CPGS - *Oryza* genome evolution
PI – R. Wing (UA); **Co-PIs** – S. Jackson (Purdue U), C. Machado (U Maryland), M. Long (U Chicago), M. Sanderson (UA); **Key Personnel** – D. Ware (CSHL/USDA-ARS), Y. Yu (UA), C. Fan (UA), A. Zuccolo (UA), A.S.S. Jetty (UA), and D. Kudrna (UA); **Key Collaborators** – O. Panaud (U Perpignan), D. Weigel (MPI)

DW to receive 1.5 FTE for work, UA to prepare Ensembl Core, UA and MPI to prepare Ensembl variation core?

ObjectivesA1 Geneome Framework Datasets for *O. glumeapatula, O. meridionalis, Leersia perrieri* (UA)
A2 Vertical dataset: sequence of chr 3 arms: *O. glumeapatula, O. meridionalis, O. granulata, Leersia perrieri.* Compile RefSeqs (UA)
A3 Horizontal dataset: sequence whole genome *O. barthii* and *O. punctata.* Assemble whole genome RefSeqs (UA)
A4 Whole transcriptome dataset cDNA and small RNA data 12 species. (UA)
A5 Population resequence data set: *O. barthii, O. glaberrima, O. sativa japonica, O. sativa indica, O. rufipogon, O. nivara, O. longistaminata, O. glumaepatula, O. meridionalis* (UA)
Automated Baseline AnnotationB Genome Annotation, comparative analysis, gene prediction, repeat annotation, SNP variation, transcriptome mapping and all *Oryza* visualization. (CSHL)

II. Specific Aims and *Research Focus*

Comparative genomics is a powerful tool for understanding how genes and genomes function and evolve. The vast majority of comparative genomics studies in plants published to date have focused on species from different genera (e.g., *Zea mays* [corn] vs. *Oryza sativa* [rice]). Although these studies have resulted in key discoveries, comparisons among closely related species are critical for a deeper understanding of speciation, polyploidization, domestication and gene regulation. The two most advanced genus level comparative systems available for multicellular organisms are the *Drosophila* comparative sequencing project [11] (where genomes from 12 *Drosophila* species were draft-sequenced and compared to the reference *D. melanogaster* and *D. pseudoobscura* genomes), and **OMAP** [10].

The long-term goal of our research is to establish *Oryza* as a key biological system to address fundamental questions in genome biology, evolution, domestication, and crop improvement. We envisage a backbone of high-quality RefSeqs for representatives of all 24 species of *Oryza*, on which successive "omics" data (transcriptome, metabolome, epigenome, phenome, etc) can be layered. All data types would be interpreted on the background of a rigorously assembled phylogeny, thus providing the evolutionary context to address wide

ranging issues, from basic plant biology to feeding the world in a sustainable manner. Such a comprehensive system would also serve as a model for other plant genera that contain important food, fiber and energy crops.

In addition to providing new genomic and breeding tools to improve cultivated rice, the goal of this proposal is to expand the OMAP comparative genomics system to solve important evolutionary questions relevant to the functional biology of rice and other plants. This proposal builds upon previous NSF-funded work by the PIs to sequence the rice and maize genomes, to create and exploit the OMAP system and diverse interspecific germplasm resources, to create comprehensive databases and collections of analytical tools to analyze these data types, and to efficiently and cost effectively sequence targeted genomic regions and whole genomes with new sequencing technologies.

To accomplish this goal we will: A) generate five publicly available primary datasets and resources for the genus *Oryza,* unprecedented in their scope, utility, and richness of data; B) perform baseline annotation on all sequence and comparative datasets; and C) perform computational and experimental analyses to address specific questions in evolutionary and comparative genomics (described below).

A) Data and resource generation: We propose to generate a set of publicly available comparative and functional genomics tools and datasets that will unite the international scientific community in addressing fundamental questions in basic and applied science. Specifically we will produce the following:

1) Genome Framework Dataset: BAC end sequence/fingerprint physical maps for two AA genome species - *O. glumaepatula*, *O. meridionalis*, and one outgroup *Leersia perrieri*. These species are endemic to S. America, Australia and Africa, respectively, and contain agronomically desirable traits (i.e. drought tolerance [23,24], elongation ability [23,24], resistance to diseases [23-26] and pests [24,25]). The AA genome physical maps will complete the OMAP framework datasets for all the AA genome species, and will provide the resources necessary to generate RefSeqs for these three genomes through I-OMAP. The *L. perrieri* map will provide a critical outgroup framework map for the entire *Oryza* genus.

2) Vertical Chr3S Dataset: RefSeqs for Chr3S of *O. glumaepatula*, *O. meridionalis*, *O. granulata* and *L. perrieri.* These RefSeqs combined with previous data will provide a complete vertical dataset of thirteen Chr3 short arm sequences for seven AA and 1 each of the BB, CC, BBCC, FF, & GG genome types and the outgroup *L. perrieri*. Why chromosome 3? Rice Chr3 is one of the most highly conserved chromosomes among the cereals, and it contains several important QTL [27]. From an evolutionary standpoint, it harbors one of the key chromosomal rearrangements that correlate with the phylogenetic origin of the Panicoid clade of grasses (maize, sugarcane and sorghum) [28-30]. Chr3 was sequenced by PI Wing and colleagues as part of the IRGSP [3] and is now a target for complete functional characterization as part of an international effort to characterize all rice genes by 2020 [4].

3) Horizontal Dataset: *O. barthii* and *O. punctata* whole genome RefSeqs. *O. barthii* is one of eight species selected by I-OMAP for whole genome sequencing.

As a progenitor of West African cultivated rice (*O. glaberrima*), it is a potential donor for resistance to bacterial blight, bacterial leaf streak, blast and green leafhopper [23,24,31] and for drought tolerance [23,24]. The independent domestication of Asian rice (~10kYA) [32-37] and African rice (~3kYA) [38-41] should lead to interesting comparisons between these genomes. Genomic regions selected in both species will be particularly interesting, as will be species-specific selected regions – perhaps as a result of regional adaptation, or due to distinct grain quality traits. When combined with the *O. sativa* ssp *japonica* and *indica*, *O. glaberrima*, and forth coming *O. longistaminata*, *O. rufipogon*, *O. nivara* and *O. brachyantha* RefSeqs, we will have at least nine whole genome *Oryza* RefSeqs for comparative analyses.

Not only is the wild BB species *O. punctata* a serious invasive weed in East Africa, but the accession selected for sequencing also contains important stress tolerance traits that could be used to improve cultivated rice (e.g., shade tolerance [42], resistance to stem rot [43], brown plant hopper [23,24,44,45], whorl maggot [46], and zigzag leaf hopper [23,24]). This species will also serve as a close outgroup for all the AA genomes – a critical component of the evolutionary analyses we propose.

4) Whole Transcriptome Dataset: Deep sequencing of mRNA and small RNAs for the 11 *Oryza* species and *L. perrieri* in the **Horizontal** and **Vertical** datasets. This resource will be used to enhance both the annotation and comparative analyses of transcripts and small RNAs across the *Oryza.*

5) Population Dataset: 15xwhole genome sequencing data for 77-140 AA genome accessions: We will resequence 14-21 accessions each of *O. glaberrima* and *O. barthii* and 7-14 accessions each of *O. sativa* ssp. *indica*, *O. sativa* ssp. *japonica*, *O. nivara*, *O. rufipogon*, *O. longistaminata*, *O. glumaepatula*, and *O. meridionalis* for a total of 77 to 140 AA genome strains (the different numbers of accessions are dependent upon expected increased sequencing throughputs). This resource will be used to add additional statistical power to all proposed analyses in genomic and molecular evolution, phylogenomics, and population genetics. Additionally, all high quality SNPs and SNVs derived from this data set will be publicly available for future marker assisted selection and association genetics studies.

B) Baseline Annotation: We will use automated methods to annotate the genomic and transcriptome resources from a comparative evolutionary perspective. Initially all sequences will be run through an automated annotation pipeline to obtain baseline information across the Vertical and Horizontal RefSeqs, including automated gene and SNP calling, transcript mapping and small RNA identification and mapping.

B) GENOME ANNOTATION, COMPARATIVE ANALYSIS AND VISUALIZATION:

The integration of many different types of data both within and between species aids biologically significant discoveries. As part of a collaborative effort among projects under the purview of the Ware Lab - notably Gramene [72] and the Maize Sequencing Project [96] - a robust software infrastructure has been established for streamlining and automating whole-genome analysis, primary annotation and visualization. The infrastructure extends the Ensembl pipeline by providing plant-specific parameters and modules. While there is a common overall structure of the pipeline, individual modules and dependencies can be customized. For this project, we propose to run the following to provide baseline annotations of repeats, genes (where there are no community provided annotations), and comparative analysis.

Automated baseline annotation for repeats and genes (CSHL Gramene): Upon receipt of Ensembl core databases from AGI, the annotation process will begin with determining the repetitive content using RepeatMasker [97], Electronic Simple Sequence Repeats [98] and Mathematically-Defined Repeats [99,100]. We currently use repeat libraries for rice and maize that are keyed to classification ontologies, allowing the projection of major TE classes and families onto masked regions [101-103]. The combined TE/repeat annotations will serve as the starting point for the analysis of the role of repetitive DNA in genome evolution (C4).

As a first pass for gene annotation, genes will be called using an *ab initio* method, Fgenesh [104]. For all genomes that have no community available annotations, we will provide Fgenesh predictions. We will subsequently provide evidence-based gene-builds, using a method that draws on extensive collections of cross-species ESTs and other evidence types [73] for up to 3 genomes. The selection of genomes to annotate will be done in consultation with all project participants. Predictions are filtered to distinguish TE-related and hypothetical genes from those encoding proteins similar to known proteins. Further annotations are provided by modules that assign InterPro functional domains [105,106], GO terms [107,108], and relationships to proteins of known function in UniProtKb/Swiss-Prot [109].

Comparative Analysis (CSHL Gramene): Intergenomic-comparisons with Ensembl Compara modules [76,110] will provide high-quality comparative maps, gene family relationships, and additional support for analysis of genome evolution. Annotated genes will be incorporated into the EnsemblCompara GeneTree pipeline [111], building on a comprehensive phylogenetic resource that includes other grass and eudicot genomes hosted by Gramene. This pipeline also gives putative ortholog and paralog assignments, which are used to build gene-level synteny maps [112]. Pair-wise whole genome alignments [101] to *O. sativa* will provide complementing nucleotide-level maps. The most recent development from Ensembl, EPO [76], elucidates the history of genome rearrangements across multiple species and reconstructs the ancestral genome. The EPO pipeline has of three stages: definition of large-scale colinear homologous blocks (Enredo), consistency based multiple aligning (Pecan), and ancestral genome reconstruction (Ortheus) [113,114]. It also supports segmental duplications. Multiple genome alignments will provide a basis for detecting

evolutionarily constrained elements within coding and non-coding regions using GERP [115,116], which calculates conservation score at each aligned site and is implemented as a Compara module. These analysis tools will be applied to support analysis of structural variation (C1), phylogenomics (C2), evolution of new genes (C3), and identification of regulatory elements (C5).

Genome Display (CSHL Gramene): An OMAP project entry page added to Gramene will provide key entry points to maps, sequences, annotations, and variation data. Each species' genome will have its own browser that can be navigated by physical map and would display all sequence-anchored annotations. Comparative annotations will provide links that promote "inter-species browsing". For example, whole genome alignment tracks would invite users to "jump" to corresponding regions in different species' RefSeqs. The Compara TreeView allows users to browse phylogeny by traversing color-coded speciation and duplication nodes. From here users can view multiple sequence alignments, nucleotide-alignments, or launch other genome browsers. Other points of entry include Gramene search pages, which support mining of SNPs, comparative maps ontologies, and include a highly configurable BLAST tool. In addition to displaying annotations generated from the pipeline, the browser can be used to display community annotations using the Distributed Annotation Server [116].

Variation (MPI/UA/CSHL Gramene): Single Nucleotide Polymorphisms (SNPs) and structural variants (SVs) including indels, inversions and copy number variants (CNVs) will be identified from the population dataset using the *SHORE* pipeline [89,90] originally developed for the *Arabidopsis thaliana* 1001 Genomes Project. *SHORE* is using a custom read quality filter and the highly sensitive short read aligner *GenomeMapper* [117] to perform gapped alignments against a reference sequence.

Homo- and heterozygous SNPs and short indels are predicted using a consensus calling algorithm that exploits base and alignment quality, read support and concordance, coverage uniformity, repetitiveness, sequence complexity and vicinity consistency. Features are currently scored using an empirically defined scoring matrix (or optionally scoring by a support vector machine) [118]. SHORE's consensus caller is sensitive enough to predict SNPs in pooled samples down to a minor allele frequency of just 1% [119].

For predicting SVs, *SHORE* provides a statistical framework exploiting read pair distance and orientation of mapped pairs. Genomic regions with read pairs significantly deviating from the insert size distribution are detected. Length and p-value are assigned to each SV based on all reads overlapping the respective region. Other predictors, e.g., read orientation, orphaned reads (missing their mate) and unique reads with multiple mates are used to detect more complex rearrangements such as inversions and translocations [120]. Additionally, analysis of read coverage reveals CNVs by virtue of zero coverage (long deletions) or significantly increased coverage (duplications). To analyze long insertions -and highly polymorphic regions not spanned by read pairs- *SHORE* applies a local homology-guided assembly.

Integration of these analyses will create a standardized baseline of comparative maps and genome annotations across *Oryza*. The consistency among maps and annotation of several cereal genomes including *Oryza* will provide the infrastructure to identify conserved genes and functional non-coding sequences, to interpret genetic diversity data, and to study evolution. All annotations will be accessible to project members via the Ensembl API, and, where appropriate, in flat file formats as soon as they are complete. Community access will be via Ensembl web display and associated tools, starting in year two after the data has been quality reviewed. Updates will occur on Gramene's biannual release schedule.

Table 1: Sequencing status of 11 *Oryza* species and *L. perrieri*

Species [GT]	GS (Mb)	Chr3S	Sequencing Lab (Whole Genome)	2009 (November) progress
O. rufipogon [AA]	448	+	NCGR China	9/2010 completion date
O. nivara [AA]	439	+	AS, Taiwan	9/2010 completion date
O. glaberrima [AA]	358**	+	AGI, USA	3/2010 completion date
O. barthii [AA]	382**	+	AGI, USA	*2010/2011, this proposal*
O. longistaminata [AA]	376*	NA	CASK/BGI China	3/2010 completion date
O. glumaepatula [AA]	464*	*this proposal*	FUP, Brazil	pending
O. meridionalis [AA]	390*	*this proposal*	SC, Australia	pending
O. punctata [BB]	425**	+	AGI, USA	*2010/2011, this proposal*
O. officinalis [CC]	651**	+	NIG, Japan	pending
O. brachyantha [FF]	362**	+	CASB/BGI China	3/2010, completion date
O. granulata [GG]	882**	*this proposal*	No Commitment	NA
L. perrieri [outgroup]	323*	*this proposal*	No Commitment	NA

Abbreviations: NCGR = National Center for Gene Research, Shanghai, PRC; AS = Academia Sinica, Taipei, Taiwan; AGI = Arizona Genomics Institute, Tucson, USA; FUP = Federal University of Pelotas, Brazil; SC = Southern Cross University, Australia; CASB = Chinese Academy of Sciences Beijing, PRC; BGI =

Beijing Institute of Genomics, Shenzhen, PRC; CASK = Chinese Academy of Sciences Kunming, PRC; NIG = National Institute of Genetics, Mishima, Japan; NA = not applicable; GT = Genome Type; GS = Genome Size
*From Ammiraju et al., 2006 [22], **unpublished estimates (Wing, Arumuganathan et al.)