**PROJECT SUMMARY** **ABI Sustaining:** “**VirtualPlant: A Software Platform For Systems Biology**".

**1. List of senior personnel**. (PI, co-PIs, Key collaborators) with home institution

**PI**: Gloria Coruzzi, NYU Center for Genomics & Systems Biology

**Co-PIs**: Dennis Shasha, NYU Courant Institute of Mathematical Sciences,

Manpreet Katari, NYU Biology, Center for Genomics & Systems Biology.

**Senior Personnel**: Kranthi Varala, NYU Center for Genomics & Systems Biology

**2. Intellectual merit of the proposed activity.**  Our vision is to enable Systems Biology studies across plant species by sustaining our VirtualPlant project. The VirtualPlant software platform ([www.virtualplant.org](http://www.virtualplant.org)) was developed under an NSF Arabidopsis 2010 Grant (DBI-0445666) and includes tools for data analysis, integration and visualization such as the Arabidopsis multinetwork data, [missing word(s) here] analysis, and other data integration tools [Katari et al., 2010]. VirtualPlant has played an integral part in enabling plant systems biology research in many plant labs around the world. Our recent expansion of VirtualPlant to include important crops, such as Rice and Maize, will allow the platform to reach a much wider community of plant biologists, and will, through comparisons to Arabidopsis, support translational research through this sustaining grant. For example, VirtualPlant can allow researchers to predict how an interacting network of genes/products in crop genomes will react *as a system* in response to an environmental change or genetic modifications. As demonstrated in the case study (Section 2), we are able to achieve this ambitious goal by using VirtualPlant for Maize by using data from a maize experiment, incorporating the Arabidopsis multinetwork, and using homology. The results of such analyses generate a set of testable hypotheses for translational studies in future experiments. Creation of a VirtualPlant database and interface for a new species is now semi-automated, allowing us to extend VirtualPlant to many more species such as *Medicago truncatula* and *Glycine max,* and other crop genomes that come on line in the near future. **We divide the work into three sections:**

**Section 1: Scientific value of Virtual Plant to enabling Systems Biology studies within the plant genomics community.** Virtual Plant enables seamless integration of data and tools into a single operating environment. The “Gene Cart” enables researchers to store results, enabling iterative cycles of analysis, a highlight of Systems Biology research. VP’s Biologist Friendly GUIs enable plant biologists, with no computer training to analyze their genomic data providing biological insights. Since its publication in Feb 2010, VirtualPlant and its tools have been cited more than 50 times. In Section 1, we demonstrate how the community is using VirtualPlant and highlight the diverse profile of our userbase. VP currently has 1,033 registered users from 42 different countries spanning academe and industry.

**Section 2. VirtualPlant: Enabling Translational Research**. Section 2 describes a case study of how a Maize researcher can use VirtualPlant to build a hypothesis regarding nitrogen responsive gene networks by transferring knowledge from Arabidopsis to Maize. With the advent of Next-generation sequencing, we expect an increase in the amount of experimental data available for additional crop plant species such as Maize, Rice, Soybean, etc. VirtualPlant facilitates such translational research from Arabidopsis and other well-studied species to new species.

**Section 3: VirtualPlant: Dissemination and User support/interaction**: To enable translational systems biology studies in crops, we will work closely with the crop plant communities, so that researchers can seamlessly access important databases like Gramene (Maize) and Soybase (Soy) through VirtualPlant. We will also provide access to VirtualPlant tools from the existing community websites such as Gramene, Soybase etc. Further, we will attend the annual meetings and conduct VP workshops to support and assess the needs of the crop user communities. Finally, video tutorials will be hosted on the VirtualPlant site to introduce new users to the interface and tools available. We will also continue to support our existing user base through the help section on our website and a mailing list where users can ask questions.

**Justification for ABI Goals**: Sustaining Virtual Plant will meet the ABI goals of supporting the analysis and dissemination of biological knowledge through informatics tools and resources that support transformative research in plant systems biology the biological community.

**3. Broader impacts of the proposed research.** This project is the result of a long-standing and highly successful collaboration between biologists at NYU and elsewhere, and computer scientists at NYU's Courant Institute of Mathematical Sciences. The Systems Biology tools resulting from this project are written in a species-independent manner so that they can empower biologists to use genomic data to predict a spectrum of gene networks in biology with broad applications to agriculture, the environment, and human health. In addition to scientific results, this collaboration extends to joint training of biologists and computer scientists in the field of Systems Biology.

**RESULTS FROM PRIOR NSF SUPPORT:** This NSF ABI Sustaining grant is most closely related to and leverages an NSF DBI Grant entitled “Conceptual Data Integration for the Virtual Plant” (NSF Database Activities: DBI-0445666). The VirtualPlant software platform (www.virtualplant.org) [Katari 2010] developed in that parent grant integrates genome-wide data from Arabidopsis concerning the known and predicted relationships among genes, proteins and molecules, as well as genome-scale experimental measurements. VirtualPlant also provides tools that render multivariate information into integrated visual displays including networks, directed acyclic graphs, and other novel visualization formats to highlight biological implications. Tools embodied in the VirtualPlant system have been used by us and other notable Arabidopsis research teams to generate hypotheses in Arabidopsis that were subsequently experimentally validated [Gutierrez 2007, 2008, Gifford 2008, Brady 2011, Wildermuth 2010, Ford 2010, Chandran 2010, Chandran 2009].

M.C. **Wildermuth** (2010) Modulation of host nuclear ploidy: a common mechanism of obligate plant

biotrophs. ***Current Opinion in Plant Biology***13: 449-58.

Ford, K.A., Casida, J.E., Chandran, D., Gulevich, A.G., Okrent, R.A., Durkin, K.A., Sarpong, R.,

Bunnelle, E.M.,  and M.C. **Wildermuth** (2010) Neonicotinoid insecticides induce salicylate-

associated plant defense responses. ***Proceedings of the National Academy of Sciences***107:

17527-32.

Chandran, D., Inada, N., Hather, G., Kleindt, C.K. and M.C. Wildermuth (2010) Laser microdissection of

Arabidopsis cells at the powdery mildew infection site reveals site-specific processes and

regulators. ***Proceedings of the National Academy of Sciences***107: 460-5.

 Chandran, D., Tai, Y.C., Hather, G., Dewdney, J., Denoux, C. Burgess, D.G., Ausubel, F.M., Speed,

T.P., and M.C. Wildermuth (2009) Temporal global expression data reveals known and novel

salicylate-impacted processes and regulators mediating powdery mildew growth and reproduction

on Arabidopsis.***Plant Physiology***149: 1435-1451.

***Our parent NSF VirtualPlant grant had three major goals***: **Integration & Prediction**, **Visualization**, and **Synthesis**, which we have accomplished, as outlined below.

**Aim 1.** **Data** **Integration & Prediction**: ***The Arabidopsis Multinetwork*.** Our VirtualPlant project included assembling the first multinetwork for Arabidopsis, a first step towards a molecular wiring diagram of the plant cell [Gutierrez 2007 Genome Biology, Katari 2010]. The Arabidopsis multinetwork in VirtualPlant has 16,562 nodes (of which 13,960 are genes) and 97,423 interactions (Fig. XB, & Table X). The multinetwork enables researchers to interpret transcriptome data in the context of all known sources of interaction including protein, DNA, RNA, etc. The multinetwork has been used to integrate plant gene/protein interaction data from a number of sources including protein:DNA interaction data [Brady 2011], protein:protein data [Popescu], and more recently the Arabidopsis interactome data [Vidal & Ecker]. In one example, a query against the Arabidopsis multinetwork with 834 nitrogen-regulated genes resulted in a sub-network of 369 genes connected by one (or more) “expression correlation edges” [Gutierrez 2008]. At the top of the resulting list of network TF “hubs” (with 47 connections to targets in the N-regulatory network) was the central clock control gene CCA1, a Myb family transcription factor (TF) [Gutierrez 2008]. This discovery enabled us to derive and validated the novel hypothesis that nitrogen-regulation of CCA1 mRNA expression sets the circadian clock. Other examples of hypotheses derived and validated using the VirtualPlant multinetwork are reported in [Gutierrez 2007, 2008, Gifford 2008, Ruffel 2011].

**Aims 2 & 3. Synthesis and Visualization: *VirtualPlant’s primary analysis tools and functions.*** In addition to the Arabidopsis multinetwork, the VirtualPlant platform houses other Systems Biology tools for data analysis, integration and visualization. Below is a list of some of the most popular tools deployed through VirtualPlant.

**The "GeneCart" Function**: A key challenge to analyzing genomic data is the complex analysis workflow required by currently available software. VirtualPlant solves this problem by integrating multiple tools into a single platform that standardizes the representation of their inputs and outputs so that the output of almost any analysis can be stored in the “GeneCart” and later serve as the input to any VirtualPlant analysis tool. The GeneCart can store lists of genes, microarray and RNA-seq experiments, and gene networks. An item or items can be used as input for an analysis tool and the output of the tool can be stored in the GeneCart. This feature facilitates Systems Biology's iterative cycles of data analysis and experimentation [Ideker et al., 2001; Gutierrez et al., 2005]. Three working examples described in [Katari et al 2010], illustrate how VirtualPlant can be used to perform iterative data analyses that build and refine testable biological hypotheses.

**Set Operations.** VirtualPlant supports the set operations union, intersect, and symmetric difference. In addition to the combinatorial operations, sets can also be renamed, deleted, or assigned to a “group”. Groups, functioning like folders in a file system, can be arbitrarily deeply nested and are used to organize gene sets. This is fundamental but important feature as one frequent user writes: “*It is surprisingly difficult to get a simple output of the intersection of two gene lists using other available resources*, *VirtualPlant rapidly allows one to find intersections, combine lists (removing redundant genes), and store the output for further analysis.”(M. Wildermuth, UC Berkeley)*

**BioMaps (Fig. X)**: BioMaps takes one or more sets of genes and determines which functional terms (GO or MIPS are statistically over-represented in each set, with respect to a background population (e.g. Arabidopsis genome). The output is presented in either a tabular format that can be downloaded to Microsoft Excel or in an interactive graphical representation of the GO tree and the significant terms. Users can zoom the GO directed acyclic graph to any level of detail, and select GO terms. They can also query the GO terms and use union, intersect, and difference set operations to form new genelists and send them to the cart. Also, they can adjust a slider that sets a p-value threshold to which hides GO terms whose p-values exceed the threshold. A png image of the biomaps DAG can be saved.

**Sungear**: Sungear is a visually interactive and biologist-driven exploration of comparisons of the results of many experiments on a genomic scale. Sungear can represent an arbitrary number of experiments/lists, all of their disjoint intersections, and their related ontological terms. The position of a circle and arrows emanating from it, indicate the input lists of which it is a subset. The size of a circle is proportional to the number of genes in the intersection of those lists (see [Poultney]). Many biologists find Sungear to be an extremely powerful and interactive tool for analyzing the interrelationships between sets of genes [Gutierrez 2007- Exp Bot with Crawford].

**GeneSect:** Genesect is a non-parametric randomization test to determine whether the overlap between two gene lists is higher/lower than expected by chance. The test consists of 1,000 tests, each of which randomly and uniformly selects two lists, out of the user-specified background population, where the lists are the same size as the observed lists. Thus, the algorithm counts the number (n) of times that the intersection size of the random lists is equal or greater/less than the intersection observed for the two tested gene lists. A p-value is thus generated equal to n/1,000.

**Microarray and RNA-seq Analysis Tools:** Transcriptome analysis is an important part of systems biology today. One of the main advantages of RNA-seq is that it does not require knowledge of the reference genome. For this purpose we expect a higher influx of transcriptome data in the near future. To accommodate transcriptome data we allow users to upload both their microarray data (CEL files or normalized matrix) and RNA-seq, data represented by a matrix containing the gene name and the number of reads that mapped to the genes. For microarray analysis we have a set of common statistical methods for determining probes that are **differentially expressed genes** including, T-test, Log Ratio Cutoff, RankProduct, and ANOVA. For RNA-seq, we use Deseq to determine genes that are differentially expressed. In all cases the result is added to the user’s gene cart as gene lists.

**Gene Network Analysis Tools: Gene Network** analysis allows users to query our Multinetwork data and displays the results in a graph using Cytoscape, an open source project that we have built upon. The output of the analysis is a subgraph of the Multinetwork containing only the genes that are in the query list. The tool also allows users to include and exclude specific interactions before displaying the graph [Katari et al. 2010, Gutierrez et al., 2007]. **Supernode Network** analysis can be used to help summarize the results of a Gene Network analysis. The genes in the gene network are grouped into “Supernodes” based on their functional annotation and they are associated with other Supernodes with edges determined from the Gene Network data. A gene can be present in more than one Supernode. The number of genes it contains determines the size of the Supernode. VirtualPlant also allows calculates simple **network statistics** such as gene “hubbiness”, which helps suggest candidates for future studies.

**Cross-Species Analysis Tools:** There are many methods and databases that can be used to identify homologous genes from one species to another. **Homology Mapping** allows Biologists to select the method or database they prefer to map their gene list of interest to another species. Biologists can also use the tool to make Gene Networks as described in the case study below. Alternatively, Biologists can align their sequence of interest to one of the fully sequenced genomes using **BLAST** to identify the homologs.The BLAST results can be added to the gene cart as a gene list.

**VirtualPlant Flexibility**: VirtualPlant is modular, not only because it can support different types of data and tools, but also because VirtualPlant allows different species to be added to the VirtualPlant database. Resources such as PlantGDB and Phytozome provide data for different species in a specific format. We have created functions parse this format and can automatically create a VirtualPlant database for a species hosted by Phytozome.

The VirtualPlant database contains gene annotations and many common ontology terms such as Gene Ontology and Plant Ontology. The Arabidopsis multinetwork includes metabolic pathways from KEGG and ARACYC, protein-protein interactions from BIND and Interolog databases, (see Table X for a complete listing of data sources). The database also contains publicly available Microarray experiments obtained from NASC and GEO.

**Software and Data Availability**: VirtualPlant is accessible via the website www.virtualplant.org. Registered users (currently > 1,000 users) store their data sets and use many tools to analyze their genomic data. The website does not require a password and free when used for non-for-profit purposes.

**Publications: Peer reviewed journal articles, chapters and books.**

**VirtualPlant: Tool development for Plant Systems Biology**

Katari M, Nowicki S, Aceituno F, Nero D, Kelfer J, Thompson L, Cabello J, Davidson R,

Goldberg A, Shasha D, Coruzzi G, Gutierrez R (2010) "VirtualPlant: A software platform

to support Systems Biology research. ***Plant Physiol***. Feb;152(2):500-15

Nero D, Kelfer J, Katari M, Tranchina D, Coruzzi G (2009) In Silico Evaluation of Predicted

Regulatory Interactions in Arabidopsis thaliana. ***BMC Bioinformatics***. Dec 21;10(1):435.

Poultney C, Gutierrez R, Katari M, Gifford M, Paley W, Coruzzi G and Shasha D (2007) "Sungear:Interactive visualization, exploration & functional analysis of genomic

datasets". ***Bioinformatics***, 23:259-61.

Ferro A, Giugno R, Pigola G, Pulvirenti A, Skripin D, Bader G, Shasha D, "NetMatch:

a Cytoscape Plugin for Searching Biological Networks" ***Bioinformatics***,

2007 23(7):910-912.

**Applications of VirtualPlant: Hypothesis Generation and Testing**

Krouk G, Tranchina D, Lejay L, Cruikshank A, Shasha D, Coruzzi G, and Gutierrez R (2009) “A

systems approach uncovers restrictions for signal interactions regulating genome-wide responses to nutritional cues in Arabidopsis.” ***PloS Comp Biol***. Mar;5(3):e1000326. (Highly Accessed).

Gutierrez R, Stokes T, Thum K, Xu X, Obertello M, Katari M, Tanurdzic M, Dean A, Nero D,

McClung R & Coruzzi G (2008) "Systems approach identifies an organic nitrogen-

responsive gene network that is regulated by the master clock control gene CCA1"

***Proc. Natl Acad Sci USA*** 105, 4939-4944. (Faculty of 1000 recommended: Factor 3)

Gutierrez R, Gifford M, Poultney C, Wang R, Shasha D, Coruzzi G, Crawford N (2007) "Insights

into the genomic nitrate response using genetics and the Sungear Software System"

***Journal of Experimental Botany*** doi: 10.1093/jxb/erm079

Gutierrez R, Lejay L, Chiaromonte F, Shasha D, Coruzzi G (2007) "Qualitative network models

and genome-wide expression data define carbon/nitrogen-responsive biomodules in

Arabidopsis" ***Genome Biology***, 8: R7. Faculty 1000 (Must Read: Factor 6)

**Plant Systems Biology: Reviews, Books and Outreach**

Ruffel S, Krouk G, Coruzzi G (2009). "A Systems View of Responses to Nutritional Cues in

Arabidopsis: Towards a Paradigm Shift for Predictive Network Modeling”.

***Plant Physiol***. Nov 25 [epub ahead of print]

Gutierrez R, Coruzzi G., Eds (2009) Book: “***Plant Systems Biology***”, Annual Plant Reviews;

Blackwell Publishing: Oxford, UK, 2009, Vol. 35. 360 pages.

Coruzzi GM, Burga A, Katari MS, and Gutierrez RA (2009) “Systems Biology: Principles and

Applications in Plant Research”. In ***Plant Systems Biology***, Annual Plant Reviews;

Blackwell Publishing: Oxford, UK, 2009, Vol. 35. Pgs 3-31. Book Chapter.

Gifford M, Gutierrez R, and Coruzzi G (2006) "Modeling the Virtual Plant: A Systems Approach

to Nitrogen-Regulatory Gene Networks". Essay 12.2 Chapter 12. Assimilation of mineral

nutrients;***In: A companion to Plant Physiology,*** Fourth Edition, Lincoln Taiz and

Eduardo Zeiger, http://4e.plantphys.net/article.php?ch=12&id=352

Gutierrez R, Shasha D and Coruzzi G. (2005) "Systems Biology for the Virtual Plant". ***Plant***

***Physiol.*** Vol 138, pp 550-554.

**EDUCATION AND TRAINING**: The development of the Systems Biology tools and the Virtual Plant software platform has trained undergraduates, MS and PhD students in Systems Biology. Students trained include **Undergraduates**: Steve Nowicki (NYU CAS), Varuni Prabhakar (Barnard College), Rebecca Davidson (BS Computer Science); **Masters Students**: Ana F. Arroja (MS student, NYU Courant), Ranjita Iyer (MS Computer Science) Jonathan Kelfer (MS Computer Science), Lee Parnell (MS Computer Science), (Jarod Wang, MS Computer Science); **PhD Students**: Chris Poultney (PhD student, NYU Courant), Jason Reisman (PhD student, NYU Courant), Saurabh Kumar (PhD student, NYU Courant). These students have gone on to PhD programs (Prabhakar, Parnell) as well as to industry (Kelfer, Bloomberg).

-----------------------------------------------------------------------------------------------------------------

**RESEARCH PLAN:** With the advent of Next-Gen sequencing, Systems Biology approaches and “network thinking” will become increasingly important especially across species. In addition, crop species should be able to build on the “network knowledge” from model species such as Arabidopsis, to enable translational research. The VirtualPlant software platform/integration environment supports Systems Biology studies across multiple species and allows integration of knowledge across experiments, data types and perhaps most importantly across species.

**Comparison to existing tools**. Besides the tools it provides, VirtualPlant’s two main unique features are i) Gene Cart and ii) the ability to build a network from scratch. The Gene Cart supports iterative exploration of data sets by allowing a researcher to apply a tool, return the result to the Gene Cart, then apply a different tool on the result and this can keep continue repeatedly. [Example would help.] The Gene Cart can hold gene lists, experiments, and networks. The ability to build a network from scratch is important, because biological interactions are context-specific. Depending on the developmental stage and environmental factors, only certain interactions will be observed. For example, as we show later, starting from nitrogen experiments in Maize and Arabidopsis, we create a synthesized network in Maize that …. The incorporation of experimental data with static interaction data gives us a network that better models the molecular circuitry underlying the observed response as compared to a static network of interactions alone. (Gloria does not understand this explanation. If I have microarray data, I should be able to make networks with the other tools too) In Table X, we compare plant network analysis tools and databases that perform network analysis. VirtualPlant provides complementary tools to those of other systems.

**Table X**. **Comparison of network analysis platforms complementary to VirtualPlant**. Most platforms listed feature a pre-constructed network that a user can query with one or more genes. PlaNet [Mutwil et al The Plant Cell March 2011], GeneMania [Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q (2008) Genome Biology 9: S4], and VP host data from multiple species, but VP allows network merges across species through homology mapping. VirtualPlant is centered on user-generated data and is the only tool that can be used to build a context specific network (Build Your Own Network) from a user-selected set of experiments. [dennis asks: are we absolutely sure about this?]

**Section 1: Scientific value of Virtual Plant to enabling Systems Biology studies across the plant genomics community. *Rationale***: In this aim, we discuss the prior and anticipated impact of VirtualPlant on enabling Systems Biology research within the plant genomics community.

**1.** **VirtualPlant (VP) Users**. VirtualPlant was specifically designed to enable plant biologists with no computer expertise to analyze their own data - enabling valuable insights from a plant biologist perspective. This user base includes researchers from academe and industry. VirtualPlant keeps track of “function calls” and “user activity” to better understand which features are most useful to our user base. From these statistics we can also determine how many users are actively using VirtualPlant. As of July 2012, there are 1,033 registered VirtualPlant accounts. In Fig. X, we present VP user demographics as revealed by the registered email addresses. Among the geographically discernible domains, the largest group of users are from .edu domain, which generally represents accounts from users based in the US academia. Following the US (.com and .edu), uk (United Kingdom), fr (France), de (Germany), cl (Chile), and au (Australia) are the most frequent users. Registration is not mandatory to access VirtualPlant tools, and 44% of function calls are made by anonymous users. However only registered users can save their data in their “Gene Cart” and analyze their experiments. For this reason, registered email addresses represent the more active, recurrent VirtualPlant users. The .com domain, in addition to personal accounts of unknown users, includes accounts from companies such as Monsanto and Syngenta (data not shown to preserve privacy). Within the United States, researchers from numerous prestigious universities currently use VirtualPlant (Fig. X). In addition to the host institution, New York University, the largest numbers of VP users in academe are from University of California at Davis, Stanford University, Duke University, Iowa State University, and other first-rate universities. [The number of new users is actually going down ☹ ]

**2. VirtualPlant Usage.** VirtualPlant includes a suite of tools that are useful to analyze –omics datasets. Fig. X shows a barplot of the different tools and the frequency of their usage by a registered user. Since some of the tools are recent additions , only activity since the start of 2012 is shown. Here we see that BioMaps (gene list enrichment analysis of GO and MIPS terms) and set operations (intersection, union of genesets) are the most highly accessed features of VirtualPlant. These popular tools are followed by the number of times registered users have logged in to use “Sungear” (an interactive java application for comparing multiple gene sets) [Poultney et al ] or “GeneSect” (statistical evaluation of the overlap in gene sets). These results demonstrate that the most common tools in genomic research today are the ability to compare gene lists, most likely derived from different conditions, experiments or literature, and enrichment analysis such as GO or MIPS terms. These simple operations comparing gene lists allow the biologist to qualitatively and quantitately measure the effect of a treatment or a gene knockout. Another observation based on tool usage is the “Network Analysis” feature in VirtualPlant is used more often compared to “Microarray Processing”, implying that VirtualPlant is very valuable as a post-microarray processing analysis tool to investigate relationships between genes and to derive hypothesis from networks formed. [Redundant- which section is this better in?] Importantly, VP is the only available tool that enables researchers to generate networks “on-the-fly”. Biological networks are not static, they change in reponse to environmental and developmental signals. VirtualPlant allows Biologists to observe interactions in a specific condition by allowing users to create their own network, for example a correlation network based only on Nitrogen treated experiments.

**3. VirtualPlant Citations.** VirtualPlant and its tools have been cited or referenced in 50 publications since June 2006, when it was made public during the International Conference of Arabidopsis Research (ICAR). Fig. X shows a linear trend representing an increase in the number of publications that cite VirtualPlant, thus representing the growing impact VirtualPlant is having on the plant community since its release. Data for the graph was collected from Google Scholar. All publications that cite VirtualPlant [Katari et al 2010], Sungear [Poultney et al 2007] or simply use “www.virtualplant.org” in their text were counted. A closer look identified which tool on VirtualPlant was used in the analysis, and this information is represented in the color codes shown in Fig. X. A testimontial from the most frequent user not associated with our lab: “My laboratory and I are active and regular users of VirtualPlant, with many periods where we use it daily. Its ease of use, inclusion of appropriate statistics in analyses, functionality, and versatility are exceptional and often unique.  I have come to depend on VirtualPlant and its importance to my laboratory's research is evidenced by its use in 4 published papers (below) and two papers under review.  **Mary Wildermuth, UC Berkeley (2nd top user of VirtualPlant)**

REFS-

[M.C. **Wildermuth** (2010) Modulation of host nuclear ploidy: a common mechanism of obligate plant

biotrophs. ***Current Opinion in Plant Biology***13: 449-58.]

[Ford, K.A., Casida, J.E., Chandran, D., Gulevich, A.G., Okrent, R.A., Durkin, K.A., Sarpong, R.,

Bunnelle, E.M.,  and M.C. **Wildermuth** (2010) Neonicotinoid insecticides induce salicylate-

associated plant defense responses. ***Proceedings of the National Academy of Sciences***107:

17527-32.]

[Chandran, D., Inada, N., Hather, G., Kleindt, C.K. and M.C. Wildermuth (2010) Laser microdissection of Arabidopsis cells at the powdery mildew infection site reveals site-specific processes and

regulators. ***Proceedings of the National Academy of Sciences***107: 460-5.]

 [Chandran, D., Tai, Y.C., Hather, G., Dewdney, J., Denoux, C. Burgess, D.G., Ausubel, F.M., Speed,

T.P., and M.C. Wildermuth (2009) Temporal global expression data reveals known and novel

salicylate-impacted processes and regulators mediating powdery mildew growth and reproduction

on Arabidopsis.***Plant Physiology***149: 1435-1451].

**Expected Scientific Value for the Future**: With the inclusion of more species, we expect VirtualPlant to have an even broader impact to the plant community. Tools created in VirtualPlant are being designed to be modular such that they can be used independently. The advantage of this approach is that it makes it easy for us to share and make the popular tools available in other forms. For example, we recently worked with iPlant to create a stand-alone version of BioMaps to be integrated into the iPlant Discovery Environment where the tools will be available to all Discovery Environment users. The second advantage of the modular nature of the VirtualPlant platform is the separation of “Microarray Processing” from “Network Analysis”, which has enabled VP to grow and thrive as new forms of expression data evolve (e.g. RNA-Seq), and will enable it to support future changes in data types.

**Section 2: VirtualPlant: Enabling Translational Research. *Rationale:*** Agronomically important plant species have now been inducted into VirtualPlant including Rice and Maize. While integrating this data, we have created new ways to support plant biologists to perform cross-species comparisons. Such comparisons are critical to leverage information available in some of the model plant species, such as *Arabidopsis thaliana* and *Oryza sativa*. To demonstrate how these new databases and tools can be used to translate knowledge from one species to another, especially to leverage Arabidopsis data for translational studies in crops, we perform a case study using a publicly available microarray dataset in Maize. We selected the dataset from Yang et al. 2011 [Yang XS, Wu J, Ziegler TE, Yang X et al. *Plant Physiol* 2011 Dec;157(4):1841-52] that discovered sensitive gene expression biomarkers for the *in planta* nitrogen status of maize.  To illustrate the cross-species functionality, we present a concise step-by-step walkthrough of a VirtualPlant analysis for this Maize dataset and its comparison to Arabidopsis.

**Case study of Nitrogen-Response across species. (GLORIA IS NOT REALLY ABLE TO EDIT THIS SECTION….KRANTHI PLEASE CLARIFY….MAYBE GLORIA WILL HAVE TO REWRITE WITH MANNY ON SATURDAY)**

1. **The Maize Data:** The *Zea mays* VirtualPlant database was created using two different publicly available sources (Phytozome v8.0 [http://www.phytozome.net] and Maize Sequence [http://maizesequence.org]). The Maize microarray dataset used in this study contains a total of 90 samples from nitrogen-treated plants [Yang 2011]. The experimental design of this study includes metadata such as time-of-day, genotype, and nitrogen-treatment, allowing for the investigation of multiple questions. For the purposes of this case study, we are interested in genes that are differentially expressed in response to nitrogen-treatment at two different time points.
2. **The Maize Correlation Network:** The maize microarray data was normalized using the RMA [Wu et al Journal of American statistical association 99:468 2004] method, and then a 2-way ANOVA analysis was performed to identify genes that are differentially expressed genes by either a Nitrogen effect, a Time effect, or the interaction of Nitrogen and Time. Probes with an FDR adjusted p-value < 0.05 were considered to be differentially expressed. These probes were then mapped to Maize annotated genes based on BLAST alignments of probe sequences to the Maize annotated transcripts. [sounds like development: Currently VirtualPlant performs the analysis at the gene locus level, but in the future we will be able to support distinct transcripts from a single locus, as such information can be obtained from Next-gen sequences.] We then calculated pairwise gene correlation using the Pearson method and all gene pair correlation values with a p-value < 0.05 were used to create a maize correlation network.
3. **The Arabidopsis Data:** The *Arabidopsis thaliana* VirtualPlant database has been created using gene interaction data from many different sources (See Table X). The Arabidopsis microarray dataset we selected to compare with the Maize N-treatment dataset is from our highly accessed Arabidopsis study which identified N-regulatory networks that were subsequently validated [Gutierrez 2008].
4. (Gloria is lost- and cannot understand this section)…..Kranthi….can you clarify?????

**A Comparison of Maize to the Arabidopsis Network:** The 863 N-regulated genes in Arabidopsis [Gutierrez 2008] were used to create an Arabidopsis correlation network. The same set of genes were used to query the Arabidopsis MultiNetwork, which includes multiple edge types (Table X), to create a sub-network containing only the 863 N-regulated genes. An intersection of these two networks created a network with 613 genes by dropping genes whose edges were not conserved across both. This Arabidopsis N-regulated network was compared to the Maize correlation network (from step 2), based on a homology map. We find that 42 interactions are conserved across the two species. These interactions are therefore present in the static Arabidopsis MultiNetwork, the Arabidopsis N-regulated network [Gutierrez 2008], and in the N-treated Maize correlation network [Yang 2011].

1. **Literature Validation:** The highly conserved network identified above can be characterized based on the Arabidopsis homologs. This fairly small list contains genes that known to be involved in circadian rhythm (CCA1 and LHY) and nitrogen assimilation (Nitrate Transporter 1.5 and Nitrate reductase) [dennis thinks we should say if genes of unknown function are also found]. These results show that comparative network analysis is beneficial for crop research. By intersecting the two different datasets, we are able to identify the most important players of nitrogen regulation and it is encouraging to note that the role of these genes in Nitrogen assimilation has already been experimentally validated [Gutierrez 2008]. Conserved networks are analogous to conserved sequences – the more distant the species, the less conservation one expects to find. However the conserved networks are very likely to have a function.

A biological network is a representation of the molecular interactions that exist in the cell. Biological networks are highly responsive to signals from within the organism and from the environment. Thus, to best model the relevant Biological network, VirtualPlant allows the Biologists to create a network based on the conditions that they are interested in. The overlap between the Maize and the Arabidopsis Nitrogen networks may be small but the interactions represent conserved Nitrogen and circadian regulatory interactions, which is one of the questions the Maize experimental dataset was designed to discover.

**Expected outcomes of Aim 2:** The case study demonstrates how VirtualPlant enables Systems Biology research in crop species, and can be augmented by what is already known in model plant species. The availability of such a tool will empower not only academic reearchers, but also the agricultural companies in identifying master regulator genes of plant response to stress or infection, and also genes that can be used as biomarkers to test a given crop’s status. VirtualPlant will be kept updated by including more and more species as their genomes and annotations reach maturation and as more global expression assays are developed for each of the newer species. Examples of crop genomes that will be included during this sustaining grant are: Glycine max, Medicago, Sorghum, etc. Our goal is to ultimately reach a fair degree of automation in this process by relying on standards of genome and annotation data that are evolving at data-generating centers such as JGI and plantGDB.

**Section 3: VirtualPlant: Dissemination and User support/interaction:** The VirtualPlant platform has been disseminated to the Arabidopsis community on numerous occasions including its first reveal at the ICAR (International Conference on Arabidopsis Research) meeting in 2006 Please see below a list of presentation about VirtualPlant made by PI and Co-PIs. With the introduction of Corn and Soybean into VirtualPlant, we intend to inform and involve these new plant communities over the course of this renewal.

Workshops: An interactive workshop to introduce VirtualPlant and publicize its features will be held at the upcoming Maize genetics conferences (2013-2015) and at the Molecular & Cellular Biology of the Soybean Conference 2014. These brief (30 minutes- 1 hour) workshops, will be led by the Co-PI Manpreet Katari and aim to provide Corn and Soybean geneticists with a hands-on experience of uploading data and analyzing test data (or their own experimental data) in the VirtualPlant environment. In addition to helping advertise the VirtualPlant platform, these sessions will give us crucial feedback on the usefulness of VirtualPlant features and allow us to make support decisions.

Seminars/Webinars: VirtualPlant and its tools are regularly presented at major plant biology conferences such as ICAR. In addition, open webinars followed, by discussion sessions, will be held annually to better communicate with and support the users.

Support and feedback: Continued support is provided to the users through the help section of the VirtualPlant website and a VP users mailing list (virtualplant-discuss@googlegroups.com). The VirtualPlant site will soon feature a video tutorial to give new users an overview of the platform’s features and user interface. The help section of VirtualPlant also contains a feature request page through which our users can submit requests to add new features or modify existing ones. In the past, we have added many tools, such as the network view of BioMaps and GeneSect, as a response to user needs. In its current form VirtualPlant is a powerful and robust platform, but we anticipate that small changes to existing tools or modifications to make them work for the newer species will be needed and we intend to address such issues in a timely manner.

**Seminar presentations on VirtualPlant have been disseminated via the following:**

Coruzzi (PI):

CSHL Plant Genome Course: Plant Systems Biology, July 2010

NSF US-EU Taskforce on Plant Biotechnology: Speaker, June 2, 2010

CSHL Plant Genomes: Genes, Networks and Applications, March 4-7, 2009.

19th International Conference on Arabidopsis Research, Montreal, July 23-27, 2008 (Session Chair and Plenary speaker: Systems Biology Plenary Session).

Society for Experimental Biology (UK) Symposium on Systems Biology the Society.

Marseille, France July 7-9, 2008 (Keynote speaker).

NSF iPlant Meeting, CSHL, May 2008 (Session moderator; Systems Biology).

6th Annual Keen Lecture, UC Riverside, Genome Center Jan 18, 2008.

Keystone Symposium, Systems Biology and Regulatory Networks, Mar 22-27, 2007, Steamboat Springs (Plenary Speaker).

CSHL 5th meeting on Systems Biology: Global Regulation of Gene Expression March 28 - April 1, 2007. (Invited Speaker).

**“Systems biology approaches to analysis of metabolic and regulatory networks of** Arabidopsis”at the 17th International Conference on Arabidopsis Research 2006. Madison, WI.

ISPMB Meeting; Plenary Speaker, Systems Biology, Adelaide, Australia Aug. 20-25, 2006

Society for Developmental Biology, Ann Arbor MI, June 17-19, 2006 (Plenary Speaker).

Systems Biology Symposium, Plant Biotech Denmark, Nov 2005 (Plenary Speaker).

2nd Tri-National Arabidopsis Meeting: Neuchåtel, Switzerland, Aug 24-27, 2005.

Annual ASPB Meeting, New Approaches for Integrating Plant Genomes & Function. Seattle, WA, July 2005 (Plenary Speaker).

CSHL Arabidopsis Genome Course, July 2005 (Lecturer).

Frontiers in Plant Biology: Genomics & Beyond: Missouri Symposium, April 27-30, 2005.

Katari (Project Director):

Systems Biology Workshop at Cold Spring Harbor Laboratories

Poster Presentation at New York University Tech Expo

A Systems Approach to Nitrogen Regulatory Networks and the VirtualPlant. Network Biology, Hinxton, UK. August 2008.

VirtualPlant: A Software platform to support Systems Biology research in the post genomic era.New York Area Plant Molecular Biology, Adelphi University, New York, June 2008.

Nitrogen Regulatory networks and plant systems biology. Department of Energy GTL Systems Biology Network/Knowledgebase Workshop, Bethesda MD, May 2008.

iPLANT Collaborative Inaugural Symposium, Cold Spring Harbor Laboratory, NY, May 2008.

18th International Conference on Arabidopsis Research 2007. Beijing, China. (Workshop presentations).

17th International Conference on Arabidopsis Research 2006. Madison, WI. (Session and workshop presentations).

**Impacting a broader scientific community:**  In order make our tools available to a broader systems biology community we are collaborating with developers of other software systems to allow us to share our tools. For example we have submitted our most popular tool (BioMaps) to iPlant Discovery Environment and are awaiting approval. We are also working with developers of Gaggle to integrate VirtualPlant and its tools with the Gaggle system, which will expose VirtualPlant to systems biologists who do not necessarily work on plants, thus making an impact on a broader scientific community.

**Timeline:**

**Year 1**: Provide FREE online Webinars, Organize VirtualCorn workshop at the 2013 Maize genetics meeting, Launch VirtualSoy and VirtualMedicago. Update data for all supported species.

**Year 2: Provide more FREE online Webinars. Organize VirtualSoy workshop at the 15th Biennial Soybean Conference. Launch VirtualGrape and VirtualPopulus. Update data for all supported species.**

**Year 3**: Provide Free online Webinars, Organize a VirtualPlant workshop at 2015 Plant and Animal Genomes meeting. Update data for all supported species.

**PLAN TO INTEGRATE RESEARCH AND EDUCATION**: (IS THIS PART OF 15 pages?)

**Cross Training of Biologists and Computer Scientist in Systems Biology**. The development of Systems Biology tools in this project has and will involve biologists teaching computer scientists about topics like genetics, experimental genomics, and the computational challenges of analyzing genomic data. We do this informally at our weekly joint lab meetings at which graduate students and post docs from NYU Biology and NYU Courant each present their work to the group. This project involves a resident full-time systems administrator (Roberto Jimeno) working within a Biology lab, interacting closely with wet-bench biologists. The PI computer scientists (Shasha and Katari), are also involved in training and engaging computer scientist students at all levels in the emerging field of Systems Biology. In the last year, they have trained two PhD students, two interns and two MS students from Courant working in this environment. For a complete listing of students trained in the past 4.5 years, see Education and Training section in Results from Prior support.

**Workshops and Classroom Training in Genomics and Systems Biology**: We also provide formal training in the form of workshops and classes to enable Systems Biology. Examples of this include a once-a-year “R-Boot camp”, which involves a weekly software workshop in “R”, which aims to teach biologists how to analyze their own genomic data. A workshop on Virtual Plant has been taught two times, once by Jonathan Kelfer, a MS student working on the project and most recently by Manpreet Katari, co-PI. Students have included several faculty on sabbatical at NYU including most recently: Mary Lou Guerinot and Rob McClung of Dartmouth. Students will be exposed to Genomics and Systems Biology also through a series of formal courses offered by faculty at NYU’s Center for Genomics and Systems Biology including: BIOL-GA-1128 Systems Biology; BIOL-GA-1130 Applied Genomics: Introduction to Bioinformatics & Network Modeling; BIOL-GA-1127 Bioinformatics & Genomes; BIOL-GA-2030 Statistics in Biology; BIOL-GA-1007 Bioinformatics for Biologists; BIOL-GA-1009 Biological databases and datamining; BIOL-UA-103 Bioinformatics in Medicine and Biology; BIOL-UA-38 – Introduction to Genomics and Bioinformatics. PhD students have and will continue to present their work in the weekly PhD seminar series hosted by the Biology Department. Computational students will be involved in constructing the pipeline and making it perform through the use of parallelization. Such students will also help to develop and test optimization and machine learning algorithms for network inference.

**PLAN TO INTEGRATE DIVERSITY**: We are committed to training scientists at the graduate and postdoctoral levels who can do independent research that cuts across fields and expertise in genomics.  Our research team is also committed to diversity.  Researchers in our current and previous NSF grants included Hispanic and African-American scientists.  We will continue to actively seek out and recruit scientists from under-represented minorities to participate in our research in our continuing commitment to increase diversity in our research program. Five female scientists are associated with this project: Coruzzi (co-PI); Rebecca Davidson (Programmer); Varuni Prabhakar (UG Programmer); Ana Arroja (MS); Ranjita Iyer (MS Courant) Stuti Srivastava (MS Biology). Damion Nero a minority, recently graduated PhD student, has written programs contributing to the Virtual Plant project. Roberto Jimeno (Systems Admin) associated with this project is of Hispanic origin.

**SHARING OF RESULTS**:

**Publications:** The results of our analysis of the data we generate will be made available through peer- reviewed literature as it is the most appropriate way to make this information available.

**MANAGEMENT PLAN**: To coordinate and facilitate interactions between individuals, Gloria Coruzzi, the PI (NYU Biology) will also serve as the overall Project Manager. Dennis Shasha (NYU Courant) will serve as a computationl advisor . The role of the Project Manager is to oversee the daily operations of the project and ensure that the needs and concerns of the participants are addressed on a day-to-day basis between the participants involved. We will also schedule day-long meetings twice a semester with our collaborator (Rodrigo Gutierrez, Chile), to do evaluation of work status and long-term planning.

**Bioinformatics manager: Dr.** **Manpreet Katari** (NYU Biology) will be in charge of the bioinformatics data and software. To enable efficient information exchange of raw and processed data, a file server has been set up at the NYU to store and distribute data and its analysis among users at NYU Biology and NYU Courant. This will be maintained by **Roberto Jimeno**, the Systems Administrator for this project, who will also maintain the web server, database server, and update the multinetwork databases. Kranthi Varala, a bioinformatian with a vast experience in crop genomics, will be the liaison for communicating with the different plant communities.

**Principal Investigators:** Shasha and Coruzzi will each supervise personnel, organization, intellectual developments and contributions.

**Role of Participants:**

|  |  |  |
| --- | --- | --- |
| **Name** | **Institution** | **Role** |
| ***Gloria Coruzzi***-PI | NYU Biology | Project Leader: Biological |
| ***Dennis Shasha***-Co-PI | NYU Courant | Co-leader: Computational |
| ***Manpreet Katari***-Co-PI | NYU Biology | Bioinformatics Manager |
| ***Kranthi Varala-*** *Senior Personnel* | NYU Biology | Bioinformatician |

**COORDINATION WITH OUTSIDE GROUPS:**

**Please see attached letter of collaboration:**

**Rodrigo Gutierrez (U Catolica, Chile)** Dr. Gutierrez, the creator of the Arabidopsis multinetwork (Gutierrez et al 2007) will assist in the assembly of multi-networks for crop species including Vitis (Grape), and Medicago, truncatula, and Zea mays.