Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes

Anuphap Prachumwat and Wen-Hsiung Li

Genome Res. 2008 18: 221-232 originally published online December 14, 2007
Access the most recent version at doi:10.1101/gr.7046608

Letter

Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes

Anuphap Prachumwat and Wen-Hsiung Li[1]

Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA

Where did vertebrate genes come from? Here we address this question by analyzing eight completely sequenced land vertebrate genomes and six completely sequenced invertebrate genomes. Approximately 70% of the vertebrate genes can be found in the six invertebrate genomes with the standard homology search criteria (denoted as *V.MCL*), another ~6% can be found with relaxed search criteria, and an additional ~2% can be found in sequenced fungal and bacterial genomes. Thus, a substantial proportion of vertebrate genes (~22%) cannot be found in the nonvertebrate genomes studied (denoted as *Vonly*). Interestingly, genes in *Vonly* are predominantly singletons, while the majority of genes in the other three groups belong to gene families. The proteins of *Vonly* tend to evolve faster than those of *V.MCL*. Surprisingly, in many cases the family sizes in *V.MCL* are only as large as or even smaller than their counterparts in the invertebrates, contrary to the general perception of a larger family size in vertebrates. Interestingly, in comparison with the family size in invertebrates, vertebrate gene families involved in regulation, signal transduction, transcription, protein transport, and protein modification tend to be expanded, whereas those involved in metabolic processes tend to be contracted. Furthermore, for almost all of the functional categories with family size expansion in vertebrates, the number of gene types (i.e., the number of singletons plus the number of gene families) tends to be over-represented in *Vonly*, but under-represented in *V.MCL*. Our study suggests that gene function is a major determinant of gene family size.

[Supplemental material is available online at www.genome.org.]

The invertebrate-to-vertebrate transition was a major evolutionary event in the evolution of the animal kingdom, so there has been much interest in the genetic differences between vertebrates and invertebrates. Ohno (1970) had stressed the contribution of gene duplication to this transition and postulated that two rounds of whole-genome duplication (WGD) occurred in the early evolution of the vertebrates. The existence of multiple copies for many types of genes in vertebrate genomes has been taken as evidence for two rounds of WGD (e.g., the *HOX* gene clusters and genes near the major histocompatibility complex regions) (Holland et al. 1994; Kasahara et al. 1996). However, some authors argued for only one round of WGD and a few authors totally rejected the WGD hypothesis (for reviews, see Meyer and Van de Peer 2005; Panopoulou and Poustka 2005). Recently, Blomme et al. (2006) and Nakatani et al. (2007) have provided further evidence for two rounds of WGD. Regardless of how many rounds of WGD had actually occurred, a significant increase in gene number in vertebrate genomes in comparison to invertebrate genomes is well documented. This gene number expansion contributed to the evolution of vertebrate genomes, including our own.

The question is then, "what are the relationships between vertebrate and invertebrate genes?" To address this question we conduct the following analyses. First, we divide the genes in the eight completely sequenced vertebrate genomes into four groups: (1) *V.MCL*: Genes that can be found, with standard homology search criteria, in one or more of the six invertebrate genomes that have been completely sequenced with good annotation. (2) *V.A50*: Genes that are not in *V.MCL*, but can be found when the homology search criteria are relaxed. (3) *VFProk*: Genes that are in neither *V.MCL* nor *V.A50*, but can be found in one or more of the available fungal or prokaryotic genomes. (4) *Vonly*: Genes that are none of the above, so that they are putatively vertebrate-specific genes. This analysis may give us a rough idea about the origins of vertebrate genes. Second, we compare the family size distributions for the above four groups of vertebrate genes. This analysis may reveal which groups of genes tend to be in a gene family and which groups of genes tend to exist in a single copy (singleton). Third, for *V.MCL* genes, we conduct a linear regression analysis of vertebrate gene family sizes on the sizes of the orthologous gene families in invertebrates. This analysis may reveal overall gene number expansions and contractions in the vertebrate genomes. Moreover, a Gene Ontology term analysis may reveal whether there is functional bias of the gene families that have been expanded or contracted and whether functional bias exists among the four groups of genes. Fourth, we compute the relative rates of amino acid substitution for proteins in the four groups of genes. This may shed light on whether a fast rate of amino acid substitution is a possible reason why some of the vertebrate proteins cannot be found by a homology search in the invertebrate genomes studied. Finally, we examine the invertebrate genes that are not found to have homologs in the eight vertebrate genomes studied. Blomme et al. (2006) have studied gene losses and gene gains during the evolution of vertebrates, especially with respect to gene number differences between paleotetraploid fish and land vertebrate genomes. In contrast, we examine gene number expansions and contractions in genomes of land vertebrates relative to gene copy numbers in invertebrate

[1]Corresponding author.
E-mail whli@uchicago.edu; fax (773) 702-9740.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.7046608.

genomes. Moreover, while they excluded vertebrate-specific genes, we include such genes and look into their function and evolution.

Results

Vertebrate and invertebrate gene copies and gene types

There are 20,404 gene types (the sum of the number of singletons and the number of gene families) found in the eight land vertebrate genomes (Table 1). Surprisingly, it is ~11,000 types fewer than that in the six invertebrate genomes, but the difference is largely due to a large number of genes found only in a single invertebrate genome (i.e., putatively species-specific genes) (Table 1). On the other hand, an average vertebrate genome has ~22% more gene copies than an invertebrate genome, as indicated by a higher average gene copy number per genome of vertebrates (19,863) than that of invertebrates (16,383).

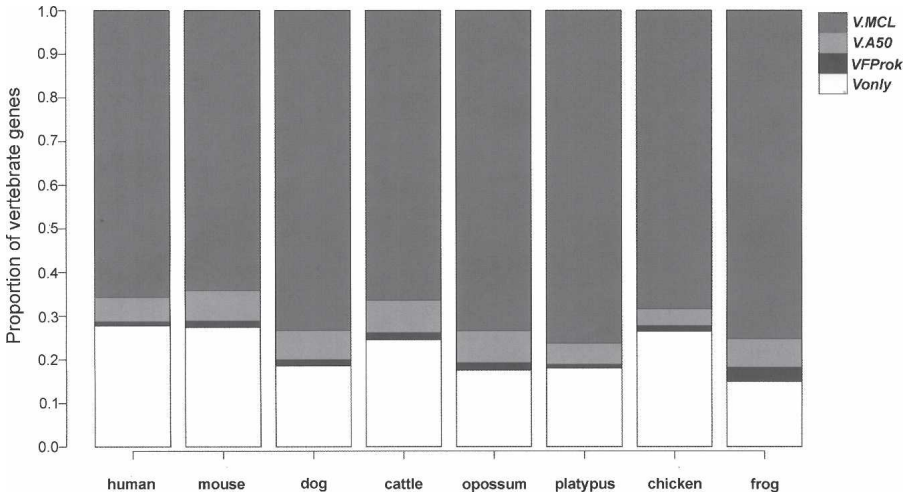

Figure 1. Proportions of vertebrate genes in the four gene groups. For each vertebrate genome on the *X*-axis, the gene groups are represented, from *top* to *bottom*, by segments of a bar for the genes in *V.MCL* (genes can be found in at least one of the six invertebrate genomes with standard homology search criteria), in *V.A50* (genes are not in *V.MCL*, but can be found in the eight invertebrates when the homology search criteria are relaxed), in *VFProk* (genes are in neither *V.MCL* nor *V.A50*, but can be found in fungal or prokaryotic genomes), and in *Vonly* (genes are in none of the above three groups).

Proportion of vertebrate genes found in nonvertebrate genomes

Figure 1 shows that under the standard homology search criteria (see Methods), on average, 70% of the genes that are present in one or more of the eight vertebrate genomes have homologs in at least one of the six invertebrate genomes (denoted as *V.MCL* genes). These *V.MCL* genes belong to 6264 gene types (Table 1). Under less-stringent criteria (see Methods), an additional 6.2% of the vertebrate genes have homologs in at least one of the eight invertebrate genomes (*V.A50*), and an additional 1.5% vertebrate genes have homologs in either fungal or prokaryotic genomes, but not in the invertebrate genomes (*VFProk*). This leaves ~22% of the vertebrate genes with no homolog in the nonvertebrate genomes studied (*Vonly*) (Fig. 1; Tables 1, 2). For the *V.A50* and *VFProk* groups, almost all of the families have only a single vertebrate gene copy (or ≥2 gene copies in only a single vertebrate genome) with a homolog hit in nonvertebrate genomes (data not shown). The proportion of *V.MCL* genes presented here is similar to that from a recent analysis using a different method (Blomme et al. 2006) and from two versions of Ensembl Compara (v.35 and v.42).

In the six invertebrate genomes, on average, ~60% of invertebrate genes have homologs in at least one of the 12 vertebrate genomes (denoted as *I.MCL* and *I.A50* genes; Supplemental Fig. 1S). However, a large proportion (62%) of worm genes cannot be found in the 12 vertebrate genomes studied (denoted as *Ionly* genes; Supplemental Fig. 1S; Supplemental Table 1S). When the worm genome is excluded, ~66% of invertebrate genes are found in at least one of the 12 vertebrate genomes. This proportion remains lower than that of vertebrate genes: The sum of *I.MCL* and *I.A50* genes is 60%–66%, while the sum of *V.MCL* and *V.A50* genes is 76%.

Table 1. Numbers of gene types and gene copies in the vertebrate and invertebrate genomes

	Vertebrates (eight genomes)			Invertebrates (six genomes)		
Gene classification	Vertebrate gene group	No. of gene types[a]	No. of gene copies	Invertebrate gene group	No. of gene types[a]	No. of gene copies
Genes shared by both vertebrate and invertebrate genomes	*V.MCL*	6264	111,018	*I.MCL*	6372[b]	58,851[b]
	V.A50	572	9985	*I.A50*	346	1568
Genes without a homolog between vertebrate and invertebrate genomes studied	*VFProk*	530	2455		NA	NA
	Vonly genes shared by ≥2 vertebrate genomes	2881	24,068	*Ionly* genes shared by ≥2 invertebrate genomes	1730	6987
	Vonly genes found in only a single vertebrate genome	10,157	11,375	*Ionly* genes found in only a single invertebrate genome	23,071	30,891
All genes		20,404	158,901		31,519	98,297

[a]Number of gene types is the sum of the number of singletons and the number of gene families.
[b]A total of 349 invertebrate gene copies are in the 170 families that are homologous to the four paleotetraploid fishes genomes but are not found in the eight tetrapod genomes under study.
NA, Not applicable.

Table 2. Numbers of genes (and gene types) in different vertebrate gene groups

Vertebrate gene group	Human	Mouse	Dog	Cattle	Opossum	Platypus	Chicken	Frog
V.MCL	14,698 (6069)	15,572 (6082)	14,125 (6037)	15,238 (5963)	14,297 (6005)	11,987 (5242)	11,376 (5524)	13,727 (5652)
V.A50	1250 (443)	1702 (432)	1287 (417)	1714 (415)	1435 (402)	761 (323)	640 (353)	1196 (380)
VFProk	219 (77)	348 (71)	273 (68)	361 (72)	324 (60)	136 (53)	209 (60)	585 (487)
Vonly	6204 (4738)	6646 (4580)	3578 (2668)	5635 (4316)	3424 (2325)	2828 (1559)	4397 (3536)	2729 (1878)
Vonly genes shared by ≥2 vertebrate genomes	3604 (2470)	4065 (2421)	3289 (2397)	3609 (2347)	2991 (1927)	2648 (1387)	2073 (1425)	1787 (1069)
All genes	22,371 (11,327)	24,264 (11,165)	19,263 (9190)	22,948 (10,766)	19,480 (8792)	15,712 (7177)	16,622 (9473)	18,237 (8397)

The number of gene types is the sum of the number of singletons and the number of gene families.

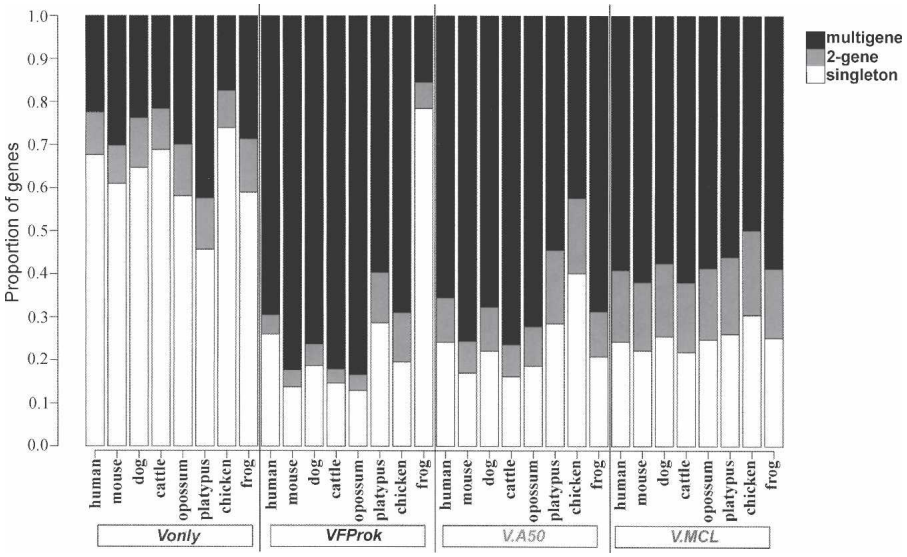

Figure 2. Proportions of the vertebrate *V.MCL*, *V.A50*, *VFProk*, and *Vonly* genes that are singletons, in two-gene families, or in multigene families.

Among the 13,038 *Vonly* gene types (35,443 gene copies) (Table 1), ~22% are found in ≥2 vertebrate genomes (2881 gene types with a total of 24,068 gene copies) and the rest (~78%, 10,157 gene types with 11,375 gene copies) are found in only one of the eight vertebrate genomes under study. For a vertebrate genome under study, the proportion of *Vonly* gene types that are shared with another vertebrate genome is, on average, ~65%; in some genomes, the vast majority of the *Vonly* genes (>80%; e.g., dog, opossum, and platypus) are found in the other vertebrate genomes (Table 2). On the other hand, the invertebrate genes that are not found in the 12 vertebrate genomes under study (i.e., *Ionly* genes) have only ~7% gene types that can be found in ≥2 invertebrate genomes (1730 gene types with a total 6987 gene copies); the other 23,071 gene types (a total of 30,891 genes) are found in only one of the six invertebrate genomes under study (Table 1). Similarly, a small proportion of *Ionly* genes within an invertebrate genome are shared with another invertebrate genome (Supplemental Table 1S). In particular, the proportions for the two-mosquito and the fruitfly genomes are ~25%–48%, whereas those in the other genomes are <5%. Almost all (~99%) of *Ionly* genes in the worm genome are putatively worm-specific genes. Therefore, the worm genome contributes ~21% to the total *Ionly* gene types, while the other five invertebrate genomes each contribute only ~15% (Supplemental Table 1S).

Family size distribution of vertebrate genes

Figure 2 shows that the *Vonly* genes are mainly singletons, but only a small proportion of the *V.MCL* genes are singletons (~62% vs. ~25%). The proportion of singletons in *Vonly* increases to >85% for those found in a single vertebrate genome and decreases to ~55% for those found in ≥2 vertebrate genomes. On average, 14% of gene copies in a vertebrate genome are *Vonly* singletons; the human, mouse, cow, and chicken genomes have a larger proportion (17%–20%) than do the dog, opossum, platypus, and frog genomes (8%–12%). The distribution patterns of the *V.A50* and the *VFProk* genes resemble that of the *V.MCL* genes in that a large proportion of the genes are in multigene families (63%–65%). However, a large proportion (80%) of frog's *VFProk* genes are singletons (Fig. 2), and the proportion of frog's *VFProk* genes is larger than that of the other vertebrate genomes (Fig. 1).

Family size distribution of invertebrate genes

Similar to the vertebrate genomes, the majority of the *Ionly* genes, which are genes that are not found in vertebrates, are singletons, whereas only small proportions of the *I.MCL* and *I.A50* genes are singletons (Supplemental Fig. 2S). For the corresponding vertebrate and invertebrate gene groups (e.g., *Vonly* vs. *Ionly* and *V.MCL* vs. *I.MCL*), the proportions of singletons in the vertebrate genomes tend to be smaller than those in the invertebrate genomes (e.g., 25% *vs.* 33% for *V.MCL vs. I.MCL*). However, for the worm and the sea urchin genomes, the proportions of singletons are smaller than those of the other invertebrate genomes, but similar to those of the vertebrate genomes (Fig. 2; Supplemental Fig. 2S). In particular, a larger proportion of *Ionly* genes in the worm and the sea urchin genomes (36% and 30%, respectively) form multigene families than do the other invertebrate genomes (14%–17%).

Gene copy number difference between vertebrates and invertebrates

A positive correlation in average family size between vertebrates (n_v) and invertebrates (n_i) is observed with an estimated slope (β_1) = 1.15 (R^2 = 0.63) and 0.92 from the simple linear regression model and the robust linear regression model, respectively (Fig. 3). The n_i was set to 1 for those families whose average family size over the six invertebrates is <1, because at least one gene copy should have existed in the common ancestor of vertebrates. The estimated slopes are ~1 and the intercepts are close to 1, so that the regression lines are close to the line $n_v = n_i$. This observation suggests that the family size in the vertebrates is largely determined by the family size in the invertebrates and that in many cases the family sizes in the vertebrates are smaller than those in the invertebrates. However, 46% of the points are located above the $n_v = n_i$ line, while only 39% are below the line, so that the average family size in the vertebrates is larger than that in the invertebrates (the means of n_v and n_i are 2.21 and 1.77, respectively; $P = 7.77 \times 10^{-26}$, the Student's two-sample paired *t*-test).

Vertebrate gene expansion vs. the number of invertebrate genomes sharing homologous genes

Figure 4 shows the distribution of average family sizes in the eight vertebrate genomes (n_v) of the *V.MCL* families that are grouped by the largest family size of homologous genes in the six invertebrate genomes (singleton, two-gene family, or multigene family) and by the number of the invertebrate genomes in which the homologs can be found. On average, family size has expanded more for the vertebrate families homologous to invertebrate multigene families than for the families homologous to invertebrate singletons. Also, vertebrate family size has expanded more for the families that have homologs present in a larger

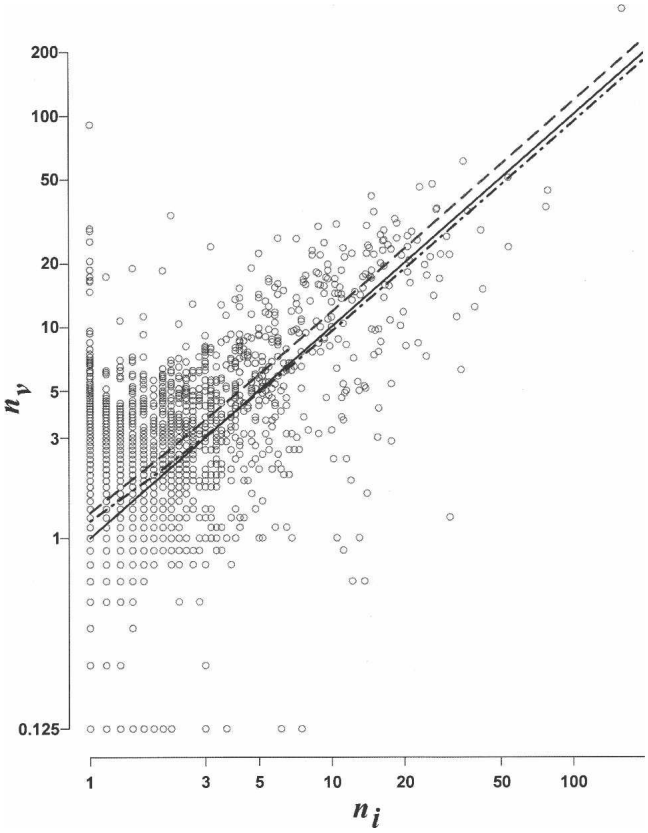

Figure 3. The family-size correlation between the vertebrate and invertebrate genomes. Scatter plots, on the log-scale axes, show the average vertebrate family size (n_v) against the average invertebrate family size (n_i) for each *V.MCL* gene family. The black solid line represents $n_v = n_i$, while the one-dashed and two-dashed lines represent the fitted lines from the simple linear regression model and the robust linear regression model, respectively ($n_v = \beta_0 + \beta_1 n_i$). The slopes (i.e., β_1 estimates) for the one-dashed and two-dashed fitted lines are 1.15 ($R^2 = 0.63$) and 0.92, with the bootstrapped 95% confidence intervals of β_1 estimates being (0.72, 1.54) and (0.78, 1.09), respectively. The bootstrapped *P*-values are 0.73 and 0.21 for the null hypothesis of $\beta_1 = 1$ from the simple regression model and the robust linear regression model, respectively, so that the slopes are not significantly different from 1.

number of the invertebrate genomes. In particular, a larger vertebrate family size expansion is observed for those gene families that have homologous multigene families in the invertebrate genomes. For the vertebrate genes that are homologous to invertebrate multigene families, the mean (median) of n_v is 15.6 (11.6) for the vertebrate genes with homologs found in all six invertebrate genomes, but is only 2.1 (1.4) for the vertebrate genes with homologs found in ≤2 invertebrate genomes.

Gene copy number difference between vertebrates and invertebrates by functional categories

As noted above, the average family sizes are, in many cases, similar for vertebrate and invertebrate genes (Fig. 3). However, we now investigate whether a difference in family size between vertebrates and invertebrates exists in some functional categories. The slope of a linear regression model for a specific functional category is an indicator of how the family sizes in the vertebrate genomes differ from those of the invertebrates. The slopes from linear regression models between n_v and n_i are indeed ~1 for most of the biological processes. However, slopes >1 are observed for the genes involved in transcription, regulation of biological process, signal transduction, protein transport, protein modification, organelle organization and biogenesis, and cellular component organization and biogenesis (Fig. 5). Although the 95% confidence interval of the bootstrapped slopes for some of these categories includes 1, the mean of the bootstrapped slopes is >1. Some functional categories (e.g., cell cycle and multicellular organization and development) have a relatively high mean of bootstrapped slopes, though the 95% confidence intervals include 1 in both simple and robust linear regression models. On the other hand, some processes (e.g., metabolic processes and electron transport) have a slope <1. Similarly, the proteins localized to nucleus, chromosome, intracellular, and protein complex have a slope >1 in the linear regression model, while the proteins localized to mitochondrion or annotated with a "cell" term have a slope <1 (Supplemental Fig. 3S). The pattern for molecular function categories, in general, recapitulates the categories in the biological processes (data not shown).

Functional bias in vertebrate gene groups

The patterns of functional bias for the number of gene types and for the number of gene copies are similar, though more significantly differential categories are observed for gene copy number than for gene type number, because there are more gene copies than gene types (see Supplemental materials). Here, we present the patterns of functional bias for the number of gene types and indicate those cases for the number of gene copies where they are different (for the complete results on the number of gene copies, see Supplemental materials and Supplemental Figs. 4S and 6S). In general, the functional categories that are significantly over-represented (under-represented) in the *Vonly* genes are significantly under-represented (over-represented) in the *V.MCL* genes (χ^2 test, FDR <0.05) (Fig. 6; Supplemental Figs. 4S, 6S). The top eight biological processes that the number of gene types are significantly over-represented in *Vonly* are involved in signal transduction, regulation of biological process, transcription, ion transport, responses to external, biotic and abiotic stimuli, response to stress, and multicellular organismal development (Fig. 6). The gene types in these functional categories are significantly under-represented in the *V.MCL* genes. Note that, contrary to functional bias for the number of gene types, the number of gene copies involved in transcription is significantly over-represented in *V.MCL* and under-represented in *Vonly* (see Supplemental materials). On the other hand, the genes that are involved in translation, protein modification, biosynthesis, cellular component organization and biogenesis, and metabolic processes tend to be over-represented in *V.MCL* but under-represented in *Vonly*. For *V.A50* and *VFProk*, a significant functional bias for the number of gene types is found in only a few functional categories. For example, the *V.A50* genes involved in signal transduction and cell growth are over-represented. A significant functional bias for the *VFProk* genes is found in similar biological process categories as those for the *V.MCL* genes, except for those involved in transport and protein modification.

In terms of cellular localization, the number of gene types of proteins localized to extracellular region, extracellular space, plasma membrane, and chromosomes are over-represented in *Vonly*, but under-represented in *V.MCL* (Supplemental Fig. 5S). On the other hand, the number of gene types of those proteins localized to intracellular space, cytoplasm, mitochondrion, ribosomes, and nucleus are over-represented in *V.MCL*, but under-represented in *Vonly*. Similar to the biological process, a signifi-

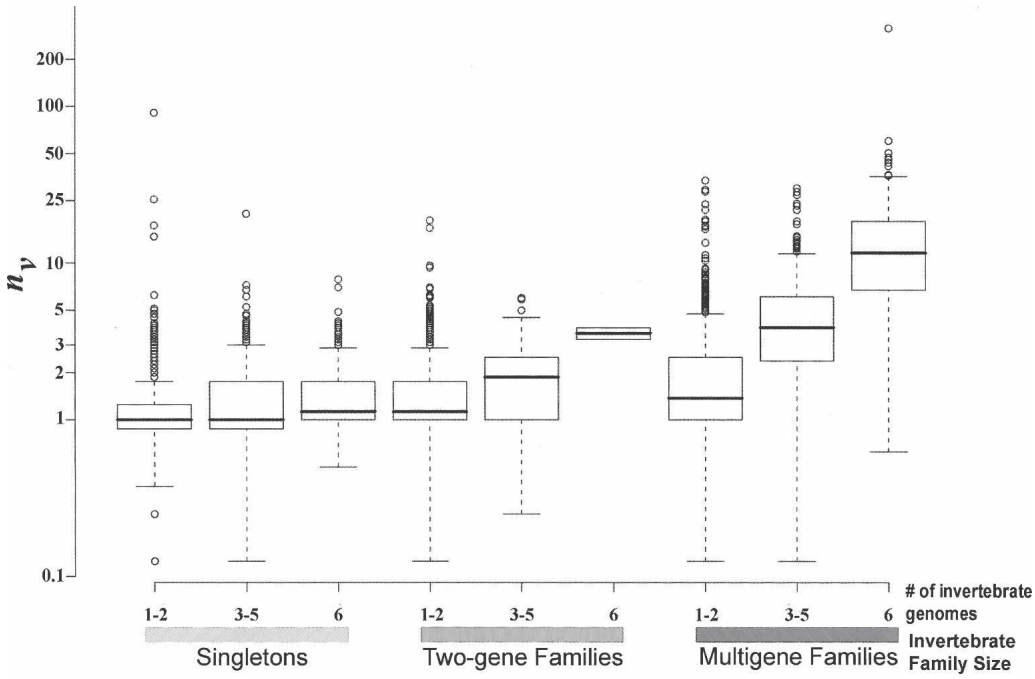

Figure 4. The boxplot of the average vertebrate family sizes (n_v) grouped by the largest family size of homologous genes in the six invertebrate genomes and by the number of the invertebrate genomes where homologous copies can be found. The *X*-axis shows the largest family size of homologous genes in the six invertebrate genomes (singletons, two-gene families, and multigene families are represented by light-gray, medium gray, and dark-gray, respectively) and also the number of the invertebrate genomes in three groups (present in only one or two invertebrate genomes, in three to five invertebrate genomes, or in all six invertebrate genomes).

cant functional bias for the number of gene types is seen only in a few cellular component categories for *V.A50* and *VFProk* (for additional significant functional bias categories in these two groups, see Supplemental materials and Supplemental Fig. 6S). The over-represented molecular function categories in general recapitulate the categories in the biological processes (data not shown).

Functional bias in invertebrate gene groups

Similar to the pattern of vertebrate genes, the functional categories that are significantly over-represented (under-represented) in *Ionly* are significantly under-represented (over-represented) in the *I.MCL* genes (χ^2 test, FDR < 0.05; Supplemental Figs. 7S–10S). Almost all functional bias categories are similar to those in the vertebrate genomes; however, the invertebrate genes involved in transcription do not show a significant difference in the number of gene types from the overall average for any of the invertebrate gene groups, but do show a significant difference in the number of gene copies (see Supplemental Fig. 8S). Moreover, the number of invertebrate gene types involved in multicellular organismal development and anatomical structure morphogenesis is significantly under-represented in *Ionly*, but over-represented in *I.MCL*; as noted above, these functional categories are over-represented in *Vonly* (Supplemental Fig. 7S; Fig. 6). However, the number of invertebrate gene copies of these two functional categories is significantly over-represented in *Ionly* and under-represented in *I.MCL* (Supplemental Fig. 8S; also see Supplemental materials).

Faster evolution of *Vonly* genes

As described above, we found that ~8% of vertebrate genes have homologs in nonvertebrate genomes only when queried under less-stringent criteria than standard searches (the sum of the *V.A50* and *VFProk* genes) (Fig. 1; Tables 1, 2). This suggests that many vertebrate genes might be fast evolving, and therefore are not found to have homologs in invertebrates according to the standard family construction method. Comparing the protein distances of the human–mouse ortholog pairs, we find that although the distributions of the uncorrected protein distances (i.e., the proportion of mismatched residues) in the four vertebrate gene groups are all positively skewed, these distributions are significantly different from one another ($P << 0.05$, Kolmogorov-Smirnov [KS] test), except between *V.A50* and *VFProk* proteins and between *Vonly* and *VFProk* proteins ($P >> 0.05$, KS test). The proportions of the ortholog pairs with protein distances ≤0.25 are 86%, 67%, 37.5%, and 44% for the *V.MCL*, *V.A50*, *VFProk*, and *Vonly* proteins, respectively (Fig. 7). The distribution of the protein distances for the *Vonly* group has the largest mean and median (0.29 and 0.27), while that for the *V.MCL* group has the smallest values (0.15 and 0.13). The protein distance values for ortholog pairs in *Vonly* and *VFProk* (or in *V.A50*) are approximately two times (or ~1.4 times) larger than those for ortholog pairs in *V.MCL*. The difference in the protein distance distributions among the vertebrate homologous families suggests that, on average, the *V.MCL* proteins have evolved at the slowest rate, whereas the *Vonly* proteins have evolved at the fastest rate.

Discussion

Origins of vertebrate genes

In our analysis, the majority of vertebrate genes can be traced back to nonvertebrate genomes. Indeed, 76% of the vertebrate

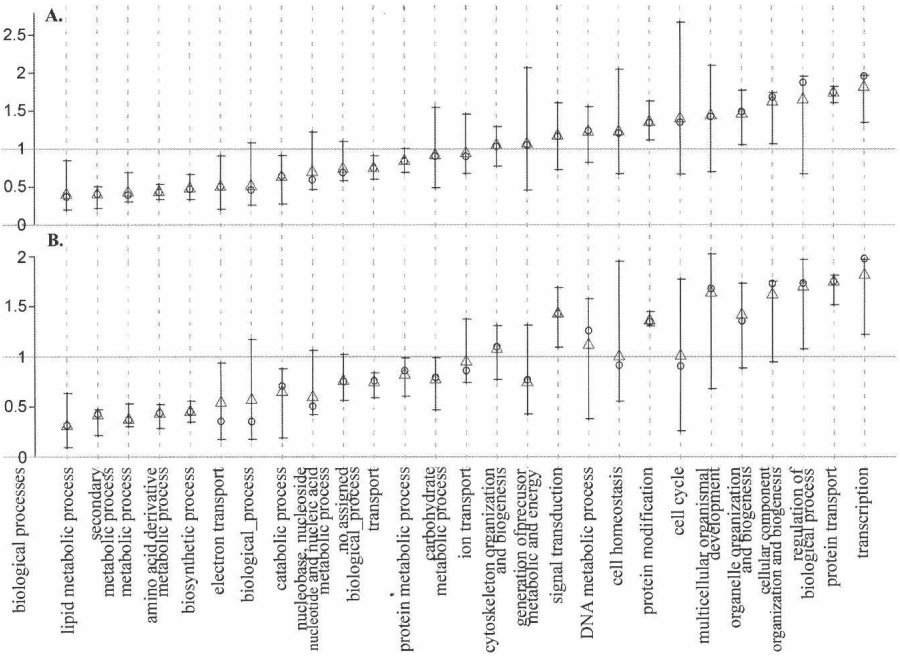

Figure 5. The slopes from a simple linear model (*A*) and a robust linear model (*B*) for the regression of the average vertebrate family size (n_v) against the average invertebrate family size (n_i) for each GOSlim biological process category in the *V.MCL* gene families. The error bar represents the 95% confidence interval of a slope from 1000 bootstrap replicates (see Methods). The slope from the original data and the mean of the bootstrapped slopes for each category are indicated by the circular and triangular points, respectively. The *X*-axis shows the biological process categories with the *P*-value < 0.01 for the null hypothesis of the estimated slope = 0 in either the simple linear model or the robust linear model, indicating a significant correlation between n_v and n_i. These categories are ordered by the mean of their bootstrapped slopes in *A*. Most of the proteins with GOSlim in the "biological_process" category are those involved in cell adhesion, response to stimulus, sensory perception of smell, immune response, homophilic cell adhesion, and defense response. Furthermore, most of these gene families are also assigned to other GOSlim categories (mainly to signal transduction, regulation of biological process, response to stress, or multicellular organismal development).

genes have homologs in the invertebrate genomes (i.e., the *V.MCL* and *V.A50* groups). In addition, ~1.5% of the vertebrate genes have homologs in the fungal or prokaryotic genomes (i.e., the *VFProk* group), though they do not have a homolog in the invertebrate genomes studied. For the genes in *VFProk*, the absence of a homolog in the invertebrate genomes studied may reflect the absence of a true homolog due to gene loss in invertebrates (probably because they are nonessential in these invertebrate genomes under study). Or, it may be due to faster sequence evolution in invertebrate lineages, so that homology has become too low to be recognized by the search methods used. Indeed, protein sequences in the *VFProk* group tend to have a higher amino acid substitution rate than those in the *V.MCL* group (Fig. 7). A third possibility is lateral gene transfer (LGT) from fungal or prokaryotic genomes to vertebrates, but this possibility may not be important because LGT from prokaryotes to the animal germ line is extremely rare. Moreover, some genes in either *V.A50* or *VFProk* groups might have experienced domain shuffling, leading to group misclassifications.

The remaining genes (i.e., the *Vonly* group) have no homolog found in the nonvertebrate genomes studied. Some of these genes might be artifacts of gene annotation. However, if a *Vonly* gene is an annotation artifact, it is likely to be "species specific." But, the majority (68%) of the *Vonly* gene copies (~64% of the *Vonly* gene types in an average vertebrate genome) (Tables 1, 2) can be found in at least two of the eight vertebrate genomes studied, and transposable element-like genes were excluded in our analysis. Thus, annotation artifacts probably contribute to only a small fraction of these genes. A second possible reason is sequence divergence. Indeed, among the four groups, proteins in *Vonly* have, on average, the highest amino acid substitution rate. Some of the *Vonly* proteins will likely be moved to the other groups when more nonvertebrate genomes are sequenced. Finally, some of the *Vonly* genes may have arisen de novo in vertebrates, though de novo gene creation is rather rare (Long et al. 2003) and new vertebrate genes have arisen mainly from retrotransposition of extant genes (Marques et al. 2005). So, there are still many uncertainties about the origins of *Vonly* genes.

Family size distribution

It was noted above that the *Vonly* genes are predominately singletons, whereas the genes in the other three gene groups mainly belong to gene families (Fig. 2). This suggests that the *Vonly* genes have a lower gene duplicability than do the non-*Vonly* genes, or that some *Vonly* genes are too young to have accumulated duplicates (the proportion of duplicates is 38% vs. 73%–76% for *Vonly* and non-*Vonly*, respectively).

The family size distribution differences among the four vertebrate gene groups may be in part produced by accelerated evolution right after duplication (Lynch and Conery 2000; Kondrashov et al. 2002) and asymmetric evolution between the two duplicate genes. Asymmetric evolution is observed for many young duplicates in human and fish genomes, where one copy has evolved fast, while the other shows a slow rate (Conant and Wagner 2003; Zhang et al. 2003; Steinke et al. 2006). Duplicate copies that have resulted from ancient gene-duplication events may now have sequences that are no longer recognizably similar, and thus are defined as single-copy genes in a genome. This substantial change in sequences may further lead to undetectable homology in nonvertebrate genomes (i.e., *Vonly* singletons). Some genes also lost sequence homology in other vertebrate genomes, as seen by a large proportion (78%) (Table 1) of the *Vonly* gene types found in only one of the eight vertebrate genomes. Almost all (>90%) of these gene types are singletons, making up approximately one-fifth of genes in some genomes (see Results). This pattern may also be in part due to incomplete genome assembly and inaccurate annotation, although these two factors may not be important, because the majority of these *Vonly* genes are found in more than one of the vertebrate genomes studied (see above). Some *Vonly* genes formed multigene families by subsequent duplications. The larger proportion of multigene-family genes in *VFProk* and *V.A50* than in *Vonly* might be in part due to a slower rate of *VFProk* and

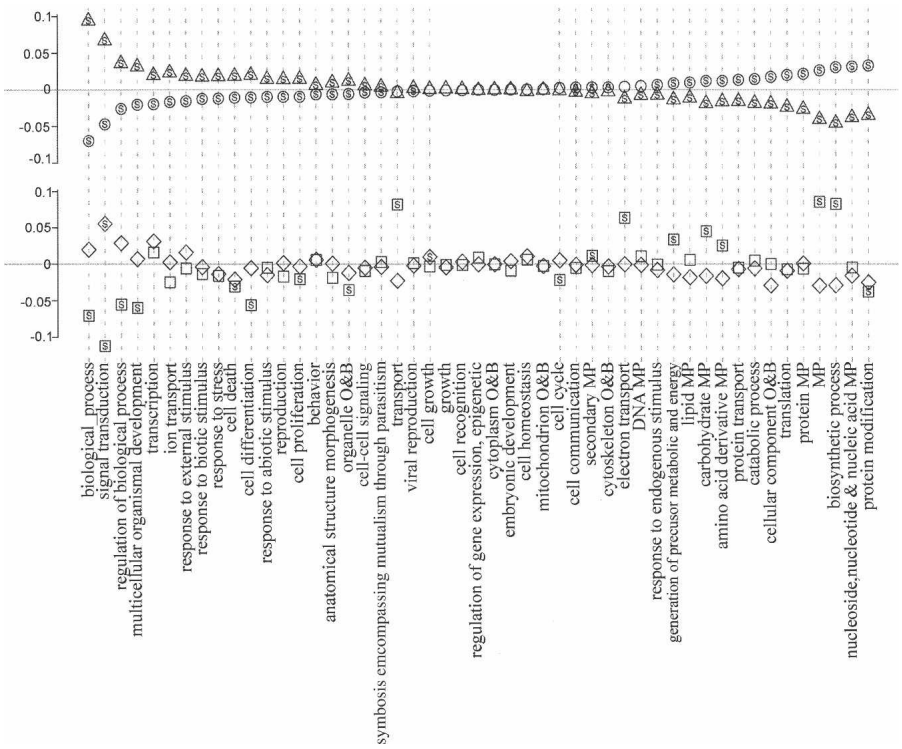

Figure 6. Functional bias in each GOSlim biological process category is shown for the *V.MCL* and *Vonly* groups (represented, respectively, by the circular and triangular points) at *top* and for the *V.A50* and *VFProk* gene groups (represented, respectively, by the diamond and square points) at *bottom*. The significant functional bias categories at the 5% false discovery rate are marked by S. The magnitude for the under- or over-represented families (below or above 0, respectively) from the overall average is indicated on the *Y*-axis. Most of the proteins with GOSlim in the "biological_process" category are those involved in cell adhesion, response to stimulus, sensory perception of smell, immune response, homophilic cell adhesion, and defense response. Furthermore, most of these gene families are also assigned to other GOSlim categories (mainly to signal transduction, regulation of biological process, response to stress, or multicellular organismal development). (MP) Metabolic process; (O & B) organization and biogenesis.

Gene family expansions and contractions

Contrary to the common view of a larger family size in vertebrates than in invertebrates, our linear regression analysis reveals that many gene families in the vertebrates are smaller than their invertebrate counterparts and that a large gene family in the vertebrates was often already a large gene family in the invertebrates (Fig. 3). However, the number of families above the $n_v = n_i$ line is larger than the number below the line, suggesting that a number of gene families have expanded in vertebrates. This observation is supported by: (1) a larger average family size in the vertebrates than in the invertebrates (the means of n_v and n_i are 2.21 and 1.77, respectively), and (2) the distribution of n_v/n_i is skewed toward a larger number of gene copies in the vertebrates than in the invertebrates, although the peak is at $n_v/n_i = 1$ (data not shown). These results are consistent with previous observations using raw counts of gene copies in a vertebrate and an invertebrate (e.g., Lander et al. 2001; Venter et al. 2001). As previously reported, many duplicates in extant vertebrate genomes have ancient origins in the common ancestor of tetrapods and fish (paleoploidy events) (e.g., Vandepoele et al. 2004; Dehal and Boore 2005; Panopoulou and Poustka 2005; Blomme et al. 2006). Thus, for the majority of genes, the family size is similar in vertebrate and invertebrate lineages, though some genes have a larger family size in vertebrates than in invertebrates. This observation implies that if one or two rounds of WGD had indeed occurred in the early vertebrates, then most of the duplicate genes had become lost.

V.A50's protein sequence evolution and more duplications of these genes than that of *Vonly*. Clearly, however, most members of the homologous families in *V.A50* and *VFProk* have greatly diverged from homologs in the non-vertebrate genomes, because under less-stringent homology search criteria, almost all of the families in these groups have a single vertebrate gene copy with a homolog in nonvertebrate genomes.

Similar to the *Vonly* genes, the majority of the *Ionly* genes are singletons, except for those in the sea urchin and the worm genomes (Supplemental Fig. 2S). The large proportion of the multigene-family *Ionly* genes in the sea urchin and the worm genomes have likely resulted from lineage-specific expansions in these two lineages (Supplemental Table 1S). Our results clearly suggest that, unlike the *Vonly* genes, a large proportion of the *Ionly* genes in each invertebrate genome are putatively species-specific genes with the largest proportion of the *Ionly* gene types from the worm genome (Supplemental Table 1S). Though the observed difference between the *Ionly* and *Vonly* genes is expected in view of the larger time depth of divergence among invertebrate genomes than that among vertebrate genomes, many of the *Ionly* genes may have lost (because they are nonessential) or become specialized in invertebrate genomes. Also, incompleteness and errors in genome annotation can contribute to this pattern.

Our result reveals that expansion of vertebrate genes has occurred more frequently for vertebrate genes that are homologous to invertebrate multigene families, whereas vertebrate genes that are homologous to invertebrate singletons tend to be singletons too (Fig. 4). Duplicability of some gene types may be under functional constraints, and these genes exist as singletons. Interestingly, a larger average vertebrate family size expansion is observed for the vertebrate genes that are present in a larger number of the invertebrate genomes studied (Fig. 4).

Functional bias is observed among the four vertebrate gene groups and is associated with family size differences between vertebrates and invertebrates (Figs. 5,6; Supplemental Figs. 3S–10S). Genes with a larger family size in the vertebrates than in the invertebrates (i.e., those functions with a slope >1) are involved in the organelle organization and biogenesis, cellular component organization and biogenesis, protein transport, protein modification, transcription, regulation of biological process, and signal transduction along with those localized to nucleus, chromosome, and protein complexes. On the other hand, those genes with a smaller family size in the vertebrates than in the inverte-

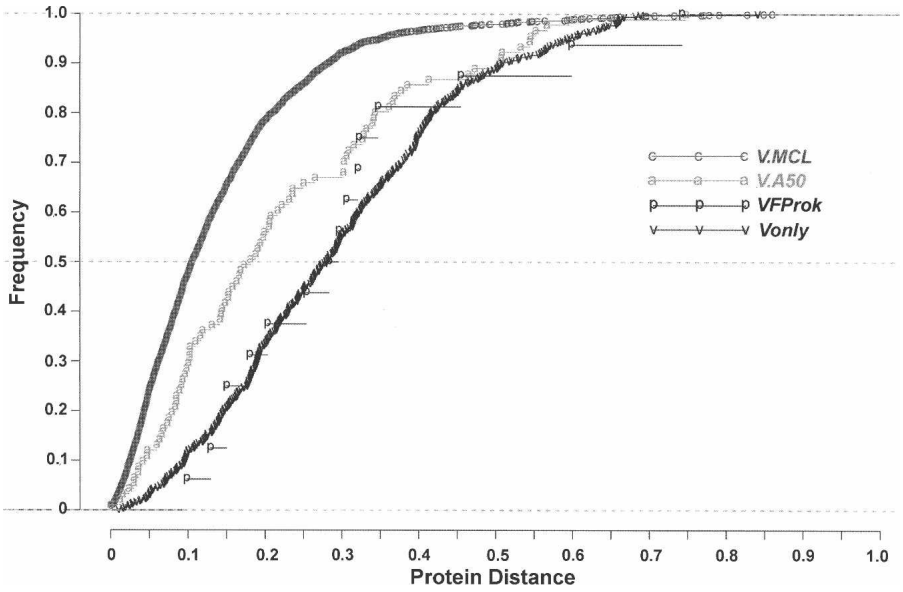

Figure 7. The cumulative distributions of the uncorrected protein distances from the human–mouse ortholog pairs in the *V.MCL*, *V.A50*, *VFProk*, and *Vonly* gene groups.

brates (with a slope <1) belong to the electron transport and metabolic processes, and they are largely localized to mitochondrion. These genes generally perform cellular housekeeping tasks. Clearly, these results imply strong functional bias of differences in gene duplicability between vertebrates and invertebrates and suggest that gene function is an important determinant of gene duplicability in these genomes. Furthermore, our finding suggests that different gene duplicabilities between vertebrates and invertebrates are better described by gene function than by organismal complexity (Yang et al. 2003).

Some functional categories that have a higher gene duplicability in vertebrates than in invertebrates (e.g., signal transduction and multicellular organismal development) are significantly over-represented in *Vonly* (but under-represented in *V.MCL*), while others are significantly under-represented in *Vonly* (but over-represented in *V.MCL*). On the other hand, the functional categories that have a lower gene duplicability in the vertebrates than in the invertebrates are over-represented in *V.MCL* (e.g., metabolic processes). Based on a scenario of accelerated evolution of duplicates, some of the genes in certain functional categories with a high gene duplicability in the vertebrates may have experienced rapid evolution, and thus, these sequences have become much diverged from the homologous sequences in invertebrate genomes. Such genes may be involved in transcription, regulation of biological process, signal transduction, response to various stimuli, or multicellular organismal development. Also, rapid protein sequence evolution of invertebrate genes in some corresponding functional categories (e.g., signal transduction or response to various stimuli) has likely occurred because such categories are over-represented in *Ionly* (see Results). Therefore, this process might have created gene families that are (functionally) specific to each taxon. A similar scenario was observed for the formation of lineage-specific gene families in mammalian genomes through rapid divergence of duplicates from previously existing families (Demuth et al. 2006). The over-representation of similar GO categories in both *Vonly* and *Ionly* could also be an indication that the failure to detect homology between these two groups was in part due to faster rates of sequence evolution in these GO categories than in others (such as metabolism). However, vertebrate genes involved in translation, protein modification, cellular component organization, and biogenesis processes may have protein sequences that evolve at a slow rate (possibly under functional constraints), and thus the majority of these genes can still be found in nonvertebrate genomes (i.e., the *V.MCL* genes). Furthermore, it is of interest to note that gene types for some of the functional categories that have a higher gene duplicability in vertebrates than in invertebrates are under-represented in *Ionly* and are over-represented in *I.MCL*, but the number of gene copies in these functional categories is over-represented in *Ionly* and under-represented in *I.MCL* (see Supplemental Figs. 7S, 8S; e.g., multicellular organismal development and transcription). This suggests that most of the invertebrate genes in these functional categories, relative to those of vertebrates, might have also expanded (relatively high number of gene copies in *Ionly*) but have not diverged much (relatively low number of gene types in *Ionly*). Based on our results, we favor the explanation of rapid divergence of duplicates of the families in these functional categories with a higher gene duplicability in vertebrates than in invertebrates. However, further analyses are needed to address this issue.

A higher gene duplicability in vertebrates than in invertebrates for these functional categories (e.g., transcription, transcription regulation, developmental control, and signal transduction pathways) may signify their contribution to an evolutionary leap in organismal or morphological complexity from invertebrates to vertebrates. For example, correlation between changes in the *Hox* gene numbers and the diversity in structures from the head to tail axis between amphioxus and large vertebrates is often cited as a potential example of this phenomenon (Holland and Garcia-Fernandez 1996). Recently, a significant retention of duplicate genes involved in transcriptional regulation was suggested to have contributed to the organismal complexity of vertebrates (Blomme et al. 2006). Further, the increase in the copy number of transcription factor genes was found to be much larger than that in all gene numbers in bacterial and eukaryotic genomes (van Nimwegen 2003). For the ray-finned fish lineage, many of the transcription factor and ATP-binding genes have expanded and evolved at particularly fast rates, where one paralog has undergone adaptive changes (Steinke et al. 2006). Moreover, positive selection has been found to act on genes in some functional categories. Such genes, for example, are involved in the immune defense system (Ig heavy chain 22, major histocompatibility complexes; Hughes and Nei 1988), external environmental sensing (e.g., olfactory receptors; Sharon et al. 1999), apoptosis (caspase proteins; Vallender and Lahn 2006), and eye lenses (ε-crystalline; Hughes 1994).

For the genes that are highly expanded in vertebrates (e.g., transcription, transcription regulation, signal transduction, and development), most have been retained from ancient duplica-

tions in the lineage leading to vertebrates (possibly from whole-genome duplication events; see Gu et al. 2002a; McLysaght et al. 2002; Blomme et al. 2006). This may suggest that most gene families in these gene functions became diversified and stabilized during the early evolution of vertebrates. Large expansions of gene families have often been interpreted as cases of positive (adaptive) selection, though an expansion could be the result of relaxed selection. Recently, Shiu et al. (2006) provided evidence that positive selection may have played a more important role in duplicate retention than the neutral process via duplication-degeneration-complementation (DDC) model (for the DDC model, see Force et al. 1999). Nonetheless, the retention of some vertebrate duplicates has been suggested as evidence for the DDC model (see Prince and Pickett 2002). While this issue awaits further investigation, adaptive evolution may have played an important role in determining different gene duplicabilities between vertebrates and invertebrates.

Methods

Genomic data and annotation

Vertebrates, invertebrates, fungi, and prokaryotes (Archaea and Eubacteria) have 12, 8, 11, and 436 completely sequenced genomes, respectively. Annotated genomes of 12 vertebrates (eight tetrapods: *Homo sapiens* [human], *Mus musculus* [mouse], *Canis familiaris* [dog], *Bos Taurus* [cattle], *Monodelphis domestica* [opossum], *Ornithorhynchus anatinus* [platypus], *Gallus gallus* [chicken], and *Xenopus tropicalis* [frog]; and four fishes: *Danio rerio* [zebrafish], *Gasterosteus aculeatus* [stickleback], *Takifugu rubripes* [fugu], *Tetraodon nigroviridis* [pufferfish]) and five invertebrates (*Caenorhabditis elegans* [worm], *Aedes aegypti* [yellow fever mosquito], *Anopheles gambiae* [African mosquito], *Drosophila melanogaster* [fruitfly], and *Ciona intestinalis* [ciona]) were downloaded from Ensembl v.42 (December 2006, http://www.ensembl.org/; Hubbard et al. 2007). The annotated sea urchin (*Strongylocentrotus purpuratus*; Spur_v2.1) and honey bee (*Apis mellifera*; Amel_4.0) genomes, as well as all Archeae and Eubacteria genomes, were obtained from GenBank (January 16, 2007; ftp://ftp.ncbi.nih.gov/genomes/). The 11 fungi genomes were downloaded from GenBank and their respective genome project websites (http://agd.unibas.ch; http://cbi.labri.fr/Genolevures/index.php; http://www.yeastgenome.org/). If a gene had splice variants, only the longest polypeptide was used. The chimpanzee (*Pan troglodytes*) and rat (*Rattus norvegicus*) genomes are not included in the analyses, because the chimpanzee sequence is not of high quality yet and is very close to that of human, and because the quality of the rat genome sequence is not as high as that of mouse. Since the number of annotated genes in a genome is dependent on the quality of the genome assembly, the other mammalian genomes that have a lower sequencing coverage *Loxodonta africana* (African savanna elephant) and *Oryctolagus cuniculus* (rabbit) are not included in our analyses. For simplicity, the results of our analysis of the four paleotetraploid fish genomes are not presented, but our conclusions do not change qualitatively when they are included.

Gene ontology (GO) (http://www.geneontology.org; Gene Ontology Consortium 2006) and InterPro annotations for the vertebrate genomes were downloaded from Ensembl. For genes without any GO annotation, the functional assignment was derived, when available, from InterPro2GO (downloaded on February 15, 2007). Because gene functions of two duplicates may rapidly diverge, potentially caused by as few as a single codon change, assigning functions to a gene family as a unit may not capture such rapid functional divergence of duplicates. To ensure that the functional categories assigned to a gene family reflect ancestral or common functions of the entire family, we used the broader functional classification (GOSlim; downloaded on February 15, 2007) of a gene family and required ≥25% of the genes within a family to be associated with that functional category. For some proteins, their GOSlim classification is mapped to the top most level (e.g., the biological_process category of the GO biological process aspect; see respective figure legends for detail).

Homology searches and gene family construction

An all-against-all BLAST search of 354,468 proteins (excluding mitochondrial genes) from the genomes of the 12 vertebrates (the eight tetrapods and the four fishes) and six invertebrates (ciona, sea urchin, fruitfly, African mosquito, yellow fever mosquito, and worm) was performed in which the query sequences were masked by CAST (Promponas et al. 2000). Hits with $E < 10^{-5}$ were retained for gene family clustering using the Markov Cluster (MCL) algorithm (van Dongen 2000; Enright et al. 2002) with Ensembl parameter settings. The inflation parameter (I), which determines the tradeoff between the number of families and the size of families, was set at 2.3, as commonly used in Ensembl; similar results were observed for a range of other values. Using known transposable element sequences (downloaded from Swiss-Prot Protein Knowledgebase, http://expasy.org/sprot/, on June 2, 2007) to BLAST against the protein database of vertebrate and invertebrate genomes at $E < 10^{-5}$, we excluded any protein that has a homology with a transposable element if the alignable region to transposable element is ≥50% of the length of the protein.

Vertebrate genes or families with no homolog in the six invertebrate genomes were subjected to further searches for homologs in nonvertebrate genomes (all eight invertebrate, 11 fungal, and 436 prokaryotic genomes were used) by less-stringent search criteria than above: homology is assumed without a limit on *E*-value from BLAST hits if any protein in such a family can be aligned via BLAST with a protein in the nonvertebrate genomes for ≥50% of the longer protein length and if an identity score of the two aligned proteins is at least as large as that defined by Gu et al. (2002b) (at least 30%). By the MCL and less-stringent homology searches, the vertebrate genes are divided into four gene groups: (1) Genes that can be found in one or more of the six invertebrate genomes by the MCL analysis (denoted as *V.MCL*); (2) Non-*V.MCL* genes that can be found in one or more of the eight invertebrates under the less-stringent homology search criteria (denoted as *V.A50*); (3) Genes that are in neither *V.MCL* nor *V.A50*, but can be found in one or more of the fungal or prokaryotic genomes (denoted as *VFProk*); and (4) Genes that are in none of the above three gene groups (denoted as *Vonly*), i.e., putative vertebrate-specific genes. Similarly, the genes in the six invertebrate genomes are divided into three groups: (1) Genes that have homologs in *V.MCL* or can be found in the four paleotetraploid fish genomes (denoted as *I.MCL*); (2) Non-*I.MCL* genes that have at least one member homologous to vertebrate *V.A50* genes (denoted as *I.A50*); and (3) Genes that are in neither *I.MCL* nor *I.A50* (denoted as *Ionly*).

Family size analysis

For a genome of interest, a gene type (a single-copy gene or a gene family) was classified into one of the following three groups by its size in the genome: (1) a singleton is a single-copy gene, (2) a two-gene family has two copies, or (3) a multigene family has ≥3 copies. Note that even a singleton is regarded as a gene family in this study.

For each *V.MCL* family, we calculated the average family size across the eight vertebrate genomes (n_v) and across the six invertebrate genomes (n_i). In the families with average family size among the six invertebrate genomes <1, the value of n_i was set to 1 in the regression analysis for the reason that at least one copy of a *V.MCL* gene must be present in the common ancestor of all vertebrates. That is, we used each n_i to estimate the number of genes in that gene family in the common ancestor of extant vertebrates. Therefore, a linear regression of n_v on n_i is equivalent to a regression of n_v on the gene number in the common ancestor of extant vertebrates. The slope of a linear regression model indicates whether a relationship exists between the family sizes of vertebrates and invertebrates and can also be used as a measure of how the family sizes in the extant vertebrate genomes change, on average, when compared with those of the invertebrates. We used both a simple linear model and a robust linear model (an M-estimator with the Huber weight method) where $n_v = \beta_0 + \beta_1 n_i$ in the R statistics package (version 2.1.1) (R Development Core Team 2006). The simple model assumes a normal distribution, while the robust model minimizes poor performance of the mean least square error in the simple linear model when outliers are present. We generated 1000 (for families grouped by GOSlim) or 10,000 (for all families) bootstrap replicates, where each replicate is generated by sampling $\{n_v^*, n_i^*\}$ pairs from $\{n_v, n_i\}$ with replacement (i.e., random-*x* resampling). Estimates of β_0 and β_1 for each bootstrap replicate (denoted as β_0^* and β_1^*, respectively) were used to obtain 95% confidence intervals (the percentile method) and *P*-values of these estimates. The *P*-values were obtained by increasing the $(1 - \alpha)$ confidence interval until the interval is just large enough to include the null hypotheses of $\beta_0 = a$ or $\beta_1 = a$, where a is 0 or 1. Thus, the *P*-values = α are obtained by

$$p^*_{\beta_0} = 1 - |2((1/B)\sum_{b=1}^{B} I(\beta^*_{0,b} \le a) - 1)|$$

$$\text{and } p^*_{\beta_1} = 1 - |2((1/B)\sum_{b=1}^{B} I(\beta^*_{1,b} \le a) - 1)|$$

for the null hypotheses of $\beta_0 = a$ and $\beta_1 = a$, respectively; where $I(\cdot)$denotes the indicator function and B is the number of bootstrap replicates.

Functional bias in vertebrate or invertebrate gene groups

For each vertebrate (or invertebrate) gene group (e.g., *V.MCL*, *V.A50*, *VFProk*, or *Vonly*), a bias in genes represented in a functional category of interest (namely function g) is defined by $r_g - R_g$, where r_g is the proportion of the gene types that are associated with function g in the specific vertebrate (or invertebrate) gene group, and R_g is the proportion of the all vertebrate (or invertebrate) gene types associated with function g. Similarly, the number of gene copies from all vertebrate (or all invertebrate) genomes is also considered; thus, r_g is the proportion of the gene copies that are associated with function g in the specific vertebrate (or invertebrate) gene group, and that R_g is the proportion of the all vertebrate (or invertebrate) gene copies associated with function g. The statistical significance of a functional bias is assessed by the χ^2 test with the false discovery rate (FDR) of 0.05. The FDR is obtained by the QVALUE software library (Storey and Tibshirani 2003) in R.

Distance between two orthologous proteins

For gene families with homologs in both human and mouse (but no more than 20 homologs in either genome, a condition set to avoid alignment difficulties), the human–mouse orthologs were retrieved from Ensembl Compara (v. 42) and aligned with MUSCLE (Edgar 2004). Pairwise uncorrected protein distances were obtained by the distmat program in the EMBOSS package (Rice et al. 2000).

Acknowledgments

We thank Henry Lu for advice in statistical analyses; A. Meyer, J. Rest, Y. Van de Peer, K. Vandepoele, G. Amoutzias, and two anonymous reviewers for valuable comments on the manuscript; S. van Dongen for suggestions on the MCL algorithm; and the Ensembl Helpdesk for help in genomic data. This study was supported by NIH and Balzan grants to W.H.L.

References

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer, Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* **7:** R43. doi: 10.1186/gb-2006-7-5-r43.

Conant, G.C. and Wagner, A. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Res.* **13:** 2052–2058.

Dehal, P. and Boore, J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3:** e314. doi: 10.1371/journal.pbio.0030314.

Demuth, J.P., De Bie, T., Stajich, J.E., Cristianini, N., and Hahn, M.W. 2006. The evolution of Mammalian gene families. *PLoS ONE* **1:** e85. doi: 10.1371/journal.pone.0000085.

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792–1797.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30:** 1575–1584.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.

Gene Ontology Consortium. 2006. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.* **34:** D322–D326.

Gu, X., Wang, Y., and Gu, J. 2002a. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat. Genet.* **31:** 205–209.

Gu, Z., Cavalcanti, A., Chen, F.C., Bouman, P., and Li, W.H. 2002b. Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol. Biol. Evol.* **19:** 256–262.

Holland, P.W. and Garcia-Fernandez, J. 1996. Hox genes and chordate evolution. *Dev. Biol.* **173:** 382–395.

Holland, P.W., Garcia-Fernandez, J., Williams, N.A., and Sidow, A. 1994. Gene duplications and the origins of vertebrate development. *Dev. Suppl.* **1994:** 125–133.

Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., et al. 2007. Ensembl 2007. *Nucleic Acids Res.* **35:** D610–D617.

Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc. Biol. Sci.* **256:** 119–124.

Hughes, A.L. and Nei, M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335:** 167–170.

Kasahara, M., Hayashi, M., Tanaka, K., Inoko, H., Sugaya, K., Ikemura, T., and Ishibashi, T. 1996. Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. *Proc. Natl. Acad. Sci.* **93:** 9096–9101.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. 2002. Selection in the evolution of gene duplications. *Genome Biol.* **3:** research0008.1–research0008.9. doi: 10.1186/gb-2002-3-2-research0008.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Long, M., Betran, E., Thornton, K., and Wang, W. 2003. The origin of new genes: Glimpses from the young and old. *Nat. Rev. Genet.* **4:** 865–875.

Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.

Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A., and Kaessmann, H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.* **3:** e357. doi: 10.1371/journal.pbio.0030357.

McLysaght, A., Hokamp, K., and Wolfe, K.H. 2002. Extensive genomic duplication during early chordate evolution. *Nat. Genet.* **31:** 200–204.
Meyer, A. and Van de Peer, Y. 2005. From 2R to 3R: Evidence for a fish-specific genome duplication (FSGD). *Bioessays* **27:** 937–945.
Nakatani, Y., Takeda, H., Kohara, Y., and Morishita, S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* **17:** 1254–1265.
Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin, Germany.
Panopoulou, G. and Poustka, A.J. 2005. Timing and mechanism of ancient vertebrate genome duplications–the adventure of a hypothesis. *Trends Genet.* **21:** 559–567.
Prince, V.E. and Pickett, F.B. 2002. Splitting pairs: The diverging fates of duplicated genes. *Nat. Rev. Genet.* **3:** 827–837.
Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Sander, C., and Ouzounis, C.A. 2000. CAST: An iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics* **16:** 915–922.
R Development Core Team. 2006. An Introduction to R. http://www.r-project.org/.
Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16:** 276–277.
Sharon, D., Glusman, G., Pilpel, Y., Khen, M., Gruetzner, F., Haaf, T., and Lancet, D. 1999. Primate evolution of an olfactory receptor cluster: Diversification by gene conversion and recent emergence of pseudogenes. *Genomics* **61:** 24–36.
Shiu, S.H., Byrnes, J.K., Pan, R., Zhang, P., and Li, W.H. 2006. Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proc. Natl. Acad. Sci.* **103:** 2232–2236.
Steinke, D., Salzburger, W., Braasch, I., and Meyer, A. 2006. Many genes in fish have species-specific asymmetric rates of molecular evolution. *BMC Genomics* **7:** 20. doi: 10.1186/1471-2164-7-20.
Storey, J.D. and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100:** 9440–9445.
Vallender, E.J. and Lahn, B.T. 2006. A primate-specific acceleration in the evolution of the caspase-dependent apoptosis pathway. *Hum. Mol. Genet.* **15:** 3034–3040.
Vandepoele, K., De Vos, W., Taylor, J.S., Meyer, A., and Van de Peer, Y. 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.* **101:** 1638–1643.
van Dongen, S. 2000. "Graph clustering by flow simulation." Ph.D. thesis, University of Utrecht, The Netherlands.
van Nimwegen, E. 2003. Scaling laws in the functional content of genomes. *Trends Genet.* **19:** 479–484.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Yang, J., Lusk, R., and Li, W.H. 2003. Organismal complexity, protein complexity, and gene duplicability. *Proc. Natl. Acad. Sci.* **100:** 15661–15665.
Zhang, P., Gu, Z., and Li, W.H. 2003. Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol.* **4:** R56. doi: 10.1186/gb-2003-4-9-r56.

Received August 18, 2007; accepted in revised form November 14, 2007.