# Evolutionary Relationships in *Oryza* Inferred from the Gene that Encodes the 10 kDa Prolamin (Seed Storage Protein) Polypeptide

Irene M. Mullins

Master's thesis submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master's of Science
in
Biology

Khidir W. Hilu, Chair
Charles L. Rutherford
Bruce J. Turner

Dec. 15, 1999

Keywords: *Oryza,* rice, prolamin, systematics, phylogeny, concerted evolution

Evolutionary Relationships of *Oryza* Inferred from the Gene that Encodes the 10 kDa Prolamin (Seed Storage Protein) Polypeptide

Irene M. Mullins

**(ABSTRACT)**

Cereals represent one of the most important food crops in the world. Rice (*Oryza sativa*) is one of the most commonly consumed cereal grains, and as a result, has considerable economic and agricultural importance, despite their potential as a source of genetic material. The phylogenetic relationships among rice and its wild species are not well understood. The objectives of this study are to evaluate the rate and type of molecular variation present in the gene that encodes the 10 kDa prolamin polypeptide in *Oryza* (Poaceae), and use this information in understanding the evolution of the *Oryza* genus. Our principle hypothesis is that genetic diversity exists at the molecular level in wild species of *Oryza*, and that this diversity can provide useful information regarding the phylogenetic relationships among rice species and permit a more theoretical examination of the evolutionary processes, such as concerted evolution, within *Oryza* and its nine genomes. A phylogeny of *Oryza* is presented, and modes of evolution are discussed.

**Grant Information**

This work was funded by grants from The Virginia Academy of Science; *Sigma Xi*; the Graduate Research Development Project, Virginia Polytechnic Institute and State University; and the Biology Department; Virginia Polytechnic Institute and State University; and the Jeffrees Foundation.

## Dedication

Special thanks go to my family who have supported me throughout my academic endeavors, and to David, my husband and rock of support, who has been both my best friend and mentor *par excellance*.

**Acknowledgments**

I would like thank my committee for their advice and constructive critique of my work. Special thanks go to Cathy Aime for reading portions of these chapters, and to Tom Weiboldt for very patiently teaching me the art of herbarium voucher preservation.

## Table of Contents

## Multi-Media Objects

**I. Chapter 1. Literature Review**



**II. Chapter 2.Variation in the gene encoding the 10 kDa prolamin in *Oryza* and its phylogenetic implications**



**III. Chapter 3. Genomic origin of *Oryza* polyploid species based on the genes encoding the 10 kDa prolamin: A case of concerted evolution.**

## -Chapter 1.-

## Literature Review

***Objective of Study***

The objectives of this study are to evaluate the rate and type of molecular variation present in the gene that encodes the 10 kDa prolamin polypeptide in *Oryza* (Poaceae), and use this information in understanding the evolution of the *Oryza* genus. Our hypothesis is that genetic diversity exists in this gene at the molecular level in the wild species of *Oryza*, and that this diversity can provide useful information regarding the phylogenetic relationships among rice species and permit a more theoretical examination of evolutionary and phylogenetic relationships among the nine *Oryza* genomes.

***Rationale for Study***

Cereals comprise the most important food crops in the world, providing about 80% of our food at the global level (USDA, 1991). Rice (*Oryza sativa*) is one of the most commonly consumed cereal grains. Consequently, considerable economic and agricultural importance has been placed on improving this crop. In spite of this economic importance, however, the genetic relationships among its species are not well understood. Data obtained in this study will provide valuable information for two areas of research. First, sequence data will aid in the determination of the breadth of genetic variation present in the gene that encodes the 10 kDa prolamin (seed storage protein) polypeptide. These sequence data will provide insight into addressing two questions: (1) do particular species of rice contain prolamins that are richer in essential amino

acids than others, and (2) if such species exist, can the sequence information be used to identify desirable amino acids, thereby improving the nutritional value of the domesticated/cultivated species. Second, comparative nucleotide sequence information of all *Oryza* species will permit a better understanding of: (1) the origin and relationships among the species in this genus, (2) the genetic relationships among the various genome types, and (3) the evolutionary mechanisms by which those genomes arose.

***Socioeconomic Importance of Rice.***

Cereals comprise the most important food crop in the world. Rice (*Oryza sativa*) is one of the most common cereal grains consumed, which gives it considerable economic and agricultural importance. The global, annual rice production is 562, 260 thousand metric tons (tmt), a yield that is a close third to wheat (584, 874 tmt) and maize (576, 821tmt) (Fig.1) (Leviton & McMahon, 1996; World Almanac, 1998). However, unlike wheat and corn, rice is almost entirely consumed by humans.

Cereals provide a major source (more than 60%) of calories and proteins for many Asian countries (Bhatia & Rabson, 1987). For example, Japanese diets receive approximately 19% of their protein from rice (Iida *et al.*, 1993). One concern among nutritionists is that protein deficiencies arise because a "sufficient bulk" of grain cannot be consumed to meet dietary protein needs, especially in young children (Bhatia & Rabson, 1987; Coffman & Juliano, 1987) and women (de Lumen *et al.*, 1991).

Recent estimates, however, have indicated that despite its role as a major food staple, rice production is slowing down as a result of: (1) land conversion for urban use, (2) salinity build-

up, (3) declining soil nutrients, and (4) pesticide accumulation (Pignali *et al*., 1997). In 1991, rice production exceeded maize production by 41, 094 tmt (Leviton & McMahon, 1996); however, by 1996 this trend was reversed, and maize production exceeded rice production by 14, 561 tmt (World Almanac, 1998). Moreover, maize production experienced a 12.3% greater increase than rice between the years 1991-1996 (Fig. 2) (Leviton & McMahon, 1996; World Almanac, 1998).

The decrease in rice production has not reflected a decrease in need. By the year 2020, the predicted world population will be over 7.593 billion, which represents a 29.7% increase over the 1997 world population census of 5.851 billion (The World Almanac, 1998) (Fig. 2). It has been predicted that by the year 2025 it will be necessary for annual rice production to increase 800 million tons over the 480 million tons of 1990 to keep pace with the worldwide demand (Pignali *et al*., 1997). In Asia alone, the 1996 population of about 3. 477 billion is expected to increase by 29.2% to about 4. 492 billion by 2020, and by 51.4% to about 5.267 billion in 2050 (The World Almanac, 1998) (Fig. 2), thereby further straining current agricultural resources.

Asia, however, is not the only rice-consuming region of the world that will experience a drastic increase in its population. It is predicted that by 2020 Africa will experience a 62.8% increase in its population from its 1997 census of 7.5 million to 1.22 billion people (The World Almanac, 1998). As a continent, Africa will experience the largest percent in population increase compared to Asia, Europe, South America, North America, Former Soviet Union, and Australia (The World Almanac, 1998). By the year 2050, the predicted African population will total 2,049,953, 000, an increase of 67.8% from the 2020 census, and 173.2% from the 1997 census (The World Almanac, 1998). Despite this explosive increase in population, the 1996 rice

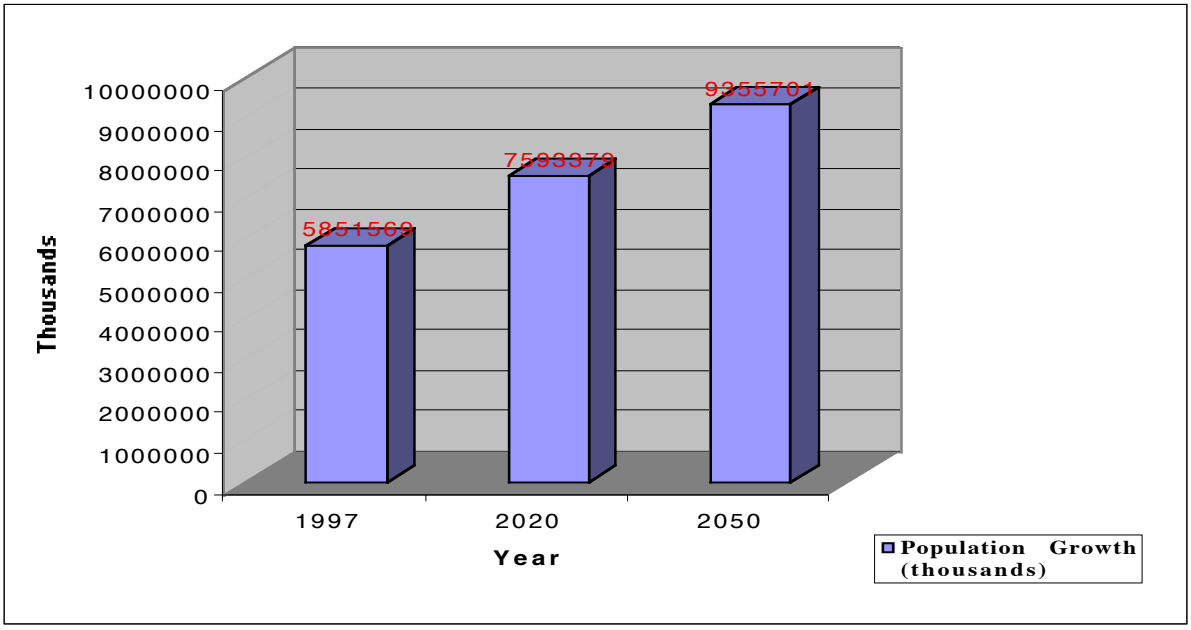



**Fig. 1. World Population Growth: Current and Predicted (*The World Almanac*, 1998).**

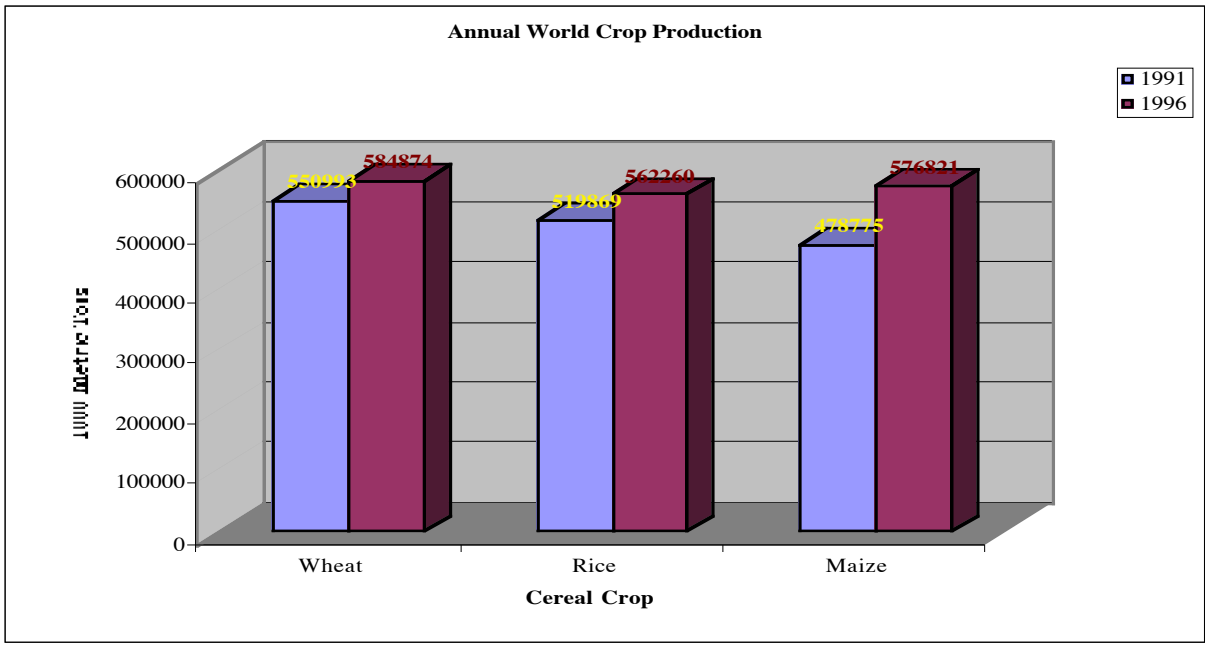



**Fig. 2. Annual World Crop Production for 1991 (Leviton & McMahon, 1996) and 1996 (*The World Almanac*, 1996).**

yield for the few reported African countries recorded by the World Almanac (1998) represented approximately 1.03% of the world rice production, while rice provides, on average, 174.9 cal/day (FAO, 1997), which represents 8.3% of the average sub-Saharan African diet (FAO, 1988/90).

***Nitrogen uptake and its role in crop improvement***

The observation that grains are excellent suppliers of carbohydrates but poor suppliers of proteins has led towards many attempts to increase the protein content in rice via genetic engineering (reviewed in Toenniessen, 1991). One proposed strategy is the insertion or mutation of pre-existing genes in the hyper-variable regions of the sequence (Argos *et al.*, 1985). Such insertions in the hypervariable region of glutelin have been tolerated (Argos *et al.* 1985, regiewed in Okita, 1991). To aid in the insertion of foreign genetic material, "safe" nucleotide substitution matrices have been calculated and deemed to cause the least disruption in the protein structure (Bordo & Argos, 1991).

Knowledge of the nucleotide sequences of prolamin genes, such as the gene that encodes the 10 kDa prolamin in the cultivated and wild members of *Oryza*, will permit a wide-sweeping comparison of the deduced amino acid sequences, thereby elucidating potential variable/target regions within the glutelin gene, and providing mechanisms by which the gene that encodes prolamin can likewise be manipulated.

One potential problem associated with this scheme of protein enhancement is that an increase in nitrogen uptake is requisite with an increase in protein content (reviewed in Bhatia & Rabson, 1987). Specifically, a 1% increase in protein content requires an additional 6-11% nitrogen (Bhatia & Rabson, 1987). This increase in demand for nitrogen is met either through direct uptake from the soil (which would most likely require the application of fertilizer) or

through reallocation of nitrogen that has been stored in leaves (Bhatia & Rabson, 1987). This reallocation of foliage nitrogen results in a negative correlation between grain yield and the protein content (Bhatia & Rabson, 1987). With the exception of the "Fukunishiki" cultivars, it was found that a negative correlation between protein content and grain yield is present in rice after the application of heavy amounts of nitrogen (Higashi *et al*., 1974). Thus, a decrease in yield could be encountered if farmers used bio-engineered rice seeds without a con-comittant high input of nitrogen.

The use of symbiotic fungus to aid in the uptake of nitrogen has been considered a target for the improvement of grain quality and quantity where the application of chemical fertilizers would be otherwise cost-prohibitive (reviewed in: Toenniessen, 1991). Early experiments that examined the presence of VAM (vesicular-arbuscular mycorrhizal) fungi in rice fields indicated that (1) weeds endemic to rice fields promote the survival of VAM, and fields in which they are plowed-under during fallow periods left few hosts for the survival of the mycorrhizal fungi and resulted in less innoculum for the ensuing crops, (2) soil preparation methods such as rototillage, plowing, and harrowing "dilute" the VAM populations, (3) flood conditions in rice fields create low oxygen levels which results in unfavorable conditions for VAM populations, and (4) natural predators, such as mycophagous nematodes, fungiforic mites and soil microarthoropods, reduce VAM populations (Ilag *et al*., 1987).

In addition, the effect of arbuscular mycorrhizal fungi (AMF) inoculation on rice seedling growth was examined by Solaiman and Hirata (1997). It was found that rice seedlings that were grown in a wet nursery without AMF inoculation yielded 1643 unhulled grains/field; whereas the seedlings that were inoculated with AMF exhibited 1930 unhulled grains/ field, representing a

17% increase in grain yield through the inoculation of arbusucular mycorrhizal fungi (Solaiman & Hirata, 1997).

**Starch Synthesis: Its correlation with Prolamin**

It has also been postulated that prolamin and starch synthesis are dependent upon each other in developing grains (Bhatia & Rabson, 1987). Normal starch synthesis appears to be dependent on the presence of prolamin in maize. Tsai *et al*. (1980; reviewed in Bhatia & Rabson, 1987) hypothesized that zein prolamin and glutelin fractions in maize serve as nitrogen sinks which aid in the accumulation of photosynthates into the kernels. They further concluded that impaired prolamin synthesis (using an *o2* [2] mutant that has reduced a prolamin content) results in the accumulation of free amino acids and disrupts solute movement into the kernel (Tsai *et al*., 1980).

Studies have been conducted that have identified rice mutants that contain both high and low amounts of the different weight classes of prolamin (Kanamaru *et al*, 1988); however further analysis must be completed to determine the effect of prolamin synthesis inhibition on the formation of starch in rice. Specifically, research protocols that knock-out the gene that encodes the 10 kDa prolamin polypeptide in *Oryza*, thereby reallocating nitrogen for glutellin synthesis, may encounter disrupted grain starch synthesis as a result of these genetic manipulations. The data presented in the subsequent chapters address the issue of amino acid variability within prolamin, should a means by which this protein could be better utilized by humans arise.

***The Prolamin***

The prolamin is a class of seed storage proteins that is unique to the grass family (Shewry *et al.*, 1995). The prolamin class is distinguished by its alcohol-solubility from the three other classes of seed storage proteins, the albumins (water-soluble), globulins (salt-soluble) and glutelins (dilute alkali or acid-soluble) (Shewry *et al.*, 1995). Of the three categories of proteins, glutelins and prolamins are the most abundant in rice, comprising nearly 90% of its total seed protein (reviewed in Chen *et al.*, 1995).

The primary function of storage proteins is the accumulation of nitrogen and sulfur, two elements required for seed germination (reviewed by Zhou *et al.*, 1993; Higgins, 1984). Prolamin is the main type of storage protein found in cereals, with the exception of rice and oats. Between 30-60% of the total protein content in maize, sorghum, barley, and wheat consists of prolamin (Bhatia & Rabson, 1997). The primary storage protein in rice is a glutelin-like protein (Bhatia & Rabson, 1987), whereas prolamin is present in lower quantities. The reported values of prolamin quantity in the protein fraction of rice, however, have been dependant on the extraction techniques utilized; as a result, these reported values have fluctuated over the past decade: 5% (Masumura *et al.*, 1991)**;** 20-25% (Ogawa *et al.*, 1987), 18% (Li and Okita, 1993) and 35% (Krishnan & White, 1995).

Prolamin is rich in hydrophobic and uncharged amino acids, which includes glutamine (reviewed in Chen *et al.*, 1995) and contains fewer essential charged amino acids such as lysine and tryptophan (Coffman & Juliano, 1987). Consequently, cereal crops are considered to have poor nutritional quality due to their high (up to 60%) content of prolamin (Bhatia & Rabson, 1987).

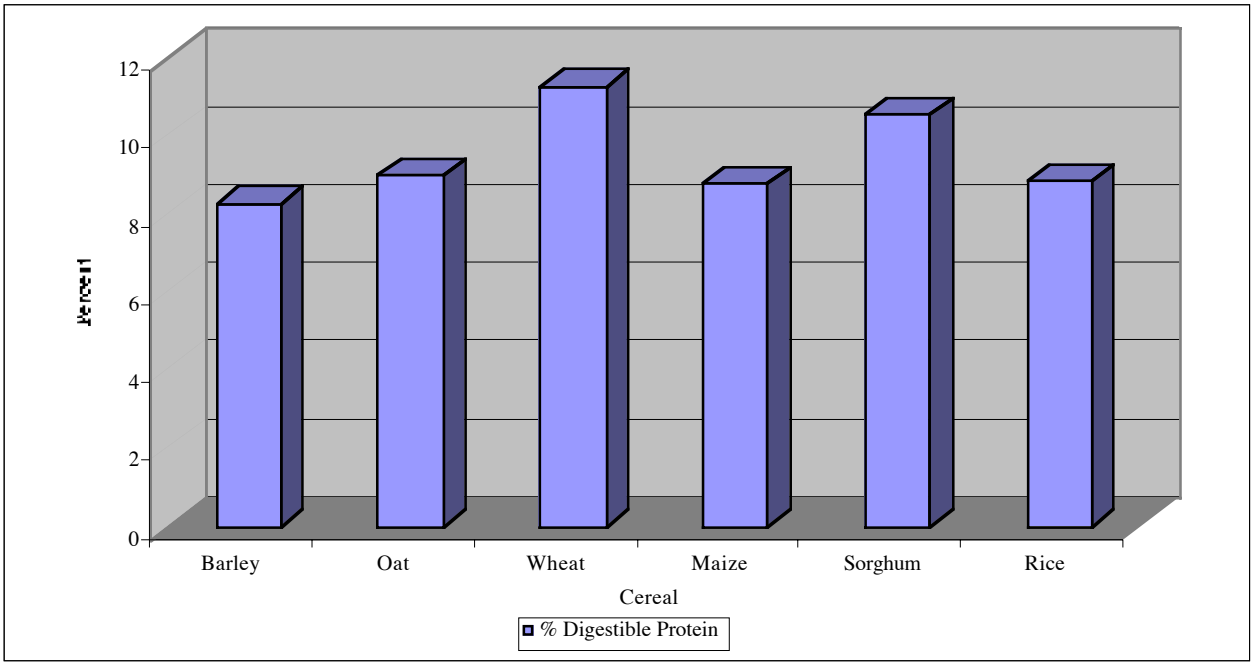



**Fig. 3. Percent digestible protein in the major cereal grains (Bhatia & Rabson, 1987).**

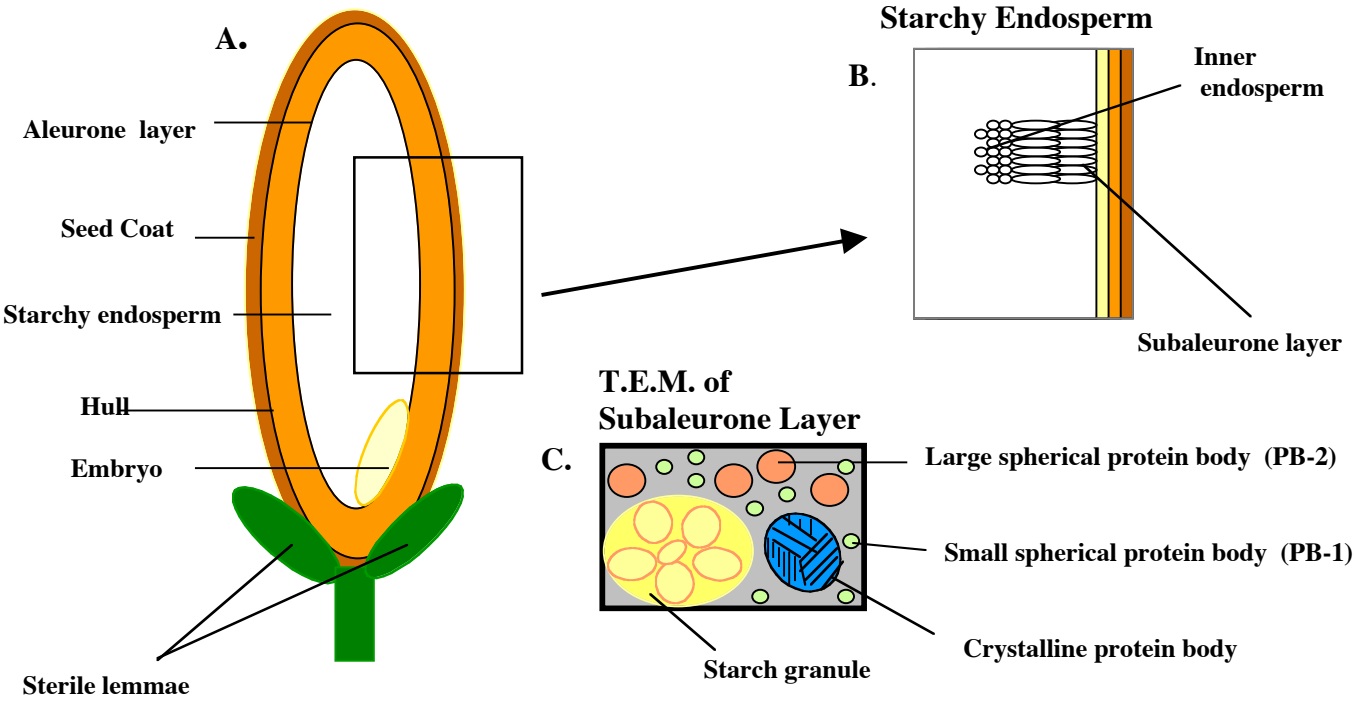



**Fig. 4. (A) Cross-section of a rice grain indicating the (B) subaleurone layer of the starchy endosperm where PB-1 is sequestered. (C) Schematic view of a transmission electron micrograph of subaleurone layer (Coffman & Juliano, 1987).**

Rice has one of the highest quality proteins among cereals because of its *low* prolamin content (Coffman & Juliano. 1987: 105). It should, none-the-less, be noted that the 10 kDa prolamin polypeptide in rice is rich in the sulfur-containing amino acids cysteine (6.8 mol/ 100mol) and methionine (12.63 mol/100 mol), and is high in proline and threonine (8.59 and 8.64 mol/ 100mol respectively) (Hibino *et al.*, 1989).

In terms of quantity, the percent of digestible protein in dry rice matter is 8.9 (Bhatia & Rabson, 1987). This value ranks closely to the 9.0% of oat and the 8.8% of maize (Bhatia & Rabson, 1987), which leaves only wheat (11.3%) and sorghum (10.6%) having significantly greater percents of digestible-in-dry-matter protein than rice (Fig.3).

The prolamin polypeptide is sequestered within a type 1 protein body (PB-1) in the starchy endosperm of the rice grain (Fig. 4) (Tanaka *et al.*, 1980; Ogawa *et al.*, 1987). PB-1 cannot be easily digested in the human body because it is resistant to the action of proteolytic enzymes such as pepsin (Ogawa *et al.*, 1987). In contrast, type 2 protein body (PB-2), the storage unit for glutelin, can be easily digested (reviewed by Iida *et al.*, 1993). The signal peptide regions, by which it is speculated that these two proteins are routed to differing storage units, have been identified for the genes that encode the 10 kDa prolamin (Masumura *et al.*, 1989a) and glutelin (Masumura *et al.*, 1989b) polypeptides in rice. This opens the possibility of re-packaging scenarios of the 10 kDa prolamin polypeptide into a more digestible form (Toenniessen, 1991).

The molecular weights of prolamin in the grass family range from 10-65 kilo Daltons (kDa) (Hilu & Esen, 1988). Rice contains prolamins of low molecular weight, which are organized into three size class, the 10, 13, and 16 kDa (Barbier & Ishihama, 1990). It is

currently thought that these three size classes of protein are encoded by three multi-gene families (Kim and Okita, 1988a). Other members of the grass family, such as wheat (60-100 kDa), oats (22.5 kDa), teosinte (27.7 kDa), and sorghum (22kDa) all contain prolamins of higher molecular weights (Bietz, 1982).

The prolamins of rice are thought to be related to those found in the Triticeae (Shewry *et al.*, 1995). Prolamins of the Triticeae (such as barley, wheat, and rye) can be classified into three structural groups based on their amino acid composition, the: (1) sulfur-rich (S-rich), (2) sulfur-poor (S-poor), and (3) high molecular weight (HMW) (Shewry *et al.*, 1995). The presence of the S-poor and S-rich groups ensures that protein synthesis will occur, regardless of the availability of sulfur (Randall *et al.*, 1979; discussed in Chandler *et al.*, 1983).

Sulfur-rich (S-rich) prolamins constitute 80-90% of prolamin fractions (Shewry *et al.*, 1995), and are comprised of a N-terminus with repeated sequences and a C-terminus with unique sequences. The repetitive sequences are rich in proline and glutamine, which form secondary structures of B-reverse turns and poly-L proline II helix (Shewry *et al.*, 1995). The unique C-terminus, in contrast, is rich in alpha helices, which contain a large percentage of cysteines. Sulfur-poor prolamins, in general, cannot form oligomers or polymers because they lack cytsteine residues (Shewry *et al.*, 1995). Encoded S-poor prolamins found in barley and rye are comprised of octapeptide repeats: Pro-Gln-Gln-Pro-Phe-Pro-Gln-Gln. The N-terminal of this repeat is flanked by short, 12-residue sequences, whereas the C-terminal side is flanked by short, 6-4 residue sequences (Shewry *et al.*, 1995).

High molecular weight prolamins consist of repeated domains, which are capped at the N and C termini by non-repetitive sequences. Like S-poor prolamins, high molecular weight prolamins consist of a spiral secondary structure, which results in a "rod-shaped" molecule

(Shewry *et al.*, 1995). Cysteine residues are largely confined to the amino and carboxyl terminals, which provide for the formation of a disulfide bond-stabilized elastic network within the alpha helix.

***The Gene Encoding the 10 kDa Prolamin Polypeptide***

It has been hypothesized that structural constraints in seed proteins are high because the protein encoded by its respective gene must: (1) have recognition sequences that are responsible for the deposition of the protein into protein bodies, (2) be able to withstand changes in osmotic conditions during the drying process of the seed, and (3) provide sufficient nitrogen stores for the seedling (reviewed in Higgins, 1984). Therefore, if constraints exist that prohibit or decrease the likelihood that a significant change in the protein structure of this polypeptide will occur, then the nucleic acid sequence coding for the amino acids will likewise be conserved (Higgins, 1984). The prime question for systematists, therefore, becomes how useful the prolamin gene sequence data will be for examining relationships among taxa. In terms of systematic/evolutionary utility, the degree to which nucleotide heterogeneity exists (which ultimately code for amino acids in the degenerate code) becomes the most important question for examining phylogenetic relationships at various taxonomic levels.

Masamura *et al*. (1989) isolated cDNA that codes for a sulfur-rich 10 kDa rice prolamin. These sulfur-rich prolamins contain levels of methionine and cysteine that are adequate from a nutritional standpoint; however, the proteins are not well-digested by monogastric animals (reviewed in Okita, 1991). Based on the nucleotide sequence, a polypeptide consisting of 134 amino acids was deduced. Unlike all other cereals, no major repetitive sequences were observed.

Barbier and Ishihama (1990) examined the variation present in the DNA sequence of a prolamin gene family that codes for the 10 kDa prolamin polypeptide in *Oryza rufipogon* by PCR amplification and direct sequencing. The advantage of this technique over cloning (where a single gene is sequenced) were considered to be twofold: (1) it is more expedient, and (2) it permits the revelation of "heterozygosity, heteroplasmy, and heterogeneity" within gene families. Barbier and Ishihama (1990) further noted that the PCR products yielded single bands that were 400 base pairs (bp) long, and that (1) the gene that encodes the 10 kDa prolamin has no intron, and (2) the gene copies that were amplified were homogeneous in size. A comparison of the nucleotide sequences for eight strains of *O. rufipogon* (genome type=AA) and one strain of *O. longisteminata* (AA) yielded nearly identical matches to *O. sativa* (AA).

In a study of the gene that encodes the 10 and 16 kDa prolamin polypeptides in African rice (*Oryza glaberrima*) and its related species (*O. barthii*)), Hilu and Sharova (unpublished data) concluded that the genes that encode the 10 kDa prolamin in *O. barthii*, *O. glaberrima* and *O. sativa* possess limited heterogeneity at the DNA level. Sequence similarities of 99% were found between *O. barthii* and *O. glaberrima*. In contrast, a comparison of sequences for the gene that encodes the 16 kDa polypeptide among the same three species yielded 76.0-100% similarity.

Masumura *et al.* (1990) cloned and characterized a cDNA that coded for an *Oryza sativa L.* japonica cultivar 13 kDa prolamin polypeptide. In comparison to the sequenced gene that encodes the 10 kDa polypeptide, the 13 kDa polypeptide (which constitutes the majority of prolamin polypeptides) consisted of 156 amino acids (Shuyr *et al.*, 1992) with a high content of glutamic acid/ glutamine, aspartic acid/ asparagine, and leucine (Hibino *et al.*, 1988). The clones

exhibited a great amount of sequence heterogeneity, as well as a large copy number, indicating that the prolamins were encoded by multi-gene families.

Kim and Okita (1988a) screened two cDNA libraries of *Oryza sativa* L., isolating clones that encode rice prolamins 15.5 kb in length (1988a). Clones were divided into two classes, which differed mainly by point mutations and insertions or deletions of nucleotides. When the coding sequences for genomic and cDNA were compared, they exhibited a high degree of similarity, except for the 5' ends, which were divergent. Kim and Okita (1988a) further noted that unlike most cereal prolamins, rice prolamins lack repetitive sequences.

***Evolution of the Prolamin Polypeptide***

Data regarding the tertiary structure and sequence of rice seed prolamins have been forthcoming over the past decade. Amino acid analysis of prolamins was used to examine the evolution of several types of cereals by Bietz (1982), who concluded that mutations and duplications of a single ancestral gene gave rise to modern cereal varieties. Specifically, Bietz noted that there were small amino acid sequence differences within genera (such as in corn and gamma grass), but larger differences between other tribes such as the Triticeae and Aveneae (oat). These data imply that prolamins originated from a single ancestral gene that was duplicated and mutated among different species. Bietz (1982) further predicted that prolamins might be useful in reconstructing evolutionary relationships in cereals based on studies of sequence similarities.

Based on patterns observed in the amino acid sequences, Shewry *et al.* (1995) concluded that "members" within the Cereal Prolamin Superfamily arose from a single ancestral protein.

Evidence supporting this notion comes from the presence of three related conserved regions in rice, oat, maize, wheat, and barley, which are believed to have arisen through the triplication of a single domain in the ancestral protein (Shewry *et al.*, 1995). The amino acid variability present within the gliadin (S-rich) is due to the insertion of variable regions between the three conserved regions of the protein. Due to their small size, *Oryza* prolamins may be closely related to the ancestral protein from which the cereal prolamin Superfamily arose.

***Evolution of the Genes that Encode Prolamin***

Combined with the absence of sequence similarity between rice, wheat (gliadin) or maize (zein) prolamins, it was concluded that rice prolamin genes evolved from a segment of DNA that was distinct from other major cereals, such as wheat or corn (Kim & Okita, 1988b), and that rice may have diverged prior to the rest of the grasses (Kim & Okita, 1988a). Sequence analysis of a rice cDNA clone (~15.4 kDa) indicated the presence of a single, long open reading frame which coded for 149 amino acids (1988b). Shuyr *et al*. (1992) sequenced a cDNA copy of the 16 kDa prolamin which yielded a polypeptide that consisted of 158 amino acids. The gene that encodes the 16 kDa prolamin gene in *O. barthii* and *O. glaberrima* was sequenced by Hilu & Sharova (unpublished data). The deduced amino acid sequences were approximately 137 amino acids in length for both species; however, the sequences themselves were variable.

***Systematics Implications of the Prolamin Polypeptide***

Hilu and Esen (1988) used prolamins as molecular characters for systematic and evolutionary studies of the grass family Poaceae. They concluded that prolamin molecular weights range from 65 kDa in the Pooideae, to 20-26 kDa in the Panicoideae, Arundinoideae,

and the Chloridoideae, and 15 kDa in the Oryzoideae and the Bambusoideae. From these data, three major groups of grasses were defined based on prolamin size, the: (1) Pooideae, (2) Panicoideae-Chloridoideae- Arundinoideae, and (3) Bambusiudeae-Oryzoideae. Recent molecular studies have confirmed the presence of these evolutionary lineages (Hilu and Alice, in press)**.**

***Rice Taxonomy***

The genus *Oryza* contains approximately 22 species, encompasses nine genomes, and is a member of the Poaceae (grass family), subfamily Oryzoideae (Aggarwal, *et al*., 1997; Vaughan, 1994*)* (Table 1). *Oryza sativa* and *O. glaberrima* are the two cultivated species; the remaining twenty are wild. Systematically, *Oryza* species have been placed within four species complexes, the: *O. sativa* complex, *O. officinalis* complex, *O. ridleyi* complex, and *O. meyeriana* complex (Vaughan, 1994).

Lu's modification (B-R, Lu, personal correspondence, 1999) of Vaughan's (1994) designation of the *Oryza* complexes consists of: Section padia (Series 1. Meyerianeae: *O. meyeriana*, *O. granulata, O. neocaledoica*; Series 2. Ridleyanae: *O. longiglumis*, *O. ridleyi;* Series 3. *O. Schlechterianae*:), Section brachyantha (Series 4. brachyantheae), and Section Oryza (Series 5. latifoliae: *O. alta, O. eichingeri*, *O. grandiglumis*, *O. latifolia*, *O. minuta*, *O. officinalis*, *O. punctata*, *O. rhizomatis*; Series 6. Australiensis: *O. australiensis*; Series 7: Sativae: *O. barthii*, *O. glaberrima, O. glumaepatula*, *O. longisteminata*, *O. meridionalis*, *O. nivara*, *O. rufipogon*, *O. sativa)* (Table 2). A comparison of these two taxanomic treatments (Vaughan, 1994; Lu personal communication) is presented in Chapter 2.

*Oryza* species are worldwide in distribution and inhabit a wide variety of habitats, such as savanna, woodland, swamps, evergreen and deciduous forests, fresh-water lagoons, deep water, shallow water, stagnant pools, and slowly moving water (Vaughan, 1994). Rice species have adapted to these environments by growing to a variety of heights (<1M to >3M), and tolerating extremes in moisture levels and sun exposure from full sun to full shade (Vaughan, 1994).

*Oryza* is a member of the tribe *Oryzeae*, which contains twelve genera (Vaughan, 1994). Among these genera are three species: *Leersia virginica*, *Hygrorhyza aristita*, and *Zizania aquatica* (American wild rice) which were selected as outgroup taxa (representatives of the closest relatives outside the genus under examination which link that genus to the overall tree of life) in this analysis. *Oryza* is thought to be closely related to bamboo (Bambusoideae) (Clark *et al.*, 1995; Esen and Hilu, 1989; Tzvelev, 1989; Hilu and Esen, 1988).

***Oryza and its Genomes.***

The somatic chromosome numbers of the diploid and allotetraploid *Oryza* species are 2n=24 and 2n=48 respectively, with a basic chromosome number of x=12 (Vaughan, 1994). Nine genomes (AA, BB, CC, BBCC, CCDD, EE, FF, GG, HHJJ) have been identified in *Oryza* (Table 1) (reviewed in Vaughan, 1994; Aggarwal *et al.*, 1997) (Fig. 5). These genome designations have been based on chromosome pairing of hybrid plants (Morinaga, 1943; Morinaga & Kuriyama, 1959; Aggarwal et al., 1997). Species that contain: the AA genome consist of: *Oryza sativa*:*O. nivara*, *O. rufipogon*, *O. barthii*, *O. perennis*, *O. glaberrima,* and *O.*

Table. 1. *Oryza* species examined in this study grouped by genome type (reviewed in Vaughan, 1994; Aggarwall et al., 1997). USDA= United States Department of Agriculture; IRRI= International Rice Research Institute, Philippines; and the Laboratory of Dr. Y. Sano, Hokkaido University, Japan all provided DNA material for this analysis.

| Species Name | Genome | Source and Geographic Origin |
|---|---|---|
| *Oryza sativa* | AA | U.S.A. (Pioneer Valley Seed) |
| *O. barthii* | AA | Sierra Leone (USDA), Guinea (USDA) |
| *O. rufipogon* | AA | Taiwan (USDA), Myanmar (USDA), India (USDA) |
| *O. glaberrima* | AA | Nigeria (USDA), Ghana (USDA) |
| *O. nivara* | AA | India (USDA) |
| *O. meridionalis* | AA | Australia (IRRI) |
| *O. officinalis* | CC | Philippines (USDA); DNA, Y. Sano |
| *O. rhizomatis* | CC | Sri Lanka (IRRI) |
| *O. eichingeri* | CC | Sri Lanka (IRRI); DNA, Y. Sano |
| *O. latifolia* | CCDD | Guatemala (IRRI) |
| *O. alta* | CCDD | no source (IRRI) |
| *O. grandiglumis* | CCDD | Brazil (IRRI) |
| *O. punctata* | BBCC, BB | DNA, Y. Sano |
| *O. minuta* | BB | Philippines (IRRI) |
| *O. australiensis* | EE | Australia (IRRI); DNA, Y. Sano |
| *O. brachyantha* | FF | Cameroon (IRRI) |
| *O. meyeriana* | GG | Philippines (IRRI) |
| *O. granulata* | GG | Laos (IRRI) |
| *O. ridleyi* | HHJJ | Thailand (IRRI) |
| *O. longiglumis* | HHJJ | Indonesia (IRRI) |
| *O. schlechteri* | unknown | not sampled |

Table 2. Lu's (1999) modification of Vaughan's (1994) classification of *Oryza*.

| **Section** | **Series**<br>*Species* |
|---|---|
| **Padia (Zoll. et Mor)** | ***Meyerianae*** Sharma et Shastry<br>*O. granulata*<br>*O. meyeriana* (Roll. et Mor. ex Steud.) Baill.<br>*O. neocaledonica* Morat |
| | ***Ridleyanae*** Sharma et Shastry<br>*O. longiglumis* Jansen<br>*O. ridleyi* |
| | ***Schlechterianae*** Sharma et Shastry<br>*O. schlechterii* |
| **Brachyantha** | ***Branchyanthae***<br>*O. brachyantha* |
| **Oryza** | ***Latifoliae***<br>*O. alta*<br>*O. eichingeri*<br>*O. grandiglumis*<br>*O. latifolia*<br>*O. minuta*<br>*O. officianlis*<br>*O. punctata*<br>*O. rhizomatis* |
| | ***Australiensis***<br>*O. australiensis* |
| | ***Sativae***<br>*O. barthii*<br>*O. glaberrima*<br>*O. longiglumis*<br>*O. meridionalis*<br>*O. rufipogon*<br>*O. sativa* |

*meridionalis;* the CC genome: *O. rhizomatis* and *O. eichingeri*; the BBCC: *O. minuta* and *O. punctata*; the CCDD genome: *O. latifolia*, *O. alta*, *O. grandiglumis*; the EE*: O. australiensis*; the

FF: *O. brachyantha*; the GG: *O. granulata* and *O. meyeriana*; and the HHJJ*: O. longiglumis*, and *O. ridleyi* (Table 1).

Chromosome pairing analyses of the tetraploid species *O. minuta* (2n=48) designated it as an allopolyploid hybrid between *O. sativa* (AA) and *O. officinalis* (CC) (reviewed in Nayar, 1975). Experimental evidence supports the hypothesis that *O. minuta* (BBCC) could have arisen *in situ* through the "interracial hybridization" of *O. officinalis* strains as demonstrated by Hu and Chang (1967) when they crossed *O. officinalis* with seven geographically distinct strains of the same species (*O. officinalis*) and isolated sterile hybrids with "48 or more chromosomes" at a ratio of 2:31 (reviewed in Nayar, 1973). A numerical taxanomic analysis (Morishima and Oka, 1959) of 16 *Oryza* species further supports the body of evidence that indicate a close relationship between *O. minuta* and *O. officinalis* exists. A phenetic analysis by McIntyre *et al.* (1992) provides evidence that a close relationship between *O. minuta* (BBCC) and *O. officinalis* (CC) is present, and further indicated that the tetraploid species *O. punctata* (BBCC) forms a cluster with the above mentioned species.

The genomic donors of the polyploid (CCDD) species of South America (*O. latifolia, O. grandiglumis*, and *O. alta*) have been subjects of investigation for several decades. Chromosome pairing experiments of hybrids between *O. minuta* (BBCC) and *O. latifolia* (CCDD) indicate that they share one "set" of similar chromosomes (Morinaga, 1943); however, no diploid DD parental species has ever been identified. It should be noted that the DD chromosomes have been identified from a CCDD genome using florescent *in situ* hybridization (FISH) (Fukui *et al.*, 1997). *Oryza punctata* (BB-Africa) and *O. officinalis* (CC-Asia) have been proposed as parental donors for *O. latifolia* (Gopalakrishnan and Sampath, 1966; reviewed in Nayar, 1973)**.** Gopalakrishnan and Sampath (1967) further proposed that the tetraploid CCDD species arose "*in*

*situ* by polyploidy" as a result of a hybridization event involving *O. officinalis* or *O. punctata* that were isolated in South America due to continental drift (reviewed in Nayar, 1975). The other parental species of allotetraploid CCDD species is unknown.

***Evolutionary Patterns of Genome Relationships***

The mechanism by which the gene copies that encode the 10 kDa prolamin in *Oryza* are homogenized to one form is speculated in this analysis to have been a result of concerted evolution. Concerted evolution (Zimmer *et al.*, 1980) refers to the observation that individual genes in multi-copy gene families tend to homogenize so that all gene copies are identical (Zimmer *et al.*, 1980; Arnheim *et al.*, 1980; Zimmer *et al.*, 1988; Hillis *et al.*, 1991; Wendel *et al.*, 1995; Liao, 1999). In other words, the multiple gene copies within an individual are more alike than those among species (Zimmer *et al.*, 1980). The molecular-drive mechanisms by which the process of concerted evolution is thought to operate are biased gene conversion and unequal crossing over, among others (reviewed in Elder & Turner, 1995) .

Zimmer *et al.* (1980) were the first to coin the phrase "concerted evolution". Their work on the region coding the alpha ($\alpha$) and beta ($\beta$) chain in human hemoglobin showed that a lower rate of concerted evolution takes place in the gene coding for the $\beta$ chain of adult hemoglobin than for the $\alpha$ chain. Moreover, the genes that encode the $\alpha$ hemoglobin are more alike within each species than between species, which is the hallmark of concerted evolution. One explanation for these disparate rates of concerted evolution lies in the presence of longer, nonhomologous "flanking sequences" in the $\beta$ gene that act as barriers to intergenic recombination. This is thought to provide an evolutionary advantage by permitting "selection

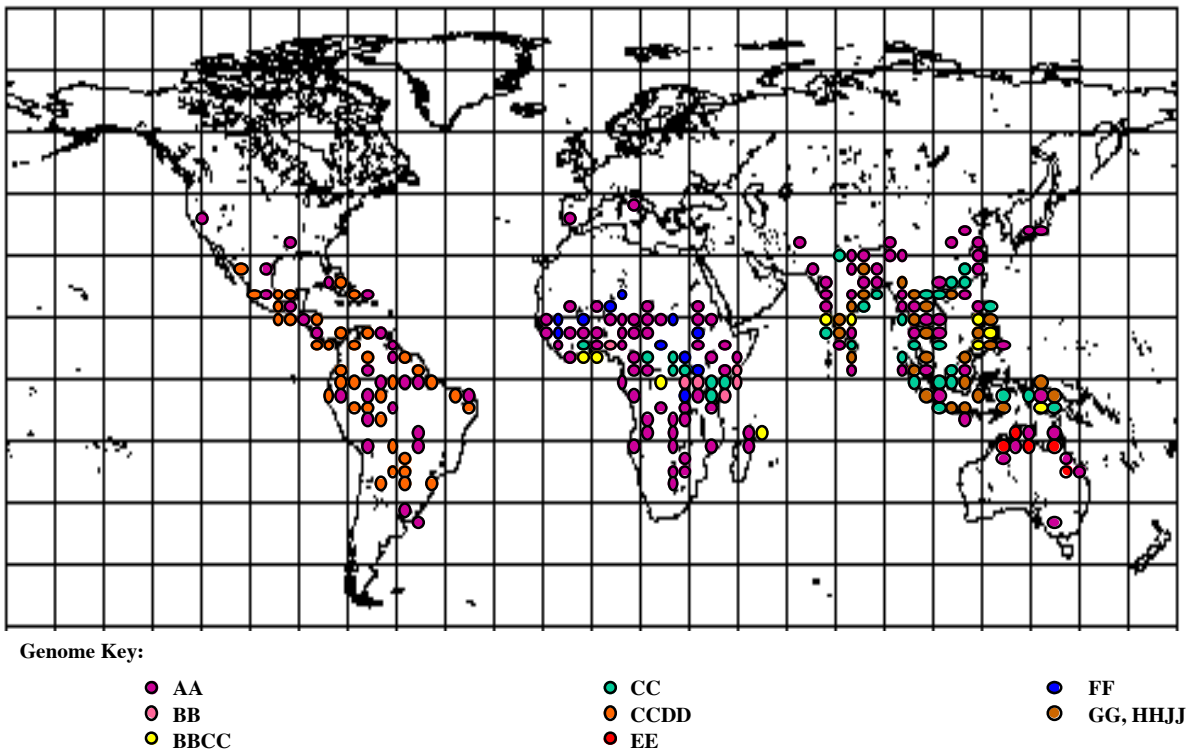



**Fig. 5. Biogeographic distribution of *Oryza* based on genome composition.**

pressure on the rate of hemoglobin synthesis" which is controlled by the number of α genes, yet keeping the rate of unequal crossing over (which is thought to be disadvantageous) in the β chain to a minimum.

An examination of the structure and variation of nuclear genes (>10,000 copies/ haploid genome) that code for ribosomal DNA in maize and teosinte revealed that coding regions of the rDNAs exhibit highly conserved restriction sites (Zimmer et al., 1988). In contrast, polymorphic restriction site variation was observed between $F_1$ and $F_2$ individuals of an inbred line B37N. Therefore, in terms of *individual* rDNA chromosomal arrays, heterogeneity in cleavage site

patterns were observed, indicating that concerted evolution among those arrays has not taken place. In contrast, the intergenic spacer (IGS), a non-coding region of the rDNA units, exhibited the "majority of the restriction site changes". These data indicate that the coding regions of the rDNA in maize and teosinte evolve at slower rates than non-coding regions such as the IGS.

Wendel *et al*. (1995) detected concerted evolution in his study of the genus *Gossypium* (cotton). Their findings indicate that not only does concerted evolution indeed occur in plants, but that it can occur in a bi-directional fashion on different rDNA chromosomal loci (Wendel *et al*., 1995). They examined nucleotide sequences for five allopolyploid cotton species: *G. tometntosum* (AD); *G. hirsutum*, *G. darwinii* (AD*), G. barbadense* (AD), *G. mustelinum* (AD), and the proposed progenitors of these allopolyploids*: G. raimondii* (D), *G. thurberi* (D), *G. arboreum* (A) and *G. herbaceum* (A). It was found that four of the five allopolyploid species formed a clade with the D progenitor species, while one of the allopolyploids (*G. mustelinum*) formed a clade with the A progenitor species. The significance of these findings indicate that multi-copy genes can homogenize to either parental donor set, and that this homogenization can take place between genes at different chromosomal loci.

**Acknowledgments**

Special thanks are extended to Kumiko Koyama for her translation of Higashi *et al*.'s (1974) article from Japanese into English.

## References


Aggarwal, K.; Brar, D.S.; Khush, G.S. 1997. Two new genomes in the *Oryza* complex identified on the basis of molecular divergence analysis using genomic DNA hybridization. *Mol. Genet. Genet.* (254): 1-12.

Argos, P.; Narayana, S.V.L.; Nielson, N.C. 1985. Structural similarity between legumin and vicillin storage proteins for legumes. *EMBO J.* (4): 1111-1117.

Arnheim, N.; Krystal, M.; Schmickel, R.; Wilson, G.; Ryder, O.; Zimmer, E. 1980. Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. *Proc. Natl. Acad. Sci. USA.* (77): 7323-7327.

Barbier, P.; Ishihama, A. 1990. Variation in the nucleotide sequence of a prolamin gene family in wild rice. *Plant Mol. Biol.* (15): 191-195.

Bhatia, C.R.; Rabson, R. 1987. Relationship of grain yield and nutritional quality. *Nutritional quality of cereal grains: genetic and agronomic improvements*: ASA-CSSA-SSSA, Madison: WI, (28): 11-131.

Bietz, A. 1982. Cereal prolamin evolution and homology revealed by sequence analysis. *Biochem. Genet.* (20): 1039-1053.

Bordo, D.; Argos, P. 1991. Suggestions for "safe" residue substitutions is site-directed mutagenesis. *J. Mol. Biol.* (217): 721-729.

Chandler, P. M.; Higgins, T.J.V.; Randall, P.J.; Spencer, D. 1983. Regulation of legumin levels in developing pea seeds under conditions of sulfur deficiency. *Plant Physiol.* (21) 47-54.

Chen, J.-M.; Lin, J.J.; Jiang, S-T.; Chen, C-S. 1995. Characterization of rice prolamin genes.

*Czech-Taiwan Symposium on Biotechnology*, Prague: June 5-8.

Coffman, W.R.; Juliano, B.O. 1987. Rice. *Nutritional quality of cereal grains: genetic and agronomic improvement*. ASA-CSSA-SSSA: Madison, WI (28): 101-131.

deLumen, B.O.; Chow, H. 1991. Nutritional quality of rice endosperm. *Rice*. Van Nostrand Reinhold: New York: (2): 363-395.

Elder, J.F.; Turner, B.J. 1995. Concerted evolution of repetitive DNA sequences in eukaryotes. *Quart. Rev. Biol.* (70): 297-320.

Fukui, K.; Shishido, R.; Kinoshita, T. 1997. Identification of the rice D-genome chromosomes by genomic *in situ* hybridization. *Theor. Appl. Genet.* (95): 1239-1245.

Gopalakrishnan, R.; Sampath, S. 1966. The American species of *Oryza*. *Oryza*. (3): 35-40.

Golpalakrishnan, R.; Sampath, S. 1967. Taxanomic status and origin of American tetraploid species of the series *Latifoliae* Tateoka in the genus *Oryza*. *Indian J. Agr. Sci.* (37): 466-475.

Hibino, T.; Kidzu, K.; Masumura, T.; Ohtsuki, K.; Tanaka, K.; Kawabata, M.; Fujii, S.; 1989. Amino acid composition of rice prolamin polypeptides. *Agric. Biol. Chem.* (53): 513-518.

Higashi, T.; Kushibuchi, K.; Ito, R. 1974. (in Jap.) Studies on breeding for high protein rice. I. Protein content of different rice varieties and their relations with some agranomic traits including yield. *Japan. J. Breed.* (24): 88-96.

Higgins, T.J.V. 1984. Synthesis and regulation of major proteins in seeds. *Ann. Rev. Plant Physiol.* (35): 191-221.

Hillis, D.M.; Moritz, C.; Porter, C.A.; Baker, R.J. 1991. Evidence for biased gene conversion in concerted evolution of ribosomal DNA. *Science*. (252): 308-310.

Hilu, K.W.; Esen, A. 1988. Prolamin size and diversity in the Poaceae. *Biochem. Syst. and Ecol.* (16): 457-465.

Hilu, K.W.; Alice, L.A.; Liang, H. in press. Phylogeny of Poaceae inferred from *mat*K sequences. *Ann. Missouri Bot. Gard.* (86): 835-851.

Hu, C-H.; Chang, C-C. 1967. Cytogenetic studies of *Oryza officinalis* complex-I. $F_1$ hybrid sterility in geographical races of *O. officinalis*. *Bot. Bull. Acad. Sin.* (8): 8-19.

Iida, S.; Amano, E.; Nishio, T. 1993. A rice (*Oryza sativa* L.) mutant having a low content of glutelin and a high content of prolamine. *Theor. Appl. Genet.* (87): 374-378.

Ilag, L.L.; Rosales, A.M.; Elazegui, F.A.; Mew, T.W. 1987. Changes in the population of ineffective endomycorrhizal fungi in a rice-based cropping system. *Plant and Soil*. (103): 67-73.

Kim, W.T.; Okita, T.W. 1988a. Structure, expression, and heterogeneity of the rice seed prolamines. *Plant Physiol.* (88) 649-655.

Kim, W.T.; Okita, T.W. 1988b. Nucleotide and primary sequence of a major rice prolamine. *FEB* (321): 308-310.

Kumamaru, T.; Satoh, H.; Iwata, N.; Omura, T.; Ogawa, M.; Tanaka, T. 1988. Mutants for rice storage proteins. *Theor. Appl. Genet.* (76) 11-16.

Krishnan, H.; White, J.A. 1995. Morphometric analysis of rice seed protein bodies. *Plant*

*Physiol*. (109): 1494-1495.

Leviton, E.; McMahon , K. 1996. Plants as a source of food. *Plants and Civilization*. Wm. C. Brown Publishers: Dubuque, IA: 176.

Li, X.; Okita, T.W. 1993. Accumulation of prolamines and glutelins during rice seed development: a quantitative evaluation. *Plant Cell Physiol*. (34): 385-390.

Liao, D. 1999. Molecular evolution '99. Concerted evolution: molecular mechanism and biological implications. *Am. J. Hum. Genet*. (64): 24-30.

Masumura, T.; Shibata, D.; Hibino, T.; Kato, T.; Kawabe, K.; Takeba, G.; Tanaka, K.; Fujii, S. 1989a. cDNA cloning of an mRNA encoding a sulfur-rich 10 kDa prolamin polypeptide in rice seeds. *Plant Mol. Biol*. (12): 123-130.

Masumura, T.; Kidzu, K.; Sugiyama, Y.; Mitsukawa, N.; Hibino, T.; Tanaka, K.; Fujii, S. 1989b. Nucleotide sequence of a cDNA encoding a major rice glutelin. *Plant Mol. Biol*. (11): 449-461.

Masumura, T.; Hibino, T.; Kidzu, K.; Mitsukawa, N.; Tanaka, K.; Fujii, S. 1990. Cloning and characterization of a cDNA encoding a rice 13kDa prolamin. *Mol. Gen. Genet*. (221): 1-7.

Masamura, T.; Mitsukawa, N.; Tanaka, K.; Fuji, S. 1991. Rice storage proteins: genetic analysis of accumulation process. *Biotechnology in agriculture and forestry*. Springer-Verlag: Heidelberg. (14): 495-507.

McIntyre, C.L.; Winberg, B.; Houchins, K.; Appels, R.; Baum, B.R. 1992. Relationships between *Oryza* species (Poaceae) based on 5S DNA sequences. *Pl. Syst. Evol*. (183): 249-264.

Morinaga, 1943. Cytogenetical studies *of Oryza sativa* L. VI. The cytogenetics of $F_1$ hybrids of *O. minuta Presl.* and *O. latifolia* Desv. *Jap. J. Bot.* (12): 347-357.

Morinaga, T.; Kuriyama, H. 1959. A note on the cross results of diploid and tetraploid rice plants. *Jap. J. Breed.* (9): 187-193.

Morishima, H.; Oka, H.I. 1959. The pattern of interspecific variation in the genus *Oryza*: its quantitative representation by statistical methods. *Evolution.* (14): 153-165.

Nayar, N.M. 1973. Origin and cytogenetics of rice. *Advances in genetics.* New York: Academic Press. (14): 153-292.

Ogawa, M.; Kumamaru, T.; Satoh, H.; Iwata, N.; Omura, T.; Kasai, Z.; Tanaka, K. 1987. Purification of protein body-1 or rice seed and its polypeptide composition. *Plant Cell Physiol.* (28): 1517-1527.

Okita, T. W. 1991. The identification and characterization of rice nuclear genes. Rice *Biotechnology*. CAB International: Oxon, U.K.: 199-224.

Pingali, P.L.; Hossain, M.; Gerpacio, R.V. 1997. Introduction: The state of rice in post-green revolution Asia. *Asian rice bowls: the returning crisis*? CAB International: 1-11.

Randall, P.J.; Thomson, J.A.; Schroeder, H.E. 1979. Cotyledonary storage proteins in *Pisum sativum.* IV. Effects of sulfur, phosphorus, potassium, and magnesium deficiencies. *Aust. J. Plant. Physiol.* (6): 11-24.

Shewry , P.R.; Napier, J.A.; Tatham, A.S. 1995. Seed storage proteins: structures and biosynthesis. *Plant Cell.* (7): 945-946.

Shuyr, L.-F.; Wen, T.-N.; Chen, C.-S. 1992. cDNA cloning and gene expression of the major prolamins of rice. *Plant Mol. Biol.* (20): 323-326.

Solaiman, M.Z.; Hirata, H. 1997. Effect of arbuscular mycorrhizal fungi innoculation of rice seedlings at the nursery stage upon performance in the paddy field and greenhouse. *Plant and Soil.* (191): 1-12.

Tanaka, K.; Sugimoto, T.; Ogawa, M.; Kasai, Z. 1980. Isolation and characterization of two types of protein bodies in the rice endosperm. *Agric. Biol. Chem.* (44): 1633-1639.

Toenniessen, G.H. 1991. Potentially useful genes for rice genetic engineering. *Rice Biotechnology.* CAB International: Oxon, UK: 253-280.

Tsai, C.Y; Huber, D.M.; Warren, H.L. 1980. A proposed role of zein and glutelin as N sinks in maize. *Plant Physiol.* (66): 330-333.

Tzvelev, N. N. (1989). The system of grasses (Poaceae) and their evolution. *The Bot. Rev.* (55): 141-204.

Vaughan, D. 1994. *The wild relatives of rice: A genetic resources handbook.* International Rice Research Institute, Los Baños, Philippines: 1-101.

Watson, L.; Clifford, H.T.; Dallwitz, M.J. 1985. The classification of Poaceae: Subfamilies and supertribes. *Aust. J. Bot.* (33): 433-484.

Wendel, J.F.; Schnabel, A.; Seelanan, T. 1995. Bidirectional concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proc. Natl. Acad. Sci. USA.* (92): 280-284.

Zhou, X.; Fan, Y.-L. 1993. The endosperm-specific expression of a rice prolamin chimaeric

gene in transgenic tobacco plants. *Transgenic Res*. (2): 141-146.

Zimmer, E.A.; Martin, S.L.; Beverly, S.M.; Kan, Y.W.; Wilson, A.C. 1980. Rapid duplication and loss of genes coding for the α chains of hemoglobin. *Proc. Natl. Acad. Sci. USA*. (77): 2158-2162.

Zimmer, E.A.; Jupe, E.R.; Walbot, V. 1988. Ribosomal gene structure, variation and inheritance in maize and its ancestors. *Genetics*. (120): 1125-1136.

**-Chapter 2-**

# VARIATION IN THE GENE ENCODING THE 10 kDa PROLAMIN IN *ORYZA* (Poaceae) AND ITS PHYLOGENETIC IMPLICATIONS

Irene M. Mullins and Khidir W. Hilu

Virginia Polytechnic Institute and State University

## Introduction

The genus *Oryza* contains approximately 22 species, encompasses nine genomes, and is a member of the Poaceae (grasses) (Aggarwal *et al*., 1997; reviewed in Vaughan, 1994*)*. *Oryza sativa* and *O. glaberrima* are the two cultivated species; the remaining twenty are wild. All *Oryza* species have been placed within three sections, which have been further divided into four species complexes, the: (1) O. sativa complex, (2) O. officinalis complex, (3) O. ridleyi complex, and (4) O. meyeriana complex (Vaughan, 1994) (Table 1). This classification system has been recently modified (Lu, personal correspondence, 1999) (Table 2), to reflect standard nomenclature terminology which does not utilize the "complex" epithet, among other changes. Currently, no phylogenetic examination that is inclusive of all *Oryza* genomes exists using a molecular approach at the gene level.

The objective of this study is to evaluate the molecular variation present in the gene that encodes the 10 kDa prolamin polypeptide in *Oryza*. Our principle hypothesis is that genetic diversity exists at the molecular level in the wild species of the genus *Oryza*, and that this diversity can provide basic information regarding intragenomic relationships among these

Table 1. Classification of *Oryza* based on Vaughan (1994). Note that two species, *O. brachyantha* and *O. schlechteri*, have not been placed within a complex.

| ***Section*** | Complex | *Species* |
|---|---|---|
| ***Oryza*** | O. sativa complex | *O. sativa* L |
| | | *O. nivara* (Sharma et Shastry) |
| | | *O. rufipogon* Griff. |
| | | *O. glaberrima* Steud. |
| | | *O. barthii* A. Chev. |
| | | *O. longisteminata* Chev. et Roehr. |
| | | *O. meridionalis* Ng |
| | O. officinalis complex | *O. officinalis* Wall ex Watt |
| | | *O. minuta* Presl. et Presl. |
| | | *O. rhizomatis* Vaughan |
| | | *O. eichingeri* Peter |
| | | *O. punctata* Kotschy ex Steud. |
| | | *O. latifolia* Desv. |
| | | *O. alta* Swallen |
| | | *O. grandiglumis* (Doell) Prod |
| | | *O. australiensis* Domin |
| ***Ridleyanae*** Tateoka | | *O. brachyantha* Chev. et Roehr |
| | | *O. schlechteri* Pilger |
| | O. ridleyi complex | *O. ridleyi* Hook. f. |
| | | *O. longiglumis* Jansen |
| ***Granulata*** Roschev. | O. meyeriana complex | *O. meyeriana* (Zoll. et Mor. ex Steud.) Baill. |
| | | *O. granulata* Nees et Arn. ex Watt. |

species. Specifically, the gene that encodes the 10 kDa prolamin polypeptide was used as a molecular marker in this study to examine phylogenetic relationships within *Oryza*. The gene has been shown to be conserved, and thus useful at the generic level.

***Prolamin and their encoding genes***

Isolation and characterization of the gene used in this phylogenetic analysis was performed by Masumura *et al*. (1989) who isolated cDNA that coded for a sulphur-rich 10 kDa

rice prolamin. The nuclear gene that encodes the 10 kDa prolamin polypeptide is 402 base pairs (bp) in length (Masumura *et al.*, 1989a), and has no intron (Barbier & Ishihama, 1990). The coding region of the gene consists of a signal peptide region, which is comprised of 84 bp (Masumura *et al.*, 1989a). In addition, this gene exists in multiple copies, perhaps as many as 80-100 copies per haploid genome (Kim & Okita, 1988), which codes for a polypeptide consisting of 134 amino acids that lacks major repetitive sequences (Masumura *et al*, 1989).

Barbier and Ishihama (1990) further examined the variation present in the DNA sequence of a prolamin gene family that codes for the 10 kDa prolamin polypeptide within *O. rufipogon* by PCR amplification and directly sequencing. Little nucleotide variation was found among *O. rufipogon* strains or between *O. rufipogon* and *O. longisteminata* utilizing this method, indicating homogeneity at the locus level.

Prolamins are a class of storage protein found in the starchy endosperm of the grain. The molecular weights of prolamin in the grass family range from 10-65 kDa (Hilu & Esen, 1988). Rice contains prolamins of low molecular weight, which are organized into three size classes, the: 10, 13, and 16 kDa (Barbier & Ishihama, 1990). It is currently thought that these three protein classes are encoded by three multi-gene families (Kim and Okita, 1988a). Other members of the grass family, such as oats (22.5 kDa), teosinte (27.7 kDa), and sorghum (22 kDa) all contain prolamins of higher molecular weights than those found in rice (Bietz, 1982).

Prolamins are characterized by their solubility in alcohol-water solutions unlike the other classes of seed storage proteins which each contain their own solubility characteristics: albumins (water soluble), globulins (salt soluble), and glutelins (dilute alkali or acid soluble) (Shewry *et*

Table 2. Classification of *Oryza* based on Lu (personal correspondence, 1999).

| ***Section*** | Series | *Species* |
|---|---|---|
| ***Padia*** (Zoll. et Mor.) Baill. | Series Meyerianae | *O. granulata*<br>*O. meyeriana* (Roll. et Mor. ex Steud.) Baill.<br>*O. neocaledonica* Morat |
| | Series Ridleyannae | *O. longiglumis* Jansen<br>*O. ridleyi* |
| | Series Schlechterianae | *O. schlechterianae* |
| ***Brachyantha*** | Series Brachyanthae | *O. brachyantha* |
| ***Oryza*** | Series Latifoliae | *O. alta*<br>*O. eichingeri*<br>*O. grandiglumis*<br>*O. latifolia*<br>*O. minuta*<br>*O. officinalis*<br>*O. punctata*<br>*O. rhizomatis* |
| | Series Australiensis | *O. australiensis* |
| | Series Sativae | *O. barthii*<br>*O. glaberrima*<br>*O. glumipaela*<br>*O. longiglumis*<br>*O. meridionalis*<br>*O. nivara*<br>*O. rufipogon*<br>*O. sativa* |

*al.*, 1995)**.** The function of prolamins is to act as a storage source for excess nitrogen and sulphur, two compounds that are required during seed germination (reviewed in Zhou *et al.*, 1993; Higgins, 1984). Prolamins accumulate after other metabolic requirements are met, and are

deposited in protein bodies (PB-1) within the starchy endosperm of the rice grain (Tanaka *et al*., 1980) (Bhatia & Rabson, 1987).

The quantity of prolamin assigned to rice has fluctuated over the last decade as a result of differing extraction methods. Prolamin constitutes 5% of the protein fraction of rice endosperm when 70% ethanol is used as an extraction method (reviewed in Masumura *et al*., 1991) and 20% when 55% n-propanol is used as a solvent (Masumura *et al*., 1989). In contrast to these earlier estimates, a recent morphometric analysis has placed the content of prolamin in rice seeds as high as 35% (Krishnan & White, 1995).

Rice is considered to have "one of the highest quality proteins among cereals" because of its low prolamin content (Coffman & Juliano, 1987), even though the 10 kDa prolamin polypeptide in rice is rich in cysteine (6.8 pmol/100 mol) and methionine (12.63 mol/100mol) (both are sulphur-containing amino acids), and has high contents of proline and threonine (8.59 and 8.64 mol/100 mol respectively) (Hibino *et al*., 1989). Other cereals that have higher quantities of prolamin are considered to be less nutritious than rice because of the low nutrition value of prolamin. Prolamin is rich in hydrophobic and uncharged amino acids, which includes glutamine (reviewed by Chen *et al*., 1995); however, it is poor in essential charged amino acids such as lysine and tryptophan (Coffman & Juliano, 1987). This lack of charged amino acids results in the poor nutritional quality for grains that are composed primarily of prolamin (maize, sorghum, wheat, and barley), and consequently accounts for the low overall nutritional quality of the major cereal grains compared to rice.

Bietz (1982) examined amino acid sequences of several types of cereals and concluded that prolamins likely originated from a single ancestral gene that had undergone various types of duplications and mutations among different species; however, rice yielded poor sequence data

due to blocked amino terminal residues. Based on patterns observed in amino acid sequences, Shewry *et al.* (1995) likewise concluded that "members" within the Cereal Prolamin Superfamily arose through the evolution of a single ancestral protein. Evidence they presented in support of this notion are the presence of three related conserved regions in rice, oat, wheat, and barley, which they consider to have arisen through the triplication of a single domain within the ancestral protein (Shewry *et al.*, 1995). *Oryza* prolamins are the smallest among cereal crops, indicating that they may be closely related to the ancestral protein from which the Prolamin Superfamily later arose.

Previous phenetic and cladistic treatments of *Oryza* species have established patterns in species relationships (Morishima & Oka, 1960; Second, 1985; McIntyre *et al.*, 1992). In a phenetic analysis based on 42 morphologic (non-molecular) characters, Morishima & Oka (1960) demonstrated the presence of two clusters, one containing: *O. granulata*, *O. minuta*, *O. officinalis, O. malabarensis*, *O. alta*, *O. latifolia*, *O. eichingeri*, and *O. australiensis*, and the other containing: *O. brevilugata*, *O. glaberrima*, *O. sativa*, *O. perennis*, *and O. sativa spontanea*. Interestingly, *O.ridleyi* (O. ridleyi complex *sensu* Vaughan, 1994), *O. brachyatha*, and *O. subulata* did not form a cluster with the other *Oryza* species.

Second (1985) conducted a multivariate analysis utilizing isozyme polymorphism of the Sativa (AA) *group (O. sativa, O. glaberrima, O. rufipogon, O. barthii,* and *O. brevigulata*). It was concluded that the group was composed of a complex species of *O. rufipogon* with South Asian, Chinese, New Guinean, Australian and American forms, and two African-endemic species*: O. longisteminata* (*O. barthii*) and *O. brevilgulata* (*O. barthii*). These results were congruent with Vaughan's classification of the Sativa Complex (1994).

Relationships among *Oryza* species were also examined by McIntyre *et al.* (1992) based on 5S DNA sequences using both phenetic and cladistic approaches. Sequence similarities were determined for nine *Oryza* species, representing six genomes. *Oryza brachyantha* formed a distinct group and the remaining species fell within: (1) an AA genome cluster, or (2) a cluster that contained the B/C/D/E genomes (McIntyre *et al.*, 1992). A cladistic analysis of the same data set, rooted with *O. australiensis,* resulted in the formation of three clades: (1) *O. latifolia*, *O. minuta*, (2) *O. nivara*, *O. rufipogon, O. sativa* (Sativa complex species), and (3) *O. officinalis*, *O. punctata*, and *O. brachyantha*. Members of the O. officinalis complex (*sensu* Vaughn, 1994) did not appear monophyletic.

While all of these studies have examined interspecific relationships among *Oryza* species, none have cladistically examined a gene-derived molecular data set that encompasses the majority of the genus. Our molecular analysis includes all *Oryza* species (except for *O. ridleyi and O. schelchterii)*, and critically examines past analyses for congruence and conflict.

## Materials and Methods

### ***Seed Acquisition and Plant Growth***

Nineteen species, represented by 24 accessions, were examined in this study. The species name, accession number, source of material, genomic constitution, and geographic origin are cited in Table 3. All rice grain for this study (except for *O. sativa*) was supplied by the International Rice Research Institute (IRRI) in the Philippines, or by the United States

Department of Agriculture (USDA) (Table 3). *Oryza sativa* grain was supplied by the Pioneer Valley Seed Company. Grains were germinated on moist, filter paper-lined Petri plates until seedling growth appeared vigorous, at which time seedlings were transferred to individual pots containing a clay/ organic soil mixture. Pots were placed in 3 inch deep pans of water in a greenhouse, and were treated every two weeks with acid fertilizer (MirAcid, 30:10:10, 2 teaspoons [tsp.]/ gallon) and iron supplements (Earthcare Secuestren Iron, 1/4 teaspoon/ gallon). Voucher specimens are located at IRRI (Philippines) and the Virginia Tech Massey Herbarium (VT). All IRRI seed lots used in this project have been positively identified by Dr. B.R. Lu, a germplasm specialist at IRRI, with vouchers at IRRI.

***Genomic DNA Extraction***

DNA was isolated from leaf material via Hilu's modification (1994) of the CTAB extraction method of Saghai-Maroof *et al.* (1984). Briefly, 700 microliters (µls) of CTAB (100mM Tris HCl, 2% CTAB, 20 mM EDTA, 1.4 M NaCl) were heated at 65˚C in a 1.5 ml Eppendorf tube with 1.5 µls Beta-Mercaptoethanol (BME) (Fisher Scientific). Approximately one gram of fresh leaf material was ground in a chilled mortar and pestle with liquid nitrogen, placed in the CTAB+ BME solution, and incubated at 65˚C for thirty minutes (min.). Seven hundred µls of chloroform were added, and then the tubes were spun at 15K, room temperature (rt) for 15 min. For some samples a second chloroform extraction was necessary. Four hundred µls of isoproponol were added, and the samples were then precipitated at -20˚C overnight, then centrifuged at 4˚C for 23 min. at 15K. Pellets were rinsed with 350 µl of 70% ethanol, and

Table 3. Description of *Oryza* accessions examined in this study. Seed accessions obtained from IRRI (=the International Rice Research Institute, Philippines) are abbreviated IRRI, those obtained from the United States Department of Agriculture (USDA) are abbreviated PI =Plant Introduction Number. DNA obtained from the laboratory of Dr. Y. Sano, Hokkaido University, Japan, is abbreviated W. Lyophilized plant material obtained from the University of Bonn, Germany, is abbreviated B.

| ***Taxon*** | **Genome** | **KH #** | **Accession #** | **Geographic Origin** | **Primer Pair used for sequencing** |
|---|---|---|---|---|---|
| *Oryza sativa* | AA | KH7000 | ---------- | United States | PR10.1F2/PR10.3RV |
| *O. nivara* | AA | KH7035 | IRRI 101524 | India | PR10.1F2/PR10.3RV |
| *O. rufipogon* | AA | KH7031 | IRRI 100907 | Taiwan | PR10.1F2/PR10.3RV |
| *O. rufipogon* | AA | KH7032 | PI 590418 | Myanmar | PR10.1F2/PR10.3RV |
| *O. rufipogon* | AA | KH7037 | PI 239671 | India | PR10.1F2/PR10.3RV |
| *O. glaberrima* | AA | KH7004 | PI 450198 | Nigeria | PR10.1F2/PR10.3RV |
| *O. meridionalis* | AA | KH7015 | IRRI 101145 | Australia | PR10.1F2/PR10.3RV |
| *O. meridionalis* | AA | KH7027 | IRRI 105279 | Australia | PR10.1F2/PR10.3RV |
| *O. officinalis* | CC | ---------- | W0002 | Thailand | PR10.1F2/PR10.3RV |
| *O. rhizomatis* | CC | KH7020 | IRRI 103421 | Sri Lanka | PR10.1F2/PR10.3RV |
| *O. rhizomatis* | CC | KH7019 | IRRI 103410 | Sri Lanka | PR10.1E/PR10.2Rb |
| *O. eichingeri* | CC | ---------- | W1522 | Uganda | PR10.1F2/PR10.3RV |
| *O. minuta* | BBCC | KH7014 | IRRI 101128 | Philippines | PR10.1F2/PR10.3RV |
| *O. minuta* | BBCC | KH7013 | IRRI 101082 | Philippines | PR10.1F2/PR10.3RV |
| *O. punctata* | BB | ---------- | W1514 | Kenya | PR10.1F2/P10.3RV |
| *O. latifolia* | CCDD | KH7040 | IRRI 100165 | Guatemala | PR10.1E/PR10.2Rb |
| *O. alta* | CCDD | KH7016 | IRRI 101395 | no source | PR10.1F2/PR10.3RV |
| *O. grandiglumis* | CCDD | KH7017 | IRRI 101405 | Brazil | PR10.1F2/PR10.3RV |
| *O. australiensis* | EE | KH7043 | IRRI 101144 | Australia | PR10.1E/PR10.2RB |
| *O. australiensis* | EE | ---------- | W0008 | N. Australia | PR10.1E/PR10.3RV |
| *O. brachyantha* | FF | KH7024 | IRRI 105171 | Cameroon | PR10.1E/PR10.2Rb |
| *O. brachyantha* | FF | KH7025 | IRRI 105172 | Cameroon | PR10.1F2/PR10.3RV |
| *O. meyeriana* | GG | KH7030 | IRRI 106474 | Philippines | PR10.1F2/PR10.3RV |
| *O. granulata* | GG | KH7029 | IRRI 106468 | Laos | PR10.1E/PR10.2RB |
| *Leersia virginica* | | | | United States | |
| *Hygrorhyza aristita* | | | B14589 | | |
| *Zizania aquatica* | | | | | |

lyophilized in a speedvac. DNA was resuspended in 40 µls of 1X TE (10mM Tris Cl [pH 8.0], 1 mM EDTA [pH 8.0]). DNA was quantified against a Lambda (λ) *Hind*III digest via

electrophoresis using a 1.0% agarose gel (Sigma, St. Louis, MO) at 65V for 2.5 hours and visualized with ethidium bromide staining by immersing gel in an ethidium bromide (0.5.µg ethidium bromide/ ml 1X TAE buffer) bath for 25 min., room temperature (rt).

***PCR amplification and alcohol precipitation***

Two sets of primers were used in this study to amplify the gene that encodes the 10 kDa prolamin polypeptide. In most cases, the primers PR10.1F2 (5' ACG TGA ATT CCA CCA TCT GGA ATC TGG 3') and PR10.3RV (ACG TTC TAG AG TGT TTG CAC ACG ATA GTA 3') (primers synthesized by Operon, Inc.) (Fig. 1) were used. In cases when these two primers did not amplify, PR10.1E (5' ACG TGA ATT CAT GGC AGC ATA CAC CAG CAA G 3') and PR10.2RB (ACG TGG ATC CAA CCA CAG GAA GAG AGT TGG 3') were used (Hilu & Sharova, unpublished data). Primers were designed based on the Masumura *et al.* (1989) nucleotide sequence of the gene that encodes the 10 kDa prolamin polypeptide.

The forward primer PR10.1E and reverse primer PR10.2RB are located within the conserved coding region, at the beginning of the gene, and were utilized to amplify species that could not otherwise be amplified with PR10.1F2 and PR10.3RV (Fig.1). Because these primers are located within the gene itself, complete forward and reverse sequence information could not be obtained. Genomic DNA was amplified using and the polymerase chain reaction (PCR) and Taq polymerase (Promega). The amplification conditions used were 40 cycles of: 0.75 min., 94˚C (denaturation); 1 min., X˚C (annealing); and 1 min., 72˚C (extension). Annealing temperatures for the following samples were: (1) *Oryza rhizomatis:* 40˚C, (2) *O. latifolia*: 48˚C (3) *O. australiensis:* 48˚C, (4*) O. brachyantha*: 40˚C, (5) *O. granulata*: 40˚C, (6*) Leersia*

*virginica:* 42˚C, (7) *Zizania aquatica*: 50˚C, and (8) *Hygrorhyza aristita:* 46˚C, and (9) *O. meyeriana* (48˚)**.**

The forward primer PR10.1F2 and the reverse primer PR10.3RV are located outside the coding region of the 10 kDa gene, and therefore provide complete sequence information for the coding region and signal peptide of the gene (Fig 1). The amplification conditions used were 40 cycles of: 0.5 min., 94˚C (denaturation); 1.5 min., 48˚C (annealing); and 1 min., 72˚C (extension) for all of the following samples: (1) *Oryza nivara*, (2) *O. rufipogon* (KH 7030, 7032, 7037), (3) *O. glaberrima*, (4) *O. meridionalis* (KH 7015, 7027), (5) *O. officinalis*, (6) *O. rhizomatis* (KH 7020), (7) *O. eichingeri*, (8) *O. minuta* (KH7013, 7014), (9) *O. punctata*, (10) *O. alta*, (11*) O. grandiglumis*, (12) *O. brachyantha*, and (13) *O. longiglumis.* An annealing temperature of 50˚C was used for *Oryza sativa*.


Fig. 1. A diagram showing the 402 bp gene that encodes the 10 kDa prolamin seed storage protein. The position of the forward/ reverse primer pair PR10.1E/ PR10.2RB or PR10.1F2/ PR10.3RV, signal peptide, and coding region are indicated as shown.

For each species, multiple PCR reaction replicates were pooled and their volume was reduced via a sodium acetate/ ethanol precipitation after the PCR products had been quantified on a 1.% agarose gel (Sigma). The DNA samples were electrophoresed on 1.5% agarose gels for

several hours in order to isolate individual bands. PCR products for *Leersia*, *Hygrorhyza*, and *Zizania aquatica* were pooled as described already and electrophoresed in a 3.0% Metaphor (FMC, Rockland, ME) high-resolution agarose gel at 67V for 8.5 hours (hrs). The appropriate bands were cut out and the DNA was cleaned using the Quiagen Gel Extraction Kit (Quiagen Inc., Valencia, CA).

***Cloning***

PCR products for *Zizania aquatica*, *Leersia virginica*, and *Hygrorhyza aristita* were cloned using the pGEM-T Easy Vector System (Promega, Madison, WI). Two µls of the standard ligation reactions (2:1 insert to vector [pGEM-T Easy] ratio) were transformed into JM109 High Efficiency Competent Cells (Promega, Madison, WI) and incubated in 950 µls of SOC medium (2.0 g Bacto-tryptone; 0.5g Bacto yeast extract; 1 ml 1M NaCl; 0.25 ml 1M KCl; 1 ml 2M $Mg^{2+}$ stock; filter-sterilized [20.33g MgCl2, 24.6g MgSO4, distilled water to 100 ml volume]; 1 ml 2M glucose, filter-sterilized) for 1.5 hrs. at 37˚C, shaking.

One hundred µls of each transformation culture was plated onto LB(per liter: 10g Bacto-tryptone, 5g Bacto-yeast extract, 5g NaCl, pH to 7.0 with NaOH, 15g agar)/ ampicillin (100 µg/ml)/ IPTG (0.5mM)/ X-Gal (80 µg/ml) agar plates, in duplicate. Control plates were prepared as indicated by Promega. Plates were incubated overnight (16-24 hrs.) at 37˚C.

One milliliter (ml) LB medium (10g Bacto-tryptone, 5g bacto-yeast extract, 5g NaCl; pH 7.0) plus ampicillin (150 µg/ml) was inoculated with a white colony selected from the standard reaction plates and incubated for 4 hrs., which was then transferred to 4 mls of LB medium and incubated overnight (12-16hrs) at 37˚C. Tubes were spun at 10K for 5 min., medium was decanted, and tubes were blotted to remove excess liquid. Plasmids were isolated using the

Wizard Plus SV Miniprep System (Promega, Madison, WI) with an additional sodium acetate (7.5M)/ ethanol precipitation. Purified plasmid DNA was eluted in 15 µls Nuclease-Free water (Promega, Madison, WI), and stored at -20˚C.

Plasmid concentration was determined via electrophoresis and comparison with a λ*Hind*III standard. M13 forward (5'-TGTAAAACGACGGCCAG-3' [Operon]) and reverse (5' ATTTAGGTGACACTATAGAATA-3'), plasmid-specific primers were used to PCR amplify the vector and ensure that only plasmids containing inserts were sequenced.

***Nucleotide Sequencing of PCR products***

Approximately 30-60 ng of column-cleaned PCR product (Quiagen. Valencia, CA) were used to prepare half-volume sequencing reactions using the ABI Prism Dye Terminator Cycle Sequencing Ready Reaction Kit with AmpliTaq DNA polymerase, FS or the ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction Kit (Perkin Elmer, Foster City, CA) with an annealing temperature of 50˚C. Samples were prepared according to the manufacturer's protocol. Sequencing reactions were alcohol precipitated with either an ethanol/ sodium acetate or 95% isoproponol protocol as indicated by Perkin Elmer.

DNA was sequenced at the University of Maine (Orono, ME), University of Florida (Gainesville, FL), Virginia Tech Biology Department (Blacksburg, VA) or Virginia/ Maryland Regional College of Veterinary Sciences (Blacksburg, VA) sequencing facilities.

***Nucleotide Sequencing of Isolated Plasmid DNA***

Approximately 500 ng of column-cleaned plasmid DNA (Wizard Plus, Promega, Madison, WI) were used in 20 µL sequencing reactions using the ABI Prism Big Dye

Terminator Cycle Sequencing Ready Reaction Kit (Perkin Elmer) with an annealing temperature of 50˚C. Sequencing reactions were alcohol-precipitated using 95% isoproponol and sequenced at the Virginia Tech Biology Department or University of Maine sequencing facilities.

***Sequence Data Analysis***

DNA sequences were manually edited and aligned using the Sequence Navigator program (Applied BioSystems, Perkin Elmer), given the unambiguous nature of the nucleotide sequence data. Both forward and reverse sequences were examined for the presence of polymorphic signal and the forward sequence was used in the alignment. Gaps were introduced by inserting dashes. Ambiguous nucleotides were coded as: k (G/T), m (A/C), r (G/A), s (C/G), and y (C/T) using standard IUPAC symbols.

Edited sequences were phylogenetically analyzed using Fitch parsimony in PAUP 4.0b3a (Swofford, 1992). Heuristic searches were performed with 1000 random stepwise-addition replicates, "MulTrees" on, and the TBR (tree-bisection-reconnection) branchswapping algorithm for two data sets: one that included the signal peptide and the coding region of the gene, and another that included the gene coding region only. Gaps were treated as missing data except for one unique ATG indel that was manually coded for within the alignment. A strict consensus tree was constructed from the equally parsimonious trees. The one gap that was parsimony-informative was mapped onto the strict consensus tree.

$G_1$ values were created for 1000 random trees using the "evaluate random trees" option in PAUP 4.0b3a (Swofford, 1999) and compared to the values given for 25 taxa with 250 characters (variable positions) with a *P* value of 0.01 (Hillis and Huelsenbeck, 1992) to test for the presence of non-random structure in the data set. Decay indices (Bremmer, 1988; Donaghue

*et al*., 1992), and bootstrap values for 100 replicates, were calculated in order to measure support for individual clades. AutoDecay (Eriksson, 1998) was used to perform the decay analysis.

A Neighbor-Joining analysis was also performed utilizing the Kimura 2-perameter distance estimate on the data set (Paup 4.0b3a [Swofford, 1999]).

Deduced amino acids were determined using Lasergene Navigator (DNA Star). The open reading frame for each of the sequences was determined to exclude the signal peptide region and then translated using the standard genetic code.

## Results

**Nucleic acid analysis of the gene that encodes the 10 kDa prolamin polypeptide.**

The gene that encodes the 10 kDa prolamin polypeptide consists of approximately 402 bases, which includes a signal peptide region consisting of 84 bases. The open reading frame for all *Oryza* species that contain the AA genome is 324 bases. All remaining *Oryza* species contain an ORF of 327 bases (Fig 1). Several species did not have an ATG start codon at position 121 in the sequence layout, rather: *Oryza australiensis* had a GCG, *O. meyeriana* and *O. granulata* had a GTG, and *O. brachyanta* had GTT (Fig. 1). The nucleotide sequence terminated with TGA for all species except for *O. latifolia* (GGA), *O. australiensis* 7043 (GGA), and *O. meyeriana* (GGA). The sequence for *O. granulata* was incomplete and terminated approximately 24 bases prior to the remainig *Oryza* sequences.

*Signal Peptide Included*

The twenty six aligned sequences for the gene and the signal peptide contained 430 characters (Fig 3A), of which 130 were variable (30%). Of the 130 variable characters, 69 were parsimony informative (53%). The –1.1946 $g_1$ value of the *Oryza* data set provides significant ($P$<0.01) evidence of non-random structure. The cladistic analysis yielded 24 most parsimonious trees that were 173 steps in length. The Consistency Index (CI) and Retention Index (RI) were 0.8056 and 0.8833 respectively (Fig 3A). The predicted CI based on the polynomial regression analysis of Sanderson and Donoghue (1989) is 0.3279, which indicates that there are low levels of homoplasy in this data set. The matrix that included the signal peptide will be the one that is highlighted and discussed in this analysis.

*Signal Peptide Excluded*

As described for the matrix that included the signal peptide, 26 sequences were aligned for the gene region excluding the signal peptide, which consisted of 349 total characters (Fig. 2). Fifty-one (53.6%) parsimony informative characters were present among 95 total variable characters (27.2%). Seventy-two most parsimonious trees were found in a cladistic analysis, which were 121 steps in length. The CI and RI values were 0.8267 and 0.9023 respectively (Fig. 3B). The predicted CI is 0.0.3279, indicating that there are low levels of homoplasy in this data set (Sanderson and Donoghue, 1989). The -1.4321 $g_1$ value indicated that there is little ($P$ < 0.01) random structure in the data set.

**Insertions and deletions (Indels) in the gene that encodes the 10 kDa prolamin polypeptide.**

A three base pair indel was present in all AA genome sequences for the gene that encodes the 10 kDa prolamin polypeptide (Fig 2). Two 9 bp gaps were inserted into the *Oryza* sequences to align them to *Phyllostachys aurea* (Bambusoideae).

**Species relationships in *Oryza***

*DNA sequences*

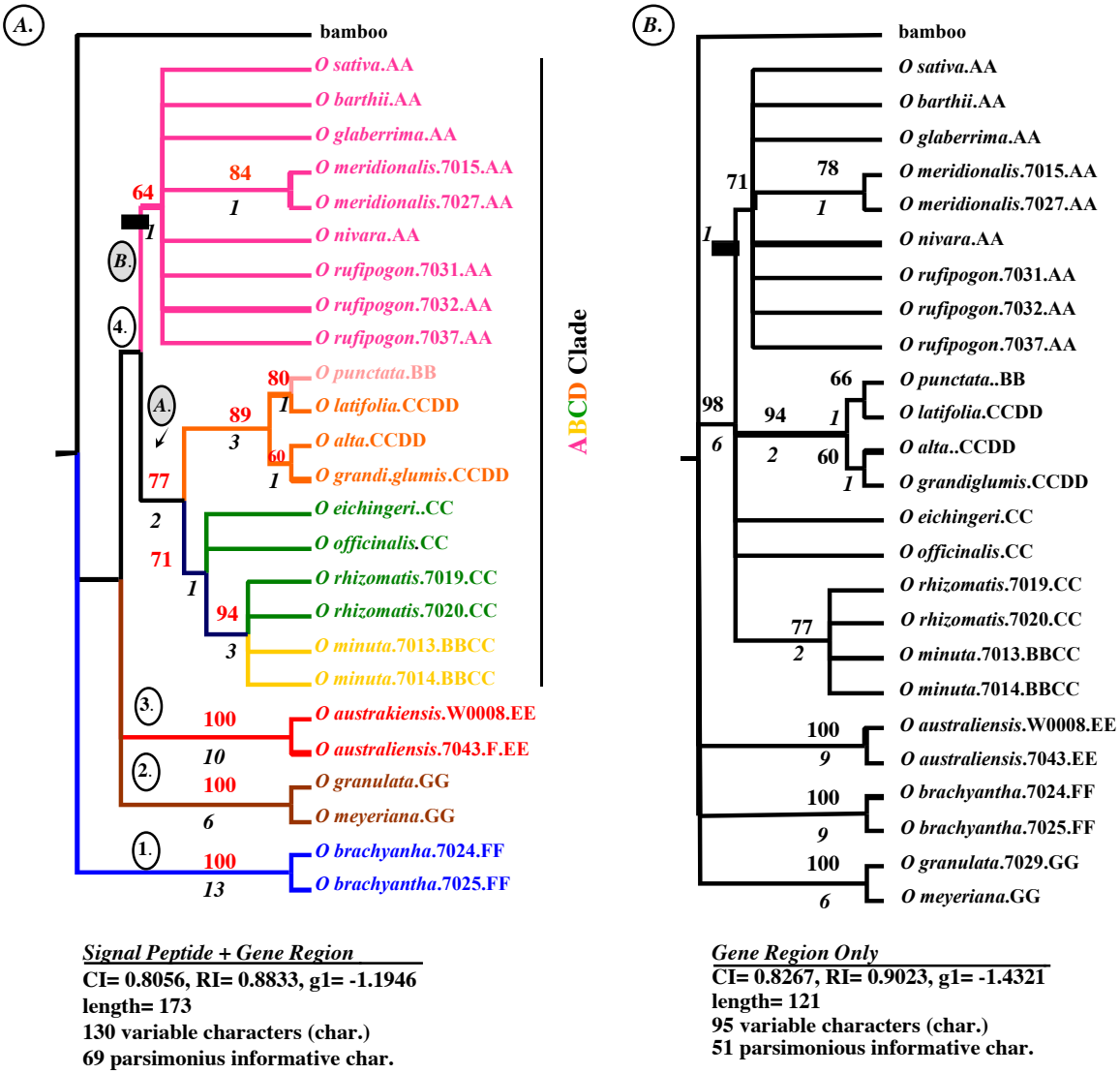
The gene that encodes the 10 kDa prolamin polypeptide in *Phyllostachys aurea* (Bambusoideae) (Hilu and Sharova, 1999) was used as an outgroup in these analyses (Figs 2 & 3). Candidate outgroup taxa originally included *Zizania aquatica* (American wild rice) *Leersia virginica*, and *Hygrorhya aristeta*; however, direct sequencing of PCR products indicated that multiple genetic species for the gene that encodes the 10 kDa prolamin polypeptide were present in each taxon. Individual clones of the PCR products for each of the three species were therefore sequenced. None of the clones were alignable with those of the *Oryza* species (data not shown).

The most basal taxon in this analysis was *O. brachyantha* (FF), which formed a well-supported clade (bootstrap 100%, decay of 13). Four major clades emerged in this *Oryza* phylogeny (Fig 3A 1-4). These clades can be delineated based on the genome composition of the species within it, the: (1) FF clade (100% bootstrap, decay pf 130, clade, (2) the GG (100% bootstrap, decay of 6), (3) EE (100% bootstrap, decay of 10), and (4) ABCD clade (100% bootstrap and a decay of 9) (Fig 3A).

Two subclades emerged within the larger ABCD clade: the (A) BCD species subclade (*O. punctata, O. latifolia, O. alta, O.grandiglumis, O. eichingeri, O. officinalis, O. rhizomatis,*

-click here to view figure-

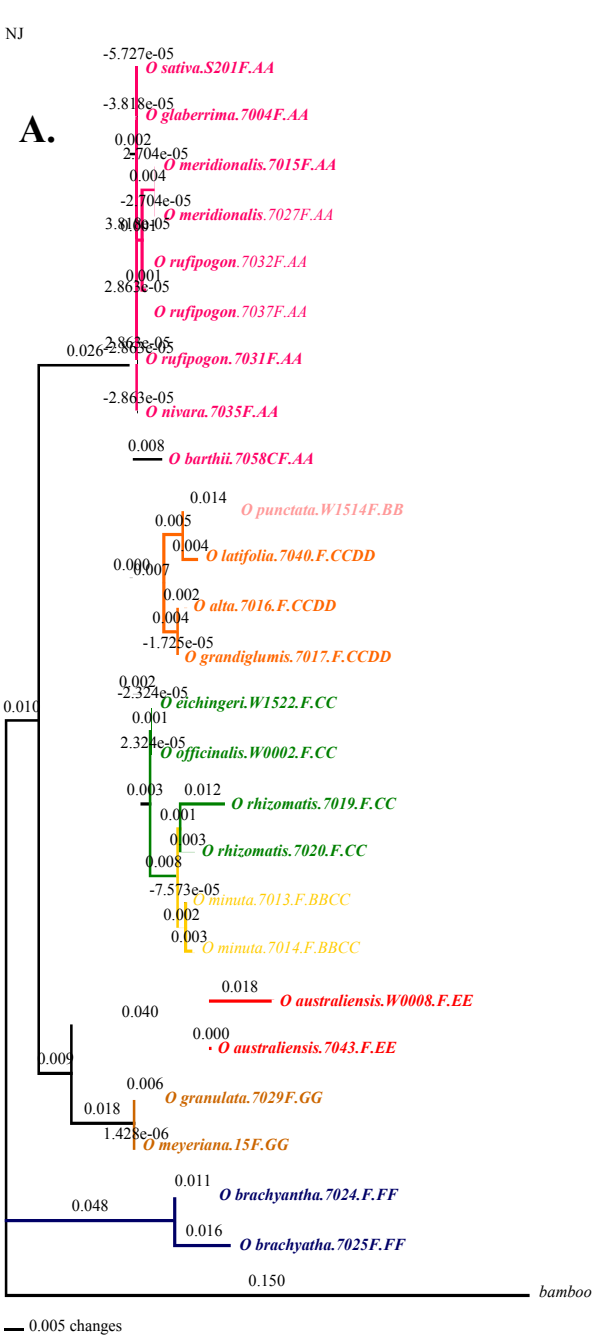
Fig. 2. Nucleotide alignment of edited *Orya* sequences. Dashes indicate gaps. Question marks indicate unknown data. Ambiguous nucleotides are codes as: K (G/T), M (A/C), R (G/A), S (C/G), and Y (C/T). The 5' end of the signal peptide is indicated by an arrow. The 5' end of the remaining part of the gene that encodes the 10 kDa prolamin polypeptide is likewise indicated by a labeled arrow.

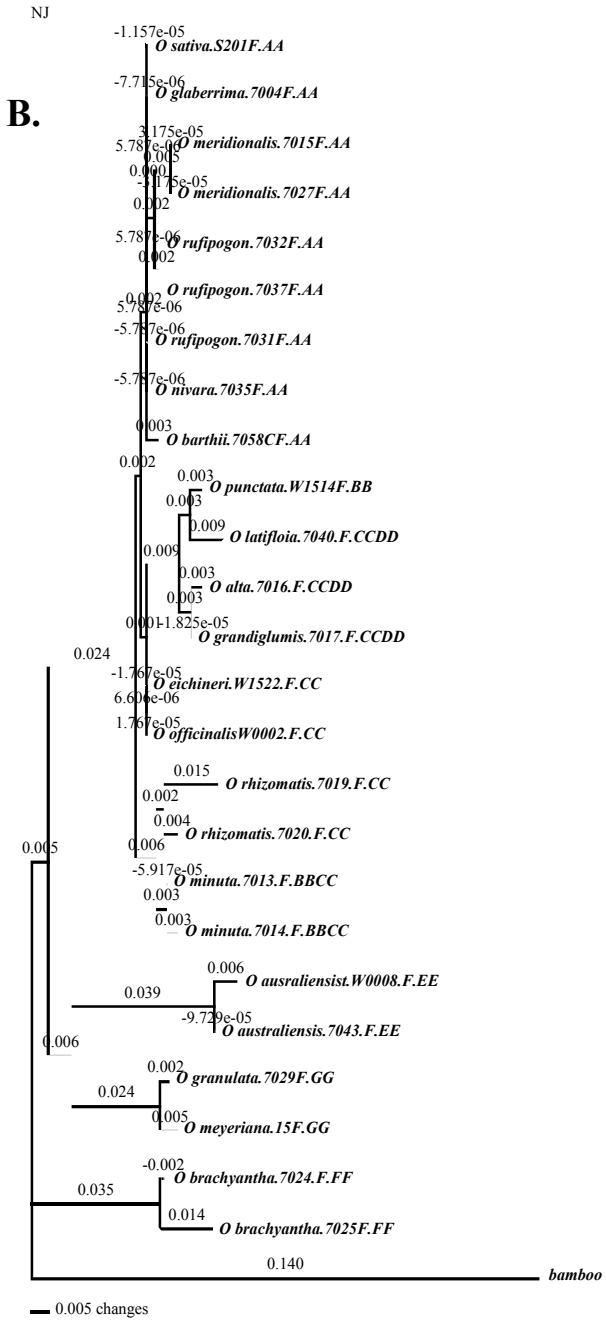



***Signal Peptide + Gene Region***
**CI= 0.8056, RI= 0.8833, g1= -1.1946**
**length= 173**
**130 variable characters (char.)**
**69 parsimonius informative char.**

***Gene Region Only***
**CI= 0.8267, RI= 0.9023, g1= -1.4321**
**length= 121**
**95 variable characters (char.)**
**51 parsimonious informative char.**

Fig. 3. Strict consensus tree of *Oryza* for the gene that encodes the 10 kDa prolamin polypeptide: (A) based on nucleotide sequence information from the signal peptide and coding regions, and (B) the coding region of the gene minus the signal peptide. Bootstrap values are indicated above each of the branches. Decay values are indicated in italics below the branches. The black vertical bar indicates a single indel event that was individually coded for in the matrix. Color coding for each of the genomes matches the biogeographic map of genome distribution (Fig.5, Chapt. 1). Circled numbers define individual clades. Circled letters indicate major subclades.




**Neighbor Joining Analysis**
**Gene + Signal Peptide Region**
**Kimura 2 Parameter Distance Estimate**

**Neighbor Joining Analysis**
**Gene Region Only**
**Kimura 2 Parameter Distance Estimate**

Fig. 4. Neighbor Joining Analysis using the Kimura 2 parameter distance estimate of the *Oryza* data set for: (A) the gene that encodes the 10 kDa prolamin polypeptide and the signal peptide region, and (B) the coding region of the gene excluding the signal peptide. Species genotypes are color-coded as presented in Chapt. 1.

and *O. minuta)* and, (B) the AA species subclade (*O. sativa, O. barthii, O. glaberrima, O. meridionalis, O. nivara,* and *O. rufipogon)* (Fig 3A). It should be noted, however, that an ATG deletion distinguished the AA genome-containing species from amongst the rest in the genus. As demonstrated in the strict consensus tree, intragenomic relationships for the *Oryza* species that contain the AA genome were not well resolved in this phylogeny (64% bootstrap and a decay of 1).

Strong support (89% bootstrap and a decay of 3) separated the South American-endemic CCDD genome species (*O. latifolia, O. alta, and O. grandiglumis*) subclade from those that contain the BBCC and CC genomes (Fig 3A). *Oryza punctata*, which contains the BB genome and is endemic to Africa, formed a subclade with *O. latifolia* (64% bootstrap), a member of the CCDD genome. *Oryza alta* (CCDD) formed a weakly supported clade (60% bootstrap) with *O. grandiglumis* (CCDD). *Oryza minuta*, which is comprised of the tetraploid genome BBCC, did not appear in the same clade as *O. punctata* (BB).

*Oryza rhizomatis* (CC) and *O. minuta* (BBCC) formed a subclade with strong bootstrap support (94%) and a decay of 3, within the BC subclade. *Oryza eichingeri* (CC) and *O. officinalis* (CC), which both share a genome with *O. rhizomatis*, did not share the subclade with *O. minuta* (BBCC) and *O. rhizomatis* (CC), indicating that the sequence for *O. rhizomatis* was divergent from the other CC genome taxa (Fig 3A).

**Neighbor Joining Analysis of *Oryza* data set**

Trees, both including and exluding the signal peptide, produced by the neighbor joining analyses based on Kimura 2 parameter distances (Fig 4A &B) were topologically similar to each

other and to the parsimony-based phylogeny described earlier (Fig 3A& B). Exceptions to this was the placement of *O. barthii* (AA), which did not form a subclade with the remaining species that contained the AA genome in the NJ analysis of the data set that contained the signal peptide and gene region (Fig. 4A).

**Deduced amino acid analysis of the gene that encodes the 10 kDa prolamin polypeptide.**

In the phylogenetic analysis of the gene that encodes the 10 kDa prolamin polypeptide (excluding the signal peptide region), the first codon position provided 13 of the 50 variable parsimony informative characters, the second position provided 13 and the third provided 24; which represent 26%, 26%, and 48% respectively of the parsimony informative variable characters within the matrix. These data indicate that 26 of the 50, or 52%, of the variable, parsimony-informative characters within the gene that encodes the 10 kDa prolamin are in candidate positions for amino acid substitutions.

The open reading frame of the gene that encodes the 10 kDa prolamin polypeptide was therefore translated into its deduced amino acids (Table 4). Residues with the greatest variability among species include: glutamine (Q), leucine (L), methionine (M), threonine (T), asparagine (N), cysteine ( C ), and serine (S) (Fig 4). The *Oryza* species that exhibit the greatest variability in amino acid composition relative to *Oryza sativa* (cultivated rice) are: *O. punctata* (7 substitutions [s]), *O. minuta* (7s), *O. rhizomatis* (7s), *O. alta* (6s), *O. latifolia* (8s), and *O. grandiglumis* (5s) (Table 1). Thus, the genomes/clades that exhibited the greatest variability in amino acid composition relative to the cultivated *sativa* AA genome are the South American-

endemic CCDD genome/ *O. punctata* (BB genome) subclade, and the BBCC/CC subclade of *O. minuta* and *O. rhizomatis* (Fig 5).

Table 4. Deduced amino acid distribution (in single letter code) of *Oryza* species for the gene that encodes the 10 kDa prolamin polypeptide.

| | **Amino Acids** | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Oryza* species** | **A** | **R** | **Q** | **E** | **G** | **H** | **L** | **K** | **M** | **F** | **T** | **W** | **N** | **D** | **C** | **I** | **P** | **S** | **Y** | **V** |
| *O.sativa* | 6 | 1 | 19 | 0 | 8 | 1 | 7 | 1 | 22 | 3 | 5 | 0 | 2 | 1 | 11 | 1 | 7 | 7 | 3 | 2 |
| *O. merid* | 6 | 2 | 19 | 0 | 8 | 1 | 7 | 1 | 22 | 3 | 4 | 0 | 2 | 1 | 11 | 1 | 7 | 7 | 3 | 2 |
| *O. rufip.* | 6 | 1 | 19 | 0 | 8 | 1 | 7 | 1 | 22 | 3 | 5 | 0 | 2 | 1 | 11 | 1 | 7 | 7 | 3 | 2 |
| *O. glabb.* | 6 | 1 | 19 | 0 | 8 | 1 | 7 | 1 | 22 | 3 | 5 | 0 | 2 | 1 | 11 | 1 | 7 | 7 | 3 | 2 |
| *O. nivara* | 6 | 1 | 19 | 0 | 8 | 1 | 7 | 1 | 22 | 3 | 5 | 0 | 2 | 1 | 11 | 1 | 7 | 7 | 3 | 2 |
| *O. punct.* | 5 | 2 | 18 | 0 | 8 | 1 | 8 | 1 | 23 | 3 | 6 | 0 | 1 | 1 | 11 | 1 | 7 | 7 | 3 | 2 |
| *O. min.* | 6 | 1 | 18 | 0 | 8 | 1 | 8 | 1 | 21 | 3 | 6 | 0 | 3 | 1 | 11 | 1 | 7 | 6 | 3 | 3 |
| *O. eich.* | 6 | 1 | 18 | 0 | 8 | 1 | 8 | 1 | 23 | 3 | 5 | 0 | 2 | 1 | 11 | 1 | 7 | 7 | 3 | 2 |
| *O. offic.* | 6 | 1 | 18 | 0 | 8 | 1 | 8 | 1 | 22 | 3 | 5 | 0 | 2 | 1 | 11 | 1 | 7 | 7 | 3 | 2 |
| *O. rhizom* | 6 | 1 | 18 | 0 | 8 | 1 | 8 | 1 | 21 | 3 | 5 | 0 | 1 | 1 | 11 | 2 | 7 | 8 | 3 | 3 |
| *O. alta* | 5 | 1 | 18 | 0 | 8 | 1 | 10 | 1 | 23 | 2 | 6 | 0 | 2 | 1 | 11 | 1 | 6 | 7 | 3 | 2 |
| *O. lat.* | 5 | 1 | 17 | 0 | 8 | 2 | 7 | 3 | 23 | 3 | 6 | 1 | 1 | 1 | 9 | 2 | 7 | 7 | 3 | 2 |
| *O. grandi* | 5 | 1 | 18 | 0 | 8 | 1 | 9 | 1 | 23 | 3 | 6 | 0 | 2 | 1 | 11 | 1 | 6 | 7 | 3 | 2 |
| *O. brachy* | 6 | 1 | 18 | 0 | 8 | 1 | 8 | 1 | 23 | 3 | 5 | 0 | 2 | 1 | 11 | 1 | 7 | 7 | 3 | 2 |

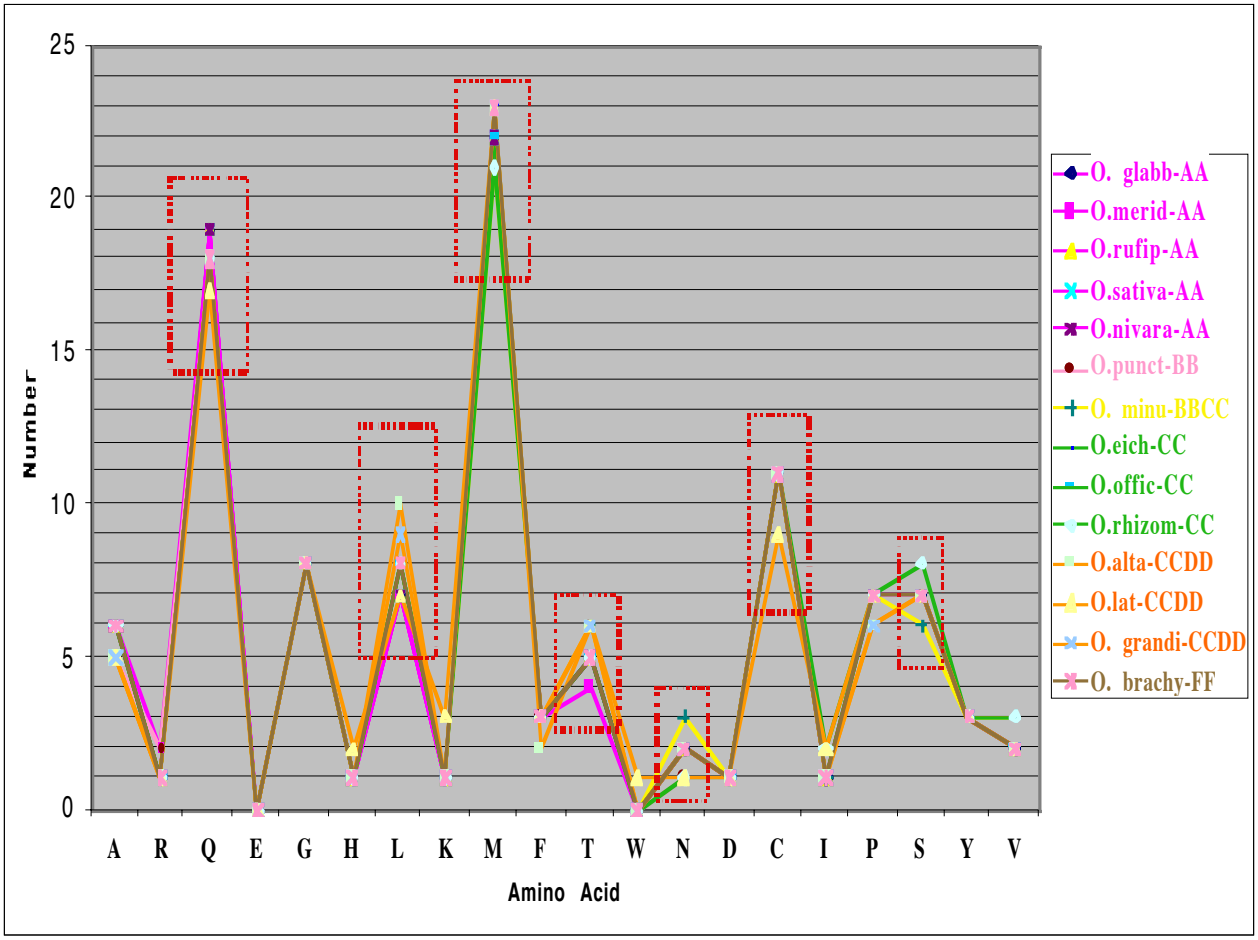



Fig. 5. Amino acid distribution among *Oryza* species. Boxed regions indicate residues with greater variability. Abbreviated species names and genomes are indicated in the key.

## Discussion

A comparative analysis of previous phenetic (Morishima & Oka, 1969; McIntyre *et al.*, 1992) and cladistic (McIntyre *et al.*, 1992) approaches towards tree construction of the *Oryza* genus demonstrates a large degree of consistency between the two methods. In a phenetic

analysis based on 42 morphological characters, Morishima and Oka (1960) concluded that *Oryza* species with the AA genome formed a cluster, indicating that these species were similar. They also showed that *Oryza granulata*, *O. minuta*, *O. officinalis*, *O. alta*, *O. latifolia*, and *O. eichingeri* clustered together, indicating that *O. granulata* (GG) is similar phenetically to species with the BBCC, CC, and CCDD genomes (Morishima & Oka, 1959). *Oryza brachyantha* and *O. ridleyi* also formed a cluster, indicating that they are more similar to each other but less similar to the other above mentioned species. These non-molecular data are consistent with those presented in this phylogeny (Fig. 3A).

McIntyre *et al.* (1992) examined the relationships between nine *Oryza* species using 5S DNA sequences. A cladistic analysis of the consensus sequences yielded 18 trees that were equally parsimonious. Using *O. australiensis* as an outgroup taxon, three clades emerged in the consensus tree: (1) *O. brachyantha, O. officinalis, O. punctata* (100% bootstrap); (2) *O. nivara, O. rufipogon*, *O. sativa* (100% bootstrap); and (3) *O. latifolia, and O. minuta* (100% bootstrap). The only significant difference between this phylogeny and the prolamin gene tree (Fig. 3A) is that *O. brachyantha* (FF-Africa) does not appear within the lineage containing a BB or CC genome, rather the two accessions of this species form the most basal clade in the present study (Fig. 3A). This implies that *O. brachyantha* is one of the most basal taxa in *Oryza*, which counters an Asian origin of cultivated rice (Chang, 1976).

The designation of *Oryza* into three sections: Oryza, Ridleyaneae, and Granulata (Vaughan, 1994; Table 2) is partially consistent with the results presented here (Fig 3A). Section Oryza is further divided into two complexes: (1) the Sativa Complex (all AA genome species), and (2) the Officinalis Complex (B,C,D, and E genomes) (Vaughan, 1994). The Sativa (64% bootstrap and a decay of 1) is weakly supported in this analysis, while the inclusion of the E

genotype in the Officinalis complex is not supported in this analysis (Fig. 3A). Vaughan's (1994) section Ridleyaneae Tateoka consists of the ridleyi complex and *O. brachyantha* and *O. schlechteri*. While *O. brachyantha* falls within this section (Vaughan, 1994), no "complex" designation has been given. In this analysis (Fig. 3A), *O. brachyantha* (FF) forms its own strongly supported clade (Fig. 3A). The third section, *sensu* Vaughan (1994), Granulata Roschev contains the O. meyeriana complex (Vaughan, 1994). Two species, *O. meyeriana* (GG) and *O. granulata* (GG) belong to the O. meyeriana complex. These two species formed a well-supported clade in the present analysis (Fig. 3A).

B.R. Lu's revision (Table 2) of Vaughan's (1994) classification system (Table 1) (personal correspondence, 1999) deviates from the latter in that it does not utilize the rank "complex", as this rank is not in use by the International Code of Botanic Nomenclature. Rather, Lu utilizes three sections to classify Oryza: (1) Sect. *Padia* (3 series, 6 species); (2) Sect. *Oryza* (3 series, 17 species); and (3) Sect. *Brachyantha* (1 series, 1 species). Section Padia contains: *O. granulata, O. meyeriana*, *O. longiglumis*, *O. ridleyi*, and *O. schlechteri*. Section *Brachyantha* contains *O. brachyantha*. Section *Oryza* contains the remaining species (Table 2, Fig 3). The data presented here support Lu's revised classification in that within Section Oryza, Series Latifoliae contains: *O. alta, O. eichingeri, O. grandiglumis, O. latifolia, O. minuta, O. officinalis, O. punctata, and O. rhizomatis* (this series proposed by Lu forms a well-supported subclade within the ABCD clade, and unlike Vaughan's *O. officinalis* complex, does not contain *O. australiensis* [EE]); Series Australiensis contains *O. australiensis*; and Series Sativae contains: *O. barthii, O. glaberrima, O. longisteminata, O. meridionalis, O. nivara, O. rufipogon, and O. sativa*. (Fig 3A). In congruence with Lu's system, *O. granulata* (GG) forms a clade with *O. meyeriana* (GG) (Fig 3A).

It is noteworthy to emphasize that *O. punctata* (BB) forms a well-supported (80% bootstrap) subclade with the three South American-endemic (CCDD) species *(O. latifolia, O. grandiglumis,* and *O. alta*) (Fig. 3A). No species containing the DD genome has been identified in the wild, although the DD chromosomes have been labeled in a CCDD tetraploid using florescent *in situ* hybridization (FISH) (Fukui *et al*., 1997). These data indicate that *O. punctata* could be related to the species that originally donated the DD genome. Based on dot blot analyses of genomic *Oryza* DNA using rDNA insert probes, McIntyre and Winberg (1998) concluded that *O. officinalis* (CC), *O. minuta* (BBCC), and *O. punctata* (BBCC) probes did not hybridize to atleast one of the cloned rDNA size fragments of *O. latifolia* (CCDD), indicating that the *O. latifolia* size fragment may have originated from the DD genome. A more detailed analysis of the allotetraploid *Oryza* species is presented in Chapter 3.

Socioeconomic implications of the amino acid variability within the gene that encodes the 10 kDa prolamin polypeptide arise in that several *Oryza* species have greater amounts of essential amino acids such as: tryptophan, methionine, valine, threonine, phenylalanine, leucine, isoleucine, and lysine (amino acids that are not synthesized by humans) than others (Table 4, Fig 5). Specifically, *Oryza alta*, *O. punctata* and *O. grandiglumis* contain more leucine, methionine, and threonine than cultivated rice (*O. sativa*). Given this variation in protein quality and quantity (Table 4), the wild species of rice are good candidates for genetic material transfer to enhance the quality of protein in cultivated rice.

**Acknowledgements**

This work was funded through grants from the Virginia Academy of Science, *Sigma Xi*, the Graduate Research and Development Project, the Department of Biology, Virginia Tech, and a grant from the Jeffrees Foundation. DNA was generously provided by Dr. Yoshio Sano, Hokaido University, Soporro Japan. Dried plant material was provided by Dr. Chrostoph Neihous, University of Bonn, Germany. Special thanks are also extended to D. Wiley for her skills in growing and maintaining the *Oryza* greenhouse collections and to T. Weiboldt for his assistance in herbarium voucher preparation.

**References**

Aggarwal, K.; Brar, D.S.; Khush, G.S. 1997. Two new genomes in the *Oryza* complex identified on the basis of molecular divergence analysis using genomic DNA hybridization. *Mol. Gen. Genet.* (254): 1-12.

Barbier, P.; Ishihama, A. 1990. Variation in the nucleotide sequence of a prolamin gene family in wild rice. *Plant Mol. Biol.* (15): 191-195.

Bhatia, C.R.; Rabson, R. 1987. Relationship of grain yield and nutritional quality. *Nutritional quality of cereal grains: genetic and agronomic improvements.* ASA-CSSA-SSSA, Madison: WI. (28): 11-131.

Bietz, A. 1982. Cereal prolamin evolution and homology revealed by sequence analysis. *Biochem. Genet.* (20): 1039-1053.

Bremmer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution.* (42): 795-803.

Chang, T.-T. 1976. The origin, evolution, cultivation, dissemination, and diversification of

Asian and African rices. *Euphytica*. (25): 425-441.

Chen, J-M.; Lin, J-J.; Jiang, S-T.; Chen, C-S. 1995. Characterization of rice prolamin genes. *Czech-Taiwan Symposium on Biotechnology*, Prague: June 5-8.

Coffman, W.R.; Juliano, B. O. 1987. Rice. *Nutritional quality of cereal grains: genetic and agronomic improvement*. ASA-CSSA-SSSA: Madison, WI. (28): 101-131.

Donoghue, M.J.; Olmstead, R.G.; Smith, J.F.; Palmer, J.D. 1992. Phylogenetic relationships of Dipsacales based on rbcL sequences. *Ann. Missouri Bot. Gard*. (79): 333-345.

Eriksson, T. 1998. AutoDecay ver. 4.0 (program distributed by author). Department of Botany, Stockholm University, Stockholm.

Hibino, T.; Kidzu, K.; Masumura, T.; Ohtsuki, K.; Tanaka, K.; Kawabata, M.; Fujii, S. 1989. Amino acid composition of rice prolamin polypeptides. *Agric. Biol. Chem*. (53): 513-518.

Higgins, T.J.V. 1984. Synthesis and regulation of major proteins in seeds. *Ann. Rev. Plant. Physiol*. (35): 191-221.

Hillis, D.M.; Huelsenbeck, J.P. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. *J. Heredity* (83): 189-195.

Hilu, K.W.; Esen, A. 1988. Prolamin size and diversity in the Poaceae. *Biochem. Syst. and Ecol*. (16): 457-465.

Hilu, K.W. 1994. Evidence from RAPD markers in the evolution of Echinochloa millets (Poaceae). *Plant Syst. and Evol*. (189): 247-257.

Hilu, K.W.; Sharova, L.V. 1998. Characterization of 10 kDa prolamin genes in *Phyllostachys aurea* (Bambusoideae, Poaceae). Amer. J. Bot. (85): 1033-1037.

Kim, W.T.; Okita, T.W. 1988. Structure, expression, and heterogeneity of the rice seed prolamines. *Plant Physiol*. (88): 649-655.

Krishnan, H.; White, J.A. 1995. Morphometric analysis of rice seed protein bodies. *Plant Physiol*. (109): 1494-1495.

Masumura, T.; Shibata, D.; Hibino, T.; Kato, T.; Kawabe, K.; Takeba, G.; Tabaka, K.; Fuji, S. 1989. cDNA cloning of an mRNA encoding a sulfur-rich 10 kDa prolamin polypeptide in rice seeds. *Plant Mol. Biol*. (12): 123-130.

Masumura, T.; Mitsukawa, N.; Tanaka, K.; Fuji, S. 1991. Rice storage proteins: genetic analysis of accumulation process. *Biotech. in Agric. and Forr*. Springer-Verlag: Berlin. (14): 495-507.

McIntyre, C.L.; Winberg, B.; Houchins, K.; Appels, R.; Baum, B.R. 1992. Relationships between *Oryza* species (Poaceae) based on 5S DNA sequences. *Pl. Syst. Evol*. (183): 249-264.

McIntyre, C.L.; Winberg, B.C. 1998. A rapid means of identifying wild rice species DNA using dot blots and genome specific rDNA probes. *Genome* (41): 391-395.

Morishima, H.; Oka, H.I. 1959. The pattern of interspecific variation in the genus *Oryza:* its quantitative representation by statistical methods. *Evolution*. (14): 13-165.

Saghai-Maroof, M.A.; Soliman, K.M.; Jorgensen, R.A.; Allard, R.W. 1984. Ribosomal DNA spacer-length polymorphism in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci.* USA (81): 8014-8018.

Sanderson, M.J.; Donoghue, M.J. 1989. Patterns of variation in levels of homoplasy. *Evolution*. (8): 1781-1795.

Second, G. 1985. Evolutionary relationships in the Sativa group of *Oryza* based on isozyme data. *Génét. Sél. Evol.* (17): 89-114.

Shewry, P.R.; Napier, J.A.; Tatham, A.S. 1995. Seed storage proteins: Structures and biosynthesis. *Plant Cell*. (7): 945-956.

Tanaka, K.; Sugimoto, T.; Ogawa, M. Kasai, Z. 1980. Isolation and characterization of two types of protein bodies in the rice endosperm. *Agric. Biol. Chem.* (44): 1633-1639.

Vaughan, D. 1994. *The wild relatives of rice: A genetic resources handbook.* International Rice Research Institute, Los Baños, Philippines: 1-101.

Zhou, X.; Fan, Y-L. 1993. The endosperm-specific expression of a rice prolamin chimaeric gene in transgenic tobacco plants. *Transgenic Res.* (2): 141-146.

**-Chapter 3.-**

# GENOMIC ORIGIN OF *ORYZA* POLYLPLOID SPECIES BASED ON THE GENES ENCODING THE 10 KDA PROLAMIN: A CASE OF CONCERTED EVOLUTION

Irene M. Mullins & Khidir W. Hilu
Virginia Polytechnic Institute and State University

## Introduction

While cereals comprise the most important food crop in the world, and rice (*Oryza sativa* 2n=24) is one of the most common cereal grains consumed (providing 80% of our food at the global level [USDA, 1991]), neither the intraspecific nor intergenomic genetic relationships in *Oryza* are well understood.

The 22 species of *Oryza* fall into nine genomes (AA, BB, CC, DD, EE, FF, GG, HH, JJ) based on chromosome pairing (reviewed in Vaughan, 1994) and total genomic DNA hybridization (Aggarwal *et al.,* 1997) experiments. In addition, this genus contains both diploid and allotetraploid taxa, which contain varying basic chromosome numbers. This paper focuses on the genomic origin of the polyploid species and their genetic progenitors.

Previous work has determined the genome composition of the *Oryza* allotetraploid species, and candidate progenitor taxa have been proposed; however, no cladistic examination of the majority of the genus *Oryza* founded on a gene-based molecular analysis has been conducted. Specifically, chromosome pairing analyses of the allotetraploid wild species *O. minuta* (BBCC, 2n=48) designated it as an allopolyploid hybrid between *O. sativa* (AA) and *O. officinalis* (CC) (reviewed in Nayar, 1975). A numerical taxanomic phenetic analysis using nonmolecular characters in *Oryza* provided additional evidence in support of a close relationship between *O.*

*minuta* and *O. officinalis* (Morishima & Oka, 1960). The genomic donors of another allotetraploid *Oryza* genome, CCDD (which includes *O. latifolia, O. grandiglumis,* and *O. alta*), have not been identified, although *O. officinalis* (CC-Asia) and *O. punctata* (BB-Africa) have been proposed as candidates (Gopalakrishnan and Sampath, 1966a,b).

The gene that encodes the 10 kDa prolamin polypeptide, a seed storage protein, was used in this study to examine species relationships and to address questions regarding genome origin and evolution in *Oryza*. Previous studies have identified and characterized this gene. Specifically, Masumura *et al.* (1989) isolated cDNA that coded for a sulfur-rich 10 kDa rice prolamin. Barbier and Ishihama (1990) examined the variation present in the DNA sequence of a prolamin gene family that codes for the 10 kDa prolamin polypeptide within the species *Oryza rufipogon* by PCR amplification and direct sequencing, noting that: (1) the gene that encodes the 10 kDa prolamin has no intron, and (2) the gene copies that were amplified were homogeneous in size.

Our principle hypothesis is that genetic diversity is present in the gene that encodes the 10 kDa prolamin polypeptide (discussed in greater detail in chapter 2), which can provide basic information regarding the phylogenetic relationships among rice species and permit a more theoretical examination of evolutionary processes (i.e. concerted evolution) within several of the allotetraploid *Oryza* species.

The process by which the multiple gene copies that encode the 10 kDa prolamin in *Oryza* have been homogenized to one form is speculated to be concerted evolution. Concerted evolution (Zimmer *et al.*, 1980) refers to the observation that individual genes in multi-gene families tend to homogenize so that apparently all gene copies are identical (Zimmer *et al.*, 1980; Arnheim *et al.*, 1980; Riesberg *et al.*, 1990; Hillis *et al.*, 1991; Wendel *et al.*, 1995; Liao, 1999).

The molecular-drive mechanisms by which the process of concerted evolution is thought to operate are biased gene conversion and unequal crossing over, among others (reviewed in Elder & Turner, 1995).

## Materials and Methods

Plant growth, genomic DNA extraction, PCR amplification, sequencing and analysis were performed as indicated in Chapter 2.

### ***Plant Growth***

Genomic DNA was isolated from 19 species of *Oryza*, and the gene encoding the 10 kDa prolamin polypeptide was amplified. All rice grains for this study (except for *O. sativa*) were supplied by the International Rice Research Institute (IRRI) in the Philippines, or by the United States Department of Agriculture (USDA) (Table 1). *Oryza sativa* grain was supplied by the Pioneer Valley Seed Company. Grains were germinated in moist, filter paper-lined Petri plates until seedling growth appeared vigorous, at which time seedlings were transferred to individual pots containing a clay/ organic soil mixture. Pots were placed into shallow, three inch-deep pans of water in a greenhouse, and were treated every two weeks with acid fertilizer (MirAcid, 30:10:10, 2 teaspoons [tsp.]/ gallon) and iron supplements (Earthcare Secuestren Iron, 1/4 teaspoon/ gallon). Voucher specimens are located at IRRI (Philippines) and the Massey Herbarium (VT). All IRRI seed lots sent to the investigators of this project have been positively identified by Dr. B.R. Lu, germplasm specialist at IRRI, with vouchers at IRRI.

***Genomic DNA Extraction***

DNA was isolated from leaf material via Hilu's modification (1994) of the CTAB extraction method (Saghai-Maroof *et al*., 1984). Briefly, 700 microliters (µl) CTAB (100mM Tris HCl, 2% CTAB, 20 mM EDTA, 1.4 M NaCl) were heated at 65˚C in a 1.5 ml Eppendorf tube with 1.5 µls Beta-Mercaptoethanol (Fisher Scientific). Approximately one gram of fresh leaf material was ground in a chilled mortar and pestle with liquid nitrogen, placed in the CTAB+ BME solution, and incubated at 65˚C for thirty minutes (min.). Seven hundred µls of

Table 1. Description of *Oryza* accessions examined in this study. Seed accessions obtained from the International Rice Research Institute, Philippines are abbreviated IRRI, those obtained from the United States Department of Agriculture (USDA) are abbreviated PI (Plant Introduction Number). DNA obtained from the laboratory of Dr. Y. Sano, Hokkaido University, Japan, is abbreviated W. Lyophilized plant material obtained from the University of Bonn, Germany, is abbreviated B.

| ***Taxon*** | **Genome** | **KH #** | **Accession #** | **Geographic Origin** | **Primer Pair used for sequencing** |
|---|---|---|---|---|---|
| *Oryza sativa* | AA | KH7000 | ---------- | United States | PR10.1F2/PR10.3RV |
| *O. nivara* | AA | KH7035 | IRRI 101524 | India | PR10.1F2/PR10.3RV |
| *O. glaberrima* | AA | KH7004 | PI 450198 | Nigeria | PR10.1F2/PR10.3RV |
| *O. meridionalis* | AA | KH7027 | IRRI 105279 | Australia | PR10.1F2/PR10.3RV |
| *O. officinalis* | CC | ---------- | W0002 | Thailand | PR10.1F2/PR10.3RV |
| *O. rhizomatis* | CC | KH7020 | IRRI 103421 | Sri Lanka | PR10.1F2/PR10.3RV |
| *O. rhizomatis* | CC | KH7019 | IRRI 103410 | Sri Lanka | PR10.1E/PR10.2Rb |
| *O. eichingeri* | CC | ---------- | W1522 | Uganda | PR10.1F2/PR10.3RV |
| *O. minuta* | BBCC | KH7014 | IRRI 101128 | Philippines | PR10.1F2/PR10.3RV |
| *O. minuta* | BBCC | KH7013 | IRRI 101082 | Philippines | PR10.1F2/PR10.3RV |
| *O. punctata* | BB | ---------- | W1514 | Kenya | PR10.1F2/P10.3RV |
| *O. latifolia* | CCDD | KH7040 | IRRI 100165 | Guatemala | PR10.1E/PR10.2Rb |
| *O. alta* | CCDD | KH7016 | IRRI 101395 | no source | PR10.1F2/PR10.3RV |
| *O. grandiglumis* | CCDD | KH7017 | IRRI 101405 | Brazil | PR10.1F2/PR10.3RV |
| *O.australiensis* | EE | ---------- | W0008 | N. Australia | PR10.1E/PR10.3RV |
| *O. brachyantha* | FF | KH7024 | IRRI 105171 | Cameroon | PR10.1E/PR10.2Rb |
| *O. meyeriana* | GG | KH7030 | IRRI 106474 | Philippines | PR10.1F2/PR10.3RV |
| *O. granulata* | GG | KH7029 | IRRI 106468 | Laos | PR10.1E/PR10.2RB |

chloroform were added to each of the tubes, and then the tubes were spun at 15K, room temperature (rt) for 15 min. For some samples a second chloroform extraction was performed as necessary. Four hundred µls of isoproponol were added; samples were then precipitated at -20˚C overnight, and centrifuged at 4˚C for 23 min., 15K. Pellets were rinsed with 350 µls of 70% ethanol, and lyophilized in a speedvac. DNA was resuspended in 40 µls of 1X TE (10mM Tris Cl [pH 8.0], 1 mM EDTA [pH 8.0]). DNA was quantified using a Lambda (λ) *Hind*III digest via electrophoresis in a 1.0% agarose gel (Sigma, St. Louis, MO) at 65V for 2.5 hours and visualized with ethidium bromide staining by immersing gel in a (0.5.µg ethidium bromide/ ml water) bath for 25 min., rt.

***PCR amplification and alcohol precipitation***

The gene that encodes the 10 kDa prolamin polypeptide was amplified using forward and reverse primers: PR10.1E (5' ACG TGA ATT CAT GGC AGC ATA CAC CAG CAA G 3') and PR10.2RB (ACG TGG ATC CAA CCA CAG GAA GAG AGT TGG 3') (primers were synthesized by Operon, Inc) (Hilu & Sharova, unpublished data). In other cases when these two primers did not amplify, another pair of primers were used: PR10.1F2 (5' ACG TGA ATT CCA CCA TCT GGA ATC TGG 3') and PR10.3RV (ACG TTC TAG AG TGT TTG CAC ACG ATA GTA 3'). Primers were designed based on the Masumura *et al*. (1989) nucleotide sequence of the gene that encodes the 10 kDa prolamin polypeptide.

The forward primer PR10.1E and reverse primer PR10.2RB are located within the conserved coding region, at the 5' end of the gene, and were utilized to amplify species that could not otherwise be amplified using PR10.1F2 and PR10.3RV. Because the primers are located within the gene itself, complete sequence information was not obtained. Genomic DNA was

amplified using Taq polymerase (Promega) and the polymerase chain reaction (PCR). The amplification conditions used were 40 cycles of: 0.75 min., 94˚C (denaturation); 1 min., X˚C (annealing); and 1 min., 72˚C (extension). Annealing temperatures (ann. temp.) for the following samples were: (1) *Oryza rhizomatis:* 40˚C, (2) *O. latifolia*: 48˚C (3) *O. australiensis*: 48˚C, (4*) O. brachyantha*: 40˚C, (5) *O. granulata*: 40˚C, and (6) *O. meyeriana: 48*˚C.

The forward primer PR10.1F2 and the reverse primer PR10.3RV are located within the signal peptide region of the 10 kDa gene, and therefore provide complete sequence information for the coding region and signal peptide of the gene. The amplification conditions used were 40 cycles of: 0.5 min., 94˚C (denaturation); 1.5 min., 48˚C (annealing); and 1 min., 72˚C (extension) were used for all of the following samples: (1) *Oryza nivara*, (2) *O. glaberrima*, (3) *O. meridionalis* (KH 7027), (4) *O. officinalis*, (5) *O. rhizomatis* (KH 7020), (6) *O. eichingeri*, (7) *O. minuta* (KH7013, 7014), (8) *O. punctata*, (9) *O. alta*, (10) *O. grandiglumis*, and (11) *O. brachyantha* (KH7024). An annealing temperature of 50˚C was used for *Oryza sativa*.

For each species, multiple PCR reaction replicates were pooled and their volume was reduced via a sodium acetate/ ethanol precipitation after the PCR products had been quantified on a 1.% agarose gel (Sigma). The DNA samples were electrophoresed on 1.5% agarose gels for several hours and visualized with ethidium bromide staining. The appropriate bands were cut out, and the DNA was cleaned using the Quiagen Gel Extraction Kit (Quiagen Inc., Valencia, CA).

***Nucleotide Sequencing of PCR products***

Approximately 30-60 ng of column-cleaned PCR product (Quiagen) were used to prepare half-volume sequencing reactions using the ABI Prism Dye Terminator Cycle Sequencing Ready Reaction Kit with AmpliTaq DNA polymerase, FS or the ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction Kit (Perkin Elmer, Foster City. CA) with an annealing temperature of 50˚C. Samples were prepared according to the manufacturer's protocol. Sequencing reactions were alcohol precipitated with either an ethanol/ sodium acetate or 95% isoproponol protocol as indicated by the manufacturer.

DNA was sequenced at the University of Maine (Orono, ME), University of Florida (Gainesville, FL), Virginia/ Maryland Regional Veterinary School (Blacksburg, VA) or the Virginia Tech Biology Department (Blacksburg, VA) sequencing facilities.

***Sequence Analysis***

DNA sequences were manually edited and aligned using the Sequence Navigator program, given the unambiguous and homogeneous nature of the nucleotide sequence data. Both forward and reverse sequences were examined for polymorphic signal and the forward sequence was used in the alignment. Gaps were introduced by inserting dashes. Ambiguous nucleotides were coded as: K (G/T), M(A/C), R(G/A), S(C/G), and Y(C/T) using standard IUPAC symbols.

Edited sequences were analyzed using Fitch parsimony in PAUP 4.0b3a with data sets that both included and excluded the signal peptide region (Fig. 1) (Swofford, 1999). Heuristic searches were performed with 1000 random stepwise-addition replicates, using the “MulTrees” option and the TBR (tree-bisection-reconnection) branchswapping algorithm. Gaps were treated as missing data except for one consistent indel event that was coded for in the species that

contain the AA genome. A strict consensus tree was constructed from the equally parsimonious trees. The single gap coded for in this analysis that was parsimony-informative was mapped onto the strict consensus tree. Decay indices (Bremmer, 1988; Donaghue *et al.*, 1992), and bootstrap values for 100 replicates, were calculated in order to measure support for individual clades. AutoDecay (Eriksson, 1998) was used to perform the decay analysis.

$G_1$ values were created for 1000 random trees using the "evaluate random trees" option in PAUP 4.0b3a (Swofford, 1999) and compared to the values given for 25 taxa with 250 characters (variable positions) with a P value of 0.01 (Hillis and Huelsenbeck, 1992) to test for the presence of non-random structure in the data set.

## Results

### Nucleic acid analysis of the gene that encodes the 10 kDa prolamin polypeptide.

A more detailed description of the phylogenetic results for all species used in this analysis has been presented in Chapter 2. The aligned sequences are presented in Fig. 1. The gene that encodes the 10 kDa prolamin polypeptide consists of approximately 402 bases, which includes a signal peptide region consisting of 84 bases. The open reading frame for the coding region of the species examined in this study is 327 bases. The phylogenies resulting from the parsimony analysis with and without the signal peptide region are shown in Fig. 2A & B respectively.

Within this data set 531, characters were aligned (Fig. 1). In the analysis that included the signal peptide, 130 characters were variable, 69 of which were parsimony informative (53%) (Fig. 2A). The -1.1946 $g_1$ value of the *Oryza* data set provides significant ($P$<0.01) evidence of

non-random structure. A cladistic analysis yielded 24 most parsimonious trees that were 173 steps in length. The Consistency Index (CI) (excluding uninformative characters) and Retention Index (RI) were 0.8056 and 0.8833, respectively. The predicted CI based on the polynomial regression analysis of Sanderson and Donoghue (1989) is 0.3279, indicating that this data set contains low levels of homoplasy.

In the analysis that excluded the signal peptide region, 95 variable characters were present, of which 51 were parsimony-informative (53.6%) (Fig. 2B). The $g_1$ value for this data set was -1.4321, indicating that the structure of the data set was non-random ($P<0.01$). Seventy-two most parsimonious trees, 121 steps in length, were generated based on a cladistic analysis of the data set. The Consistency (excluding uninformative characters) and Retention indices were 0.8267 and 0.9023, respectively. The predicted CI was 0.3279, which also indicated that there were low levels of homoplasy within the data set.

**Genome relationships of select *Oryza* allotetraploid taxa.**

*CCDD/ BB Clade*

The South American-endemic species containing the allotetraploid CCDD genome, *O. latifolia, O. grandiglumis,* and *O. alta*, formed a subclade with strong support in both consensus trees (signal peptide included and excluded) with 89 and 94% bootstraps and decays of 3 and 2 respectively (Fig. 2A & B). Within this subclade, two subclades were formed, one consisting of *Oryza latifolia* (CCDD) and the African-endemic *O. punctata* (BB genome) (80% bootstrap), and the second consisting of *O. alta* (CCDD) and *O. grandiglumis* (CCDD) (60% bootstrap) (Fig. 2A).

*BBCC/CC Clade*

The allotetraploid *O. minuta* (BBCC) formed a well-supported clade (bootstrap of 94% and 77% and decay of 3 and 2, respectively) (Fig. 2A & B) with *O. rhizomatis,* a diploid species that contains the CC genome and that is endemic to Sri Lanka. None of the other CC or BB species appeared in this subclade. Specifically, in two variable sites where *O. minuta* was variable relative to any of the other taxa, it favored the CC genome of *O. rhizomatis* over the BB genome of *O. punctata* 100% (Fig.1). The two remaining CC genome species *O. officinalis* (CC genome), and *O. eichingeri* (CC genome) do not form a clade with *O. minuta* and *O. rhizomatis*; however, they do fall within the larger B/C subclade of the ABCD clade in both the gene tree that included and excluded the signal peptide region (Fig. 2A & B). *Oryza minuta* contains sequence mutations that are more similar to *O. rhizomatis* than *O. officinalis* or *O.eichinger*i (Fig.1).

**Deduced amino acid analysis for the translated open reading frame of the coding region of the gene that encodes the 10 kDa prolamin polypeptide for select *Oryza* taxa.**

The deduced amino acid residues from the translated open reading frames for select *Oryza* taxa exhibited little heterogeneity (Fig. 3). *Oryza punctata* (BB) shares a nucleotide transversion at residue 9 from a GCC (alanine-A) to an ACC (threonine-T, polar) (Fig 1) with the CCDD taxa (*O. latifolia, O. alta,* and *O. grandiglumis)* (Fig. 3, residue 9). *Oryza minuta* shares a nucleotide transversion at position 313 to a GTG (valine-V, hydrophobic) from an ATG (methionine-M) with *O. rhizomatis* at residue 65.

-click here to view figure-

Fig. 1. Nucleotide alignment of edited *Orya* sequences. Dashes indicate gaps. Question marks indicate unknown data. Ambiguous nucleotides are codes as: K (G/T), M (A/C), R (G/A), S (C/G), and Y (C/T). The 5' end of the signal peptide is indicated by an arrow. The 5' end of the remaining part of the gene that encodes the 10 kDa prolamin polypeptide is likewise indicated by a labeled arrow.

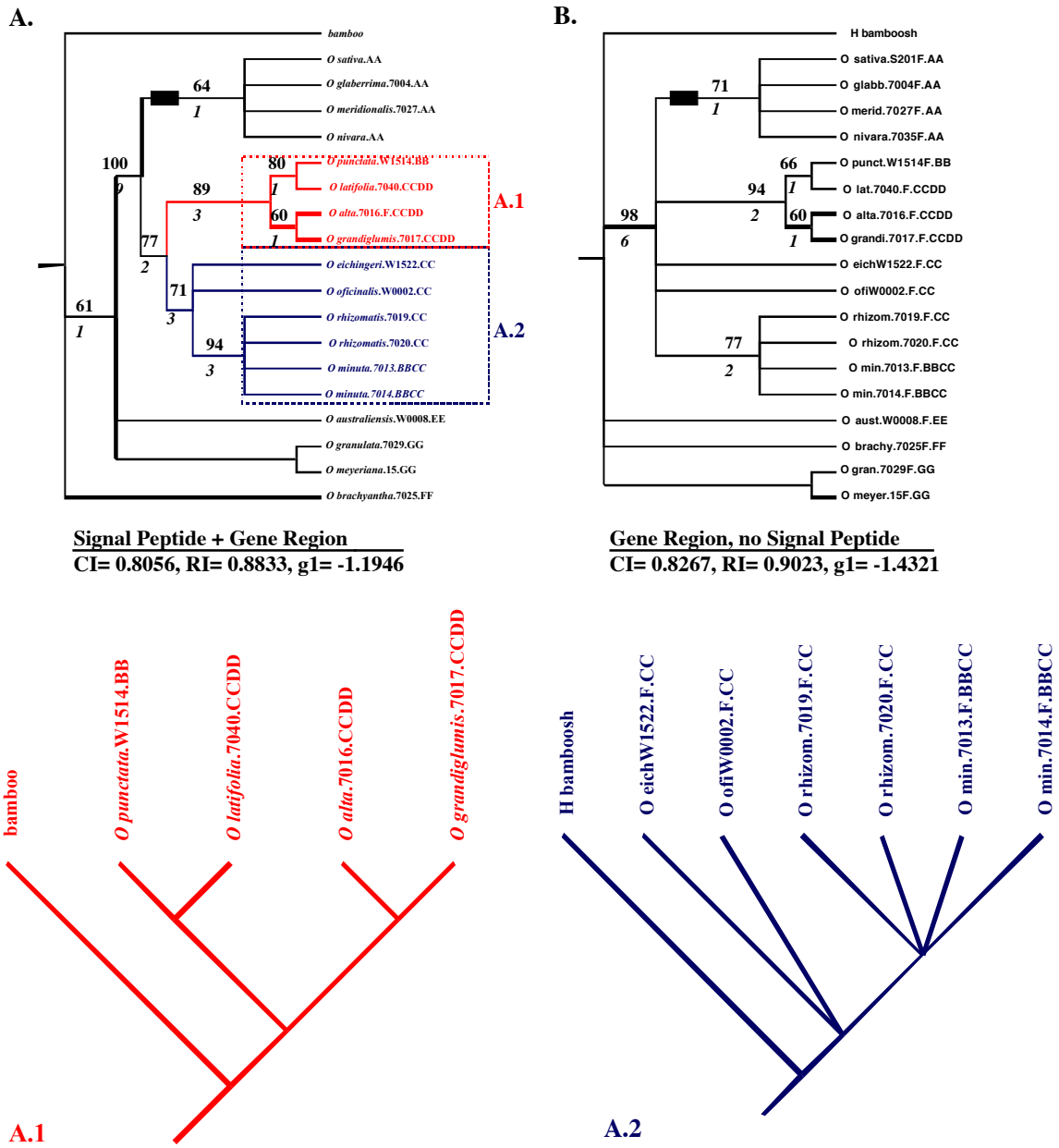



Fig. 2. (A) A strict consensus tree of *Oryza* for the gene that encodes the 10 kDa prolamin polypeptide including the signal peptide region that precedes the gene. Bootstrap values are indicated above each of the branches. Decay values are indicated in *italics* below each of the branches. The vertical bar indicates an ATG indel event that was coded within the matrix. (B) A strict consensus tree for *Oryza* based only on the gene that encodes the 10 kDa prolamin polypeptide. (A-1) Detail of the CCDD/ BB subclade containing *O. latifolia*, *O. grandiglumis*, *O. alta*, and *O. punctata*. (A-2) Detail of the CC/BBCC subclade containing *O. rhizomatis* and *O. minuta*.

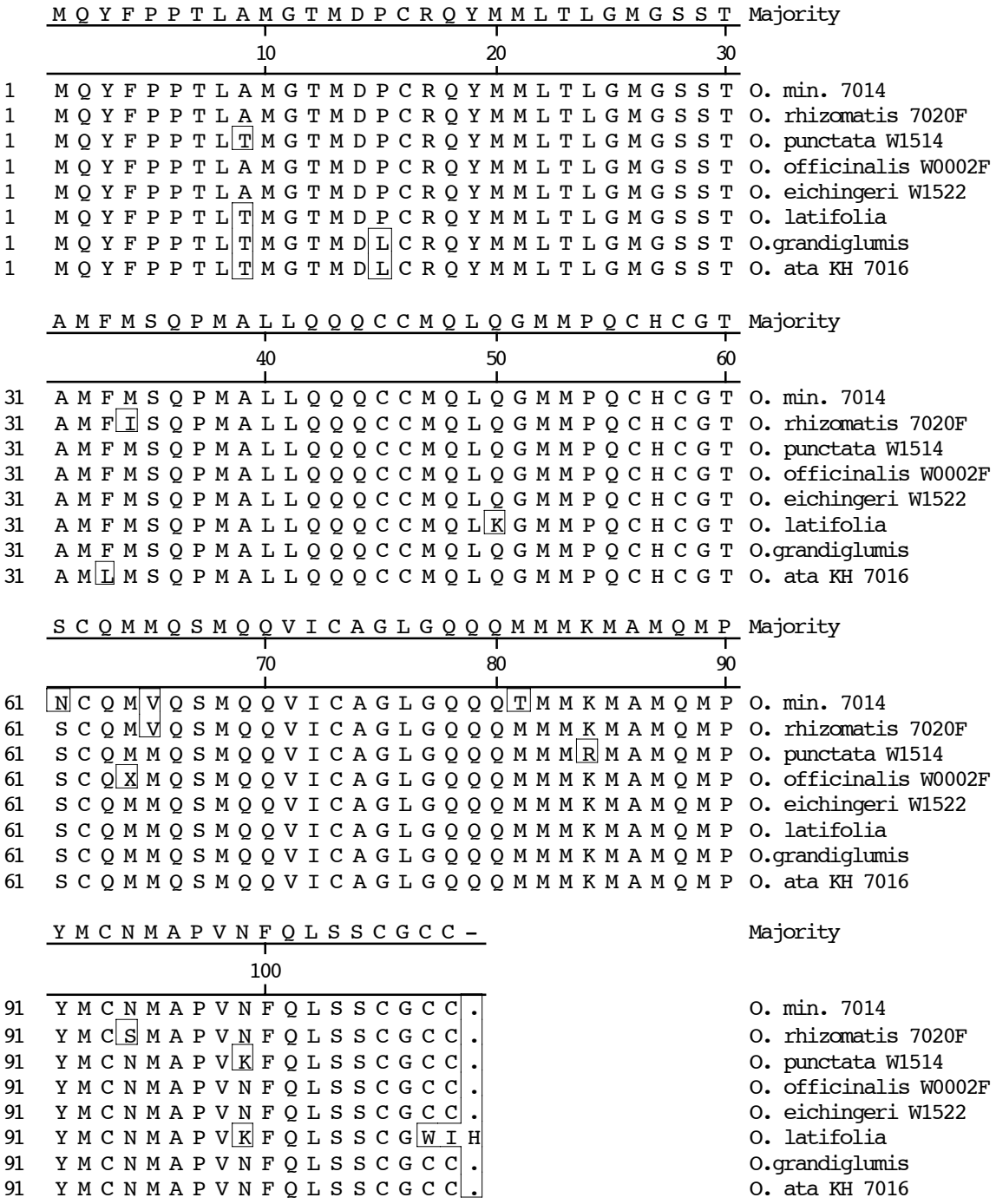



Fig. 3. Deduced amino acid residues translated from the open reading frame, excluding the signal peptide region, of select *Oryza* species. Boxed residues indicate deviations from the consensus sequence. The *Oryza latifolia* does not have a stop codon. "O. min."= *O. minuta*, "O. ata"= *O. alta*

## Discussion

The genus *Oryza* has been partially examined in several phylogenetic (Morishima & Oka, 1959; Second, 1985; McIntyre *et al.*, 1992) and cladistic (McIntyre *et al.*, 1992) analyses; however, no single molecular study has examined the bulk of the genus. Species that are not relevant to the genomic evolution analysis presented here yielded results (Fig 2A& B) that were largely congruent with the previous analyses mentioned above, as discussed in Chapter 2.

In contrast to earlier studies, a novel relationship between *O. minuta* (BBCC) and *O. rhizomatis* (CC) is presented here (Fig. 2D). Specifically, the presence of *O. minuta* (BBCC) in a well-supported clade with *O. rhizomatis* (CC) (Fig 2A) instead of *O. punctata* (BB) suggests that for this particular gene, *O. minuta* is genetically more similar to the CC genome than to the BB genome.

It is unexpected that one half of the "alleles" of the allotetraploid *O. minuta* would not group with the BB parental donor, given that it has been estimated that *Oryza* contans as many as 80-100 copies per halpoid genome (Kim & Okita, 1988a). Instead, the majority of sequences cluster with those of the CC donor. The most straightforward interpretation of these data is as a case of concerted evolution. Concerted evolution (Zimmer et al., 1980) refers to the homogenization of individual genes in multi-gene families so that all gene copies are identical (Zimmer *et al.*, 1980; Arnheim *et al.*, 1980; Zimmer *et al.*, 1988; Rieseberg *et al.*, 1990; Hillis *et al.*, 1991; Wendel *et al.*, 1995; Liao, 1999). In other words, the multiple gene copies within a species are more alike than those among species (Zimmer *et al.*, 1980). The molecular mechanisms by which the process of concerted evolution is thought to operate are biased gene conversion and unequal crossing over, among others (reviewed in Elder & Turner, 1995).

Not only has the nucleotide sequence of the gene that encodes the 10 kDa prolamin polypeptide in *O. minuta* (BBCC) been homogenized to that of *O. rhizomatis* (CC); but this gene in the CCDD taxa has homogenized to the *O. punctata* (BB) copy rather than to that of a species containing the CC genome. These data indicate that the diploid BB genome could be related to the presumed extinct species that originally donated the DD genome. This proposed relationship between *O. latifolia* (CCDD) and *O. punctata* (BB) (Fig 2A,B,D) is not novel (Golpapakrishnan & Sampath, 1966a, b). Based on dot blot analyses of genomic *Oryza* DNA using rDNA insert probes, McIntyre & Winberg (1998) concluded that atleast one of the cloned size fragments of *O. latifolia* (CCDD) did not hybridize to genomic DNA originating from *O. officinalis* (CC), *O. minuta* (BBCC), and the tetraploid form of *O. punctata* (BBCC), further indicating that *the O. latifolia* size fragment may have originated from the DD genome.

As indicated previously, chromosome pairing experiments of hybrid plants between *O. minuta* (BBCC) and *O. latifolia* (CCDD) indicate that they share a common genome (Morinaga, 1943; reviewed in Nayar, 1975); however, no diploid DD parental species has ever been identified in the wild. *Oryza punctata* (BB-Africa) and *O. officinalis* (CC-Asia) have been previously speculated to be progenitors for *O. latifolia* (CCDD) (Golpapakrishnan & Sampath, 1966a; reviewed in Nayar, 1973). The origin of the CCDD genome has further been proposed to have arisen "*in situ* by polyploidy" as a result of a hybridization event involving *O. officinalis* or *O. punctata* that were isolated in South America due to continental drift (reviewed in Nayar, 1975). In other words, the origin of the CC genome in the allotetraploid CCDD species is not in question; however, the origin of the presumed extinct DD genome is yet unresolved. It is possible, given these data (Fig 2A), that *O. punctata* (BB) is related to the extinct DD species, thereby explaining the apparent relationship between *O. punctata* and the CCDD species.

Such a continental drift scenario would indeed be possible given that diploid and allotetraploid populations of *O. punctata* (BB) currently coexist on the western coast of Central Africa (Nigeria, Cameroon, Congo, Angola, Benin, Ghana, Côte d' Ivoire) and that populations of *O. latifolia* (CCDD) are found along the eastern coast of Brazil, Central America, Cuba, Haiti, and the Dominican Republic (Vaughan, 1994). Fossil record evidence, however, suggests that grass diversity was established only 23.7 million years ago during the Miocene period (Hattersley & Watson, 1992), which is approximately 160 million years after the break up of Pangea during the Jurassic (Rast, 1997).

Were it geologically plausible, continental drift could not be invoked as easily to explain the genomic relationship that exists between *O. minuta* (BBCC) and *O. rhizomatis* populations, given the disparate geographic locations of these two species (Table 1). *Oryza rhizomatis* is endemic to Sri Lanka, while *O. minuta* is distributed throughout the Philippines and Papua New Guinea (Vaughan, 1994). Naturally occurring hybrids between these two species have not been reported. However it is likely that the diploid *O. rhizomatis* supplied the CC copy of the gene that encodes the 10 kDa prolamin polypeptide in the tetraploid *O. minuta*, given that in every variable nucleotide position within this gene, *O. rhizomatis* point mutations were predominant in *O. minuta* over any other CC or BB mutation (Fig 1).

Larger theoretical implications, such as selective advantage, arise from the homogenization of gene families resulting in patterns that suggest concerted evolution. It has been suggested that concerted evolution acts as a "'quality control'" in the production of transcripts that are homogeneous in the cell (Liao, 1999). Should this scenario be correct, two questions arise: (1) is "advantageous" residue variation present among species that contain the

CC genome, and (2) if no residue variation is present, is a genome of one species dominant over another, and why.

To address the first issue, an examination of the deduced amino acids from the gene that encodes the 10 kDa prolamin polypeptide indicates that little residue heterogeneity is present in rice and its wild relatives (Fig 3). Thus, little selective advantage could be attributed to the homogenization of *O. minuta* to the genome of *O. rhizomatis*. Moreover, given constraints on protein folding, little residue variability would be expected in the coding region of the gene.

Secondly, the issue of "genome dominance" may arise merely as a result of geographic proximity, or in other words, hybridization between two species over others that contain the same genome may occur merely because the potential diploid genome donors of the hybrid species are physically situated closer to each other. Put differently, is one genome type favored over another merely as a result of introgression?

As mentioned previously, a discussion of hybridization between the parental diploid species that formed the CCDD genome is difficult given the lack of extant diploid DD species. In the case of *O. rhizomatis* (CC- endemic to Sri Lanka) and *O. minuta* (BBCC-endemic to the Philippines and Papua New Guinea), a geographic rationale for the homogenization event between these two species is also difficult given that the *O. officinalis* species used in this analysis (CC-Thailand) is found closer to *O. minuta* (Philippines) geographically than is the *O. rhizomatis* (CC) species. The only other species that contains the CC genome, *O. eichingeri*, is found only in Sri Lanka and central Africa. It should however be noted that species of *O. eichingeri* that was used in this analysis was from Uganda, so a sample obtained from Sri Lanka would be useful for future experiments to determine whether or not it is more closely related to *O. minuta* than the *O. rhizomatis* (CC) sample from Sri Lanka.

In terms of phylogeny reconstruction, concerted evolution would be expected to have little negative effect when species under examination homogenize uniformly to one genome over another. Conversely, concerted evolution can raise questions as to the utility of some genes in phylogenetic analyses when all members of one allopolyploid genome type do not homogenize to one genome over another (Wendel, 1995). A case in point occurs in *Gossypium* (Wendel, 1995). Specifically, some allopolyploid species of *Gossypium* do not concert in a uniform fashion, i.e. the rDNA of four AD species homogenize to the D genome, whereas one species homogenizes to the A genome, thereby leading to a misleading phylogeny construction. *Oryza* does not seem to encounter the same problems as *Gossypium* in that all members of the CCDD genome have been homogenized to the BB genome copy of this gene, and that *O. minuta* has homogenized to the copies of one of the species that contains the CC genome, specifically *O. rhizomatis*. The allotetraploid form of *O. punctata* has not been sequenced in this analysis.

To rule out PCR bias of one genome over another when examining the evolutionary relationships of the allotetraploid *Oryza* species to their presumed diploid progenitors, both PCR drift and PCR selection should be considered as potential factors (Polz & Cavanaugh, 1998). In order to minimize PCR drift (stochastic fluctuations in the reaction mixture), several PCR reactions were combined, precipitated, and then sequenced in both directions.

Bias as a result of PCR selection (property of the templates) can be examined by adding known concentrations of two templates into one PCR reaction to determine if one is preferentially amplified over another. Such "genome doping" experiments using degenerate 16S rDNA primers to PCR amplify known concentrations of bacterial genomes yielded ratios of PCR products that did not correspond to the original genomic template ratios (Polz & Cavanaugh, 1998). It should be noted, however, that it was determined that the GC content of the priming

region significantly effected PCR efficiency of the templates, while gene dosage did not effect product ratios.

Barbier and Ishihama (1990) further concluded that mixtures of PCR products from the rice cultivar *O. rufipogon* Nihonbare (AA) and *O. longisteminata* (AA) (syn. *O. barthii*-AA) at a ratio of 1:10 yielded almost no sequence variation, however, at a ratio of 1:5, minor template was detected. Further doping experiments involving the remaining *Oryza* genomes can be performed to rule out bias and confirm that the process of concerted evolution has indeed occurred in the allotetraploid *Oryza* species.

To conclude, the allotetraploid species of *Oryza* (BBCC, CCDD) have been examined for trends regarding genome origin and mechanisms of evolution within those genomes. *Oryza minuta* (BBCC) is closely related to the diploid CC species *O. rhizomatis*, whereas *O. punctata* (BB) is closely related to the allopolyploid CCDD taxa *O. latifolia*, *O. alta*, and *O. grandiglumis*. Evidence presented here suggests that concerted evolution has occurred in these *Oryza* sequences.

## Acknowledgements

This work was funded through grants from the Virginia Academy of Science, *Sigma Xi*, the Graduate Research Development Project, the Department of Biology, Virginia Tech, and the Jeffrees Foundation. DNA was generously provided by Dr. Yoshio Sano, Hokaido University, Soporro, Japan. Dried plant material was provided by Christoph Neihous, University of Bonn. Special thanks are extended to Dr. B. Lu, who confirmed all rice seed lot designations with

herbarium vouchers for specimens housed at the International Rice Research Institute, Philippines.

## References


Aggarwal, K.; Brar, D.S.; Khush, G.S. 1997. Two new genomes in the *Oryza* complex identified on the basis of molecular divergence analysis using total genomic DNA hybridization. *Mol. Gen. Genet.* (254): 1-12.

Arnheim, N.M.; Krystal, M.; Schmickel, R.; Wilson, G.; Ryder, O.; Zimmer, E. 1980. Molecular evidence for genetic exchanges among ribosomal genes on non-homologous chromosomes in man and apes. *Proc. Natl. Acad. Sci. USA.* (77): 7323-7327.

Barbier, P.; Ishihama, A. 1990. Variation in the nucleotide sequence of a prolamin gene family in wild rice. *Plant Mol. Biol.* (15): 191-195.

Bremmer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution.* (42): 795-803.

Donoghue, M.J.; Olmstead, R.G.; Smith, J.F.; Palmer, J.D. 1992. Phylogenetic relationships of Dipsacales based on *rbc*L sequences. *Ann. Missouri Bot. Gard.* (79): 333-345.

Elder, J.F.; Turner, B.J. 1995. Concerted evolution of repetitive DNA sequences in eukaryotes. *The Quart. Rev. Biol.* (70): 297-320.

Eriksson, T. 1998. AutoDecay ver. 4.0 (program distributed by the author). Department of Botany, Stockholm University, Stockholm.

Golpalakrishnan, R. Sampath, S. 1966. The American species of *Oryza*. *Oryza*. (3): 35-40.

Gopalakrishnan, R., & Sampath, S. 1967. Taxanomic status and origin of American tetraploid species of the series *Latifoliae* Tateoka in the genus *Oryza*. *Indian J. Agr. Sci*. (37): 466-475.

Hillis, D.M.; Moritz, C.; Porter, C.A.; Baker, R.J. 1991. Evidence for biased gene conversion in concerted evolution of ribosomal DNA. *Science*. (252): 308-310.

Hillis, D.M.; Huelsenbeck, J.P. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. *J. Heredity* (83): 189-195.

Kim, W.T.; Okita, T.W. 1988. Structure, expression, and heterogeneity of the rice seed prolamines. *Plant Physiol*. (88): 649-655.

Liao, D. 1999. Molecular evolution '99. Concerted evolution: molecular mechanism and biological implications. *Am. J. Hum. Genet*. (64): 24-30.

Masumura, T.; Shibata, D.; Hibino, T.; Kato, T.; Kawabe, K.; Takeba, G.; Tanaka, K.; Fuji, S. 1989. cDNA cloning of an mRNA encoding a sulfur-rich 10 kDa prolamin polypeptide in rice seeds. *Plant Mol. Biol*. (12): 123-130.

McIntyre, C.L.; Winberg, B.; Houchins, K.; Appels, R.; Baum, B.R. 1992. Relationships between Oryza species (Poaceae) based on 5S DNA sequences. *Pl. Syst. Evol*. (183): 249-264.

McIntyre, C.L.; Winberg, B.C. 1998. A rapid means of identifying wild rice species DNA using dot blots and genome specific rDNA probes. *Genome*. (41): 391-395.

Morinaga, T. 1943. Cytogenetical studies on *Oryza* sativa L. VI. The cytogenetics of $F_1$

hybrids of *O. minuta Presl.* and *O. latifolia* Desv. *Jap. J. Bot.* (12): 347-357.

Morishima, H.; Oka, H.-I. 1959. The pattern of interspecific variation in the genus Oryza: its quantitative representation by statistical methods. *Evolution* (14): 153-165.

Nayar, N.M. 1973. Origin and cytogenetics of rice. *Advances in genetics.* New York: Academic Press. (V14): 153-292.

Polz, M.F.; Cavanaugh, C.M. 1998. Bias in template-to-product ratios in multi-template PCR. *Appl. and Environ. Micro.* (64): 3724-3730.

Rast, N. 1997. Mechanism and sequence of assembly and dispersal of supercontinents. *J. Geodynamics.* (23): 155-172.

Rieseberg, L.H.; Carter, R.; Zona, S. 1990. Molecular tests of hypothesized hybrid origin of the two diploid *Helianthus* species (Asteraceae). *Evolution* (44): 1498-1511.

Sanderson, M.J. & Donoghue, M.J. 1989. Patterns of variation in levels of homoplasy. *Evolution.* (8): 1781-1795.

Saghai-Maroof, M.A.; Soliman, K.M.; Jorgensen, R.A.; Allard, R.W. 1984. Ribosomal DNA spacer-length polymorphism in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci.* USA (81): 8014-8018.

Vaughan, D. 1994. *The wild relatives of rice: A genetic resources handbook.* International Rice Research Institute, Los Baños, Philippines: 1-101.

Wendel, J.F. Schnabel, A.; Seelanan, T. 1995. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proc. Natl. Acad. Sci. USA.* (92): 280-284.

Zimmer, E.A.; Martin, S.L.; Beverley, S.M.; Kan, Y.W.; Wilson, A.C. 1980. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc. Natl. Acad. Sci. USA.* (77): 2158-2162.

## Conclusion

This phylogenetic analysis of nineteen *Oryza* species was founded on the hypothesis that genetic diversity exists at the nucleotide level in wild species of *Oryza*, and that this diversity can provide information regarding species relationships and permit a more theoretical examination of genomic evolution and its mechanisms. The gene that encodes the 10 kDa prolamin polypeptide was used as a tool for investigating the phylogenetic relationships within this genus. Moreover, given that rice is an economically important crop, it was anticipated that a thorough evaluation of the deduced amino acid sequences of the coding region of the gene would provide insight into the variation present in the wild *Oryza* species, thereby permitting rational gene targeting regimes for crop protein enhancement.

Two phylogenies were constructed based on two data subsets composed of nucleotide sequences for the gene that encoded the 10 kDa prolamin polypeptide: (1) one that included the coding region as well as its signal peptide, and (2) one that included the coding region of the gene only. The resulting phylogenies were topologically different, and the trees generated from the matrix that included the signal peptide region had greater resolution. Three major clades emerged, and are correlated with genome composition: the (1) ABCD, (2) EE, and (3) GG. The ABCD clade was further divided into the A and BCD subclades, which supported Lu's (personal communication, 1999) classification of *Oryza* and deviated from Vaughan's (1994) classification that included *O. australiensis* (EE) among the BCD species. A comparative analysis of the deduced amino acid composition for the gene that encodes the 10 kDa prolamin polypeptide indicates that amino acid variability is present among rice and its wild relatives, and that the

species within the subclades containing the allotetraploid *Oryza* species exhibit greater residue variability relative to cultivated rice.

A focused analysis of the subclades containing the allotetraploid *Oryza* taxa permitted a more detailed examination of the putative origin and evolution of the hybrid species. The subclade containing the allotetraploid BBCC species *O. minuta* and the diploid *O. rhizomatis* (CC) species has strong statistical support (95% bootstrap, decay of 3) pointing to a close evolutionary relationship between these two species. Likewise, the subclade containing the allotetraploid CCDD taxa and *O. punctata* (BB) demonstrated strong support (94% bootstrap, decay of 3) for a close relationship between these geographically diverse species. The mechanism by which the multiple gene copies for the gene that encodes the 10 kDa prolamin polypeptide have been homogenized, is speculated to be concerted evolution, which refers to the observation that individual genes in multi-gene families tend to homogenize so that all of the copies are identical.

This analysis represents the first comprehensive molecular data set derived from nucleotide sequences to examine the majority of the *Oryza* genus. The resulting phylogenies provide support not found previously for hypotheses regarding the origin of *Oryza* species (CCDD taxa), and have presented relationships not previously proposed (*O. minuta*/*O. rhizomatis*). A detailed analysis has also indicated which *Oryza* species are variable in their deduced amino acid composition for the gene that encodes the 10 kDa prolamin polypeptide, thereby highlighting target taxa for future agronomic analyses.

# Vita

# Irene M. Mullins

Biology Department
2199 Derring Hall
Virginia Polytechnic Institute & State University
Blacksburg, VA 24061-0406
Tel. 540-231-5407
Fax # 540-231-9307
email: iboyle@vt.edu

## Education

M.S., candidate, Biology, Virginia Polytechnic Institute & State University; QCA: 3.6/4.0
B.A., Biology & Spanish 1997, Mount Holyoke College, GPA: 3.5/4.0
Honors Thesis: Is a molecular lamina present in higher plant cells, a structural approach.

## Awards and Recognition

Graduate Teaching Assistantship, Virginia Polytechnic Institute and State University, Fall 1997-
Graduation *cum laude* and with High Honors in Biology Mount Holyoke College, Fall 1997
Mary Lyon Scholar, Mount Holyoke College, 1997
Undergraduate Honors Thesis passed with High Honors in Biology, Mount Holyoke College, 1997
Election into *Sigma Xi*, Mount Holyoke College, Fall 1997
Bernice MacLean Award for excellence in biology during junior year, Mount Holyoke College, Spring 1996

## Grants

Graduate Research Development Project (GRDP), A comparative molecular study of protein genes in the rice genus *Oryza*. (funded-$300.00; matched by department)
Virginia Academy of Science, A comparative molecular study of the genetic variation in protein genes of rice and related species. (funded-$1221.00; $500.00 matched by department)
*Sigma Xi* Grant-in-Aid, A comparative molecular study of the genetic variation in protein genes of rice and related species. (funded-$464.00; matched by department)

## Membership in Professional Societies

*Sigma Xi* (1997-)
Virginia Academy of Science (1998-)
American Association for the Advancement of Science (1999-)

## Professional Meetings Attended

American Institute of Biological Sciences (AIBS) August 1998; Baltimore, Maryland
Missouri Botanical Garden Symposium on Plant Systematics, October, 1998, St. Louis, Missouri

## Talks Presented

Evolutionary relationships of the rice genus inferred from the prolamin (seed storage protein) gene. Botany Seminar: Department of Biology; Virginia Tech, Va, April 10, 1999.

Evolutionary relationships of the rice genus inferred from the prolamin (seed storage protein) genes. Virginia Academy of Science: Norfolk, Va: May 26-28, 1999.

Concerted evolution in the rice prolamin genes. Society of Systematic Biologists. Evolution '99 Meetings: Madison, Wisconsin, June 22-26, 1999.

## Posters Presented

Phylogeny of *Oryza*, Poaceae, inferred from prolamin (seed storage protein) gene (poster). International Botanical Congress. St. Louis, Missouri: Aug 1-7, 1999.

## Teaching Experience

General Biology/ Principles of Biology Lab Teaching Assistant, Virginia Polytechnic Institute and State University, 1997-1999
Evaluations: 3.7/4.0 (Fall 1997); 3.7/4.0 (Spring 1998); 4.0/4.0 (Summer 1998); 3.7/4.0 (Fall 1998)

Principles of Biology for Majors Lab Teaching Assistant, Virginia Polytechnic Institute and State University 1999-
Evaluations: 3.6/4.0 (Spring 1999); 3.7/4.0 (Fall 1999)