assemble about 70-80% of the expressed genes [Liu et al 2013. PLoS ONE 8(6): e66883] with particular bias towards the biosynthetic pathways active in these tissues and therefore reduces the problem of missing data in phylogenomic analysis. Coverage of the genome achieved for each species will be estimated by determining the proportion of core eukaryotic genes [Parra et al. *Nucleic Acids Research*, 37(1): 298-297 (2009)] that are fully covered. A second criteria for estimating completeness of the Gene Space, will be to calculate the coverage of all known plant secondary metabolic pathways from the KEGG database in the species. If a certain species shows significantly less coverage than the overall species set, it will be re-sequenced to attain a higher depth of coverage.

**Gene expression values**: Gene expression values will be determined by absolute counts of reads from each organ aligned to the assembled contigs and quantile normalized[Bullard et al 2010] at a species level. From this expression matrix, two measures of correlation of expression will be calculated. First a Pearson’s r measure will be calculated within the species for all-vs-all gene pairs across all sampled tissues and replicates. A second measure for correlation will be calculated, within the medicinal plant organ, between every Ortholog group, as determined in Aim 2A, across all species. In cases where a given ortholog group includes multiple genes within a species the expression values for the most correlated pair will be chosen.

**Aim 1. Expected Outcomes and Alternate Approaches.**

Aim 1 will generate the “gene space” for 50+ species of plants that have a high specificity for the treatment of infections. We have a high level of expertise in generating RNA-seq libraries (NYU), performing Illumina sequencing (CSHL), and in bioinformatic analysis (NYU/CSHL). We also have a high degree of expertise in extraction of RNA from diverse species and tissues (NYBG), so for samples recalcitrant to RNA isolation by one method (e.g. Trizol extraction), we have a battery of alternate methods on hand [Johnson et al. (2012) PLoS ONE 7(11): e50226]. In addition to the tissue collected for RNA sequencing, excess tissue will be collected and freeze dried to allow repeated compound extraction in Aim 3. Finally, these plant species will be maintained as live specimens at the New York Botanical Garden to allow resampling as and when required.

**Aim 2** **Phylogenomic Analysis: WildMedicine v1.0**.

**Rationale**: In Aim 2, we will perform phylogenetic and bioinformatics analysis of RNA-seq data for the 50 species from Aim 1 to identify hypotheses for genes and pathways associated with their medicinal trait “anti-infectious”. We will use two complementary approaches to achieve these goals. In **Aim 2A** we will build a genome-scale phylogenetic scenario for the evolution of the 50 medicinal species using the “gene space” data for the 50 species and fully sequenced non-medicinal species. This genome-scale phylogeny will enable us to identify genes and pathways that support the divergence (or convergence) of medicinal traits in specific species, and clades. To do this, we will use the BigPlant pipeline for genome-scale phylogenies that we have successfully applied to 150 plant genomes [1]. In **Aim 2B**, we will consider the expression levels of genes in correlation networks and their intersection with biochemical pathways. Finally, in **Aim 2C**, we integrate data across three levels: phylogenomic signal, network connectivity and biochemical pathway. The integration of these three levels will identify genes and pathways involved in “anti-infectious” compound synthesis that will be used to direct chemical and bioassay studies in Aim 3.

**Aim 2A. A phylogenomic analysis of the Wild Medicine collection.** In this aim, we will perform a phylogenomic analysis of the 50+ medicinal plant genomes along with representative outgroups from each clade (selected in Aim 1) and all fully sequenced land plants. Bi-annual updates of our BigPlant analysis, as part of our NSF funded effort [NSF# IOS-0922738], include all fully sequenced plant genomes and this Aim will benefit by including the resources developed therein. We will implement this using a phylogenomic pipeline we have previously constructed and validated using genome sequences from 150 species (BigPlantv1.0) [1]. This bioinformatics pipeline will enable us to identify the individual gene losses, gains and mutations leading to species divergence [1,2].

**BigPlant framework**: The “BigPlant” phylogenomic pipeline (Fig. 2) is a parsimony-based phylogeny building software that can handle large genomic datasets at an, unprecedented one- amino acid resolution level [1] and provide information about the level of support provided by each gene/amino acid to each branching point on the final tree. It has the ability to include complete and partial genomes in the same analysis and is robust to missing data. These properties are particularly amenable to this project since our transcriptome sampling will generate partial genomes with an estimate 20-30% of genes missing. The missing gene information is unlikely to be a major factor in our approach since we are specifically sampling the active plant part and therefore are very likely to capture the relevant actively expressed genes. The positive PBS values for each gene gives us a reliable measure to identify genes underlying diversification of species and/or groups and sets of such genes can be readily used in downstream over-representation analysis.

We have previously shown that such high-resolution evolutionary analysis of plant genomes is capable of identifying the genes that provide Positive Branch Support [Baker RH, DeSalle R (1997) Syst Biol 46: 654–673] at each node of the BigPlant Tree [1]. This approach has enabled us to identify specific genes and biological processes (using Gene Ontology/MIPS term analysis) that have independently evolved in multiple groups of species [1] i.e., signatures of convergent evolution. For example, our previous work on the genomes of 150 species of land plants (Fig. 3), including 5 fully sequenced genomes, was able to identify genes involved in oxidative protection as having specifically evolved in three separate clades of the Rosids and Asterids, plant species known to be rich in antioxidants (Fig. 4).

Genes/proteins yielding positive PBS contributing to the appearance of the “anti-infectious” trait on the phylogenetic tree are excellent candidates to be involved in the biosynthesis of the underlying therapeutic molecules. This approach is applicable to the identification of convergent evolution in the same biosynthetic pathway. While some of the plant species have likely converged on similar classes of compounds to resist infections, it is also possible that one or more plant groups with shared evolutionary history have developed an entirely new class of molecules for the same purpose as independent key innovations. Our phylogenetic approach enables the detection of such evolutionary physiological adaptations by finding pathways whose components evolve under specific phylogenetic regimes, as shown below. Figure 5 shows such an example from our BigPlant v1.0 tree of a small group of plants that are known to produce abundant aromatic oils while their immediate sister clade does not. The divergence between this clade and its closest relatives is strongly supported by the BigPlant v1.0 framework [1], with 3 genes crucial to the synthesis of tryptophan, the amino acid precursor of such aromatic compounds. Therefore, our approach is uniquely geared to identify pathways that are recurrently activated across plants species or that have specifically evolved in a small group by detecting genes involved in these pathways as being evolutionarily critical to their emergence.

**PhyloBrowse:** We have developed an intuitive and data-rich visualization tool called PhyloBrowse [http://phylobrowse.bio.nyu.edu] for querying support and over-representation data generated from the BigPlant pipeline. Our visualization tool is a fully capable tree viewing software but more importantly allows the user to mine the phylogenomic analysis to identify genes and pathways enriched at the nodes of their interest. This tool will be deployed to display data from this project but also from other projects such as MPGR [http:// http://medicinalplantgenomics.msu.edu/;NIH#1RC2GM092521], 1KP [http://www.onekp.com] that generate large scale genomic data from medicinal plants. All these datasets will be analysed through the BigPlant pipeline and will be available to the broader community to derive their own hypotheses regarding the traits of their interest.

The list of genes supporting each node will be queried using (KEGG, PlantCyc, Reactome etc.) for their potential role in the biosynthesis of the active molecule(s). An intermediate step in our BigPlant pipeline is the creation of gene families spanning all species analysed and provides information if any gene family shows significant expansion in the medicinal plants. The representative member of a gene family showing such expansion will be added to the list of PBS genes as evidence of evolutionary signal. We anticipate the pathway information provided by overlaying PBS genes on known pathways will be incomplete and therefore we will expand our ability to identify pathways by incorporating a co-expression based network analysis and overlaying that additional information on the biochemical pathways databases (Fig. 4) as described in Aim 2B, below.

**Aim 2B. Network-based identification of biochemical pathways associated with medicinal traits.** Deep transcriptome profiling across organs and species allows us to build networks of genes that correlate in expression and are specific to the plant tissues that produce therapeutic compounds. Our prior expertise and experience in developing multilevel networks by incorporating expression, protein interaction and metabolic information in a pipelined systems biology analysis, provide us with a means to analyse and interpret data across data-types, species, and tissues [3]. It has been shown that genes which are highly correlated in expression are likely to interact with each other or be involved in the same metabolic pathway [4]. This approach to network building – integration of co-expression, protein-protein interaction, and metabolic pathway information, has enabled us to generate new hypotheses and uncover new biology in plants, e.g., nutrient control of the circadian clock [5]. The correlation measures developed from expression datasets generated in Aim 1 for individual genes at a species level and for ortholog groups across the 50 species will form the core of the co-expression networks. Based on the phylogenomic analysis in Aim 2A, genes that drive phylogenetic divergence of medicinal plant species will be used to nucleate Node Vicnity networks (NVNs) using the correlation values from Aim 1 that pass a stringent cutoff (R>0.8, p<0.01) to create network modules. For the subset of genes that have a clear ortholog in Arabidopsis (as determined in Aim 2A), protein interaction [3] and metabolic information [KEGG, AraCyc] will be included to expand the network module. Each network module is then mapped onto known pathways using the gene-to-product associations in KEGG, Reactome and PlantCyc.

**Aim 2C. Convergence and Integration of Phylogenomic, Network-based and Pathway analyses.**

The strength of our approach lies in integrating the evolutionary signal of convergence from the phylogenetic approach (Aim 2A) and the network knowledge from correlation (Aim 2B) to narrow down the candidate genes involved in the biosynthesis of “therapeutic” compounds or their precursors. We will combine these datasets with the objective of generating a ranked list of candidate biosynthetic pathways as follows:

Step 1: Identify PBS [spell this out] genes from Aim 2A and overlay them on metabolic pathways from KEGG, PlantCyc, Reactome etc.

Step 2: For each PBS gene construct a Node vicinity network from the -co-expression of the gene with all other genes from the medicinal organ in all species. Correlation of expression (R≥0.8,p<0.01) is used to create gene pairs that are then linked by transitive closure to form modules. Modules are overlaid on metabolic pathways as in Step 1.

Step 3: Perform pathway over-representation analysis with PBS genes using a Z-score to measure the difference between observed and expected (under a hypergeometric distribution) number of genes in a pathway. Permutation tests will then be conducted to assign a p-value to each pathway. All pathways are ranked based on this p-value.

Step 4: Repeat pathway over-representation analysis with the PBS-seeded NVNs [spell out] from medicinal organs and calculate a p-value as described in Step 3. All pathways are ranked based on this p-value.

Step 5: Calculate the rank of each pathway as the product of its ranks from Step 3 and 4. This method will rank pathways higher that have a greater percentage of genes providing PBS support for medicinal plants. The pathway rank will further increase if the pathway has co-expression of genes in medicinal plant organs due to the rank product approach. This ordered list of pathways provides the best candidate for biosynthesis of therapeutic molecules and the partitions expected to contain compounds produced by such pathways will be prioritized in Aim 3 as described below.

It should be noted that certain highly correlated network modules expressed in the Medicinal plants will not map to any known pathways in standard metabolic databases. The most promising of such modules (expressed highly in medicinal tissues but not others) are then labelled putative novel pathways and members of this module will be manually curated against all known gene sequences in public databases (eg. NCBI’s Nucleotide database) with particular attention to those described in previous literature as being involved in the biosynthesis of bioactive compounds. We will also seek the advice of our natural product consultant to elucidate the class of compound that might be generated by this “pathway”.

NEED INPUT FROM MARKUS LANGE ON HOW TO TRANSLATE THE PATHWAY INFORMATION INTO LEADS FOR AIM 3.

The primary outcome of this analysis will be the identification of pathways that are associated with classes of compounds eg., alkaloids, flavonoids etc. that are produced in the medicinal tissues through the elucidation of pathways showing an evolutionary signal in the medicinal plant species and/or clades. This knowledge will greatly help focus the extraction and bioassay efforts in Aim 3. Crucially, this approach will simultaneously identify key components (through phylogenetic signal) and their pathway neighbours (through network analysis) that are responsible for the production of therapeutic compounds. These compound and pathway predictions will be verified through standard natural product chemistry and bioassays in Aim 3.

**Aim 2: Expected Outcome and Alternate Approaches**

The phylogenomic approach used in Aim 2A has been validated for 150 land plant species, and was used to derive hypotheses for genes underlying species (and trait) divergence [1]. This aim will implement the BigPlant phylogenomic framework to a set of species with “anti-infectious” activity from the “Wild Medicine” collection at NYBG and a set of outgroups. We anticipate finding genes that provide positive branch support at each node in the tree and using an over-representation of term analysis we can also estimate the “biological functions” encoded by these genes. The list of PBS genes will be expanded into functional modules by incorporating co-expression networks in Aim 2B. Overlaying the metabolic pathways on the PBS genes and the co-expression modules derived from them will provide leads for potential pathways and compounds for assays in Aim 3. This combination of data types (eg. Phylogenetic, metabolic and correlation analysis) is the breakthrough innovation of our approach and should provide a rational method for identifying bioactive compounds and their biosynthetic pathways. We anticipate two primary challenges to this approach: 1. The development of novel pathways through expansion in gene families rather than amino acid changes in existing proteins which would not be picked up by our PBS based approach and 2. Entirely novel pathways that are unknown and hence unmappable in our pathway overlay approach. Our phylogenomic framework captures the membership of gene families and can detect specific expansion in specific plant groups as was shown in our recent study of Membrane transporters [Leran et al. TiPS 2013]. Therefore members of gene families that show specific expansion that correlates with the Medicinal trait will be added to pathway over-representation analysis in Aim 2C to address challenge 1. For the second challenge we will manually curate the genes from NVNs [spell out] (Aim 2B) that correlate well with Medicinal trait by employing the expertise of our bioinformaticians in collaboration with our expert Biochemist and natural product chemists at NPDI. [This is very vague and the word “manual” and “bioinformatician” doesn’t work. How about: For the second challenge, we will identify genes that correlate with medicinal traits regardless of literature annotations.]

**Aim 3** **Natural product chemistry, biological assays and compound evaluation**

**Rationale and Overview:** Our phylogenomic approach in Aim 2A, combined with network and metabolic pathway analysis in Aims 2B and C, will enable us to identify leads on pathways and/or general classes of compounds associated with “anti-infectious” activity. Those hypotheses for compounds involved in the “anti-infectious” activity will provide leads for Natural Product Chemistry assays and BioAssays conducted in Aim 3 – which use well-tested approaches executed by the NPI team, as described below. We anticipate conducting such assays on 25-50 species in this proof-of-principle example, to demonstrate the ability of phylogenomic approaches to direct natural product discovery.

This aim attempts to verify the predictions of bioactivity within limited sub-fractions of plant extracts. Information of the “medicinally active” plant part from Aim 1, and the potential pathway/class of compound from Aim 2, will both be used to prioritize the order and limit the scope of bioassays. An iterative cycle of extractions and assaying will be conducted as described below.

***Aim 3 A. Phytochemistry: Extractions and early steps of processing***

NPDI personnel have decades of experience in the processing and extraction of plant materials and have worked extensively with collaborators at NYBG. Our approach utilizes a standard partition sequence starting with ethanol extracts of dried powdered plant followed by solvent/solvent partitioning to afford low-polarity, medium-polarity and polar subfractions [REF]. This methodology has proven effective in yielding preparations that can be reliably bio-assayed and used in the early classification of extracts [REF]; materials of this sort are suitable for a first and more detailed examination of composition.

This process, which will be carried on the 100-250 gram scale of dried plant material [REF], will be of use in the particular case of the proposed study, correlating specific plant parts with meaningful biological activity. The subfractions are further suitable for an attempt at associating activity with particular types of compounds. Leads of specific classes of compounds (e.g. terpenes, alkaloids, etc) from (again should you change this according to mike’s email?) Aim 2, will direct our natural products extraction protocols.

Medium resolution methodology that we use can point very early on to certain classes of compounds, *e.g*. polyamide SC-6 chromatography in polar solvents combined with behavior on Sephadex gels for small polyphenolics; acid-base extraction and color tests for alkaloids and the like; patterns of extraction and chromatography for small terpenoids, triterpenes and saponins; particulars of extraction, chromatographic behavior and response to mild hydrolyses for glycosides of various sorts etc. A combination of these methods should allow for early classification of bioactivities and determination of how well genomic patterns and predictions are correlated with classes of compounds.

***Dereplication procedures***: LCMS dereplication [REF] will be an essential component in any Natural Product discovery process and will be used in the early stages of evaluation of plant extracts and sub-fractions. This will be done in conjunction with our extensive in-house database [REF] as well as with commercially available materials, e.g. Chapman & Hall dictionary [REF]. NPDI has access to high-resolution equipment at the Bristol Biotech Center as well as through our close collaboration with Fundacion Medina (NEED LETTERS?)

***Aim 3B. Bioassays for Anti-infectious activity.***

Anti-infectious activity of plant extracts and sub-fractions will be evaluated and monitored in both standard agar diffusion assays [REF] and in liquid growth format [REF]. NPDI personnel have extensive experience in running such assays and in their use in tracking and characterizing compounds of interest. A panel of gram+ and gram- bacteria and two fungal species will be utilized for testing crude and partitioned extracts. The panel envisaged will include S. aureus, B. pumilis, H. influenzae, E.coli (both wild type and permeabilized strains), C. albicans and A. fumigatus. Samples dissolved in 20% DMSO are suitable for these assays; dose response in twofold dilution steps will be included.

For each selected plant of interest, bio-assay driven fractionation efforts will utilize the most 1-2 potent activities uncovered at the crude stage for the (usually) three-step sequential chromatographic steps necessary to purify a compound. The final products will be then evaluated in the complete panel. It will also be possible to characterize final products as being either cidal or static agents.

In our experience at NPDI, fully one of every four *random* plant ethanol extract produces antibacterial activity of some sort, particularly against the gram+ agents. It is thus fully anticipated that a *selected* set of plant extracts will provide plenty of biological activity that can be followed to good use. Particularly, samples that show activity from partitions containing novel classes of compounds (eg ., [An example of a class that is not commonly bactericidal]) provide exciting new leads in the development of novel drugs. Results from this Aim provide important feedback to Aims 1 and 2 (Fig 6) in two ways. First, individual species or evolutionarily related groups from the initial set, that show particularly promising activity will inform Aim 1 to select a second set of species from such clades for more extensive sampling. Extensive sampling in a specific clade increases the detectable phylogenetic signal within the clade and improves the expression knowledge of genes in that species. Second, information about the class of compounds expected to enrich in the active partition will allow a more focused analysis in network analysis Aim 2, for example by including network interaction terms from literature mining approaches [Krallinger et. Al. *Nucl. Acids Res., Web Server Issue*, 2009].

***Bioassay-directed fractionation and purification***

Every step of extraction, partition and chromatographic separation will be assessed for bioactivity to validate the specificity of the leads generated in Aim 2. This will involve the most relevant for each plant of either agar-based or liquid format antibacterial and/or antifungal activity. Chromatography proceeding through 2-3 successive steps of medium-resolution any of normal phase, reverse phase, gel filtration, ion exchange and the like will be followed up with a polishing step by preparative high performance liquid chromatography.

***Structural elucidation***

Several types of standard analytical methods (UV, IR, MS, NMR and 2-D methods for NMR) will be used in combination to a) characterize active materials down to the chemical class, and for the approximately ten most promising bioactivities down to complete structure. Data will here too be compared with in-house materials and that from commercial databases. NPDI has access to excellent spectroscopic facilities at Bristol Biotech Center and at Medina and the NPDI scientists have decades of experience in the structural elucidation of unknowns.

**Aim 3. Expected Outcomes and Alternate approaches.**

The Natural product chemistry approaches and bioassays are standard, well-tested approaches used routinely at NPDI. The novel aspect of this is the “leads” identified by the phylogenomic approach described in Aims 1 and 2. The integration of data from Aims 2 and 3, will test as proof-of-principle whether and how the phylogenomic enhanced approach to natural products, enables researchers to 1. Take a directed approach to natural products discovery, and 2. Provide mode of action information that will be required to bring such drugs to market. Characterization of rare or silent pathways that synthesize therapeutic compounds will educate and enable synthetic biology approaches for their production and testing.

**Conclusion:** Novel classes of compounds that can cure infectious diseases is both a major and urgent necessity in human medicine and historically some of our major advances in this field have come from plant-derived natural products. Unfortunately, despite the extended adoption of crude plant extracts in native communities to cure these very diseases the identification and industrialization of the active ingredients has remained technically challenging. Our proposal aims squarely to take a very novel approach to solving these challenges and provide breakthroughs in the discovery of natural products. Even more excitingly our proposal aims to develop a method to simultaneously discover the biosynthetic pathway generating the active compound. Even if we rediscover already known compounds but can elucidate their biosynthetic pathways, synthetic biology approaches can be used to scale up the production of these medicines which would in turn provide a sufficient and cheap supply of the medicine without endangering the donor species. This proposal brings together the extensive and unique expertise in ethanobotany (NYBG), phylogenomic reconstruction (AMNH), systems biology and trait-to-gene association (NYU) with an equally solid background in Biochemistry (WSU) and natural product chemistry (NPDI) to create an excellent multidisciplinary team of researchers that are capable of creating exciting breakthroughs in the very important field of natural product pharmacology.

**Time-Line**

**Year 1:** Collection of samples from medicinal plants for RNA-Seq. Creation of RNA-Seq libraries and sequencing (Aim1).

**Year 2:** Completion of bulk sample collection from medicinal plants for chemical extractions and continued creation of RNA-Seq libraries (Aim1). RNA-Seq data analysis and advancement of phylogenetic tools, development of integrated pipeline for Phylogenetic and network analysis integration (Aim2)

**Year 3:** Generate pathway leads for Aim 3, conduct first stage of extractions and bioassays. Feedback from Aim 3 to focus reanalysis in Aim 2. Repeat analysis will generate further leads in Aim 3. Iterate and generate a list of clades of interest to sample more species.

**Year 4:** Finish collection of additional species, prepare RNA-Seq libraries and sequence them. Test relevant partitions from new species. Bioassay subpartition of active fractions.

**Year 5:** Deeper transcriptome profiling of 2-5 species of greatest interest. Detailed pathway construction from Network analysis and literature mining. Structural elucidation of active compounds.