

Whole-Genome Duplications in the Ancestral Vertebrate Are Detectable in the Distribution of Gene Family Sizes of Tetrapod Species

Timothy Hughes · David A. Liberles

Received: 15 November 2007 / Accepted: 23 June 2008 / Published online: 25 September 2008
© Springer Science+Business Media, LLC 2008

Abstract A clustering of all protein coding genes from the complete genomes of five tetrapod species into gene families shows a clear deviation from the expected power-law distribution of gene family size. We hypothesize that at least part of the deviation is the result of the two whole-genome duplications (WGDs) that are now known, with reasonable certainty, to have occurred prior to the fish-tetrapod split. We build a model of homologous gene family evolution and perform simulations to show that speciations alone cannot produce a distribution that resembles the empirical data. In order to replicate the features of the empirical distribution, the simulation must incorporate two WGD events. These WGDs must be such that a significant number of the gene duplicates generated in the WGDs have a higher retention rate than they do following small-scale duplication (SSD). This requirement is consistent with what is known about duplicate retention following a WGD, namely, that genes belonging to specific functional classes, such as genes regulating transcription, are much more likely to be retained following WGD than SSD. We conclude that the deviation from the power-law that we observe in the empirical data is the result of the two WGDs that occurred in the ancestral chordate. This implies

that the two ancient WGDs continue to have a structural effect on gene families approximately 500 million years after the initial events. On the one hand, this is a surprising result, given the limited retention of duplicates generated by a WGD and the continual SSD, which further weakens the signal created by the fraction of duplicate pairs that are retained. On the other hand, WGD's capacity to fundamentally change the architecture of gene families in a profound and lasting way is consistent with the observed correlation between WGDs and important evolutionary transitions.

Keywords Gene duplication · Whole-genome duplication · Pseudogenization · Nonsynonymous substitution · Gene family size · Power-law distribution · Speciation

Introduction

The protein-coding genes of a genome are connected to each other through relationships of homology which are the result of novel genes arising through the duplication of ancestral genes. A natural way of structuring a set of proteins is therefore to divide it into subsets of homologous genes. There are different methods for building clusters of homologous genes, but a clustering of all protein coding genes in a genome, irrespective of the method used, produces many small clusters and few large clusters (Huynen and van Nimwegen 1998; Yanai et al. 2000; Harrison and Gerstein 2002), i.e., most genes have at most a few detectable homologs, whereas a few genes have very many (see Fig. 1). The functional form that best fits the data is the power-law (Huynen and van Nimwegen 1998; Luscombe et al. 2002): $N = aF^b$, where F is the family

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9145-x) contains supplementary material, which is available to authorized users.

T. Hughes (✉)
Computational Biology Unit, BCCS, University of Bergen, 5020
Bergen, Norway
e-mail: tim@bccs.uib.no

D. A. Liberles (✉)
Department of Molecular Biology, University of Wyoming,
Laramie, WY 82071, USA
e-mail: liberles@uwyo.edu

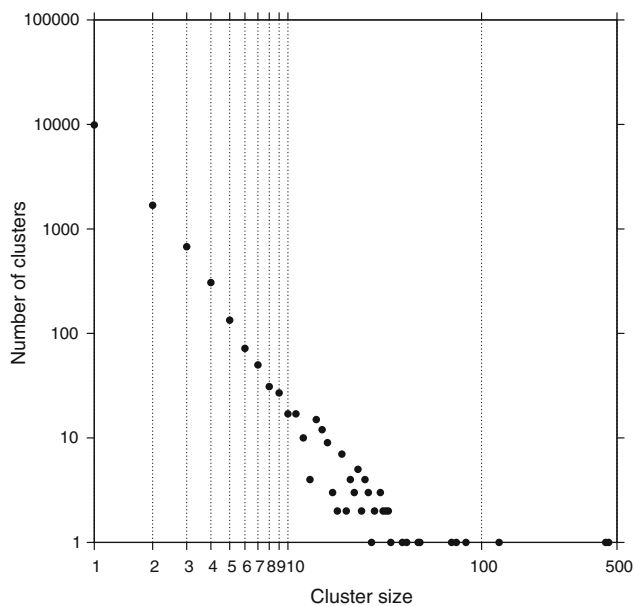


Fig. 1 Distribution of gene family size for *H. sapiens* (from a clustering of all putative genes from the fully sequenced genome)

size and N is the number of families of this size or, taking the natural logarithm, $\ln(N) = \ln(a) + b \cdot \ln(F)$, i.e., a linear relationship on a log-log plot. The exponent b is usually in the range -4.0 to -2.75 , and there is a weak positive correlation between the exponent and the logarithm of the number of genes in the genome, implying that an increase in the number of genes leads to a relative increase in the number of large clusters over the number of small clusters (Huynen and van Nimwegen 1998).

We have previously shown through simulation what are likely to be the key processes causing the emergence of a power-law (Hughes and Liberles 2008). If we assume a constant small-scale gene duplication rate (tandem and segmental duplications), and we use specifications of the pseudogenization rate and sequence divergence rate that have been validated using genomic data, the minimal requirements for such a distribution to emerge from an initial set of singleton genes are that the genes are subject to duplication and loss, and that there is heterogeneity of the rate of loss across gene families. Once a power-law distribution emerges, and assuming that there are no large-scale duplication events, the intuition as to why such a distribution is maintained is simple: genes in all families duplicate, but the vast majority rapidly pseudogenize (Lynch and Conery 2000; Lynch and Conery 2003; Hughes and Liberles 2007). Some families may have a lower pseudogenization rate, causing such families to increase in size relative to families with a higher rate, but this is a slow process due to the strength of pseudogenization. In addition, sequence divergence will be a moderating factor allowing older duplicates to accumulate sufficient

replacement substitutions and split away from their original family, thus limiting the size of large families.

Of course, lineage-specific expansions and contractions do occur in certain gene families, for example, in the human lineage, the GAGE gene family seems to be expanding, while the olfactory receptor gene family seems to be contracting (Gilad et al. 2003). This has been termed the revolving-door mechanism (Demuth et al. 2006) and is expected to be nonrandom and related to both gene function (Maere et al. 2005) and protein fold (Rastogi et al. 2006). However, this dynamic process only affects a limited number of families in any one lineage (Demuth et al. 2006).

It has also been shown that the power-law distribution applies to the clustering of genes from multiple complete genomes, if the species concerned are evolutionary distant (Enright et al. 2003). We were therefore initially intrigued by the observation of a strong deviation from the power-law when clustering the genes from the complete genomes of five tetrapod species, with clear “waves” with a period of size 5, visible for sizes 1 to 15 (see Fig. 2). These “waves” are very pronounced: not only do the frequencies not follow a linear relationship in the log-log plot, but also it is the case for several sizes that the frequency of size x is less than the frequency of size $x + 1$. However, in this case, although the species’ divergence times spanned several hundred million years, they were not as distantly related as in the Enright et al. study. Moreover, a convincing case, based on gene family phylogenetic trees and genomic map position data, has been made in favor of the

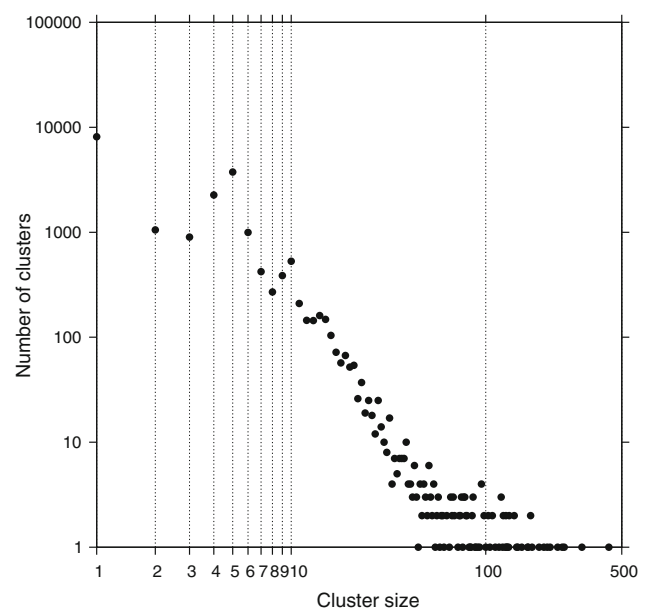


Fig. 2 Distribution of gene family size for five tetrapod species (from a clustering of all putative genes from the fully sequenced genomes of *G. gallus*, *C. familiaris*, *M. musculus*, *R. norvegicus*, and *H. sapiens*)

hypothesis that the genome of the ancestral vertebrate was subject to two whole-genome duplications (WGDs) (Dehal and Boore 2005). We hypothesize that the deviation from the power-law observed in the empirical data is at least partly the result of these two whole genome duplications.

It is obvious that if a WGD reaches fixation, it will initially dramatically alter the gene family size distribution: the frequency of families of size x will become the frequency of families of size $2x$ and odd-sized families will become effectively nonexistent. Speciation initially has a similar effect if we consider a clustering of all genes from both descendent species. However, it is less straightforward to establish the effect of small-scale duplication (SSD) and loss following the initial event: whether a power-law distribution returns (and, if so, on what timescale) and what effect a combination of WGD and multiple speciations has on the gene family size distribution.

It was Ohno (1970) who originally proposed that the ancestral chordate underwent two rounds of WGD. The advent of genomic data has made it possible to begin testing this hypothesis, but initial efforts were not conclusive, with different studies providing evidence for different numbers of WGDs: some arguing for none (Friedman and Hughes 2001, 2003), others for one (McLysaght et al. 2002) or two (Abi-Rached et al. 2002). Several of these studies centered on the size of gene families. The opponents of WGD argued that there was no signal of a WGD in this kind of data (Friedman and Hughes 2001). A recent study has produced strong evidence of 2R through the genomic mapping of paralogous regions known to have arisen before the fish-tetrapod split (Dehal and Boore 2005). However, they too claim that there is no signal of WGD in gene family size data. These findings directly contradict our hypothesis that the pattern we observe in the empirical data is caused by the ancestral WGDs. The reason that these studies found no signal of the WGDs in gene family size data is that they did not compare the empirical distribution to the distribution that would be expected in the absence of WGDs. In previous work (Hughes and Liberles 2008), we have shown that SSD and loss results in a power-law distribution and, thus, is the expected distribution in the absence of a WGD. In this paper, we extend our model of homologous gene family evolution to incorporate WGD and speciation, and use it to simulate the evolution of the distribution of gene family size under different scenarios (WGD, speciation, and WGD followed by multiple speciations). The output of the simulations shows that, in order to produce a deviation from the power-law which is consistent with the empirical data, the simulation must incorporate not only speciations but also two WGDs. We conclude that the deviation from the power-law that we observe in the empirical data is the result of the two WGDs that occurred in the ancestral chordate.

Materials and Methods

Empirical Data

Our empirical data consist of the longest protein-coding transcript sequence for every gene of the annotated genome of the following species from release 31 of Ensembl (Birney et al. 2006): *Gallus gallus*, *Canis familiaris*, *Mus musculus*, *Rattus norvegicus*, and *Homo sapiens*. First, we carry out low-complexity masking of the translated sequences using CAST (Promponas et al. 2000), and then we perform an all-against-all BLAST (Altschul et al. 1997) (substitution matrix = BLOSUM62, gap opening cost = 11, gap extension cost = 1). In order to make the output of the all-against-all BLAST manageable, the BLAST sequence pairs (query and target sequences) are filtered to remove any targets that do not satisfy all of the following criteria, which should be satisfied by even very distant homologs: 20% similarity to the query, 60% coverage of the query, and e-value $<10^{-5}$. The e-values of the retained sequences are then used as input to the MCL clustering algorithm with the inflation parameter set to 4.0 (Enright et al. 2002). This procedure produces a clustering of all genes into gene families from which the distribution of gene family size can be computed (see Fig. 2).

Simulated Data

Overview

In order to theoretically investigate the effect of either a speciation event or a WGD on the power-law distribution, we need a model of homologous gene family evolution. We use as our starting point the model developed in a previous paper (Hughes and Liberles 2008). We repeat here a description of this original model and we extend it to incorporate speciation and WGD events.

Basic Model

We model the rate of gene duplication, the rate at which replacement substitutions per replacement site accumulate between genes in a duplicate pair and the rate at which one of the genes in a pair pseudogenizes. These models are taken directly from our previous study (Hughes and Liberles 2007), which built on earlier work on the same topic (Lynch and Conery 2000, 2003). Time is measured through the accumulation of silent substitutions per silent site (S) between duplicate genes. In *H. sapiens*, under the assumption of a constant rate of small-scale duplication, we have estimated that genes duplicate at a rate of 2.07 per gene per S (all parametrizations used here are the result of fitting the models to duplicate gene pair data from the *H.*

sapiens full-genome sequence annotation). A duplicate pair i accumulates replacement substitutions per replacement site (R), according to the equation:

$$R_i = \theta_1 S_i + (\theta_2/\theta_3)(1 - \exp(-\theta_3 S_i)) + \varepsilon_i \quad (1)$$

$$\text{Var}(\varepsilon_i) = \sigma^2(\tau_1 S_i + \exp(\tau_2(1 - \exp(-\tau_3 S_i))), E(\varepsilon_i) = 0 \quad (2)$$

where the ε_i are assumed to be independent random variables for i varying from 1 to n . We use the following fitted values of the parameters (Hughes and Liberles 2007): $\theta_1 = 0.13$, $\theta_2 = 0.70$, $\theta_3 = 2.4$, $\sigma^2 = 3.55e - 5$, $\tau_1 = 229.4$, $\tau_2 = 6.32$, and $\tau_3 = 4.14$.

The probability of pseudogenization of one of the genes in a pair within Δt given that both genes are still functional at t is

$$\text{Pr}(t < T < t + \Delta t / T > t) = -\frac{Q(t + \Delta t) - Q(t)}{Q(t)}. \quad (3)$$

where $Q(t) = \text{Pr}(T > t)$ is the survival function: the probability that the time of death, T , is greater than t , i.e., the probability that both genes are still functional at time t . The hazard function $\lambda(t)$ is defined as the event (death/pseudogenization) rate at time t conditional on survival to time t or later:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{\text{Pr}(t < T < t + \Delta t / T > t)}{\Delta t} = -Q'(t)/Q(t) \quad (4)$$

We have shown that the Weibull survival function $Q(t) = e^{\rho_1 t^{\rho_2}}$ provides an excellent fit to the data (Hughes and Liberles 2007). Thus, we use this model of the survival function and S as a proxy for time:

$$\lambda(S) = -\rho_1 \rho_2 S^{\rho_2 - 1} \quad (5)$$

In *H. sapiens*, the fitted parameters are $\rho_1 = -4.1$ and $\rho_2 = 0.33$, which implies that the rate of pseudogenization of a duplicate is a decreasing function of S .

A gene in our model has two key characteristics: it is either functional or pseudogenized, and it has a measure of the number of silent and replacement substitutions per site between itself and all homologous genes, i.e., all genes that can be traced to a common ancestor through a series of duplication events. The model is initialized with a set of singleton genes, i.e., genes that have no duplicates, and therefore each forms a family of size 1. These are the “founding” genes of the homologous gene families. Because all key processes are defined in terms of S , we define a “clock” which “ticks” in increments of 0.001 unit of S . At each tick of the clock each gene’s number of silent substitutions is incremented by half a tick, so that the distance between all genes increases by one tick. For each gene, we then detect the closest nonpseudogenized homolog, which we define as the homologous gene with the

lowest R distance to the gene of interest. The S distance between the two genes is a measure of the time since the original duplication event and is used to compute the number of replacement substitutions per site the duplicate pair should be subject to in the timeframe of the current tick (Eq. 1) and the probability that one of the duplicates pseudogenizes during the current tick (Eq. 3) with S as a proxy for time and the Weibull survival function. A gene that has no homologs (such as a founding singleton before it is duplicated) is assigned an S value of 1000, which ensures that it accumulates R at a very low rate and is subject to a very low probability of pseudogenization. This is a reasonable way to model singletons, as singletons can be expected to have evolved some kind of specialized function that is under selective pressure to be retained in the genome. Finally, each gene is subject to a constant probability of duplication during each tick. The gene that results from a duplication is added to the set of homologous genes. It inherits the R and S distances to other genes from its parent and has a distance of 0 R and 0 S to its parent.

As the model stands, all genes are subject to the same rates of the three main processes. We introduce differences between genes by introducing error terms to the rate of sequence divergence and the rate of pseudogenization. Errors for the rate of sequence divergence are available, as the fitting of Eq. 1 produced residuals. The functional form of the distribution of the error term in Eq. 1 is not known so, as the second-best alternative, we draw an error term randomly from all residuals from the fitting of Eq. 1 to the *H. sapiens* gene duplicate data (Hughes and Liberles 2007). This error term can be standardized through the model of the variance as a function of S (see Eq. 2). A new gene is assigned this error term and this is used in Eq. 1 to calculate the number of replacement substitutions per replacement site the duplicate pair should be subject to in a tick (each gene in a pair is subject to half the value predicted by Eq. 1 when using the gene’s specific error term).

In order to be able to accommodate different genes having different rates of pseudogenization, we modified the definition of the probability of pseudogenization of a duplicate pair i within a time interval Δt given survival until t , by defining

$$\begin{aligned} \text{Pr}'(t < T < t + \Delta t / T > t) \\ = (1 + v_i) \text{Pr}(t < T < t + \Delta t / T > t) \end{aligned} \quad (6)$$

where $v_i \sim N(\mu, \sigma^2)$. We want to be able to control the extent to which hazard rates are correlated within families. Thus, when a new gene is created by duplication, either the error term v_i is inherited from the gene that duplicated, in which case all genes descendent from a founding singleton will have the same hazard function and the heterogeneity between families is determined by σ^2 , or a new error can be

drawn from the distribution, in which case there will be no correlation between the hazard functions of genes descended from the same singleton.

Model Validation and Results

In the original paper where we first presented this model, we successfully tested that the simulated evolution of gene duplicates using this model matches the real *H. sapiens* duplicate gene data, i.e., that the rates of pseudogenization and the rate of accumulation of replacement substitutions are the same in the simulated and empirical data.

At regular intervals during the simulation, we extract R for all duplicate pairs and use these data to compute a complete linkage clustering of all nonpseudogenized genes. We use an empirically derived maximum distance of 0.56 R between genes in the same family as the cutoff value in the clustering process (Hughes and Liberles 2008). From this clustering, we can compute the distribution of gene family size.

In our previous paper, we found that the power-law distribution of gene family size failed to emerge if all genes in all families had the same hazard function or if all genes had different hazard functions. The key conclusion was that it is necessary for v_i to be correlated within a family (inheritance of v_i) and for there to be sufficient heterogeneity between families ($v_i \sim N(0, 0.04)$). We found that, in such circumstances, a power-law distribution of gene family size had clearly emerged by $S > 2.0$. This corresponds to approximately 460 million years, given a rate of 2.20 silent substitutions per silent site per billion years for *H. sapiens* (Yang and Nielsen 1998). Thus, in all our modeling in this paper, we use this configuration of the model, which means that, in the absence of WGD, all genes in the same family have the same hazard rate. We run the simulation until $S = 2.0$, so that a power-law emerges in the distribution of gene family size. We initiate the model with 1000 singleton families, as the number of genes grows rapidly when the genome is subject to WGD and speciation, and the complete linkage clustering is compute-time intensive. Only once the power-law distribution has emerged do we disrupt it by subjecting the genome to a speciation or WGD event.

Speciation and Whole-Genome Duplication Extensions

A speciation event is modeled by copying every nonpseudogenized gene (including S and R distances to all homologs) and labeling the genes with the species of the genome in which they exist. This information is then used to limit the search for the closest nonpseudogenized homolog to genes that belong to the same species. This

ensures that both genomes evolve independently after speciation, as only genes from the same species (and not orthologues) play a role when computing the rate of sequence divergence or the probability of pseudogenization, but orthologues do get clustered together when homologous gene families are built.

A WGD is modeled by duplicating all genes of a specific species, giving them a distance of 0 R and 0 S to their parent as in SSD. Due to the one-off nature of WGD, we do not have a quantitative characterization of the rate of sequence divergence and pseudogenization as a function of time since WGD. We are thus forced to define these rates as best we can. We leave the sequence divergence rate the same as for SSDs, as this process does not appear to play a crucial role in dynamics of the power-law distribution. For the hazard rate, we implement two options. The first option is to simply consider that the duplication event to which each gene is subject in the WGD is the same as a SSD, i.e., that each gene inherits the pseudogenization error term from its parent. However, there is evidence that the retention rate is higher following many WGDs, e.g., following the fish-specific WGD (Woods et al. 2005). The most prominent models that have been put forward to explain this higher retention rate of gene duplicates that arise through WGD are dosage balance (Aury et al. 2006) and subfunctionalization (Force et al. 1999). Although the models are fundamentally very different, they both share that some genes are more highly retained following WGD and that these genes may be those that stand a below-average chance of being retained following SSD.

To incorporate such features in our model, we divide the families into two categories: those that have a high hazard rate error (defined as $v_i > 0.1$) following SSD and those with a low hazard rate (defined as $v_i < 0.1$). We choose this definition as it makes approximately one-third of the genes dosage sensitive and thus subject to a hazard shift. We have no data on the proportion of genes that may be subject to such a hazard shift but given that, following the fish-specific WGD, it was estimated that 20% of duplicates were retained (Jaillon et al. 2004; Woods et al. 2005; Brunet et al. 2006), whereas retention rates following SSD are only a few percent (Hughes and Liberles 2007), one-third is not an unreasonably large fraction. When a WGD occurs, genes with a high hazard rate error are duplicated and both the original gene and the duplicate are given a hazard rate error (v_i) drawn from $N(-0.85, 0.0025)$, which ensures a very high probability of retention. We refer to this as hazard shift. Genes with a low SSD hazard rate are duplicated and inherit the error of the duplicated gene as in SSD. When a hazard shifted gene is subsequently duplicated in a small-scale duplication event, the duplicate does not inherit the hazard shift, instead the v_i shared by all genes in the family prior to the WGD is restored.

Results

The simulations use a model of homologous gene evolution which incorporates gene duplication, sequence divergence, and pseudogenization. It is impossible to obtain accurate estimates of these processes for the 500 million years that separate us from the two WGDs; as a second option the model is parameterized using *H. sapiens* data. Time is measured in units of silent substitutions per silent site between duplicate pairs (S), and in *H. sapiens* 1 S corresponds to ~ 230 million years (Yang and Nielsen 1998). The model allows not only for SSD caused by tandem and segmental duplication, but also for speciation and WGD events (see Materials and Methods for full details). WGD events are implemented in two ways: the first implementation assumes that a gene duplicate's pseudogenization rate (or hazard rate) following a WGD is the same as following a SSD (the “no hazard shift” model); in the second implementation, gene duplicates that have a high hazard rate following a SSD are given a very low hazard rate following a WGD, while gene duplicates with a low hazard rate following a SSD maintain the same hazard rate following a WGD (the “hazard shift” model). This second implementation is consistent with models such as dosage balance (Aury et al. 2006) and subfunctionalization (Force et al. 1999) which have been developed to explain the higher retention of duplicates following a WGD compared to a SSD. There is mounting evidence that these models fit the data on the retention of duplicates generated through WGD (see the discussion for the details of this point).

Whole-Genome Duplication

Immediately following the fixation of a WGD event, the frequency of families of size x will become the frequency of families of size $2x$ and odd-sized families will become effectively nonexistent. Assuming that the distribution of gene family sizes followed a power-law prior to the WGD, then the post-WGD distribution of even-sized families should also follow a power-law with the same exponent but a higher intercept (see Fig. 3). As time passes since the WGD, if small-scale gene duplication returns, and gene loss and sequence divergence happen in an SSD manner (for all genes, irrespective of whether or not they were generated by the WGD), then we would expect to see the families of odd sizes increase in number as even-sized families increase in size by 1 through SSD and decrease in size by 1 through loss (see Fig. 3). Loss should be rampant due to all genes being recent duplicates which are subject to a high pseudogenization rate (Hughes and Liberles 2007). This should result in the return of the power-law with a similar exponent to the original power-law and an intercept lower than the intercept that prevailed immediately following the WGD but higher than the pre-WGD intercept due to the retention of some of the WGD duplicates. However, if we consider that a significant proportion of gene families is subject to a downward shift in the hazard rate following WGD and that families experiencing this shift are drawn from families that have a high hazard rate following SSD, then the return to the power-law will be slower. The reasons for this are twofold, and both are

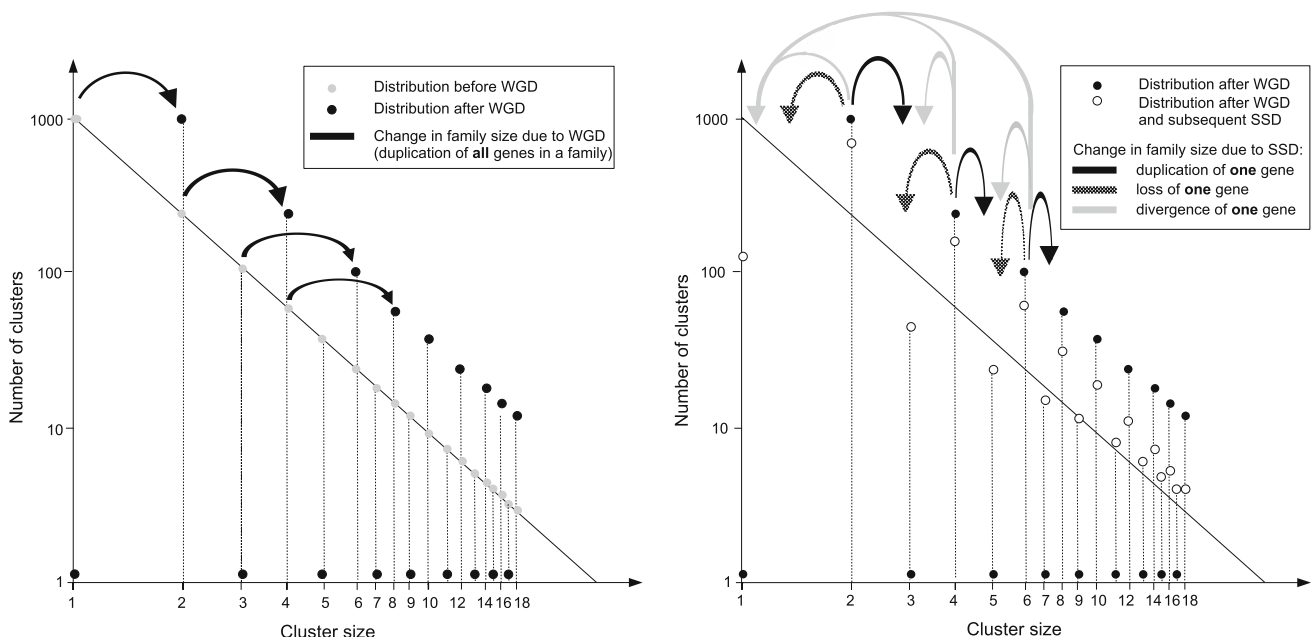


Fig. 3 Qualitative description of the immediate effect of a WGD on a power-law distribution of gene family size (first graph) and subsequent return to the power-law distribution through SSD (second graph)

connected to the tendency for the hazard-shifted families to retain their post-WGD size: they are less likely to diminish in size through loss because of the reduced pseudogenization rate that applies to the duplicates generated by WGD and less likely to increase in size due to the high hazard rate that applies to duplicates that may be generated by SSD after the WGD.

In order to verify these predictions, we carry out two simulations (see Materials and Methods for details): one where WGD duplicates behave in the same manner as SSDs (the “no hazard shift” model) and one where a certain fraction of families is subject to a hazard shift following WGD (the “hazard shift” model). In both cases, the WGD is carried out at $S = 2.0$, as it takes approximately this time for the power-law distribution to emerge from the initial state of the model through SSD and loss. We consider the power-law distribution to have returned for a certain size range when the frequency of these sizes has an approximately linear downward-sloping relationship in a log-log plot and when, for all sizes within this range, the frequency of size x is stably greater than the frequency of size $x + 1$.

As predicted, there is a relatively rapid return toward the pre-WGD distribution for the “no hazard shift” model (see supplementary Fig. 1). By $S = 2.3$ (i.e., $0.3 S$ after the WGD), the power-law has returned for all sizes for the “no hazard shift” model (see Fig. 4). In contrast, in the “hazard shift” model, the lower pseudogenization rate of the duplicates subject to a hazard shift results in a continued underrepresentation of singletons and less of a shift back toward the origin.

Speciation

A speciation event results in two independently evolving genomes, but initially these genomes will be almost identical, as the new species will inherit the genetic material of the ancestral species. Thus, if we consider the clustering of the genes of two genomes that recently speciated, the initial effect of the speciation is the same as that of a WGD. Subsequently, as after WGD, duplication and loss begin to smooth the distribution, but loss is more restricted following speciation because it is not associated with a sudden increase in the number of young duplicates which are subject to a higher pseudogenization rate (Hughes and Liberles 2007). This should slow the return to the power-law compared to both WGD models. In addition, there is expected to be little loss of singletons in each species which we model by ensuring that singleton families have a near-zero probability of losing their last member. These singletons form families of size 2 when clustering the genes from both genomes, and the restricted loss of singletons from each species means that the number of

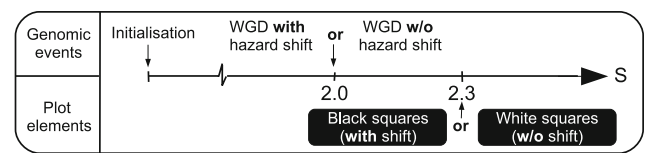
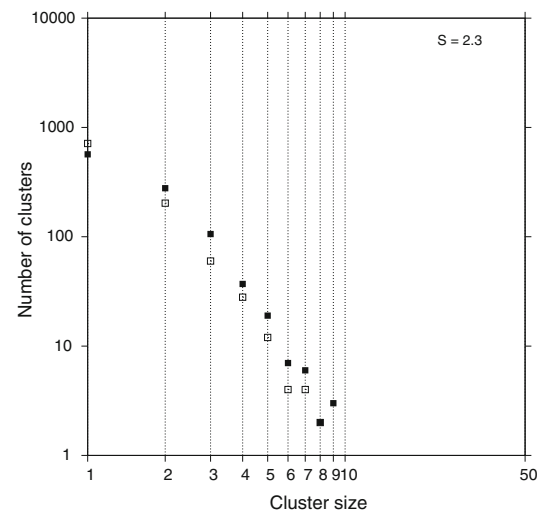


Fig. 4 Comparison of the effect of WGD with vs. without hazard shift on the distribution of gene family size. WGD with hazard shift followed by SSD results in less of a shift back toward the origin and an underrepresentation of singletons relative to a WGD without hazard shift

families of size 1 can effectively only increase through the divergence of sequences from larger-sized families causing the sequence to “break away” to form a new singleton family. Thus, the increase in the number of families of size 1 is expected to be particularly slow. Still, relative to other family sizes with fewer members than the species number, size 1 is overrepresented, indicating that sequence divergence generating new families does happen.

We run a simulation in which we perform two speciation events. The qualitative predictions are fulfilled as expected. As explained earlier, the very high underrepresentation of families of size 1 after the first speciation (and of families of sizes 1 and 2 after the second speciation) is to be expected but is perhaps exaggerated in the simulation, as our model effectively does not allow singletons to be lost from a genome (see Fig. 5).

The underrepresentation of smaller gene family sizes observed here has also been observed in the distribution of gene family sizes in gene family databases like TAED (Roth et al. 2005) that were built using entirely different methodology and, as such, is not an artifact of the gene family construction process. This may in fact reflect a set of core functions, where deletion from a genome is highly deleterious. While the core set of functions necessary for parasitic prokaryotic life has been estimated to be 500–600 genes (Koonin 2003), that set for vertebrate life might be expected to be larger. It has been suggested that

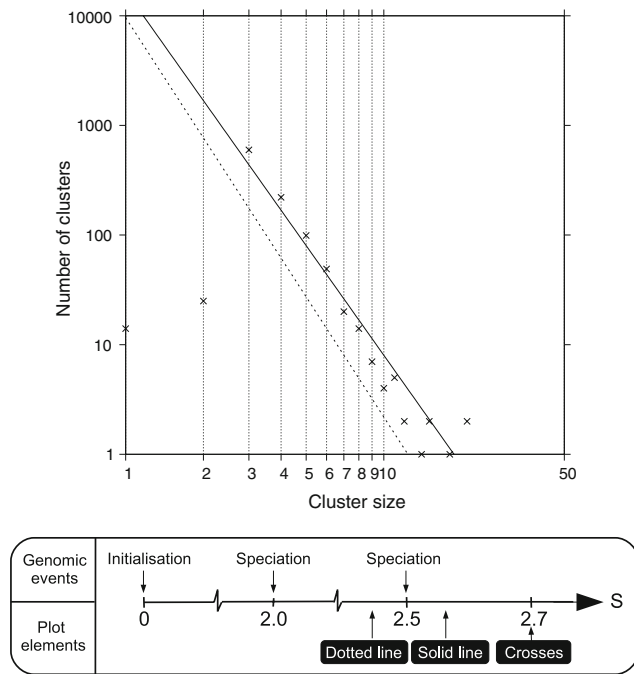


Fig. 5 Effect of two speciations on the distribution of gene family size. Dotted line: equation fitted to the distribution data immediately prior to the second speciation. Solid line: equation fitted to the distribution data immediately after the second speciation Crosses: data at $S = 2.7$. A clear return to the power-law distribution for family sizes greater than the number of species (two speciations lead to three species) and a clear underrepresentation for family sizes less than the number of species

informational genes (involved in the retention of biological information) represent a phylogenetic core in bacterial species (Rivera et al. 1998) and a similar expanded core might be expected in vertebrate species.

Whole-Genome Duplication Followed by Speciation

Our first two sets of simulations have established that (1) following a WGD without hazard shift a power-law distribution returns within a relatively short period of time for all sizes; (2) for a WGD with hazard shift, a power-law also returns, but families of size 1 are underrepresented and the intercept is higher than in the “no hazard shift” model; and (3) following multiple speciation events, a power-law returns, but it takes longer than following a WGD, the intercept is significantly higher than prespeciation, and sizes less than the number of species are underrepresented relative to the power-law defined by larger sizes. This strongly suggests that the underrepresentation of family sizes less than the number of species, which we observe in the empirical data, can be explained by the speciations, but that speciation events alone or WGD events without hazard shift cannot explain the “waves” observed in the empirical data for family sizes greater than the number of species

(see Fig. 2). In fact, given the output of the simulation of WGD with hazard shift, it is also difficult to see how it can explain a deviation from the power-law for these larger family sizes: we are able to detect a signal when comparing the output of simulations with vs. without hazard shift (see Fig. 4), but if we were to observe the output of the “hazard shift” model alone, it would be difficult to argue that a signal was still observable.

We now run simulations that combine both WGD and speciation events to investigate whether the signal that remains following a WGD with hazard shift continues to be detectable when it is followed by multiple speciation events and whether the qualitative features of the signal are consistent with our empirical observation of “waves” with a period equal to the number of species. First, we run the same two simulations as in the WGD section, i.e., with and without hazard shift, but we let the WGD be followed by two speciations. For the WGD without hazard shift, we obtain the same results as in the absence of speciation events, i.e., a return to the power-law and no detectable signal of the WGD (see supplementary Fig. 5). For the WGD with hazard shift, on the other hand, the underrepresentation of singletons (as shown in Fig. 4) is not affected but does shift to another size due to the speciation events: immediately prior to the second speciation, i.e., $0.5 S$ since the first speciation, the frequency of families of size 3 is almost equal to that of families of size 4 (see the white dot distribution in the first graph in Fig. 6). This is the result of the reduced loss from “hazard-shifted” families. Singletons are underrepresented before the first speciation due to the reduced loss, thus, following the first speciation which results in two species, families of size 2, consisting mainly of one gene from each species, are underrepresented and families of size 4, which consist mainly of a duplicate pair generated by the WGD for each species, are overrepresented. Interestingly, the second speciation event would appear to make the signal of the WGD event with hazard shift stronger (see the clear and persistent underrepresentation of families of size 5 relative to size 6 in graphs 2–4 in Fig. 6).

To understand how the frequency of families of size 5 can be less than the frequency of families of size 6 following the second speciation in the model with hazard shift, we investigate the details of the gene family size distribution immediately prior to the second speciation and compare it to the equivalent situation for the model without hazard shift (see Table 1). There are two clear differences between the simulated data for the two models prior to the second speciation. First, the frequency of size 4 in the model with hazard shift is more than double that in the model without hazard shift. Second, the species composition of size 4 is radically different between the two models, with a clearly higher frequency of two sequences from each

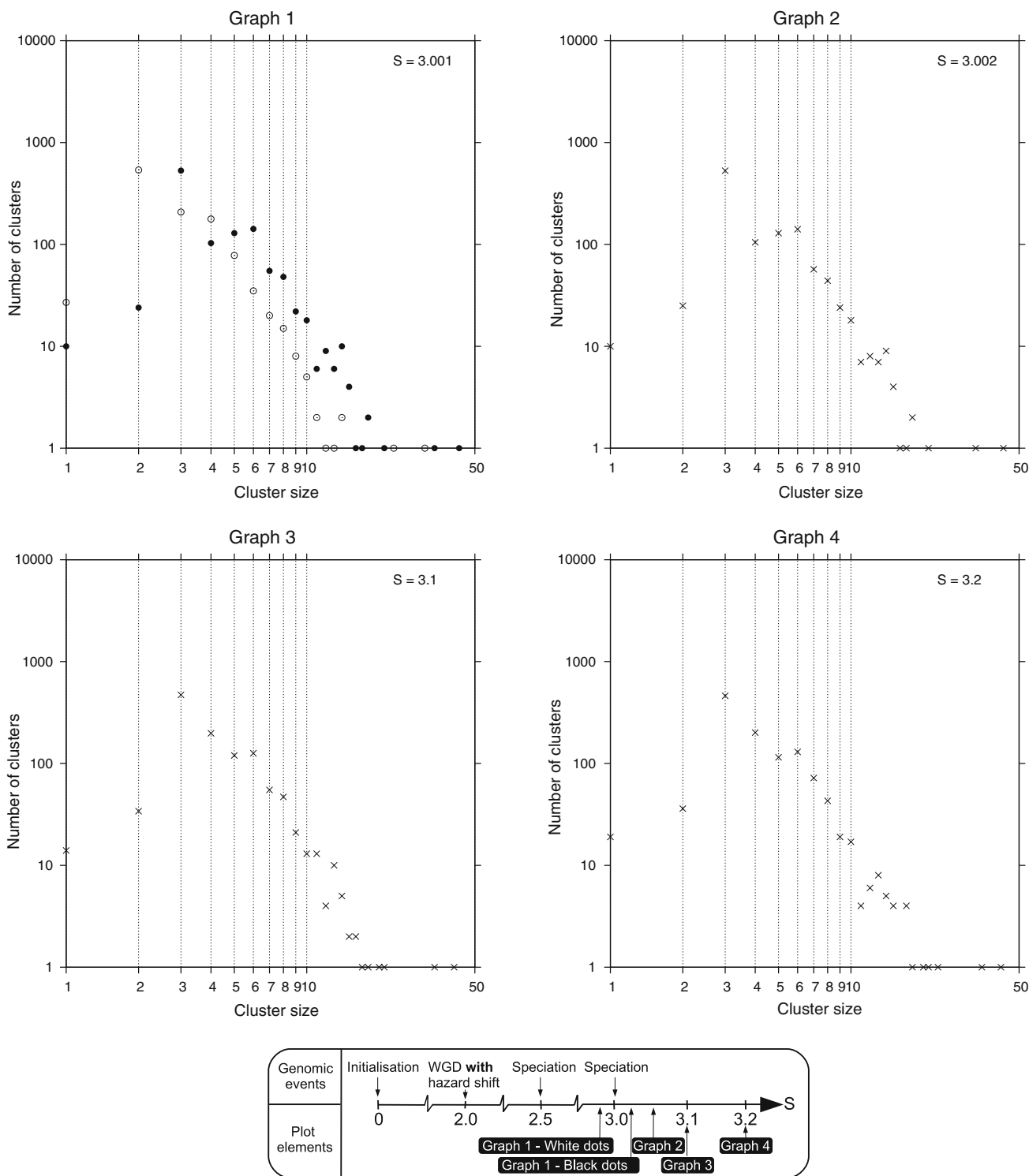


Fig. 6 Effect of a WGD with hazard shift followed by two speciations on the distribution of gene family size. Prior to the second speciation, there appears to be a slight underrepresentation of families of size 3. Following the second speciation, this translates into

a clear underrepresentation of families of sizes 4 and 5, which persists despite the SSD that returns following the WGD. The underrepresentation of families of sizes 1 and 2 is due to the speciation events and not the WGD (see Fig. 5)

species in the model with hazard shift. Since families of size 4 with two sequences from each species become families of size 6 following the second speciation, this

explains how families of size 6 become overrepresented relative to families of size 5. The speciation event is effectively combining these two signals to produce a

Table 1 Gene family size distribution for the genes of species A and B which arose at $S = 2.5$ immediately prior to a second speciation of B at $S = 3.0$: prior to the first speciation, the common ancestor of A and B was subjected to a whole-genome duplication

Family size	Prior to 2nd speciation					After 2nd speciation Family size		
	Frequency (%)		Species composition	Frequency (%)				
	No hazard shift	Hazard shift		No hazard shift	Hazard shift			
1	3.2	2.4	A	35.3	37.0	1		
			B	64.7	63.0	2		
2	59.3	48.0	AA	0.6	0.9	2		
			AB	98.6	98.1	3		
			BB	0.8	0.9	4		
			AAA	0.5	1.0	3		
3	20.3	18.6	AAB	47.7	46.2	4		
			ABB	51.4	51.4	5		
			BBB	0.5	1.4	6		
			AAAA	0.0	0.0	4		
4	6.9	15.9	AAAB	36.5	13.5	5		
			AABB	36.5	72.5	6		
			ABBB	27.0	14.0	7		
			BBBB	0.0	0.0	8		

stronger unified signal which is detectable as a clear deviation from the power-law distribution.

Finally, we run a simulation with two hazard-shifted WGDs followed by two speciations to test whether a simulation can produce a distribution that resembles the empirical data, which we know with reasonable certainty was subject to two WGDs prior to the fish-tetrapod split dated to ~ 476 million years ago (Blair and Hedges 2005). We separate the WGDs by $0.2 S$, which might correspond approximately to the time between $2 R$, but estimates of the time separating the two events vary between 10 million and 100 million years (Lundin et al. 2003). The separation of the speciation events is set arbitrarily to $0.5 S$, which corresponds to ~ 115 million years, i.e., we do not ensure that the speciation times correspond precisely to their estimates in the literature. However, the goal of this work is not to develop a model that can reproduce the exact quantitative features of the empirical distribution of gene family size but, rather, to apply the *H. sapiens* parameterizations to get an indication of whether the deviation from the power-law observed in the empirical data is the result of the two rounds of WGD. Given the above and also the raging debate about the large errors in timing major divergence times in the chordate species tree (Graur and Martin 2004; Hedges and Kumar 2004), we do not consider this discrepancy to be of importance.

The simulation produces data with clear peaks at 3, 6, 9, and 12 (see Fig. 7). Two rounds of WGD with hazard shift should result in an overrepresentation of families of size 4 and the subsequent two speciations produce three species, thus explaining the peak at 12. The peaks at 3, 6, and 9 are due to the fact that not all genes with a hazard shift will

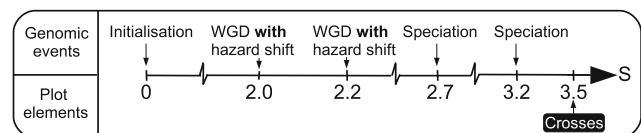
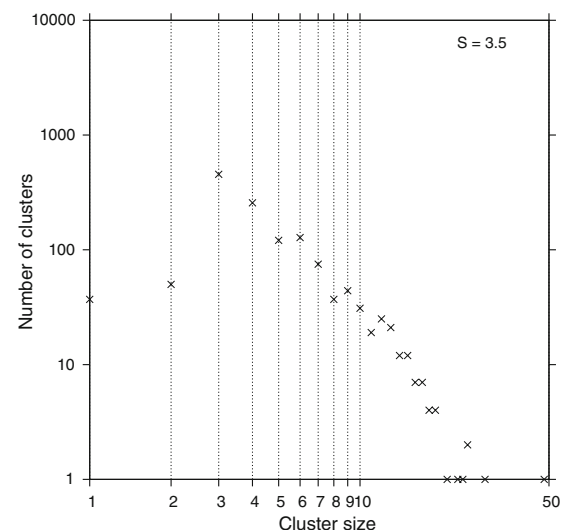


Fig. 7 Effect of two WGDs with hazard shift followed by two speciations on the distribution of gene family size. A clear wave pattern in the data with a period of three, i.e., equal to the number of species (two speciations lead to three species). This pattern is present despite $0.3 S$ having passed since the second speciation

retain all duplicate copies following the WGD: some will lose one duplicate (size 9), some will lose two (size 6), and most will lose three (size 3). This is qualitatively very similar to the empirical data in Fig. 2, with the only difference being that the data plotted in Fig. 2 are for five species instead of the three (generated by the two

speciations). The fact that the waves are not detectable at larger family sizes in the empirical data is probably due to the fact that larger families, due to their large number of genes, have a higher probability of containing a gene that duplicates. This results in a less stable size and thus a lack of conservation of the signal of the WGD.

Discussion

Small- and Large-Scale Gene Duplication

The redundancy generated by gene duplication has long been hypothesized to provide the raw material from which new function can evolve (Ohno 1970) and, as such, is of great interest. Small-scale gene duplication is known to occur in many species at a high and relatively constant level through tandem and segmental duplication (Lynch and Conery 2000, 2003). WGD is also known to occur and has the potential to rapidly and dramatically change the gene content of a genome, but can be difficult to detect if it occurred in the distant past. Probably the two most studied WGDs are those originally hypothesized by Ohno (1970) to have occurred prior to the fish-tetrapod split. There is now strong evidence for two rounds of WGD (2R), possibly in quick succession, prior to the divergence of ray-finned and lobe-finned fish (Wang and Gu 2000; Dehal and Boore 2005) and a ray-finned fish-specific WGD (3R) prior to the radiation of teleosts (Christoffels et al. 2004; Vandepoele et al. 2004). However, there has been considerable debate about the number of WGDs in the ancestral chordate (Friedman and Hughes 2001, 2003; Abi-Rached et al. 2002; McLysaght et al. 2002).

There are two main reasons why ancient large-scale duplication events such as WGDs can be difficult to detect. First, the retention rate of duplicates generated through a WGD is often not very high, although it is generally thought to be higher than the retention rate of SSD. This results in a weak signal in the genome, e.g., it is estimated that only ~20% of duplicates were retained in pufferfish and zebrafish following the fish-specific whole-genome duplication (Jaillon et al. 2004; Woods et al. 2005; Brunet et al. 2006). Second, there is a high level of SSD and loss: genes are constantly subject to a high probability of being duplicated through a SSD event and the resulting duplicates are themselves subject to a high probability of pseudogenization (Lynch and Conery 2000, 2003; Hughes and Liberles 2007).

The 2R Debate

To prove the occurrence of a WGD it is necessary to statistically test whether the null hypothesis, that the data

were produced by the background process of SSD, can be rejected in favor of the hypothesis of a large-scale duplication. To carry out such a test it is necessary to obtain estimates of the background SSD and pseudogenization rates. If the event is hypothesized to have occurred in the relatively recent past, then it is possible to tackle this problem as a significant percentage of the duplicates generated by the WGD are still functional and it is possible to get a relatively accurate estimate of the SSD and loss process since the hypothesized event (Maere et al. 2005). If, however, the hypothesized event is more ancient, obtaining an accurate estimation of the SSD and loss process that applied during the period since the hypothesized WGD and formulating a statistical test are impossible.

In the absence of a formal test, advocates and opponents of 2R have studied many features of homologous gene families in order to gather corroborative evidence for their hypotheses, e.g., the number of genes in multigene families, timing of duplications, and genomic location of paralogs. The most recent study, which uses the full-genome sequences of *Ciona intestinalis*, *Homo sapiens*, *Mus musculus*, and *Takifugu rubripes*, provides all these types of data (Dehal and Boore 2005). They find that only the genomic location data provide clear evidence of 2R. The data on the number of genes per family for a given species (as well as data on number and timing of duplications) are dismissed as uninformative, as opponents of the 2R hypothesis did previously (Hughes et al. 2001).

Although we agree with the conclusion that the ancestral vertebrate is very likely to have undergone two WGDs, we do not agree that there is no sign of these events in gene family size data.

Deviation from the Power-Law

The motivation for this study was the observation that a clustering of all genes from five tetrapod species produced a distribution of gene family sizes with “waves” with a period of five (see Fig. 2). This represents a clear deviation from the expected power-law distribution. A study of gene family phylogenetic trees from fully sequenced vertebrate genomes (Blomme et al. 2006) has shown that a large proportion of duplicated genes in extant vertebrate genomes is ancient and was created at times that coincide with the proposed WGD events. The same study also established that regulatory genes have a higher probability of being duplicated and retained through WGD than SSD, and it was noted that this is consistent with the dosage balance hypothesis. These findings suggest that WGD has the potential to fundamentally and persistently modify the distribution of gene family size.

Through several sets of simulations, we have shown that the deviation from the power-law that we observe in the

empirical data for sizes less than five can be attributed to the speciation events, but for larger sizes the “waves” are best explained by the two ancient WGDs. Moreover, we have shown that the WGD event must have a specific characteristic, namely, that a significant fraction of genes must undergo a hazard shift in WGD compared to the hazard they are subject to following SSD. Critics might be tempted to suggest that the deviations from the power-law are an artifact of the clustering algorithm. This, however, is very unlikely given that the empirical data were clustering using the MCL algorithm (Enright et al. 2002), which is the algorithm used by Ensembl (Birney et al. 2006) to produce gene families, whereas the simulated data were clustered using complete linkage.

These results are of interest for several reasons. First, they show that “simple” data on the size of gene families can provide an indication of ancient large scale duplication. Note, however, that we do not claim that gene family size provides as strong evidence for WGD as the spatiotemporal data of the Dehal and Boore (2005) study. Second, it is remarkable that the signal is still detectable given that ~500 million years of SSD and loss separate us from the WGD events. This suggests that there exists a significant fraction of genes with low retention following SSD which are subject to high retention following WGD. Third, it shows that a WGD modifies the structure of the genome’s gene content in a profound and persistent way, a finding which is consistent with the observed correlation between WGD and major evolutionary transitions.

Molecular Basis of the Model of Whole-Genome Duplication and Retention

A key component of the model needed to produce simulated data that qualitatively match the empirical data is the WGD with hazard shift. This is consistent both with the dosage balance (Aury et al. 2006) and subfunctionalization model (Force et al. 1999) and with a different mutational opportunity in the “different on arrival” mechanism.

The theory behind dosage balance is that certain categories of protein-coding genes, such as proteins forming complexes or enzymes in a metabolic pathway, are very sensitive to the stoichiometry of their interaction partners. If such a gene duplicates through a SSD, it will have a negative fitness effect by disturbing the stoichiometry and will be selected against. If, on the other hand, it is duplicated in a WGD, it is duplicated with all its interaction partners and the stoichiometry of the interacting partners is unchanged. Thus, the fitness effect of the duplicates is neutral. However, the loss of any of the duplicates will have a negative fitness effect by upsetting the stoichiometry. Under such a model there are two broad types of genes: those that are dosage sensitive and those that are

not. For those that are dosage sensitive, the probability of retention depends on whether the duplication occurred through SSD or large-scale duplication; for those that are not dosage sensitive, the scale of the duplication is immaterial. Moreover, the genes that are dosage sensitive will tend to be those with a low probability of retention following SSD, but they will have a high probability of retention following WGD.

In the subfunctionalization model, where retention is driven by temporal or spatial partitioning of expression through complementary loss of regulatory regions between a duplicate pair (Force et al. 1999), the probability of retention is an increasing function of the number of regulatory modules. Thus, as long as all regulatory modules are duplicated, the duplicate’s hazard rate should not depend on whether the duplication was small-scale or whole-genome (more on this below). In addition to the subfunctionalization of regulatory regions, another variant of the subfunctionalization model involves the partitioning of function in the coding sequence, causing, for example, the partitioning of interaction partners. Here, as a protein has many interacting partners, this provides opportunities for both subfunctionalization of interactions (proportional to the number of interactions) and neofunctionalization of each interacting partner as an independent probability. Once a copy of a duplicated interacting partner neofunctionalizes, the subfunctionalization of the interactions with each partner can then lead to fixation of the duplicates. Through this mechanism, a hazard shift is generated after WGD that is different depending on the number and nature of physical interactions with other proteins.

Another difference between SSD and WGD duplicates derives from the difference in gene structure following the initial mutational event. It is becoming increasingly evident that a gene’s regulatory modules are not necessarily located in the immediate vicinity of the gene’s promoter and may even extend into and beyond adjacent transcriptional units (Kikuta et al. 2007). Therefore, the duplicates after SSD but not after WGD may be different upon arrival, including variants such as neofunctionalized on arrival and dead on arrival. This process is also relevant to subfunctionalization, where a duplicated gene may be part of the way to subfunctionalization with some of the regulatory regions already lost. If many regulatory blocks are distant from the genes they regulate, then such genes would have a lower hazard rate if duplicated in a WGD event than if duplicated in a small-scale event. In addition, if a gene subfunctionalizes following WGD, each duplicate is left with only a subset of the ancestral regulatory regions, thus the probability of retention following a future SSD is reduced.

From a theoretical point of view, it is difficult to determine whether it is dosage balance, subfunctionalization, or “difference on arrival” that is most likely to cause

a hazard shift. Because of the negative selection against SSD for dosage-sensitive genes and the negative selection against loss of duplicates following WGD, dosage balance appears to be a strong candidate. However, this shift may only be transient because, if the negative selection against loss is stochastically overcome in small-effective-population-size vertebrates, cooperative positive selection for rapid gene loss will follow. The hazard shift associated with subfunctionalization may be of a more permanent nature.

Empirical data also support the hazard shift model: genes involved in transcription regulation are a large functional class which multiple studies have shown is preferentially retained following WGD compared to SSD (Blanc and Wolfe 2004; Maere et al. 2005; Blomme et al. 2006). But, again, this can be interpreted either as being caused by dosage balance, as regulatory genes are often functional in complexes, or as being caused by subfunctionalization or “difference on arrival,” as they are also often subject to complex regulation involving many enhancer regions.

In our model, the SSD rate is assumed to be a constant and WGD is assumed to duplicate all genes. In both cases, we assume that the duplications reach fixation. This is clearly a simplification, as all genes do not duplicate with a given frequency. However, for models of duplication and retention where the initial duplication event is neutral, such as subfunctionalization, this should be an acceptable simplification. In the case of dosage balance, the initial duplication has a negative fitness effect if the duplication was small-scale. This results in a reduced chance of fixation. Our model does not capture this, as we model differential retention rates across families exclusively through different pseudogenization rates. Thus, this feature of dosage balance is modeled through a higher than average hazard rate following SSD for a fraction of the gene families rather than a reduced duplication rate. This approach is justified by the need to build a model that is consistent with multiple modes of gene duplication and retention while restricting the modeling to the genomic level (rather than descending to the population genetic level, where we do not have estimates of the key processes).

It is important to emphasize that although the original retention of duplicates following WGD is likely to be driven by subfunctionalization or dosage balance, there is a strong possibility that the ultimate fate of at least one of the duplicates is neofunctionalization (Ohno 1970). This is because the retention mechanism, particularly in the case of subfunctionalization, may reduce the level of pleiotropic constraint that was exerted on the ancestral gene prior to duplication, thus allowing, at the very least, fine-tuning of function (Lynch and Force 2000; He and Zhang 2005; Rastogi and Liberles 2005).

Use of *Homo Sapiens* Estimates

The key processes in our model are gene duplication, gene loss, and sequence divergence. These processes are obviously variable across lineages and time. Ideally, we would have reliable estimates of these processes for the past 500 million years. However, producing estimates of these processes in the distant past is effectively impossible due to the directly counteracting nature of the processes of duplication and loss and to the saturation of silent sites with time. As the second-best alternative, we use the estimates for these processes obtained from data on recent SSD duplicates in *Homo sapiens* (Hughes and Liberles 2007). Estimates for other species are available, but these are also for recent duplicates, so we decided to use the high-quality human data rather than create a consensus between multiple species. Although, the numerical values of parameters of the equations are different across species, the functional forms are the same, thus there is no reason to believe that the functional forms were any different in the past, although the parameter values were undoubtedly different.

Again, due to the distant nature of the WGD events, we built a model that only qualitatively matches existing theories for the retention of duplicate genes. We have no basis for the numerical parametrizations of the “hazard shift” model (proportion affected by hazard shift, magnitude of hazard shift, and functional form of the hazard rate following WGD). As a result of this, we are only able to produce simulated data that qualitatively match the empirical data, as can be seen through a comparison of Figs. 2 and 7. We could have fine-tuned parameters in the model and, thus, obtained a better fit between the simulated and the empirical data. For example, by increasing the proportion of families affected by the hazard shift or increasing the size of the hazard shift, the deviation from the power-law in the simulated data would have been stronger and quantitatively more similar to the empirical data, but this would have been misleading. Moreover, it is unnecessary, as we are not claiming to have produced a precise reconstruction of the evolution of the distribution of gene family sizes in the tetrapod lineage. We aimed only to show that the deviation from the power-law in the empirical data was the result of two rounds of WGD, and we consider that our qualitative results show this to be highly probable.

Conclusion

In this study, we have used a model of gene family evolution to produce an approximate characterization of the effects of WGD and speciation on the distribution of gene family size. We find that, for our simulations to produce the kind of deviation from the power-law observed in the

empirical distribution of gene family size for several tetrapod species, it is necessary that a significant proportion of genes is subject to a high probability of retention following WGD and that these genes also have a low probability of retention following SSD. Whether this difference in probability of retention is the result of the fixation of duplication events being selected against following SSD and loss selected against following WGD (as in the dosage balance model) or whether it is due to a shift in the pseudogenization rate between SSD and WGD (as in the subfunctionalization model) is not known. Given that it is difficult to imagine what other type of genomic event would disrupt the distribution in this way and that strong evidence already exists for two WGDs in the ancestral vertebrate ~500 Mya, we find it logical to conclude that the pattern that we observe in the empirical distribution of gene family size for tetrapods is the result of the ancient WGDs. This implies that WGD may profoundly and persistently modify the distribution of gene family size.

Acknowledgment This work was funded by FUGE, the functional genomics platform of the Norwegian Research Council.

References

- Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H (2002) Evidence of en bloc duplication in vertebrate genomes. *Nature Genet* 31:100–105
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
- Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O, Billaut A, Beisson J, Blanc I, Bouhouche K, Câmara F, Duharcourt S, Guigo R, Gogendeau D, Katinka M, Keller AM, Kissmehl R, Klotz C, Koll F, Mouël AL, Lepère G, Malinsky S, Nowacki M, Nowak JK, Plattner H, Poulain J, Ruiz F, Serrano V, Zagulski M, Dessen P, Béttermier M, Weissenbach J, Scarpelli C, Schächter V, Sperling L, Meyer E, Cohen J, Wincker P (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444:171–178
- Birney E, Andrews D, Caccamo M et al (2006) Ensembl 2006. *Nucleic Acids Res* 34:D556–D561
- Blair JE, Hedges SB (2005) Molecular phylogeny and divergence times of deuterostome animals. *Mol Biol Evol* 22:2275–2284
- Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16:1679–1691
- Blomme T, Vandepoele K, Bodt SD, Simillion C, Maere S, van de Peer Y (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol* 7:R43
- Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson Rechavi M (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* 23:1808–1816
- Christoffels A, Koh EGL, Chia JM, Brenner S, Aparicio S, Venkatesh B (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol* 21:1146–1151
- Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 3:e314
- Demuth JP, Bie TD, Stajich JE, Cristianini N, Hahn MW (2006) The evolution of mammalian gene families. *PLoS ONE* 1:e85
- Enright AJ, Dongen SV, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575–1584
- Enright AJ, Kunin V, Ouzounis CA (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res* 31:4632–4638
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545
- Friedman R, Hughes AL (2001) Pattern and timing of gene duplication in animal genomes. *Genome Res* 11:1842–1847
- Friedman R, Hughes AL (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol Biol Evol* 20:154–161
- Gilad Y, Man O, Pääbo S, Lancet D (2003) Human specific loss of olfactory receptor genes. *Proc Natl Acad Sci USA* 100:3324–3327
- Graur D, Martin W (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet* 20:80–86
- Harrison PM, Gerstein M (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J Mol Biol* 318:1155–1174
- He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169:1157–1164
- Hedges SB, Kumar S (2004) Precision of molecular time estimates. *Trends Genet* 20:242–247
- Hughes AL, da Silva J, Friedman R (2001) Ancient genome duplications did not structure the human hox-bearing chromosomes. *Genome Res* 11:771–780
- Hughes T, Liberles D (2007) The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation. *J Mol Evol* 65:574–588
- Hughes T, Liberles DA (2008) The power-law distribution of gene family size is driven by the pseudogenisation rate's heterogeneity between gene families. *Gene* 414:85–94
- Huynen MA, van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol* 15:583–589
- Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biémont C, Skalli Z, Cattolico L, Poulain J, Berardinis VD, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigó R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quétiér F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Crollius HR (2004) Genome duplication in the teleost fish *tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431:946–957
- Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, Ghislain J, Pezeron G, Mourrain P, Ellingsen S, Oates AC, Thisse C, Thisse B, Foucher I, Adolf B, Geling A, Lenhard B, Becker TS (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res* 17:545–555

- Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Rev Microbiol* 1:127–136
- Lundin LG, Larhammar D, Hallböök F (2003) Numerous groups of chromosomal regional paralogs strongly indicate two genome doublings at the root of the vertebrates. *J Struct Funct Genomics* 3:53–63
- Luscombe NM, Qian J, Zhang Z, Johnson T, Gerstein M (2002) The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol* 3(8):research0040.1-0040.7. Available at: <http://www.genomebiology.com/2002/3/8/research/0040>
- Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155
- Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes. *J Struct Funct Genomics* 3:35–44
- Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–473
- Maere S, Bodt SD, Raes J, Casneuf T, Montagu MV, Kuiper M, de Peer YV (2005) Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci USA* 102:5454–5459
- McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. *Nature Genet* 31:200–204
- Ohno S (1970) *Evolution by gene duplication*. Springer-Verlag, New York
- Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, Ouzounis CA (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. *Bioinformatics* 16:915–922
- Rastogi S, Liberles DA (2005) Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol* 5:28
- Rastogi S, Reuter N, Liberles DA (2006) Evaluation of models for the evolution of protein sequences and functions under structural constraint. *Biophys Chem* 124:134–144
- Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA* 95:6239–6244
- Roth C, Betts MJ, Steffansson P, Saelensminde G, Liberles DA (2005) The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics. *Nucleic Acids Res* 33:D495–D497
- Vandepoel K, Vos WD, Taylor JS, Meyer A, de Peer YV (2004) Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci USA* 101:1638–1643
- Wang Y, Gu X (2000) Evolutionary patterns of gene families generated in the early stage of vertebrates. *J Mol Evol* 51:88–96
- Woods IG, Wilson C, Friedlander B, Chang P, Reyes DK, Nix R, Kelly PD, Chu F, Postlethwait JH, Talbot WS (2005) The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res* 15:1307–1314
- Yanai I, Camacho CJ, DeLisi C (2000) Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. *Phys Rev Lett* 85:2641–2644
- Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol* 46:409–418