Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
GENOMICS PROTEOMICS & BIOINFORMATICS

Article

Evolutionary Transients in the Rice Transcriptome

Jun Wang[1,2,3#*], Jianguo Zhang[4#], Ruiqiang Li[2,3#], Hongkun Zheng[2,3#], Jun Li[2], Yong Zhang[2], Heng Li[2], Peixiang Ni[2], Songgang Li[2], Shengting Li[2], Jingqiang Wang[1], Dongyuan Liu[1], Jason McDermott[5,6], Ram Samudrala[5], Siqi Liu[1,2], Jian Wang[1,2], Huanming Yang[1,2], Jun Yu[1*], and Gane Ka-Shu Wong[1,2,7*]

[1]*Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China;*
[2]*Beijing Genomics Institute at Shenzhen, Shenzhen 518083, China;*
[3]*Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230, Odense M, Denmark;*
[4]*Theoretical Life Research Center, Fudan University, Shanghai 200433, China;*
[5]*Computational Genomics Group, Department of Microbiology, University of Washington, Seattle, WA 98195, USA;*
[6]*Computational Biology and Bioinformatics, Pacific Northwest National Laboratory, Richland, WA 99352, USA;*
[7]*Department of Biological Sciences and Department of Medicine, University of Alberta, Edmonton AB, T6G 2E9, Canada.*

Genomics Proteomics Bioinformatics 2010 Dec; 8(4): 211-228 DOI: 10.1016/S1672-0229(10)60023-X

Abstract

In the canonical version of evolution by gene duplication, one copy is kept unaltered while the other is free to evolve. This process of evolutionary experimentation can persist for millions of years. Since it is so short lived in comparison to the lifetime of the core genes that make up the majority of most genomes, a substantial fraction of the genome and the transcriptome may—in principle—be attributable to what we will refer to as "evolutionary transients", referring here to both the process and the genes that have gone or are undergoing this process. Using the rice gene set as a test case, we argue that this phenomenon goes a long way towards explaining why there are so many more rice genes than *Arabidopsis* genes, and why most excess rice genes show low similarity to eudicots.

Key words: evolutionary transients, rice, gene duplication

Introduction

Within any given genome there is an evolutionarily stable core of genes (which in this manuscript we will limit to proteins) and a smaller subset of lineage specific genes. Here, we wish to raise awareness of an often-ignored aspect of the evolutionary process that is not only responsible for creating many lineage specific genes but also endows them with very distinctive characteristics. More specifically, we consider the canonical version of evolution by gene duplication whereby one copy is kept unaltered while the other is free to evolve. This process of evolutionary experimentation persists for many millions of years (*1*), and at any given time some number of genes must be undergoing this process. The only question is how many. Because this process is so short lived in comparison to the core genes, it is apt to use the term "evolutionary transients" for both the process and the genes that

# Equal contribution.
*Corresponding authors.
E-mail: wangj@genomics.org.cn; junyu@big.ac.cn; gksw@genomics.org.cn

have gone or are undergoing this process. We will provide evidence to indicate that, at least for rice, such transients may account for a surprisingly large fraction of the genome and the transcriptome.

The issue of gene content variation in plants arose during our sequencing of the rice genome, where even the reduced estimate of 38,000 to 40,000 genes (*2, 3*) was much larger than the 25,498 genes found in *Arabidopsis* (*4*). A vast majority of the excess rice genes show only low similarity (LS) to *Arabidopsis* (*5*) and more generally to all eudicots. Although some of the annotations in 2002 were transposable elements (TEs) (*6*), all reasonable efforts were made to correct this problem in the annotations in 2005. Subsequent analyses, with many incorporating transcript and protein data, show that the rice gene count is mostly correct and that the excess of LS genes relative to *Arabidopsis* is real. For example, two expression-microarray experiments have detected the transcripts for up to 35,970 rice genes. The first experiment studied 70-mer oligos designed from the annotations (*7*), while the second tiled all non-repetitive sequences (*8*). Even in a set of 19,079 non-redundant full-length rice cDNAs (*9*) referred to as nr-KOME, there are many LS genes to *Arabidopsis*. Data from other grass species provide even more support for this excess of LS genes: ESTs from maize (*10*) and sugarcane (*11*), methylation-filtered genome sequences from sorghum (*12*), and phylogenetic profiling of 252,383 non-redundant ESTs and proteins from 32 plant species (*13*). Finally, comparative analyses with 4 plant genomes and the transcripts of 185 plant species support 38,109 rice genes (*14*).

We have chosen to focus on a specific mechanism to explain the excess of LS genes in rice—not because it is the only explanation, but because it is so rarely considered and it is clearly a major effect. Gene duplications are an especially frequent occurrence in plants (*2*)—at all levels of organization, from whole genomes to chromosomal segments, and especially individual genes. Our hypothesis was that, if new duplications are created as rapidly as they are destroyed, then evolutionary transients can have a persistent effect on the genome and the transcriptome. The observed loss of similarity would arise from the post-duplicative degradation (*i.e.*, relaxed selection) that is expected to occur in one of the two gene copies. We would however note that if, as is often the case for rice, there were multiple duplication events since the divergence from *Arabidopsis*, the loss of similarity may have occurred in the distant past, not the most recent duplication. Our objective will not be to trace the loss of similarity to any specific duplication event, but rather to show that many LS genes have many of the characteristics expected for genes undergoing the hypothesized process, as summarized in **Table 1**. It is necessarily an underestimate, as the nature of the degradation process renders the more ancient duplication events difficult to recognize.

Table 1 Summary of HS and LS gene characteristics

	HS genes	LS genes
Non-redundant cDNA	12,528	6,551
Transcript length	1,867 (1,714)	1,435 (1,249)
Open reading frame	1,231 (1,116)	552 (414)
In a single exon	12.7%	33.9%
mRNA confirmation	85.0%	65.5%
confirmed by EST	69.1%	44.2%
confirmed by SAGE	68.3%	49.5%
Protein confirmed	20.6%	10.6%
GO function classified	57.8%	17.1%
Duplicated in rice	75.5%	35.6%
Ks average	0.515 (0.560)	0.226 (0.096)
Ka/Ks average	0.358 (0.322)	0.649 (0.680)
Ka/Ks <1.0	93.1%	61.8%
Ka/Ks <0.5	69.5%	27.2%
Conserved in maize	98.9%	41.7%
Ks average	0.631 (0.618)	0.646 (0.643)
Ka/Ks average	0.237 (0.183)	0.410 (0.357)
Ka/Ks <1.0	99.0%	93.9%
Ka/Ks <0.5	92.9%	71.1%
Conserved in sorghum	95.5%	39.7%
In maize or sorghum	99.6%	46.9%
In maize and sorghum	94.8%	34.4%
Protein disorder(2)	1.0%	19.3%
Protein disorder(3)	6.5%	40.1%

Note: HS, high similarity; LS, low similarity. "In a single exon" is computed as a percentage of the cDNAs that are at least 95% aligned. Ka/Ks data are normalized to a subset of genes, either the maximal set of homolog pairs in Figure 2 or those conserved in maize. For all other cases, the normalization is against 12,528 HS and 6,551 LS genes. Average quantities indicate mean (median). For Ks and Ka/Ks, the mean is computed by concatenating all the genes into a single long sequence. Disorder(2) genes are those for which the encoded protein is 50% covered by low-complexity sequence (LCS) or 50% covered by remark465. In the same way, disorder(3) genes combine LCS, remark465, and hot loops.

Evolutionary transients muddy the very concept of what is gene, if we include a requirement of functionality, because a duplicate that is currently in a state of degradation may still eventually develop a novel function. Given this intrinsic uncertainty, it may be difficult to ever define the precise gene number for rice. Our paper will therefore focus on the nr-KOME cDNAs, to discount annotation errors and ensure that what we study is at least transcribed, regardless of its evolutionary fate, which is unknowable. The number of functional genes in rice and *Arabidopsis* may turn out to be similar, but at this stage of our understanding it remains difficult to say.

Post-duplicative gene degradation

Gene duplications are a major source of evolutionary innovation (*15*), and they have been especially ubiquitous in plant evolution. In the canonical model, one copy is left unchanged while the other is free to evolve. The eventual outcome is either neofunctionalization, where the second copy acquires a new or modified function and is preserved in the genome, or nonfunctionalization, where it keeps degrading and is lost from the genome. There is another model called subfunctionalization, where both copies are preserved in a tissue-differentiated manner; however, comparative analysis of yeast genomes has argued that the canonical model is by far the more common fate (*16*). Much of the literature on gene duplication focuses on the eventual outcome, ignoring the issue of how one gets to that outcome. However, to understand the excess of LS genes in rice, and by extension all plants, the processes that occur in the aftermath of duplication must be considered.

Multiple lines of evidence support the following. First, changes in the expression levels are needed to compensate for the immediate doubling in gene dosage. In natural and synthetic plant polyploids, this is observed after only a few generations (*17-20*). Given the rapidity by which these changes are observed, epigenetic mechanisms are assumed to be involved, as well as genetic mechanisms. To say that one gene copy is free to evolve is equivalent to saying that one gene copy is under relaxed selection. In protein-coding regions, the signature is an increase in the ratio of non-synonymous (Ka) to synonymous (Ks) amino acid substitutions. This is too well documented, albeit mostly in animal genomes (*21-24*). All told, the estimated half-life of these post-duplicative processes range from 4 million years (My) in animal genomes to 17 My in plant genomes (*25*).

Duplications can occur on any length scale, from individual genes to multigene segments that encompass much of a chromosome (or even the whole genome). As a rule, the larger the duplication, the less frequently it occurs. Thus in comparison, one can say that duplication of multigene segments is episodic, while duplication of individual genes is continuous. Although the legacy of the post-duplicative process can be long lived, in the observed excess of LS genes, to the extent that evolutionary transients can have a persistent effect on the observed transcriptome (*i.e.*, nr-KOME cDNAs), it will be for duplications that are of sufficiently recent origins to be still transcribed at a detectible level. Thus, much of the reported effect will be due to duplication of individual genes. Since duplication history can vary with species, so will the number of LS genes. For example, analysis of 14 plant species (*26*) showed that *Arabidopsis* has the fewest tandem gene duplications of recent origin. One should not be surprised if it has fewer LS genes.

Not all gene duplicates are fated to die. A few may reemerge from the period of relaxed selection with a new or modified function that not only improves its likelihood of survival, but also results in it being maintained in the genome. Regardless, all will have experienced some period of relaxed selection, and hence similarity will be lost to varying degrees. Because the half-life concept refers to an exponential distribution, one cannot say with certainty that a gene's fate is decided when the time since its duplication greatly exceeds the nominal half-life. This is the underlying source of uncertainty as regards the gene number, since as a result one can only make probabilistic statements.

Results

Definitions of low and high similarity

For our purposes, the threshold between LS and high similarity (HS) is not a critical parameter, because our

primary objective is to explain why so many rice genes appear unrelated to *Arabidopsis* when in fact they probably share a common ancestor, through one or more duplication events. We therefore define the threshold with the BLAST family of alignment tools because it is so widely used. In contrast, when we infer biological function by similarity to previously characterized genes, we use state-of-the-art methods like PSI-BLAST (*27*), as implemented in the Bioverse pipeline (*28, 29*). By not defining the threshold at the limits of detection for sequence similarity, we allow ourselves a small glimpse at what LS genes might do biologically.

To be consistent with our previous work (*5*), we will use the same parameters. Note that what we now call LS was previously called NH. Specifically, we translate the coding sequence into protein and search the *Arabidopsis* genome in all six reading frames using TBlastN at an E-value of 10^{-7}. Successive exons are linked together by a dynamic programming algorithm and the result is only accepted if at least 50% of the protein or 100 amino acids are aligned. The optimal trade-off between false positives and negatives is found at E-values of 10^{-3} to 10^{-5}, based on an analysis for *Drosophila* that considered how reliably named versus unnamed genes are distinguished (*30*). However, we find that how we do the similarity detection has little effect on the number of LS genes. As defined, 34.3% of rice cDNAs are LS genes. Using the state-of-the-art methods in Bioverse would only have brought this estimate down to 28.5%. In other words, for the vast majority of LS genes, there is no detectable similarity to *Arabidopsis* regardless of the methodology that is used.

Applying the same definition to the 13,737 full-length *Arabidopsis* cDNAs in GenBank, we find that 20.0% are LS genes in comparison to rice. We believe the smaller percentage of LS genes reflects the prior observation that there are fewer instances of recent tandem gene duplications in *Arabidopsis*.

Other contributors to the LS effect

There are other explanations for the excess of LS rice genes, but we believe these are lesser contributors. One possibility is that a rice gene only appears to be LS because its homolog was lost from the *Arabidopsis* lineage. To assess the extent of this effect, we took the rice cDNAs and applied the same TBlastN procedure as described above to search all eudicot sequences in GenBank. Similarity to eudicots was seen for 20.3% of LS genes and 98.1% of HS genes. It suggests that those 20.3% of LS genes, 1,330 genes in all, should be redefined as HS genes. Along the same lines, the fact that using a more state-of-the- art method to detect similarity decreases the fraction of the rice cDNAs with LS status from 34.3% to 28.5% suggests that those LS genes whose status changed, 1,117 genes in all, should also be redefined as HS genes. Comparing these 1,330 and 1,117 LS genes, we find that 434 are shared. The underlying issues are intertwined, having as much to do with limits of detection for sequence similarity as with actual gene loss from a lineage. The core issue is why that similarity is being lost.

Another possible contributor to the LS effect is that some of the cDNAs might be transcribed TEs. This is a difficult issue to correct for if the TE databases are less than complete, which is often the case. However, we have developed a method called ReAS to recover all of the ancestral TEs from the raw data of a whole genome shotgun (*31*). The validity of this method for reducing TE contamination has been independently demonstrated; it gives the same results as a PFAM structure based method (*32*). For rice, comparisons to all known and ReAS-recovered TEs identify just 246 likely TEs among 19,079 cDNAs. It is also known that some TEs can incorporate and transmit nuclear gene transcripts (*33*). We used LTR_STRUC (*34*) to search for long terminal repeats (LTRs) and test if they are preferentially found near LS genes, finding that 3.0% and 3.8% of HS and LS genes, respectively, are flanked by LTRs at distances of up to 20 kb. These results indicate that TEs are not a confounding factor.

Duplications within the rice genome

If post-duplicative degradation is a major factor in the excess of LS rice genes, this should be apparent when we compare duplication histories for LS and HS genes. We have previously constructed a duplication history of the rice genome (*2*). A simplified representation is presented in **Figure 1A**. Our method differs from many others currently in the literature, as we do

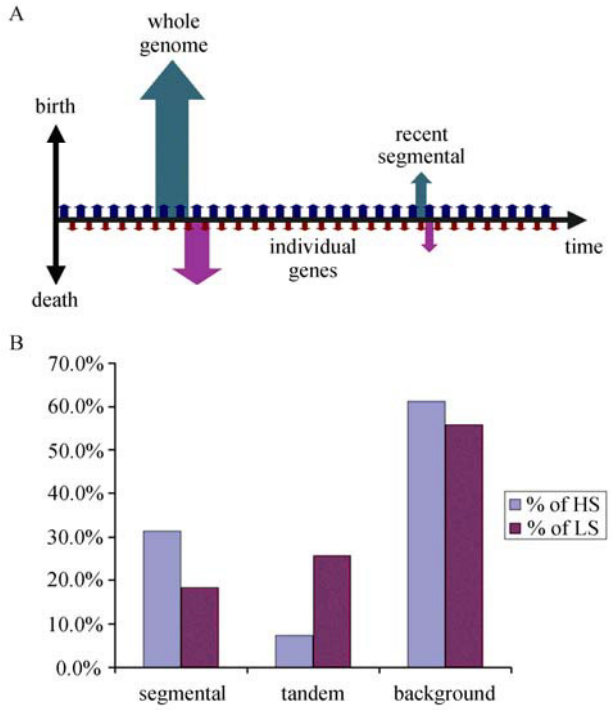

Figure 1 Rice duplication history. This figure is based on our previous paper (*2*). **A**. Every homolog pair is put into one of the three duplication categories: segmental, tandem and background. Segmental duplications are episodic. There is evidence of a whole-genome duplication before the divergence of the grasses 55 to 70 million years ago (Mya), and of a more recent sub-chromosomal duplication 21 Mya. Duplication of individual genes (*i.e.*, tandem and background) is effectively continuous. **B**. The bar chart shows the number of homolog pairs in each duplication category. Restricting to cDNAs with one-and-only-one duplicate in rice, there are 609, 311, and 1,351 homolog pairs, respectively. These are subdivided into HS and LS genes to show relative contributions.

not make use of predicted genes from official annotations. Instead, we translate the cDNA's coding sequence into protein and search the genome in all six reading frames with TBlastN, much as we do to define LS status, but now requiring that at least 50% of the protein be aligned. A "homolog pair" is defined as a cDNA and one of its many duplicates identified with the above procedure. Duplicates identified by TBlastN need not be functional genes. They can be remnants of ancient duplications that are not in the annotations. For example, 48.0% of the tandem duplications that we identified would never have been found had we used the TIGR rice genome annotations (*35*) instead. We believe this is appropriate, as our objective is not to show that a cDNA is part of a family of active genes, but rather that it has been subject to a duplication event.

One of the more striking observations is that multiple duplications of the same gene is the norm. In searching for homologs (duplicates) of a given cDNA, we either find no homologs at all, or a mean (median) of 40 (5) homologs. It befits the adage, "when it rains it pours". This has important consequences, as it suggests that similarity need not be lost in a single duplication event, but can be the cumulative effect of many duplications over time. This presents a challenge to the duplication analysis because, depending on the circumstances, it may not always be easy to determine which homolog to consider. We discuss this in greater detail in Materials and Methods, but as a rule, when in doubt we only used those cDNAs with a single homolog so that there is no ambiguity.

Every homolog pair was placed into one of the three possible duplication categories, defined in our 2005 paper as segmental, tandem, and background (*2*). Segmental duplications refer to a collinear set of genes all duplicated at the same time. The other two categories are individual gene duplications. In the tandem case, the gene and its copy are adjacent to each other on the chromosome; in the background case they are non-adjacent, and even on different chromosomes in many instances. We found 18 pairs of duplicated segments, covering 65.7% of the genome, with 17 pairs representing a whole-genome duplication (WGD) dating to a time just before the divergence of the grasses 55 to 70 million years ago (Mya) (*36*). The final pair was a recent duplication that connected chromosome 11 to 12 and dated to 21 Mya. Note that both duplication events occurred at discrete times well after the divergence of rice and *Arabidopsis* 170 to 235 Mya (*37*). In contrast, duplication of individual genes was found to occur so frequently that one can effectively think of their occurrence as a continuous activity.

To appreciate the relative importance of each duplication category, we start by considering the subset of the cDNAs with one-and-only-one homolog (duplicate) in rice. This is required because higher-order homologs, *i.e.*, those cDNAs with more than one putative duplicate, are incorporated using different rules for each of the three duplication categories. **Figure 1B** depicts a total of 609, 311, and 1,351 homolog

pairs, respectively. Decomposed into LS and HS genes, one finds that HS genes are dominant in segmental, LS genes are dominant in tandem, and neither is preferred in background duplications. This is telling us something about the relative ages of LS and HS genes, but to be precise we must compute the number of Ka and Ks changes per available site. K-estimator (*38*) is used to make the multiple substitution corrections. Ks can then be converted to the time since duplication, by assuming a neutral substitution rate of 6.5×10^{-9} per silent site per year (*39*).

Most of the following is taken from Figures 8 and 10 of Yu *et al*, 2005 (*2*), where higher-order homologs were used to increase the number of homolog pairs to 1,340, 1,685, and 1,351 in the three duplication categories. A bimodal Ks distribution was seen in the segmental duplications, with the major mode attributed to the WGD, and the minor mode attributed to the recent segmental duplication on chromosomes 11 and 12. In tandem and background duplications, Ks distributions were strongly peaked at zero and accompanied by a long exponential tail. Averaged by duplication category, the mean (median) for Ks was found to be 0.592 (0.604), 0.253 (0.118), and 0.594 (0.560), respectively. These data show that tandem duplications are of more recent origins than segmental or background duplications. The fact that LS genes are dominant in tandem duplications basically means LS genes are associated with younger duplications.

We know that younger (recent) duplications are more likely to be under relaxed selection, and this point is made explicitly by **Figure 2**, which shows Ka/Ks as a function of Ks for homolog pairs from LS and HS genes. Ka/Ks ratios are normally much less than 1.0, because most non-synonymous changes are deleterious and rapidly purified from the population. Under relaxed selection, the ratios increase toward 1.0. When greater than 1.0, the ratios are taken as evidence of positive selection. It is also possible to think of selection as acting on a subset of the protein sequence, in which case, an increase in Ka/Ks might be due to a subset being under positive selection. Regardless of duplication category, LS genes tend to have smaller Ks. Averaged over all categories, the mean (median) for Ks is 0.515 (0.560) in HS genes and 0.226 (0.096) in LS genes. The typical half-life for the post-duplicative degradation phase is 4 to 17 My. This translates into a Ks of 0.052 to 0.221, which is essentially what we find in the LS genes. The implication is that many LS genes may be under relaxed selection. What we see is that 30.5% and 6.9% of HS genes have Ka/Ks of greater than 0.5 and 1.0, respectively; in contrast for LS genes, the corresponding percentages are much larger, 72.8% and 38.2%.

Increase in Ka/Ks as Ks→0 is the defining feature of the evolutionary transients that we believe are a major contributor to the LS effect. Not many things will lead to such a signature. For example, if our gene set is contaminated by random genomic DNA, we would see an increase in both Ks and Ka/Ks. A more

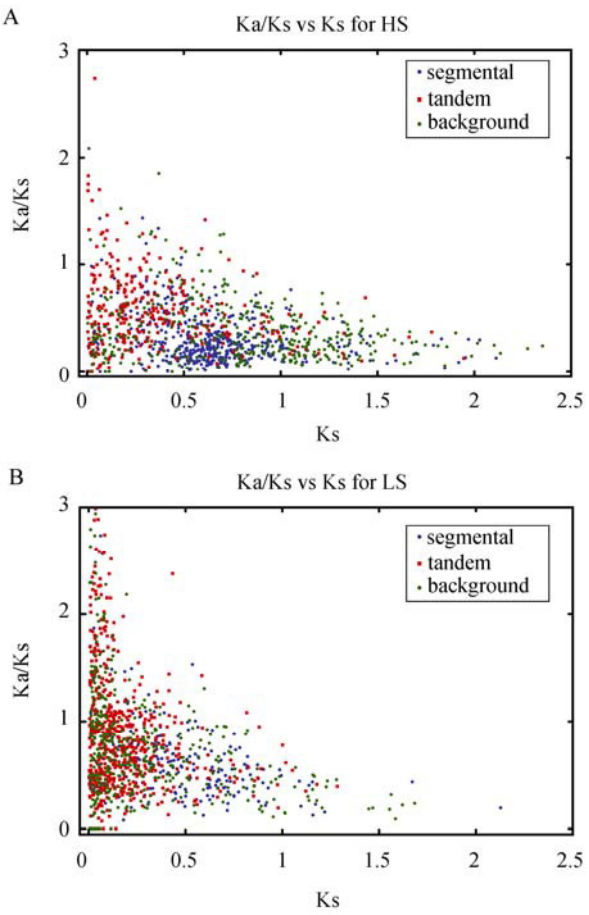

Figure 2 Ka/Ks in rice duplications. We maximize the number of homolog pairs in each duplication category, with slightly different rules in each category, resulting 1,340, 1,685, and 1,351 homolog pairs for segmental, tandem, and background duplications. Ks is the time since duplication. Ka/Ks versus Ks is shown for HS genes (**A**) and LS genes (**B**). Scatter plots like these are sensitive to the number of data points. Because there are more HS genes than LS genes, the HS plot depicts a random subset of the data equal in size to the LS plot.

subtle possibility is that the number of nucleotide substitutions available to compute Ka/Ks also goes to zero in the limit of Ks→0. Hence the observed signature may be due to statistical noise. To show that this is not the case, we counted the number of nucleotide substitutions for homolog pairs with Ka/Ks in the ranges of 0−1, 1−2, and 2−3. For LS genes, the mean (median) is 85 (33), 50 (20), and 23 (14) homolog pairs, respectively. That is more than enough to compute an increase in Ka/Ks ratio to 1 or 2 significant digits.

Conservation in other grass species

Strictly speaking, computing Ka/Ks for rice duplicates does not allow us to say which of the two copies is under relaxed selection. Is it the cDNA, the TBlastN match, or both? We want to argue that it is the cDNA, but to do so we must compare it to another species that is more closely related to rice than *Arabidopsis*. Given the plant phylogeny of **Figure 3A**, the ideal species is another grass. None of them has been completely sequenced, but what we do have are gene-enriched sequences from maize (*40, 41*) and sorghum (*12*), by methylation filtering and/or high C_0t selection. Although these data provide only fragmentary coverage for each gene, they tag almost every gene. We search the data by using TBlastN at an E-value of 10^{-7}, but considering the fragmentary nature of these data, no further conditions are imposed. **Figure 3B** shows that 98.9% and 95.5% of HS genes are conserved in maize and sorghum, while 94.8% are conserved in both. In comparison, 41.7% and 39.7% of LS genes are conserved in maize and sorghum, while 34.4% are conserved in both.

Figure 4 shows Ka/Ks and Ks values computed from the comparison of rice to maize. Note that Ks is now a measure of the time since the divergence of rice and maize, not the time since the duplication within rice. We would thus expect similar Ks distributions in HS and LS genes, and indeed that is what we see. The mean (median) for Ks is 0.631 (0.618) in HS genes and 0.646 (0.643) in LS genes. We do however expect to see much larger Ka/Ks in the LS genes, and again our expectations are met. Specifically, 7.1% and 1.0% of HS genes have Ka/Ks above 0.5 and 1.0, respectively; in contrast for LS genes, these percentages are 28.9% and 6.1%. As a point of reference, mammalian

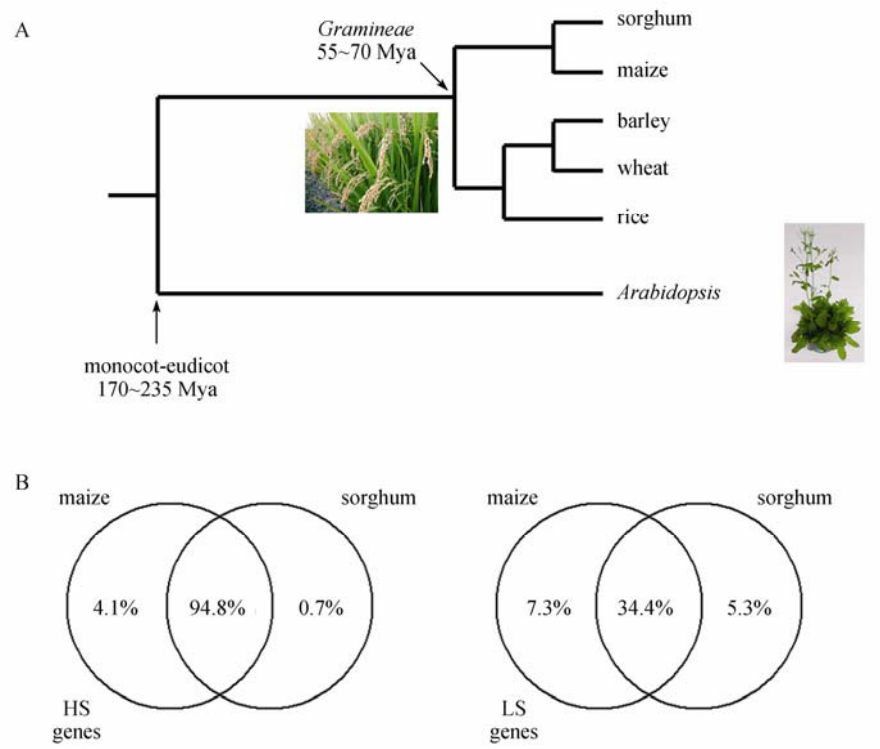

Figure 3 Cross-species conservation. **A**. A phylogeny of the Gramineae (grasses) and their relationship to the model eudicot *Arabidopsis*. **B**. Venn diagrams for percentage of HS and LS genes conserved in the genomes of maize and sorghum.

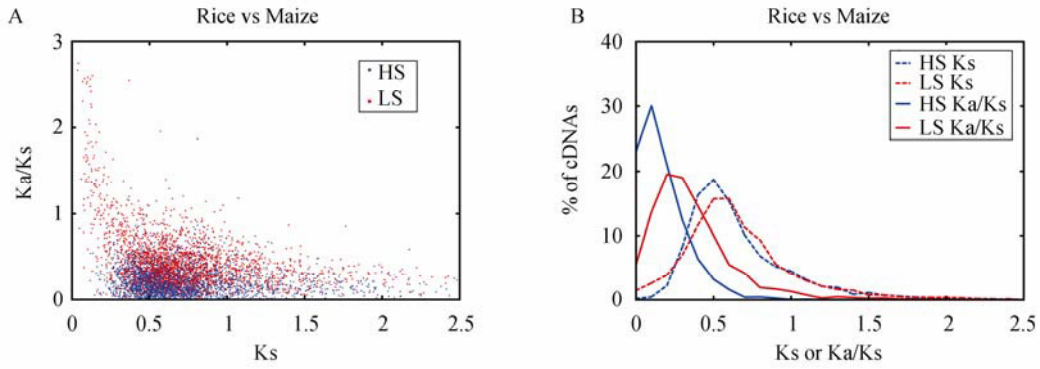

Figure 4 Ka/Ks in maize homologs. There are 12,392 HS and 2,731 LS genes for which we have maize data. Ks is the time since divergence of rice and maize. **A**. Ka/Ks versus Ks. **B**. Distributions for Ka/Ks and Ks. The distribution plots clearly show that although the Ks values are comparable, there is an increase in Ka/Ks for LS genes relative to HS genes. To equalize the datasets, the HS plot depicts a random subset of the data equal in size to the LS plot.

genomes report a mean for Ka/Ks of 0.20 in primates and 0.14 in rodents (*42*). For the rice to maize comparison, we find a mean (median) for Ka/Ks of 0.237 (0.183) in HS genes and 0.410 (0.357) in LS genes. Similar results were found with sorghum, but for brevity we omit these numbers.

More recently, the rice annotations were compared to a much larger dataset of plant transcripts (185 species in all) (*14*). This suggested that many LS genes have similarity to other grasses. We repeated that analysis using the TIGR Plant Transcript Assemblies (01/30/2007 release, 233 species), which we subdivided into three categories: non-rice Gramineae, non-rice monocot, and non-*Arabidopsis* eudicot. The searches used the same procedures as used in the definition of LS genes. Not surprisingly, we find that 99.8%, 99.8%, and 99.6% of HS genes are conserved in these three plant categories, respectively. In contrast for LS genes, we find that 45.4%, 45.6%, and 24.5% are conserved. Despite adding many more plant species, these results were remarkably stable. For example, 41.7% of LS genes are conserved in gene-enriched maize sequences alone, versus 45.6% in non-rice monocot transcripts.

If evolutionary transients are a major contributor to the LS effect, and given that the post-duplicative degradation phase lasts 4 to 17 My, we must reconcile this time scale with the observation that many LS genes are found in grass genomes that diverged 55 to 70 Mya. The resolution of this paradox lies in the fact that multiple duplications of the same gene is the norm. Recall that for a given cDNA, we either find no homologs at all, or a mean (median) of 40 (5) homologs. Hence loss of similarity to *Arabidopsis* need not occur in a single duplication event. It can be the cumulative effect of many such events. This also presents a challenge for the cross-species analysis when we assign orthologs. In maize and sorghum, where the gene set is reasonably complete, a "best hits" criterion is suitable. Nonetheless to strengthen our argument, we present additional and independent evidence to support the idea that many LS genes are in evolutionary flux.

mRNA and protein expression level

Figure 5A describes what is expected to happen in the post-duplicative degradation phase. Right away, there is a reduction in gene expression levels to compensate for the doubling in gene dosage. One should therefore expect to find lower levels of mRNA and protein expression in LS genes. The expression levels can be estimated by counting the number of times each cDNA is confirmed in EST, SAGE and proteomics. We used data from a variety of tissues and physiological conditions, as shown in Table S1. There are 104,903 ESTs and 431,853 10-mer SAGE tags that together confirm 85.0% of HS genes and 65.5% of LS genes. Note also that by finding the cDNAs in multiple independent experiments we confirm that they represent true gene transcripts. Proteomics data are less sensitive, but they confirm 20.6% of HS genes and 10.6% of LS genes. Note that we find 3,276 proteins in total, consistent with the "thousands of

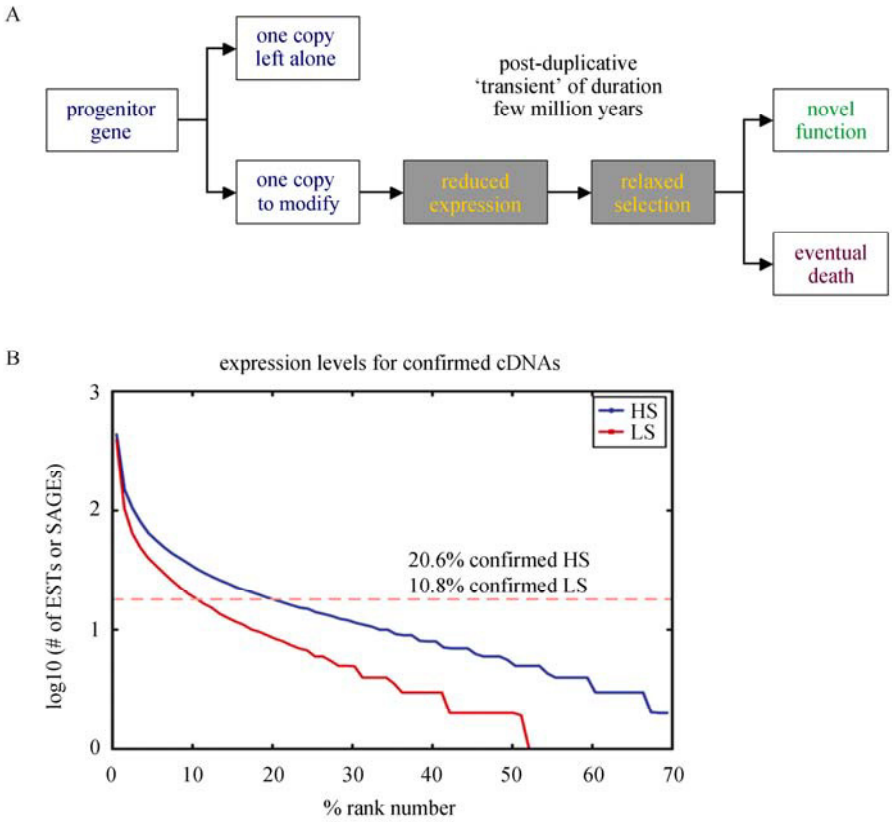

Figure 5 Post-duplicative "transients". **A**. A schematic for the most commonly observed outcome, where one of the two copies either dies or evolves a new function. Note that it is also possible for both copies to survive, via subfunctionalization. **B**. Expression level based on mRNA and proteomics data. Each gene is ranked according to the number of confirming EST or SAGE tags. The proteomics detection limit is indicated by a horizontal line intersecting the HS data at its 20.6 percentile. Extrapolation to the LS data predicts that it should be confirmed at 10.8% (versus observed rate of 10.6%).

proteins" limit for state-of-the-art proteomics (*43, 44*). Although some LS genes are just as highly expressed as the most abundant HS genes, most are expressed at systematically lower levels. This is summarized in **Figure 5B**, where we rank the cDNAs based on the number of confirming EST or SAGE tags, and plot their expression levels as a function of rank order. These low expression levels could explain why there had been so few reports of LS genes in the past. For example, before nr-KOME cDNAs became available, 10.9% of 787 full-length cDNAs in GenBank were LS genes. It is likely that LS genes remain under-represented, even with the latest data. We can also show that the mRNA and protein levels are mutually consistent. Suppose we draw a horizontal line on Figure 5B at the mRNA expression level corresponding to the upper 20.6% of HS genes that are confirmed by proteomics, the intersection of this line with the LS genes data predicts a proteomics confirmation rate of 10.8%, which is remarkably close to the actual rate of 10.6%.

Some concerns have been raised about the quality of the rice cDNA data (*45*), because it came from two different libraries, where 43.5% of 19,079 cDNAs came from FAIS, and 56.5% came from RIKEN. But if we consider only the subset of LS genes, 57.9% came from FAIS, while 42.1% came from RIKEN. There is certainly a bias in that more LS genes came from FAIS, but many came from RIKEN as well. We note, however, that the proteomics confirmation rates for LS genes from the two libraries are comparable, 10.4% for FAIS and 10.8% for RIKEN. The LS gene bias has a plausible explanation when one considers that the FAIS library was sampled from 21 tissues and conditions, versus only 3 tissues and conditions for RIKEN, as summarized in Table S2. This would be consistent with the prevalence of a form of

post-duplicative evolution called subfunctionalization, where both copies are preserved to serve different purposes in different tissues (*46*).

Mutational bias and protein disorder

There is another way to know which genes are in evolutionary flux. Rice has a mutational mechanism that leaves a distinctive signature in the protein-coding sequences of its evolving genes, as reflected by the compositional gradients found in Gramineae, but not eudicots (*47*). Compared to *Arabidopsis*, rice genes are more GC-rich in general, especially at the 5'-ends. In Figure S1 we show that the distribution of GC-contents is bimodal, with the higher GC mode favoring LS genes. The differences are mostly in the third codon position, which does not usually affect the encoded proteins. However, LS genes are also on average 5% to 10% more GC-rich than HS genes in their second codon position. Studies have been made of the amino acid substitution patterns (*48*), but to appreciate how similarity is lost, it is even more informative to consider the effects on 3D protein structure.

Many protein sequence alignment algorithms (*e.g.*, BlastP) have trouble dealing with LS genes because they contain too much low-complexity sequences (LCSs) (*49*). "AALAGKAVANAKV" and "KSAAKPKPAAASG" are two of the examples from rice. The repeated occurrence of alanine (A) is not a coincidence. It is one of the most common residues in rice, much as serine (S) is common in *Arabidopsis*. The two corresponding codons differ by only one base change in GC content, from GCN (alanine) in rice to TCN (serine) in *Arabidopsis*. **Figure 6A** compares the LCS content in rice and *Arabidopsis*, where for the latter we selected from the *Arabidopsis* cDNAs some 6,605 "best homologs" (*i.e.*, highest similarity) for the rice cDNAs. LS genes from rice show a 10% increase in LCS content over most of their coding region. High LCS content is problematic for sequence alignment because the LCS regions have to be masked out. Improving the alignment algorithms might help, but this is not easy, and in any case it is not clear that similarity would be recovered if we included these regions.

Changes in amino acid patterns can also result in disordered proteins (*50*) defined by their inability to fold into compact structural domains under normal physiological conditions. DisEMBL (*51*) predicts sites of likely disorder in three categories: Loops/coils identify all residues that are not α helix, 3_{10}-helix, or β strand (a necessary but not a sufficient condition); Hot loops identify a refined subset of loops/coils with high mobility, as determined from C_α temperature factors (B factors); Remark465 is an entry from the Protein Data Bank (PDB) that refers to missing coordinates in X-ray structure. Although it is not entirely accurate to treat LCS as yet another category of disorder, they are correlated, and all of them can help to discriminate between HS and LS genes. **Figure 6B** shows the fraction of the HS and LS genes where 50% of the residues are "disordered". First, consider each category indi-

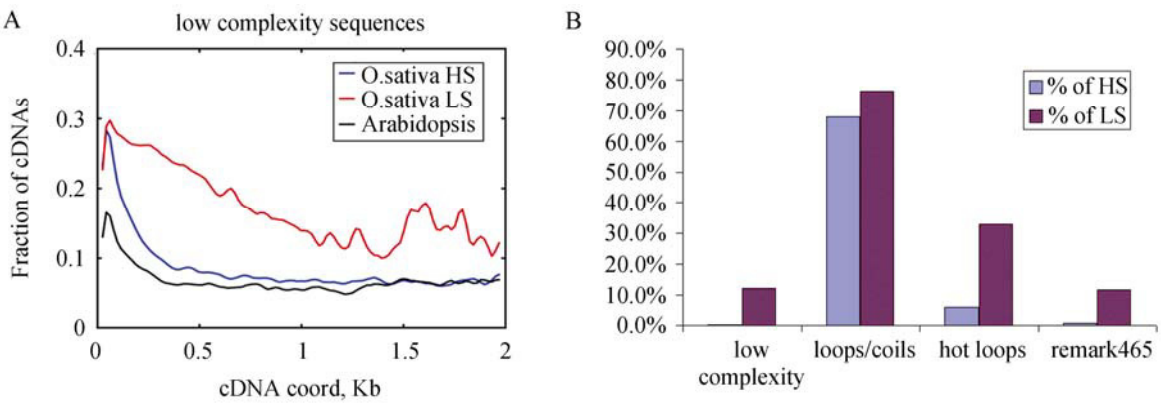

Figure 6 Protein disorder categories. **A**. LCS, as flagged by BlastP. At each position along the coding region, we determine how many genes are present, and compute their mean LCS content with a 51-bp sliding window. HS and LS genes are compared to 6,605 *Arabidopsis* cDNAs, which are called "best homologs" (*i.e.*, highest similarity) because they exhibit similarity to something in nr-KOME. **B**. The bar chart shows the number of rice cDNAs where over 50% of the protein is disordered. We plotted both the LCS category and those categories (loops/coils, hot loops, and remark465) that are predicted by the DisEMBL algorithm.

vidually, in the order of LCS, loops/coils, hot loops, and remark465. If we compute the ratio for the number of genes that are disordered, LS genes are 49.4, 1.1, 5.3, and 14.5 more disordered than HS genes, respectively. Combining LCS and remark465 as discussed in Materials and Methods, the ratio is 19.5. If we add hot loops, it becomes 6.1.

Indicators of biological functionality

The issue of functionality is intrinsically difficult to assess when the gene is in a state of evolutionary flux. Even if there is enough similarity to a previously characterized gene for us to infer a biological role, there is no assurance that the gene is fulfilling that role, because for example it could already be sufficiently down-regulated to be irrelevant for organism survival. It is worth noting that many gene knockouts create no "obvious" phenotype, even for highly expressed and well conserved genes. On the other hand, from an evolutionary perspective the whole point of having a process that includes a state of evolutionary flux is the possibility that some transients may eventually evolve biological roles. Hence, even when we cannot infer a role, we still want to know if that gene may ultimately regain a functional status. In lieu of an inordinate amount of experimentation, computational assessments will be provided here.

Using Bioverse, we inferred a biological role for 57.8% of HS genes and 17.1% of LS genes. These are listed in Table S3, organized according to Gramene (*52*) and Gene Ontology (GO) (*53*). HS genes are dominant in enzyme, carbohydrate metabolism, and protein metabolism. Given the essential nature of these functions to any cell type, one can say they are housekeeping-related genes. LS genes in contrast dominate in enzyme regulator, nucleic acid binding, cell communication, and development. Many of these genes are regulation and development related. Even if we restrict this analysis to the most reliable, highly expressed genes, such as those confirmed by proteomics, the exact same categories stand out. Previous analyses have likewise shown that different functional categories will stand out for genes that have evolved through different duplication processes (*54, 55*).

To determine if a specific cDNA is likely to have a biological role, irrespective of whether or not we know what that role is, we consider five indicators: (1) it is confirmed by EST or SAGE; (2) it is classified by GO; (3) it has an identifiable duplicate in rice; (4) it is conserved in maize or sorghum; (5) it encodes a mostly ordered protein. The last indicator runs somewhat counter to the trend in the literature for disordered proteins where much of the focus is on the few cases in which the gene is known to be functional. Here, LS gene disorder is more useful as an indicator of what is not functional. This is easy enough to understand. Most gene duplicates are eventually silenced by degenerative mutations, and disordered protein regions (*i.e.*, LCS and remark465, with an option for hot loops) are the likely reflections of this process.

Table S4 considers all pairwise combinations of these five indicators, and shows where we have a statistically significant correlation. We compute as follows. Suppose N_i and N_j are the numbers of genes that fulfill indicators i and j, while N_{ij} is the number of genes that fulfills both indicators. There are N genes in total. Let M_{ij} be the likelihood that indicator j is fulfilled, given that indicator i is fulfilled. The observed value is N_{ij}/N_i. In the absence of a correlation, the expected value is N_j/N. Fisher's exact test is used to find statistically significant increases in this observed frequency. In general, LS genes show the strongest effects. Conservation in maize or sorghum is highly correlated with mRNA confirmation by EST or SAGE, existence of GO classifications, and detection of gene duplicates in rice. Well-ordered proteins are correlated with detection of gene duplicates in rice, and conservation in maize or sorghum. As an aside, similar considerations can determine if LS genes are contaminated by retroposed pseudogenes. This is a concern as many are single exon genes. Severe contamination would be reflected in a strong anti-correlation between protein ordering and single exon status. Overall, LS genes have 33.9% single exons. Restricting to the subset of LS genes that is 2-way or 3-way ordered, one finds that 31.5% or 29.1% are single exons. The anti-correlation does exist, with P-values of 9.6×10^{-3} or 2.3×10^{-6}, respectively, but it is not the dominant effect.

We can also correlate these functional indicators with duplication category, as shown in **Table 2**. One duplication category stands out. For LS genes that are identifiably part of a segmental duplication, maize or

Table 2 Functionality correlations based on duplication category

HS genes								
	EST or SAGE data			Function classified	Duplicated in rice	Conserved in maize or sorghum	Protein ordered(2)	Protein ordered(3)
	No. of cDNA	confirmed	No. of tags					
Duplicated	8,796	85.1%	21.42 (7)	62.1%	100.0%	99.9%	99.5%	95.0%
segmental	1,002	90.3%	20.20 (8)	58.6%	100.0%	99.9%	98.9%	91.5%
tandem	512	80.1%	23.93 (6)	58.8%	100.0%	99.6%	99.2%	95.3%
background	908	90.3%	17.12 (7)	55.8%	100.0%	99.6%	98.8%	91.3%
Unique	2,850	89.9%	16.25 (7)	43.3%	0.0%	98.9%	97.6%	88.7%
Average				**57.5%**		**99.7%**	**99.0%**	**93.5%**
LS genes								
	EST or SAGE data			Function classified	Duplicated in rice	Conserved in maize or sorghum	Protein ordered(2)	Protein ordered(3)
	No. of cDNA	confirmed	No. of tags					
Duplicated	1,999	65.1%	12.88 (3)	16.4%	100.0%	61.0%	92.6%	74.6%
segmental	170	79.4%	**29.81 (5)**	24.1%	100.0%	**87.1%**	90.0%	69.4%
tandem	324	63.0%	10.86 (3)	17.0%	100.0%	51.2%	96.0%	79.6%
background	443	66.8%	10.91 (3)	14.9%	100.0%	44.2%	90.5%	67.3%
Unique	3,616	70.0%	18.01 (5)	17.5%	0.0%	44.1%	74.9%	53.0%
Average				**17.1%**		**50.1%**	**81.2%**	**60.7%**

Note: "Duplicated" refers to cDNAs with any number of detectible duplicates in rice, without regarding for classification. "Unique" refers to cDNAs with no detectible homologs. All must be 95% alignable to the rice genome. "Segmental", "tandem", and "background" refer to the maximal set of homolog pairs used in Figure 2. Mean (median) number of EST or SAGE tags is averaged over those cDNAs with at least one confirming tag. We highlight the categories that are discussed in Results.

sorghum conservation is observed 1.84 times more often than for the other two categories of duplication. Similarly, mRNA expression level based on the number of EST or SAGE tags is 2.74 times larger than normal. Most of this effect is due to the whole genome duplication. The figures do not much change even if we remove the recent segmental duplication connecting chromosome 11 to 12. They become 1.90 and 2.78, respectively. We believe this effect is mostly related to the time since duplication. Because the WGD occurred over 55 to 70 Mya, any genes that survived would likely have developed important functions.

Functional LS genes seen in UniProt

Given the uncertainty over functionality when the gene is in evolutionary flux, it is instructive to look for a few cases where the LS gene is undoubtedly functional. For our gold standard of functionality, we use the existence of a well-characterized protein in the UniProt database (*56*) (time stamp 2005-10-25). We developed a series of rules to identify cases consistent with the transient hypothesis. First, we discard mitochondrial, chloroplast, ribosomal, and fragmentary proteins. This left us with 383 functional rice proteins, of which 45 are LS genes. A search of the rice genome revealed that 22 of them have identifiable duplicates. We focus on a subset of 12 proteins (7 of which are in nr-KOME) with a complete alignment for both the gene and its most recent duplicate. Most of these genes are highly duplicated, with many putative copies, so we required the coding region size and the number of exons to be consistent. In most cases, it was easy to identify the most recent duplicate. In four recalcitrant cases, RA05-RA14-RA17-RAG2, we just used the smallest of many similar Ks in Table S5.

What stands out is that Ks is small. The mean (median) is 0.174 (0.158), with 11 of 12 duplicate pairs under 0.211, and the outlier SALT at 0.582. Ka/Ks is expected to be large, which it is at 0.561 (0.529). All are tandem or background duplications; none are segmental duplications. All but one align to the genome in 1 or 2 exons, as compared to a mean (median) of 4.4 (3) exons for all of nr-KOME. The outlier is AMC1 with 5 exons. Given the many duplicates per gene, it is unlikely that loss of similarity with

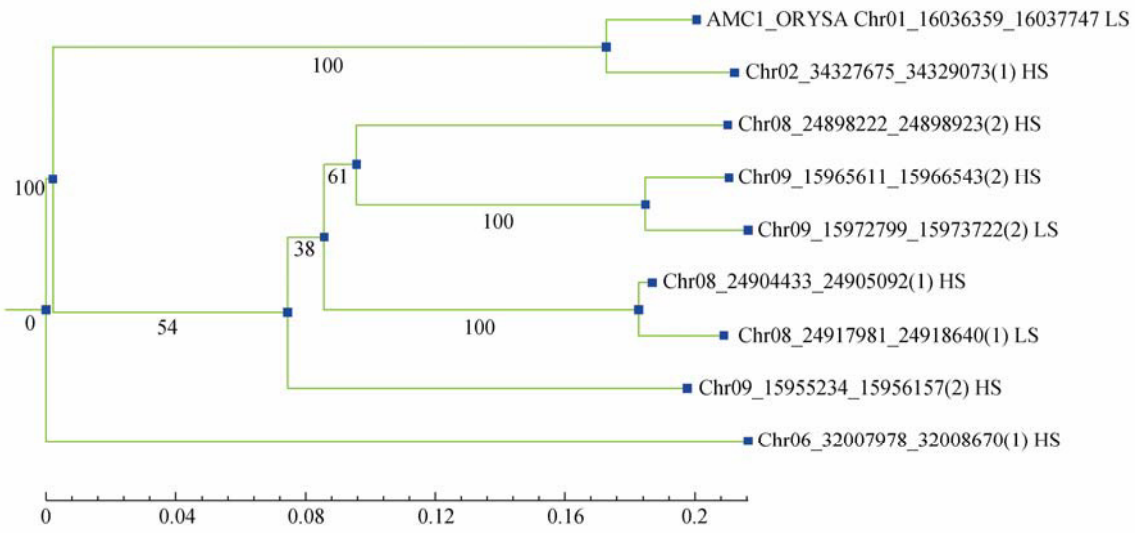

Figure 7 Duplication history of AMC1 (α-amylase isozyme C). The scale bar represents divergence in substitutions per site. Bootstrap values are shown on the branches. 100 is best. Gene names indicate chromosome position and HS/LS status.

Arabidopsis occurred on the most recent duplication. Hence we did a recursive search for duplicates of duplicates, to trace back to the HS gene. In the process, our 12 examples collapsed to 7 groupings: AMC1, DH16B- DH16C, GOS9-SALT, IBBR, PRO1, PRO2, and RA05-RA14-RA17-RAG2, which contain 9, 4, 51, 11, 4, 17, and 11 genes, respectively. We succeeded in identifying the HS gene in 2 of 7 cases, AMC1 and GOS9-SALT. **Figure 7** depicts the AMC1 analysis. The LS gene on chromosome 1 has both a protein (AMC1) and a cDNA (*AK063489*) for support. Its most recent duplicate on chromosome 2 is a HS gene with both a protein (AMY1) and a cDNA (*AK101744*). However, there is a change in the gene structure. Exon 3 in AMY1 becomes two exons and an intron in AMC1. Essentially, a 12-bp microexon breaks off the 3'-end, and a small part of the leftover exon is converted to an intron with non-canonical splice sites TC-GC.

Table S6 shows the InterPro, GO (Gramene), and UniProt descriptions for these 12 proteins. We would advise caution about over-generalizing from a few examples, but at the same time one has to at least look. AMC1 (α-amylase isozyme C) is a member of the glycosyl hydrolase 13 family and is important for the breakdown of endosperm starch during germination (*57, 58*). The rest are broadly classified under defense and stress response. One group of seven proteins (IBBR, PRO1, PRO2, RA05-RA14-RA17-RAG2) has serine-type endopeptidase inhibitor activity, an important defense mechanism against phytophagous insects and microorganisms (*59, 60*). Stress response refers to the two dehydrins (DH16B-DH16C) (*61*) and two jacalin-like lectins (GOS9-SALT) (*62*). LS genes might not be as enriched for housekeeping functions as are HS genes, but they do appear enriched for functions that while less essential are nonetheless important for how an organism adapts to different ecological niches.

Discussion

The idea that evolutionary transients can exist, while perhaps not extensively considered before, is conceptually not surprising. What is surprising is the magnitude of the effect. Although we do not claim that every LS gene is attributable to evolutionary transients, it is important to raise awareness of this phenomenon, as it may be relevant in other contexts. For example, the annotations for most sequenced genomes yield many orphan (*63*) or ORFan (*64*) proteins. Their numbers may be reduced by the constant improvements in the annotation pipeline, but they are not completely eliminated (*65*). Others have reported LS-like characteristics in the POFs (proteins with obscure features) that lack identifiable structural motifs or domains (*66*). In comparison to PDFs (proteins with defined features), POFs are more divergent (*i.e.*,

low similarity) and disordered. POFs can have mutant phenotypes, as reported in *Saccharomyces cerevisiae* studies, but they are notably under-represented in the essential gene category.

It is especially important that this phenomenon affects not only the genome (*i.e.*, gene prediction) but also the transcriptome (*i.e.*, cDNA sequence), as there have been hints of this possibility for some time in the large-scale expression studies for mouse and human. Regardless of the method used, full-length cDNAs (*67*) and tiling arrays (*68*) consistently observe more transcription than annotated by Ensembl. This discrepancy has been called "dark matter", in honor of the mystery in astrophysics, where most of the matter in the universe is of an unknown form. A comprehensive survey of the expressed transcripts from 1% of the human genome confirms this basic observation (*69*). We do not expect evolutionary transients to explain all of the dark matter. Non-coding RNA genes may indeed be as pervasive as some have speculated (*70*). Nonetheless, it was the prospect of evolutionary transients that prompted us to not use Ensembl in our initial annotation of the rice genome. At that time, the only other sequenced plant genome was *Arabidopsis*. Without a closely related plant genome to compare against, imposing a conservation rule would have compromised gene set completeness.

More generally, the LS phenomenon should be applicable beyond plants, and certainly beyond duplication of individual genes. Recent segmental duplications are 2.7% of the difference between humans and chimpanzees (*71*). Exon 2 of the *morpheus* gene on the short arm of human chromosome 16 was found to have a Ka/Ks ratio of 13.0 in a comparison of humans and Old World monkeys (*72*). Even more generally, the LS phenomenon is not just a consequence of duplications *per se*. It is a consequence of how evolution solves a fundamental problem. Namely, how does an organism maintain the integrity of its functional sequences even as it "experiments" with other sequences? Gene duplication is one solution, but so is alternative splicing. In our previous study (*5*), we claimed that plants favor gene duplication, while animals favor alternative splicing. Additional evidence has surfaced since that time to further support this idea. We review those now.

Distributions of gene family size show a striking contrast. In humans, 86.4% of the genes exist as singletons, while 1.4% exist in families with more than 5 genes. In rice and *Arabidopsis*, the average is 44.0% for singletons and 32.9% for families with more than 5 genes (*73*). Meanwhile, our perception for the frequency of alternative splicing in humans continues to grow with the accumulation of data. A little over a decade ago, the data indicated that alternative splicing occurred in fewer than 10% of human genes; but now it is believed to be at least 60%, and perhaps as much as 99% (*74*). The latest data for rice and *Arabidopsis* place this rate at 22% (*75*). But there is a crucial difference. Exon skipping is favored in humans, and intron retention is favored in plants. Artifactual instances of intron retention can arise in unprocessed pre-mRNAs; thus the corrected rate for plants is certain to be under 22%. Human analyses have also shown that there is an inverse correlation between the size of a gene family and its use of alternatively spliced isoforms (*76*). We therefore predict that many of the LS gene characteristics that we saw in rice will eventually be observed in alternatively spliced exons in humans.

Materials and Methods

The initial data from KOME had 28,444 *japonica* cDNAs with complete open reading frames (ORFs) (*9*). We aligned these to the Syngenta *japonica* genome. When two aligned regions overlapped by over 100 bp, the smaller cDNA was removed. Most of these redundancies were just minor differences in mRNA initiation and termination, as opposed to alternative splicing of internal protein-coding exons. A few cDNAs failed to align even partially to any of the available rice genomes (Beijing *indica*, Syngenta *japonica*, and IRGSP *japonica*) with a combined 22× coverage. We assumed that these were non-rice contaminants and removed them. This produced a set of 19,079 non-redundant cDNAs that we called nr-KOME.

Because the duplication history was taken from a previous analysis (*2*), we refer the reader to that publication for further details. Here we only discuss an issue that might otherwise be very confusing. In searching for duplicates for any given cDNA, one

either finds no homologs at all, or a mean (median) of 40 (5) homologs. To simplify the analysis, we do it in two phases. First we consider those cDNAs with one-and-only-one homolog in rice. This helps us identify trend lines indicative of segmental and tandem duplications. All other duplications are deemed to be background. At this point the number of homolog pairs for the three duplication categories is 609, 311, and 1,351, respectively. We then add back those higher-order homologs (*i.e.*, those cDNAs with more than one putative duplicate) that we had deferred. The rules are different for each duplication category because we wish to maximize the number of homolog pairs. The end result is 1,340, 1,685, and 1,351 homolog pairs, respectively.

Comparisons to the gene-enriched sequences of maize and sorghum require an assembled version of this fragmentary data. For maize (*40, 41*), this assembly was provided by the authors; however for sorghum (*12*), no assembly was provided, so we created our own using the algorithm RePS (*77, 78*). We considered three criteria for "best hit": the size LP of the aligned sequence, the SCORE parameter from TBlastN, and XP/LP where XP is the number of identically matched amino acids. We chose the criterion that gives the most consistent result for maize, knowing that since Ks is the time since divergence, HS and LS genes should have similar Ks. For LP, the mean Ks for HS and LS genes is 0.871 and 0.679, respectively; for SCORE, we get 0.746 and 0.648; and for XP/LP, we get 0.631 and 0.646. XP/LP is best. The mean for all genes is 0.634, smaller than the 0.688 that we got for the WGD, and consistent with the supposition that the WGD predated the divergence of the grasses. Note that the duplication analyses and grass comparisons both use GeneWise (*79*) to refine the exon-intron boundaries, and use K-estimator (*38*) to compute Ka/Ks.

We collected 200,648 ESTs by combining what we found in GenBank with our own data (*80*). Given the preponderance of low quality sequences, we filter the data aggressively. Only sequences that are 95% aligned to the rice genome are used, giving a reduced set of 104,903 ESTs, 65.0% of which have a match to nr-KOME on the criterion that 80% of the EST must align to the cDNA at 95% identity. The SAGE dataset was described in a previous publication (*81*). It was processed by SAGE300 (*82*), giving a set of 431,853 10-mer tags, of which 48.9% have an exact match to nr-KOME. The proteomics mass spectrometry (*83*) involved two series of experiments, one preceded by 2D-gel electrophoresis (2D-MS), and the other preceded by liquid chromatography (LC-MS). Altogether we confirmed 3,276 nr-KOME cDNAs, with 9.2% confirmed by 2D-MS (MASCOT $P<0.05$) (*84*) and 95.5% by LC-MS (MudPIT score>2,400) (*85*).

Low complexity sequences were identified by SEG, which is also the algorithm used by BlastP (*49*). Structural disorder (loops/coils, hot loops, and remark465) was identified by DisEMBL (*51*). Different categories of "disorder" can be combined, with LCS+remark465 and LCS+remark465+hotloops as the most useful. There are two ways to do the combining. First, we can label an amino acid as disordered if it belongs to any of the target categories and then require that 50% of the amino acids be disordered. Alternatively, we can require that this 50% rule be satisfied entirely by one target category at a time. Our results showed that the second method had better discrimination power. In particular, using the second method, the LS/HS ratio is 19.5 and 6.1 when we combine two and three disorder categories, respectively. But with the first method, these ratios are 13.6 and 3.8.

For the UniProt analysis, the duplications were mostly based on alignments to Beijing *indica*, but we also did alignments to Syngenta *japonica* and IRGSP *japonica*, to ensure that the map positions for the gene and its most recent duplicate were consistent. We did find one discrepancy. For GOS9, Beijing *indica* had two tandem duplications, on chromosomes 1 and 6; both *japonica* assemblies had only chromosome 6. This is likely to be a polymorphism. Absolutely no frame shifts were seen, but some non-synonymous changes were. GOS9 was an outlier, with 13 changes; ignoring that, the mean (median) number of changes is 1.1 (0). The phylogenetic tree was constructed by the neighbor-joining method with amino acid substitution distance (*86*).

Acknowledgements

This work was jointly supported by Chinese Academy

of Sciences (Grants No. GJHZ0701-6 and KSCX2-YWN-023), National Natural Science Foundation of China (Grants No. 30725008, 90403130, 90608010, 30221004, 90612019, and 30392130), the "973" Program (Grants No. 2006CB910400, 2007CB815701, 2007CB815703, and 2007CB815705), the "863" Program (Grants No. 2006AA02Z334, 2006AA10A121, and 2006AA02Z177), Beijing Municipal Science and Technology Commission (Grant No. D07030200740000), Danish Platform for Integrative Biology, Danish Natural Science Research Council, Danish Research Council, the Solexa Project (Grant No. 272-07-0196), National Science Foundation of USA (Grant No. DBI 0217241), Searle Scholar's Program, and Alberta Informatics Circle of Research Excellence. We thank Shoshi Kikuchi and Piero Carninci for cDNA sequences, Peter Roepstorff and Xumin Zhang for proteomics advice, and BGI's SAGE/EST/proteomics groups for additional data. Parts of the final manuscript were edited by Laurie Goodman.

Authors' contributions

JZ, RL, HZ, JL, YZ, HL, PN, Songgang L, Shengting L, Jingqiang W, DL, JM, RS, SL, Jian W, HY collected the datasets and conducted the analyses. Jun W, JY, and GKSW supervised the project and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1 Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151-1155.
2 Yu, J., *et al.* 2005. The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* 3: e38.
3 International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* 436: 793-800.
4 The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796-815.
5 Yu, J., *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79-92.
6 Bennetzen, J.L., *et al.* 2004. Consistent over-estimation of gene number in complex plant genomes. *Curr. Opin. Plant Biol.* 7: 732-736.
7 Ma, L., *et al.* 2005. A microarray analysis of the rice transcriptome and its comparison to *Arabidopsis*. *Genome Res.* 15: 1274-1283.
8 Li, L., *et al.* 2006. Genome-wide transcription analyses in rice using tiling microarrays. *Nat. Genet.* 38: 124-129.
9 Rice Full-Length cDNA Consortium. 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science* 301: 376-379.
10 Brendel, V., *et al.* 2002. Comparative genomics of *Arabidopsis* and maize: prospects and limitations. *Genome Biol.* 3: Reviews1005.
11 Vincentz, M., *et al.* 2004. Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. *Plant Physiol.* 134: 951-959.
12 Bedell, J.A., *et al.* 2005. Sorghum genome sequencing by methylation filtration. *PLoS Biol.* 3: e13.
13 Vandepoele, K. and Van de Peer, Y. 2005. Exploring the plant transcriptome through phylogenetic profiling. *Plant Physiol.* 137: 31-42.
14 Zhu, W. and Buell, C.R. 2007. Improvement of whole-genome annotation of cereals through comparative analyses. *Genome Res.* 17: 299-310.
15 Ohno, S. 1970. *Evolution by Gene Duplication.* Springer, Berlin, Germany.
16 Kellis, M., *et al.* 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428: 617-624.
17 Kashkush, K., *et al.* 2002. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* 160: 1651-1659.
18 Feldman, M. and Levy, A.A. 2005. Allopolyploidy—a shaping force in the evolution of wheat genomes. *Cytogenet. Genome Res.* 109: 250-258.
19 Comai, L., *et al.* 2000. Phenotypic instability and rapid gene silencing in newly formed *Arabidopsis* allotetraploids. *Plant Cell* 12: 1551-1568.
20 Wang, J., *et al.* 2004. Stochastic and epigenetic changes of gene expression in *Arabidopsis* polyploids. *Genetics* 167: 1961-1973.
21 Conant, G.C. and Wagner, A. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Res.* 13: 2052-2058.
22 Zhang, P., *et al.* 2003. Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol.* 4: R56.
23 Jordan, I.K., *et al.* 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol.*

Biol. 4: 22.
24 Morin, R.D., *et al*. 2006. Sequencing and analysis of 10,967 full-length cDNA clones from *Xenopus laevis* and *Xenopus tropicalis* reveals post-tetraploidization transcriptome remodeling. *Genome Res.* 16: 796-803.
25 Lynch, M. and Conery, J.S. 2003. The evolutionary demography of duplicate genes. *J. Struct. Funct. Genomics* 3: 35-44.
26 Blanc, G. and Wolfe, K.H. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16: 1667-1678.
27 Altschul, S.F., *et al*. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
28 McDermott, J. and Samudrala, R. 2003. Bioverse: functional, structural and contextual annotation of proteins and proteomes. *Nucleic Acids Res.* 31: 3736-3737.
29 McDermott, J., *et al*. 2005. Functional annotation from predicted protein interaction networks. *Bioinformatics* 21: 3217-3226.
30 Domazet-Loso, T. and Tautz, D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res.* 13: 2213-2219.
31 Li, R., *et al*. 2005. ReAS: recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput. Biol.* 1: e43.
32 *Drosophila* 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450: 203-218.
33 Jin, Y.K. and Bennetzen, J.L. 1994. Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. *Plant Cell* 6: 1177-1186.
34 McCarthy, E.M. and McDonald, J.F. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* 19: 362-367.
35 Yuan, Q., *et al*. 2005. The institute for genomic research Osa1 rice genome annotation database. *Plant Physiol.* 138: 18-26.
36 Kellogg, E.A. 2001. Evolutionary history of the grasses. *Plant Physiol.* 125: 1198-1205.
37 Yang, Y.W., *et al*. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. *J. Mol. Evol.* 48: 597-604.
38 Comeron, J.M. 1999. K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. *Bioinformatics* 15: 763-764.
39 Gaut, B.S., *et al*. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc. Natl. Acad. Sci. USA* 93: 10274-10279.
40 Palmer, L.E., *et al*. 2003. Maize genome sequencing by methylation filtration. *Science* 302: 2115-2117.
41 Whitelaw, C.A., *et al*. 2003. Enrichment of gene-coding sequences in maize by genome filtration. *Science* 302: 2118-2120.
42 Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437: 69-87.
43 Aebersold, R. and Mann, M. 2003. Mass spectrometry-based proteomics. *Nature* 422: 198-207.
44 Desiere, F., *et al*. 2004. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* 6: R9.
45 Jabbari, K., *et al*. 2004. The new genes of rice: a closer look. *Trends Plant Sci.* 9: 281-285.
46 Adams, K.L., *et al*. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. USA* 100: 4649-4654.
47 Wong, G.K., *et al*. 2002. Compositional gradients in Gramineae genes. *Genome Res.* 12: 851-856.
48 Wang, H.C., *et al*. 2004. Mutational bias affects protein evolution in flowering plants. *Mol. Biol. Evol.* 21: 90-96.
49 Wootton, J.C. and Federhen, S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* 266: 554-571.
50 Dyson, H.J. and Wright, P.E. 2005. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6: 197-208.
51 Linding, R., *et al*. 2003. Protein disorder prediction: implications for structural proteomics. *Structure* 11: 1453-1459.
52 Ware, D.H., *et al*. 2002. Gramene, a tool for grass genomics. *Plant Physiol.* 130: 1606-1613.
53 Harris, M.A., *et al*. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32: D258-261.
54 Maere, S., *et al*. 2005. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA* 102: 5454-5459.
55 Rizzon, C., *et al*. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Comput. Biol.* 2: e115.
56 Bairoch, A., *et al*. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 33: D154-159.
57 Horvathova, V., *et al*. 2001. Amylolytic enzymes: molecular aspects of their properties. *Gen. Physiol. Biophys.* 20: 7-32.
58 Rolland, F., *et al*. 2006. Sugar sensing and signaling in plants: conserved and novel mechanisms. *Annu. Rev. Plant Biol.* 57: 675-709.
59 Ryan, C.A. 1990. Protease inhibitors in plants: genes for improving defenses against insects and pathogens. *Annu. Rev. Phytopathol.* 28: 425-449.

60 Shewry, P.R. and Lucas, J.A. 1997. Plant proteins that confer resistance to pests and pathogens. *Adv. Bot. Res.* 26: 135-192.
61 Vinocur, B. and Altman, A. 2005. Recent advances in engineering plant tolerance to abiotic stress: achievements and limitations. *Curr. Opin. Biotechnol.* 16: 123-132.
62 Raval, S., *et al*. 2004. A database analysis of jacalin-like lectins: sequence-structure-function relationships. *Glycobiology* 14: 1247-1263.
63 Dujon, B. 1996. The yeast genome project: what did we learn? *Trends Genet.* 12: 263-270.
64 Siew, N. and Fischer, D. 2003. Twenty thousand ORFan microbial protein families for the biologist? *Structure* 11: 7-9.
65 Kunin, V., *et al*. 2003. Myriads of protein families, and still counting. *Genome Biol.* 4: 401.
66 Gollery, M., *et al*. 2006. What makes species unique? The contribution of proteins with obscure features. *Genome Biol.* 7: R57.
67 Hayashizaki, Y. and Carninci, P. 2006. Genome Network and FANTOM3: assessing the complexity of the transcriptome. *PLoS Genet.* 2: e63.
68 Johnson, J.M., *et al*. 2005. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* 21: 93-102.
69 ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799-816.
70 Mattick, J.S. 2004. RNA regulation: a new genetics? *Nat. Rev. Genet.* 5: 316-323.
71 Cheng, Z., *et al*. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437: 88-93.
72 Johnson, M.E., *et al*. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* 413: 514-519.
73 Lockton, S. and Gaut, B.S. 2005. Plant conserved non-coding sequences and paralogue evolution. *Trends Genet.* 21: 60-65.
74 Boue, S., *et al*. 2003. Alternative splicing and evolution. *Bioessays* 25: 1031-1034.
75 Wang, B.B. and Brendel, V. 2006. Genomewide comparative analysis of alternative splicing in plants. *Proc. Natl. Acad. Sci. USA* 103: 7175-7180.
76 Kopelman, N.M., *et al*. 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat. Genet.* 37: 588-589.
77 Wang, J., *et al*. 2002. RePS: a sequence assembler that masks exact repeats identified from the shotgun data. *Genome Res.* 12: 824-831.
78 Zhong, L., *et al*. 2003. A statistical approach designed for finding mathematically defined repeats in shotgun data and determining the length distribution of clone-inserts. *Genomics Proteomics Bioinformatics* 1: 43-51.
79 Birney, E., *et al*. 2004. GeneWise and Genomewise. *Genome Res.* 14: 988-995.
80 Zhou, Y., *et al*. 2003. Gene identification and expression analysis of 86,136 Expressed Sequence Tags (EST) from the rice genome. *Genomics Proteomics Bioinformatics* 1: 26-42.
81 Bao, J., *et al*. 2005. Serial analysis of gene expression study of a hybrid rice strain (LYP9) and its parental cultivars. *Plant Physiol.* 138: 1216-1231.
82 Lash, A.E., *et al*. 2000. SAGEmap: a public gene expression resource. *Genome Res.* 10: 1051-1060.
83 Zhao, C., *et al*. 2005. Proteomic changes in rice leaves during development of field-grown rice plants. *Proteomics* 5: 961-972.
84 Perkins, D.N., *et al*. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20: 3551-3567.
85 Washburn, M.P., *et al*. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* 19: 242-247.
86 Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406-425.

Supplementary Material

Figure S1; Tables S1-S6

DOI: 10.1016/S1672-0229(10)60023-X