Statement of Research Interests

My research activities have centered on genomics, bioinformatics and computational biology of plants and plant pathogens. A number of my projects are complementary allowing leveraging of resources between projects. While some projects are housed solely at MSU, a majority of my projects are collaborative with scientists at MSU, within the U.S. and throughout the world. This statement of research interests is focused on my accomplishments in the last 3 years although where appropriate, I have provided the historical context of my research since I began in genomics in 1999 at the Institute for Genomic Research.

Rice Genomics, Bioinformatics and Computational Biology

In 1999, I initiated a project to sequence a portion of the rice genome as a member of the International Rice Genome Sequencing Project (IRGSP), a consortium of laboratories throughout the world focused on sequencing the rice genome that culminated in a manuscript describing the map-based, finished rice genome sequence (The International Rice Genome Sequencing Project 2005). In parallel with my sequencing activities, I began a bioinformatics initiative to annotate and analyze the rice genome and to provide the community with resources for data-mining the rice genome. Software developed through our rice annotation project has been published and made publicly available [redacted] Our rice genome annotation project (currently funded by the NSF) has been highly successful in public dissemination of data, a critical component of any publicly funded resource project. Our web-based rice genome annotation data website is accessed nearly 1 million times each year [redacted]

I have also been involved in a complementary project to generate single nucleotide polymorphism (SNP) data from rice using the Perlegen hybridization-based resequencing technology. This project (PD [redacted] Colorado State University) involved SNP detection from 20 rice lines, phenotyping of these lines, and development of genetic resources from these lines. My group was instrumental in the bioinformatics component of this project [redacted] and currently provide access to the OryzaSNP diversity data through our website [redacted]

I have also pursued research activities in rice evolutionary biology, molecular biology, and comparative genomics. [redacted] we report on the evolution of intron loss in segmentally duplicated genes in which we observed that 95% of the introns were retained following segmental duplication in rice. [redacted] we analyzed alternative splicing in rice and observed that alternative splicing is not only widespread but also that a surprising number of alternative splice forms result in a significant change in coding sequence, suggesting a potential pathway for non-sense mediated decay of mRNAs in rice. With access to sequence data from 184 plant species, we investigated the conservation of predicted rice genes throughout the Plant Kingdom [redacted] Our analyses revealed support for nearly 90% of the genes in the rice genome in at least one other plant species. Although the majority of the putative homologues were obtained from Poaceae species, putative homologues were identified in dicotyledonous angiosperms, gymnosperms, and other plants such as algae, moss, and fern. We were able to identify improvements to our current rice gene structural annotation, demonstrate the utility of cross-species comparative alignments in the identification of non-coding sequences, and confirm gene nesting in rice. We have also characterized pseudogenes in the rice genome to aid in annotation and interpretation of the rice genome [redacted]

A major theme of research in my lab has been in the identification and characterization of lineage specific genes. This was initiated in rice [redacted] nd extended to Arabidopsis and the Brassicaceae [redacted] Using our computational pipeline, we were able to identify not only rice lineage specific genes, but also Poaceae specific genes. In 2009, I was successful in leveraging this into a USDA-funded project (in collaboration with [redacted] SU Horticulture) to determine the function of a subset of Poaceae-specific genes using functional genomics methods including expression profiling, over-expression, and knockout lines.

Solanaceae Genomics, Bioinformatics, and Computational Biology
My work on Solanaceae genomics, bioinformatics and computational biology has been funded since 1999 through multiple awards from the NSF and USDA. This work includes generation of primary sequence resources, fabrication of publicly available potato cDNA microarrays [redacted] sequence and analysis of resistance gene clusters, transposable elements [redacted] SNP discovery in potato and tomato, and generation of a public resource for mining Solanaceae genomic data [redacted]

A major activity in my lab for the last three years has been as a participant of the Potato Genome Sequencing Consortium [redacted] funded through the NSF. The initial stages of the PGSC employed a bacterial artificial chromosome (BAC) by BAC approach to sequence the 850 Mb potato genome [redacted] However, technical barriers due to the heterozygosity, coupled with improvements in next generation sequencing technology, made it feasible to sequence a doubled monoploid potato clone using a whole genome shotgun sequencing approach. My group has been instrumental in generation of sequence and annotation data for the potato genome and in analysis of the whole genome for publication. We have also led the efforts to characterize the transcriptome of the doubled monoploid clone through the generation of extensive RNA-Seq data which has been used for annotation and expression profiling analyses. Data from the PGSC is available to the public through a BLAST server and Genome Browser maintained by group [redacted] I am the manuscript coordinator for the PGSC, and we anticipate submitting our manuscript to a high impact journal in the coming months. [redacted] (University of Wisconsin) is a collaborator on this project and I have also continued my long-standing collaboration with [redacted] to characterize potato and other Solanaceae species using cytogenetics [redacted]

While my research on the Solanaceae has generated large-scale resources for genomic and genetic research, this has not been readily translated to improvements in agriculture. Through funding from the USDA [redacted] MSU Crop and Soil Sciences), we have initiated the Solanaceae Coordinated Agriculture Project [redacted] to translate knowledge in genomics to breeders of potato and tomato germplasm. We have identified a large number of SNPs in potato and tomato and are collaborating on development of two 10,000 SNP Infinium arrays to genotype tomato and potato germplasm relevant to U.S. agriculture. With genotype data in hand, we will link that to phenotype data (focus on carbohydrates, sugars and vitamins) to be generated in the project to identify alleles associated with desired traits.

Genomics and Bioinformatics of Plant Pathogens
My work on plant pathogens has involved genome sequencing, comparative genomics, and bioinformatics. I have lead the sequencing and analysis of multiple plant pathogen genomes [redacted] One of the pathogens we sequenced is *Pythium ultimum*, an oomycete, with a broad host range. Our analyses of the genome revealed a number of genes critical to virulence and in a novel finding, revealed that unlike related oomycetes such as *Phytophthora* species, it lacks a specific class of effectors critical to host range determination [redacted] We are currently expanding our genome and transcriptome sequencing in the genus *Pythium* to an additional five species to further understand virulence and host specificity in this species. We have also initiated a collaboration with [redacted] MSU Plant Pathology) to characterize the genome of the cucurbit downy mildew, *Pseudoperonospora cubenesis*. In July we submitted a proposal to USDA NIFA to develop an integrated management approach ([redacted]-PD, MSU Plant Pathology) for cucurbit downy mildew that includes characterization of the *P. cubensis* population structure and development of DNA-based diagnostic markers.

Through USDA funding, I have developed the web-based Comprehensive Phytopathogen Genome Resource [redacted] that provides access and analysis tools for data-mining plant pathogen genomes. In the CPGR warehouse, we provide access to ~800 plant pathogen genomes. Using a computational approach in conjunction with data within the CPGR, we were able to identify diagnostic markers for the USDA-APHIS select agent, *Xanthomonas oryzae* [redacted]. We are collaborating with [redacted] (University of Wisconsin) to further enhance the annotation

of bacterial plant pathogen genomes in the CPGR by adding manual curation and comparative genomics resources.

Genomics of Biofuel Feedstock Species

With my experience in genomics and bioinformatics, I developed with DOE/USDA funding a web-based resource for data-mining genomic sequence from biofuel feedstock species [redacted] The Biofuel Feedstock Genomics Resource (BFGR) provides a web-based portal or "clearing house" for genome sequence/annotation (structural, functional, and comparative), germplasm data, and large-scale functional genomic datasets for plant species relevant to biofuel feedstock production. Using a core annotation pipeline, we have generated centralized, uniform and integrated functional annotation data for all gene and transcript sequences for species within the BFGR. We are currently deploying a new release of the BFGR and will be preparing a manuscript describing this resource later in 2010.

I have also collaborated with [redacted] and [redacted] (University of Wisconsin) as part of the Great Lakes Bioenergy Research Center (http://glbrc.org/). Our work to date has focused on construction of a gene atlas for maize [redacted] and association of genotype with phenotype in maize lines. We are incorporating state-of-the-art sequencing technologies and will be genotyping a diverse maize panel using next generation sequencing technologies. These data will be essential for identifying genes associated with key biofuel feedstock traits. With funding for the next two years, we will continue our collaboration with on maize and expand our efforts to include switchgrass, a key future biofuel feedstock species.

Metabolomics and Systems Biology

The majority of my experience in genomics and bioinformatics has involved genome sequencing, annotation, and development of community resources for data-mining. In 2007, I realized that to address emerging questions in plant biology a systems biology approach in which "omic" data (genomic, transcriptomic, metabolomic, and interactomic) are integrated needs to be undertaken. To address this, I have initiated two projects at MSU to develop the knowledge, skills and experience in metabolomics and systems biology.

In rice, I initiated a metabolomics project to measure pigments in diverse rice lines. In collaboration with [redacted] (MSU Biochemistry & Molecular Biology), Dr. [redacted] a NSF funded Postdoctoral Fellow in my lab, has established a protocol to profile major pigments from leaves, nodes, and seeds in rice. Concurrently, we are genotyping these lines for key genes involved in pigment biosynthesis with the aim to link genotype with phenotype. While this project is serving as a pilot for metabolomics methods, data analysis approaches, and integration with other data sets, we have also been pursuing a systems biology approach to improve our understanding of gene function in rice. Through co-expression analyses, we have been able to identify gene networks in rice to layer additional data and annotation to improve our understanding of gene function in rice.

In collaboration with [redacted] (MSU Biochemistry & Molecular Biology), [redacted] and [redacted] (MSU Horticulture), we are generating transcriptome and metabolome data for 14 medicinal plants with the aim to provide resources to identify genes involved in biosynthesis of pharmaceutical and human health related compounds [redacted] This project, funded by the NIH, is in its first year with the goal to generate transcriptome and metabolome datasets within two years. We have developed a robust pipeline to analyze and annotate transcriptomes from these species and will be working with other project personnel to link transcripts with metabolites of interest thereby identifying candidate genes responsible for synthesis of the target compounds. This project complements a MSU-funded Strategic Partnerships project [redacted] Biochemistry & Molecular Biology) focused on identifying genes in Solanaceae species important in human health using transcriptomics and metabolomics. This project is synergistic with the NIH Medicinal Plants project with my role to provide sequencing and bioinformatics analysis of the transcriptomic data to aid in identification of key genes involved in the target human health related compounds.

In the coming years, I anticipate expanding the research themes in my lab to include an increased focus on functional genomics, metabolomics and systems biology. Thus, while genomics, transcriptomics, and bioinformatics will continue to be a major component of my research program, I intend to pursue funding that permits more empirical connections between sequence, annotation and phenotypes.

Teaching Philosophy

Classroom:

I have taught at both the undergraduate and graduate level. At Louisiana State University, I taught Introduction to Biology for two semesters to undergraduates and was able to implement improvements in my presentations and assessments in the second semester such that my evaluations were very positive. At Michigan State University, I taught one semester of Genetics with [redacted] in Fall 2008. This was a learning experience as it was not only a large class (340 students) but was also composed primarily of pre-med students. If I taught this course again, I would alter my instructional approach and assessments based on my experience and evaluations from students. From my three semesters of teaching large classes, it is essential to connect with the students with respect to their background, ability to learn new material, and purpose in enrollment.

In fall 2009, I taught a graduate level course on Plant Genomics as this was not covered in any substantial manner in the existing curriculum. I made the course rigorous and required extensive reading of scientific literature, discussion of scientific topics by the entire class, presentations, and preparation of a grant proposal. This was a new course for MSU and I feel that the course was highly successful. Not only were the evaluations very positive, but through my subsequent interactions with the students, it was clear that they learned not only the scientific content from the course, but also critical skills needed to be a successful independent scientist such as critique of scientific literature, presentation skills, and proposal development. In Fall 2010, I will be making improvements to the course based on my review of the lecture and lab activities and the student evaluations.

Mentoring:

My cumulative experiences in mentoring undergraduates, graduate students, postdoctoral fellows, and junior bioinformatic staff have involved assessment of the student/postdoc/staff's capabilities, development of a training program, and then adaptation of the program/curriculum to meet the goals of the individual. While this may seem like a basic strategy, I have found that failure to adapt and tailor a program or curriculum for each individual results in not only an unpleasant experience on both sides but also a reduction in the content learned. Thus, I have placed a strong emphasis in my mentoring activities in familiarizing myself with the background of the student in addition to what they want to "get" from their work/research experience. While this would not result in a major change in the content of the training, it would potentially result in an alteration of time spent on select topics, modification of the presentation style and presentation aids, and engagement in real life applications of science.

I feel I have been very effective in mentoring students and postdoctoral fellows and will continue to work with students (undergraduate and graduate) and postdocs. In today's technology driven science field, it is important that students and postdocs get a broad, yet deep training in not only technology but also application of technology to biology. Thus, I see it imperative that research projects engage multiple scientific disciplines to effectively address biological problems. For my current postdoctoral fellows (4), they all are receiving training in bioinformatics to ensure they have appropriate skills to develop an independent scientific career.