Population Level Purifying Selection and Gene Expression Shape Subgenome Evolution in Maize

Saurabh D. Pophaly*1 and Aurélien Tellier1

1Section of Population Genetics, School of Life Sciences Weihenstephan, Technische Universität München, Freising, Germany
*Corresponding author: E-mail: pophaly@wzw.tum.de.
Associate editor: Brandon Gaut

Abstract

The maize ancestor experienced a recent whole-genome duplication (WGD) followed by gene erosion which generated two subgenomes, the dominant subgenome (maize1) experiencing fewer deletions than maize2. We take advantage of available extensive polymorphism and gene expression data in maize to study purifying selection and gene expression divergence between WGD retained paralog pairs. We first report a strong correlation in nucleotide diversity between duplicate pairs, except for upstream regions. We then show that maize1 genes are under stronger purifying selection than maize2. WGD retained genes have higher gene dosage and biased Gene Ontologies consistent with previous studies. The relative gene expression of paralogs across tissues demonstrates that 98% of duplicate pairs have either subfunctionalized in a tissuewise manner or have diverged consistently in their expression thereby preventing functional complementation. Tissuewise subfunctionalization seems to be a hallmark of transcription factors, whereas consistent repression occurs for macromolecular complexes. We show that dominant gene expression is a strong determinant of the strength of purifying selection, explaining the inferred stronger negative selection on maize1 genes. We propose a novel expression-based classification of duplicates which is more robust to explain observed polymorphism patterns than the subgenome location. Finally, upstream regions of repressed genes exhibit an enrichment in transposable elements which indicates a possible mechanism for expression divergence.

Key words: whole-genome duplication, purifying selection, DoFE, gene expression.

Introduction

Whole-genome duplication (WGD) is better tolerated in plants compared with animals as evident by multiple WGDs sprinkled across the plant phylogeny (Lee et al. 2012). The family of grasses shares a round of WGD in the pragress ancestor (Paterson et al. 2004). This event was followed by another WGD in the ancestor to the genus Zea (including maize; Zea mays L. ssp. mays) (Gaut and Doebley 1997), which occurred around 5–12 Ma (Swigov et al. 2004). Most analyzed WGDs are very ancient and the evolution of resulting duplicate pairs (ohnologs) has been studied based on divergence (Ka and Ks ratios). Maize WGD forms a special case as it is relatively recent and has occurred after the divergence from the nearest independently sequenced relative (Sorghum). A whole-genome single nucleotide polymorphism (SNP) data set is available for maize in the form of a large population resequencing effort (the maize hapmap2 project) (Chia et al. 2012). Moreover, maize has extensive data for gene expression in various tissues. The evolution of ohnologous pairs can thus be studied at a recent timescale by using differences in patterns of polymorphism and gene expression.

A WGD not only duplicates every gene but also preserves the intergenic and regulatory interactions. Paralogous pairs (ohnologs) from WGD are thus “born equal,” but then they diverge in sequence, expression, and sometimes function.

After WGD a process of diploidization ensues, which is typically accompanied by rapid and massive gene loss, a process called fractionation (Langham et al. 2004). In maize, only 3,228 gene genes have retained their duplicate copies since the maize-specific WGD (Schnable et al. 2011). Fractionation is shown to be biased in many studies which means that some genomic regions of the diploidized polyploid shed more genes (sensitive subgenome) compared with other (dominant subgenome) (Thomas et al. 2006; Woodhouse et al. 2010; Murat et al. 2013). Schnable et al. (2011) individually aligned the duplicated gene copies in maize to their ortholog from sorghum and then created syntenic blocks of genes presuming that the ancestral (pre-WGD) gene arrangement in maize would be similar to that in sorghum. The blocks with higher number of deleted genes were called maize2 (the counterpart called maize1). Although seemingly artificial, this grouping has biological consequences. A maize1 gene is more likely to dominate the maize2 ohnolog in expression (Schnable et al. 2011). Schnable and Freeling (2011) also showed that genes which display visible mutant phenotype are more likely to be located in maize1 subgenome. They thus hypothesized that deletion or mutation in maize1 copy would have a higher effect on fitness. Importantly, there is no difference between inherent deletion rates between subgenomes, thus biased number of gene deletions can only be explained by differential consequences of gene deletion between subgenomes (Schnable et al. 2011).

Retention after WGD is generally seen in the light of two hypotheses. First, the gene dosage balance hypothesis (Birchler and Veitia 2012) which predicts that selection acts on maintaining the stoichiometric ratios between interacting
gene partners. Thus regulatory genes and genes involved in multiprotein complexes which typically have many interactions are more likely to be retained after WGD (Bekaert et al. 2011; Rodgers-Melnick et al. 2012). Few studies have also indicated absolute gene dosage to be an important determinant of retention after WGD (Aury et al. 2006; Chain et al. 2011; Renny-Byfield et al. 2015). Second, the subfunctionalization/neofunctionalization hypothesis which states that the fate of a retained duplicate gene pair broadly follows two known outcomes, subfunctionalization, where the ancestral function is partitioned between duplicate copies (Force et al. 1999), and neofunctionalization where one of the duplicate copy evolves a new function. Both outcomes can be achieved at the level of gene expression or protein function (Flagel and Wendel 2009). Expression-based subfunctionalization can be readily assayed by analyzing relative expression of duplicates. Duplicates can be expressed differentially across tissues, developmental stages, or environmental conditions or one copy can attain a novel expression profile (Duarte et al. 2006; Liu and Adams 2007; Tiros and Barkai 2007; Chaudhary et al. 2009). A study in Arabidopsis found that 85% of duplicate genes show evidence of regulatory subfunctionalization and/or neofunctionalization (Duarte et al. 2006). Another study (Hughes et al. 2014) quantified expression profile divergence of duplicates between two different leaf types of maize and analyzed long-term selection between ohnologs. It was shown that the purifying selection is weaker on one copy when the expression profile of duplicated pairs remains nondivergent across tissues, whereas, when the expression profile diverges in tissues, both copies are under strong purifying selection.

While analyzing divergent relative expression of duplicate pairs across many tissues, two possible patterns exist. The first scenario is a unidirectional expression divergence (UED) defined as one member of the pair being consistently highly expressed (UED-dominant) than its counterpart (UED-repressed). Such a consistent decrease in expression of one member was called as “regulatory hypofunctionalization” (Duarte et al. 2006). The second scenario is bidirectional expression divergence (BED) for which both genes are alternatively dominant and repressed in different tissues.

Another aspect of duplication lies in achieving mutational robustness via functional complementation by the duplicate copy. A study in yeast (Gu et al. 2003) showed that complementation occurs but decreases with sequence divergence between paralogs and the copy with higher dosage exhibits stronger deleterious effect when silenced. Schnable and Freeling (2011) reported cases where a mutant phenotype is catalyzed for a maize gene which has a duplicated copy present, thereby in these mutants the duplicate copy fails to complement.

The mechanisms operating behind the initial expression differences and divergence between ohnologs are not clearly understood. Expression difference between duplicates is known to be quickly established after the formation of synthetic and natural allopolyploids (Adams 2007; Wang et al. 2006). Epigenetic effects like DNA methylation have been tested as possible mechanisms but have not been proven. Parkin et al. (2014) found subgenome dominance for expression in Brassica oleracea but methylation profiles did not correlate with dominance for individual genes. No differences were found in gene body methylation between maize subgenomes (Eichten et al. 2011). Initial differences in upstream transposable elements (TEs) caused by allotetraploidy has been proposed (Schnable et al. 2011; Woodhouse et al. 2014). Repression of upstream TEs has been shown to cause an inadvertent decrease in expression of the nearby genes (Hollister and Gaut 2009; Hollister et al. 2011).

In this study, we first report similarities between duplicate pairs based on genome scale correlations for diversity, and then look at differences in purifying selection and gene ontology (GO) enrichments between sets of duplicate genes when classified by their subgenome location and relative expression (UED vs. BED). We use polymorphism data from 14 teosinte lines (wild cousins of maize; Matsuoka et al. 2002) from the maize hapmap2 project and gene expression data for various tissues from maize line B73. Finally, we test the proposed mechanism of expression divergence by assessing the enrichment of upstream TEs for our expression classified gene sets.

Results

Nucleotide Diversity between Duplicates Is Correlated

Nucleotide diversity (estimated as per-site average pairwise difference Π) was found to be correlated between duplicate gene pairs (particularly strong for introns) (table 1 and fig. 1). In addition, both nonsynonymous (Ka) and synonymous divergence (Ks) calculated with respect to sorghum orthologs were also correlated (Spearman’s rho 0.71 $P < 2.2 \times 10^{-16}$ and 0.73 $P < 2.2 \times 10^{-16}$, respectively). Interestingly, the diversity in upstream regions (2 kb) was not correlated indicating a decoupled upstream evolution of the duplicates. Hapmap2 project also provides per-site nucleotide diversity of genes calculated in a very stringent way (see Methods). These values were also correlated between duplicates (Spearman’s rho 0.28, $P = 1.1 \times 10^{-16}$).

Maize1 Subgenome Genes Are Under Stronger Purifying Selection

Different measures of purifying selection indicated a consistent higher purifying selection on the maize1 subgenome genes. The median Ka for maize1 retained copy was 0.039 versus 0.043 for maize2 retained copy ($P = 9 \times 10^{-6}$, Wilcoxon rank-sum test). The median Ks was 0.195 versus 0.193, respectively ($P = 0.3$). We then looked for occurrence of very deleterious mutations (VDMs) between subgenomes. Our definition of VDMs include frameshift, stop gained, and stop lost. Of 3,228 retained gene pairs, 1,003 in maize1 had a VDM in at least one sampled accession versus 1,140 for maize2 subgenome ($P < 0.0003$, chi-square test). The per-site nucleotide diversity for nonsynonymous sites was found to be lower for maize1 genes compared with maize2 genes (0.0017 vs. 0.0020, $P = 1 \times 10^{-6}$, Wilcoxon rank-sum test), but the difference in synonymous diversity was not found to be significant (0.0088 vs. 0.0087, $P = 0.8$, Wilcoxon rank-sum test), indicating no inherent difference in mutation rate between
subgenomes. We further used SIFT (Sorting Intolerant From Tolerant) scores (Ng and Henikoff 2001) which quantify the deleterious nature of an SNP based on the conservation at the site. We considered SIFT score < 0.01 to be deleterious in nature. A gene with more than one SNP with score < 0.01 was considered to be harboring deleterious mutations. Significantly more genes harboring deleterious mutations were found to be located in the maize2 subgenome (supplementary table S1, Supplementary Material online). A complete cross comparison of GOs between these two data sets is available in supplementary table S4, Supplementary Material online. Comparing the dosage of dominantly expressed UED and BED genes in each tissue revealed that UED-dominant genes have higher dosage than the dominantly expressed BED genes in all except one tissue (supplementary fig. S8, Supplementary Material online). Thus, the overall dosage of UED-dominant genes is higher than BED genes.

Retained Genes in Maize Show Expected Patterns of Expression and GO Enrichments

We looked for differences in the amount of gene expression (dosage) between single genes versus 3,228 gene pairs retained after WGD across 22 tissues (the list of tissues is available in supplementary table S2, Supplementary Material online). Retained genes showed consistently higher median expression in all tissues tested (supplementary fig. S2, Supplementary Material online). GO terms pertaining to regulation, transcription factors, and macromolecular complexes were enriched in retained genes (supplementary figs. 3 and 4, Supplementary Material online). For single genes only catalytic activity was enriched (supplementary fig. S5, Supplementary Material online).

Stronger Purifying Selection on Maize1 Subgenome Only Exists for BED Genes

We looked if subgenome dominance (maize1 genes having stronger purifying selection) exists for our expression-based gene classes (UED-dominant, UED-repressed, and BED). We did not find increased purifying selection (measured by ΠTn/ΠTs) in maize1 UED-dominant genes compared with maize2 UED-dominant genes and for maize1 UED-repressed genes compared with maize2 UED-repressed genes (fig. 4). We also found lower ΠTn/ΠTs than maize2 genes, but the effect is not as strong as for UED-dominant versus UED-repressed genes (fig. 3). As we present nucleotide diversity data from teosinte lines while the expression data is from maize line B73, we conform this analysis using modern inbred lines (of which B73 is a member) and found qualitatively similar results (supplementary fig. 9, Supplementary Material online).

Supplementary Table 1

<table>
<thead>
<tr>
<th>Region</th>
<th>Correlation ρ Value for Nucleotide Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intron</td>
<td>0.56*</td>
</tr>
<tr>
<td>Synonymous</td>
<td>0.24*</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>0.19*</td>
</tr>
<tr>
<td>Gene</td>
<td>0.21*</td>
</tr>
<tr>
<td>Upstream</td>
<td>-0.01 (not significant)</td>
</tr>
</tbody>
</table>

Note.—This table gives Spearman’s correlation coefficient for nucleotide diversity for duplicate gene pairs for different genic regions.

*ρ < 2.2 × 10^-16.

Most Retained Genes Are Either Diverged in Expression or Subfunctionalized

We used the relative expression of duplicate pairs in various maize tissues to classify the expression divergence as unidirectional or bidirectional. Gene pairs with at least 2-fold difference were considered to be diverged in expression. Of 3,228 retained gene pairs, 1,641 pairs showed an UED where one member of the pair has a consistently higher expression in all tissues with a 2-fold difference (UED-dominant) compared with the counterpart (UED-repressed). One thousand five hundred seventeen pairs showed a BED where one copy displayed higher expression in one tissue and the other copy in another tissue. Thus close to 98% of the retained gene pairs seem to have diverged in expression or have divergent expression pattern across tissues.

UED and BED Genes Form Distinct Subsets in GO Enrichment

We did separate GO enrichment analysis of UED and BED genes. We found transcriptional regulation to be enriched for BED genes (supplementary fig. S6, Supplementary Material online) and macromolecular complexes and structural molecule genes to be enriched for UED genes (supplementary fig. S7, Supplementary Material online). A complete cross comparison of GOs between these two data sets is available in supplementary table S4, Supplementary Material online.
expression. maize1 BED genes dominate the maize2 BED genes in expression in a larger number of tissues (fig. 5). The median number of tissues in which maize1 BED gene dominates in expression is 6 compared with 4 for a maize2 BED gene ($P = 1.8 \times 10^{-12}$, Wilcoxon rank-sum test). Overall, we suggest that dominant expression either consistently or in a larger number of tissues is a determinant for stronger purifying selection rather than the subgenome location.

Genes Displaying Mutant Phenotype Are Broadly Expressed and Are Under Purifying Selection

We looked at cases where one copy of the paralogous pair displays a mutant phenotype while the other copy has no visible mutant phenotype associated, thus the duplicate copy fails to complement. We analyzed 15 such cases and the member displaying the mutant phenotype is dominant in more tissues and has lower ratio of nonsynonymous to synonymous diversity, except in two cases (supplementary table S3, Supplementary Material online). This indicates that paralog displaying mutant phenotype is under stronger purifying selection than its counterpart and this is intertwined with the dominant expression of the copy displaying the phenotype.

Upstream Regions of Repressed Genes Are Enriched in TEs

Upstream TEs have been proposed as a mechanism for generating differences in expression between paralogs (Freeling et al. 2012; Woodhouse et al. 2014). To test this hypothesis, we looked at the fraction of upstream 2 kb region covered by annotated repeats and TEs. The mean (median) coverage for UED-repressed genes was 0.33 (0.27) compared with the coverage of UED-dominant genes being 0.26 (0.19) ($P = 9.2 \times 10^{-13}$, Wilcoxon rank-sum test) and BED genes 0.27 (0.20) ($P = 2.9 \times 10^{-10}$). However, the coverage difference between maize1 and maize2 was not found to be significant (0.28 (0.2) vs. 0.29 (0.23)) ($P = 0.02$) at $P < 0.01$. We also looked at the nearest upstream distance between the annotated transcription start site and the upstream annotated TE. The mean (median) upstream distance of an UED-repressed gene from an annotated repeat or TE was 908 (480) bp compared with UED-dominant genes 1,102 (610) bp ($P = 1.3 \times 10^{-8}$, Wilcoxon rank-sum test) and BED genes 1,019 (597) bp ($P = 4.8 \times 10^{-8}$) (fig. 6). The distribution of this distance was not found to be significantly different between UED-dominant and BED genes ($P = 0.24$). Interestingly, this distance was also not found to be significant between maize1 and maize2 retained genes, 1,039 (587) bp versus 975 (564) bp ($P = 0.06$, Wilcoxon rank-sum test). Thus, the repressed genes (UED-repressed) not only have higher fraction of upstream regions covered by TEs but also have the nearest distance to an upstream TE.

Discussion

The recent nature of WGD in maize and the availability of a comprehensive polymorphism data set give an unique opportunity to investigate the evolution of duplicate genes at a population level. We report genome scale correlations in nucleotide divergence and diversity between duplicate gene pairs (table 1 and fig. 1) but no correlation for upstream (2 kb) regions indicating that promoter evolution between duplicates is uncoupled. These correlations mean a general similarity in mutabilities, mutation rates, and at least a component of protein structural and functional constraints to be similar among duplicates. Moreover, both subgenomes share the same demography. These correlations are particularly striking given the 5–12 My of post WGD-divergent evolution and the volatile nature of maize genome with copious genome rearrangements and abundant transposon activity (Schnable et al. 2009).

Studies have predicted maize1 genes to be under stronger purifying selection because of the dominant expression (Schnable et al. 2011) and higher tendency of genes displaying...
mutant phenotypes to be located in maize1 subgenome (Schnable and Freeling 2011). Here we show that this difference is also reflected in the patterns of polymorphism. Maize1 genes exhibit less nonsynonymous divergence and diversity compared with maize2 genes, whereas the difference in synonymous divergence and diversity is not significant. DoFE analysis shows an excess of large effect mutation class (\(-NeS > 100\)) in maize1 subgenome (fig. 2). Furthermore, SIFT predictions show a higher number of genes harboring deleterious polymorphisms in the maize2 subgenome (supplementary table S1, Supplementary Material online). In the later part of the study, we show that this effect largely arises from dominant expression of the maize1 gene copies.

Next, we show that dosage seems to be a major determinant for gene retention after WGD in maize. Genes retained after WGD in maize exhibit a higher median expression than single genes across many tissues (supplementary fig. S2, Supplementary Material online). This is consistent with studies from diverse organisms (Aury et al. 2006; Chain et al. 2011; Chen and Sankoff 2014). In addition to dosage, typically networked proteins like transcription factors and macromolecular complexes are more likely to be retained after a WGD (Maere et al. 2005; Wu and Qi 2010; Bekaert et al. 2011). We find similar trends in GO analysis of retained genes (supplementary figs. 3 and 4, Supplementary Material online) indicating that established features behind gene retention after WGD are also acting in maize.

There exist two major mechanisms for ohnolog retention. First, the dosage balance model in which relative dosage of interacting genes is under purifying selection and deletion of one member of the pair creates stoichiometric imbalance in the interaction network. The second is the subfunctionalization model where ancestral function is subdivided between both members. We looked at the relative dosage between paralogous pairs in various tissues and partitioned the genes into two major categories consistent with both models. The first category being UED gene category where one paralog has consistently dominant expression (UED-dominant) across all tissues with divergent expression compared with the other paralog (UED-repressed). The second category being BED for which both paralogs have alternatively dominant and repressed expression but in different tissues. The UED genes comply with the dosage balance model where dosage of one member is consistently reduced ameliorating the dosage constraints and making the repressed gene dispensable. The BED genes comply with the tissuewise subfunctionalization model where both paralogs perform the function albeit in different tissues. Around 98% of the duplicate pairs could be classified into either BED or UED indicating that sorting of duplicate genes by expression is almost complete and most of the duplicate gene pairs have either diverged in expression consistently or in a tissuewise manner. Such rapid divergence in gene expression between duplicates has been observed for yeast and humans (Gu et al. 2002; He and Zhang 2005) and can even be established in a few generations after the formation of an allotetraploid (Adams 2007). These classifications also show strikingly different GO enrichments (supplementary table S4, Supplementary Material online). UED genes tend to be part of macromolecular complexes and structural molecules (supplementary fig. S7, Supplementary Material online), whereas BED genes are more likely to be involved in transcription regulation (supplementary fig. S6, Supplementary Material online). These results indicate that tissuewise subfunctionalization seems to largely influence the regulatory network, while consistent repression influences the protein–protein interaction network. The agreement between UED genes and relative dosage model could be further tested by overlaying the protein–protein interaction network on UED genes and checking if members of the duplicated pathways divide equally between UED-repressed and UED-dominant.

The UED-dominant genes show the strongest purifying selection followed by BED genes and then the UED-repressed genes (fig. 3), indicating that gene dosage is positively correlated with the strength of purifying selection. Because the strength of purifying selection for tissuewise subfunctionalized (BED) genes seem to be intermediate between UED-dominant and UED-repressed, we checked if the dosage of dominantly expressed BED genes in a particular tissue is less than dosage of the UED-dominant genes. We find this to be true. Dominantly expressed BED genes have lower expression than UED-dominant genes in all tissues assayed except pollen (supplementary fig. S8, Supplementary Material online).
Genes with higher dosage appear to suppress one copy consistently, whereas genes with lesser gene dosage evolve tissue-specific regulatory patterns.

The subgenome dominance in purifying selection is not significant for UED genes and only exists for BED genes (fig. 4). When the genes are classified as UED-dominant and UED-repressed, the strength of purifying selection does not appear to be different between maize1 and maize2 genes in these categories. Thus, the purifying selection results from the dominant expression and maize2 genes appear statistically to be under stronger purifying selection because there are more UED-dominant maize1 genes. The tissuewise subfunctionalized genes (BED) still show a difference in purifying selection between maize1 and maize2 genes (fig. 4).

However, we show that for BED genes a maize1 BED gene is more likely to be dominantly expressed in a larger number of tissues compared with its maize2 counterpart (fig. 5). Thus, the factors determining the strength of purifying selection are not only the dominant expression but also the number of tissues in which the gene is dominant. It could be possible that the broadly expressed copy retains the original expression pattern, whereas the other copy gains a new tissuewise expression profile or loses some part of the ancestral expression profile. It is hard to distinguish between such subfunctionalization and neo-functionalization unless the ancestral state of expression is known.

The gene pairs in which one member is associated with a mutant phenotype present an enigmatic case wherein the duplicated copy does not seem to complement the function. We show that the gene at a given pair which displays a mutant phenotype is generally dominant in a larger number of tissues and has lower ratio of nonsynonymous to synonymous diversity (supplementary table S3, Supplementary Material online) indicative of higher purifying selection acting on it. We suggest that the distribution of function among paralogs is asymmetric with one paralog handling larger component of the function. This explains why the duplicate copy fails to complement.

Nevertheless, the mechanism behind the subgenome dominance is still not understood. Dominant expression can explain that the deletion of a nondominantly expressed ohnolog incurs a lower fitness cost due to relaxed dosage constraints. However, the reason why maize1 subgenome has more dominantly expressed genes is yet to be answered. Epigenetic effects like chromatin modification and DNA methylation have not been proposed as underlying mechanisms but have not been proven. Reduced expression of genes caused by upstream TEs can cause a subgenome dominance if the two progenitor genomes in an allotetraploid differ in the number of TEs at start (Freeling et al. 2012; Woodhouse et al. 2014). We tested a part of this hypothesis by associating upstream TEs with our different expression categories and indeed found an enrichment of TEs in upstream regions of repressed genes (UED-repressed) (fig. 6). Repression of these TEs could inadvertently cause the repression of nearby genes as demonstrated in *Arabidopsis thaliana* (Henderson and Jacobsen 2008; Hollister and Gaut 2009). Note that we cannot establish causality because upstream transposon abundance can as well be a consequence of low expression and reduced purifying selection. A recent study reported an unexpected correlation between expression of nearby genes which extends to more than 100 kb (Ghanbarian and Hurst 2015) which could point to the mechanism generating subgenomes.

We conclude that majority of ohnolog pairs from the maize WGD have diverged in expression either consistently or in a tissuewise manner. This divergence seems to prevent functional complementation by the duplicate copy. Relative and absolute gene dosage and the number of tissues in which a gene is dominant is an important determinant of purifying selection. Transcription regulators are more likely to develop tissuewise subfunctionalization, while macromolecular complexes tend to suppress one duplicate copy consistently. We propose that an expression-based classification of duplicates is more biologically relevant than subgenomes. Finally, upstream divergence and TEs form a possible mechanism for the expression divergence.

### Materials and Methods

#### Obtaining SNP Data

SNP data were obtained from the maize hapmap2 project (Chia et al. 2012) which contains whole-genome SNP data for 19 teosinte (wild relatives of maize), 23 landraces, and 60 modern inbred lines. For this study, we show the data from the teosinte lines only as they are the closest approximation to a panmictic population. VCF (variant call format) file for maize hapmap2 SNPs was downloaded from the URL: data.iplantcollaborative.org/quickshare/e75bc315fc0f9fda/HapMapV2RefgenV220120328.vcf.gz. The 19 teosinte lines contain 17 lines from the sub species parviglumis and 2 from the subspecies mexicana. The lines belonging to mexicana were removed because it forms a different subspecies. Two more lines TILI04-TIP285;TEO and TILI06-TIP496; TEO were removed as they are similar lines to TILI04-TIP454;TEO and TILI06-TIP260;TEO already present and differing only in the generations of selfing. One more line TILI02 was removed due to low coverage. The resulting VCF containing 14 teosinte lines was annotated using the program..

**Fig. 4.** Ratio of nonsynonymous to synonymous nucleotide diversity (T1n/T1s) for maize subgenome 1 and 2 genes for different expression classifications. UED-D, UED-dominant; UED-R, UED-repressed. P values were calculated using Wilcoxon rank-sum test: *P = 2.9 × 10⁻¹⁰*; ns, not significant.
Calculating Nucleotide Diversity

The VCF file was converted into “hapmap” format using a perl script. Variscan (Vilella et al. 2005) was used with runmode 12 to obtain values for nucleotide diversity. The per SNP diversity (estimated as the average pairwise difference, $\Pi$) was summed over all genic (respectively intronic) SNPs and then divided by the length of the gene (respectively intron) to obtain genic (intronic) diversity. For synonymous and non-synonymous diversity, $\Pi$ was summed over all synonymous and nonsynonymous SNPs and then divided by the total number of synonymous and nonsynonymous sites in the gene. SNPs with less than 50% of genotypes called were removed from the calculations. Because we do not remove the sites with missing data in the total number of sites, we underestimate the per-site diversity but our main interest lies in the differences in diversity between categories rather than the absolute value. Moreover, we take the ratio of nonsynonymous to synonymous diversity so any difference in coverage would affect both the sites similarly due to their intertwined nature. Per-site diversity estimates for teosinte lines for the entire genes are available from the hapmap2 project (Chia et al. 2012) from the URL: figshare.com/articles/new_filesset/757736 in which the diversity was obtained by removing sites with missing data. However, this is a very stringent procedure and the median covered gene length for retained genes was 213 bp resulting in too few SNPs to calculate nonsynonymous and synonymous diversity. Note that the genic $\Pi$ values from the hapmap2 correlate nevertheless strongly with our computations (Spearman’s rho 0.59, $P < 2 \times 10^{-16}$; supplementary fig. S11, Supplementary Material online). Also, we did not find any significant difference between maize1 versus maize2 genes and UED-dominant versus UED-repressed genes for the distribution of the covered gene length from the hapmap2 project.

Calculating Sequencing Depth for Genes

Because the calculations for nucleotide diversity can be influenced by the read coverage, it was calculated for each retained gene. Hapmap2 Binary Alignment Map (BAM) files for each line used in the analysis were downloaded from the URL: http://mirrors.iplantcollaborative.org/download/ipant/home/shared/panzea/hapmap2/bam/. Bedtools (Quinlan 2002) was used to get the read depth at protein-coding region (open reading frame) for each gene for each individual line. Because coverage varies at each position for different members of sequenced lines, a “mean coverage fraction” was calculated which is the sum of the fraction of the individuals at each site with at least one read aligned, normalized by the length of the open reading frame. Mean coverage fraction for UED-dominant genes was 0.81 versus 0.79 for UED-repressed genes, which indicates slightly low coverage for repressed genes. This might be caused by higher divergence or higher content of repeats or TEs. Because we report UED-repressed genes to have higher
nucleotide diversity, our computation is an underestimate as we might miss SNPs in those genes due to the low coverage. Thus, read coverage does not alter the direction of results, on the contrary makes them more conservative.

Calculating DoFE
DoFE gives the probability of a mutation having a given deleterious effect (Eyre-Walker and Keightley 2007). The method uses the number and frequency of SNPs for two classes of sites, one assumed to be selected and another being neutral. Here, nonsynonymous sites represented the selected class, and synonymous the neutral one. The method is based on the premise that mutations on selected sites are few and kept at a lower frequency by purifying selection. The Eyre-Walker and Keightley method as implemented in the software DoFE was used (Eyre-Walker and Keightley 2009). The software was downloaded from Adam Eyre-Walker’s lab URL: http://www.lifesci.susx.ac.uk/home/Adam_Eyre-Walker/Website/Software.html. As a test this method was run on classical genes in maize versus 15,000 maize protein-coding genes with cDNA evidence (supplementary fig. S1, Supplementary Material online). Classical genes are genes which are well studied in maize and are more likely to exhibit a mutant phenotype (Schnable and Freeling 2011). The results are given in supplementary fig. S1, Supplementary Material online, where the strength of selection is represented as a product of selection coefficient and effective population size (−NeS). Higher values of −NeS indicate more purifying selection. Higher fraction of mutations in higher −NeS classes shows stronger purifying selection acting on classical genes compared with a set of random genes. This method was then run on maize1 and maize2 SNPs.

Obtaining Ka and Ks
The list of genes belonging to each subgenome with their Sorghum ortholog was downloaded from James Schnable’s webpage URL: http://skraelingmountain.com/data-630-sets.php.php. For each gene the splice variant with the longest protein-coding sequence was chosen for further analysis. Ka and Ks values for maize–sorghum orthologs were obtained from ensemble biomaRt website URL: plants.ensembl.org/biomart/martview/ (Kinsella et al. 2011). A static dump of the file is available at the URL: ftp.ensemblgenomes.org/pub/plants/release-27/mysql/plants_mart_27/zmays_eg_gene_homolog_sbicolor_eg_dm.txt.gz. Ka and Ks values <0 and >0.5 were excluded from the analysis.

Obtaining SIFT Scores
SIFT is an algorithm for predicting the deleterious effect of nonsynonymous polymorphisms (Ng and Henikoff 2001). It is based on site conservation. It gives a score between 0 and 1, with 0 being intolerant and 1 being neutral. Typically, a score between 0 and 0.05 is considered to be deleterious (Sim et al. 2012). Sift scores for maize were downloaded from the sift4g website URL: siftdb.org.

Expression Data
Expression data for maize inbred line B73 in the form of FPKM (fragments per kilobase per million mapped reads) values for diverse tissues was downloaded from the qteller website URL: http://qteller.com/qteller3/. The qteller website gives expression for maize genes combining data from various studies. The tissues with expression data under peculiar local conditions (e.g., drought) were removed from analysis. The total list of 22 tissues used is available in supplementary table S2, Supplementary Material online. Relative expression of ortholog pairs was calculated and a 2-fold expression difference was used as a threshold to define dominance in expression. If both paralogs have expression <0.5 FPKM in a particular tissue then that comparison was discarded. Combining the data from different studies might cause some differences in expression to be due to differences in studies gathering the expression data. We would like to stand by this decision due to three reasons. First and importantly, only relative expression between two duplicate copies was compared and no comparison of expression values of genes across tissues was made. Second, it is important to sample many tissues to obtain accurate assignment of BED genes which is only possible by using expression data from different studies. Third, FPKM is a normalized measure of expression which takes into account the sequencing depth. Note that a similar protocol is used for aligning reads and obtaining the FPKM values in the qteller website.

GO Analysis
GO analysis was done using AgriGo tool (Du et al. 2010). “Single enrichment analysis” option of AgriGo was used with all maize genes with syntenic orthologs as background. This list was obtained from James Schnable’s website URL: http://skraelingmountain.com/datasets/grass_syntenic_orthologs.csv.zip (Schnable et al. 2012). A false discovery rate of <0.05 was used for a given enrichment to be significant.

Upstream Transposable Elements
Information about the annotated repeats in maize AGPV2 assembly was obtained from URL: http://ftp.maize-sequence.org/release-5b/repeats/. Most repeats were annotated as TE’s. These files were combined and converted to a Browser Extensible Data (BED) format. Bedtools (Quinlan 2002) “coverage” command was run to obtain fraction of 2 kb upstream regions covered by repeats/TEs. The Bedtools “closest” command was used to obtain the distance from the transcription start site to the nearest TE.

Statistical Analysis
All statistical analyses were done using R (http://www.R-project.org). Because most parameters used in the study follow a non-normal distribution, nonparametric tests were used for significance and medians reported instead of means.
R package “boot” (https://cran.r-project.org/web/packages/boot/index.html) was used to calculate confidence intervals for statistics but they were found to be very small and were not plotted. A visual inspection of the data for extreme values was done to ensure that statistical significance was not caused due to them.

Supplementary Material
Supplementary tables S1–S4 and figures S1–S11 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustER Synbreed (synergistic plant and animal breeding, FKZ: 0315528). We thank Sandra Unterseer and Remco Stam for comments on the manuscript. We also thank James Schnable for providing convenient access to several data sets. We specially thank Jeffrey Ross-Ibarra, an anonymous reviewer, and Brandon Gaut whose comments made fundamental improvement in the quality of the article.

References
/C19


