Supporting Information

Gao and Lynch 10.1073/pnas.0911093106

SI Text

Data Resources. For each species, the genome, gene, protein sequences, and annotations were directly downloaded from the following sources: TAIR v6.0 for *A. thaliana* (1); TIGR v4.0 for *O. sativa* (2); Genoscope for *Vitis vinifera* (3); JGI v1.0 for *Sorghum bicolor* (http://genome.jgi-psf.org/Sorbi1/Sorbi1.download.ftp.html); WormBase ws174 for *C. elegans* and *C. briggsae* (www.wormbase.org); FlyBase v r5.5 for *D. melanogaster*; v r1.1 for *D. virilis* and v r1.3 for *D. willistoni* (4); UCSC Table Browser version caeJap1, March 2008 for *C. japonica* (5); and Ensembl release 50 for *H. sapiens*, *M. musculus*, and *Bos taurus* (6). For the species whose sequence data were obtained from the UCSC Table Browser (i.e., *C. japonica*), the coding sequences (CDS), exon and intron sequences were extracted by using the UCSC Table Browser tool. For *A. thaliana*, *O. sativa*, *V. vinifera*, *S. bicolor*, *H. sapiens*, *M. musculus*, and *B. taurus*, the CDS, exon and intron sequences were parsed with customized Perl scripts from genome sequences, based on corresponding gene annotation files. For *D. melanogaster*, *D. willistoni*, and *D. virilis*, the CDS, exon, and intron sequences were directly downloaded from FlyBase (4). The coding sequences of *B. taurus* were retrieved through BiomMart in Ensembl (7).

Number of Substitutions per Silent Site in Internally Duplicated Genes. Under the assumption that the number of substitutions per silent site has a linear relationship with the divergence time of duplicated regions, we quantified the substitutions per silent site (S) to estimate the age of each internal duplication event. For every gene with internal duplications, we partitioned internal duplications into two types: (*i*) both duplicated regions located in exons only and (*ii*) one sequence in the duplicated pair located in an intron and the other in an exon. For the first case, the YN00 algorithm in PAML package was used to calculate S (substitutions per synonymous site) (8). To minimize the noise from poorly aligned regions, only unambiguously aligned regions were retained and joined by using the PPAT program (Volker Brendel, personal communication). For the second case, we estimated the substitutions at neutral sites between intron and exon duplications; however, the existing methods of computing S between two coding regions cannot be applied here, because the neutral sites in exons and introns are of a different nature. Substitutions at most intron nucleotide positions are assumed to be neutral, except for splice sites and branching points (we exclude 5 nt at two ends of intron, but not any sequences around branch-point, because of a lack of knowledge of its position), but in exons, only synonymous sites can be largely assumed to be neutral. To overcome the lack of suitable methods for computing this S, we took the following approach.

The observed substitutions between exon and intron in a duplication pair can come from two possible sources: substitutions in the intron sequence and/or substitutions at the same site in exon sequences. Therefore, we have

$$2D = S + P_sS + (1 - P_s)N_{\text{exon}}, \quad [1]$$

where D is the number of substitutions per site since the split between duplicated exon and intron; $2D$ is the total divergence down both copies; S, which is what we want to estimate, is the number of substitutions per site in a duplicated intron region, also assumed to be the substitutions per synonymous site in the duplicated exon copy; N_{exon} is the number of substitutions at nonsynonymous sites; and P_s is the fraction of synonymous sites out of all nucleotide positions in the duplicated gene (typically, ≈0.25, but determined from the codon usage in actual sequences).

In Eq. **1**, the distance between an exon and an intron ($2D$) is the addition of two sources of substitutions: the substitutions estimated from all sites in an intron (S) and from all sites in the exon. The latter component is the averaged substitutions from synonymous sites, P_sS, and nonsynonymous sites, $(1 - P_s)N_{exon}$, in the exon, weighted by the fraction of synonymous (P_s) and nonsynonymous sites ($1 - P_s$). N_{exon} can be estimated from $N_{exon} = \omega S$, where ω is the nonsynonymous/synonymous rate ratio obtained when compared with an orthologous gene. Therefore, Eq. **1** can be rewritten as:

$$2D = S + P_sS + (1 - P_s)\omega S. \quad [2]$$

Then, S can be derived using the following formula:

$$S = \frac{2D}{1 + P_s + (1 - P_s)\omega}. \quad [3]$$

The parameters on the right side of Eq. **3** are computed as follows: The nucleotide divergence in a duplicated region between an exon and an intron (D) is calculated by the Kimura two-parameter substitution model (9). We used the YN00 method in the PAML package to calculate the nonsynonymous/synonymous substitution ratio ω and count the number synonymous and nonsynonymous sites (8), using orthologous genes from out-group taxa. To search for orthologous genes, we used *A. thaliana* and *O. sativa* as reference genomes to each other; human and mouse as reference genomes to each other; *D. willistoni* genome as the reference genome for *D. melanogaster*; and *C. briggsae* as the reference genome for *C. elegans*.

In some cases, one internal duplication pair in a single gene was broken into multiple duplication pairs. Such discontinuities can arise for multiple reasons: (*i*) secondary sequence rearrangement, insertions, and/or deletions may modify the duplicated sequence; (*ii*) internal duplications spanning more than one exon or intron unit may be recognized as more than one duplication pair because individual exons and introns were used as Blast query in this study; and (*iii*) artifact from BLAST programs. When there were multiple duplication pairs in a single gene, we need to decide how many internal duplication events those duplicated fragments came from and the age of each event. Here, we assumed that two pairs of duplicated regions in a single gene resulted from a single duplication event if their S estimates were not significantly different (Student's t test), and we merged those duplicated regions by taking the average of S as the final S for this duplication event. If the S values from different regions were significantly different, they were considered as independent internal duplication events in one gene. In rare cases, we found regions that have been internally duplicated multiple times. Multiple duplications (>2) in the same region complicate the inference of the real process of duplication, thus we removed those genes (18 genes in *A. thaliana*, 18 genes in *O. sativa*, none in C. *elegans* and *D. melanogaster*, 14 genes in *M. musculus*, and 35 genes in *H. sapiens*). The exclusion of such genes should not affect our conclusions, because they are only a minor portion of all internally duplicated genes.

Some internally duplicated genes may later experience another round of complete gene duplication such as through genome duplication, or segmental duplication. In such cases, the resulting paralogous genes all share the same pattern of internal

duplication, because the episode of internal duplication predates the polyploidization or segmental duplication event. To prevent an inflated counting of independent internal duplication events, we removed paralogous copies of internally duplicated genes when we calculated the age of duplications. Specifically, a single linkage clustering method was used to cluster paralogs [BLASTP (10) E $\leq 10^{-5}$; alignment coverage $\geq$80% of the length of query and hit], and only one internally duplicated gene from each cluster was retained for further demographic study.

Birth and Death Rates of Internal Duplications. The birth and death rates of internal duplications provide insight into the power of gene duplication as an evolutionary force (11). The theoretical principle for the calculation follows previously published methods

$$\ln(n_s) = S\ln(1 - D) + \ln\left(\frac{BD}{D - B}\right),$$

where n_s is the number of internal duplications at each S category, and B and D are the rates of internal gene duplication birth and loss per time interval on the scale of S. The slope of the least-squares regression of $\ln(n_s)$ over S is $\ln(1 - D)$, which is derived from plots in Fig. 1. d is the instantaneous loss rate of duplicates, defined by

$$1 - D = e^{-dS}$$

The bin size of S in Fig. 1 is 0.025. To infer instantaneous d, we now have

$$d = -\ln(1 - D_{0.025})/0.025,$$

where $D_{0.025}$ is the death rate of internal duplication per gene on a time scale of divergence at silent sites of 2.5%. The half-life of internal duplicated genes ($S_{0.5}$, the time scale at S when 50% of duplications are removed) is defined to be

$$S_{0.5} = -\ln(0.5)/d.$$

With a calibrated molecular clock (12, 13), $S_{0.5}$ can be translated into units of absolute time: $t_{0.5} = S_{0.5}/2\mu$, where μ is the number of neutral substitutions per site per MY.

The birth rate of internal gene duplications is estimated from n_B (the observed number of duplications in the most recent time span S). The value of the birth rate at time span S can be estimated as

$$B = \frac{n_B dS}{n(1 - e^{-dS})},$$

where n is the total number of genes in the dataset. The birth and death rates of complete gene duplications reported in preceding papers were estimated on a time scale for which $S = 0.01$ (11, 12). To make a meaningful comparison, we estimate the birth (B) and death (D) rates of internal gene duplications based on the same time span ($S = 0.01$). We have $D_{0.01} = 1 - e^{-d\cdot 0.01}$ and

$$B_{0.01} = \frac{n_{B_{0.01}} d\cdot 0.01}{n(1 - e^{-d\cdot 0.01})},$$

where $n_{B_{0.01}}$is the observed number of duplications in the most recent time span at which $S = 0.01$. $D_{0.01}$ and $B_{0.01}$, referred to as D and B in *Results* and Table 2, are the death and birth rates of internal duplications per gene on a time scale of divergence at silent sites of 1% ($S = 0.01$), respectively. Standard errors of parameters in Table 2 (i.e., B, D, d, and $t_{0.5}$) were estimated with the delta method (14).

Identification of New Introns with cDNA Evidence. We searched for cDNA evidence to confirm that new introns created by the internal gene duplication were indeed functioning as introns, and not simply annotation artifacts. cDNA sequences (including ESTs) from the six species examined were downloaded from GenBank (v162) (15). The exon nucleotide sequences flanking each new intron were joined and searched against the cDNAs with BLASTN. Intron splicing was confirmed if the joint sequence of both flanking exons had a cDNA matched (identity $\geq$98%, number of gaps $<$2).

Purifying Selection Analysis of Internally Duplicated Genes. For each internally duplicated gene with new intron, we used its protein sequence alignment with its orthologous genes in other two close related species, and converted them into corresponding codon alignment with the PAL2NAL program (16). The unrooted tree of these three gene members was constructed. Codeml program in the PAML package was applied to compare two models: The null hypothesis model is that the N/S of the internally duplicated gene is fixed to one; and the alternative model is that the N/S of the internally duplicated gene is estimated from data and less than one. The likelihood ratio test was performed to evaluate these two models (17).

1. Swarbreck D, et al. (2008) The *Arabidopsis* Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Res* 36:D1009–D1014.
2. Yuan Q, et al. (2005) The institute for genomic research *Osa*1 rice genome annotation database. *Plant Physiol* 138:18–26.
3. Jaillon O, et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467.
4. Wilson RJ, Goodman JL, Strelets VB (2008) FlyBase: Integration and improvements to query tools. *Nucleic Acids Res* 36:D588–D593.
5. Karolchik D, et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32:D493–D496.
6. Hubbard TJ, et al. (2007) Ensembl 2007. *Nucleic Acids Res* 35:D610–D617.
7. Smedley D, et al. (2009) BioMart: Biological queries made easy. *BMC Genomics* 10:22.
8. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17:32–43.
9. Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217:624–626.
10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
11. Lynch M (2007) *The Origins of Genome Architecture* (Sinauer, Sunderland, MA).
12. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.
13. Li W-H (1999) *Molecular Evolution* (Sinauer, Sunderland, MA).
14. Lynch M, Walsh JB (1998) *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, MA).
15. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. *Nucleic Acids Res* 36:D25–D30.
16. Suyama M, Torrents D, Bork P (2006) PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34:W609–W612.
17. Yang Z, Nielsen R, Hasegawa M (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol* 15:1600–1611 .
18. Guyot R, et al. (2009) Microcollinearity in an ethylene receptor coding gene region of the *Coffea canephora* genome is extensively conserved with *Vitis vinifera* and other distant dicotyledonous sequenced genomes. *BMC Plant Biol* 9:22.
19. Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc Natl Acad Sci USA* 86:6201–6205.
20. Cutter AD (2008) Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. *Mol Biol Evol* 25:778–786.
21. Hahn MW, Han MV, Han SG (2007) Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet* 3:e197.
22. Waterston RH, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562 .
23. Chaw SM, Chang CC, Chen HL, Li WH (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *J Mol Evol* 58:424–441.
24. Li WH, Tanimura M, Sharp PM (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J Mol Evol* 25:330–342.

Table S1. Collection of internally duplicated genes in each species

Section	Filtering process	*A. thaliana*	*O. sativa*	*C. elegans*	*D. melanogaster*	*M. musculus*	*H. sapiens*
Total genes	Total transcripts	34,303	63,514	23,258	20,914	44,123	61,043
	After removing TE genes	31,791	50,307	23,237	20,914	42,516	55,254
	Total loci	30,359	43,475	20,101	14,141	26,772	30,451
	Protein-coding genes	26,514	42,802	17,830	9,221	23,739	21,481
Internally duplicated genes	Internal duplicated transcripts (repeat masked)	3,423	7,834	2,197	2,920	4,779	7,487
	Internal duplicated loci	3,073	6,966	1,791	1,678	2,747	3,855
	Protein-coding genes	2,868	6,952	1,478	978	2,703	3,571
Ratio (internally duplicated protein-coding genes/total protein-coding genes), %		10.8	16.2	8.3	10.6	11.4	16.6

The technical details for obtaining the data in this table are described in *Materials and Methods*. In the section of total genes, the numbers of genes/transcripts after each filter process are shown. Annotated gene transcripts in each species were counted.Transposable element (TE) genes were excluded. Total gene loci were counted after removing alternative transcripts. Protein-coding genes are identified based on gene annotation. In the section of internally duplicated genes, internally duplicated genes were identified when there were homologous regions within the gene itself (homology between exon–exon and/or exon-intron), and the same filter process was applied as to the section of total genes, such as masking out repetitive sequences, and removing alternative transcripts and noncoding genes. The ratio of genes with internal duplications to total genes in each genome is based on protein-coding genes.

Table S2. The out-group species searched for orthologous genes of internally duplicated genes

Species	*A. thaliana*	*O. sativa*	*C. elegans*	*D. melanogaster*	*M. musculus*	*H. sapiens*
1 (close)	V. *vinifera* (108–117 MY) (18)	*S. bicolor* (50 MY) (19)	*C. briggsae* (18 MY) (20)	*D. willistoni* (60 MY) (21)	*H. sapiens* (75 MY) (22)	*M. musculus* (75 MY) (22)
2 (distant)	*O. sativa* (140–200 MY) (19, 23)	*A. thaliana* (140–200 MY) (19, 23)	*C. japonica* (30 MY) (20)	*D. virilis* (65 MY) (21)	*B. taurus* (80 MY) (24)	*B. taurus* (80 MY) (24)

The speciation times are in parentheses.

Table S3. Summary statistics of three classes of duplications in six genomes

Classes of duplications	*A. thaliana*	*O. sativa*	*C. elegans*	*D. melanogaster*	*M. musculus*	*H. sapiens*
Exon–exon	2,868	6,192	1,556	1,571	7,154	3,096
Exon–intron	379	1,290	445	200	852	1,233
Intron–intron	1,830	4,742	3,426	714	7,759	3,925
Total loci with internal duplication	3,773	10,563	4,675	2,167	9,058	5,175
Ratio, %	12.4	24.3	23.3	15.3	33.8	17.0

In this study, we did not include intron–intron homology as internal gene duplication to avoid the possible contamination by the unidentified repetitive elements. Observing such unidentified repeat elements in exon regions is assumed to be much less likely because selection pressure would preserve gene integrity. Ratio = total loci with internal duplication/total loci in genome (see Table S1).

Other Supporting Information Files

Dataset S1 (PDF)