Modeling gene and genome duplications in eukaryotes

Steven Maere*, Stefanie De Bodt*, Jeroen Raes, Tineke Casneuf, Marc Van Montagu, Martin Kuiper, and Yves Van de Peer†

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

Contributed by Marc Van Montagu, February 9, 2005

Recent analysis of complete eukaryotic genome sequences has revealed that gene duplication has been rampant. Moreover, next to a continuous mode of gene duplication, in many eukaryotic organisms the complete genome has been duplicated in their evolutionary past. Such large-scale gene duplication events have been associated with important evolutionary transitions or major leaps in development and adaptive radiations of species. Here, we present an evolutionary model that simulates the duplication dynamics of genes, considering genome-wide duplication events and a continuous mode of gene duplication. Modeling the evolution of the different functional categories of genes assesses the importance of different duplication events for gene families involved in specific functions or processes. By applying our model to the *Arabidopsis* genome, for which there is compelling evidence for three whole-genome duplications, we show that gene loss is strikingly different for large-scale and small-scale duplication events and highly biased toward certain functional classes. We provide evidence that some categories of genes were almost exclusively expanded through large-scale gene duplication events. In particular, we show that the three whole-genome duplications in *Arabidopsis* have been directly responsible for >90% of the increase in transcription factors, signal transducers, and developmental genes in the last 350 million years. Our evolutionary model is widely applicable and can be used to evaluate different assumptions regarding small- or large-scale gene duplication events in eukaryotic genomes.

Arabidopsis | functional categories | gene retention

Thirty-five years ago, Susumu Ohno (1) outlined the potential role of gene duplication as the driving force behind the evolution of increasingly complex organisms. Recent analysis of complete eukaryotic genome sequences has revealed that gene duplication has indeed been rampant (2–4). Furthermore, many eukaryotic organisms had their whole genome duplicated, sometimes more than once (5, 6). In particular such large-scale gene duplication events have been considered of major importance for evolution and increase in biological complexity (1, 7–10).

Lynch and Conery (2) were among the first to investigate the overall degree of gene duplication and gene loss in completely sequenced genomes. When the number of duplicated pairs of genes is plotted against their age, inferred from the number of synonymous substitutions per synonymous site (K_S), the resulting age distributions exhibit a typical L shape, with many recently duplicated genes and much fewer older duplicates. Based on these age distributions, Lynch and Conery (2) suggested a steady-state stochastic birth–death model for the dynamics of duplicate populations, from which they inferred the overall rate of gene duplication and gene loss. However, the gene birth and death model proposed by Lynch and Conery (2) does not take into account larger-scale gene duplication events, such as paleopolyploidy events.

Here, we propose a generally applicable evolutionary model that simulates the birth and death of genes based on observed age distributions of duplicates, considering small-scale, continuously occurring local duplication events (hereafter referred to as 0R) and duplication events affecting the whole genome. In the present study, this model is applied to the *Arabidopsis* genome. There is compelling evidence based on the identification and delineation of intergenomic homology and phylogenetics that the *Arabidopsis* genome has been duplicated three times (events hereafter referred to as 1R, 2R, and 3R) during the last ≈350 million years (11–14). Because *Arabidopsis* has undergone several well documented rounds of genome duplication, it is an ideal model system to study gene retention that occurs after ancient polyploidy events versus small-scale gene duplication events. Furthermore, by applying this computational model to different functional categories of genes, we can assess the importance of different gene duplication events for the evolution of specific gene functions or biological processes and pathways.

The aims of our study were fivefold: (*i*) to develop an evolutionary model that can take into account whole-genome duplication events in addition to the continuous mode of duplication, (*ii*) to use this model to investigate whether there is a difference in gene loss for genes created during small-scale (continuous) or large-scale (global) duplication events, (*iii*) to investigate whether duplicated genes indeed form a functionally biased set in small-scale and large-scale gene duplication events, (*iv*) to investigate whether gene decay and gene retention were similar for the successive whole-genome duplication events in *Arabidopsis*, and (*v*) to infer the number of *Arabidopsis* genes before the gene and genome duplication events considered in the present study.

Methods

Identification of Paralogs. An all-against-all protein sequence similarity search was performed by using BLASTP (with an *E*-value cutoff of e^{-10}) (15). Sequences alignable over a length of 150 amino acids with an identity score of 30% were defined as paralogs, according to ref. 16. Gene families were built through single-linkage clustering.

Dating of Paralogous Gene Pairs. Synonymous substitutions do not result in amino acid replacements and are, in general, not under selection. Consequently, the rate of fixation of these substitutions is expected to be relatively constant in different protein-coding genes and, therefore, to reflect the overall mutation rate. As a result, the fraction of synonymous substitutions per synonymous site (K_S) is used to estimate the time of duplication between two sequences. All pairwise alignments of the paralogous nucleotide sequences belonging to a gene family were made by using CLUSTALW (17), with the corresponding protein sequences as alignment guides. Gaps and adjacent divergent positions in the alignments were removed. K_S estimates were obtained with the CODEML program (18) of the PAML package (19). Codon frequencies were calculated from the average

Abbreviation: GO, Gene Ontology.

*S.M. and S.D.B. contributed equally to this work.

†To whom correspondence should be addressed. E-mail: yves.vandepeer@psb.ugent.be.

www.pnas.org/cgi/doi/10.1073/pnas.0501102102

nucleotide frequencies at the three codon positions (F3 × 4), whereas a constant K_N/K_S (nonsynonymous substitutions per nonsynonymous site over synonymous substitutions per synonymous site, reflecting selection pressure) was assumed (codon model 0) for every pairwise comparison. Calculations were repeated five times to avoid incorrect K_S estimations because of suboptimal local maxima.

Building Age Distributions of Duplicated Genes in *Arabidopsis*. Only gene pairs with a K_S estimate of <5 were considered for further evaluation. Large gene families were subdivided into subfamilies for which K_S values between genes did not exceed a value of 5. It is assumed that a gene family of n members originates from $n - 1$ retained single gene duplications, whereas the number of possible pairwise comparisons (K_S measurements) within a gene family is $[n(n - 1)]/2$. To correct for the redundancy of K_S values when building the age distribution for duplicated genes, we use an approach similar to that adopted by Blanc and Wolfe (20) (*Supporting Methods*, which is published as supporting information on the PNAS web site).

Functional Classification of the Paranome. The Gene Ontology (GO) annotation for *Arabidopsis thaliana* was downloaded from The Arabidopsis Information Resource (www.arabidopsis.org; version April 10, 2004) and remapped to the plant-specific GO Slim ontology (www.geneontology.org) (21). A few extra subdivisions were added to the GO Slim "structural molecule activity" and "transporter activity" categories (see Fig. 5, which is published as supporting information on the PNAS web site). Genes mapped to a particular GO Slim category were also explicitly included into all parental categories. Individual gene family K_S distributions were only added to a particular GO Slim category K_S distribution if >20% of the genes in the family were annotated to that category (*Supporting Methods*, Figs. 5, 6, and 7, and Table 1, which are published as supporting information on the PNAS web site). GO Slim categories containing <50 retained duplicates (i.e., very sparse distributions) were *a priori* discarded as candidates for further modeling. After modeling, some other categories were removed for interpretation and discussion because of low-confidence parameter estimates (*Supporting Methods* and Table 2, which is published as supporting information on the PNAS web site).

Population Dynamics Model for Duplicate Genes in *Arabidopsis*. Our model simulates the dynamics of a population of duplicated genes, as reflected by their K_S age distribution, in 50 time steps, each time step corresponding to an average K_S interval of 0.1 (Fig. 1). The principal equations of the model are summarized below.

$$D_0(1, t) = \nu\left[\sum_{x'=1}^{\infty} D_{\text{tot}}(x', t - 1) + G_0\right]$$

$$D_i(1, t) = \left[\sum_{x'=1}^{\infty} D_{\text{tot}}(x', t - 1) + G_0\right]\delta(t, t_i) \quad i = 1, 2, \text{ or } 3$$

$$D_i(x, t) = D_i(x - 1, t - 1)[x/(x - 1)]^{-\alpha_i}\, x > 1 \quad i = 0, 1, 2, \text{ or } 3$$

$$D_{\text{tot}}(x, t) = \sum_i D_i(x, t) \qquad [1]$$

In this set of equations, $D_i(x, t)$ stands for the number of retained duplicates in the ith duplication mode ($i = 0$ for the 0R, $i = 1$, 2, and 3 for 1R, 2R, and 3R, respectively) having an age x (measured in 0.1 synonymous substitutions per synonymous site equivalents) at time step t in the simulation. $D_{\text{tot}}(x, t)$ is the total number of duplicates of age x at time step t, which is fed back to time step $t + 1$. G_0 represents the number of ancestral genes at $K_S = 5$ (see *Supporting Methods* for details). The first equation describes the birth of duplicates in the continuous mode at a birth rate of ν duplicates per gene and per time step. Because the birth rate can be assumed to be the same for all GO categories, ν was estimated once from the category with the highest resolution, namely the whole-paranome category (see *Results and Discussion*). The same birth rate was then used throughout all simulations for all functional categories, reducing the number of parameters that needed to be optimized by one. The second equation models the discrete (hence the δ function) large-scale duplication events at time steps t_i. The third equation models the loss of duplicates from one time step to the next, with power-law decay constants α_i. The last equation ensures the coupling between all duplication modes.

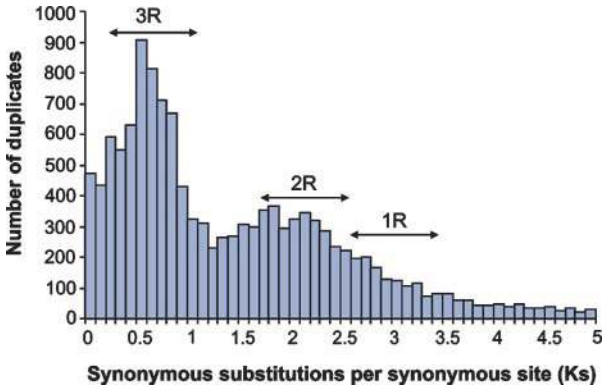


Fig. 1. Age distribution of the *Arabidopsis* paranome based on K_S values. 1R, 2R, and 3R refer to the three genome-wide duplication events that have occurred in *Arabidopsis* or its predecessors (12, 13).

The equations (Eq. **1**) are recursively evaluated 50 times in the course of a single simulation. The resulting distribution $D_{\text{tot}}(x, 50)$ is the simulated present-day age distribution of the duplicate population for a given choice of parameters α_i, which are the parameters to be optimized. However, $D_{\text{tot}}(x, 50)$ is an age distribution featuring discrete large-scale duplication peaks as opposed to the relatively wide peaks observed in the K_S distributions. The modeled age distribution of retained duplicates $D_{\text{tot}}(x, 50)$ is converted to a K_S distribution by Poisson distributing the duplicate count of each age bin (see *Supporting Methods*). The net effect is a broadening of discrete peaks in the modeled age spectra, increasing with age, as observed in the initially obtained K_S distributions (Fig. 1). The modeled K_S distribution is calculated from the modeled age-distribution as follows:

$$D'(x, \boldsymbol{\alpha}) = \sum_{\lambda=1}^{\infty} D_{\text{tot}}(\lambda, 50)\cdot\lambda^x e^{-\lambda}/x!, \qquad [2]$$

where x is the K_S bin, λ is the age bin, $D_{\text{tot}}(\lambda, 50)$ is the modeled age-distribution after 50 time steps and $D'(x, \boldsymbol{\alpha})$ is the corresponding model K_S distribution after Poisson smoothing, with decay parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$. The model parameters α_i are optimized to give the best possible fit of $D'(x, \boldsymbol{\alpha})$ to the observed K_S distribution. A classic Monte Carlo Simulated Annealing optimization strategy was used with an exponential temperature decay (22, 23) (see *Supporting Methods* and Fig. 8, which is published as supporting information on the PNAS web site). The parameters α_i were optimized 10 times for each functional category to monitor the convergence of the parameter

EVOLUTION

estimates. Confidence intervals for the parameters α_i were calculated based on the covariance matrix for the best fit (see *Supporting Methods* and Table 2). GO Slim categories with more than two low-confidence parameter estimates were discarded in all further analyses (colored gray in Figs. 5 and 6; see also Table 2).

Results and Discussion

The age distribution of all duplicated genes of *Arabidopsis*, including all 3,472 gene families (see Table 1), clearly shows two peaks or waves (Fig. 1), of which the youngest can be attributed to the youngest duplication event (12–14), whereas the second wave corresponds to the two older genome duplications (12, 13) that have become almost indistinguishable (see below). In previous studies, the second wave had been missing mainly either because large multigene families had been excluded from the analyses (2) or because only small K_S values had been considered (20). As shown earlier, many of the genes in these waves lie in so-called paralogons, i.e., intragenomic homologous segments (12–14). However, many duplicates that originated from large-scale duplication events are found outside those paralogons, particularly for the older genome duplication events, because of gene translocation events. These duplicates were largely ignored in previous studies (24, 25) because they cannot be distinguished from duplicates generated in the continuous mode. In our model, this problem is circumvented by simulating, rather than enumerating, the number of duplicates generated in each duplication mode, regardless of whether they belong to paralogons.

The Functional Landscape of the *Arabidopsis* Paranome. To investigate the relative impact of small-scale and large-scale gene duplications on different functional categories of genes in *Arabidopsis*, we subdivided the global K_S distribution according to the GO Slim ontology (21). Based on the current status of the GO annotations and on the robustness of the age distributions for different thresholds (see *Supporting Methods* and Fig. 7), we chose to add individual gene families to a particular GO Slim category distribution if >20% of the genes in the family were assigned to that category. Despite using a 20% threshold for individual gene families, the minimum overall percentage of genes in a GO Slim class distribution that are annotated accordingly in GO is 58% (for the "carbohydrate binding" category) (Table 1). We do recognize the risk of assigning gene families to a particular GO Slim function or process that are only partially involved in that function or process. Although we found no direct evidence of such cases, the K_S distribution for, e.g., the "response to abiotic stimulus" category should be considered as the K_S distribution for gene families that during their history have been important in the evolution of the response to abiotic stimulus rather than the distribution for duplicate genes involved in the response to abiotic stimulus *sensu stricto*. The size of the gene families, the total number of genes ascribed to a functional category based on these gene families, the proportion of those genes directly annotated by GO to that functional category, and the number of retained duplicates and the estimated number of ancestral genes for that functional category can be found in Table 1.

Modeling Gene and Genome Duplications. To quantify the differences in K_S distribution between the GO categories, a population dynamics model was developed that is able to accurately reproduce the observed K_S distributions and characterize them in terms of only a few parameters. The model itself is described in detail in *Methods*, but the principal assumptions and potential shortcomings of our model will be considered here. Because the calibration of time since duplication versus K_S is controversial [see, for example, Lynch and Conery (2) and Koch *et al*. (26), who propose quite different rates of synonymous substitutions in dicots], all calculations were performed based on K_S time equivalents without explicit conversion to real time (*Supporting Methods*). Throughout the manuscript, time since duplication is therefore expressed in K_S time equivalents. The simulation starts at time step 1 (5.0 K_S time equivalents ago) from a number of ancestral genes G_0 (*Supporting Methods* and Table 1) and evolves this ancestral genome to the present-day size by gene duplication and gene loss, thereby creating a simulated K_S distribution. Four distinct modes of gene duplication are included, namely a continuous mode of small-scale gene duplication (0R) and three large-scale duplication modes (1R, 2R, and 3R). We assume that small-scale duplications in the continuous mode occur at a constant birth rate ν (see *Supporting Methods*). Local fluctuations of the birth rate ν with time are averaged out over longer time periods. Systematic deviations from a constant birth rate (e.g., systematic increase of birth rate with time) or prolonged time periods with a significantly altered birth rate would be reflected by the inability of our model to reproduce the observed K_S distribution. In our case, it proved to be unnecessary to make more elaborate assumptions (Occam's razor). The average birth rate ν of new duplicates was estimated to be 0.03 per gene and per 0.1 K_S time equivalent based on optimization of the model fit to the whole paranome K_S distribution for several values of ν (Fig. 9, which is published as supporting information on the PNAS web site). Our estimate is about twice as high as the one proposed by Lynch and Conery (27).

On top of the continuous duplication mode, we have modeled three whole-genome duplications occurring at time steps $t_i = 20$, 31, and 44 in the simulation (respectively 3.1, 2.0, and 0.7 K_S time equivalents ago). These values correspond to the three previously described large-scale duplication events in the evolutionary past of *Arabidopsis* (12, 13). The ages of the whole-genome duplications were estimated through simulations of the duplication history of the whole paranome for different age values. These ages were subsequently used throughout the simulations for all GO Slim categories. A model based on only two large-scale duplications, assuming that 1R did not take place, gave considerably worse fits (Fig. 2 *A* and *B*), again providing evidence that three large-scale duplications have, indeed, occurred in the evolutionary past of *Arabidopsis*. The model is able to compensate in part for the lack of genes created by 1R by increasing the retention of duplicates in the continuous mode (lower decay parameters α_0), especially for GO categories with moderate to low retention after 1R, such as the "whole paranome" category. However, categories with a high retention subsequent to 1R, such as "development," show pronounced bias in the residuals. We also assumed that the three large-scale duplication events were complete genome duplications. Although for the youngest event there is substantial evidence that at least 80% of the genome was duplicated (12–14), it is very difficult to assess whether the older large-scale duplication events were also genome-wide. The validity of our assumption can, at least to some extent, be examined by modeling alternative assumptions. For example, if we assume that the second large-scale event (2R) only affected half of the genome, the effects thereof will propagate to later time points (smaller K_S), by means of the coupling of all duplication modes. More specifically, the continuous mode of duplication will then have acted on considerably less genetic material right after 2R, resulting in the inability of the model to reproduce the duplicate count observed in the actual K_S distribution between $K_S = 1.0$ and 2.0, after 2R (Fig. 2*C*). This effect is more pronounced for GO categories with a low decay rate (or high retention) in the continuous mode. The 2R peak itself ($K_S > 2.0$) is still fitted reasonably well by lowering the 2R decay parameter α_2.

The duplicates created during the whole-genome duplication events and the continuous mode of duplication are lost with mode-specific time-dependent decay rates α_i/t ($i = 1$ for 1R, $i =$

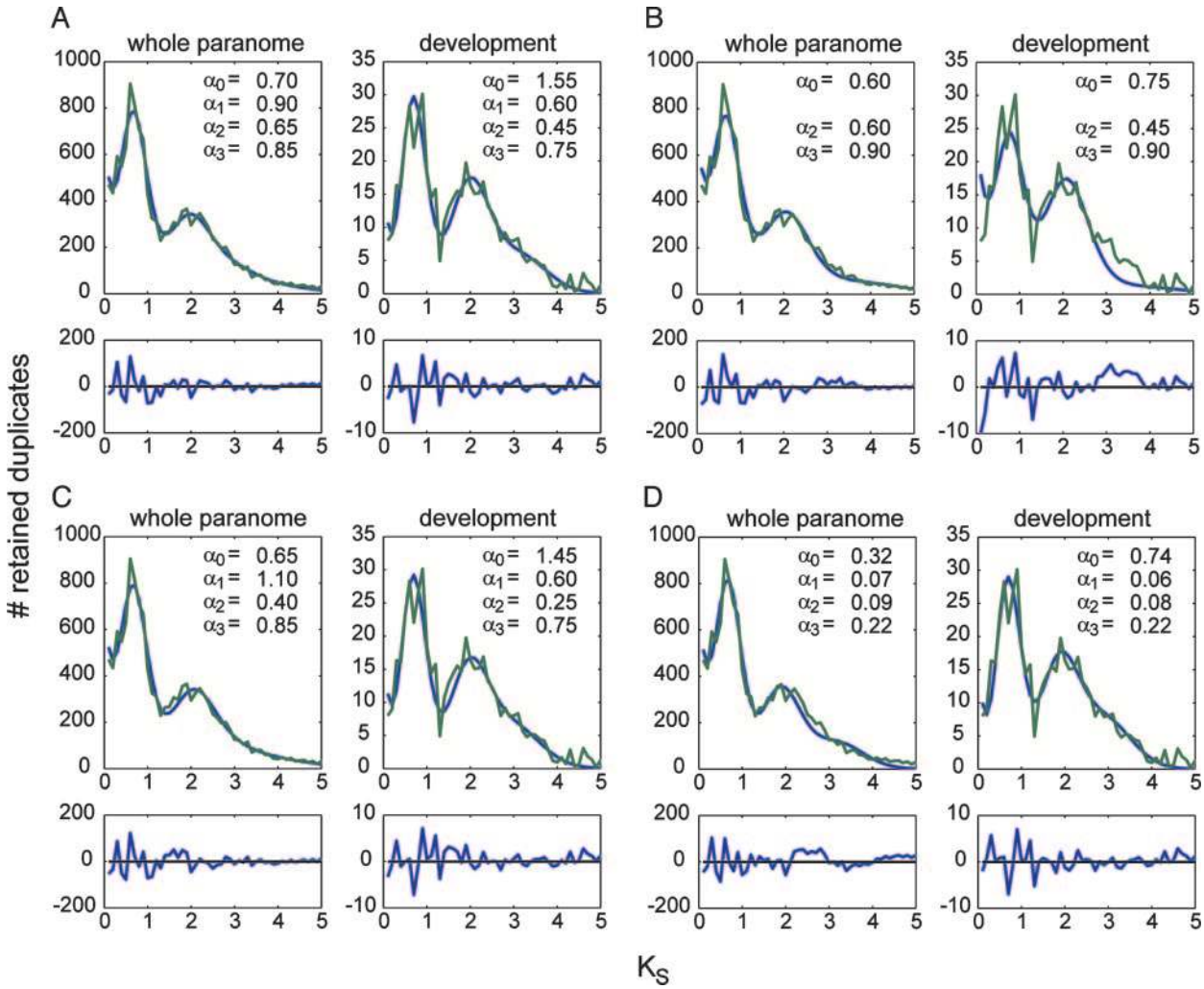


Fig. 2. Optimal fits and parameters α_i (*Upper*) and residual errors (*Lower*) for the ''whole paranome'' and ''development'' GO categories, simulated under various model assumptions. (*Upper*) The green curves show the observed K_S distributions, and the blue curves represent the simulated K_S distributions. (*Lower*) The residual error is defined as the difference between the observed and the simulated distributions. Biased residual errors, meaning that they are consistently positive or negative for prolonged K_S intervals, hint at unrealistic model assumptions. (*A*) Model fits under the assumption that there were three whole-genome duplications and that gene decay follows a power law. The residual errors show very little bias. (*B*) Model fits under the assumption that 1R did not occur. (*C*) Model fits under the assumption that 2R was partial and involved only 50% of the genome. (*D*) Model fits under the assumption that the number of retained duplicates decays exponentially.

2 for 2R, and $i = 3$ for 3R) and α_0/t (0R), respectively. A decay rate α_i/t leads to a decay of the power-law form: $D_i(t) = D_i(0)t^{-\alpha_i}$, where $D_i(t)$ represents the number of duplicates in the ith duplication mode after a time t. Compared to an exponential decay with a constant decay rate α_i, as suggested by Lynch and Conery (2), a power-law decay exhibits a flattened tail. We observed that an exponential decay model could not adequately reproduce the observed K_S distributions, in particular for high K_S values (Fig. 2*D*). Also, decay parameters α_i obtained with the exponential model steadily increase with the decreasing age of the duplication mode ($\alpha_1 < \alpha_2 < \alpha_3 < \alpha_0$), which cannot be biologically motivated. Indeed, a constant decay rate is unrealistic from a biological viewpoint. If duplicates have been retained for a longer time, it is more probable that they confer added value or fitness to the organism, which reduces their chance of being lost (28). In other words, the decay rate should asymptotically tend to zero for increasing time since duplication. This scheme allows for rapid initial gene loss that gradually evolves toward a preferential retention of older duplicates under selective constraints.

Small-Scale Versus Large-Scale Duplications and Biased Retention of Duplicates. Gene decay rates were estimated by the model through fitting of the age distributions drawn for the different functional categories (Figs. 5 and 6). Fig. 3 shows examples of the four different decay parameters, namely those for 0R, 1R, 2R, and 3R, for some specific GO classes, such as transcription, development, and secondary metabolism. A table with the decay parameters for other functional categories and for confidence values for these parameters can be found in Table 2. A clustered color representation of gene decay is shown in Fig. 4 for all GO classes that could be modeled adequately (evaluated based on confidence intervals; see Table 2).

One of the most striking observations is that, for many functional categories, gene decay rates differ considerably for genes created during large-scale (1R, 2R, or 3R) and small-scale (0R) duplication events. As a matter of fact, for a majority of GO Slim categories, an almost opposite picture is obtained for genes created during whole-genome or small-scale duplication events. Probably most prominently, gene decay is low for genes involved in kinase activity, transcription, protein binding and modifica-

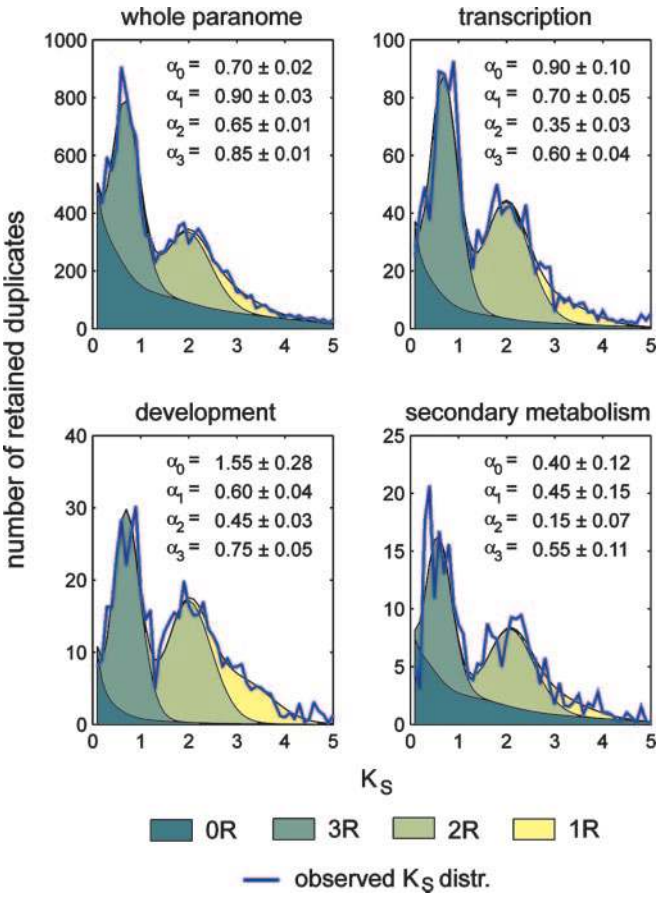


Fig. 3. Observed (blue line) versus simulated (green and yellow surface areas) K_S distributions for some GO classes discussed in the text. The parameters in the upper right corners of each graph specify the simulated decay rates for the continuous mode of gene duplication (α_0) and for the whole-genome duplications 1R (α_1), 2R (α_2), and 3R (α_3) and their confidence intervals (Table 2). The colored areas show the simulated fraction of retained duplicates created by each duplication mode as a function of K_S. Similar graphs for other functional classes can be found in Fig. 10, which is published as supporting information on the PNAS web site.

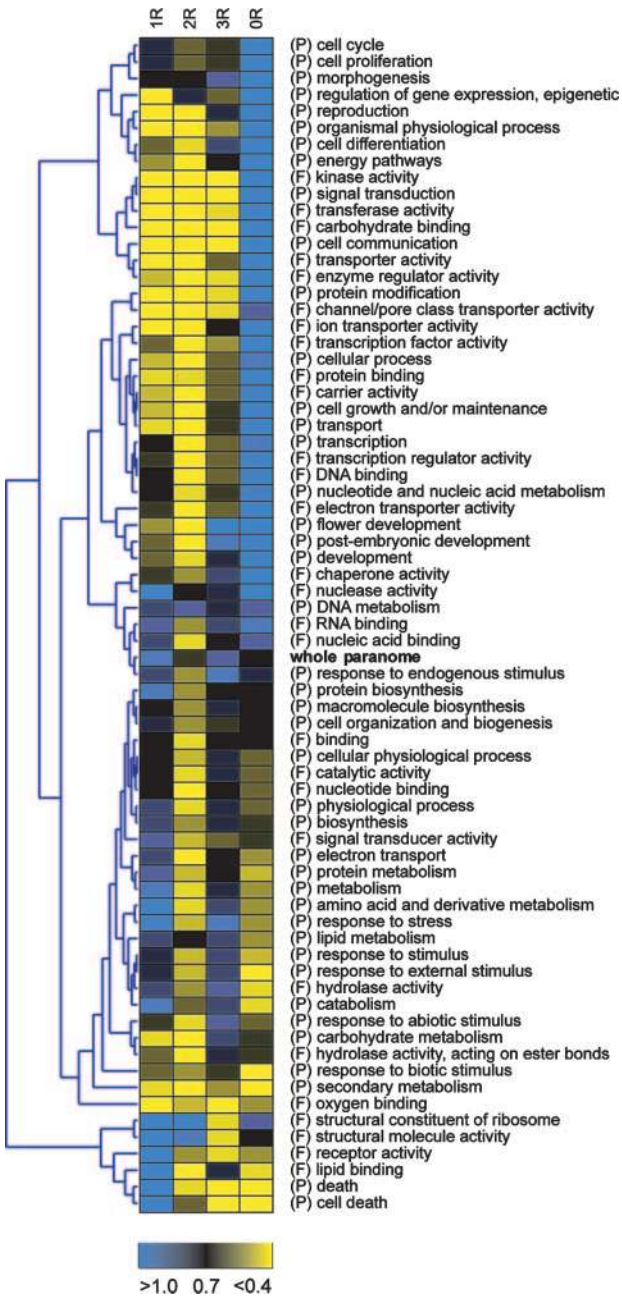


Fig. 4. Clustered color representation of the decay parameters for all duplication modes and GO Slim categories. Light blue corresponds to high gene decay or low retention, and bright yellow corresponds to low decay or high gene retention. The numerical values and confidence intervals of the decay parameters can be found in the supporting information. The decay parameter of 0.70 (black) was chosen to match the continuous-mode decay for the whole paranome. P denotes the Biological Process categories, and F denotes the Molecular Function categories.

tion, and signal transduction pathways when created in large-scale gene duplication events, whereas gene decay is very high for such genes when created by individual, small-scale duplication events (Fig. 4). Accordingly, Blanc and Wolfe (24), considering only the most recent polyploidy event in *Arabidopsis*, also observed a high retention of genes with regulatory functions, such as transcription factors, kinases, phosphatases, and calcium-binding proteins. Seoighe and Gehring (25) also found that genes involved in transcription regulation and signal transduction had a significantly higher survivability after genome duplication than other functional categories. Rapid loss of these duplicated genes after small-scale gene duplication events may be explained by the fact that regulatory genes involved in signal transduction and transcription tend to show a high dosage effect in multicellular eukaryotes (29). That transcription factors and kinases are often active as protein complexes and need to be present in stoichiometric quantities for their correct functioning is congruent with their high retention rate after whole-genome duplication events in contrast to small-scale duplication events (30, 31). On the other hand, genes belonging to other functional categories show a markedly different behavior and are retained in excess after large-scale and small-scale duplication events. Examples are genes involved in secondary metabolism and response to biotic stimulus. Because plants are sessile organisms, secondary metabolite pathways and genes governing the response to biotic stimulus have been crucial to develop survival strategies against herbivores, insects, snails, and plant pathogens (32). The low decay rate of these genes in small- and large-scale duplication modes (Fig. 4) furthers the evidence that secondary metabolites represent important adaptive traits that are heavily selected for during evolution to protect plants against a wide variety of enemies imposing a constant need for adaptation. Genes in-

volved in conserved biological processes are generally little retained (Fig. 4). Examples are DNA metabolism genes (which includes DNA repair, DNA replication, and DNA recombination), ribosomal genes (except for 3R), nucleases, RNA binding genes, and (to a lesser extent) cell cycle genes and protein and macromolecule biosynthesis genes. Our model also shows that gene decay is not the same for different whole-genome duplication events, although the general trends are similar. For instance, gene decay occurring after the youngest duplication event (3R) seems to be higher (Fig. 4, blue coloring in the whole paranome row at column 3R) and less biased toward functional class (Fig. 4, less deviation from the mean reflected by an overall darker coloring in column 3R) than for 1R and 2R. In particular, genes encoding transcriptional regulators and genes involved in development are better retained after the second genome duplication event than after the other duplication events. This finding seems to be congruent with what is known about the rise and early diversification of the angiosperms, but this result will be discussed elsewhere.

The impact of small- and large-scale duplications on the expansion of specific functional categories of genes becomes even clearer when we consider the actual numbers of genes retained subsequent to 0R, 1R, 2R and 3R. Based on integration of the mode-specific K_S distributions (Fig. 3, colored areas), we estimate that the three genome duplication events are directly responsible for ≈90% of all transcription factors in higher plants created in the last ≈350 million years (roughly corresponding to $K_S = 5.0$) (Table 3, which is published as supporting information on the PNAS web site). Similarly, we estimate that 1R, 2R, and 3R taken together account for 92% of all developmental genes and 99% of the kinases and genes involved in signal transduction created since the time corresponding with a K_S value of 5.0. For most categories related to metabolism, stress response, or cell death, the percentage of large-scale gene duplicates ranges from 50% to 70%, reflecting the fact that these categories show relatively higher gene retention after small-scale gene duplication events.

From the simulation results, we can also infer the number of genes that was initially created in each mode. We estimate that 17,193 duplicates were created by 1R, of which 771 (or 4.4%) duplicates have been retained; 20,316 duplicates were created by 2R, of which 2,765 (13.6%) were retained; and 24,351 duplicates were created by 3R, of which 3,947 (16.2%) duplicates have survived. In contrast, 0R created 33,182 duplicates in the last 350–400 million years (12, 13) and is responsible for 5,266 (15.8%) retained duplicates (see Table 3). It is clear from these numbers that, although a considerable number of genes has been retained after gene duplication, gene loss is by far the most likely fate of duplicate genes. Overall, the three genome duplications in *Arabidopsis* have been directly responsible for ≈59% of the total number of duplicates that have been retained during the last ≈350 million years, which means that more than half of the *Arabidopsis* genome expansion, from ≈14,800 genes in the ancestral genome at time point $K_S = 5.0$ (G_0 for the whole paranome in Table 1) to ≈27,500 genes now (from GO; Table 1), is directly caused by genome duplications. Still, ≈40% of the genome expansion is caused by gradual accumulation of small-scale gene duplicates.

In conclusion, we have developed an evolutionary model that simulates the population dynamics of duplicate genes created by small- and large-scale duplication events based on their age distribution in a genome. One of the main advantages of our modeling approach is that it provides a means to study gene retention occurring after genome duplications without the need to attribute every gene to a particular duplication event. Applying our model to the *Arabidopsis* genome shows that much of the genetic material in extant plants, i.e., ≈60%, has been created by ancient genome duplication events. More importantly, it seems that a major fraction of that material could have been retained only because it was created through large-scale gene duplication events (Figs. 3 and 4). In particular, transcription factors, signal transducers, and developmental genes have been retained subsequent to large-scale gene duplication events, in particular, to the second genome duplication (2R), whereas the contribution of small-scale gene duplications to the increase of regulatory and developmental genes has been very limited. Because the divergence of regulatory genes is being considered necessary to bring about phenotypic variation and increase in biological complexity, it is tempting to conclude that such large-scale gene duplication events have indeed been of major importance for evolution in general, as suggested in refs. 1, 7, 9, 10, and 33.

We thank Ken Wolfe, Axel Meyer, Cathal Seoighe, Dirk Aeyels, and Dirk Inzé for critical comments on the manuscript. S.M. is a Research Fellow of the Fund for Scientific Research (Flanders, Belgium). S.D.B. and J.R. are indebted to the Institute for the Promotion of Innovation by Science and Technology (Flanders, Belgium) for a predoctoral and postdoctoral fellowship, respectively.

1. Ohno, S. (1970) *Evolution by Gene Duplication* (Springer, New York).
2. Lynch, M. & Conery, J. S. (2000) *Science* **290,** 1151–1155.
3. Lynch, M. & Conery, J. S. (2003) *J. Struct. Funct. Genomics* **3,** 35–44.
4. Li, W.-H., Gu, Z., Cavalcanti, A. R. O. & Nekrutenko, A. (2003) *J. Struct. Funct. Genomics* **3,** 27–34.
5. Wolfe, K. H. (2001) *Nat. Rev. Genet.* **2,** 333–341.
6. Van de Peer, Y. (2004) *Nat. Rev. Genet.* **5,** 752–763.
7. Otto, S. P. & Whitton, J. (2000) *Annu. Rev. Genet.* **34,** 401–437.
8. Wendel, J. F. (2000) *Plant. Mol. Biol.* **42,** 225–249.
9. Holland, P. W. (2003) *J. Struct. Funct. Genomics* **3,** 75–84.
10. Aburomia, R., Khaner, O. & Sidow, A. (2003) *J. Struct. Funct. Genomics* **3,** 45–52.
11. Vision, T. J., Brown, D. G. & Tanksley, S. D. (2000) *Science* **290,** 2114–2117.
12. Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M. & Van de Peer, Y. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 13627–13632.
13. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. (2003) *Nature* **422,** 433–438.
14. Blanc, G., Hokamp, K. & Wolfe, K. H. (2003) *Genome Res.* **13,** 137–144.
15. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
16. Li, W.-H., Gu, Z., Wang, H. & Nekrutenko, A. (2001) *Nature* **409,** 847–849.
17. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
18. Goldman, N. & Yang, Z. (1994) *Mol. Biol. Evol.* **11,** 725–736.
19. Yang, Z. (1997) *Comput. Appl. Biosci.* **13,** 555–556.
20. Blanc, G. & Wolfe, K. H. (2004) *Plant Cell* **16,** 1667–1678.
21. The Gene Ontology Consortium (2000) *Nat. Genet.* **25,** 25–29.
22. Metropolis, N. & Ulam, S. (1949) *J. Am. Stat. Assoc.* **44,** 335–341.
23. Kirkpatrick, S., Gelatt, C. D., Jr., & Vecchi, M. P. (1983) *Science* **220,** 671–680.
24. Blanc, G. & Wolfe, K. H. (2004) *Plant Cell* **16,** 1679–1691.
25. Seoighe, C. & Gehring, C. (2004) *Trends Genet.* **20,** 461–464.
26. Koch, M. A., Haubold, B. & Mitchell-Olds, T. (2000) *Mol. Biol. Evol.* **17,** 1483–1498.
27. Lynch, M. & Conery, J. S. (2001) *Science* **293,** 1551a.
28. Long, M. & Thornton, K. (2001) *Science* **293,** 1551a.
29. Birchler, J. A., Bhadra, U., Bhadra, M. P. & Auger, D. L. (2001) *Dev. Biol.* **234,** 275–288.
30. Papp, B., Pál, C. & Hurst, L. D. (2003) *Nature* **424,** 194–197.
31. Krylov, D. M., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. (2003) *Genome Res.* **13,** 2229–2235.
32. Chen, F., Tholl, D., D'Auria, J. C., Farooq, A., Pichersky, E. & Gershenzon, J. (2003) *Plant Cell* **15,** 481–494.
33. Postlethwait, J., Amores, A., Cresko, W., Singer, A. & Yan, Y. L. (2004) *Trends Genet.* **20,** 481–490.

Supporting Methods

Correction for redundant K_S values

A gene family of *n* members originates from *n-1* retained single gene duplications, whereas the number of possible pairwise comparisons (K_S measurements) within a gene family is *n(n-1)/2*. To correct for the redundancy of K_S values when building the age distribution for duplicated genes, we constructed tentative phylogenetic trees for each gene family with an average linkage clustering algorithm using K_S as a distance measure, similar to the approach adopted by Blanc and Wolfe (1). Starting from each gene as a separate cluster, the two clusters with the lowest mean inter-cluster K_S value (i.e. the mean of all observed K_S values (edges) between two clusters) were iteratively merged. The splits in the resulting average linkage tree represent the *n-1* retained duplication events. For each split, the *m* K_S measurements between the two merged gene clusters were added to the K_S distribution with a weight 1/*m*. In other words, all K_S estimates for a particular duplication event were added to the K_S distribution, while the total weight of a single duplication event sums up to one.

Assignment of gene families to GO Slim category K_S distributions

In order to investigate the relative impact of small-scale and large-scale gene duplications on different functional categories of genes in *Arabidopsis*, we subdivided the global K_S distribution according to the Gene Ontology (GO) annotation, which provides a standardized and hierarchical vocabulary to describe the function of genes (2). Individual gene families were included in one or more K_S distributions depending on their GO Slim annotation. In this GO Slim ontology, categories close to the leaves of the GO hierarchy are mapped onto the more general parental categories. As such, these GO Slim categories generally contain enough duplicated genes to construct reliable K_S distributions and to model their duplication history. A visual

representation of the K_S distributions for the various functional classes mapped onto the GO Slim hierarchy can be found in Figs. 5 and 6.

To assign a given gene family to a certain GO Slim class, a threshold was used that is expressed as a percentage of genes in the family assigned to that class. Using no threshold, meaning that a gene family is included in the K_S distribution of a GO Slim class as soon as one gene is assigned to that class, would be unacceptable because of the high false-positive rate in the GO annotations. Although the GO annotation for *Arabidopsis* genes has recently been improved considerably (3), it still contains errors (false positives) while for many genes the annotation is missing or incomplete (false negatives). On the other hand, a high threshold would discard too many families with incomplete annotations, leading to sparser distributions and lower sensitivity. A threshold of 40%, for example, would already require two genes in a family of three to have the same annotation, a number that is hard to reach given the current status of GO annotation and knowledge about gene function in *Arabidopsis*, especially for Biological Process categories. To decide which gene families to assign to which GO classes, we compared three thresholds, namely 10%, 20%, and 30% (Fig. 7). As can be observed, GO Molecular Function categories tend to be relatively indifferent to threshold changes (Fig. 7*C*), reflecting the fact that most genes in a family are, often electronically, annotated to the same Molecular Function. The same holds true for general Biological Process categories, such as metabolism (Fig. 7*A*). For more specific Biological Process categories, such as development, the distribution degrades more quickly when the threshold is raised, although the shape of the distribution, which is our main concern, is largely preserved (Fig. 7*B*).

Estimation of the number of ancestral genes G_0

For each GO Slim class, the number of ancestral genes G_0 existing at time point K_S = 5 was estimated as follows. Each gene family (i.e. subfamily where K_S measurements between genes do not exceed 5) in the GO Slim category K_S distribution is the progeny of a single ancestor

gene that existed at K_S = 5, which sets the ancestor count at the number of gene families included in the GO Slim distribution. To this number, we added the number of singletons (i.e. genes that did not retain any duplicates after K_S = 5) annotated to the GO Slim class to get the final estimate of G_0.

Age versus K_S distributions

An issue that needs consideration is the difference between age distributions and K_S distributions. The distribution that is initially simulated in our population dynamics model is an age distribution, featuring discrete large-scale duplication peaks as opposed to the smooth peaks, widening with time, observed in the K_S distributions. In order to fit our model to the observations, the simulated age distribution needs to be converted to a K_S distribution, which implies that we have to consider the processes that cause the growing uncertainty in K_S as a function of age. The basic process responsible for peak broadening in K_S spectra is the process of synonymous substitution, used to infer K_S values from sequence data. A given site has a particular probability per unit of time of undergoing a synonymous substitution (4, 5). For example, in an ensemble of sequence pairs with say L completely unrestricted synonymous sites, the number of synonymous substitutions after any given length of time λ (measured in K_S time equivalents, see below) will be Poisson-distributed with mean λL and standard deviation $\sqrt{\lambda L}$. Consequently, the corresponding K_S distribution will be a scaled Poisson distribution with mean λ and standard deviation $\sqrt{\lambda / L}$. As the peak width varies with the length of the sequences, the distribution of K_S values for a set of sequence pairs of varying length will be a superposition of several such scaled Poisson distributions. Furthermore, factors that impose selective constraints on synonymous sites increase the basic peak width by lowering the 'effective number' L of synonymous sites (6). E.g., sites which are only twofold degenerate are effectively counted as one-third of a synonymous site in the calculation of K_S values (6). Other

important factors influencing the effective number of synonymous sites include, but are not limited to, codon bias and RNA secondary structure constraints. Peak widening is also enhanced by errors in K_S measurement and correction for multiple substitutions. It is virtually impossible to take into account the influence of all these factors in detail. Instead, we found that, phenomenologically, the K_S distribution of sequence pairs of age λ (measured in K_S time equivalents) can be approximated by a scaled Poisson distribution with mean λ and standard deviation $\sqrt{\lambda/10}$, suggesting that the number of effective synonymous sites in an average Arabidopsis gene is only of the order of 10 (neglecting the effect of measurement errors). Please note that when λ is measured in 0.1 K_S time equivalent units instead of K_S time equivalents (which boils down to multiplying the above values for mean and standard deviation by 10 and substituting 10λ by λ), the scaled Poisson distribution reduces to a Poisson distribution with mean λ and standard deviation $\sqrt{\lambda}$, as in Eq. **2** in the article.

Another issue is the calibration of K_S versus time. Because the calibration of time since duplication versus K_S is controversial (see for example Lynch and Conery (7) and Koch *et al.* (8) who propose quite different rates of synonymous substitutions in dicots), we deliberately chose to perform all calculations based on 'K_S time equivalents' without explicit conversion to real time. A K_S time equivalent is defined as the time needed to produce an average K_S difference of 1. Working with K_S time equivalents also solves some other issues related to modeling duplication dynamics. For instance, rates such as the birth rate of duplicates in the continuous mode or the synonymous substitution rate cannot be assumed constant over physical time at an evolutionary time scale (9). They depend for example on the generation time of the *Arabidopsis* ancestors in the course of evolution. However, when measuring in K_S time equivalents, the effect of generation time on the birth rate of new duplicates is largely cancelled out, because the relation of K_S time equivalents to physical time depends on the generation time in the same fashion. Of course, this does not validate our constant birth rate assumption entirely, because other factors

(e.g. effective population size) could influence the birth rate of new duplicates and the rate of synonymous substitution in different ways. Systematic deviations from a constant birth rate (e.g. systematic increase of birth rate with K_S based time) or prolonged time periods with a significantly altered birth rate would be reflected in the inability of our model to reproduce the observed K_S distribution. In our case, it proved to be unnecessary to make more elaborate assumptions (Occam's razor).

Simulation and Optimization strategy

Our model dynamically simulates the K_S distribution of duplicated genes for a given functional category in 50 time steps, each time step corresponding to an average K_S interval of 0.1. This sampling rate gives us sufficiently high resolution with respect to the features that we want to model while keeping the computational cost minimal. Next to the continuous duplication mode (0R), three whole-genome duplications were modeled at time-steps t_i = 20, 31, and 44 in the simulation (respectively 3.1, 2.0, and 0.7 K_S time equivalents ago). The ages of the whole-genome duplications were estimated through simulations of the duplication history of the whole paranome for different age values. The resulting estimates were subsequently used throughout the simulations for all GO Slim categories although they were allowed to deviate slightly ($\pm$1 time step) during the course of a simulation. In other words, we used a tolerance of $\pm$1 time step on the age of the whole-genome duplications. This improved the ability of our model to fit the large-scale duplication peaks for different classes as well as the performance of our optimization procedure (better convergence towards global minimum, improved ability to overcome local minima).

A classic Monte Carlo Simulated Annealing optimization strategy was used with exponential temperature decay. Starting from an initial (random) guess for the parameters α,

random steps are taken in parameter space. In practice, a step size of 0.05 was employed. A step is accepted if

$$\text{rand}(1) < \exp(-\Delta\chi^2 / kT), \qquad [1]$$

with rand(1) a random number drawn uniformly from the interval [0,1], $\Delta\chi^2$ the change in optimization potential and kT the simulated annealing parameter (temperature), which gradually decreases over four orders of magnitude (from kT = 10 to kT = 0.001) during the course of the optimization, according to the exponential scheme $kT_i = 0.995\ kT_{i-1}$.

The optimization potential is defined by the reduced χ^2 (goodness-of-fit) statistic

$$\chi^2(\boldsymbol{\alpha}) = \sum_{x=1}^{50}\left[\frac{F(x) - D'(x,\boldsymbol{\alpha})}{\sigma(x)}\right]^2 \Big/ 46, \qquad [2]$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ is the vector of parameter estimates, x the K_S bin, $F(x)$ the observed K_s distribution, $D'(x,\boldsymbol{\alpha})$ the simulated K_S distribution with parameters $\boldsymbol{\alpha}$, and $\sigma(x)$ the standard deviation for bin x. The resulting χ^2 values are divided by a normalization factor (46) defined by the number of error degrees of freedom (50) minus the number of free parameters in the model (4 α's). Good fits should have reduced χ^2 values in the order of magnitude of 1. The standard deviations $\sigma(x)$ were estimated by constructing a cubic smoothing spline $S(x)$ to the observed K_s distribution $F(x)$, with a smoothing parameter of 0.3 (csaps function in MATLAB Spline Toolbox). $\sigma(x)$ is then approximated by $\sqrt{S(x)}$ (Fig. 8). The parameters α_i were optimized 10 times for each functional category in order to monitor the convergence of the parameter estimates.

Parameter confidence intervals

In order to calculate confidence intervals for the parameters α_i, we first calculated the covariance matrix for the best fit (simulation with the lowest χ^2 and parameters $\boldsymbol{\alpha}$'):

$$[C] = [A]^{-1} \qquad [3]$$

with

$$A_{ij} = \sum_{x=1}^{50} \frac{1}{\sigma^2(x)} \left[\frac{\partial D'(x,\boldsymbol{\alpha}')}{\partial \alpha_i} . \frac{\partial D'(x,\boldsymbol{\alpha}')}{\partial \alpha_j} \right] \qquad [4]$$

Approximate confidence intervals for the parameters α_i can then be calculated as

$$\delta\alpha_i = \pm\sqrt{\Delta\chi^2_\nu}\sqrt{C_{ii}}\,, \qquad [5]$$

where $\delta\alpha_i$ represents the 68% confidence interval for α_i, C_{ii} is the *i*th diagonal element of the covariance matrix, and $\Delta\chi^2_\nu$ is the 68th percentile of the χ^2 distribution with ν degrees of freedom. Because we calculate the confidence intervals in each parameter separately, ν = 1 and $\Delta\chi^2_\nu$ = 1. More background about these procedures can be found in ref. 10.

The minimum χ^2 values for all classes, the optimized parameters $\boldsymbol{\alpha}$ and their confidence intervals are summarized in Table 2. In general, a parameter α was considered reliable if its (one-sided) 68% confidence interval did not exceed 0.10, or if its relative confidence (i.e. the 68% confidence interval of α divided by α) was <20% (for higher parameters). For very high α's (>1), however, the calculated confidence intervals tend to be very large. This is mainly due to

the fact that for very high parameters, the newborn duplicates are quickly lost and have less influence on the course of the distribution. As a consequence, small changes in these parameters have virtually no effect on the modeled distribution or the χ^2, which leads to unnaturally large confidence intervals when using Eqs. **3**, **4**, and **5** (these equations only take into account the local environment of α). More accurate confidence intervals could be obtained by varying the α under study while optimizing for all other parameters until a given $\Delta\chi^2_\nu$ is reached (10). Unfortunately, the time required to do the necessary simulations is prohibitive. Instead, parameters >1 were considered to be reliably high regardless of their calculated confidence interval.

1. Blanc, G. & Wolfe, K. H. (2004) *Plant Cell* **16**, 1667-1678.

2. The Gene Ontology Consortium (2000) *Nat. Genet.* **25,** 25-29.

3. Berardini, T. Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L. A., Yoon, J., Doyle, A., Lander, G., *et al.* (2004) *Plant Physiol.* **135**, 745-755.

4. Zuckerkandl, E. & Pauling, L. (1965) in *Evolving Genes and Proteins*, eds. Bruson, V., Vogel, H. J. (Academic, New York), pp. 97-166.

5. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. N. (Academic, New York), pp. 21-132.

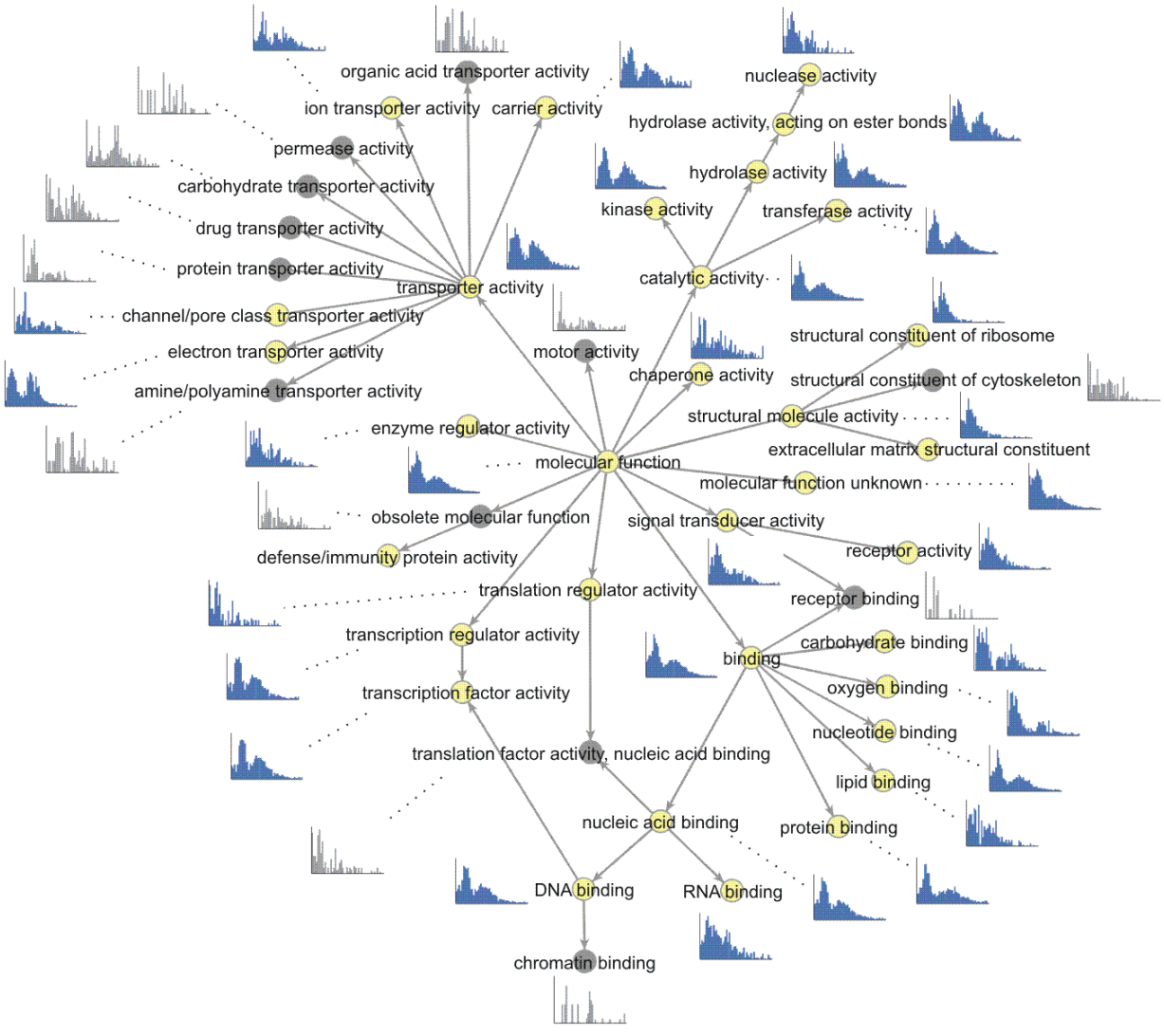
6. Li, W.-H. (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).

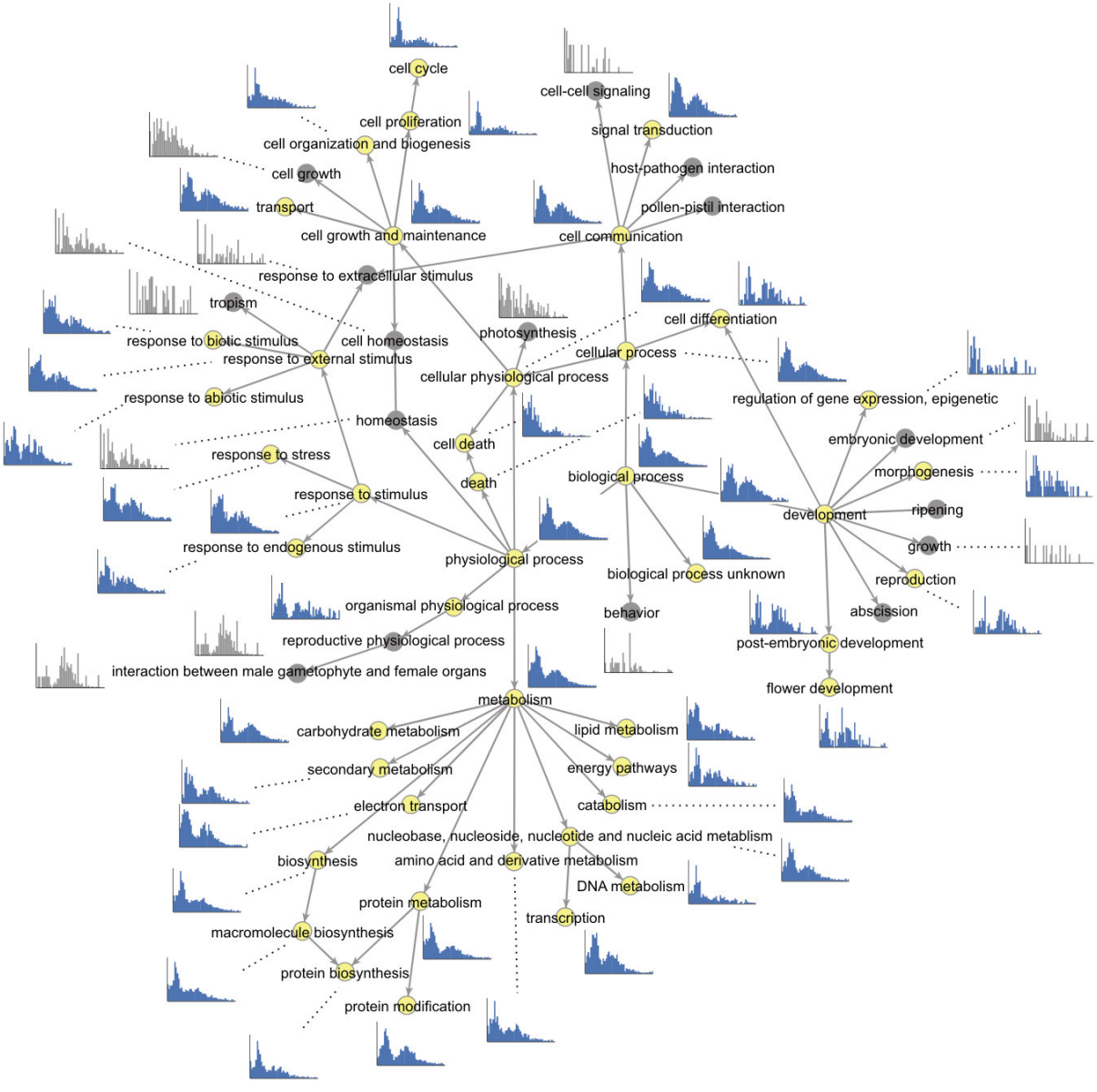
7. Lynch, M. & Conery, J. S. (2000) *Science* **290**, 1151-1155.

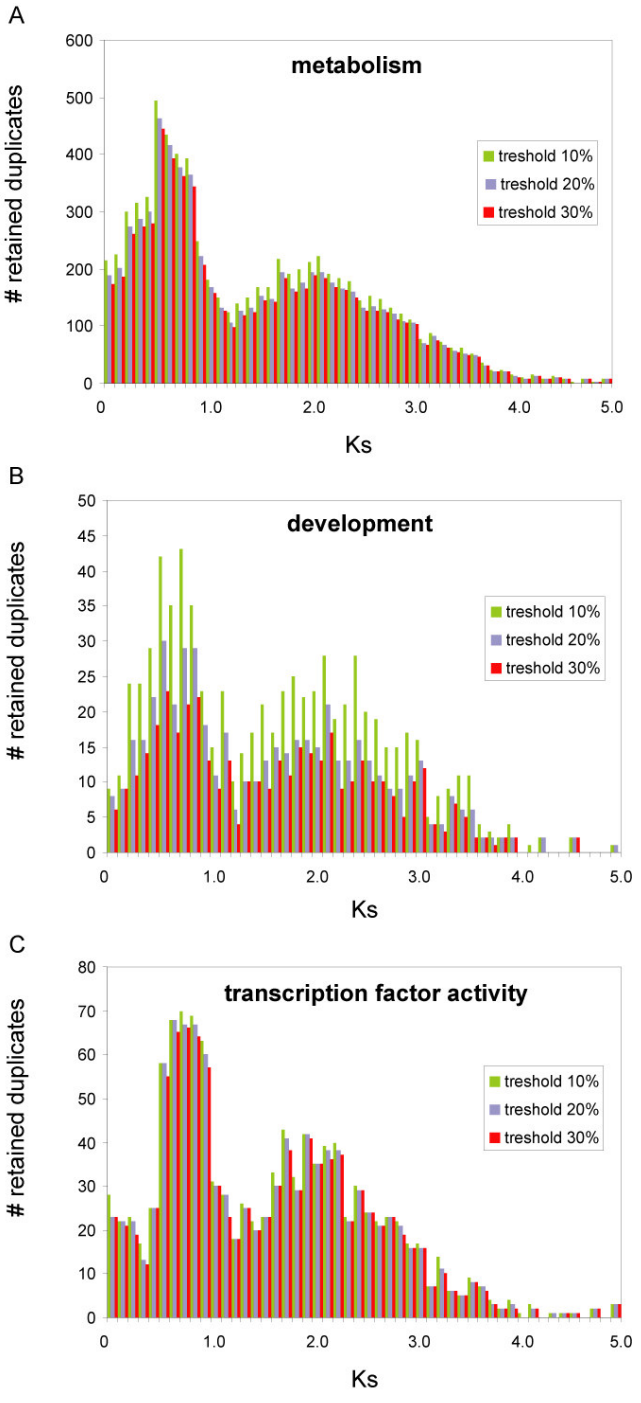
8. Koch, M. A., Haubold, B. & Mitchell-Olds, T. (2000) *Mol. Biol. Evol.* **17**, 1483-1498.

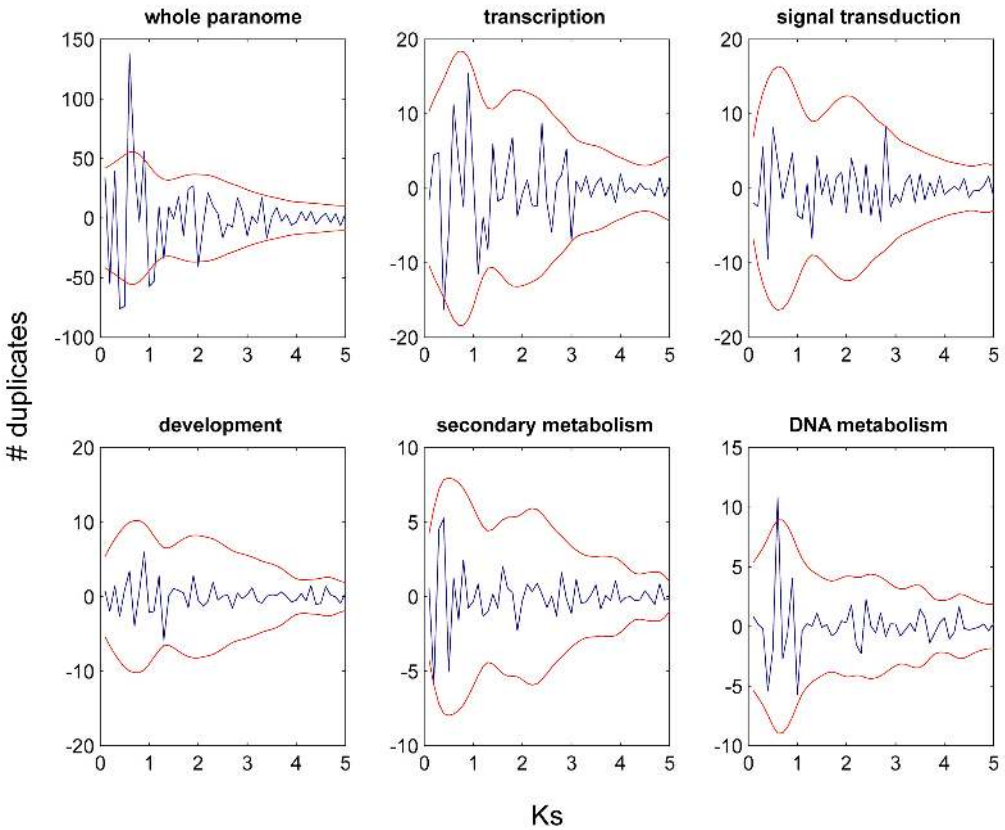
9. Seo, T.-K., Kishino, H. & Thorne, J. L. (2004) *Mol. Biol. Evol.* **21**, 1201-1213.

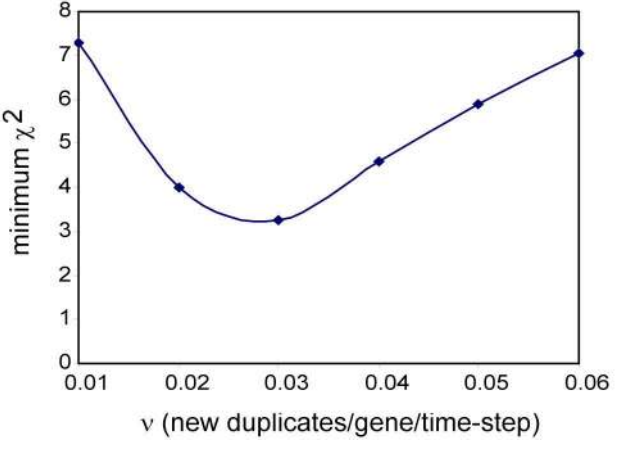
10. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992) in *Numerical Recipes in C: The Art of Scientific Computing*, eds. Cowles, L., & Harvey, A. (Cambridge University Press, Cambridge), 2nd Ed., pp. 656-706.

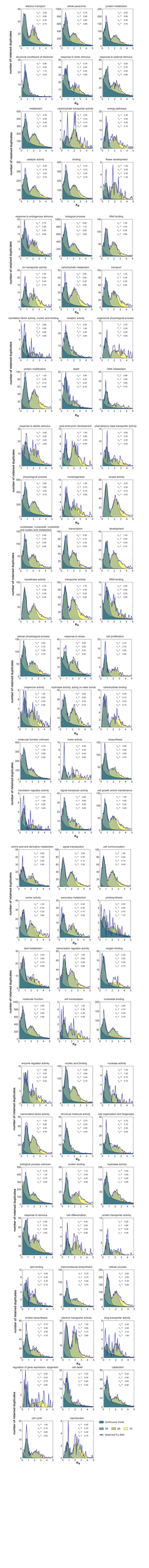

organic acid transporter activity
ion transporter activity
carrier activity
permease activity
carbohydrate transporter activity
drug transporter activity
protein transporter activity
transporter activity
channel/pore class transporter activity
electron transporter activity
amine/polyamine transporter activity
nuclease activity
hydrolase activity, acting on ester bonds
hydrolase activity
kinase activity
transferase activity
catalytic activity
motor activity
chaperone activity
structural constituent of ribosome
structural constituent of cytoskeleton
structural molecule activity
extracellular matrix structural constituent
enzyme regulator activity
molecular function
molecular function unknown
obsolete molecular function
signal transducer activity
defense/immunity protein activity
receptor activity
translation regulator activity
receptor binding
transcription regulator activity
carbohydrate binding
binding
oxygen binding
transcription factor activity
nucleotide binding
translation factor activity, nucleic acid binding
lipid binding
nucleic acid binding
protein binding
DNA binding
RNA binding
chromatin binding


cell cycle
cell proliferation
cell organization and biogenesis
cell growth
transport
cell growth and maintenance
cell-cell signaling
signal transduction
host-pathogen interaction
pollen-pistil interaction
cell communication
response to extracellular stimulus
tropism
response to biotic stimulus
response to external stimulus
response to abiotic stimulus
cell homeostasis
homeostasis
response to stress
response to stimulus
response to endogenous stimulus
photosynthesis
cellular process
cellular physiological process
cell differentiation
regulation of gene expression, epigenetic
embryonic development
morphogenesis
ripening
growth
reproduction
abscission
post-embryonic development
flower development
development
cell death
death
biological process
biological process unknown
behavior
physiological process
organismal physiological process
reproductive physiological process
interaction between male gametophyte and female organs
metabolism
carbohydrate metabolism
lipid metabolism
secondary metabolism
energy pathways
electron transport
catabolism
nucleobase, nucleoside, nucleotide and nucleic acid metablism
biosynthesis
amino acid and derivative metabolism
DNA metabolism
protein metabolism
transcription
macromolecule biosynthesis
protein biosynthesis
protein modification


A
metabolism
retained duplicates
Ks
treshold 10%
treshold 20%
treshold 30%
B
development
retained duplicates
Ks
treshold 10%
treshold 20%
treshold 30%
C
transcription factor activity
retained duplicates
Ks
treshold 10%
treshold 20%
treshold 30%


whole paranome
transcription
signal transduction
development
secondary metabolism
DNA metabolism
duplicates
Ks


8
7
6
5
4
3
2
1
0
minimum χ2
0.01
0.02
0.03
0.04
0.05
0.06
ν (new duplicates/gene/time-step)


electron transport
whole paranome
protein metabolism
structural constituent of ribosome
response to biotic stimulus
response to external stimulus
metabolism
carbohydrate transporter activity
energy pathways
catalytic activity
binding
flower development
response to endogenous stimulus
biological process
DNA binding
ion transporter activity
carbohydrate metabolism
transport
translation factor activity, nucleic acid binding
receptor activity
organismal physiological process
protein modification
death
DNA metabolism
response to abiotic stimulus
post-embryonic development
channel/pore class transporter activity
physiological process
morphogenesis
kinase activity
nucleobase, nucleoside, nucleotide and nucleic acid metabolism
transcription
development
translational activity
transporter activity
RNA binding
cellular physiological process
response to stress
cell proliferation
chaperone activity
hydrolase activity, acting on ester bonds
carbohydrate binding
molecular function unknown
motor activity
biosynthesis
translation regulator activity
signal transducer activity
cell growth and/or maintenance
amino acid and derivative metabolism
signal transduction
cell communication
carrier activity
secondary metabolism
photosynthesis
lipid metabolism
transcription regulator activity
oxygen binding
molecular function
cell homeostasis
nucleotide binding
enzyme regulator activity
nucleic acid binding
nuclease activity
transcription factor activity
structural molecule activity
cell organization and biogenesis
biological process unknown
protein binding
hydrolase activity
response to stimulus
cell differentiation
protein transporter activity
lipid binding
macromolecule biosynthesis
cellular process
protein biosynthesis
electron transporter activity
drug transporter activity
regulation of gene expression, epigenetic
cell death
catabolism
cell cycle
reproduction
number of retained duplicates
K_S
continuous mode
3R
2R
1R
observed K_S distr.