Molecular Diversity, Structure and Domestication of Grasses

EDWARD S. BUCKLER IV1,2*, JEFFRY M. THORNSBERRY2 AND STEPHEN KRESOVICH3
1 USDA-ARS, Plant Science Research Unit
2 Dept. of Genetics, North Carolina State University, Raleigh, NC 27965-7614
3 Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853-2703

Over the last 10,000 years, crop domestication has been the single most important human cultural development. Grasses are prominent among these crops, and provide the vast majority of the world’s food. Similar traits have been selected during the domestication and breeding of these critically important grasses, and since they share a similar complement of genes, the same set of genes may have been selected. Even though the process of domestication occurred over the same 5000 to 10,000 year period, the domesticated grasses have major differences in genome structure, diversity, and life history. Molecular investigations of grass domestication have succeeded in identifying progenitor species and are beginning to catalog genetic resources. Additionally, research is now elucidating some of the basic processes by which crops have evolved over the last few millennia. In this review, we discuss our present knowledge of molecular diversity among the grass crops and relate that diversity to the genes involved in domestication and to yield gains. Understanding the connection between diversity and genome structure will be critical to future crop breeding.

Genome Structure

Three major grass crops – maize, wheat, and rice – share a common ancestor within the last 55–70 million years (Kellogg, 2001). Despite the differentiation that has occurred over this time period, chromosomal synteny between each of these genomes remains (Ahn et al., 1993). Recently, genome sequencing has clarified the question of whether synteny is maintained over smaller regions (Bennetzen, 2000; Chen et al., 1997; Tarchini et al., 2000). These investigations have highlighted the preservation of colinear regions across grass species, but there are also many cases where individual genes are missing or locally rearranged. In addition, the amount of repetitive DNA found in genes conserved between species varies greatly, due in large part to retrotransposons (SanMiguel et al., 1996). For example, the maize genomic distance between sh2 and a1 is 140 kb, while in rice and sorghum the distance is roughly 19 kb (Chen et al., 1997).

In general, basic gene order has been preserved. An interesting contrast arises with respect to differences in gene copy number (Hancock, 1992). Relative to other grasses, rice appears to have few duplicated genomic regions, although genome sequencing has and will continue to uncover some regions of genome duplication. Wheat, a recent hexaploid, incorporates three genomes with varying levels of divergence, while maize appears to have undergone an ancient tetraploidization event roughly 10 to 20 million years ago (Gaut and Doebley, 1997). Recent statistical searches for map colinearity indicate that many parts of the maize genome may have actually been duplicated, triplicated, and quadruplicated (Gaut, 2001). Overall, it appears that 60–80% of the maize genome is duplicated. Across these important domesticates, there is a large range of variation in terms of gene copy number.

These genome duplications provide new mutational opportunities for creating greater phenotypic diversity. In maize, many of these duplicated genes are expressed at slightly different times in development (Van der Meer et al., 1993). Diversity in expression and copy number is also possible through the elimination of DNA sequences in polyploids. Relative to its progenitors, hexaploid wheat appears to have deleted many low copy DNA sequences since the polyploidization event (Feldman et al., 1997).
Mutational inactivation of paralogous loci also provides less severe phenotypes than might normally be found in the diploid state. For example, wheat breeders have been able to use mutants of the duplicated gibberellin insensitivity loci to design wheat with specific heights (Peng et al., 1999). Further studies contrasting the relative rates of insertion and deletion events should soon shed light on the reasons for the enormous variation in c-value among plant species, as recently argued for invertebrate genome evolution (Petrov et al. 2000).

**What genes have been the targets of domestication and breeding?**

Despite the independent domestication of the four major cereal complexes, (maize in America; wheat, barley, oats and rye in the Near East; rice in Asia; sorghum and millet in Africa), the earliest plant ‘selectors’ desired the same sets of traits. Wild grasses that flowered in short-days and produced small, naturally dispersed seeds were transformed into domesticates in which flowering time was unaffected by day length, and which produced large seeds necessitating human planting and harvesting.

The probable mechanism of this convergent domestication across grasses was selection at a common set of loci. Quantitative trait loci (QTL) for seed size, seed dispersal (shattering), and photoperiod have been mapped in maize, rice, and sorghum (Paterson et al., 1995). These QTL correspond to homologous regions between taxa more often than would be expected by chance, suggesting that homologous loci/genes may be involved in the evolution of these phenotypes. Three QTL that affect seed size correspond closely in sorghum, rice, and maize (Paterson et al., 1995) and explain large portions of the phenotypic variation in seed size when the taxa are compared in a pairwise fashion. A single seed dispersal locus was mapped in sorghum that corresponds to a single rice QTL on chromosome 9 and to maize QTL on the duplicated regions of chromosomes 4 and 1. QTL relating to flowering time and photoperiod also show correspondence (Lin et al., 1995). For example, a QTL on chromosome 10 of maize corresponds to a region of the sorghum genome bearing Mal, the locus responsible for 85% of the flowering time variation in sorghum. Lin et al. (1995) also demonstrated that this correspondence of QTL could be extended to wheat and barley populations. Overall, these QTL correspondences suggest that domestication of these grasses was the result of mutations in a small number of genes with potentially large effects.

Due to the difficulties of map-based cloning of QTL, it is unclear if the genes underlying these corresponding QTL are identical. However, some evidence indicates that the targets of breeding and domestication continue to be the same genes across the species. For example, mutations of the wheat RhtD1a and rice Gai orthologues both affect plant height and flowering time (Peng et al., 1999). This gene appears to have played a role in the “Green Revolution” varieties of wheat that increased yield greatly in the 1960s and 1970s. Association and selection tests of the maize orthologue dwarf8 also suggest that this locus is currently a target of selection and adaptation of maize to various flowering times (Thornsberry et al., submitted). Orthologous loci have also been targeted for grain processing; for example, low amylose was selected for in both rice and maize by using the starch synthase-encoding waxy loci (Ishikii et al., 1998; Shure et al., 1983).

**How has diversity changed during domestication?**

Numerous studies have examined changes in diversity between wild relatives and domesticated grass species. Most of these studies were done with isozymes, SSRs, or RFLPs. These markers have been very useful for addressing evolutionary relationships and comparing diversity within species. However, differences in experimental systems make comparisons between species difficult. Nucleotide diversity studies have the primary advantage of being more comparable between laboratories and experimental systems. The main limitation with any type of diversity survey is that there can be a wide variance in diversity between loci, and only in maize have a large number of loci been examined thus far (Gaut et al., 2000).

Maize (Zea mays ssp. mays) nucleotide diversity at silent sites averages 1-6% for genes that appear to be behaving neutrally, while the diversity in maize’s wild relative Z. mays ssp. parviglumis is roughly 2% (Gaut et al., 2000; White and Doebley, 1999). At individual loci, diversity estimates have ranged from 0-2% to 5% (White and Doebley, 1999). There has been roughly a 30% drop in diversity at the average locus from maize’s wild relatives. Relative to other grasses, maize and its wild relatives appear to have high levels of genetic diversity, which is probably the result of high levels of outcrossing. Maize is a monoecious species with male and female reproductive parts that are physically separated, which facilitates outcrossing. Long-term effective population size is estimated at roughly a million plants, although smaller populations could have persisted for shorter times (Eyre-Walker et al. 1998).

The drop in diversity is substantially greater at genes involved in domestication. One maize gene involved in domestication, teosinte branched1, controls tillering and apical dominance, and was key in converting maize from a plant with multiple stalks...
(tillers) to a plant with a single tiller (Doebley et al., 1997). The promoter of this locus has 61-fold lower diversity in the crop than it does in the closest wild relative (Wang et al., 1999). Interestingly, this drop in diversity does not extend for the entire length of the gene, as the coding region has levels of diversity similar to those at neutral loci. Analysis of this data assuming a reasonable set of population genetic parameters suggests that the process of domestication could have taken at least hundreds of years for Adh1 with only modest levels of selection (Wang et al., 1999). Other surveys in maize kernel starch accumulation have found additional genes involved in domestication and breeding (Buckler in preparation).

At these loci, low levels of diversity generally persist for the entire length of the gene indicating selection intensity or recombination rates could vary dramatically for each domestication locus. Currently, we do not have a good estimate of how many genes were involved in the domestication of maize.

In sorghum (Sorghum bicolor), surveys of nucleotide diversity of the members of the phytoclone gene family (phyA, phyB and phyC) in a set of wild subspecies, including the progenitors for cultivated sorghum, show an average diversity of 0.35%.

However the diversity is heterogeneous between the gene family members, ranging from 0.14% (phyA) to 0.5% (phyC) (G. White, in preparation). Isozyme surveys suggest that domesticated Sorghum bicolor also has about two-thirds of the diversity of its wild relative (Morden et al., 1990). A single survey of Adh1 nucleotide diversity in millet (Pennisetum glaucum) indicates that cultivated type has modest diversity of 0.24% while the wild progenitor is at 0.36% (Gaut and Clegg, 1993b).

Wheat (Triticum aestivum) is a hexaploid (designated AABBDD). If wheat were the product of a single polyploidization event (AB x D), the likely result would have been a complete genetic bottleneck.

However, this model is not consistent with some of the molecular surveys for the wheat genome. For example, the A1 locus of the D genome has two very distinct haplotypes, which are found in both wheat and Triticum tauschii (D genome progenitor) (Talbert et al., 1998). Overall diversity was roughly 1% at the locus in both the domesticated T. aestivum and the wild T. tauschii (Talbert et al., 1998). A wider survey of the nucleotide diversity of the B and D genomes at anonymous loci suggests that average diversity in Triticum aestivum is 0.6% for the B genome and 0.4% for the D genome (Blake et al., 1999). In wheat, estimates of RFLP diversity at RbcS have indicated that hexaploid wheat has perhaps 30% of the diversity levels found in its diploid relatives, but there are substantial differences between the A, B, and D genomes (Galili et al., 2000).

In barley (Hordeum vulgare), the three examined loci exhibit very different patterns of diversity. Surveys of Adh1 and Bkn-3 found low diversity levels of roughly 0.1% to 0.2% (Badr et al., 2000; Cummings and Clegg, 1998; Petersen and Seberg, 1998). Diversity levels were roughly the same in both the domesticated and wild subspecies for these loci. In contrast, the Adh3 locus in the wild H. vulgare ssp. spontaneum had diversity 10 fold higher at roughly 2.2% (Lin et al., 2001). This high level of diversity results from two very different alleles that were distributed along geographic lines, unlike Adh1. Adh3 diversity may be the product of ancient geographic separation combined with introgression and selection in this predominantly selfing subspecies. More surveys would be needed to clarify this fascinating situation in Hordeum.

Nucleotide diversity studies have not been carried out in hexaploid oat (Avena sativa), but isozyme surveys suggest that the domesticate has roughly two-thirds of the diversity in the hexaploid wild relative, Avena sterilis (Murphy and Phillips, 1993).

In rice (Oryza sativa), few nucleotide diversity studies have been conducted. A survey of a phytochrome intron in rice’s wild relative Oryza rufipogon indicated the nucleotide diversity was 0.35% (Barbier et al., 1991), while diversity at the multicycloc prolamain family found silent diversity of 0.73% (Barbier and Ishihama, 1990). The evolutionary dynamics of this multicycloc gene family are not comparable to the single copy loci. Divergence between O. sativa ssp. japonica and indica at the waxy locus promoter is 0.83% (Hirano et al., 1998). Isozyme data suggests that domesticated rice has roughly 71% of the diversity of its wild relatives (Oka, 1988).

Overall, it is rather surprising how much diversity has been maintained in these grasses relative to their wild progenitors. In general, the domesticated relatives have two-thirds of the diversity found in wild relatives (Table 1). There has probably been a greater loss in terms of alleles for agronomic use, as nucleotide diversity estimates are relatively insensitive to the loss of rare alleles. The most likely factor for the maintenance of diversity in domesticated grasses is that they are generally used as a basis for subsistence. Large quantities of grass (and grain) are required before they are useful. For example, if 10 people derived 10% of the calories from wild Triticum or Zea, they would have to grow roughly 1–6 ha of plants (Hillman and Davies, 1990). Roughly 250,000–350,000 plants would have to be grown annually. Even in limited geographic regions, many millions of plants would have been grown during cultivation and the early stages of domestication. Although people likely applied strong selection during various phases of domestication, the bottlenecks were likely less severe in the grasses because of the need to grow large numbers for subsistence. The relatively small drops in diversity at neutral loci would correlate with domi-
Yield estimates are from FAO average yields in France from Zohary, estimates are higher than would be expected in the wild. When wild estimates were unavailable, estimates from landrace domestication. This dramatic rise in yield from a maize yield has increased 55 fold through its history of (Evans of the yield gain in maize, wheat, and barley over the

recent yield gain 3–60 fold, and this increase is continuing as most of the major grass species have exhibited two to three fold increases in yield over the last 40 years. Although some of this gain is through intensification of agricultural inputs, a substantial portion reflects gains by genetics. Generally, estimates suggest that 40–80% of the yield gain in maize, wheat, and barley over the last century has been from genetic improvements (Evans et al., 1993; Hallauer et al., 1988). Impressively, maize yield has increased 55 fold through its history of domestication. This dramatic rise in yield from a physiological perspective probably reflects the poor yield of maize’s wild relatives. But from a genetic perspective, this increase was likely possible only through maize’s high diversity, which is 3–10 fold higher than other grasses (Table 1).

Domestication has radically changed many grass species, and in many cases human selection has moved these species towards similar phenotypes and adaptations. It is likely that many of the same genes have been involved in all of these domestication events, but the origin of the genetic variation, which allows for the modification of these species, still remains unclear. The variation necessary for selection may come from genome duplications, large effective population size, and/or high mutation rates. The two most productive domesticates (maize and wheat) both have many duplicated loci, yet maize has higher diversity compared to wheat. To more fully understand the importance of genome structure and diversity in the grasses it will be necessary to sample nucleotide diversity and genome structure from some of the domestication failures. For example, prior to the shift to maize, Mexican Setaria species appear to have been more productive than maize and more widely used throughout Mexico, but domestication does not appear to have occurred (Callen, 1967). Was this a result of the species not responding to selection? One method of addressing this question would be to contrast nucleotide diversity at a sample of unlinked loci from several cultivars of unsuccessful grasses with nucleotide diversity in successful grass domesticates. It would be important to do this for loci that are not expected to play a role in domestication, as selection

Table 1. Ploidy, nucleotide diversity, and yield differences between domesticated crops and wild progenitors

<table>
<thead>
<tr>
<th>Ploidy</th>
<th>Zea mays</th>
<th>Sorghum bicolor</th>
<th>Orzya sativa</th>
<th>Arena sativa</th>
<th>Hordeum vulgare</th>
<th>Triticum aestivum</th>
<th>Pennisetum glaucum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Progenitor Diversity</td>
<td>0.0210</td>
<td>0.0035</td>
<td>0.0035</td>
<td>Hexaploid</td>
<td>0.0014</td>
<td>0.0070</td>
<td>0.0003</td>
</tr>
<tr>
<td>Cultivar Diversity</td>
<td>0.0163</td>
<td>0.0035</td>
<td>0.0035</td>
<td>Hexaploid</td>
<td>0.0016</td>
<td>0.0050</td>
<td>0.0002</td>
</tr>
<tr>
<td>Cult./Prog. Diversity</td>
<td>0.16*</td>
<td>&lt;0.60*</td>
<td>1.12*</td>
<td>2.93*</td>
<td>0.65*</td>
<td>0.65&lt;0.55</td>
<td></td>
</tr>
<tr>
<td>Wild Yield</td>
<td>0.65*</td>
<td>2.4</td>
<td>4.1</td>
<td>1.8</td>
<td>2.4</td>
<td>2.4</td>
<td>1.1</td>
</tr>
<tr>
<td>Yield in 1961</td>
<td>2.5</td>
<td>2.4</td>
<td>4.1</td>
<td>1.8</td>
<td>2.4</td>
<td>2.4</td>
<td>1.1</td>
</tr>
<tr>
<td>Yield in 2000</td>
<td>9.1</td>
<td>5.9</td>
<td>5.6</td>
<td>4.5</td>
<td>6.4</td>
<td>7.1</td>
<td>1.5</td>
</tr>
<tr>
<td>Recent Yield Gain</td>
<td>3.6</td>
<td>2.5</td>
<td>1.4</td>
<td>2.5</td>
<td>2.7</td>
<td>3.0</td>
<td>1.4</td>
</tr>
<tr>
<td>Gain Over Progenitor</td>
<td>55.9</td>
<td>&gt;99</td>
<td>5.0</td>
<td>1.5</td>
<td>9.8</td>
<td>11.0</td>
<td>&gt;27</td>
</tr>
</tbody>
</table>

1. Maize has many duplications and may be an ancient tetraploid.
2. Ratio of cultivar to wild progenitor nucleotide diversity. If nucleotide diversity estimates were missing, isozyme data was substituted. References for diversity estimates can be found in the text. Averages based on only one or two loci are indicated by *
3. Estimates of yield for wild progenitors come from both the wild (*) and agricultural field stations (*). The field station estimates are higher than would be expected in the wild. When wild estimates were unavailable, estimates from landrace cultivars (*) in the region of domestication were substituted (ICRISAT, Murphy and Frey, 1984; Oka, 1988; Wilkes, 1967; Zohary, 1969).
4. Yield estimates are from FAO average yields in France from 1961 and 2000.

What gains in yield have been realized from grass domestication and breeding?

The first targets of domestication were likely genes that made harvesting easier and allowed expansion into new environments. Soon after, yield must have become a primary objective as greater yield would have substantially reduced the labor input and land needs. Domestication and breeding have been remarkably successful at increasing yield in the grasses for a small group of today’s cultivars. For the most successful cultivars, the average yields of wild stands were probably under 1 Mg/ha compared to today’s yield of 4–9 Mg/ha (Table 1).

This represents a gain over the wild yield of roughly 5–10 fold, and this increase is continuing as most of the major grass species have exhibited two to three fold increases in yield over the last 40 years. Although some of this gain is through intensification of agricultural inputs, a substantial portion reflects gains by genetics. Generally, estimates suggest that 40–80% of the yield gain in maize, wheat, and barley over the last century has been from genetic improvements (Evans et al., 1993; Hallauer et al., 1988). Impressively, maize yield has increased 55 fold through its history of domestication. This dramatic rise in yield from a
at specific loci can greatly skew diversity estimates. The research at the \textit{Adh1} locus for three grasses has already been very informative in suggesting how different life history traits may relate to diversity (Cummings and Clegg, 1998; Gaut and Clegg, 1993a; Gaut and Clegg, 1993b), but with high throughput sequencing becoming more accessible, surveys across multiple taxa and loci are now becoming feasible.

As the human population continues to grow and arable land becomes limited, it is critical that substantial yield increases continue. The grasses are key to meeting these needs, and their similar genomes provide a unique opportunity to use them as a single evolutionary genetic and functional genomic system (Freeling, 2001). Understanding why grass domestication has succeeded or failed in the past should provide important knowledge on how to exploit diversity and genome structure for future agricultural improvement.

We thank Gemma White for contributing unpublished data. We also thank Carlyn Keith Buckler, Julie Ho, Sherry Whitt, Ken Olsen, and Greg Gibson for commenting on this manuscript. JMT was supported with support by NSF grant DBI-9872631.

References


