Cell Reports
Article

On the Expansion of "Dangerous" Gene Repertoires by Whole-Genome Duplications in Early Vertebrates

Param Priya Singh,[1,3] Séverine Affeldt,[1,3] Ilaria Cascone,[2] Rasim Selimoglu,[2] Jacques Camonis,[2] and Hervé Isambert[1,*]
[1]CNRS UMR168
[2]INSERM U830
UPMC, Institut Curie, Research Center, 26, rue d'Ulm, 75248 Paris, France
[3]These authors contributed equally to this work
*Correspondence: herve.isambert@curie.fr
http://dx.doi.org/10.1016/j.celrep.2012.09.034

SUMMARY

The emergence and evolutionary expansion of gene families implicated in cancers and other severe genetic diseases is an evolutionary oddity from a natural selection perspective. Here, we show that gene families prone to deleterious mutations in the human genome have been preferentially expanded by the retention of "ohnolog" genes from two rounds of whole-genome duplication (WGD) dating back from the onset of jawed vertebrates. We further demonstrate that the retention of many ohnologs suspected to be dosage balanced is in fact indirectly mediated by their susceptibility to deleterious mutations. This enhanced retention of "dangerous" ohnologs, defined as prone to autosomal-dominant deleterious mutations, is shown to be a consequence of WGD-induced speciation and the ensuing purifying selection in post-WGD species. These findings highlight the importance of WGD-induced nonadaptive selection for the emergence of vertebrate complexity, while rationalizing, from an evolutionary perspective, the expansion of gene families frequently implicated in genetic disorders and cancers.

INTRODUCTION

Just as some genes happen to be more "essential," owing to their deleterious loss-of-function or null mutations, some genes can be classified as more "dangerous," due to their propensity to acquire deleterious gain-of-function mutations. This is, in particular, the case for oncogenes and genes with autoinhibitory protein folds, whose mutations typically lead to constitutively active mutants with dominant deleterious phenotypes (Pufall and Graves, 2002).

Dominant deleterious mutations, that are lethal or drastically reduce fitness over the lifespan of organisms, must have also impacted their long term evolution on timescales relevant for genome evolution (e.g., >10–100 million years [MY]). In fact, dominant disease genes in humans have been shown to be under strong purifying selection (Furney et al., 2006; Blekhman et al., 2008; Cai et al., 2009). Yet, "dangerous" gene families implicated in cancer and severe genetic diseases have also been greatly expanded by duplication in the course of vertebrate evolution. For example, the single orthologous locus, *Ras85D* in flies and *Let-60* in nematodes, has been duplicated into three *RAS* loci in typical vertebrates, *KRAS*, *HRAS*, and *NRAS*, that present permanently activating mutations in 20%–25% of all human tumors, even though *HRAS* and *NRAS* have also been shown to be dispensable for mouse growth and development (Ise et al., 2000; Esteban et al., 2001).

While the maintenance of essential genes is ensured by their lethal null mutations, the expansion of dangerous gene families remains an evolutionary puzzle from a natural selection perspective. Indeed, considering that many vertebrate disease genes are phylogenetically ancient (Domazet-Loso and Tautz, 2008; Cai et al., 2009; Dickerson and Robertson, 2012), and that their orthologs also cause severe genetic disorders in extant invertebrates (Berry et al., 1997; Ciocan et al., 2006; Robert, 2010), it is surprising that dangerous gene families have been duplicated more than other vertebrate genes without known dominant deleterious mutations. While gene duplicates can confer mutational robustness against loss-of-function mutations, multiple copies of genes prone to gain-of-function mutations are expected to lead to an overall aggravation of a species' susceptibility to genetic diseases and thus be opposed by purifying selection.

Two alternative hypotheses can be put forward to account for the surprising expansion of dangerous gene families. Either, the propensity of certain genes to acquire dominant deleterious mutations could be a mere by-product of their presumed advantageous functions. In that case, only the overall benefit of gene family expansion should matter, irrespective of the mechanism of gene duplication. Alternatively, gene susceptibility to dominant deleterious mutations could have played a driving role in the striking expansion of dangerous gene families. But what could have been the selection mechanism?

In this article, we report converging evidences supporting the latter hypothesis and propose a simple evolutionary model to explain the expansion of such dangerous gene families. It is based on the observation that the majority of human genes prone to dominant deleterious mutations, such as oncogenes and genes with autoinhibitory protein folds, have not been duplicated through small scale duplication (SSD). Instead, the expansion of these dangerous gene families can be traced back to two rounds of whole-genome duplication (WGD), that occurred at the

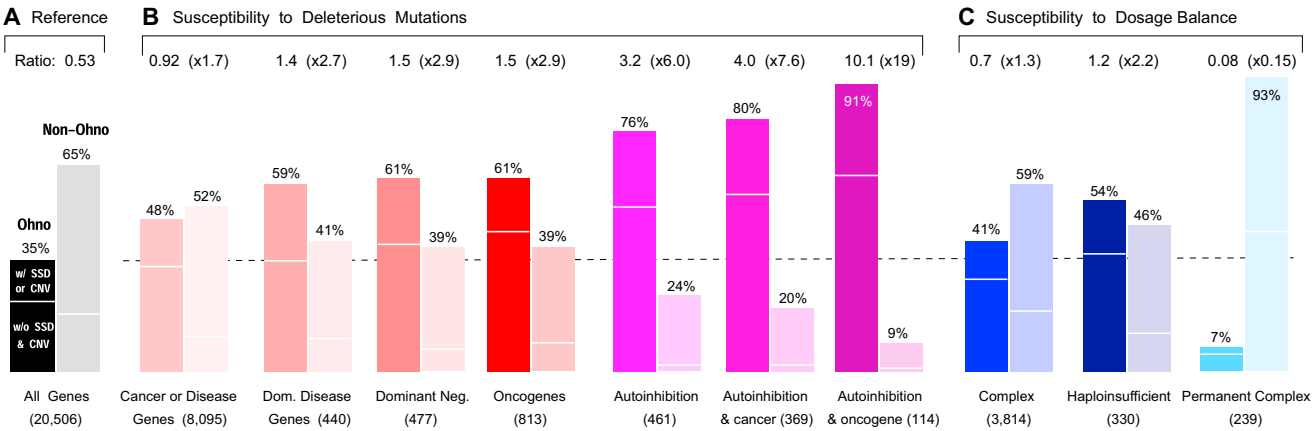

Figure 1. Prevalence of Retained Ohnologs in the Human Genome within Different Gene Classes

(A and B) Prevalence of retained ohnologs either "w/ SSD or CNV" or "w/o SSD & CNV" for all 20,506 human protein-coding genes (A), and gene classes susceptible to deleterious mutations (B). Note that gene classes with higher susceptibility to deleterious mutations retained more ohnologs.

(C) Ohnolog retention in gene classes susceptible to dosage balance constraints. Fold changes in ohnolog/nonohnolog ratios are given relative to the reference from all human genes in (A).

See also Figure S1.

onset of jawed vertebrates, some 500 MY ago (Ohno, 1970; Putnam et al., 2008).

These two rounds of WGD in the early vertebrate lineage are frequently credited with creating the conditions for the evolution of vertebrate complexity. Indeed, WGD-duplicated genes, so-called "ohnologs" in honor of Susumu Ohno (Ohno, 1970; Wolfe, 2000), have been preferentially retained in specific gene classes associated with organismal complexity, such as signal transduction pathways, transcription networks, and developmental genes (Maere et al., 2005; Blomme et al., 2006; Freeling and Thomas, 2006; Sémon and Wolfe, 2007; Makino and McLysaght, 2010; Huminiecki and Heldin, 2010). By contrast, gene duplicates coming from SSD are strongly biased toward different functional categories, such as antigen processing, immune response, and metabolism (Huminiecki and Heldin, 2010). SSD paralogs and WGD ohnologs also differ in their gene expression and protein network properties (Hakes et al., 2007; Guan et al., 2007). Furthermore, recent genome-wide analysis have shown that ohnologs in the human genome have experienced fewer SSD than "nonohnolog" genes and tend to be refractory to copy number variation (CNV) caused by polymorphism of small segmental duplications in human populations (Makino and McLysaght, 2010). These antagonist retention patterns of WGD and SSD/CNV gene duplicates in the human genome have been suggested to result from dosage balance constraints (Makino and McLysaght, 2010) on the relative expressions of multiple protein partners (Veitia, 2002), as proposed earlier for other organisms like yeast (Papp et al., 2003) and the paramecium (Aury et al., 2006).

In this article, we investigate the evolutionary causes responsible for the expansion of gene families prone to deleterious mutations in vertebrates and propose a simple evolutionary model accounting for their antagonistic retention pattern after WGD and SSD events. The retention of ohnologs in the human genome is shown to be more strongly associated with their susceptibility to deleterious mutations, than their functional importance or "essentiality." We also demonstrate using a causal inference analysis, that the retention of many ohnologs suspected to be dosage balanced is in fact an indirect effect of their higher susceptibility to deleterious mutations. We argue that the enhanced retention of dangerous ohnologs is a somewhat counterintuitive yet simple consequence of the speciation event triggered by WGD and the ensuing purifying selection in post-WGD species.

These findings rationalize, from an evolutionary perspective, the WGD expansion of gene families frequently implicated in genetic disorders, such as cancer, and highlight the importance of nonadaptive selection on the emergence of vertebrate complexity.

RESULTS

Genes Prone to Deleterious Mutations Retain More Ohnologs

We first analyzed a possible association between the susceptibility of human genes to deleterious mutations and their retention of ohnologs, as proposed in Gibson and Spring (1998) for multidomain proteins. To this end, we considered multiple classes of genes susceptible to deleterious mutations from experimentally verified databases and literature. These classes include cancer genes (from multiple sources including COSMIC [Forbes et al., 2011] and CancerGenes [Higgins et al., 2007]), genes mutated in other genetic disorders, dominant negative genes from OMIM, and genes with autoinhibitory protein folds (Experimental Procedures). We looked at the relative contributions of WGD and SSD in the expansion of these "dangerous" gene classes.

The results, depicted in Figures 1 and S1, demonstrate indeed a strong association between the retention of human ohnologs from vertebrate WGD and their reported susceptibility to deleterious mutations, as compared to nonohnologs, whereas an

opposite pattern is found for SSD/CNV gene duplicates. Overall, the 8,095 human genes associated with the occurrence of cancer and other genetic diseases have retained significantly more ohnologs than expected by chance, 48% versus 35% (48%; 3,844/8,095; $p = 1.3 \times 10^{-129}$, χ^2 test). Furthermore, these associations, which do not take into account the actual severity of the gene mutations, are clearly enhanced when the analysis is restricted to genes with direct experimental evidence of dominant deleterious mutations, such as dominant disease genes (59%; 261/440; $p = 1.7 \times 10^{-27}$, χ^2 test), dominant negative mutants (61%; 292/477; $p = 3.9 \times 10^{-34}$, χ^2 test), oncogenes (61%; 493/813; $p = 1.4 \times 10^{-54}$, χ^2 test), or genes exhibiting autoinhibitory constraints (76%; 350/461; $p = 2.7 \times 10^{-77}$, χ^2 test). The biased retention of ohnologs is even stronger for genes combining several factors associated with an enhanced susceptibility to deleterious mutations, such as cancer genes with autoinhibitory folds, (80%; 294/369; $p = 1.0 \times 10^{-73}$, χ^2 test), or oncogenes with autoinhibitory folds, (91%; 104/114; $p = 6.9 \times 10^{-37}$, χ^2 test).

This retention of dangerous ohnologs is illustrated on Table 1 that presents an up-to-date list of 76 hand-curated gene families of up to four ohnologs, exhibiting both autoinhibitory folds and oncogenic properties (see Table S1 for oncogenic and autoinhibitory details and references). These dangerous ohnologs are typically found along signal transduction cascades, from receptor tyrosine kinases and cytoplasmic or nuclear kinases to guanine exchange factors (GEF), GTPase activating proteins (GAP), and transcription factors (Table 1, gene classes A–E). In addition, autoinhibited oncogenes are also found in other ohnolog families with diverse functions (Table 1, gene class F). By contrast, we obtained a hand-curated list of only ten nonohnolog genes exhibiting both autoinhibitory and oncogenic properties, Table 1, gene class G (see Table S2 for oncogenic and autoinhibitory details and references). Interestingly, half of them (4/10) can be traced back to SSD events, which occurred after or at the same period of the two WGD in early vertebrate lineages (Table S2). All in all, this implies that >90% of known oncogenes with autoinhibitory folds have retained at least one ohnolog pair in the human genome (as well as, possibly, a few additional duplicates from more recent SSD events).

Ohnologs Are Conserved but More "Dangerous" than "Essential"

We then investigated whether the susceptibility of ohnologs to deleterious mutations could be directly quantified through comparative sequence analysis. We used Ka/Ks ratio estimates, which measure the proportion of nonsynonymous substitutions (Ka) to the proportion of synonymous substitutions (Ks) (Extended Results and Table S3). Ohnologs exhibit statistically lower Ka/Ks ratios, Figures 2, S2, and S3, which provides direct evidence of strong conservation, consistent with a higher susceptibility of ohnologs to deleterious mutations. Similar trends have also been reported for ohnologs specific to teleost fishes (Brunet et al., 2006) or to the more recent WGD in *Xenopus laevis* lineage (Sémon and Wolfe, 2008). Note, however, that the functional consequences of such deleterious mutations, leading either to a gain or a loss of function, cannot be directly inferred from Ka/Ks distributions. Yet, as outlined below, we found marked differences in the retention of "dangerous" ohnologs prone to dominant gain-of-function mutations and "essential" ohnologs exhibiting lethal loss-of-function or null mutations.

While autosomal-dominant disease genes exhibit a strong ohnolog retention bias (Figure 1B), 59% versus 35% (59%; 261/440; $p = 1.7 \times 10^{-27}$, χ^2 test), autosomal-recessive disease genes are not significantly enriched in ohnologs 37% versus 35% (37%; 221/598; $p = 0.24$, χ^2 test). Similarly, human orthologs of mouse genes, reported as being "essential" genes from large-scale null mutant studies in mouse, are not strongly enriched in ohnologs 56% versus 54% (56%; 1,537/2,729; $p = 3.8 \times 10^{-3}$, χ^2 test), where 54% = 3,190/5,956 is the global proportion of ohnologs among the 5,956 genes tested for null mutation in mouse (Experimental Procedures). In fact, this small enrichment becomes even nonsignificant once genes with dominant allelic mutants are removed from the list of 5,956 genes tested for essentiality in mouse, i.e., 50% versus 48% (50%; 760/1,525; $p = 0.09$, χ^2 test), where 48% = 1,782/3,739 is the global proportion of ohnologs among the 3,739 genes tested for essentiality in mouse, after removing dominant disease genes, oncogenes, and genes with dominant negative mutations or autoinhibitory folds.

All in all, this shows that the retention of ohnologs has been most enhanced for genes prone to autosomal-dominant deleterious mutations and not autosomal-recessive deleterious mutations. This suggests that the retention of ohnologs is more strongly related to their "dangerousness," as defined by their high susceptibility to dominant deleterious mutations, than their functional importance or "essentiality," as identified through large-scale null mutation studies in mouse.

Ultimately, we will argue that the "dangerousness" of ohnologs effectively controls their individual retention in the genomes of post-WGD species, as will be shown below in the section Model for the Retention of Dangerous Ohnologs.

Mixed Susceptibility of Human Ohnologs to Dosage Balance

An alternative hypothesis, focusing instead on the collective retention of interacting ohnologs, has been frequently invoked to account for the biased retention of ohnologs in unicellular organisms like yeast (Papp et al., 2003) or the paramecium (Aury et al., 2006) and in higher eukaryotes (Birchler et al., 2001; Makino and McLysaght, 2010).

This "dosage balance" hypothesis posits that interacting protein partners tend to maintain balanced expression levels in the course of evolution, in particular, for protein subunits of conserved complexes (Birchler et al., 2001; Veitia, 2002; Papp et al., 2003; Veitia, 2010; Makino and McLysaght, 2010). Thus, SSD of dosage balanced genes are thought to be generally detrimental through the dosage imbalance they induce, thereby raising the odds for their rapid nonfunctionalization (Papp et al., 2003; Maere et al., 2005). By contrast, rapid nonfunctionalization of ohnologs after WGD has been suggested to be opposed by dosage effect, in particular, for highly expressed genes and genes involved in protein complexes or metabolic pathways (Aury et al., 2006; Evlampiev and Isambert, 2007; Gout et al., 2010; Makino and McLysaght, 2010). This is because WGD initially preserves correct relative dosage between

Table 1. Ohnolog Families with Both Autoinhibitory and Oncogenic Properties

A. Ohnolog Receptor Tyrosine Kinases and Other Receptor Kinases							
ALK	*LTK*			*KIT*	*CSF1R*	*FLT3*	
EGFR	*ERBB2*	*ERBB3*	*ERBB4*	*MET*	*MST1R*		
FGFR1	*FGFR2*	*FGFR3*	*FGFR4*	*NPRA*	*NPRB*		
IGF1R	*INSR*	*INSRR*		*PDGFRA*	*PDGFRB*		
B. Ohnolog Cytoplasmic and Nuclear Protein Kinases							
ABL1	*ABL2*			*PKN1*	*PKN2*	*PKN3*	
ARAF	*BRAF*	*RAF1*		*PRKAA1*	*PRKAA2*		
AKT1	*AKT2*	*AKT3*		*PRKCA*	*PRKCB*	*PRKCG*	
CAMK1	*CAMK1D*	*CAMK1G*	*PNCK*	*PRKCE*	*PRKCH*		
CAMKK1	*CAMKK2*			*PRKCI*	*PRKCZ*		
CSNK1D	*CSNK1E*			*PRKD1*	*PRKD2*	*PRKD3*	
GSK3A	*GSK3B*			*PRKG1*	*PRKG2*		
GRK4	*GRK5*	*GRK6*		*PTK2*	*PTK2B*		
JAK1	*JAK2*	*JAK3*	*TYK2*	*RSK1*	*RSK2*	*RSK3*	*RSK4*
SRC	*FGR*	*FYN*	*YES1*	*MSK1*	*MSK2*		
HCK	*LCK*	*BLK*	*LYN*	*NDR1*	*NDR2*		
MKNK1	*MKNK2*			*SYK*	*ZAP70*		
NEK6	*NEK7*						
C. Ohnolog GEF							
ARHGEF3	*NET1*			*RALGDS*	*RGL1*	*RGL2*	*RGL3*
ARHGEF6	*COOL1*			*SOS1*	*SOS2*		
DBL	*DBS*	*MCF2L2*		*TIAM1*	*TIAM2*		
FGD1	*FGD2*	*FGD3*	*FGD4*	*TIM*	*WGEF*	*SGEF*	*NGEF*
PDZ-RHOGEF	*LSC*	*LARG*		*VAV1*	*VAV2*	*VAV3*	
P114-RHOGEF	*GEF-H1*						
D. Ohnolog GAP							
ASAP1	*ASAP2*	*ASAP3*		*PLXNA1*	*PLXNA2*	*PLXNA3*	*PLXNA4*
IQGAP1	*IQGAP2*	*IQGAP3*		*PLXNB1*	*PLXNB2*	*PLXNB3*	*PLXND1*
E. Ohnolog DNA Binding and Transcription Factors							
CEBPA	*CEBPB*	*CEBPE*		*IRF4*	*IRF8*	*IRF9*	
CUX1	*CUX2*			*MEIS1*	*MEIS2*	*MEIS3*	
ELK1	*ELK3*	*ELK4*		*p53*	*p63*	*p73*	
ETS1	*ETS2*			*RUNX1*	*RUNX2*	*RUNX3*	
ETV1	*ETV4*	*ETV5*		*SOX1*	*SOX2*	*SOX3*	
ETV6	*ETV7*						
F. Other Ohnolog Genes with Both Autoinhibitory and Oncogenic Properties							
ANP32A	*ANP32B*	*ANP32E*		*nNOS*	*eNOS*		
ATP2B1	*ATP2B2*	*ATP2B3*	*ATP2B4*	*NOTCH1*	*NOTCH2*	*NOTCH3*	
cIAP1|2	*XIAP*			*PLCB1*	*PLCB2*	*PLCB3*	
CCNT1	*CCNT2*			*PLCD1*	*PLCD3*	*PLCD4*	
FLNA	*FLNB*	*FLNC*		*PLCG1*	*PLCG2*		
FURIN	*PCSK4*			*PTPN1*	*PTPN2*		
KPNA2	*KPNA7*			*SMURF1*	*SMURF2*		
NEDD4	*NEDD4L*			*TRPV1|3*	*TRPV2*	*TRPV4*	*TRPV5|6*
NOXA1	*NOXA2*						
G. Nonohnolog Genes with Both Autoinhibitory and Oncogenic Properties							
CAMK4	*ELF3*	*MELK*	*MOS*	*PDPK1*	*BRK*	*PTPN11*	*RET*	*RPS6KB1*	*TTN*

GEF, guanine exchange factors; GAP, GTPase activating proteins.
See also Tables S1 and S2.

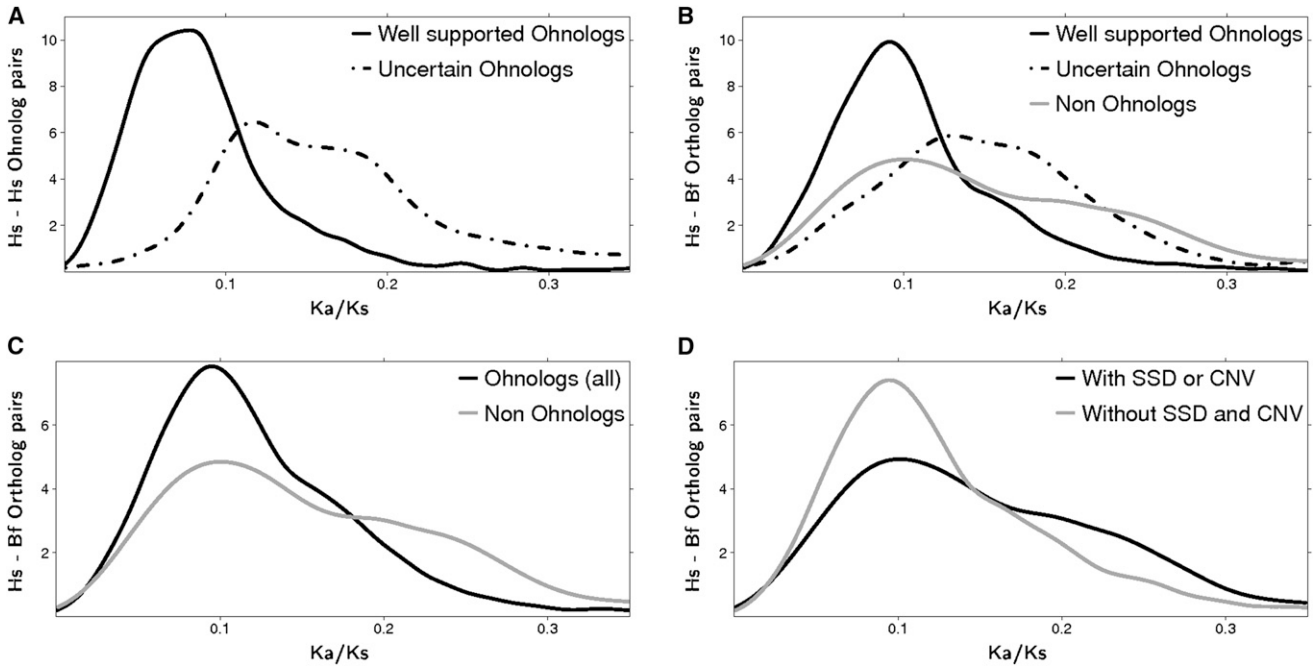

Figure 2. Ka/Ks Distributions for WGD and SSD or CNV Duplicates in the Human Genome

(A–D) Ka/Ks distributions for human-human (Hs-Hs) ohnolog pairs (A) and human-amphioxus (Hs-Bf) ortholog pairs (B) with different confidence status (see Extended Results). Ka/Ks distributions for human-amphioxus (Hs-Bf) ortholog pairs involving a human ohnolog (C) and for human-amphioxus (Hs-Bf) ortholog pairs exhibiting either SSD or CNV (D).

See also the Extended Results, Figures S2 and S3, and Table S3 for statistical significance and comparison with other invertebrate outgroups.

expressed genes, whereas subsequent random nonfunctionalization of individual ohnologs disrupts this initial dosage balance. For instance, yeast *Saccharomyces cerevisiae* has retained 76% of its ribosomal gene ohnologs from a 150 MY old WGD (Kellis et al., 2004; Lin et al., 2007), although the maintenance of these ohnologs has been suggested to require frequent gene conversion events (Kellis et al., 2004; Evangelisti and Conant, 2010) as well as fine-tuned dosage compensation to ensure a balanced expression with the remaining 24% ribosomal genes having lost their ohnologs (Zeevi et al., 2011).

Following on this dosage balance hypothesis, we performed statistical analysis on multiprotein complexes from HPRD (Keshava Prasad et al., 2009) and CORUM (Ruepp et al., 2010) databases and a hand-curated list of permanent complexes (Zanivan et al., 2007) (Experimental Procedures) to investigate for a possible association between the retention of human ohnologs and their susceptibility to dosage balance constraints.

The results depicted in Figure 1C demonstrate, in agreement with (Makino and McLysaght, 2010), that genes implicated in multiprotein complexes have retained significantly more ohnologs than expected by chance, 41% versus 35% (41%; 1,567/3,814; $p = 8.7 \times 10^{-17}$, χ^2 test). This trend is also enhanced when focusing on haploinsufficient genes, that are known for their actual sensitivity to dosage balance constraints (Qian and Zhang, 2008) (54%; 179/330; $p = 8.0 \times 10^{-14}$, χ^2 test).

Yet, surprisingly, an opposite trend corresponding to the elimination of ohnologs is observed for genes implicated in permanent complexes, that are presumably strongly sensitive to dosage balance constraints (7.5%; 18/239; $p = 1.2 \times 10^{-18}$, χ^2 test) (Figure 1C). In fact, looking more closely at the few human ohnologs, that have not been eliminated from permanent complexes (Table 2), we found that they are likely under less stringent dosage balance constraints than most proteins in permanent complexes, as they typically coassociate with mitochondrial proteins or form large multimeric subcomplexes with intrinsic stoichiometry disequilibrium.

This suggests that the elimination of most ohnologs from permanent complexes is, in fact, strongly favored under dosage imbalance and becomes likely inevitable once a few of those ohnologs have been accidentally lost following WGD. Indeed, the uneven elimination of ohnologs in permanent complexes is expected to lead to the assembly of nonfunctional, partially formed complexes detrimental to the cell, unless dosage compensation mechanisms effectively re-establish proper dosage balance at the level of gene regulation (Birchler et al., 2001), as for yeast ribosomal proteins (Zeevi et al., 2011). By contrast, transient complexes, which are typically more modular than permanent complexes, are expected to accommodate such dosage changes more easily, as they do not usually require the same strict balance in the expression levels of their protein partners.

These findings on the differences in retention of human ohnologs between permanent and more transient complexes suggest the relevance of different underlying causes. Although dosage balance presumably remains the primary evolutionary constraint in permanent complexes (<2% of human genes), which lead to the elimination of ohnologs in permanent complexes in

Table 2. Low Retention of Ohnologs in Permanent Complexes

Permanent Complexes[a]	Number of Ohnologs	Intrinsic Stoichiometry Disequilibrium of Ohnologs in Permanent Complexes
ATP F0	3/12	the 3 ohnologs ATP5G1-3 form the 10-mer C-ring of the F-type ATP synthase
ATP F1	0/5	
COX	2/11	the 2 ohnologs COX4I1,2 coassemble with 3 mitochondrial encoded genes
SRS	2/32	Ohnologs are X-linked RPS4X (with no X-inactivation) and Y-linked RPS4Y1
Mitochondrial SRS	0/30	
LRS	2/50	RPL3 and RPL39 have ohnologs RPL3L and RPL39L with unknown functions
Mitochondrial LRS	0/48	
Proteasome	2/31	ohnologs PSMA7 or PSMA7L are included in the 2 rings of 7 α subunits
Pyruvate dehydrogenase	0/5	
RNA Pol II	0/12	
RNA Pol III	0/9	

COX, cytochrome c oxidase; LRS, large ribosomal subunit; SRS, small ribosomal subunit.
[a]Zanivan et al., 2007.

vertebrate genomes, gene susceptibility to deleterious mutations may be more relevant for the retention of ohnologs within the 17% of human genes participating in more transient complexes. For instance, transient complexes involved in phosphorylation cascades or GTPase signaling pathways are known to be more sensitive to the level of activation of their protein partners than to their total expression levels. Thus, although the active forms of multistate proteins typically amount to a small fraction of their total expression level, hence providing a large dynamic range for signal transduction, it also makes them particularly susceptible to gain-of-function mutations. Such mutations can shift protein activation levels 10- to 100-fold without changes in expression levels and likely underlie stronger evolutionary constraints than the 2-fold dosage imbalance caused by gene duplication.

Indirect Cause of Ohnolog Retention in Protein Complex

To further investigate the relative effects of dosage balance and gene susceptibility to deleterious mutations, we analyzed whether the overall enhanced retention of ohnologs within multiprotein complexes (Figure 1C) could indirectly result from an enhanced susceptibility to deleterious mutations. Indeed, as outlined in Figure 3A, cancer and disease genes are more prevalent within complexes than expected by chance, 29% versus 19% (29%; 2,362/8,095; $p = 3.7 \times 10^{-132}$, χ^2 test) and this trend is enhanced for genes with stronger susceptibility to deleterious mutations, such as oncogenes (39%; 320/813; $p = 2.9 \times 10^{-52}$, χ^2 test) or oncogenes with autoinhibitory folds (59%; 67/114; $p = 2.9 \times 10^{-28}$, χ^2 test). By contrast, ohnologs are only slightly, although significantly, more prevalent in complexes than expected by chance, 22% versus 19% (22%; 1,567/7,110; $p = 9.0 \times 10^{-14}$, χ^2 test), whereas the proportion implicated in cancer or disease genes is clearly enhanced 54% versus 39% (54%; 3,844/7,110; $p = 9.5 \times 10^{-140}$, χ^2 test).

To go beyond these simple statistical associations and quantify the direct versus indirect effects of deleterious mutations and dosage balance constraints on the biased retention of human ohnologs, we have performed a Mediation analysis following the approach of Pearl (Pearl, 2001, 2011). The Mediation framework, developed in the context of causal inference analysis, aims at uncovering, beyond statistical correlations, causal pathways along which changes in multivariate properties are transmitted from a cause, *X*, to an effect, *Y*. More specifically, a Mediation analysis assesses the importance of a mediator, *M*, in transmitting the indirect effect of *X* on the response $Y \equiv Y(x,m(x))$ (Figure 3B).

Mediation analyses have been typically used in social sciences research (Baron and Kenny, 1986) as, for instance, in the context of legal disputes over alleged discriminatory hiring. In such cases, the problem is to establish that gender or race (*X*) have directly influenced hiring (*Y*) and not simply indirectly through differences in qualification or experience (*M*). Mediation analyses have also been used in epidemiology, as in a formal study (Robins and Greenland, 1992) that establishes the direct effect of smoking (*X*) on the incidence of cardiovascular diseases (*Y*), while taking into account the indirect effect of other aggravating factors, such as hyperlipidemia (*M*).

In this report, we have applied the Mediation analysis to genomic data to discriminate between direct effect (*DE*) and indirect effect (*IE*) of deleterious mutations (*X* or *M*) and dosage balance constraints (*M* or *X*) on the biased retention of human ohnologs (*Y*). The results, derived in Extended Experimental Procedures (Table S4) and summarized in Figure 3C and Table S5, demonstrate that the retention of ohnologs in the human genome is more directly caused by their susceptibility to deleterious mutations than their interactions within multiprotein complexes.

Indeed, the direct causal effect of a change from "noncomplex" to "complex" proteins only accounts for 23% of a small total effect (*TE*) of complex on the retention of ohnologs (*DE*/*TE* = 23% with *TE* = 0.079), whereas 82% of this small total effect is indirectly mediated by their susceptibility to deleterious mutations (*IE*/*TE* = 82% with 5% nonlinear coupling between direct and indirect effects) (Extended Results). By contrast, the alternative hypothesis, assuming a direct effect of deleterious mutations, accounts for 99% of a three times larger total effect on ohnolog retention (*DE*/*TE* = 99% with *TE* = 0.23), whereas the "complex" versus "noncomplex" status of human genes

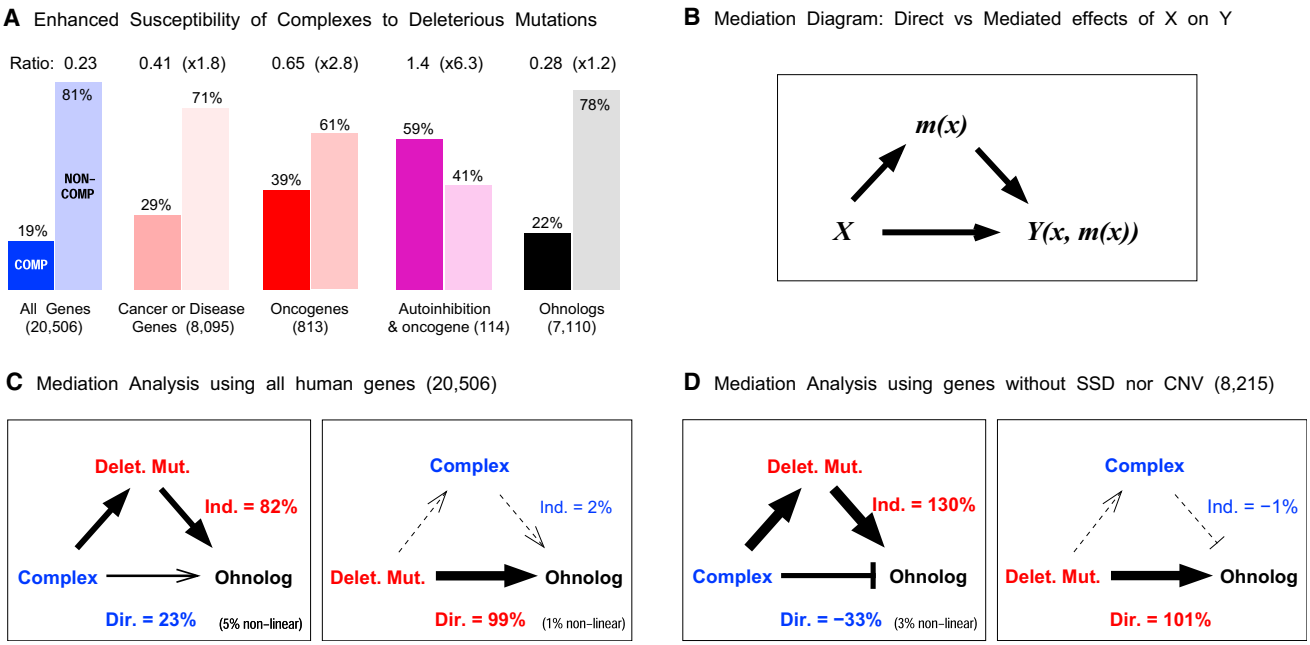

Figure 3. Mediation Analysis of the Indirect Effect of Deleterious Mutations on the Retention of Ohnologs in Multiprotein Complexes

(A) Enhanced susceptibility of complexes to deleterious mutations.

(B) Mediation diagram depicting the direct versus indirect (i.e., mediated) effects of the cause *X* on the outcome *Y*(*x*,*m*(*x*)) (Pearl, 2011). See also Extended Experimental Procedures.

(C and D) Quantitative Mediation analysis of direct versus indirect effects of deleterious mutations and dosage balance on the retention of human ohnologs using (C) all human genes (20,506) or (D) all human genes without SSD nor CNV (8,215). The thickness of the arrows outlines the relative importance of the corresponding direct or indirect effects. These results are consistent with those obtained from partial correlation analysis.

See also the main text, Extended Results, and Tables S4, S5, and S6.

has a negligible indirect effect on ohnolog retention in this case (*IE*/*TE* = 2%) (Extended Results). These trends are also further enhanced when the analysis is restricted to the 40% of human genes (8,215) without SSD and CNV duplicates (Figure 3D; Table S5; Extended Results). In fact, the direct effect of multiprotein complexes then tends to oppose the retention of ohnologs (*DE*/*TE* = −33% with *TE* = 0.064), as in the case of permanent complexes detailed above, but on an increased sample size of 8,215 genes without SSD or CNV duplicates (i.e., more than a third of human genes) in place of 239 genes from permanent complexes. By contrast, there is a five times larger total effect due to the direct effect of deleterious mutations on the retention of ohnologs (*DE*/*TE* = 101% with *TE* = 0.32), Figure 3D. This is an instance of Simpson's paradox, where two effects oppose each other, thereby, revealing the existence of conflicting underlying causes, namely, a strong positive effect of deleterious mutations and a small negative effect of dosage balance constraints on the retention of human ohnologs without SSD and CNV duplicates.

We have also examined the effects of other alternative properties on the retention of ohnologs (Extended Results; Table S5). In particular, we have found that gene expression levels and Ka/Ks ratios do not significantly mediate the effect of deleterious mutations on the retention of ohnologs. In fact, gene expression levels (Extended Experimental Procedures) have a negligible total effect on the retention of human ohnologs (*TE* = 0.003), by contrast to what has been reported for the paramecium (Gout et al., 2009). The total effects of Ka/Ks on ohnolog retention are also lower than the total effects of deleterious mutations, as *TEs* from deleterious mutations are ~2- to 3-fold stronger than *TEs* from Ka/Ks and become >10-fold stronger for genes without SSD and CNV (Extended Results).

In addition, we have performed a complementary systematic study of all these genomics properties using partial correlation analysis, which aims at "removing" the effect of a third property (*Z*) on the standard pair correlations between two variables (*X*) and (*Y*). The results detailed in Extended Results and Table S6 are entirely consistent with those obtained through mediation analysis, although the two approaches are not equivalent. Indeed, although mediation effects require partial correlation, partial correlation does not imply mediation, in general (Extended Results).

All in all, these results support the fact that the retention of ohnologs in the human genome is more strongly associated with their "dangerousness" (i.e., susceptibility to dominant deleterious mutations) than with their functional importance ("essentiality"), sensitivity to dosage balance, absolute expression levels or sequence conservation (i.e., Ka/Ks).

Model for the Retention of "Dangerous" Ohnologs

As demonstrated above, human genes with a documented sensitivity to dominant deleterious mutations have retained statistically more ohnologs from the two WGD events at the

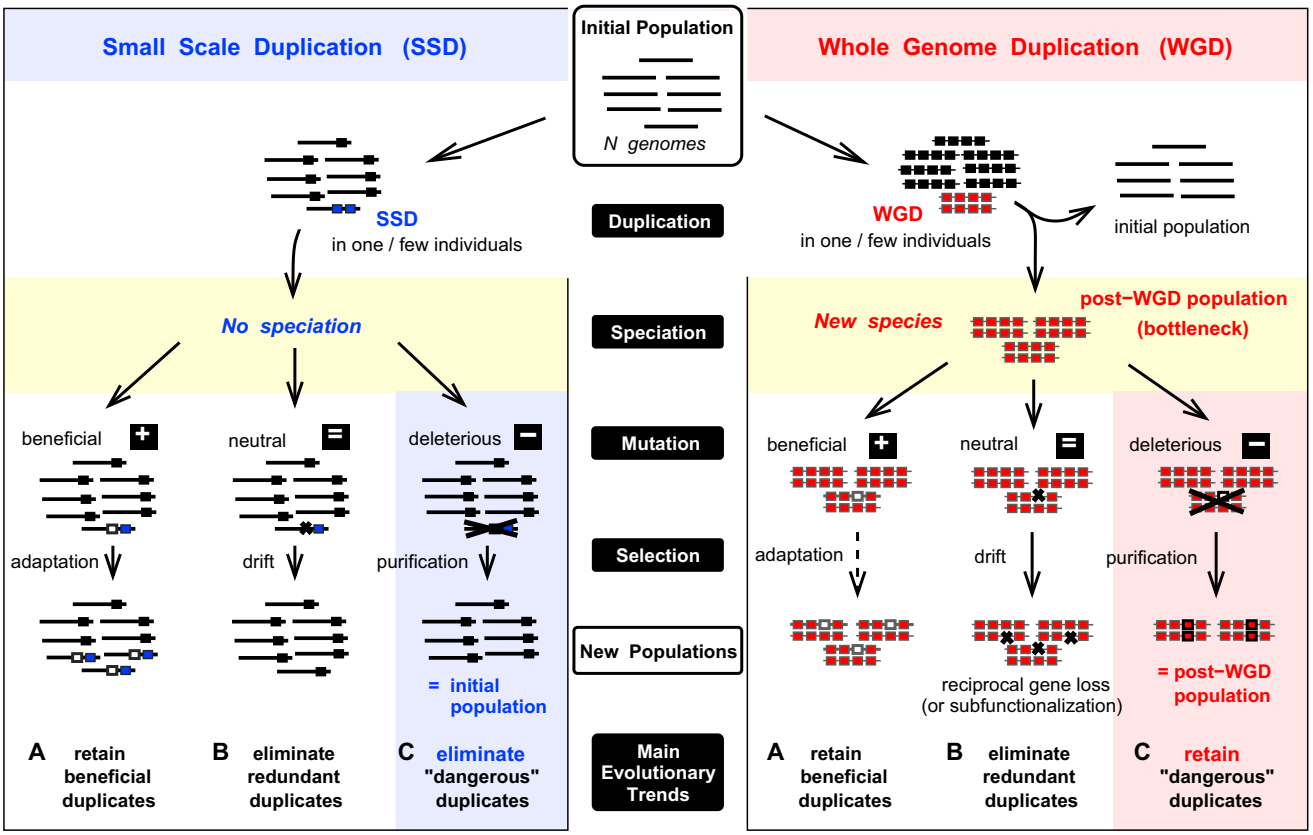

Figure 4. Evolutionary Trends of Duplicated Genes following SSD or WGD

(A–C) Horizontal lines represent the genome of different individuals. Square blocks symbolize the genes, duplicated (SSD: blue; WGD: red) or not (black). Black crosses highlight the loss of one gene (small crosses) or the elimination of an individual (larger crosses), whereas bordered square blocks emphasize retained mutated copies. Evolutionary scenarios are depicted at the population genetics level following either a SSD (left panel) or a WGD (right panel) in one or a few individuals of an initial population. Unlike SSD, WGD is invariably coupled to a speciation event, owing to the difference in ploidy between pre- and post-WGD individuals. Three possible scenarios—beneficial (A), neutral or nearly neutral (B), or deleterious mutations (C) in one gene duplicate—are outlined in post-SSD and post-WGD populations. The main difference concerns the mutation/selection process of "dangerous" genes, i.e. genes prone to autosomal-dominant deleterious mutations (C). See main text for a detailed description.

onset of jawed vertebrates. This suggests that ohnologs have been retained in vertebrate genomes, not because they initially brought selective advantages following WGD, but because their mutations were more likely detrimental or lethal than nonfunctional, thereby preventing their rapid elimination from the genomes of surviving individuals following WGD transitions, as outlined in the evolutionary model depicted in Figure 4.

For completeness and clarity, Figure 4 examines all possible evolutionary scenarios following either a SSD or a WGD duplication event in the genome of one or a few individuals in an initial population. The first and critical difference between SSD and WGD duplication events occurs at the population genetics level with an obligate speciation following WGD event, owing to the difference in ploidy between pre- and post-WGD individuals. As a result, all individuals in the post-WGD population carry twice as many genes as their pre-WGD relatives, whereas only a few individuals in the post-SSD population carry a single small duplicated region. Figure 4 then outlines the three mutation/selection scenarios focusing on a single gene duplicate in the genomes of post-SSD or post-WGD populations: (A) Beneficial mutations after SSD or WGD are expected to spread and become eventually fixed in the new populations, although the bottleneck in population size following WGD limits in practice the efficacy of adaptation in post-WGD species. (B) Neutral or nearly neutral mutations mainly lead to the random nonfunctionalization of one copy of most redundant gene duplicates and, therefore, to their elimination following both SSD and WGD events. In post-WGD populations, this results in the "reciprocal gene loss" of most gene duplicates, which is also known to lead to further speciations in post-WGD species, owing to the interbreeding incompatibility between post-WGD individuals with different "reciprocal gene loss" pattern (Lynch and Force, 2000a). Alternatively, neutral or nearly neutral mutations can also result in the eventual retention of both duplicate copies through subfunctionalization (Hughes, 1994; Lynch and Force, 2000b), that is, by rendering each duplicate copy unable to perform all the functions of their ancestral gene (see Discussion). (C) Finally, dominant deleterious mutations favor the elimination of the individuals

(or their descendants) harboring them through purifying selection. However, this typically leads to opposite outcomes in post-SSD and post-WGD populations. In post-SSD populations, dominant deleterious mutations will tend to eliminate SSD duplicates before they have the time to reach fixation (see below). By contrast, in post-WGD populations, where all ohnologs have been initially fixed through WGD-induced speciation, purifying selection will effectively favor the retention of dangerous ohnologs, as all surviving individuals still present (nondeleterious) functional copies of these dangerous genes.

Note, in particular, that this somewhat counterintuitive evolutionary model for the retention of ''dangerous'' ohnologs hinges on two unique features:

(1) It requires an autosomal dominance of deleterious mutations, in agreement with our observation, above, that retained ohnologs are more ''dangerous'' than ''essential.''
(2) It relies on the fact that successful WGD events start with a concomitant speciation event, which immediately fixes all ohnolog duplicates in the initial post-WGD population (Figure 4).

Note, also, that the same evolutionary trend is expected for dangerous SSD duplicates that would have the time (t) to become fixed through genetic drift in a population of size N before deleterious mutations can arise at a rate K, i.e., $t = 4N < 1/K$. This corresponds to a population bottleneck effect with $N < 1/(4K) \approx 5{,}000–10{,}000$ for typical vertebrates.

DISCUSSION

Beyond human and vertebrate genomes, WGD events have now been established in all major eukaryote kingdoms (Sémon and Wolfe, 2007; Evlampiev and Isambert, 2007). Unlike SSD events, WGD transitions provide a unique evolutionary mechanism, enabling the simultaneous duplication of entire genetic pathways and multiprotein complexes, followed by long periods of functional divergence and extensive loss of ohnologs (Aury et al., 2006). Moreover, although both WGD and SSD events have expanded the gene repertoires and resulting protein networks (Evlampiev and Isambert, 2007; Evlampiev and Isambert, 2008) of eukaryotes, it has become increasingly clear that WGD and SSD events actually lead to the expansion of different gene classes in the course of evolution, (Maere et al., 2005; Aury et al., 2006; Sémon and Wolfe, 2007; Makino and McLysaght, 2010; Huminiecki and Heldin, 2010; and this study).

In this article, we report that WGD have effectively favored the expansion of gene families prone to deleterious mutations in the human genome, such as genes implicated in cancer and genes with autoinhibitory interactions. In particular, we found that the retention of many ohnologs suspected to be dosage balanced is in fact indirectly mediated by their susceptibility to deleterious mutations.

From a broader perspective, a number of studies have now shown that many genomic properties, such as gene essentiality, duplicability, functional ontology, network connectivity, expression level, mutational robustness, divergence rates, etc., all appear to be correlated to some extent. In the light of the present study, we expect that many of these statistically significant correlations mainly result from indirect rather than direct associations, which may even frequently oppose each other. This highlights the need to rely on more advanced inference methods to analyze the multiple, direct, and indirect causes underlying the evolution of specific gene repertoires.

In the present study, we have quantitatively analyzed the direct versus indirect effects of the susceptibility of human genes to deleterious mutation and dosage balance constraints on the retention of ohnologs and proposed a simple evolutionary mechanism to account for the initial retention of ''dangerous'' ohnologs after WGD (Figure 4). On longer timescales, we expect that this initial retention bias of ''dangerous'' ohnologs effectively promote a prolonged genetic drift and, thus, a progressive functional divergence between ohnolog pairs. This eventually favors the subfunctionalization (Hughes, 1994; Lynch and Force, 2000b) of ancestral functions between ohnolog pairs, which ultimately warrants their long-term maintenance following WGD events.

Note, however, that this subfunctionalization process requires that the expression of ohnologs is not rapidly suppressed by large-scale deletion or silencing mutations in regulatory regions. As ohnolog pairs are not arranged in tandem, large-scale deletions through unequal crossing-over cannot typically remove entire ohnolog duplicates while preserving the integrity of nearby genes. Furthermore, as the size of promoter or enhancer regions is typically much smaller than UTRs and coding regions, one expects that the rate of transcriptional silencing does not exceed the rates of functional silencing and divergence in UTRs and coding regions. In fact, early estimates (Nadeau and Sankoff, 1997) showed that gene loss and functional divergence after genome duplications in early vertebrates occurred at comparable rates in gene families including at least two ohnologs. This is also directly evidenced by pseudotetraploid species like the vertebrate *Xenopus laevis*, which still retains ~40% of its initial ohnologs from a 30-million-year-old WGD (Sémon and Wolfe, 2008). All in all, this suggests that ohnologs prone to dominant deleterious mutations have at least a few million years to diverge and become nonredundant genes before they have a chance to be deleted or transcriptionally silenced.

Yet, we found that the retention of these dangerous ohnologs remains intrinsically stochastic by nature as many of them have also been eliminated following WGD events. This presumably occurred through loss-of-function mutations, transcriptional silencing, or large-scale deletion before ohnolog pairs could diverge and become nonredundant genes. More quantitatively, a simple theoretical estimate, derived from the long-term retention statistics of Figure 1, shows that only 6%–10% of the initial ohnolog duplicates have been retained on average at each round of WGD, Figure 5 (see Extended Results for details). By comparison, ~23%–30% of the initial ohnologs prone to gain-of-function mutations have been retained on average at each WGD. This implies that genes susceptible to deleterious mutations are two to five times more likely to retain ohnologs on long evolutionary timescales. Moreover, genes combining several factors associated with enhanced susceptibility to autosomal-dominant deleterious mutations are shown to be more than ten times more

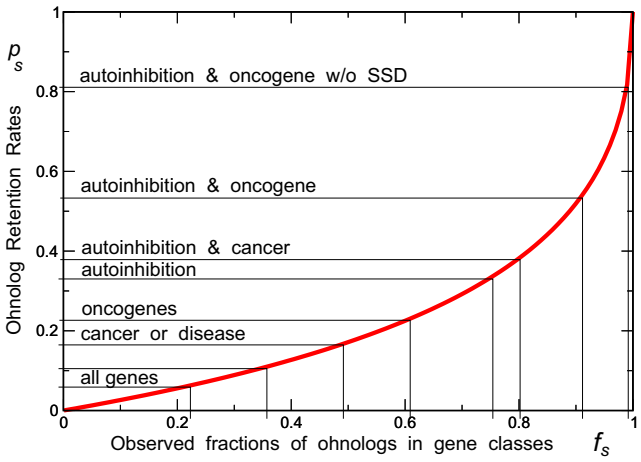

Figure 5. Estimates of Ohnolog Retention Rates

Estimates of ohnolog retention rates p_s in early vertebrates from the observed fraction f_s of ohnologs in the human genome for gene classes, s, with increasing susceptibility to deleterious mutations. The theoretical estimate (red curve) is obtained assuming that the retentions of ohnologs were comparable for each of the two WGD at the onset of vertebrates, and reads $P_s = 2/f_s - 1 - \sqrt{(2/f_s - 1)^2 - 1}$ as detailed in the Extended Results and Tables S7 and S8.

likely to retain ohnologs than genes lacking gain-of-function mutations (Figure 5), as illustrated on the examples of oncogenes with autoinhibitory folds (Table 1).

In turn, the elimination of ohnologs has been shown to drive further speciation events within post-WGD (sub)populations, due to the emergence of recombination barriers from the accumulation of differences in ohnolog deletion patterns between post-WGD individuals (Lynch and Force, 2000a). The resulting fragmentation of post-WGD subpopulations is then expected to sustain negative selection pressure that favors the retention of the remaining ohnolog pairs prone to deleterious mutations, as outlined in Figure 4. Hence, although most WGDs are unlikely to bring much fitness benefit on short evolutionary timescales (if only due to the population bottlenecks associated with WGD-induced speciations; Figure 4), they provide a unique evolutionary mechanism to experiment virtually unlimited combinations of regulation/deletion patterns from redundant ohnolog genes. Over long timescales (>100–500 MY), such trial and error combinations have visibly led to the evolutionary success and radiation of WGD species.

In summary, we present evidence supporting an evolutionary link between the susceptibility of human genes to dominant deleterious mutations and the documented expansion of these "dangerous" gene families by two WGD events at the onset of jawed vertebrates. We propose that deleterious mutations, responsible for many cancers and other severe genetic diseases on the lifespan of human individuals, have also underlain purifying selection over long evolutionary timescales, which effectively favored the retention of vertebrate ohnologs prone to dominant deleterious mutations, as outlined in Figure 4. From a population genetics perspective, we argue that this counterintuitive retention of dangerous ohnologs hinges in fact on WGD-induced speciation events, which are largely credited for the genetic complexity and successful radiation of vertebrate species.

These findings highlight the importance of purifying selection from WGD events on the evolution of vertebrates and, beyond, exemplify the role of nonadaptive forces on the emergence of eukaryote complexity (Fernández and Lynch, 2011).

EXPERIMENTAL PROCEDURES

WGD Duplicated Genes or "Ohnologs"

Human ohnolog genes were obtained from (Makino and McLysaght, 2010). Makino and McLysaght compared different vertebrate and six nonvertebrate outgroup genomes to identify ohnologs in the human genome. The final data set consists of 8,653 ohnolog pairs and 7,110 unique ohnologs. We further divided ohnologs into well supported (3,963), plausible (894), and more uncertain (2,253) ohnologs (see Extended Experimental Procedures).

SSD Duplicated Genes

We identified paralogous genes within the human genome from sequence similarity search. We obtained a total of 11,185 SSD genes. In particular, paralogs that were not annotated as ohnologs were taken to be SSD genes (see Extended Experimental Procedures).

Genes with CNV

CNV regions were obtained from Database of Genomic Variants (Zhang et al., 2006). A total of 5,709 genes were identified to be CNV genes as their entire coding sequence fell within one of the CNV regions.

Cancer and Disease Genes

We obtained cancer genes from multiple databases, including COSMIC (Forbes et al., 2011) and CancerGenes (Higgins et al., 2007), listed in Table S7. The detailed list of 6,917 cancer genes is given in Table S8 with a hand-curated list of 813 verified or predicted (Bozic et al., 2010) oncogenes (see Extended Experimental Procedures). We obtained 2,580 disease genes from the "Morbidmap" database of OMIM and hand curated subsets of 440 autosomal-dominant and 598 autosomal-recessive disease genes from Blekhman et al. (2008).

Genes with Autoinhibitory Folds

To obtain genes coding for proteins with autoinhibitory folds we searched PubMed with keyword "autoinhibitory domain" and retrieved relevant autoinhibitory genes and domains manually. Further gene candidates with autoinhibitory folds were obtained from databases, OMIM, SwissProt, NCBI Gene, and GeneCards using the parsing terms: auto/self-inhibit*. Careful manual curation of this list of gene candidates with the available literature finally yielded a total of 461 genes with autoinhibitory folds (94% of initial candidates).

Essential Genes

Mouse essential genes were obtained from Mouse Genome Informatics database. Essential genes were defined as genes having lethal or infertility phenotypes on loss-of-function or knockout mutations (2,729 genes) (see Extended Experimental Procedures).

Genes in Complexes and Permanent Complexes

Protein complexes were obtained from Human Protein Reference Database (HPRD) (Keshava Prasad et al., 2009) and CORUM database (Ruepp et al., 2010). In addition, a manually curated data set of permanent complexes (239 genes) was obtained from Zanivan et al. (2007). The final data set consists of 3,814 protein complex genes (see Extended Experimental Procedures).

Haploinsufficient and Dominant Negative Genes

Haploinsufficient and dominant negative candidate genes were obtained from parsing OMIM text files with Perl regular expressions. The resulting gene lists were manually curated with the available literature, yielding a total of

330 haploinsufficient genes (80% of initial candidates) and 477 dominant-negative genes (63% of initial candidates).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Results, Extended Experimental Procedures, three figures, and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2012.09.034.

LICENSING INFORMATION

ACKNOWLEDGMENTS

P.P.S. acknowledges a PhD fellowship from Erasmus Mundus and Université Pierre et Marie Curie; S.A. acknowledges a PhD fellowship from Ministry of Higher Education and Research, France; I.C. acknowledges postdoctoral support from ANR (grant ANR-08-BLAN-0290); R.S. acknowledges PhD fellowships from INCa and ARC; H.I. and J.C. acknowledge funding from Foundation Pierre-Gilles de Gennes. We thank H. Roest Crollius, L. Peliti, S. Coscoy and V. Hakim for discussions.

Received: April 12, 2012
Revised: September 17, 2012
Accepted: September 27, 2012
Published: November 15, 2012

REFERENCES

Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature *444*, 171–178.

Baron, R.M., and Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. *51*, 1173–1182.

Berry, L.W., Westlund, B., and Schedl, T. (1997). Germ-line tumor formation caused by activation of glp-1, a Caenorhabditis elegans member of the Notch family of receptors. Development *124*, 925–936.

Birchler, J.A., Bhadra, U., Bhadra, M.P., and Auger, D.L. (2001). Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. Dev. Biol. *234*, 275–288.

Blekhman, R., Man, O., Herrmann, L., Boyko, A.R., Indap, A., Kosiol, C., Bustamante, C.D., Teshima, K.M., and Przeworski, M. (2008). Natural selection on genes that underlie human disease susceptibility. Curr. Biol. *18*, 883–889.

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.

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K.W., Vogelstein, B., and Nowak, M.A. (2010). Accumulation of driver and passenger mutations during tumor progression. Proc. Natl. Acad. Sci. USA *107*, 18545–18550.

Brunet, F.G., Roest Crollius, H., Paris, M., Aury, J.M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol. Biol. Evol. *23*, 1808–1816.

Cai, J.J., Borenstein, E., Chen, R., and Petrov, D.A. (2009). Similarly strong purifying selection acts on human disease genes of all evolutionary ages. Genome Biol. Evol. *1*, 131–144.

Ciocan, C.M., Moore, J.D., and Rotchell, J.M. (2006). The role of ras gene in the development of haemic neoplasia in Mytilus trossulus. Mar. Environ. Res. Suppl. *62*, S147–S150.

Dickerson, J.E., and Robertson, D.L. (2012). On the origins of Mendelian disease genes in man: the impact of gene duplication. Mol. Biol. Evol. *29*, 61–69.

Domazet-Loso, T., and Tautz, D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. Mol. Biol. Evol. *25*, 2699–2707.

Esteban, L.M., Vicario-Abejón, C., Fernández-Salguero, P., Fernández-Medarde, A., Swaminathan, N., Yienger, K., Lopez, E., Malumbres, M., McKay, R., Ward, J.M., et al. (2001). Targeted genomic disruption of H-ras and N-ras, individually or in combination, reveals the dispensability of both loci for mouse growth and development. Mol. Cell. Biol. *21*, 1444–1452.

Evangelisti, A.M., and Conant, G.C. (2010). Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. Genome Biol. Evol. *2*, 826–834.

Evlampiev, K., and Isambert, H. (2007). Modeling protein network evolution under genome duplication and domain shuffling. BMC Syst. Biol. *1*, 49.

Evlampiev, K., and Isambert, H. (2008). Conservation and topology of protein interaction networks under duplication-divergence evolution. Proc. Natl. Acad. Sci. USA *105*, 9863–9868.

Fernández, A., and Lynch, M. (2011). Non-adaptive origins of interactome complexity. Nature *474*, 502–505.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*(Database issue), D945–D950.

Freeling, M., and Thomas, B.C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. *16*, 805–814.

Furney, S.J., Albà, M.M., and López-Bigas, N. (2006). Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics *7*, 165.

Gibson, T.J., and Spring, J. (1998). Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet. *14*, 46–49, discussion 49–50.

Gout, J.F., Duret, L., and Kahn, D. (2009). Differential retention of metabolic genes following whole-genome duplication. Mol. Biol. Evol. *26*, 1067–1072.

Gout, J.F., Kahn, D., and Duret, L.; Paramecium Post-Genomics Consortium. (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet. *6*, e1000944.

Guan, Y., Dunham, M.J., and Troyanskaya, O.G. (2007). Functional analysis of gene duplications in Saccharomyces cerevisiae. Genetics *175*, 933–943.

Hakes, L., Pinney, J.W., Lovell, S.C., Oliver, S.G., and Robertson, D.L. (2007). All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol. *8*, R209.

Higgins, M.E., Claremont, M., Major, J.E., Sander, C., and Lash, A.E. (2007). CancerGenes: a gene selection resource for cancer genome projects. Nucleic Acids Res. *35*(Database issue), D721–D726.

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

Huminiecki, L., and Heldin, C.H. (2010). 2R and remodeling of vertebrate signal transduction engine. BMC Biol. *8*, 146.

Ise, K., Nakamura, K., Nakao, K., Shimizu, S., Harada, H., Ichise, T., Miyoshi, J., Gondo, Y., Ishikawa, T., Aiba, A., and Katsuki, M. (2000). Targeted deletion of the H-ras gene decreases tumor formation in mouse skin carcinogenesis. Oncogene *19*, 2951–2956.

Kellis, M., Birren, B.W., and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature *428*, 617–624.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human Protein Reference Database—2009 update. Nucleic Acids Res. *37*(Database issue), D767–D772.

Lin, Y.S., Hwang, J.K., and Li, W.H. (2007). Protein complexity, gene duplicability and gene dispensability in the yeast genome. Gene *387*, 109–117.

Lynch, M., and Force, A. (2000a). Gene duplication and the origin of interspecific genomic incompatibility. Am. Nat. *156*, 590–605.

Lynch, M., and Force, A. (2000b). The probability of duplicate gene preservation by subfunctionalization. Genetics *154*, 459–473.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. USA *102*, 5454–5459.

Makino, T., and McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc. Natl. Acad. Sci. USA *107*, 9270–9274.

Nadeau, J.H., and Sankoff, D. (1997). Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics *147*, 1259–1266.

Ohno, S. (1970). Evolution by Gene Duplication (New York: Springer-Verlag).

Papp, B., Pál, C., and Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. Nature *424*, 194–197.

Pearl, J. (2001). Direct and indirect effects. Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence, San Francisco, CA: Morgan Kaufmann, 411–420.

Pearl, J. (2011). The Mediation Formula: A guide to the assessment of causal pathways in nonlinear models. In Causality: Statistical Perspectives and Applications, C. Berzuini, P. Dawid, and L. Bernardinelli, eds. (United Kingdom: John Wiley & Sons), pp. 151–175.

Pufall, M.A., and Graves, B.J. (2002). Autoinhibitory domains: modular effectors of cellular regulation. Annu. Rev. Cell Dev. Biol. *18*, 421–462.

Putnam, N.H., Butts, T., Ferrier, D.E., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.K., et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. Nature *453*, 1064–1071.

Qian, W., and Zhang, J. (2008). Gene dosage and gene duplicability. Genetics *179*, 2319–2324.

Robert, J. (2010). Comparative study of tumorigenesis and tumor immunity in invertebrates and nonmammalian vertebrates. Dev. Comp. Immunol. *34*, 915–925.

Robins, J.M., and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. Epidemiology *3*, 143–155.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. *38*, 497–501.

Sémon, M., and Wolfe, K.H. (2007). Consequences of genome duplication. Curr. Opin. Genet. Dev. *17*, 505–512.

Sémon, M., and Wolfe, K.H. (2008). Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. Proc. Natl. Acad. Sci. USA *105*, 8333–8338.

Veitia, R.A. (2002). Exploring the etiology of haploinsufficiency. Bioessays *24*, 175–184.

Veitia, R.A. (2010). A generalized model of gene dosage and dominant negative effects in macromolecular complexes. FASEB J. *24*, 994–1002.

Wolfe, K. (2000). Robustness—it's not where you think it is. Nat. Genet. *25*, 3–4.

Zanivan, S., Cascone, I., Peyron, C., Molineris, I., Marchio, S., Caselle, M., and Bussolino, F. (2007). A new computational approach to analyze human protein complexes and predict novel protein interactions. Genome Biol. *8*, R256.

Zeevi, D., Sharon, E., Lotan-Pompan, M., Lubling, Y., Shipony, Z., Raveh-Sadka, T., Keren, L., Levo, M., Weinberger, A., and Segal, E. (2011). Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters. Genome Res. *21*, 2114–2128.

Zhang, J., Feuk, L., Duggan, G.E., Khaja, R., and Scherer, S.W. (2006). Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet. Genome Res. *115*, 205–214.

Cell PRESS
Open ACCESS

EXTENDED RESULTS

Analysis of Ohnologs Conservation Using Ka/Ks Ratios

We quantify in this extended result the susceptibility of ohnologs versus nonohnologs to deleterious mutations through comparative sequence analysis, using Ka/Ks ratio estimates (Yang and Nielsen, 2000), which measure the proportion of nonsynonymous substitutions (Ka) to the proportion of synonymous substitutions (Ks).

Ka/Ks ratios were calculated to evaluate the selection pressure on human ohnolog and nonohnolog genes and their respective invertebrate orthologs as follows. Protein sequences for *B. floridae, C. intestinalis, C. savignyi, D. melanogaster* and *C. elegans* were obtained from Ensembl release 61. *S. purpuratus* proteins were downloaded from NCBI Genome FTP. Human-invertebrate orthologs were identified using BLASTp (E-value: $< 10^{-7}$). The best human-to-invertebrate hits were identified and used for Ka/Ks ratio calculations (Yang and Nielsen, 2000). Each human ohnolog pair and human-invertebrate ortholog pair was aligned using clustalW (Larkin et al., 2007). Ka and Ks values (Yang and Nielsen, 2000) have been calculated using the KaKs_Calculator 2.0 (Wang et al., 2010). Ohnolog or ortholog pairs for which Ka/Ks ratio could not be calculated (0.1%–0.7% of pairs) or having saturated Ks values (2%–7% of pairs) were discarded from the analysis, but did not significantly affect the overall results. The status of human ohnolog pairs were defined as the number of outgroups supporting them (1-6), Extended Experimental Procedures. For human-invertebrate ortholog pairs, the ohnolog status of the human gene was defined as the maximum outgroup (1-6) over its human ohnolog pairs, restricted to its ohnolog partners with the same ortholog best hit.

Human genes identified as ohnologs by more than half of the invertebrate outgroups (>3) were considered well supported ohnologs, whereas genes identified as possible ohnologs by less than half of the invertebrate outgroups (<3) were regarded as more uncertain. We found that human ohnolog (*Hs-Hs*) pairs with well supported status (2,984) have statistically lower Ka/Ks ratios than *Hs-Hs* pairs with more uncertain ohnolog status (4,472), Figure 2A (Mann-Whitney-Wilcoxon (MWW) nonparametric test: $p < 2 \times 10^{-308}$, Table S3). Similarly, comparing human-amphioxus (*Hs-Bf*) ortholog pairs, we found that *Hs-Bf* pairs involving a human ohnolog with a well-supported status (3,151) have statistically lower Ka/Ks ratios than *Hs-Bf* pairs involving a human gene with a more uncertain ohnolog status (1,292; MWW test: $p = 2.2 \times 10^{-125}$, Table S3) or than *Hs-Bf* pairs involving a nonohnolog human gene (9,406; MWW test: $p = 6.1 \times 10^{-234}$, Table S3), Figure 2B. Similar trends are found with other invertebrate outgroups (Figures S2, S3,and Table S3) and directly from Ka distributions, that are unaffected by possible Ks saturation. Similar trends have also been reported for ohnologs specific to teleost fishes (Brunet et al., 2006) or to the more recent WGD in *Xenopus laevis* lineage (Sémon and Wolfe, 2008).

By contrast, genes with SSD or CNV duplicates in the human genome exhibit significantly higher Ka/Ks ratios (Figure 2D, MWW test: $p = 1.8 \times 10^{-92}$, Table S3) than ohnologs (Figure 2C, MWW test: $p = 5.1 \times 10^{-142}$, Table S3). This suggests different underlying evolutionary constraints on the sequences of ohnologs and SSD or CNV duplicates.

As lower Ka/Ks and Ka distributions provide direct evidence for stronger negative selection pressure, these results are consistent with a higher susceptibility of ohnologs to deleterious mutations. However, the functional consequences of such deleterious mutations, leading either to a gain or a loss of function, cannot be directly inferred from Ka/Ks and Ka distributions. Yet, as shown in the main text, we observed marked differences in the retention biases of ''dangerous'' ohnologs prone to dominant gain-of-function mutations and ''essential'' ohnologs exhibiting lethal loss-of-function or null mutations.

Details of Mediation Analysis Results

In this extended result section we have performed a series of Mediation Analyses concerning the direct and indirect effects on ohnolog retention ('ohno') caused by gene susceptibility to deleterious mutations ('delet. mut') and dosage balance constraints ('dosage bal.'), where,

- 'delet. mut.' genes (7,227) include 6,917 'cancer' genes, 440 'dominant disease' genes, 477 'dominant negative' genes and 461 'autoinhibited' genes
- 'dosage. bal.' genes (4,003) include 3,814 protein 'complex' and 330 'haploinsufficient' genes.

We also study the effect of protein 'complex' genes independently from 'haploinsufficient' genes.

The results are summarized in Table S5. See full mediation analyses below for details.

We first find that the global 'complex' ⇒ 'ohno' link is strongly mediated by 'delet. mut.' genes (Table S5 1.a: *IE/TE* = 82.0%), whereas 'complex' genes have a negligible indirect effect on the global 'delet. mut.' ⇒ 'ohno' link (Table S5 2.a: *IE/TE* = 1.8%).

Adding 'haploinsufficient' genes (330) to 'complex' genes (3,814), diminishes slightly the mediation by 'delet. mut.' genes but does not significantly change the causal relations, as we find that the global 'dosage. bal.' ⇒ 'ohno' link remains strongly mediated by 'delet. mut.' genes (Table S5 1.b: *IE/TE* = 74.1%), whereas 'dosage. bal.' genes retain a negligible indirect effect on the global 'delet. mut.' ⇒ 'ohno' link (Table S5 2.b: *IE/TE* = 2.6%). In fact, 'haploinsuffficient' genes are also bona fide dominant deleterious genes and presumably contribute as other genes prone to dominant deleterious mutations to enhance the retention of ohnologs. Hence, 'haploinsufficient' genes could equally be associated to 'delet. mut.' genes or to 'dosage. bal.' genes or can be removed altogether from the mediation analysis as in Table S5 (1.a, 1.a′, 2.a, 2.a′) and in Figures 3C and 3D in the main text.

These trends are further enhanced when excluding 5,709 'CNV' and 9,917 'SSD' genes (Table S5 1.a′, 1.b′, 2.a′, 2.b′). There is even a negative direct effect of 'complex' on ohnolog retention, when excluding 'SSD' genes, *i.e.* a Simpson's paradox (see main text).

This supports a direct effect of deleterious mutations on the retention of ohnologs (Table S5 2.a, 2.a′, 2.b and 2.b′) and a largely indirect effect for genes susceptible to dosage balance (Table S5 1.a, 1.a′, 1.b and 1.b′).

To confirm these results about the prevalence of deleterious mutations on the retention of ohnologs, we have also quantified the influence of other genomic properties that have been suggested to impact the retention of ohnologs, such as gene expression levels (Gout et al., 2009) and Ka/Ks values (Brunet et al., 2006; and this study).

To enable direct comparisons with the results obtained for the binary properties, dosage balance and deleterious mutations, above, gene expression levels and Ka/Ks values have also been binarized with the threshold 0/1 at their median across all human genes with measured expression levels or available invertebrate orthologs for Ka/Ks estimates, respectively (See Extended Experimental Procedures and Extended Results).

The results, outlined in Table S5, demonstrate that gene expression levels (Table S5 3a and 3a′) and Ka/Ks values estimated from 6 different invertebrate genomes (Table S5 3b-g and 3b′-g′) do not significantly mediate the total effect of deleterious mutations on the retention of human ohnologs, irrespective of the inclusion of SSD and CNV genes or not.

Similarly and consistently with the results obtained for the effect of dosage balance constraints above, we also found that gene expression levels (Table S5 4a) and Ka/Ks values (Table S5 4b–g) exhibit smaller total effects on the retention of ohnologs than their susceptibility to deleterious mutations (Table S5 2a and 3b–g).

In fact we found that gene expression levels do not have a significant total effect on the retention of human ohnologs ($TE = 0.003$), by contrast to what has been reported for the paramecium (Gout et al., 2009).

The total effects of Ka/Ks on ohnolog retention are also lower (Table S5 4b–g; $TE = 0.08 \pm 0.04$) than the total effects of 'delet. mut.' restricted to human genes with orthologs in the corresponding invertebrate genomes (Table S5 3b–g; $TE = 0.18 \pm 0.01$) or across all the human genes (Table S5 2a; $TE = 0.23$). Note that this trend is also stronger for genes without SSD & CNV, as the *TEs* from Ka/Ks decrease (Table S5 4b′–g′; $TE = 0.02 \pm 0.07$) while the *TEs* from 'delet. mut.' increase (Table S5 3b′–g′; $TE = 0.2 \pm 0.02$) and (Table S5 2a′; $TE = 0.32$). Hence, *TEs* from deleterious mutations are about 2 to 3-fold stronger than *TEs* from Ka/Ks and become more than 10-fold stronger for genes without SSD & CNV.

All in all, these results are consistent with our finding (see main text) that the retention of ohnologs is more strongly associated with their "dangerousness" (i.e., susceptibility to dominant deleterious mutations) than with their functional importance ("essentiality"), sensitivity to dosage balance, absolute expression levels or their sequence conservation (i.e., Ka/Ks).

Case of Mediation of 'Complex' ⇒ 'Ohno' Link by 'Delet. Mut.' Genes, Table S5, 1.a

We analyze the association table between the three binary categories 'complex', 'delet. mut.' and 'ohno' genes. 'Delet. mut.' genes are genes prone to deleterious mutations, *i.e.* cancer, dominant disease, dominant negative genes and genes with autoinhibitory protein folds.

As derived in Table S4, we find,

- A significant global effect of 'complex' on 'ohno' retention: 41.1% versus 34.7% ($p = 8.7 \times 10^{-17}$, χ^2 test).
- A strong effect of 'complex' on 'delet. mut.' genes: 58.2% versus 35.2% ($p = 3.2 \times 10^{-193}$, χ^2 test).
- A significant effect of 'delet. mut.' on 'ohno' retention for 'complex' genes: 50.1% versus 41.1% ($p = 8.0 \times 10^{-18}$, χ^2 test).
- A strong effect of 'delet. mut.' on 'ohno' retention for non-'complex' genes: 49.3% versus 33.2% ($p = 8.0 \times 10^{-128}$, χ^2 test).

The Mediation analysis (Extended Experimental Procedures) then yields, $TE = 0.0788$, $DE = 0.0183$, $IE = 0.0646$ and the relative direct and indirect effects:

$DE/TE = 23.2\%$, $IE/TE = 82.0\%$;
$1-IE/TE = 18.0\%$, $1-DE/TE = 76.8\%$
and 5.2% of a non-linear combination of direct and indirect effects.

This implies that,

- Only 18.0% of the global effect on 'ohno' retention is owed to the direct link: 'complex' → 'ohno' (*i.e.* global expected effect if the mediation by 'delet. mut.' were 'deactivated' [Pearl, 2001, 2011]).
- 76.8% of the global effect on 'ohno' retention is owed to the indirect mediation by 'delet. mut.' (*i.e.* global expected effect if the direct link were 'deactivated' [Pearl, 2001, 2011]),

hence,

- Mediation by 'delet. mut.' is sufficient (as sole cause) to account for 82.0% of the 'complex' ⇒ 'ohno' link.
- Mediation by 'delet. mut.' is necessary (as complementary cause) to account for 76.8% of the 'complex' ⇒ 'ohno' link.

Case of Mediation of 'Dosage Bal.' ⇒ 'Ohno' Link by 'Delet. Mut.' Genes, Table S5, 1.b

We analyze the association table between the three binary categories 'dosage. bal.', 'delet. mut.' and 'ohno' genes. 'Delet. mut.' genes are genes prone to deleterious mutations, *i.e.* cancer, dominant disease, dominant negative genes and genes with autoinhibitory protein folds.

As derived in Table S4, we find,

- A significant global effect of 'dosage bal.' on 'ohno' retention: 41.9% versus 34.7% ($p = 4.2 \times 10^{-22}$, χ^2 test).
- A strong effect of 'dosage bal.' on 'delet. mut.' genes: 59.0% versus 35.2% ($p = 7.8 \times 10^{-218}$, χ^2 test).
- A significant effect of 'delet. mut.' on 'ohno' retention for 'dosage bal.' genes: 50.8% versus 41.9% ($p = 3.1 \times 10^{-18}$, χ^2 test).
- A strong effect of 'delet. mut.' on 'ohno' retention for non-'dosage bal.' genes: 48.9% versus 32.9% ($p = 2.2 \times 10^{-124}$, χ^2 test).

The Mediation analysis (Extended Experimental Procedures) then yields, $TE = 0.0903$, $DE = 0.0266$, $IE = 0.0670$ and the relative direct and indirect effects:

$DE/TE = 29.4\%$, $IE/TE = 74.1\%$;
$1-I\,E/TE = 25.9\%$, $1-DE/TE = 70.6\%$,
3.5% of a non-linear combination of direct and indirect effects.

This implies that,

- Only 25.9% of the global effect on 'ohno' retention is owed to the direct link: 'dosage bal.' → 'ohno' (*i.e.* global expected effect if the mediation by 'delet. mut.' were 'deactivated' [Pearl, 2001, 2011]).
- 70.6% of the global effect on 'ohno' retention is owed to the indirect mediation by 'delet. mut.' (*i.e.* global expected effect if the direct link were 'deactivated' [Pearl, 2001, 2011]),

hence,

- Mediation by 'delet. mut.' is sufficient (as sole cause) to account for 74.1% of the 'dosage bal.' ⇏ 'ohno' link.
- Mediation by 'delet. mut.' is necessary (as complementary cause) to account for 70.6% of the 'dosage bal.' ⇏ 'ohno' link.

Case of Mediation of 'Complex' ⇒ 'Ohno' Link by 'Delet. Mut.' Genes, Excluding 'CNV+SSD' Genes, Table S5, 1.a′

We analyze the association table between the three binary categories 'complex', 'delet. mut.' and 'ohno' genes, excluding 'CNV+SSD' genes. 'Delet. mut.' genes are genes prone to deleterious mutations, *i.e.* cancer, dominant disease, dominant negative genes and genes with autoinhibitory protein folds.

As derived in Table S4, after excluding 'CNV+SSD' genes we find,

- A small global effect of 'complex' on 'ohno' retention: 60.5% versus 55.5% ($p = 1.7 \times 10^{-5}$, χ^2 test).
- A strong effect of 'complex' on 'delet. mut.' genes: 59.0% versus 38.9% ($p = 1.3 \times 10^{-69}$, χ^2 test).
- A significant effect of 'delet. mut.' on 'ohno' retention for 'complex' genes: 74.1% versus 60.5% ($p = 5.6 \times 10^{-20}$, χ^2 test).
- A strong effect of 'delet. mut.' on 'ohno' retention for non-'complex' genes: 75.8% versus 54.1% ($p = 2.2 \times 10^{-89}$, χ^2 test).

The Mediation analysis (Extended Experimental Procedures) then yields, $TE = 0.0644$, $DE = -0.0215$, $IE = 0.0840$ and the relative direct and indirect effects:

$DE/TE = -33.4\%$, $IE/TE = 130.5\%$;
$1-IE/TE = -30.5\%$, $1-DE/TE = 133.4\%$, and
-2.9% of a non-linear combination of direct and indirect effects.

This implies that, excluding 'CNV+SSD' genes,

- -30.5% of the global effect on 'ohno' retention is owed to the direct link: 'complex' → 'ohno' (*i.e.* global expected effect if the mediation by 'delet. mut.' were 'deactivated' [Pearl, 2001, 2011]). $DE < 0$ and $TE > 0$ corresponds to a Simpson's paradox ($DE < 0$ = anticorrelation).
- 133.4% of the global effect on 'ohno' retention is owed to the indirect mediation by 'delet. mut.' (*i.e.* global expected effect if the direct link were 'deactivated' [Pearl, 2001, 2011]),

hence,

- Mediation by 'delet. mut.' is sufficient (as sole cause) to account for 130.5% of the 'complex' ⇒ 'ohno' link.
- Mediation by 'delet. mut.' is necessary (as complementary cause) to account for 133.4% of the 'complex' ⇒ 'ohno' link.

Case of Mediation of 'Dosage Bal.' ⇒ 'Ohno' Link by 'Delet. Mut.' Genes, Excluding 'CNV+SSD' Genes, Table S5, 1.b′

We analyze the association table between the three binary categories 'dosage. bal.', 'delet. mut.' and 'ohno' genes, excluding 'CNV+SSD' genes. 'Delet. mut.' genes are genes prone to deleterious mutations, *i.e.* cancer, dominant disease, dominant negative genes and genes with autoinhibitory protein folds.

As derived in Table S4, after excluding 'CNV+SSD' genes we find,

- A small global effect of 'dosage bal.' on 'ohno' retention: 61.4% versus 55.5% ($p = 1.9 \times 10^{-7}$, χ^2 test).
- A strong effect of 'dosage bal.' on 'delet. mut.' genes: 59.8% versus 38.9% ($p = 4.0 \times 10^{-79}$, χ^2 test).
- A significant effect of 'delet. mut.' on 'ohno' retention for 'dosage bal.' genes: 74.6% versus 61.4% ($p = 5.0 \times 10^{-20}$, χ^2 test).
- A strong effect of 'delet. mut.' on 'ohno' retention for non-'dosage bal.' genes: 75.6% versus 53.7% ($p = 8.6 \times 10^{-88}$, χ^2 test).

The Mediation analysis (Extended Experimental Procedures) then yields, $TE = 0.0771$, $DE = -0.0123$, $IE = 0.0887$ and the relative direct and indirect effects:

$DE/TE = -15.9\%$, $IE/TE = 115.0\%$;
$1-IE/TE = -15.0\%$, $1-DE/TE = 115.9\%$, and
-0.9% of a non-linear combination of direct and indirect effects.

This implies that, excluding 'CNV+SSD' genes,

- -15.0% of the global effect on 'ohno' retention is owed to the direct link: 'dosage bal.' → 'ohno' (*i.e.* global expected effect if the mediation by 'delet. mut.' were 'deactivated' [Pearl, 2001, 2011]). $DE < 0$ and $TE > 0$ corresponds to a Simpson's paradox ($DE < 0$ = anticorrelation).
- 115.9% of the global effect on 'ohno' retention is owed to the indirect mediation by 'delet. mut.' (*i.e.* global expected effect if the direct link were 'deactivated'[Pearl, 2001, 2011]),

hence,

- Mediation by 'delet. mut.' is sufficient (as sole cause) to account for 115.0% of the 'dosage bal.' ⇒ 'ohno' link.
- Mediation by 'delet. mut.' is necessary (as complementary cause) to account for 115.9% of the dosage. bal. ⇒ 'ohno' link.

Case of Mediation of 'Delet. Mut.' ⇒ 'Ohno' Link by 'Complex' Genes, Table S5, 2.a

We analyze the association table between the three binary categories 'delet. mut.', 'complex' and 'ohno' genes. 'Delet. mut.' genes are genes prone to deleterious mutations, *i.e.* cancer, dominant disease, dominant negative genes and genes with autoinhibitory protein folds.

As derived in Table S4, we find,

- A strong global effect of 'delet. mut.' on 'ohno' retention: 49.5% versus 34.7% ($p = 9.6 \times 10^{-155}$, χ^2 test).
- A strong effect of 'delet. mut.' on 'complex' genes: 30.7% versus 18.6% ($p = 4.0 \times 10^{-154}$, χ^2 test).
- No effect of 'complex' on 'ohno' retention for 'delet. mut.' genes: 50.1% versus 49.5% ($p = 0.6$, χ^2 test).
- No significant effect of 'complex' on 'ohno' retention for non-'delet. mut.' genes: 28.6% versus 26.6% ($p = 0.07$, χ^2 test).

The Mediation analysis (Extended Experimental Procedures) then yields, $TE = 0.2291$, $DE = 0.2276$, $IE = 0.0042$ and the relative direct and indirect effects:

$DE/TE = 99.3\%$, $IE/TE = 1.8\%$;
$1-IE/TE = 98.2\%$, $1-DE/TE = 0.7\%$, and
1.2% of a non-linear combination of direct and indirect effects.

This implies that,

- 98.2% of the global effect on 'ohno' retention is owed to the direct link: 'delet. mut.' → 'ohno' (*i.e.* global expected effect if the mediation by 'complex' were 'deactivated' [Pearl, 2001, 2011]).
- Only 0.7% of the global effect on 'ohno' retention is owed to the indirect mediation by 'complex' (*i.e.* global expected effect if the direct link were 'deactivated' [Pearl, 2001, 2011]),

hence,

- Mediation by 'complex' is sufficient (as sole cause) to account for only 1.8% of the 'delet. mut.' ⇒ 'ohno' link.
- Mediation by 'complex' is necessary (as complementary cause) to account for only 0.7% of the 'delet. mut.' ⇒ 'ohno' link.

Quantitative Partial Correlation Analysis

We outlined in this section quantitative results obtained using partial correlation analysis. The partial correlation coefficient $r_{XY \cdot Z}$ between two variables X and Y aims at ''removing'' the possible influence of a third variable Z on the global correlation, r_{XY}, between X and Y. The partial correlation coefficient $r_{XY \cdot Z}$ is defined from the pair correlation coefficients between the three variables X, Y and Z, as

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ}.r_{YZ}}{\sqrt{1 - r_{XZ}^2} \times \sqrt{1 - r_{YZ}^2}}$$

where the pair correlation coefficient r_{XY} is given by

$$r_{XY} = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i (X_i - \overline{X})^2} \times \sqrt{\sum_i (Y_i - \overline{Y})^2}}$$

and similarly for r_{XZ} and r_{YZ}. In particular, for binary properties, r_{XY} can be expressed in terms of the counts n_1, n_2, ..., n_8, for the 8 combinations of the binary variables $X = \{0,1\}$, $Y = \{0,1\}$ and $Z = \{0,1\}$ with $n_i = n_{xzy}$ in binary order (see Quantitative Mediation Analysis in Extended Experimental Procedures). Then r_{XY} reads,

$$r_{XY} = \frac{(n_1 + ... + n_8)(n_6 + n_8) - (n_5 + ... + n_8)(n_2 + n_4 + n_6 + n_8)}{\sqrt{(n_1 + ... + n_4)(n_5 + ... + n_8)(n_2 + n_4 + n_6 + n_8)(n_1 + n_3 + n_5 + n_7)}}$$

where r_{XZ} [and r_{YZ}] can be deduced by exchanging (n_2, n_6) and (n_3, n_7) [and (n_3, n_4) and (n_5, n_6) respectively] (see Quantitative Median Analysis in Extended Experimental Procedures).

Analyzing the correlations and partial correlations between the same properties of interest studied through Mediation Analysis in Extended Results, we found consistent results underlying the significance of the statistical association between the retention of ohnologs and their susceptibility to deleterious mutations (see Table S6).

In particular, Table S6 (1a & b and 1a′ & b′) shows that the weak correlation ($r_{XY} \sim 0.05$-0.07) between complex/dosage balance effects (X) and the retention of ohnologs (Y) can be largely accounted for by their susceptibility to deleterious mutations (Z), since $r_{XY.Z} \sim 0.01$-0.02 corresponds to a three- to seven-fold reduction of r_{XY} (*i.e.,* $r_{XY.Z} \sim 1/3 - 1/7\ r_{XY}$). Conversely, we found larger correlations between the retention of ohnologs (Y) and their susceptibility to deleterious mutations (X) with $r_{XY} > 0.2$ (Table S6 2a & b), while partial correlations are hardly affected by complex/dosage balance constraints (Z), *i.e.* no fold change $r_{XY.Z} \sim r_{XY}$. This trend is further enhanced for genes without SSD and CNV as $r_{XY} > 0.3$ and $r_{XY.Z} \sim r_{XY}$ in this case (Table S6 2a′ & b′).

Similarly and consistently with the results obtained with the Mediation Analysis, we also found that the statistically significant correlations between ohnolog retention (Y) and their susceptibility to deleterious mutations (X) are hardly affected by other possible intervening properties (Z) like expression levels (Table S6 3a and 3a′) or Ka/Ks values (Table S6 3b–g and 3b′–g′), as no change in partial correlations $r_{XY.Z}$ is found relative to r_{XY} in these cases (i.e., $r_{XY.Z} \sim r_{XY}$ in the Table S6). Expression levels and Ka/Ks values have been binarized in Z with the threshold 0/1 at their median values. The rationale for choosing such a binarization is to enable a direct comparison with the results from the binary Mediation Analysis in Extended Results. However, keeping continuous values to estimate correlations gave very similar results.

By contrast, the correlation between expression levels (X) and ohnolog retention (Y) was found to be statistically non-significant, *i.e.,* $r_{XY} \sim 0.003$ ($p = 0.69$), as shown in Table S6 (4a). This result is consistent with earlier observations of weak correlations between expression levels and Ka/Ks ratios for mammalian proteomes (Liao et al., 2010) unlike previous observations in yeast showing stronger correlations (Drummond et al., 2005; Drummond and Wilke, 2010). In fact, we also found that the correlations and partial correlations between Ka/Ks values (X) and ohnolog retention (Y) are weaker (Table S6 4b–g; $r_{XY} = -0.08 \pm 0.04$) than the correlations with their susceptibility to deleterious mutations (Table S6 2a & b; $r_{XY} = 0.23$ and 3b-g; $r_{XY} = 0.18 \pm 0.01$). These trends are also enhanced for genes without SSD and CNV (Table S6 4b′–g′; $r_{XY} = -0.03 \pm 0.06$ versus 2a'&b'; $r_{XY} = 0.32$ and 3b'-g'; $r_{XY} = 0.22 \pm 0.01$ above).

Hence, the results obtained through Mediation analysis (detailed in Extended Results) are entirely consistent with those obtained through partial correlation analysis, although the two approaches are not equivalent. In particular, while mediation effects require partial correlation, partial correlation does not imply mediation in general.

Indeed, partial correlation can exist in absence of mediation, for instance, if the intermediate variable Z is correlated with the outcome Y ($r_{YZ} \neq 0$) but not with the putative 'cause' X ($r_{XZ} = 0$). Then, the mediation by Z vanishes, regardless of its correlation with Y (r_{YZ}), *i.e.*,$IE = Y(x, z(x')) - Y(x, z(x)) = 0$, as Z is statistically independent of X (*i.e.,* $r_{XZ} = 0$).

By contrast, the coefficient of partial correlation $r_{XY.Z}$ remains affected by Z, *i.e.*, $r_{XY.Z} - r_{XY} \neq 0$, if $r_{XZ} = 0$ and $r_{YZ} \neq 0$, as

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ}.r_{YZ}}{\sqrt{1 - r_{XZ}^2} \times \sqrt{1 - r_{YZ}^2}} = \frac{r_{XY}}{\sqrt{1 - r_{YZ}^2}} \neq r_{XY}$$

A similar result is found if the intermediate variable Z is correlated with the putative 'cause' X ($r_{XZ} \neq 0$) but not with the outcome Y ($r_{YZ} = 0$). Hence, while mediation implies partial correlation, the opposite is not necessary the case.

Estimates of Ohnolog Retention Rates

We estimate, in this extended result, the retention rates of different classes of ohnologs from the two rounds of WGD that occurred in early vertebrates (Ohno, 1970; Putnam et al., 2008), using the observed fraction f_s of ohnologs in the human genome for gene classes, s, with increasing susceptibility to deleterious mutations, Figure 1B, in the main text.

We note, p_1 and p_2, the probabilities to retain an ohnolog pair after the first and second WGD events, respectively. Then, $(1-p_1)$ and $(1-p_2)$, are the probabilities to eliminate one of the first or second duplicates, respectively, assuming that at least one duplicate copy is always retained after either WGD event, due to functional constraints (*i.e.,* $p_1+(1-p_1)=1$ and $p_2+(1-p_2)=1$.

The resulting expansion rates for each WGD then become $r_1=2p_1+(1-p_1)=1+p_1$ and, $r_2=2p_2+(1-p_2)=1+p_2$ leading to an effective expansion for the global genome of $r_g^2=r_1r_2=(1+p_1)(1+p_2)$. Similarly, the expansion rates of ohnolog pairs can be estimated as follows: it is $2p_1(1+p_2)$, if the first ohnolog pair is retained, and $(1-p_1)2p_2$ otherwise, leading to an effective expansion rate for ohnologs of $r_0^2=2(p_1+p_2)$ after the two rounds of WGD.

Hence, noting f_s, the observed fraction of retained ohnologs in (sub)category 's' in the human genome ('s' = 'all genes', 'cancer', 'oncogene', 'autoinhibition', etc.), yields the following equation,

$$f_s=\frac{r_0^2}{r_g^2}=\frac{2(p_1+p_2)}{(1+p_1)(1+p_2)}$$

with $f_s=1$ for $P_1=1$ or $P_2=1$, as expected.

Then, assuming, for simplicity, that the retentions of ohnologs were comparable for each of the two WGDs at the onset of vertebrates, *i.e.,* $p_1=p_2=p_s$ yields the following result (other models with $p_1\neq p_2$ give qualitatively similar results),

$$f_s=\frac{r_0^2}{r_g^2}=\frac{4p_s}{(1+p_s)^2}\text{ and }P_s=\frac{2}{f_s}-1-\sqrt{\left(\frac{2}{f_s}-1\right)^2-1}$$

which leads to the plot presented in Figure 5 in the main text, corresponding to the following ohnolog retention rates for gene (sub)categories of interest:

- **'all genes'** f_s = 22–35%; p_s = 6–10% (Reference; with 'well supported' versus 'all' ohnologs)
- **'cancer or disease'** f_s = 49%; p_s = 17% (*i.e.* x1.7–3 fold change in p_s)
- '**oncogenes**' f_s = 61%; p_s = 23% (*i.e.* x2.3–4 fold change in p_s)
- '**autoinhibition**' f_s = 76%; p_s = 33% (*i.e.* x3.3–5.5 fold change in p_s)
- '**autoinhibition + cancer**' f_s = 80%; p_s = 38% (*i.e.* x3.8–6.3 fold change in p_s)
- '**autoinhibition + oncogene**' f_s = 91%; p_s = 53% (*i.e.* x5.3–8.8 fold change in p_s)
- '**autoinhibition + oncogene - SSD**' f_s = 99%; p_s = 81% (*i.e.*x8.1–13.5 fold change in p_s)

EXTENDED EXPERIMENTAL PROCEDURES

WGD Duplicated Genes or 'Ohnologs'

Ohnologs were obtained from (Makino and McLysaght, 2010) who used Ensembl release 52. These ohnolog pairs were mapped to Ensembl release 61 using BioMart. To identify potential ohnolog pairs, Makino and McLysaght (2010) compared vertebrate genomes, human, zebrafish (*D. rerio*), tetraodon (*T. nigroviridis*), stickleback (*G. aculeatus*), medaka (*O. latipes*) and fugu (*T. rubripes*) to six different non-vertebrate outgroup species, amphioxus (*B. floridae*), the ascidians (*C. intestinalis* and *C. savignyi*), sea urchin (*S. purpuratus*), fly (*D. melanogaster*) and worm (*C. elegans*). To test the robustness of our observations against the possibility of ohnolog false positive annotation, we used a confidence index for human ohnologs, defined as the number of invertebrate outgroups (1-6) supporting their ohnolog status. We defined the outgroup support of individual ohnologs as the maximum outgroup support among their ohnolog pairs, which yielded 7,110 ohnologs supported by, 1 (1,300), 2 (953), 3 (894), 4 (915), 5 (1,282) and 6 (1,766) outgroups, respectively. Human genes identified as ohnologs by more than 3 invertebrate outgroups were considered well supported ohnologs (3,963), whereas genes identified by 3 or less than 3 invertebrate outgroups were regarded as plausible (894) and more uncertain (2,253) ohnologs, respectively.

SSD Duplicated Genes

To identify genes duplicated by small scale duplication (SSD), we ran an all-against-all BLASTp (Altschul et al., 1997) using human protein sequences from Ensembl 61 and identified the best non-self hits (E-value < 10^{-7}). For all ohnolog genes, we assess whether their best non-self hit corresponds to (one of) their ohnolog partner(s). If it is the case, the ohnolog is regarded as a non-SSD ohnolog (5,743), otherwise it is considered to have been duplicated by SSD (1,367). For all nonohnolog genes, if they have a significant best hit paralog, they are considered to have experienced a SSD (9,818) or else they are counted as non-SSD genes (4,945).

Cancer Genes

We obtained candidate cancer genes from multiple databases detailed in Table S7. In the case of COSMIC database (Forbes et al., 2011), we manually curated the downloaded database to restrict cancer gene candidates to genes with at least 2 non-synonymous mutations in at least two different mutated samples (4,237 COSMIC cancer genes). Altogether, these databases led us to 6,917 cancer genes identified either through COSMIC database or text searches using different keywords in other databases (Table S8).

We further refined the above data set to identify a 'core' data set with manually-curated cancer genes and their molecular genetics status (1,139). This includes genes from CGC (420) and AOGIC (77) databases and a refined set of COSMIC genes harboring at least 5 non-synonymous mutations (784). We obtained 813 oncogenes from the core data set as annotated in CGC (327), AOGIC (77) and COSMIC core (517) databases. Oncogenes status from COSMIC core were predicted following the procedure described in Bozic et al. (2010).

Essential Genes

We used phenotype section (http://www.informatics.jax.org/phenotypes.shtml) of Mouse Genome Informatics (MGI) database version MGI_4.42 (Eppig et al., 2012). All mouse mutant alleles matching mammalian phenotype ontology terms: prenatal lethality (MP:0002080), perinatal lethality (MP:0002081), post natal lethality (MP:0002082), premature death (MP:0002083) and infertility (MP:0001924) generated by either knock-out, random gene disruption or gene trap mutagenesis were considered as essential genes. Alleles in which the above loss-of-function mutations resulted in other phenotypes (neither lethal nor infertile) were considered to be non-essential genes. Human 1-to-1 orthologs of these mouse genes were obtained using BioMart from Ensembl v-61. Discarding genes without a clear 1-to-1 mouse orthologs we obtained 2,729 essential and 3,227 non-essential mouse ortholog genes.

Genes in Complexes and Permanent Complexes

We obtained protein complexes from Human Protein Reference Database (HPRD) (Keshava Prasad et al., 2009) and CORUM database (Ruepp et al., 2010). As no distinction between transient and permanent complexes is made in these two databases, we obtained a manually curated data set of permanent complexes from Zanivan et al. (2007). For all these three data sets, genes were mapped to Ensembl IDs using BioMart, NCBI gene IDs and gene names. Complex partners without a mapping were discarded. We obtained in total 2,721 and 2,500 protein complex genes from HPRD and CORUM respectively, along with 239 permanent protein complex genes from Zanivan et al. (2007). Genes are considered as ''complex'' genes if they are included in any one of the three data sets (total 3,814 complex genes).

Gene Expression Level

Gene expression values were downloaded from BioGPS (Wu et al., 2009), which provides GC content adjusted-robust multi-array (GC-RMA) normalized expression values (U133A and GNF1H microarrays) for 78 healthy human tissues and cell types (Su et al., 2004). Affimetrix tags were mapped to Ensembl gene IDs using BioMart and annotation provided by BioGPS. Tags which are known to bind to multiple genes were removed from the final analysis, and expression levels from different tags for the same gene were averaged. The expression levels correspond to the median values of expression across 78 tissues/cell types (13,372 genes in total).

Quantitative Mediation Analysis

We have performed a Mediation analysis following the approach of (Pearl, 2001, 2011). The Mediation framework, developed in the context of causal inference analysis, (Pearl, 2009) aims at uncovering, beyond statistical correlations, *causal* pathways along which changes in multivariate properties are transmitted from a cause, X, to an effect, Y. More specifically, a Mediation analysis assesses the importance of a mediator, M, in transmitting the indirect effect of X on the response $Y \equiv Y(x, m(x))$, see the Mediation diagram (Figure 3).

The direct and M - mediated effects on Y to a change $x \rightarrow x'$ are then quantified using counterfactual expressions, which formally enable to decouple the direct (x_1) and indirect (x_2) conditions on X, seen by the outcome Y, i.e., $Y(x_1, m(x_2))$. Hence, the direct effect of X on Y ($DE_{xx'}$) can be defined as the hypothetical change that Y would experience, if x could be changed to x' while keeping the mediator M to its original value $m(x)$. Likewise, the indirect effect of X on Y, $IE_{xx'}$, corresponds to the hypothetical change that Y would experience, if the mediator M could be changed to $m(x')$, while keeping the direct influence of X on Y to its original value x. This yields the following counterfactual definitions (Pearl, 2001) for the direct ($DE_{xx'}$), indirect ($IE_{xx'}$) and total ($TE_{xx'}$) effects on Y to a change $x \rightarrow x'$,

$$DE_{xx'} = Y(x', m(x)) - Y(x, m(x))$$

$$IE_{xx'} = Y(x, m(x')) - Y(x, m(x))$$

$$TE_{xx'} = Y(x', m(x')) - Y(x, m(x))$$

where $DE_{xx'}$, $IE_{xx'}$ and $TE_{xx'}$ can be evaluated within the framework of Bayesian statistics (Pearl, 2001, 2009) using the counterfactual expressions, $Y(x_1, m(x_2)) = \sum_m E(Y|x_1, m)P(M|x_2)$,

$$DE_{xx'} = \sum_m [E(Y|x', m) - E(Y|x, m)]P(M|x)$$

$$IE_{xx'} = \sum_m E(Y|x, m)[P(M|x') - P(M|x)]$$

$$TE_{xx'} = E(Y|x') - E(Y|x)$$

Note, in particular, that total effect, $TE_{xx'}$, is not, in general, the simple sum of direct and indirect effects. Indeed, $TE_{xx'} = DE_{xx'} - Y(x', m(x)) + Y(x', m(x')) \neq DE_{xx'} + IE_{xx'}$ for non-linear systems. Instead, the total effect $TE_{xx'}$ is related to the indirect effect of the *reversed* transition changing, x' into x i.e., $IE_{x'x} = Y(x', m(x)) - Y(x', m(x'))$, that is,

$$TE_{xx'} = DE_{xx'} - IE_{x'x}$$

Yet, although, $DE + IE \neq TE$ owing to possible non-linear couplings between direct and indirect causal effects (Pearl, 2001), the Mediation analysis can be interpreted in terms of proportions of the total effect, *DE*/*TE* and *IE*/*TE*, as well as necessary and sufficient contributions from the direct and indirect causal pathways (Pearl, 2011). Namely, for a total effect *TE* of *X* on *Y*, the Mediation Analysis assesses that,

- the direct effect $X \rightarrow Y$ is *sufficient* as a sole cause to account for a proportion *Dir.* = *DE*/*TE* of the total effect, while,
- the indirect effect $X \rightarrow M \rightarrow Y$ is *necessary* as complementary cause to account for a proportion *1-DE*/*TE*,

and likewise,

- the indirect effect $X \rightarrow M \rightarrow Y$ is *sufficient* as a sole cause to account for a proportion *Ind.* = *IE*/*TE* of the total effect, while,
- the direct effect $X \rightarrow Y$ is *necessary* as complementary cause to account for a proportion *1-IE*/*TE* of the total effect.

We have applied the Mediation analysis to genomic data using binary categories (Pearl, 2011) $(X, M, Y = 0/1; DE, IE, TE \in [-1, 1])$ to discriminate between direct and indirect effects of deleterious mutations (*X or M*) and dosage balance constraints (*M or X*) on the biased retention of human ohnologs (*Y*). In the case of binary categories, Pearl's Mediation Formulae (Pearl, 2011) for direct effect (*DE*), indirect effect (*IE*) and total effect (*TE*) simply read,

$$DE = (g_{10} - g_{00})(1 - h_0) + (g_{11} - g_{01})h_0$$

$$IE = (h_1 - h_0)(g_{01} - g_{00})$$

$$TE = f_1 - f_0 = g_{11}h_1 + g_{10}(1 - h_1) - [g_{01}h_0 + g_{00}(1 - h_0)]$$

where f_0, f_1, g_{00}, g_{01}, g_{10}, g_{11}, h_0 and h_1 are obtained from the association table of the binary variables *X*, *M* and *Y* (Table S4).

SUPPLEMENTAL REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. (2005). Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. USA. *4;102(40)*, 14338-43.

Drummond, D.A. and Wilke, C.O. (2010). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 25;134(2), 341–52.

Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., and Richardson, J.E.; Mouse Genome Database Group (2012). The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res. *40*(Database issue), D881–D886.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Liao, B.Y., Weng, M.P., and Zhang, J. (2010). Impact of extracellularity on the evolutionary rate of mammalian proteins. Genome Biol. Evol. *2*, 39–43.

Pearl, J. (2009). Causality: models, reasoning and inference (New York: Cambridge University Press).

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA *101*, 6062–6067.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics *8*, 77–80.

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C.L., Haase, J., Janes, J., Huss, J.W., 3rd, and Su, A.I. (2009). BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol. *10*, R130.

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. *17*, 32–43.

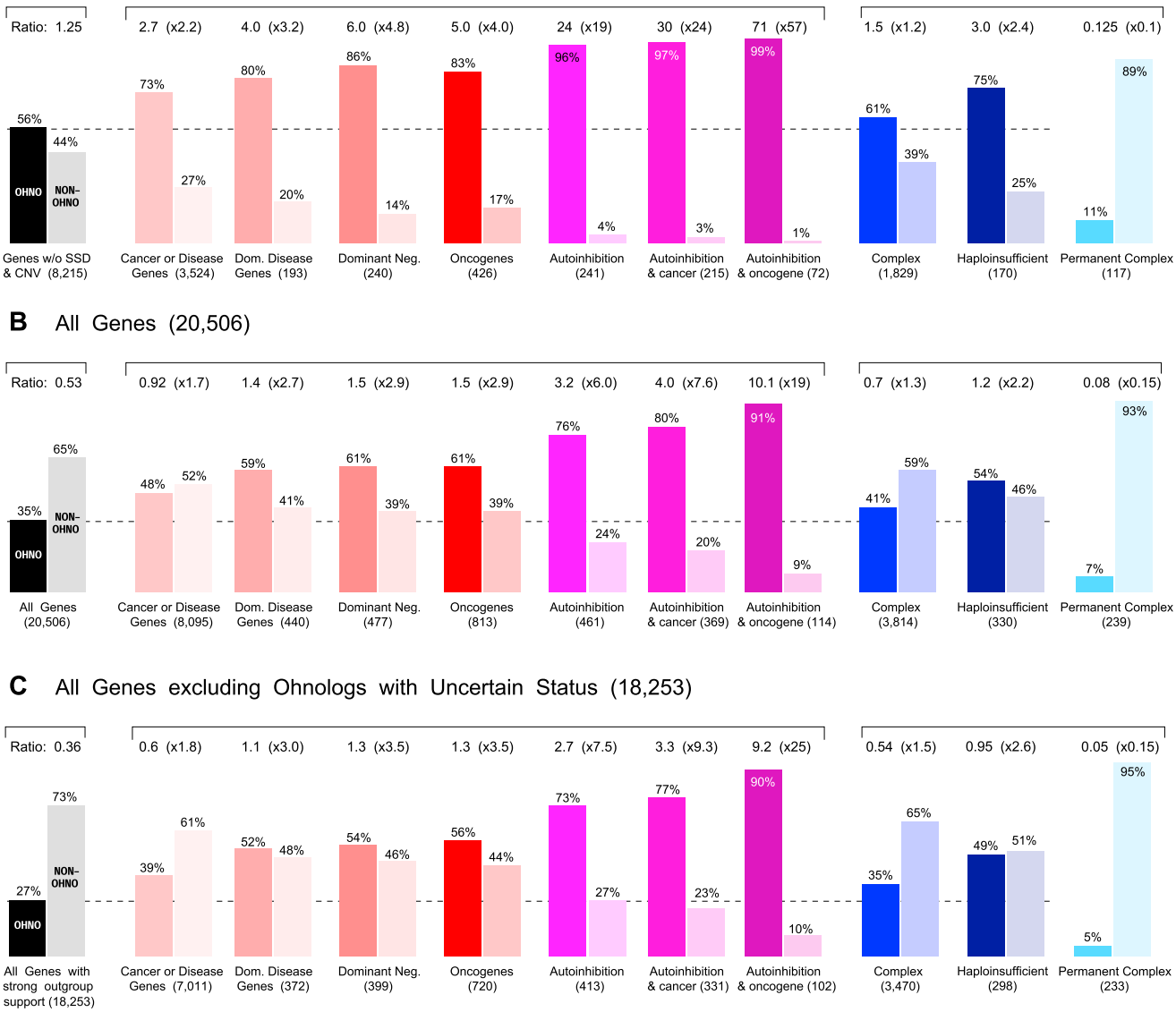

Figure S1. Prevalence of Retained Ohnologs in Gene Classes Prone to Deleterious Mutations or Dosage Balance Constraints, Related to Figure 1

(A and B) Prevalence of retained ohnologs in the human genomes for (**A**) all genes without SSD & CNV duplicates (8,215), (B) all human protein-coding genes (20,506). Note that the statistics, based on the entire human genome (B) are strongly enhanced on the 40% of human genes (8,215) without SSD and CNV duplicates (A).

(C) To further test the robustness of these observations against the possibility of ohnolog false positive annotation, we also display the prevalence of retained ohnologs in the human genomes for all genes excluding ohnologs with uncertain status (18,253). The confidence index for human ohnologs is defined as the number of invertebrate outgroups (1-6) supporting their ohnolog status (Extended Experimental Procedures). Briefly, human genes identified as ohnologs by more than half of the invertebrate outgroups (>3) were considered well supported ohnologs, whereas genes identified as possible ohnologs by less than half of the invertebrate outgroups (<3) were regarded as more uncertain. Discarding the 11% of human genes with somewhat uncertain ohnolog status (C) confirmed that the biased retention of actual ohnologs is strongly associated with their susceptibility to deleterious mutations.

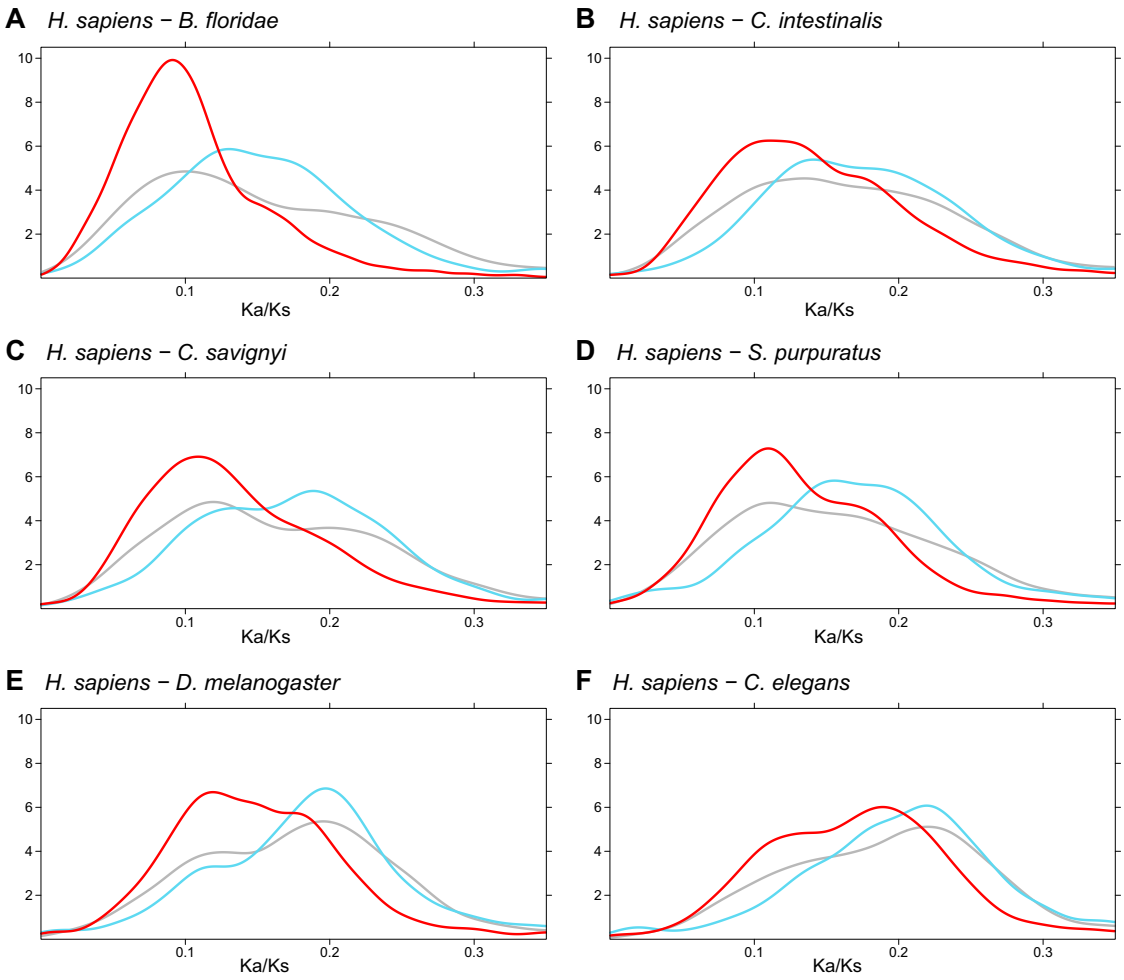

Figure S2. Comparison of Ka/Ks Distributions for *Human– Invertebrate* Ortholog Pairs, Related to Figure 2

(A–F) (A) with *B. floridae*, (B) *C. intestinalis*, (C) *C. savignyi*, (D) *S. purpuratus*, (E) *D. melanogaster* and (F) *C. elegans*. Human-invertebrate ortholog pairs involving human ohnologs with a well-supported status (Extended Results) have statistically lower Ka/Ks ratios (red) than pairs involving human orthologs with a more uncertain ohnolog status (blue) or than pairs involving human nonohnologs (gray). While these Ka/Ks distributions present significant overlaps, the lower averaged Ka/Ks ratios for well supported human ohnologs (red) are highly significant, as demonstrated using Mann-Whitney-Wilcoxon non-parametric test. By contrast, the differences between uncertain ohnolog (blue) and nonohnolog (gray) Ka/Ks distributions are much less significant, as expected (see Table S3).

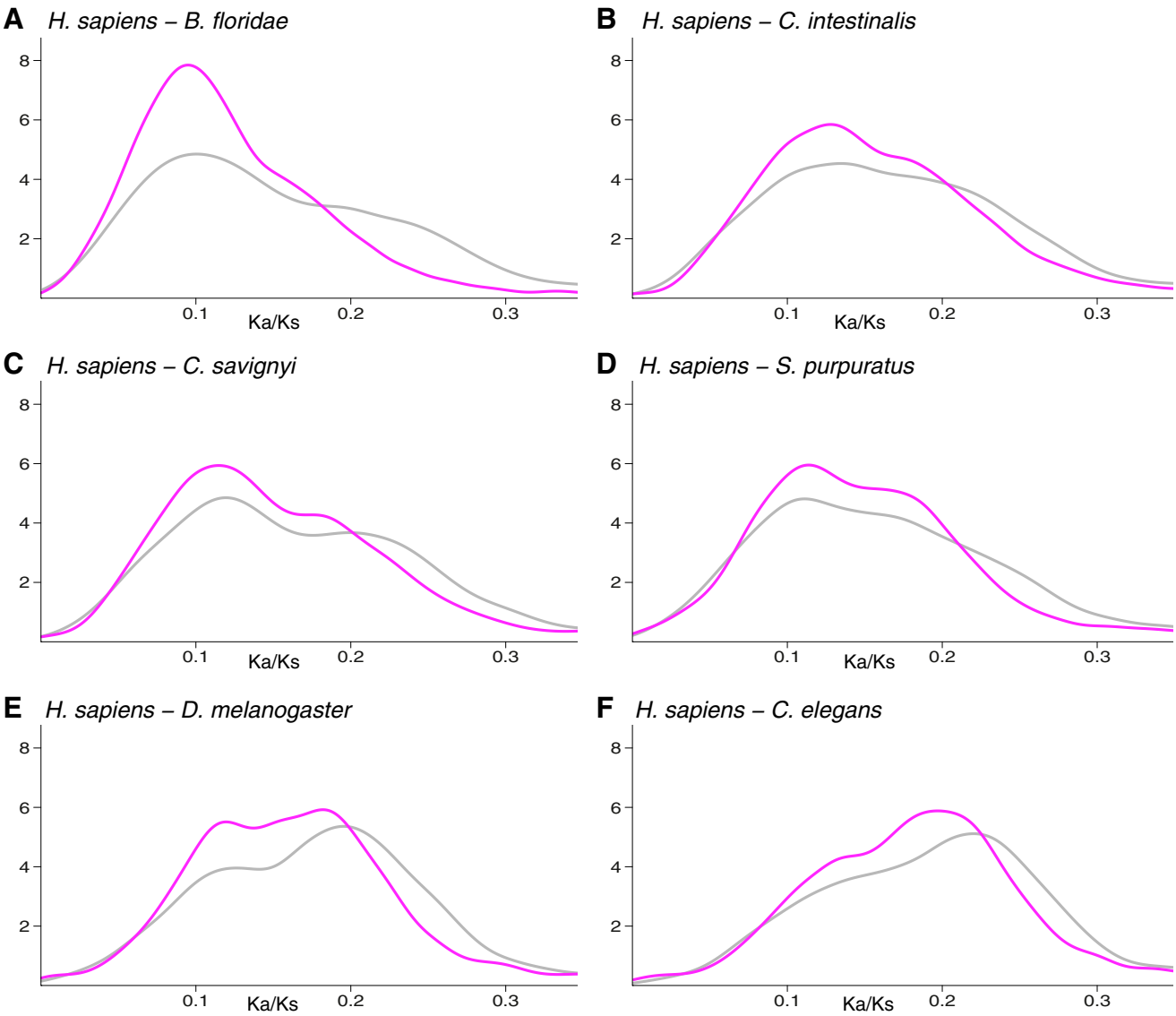

Figure S3. Comparison of Ka/Ks Distributions for *Human–Invertebrate* Ortholog Pairs, Related to Figure 2

(A–F) (A) with *B. floridae*, (B) *C. intestinalis*, (C) *C. savignyi*, (D) *S. purpuratus*, (E) *D. melanogaster* and (F) *C. elegans*. Human-invertebrate ortholog pairs involving human ohnologs including all outgroup supports (from 1 to 6; Extended Results) have statistically lower Ka/Ks ratios (magenta) than pairs involving human nonohnologs (gray). While these Ka/Ks distributions present significant overlaps, the lower averaged Ka/Ks ratios for human ohnologs (magenta) are still significant, as demonstrated using Mann- Whitney-Wilcoxon non-parametric test (see Table S3).