"Out of Pollen" Hypothesis for Origin of New Genes in Flowering Plants: Study from *Arabidopsis thaliana*

Dong-Dong Wu[1,3,†], Xin Wang[2,†], Yan Li[1,3,†], Lin Zeng[1,3], David M. Irwin[1,4,5], and Ya-Ping Zhang[1,2,3,*]

[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

[2]Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, Kunming, China

[3]Kunming College of Life Science, University of Chinese Academy of Sciences, Kunming, Yunnan, China

[4]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada

[5]Banting and Best Diabetes Centre, University of Toronto, Toronto, Canada

*Corresponding author: E-mail: zhangyp@mail.kiz.ac.cn.
†These authors contributed equally to this work.

Accepted: September 10, 2014

Abstract

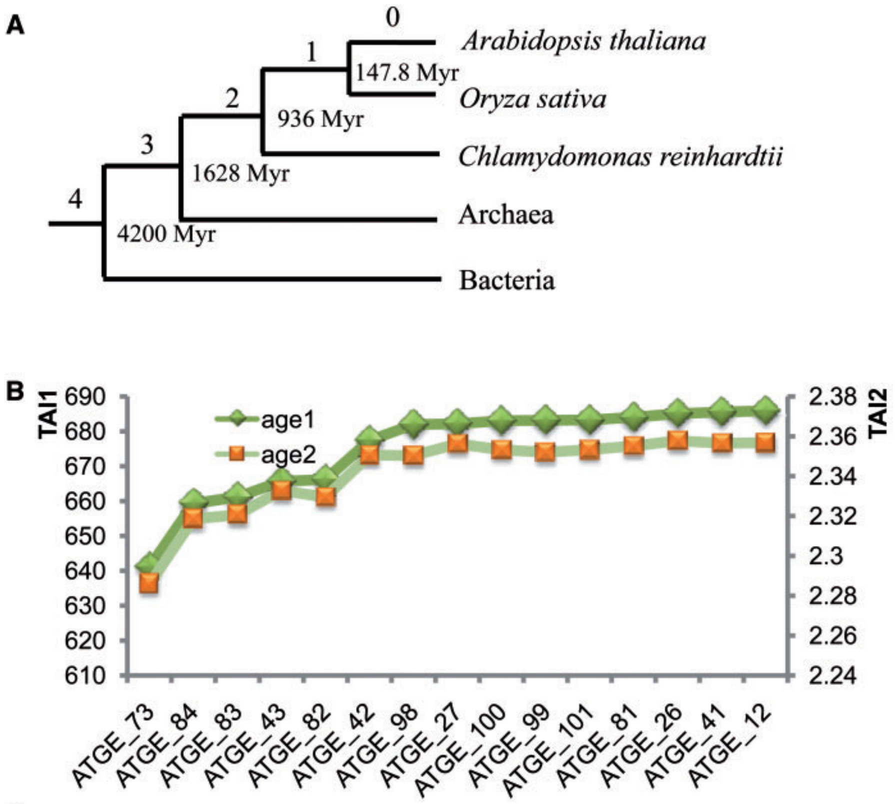
New genes, which provide material for evolutionary innovation, have been extensively studied for many years in animals where it is observed that they commonly show an expression bias for the testis. Thus, the testis is a major source for the generation of new genes in animals. The source tissue for new genes in plants is unclear. Here, we find that new genes in plants show a bias in expression to mature pollen, and are also enriched in a gene coexpression module that correlates with mature pollen in *Arabidopsis thaliana*. Transposable elements are significantly enriched in the new genes, and the high activity of transposable elements in the vegetative nucleus, compared with the germ cells, suggests that new genes are most easily generated in the vegetative nucleus in the mature pollen. We propose an "out of pollen" hypothesis for the origin of new genes in flowering plants.

Key words: "Out of pollen" hypothesis, young gene evolution, *Arabidopsis thaliana*.

New genes, which provide materials for evolutionary innovation, have been studied for many years in viruses, bacterium, yeast, plants, and animals, where the remarkable contribution of these new genes to phenotypic evolution has been appreciated (Kaessmann 2010; Chen et al. 2013; Wu and Zhang 2013). Most of the findings concerning the evolutionary patterns of new genes are derived from studies on animals, particularly *Drosophila melanogaster* (Betran et al. 2002; Wang et al. 2002, 2004; Domazet-Loso and Tautz 2003; Levine et al. 2006; Zhou et al. 2008; Vibranovski et al. 2009; Chen et al. 2010; Zhang et al. 2010) and primates (e.g., human) (Marques et al. 2005; Vinckenbosch et al. 2006; Zhang et al. 2010; Wu et al. 2011; Xie et al. 2012). These studies with animal genomes have led to the proposed "out-of-testis" hypothesis, where the testis, a male reproductive organ, is the catalyst for the birth and evolution of new genes (Kaessmann 2010). New genes, including both duplicated and de novo genes, show a male or testis-bias in expression in animals, and many of them have biological functions associated with reproduction (Kaessmann 2010; Chen et al. 2013; Long et al. 2013). Chromatin in spermatocytes and spermatids shows widespread demethylation of the CpG-enriched promoter sequences and contain modified histones (Kleene 2001; Soumillon et al. 2013), which causes an elevation in the levels of components of the transcriptional machinery and permit promiscuous transcription of nonfunctional sequences, and facilitates the initial transcription of newly arisen genes (Kaessmann 2010).

In contrast to animals, studies on the origin of new genes in plants has received much less attention, although there have been, over the years, many studies identifying new genes (Jiang et al. 2004; Zhang et al. 2005; Wang et al. 2006; Fan et al. 2007; Hanada et al. 2008; Zhu et al. 2009; Zou et al. 2009). To date, it is unclear which tissue/organ/developmental stages are the pools from which new genes are generated.

Here, we retrieved microarray expression data from a total 79 samples representing different tissues and developmental stages in *Arabidopsis thaliana*. These data were used for the following analysis of the expression patterns of young genes. First, we calculated the transcriptome age index (TAT) for each



C

ATGE_73	mature pollen
ATGE_84	seeds, stage 10, w/o siliques; green cotyledons embryos
ATGE_83	seeds, stage 9, w/o siliques; curled cotyledons to early green cotyledons embryos
ATGE_43	flowers stage 15, stamen
ATGE_82	seeds, stage 8, w/o siliques; walking-stick to early curled cotyledons embryos
ATGE_42	flowers stage 15, petals
ATGE_98	root
ATGE_27	stem, 2nd internode
ATGE_100	seedling, green parts
ATGE_99	root
ATGE_101	seedling, green parts
ATGE_81	seeds, stage 7, w/o siliques; late torpedo to early walking-stick embryos
ATGE_26	cauline leaves
ATGE_41	flowers stage 15, sepals
ATGE_12	rosette leaf # 2

FIG. 1.—(*A*) Phylogenetic tree used for deducing the ages of genes using ProteinHistorian. Age indices 0–4 are presented above each branch. The actual ages are presented beside the ancestral nodes: 1,47.8, 936, 1,628, and 4,200 Myr. (*B*) Top 15 tissues with lowest TAI values. TAI = ∑E*AGE/∑E, where E, and AGE are the expression value, and age of each gene, respectively. TAI1 is calculated using the true phylogenetic time (Myr) as AGE, TAI2 is calculated by using the phylostratum number as described in Domazet-Loso et al. (2007) to represent AGE. The full data is in supplementary table S1, Supplementary Material online.

tissue/developmental stage using the formula TAI=∑(E*A)/∑E, where E and A are the expression value and age of each gene (fig. 1*A*), respectively, as previously described (Domazet-Loso et al. 2007; Domazet-Loso and Tautz 2010). The tissue with the lowest value for the TAT is mature pollen, suggesting that mature pollen expresses relative more young genes than other tissues (fig. 1*B* and supplementary table S1, Supplementary Material online). This observation is consistent with a previous study on the transcriptomic hourglass model in plant embryogenesis, which found young features in the mature stage (Quint et al. 2012). In addition, we also used transcriptome data generated by RNA-sequencing, and also find that the lowest TAI value is for pollen (supplementary fig. S1, Supplementary Material online).

A weighted gene coexpression network analysis was then performed based on the microarray expression data to identify distinct coexpression modules corresponding to the clusters of correlated genes (Langfelder and Horvath 2008). A total of 29 modules, that is, gene coexpression networks and labeled by different colors, were constructed (fig. 2*A*), with many of the

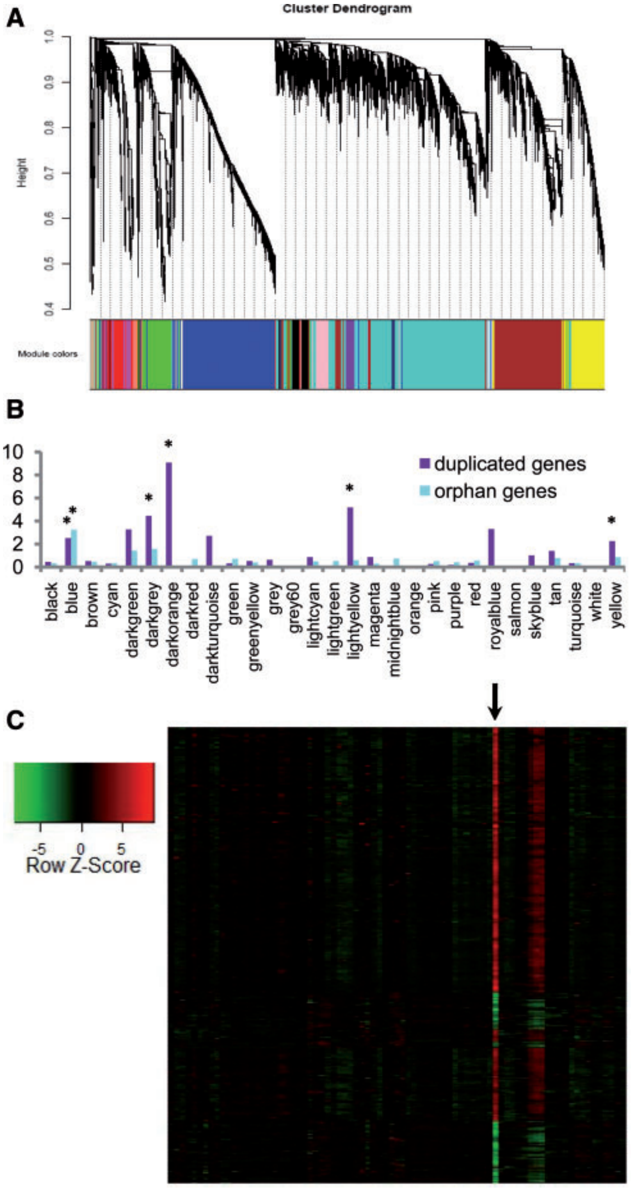


Fig. 2.—(*A*) Dendrogram showing clustering of genes by WGCNA. Modules labeled by different colors. (*B*) Enrichment of duplicated genes and orphan genes among different modules. Stars show modules with significant enrichments. The enrichment score of young genes in each module is calculated by the proportion of young genes in the module divided by the proportion of other genes in the module. (*C*) Heatmap of the expression of genes in the blue module. Arrow shows mature pollen. *Z* score is defined as "actual value" minus the mean of the group divided by the standard deviation.

modules showing significant correlation of their expression with specific stages/tissues (supplementary fig. S2, Supplementary Material online). A total of 562 young genes that likely originated by duplication (which are named young duplicated genes in the following) and 491 orphan genes in the *A. thaliana* were found in the coexpression modules. The relative expression level of these young genes in each sample was calculated by dividing the mean expression value of the young genes in the sample by the mean expression value of the other genes in the sample. As expected, young genes showed the highest relative expression level in mature pollen (supplementary tables S2 and S3, Supplementary Material online). Five gene coexpression modules, labeled by blue, dark gray, dark orange, light yellow, and yellow, were found to be significantly enriched with genes that were specifically duplicated in *A. thaliana* (fig. 2*B*). Because only blue module show consistent enrichment of both duplicated new genes and orphan genes, we focused on this module subsequently. The blue module contained the greatest number of young duplicated genes (277, 49.3%), and is significantly associated with mature pollen tissue ($P = 6\text{E-}13$, supplementary fig. S2, Supplementary Material online). Consistent with this, young orphan genes (using an *E* value of 10^{-10} in the BLASTP analysis) (309, 62.9%) were extremely significantly enriched only in the blue module ($P = 3.60\text{E-}124$). Similarly, when an *E* value of 10^{-5} was used for the BLASTP analysis, 179 (69.11%) orphan genes are significantly enriched in the blue module (supplementary fig. S3, Supplementary Material online). A heatmap analysis also revealed high expression levels of these genes in mature pollen (fig. 2*C*) supporting the conclusions that young new genes show a bias in expression to mature pollen. The three other columns showing higher expression levels (fig. 2*C*) are seeds (stage 8, w/o siliques; walking-stick to early curled cotyledons embryos), seeds (stage 9, w/o siliques; curled cotyledons to early green cotyledons embryos), and seeds (stage 10, w/o siliques; green cotyledons embryos respectively. There may be similar expression patterns among these tissues. Actually, the seeds also show lower values of TAT following the mature pollen.

The process of pollen development may provide crucial information for understanding the generation of new genes. In flowering plants, male gametophyte (or pollen) development is a complex process requiring the coordination of various cell and tissue types. Pollen development consists of two major phases—microsporogenesis and microgametogenesis. During microsporogenesis, microsporocytes undergo a meiotic division to give rise to a tetrad of haploid microspores. In microgametogenesis, the released microspores undergo an asymmetric mitotic division (Pollen Mitosis I, PMI), to give rise to a bicellular pollen grain, with a germ cell engulfed within the cytoplasm of a larger vegetative cell. The germ cell then undergo an additional mitotic division (Pollen Mitosis II, PMII) to produce two sperm cells (SCs) (McCormick 2004; Honys et al. 2006; Borg et al. 2009).

The expression patterns of new genes during pollen development, covering four main stages of male gametophyte development (uninucleate microspores, bicellular pollen, immature tricellular pollen, and mature pollen [Honys and Twell 2004]) were then studied in more detail.

Relative expression levels of duplicated genes and orphan genes are higher in the immature tricellular pollen and mature pollen, than in the uninucleate microspores and bicellular pollen (fig. 3*A*). When TAI values were examined, mature pollen showed the lowest TAI value followed by immature tricellular pollen (fig. 3*B*). These data support a mature pollen bias expression of new genes.

Mature pollen is composed of three cells, a vegetative nucleus (VN) and two identical SCs. SCs have condensed chromatin, and provide the paternal genetic contribution to the zygote. The chromatin of the VN is less condensed, compared with SC, and does not contribute DNA directly to the progeny, but functions to control delivery of the sperm (McCormick 1993). Which of these cells are more helpful for the generation of new genes?

Transposable elements (TEs) make up a large proportion of many plant genomes. Transposition, mediated by these TEs, is a very important mechanism in the generation of new duplicated genes in plants, with exaptation of TEs also being an important contributor to orphan genes (Bennetzen 2000, 2005; Jiang et al. 2004; Morgante et al. 2005; Volff 2006; Wissler et al. 2013). A significant enrichment of TEs is observed within young genes, including both duplicated and orphan genes, in *A. thaliana* (fig. 3*C*, $P = 2.00E\text{-}42$; fig. 3*D*, $P = 1.05E\text{-}37$ by Mann–Whitney *U* test). The proportion of the regions covered by TE was significantly higher upstream (0–2 kb) and downstream (0–2 kb) of young genes (21.87% in young duplicated genes and 17.91% in orphan genes) than for the coding region of these young genes (1.50% in young duplicated genes and 2.18% in orphan genes) ($P = 1.03E\text{-}58$, 1.14E-113, respectively, fig. 3*E*), which suggested that these genes were unlikely to be misannotated TEs, but suggest that TEs facilitated the generation of these new genes. The pattern is still significant for the orphan genes identified by BLASTP with *E* value cutoff of 1E-5 (supplementary figs. S4 and S5, Supplementary Material online). TEs are largely quiescent, being both transpositionally and transcriptionally inactive, due to epigenetic silencing (Lisch 2009). TEs are typically highly methylated in plants (Furner and Matzke 2011); however, it had been found that TEs have reduced DNA methylation, higher transcriptional activity, and higher mobility in VN, whereas SCs do not show this increase (Johnson and Bender 2009; Slotkin et al. 2009). Extensive histone replacement, with loss of many canonical histones, also occurs in VNs, which may also contribute to the activation of TEs (Schoft et al. 2009; Berger and Twell 2011). In contrast, in SCs chromatin is condensed and CG methylation is maintained, which may account for the prevalence of epigenetic inheritance in plants compared with mammals (Calarco et al. 2012). As expected, we observed significantly lower levels of CG methylation at new gene regions in the VN compared with SCs (fig. 3*F*, $P = 5.24E\text{-}62$; fig. 3*G*, $P = 1.90E\text{-}108$ by Wilcoxon Signed-Ranks Test). After excluding genes that are overlapped with TEs, the comparison remains statistical significant ($P = 1.68E\text{-}10$, 3.11E-18 by Wilcoxon Signed-Ranks Test, supplementary fig. S6, Supplementary Material online). However, no significant difference of the level of CHG and CHH methylation is observed between VN and SCs after excluding genes overlapping with TEs. The different patterns are likely attributable to different mechanisms. Different mechanisms and molecules are required for regulating CG, CHG, and CHH methylation. CG methylation can be maintained during replication by the DNA MET1, whereas CHG maintenance requires the activity of CMT3, which recognizes H3K9me2 in a self-reinforcing loop mechanism, and CHH methylation is asymmetric and must be reestablished de novo after each cell division, and is directed via siRNAs and is dependent on the activity of DNA methyltransferase DRM2 (Borges et al. 2012). These chromatin changes should support a higher transpositional and transcriptional activity of TEs in the VN and thus would facilitate the generation of new genes. Consistently, gene ontology analysis revealed a significant enrichment of genes involved in the transcription in the blue module, suggesting enhanced transcriptional activity in mature pollen (table 1).

It is paradoxical that new genes are more likely generated from the VN that does not contribute DNA directly to the progeny, rather than the SCs that provide the paternal genetic contribution to the zygote in plants. In animals, the "out-of-testis" hypothesis proposes that the testis, a male reproductive organ, is the catalyst for the birth and evolution of new genes. However, only a few new genes were found that have functions involved with the reproductive system in plants, with greater numbers with functions in adaptive responses to environmental stimuli (Hanada et al. 2008; Zou et al. 2009). Consistently, genes in the blue module, where new genes are overrepresented, showed significant enrichment in the response to stimuli.

TEs facilitate the generation of new protogenes and are most likely to be transcribed in the VN during the development of pollen. Most protogenes would be lost, but some might interact slightly with other older genes, raising the possibility of gaining a very minor role in a pathway or phenotype, such as response to stimuli. If the minor functions of the protogenes increase the fitness of an organism, then the protogenes would have a greater chance of being retained due to natural selection.

Conclusion

Here, we found an expression bias of new genes to mature pollen, and propose an "out of pollen" hypothesis for the origin and evolution of new genes in flowering plants. Enhanced activity of TEs due to epigenetic shifts in the VN of mature pollen facilitates the generation of new genes.

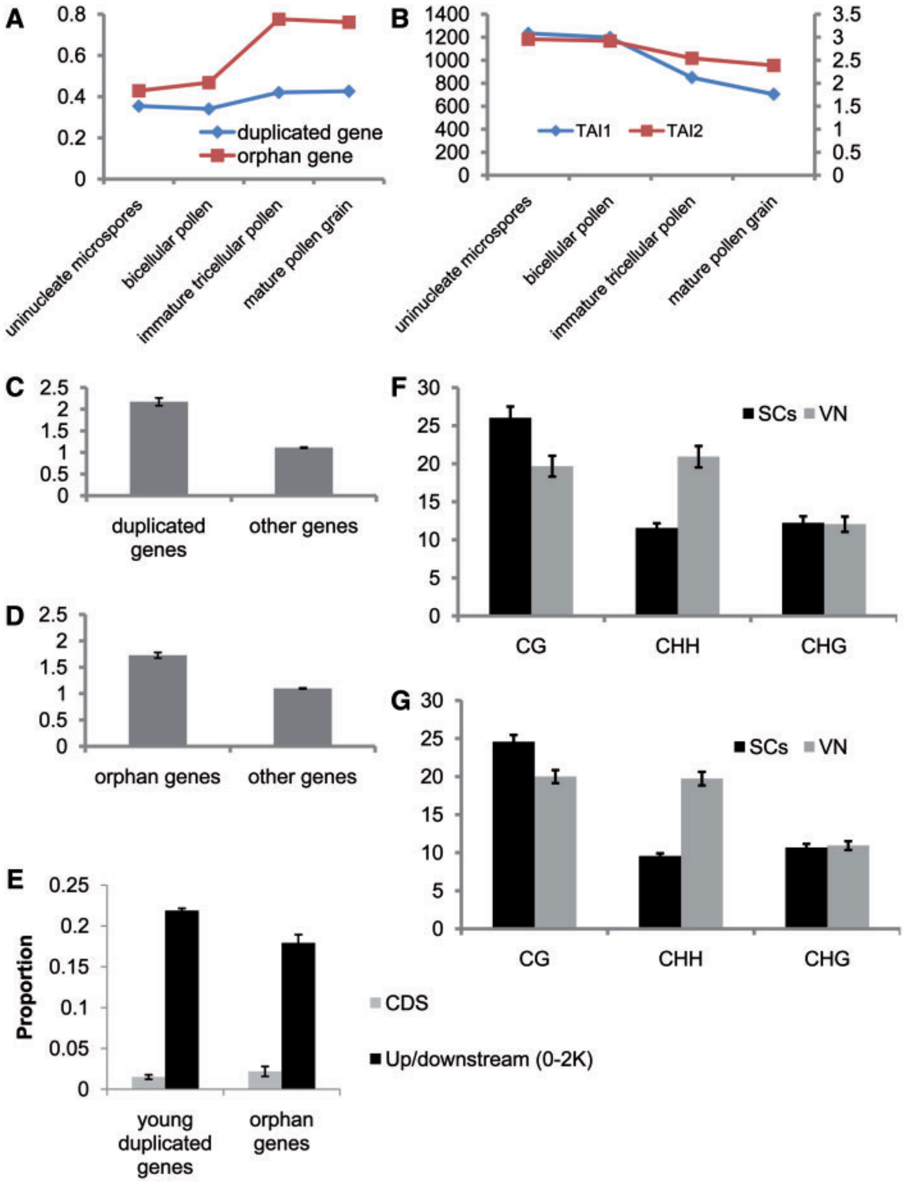


Fig. 3.—(*A*) Relative expression level of new genes at the four stages of pollen development. Relative expression level was calculated by the mean expression level of new genes divided by the mean expression level of genome wide genes for each stage. (*B*) TAI values of the four stages. (*C* and *D*) Enrichment of TEs at regions of duplicated genes, orphan genes. (*E*) Proportion of the regions covered by transposons upstream (0–2 kb) and downstream (0–2 kb) of young genes and in the coding regions of young duplicated genes and orphan genes. (*F* and *G*) Levels of DNA methylation at the CG, CHH, and CHG sites (H = A, C, or T) of duplicated genes and orphan genes in the VN and SCs.

Materials and Methods

The ProteinHistorian database (http://lighthouse.ucsf.edu/ProteinHistorian/, last accessed September 19, 2014) was searched for the ages of *A. thaliana* genes. The full pipeline for dating protein-coding genes is described in (Capra et al. 2012). The pipeline classified *A. thaliana* genes into five different age groups according to their phylogenetic age (fig. 1*A*). Duplicated genes in *A. thaliana* were retrieved by Biomart from EnsemblPlants (http://plants.ensembl.org/, last accessed September 19, 2014), which included 2,312 genes. These genes were those entries related with *A. thaliana* specific duplication events rather than all duplicated genes encoded by *A. thaliana* genome. In other words, EnsemblPlants dates when the duplication occurred, but does not provide information on the direction, namely, which gene is the source copy and which is the derived

Table 1
Gene Ontology Analysis of Genes in the Blue Module

GO Term	Gene Count	*P* Value	Benjamini–Hochberg Corrected *P* Value
GO:0045449~regulation of transcription	304	3.67E-10	5.66E-07
GO:0006355~regulation of transcription, DNA-dependent	182	1.19E-09	9.14E-07
GO:0051252~regulation of RNA metabolic process	182	1.84E-09	9.48E-07
GO:0006350~transcription	205	2.52E-08	9.70E-06
GO:0032870~cellular response to hormone stimulus	84	1.28E-07	3.94E-05
GO:0009755~hormone-mediated signaling	84	1.28E-07	3.94E-05
GO:0006508~proteolysis	165	2.31E-07	5.94E-05
GO:0006511~ubiquitin-dependent protein catabolic process	60	3.25E-07	7.17E-05
GO:0044257~cellular protein catabolic process	100	5.66E-07	9.70E-05
GO:0051603~proteolysis involved in cellular protein catabolic process	99	6.29E-07	9.71E-05
GO:0043632~modification-dependent macromolecule catabolic process	98	7.63E-07	1.07E-04
GO:0019941~modification-dependent protein catabolic process	98	7.63E-07	1.07E-04
GO:0007242~intracellular signaling cascade	131	5.58E-07	1.08E-04
GO:0044265~cellular macromolecule catabolic process	102	1.71E-06	2.20E-04
GO:0030163~protein catabolic process	100	1.94E-06	2.30E-04
GO:0009057~macromolecule catabolic process	113	3.90E-06	4.30E-04
GO:0030528~transcription regulator activity	287	1.43E-05	0.012428
GO:0009725~response to hormone stimulus	134	1.61E-04	0.016387
GO:0006468~protein amino acid phosphorylation	153	1.88E-04	0.017933
GO:0048545~response to steroid hormone stimulus	11	2.27E-04	0.02042
GO:0009742~brassinosteroid mediated signaling	11	2.27E-04	0.02042
GO:0043401~steroid hormone mediated signaling	11	2.27E-04	0.02042
GO:0003700~transcription factor activity	252	5.08E-05	0.021912
GO:0000160~two-component signal transduction system (phosphorelay)	42	3.05E-04	0.025773
GO:0009719~response to endogenous stimulus	140	3.75E-04	0.029989

copy. Here, we only focused the young genes, so we excluded the genes with age >0 deduced by the ProteinHistorian pipeline from the 2,312 duplicated genes, which resulted in retaining 562 young genes that are most likely derived copies generated by duplication. *Arabidopsis thaliana* protein sequences from EnsemblPlants release 17 were BLASTP against the proteins of *A. lyrata, Brachypodium distachyon, Oryza indica, Oryza glaberrima, Oryza sativa, Physcomitrella patens, Populous trichocarpa, Sorghum bicolor,* and *Vitis vinifera, Zea mays*, with an *E* value lower than 10^{-10} used to designate homologs. The genes in *A. thaliana* having no homologous hits in the other species were treated as orphan genes. Sequences with amino acid lengths shorter than 100 residues were discarded. A total 1,527 orphan genes remained after excluding young duplicated genes. A total 491 orphan genes have expression data and remain in the gene coexpression network. We also performed a BLASTP analysis using an *E* value of 10^{-5}.

Microarray expression data from 79 samples (including different tissues and developmental stages) in the study by (Schmid et al. 2005) were downloaded from http://www.webcitation.org/getfile?fileid=73b38cfb54fd3dba9bc1299632d18aab1228162b (last accessed September 19, 2014). The microarray expression data during pollen development, covering four main stages of male gametophyte development (uninucleate microspores, bicellular pollen, immature tricellular pollen and mature pollen) were from the study by Honys and Twell (2004). Weighted gene networks were constructed using the WGCNA (weighted gene coexpression network analysis) package (Langfelder and Horvath 2008). We also evaluated the expression pattern using transcriptome data generated by RNA-sequencing. Data were downloaded from NCBI SRA (supplementary table S4, Supplementary Material online, http://www.ncbi.nlm.nih.gov/sra/, last accessed September 19, 2014). Reads of the RNA-seq data were mapped by Tophat (Trapnell et al. 2009), and Cufflinks (Trapnell et al. 2010) and was used to calculate the expression value for each gene in each tissue.

Gene ontology analysis was performed using the Database for Annotation visualization and Integrated Discovery. TEs data was downloaded from TAIR (www.arabidopsis.org). DNA methylation data were from the study (Calarco et al. 2012) and was downloaded from the GEO data set (GSE40501).

Supplementary Material

Supplementary tables S1–S4 and figures S1–S6 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (31271339, 31321002).

Literature Cited

Bennetzen JL. 2000. Transposable element contributions to plant gene and genome evolution. Plant Mol Biol. 42:251–269.

Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev. 15: 621–627.

Berger F, Twell D. 2011. Germline specification and function in plants. Ann Rev Plant Biol. 62:461–484.

Betran E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophila*. Genome Res. 12:1854–1859.

Borg M, Brownfield L, Twell D. 2009. Male gametophyte development: a molecular perspective. J Exp Bot. 60:1465–1478.

Borges F, Calarco JP, Martienssen RA. 2012. Reprogramming the epigenome in *Arabidopsis* pollen. Cold Spring Harb Symp Quant Biol. 77:1–5.

Calarco JP, et al. 2012. Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. Cell 151: 194–205.

Capra JA, Williams AG, Pollard KS. 2012. ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. PLoS Comput Biol. 8: e1002567.

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. Nat Rev Genet. 14:645–660.

Chen S, Zhang YE, Long M. 2010. New genes in *Drosophila* quickly become essential. Science 330:1682–1685.

Domazet-Loso T, Brajkovic J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 23:533–539.

Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. Genome Res. 13:2213–2219.

Domazet-Loso T, Tautz D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature 468: 815–818.

Fan C, Vibranovski MD, Chen Y, Long M. 2007. A microarray based genomic hybridization method for identification of new genes in plants: case analyses of *Arabidopsis* and *Oryza*. J Integr Plant Biol. 49: 915–926.

Furner IJ, Matzke M. 2011. Methylation and demethylation of the *Arabidopsis* genome. Curr Opin Plant Biol. 14:137–141.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. 148: 993–1003.

Honys D, Renak D, Twell D. 2006. Male gametophyte development and function. Floriculture, ornamental and plant biotechnology: advances and topical issues, 1st ed. p. 76–87.

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

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569–573.

Johnson MA, Bender J. 2009. Reprogramming the epigenome during germline and seed development. Genome Biol. 10:232.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res. 20:1313–1326.

Kleene KC. 2001. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mech Dev. 106: 3–23.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci U S A. 103:9935–9939.

Lisch D. 2009. Epigenetic regulation of transposable elements in plants. Ann Rev Plant. 60:43–66.

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we know. Ann Rev Genet. 47: 307–333.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 3:e357.

McCormick S. 1993. Male gametophyte development. Plant Cell 5: 1265–1275.

McCormick S. 2004. Control of male gametophyte development. Plant Cell 16:S142–S153.

Morgante M, et al. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet. 37:997–1002.

Quint M, et al. 2012. A transcriptomic hourglass in plant embryogenesis. Nature 490:98–101.

Schmid M, et al. 2005. A gene expression map of Arabidopsis development. Nat Genet. 37:501–506.

Schoft VK, et al. 2009. Induction of RNA-directed DNA methylation upon decondensation of constitutive heterochromatin. EMBO Rep. 10: 1015–1021.

Slotkin RK, et al. 2009. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell 136:461–472.

Soumillon M, et al. 2013. Cellular Source and Mechanisms of High transcriptome complexity in the mammalian testis. Cell Rep. 3: 2179–2190.

Trapnell C, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28:511–515.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111.

Vibranovski MD, Zhang Y, Long M. 2009. General gene movement off the X chromosome in the *Drosophila* genus. Genome Res. 19: 897–903.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A. 103:3220–3225.

Volff JN. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28: 913–922.

Wang W, Brunet FG, Nevo E, Long M. 2002. Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 99:4448–4453.

Wang W, et al. 2006. High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18:1791–1802.

Wang W, Yu H, Long M. 2004. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. Nat Genet. 36:523–527.

Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. 2013. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol Evol. 5:439–455.

Wu D-D, Irwin DM, Zhang Y-P. 2011. De novo origin of human protein-coding genes. PLoS Genet. 7:e1002379.

Wu D-D, Zhang Y-P. 2013. Evolution and function of de novo originated genes. Mol Phylogenet Evol. 67:541–545.

Xie C, et al. 2012. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. PLoS Genet. 8:e1002942.

Zhang Y, Wu Y, Liu Y, Han B. 2005. Computational identification of 69 retroposons in *Arabidopsis*. Plant Physiol. 138: 935–948.

Zhang YE, Vibranovski MD, Krinsky BH, Long M. 2010. Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. Genome Res. 20:1526–1533.

Zhou Q, et al. 2008. On the origin of new genes in *Drosophila*. Genome Res. 18:1446–1455.

Zhu Z, Zhang Y, Long M. 2009. Extensive structural renovation of retrogenes in the evolution of the Populus genome. Plant Physiol. 151: 1943–1951.

Zou C, Lehti-Shiu MD, Thomashow M, Shiu S-H. 2009. Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. PLoS Genet. 5:e1000581.

Associate editor: Michael Purugganan