Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes

Citation	Chettoor, A. M., Givan, S. A., Cole, R. A., Coker, C. T., Unger-Wallace, E., Vejlupkova, Z., ... & Evans, M. (2014). Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. Genome Biology, 15(7), 414. doi:10.1186/s13059-014-0414-2
DOI	10.1186/s13059-014-0414-2
Publisher	BioMed Central Ltd.
Version	Version of Record
Terms of Use	http://cdss.library.oregonstate.edu/sa-termsofuse

Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes

Antony M Chettoor[1]
Email: chettoor@stanford.edu

Scott A Givan[2]
Email: givans@missouri.edu

Rex A Cole[3]
Email: colerex@science.oregonstate.edu

Clayton T Coker[1]
Email: ccoker@stanford.edu

Erica Unger-Wallace[4]
Email: eunger@iastate.edu

Zuzana Vejlupkova[2]
Email: zuzanav@science.oregonstate.edu

Erik Vollbrecht[4]
Email: vollbrec@iastate.edu

John E Fowler[3]
Email: fowlerj@science.oregonstate.edu

Matthew MS Evans[1*]
[*] Corresponding author
Email: mevans@carnegiescience.edu

[1] Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 94305, USA

[2] Informatics Research Core Facility, University of Missouri, Columbia, MO 65211, USA

[3] Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, USA

[4] Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA

Abstract

Background

Plant gametophytes play central roles in sexual reproduction. A hallmark of the plant life cycle is that gene expression is required in the haploid gametophytes. Consequently, many mutant phenotypes are expressed in this phase.

Results

We perform a quantitative RNA-seq analysis of embryo sacs, comparator ovules with the embryo sacs removed, mature pollen, and seedlings to assist the identification of gametophyte functions in maize. Expression levels were determined for annotated genes in both gametophytes, and novel transcripts were identified from *de novo* assembly of RNA-seq reads. Transposon-related transcripts are present in high levels in both gametophytes, suggesting a connection between gamete production and transposon expression in maize not previously identified in any female gametophytes. Two classes of small signaling proteins and several transcription factor gene families are enriched in gametophyte transcriptomes. Expression patterns of maize genes with duplicates in subgenome 1 and subgenome 2 indicate that pollen-expressed genes in subgenome 2 are retained at a higher rate than subgenome 2 genes with other expression patterns. Analysis of available insertion mutant collections shows a statistically significant deficit in insertions in gametophyte-expressed genes.

Conclusions

This analysis, the first RNA-seq study to compare both gametophytes in a monocot, identifies maize gametophyte functions, gametophyte expression of transposon-related sequences, and unannotated, novel transcripts. Reduced recovery of mutations in gametophyte-expressed genes is supporting evidence for their function in the gametophytes. Expression patterns of extant, duplicated maize genes reveals that selective pressures based on male gametophytic function have likely had a disproportionate effect on plant genomes.

Background

The plant life cycle has genetically active diploid and haploid phases, called the sporophyte and gametophyte respectively [1]. In angiosperms the gametophytes are highly reduced, are dependent on the parent sporophyte, and develop embedded within the diploid sporophyte tissues, with a three-celled male gametophyte and a female gametophyte consisting of as few as seven cells.

To produce the female gametophyte, or embryo sac, after meiosis, one spore undergoes three rounds of synchronous divisions to produce an 8-nucleate syncytium with micropylar and chalazal clusters of four nuclei each [2]. Cellularization then produces seven cells: two synergids, the egg cell, the bi-nucleate central cell, and three antipodal cells [3]. In maize, the antipodal cells continue to divide during embryo sac maturation, reaching a final number of 20 to 100 cells. The male gametophyte, or pollen grain, has an even more reduced phase of growth. Each microspore first undergoes an asymmetric cell division to produce the

vegetative cell and the generative cell. The generative cell then divides once to produce the two sperm cells, which are carried within the vegetative cell. In addition to expressing functions required for pollen grain development, the vegetative cell must also generate the tip-growing pollen tube that navigates through the pistil tissues to reach the embryo sac and deliver the sperm cells [4].

Mutations in genes required in the gametophytes result in characteristic fertility phenotypes and modes of transmission that have formed the basis of many mutant screens [5-9]. When heterozygous, mutations affecting the embryo sac are expected to have reduced fertility and seed set, because half of the ovules contain mutant embryo sacs and so often fail to produce seed. Mutations affecting the male gametophyte do not cause reduced seed set, because both wild-type and mutant pollen from heterozygotes enter the pistil. However, for mutations affecting male and/or female gametophytes, the mutant allele (and the alleles of loci linked to it) is found at a reduced frequency in progeny when the defective gamete is involved (*i.e.* male gametophyte mutants are recovered poorly when heterozygotes are crossed as males). This characteristic reduced transmission also prevents, or makes very difficult, the generation of mutant homozygotes. Note that genetic redundancy can facilitate the recovery of mutations in genes active in the gametophytes but also can complicate recognizing them as such, given generally weaker phenotypes. Maize, as an ancient allotetraploid constituted by two progenitor genomes (subgenomes 1 and 2), has a mix of genes present as either duplicated pairs (homeologs), or as singletons, due to gene loss [10]. Notably, subgenome 2 is characterized by lower levels of gene expression and higher rates of gene loss than subgenome 1 [11].

Because of the poor recovery of gametophyte-lethal mutants, additional strategies (e.g., transcriptome profiling) have been utilized to identify gametophyte active genes in several species. Microarrays were used first to assess the transcriptomes of pollen [12-15] and embryo sacs (by comparing ovules with and without embryo sacs) [16-21] in Arabidopsis. These studies identified up to ~14,000 genes as expressed at some point in pollen development [13], with ~6500-7200 expressed at the mature pollen stage [13,15] and 1200 embryo sac-expressed genes. The identification of more expressed genes in pollen is likely due to the ease of isolating large populations of relatively pure material. Sperm cell purification and assessment of growing pollen tubes have extended these studies to provide additional details of male gametophytic transcriptomes [22-25].

Enrichment of embryo sac cells (e.g., by cell wall digestion and dissection [26] or Laser Capture Microdissection [27]) facilitated the identification of additional genes expressed in the embryo sac. Isolation of gametophyte cells for EST sequencing or microarray hybridization in maize, rice, and wheat [26,28-31] identified greater complexity for the egg transcriptome than that of sperm, and a preponderance of unknown, hypothetical, and novel proteins encoded by these transcripts. Three of the cell types in the mature Arabidopsis embryo sac (the egg cell, the central cell and the synergids; but not the antipodals) were analyzed by microarray, with 8,850/20,777 of the genes on the ATH1 chip identified as expressed [27], a number comparable to mature pollen. RNA-Seq analysis removes some of the limitations associated with sequencing individual cDNA clones or microarray technology (*e.g.* not all of the genes are present on the microarray), revealing both the expression of a higher fraction of known transcripts in the gametophytes and the existence of new genes and transcript isoforms in mature pollen and the central cell of the female gametophyte [32-34]. RNA-Seq has also identified gene families enriched in the central cell that were missed in microarray studies [35].

These studies have revealed a few broad themes. The pollen transcriptome is the most distinctive, although all gametophytic transcriptomes have some similar features to one another. Of sporophytic transcriptomes, the early embryo (heart and globular stages) is most similar to the gametophytes [15,24,27]. Some parallels for plant egg and sperm cell transcriptomes with animal gamete transcriptomes have also been detected, particularly with regards to the epigenetic regulation of gene function through small RNA pathways [27]. Within the embryo sac, the egg and central cell transcriptome are more similar to one another than to the synergids. Small signaling peptides of the DEFENSIN/LURE (DEFL) family are overrepresented in the female gametophyte (particularly the central cell), although only a subset of these was assayed in the whole embryo sac. Pollen grain transcriptomes are enriched for GO terms related to signaling, vesicle trafficking, cell wall functions, and cytoskeletal functions thought to be important for tip growth [15,25]. Finally, putative connections between epigenetic regulation, small RNA pathways, and reactivation and silencing of transposable elements have been observed in gametophyte transcriptomes [23,27,36]. Preliminary RNA-Seq analysis is available for maize mature pollen [37], which, as is the case in Arabidopsis, is very different from sporophytic tissues. Use of *de novo* transcript assembly of RNA-Seq reads in maize has also been used to study long non-coding RNAs in reproductive and vegetative tissues [38]. Reproductive tissues, including male and female gametophytes, express more lncRNA loci than vegetative tissues.

Here the first detailed, replicated, RNA-Seq-based analysis of both male and female gametophytic transcriptomes of maize (or any monocot) is used to identify genome features with differential expression between the gametophytes and sporophytic tissues, including protein-coding gene families, duplicated genes, and previously unannotated genes. These studies identify small signaling peptides and several transcription factor gene families as being overrepresented in gametophyte transcriptomes. The first genome-wide comparison of gene expression patterns on duplicate gene retention also reveals an effect of pollen gene function on genome evolution. This study also provides the first evidence for transposon expression in the male and female gametophytes of a plant with a large, complex genome containing many active transposon classes.

Results

Production of RNA-seq and mapping of reads to the B73 reference genome

To define the transcriptome of mature maize male and female gametophytes, RNA-Seq was performed on four tissue types: nine-day old, above-ground seedling (S); mature pollen (MP); embryo-sac-enriched samples with some remaining nucellar cells (ES); and ovules with embryo sacs removed (Ov). We generated between ~54 million to ~195 million mappable Illumina reads per B73 sample. The ES samples, which had the lowest amount of starting material and required additional amplification before sequencing, had the lowest percentage of reads that could be mapped back to the reference genome, ranging from 54% to 62% of the total reads per replicate. Before mapping reads to the maize genome, reads were compared to the available maize repeat database to remove reads with a high confidence match to maize repetitive elements [39]. Remaining reads were mapped to the maize genome sequence in two ways: (1) to the existing gene models to determine expression levels for annotated genes and (2) to the reference genome sequence independently of gene models to build empirical transcripts to aid the identification of novel genes. There are three gene sets for the maize B73 RefGen_v2: the Filtered Gene Set (FGS), composed of high-confidence gene models;

the Rejected Gene Set (RGS), composed of lower-confidence gene models that include likely pseudogenes and transposons; and the Working Gene Set (WGS), which encompasses both FGS and RGS. To insure that unknown gametophytic transcripts were not missed in this analysis, B73 RNA-Seq reads were mapped to both the WGS (Additional file 1: Table S1) and FGS (Additional file 2: Table S2) gene models (summarized in Table 1).

Table 1 RNA-Seq reads mapped backed to the maize genome (5a.59)

	B73 Seedling	B73 Mature Pollen	B73 Embryo Sac	B73 Ovule without Embryo Sac	W23 Embryo Sac	W23 Ovule without Embryo Sac
Percentage of reads mapped back to nuclear genome	82%	85%	61%	81%	nd	nd
Total mapped reads	91,076,832	123,536,281	50,435,150	140,282,423	nd	nd
Reads mapped to FGS low-copy exons	46,765,091	100,723,578	34,727,288	117,425,717	17,569,670	16,837,249
Reads mapped to RGS Low-copy exons	2,099,557	1,822,553	1,602,788	3,873,497		
Intron	3,214,112	367,752	1,110,429	2,277,426	nd	nd
Intergenic	1,522,581	2,732,794	5,583,100	4,199,131	nd	nd
Ribosomal RNA genes	21,039,136	4,306,489	729,102	4,224,584	nd	nd
Transposons and other repeats	8,256,280	12,504,295	3,590,804	3,629,696	nd	nd
Mitochondrial genes	3,442,450	49,613	1,399,727	2,450,967	nd	nd
Chloroplast genes	4,176,008	771	988,470	1,592,112	nd	nd
Genes in FGS with average expression > 0.1 FPKM (39,635 total)	27,564	14,591	27,530	25,971	20,857	20,539
Genes in RGS with average expression > 0.1 FPKM (69,689 total)	8,165	4,335	17,751	11,933	nd	nd
Percentage of FGS Expressed Genes (>0.1 FPKM) in only one replicate	9%	18%	16%	11%	19%	9%
Percentage of FGS Expressed Genes (>0.1 FPKM) in two replicates	8%	13%	17%	8%	15%	14%
Percentage of FGS Expressed Genes (>0.1 FPKM) in all Three Replicates	83%	69%	67%	81%	66%	77%

nd = not determined.

The variability of samples can be seen in the percentages of gene models expressed above an arbitrary threshold (0.1 fragments per kilobase per million reads - FPKM) that are shared between replicates (Table 1 and Figure 1). ES was the most variable. The percentage of genes shared across all three ES samples is lower than in the other tissues (67% vs. 69-83%). Because of the variability of the ES samples, an additional set of ES and comparator Ov samples were analyzed to improve identification of genes enriched in the embryo sac. RNA-Seq was performed on a set of samples from a different inbred line, W23, using the ABI SOLiD platform. These reads were mapped against the FGS genes, and FPKM values calculated for each gene (Additional file 2: Table S2). The W23 ES samples had similar variability between replicates as the B73 ES (66% of genes above 0.1 FPKM are shared in all three). For those FGS genes with an average signal above threshold in the W23 samples, there was a strong concordance with the same characteristic in B73 (94% for ES, 93% for Ov), arguing that the samples from the two inbred lines are indeed comparable. All subsequent analysis of FGS gene expression values used a modified average of the W23 and B73 ES and Ov samples derived as described in the Materials and Methods.

Figure 1 Similarity between replicates. The lists of genes with expression above 0.1 FPKM for each sample was compared between biological replicates. Overlaps between replicates within each tissue type are shown. The number of genes with an expression of at least 0.1 FPKM within each set or overlap between sets is indicated. The samples with the least overlap are the B73 Embryo Sacs. **(A)** B73 Seedlings; **(B)** B73 Mature Pollen; **(C)** B73 Embryo sac enriched; **(D)** B73 Ovules without embryo sacs; **(E)** W23 Embryo sac enriched; **(F)** W23 Ovules without embryo sacs; **(G)** Overlap between lists of FGS genes with an average expression above 0.1 FPKM for each tissue type. FGS = filtered gene set.

Several trends involving genomic regions were revealed when reads were mapped to the whole genome (Table 1 and Figure 2). For example, the reads mapping to ribosomal sequences and the chloroplast genome were highest in seedling, whereas MP was nearly devoid of reads from both the mitochondrial and chloroplast genomes. Reads classified as intronic were also notably less frequent in MP (0.3%) than in the other tissues tested (2% to 3%). This is in contrast to Arabidopsis pollen, which has a high frequency of intron reads [33]. One possible explanation for the low representation of introns in MP is that, in contrast to the other sample types, mature pollen is in a somewhat quiescent state prior to contact with female tissue. Thus, this observation is consistent with the view that the vast majority of mRNAs in MP are fully mature (i.e., completely spliced), and stored for rapid translation upon pollen tube germination [40].

Figure 2 Distribution of RNA-Seq reads to different genomic features in maize. The frequency of the reads mapping to transposable elements and other intergenic sequences (TE & other intergenic), ribosomal RNA genes (rRNA), other nuclear annotated gene model exons (exons), annotated gene introns, mitochondrial genes, and chloroplast genes. Dramatic differences were seen with rRNA reads most abundant in the seedling tissue, TE & other intergenic transcripts lowest in the Ov, chloroplast and mitochondrial transcripts lowest in MP, and TE & other intergenic transcripts most abundant in ES.

Non-exonic reads (intron, transposable element and other intergenic) were overrepresented in ES samples relative to the other transcriptomes (approximately 50% more frequent than in seedling and MP and 3-fold more frequent than in surrounding Ov), raising the possibility that ES-specific genes have been systematically missed in the current WGS and FGS predictions. To identify genes absent from the current WGS and FGS predictions, all reads were used to build empirical transcript models. The resulting dataset contains 31,015 models longer than 100 bp that are completely intergenic relative to the existing WGS gene models and are detected above 0.1 FPKM (Additional file 3: Tables S3). However, 27,685 of these intergenic models (89.3% of the total) were classified as transposable element-related or other repeat-related via BLAST, using previously validated parameters [41], or via RepeatMasker (see Materials and Methods) (Additional file 3: Table S3C). A small number of these repeat-related transcripts (1,174; 4.2%) overlap with long non-coding RNA (lncRNA) loci [38]; a larger percentage of the non-repeat related intergenic gene models (648 out of 3,330; 19.5%) show lncRNA locus overlap (Additional file 3: Table S3A).

Thus, most of the 3,330 non-repeat related intergenic transcript models (Additional file 3: Table S3B) represent potential novel protein-coding genes. Although many of these models are small (100–200 bp, possibly incomplete transcripts), the overall average is 546 bp, with lengths extending up to 3.4 kb. The largest category of these did not show enriched expression in any one tissue (Table 2). However, both embryo sac and pollen samples were associated with significantly higher counts of tissue-enriched (i.e., 2-fold higher than any

other sample) intergenic transcript models than either sporophytic sample assessed. Non-enriched, ES-enriched, and MP-enriched transcript models show a similar likelihood to encode proteins, based on BLAST and InterProScan assessment (ES, 25.9%; MP, 24.2%; non-enriched, 29.8% - Additional file 3: Table S3A).

Table 2 Novel gene models identified by transcript assembly from RNA-Seq data

	Non-TE/Repeat Related Gene Models	Average Length (bp)	TE/Repeat-Related Gene Models	Average Length (bp)
Seedling enriched (2x higher than other three tissues)	26	515	74	816
Pollen enriched (2x higher than other three tissues)	376	679	1,133	836
Embryo Sac enriched (2x higher than other three tissues)	622	484	4,395	881
Ovule (w/o Embryo Sac) enriched (2x higher than other three tissues)	37	643	67	1,486
Not specific to any one tissue	2,269	540	22,016	939

Transcripts from transposable element (TE)-related sequences were detected at a higher level in gametophytes than in sporophytic tissues (Table 2), consistent with results in Arabidopsis [36]. This is despite filtering out a large number of repeat-matching reads prior to transcript assembly. To further assess this trend, gene models in both the FGS and the WGS annotated as “probable transposon” were evaluated for expression in different tissues. Analysis of RNA-Seq reads mapped to these gene models revealed that the gametophytes (particularly the embryo sac) are significantly more likely than the sporophytic tissues to express one of these (Table 3). A similar bias is found when focusing on gene models enriched in one tissue (defined as a two-fold signal increase over the other three tissues), with the set of ES-enriched genes overrepresented for “probable transposon” genes compared to the two sporophyte samples and the total gene set.

Table 3 Percentage of genes annotated as probable transposon genes expressed in gametophyte and sporophyte samples

	Probable Transposon Genes in the Filtered Gene Set Expressed Above 0.1 FPKM (1.2% of total (456/39,635))	Probable Transposon Genes in the Working Gene Set Expressed Above 0.1 FPKM (3.3% of total (3,692/109,324))
All Seedling Expressed Genes	0.9% (259/27,564)	1.3% (426/33,528)
All Pollen Expressed Genes	1.0% (143/14,591)	1.4% (251/17,314)
All Embryo Sac Expressed Genes	1.1% (308/28,489)	2.3% (964/42,672)^{s,p,o}
All Ovule (w/o Embryo Sac) Expressed Genes	1.0% (263/26,338)	1.6% (560/35,727)
Seedling enriched (2x higher than other three tissues)	0.7% (58/8,066)	0.8% (65/8,335)
Pollen enriched (2x higher than other three tissues)	1.3% (30/2,224)^{s}	2.3% (82/3,526)^{s,o}
Embryo Sac enriched (2x higher than other three tissues)	1.6% (83/5,011)^{s,o,t}	4.4% (315/7,097)^{s,p,o,t}
Ovule (w/o Embryo Sac) enriched (2x higher than other three tissues)	1.1% (19/1,751)	1.5% (85/5,475)
Dual Gametophyte enriched (Pollen and Embryo Sac each 2x higher than both sporophyte tissues)	2.0% (12/591)^{s,o,t}	2.3% (10/434)^{s}

^{s} = higher than equivalent seedling frequency at $p \leq 0.01$.
^{s} = higher than equivalent seedling frequency at $p \leq 0.05$.
^{p} = higher than equivalent pollen frequency at $p \leq 0.01$.
^{o} = higher than equivalent ovule frequency at $p \leq 0.01$.
^{o} = higher than equivalent ovule frequency at $p \leq 0.05$.
^{t} = higher than total in gene set at $p \leq 0.01$.
^{t} = higher than total gene set at $p \leq 0.05$.

Validation of RNA-Seq by quantitative RT-PCR

To verify the differential expression detected by the Illumina RNA-Seq data, quantitative RT-PCR (qRT-PCR) was performed on a new set of 3 replicates for all four tissue types. A set of 46 genes was chosen randomly, based on the availability of *Ds* insertion alleles (see below). These genes had a range of average expression levels for each tissue: 0.16 to 537 FPKM for seedling; 0 to 6635 FPKM for MP; 0.03 to 815 FPKM for Ov; and 0 to 417 FPKM for ES samples. One concern of the RNA-Seq analysis was that amplification of cDNA of the ES and Ov samples prior to Illumina library construction may have introduced biases in composition of the library. To test the potential for biases, cDNA was prepared for all four tissues using similar quantities of RNA as was used in the original ES and Ov samples. The cDNA from these samples was then amplified prior to quantitative RT-PCR. The qRT-PCR analysis from these samples was then compared to the RNA-Seq expression data for all four tissue types. To corroborate the expression levels measured by RNA-Seq, the ratio of expression levels between tissues using RNA-Seq was compared to the ratio of expression as measured by qRT-PCR. For all genes, expression of genes in seedling, Ov, and ES were measured relative to their expression in MP, since MP is the least complex tissue of the four samples on a cellular level. As can be seen from the R^2 values (0.83, 0.82 and 0.72), the ratios of gene expression measured between tissues by RNA-Seq and by qRT-PCR are highly correlated (Figure 3). The lowest correlation was seen with the ES samples, which had the least amount of starting material for RNA-Seq, suggesting that there is some loss of fidelity

with amplification from small amounts of starting RNA. However, these validation experiments support the reliability of the relative values provided by the RNA-Seq analysis.

Figure 3 Verification of RNA-Seq with quantitative RT-PCR. The $\log_2$ (expression relative to *actin* and *ubiquitin* in test tissue by qRT-PCR/expression relative to *actin* and *ubiquitin* in pollen by qRT-PCR) is plotted on the y-axis and the $\log_2$ (expression in test tissue by RNA-Seq(FPKM)/expression in pollen by RNA-Seq(FPKM)). Although the slopes of the lines are not equal to one, the R^2 values show good correlation between measurements using both types of data with the lowest correlation for the embryo sac samples. For RNA-Seq values expression values were from B73 Pollen, B73 Seedling, combined W23-B73 Embryo Sac, and combined W23-B73 Ovules-without-embryo Sacs. For RT-PCR expression values only W23 Embryo Sac and Ovules-without-embryo Sacs were used.

Comparison of gametophytic and sporophytic gene expression programs

Comparison of the lists of FGS gene models above an average expression threshold of 0.1 FPKM in each sample type (Additional file 4: Table S4) revealed a number of features (Figure 1G). The largest set, 12,062 genes, shows expression above threshold in all four tissue types. Seedling had the highest percentage of genes in its transcriptome above 0.1 FPKM that are not shared with any of the other samples at 8.4%, compared to ES samples at 6.0%, MP at 5.0%, and Ov with the lowest frequency of unique genes at 1.9%. The lower numbers of unique genes for Ov are not surprising given that the ES samples also contain small amounts of contaminant nucellus cells. Corroborating earlier studies on maize pollen mRNA diversity [42], and similar to Arabidopsis [33], the MP transcriptome is the least complex of the four with half as many of the FGS genes expressed above a threshold of 0.1 FPKM as the other tissues (Table 2). This is consistent with the view that MP is highly specialized compared to the other three tissue types, as 10,662 genes shared by the other three tissue types are not detected in MP (Figure 1G). Thus, a picture emerges of a relatively large core of genes expressed across all four developmental stages, with functional specialization potentially due to the combination of differences in expression level for this core set, plus developmentally-specific expression of a smaller set of genes.

To determine the similarity of the transcriptome of different tissues to one another, including between different inbred lines (B73 and W23), hierarchical clustering was used to compare the 18 replicates across 6 tissues and/or genotypes (Figure 4). The first analysis used all FGS genes, only excluding genes that were below threshold in all samples (Figure 4A). Due to the possibility that polymorphisms between W23 and B73 could lead to inaccurate measurement of expression of some W23 genes, a second comparison was also made. This second analysis (Figure 4B) excluded the ~6,000 genes for which no reads were detected in any of the six W23 (ES and Ov) samples. As in Arabidopsis [14], these analyses supported the view that the MP transcriptome is the most distinct, clustering away from the other samples regardless of which gene set was used.

Figure 4 Hierarchical clustering of replicates based on expression profiles using the R statistical package. Genes are organized vertically based on expression in W23 ES sample 1. **(A)** Clustering based on FGS gene expression FPKM except for genes with 0 FPKM in all 18 samples. **(B)** Same as in **(A)** except that genes having 0 FPKM in all 6 W23 ES and Ov samples but having reads above 0 in the B73 ES or Ov samples were also omitted to remove possible artifacts caused by read mapping difficulties.

Relationship between gene expression pattern and duplicate gene retention

The maize genome consists of two subgenomes as a consequence of an ancient allotetraploidy event, and thus genes in the modern genome can be classified as either singletons (if the corresponding homeolog has been lost since tetraploidization), or duplicates (if both genes have been retained) [11]. Subgenome 2 is characterized by higher gene loss and lower gene expression of retained genes than subgenome 1. To determine if expression in the gametophytes is distributed differently in the two subgenomes, two sets of gene lists were developed for each sample type from B73. The first set (the total transcriptomes) included all FGS gene models above a threshold of 0.1 FPKM in that sample type; the second set (the tissue-enriched transcriptomes) included only those gene models from the total transcriptomes that were at least 2-fold higher in a sample type relative to all three other sample types (see Materials and Methods). These gene model lists were then mapped to high-confidence subgenome 1 and 2 sets [11] (Table 4). As expected, in all four tissues the percentage of genes expressed above the threshold is higher for subgenome 1 than subgenome 2. However, for both the total MP transcriptome and for the MP-enriched gene list, the percentage of genes in subgenome 2 is significantly higher than it is for the total gene list, or for the other tissue-focused gene lists. None of the other tissue transcriptomes show overrepresentation of subgenome 2 compared to the whole genome. A breakdown of how the tissue-enriched genes are distributed in the subgenomes (as either singletons, or part of a retained duplicate pair) reveals the basis for this difference (Table 5 and Additional file 5: Table S5). Relative to the other expression categories, and to the entire FGS set, MP-enriched genes are more likely to be duplicates, retained in both subgenomes. Furthermore, the MP-enriched set has a significantly lower distribution of subgenome 1 singletons, supporting the idea that this set of genes is less likely to have lost subgenome 2 homeologs. Finally, when focusing on the retained duplicate pairs in the four tissue-enriched gene sets, the MP set has a significantly greater proportion of pairs in which both members are represented; in fact, one-quarter of the gene models in the MP-enriched set that could be assessed via subgenome mapping are a member of an expressed pair. These data are consistent with the idea that pollen places some exceptional requirement on gene function, such that selection pressure results in retention and expression in pollen of a higher proportion of both genes of a duplicate pair.

Table 4 Expression of subgenome 1 and subgenome 2 assigned genes in gametophyte and sporophyte samples

	Subgenome 2 to Subgenome 1 Ratio (63.1% for all subgenome2/all subgenome1 (7,118/11,282))
All Seedling Expressed Genes	62.6% (6,047/9,657)
All Pollen Expressed Genes	67.5% (3,421/5,066)[s,e,o,t]
All Embryo Sac Expressed Genes	63.3% (5,820/9,195)
All Ovule (w/o Embryo Sac) Expressed Genes	63.6% (5,541/8,716)
Seedling enriched (2x higher than other three tissues)	57.3% (1,749/3,054)
Pollen enriched (2x higher than other three tissues)	73.0% (465/637)[s,t]
Embryo Sac enriched (2x higher than other three tissues)	63.4% (645/1,017)
Ovule (w/o Embryo Sac) enriched (2x higher than other three tissues)	62.7% (207/330)

[s] = higher than equivalent seedling frequency at $p \leq 0.01$.
[e] = higher than equivalent embryo sac frequency at $p \leq 0.05$.
[o] = higher than equivalent ovule frequency at $p \leq 0.05$.
[t] = higher than total gene set at $p \leq 0.05$.

Table 5 Tissue-enriched gene models mapped to singletons and duplicates in the maize subgenomes

	Singleton – Subgenome 1	**Singleton – Subgenome 2**	**Duplicate – Subgenome 1**	**Duplicate – Subgenome 2**	**Total # Mapped to Subgenomes**	**# of Duplicate Pairs, both represented in the enriched set (Percent of Set)**
Seedling enriched (2x higher than other three tissues)	41.0% (1369)	16.9% (564)	21.3% (712)	20.8% (693)	3338	338 (20.3%)
Pollen enriched (2x higher than other three tissues)	32.6% (259)	17.9% (142)	23.1% (184)	26.4% (210)	795	99 (24.9%)
Embryo Sac enriched (2x higher than other three tissues)	39.4% (443)	17.6% (198)	21.9% (246)	21.2% (238)	1125	55 (9.8%)
Ovule (w/o Embryo Sac) enriched (2x higher than other three tissues)	43.4% (155)	17.4% (62)	20.7% (74)	18.5% (66)	357	8 (4.5%)
Filtered Gene Set	37.5% (4860)	16.7% (2161)	23.0% (2982)	22.7% (2945)	12948	

For the 4x4 categorical comparison of the expression sets vs. the subgenome mapping characters, the chi-square value is 25.88, for $p < 0.005$, indicating a significant difference in the distributions, due to the pollen set. No significant difference is present comparing only seedling, embryo sac and ovule sets (χ^2 2.61; $p > 0.5$).

Gene ontology functional category enrichment

Functional enrichment of Gene Ontology terms was performed for lists of genes with particular expression patterns using the online Agrigo GO Analysis Toolkit [43] using the modified average expression values of the FGS genes. GO term overrepresentation was performed for the full transcriptome above 0.1 FPKM for each tissue, and for the tissue-enriched gene lists (Additional file 6: Table S6) (see Materials and Methods for description).

Comparison of the GO terms overrepresented in the full transcriptome of each of the four tissue samples revealed that the MP samples had the most GO terms (114) that were not shared with the other samples ((Additional file 7: Table S7 and Additional file 8: Figure S1). The largest number of overrepresented GO terms were in the ES and Ov samples, but many of these were shared, with GO terms unique to the ES in large part related to the DEFENSIN/LURE (DEFL) family (see below). By far, the largest group of GO categories overrepresented in each of the four full transcriptome gene lists were shared by all four samples (212 GO terms), followed by the number shared by ES, Ov, and seedling (90 GO terms). Thus, the overall analysis suggests that the distinctiveness of the pollen transcriptome is extended to the functional level.

Second, analysis of overrepresented GO terms was performed for tissue-enriched gene lists to identify potential tissue-specific functions (Additional file 9: Table S8). A dual gametophyte-enriched gene set (enriched in both MP and ES, relative to both Ov and Seedling) was also identified. The Ov sample had the fewest overrepresented GO terms of all four tissue-enriched gene lists, with apoptosis related terms being significantly increased (Additional file 9: Table S8D). In Seedling genes, GO categories related to photosynthetic functions and environmental responses were overrepresented (Additional file 6: Table S7A and Additional file 9: Table S8A).

For the MP-enriched genes, the most significantly overrepresented GO categories include functions related to the actin cytoskeleton, and GO terms potentially related to pollen tube growth and penetration of the pistil (e.g., the cell wall-loosening expansins; pectinesterases and glycosidases) (Additional file 9: Table S8B). Additionally, there is significant overrepresentation of post-translational protein modification, driven in large part by an abundance of Protein Kinases. A few members of the DEFL family are also specifically overrepresented in the MP transcriptome. MP-enriched genes in subgenome 2 (Tables 4 and 5), were also examined for overrepresented GO terms (Additional file 10: Table S9). In this subset of subgenome 2 genes, functions related to localization and transmembrane transport, as well as pectinesterase activity, were the most significantly overrepresented GO terms.

For the ES, biological processes and molecular functions related to transcriptional regulation were the most highly overrepresented (Additional file 9: Table S8C). Interestingly, as in the MP transcriptome, expansin gene expression is significantly overrepresented in the ES transcriptome, although a different set of expansins from those found in MP. These genes may facilitate the rapid expansion of the embryo sac within the surrounding nucellus. The other most significant GO terms in the ES-enriched genes include nucleotide metabolic processes. Enrichment of this category is entirely driven by the high number of ES-enriched members of the DEFENSIN/LURE (DEFL) family, as these small proteins contain a Knottin fold with a dinucleoside diphosphate kinase core. The shared gametophyte-enriched gene set shows similar GO category overrepresentation, again driven by DEFL proteins. Thus, in total, three different sets of DEFL genes were found, each overrepresented among the transcripts showing that expression character: ES-enriched; MP-enriched; and dual gametophyte-enriched. Some members of this family have previously been shown to be expressed in synergids in maize and to function as pollen tube attractants in Torenia [28,44-47].

Analysis of transcription factor gene families

Because transcriptional regulation GO terms were significantly overrepresented in the ES-enriched gene list, all known transcription factors in maize from the Grass Transcription

Factor Database [48,49] were assayed for tissue-enriched expression in the gametophyte and sporophyte tissues (Additional file 11: Table S10). Using this more comprehensive TF list, significant overrepresentation for the aggregate of all transcription factor families was detected in the embryo sac. Five separate TF gene families showed significant overrepresentation ($p < 0.05$) in the ES-enriched gene list, in order of significance: AP2-EREB, WRKY, MYB-RELATED, NAC, and MADS-box; and no TF families were significantly under-represented in the embryo sac. This contrasts with MP, seedling, and Ov, where there was a global underrepresentation of TFs. In the MP-enriched gene list, only Orphan TF genes and MADS-box genes (including a previously-identified MADS box gene specific to pollen [50]) are overrepresented. Neither seedling nor Ov had any TF gene families overrepresented by the criteria used.

Because the MADS gene family appeared in both MP and ES gene lists, it was analyzed in greater detail (Additional file 12: Figure S2). Ten MADS genes are in the MP-enriched gene set, and four are present in the dual gametophyte-enriched gene set (although three of these four are significantly higher in MP than ES and so are also present in the MP-enriched set). Enrichment for different MADS family members in the ES and the MP is reminiscent of the distribution of ES-specific and MP-specific MADS genes in Arabidopsis [17,27,51-54]. In Arabidopsis, MIKC* MADS genes are overrepresented in the MP [13,15,55], whereas the Type I Class α and ß genes are overrepresented in the ES [53]. In maize, all MIKC* MADS genes are enriched in MP, supporting the conclusion for an ancient role for these genes in the male gametophyte [56]. Other MP-enriched maize MADS genes fall into the MIKC and the Type 1 Class α groups. Maize ES-enriched genes fall into the MIKC and the Type 1 Class α and γ groups, which is somewhat distinct from the pattern in Arabidopsis. There are no clear Type 1 Class ß MADS genes in maize, just as there are none reported in rice [57].

The phylogenetic relationships between ES-enriched genes were also determined for the other TF families overrepresented in the embryo sac. The NAC gene family was particularly striking, with 25 of 25 genes in one clade being ES-enriched and only one of the 109 genes in the other clade being ES-enriched (Additional file 13: Figure S3). For the AP2-EREB, WRKY and MYBR families, ES-enriched genes were broadly distributed across most clades, although differences between subgroups exist (e.g., local over-representation of a few closely related genes) (Additional file 14: Figure S4, Additional file 15: Figure S5, Additional file 16: Figure S6). For many of these sub-family enrichments of TFs, the shared ES expression patterns are associated with syntenic regions, rather than tandem duplications and may reflect an ancestral embryo sac function for these branches of the gene family.

Analysis of small peptide gene family expression

The expression pattern of small peptide gene families was investigated in greater detail based on three reasons: (1) GO analysis highlighted small peptide DEFL genes as overrepresented in all three gametophyte-enriched gene sets; (2) shorter transcripts were more prevalent in the ES transcriptome compared to the other three tissues (data not shown); and (3) probes for small peptide genes were often omitted from earlier microarray studies. Characterization focused on two families with known gametophyte members: DEFENSIN/LURE (DEFL) [44], and Zm Egg Apparatus1 (ZMEA1)-LIKE (EAL) [58-61]; and two families that had not previously been shown to have gametophyte-expressed members: CLAVATA3-ESR (CLE) [62], and LITTLE ZIPPER (ZPR) [63] (Figures 5, 6, Additional file 17: Figure S7, and Additional file 18: Figure S8).

Figure 5 Phylogeny and expression of maize DEFL genes. Gene names in blue are part of the MP-enriched gene set. Gene names in red are part of the ES-enriched gene set. Gene names in magenta are part of the dual gametophyte-enriched gene set. Expression levels are indicated by color of the letter of each sample type with red meaning >10 FPKM, orange between 1 and 10 FPKM, green between 0.1 and 1 FPKM, blue greater than zero but less than 0.1 FPKM, and black having 0 reads. E = embryo sac expression; O = ovule without embryo sac expression; S = seedling expression; P = mature pollen expression. Torenia LURES are included for reference. Posterior probability values are given at node positions.

Four maize DEFL genes (*ZmES1*, *2*, *3*, and *4*) had previously been identified in the A188 inbred line, and characterized as embryo sac-expressed [28]. In the B73 genome these four genes correspond to three tandemly duplicated genes, which we have termed *ZmES1*, *ZmES3*, and *ZmES2/4*. One likely explanation for the discrepancy is that an additional duplication exists in A188. Using BLAST to identify similar genes in the B73 genome identified 39 DEFL gene models in the B73 v5a Working Gene Set, a larger DEFL family than in the AgriGO database. Expression analysis shows clear bias for expression of these genes in the embryo sac. Twenty of these are expressed above 1 FPKM in ES, compared to five above 1 FPKM in MP, six in seedling, and four in Ov. In all, 23 of the 39 DEFL genes have tissue-enriched expression in one or both of the gametophytes. The strong embryo sac enrichment for DEFL gene expression contains genes in three clades within the DEFL family, one group including *ZmES1* through *ZmES2/4*, one clade with all members in either the ES-enriched or dual gametophyte-enriched gene set, and a third clade more divergent from the rest of the DEFL genes, including the endosperm-expressed *ESR6* gene (Figure 5). Many of the ES-enriched DEFL family members are found in tandem clusters of recently duplicated family members, as exemplified by the *ZmES1-3-2/4* cluster. The relationships within these clusters are more robust than those between the less recently diverged groups. Of the 19 DEFL genes in the ES, ten are also dual gametophyte-enriched, although the level of expression in the MP is consistently lower than in the embryo sac.

The *EAL* family was founded by the maize embryo sac-specific gene *Zm Egg Apparatus1* (corresponding to GRMZM2G456746) that functions in the embryo sac as a pollen tube attractant [58]. This family is characterized by an EA1 box near the C-terminus [59]. Three additional small peptide *EA1 like* (*EAL*) genes have been described: *ZMEAL1* (transcript maps upstream of and includes GRMZM2G576769) [64], *ZMEAL2* (GRMZM2G157505) and GRMZM2G180950. BLAST querying for other small peptide *ZmEA1* homologs in the B73 genome identified six additional genes (Figure 6). *ZMEAL1* is expressed in the embryo sac and required for normal antipodal cell development [64]. Like the *DEFL* family, *EAL* genes also show family-wide enrichment in the embryo sac, with eight above 1 FPKM in ES, none in MP, two in seedling, one in Ov, and one with expression below 1 FPKM in all tissues tested. *ZMEA1*, *EAL1*, and *EAL2* are part of a cluster of four tandemly duplicated genes on chromosome 7 with the fourth gene adjacent to and nearly identical to *EAL1* but with much lower expression. All members of this cluster are preferentially expressed in the embryo sac, albeit at different levels. A second clade including four tandemly duplicated *EAL* genes located on chromosome 8 also has every member in the ES-enriched gene set.

Figure 6 Phylogeny and expression of maize EAL genes. Gene names in blue are part of the MP-enriched gene set. Gene names in red are part of the ES-enriched gene set. Gene names in magenta are part of the dual gametophyte-enriched gene set. Expression levels are indicated by color of the letter of each sample type with red meaning >10 FPKM, orange between 1 and 10 FPKM, green between 0.1 and 1 FPKM, blue greater than zero but less than 0.1 FPKM, and black having 0 reads. E = embryo sac expression; O = ovule without embryo sac expression; S = seedling expression; P = mature pollen expression. Rice genes from Krohn et al. [64] are included for comparison. Posterior probability values are given at node positions.

In contrast, the CLE and ZPR families do not show family-wide enrichment for embryo sac expression. Twenty-six and eight genes were identified for the CLE and ZPR families, respectively (Additional file 17: Figure S7 and Additional file 18: Figure S8). For the CLE family there were nine above 1 FPKM in ES, two in MP, seven in the seedling, and four in Ov. The CLE family was almost completely absent in MP, with 24 of the 26 members having no reads in the MP. For the ZPR family there were two above 1 FPKM in ES, one in MP, three in the seedling, and one in Ov. Some of the ZPR family members are characterized by low expression in the ovule and no expression in the ES, suggesting they are expressed in portions of the ovule excluded from the ES samples (e.g., integuments). The expression of small peptides in the gametophytes is therefore not a general phenomenon; rather the DEFL and EAL families are likely enriched in these tissues for critical roles in gametophyte biology.

Test of gametophyte expressed genes for gametophyte function

Genes with expression in the gametophytes should be enriched for genes with gametophyte-critical functions. Such a function can be confirmed by observing reduced transmission of a mutation in that gene through the relevant gametophyte to the next generation. This reduced transmission is also predicted to result in reduced recovery of mutations in gametophyte essential genes. Thus, there should be a bias against recovering mutations in the sets of MP-enriched and ES-enriched genes compared to sporophyte-enriched genes identified in this study. The large collections of sequence-indexed transposon insertions for both the *Mutator* (*UniformMu* and the Photosynthetic Mutant Library [65-67]) and *Ac*/*Ds* [68] systems available in maize allowed a test of this prediction.

A baseline for transposon insertion rates in these collections was generated by assessing the frequencies for insertions into particular regions of the gene models of the Filtered Gene Set (FGS) and the Rejected Gene Set (RGS). The regions were assessed separately given the known bias in certain transposon insertion patterns (e.g., *Mu* is targeted near transcription start sites [69]), and the presumed likelihood of affecting gene function (e.g., exons in coding sequence vs. introns). As expected, the FGS was associated with significantly higher rates of transposable element insertion than the RGS (which is also associated with significantly higher methylation [70]) in the insertion collections assessed (Additional file 19: Table S11). The higher methylation may be associated with a relative decrease in accessibility for these sequences, and thus a decrease in transposon insertion rates [69]. Notably, a set of FGS gene sequences identified as containing TE/Repeat sequences (see Materials & Methods) was also associated with a bias toward fewer insertions relative to the non-TE FGS gene models (Additional file 19: Table S11). Therefore, these TE/Repeat-related gene models in the FGS were left out of further analyses of insertion frequency.

The frequency of associated *Mu* and *Ac/Ds* insertions was then calculated for the seedling, mature pollen, and embryo sac sets of tissue-enriched gene models (Tables 6 and Additional file 20: Table S12). The *UniformMu* population is the largest currently available, with 41,543 flanking sequence locations (April 2012, release 5); in addition, the propagation scheme for this population relies on self- and sib-pollination, imposing selection against both male and female gametophytic functions. Consistent with the predicted bias, in this population *Mu* insertions into the MP-enriched and ES-enriched gene sets were significantly less common, relative to the seedling gene set. The decreased prevalence of *Mu* insertions could not be explained solely by differences in gene size among the gene sets, as the bias remains detectable when normalized based on average size in bp for each region (Table 6). Although flanking sequence data for the Photosynthetic Mutant Library population (May 2013) is only approximately one-fourth that available for *UniformMu*, a similar, significant decrease in *Mu* insertions for the MP-enriched and ES-enriched sets is also discernible in this population (Additional file 20: Table S12). Notably, the deficit appears to be strongest in the MP-enriched gene set for insertions in exons in both populations, consistent with an effect associated with gene function. For the ES-enriched gene set, the strongest decreases appear to be in introns and the proximal promoter, in addition to exons, suggesting that factors in addition to gene function play a role in influencing insertion likelihood.

Table 6 Reduced frequency of insertion mutants in gametophyte vs seedling enriched genes

Expression characteristic	**Total gene models tested**	**Percent of gene models tested with a coding sequence *Ac/Ds* insertion**	**Percent of gene models tested with a coding sequence *Mu* insertion**
Seedling enriched genes	7,385	1.6% (0.70%)	19.9% (9.3%)
Pollen enriched genes	2,042	1.9% (0.90%)	12.7% (6.1%)*
Embryo sac enriched genes	4,238	0.8% (0.43%)*	13.2% (7.1%)*

Percentages in parentheses show the frequency of insertion per gene normalized for gene size. *Significantly lower than frequency of insertions in seedling enriched genes, $p < 0.01$.

The smaller number of available mapped *Ac/Ds* insertion locations limits the power to detect bias, but the largest population available (*Ds* Mutagenesis, 1,969 flanking sequence locations) is a useful comparison to the *Mu* populations as new insertions are selected and propagated solely through the female. Therefore, male-specific gametophytic insertions should not be selected against in this population, in contrast to insertions in female gametophyte genes. Consistent with this prediction, a significant bias against *Ac/Ds* insertions is found associated with the ES-enriched gene set in exons, introns and the 3′ end of predicted transcripts. Further, no significant difference in insertion bias was found between the seedling- and MP-enriched gene sets. However, the MP-enriched gene set is approximately half the size of the ES-enriched gene set, raising the possibility that the limited size prevents a robust assessment of any differences.

To address gametophyte function among these gene sets more directly, we also examined a set of 27 *Ds* insertions from the *Ds* Mutagenesis population in genes with a range of expression levels to see if the expression pattern would predict whether or not they would have transmission defects (Table 7 and Additional file 21: Table S13). Heterozygous plants

carrying the mutations were crossed reciprocally with homozygous wild type and their progeny tested for the presence of the *Ds* insertion using PCR. Transmission of the *Ds* insertion was called as reduced if the frequency in progeny was significantly less than 50% using a χ^2 test with a cutoff of $p < 0.05$. Nine of the genes had highest expression in the MP (eight of these were in the MP-enriched set), eight of the test genes had highest expression in the embryo sac (four of them in the ES-enriched set), and the remaining ten genes had their highest expression in one of the sporophyte samples (six in one of the sporophyte-enriched sets). Note that the analysis of transposon insertion patterns shows that this population (Tables 7 and Additional file 21: Table S13) is biased against recovery of *Ds* insertions in genes highly-expressed in the embryo sac; due to the propagation scheme for this population, mutations with strong female transmission defects are likely to be systematically excluded. All 27 mutations were tested for transmission as females; two of the eight *Ds* insertions in the genes with highest expression in the ES had slightly reduced transmission through the female, whereas none of the other 19 tested had reduced female transmission. Twenty-two were tested as males; of these 9 were in the MP-enriched list and 13 were not. Two of the nine mutations in MP-enriched genes had significantly reduced pollen transmission, whereas none of the other 13 did. Notably, the two mutations with reduced male transmission were in genes likely associated with cytoskeletal and signaling functions crucial for pollen: *profilin3* [71] and a potential calcium-binding (C2 domain) protein. The roles of the two genes associated with the female transmission defects, encoding a RING finger protein and a hypothetical protein, are less clear. Taken together, four of 17 tests of mutations in genes with highest expression in one of the gametophytes showed reduced transmission through that gametophyte, whereas none of the 32 tests without gametophyte enrichment of expression showed reduced transmission through that gametophyte, confirming that the probability of a gene being required for function in the gametophyte can be predicted on the basis of the relative expression between tissues.

Table 7 Transmission frequency of *Ds* insertions in genes with high and low gametophyte expression

	***Ds* insertions with reduced transmission**
Female transmission of *Ds* insertions in genes with highest expression in the Embryo Sac	2/8
Male transmission of *Ds* insertions in genes with highest expression in the Pollen	2/9
Transmission through the opposite gametophyte for genes with highest expression in the Embryo Sac or Pollen	0/16
Transmission through either gametophyte for genes with highest expression in one of the sporophyte tissues	0/16 (or 1/16*)

Reciprocal crosses were made between wild-type W22 plants and plants carrying *Ds* insertions in genes with varying expression patterns. *Ds* insertions that were recovered in fewer than 50% of the progeny ($p < 0.05$) were scored as having reduced transmission. See Additional file 21: Table S13 for supporting data.

*(One of these *Ds* lines is on the borderline of being significantly lower than 50% ($p = 0.0474$).

The frequency of *Ds* insertions with reduced transmission is significantly higher for genes with the highest expression in the gametophyte tested (4/17) than other genes (Fisher's Exact Test $p = 0.011$, for 0/32 non-gametophyte genes, and Fisher's Exact Test, $p = 0.043$, for 1/32 non-gametophyte genes).

Discussion

The ultimate function of the gametophyte is the production of viable offspring through the fusion of the male and female gametes. The process of double fertilization is unique to flowering plants and results in the formation of a diploid (1 maternal: 1 paternal) embryo and typically triploid (2 maternal: 1 paternal) endosperm. Similarities between the male and female gametophytes may result from conserved functions in gamete production or may have arisen from the inheritance of an ancestral condition of bisexual gametophytes found in many non-seed plants (*e.g.* Physcomitrella) [72]. However, the developmental patterns and cellular functions of the gametophytes are quite distinct. Identification of the genes active in the gametophyte generation provides a better understanding of their function, similarities, and uniqueness. To better understand the function of the maize gametophyte generation we have performed a full transcriptome analysis of mature male and female gametophytes using RNA-Seq.

Genome wide expression analysis reveals several implications for maize genome organization. Analysis of expression of genes annotated as transposon-related, as well as analysis of intergenic transcript models with similarity to repeat sequences, reveals that repetitive DNA elements are more likely to be expressed in both the male and female gametophytes than in sporophytic tissues. These data agree with results in Arabidopsis that gametophytes produce RNA from highly repetitive DNA elements [23,27,36]. Perhaps, as in Arabidopsis, in maize this is done as a means for silencing mobile elements in the germline, although the data here do not resolve in which cells these transcripts accumulate or are synthesized. Future experiments are necessary to determine if these transcripts are present in the gametes, whether or not they are transcribed in the gametes themselves, or if, as is the case in Arabidopsis pollen, they are transcribed in subsidiary cells (*i.e.* the antipodal cells and synergids of the female gametophyte and the vegetative cell of the pollen grain). Expression of repetitive elements is not identical between the male and female gametophytes with a greater likelihood for their expression in the female than in the male.

In Arabidopsis central cells, non-exonic transcripts, including known transposon and other intergenic transcripts, are more common than in other tissues – approximately 2- to 4-fold more non-exonic transcripts are in central cells than in seedlings or immature floral buds [34,73,74] – raising the possibility that transcriptional activity in 'intergenic' regions is a common feature of angiosperm gametophytes. Arabidopsis pollen also has a high frequency of intron reads [33] as well as expression of TEs [36]. In Arabidopsis, like maize, the majority of the predicted intergenic transcripts in the gametophytes are less than 500 bp [33]. However, the majority of the non-exonic Arabidopsis central cell reads were intronic, suggesting that this is driven in large part by incomplete annotation [34]. In contrast, in maize ≥90% of these reads are intergenic, suggesting that both incomplete annotation and TE transcripts are responsible for the exceptional ES transcriptome. Consequently, true intergenic transcriptional activity may vary between species. The higher expression of transposons and other intergenic sequences in maize embryo sacs may reflect either a higher activity of maize transposons than of those in Arabidopsis or the difference between sampling the whole embryo sac in maize vs. the central cell in Arabidopsis. Cell-specific analysis of these transcripts in maize is needed to resolve whether it is one of these two alternatives or a combination of the two. Two classes of transposons are also expressed in rice ovules but it is not known if these are in the embryo sac or the surrounding ovule tissue [75].

Like the pattern of TE transcripts, intergenic, non-repeat transcripts are more common in ES samples than other tissues. Potential novel genes were defined as gene models assembled directly from the RNA-Seq data that lacked homology to known transposable elements and other repeats. More potential protein-coding novel genes were identified in ES-enriched and MP-enriched gene sets than in sporophyte-enriched sets, with the greatest number present in the embryo sac. The relative inaccessibility of this tissue may have caused embryo-sac-specific transcripts to be underrepresented in the expression data used to help build maize gene models, and thus be omitted from annotated gene sets. The high number of gametophyte transcripts intergenic to the WGS may be an additional consequence of the genome-wide relaxation of silencing of repetitive elements (and sequences adjacent to repetitive elements) in the gametophytes compared to the sporophyte. RNA-Seq transcript assembly, including the samples in this study, identified long non-coding RNA (lncRNA) genes in the maize genome, and many of these were also found to be intergenic to WGS gene models [38]. Interestingly, reproductive tissues, including pollen and embryo sac, had more examples of lncRNA expression than any other tissues characterized.

The pollen transcriptome is also notable for its unusual representation in the two subgenomes of maize. Maize consists of two subgenomes from an ancient allotetraploidy event, with subgenome 2 characterized by reduced expression and reduced gene retention rates relative to subgenome 1 [11]. However, relative to the other three tissues assessed (which conform to expectations), pollen is associated with a significantly greater proportion of expression associated with genes of subgenome 2. This increase in subgenome 2 expression is not due to over-representation of pollen singleton genes in subgenome 2 (i.e., genes for which the corresponding subgenome 1 duplicate has been lost over evolutionary time), but rather due to a retention of more duplicate pairs (i.e., both subgenome 1 and 2 genes are retained in the genome) and correspondingly fewer pollen singleton genes in subgenome 1. Moreover, both members of a duplicate pair are more likely to be in the MP-enriched transcriptome than duplicates are to be in the other three tissues, consistent with the idea that expression of both plays a functional role in pollen. Thus, selection could be acting to maintain functional copies of both members of pollen-expressed genes following tetraploidization.

The gene balance hypothesis, which emphasizes that the expression dosage of genes encoding members of multi-subunit complexes, components of signal transduction pathways, or transcription factors needs to be maintained for correct function, has been invoked as an explanation for the retention of duplicates in genomes [76,77]. In one view, this balance would be even more critical in the male gametophyte, and therefore may result in a greater proportion of duplicate retention. First, the male gametophyte is haploid, so loss of one gene copy via mutation after tetraploidization reduces expression by half in the first generation, rather than by one-quarter, as would occur in the diploid. Second, differentiating it from the female gametophyte (which did not show such preferential retention), in an outcrossing species such as maize, pollen and the pollen tube are potentially under more stringent selection than other phases of the life cycle, via intense competition as a haploid for efficient pollen tube germination, tip growth and fertilization processes. Consistent with this idea, pollen-specific genes in an outcrossing relative of Arabidopis (*Capsella grandiflora*) are associated with stronger purifying selection and greater proportion of adaptive substitutions than sporophyte-specific genes [78]. In this interpretation of gene balance, one would expect to see a larger percentage of pollen-critical genes to be retained as duplicates in maize, and furthermore, that mutation of either copy should result in a deleterious phenotype. At least one such example has already been described, the *rop2*/*rop9* duplicate pair, although the deleterious effect of *rop2* mutation is only revealed when competing with wild-type pollen

[79]. This interpretation thus predicts that the MP-enriched duplicate genes identified herein are more likely to be associated with such competitive defects. Consequently, it also suggests that the overrepresented GO category processes identified in the MP-enriched subgenome 2 set (localization, transmembrane transport, and pectinesterase activity) are more likely subject to such dosage sensitivity.

Analysis of Gene Ontology categories confirms previous results (e.g., [10-13]) that regulation of a dynamic cytoskeleton is an important aspect of pollen biology. Additionally, post-translational modification is also overrepresented in the pollen transcriptome. Protein modification, *e. g.* protein phosphorylation, may facilitate the rapid growth reorientations in response to local cues necessary for pollen tube function. In the ES-enriched gene set, regulation of transcription and small peptide DEFLs were overrepresented. Because of the presence of the DEFL gene family in the embryo sac transcriptome additional small peptide gene families were also analyzed, since they were mostly not included in the GO term analysis. A second family of small signaling peptides, the EAL family, is also overrepresented in the ES transcriptome. Some members of both of these families have previously been shown to have female gametophyte expression [28,58], and to be involved in cell identity [64] and species-specific interactions with the pollen tube [47,60,61]. Here we have expanded the analysis of these gene families and shown that many members are enriched in the female gametophyte transcriptome. Certain DEFL genes show enriched expression in Arabidopsis central cells [34], suggesting that at least some of the DEFL enrichment reported here is associated with the central cell of maize. Correlations of gametophyte expression with phylogenetic relationships, including their location in tandem arrays, suggests that female gametophyte expression is an ancestral feature of some branches of both the DEFL and EAL gene families. The DEFL family also has members enriched in both male and female gametophyte transcriptomes. Mirrored expression of these small peptides in the two gametophytes may indicate a mechanism for reciprocal signaling between them. Shared and reciprocal signaling pathways of the male and female gametophyte will be easier to identify and resolve once it is known how cells perceive and respond to these small peptides.

Enrichment for transcriptional regulation in the embryo sac transcriptome was concentrated in five gene families: MADS, NAC, AP2/EREB, MYB-R, and WRKY. The MADS box gene family is also over-represented in Arabidopsis gametophyte transcriptomes. Maize and Arabidopsis both show a prevalence of pollen expressed genes in the MIKC* family suggesting that pollen function for MIKC* genes may predate the split between monocots and eudicots. Both maize and Arabidopsis also have members of the Type 1 Class α MADS genes. However, while in Arabidopsis the Type1 Class ß genes are overrepresented in female gametophytes, this clade is absent in maize. In maize, these functions may be taken over by other MADS gene clades (*e.g.* the MIKC class, present in the ES-enriched gene set of maize, but not of Arabidopsis). The NAC, AP2, MYB-R, and WRKY gene families are also over-represented in the ES-enriched gene set. An overlapping set of transcription factor families are over-represented in the transcriptome of whole rice ovules, including not only the AP2/EREB and MADS families but also the ABI3, AP2, YABBY, C2H2, HSF, LFY, MYB, and ZfHD families [75]. Many of these differences likely arise from the inability to compare the embryo sac to its surrounding ovule tissue in the rice study, but the shared groups may reflect gametophyte functions in the ancestor of maize and rice.

Mapping expression patterns on a gene phylogeny assists in evolutionary analyses, as a shared expression pattern by multiple members of a clade provides a hypothesis for the

expression pattern of the common ancestor of that clade. The notable example of this is in the *NAC* transcription factor family. A large ES-enriched clade includes duplicate genes from the ancestral maize allotetraploidization, as well as from older expansions of this gene family. In other cases, conserved genes with shared female gametophyte expression are part of a tandem cluster of genes with high similarity, suggesting more recent family expansion. This is seen for clusters of genes in the *DEFL* and *EAL* gene families, in which most or all of the genes in the cluster are expressed in the embryo sac. In fact, based on the phylogenetic analyses, the enrichment for female gametophyte expression of these families is apparently largely driven by expansion through tandem duplication. In some cases these tandem arrays are present in multiple grass lineages, as suggested by the maize *EA1* and *EAL1* genes being less similar to each other than to their rice homologs, which are also present as tandem duplications. In support of this hypothesis, the only one of the three rice EA1/EAL1 genes in the cluster that was assayed by microarray hybridization was expressed in both the egg and synergids, supporting the model that gametophyte expression of these genes reflects shared ancestral gene regulation [80].

Analysis of mutants and mutant frequencies show that genes significantly enriched in the gametophyte transcriptomes are more likely to be required in the gametophyte than other genes. Mutant frequencies and transmission rates confirm that gametophyte-enriched expression is predictive of a requirement for gametophyte function without making additional accommodations for genetic redundancy. Consequently, the entire transcriptomic dataset is expected to prove useful for identification of candidates for gametophyte mutants, as well as for additional broader analysis of gametophyte functions.

Conclusions

The gametophyte transcriptomes, particularly that of the male gametophyte, are distinct from those of sporophytic tissues, in agreement with results in Arabidopsis [12-15]. Analysis of RNA-Seq data is useful for identifying previously unrecognized genes with gametophyte expression, particularly for the less accessible female. The male and female gametophyte transcriptomes are quite distinct from one another in the specific content of expressed genes, but some similarities in trends can be detected. Both gametophytes are more likely to express transposons/repetitive DNA than the sporophytic tissues examined, a phenomenon that has been reported previously in the pollen grain in maize [23] and Arabidopsis [36]. Male and female gametophytes are also both enriched compared to sporophyte tissues for expression of MADS box transcription factors and small DEFL signaling peptides. Whether these shared patterns reflect conserved haploid generation functions or convergence of function is unclear. Reduced mutation frequency in gametophyte expressed genes also confirms the utility of these expression-based gene sets in identifying genes that are critical for gametophyte function and/or development. Comparison of retention rates for duplicate genes expressed in the pollen grain vs. other tissues suggests that pollen function and competitiveness are more sensitive to gene balance, affecting evolution of gene pairs after genome duplication events.

Materials and methods

Sample preparation and RNA isolation

Plants for RNA were grown under long day conditions in the greenhouse in Stanford, CA or in summer field conditions in Corvallis, OR. Samples were collected between 11:00 and

11:30 am. Fresh mature pollen was collected upon shedding. For mature female gametophytes, ovules were dissected from ears and subjected to cell wall digesting enzymes to facilitate isolation of embryo sac tissue. ES and Ov samples were paired (*i.e.* the ovule samples were produced from the tissue left over from embryo sac isolation). Three replicate RNA samples of each type were used to prime cDNA synthesis and amplification, with a slightly modified protocol for embryo sac and ovule tissue, due to the limited amount of starting material. We constructed libraries and produced paired-end and single-end sequence reads on the Illumina or SOLiD platform.

For isolation of ovule and embryo sac tissue whole ears were processed in the lab under a dissecting microscope. Ovules were isolated from ear florets with a silk length of ~10 cm by removing the silk and ovary wall with forceps and cutting the ovule at its base from the floret. Each ovule was immediately placed in a Petri dish in a cell wall enzyme digesting mix of 0.75% pectinase, 0.25% pectolyase, 0.5% cellulase, 0.5% hemicellulase buffered in 0.55 M mannitol pH 5.0 for one hour after collecting the last ovule at 24 ± 1.0 ° C before embryo sac isolation according to Kranz et al. [81] and Yang et al. [26]. Embryo sacs (with some attached nucellus cells) were mechanically extracted from ovules using dissecting needles. The embryo sac samples and remaining ovule tissue (now lacking an embryo sac) were placed in separate microfuge tubes containing 500 μL of 0.55 M Mannitol pH 5.0 until 15 to 20 embryo sacs and ovules lacking embryo sacs had been collected, and then samples were spun at 3000 rpm for 1 minute and excess Mannitol removed. Samples were homogenized in 400 μL of Trizol (Invitrogen) on a MixerMill300 (Qiagen) with a tungsten-carbide bead (Qiagen) at high speed for 3 minutes, and RNA extracted according to manufacturer's specifications to isolate total RNA. Mature, freshly shed pollen was collected from field-grown B73 plants, frozen in liquid Nitrogen, and RNA extracted with Trizol (Invitrogen) and purified from the aqueous phase using RNEasy MinElute columns (Qiagen). For whole seedling samples, all shoot tissue above the first leaf node was collected from 9-day old B73 plants on the same day in liquid Nitrogen, and RNA was isolated as for mature pollen.

cDNA libraries were generated from 0.5 to 20 μg total RNA. First strand cDNA was synthesized using the SMART PCR cDNA Synthesis Kit with SMART MMLV Reverse Transcriptase (Clontech Laboratories, Inc.) for pollen and seedlings or the SMARTer PCR cDNA Synthesis Kit with SMARTScribe Reverse Transcriptase (Clontech Laboratories, Inc.) for embryo sacs and ovules. The second strand was synthesized with the Advantage 2 PCR kit (Clontech Laboratories, Inc.). After second strand synthesis, cDNAs from the seedling and pollen samples (15–17 cycles), and the ovule and embryo sac samples (26 cycles), were amplified using the Advantage 2 PCR kit (Clontech) to produce sufficient cDNA for generating Illumina libraries. To identify embryo sac/ovule sample pairs that had no contaminating post-fertilization (endosperm) tissue in any samples and no contaminating embryo sac tissue in the ovule samples the resultant amplified cDNA was tested for presence and/or absence of several test genes for both B73 and W23 samples. PCR was performed with primers for the *Embryo Sac1* (*ES1*) (an embryo sac specific gene) [28], *Embryo surrounding region1* (*ESR1*) (an endosperm specific gene) [82], *EBE2* (a central cell and endosperm specific gene) [83], *ubiquitin* (a constitutive gene), and *knox6* (a constitutive gene) [84]. Samples with no detectable *ESR1* transcripts in the ES or Ov samples or detectable *ES1* or *EBE2* transcripts in the Ov samples were used for sequencing. The cDNA libraries from B73 inbred samples were prepared for Illumina sequencing using a nebulizer for fragmentation and the Illumina Paired-End Sequencing preparation kit per manufacturer's protocol (Illumina cat. # PE-102-1001 and cat # PE-102-1002). Illumina sequencing was performed at the Oregon State University Center for Genome Research and Biocomputing.

cDNA of the W23 samples was then used to prepare libraries and sequenced using the ABI SOLiD platform by Seqwright DNA Technology Services (Houston, USA). Following mapping of reads to the maize genome and generation of FPKM values for all maize genes, a final round of quality control of the embryo sac samples was performed. Because of the potential for variability introduced by the amplification of cDNA before sequencing and by variation in the amount of nucellar tissue left attached to the embryo sacs, a set of high confidence embryo sac-specific genes selected from the literature were examined in all the embryo sac samples to determine which were sufficiently robust for further analysis (Additional file 22: Table S14). These high confidence embryo sac-specific genes have been confirmed as embryo sac specific in the context of the ovule either by *in situ* hybridization or by transgenic reporter analysis.

Sequence analysis

80-mer paired-end reads were processed using the Illumina Genome Analysis Pipeline, version 1.5.0. TopHat, version 1.0.13, was used to align the RNA-Seq reads to the maize genome (version ZmB73_5a.59) following several preprocessing steps, which included primer trimming, quality control filtering and length sorting. Prior to aligning reads to the maize genome reads matching maize repetitive sequences were filtered using the list available from the maize TE database [39]. Reads were aligned in paired-end mode when both reads of a pair passed all preprocessing steps, otherwise reads were aligned as singles. Empirical transcripts (etranscripts) were assembled from aligned data using Cufflinks, version 0.8.1, and FPKM expression data were generated using Tophat and Cufflinks. All reads were then loaded into the gbrowse genome browser and a novel gbrowse plugin, QuantDisplay, was used to visualize the data. The sequence data are available at the Sequence Read Archive at NCBI, accession number SRP006965.

Given that TE databases have improved since the initial RefGen annotations were generated, BLAST Best Hits was used to identify additional TE-related gene models in the FGS, WGS, and empirical transcript models. All transcript model sequences were BLASTed against three different maize repeat databases (the MIPS Repeat Database, an updated version of the MTEC Transposable Element database [85], and the UTE database of unique TE sequences) [86]. The BLAST Best Hit (ranked by bit score) for each was used to define whether a particular transcript model included TE-related sequences using a previously validated threshold (minimum hit length 50 bp, minimum identity 85%, minimum bit score 50) [41]. Sequences in the empirical transcript set not recognized as TE-related by this set of parameters were further screened by the RepeatMasker tool [87,88], which also detects simple sequence repeats. Sequences with lengths greater than 20% repetitive, or with >240 Smith-Waterman match score, were also classified as TE/Repeat-related. Empirically-predicted transcripts which did not correspond to annotated gene models and also were not recognized as TE- or repeat-related were subsequently analyzed by the BLAST2GO tool, to assess their potential protein coding capacity, either via BLAST or via a scan of the InterPro collection of protein signature databases [89].

Quantitative RT-PCR analysis

Two control primer pairs were chosen, one each for *ubiquitin* and *actin* transcripts, each of which would target cDNA from multiple genes of each class to minimize effects of tissue-specific isoforms. All primer pairs were designed using Primer Select of the DNASTAR software package. Primer pairs were selected based on the following criteria; they had to: (1)

amplify within the last two or three exons to avoid potential problems from truncated, non-full-length cDNA; (2) span an intron to distinguish cDNA from genomic DNA amplicons; (3) have amplicons less than 150 bp to increase efficiency; (4) and have a Tm of 60.0 ± 1.5°C. Primer pairs were then tested for efficiency on a pool of cDNA from all 12 samples. Primer pairs were only selected for further analysis if they produced a single amplicon and had an efficiency between 1.8 and 2.0. This produced a set of 22 genes for verification by qRT-PCR (genes and primers in Additional file 23: Table S15).

Analysis of gene sets based on expression

For identifying tissue-expressed or tissue-specific lists, a five read minimum per replicate and 0.1 FPKM minimum average were used. For the Pollen and Seedling tissues the average of the three B73 replicates was used to calculate a tissue-specific expression level. For the ES and Ov samples a more complicated method was used to combine data from three B73 replicates and three W23 replicates for each tissue type. The rationale for combining the B73 and W23 ES lists was supported by the analysis of the high confidence embryo sac specific gene list that showed some genes had more robust comparisons with the B73 samples and others with the W23 samples (Additional file 22: Table S14). Fewer genes were detected above 0.1 FPKM (Additional file 2: Table S2) in the W23 samples than in the analogous B73 samples (24% fewer in ES, 21% in Ov). This is due in large part to the approximately 6,000 above-threshold gene models in the B73 FGS (some above 1,000 FPKM) that are associated with no reads in any of the W23 replicates. Many of the expression differences among these genes are likely caused by polymorphisms between W23 and B73 (indels or SNPs) that prevent mapping of the reads to the appropriate gene model; these polymorphisms may also include complete absence of these genes from W23. Presence/absence variation between maize inbreds can involve several thousand of sequences [90]. For all genes that had reads in the W23 samples, the average of the 6 samples (3 W23 and 3 B73) was used to produce an Embryo Sac or Ovule expression value. However, it is possible because of polymorphisms between W23 and B73 that some genes from the W23 samples would erroneously be assigned a FPKM value of zero because the reads do not match the reference B73 genome. For genes that had reads for either B73 Embryo Sac or B73 Ovule samples but zero reads in any of the six W23 samples, the B73 average was used instead of the average of the W23 and B73 samples together. This adjustment in the average FPKM was done for both tissue types with W23 and B73 samples: the ES and Ov samples. Consequently, the gene expression set for ES and Ov sample consisted of a hybrid of genes with an average over all six replicates and genes with an average over 3 B73 samples.

To identify tissue-enriched genes, pairwise comparisons between tissue expression levels were made for each tissue combination to identify genes with a 2-fold expression difference between tissues. Tissue-enriched genes for this study were defined as genes 2-fold higher in one tissue than all three other tissues and above a threshold of 0.1 FPKM. They were identified by determining the overlap between the gene lists of the three independent comparisons (*e.g.* W23-B73 ES to W23-B73 Ov plus W23-B73 ES to B73 Seedling plus W23-B73 ES to B73 MP) (Additional file 7: Table S7). To determine if there were any common functions in the gametophytes distinct from functions in the sporophyte we identified a common gametophyte-enriched gene set – genes that were 2-fold higher (threshold of 0.1FPKM) for both gametophytes vs. both sporophyte samples. First we identified the genes 2-fold higher in the ES vs. the Ov samples and 2-fold higher in the ES vs. the seedling samples. We similarly identified the genes 2-fold higher in the MP vs. the Ov and 2-fold higher in the MP vs. the seedling. Then the genes in common between these two

sets were identified as a potential core gametophyte-enriched gene set of 591 genes (dual gametophyte-enriched) (Additional file 7: Table S7).

Overlapping and tissue-exclusive gene sets were identified using the Venny online tool [91], and proportional Venn Diagrams were produced using the online tool from BioInfoRx [92]. Gene Ontology terms over-represented in the full transcriptomes of each tissue type and in the sets of tissue-enriched genes were identified using the online Agrigo GO Analysis Toolkit and Database for Agricultural Community [43,93], using the Maize ssp V5a gene ID settings.

To identify Transcription Factor gene families overrepresented in tissue-enriched gene lists, the fraction of each tissue-enriched (*i.e.* 2-fold higher than the other three tissues) gene list made up of each Transcription Factor (TF) family was compared to the expected value in the gene list based on the fraction of the Filtered Gene Set made up each TF family (Additional file 11: Table S10). Chi-square values were calculated for each comparison between the observed and the expected number of TF family members, and TF families with significant enrichment were confirmed using a Fisher Exact test for the families with fewer than 200 members. Only TF families with an expected number above four were assayed and families with a $p < 0.05$ were considered significantly different from background.

To identify small peptide genes present in the WGS gene set but not annotated as being in these families, the Working Gene Set Peptide database was queried using BLAST at MaizeSequence [94] starting with the published founding family members. Transcription Factor family lists were taken from the Grass Transcription Factor Database [48,49]. For all phylogenetic analyses, alignments were made using the ClustalW algorithm in MegAlign (DNASTAR). Phylogenies were produced from these alignments using MrBayes v3.2.0 using default settings for amino acid analysis [95]. Each analysis was performed for 100,000 generations or until the standard deviation of the split frequencies dropped below 0.05. The CLE, EAL, and ZPR gene families were each run for 100,000 generations; the DEFL family was run for 1,000,000 generations; the MADS family was run for 3,300,000 generations; the NAC family was run for 750,000 generations; the AP2-EREB family was run for 900,000 generations; the MYBR family was run for 950,000 generations; and the WRKY family was run for 350,000 generations. Phylogenetic trees were drawn from the MrBayes files using FigTree v1.4.0 [96].

For comparison of insertion frequencies in the tissue-enriched gene sets (seedling, pollen and embryo sac), datasets with the insertion locations for each of the three transposable element populations assessed were obtained from MaizeGDB [97] and imported into a Filemaker Pro database also containing the B73 Refgen v2 WGS and FGS feature locations (e.g. exons, introns, CDS). Each insertion location was subsequently mapped relative to these features in the WGS, and categorized based on this location (e.g., Promoter −500 to −1, CDS_Exons, CDS_Introns). The number of insertions in each category, and the average sizes in bp for each category, were then derived by cross-referencing the expression sets with this insertion database.

Abbreviations

DEFL: Defensin/Lure; EAL: Egg Apparatus1 Like; ES: Embryo Sac; FGS: Filtered Gene Set; FPKM: Fragments per Kilobase per Million reads; lncRNA: long noncoding RNA; MP: Mature Pollen; Ov: Ovule with the embryo sac removed; RGS: Rejected Gene Set; RNA-seq: RNA based next generation sequencing; S: Seedling shoot; TE: Transposable Element; TF: Transcription Factor; WGS: Working Gene Set

Competing interest

The authors declare that they have no competing interests.

Authors' contributions

MMSE, JEF, SAG, and EV conceived of the project idea, and MMSE and JEF supervised the project. MMSE performed the analysis of GO terms and small peptide genes and the phylogenetic analyses of gene families. JEF performed analysis of the mutation frequency and subgenome expression comparisons. AMC prepared and tested the RNA and cDNA from the embryo sac and ovule samples, and RAC and ZV prepared and tested the RNA and cDNA from the pollen and seedling samples. SAG performed transcriptome mapping to the maize genome and transcript assembly and statistical analysis. CTC performed qRT-PCR experiments. EUW and EV performed Ds insertion allele transmission analysis, and EV and MMSE performed analysis of transcription factor families. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank William Nelson and Bindu Joseph for help with analysis. We also thank M. Dasenko and C. Sullivan at the OSU Center for Genome Research and Biocomputing for assistance with DNA sequencing and computational support. This work was supported by the National Science Foundation Plant Genome Program Grant DBI-0701731.

References

1. Walbot V, Evans MMS: **Unique features of the plant life cycle and their consequences.** *Nat Rev Genet* 2003, **4:**369–379.

2. Evans MMS, Grossniklaus U: **The maize megagametophyte**. In *Handbook of Maize: Its Biology.* Edited by Bennetzen JL, Hake S. New York: Springer; 2009:79–104.

3. Drews GN, Yadegari R: **Development and function of the angiosperm female gametophyte.** *Annu Rev Genet* 2002, **36:**99–124.

4. Bedinger PA, Fowler JE: **The maize male gametophyte**. In *Handbook of Maize: Its Biology.* Edited by Bennetzen JL, Hake SC. New York: Springer; 2009:77. 57–77.

5. Feldmann KA, Coury DA, Christianson ML: **Exceptional segregation of a selectable marker (KanR) in *Arabidopsis* identifies genes important for gametophytic growth and development.** *Genetics* 1997, **147:**1411–1422.

6. Christensen CA, Subramanian S, Drews GN: **Identification of gametophytic mutations affecting female gametophyte development in Arabidopsis.** *Dev Biol* 1998, **202:**136–151.

7. Howden R, Park SK, Moore JM, Orme J, Grossniklaus U, Twell D: **Selection of T-DNA-tagged male and female gametophytic mutants by segregation distortion in Arabidopsis.** *Genetics* 1998, **149:**621–631.

8. Boavida LC, Shuai B, Yu HJ, Pagnussat GC, Sundaresan V, McCormick S: **A collection of Ds insertional mutants associated with defects in male gametophyte development and function in Arabidopsis thaliana.** *Genetics* 2009, **181:**1369–1385.

9. Pagnussat GC, Yu HJ, Ngo QA, Rajani S, Mayalagu S, Johnson CS, Capron A, Xie LF, Ye D, Sundaresan V: **Genetic and molecular identification of genes required for female gametophyte development and function in Arabidopsis.** *Development* 2005, **132:**603–614.

10. Gaut BS, Doebley JF: **DNA sequence evidence for the segmental allotetraploid origin of maize.** *Proc Natl Acad Sci U S A* 1997, **94:**6809–6814.

11. Schnable JC, Springer NM, Freeling M: **Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss.** *Proc Natl Acad Sci U S A* 2011, **108:**4069–4074.

12. Honys D, Twell D: **Comparative analysis of the Arabidopsis pollen transcriptome.** *Plant Physiol* 2003, **132:**640–652.

13. Honys D, Twell D: **Transcriptome analysis of haploid male gametophyte development in Arabidopsis.** *Genome Biol* 2004, **5:**R85.

14. Becker JD, Boavida LC, Carneiro J, Haury M, Feijo JA: **Transcriptional profiling of Arabidopsis tissues reveals the unique characteristics of the pollen transcriptome.** *Plant Physiol* 2003, **133:**713–725.

15. Pina C, Pinto F, Feijo JA, Becker JD: **Gene family analysis of the Arabidopsis pollen transcriptome reveals biological implications for cell growth, division control, and gene expression regulation.** *Plant Physiol* 2005, **138:**744–756.

16. Jones-Rhoades MW, Borevitz JO, Preuss D: **Genome-wide expression profiling of the Arabidopsis female gametophyte identifies families of small, secreted proteins.** *PLoS Genet* 2007, **3:**1848–1861.

17. Steffen JG, Kang IH, Macfarlane J, Drews GN: **Identification of genes expressed in the Arabidopsis female gametophyte.** *Plant J* 2007, **51:**281–292.

18. Kasahara RD, Portereiko MF, Sandaklie-Nikolova L, Rabiger DS, Drews GN: **MYB98 is required for pollen tube guidance and synergid cell differentiation in Arabidopsis.** *Plant Cell* 2005, **17:**2981–2992.

19. Johnston AJ, Meier P, Gheyselinck J, Wuest SE, Federer M, Schlagenhauf E, Becker JD, Grossniklaus U: **Genetic subtraction profiling identifies genes essential for Arabidopsis reproduction and reveals interaction between the female gametophyte and the maternal sporophyte.** *Genome Biol* 2007, **8:**R204.

20. Drews GN, Wang D, Steffen JG, Schumaker KS, Yadegari R: **Identification of genes expressed in the angiosperm female gametophyte.** *J Exp Bot* 2011, **62:**1593–1599.

21. Yu HJ, Hogan P, Sundaresan V: **Analysis of the female gametophyte transcriptome of Arabidopsis by comparative expression profiling.** *Plant Physiol* 2005, **139:**1853–1869.

22. Qin Y, Leydon AR, Manziello A, Pandey R, Mount D, Denic S, Vasic B, Johnson MA, Palanivelu R: **Penetration of the stigma and style elicits a novel transcriptome in pollen tubes, pointing to genes critical for growth in a pistil.** *PLoS Genet* 2009, **5:**e1000621.

23. Engel ML, Chaboud A, Dumas C, McCormick S: **Sperm cells of Zea mays have a complex complement of mRNAs.** *Plant J* 2003, **34:**697–707.

24. Borges F, Gomes G, Gardner R, Moreno N, McCormick S, Feijo JA, Becker JD: **Comparative transcriptomics of Arabidopsis sperm cells.** *Plant Physiol* 2008, **148:**1168–1181.

25. Wang Y, Zhang WZ, Song LF, Zou JJ, Su Z, Wu WH: **Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in Arabidopsis.** *Plant Physiol* 2008, **148:**1201–1211.

26. Yang H, Kaur N, Kiriakopolos S, McCormick S: **EST generation and analyses towards identifying female gametophyte-specific genes in Zea mays L.** *Planta* 2006, **224:**1004–1014.

27. Wuest SE, Vijverberg K, Schmidt A, Weiss M, Gheyselinck J, Lohr M, Wellmer F, Rahnenfuhrer J, von Mering C, Grossniklaus U: **Arabidopsis female gametophyte gene expression map reveals similarities between plant and animal gametes.** *Curr Biol* 2010, **20:**506–512.

28. Cordts S, Bantin J, Wittich PE, Kranz E, Lorz H, Dresselhaus T: **ZmES genes encode peptides with structural homology to defensins and are specifically expressed in the female gametophyte of maize.** *Plant J* 2001, **25:**103–114.

29. Le Q, Gutierrez-Marcos JF, Costa LM, Meyer S, Dickinson HG, Lorz H, Kranz E, Scholten S: **Construction and screening of subtracted cDNA libraries from limited populations of plant cells: a comparative analysis of gene expression between maize egg cells and central cells.** *Plant J* 2005, **44:**167–178.

30. Sprunck S, Baumann U, Edwards K, Langridge P, Dresselhaus T: **The transcript composition of egg cells changes significantly following fertilization in wheat (Triticum aestivum L.).** *Plant J* 2005, **41:**660–672.

31. Abiko M, Maeda H, Tamura K, Hara-Nishimura I, Okamoto T: **Gene expression profiles in rice gametes and zygotes: identification of gamete-enriched genes and up- or down-regulated genes in zygotes after fertilization.** *J Exp Bot* 2013, **64:**1927–1940.

32. Schmidt A, Schmid MW, Grossniklaus U: **Analysis of plant germline development by high-throughput RNA profiling: technical advances and new insights.** *Plant J* 2012, **70:**18–29.

33. Loraine AE, McCormick S, Estrada A, Patel K, Qin P: **RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing.** *Plant Physiol* 2013, **162:**1092–1109.

34. Schmid MW, Schmidt A, Klostermeier UC, Barann M, Rosenstiel P, Grossniklaus U: **A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing.** *PLoS One* 2012, **7:**e29685.

35. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU: **A gene expression map of Arabidopsis thaliana development.** *Nat Genet* 2005, **37:**501–506.

36. Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD, Feijo JA, Martienssen RA: **Epigenetic reprogramming and small RNA silencing of transposable elements in pollen.** *Cell* 2009, **136:**461–472.

37. Davidson RM, Hansey CN, Gowda M, Childs KL, Lin H, Vaillancourt B, Sekhon RS, de Leon N, Kaeppler SM, Jiang N, Buell CR: **Utility of RNA sequencing for analysis of maize reproductive transcriptomes.** *Plant Genome* 2011, **4:**191–203.

38. Li L, Eichten SR, Shimizu R, Petsch K, Yeh CT, Wu W, Chettoor AM, Givan SA, Cole RA, Fowler JE, Evans MMS, Scanlon MJ, Yu J, Schnable PS, Timmermans MC, Springer NM, Muehlbauer GJ: **Genome-wide discovery and characterization of maize long non-coding RNAs.** *Genome Biol* 2014, **15:**R40.

39. **Maize Transposable Element Database.** [http://maizetedb.org/~maize].

40. Hafidh S, Capkova V, Honys D: **Safe keeping the message: mRNP complexes tweaking after transcription.** *Adv Exp Med Biol* 2011, **722:**118–136.

41. Estep MC, DeBarry JD, Bennetzen JL: **The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution.** *Heredity (Edinb)* 2013, **110:**194–204.

42. Willing RP, Bashe D, Mascarenhas JP: **An analysis of the quantity and diversity of messenger RNAs from pollen and shoots of Zea mays.** *Theor Appl Genet* 1988, **75:**751–753.

43. **AgriGO GO Analysis Toolkit and Database for Agricultural Community.** [http://bioinfo.cau.edu.cn/agriGO/analysis.php].

44. Okuda S, Higashiyama T: **Pollen tube guidance by attractant molecules: LUREs.** *Cell Struct Funct* 2010, **35:**45–52.

45. Okuda S, Tsutsui H, Shiina K, Sprunck S, Takeuchi H, Yui R, Kasahara RD, Hamamura Y, Mizukami A, Susaki D, Kawano N, Sakakibara T, Namiki S, Itoh K, Otsuka K, Matsuzaki M, Nozaki H, Kuroiwa T, Nakano A, Kanaoka MM, Dresselhaus T, Sasaki N, Higashiyama T: **Defensin-like polypeptide LUREs are pollen tube attractants secreted from synergid cells.** *Nature* 2009, **458:**357–361.

46. Goto H, Okuda S, Mizukami A, Mori H, Sasaki N, Kurihara D, Higashiyama T: **Chemical visualization of an attractant peptide, LURE.** *Plant Cell Physiol* 2011, **52:**49–58.

47. Amien S, Kliwer I, Marton ML, Debener T, Geiger D, Becker D, Dresselhaus T: **Defensin-like ZmES4 mediates pollen tube burst in maize via opening of the potassium channel KZM1.** *PLoS Biol* 2010, **8:**e1000388.

48. **GRASSIUS Grass Regulatory Information Server.**
[http://grassius.org/tf_browsefamily.html?species=Maize].

49. Yilmaz A, Nishiyama MY Jr, Fuentes BG, Souza GM, Janies D, Gray J, Grotewold E: **GRASSIUS: a platform for comparative regulatory genomics across the grasses.** *Plant Physiol* 2009, **149:**171–180.

50. Heuer S, Lorz H, Dresselhaus T: **The MADS box gene *ZmMADS2* is specifically expressed in maize pollen and during maize pollen tube growth.** *Sex Plant Reprod* 2000, **13:**21–27.

51. Wang D, Zhang C, Hearn DJ, Kang IH, Punwani JA, Skaggs MI, Drews GN, Schumaker KS, Yadegari R: **Identification of transcription-factor genes expressed in the Arabidopsis female gametophyte.** *BMC Plant Biol* 2010, **10:**110.

52. Steffen JG, Kang IH, Portereiko MF, Lloyd A, Drews GN: **AGL61 interacts with AGL80 and is required for central cell development in Arabidopsis.** *Plant Physiol* 2008, **148:**259–268.

53. Bemer M, Heijmans K, Airoldi C, Davies B, Angenent GC: **An atlas of type I MADS box gene expression during female gametophyte and seed development in Arabidopsis.** *Plant Physiol* 2010, **154:**287–300.

54. Bemer M, Wolters-Arts M, Grossniklaus U, Angenent GC: **The MADS domain protein DIANA acts together with AGAMOUS-LIKE80 to specify the central cell in Arabidopsis ovules.** *Plant Cell* 2008, **20:**2088–2101.

55. Verelst W, Twell D, de Folter S, Immink R, Saedler H, Munster T: **MADS-complexes regulate transcriptome dynamics during pollen maturation.** *Genome Biol* 2007, **8:**R249.

56. Kwantes M, Liebsch D, Verelst W: **How MIKC* MADS-box genes originated and evidence for their conserved function throughout the evolution of vascular plant gametophytes.** *Mol Biol Evol* 2012, **29:**293–302.

57. Parenicova L, de Folter S, Kieffer M, Horner DS, Favalli C, Busscher J, Cook HE, Ingram RM, Kater MM, Davies B, Angenent GC, Colombo L: **Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world.** *Plant Cell* 2003, **15:**1538–1551.

58. Marton ML, Cordts S, Broadhvest J, Dresselhaus T: **Micropylar pollen tube guidance by egg apparatus 1 of maize.** *Science* 2005, **307:**573–576.

59. Gray-Mitsumune M, Matton DP: **The *Egg apparatus1* gene from maize is a member of a large gene family found in both monocots and dicots.** *Planta* 2006, **223:**618–625.

60. Marton ML, Fastner A, Uebler S, Dresselhaus T: **Overcoming hybridization barriers by the secretion of the maize pollen tube attractant ZmEA1 from Arabidopsis ovules.** *Curr Biol* 2012, **22:**1194–1198.

61. Uebler S, Dresselhaus T, Marton M: **Species-specific interaction of EA1 with the maize pollen tube apex.** *Plant Signal Behav* 2013, **8:**e25682.

62. Cock JM, McCormick S: **A large family of genes that share homology with CLAVATA3.** *Plant Physiol* 2001, **126:**939–942.

63. Wenkel S, Emery J, Hou BH, Evans MM, Barton MK: **A feedback regulatory module formed by LITTLE ZIPPER and HD-ZIPIII genes.** *Plant Cell* 2007, **19:**3379–3390.

64. Krohn NG, Lausser A, Juranic M, Dresselhaus T: **Egg cell signaling by the secreted peptide ZmEAL1 controls antipodal cell fate.** *Dev Cell* 2012, **23:**219–225.

65. Settles AM, Holding DR, Tan BC, Latshaw SP, Liu J, Suzuki M, Li L, O'Brien BA, Fajardo DS, Wroclawska E, Tseung CW, Lai J, Hunter CT 3rd, Avigne WT, Baier J, Messing J, Hannah LC, Koch KE, Becraft PW, Larkins BA, McCarty DR: **Sequence-indexed mutations in maize using the UniformMu transposon-tagging population.** *BMC Genomics* 2007, **8:**116.

66. Williams-Carrier R, Stiffler N, Belcher S, Kroeger T, Stern DB, Monde RA, Coalter R, Barkan A: **Use of Illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy Mutator lines of maize.** *Plant J* 2010, **63:**167–177.

67. McCarty DR, Suzuki M, Hunter C, Collins J, Avigne WT, Koch KE: **Genetic and molecular analyses of UniformMu transposon insertion lines.** *Methods Mol Biol* 2013, **1057:**157–166.

68. Vollbrecht E, Duvick J, Schares JP, Ahern KR, Deewatthanawong P, Xu L, Conrad LJ, Kikuchi K, Kubinec TA, Hall BD, Weeks R, Unger-Wallace E, Muszynski M, Brendel VP, Brutnell TP: **Genome-wide distribution of transposed Dissociation elements in maize.** *Plant Cell* 2010, **22:**1667–1685.

69. Liu S, Yeh CT, Ji T, Ying K, Wu H, Tang HM, Fu Y, Nettleton D, Schnable PS: **Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome.** *PLoS Genet* 2009, **5:**e1000733.

70. Eichten SR, Swanson-Wagner RA, Schnable JC, Waters AJ, Hermanson PJ, Liu S, Yeh CT, Jia Y, Gendler K, Freeling M, Schnable PS, Vaughn MW, Springer NM: **Heritable epigenetic variation among maize inbreds.** *PLoS Genet* 2011, **7:**e1002372.

71. Staiger CJ, Goodbody KC, Hussey PJ, Valenta R, Drobak BK, Lloyd CW: **The profilin multigene family of maize: differential expression of three isoforms.** *Plant J* 1993, **4:**631–641.

72. Cove D: **The Moss, *Physcomitrella patens*.** *J Plant Growth Regul* 2000, **19:**275–283.

73. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC: **Genome-wide mapping of alternative splicing in Arabidopsis thaliana.** *Genome Res* 2010, **20:**45–58.

74. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in Arabidopsis.** *Cell* 2008, **133:**523–536.

75. Kubo T, Fujita M, Takahashi H, Nakazono M, Tsutsumi N, Kurata N: **Transcriptome analysis of developing ovules in rice isolated by laser microdissection.** *Plant Cell Physiol* 2013, **54:**750–765.

76. Birchler JA, Veitia RA: **Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines.** *Proc Natl Acad Sci U S A* 2012, **109:**14746–14753.

77. Birchler JA, Veitia RA: **The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution.** *New Phytol* 2010, **186:**54–62.

78. Arunkumar R, Josephs EB, Williamson RJ, Wright SI: **Pollen-Specific, but Not Sperm-Specific, Genes Show Stronger Purifying Selection and Higher Rates of Positive Selection Than Sporophytic Genes in Capsella grandiflora.** *Mol Biol Evol* 2013, **30:**2475–2486.

79. Arthur KM, Vejlupkova Z, Meeley RB, Fowler JE: **Maize ROP2 GTPase provides a competitive advantage to the male gametophyte.** *Genetics* 2003, **165:**2137–2151.

80. Ohnishi T, Takanashi H, Mogi M, Takahashi H, Kikuchi S, Yano K, Okamoto T, Fujita M, Kurata N, Tsutsumi N: **Distinct gene expression profiles in egg and synergid cells of rice as revealed by cell type-specific microarrays.** *Plant Physiol* 2011, **155:**881–891.

81. Kranz E, Bautor J, Lorz H: **In vitro ferilization of single, isolated gametes of maize mediated by electrofusion.** *Sex Plant Reprod* 1991, **4:**12–16.

82. Opsahl-Ferstad HG, Le Deunff E, Dumas C, Rogowsky PM: ***ZmEsr*, a novel endosperm-specific gene expressed in a restricted region around the maize embryo.** *Plant J* 1997, **12:**235–246.

83. Magnard JL, Lehouque G, Massonneau A, Frangne N, Heckel T, Gutierrez-Marcos JF, Perez P, Dumas C, Rogowsky PM: **ZmEBE genes show a novel, continuous expression pattern in the central cell before fertilization and in specific domains of the resulting endosperm after fertilization.** *Plant Mol Biol* 2003, **53:**821–836.

84. Evans MMS: **The indeterminate gametophyte1 Gene of Maize Encodes a LOB Domain Protein Required for Embryo Sac and Leaf Development.** *Plant Cell* 2007, **19:**46–62.

85. Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, Sanmiguel PJ, Bennetzen JL: **Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome.** *PLoS Genet* 2009, **5:**e1000732.

86. Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J: **Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians.** *Genome Biol Evol* 2011, **3:**219–229.

87. **RepeatMasker.** [http://www.RepeatMasker.org].

88. Tarailo-Graovac M, Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences.** *Curr Protoc Bioinformatics* 2009, **Chapter 4:**Unit 4 10.

89. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A: **High-throughput functional annotation and data mining with the Blast2GO suite.** *Nucleic Acids Res* 2008, **36:**3420–3435.

90. Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D, Schnable PS: **Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content.** *PLoS Genet* 2009, **5:**e1000734.

91. Oliveros JC: **VENNY. An interactive tool for comparing lists with Venn Diagrams.** [http://bioinfogp.cnb.csic.es/tools/venny/index.html].

92. **Bioinforx Convenient Research Tools.**
[http://apps.bioinforx.com/bxaf6/tools/app_overlap.php].

93. Du Z, Zhou X, Ling Y, Zhang Z, Su Z: **agriGO: a GO analysis toolkit for the agricultural community.** *Nucleic Acids Res* 2010, **38:**W64–W70.

94. **Maize Sequence.** [http://www.maizesequence.org].

95. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754–755.

96. **Molecular Evolution, Phylogenetics and Epidemiology.**
[http://tree.bio.ed.ac.uk/software/figtree/].

97. Sen TZ, Andorf CM, Schaeffer ML, Harper LC, Sparks ME, Duvick J, Brendel VP, Cannon E, Campbell DA, Lawrence CJ: **MaizeGDB becomes 'sequence-centric'.** *Database* 2009, **2009:**bap20.

Additional files

Additional_file_1 as XLSX

Additional file 1: Table S1. Working Gene Set FPKM for all replicates and average FPKM per tissue type.

Additional_file_2 as XLSX

Additional file 2: Table S2. Filtered Gene Set FPKM for all replicates and average FPKM per tissue type.

Additional_file_3 as XLSX

Additional file 3 Table S3. Predicted Intergeinc Gene Models **S3A.** Summaries of Characteristics of Intergenic GeneModels (XLOCs). **Table S3B**. Expression and position of non-repeat-related intergenic gene models. **Table S3C**. Expression and position of repeat-related intergenic gene models. **Table S3D**. Detailed BLAST2GO Output for Non Repeat-related Intergenic Gene Models.

Additional_file_4 as XLSX

Additional file 4: Table S4. All FGS genes with expression greater than or equal to 0.1 FPKM for each tissue, sorted highest to lowest as extracted from Additional file 2: Table S2.

Additional_file_5 as XLSX

Additional file 5: Table S5. Genes with duplicates in subgenome 1 and 2 and singleton genes with enrichment (2-fold) in one tissue type vs other three.

Additional_file_6 as XLSX

Additional file 6: Table S6. Genes above 0.1 FPKM that are enriched in each tissue (i.e. 2 fold higher in the test tissue(s) than the other tissues).

Additional_file_7 as XLSX

Additional file 7 Table S7. GO Terms Overrepresented in the Full Transcriptome of Each Tissue above 0.1 FPKM. **Table S7A.** GO terms Overrepresented in the full B73 Seedling transcriptome for genes with average FPKM, above 0.1; total of 27,564 genes, agrigo performed on 3-1-13. **Table S7B**. GO terms Overrepresented in full B73 Pollen transcriptome for genes with average FPKM above 0.1; total of 14,591 genes, agrigo performed on 3-1-13. **Table S7C**. GO terms Overrepresented in full B73/W23 combined Embryo Sac transcriptome for genes with average FPKM above 0.1; total of 28,489 genes, agrigo performed on 3-1-13. **Table S7D**. GO terms overrepresented in the full B73/W23 combined Ovules without Embryo Sacs transcriptome for genes with average FPKM above 0.1; total of 26,338 genes, agrigo performed on 3-1-13.

Additional_file_8 as PNG

Additional file 8: Figure S1. Comparison of GO terms overrepresented in the full transcriptome of all four tissue samples. Overrepresented GO terms from the full transcriptome of each tissue type in Additional file 7: Table S7 were compared to identify which GO terms are unique to each tissue and which are shared between tissues.

Additional_file_9 as XLSX

Additional file 9: Table S8. GO terms Overrepresented in each list of tissue enriched genes. **Table S8A.** GO terms Overrepresented in genes Seedling enrichment (2-fold higher in Seedling compared to all other tissue types) and expressed above 0.1 FPKM; total of 8,066 genes, agrigo performed on 2-26-13. **Table S8B**. GO terms Overrepresented in genes with Pollen enrichment (2-fold higher in Pollen compared to all other tissue types) and expressed above 0.1 FPKM; total of 2,224 genes, agrigo performed on 2-26-13. **Table S8C**. GO terms Overrepresented in genes with B73/W23 Embryo Sac-enrichment (2-fold B73/W23 combined Embryo Sac compared to all other tissue types) and expressed above 0.1 FPKM; total of 5,011 genes, agrigo performed on 2-26-13. **Table S8D**. GO terms Overrepresented in genes with combined B73/W23 Ovules enrichment (2-fold in B73/W23 Ovules without Embryo Sacs compared to all other tissue types) and expressed above 0.1 FPKM; total of 1,770 genes, agrigo performed on 2-26-13. **Table S8E**. GO terms Overrepresented in genes with enrichment in both B73 pollen and B73/W23 combined Embryo Sacs (2-fold compared to both seedling and ovule-without-embryo-sacs) and expressed above 0.1 FPKM; total of 591 genes, agrigo performed on 2-26-13.

Additional_file_10 as XLSX

Additional file 10: Table S9 Analysis of Pollen-expressed Subgenome 2 genes.

Additional_file_11 as XLSX

Additional file 11: Table S10. Representation of Transcription Factor Families Within Each Tissue-Enriched Gene List.

Additional_file_12 as PNG

Additional file 12: Figure S2. Phylogeny and expression of maize and Arabidopsis MADS transcription factor genes. Gene names in blue are part of the MP-enriched gene set. Gene names in red are part of the ES-enriched gene set. Gene names in magenta are part of the dual gametophyte-enriched gene set. Arabidopsis genes expressed in the Embryo Sac and Pollen are from published reports [13,15,17,27,51-55]. Indication of gametophyte expression is different for maize and Arabidopsis. For maize it is called as positive if the gene is two-fold higher vs the three other tissues while in Arabidopsis it is measured as detectable expression, often by use of transgenic reporters. The yellow bar indicates the MIKC* group. The blue bar indicates the Type 1 Class γ group. The aqua bar indicates the Type 1 Class ß group with no maize genes. The purple bar indicates the MIKC group. The green bar indicates the Type 1 Class α group. Posterior probability values are given at node positions. Arabidopsis genes begin with At.

Additional_file_13 as PNG

Additional file 13: Figure S3. Phylogeny and expression of maize NAC transcription factor genes. Gene names in red are part of the ES-enriched gene set. Posterior probability values are given at node positions.

Additional_file_14 as PNG

Additional file 14: Figure S4. Phylogeny and expression of maize AP2-EREB transcription factor genes. Gene names in red are part of the ES-enriched gene set. Posterior probability values are given at node positions.

Additional_file_15 as PNG

Additional file 15: Figure S5. Phylogeny and expression of maize MYBR transcription factor genes. Gene names in red are part of the ES-enriched gene set. Posterior probability values are given at node positions.

Additional_file_16 as PNG

Additional file 16: Figure S6. Phylogeny and expression of maize WRKY transcription factor genes. Gene names in red are part of the ES-enriched gene set. Posterior probability values are given at node positions.

Additional_file_17 as PNG

Additional file 17: Figure S7. Phylogeny and expression of maize CLE genes. Gene names in blue are part of the MP-enriched gene set. Gene names in red are part of the ES-enriched gene set. Gene names in magenta are part of the dual gametophyte-enriched gene set. Expression levels are indicated by color of the letter of each sample type with red meaning >10 FPKM, orange between 1 and 10 FPKM, green between 0.1 and 1 FPKM, blue greater than zero but less than 0.1 FPKM, and black having 0 reads. E = embryo sac expression; O = ovule without embryo sac expression; S = seedling expression; P = mature pollen expression. CLV3 of Arabidopsis is included for reference. Posterior probability values are given at node positions.

Additional_file_18 as PNG

Additional file 18: Figure S8. Phylogeny and expression of maize ZPR genes. Gene names in blue are part of the MP-enriched gene set. Gene names in red are part of the ES-enriched gene set. Gene names in magenta are part of the dual gametophyte-enriched gene set. Expression levels are indicated by color of the letter of each sample type with red meaning >10 FPKM, orange between 1 and 10 FPKM, green between 0.1 and 1 FPKM, blue greater than zero but less than 0.1 FPKM, and black having 0 reads. E = embryo sac expression; O = ovule without embryo sac expression; S = seedling expression; P = mature pollen expression. Both Arabidopsis (At) and rice (Os) genes are included for reference. Posterior probability values are given at node positions.

Additional_file_19 as XLSX

Additional file 19: Table S11. Baseline Frequency of Gene Models in the whole genome with at least one transposon insertion mapping to a particular region of the gene model.

Additional_file_20 as XLSX

Additional file 20: Table S12. Frequency of Tissue-enriched Gene Models with at least one transposon insertion mapping to a particular region of the gene model.

Additional_file_21 as XLSX

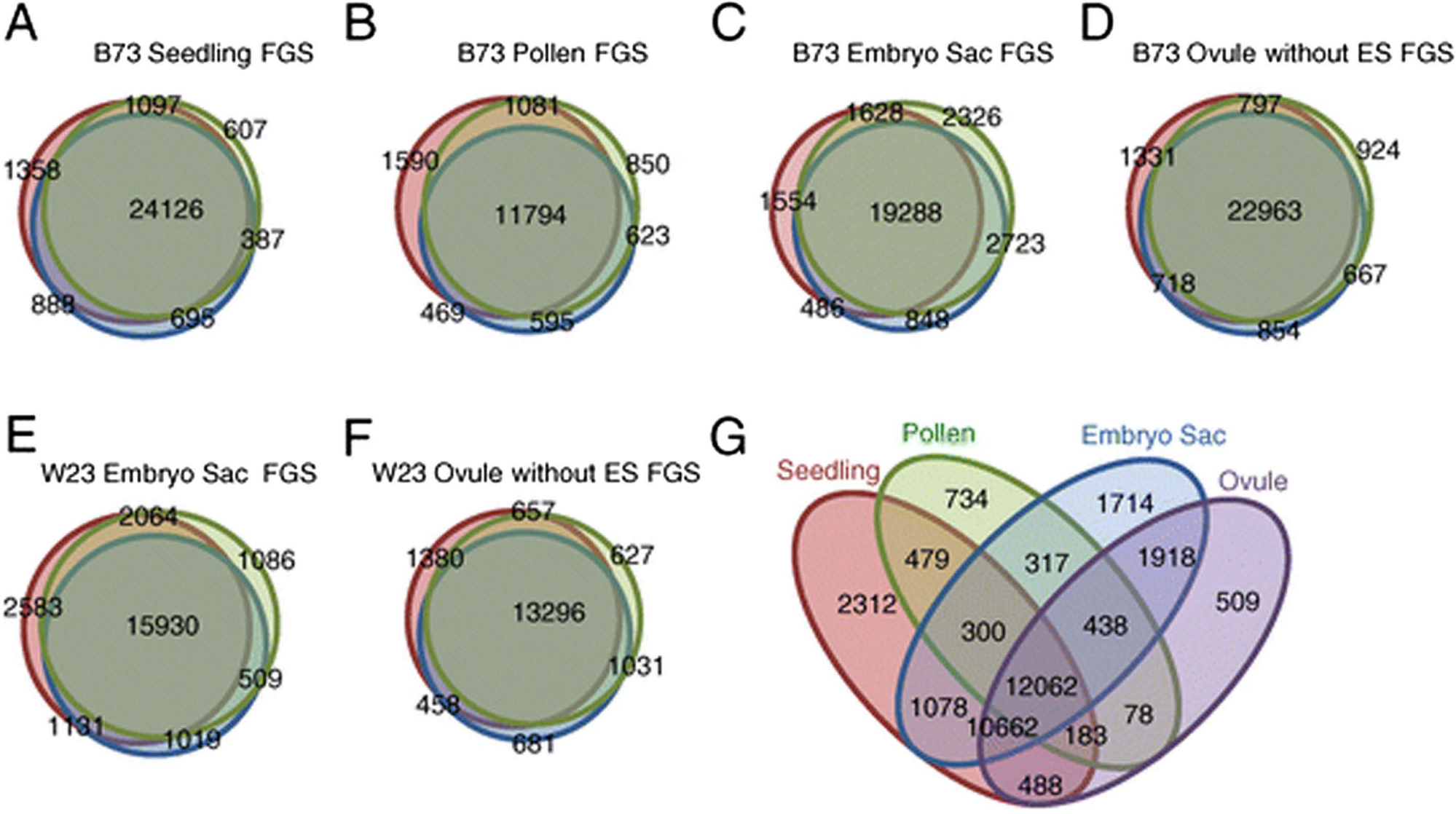
Additional file 21: Table S13. Transmission Data for Ds insertions in genes with and without enrichment in a gametophyte tissue.

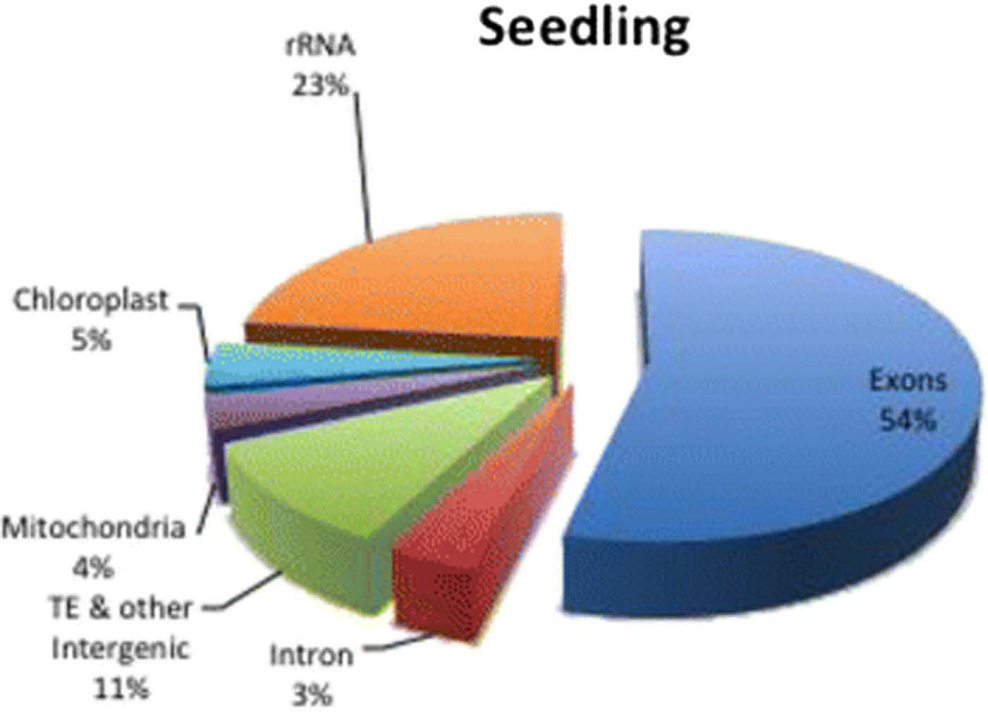
Additional_file_22 as DOCX

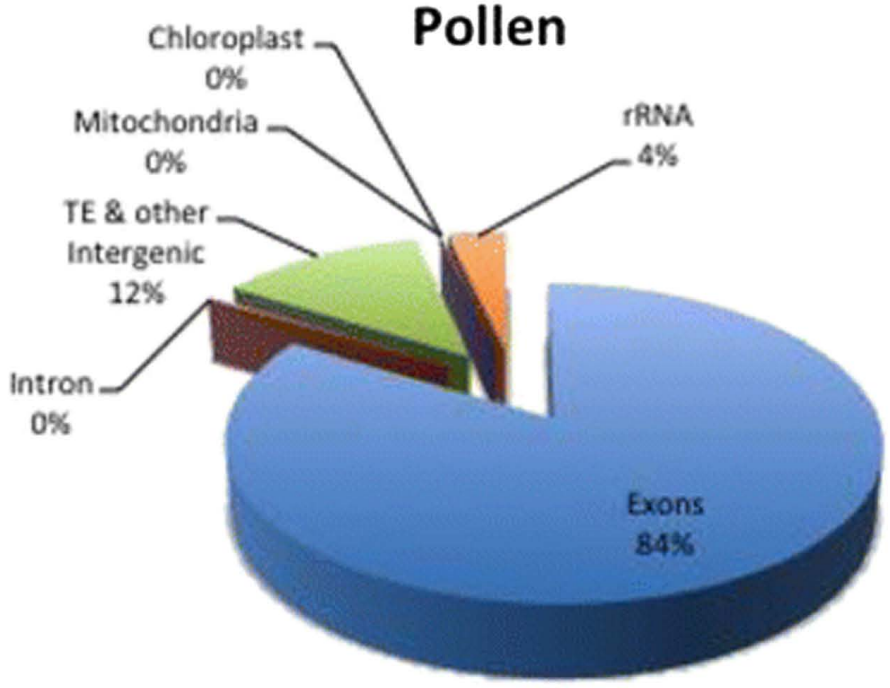
Additional file 22: Table S14. Verifying RNA-Seq Quality of Embryo Sac Samples Using High Confidence Embryo Sac Specific Genes.

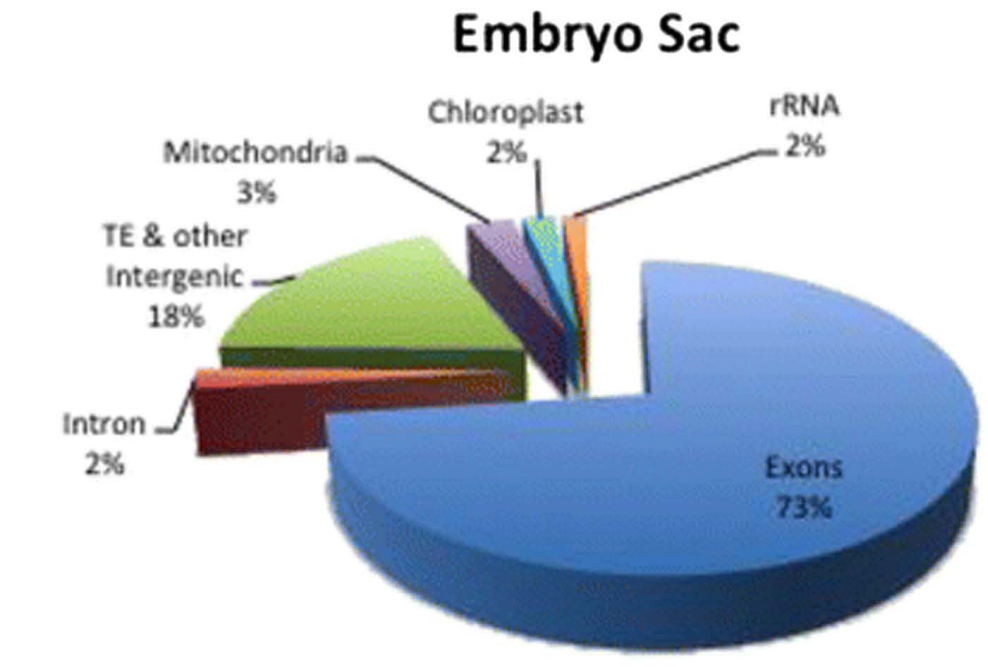
Additional_file_23 as XLSX

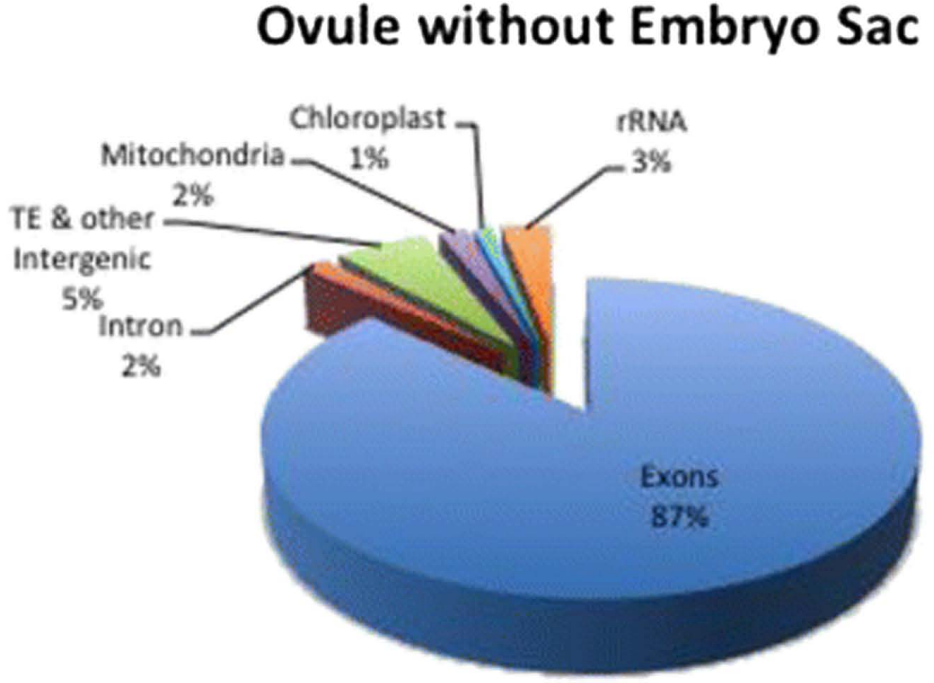
Additional file 23: Table S15. Primers for RT-PCR of test genes.


A
B73 Seedling FGS
1097
607
1358
24126
387
888
695
B
B73 Pollen FGS
1081
1590
850
11794
623
469
595
C
B73 Embryo Sac FGS
1628
2326
1554
19288
2723
486
848
D
B73 Ovule without ES FGS
797
1331
924
22963
718
667
854
E
W23 Embryo Sac FGS
2064
1086
2583
15930
509
1131
1019
F
W23 Ovule without ES FGS
657
1380
627
13296
1031
458
681
G
Seedling
Pollen
Embryo Sac
Ovule
734
1714
479
317
1918
2312
509
300
438
12062
1078
10662
183
78
488

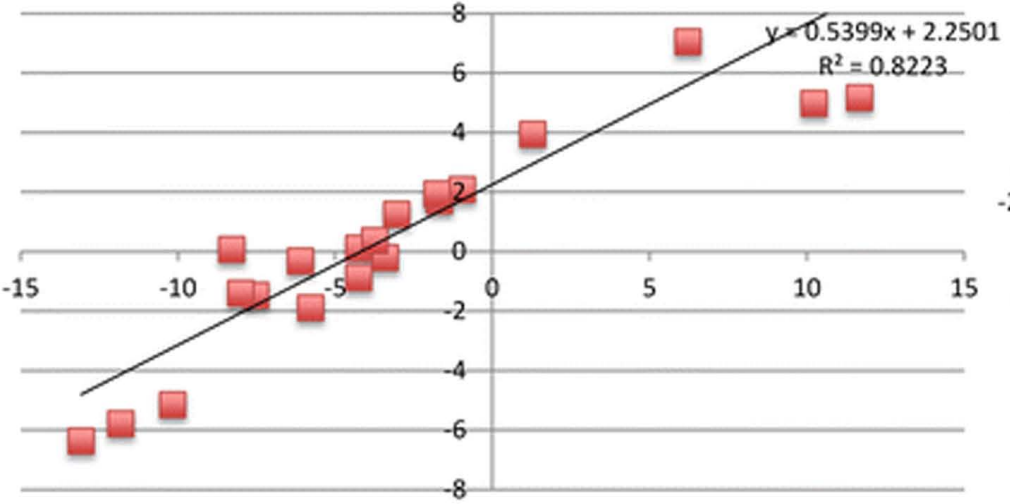

Seedling
rRNA
23%
Chloroplast
5%
Mitochondria
4%
TE & other
Intergenic
11%
Intron
3%
Exons
54%


Pollen
Chloroplast
0%
Mitochondria
0%
TE & other
Intergenic
12%
Intron
0%
rRNA
4%
Exons
84%

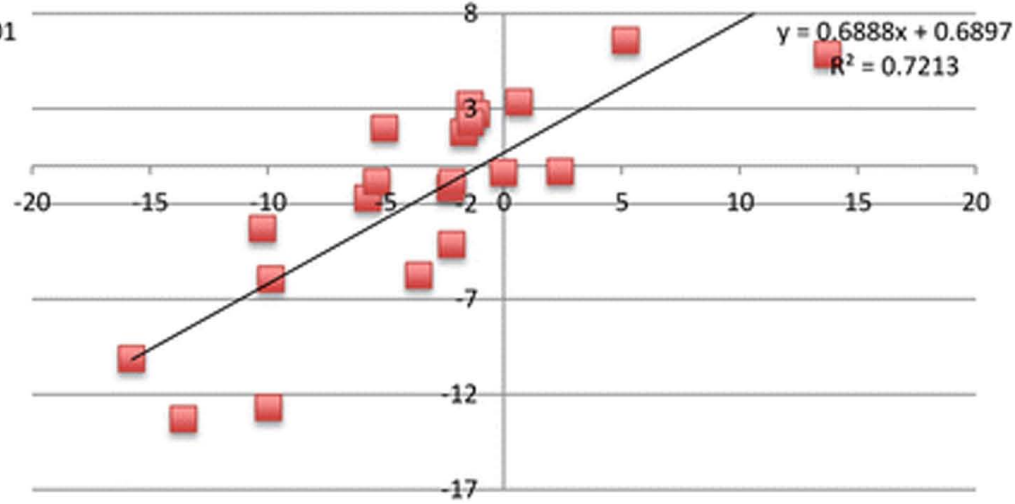

Embryo Sac
Chloroplast
2%
rRNA
2%
Mitochondria
3%
TE & other
Intergenic
18%
Intron
2%
Exons
73%


Ovule without Embryo Sac
Chloroplast
1%
rRNA
3%
Mitochondria
2%
TE & other
Intergenic
5%
Intron
2%
Exons
87%

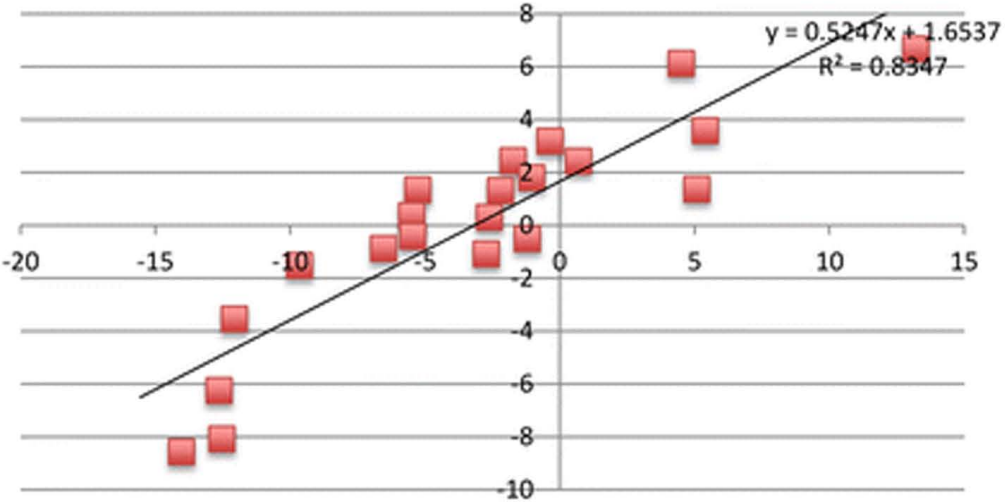
Seedling vs Pollen

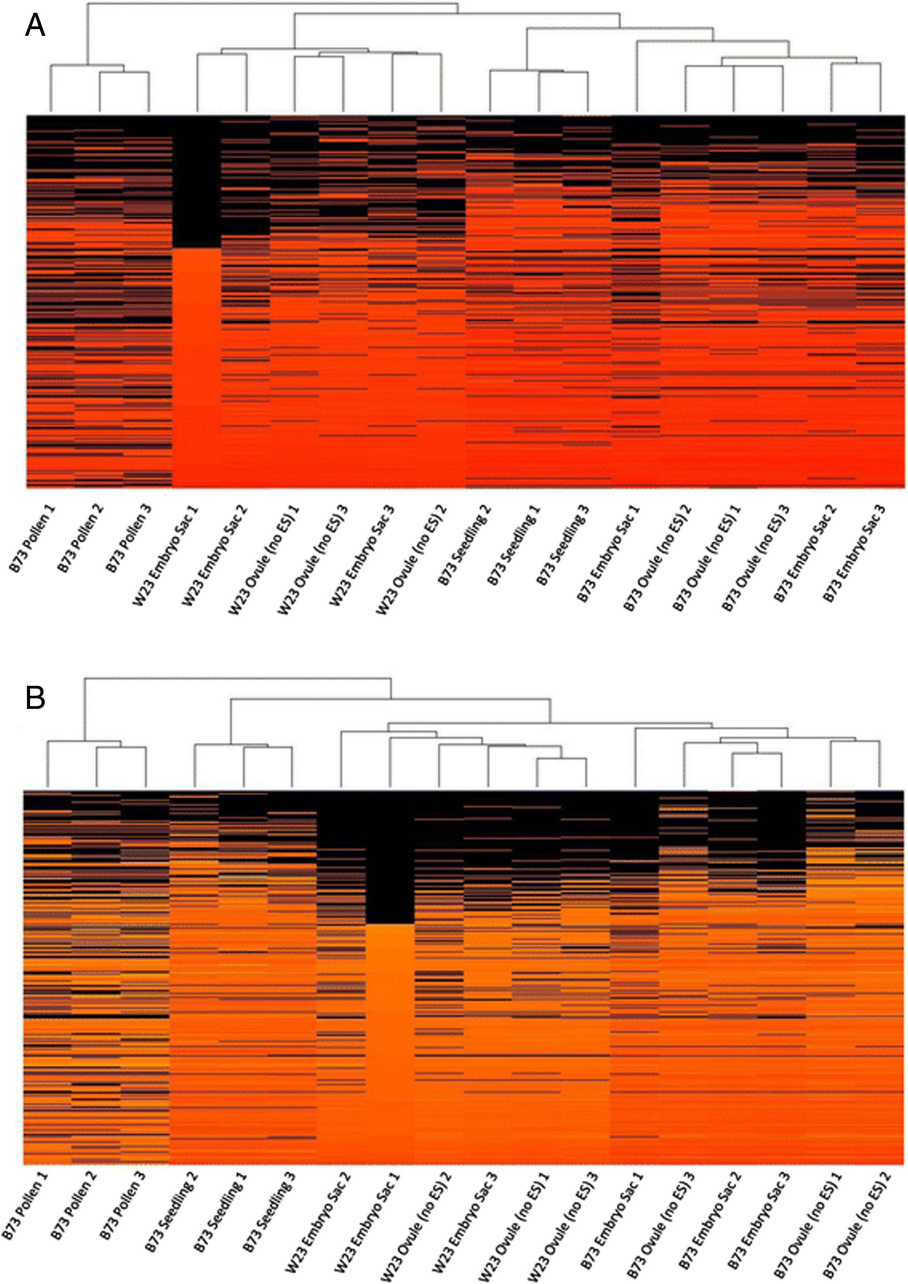


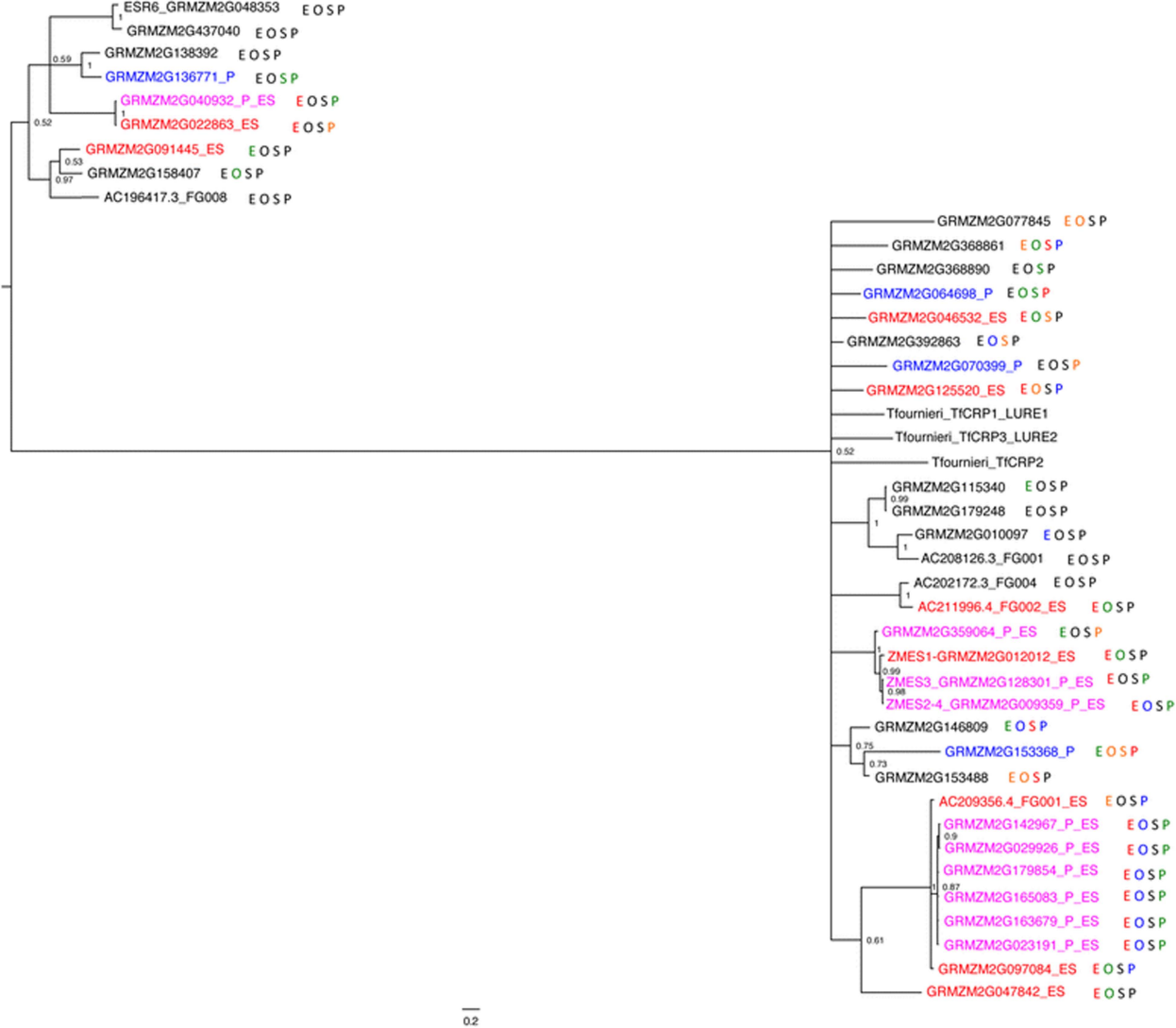
Embryo Sac vs Pollen

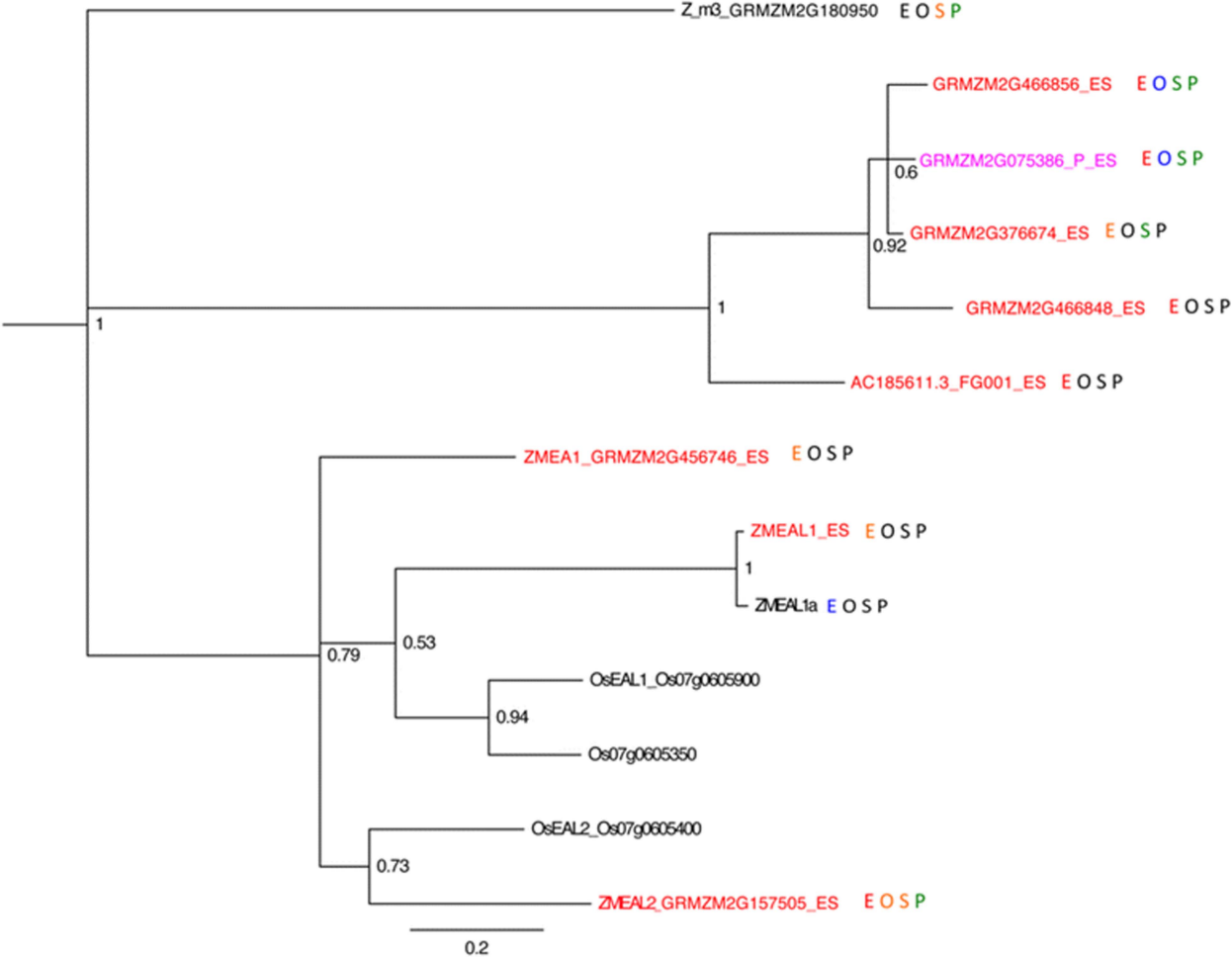


Ovule vs Pollen




A
B73 Pollen 1
B73 Pollen 2
B73 Pollen 3
W23 Embryo Sac 1
W23 Embryo Sac 2
W23 Ovule (no ES) 1
W23 Ovule (no ES) 3
W23 Embryo Sac 3
W23 Ovule (no ES) 2
B73 Seedling 2
B73 Seedling 1
B73 Seedling 3
B73 Embryo Sac 1
B73 Ovule (no ES) 2
B73 Ovule (no ES) 1
B73 Ovule (no ES) 3
B73 Embryo Sac 2
B73 Embryo Sac 3
B
B73 Pollen 1
B73 Pollen 2
B73 Pollen 3
B73 Seedling 2
B73 Seedling 1
B73 Seedling 3
W23 Embryo Sac 2
W23 Embryo Sac 1
W23 Ovule (no ES) 2
W23 Embryo Sac 3
W23 Ovule (no ES) 1
W23 Ovule (no ES) 3
B73 Embryo Sac 1
B73 Ovule (no ES) 3
B73 Embryo Sac 2
B73 Embryo Sac 3
B73 Ovule (no ES) 1
B73 Ovule (no ES) 2


ESR6_GRMZM2G048353 E O S P
GRMZM2G437040 E O S P
GRMZM2G138392 E O S P
GRMZM2G136771_P E O S P
GRMZM2G040932_P_ES E O S P
GRMZM2G022863_ES E O S P
GRMZM2G091445_ES E O S P
GRMZM2G158407 E O S P
AC196417.3_FG008 E O S P
GRMZM2G077845 E O S P
GRMZM2G368861 E O S P
GRMZM2G368890 E O S P
GRMZM2G064698_P E O S P
GRMZM2G046532_ES E O S P
GRMZM2G392863 E O S P
GRMZM2G070399_P E O S P
GRMZM2G125520_ES E O S P
Tfournieri_TfCRP1_LURE1
Tfournieri_TfCRP3_LURE2
Tfournieri_TfCRP2
GRMZM2G115340 E O S P
GRMZM2G179248 E O S P
GRMZM2G010097 E O S P
AC208126.3_FG001 E O S P
AC202172.3_FG004 E O S P
AC211996.4_FG002_ES E O S P
GRMZM2G359064_P_ES E O S P
ZMES1-GRMZM2G012012_ES E O S P
ZMES3_GRMZM2G128301_P_ES E O S P
ZMES2-4_GRMZM2G009359_P_ES E O S P
GRMZM2G146809 E O S P
GRMZM2G153368_P E O S P
GRMZM2G153488 E O S P
AC209356.4_FG001_ES E O S P
GRMZM2G142967_P_ES E O S P
GRMZM2G029926_P_ES E O S P
GRMZM2G179854_P_ES E O S P
GRMZM2G165083_P_ES E O S P
GRMZM2G163679_P_ES E O S P
GRMZM2G023191_P_ES E O S P
GRMZM2G097084_ES E O S P
GRMZM2G047842_ES E O S P
0.2


Z_m3_GRMZM2G180950 E O S P
GRMZM2G466856_ES E O S P
GRMZM2G075386_P_ES E O S P
0.6
GRMZM2G376674_ES E O S P
0.92
GRMZM2G466848_ES E O S P
1
1
AC185611.3_FG001_ES E O S P
ZMEA1_GRMZM2G456746_ES E O S P
ZMEAL1_ES E O S P
1
ZMEAL1a E O S P
0.53
0.79
OsEAL1_Os07g0605900
0.94
Os07g0605350
OsEAL2_Os07g0605400
0.73
ZMEAL2_GRMZM2G157505_ES E O S P
0.2