Adaptive evolution of young gene duplicates in mammals

Mira V. Han, Jeffery P. Demuth, Casey L. McGrath, et al.

Genome Res. 2009 19: 859-867
Access the most recent version at doi:10.1101/gr.085951.108

Letter

Adaptive evolution of young gene duplicates in mammals

Mira V. Han,[1] Jeffery P. Demuth,[1,2,3] Casey L. McGrath,[2] Claudio Casola,[1,2] and Matthew W. Hahn[1,2,4]

[1]*School of Informatics, Indiana University, Bloomington, Indiana 47405, USA;* [2]*Department of Biology, Indiana University, Bloomington, Indiana 47405, USA*

Duplicate genes act as a source of genetic material from which new functions arise. They exist in large numbers in every sequenced eukaryotic genome and may be responsible for many differences in phenotypes between species. However, recent work searching for the targets of positive selection in humans has largely ignored duplicated genes due to complications in orthology assignment. Here we find that a high proportion of young gene duplicates in the human, macaque, mouse, and rat genomes have experienced adaptive natural selection. Approximately 10% of all lineage-specific duplicates show evidence for positive selection on their protein sequences, larger than any reported amount of selection among single-copy genes in these lineages using similar methods. We also find that newly duplicated genes that have been transposed to new chromosomal locations are significantly more likely to have undergone positive selection than the ancestral copy. Human-specific duplicates evolving under adaptive natural selection include a surprising number of genes involved in neuronal and cognitive functions. Our results imply that genome scans for selection that ignore duplicated loci are missing a large fraction of all adaptive substitutions. The results are also in agreement with the classical model of evolution by gene duplication, supporting a common role for neofunctionalization in the long-term maintenance of gene duplicates.

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

Recently duplicated loci suffer one of two long-term fates: maintenance or loss (Ohno 1970; Walsh 1995). While pseudogenization is the more likely fate of recently duplicated genes, many models have been proposed that could lead to the long-term maintenance of multiple paralogs (Spofford 1969; Ohno 1970; Dykhuizen and Hartl 1980; Hughes 1994; Force et al. 1999; Stoltzfus 1999). The maintenance of duplicates can be a by-product of neutral evolution (Dykhuizen and Hartl 1980; Force et al. 1999; Stoltzfus 1999), or there can be adaptive substitutions either during (Spofford 1969; Ohno 1970) or after (Ohno 1970; Hughes 1994) the fixation of the duplicated locus. Previous studies have found signatures of adaptive evolution among individual duplicated genes, suggesting that selection for new functions ("neofunctionalization") is the mechanism acting to retain new paralogs (Zhang et al. 1998; Merritt and Quattro 2001; Betran and Long 2003; Moore and Purugganan 2003; Rodriguez-Trelles et al. 2003; Thornton and Long 2005). While these studies support the neofunctionalization model, the genome-wide proportion of all duplicates fixed and maintained by natural selection is still not known (Hahn 2009).

Gene duplication supplies the raw material necessary to evolve novel functions and is therefore a source of adaptive change. Previous studies in mammals have searched for positively selected genes in the hope of identifying the nucleotide substitutions that underlie phenotypic divergence between species, but these genome-wide scans have intentionally ignored duplicated loci in order to avoid problems in the assignment of orthology (Clark et al. 2003; Nielsen et al. 2005; Bakewell et al. 2007; Kosiol et al. 2008). This oversight is unfortunate, as many cases of adaptive evolution of individual gene duplicates are known (see above). These previous results from studies of individual gene families imply that by neglecting duplicated loci we are missing a substantial fraction of the adaptive events that differentiate species. Ignoring patterns of selection on duplicated loci in humans may be particularly shortsighted, as the rate of gene duplication has increased in our recent past (She et al. 2006; Hahn et al. 2007).

Here we study the evolutionary forces acting on recent gene duplications in four mammalian genomes: human, macaque, rat, and mouse. By focusing on young duplicates we hope to capture the mechanisms responsible for the initial maintenance of new genes. We use codon-based likelihood models implemented in the PAML package (Yang 2007) to test for adaptive evolution shortly after the duplication. To ensure the accuracy of our results we also use non-likelihood-based methods and conduct a number of checks on our results. We find that a larger fraction of young duplicates have experienced positive selection than have a comparable set of single-copy orthologs. We also observe that, among duplicates, new paralogs that have moved to a different genomic location are more likely to experience adaptive evolution than are the copies in the original location.

Results

To restrict our study to recently duplicated genes, we obtained lineage-specific duplicates from the human, macaque, mouse, and rat genomes (Methods). Although there is little statistical power to detect positive selection in recent duplicates because of the small number of interparalog substitutions, we expect these genes to reveal the most about the selective forces that maintain duplicates.

[3]Present address: Department of Biology, University of Texas, Arlington, Texas 76019, USA.
[4]Corresponding author.
E-mail mwh@indiana.edu; fax (812) 855-6705.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.085951.108.

Briefly, we constructed gene trees for every gene family in the four lineages together and extracted all duplication events specific to each species using gene tree–species tree reconciliation (Chen et al. 2000). To avoid cases where multiple lineage-specific losses might result in incorrect inferences of lineage-specific duplications, we then partitioned the gene trees into subtrees that only include duplication events that occurred subsequent to the most recent speciation event. A subtree is defined as the phylogeny of a set of paralogs for which the nucleotide divergence (d_S) between any pair is less than twice the time since the last speciation (see Methods). Our methods identified 893, 436, 1723, and 1428 lineage-specific duplications in human, macaque, mouse, and rat, respectively (Fig. 1; Table 1).

We chose to analyze lineage-specific duplicates because these offer independent instances of evolution by gene duplication. Incorrectly identifying duplicates as lineage-specific using the tree-based and d_S-based methods described above will not affect our inferences of adaptive natural selection, though it may affect our estimates of the timing of selection. Therefore, to further validate the duplications we found, we checked the duplicated genes against experimental data. Low-quality sequencing and assembly is known to affect duplication content by collapsing the duplicates into one gene. This can result in two types of errors, underestimation of lineage-specific genes in a low-quality genome, or overestimation of lineage-specific genes in species sister to low-quality genomes. Since the macaque genome has the lowest quality among the species considered here, we focused on estimates of macaque-specific duplicates and the human-specific duplicates inferred using macaque as the outgroup. First, we confirmed that the duplicates we have identified are supported by synteny in the flanking regions. While 14% of the macaque duplicates were on problematic contigs—short contigs that are not included in the final assembly—and thus could not be confirmed by alignment of the flanking region, 100% of the human duplicates had unambiguous alignment in the flanking regions. Second, we used two different data sets that identified macaque duplicates using experimental methods to confirm our inferences of lineage-specific duplications. Using array comparative genomic hybridization (aCGH) data (Gibbs et al. 2007), we found that seven out of 355 gene families identified as having human-specific duplicates had a macaque duplication confirmed by aCGH but missing in the assembly. These cases may therefore indicate either duplicates incorrectly inferred as being human-specific when in fact they predate the human-macaque split, or cases of parallel lineage-specific duplications in the two lineages. In general, however, there is good agreement between aCGH experiments and computational predictions of duplication (Hahn et al. 2007). Using sequence read-depth data from the shotgun assembly of the macaque genome to identify unassembled duplicates (i.e., the "WSSD" method of Bailey et al. 2002), we found 21 out of 355 families with human duplicates have a macaque gene with a missing duplication discovered by WSSD. Again, these may represent either human duplicates that are not truly lineage-specific or parallel duplication events. In any event, both results show that the majority of our duplicates are not incorrectly identified as lineage specific. More importantly, either type of error produced by the collapsing of genes should not have a directional effect on the detection of positive selection; we show below that having a collapsed gene among the duplicates does not lead to more detection of positive selection.

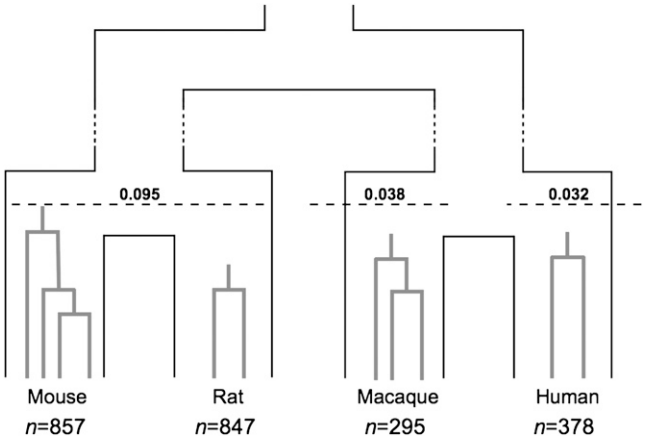

Figure 1. Experimental design. Lineage-specific duplicates were extracted from each of the four lineages shown, using the d_S values indicated. The number of lineage-specific subtrees containing paralogs differing by at least two substitutions is given for each genome.

Table 1. Number of lineage-specific duplications

		Human	Macaque	Mouse	Rat	Total
Reconciliation	Genes	2262	2463	3371	3273	11,369
	Trees	748	842	1084	1072	3746
	Duplications	1514	1621	2287	2201	7623
d_S filtering	Genes	1417	782	2701	2371	7271
	Trees	524	346	978	943	2791
	Duplications	893	436	1723	1428	4480
PAML	Genes	1051	678	2427	2159	6315
	Trees	378	295	857	847	2377
	Duplications	673	383	1570	1312	3938

Reconciliation denotes duplicates identified by gene–species tree reconciliation. d_S filtering denotes duplicates that have distances consistent with the age of the species. PAML denotes duplicates with enough statistical power that were used in the likelihood analyses.

For the 2377 total subtrees containing lineage-specific duplicates that also had more than two nucleotide differences between paralogs (Table 1; cf. Nielsen et al. 2005), we tested for evidence of positive natural selection by estimating the nonsynonymous/synonymous substitution rate ratio (d_N/d_S) distribution among amino acid sites. Specifically, we used the codon-substitution models implemented in PAML (Yang 2007) to estimate the likelihoods of each subtree when d_N/d_S is allowed to exceed 1 (M2a) versus a model where d_N/d_S is constrained to ≤ 1 (M1a). As we show below using simulations—and as has been shown before (Wong et al. 2004)—comparing twice the difference in log-likelihoods between models to a χ^2 distribution with two degrees of freedom is a conservative test for positive selection.

In each of the four lineages at least 8.8% of subtrees show significant ($P < 0.05$) evidence for positive selection (Fig. 2). Furthermore, since the divergence of humans from macaque, 13.8% of all subtrees containing human-specific duplications appear to have evolved under positive selection. All subtrees across all four genomes with significant evidence for positive selection also show $d_N/d_S > 2$ on a subset of codons. After excluding the subtrees containing misannotated genes—genes that are removed in the more recent version of Ensembl—there were 18 human subtrees significant at a false discovery rate (FDR) <0.10 (Table 2). To directly compare our results to the proportion of single-copy genes experiencing positive selection, we carried out the same analyses

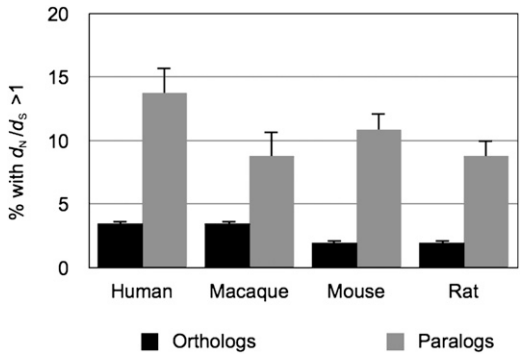

Figure 2. Positive selection on gene duplicates. The proportion of lineage-specific subtrees and orthologs with evidence for positive selection using a likelihood ratio test between models M1a and M2a ($P < 0.05$). Standard errors are shown.

on 10,376 single-copy orthologs between human and macaque and 8631 orthologs between mouse and rat. The proportion of single-copy orthologs where we detect positive selection ($d_N/d_S > 1$) is consistent with previously published results for these mammals, and is significantly lower than the proportion of duplicates under positive selection (Fig. 2). Because these orthologs are as old as or older than the lineage-specific paralogs considered here, we actually expect that there is more statistical power to detect selection in the orthologs because of the higher levels of interortholog divergence. No significant differences have been found between the functions of single-copy genes and genes duplicated among the mammals (Demuth et al. 2006), so this cannot explain our results either. We next consider a number of other possible conflating factors.

We conducted a number of checks on our data and analyses to eliminate potential sources of error or biases in our results. (1) To ensure that there were no false positives due to misalignment, we used stringent criteria to mask regions where alignments were of poor quality (Methods). (2) To rule out misannotations or pseudogenes, we checked a more recent version of the Ensembl database (v48) and removed any duplicates that had been modified since the version our data are based on (v41); there was no qualitative difference in results (Supplemental Fig. 1). (3) Importantly, the likelihood ratio test implemented here does not give false evidence for positive selection even if pseudogenes are inadvertently included (Wong et al. 2004). To demonstrate this we simulated 1000 data sets at five different values of d_S with $d_N/d_S = 1.0$. The largest fraction of significant genes in these simulations

Table 2. Hominid-specific gene families under positive selection

Family ID	Description	*P*-value	FDR
ENSF00976	Neuroblastoma breakpoint family NBPF (NBPF14, KIAA1245)	4.24×10^{-11}	1.59×10^{-8}
ENSF00597	Golgin subfamily A (ENSP00000289798)	1.77×10^{-6}	2.22×10^{-4}
ENSF01738	FAM75A (FAM75A6)	4.19×10^{-5}	0.0036
ENSF06756	Similar to nuclear pore membrane protein 121	5.33×10^{-5}	0.0036
ENSF02009	RNA exonuclease 1 homolog	5.75×10^{-5}	0.0036
ENSF04148	Williams Beuren syndrome region 19	8.51×10^{-5}	0.0046
ENSF00900	Ankyrin repeat domain containing (ANKRD36)	0.0001	0.0059
ENSF11001	Uncharacterized (ENSP00000369576)	0.0003	0.0127
ENSF01499	Serum Amyloid A (SAA1)	0.0009	0.0304
ENSF00664	Ankyrin repeat domain containing (ENSP00000326572)	0.0009	0.0304
ENSF00664	Ankyrin repeat domain containing (ENSP00000340206)	0.0019	0.0604
ENSF01546	Ubiquitin carboxyl-terminal hydrolase 17	0.0032	0.0798
ENSF07469	Ubiquitin-specific peptidase (USP41)	0.0032	0.0798
ENSF04738	Morpheus (K0220)	0.0034	0.0808
ENSF04835	FAM90A	0.0037	0.0822
ENSF01431	Nuclear receptor corepressor 1 (C20orf191)	0.0042	0.0880
ENSF00061	Serpin (SERPINB4)	0.0046	0.0909
ENSF00449	Haptoglobin	0.0053	0.0998
ENSF01302	RAN binding	0.0063	0.1083
ENSF09825	Chondrosarcoma associated (CSAG1)	0.0122	0.1812
ENSF02496	Chaperonin containing T-complex protein (CCT8L1)	0.0125	0.1812
ENSF02304	Mitochondrial peptide chain release factor (ENSP00000318184)	0.0181	0.2258
ENSF00664	Ankyrin repeat domain containing	0.0186	0.2258
ENSF00597	Golgin subfamily A (ENSP00000327024)	0.0186	0.2258
ENFS00392	PRAME family member (PRAMEF11)	0.0214	0.2302
ENFS12591	Uncharacterized	0.0215	0.2302
ENSF01435	Methyltransferase (METTL2B)	0.0219	0.2302
ENSF05738	AMAC homologs (AMAC1,AMAC1L1)	0.0227	0.2308
ENSF00096	UDP glucuronosyltransferase precursor	0.0259	0.2413
ENSF09690	Uncharacterized	0.0273	0.2426
ENSF00864	Sialic acid binding Ig lectin precursor	0.0284	0.2426
ENSF03375	G antigen	0.0297	0.2426
ENSF00011	Olfactory receptor	0.0307	0.2426
ENSF07466	Methyl CpG binding domain	0.0310	0.2426
ENSF00203	Olfactory receptor OR1 (OR2T33)	0.0332	0.2548
ENSF01533	RIM binding, PBR associated	0.0436	0.3241
ENSF08345	Variable charge X-linked	0.0440	0.3241
ENFS00221	Alpha amylase precursor (AMY2B)	0.0482	0.3488

Subtrees with $d_N/d_S > 1$ by M1a/M2a. Family IDs and descriptions are from Ensembl v41. Proteins identified as under positive selection by the branch-site test with *P*-value < 0.05 are included in parentheses.

was 3.3%, and mostly much lower than this, indicating that our tests are indeed conservative (Supplemental Fig. 2). (4) We checked for convergence of the parameter estimates in PAML by carrying out at least two runs for each subtree: If the difference in likelihoods between runs was >0.001, the subtrees were discarded. (5) To assess possible biases in the likelihood method used or violations of the assumptions of the likelihood model in our data, we estimated d_N/d_S using the simpler Nei–Gojobori method; consistent with the likelihood method, between 6.7% and 16.8% of duplicates had $d_N/d_S > 1$ using this method (Supplemental Fig. 3). (6) To address the concern that collapsed duplicates could result in spurious detection of positive selection, we compared our positively selected subtrees against the collapsed duplicates identified by the WSSD method using read-depth data (Bailey et al. 2002). Of the 42 genes we called under positive selection in macaque, only one showed increased read-depth, indicative of a collapsed duplicate. Removing both these families and those found on unassembled contigs or identified by aCGH to have additional copies in macaque (99 macaque families and 46 related human families) did not lower the proportion of positively selected duplicates in these lineages (11 macaque and four human subtrees with positive selection were removed). (7) Another possible explanation for the high proportion of positive selection detected among duplicates is that tests using paralogs can have more power to detect selection than using pairs of orthologs, simply because some of the subtrees have more than two genes. We therefore restricted our analyses to only cases with two paralogs; the proportion of genes with $d_N/d_S > 1$ was slightly lower but still significantly higher compared to the proportion of ortholog pairs (Supplemental Fig. 4). (8) Finally, because gene conversion may result in false positives when using the current likelihood ratio test (Anisimova et al. 2003; Casola and Hahn 2009) we searched for gene conversion among all of the paralogs and removed those with evidence for conversion from our analyses. Again, we still had between 5.6% and 9.7% of duplicates with evidence of positive selection in all four species (Supplemental Fig. 5). After verifying our results in these ways, we are confident that the abundance of adaptive natural selection found in young duplicates is a real biological pattern.

For the subtrees of lineage-specific duplicates that had a signal of positive selection, we were interested in which particular gene was undergoing adaptive evolution. We can identify these genes using the branch-site model implemented in PAML ("test 2") (Zhang et al. 2005). When an outgroup sequence was available we added it to the unrooted phylogeny of each subtree, and all the branches in the tree were tested except for the branch leading to the outgroup gene. Of the 245 total subtrees identified as evolving under positive selection using the M1a/M2a comparison, 61% also had at least one branch with $d_N/d_S > 1$ at $P < 0.05$ and 70% at $P < 0.10$ using test 2. Across all four genomes, 16.1%–21.1% of the branches tested show a significant signature of positive selection using test 2 ($P < 0.05$ or 10.2%–18.4% with FDR < 0.10; Supplemental Fig. 6). Along the human lineage we found significant evidence for positive selection on 37 out of the 230 total branches considered ($P < 0.05$, or 25 with FDR < 0.10).

Thus far we have considered genes that have duplicated along the human lineage since the split from macaque. When we further restricted our analyses to even more recently derived human-specific duplicates—those arising since the split with chimpanzee—we found 434 new genes that have appeared in the last five million years. Previous work estimated that 3% of both genomic DNA (Cheng et al. 2005) and genes (Demuth et al. 2006; Hahn et al. 2007) have duplicated along the human lineage since the split with chimpanzee. Our current findings confirm these results and should be considered a high-confidence set of the new duplicates to appear in the human genome. Testing for adaptive natural selection among the 125 subtrees containing these duplicates with enough power to detect positive selection, we identified nine human-specific subtrees with $d_N/d_S > 1$ at $P < 0.05$, three of them with FDR < 0.10. Using the branch-site test on the nine significant families, we identified four branches that were positively selected with FDR < 0.10 (Table 3).

Identifying the branch responsible for the signal of positive selection also lets us explore the fate of each duplicate and, consequently, the effect of genomic context on the diversification of gene copies. We used the locations of single-copy orthologs in other species to polarize lineage-specific duplicates as either parent copies (in the same location) or daughter copies (in a different location) in each genome (Han and Hahn 2009). Due to ambiguities in the assignment of tandem duplicates as either parent or daughter, we removed all such pairs from the analysis. Across the four species, we could unambiguously identify the parent copy for a total of 641 families by finding extended synteny with the outgroup gene. We restricted our analyses to those cases with a one-to-one relationship of parent and daughter (Supplemental Table 1). Among these genes, 66 had a d_N/d_S significantly greater than 1 by the branch-site test ($P < 0.05$). When we looked at the relationship between the two variables—the class of the gene

Table 3. Human-specific duplicates under positive selection

M1a/M2a[a]			
Family ID	**Description**	***P*-value**	**FDR**
ENSF01738	Similar to FAM75A1 (also known as C9orf36)	4.19×10^{-5}	0.0051
ENSF04738	Uncharacterized (NPIPL3 [also known as A8MRT5])	0.0003	0.0177
ENSF04835	FAM90	0.0013	0.0534
ENSF09825	Chondrosarcoma associated (CSAG1)	0.0122	0.3742
ENSF00841	Tripartite motif containing (TRIM64 [also known as AP004607])	0.0175	0.3886
ENSF08345	Variable charge X-linked (VCX)	0.0221	0.3886
ENSF00900	Prostate, ovary, testis-expressed (POTEH [also known as A26C3])	0.0409	0.5958
ENSF01533	RIM binding, PBR associated, RIMBP3B, RIMBP3C	0.0436	0.5958
Branch-site test[b]			
Protein ID	Description	*P*-value	FDR
A8MRT5	Uncharacterized protein	3.68×10^{-5}	0.0018
CSAG1	Chondrosarcoma associated gene 1	0.0012	0.029
VCX	Variable charge X-linked protein 1	0.003	0.0481
A26C3	ANKRD26-like family C, member 3	0.0047	0.0563
AP004607	Similar to tripartite motif protein 39	0.0394	0.3788

[a]Subtrees with $d_N/d_S > 1$ by M1a/M2a. Ensembl v41 gene families IDs are provided. Proteins identified as under positive selection by the branch-site test with *P*-value < 0.05 are included in parentheses.
[b]Branches within the subtrees in (a) with $d_N/d_S > 1$ by the branch-site test.

(parent vs. daughter) and whether the gene experienced positive selection—there was a clear association ($P = 3.7 \times 10^{-7}$, Fisher's exact test; Fig. 3). Only 13 out of 66 positively selected genes were the parent copy, while 53 out of 66 were the daughter copy located in a new genomic region. This pattern was consistent across all lineages (Fig. 3).

To examine genomic features that could affect the rate of evolution between parents and daughters we compared GC content, recombination rate, and density of conserved elements (as defined by phastCons) (Siepel et al. 2005) in the regions surrounding the pairs of genes. We found that GC content is significantly higher in the regions flanking all parent genes in every species examined ($P < 0.01$; Supplemental Table 2), though it is only significantly higher in mouse and rat when restricting the comparison to pairs in which the daughter is under positive selection. This result is consistent with a previous study that found lower levels of histone modifications in daughter regions (Zheng 2008). Unlike in previous studies of yeast (Zhang and Kishino 2004), we did not find any differences in the recombination rate between the positively selected daughters and their parents. There were also no differences in the number or density of conserved elements in the regions flanking the two sets of genes (Supplemental Table 2).

Discussion

Positive selection on young duplicates

There have been a number of previous studies on gene duplicates in mammals that have measured d_N/d_S between paralogs, even if they have not explicitly tested for $d_N/d_S > 1$ (Lynch and Conery 2000; Kondrashov et al. 2002; Zhang et al. 2003). Though the pattern found here is obvious in retrospect when examining the results from previous studies, the excess of young duplicates with $d_N/d_S > 1$ was largely overlooked because the overwhelming majority of older paralogs ($d_S > 1$) that were also being considered do not show signatures of selection. A recent study comparing the frequency of positive selection among duplicates and single-copy genes found no differences (Studer et al. 2008) but was limited to comparisons using duplication events that occurred hundreds of millions of years ago. Positive selection may often act on only a small number of amino acids in a very short period of time after duplication. By capturing the earliest stages of gene duplicate evolution, we are able to observe evidence for positive selection before it is masked by the purifying selection that follows. The high level of positive selection acting on amino acids in these very young duplicates also implies that these changes are the ones responsible for the maintenance of the paralogs and are not just substitutions occurring after the maintenance has been established by other means. It must also be pointed out that the power of our test for selection is extremely low. Even if 20% of codons in every single duplicate along the human lineage were undergoing adaptive natural selection with $d_N/d_S = 5$, simulations show that we would only detect 25% of paralagous pairs with d_N/d_S significantly greater than 1 (Supplemental Fig. 7). This suggests that a much larger proportion of duplicate genes may be under positive selection than we are able to detect.

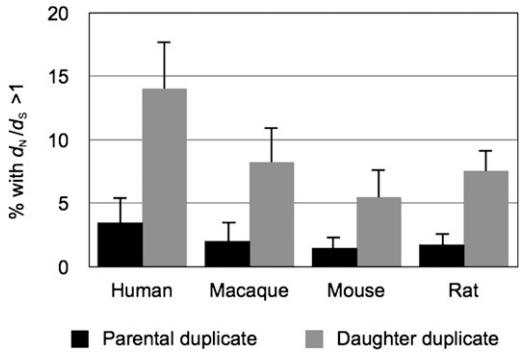

Figure 3. Positive selection on transposed duplicates. The numbers of lineage-specific duplicates with evidence for positive selection using test 2 ($P < 0.05$) are shown, for cases where one paralog has been classified as "parental" and one as "daughter" (see text for details).

A large number of the families we identified as evolving by positive selection in humans also overlap with gene families previously shown to have rapidly expanded in copy number in primates (Demuth et al. 2006); over all lineages, families that have undergone rapid lineage-specific expansions also showed a significant increase in adaptive evolution of the constituent protein sequences ($P = 0.009$). This pattern supports the hypothesis that positive selection is responsible for the retention of duplicated genes and the consequent expansion of gene families.

Genes with a history of adaptive evolution along the human lineage

Along the human lineage since the split with macaque, we found a number of families previously identified as having evolved by positive selection: the DUF1220-containing NBPF family (Vandepoele et al. 2005; Popesco et al. 2006), the morpheus gene family (Johnson et al. 2001), chorionic gonadotropin (Maston and Ruvolo 2002), the PRAME gene family (Birtle et al. 2005; Gibbs et al. 2007), and the SERPIN family (Seixas et al. 2007). We also identified a number of new targets of selection (Table 2). These include two genes that are involved in neurotransmission (*RIMBP3B*, *RIMBP3C*), as well as multiple genes involved in immune response and response to inflammation or stress (*SAA1*, haptoglobin, *SERPINB4*). Both serum amyloid A genes (*SAA1* and *SAA2*) are induced by cytokines, but only *SAA1* is enhanced by glucocorticoids (GCs) while *SAA2* is unresponsive to GCs (Thorn and Whitehead 2002). This difference is interesting considering the negative effect of GCs when secreted chronically (Sapolsky 2005). Along with divergence in regulatory control, our results suggest additional divergence in the function of the SAA1 protein itself. We also found many families under positive selection that have not been characterized but may warrant further study (Table 2).

Considering only those genes that duplicated along the human lineage after the split with chimpanzee, the nine families identified as evolving by positive selection again include *RIMBP3B*, *RIMBP3C*, as well as *CSAG1* (a chondrosarcoma-associated protein), *VCX*, and a number of other genes (Table 3). *RIMBP3B* and *RIMBP3C* are members of the RIM-binding protein family. Both are paralogous to the *RIMBP3* gene that exists in all mammals, but these two genes are only found in humans indicating two duplication events along the human lineage (Mittelstaedt and Schoch 2007). RIMBPs are known to interact with both Ca^{2+} channels and presynaptic active zone proteins (RIMs), suggesting a role in the controlled release of neurotransmitters. Interestingly, members of the RIMBP family have been found to interact with the TSPO (formerly known as BZRP) protein—a peripheral receptor of benzodiazepines—in a yeast two-hybrid system (Galiegue et al. 1999). *VCX* is a member of the Variable Charge

VCX/Y gene family that is found on the X and Y chromosomes. Previous studies have shown that there are more nonsynonymous than synonymous substitutions between genes *VCX* and *VCY* (Lahn and Page 2000). *VCX* is expressed exclusively in germ-line cells and, while the function of *VCX* is not known, studies have suggested an intriguing association with cognitive development (Van Esch et al. 2005). Though it is not significant, it is nonetheless interesting to note the excess of adaptively evolving genes we found that are involved in neuronal and cognitive functions, as well as being used in response to social stresses among primates (Sapolsky 2005). Regardless of the exact selective forces responsible for these changes, equivalent genome-wide analyses of orthologs have only identified between two and 10 genes under positive selection in humans since our split with chimpanzee at similar levels of statistical stringency (Bakewell et al. 2007; Gibbs et al. 2007; Kosiol et al. 2008); our analysis of gene duplicates has therefore added substantially to the total number of human-specific genic adaptations.

Effect of genomic context on the evolution of duplicated genes

Duplication has an important role in the creation of novel genes. Through the redundancy generated by duplication, one of the paralogous copies can escape the pressure of negative selection and accumulate mutations that can establish a new function. A number of previous studies have found asymmetries in the rates of evolution among paralogs (Kondrashov et al. 2002; Zhang et al. 2002; Conant and Wagner 2003), though to our knowledge only one has been able to polarize mutations between parent and daughter copies (Cusack and Wolfe 2007). This study found that, while daughter genes appeared to be evolving faster overall, the pattern was driven solely by those genes relocated by retrotransposition; they also did not explicitly test for positive selection among paralogs (Cusack and Wolfe 2007). Our study is the first attempt to classify all duplicated genes, created both by retrotransposition and by DNA-based duplication, into parents and daughters and to find differences in the form of natural selection each has experienced.

The pattern we observed is striking. Among the duplicates for which we were able to polarize the parent and daughter copies, a significantly higher fraction of daughter copies experienced adaptive evolution in all lineages. In >80% of parent–daughter pairs that show $d_N/d_S > 1$, it is the duplicate that has moved to a new genomic location (the "daughter") that has experienced positive selection. Some of these daughter copies are single-exon genes possibly created through retrotransposition, but a considerable number of them are genes with multiple exons (32 out of 53), suggesting that this pattern is valid across all duplicates regardless of their mechanism of creation.

Our results support the original model of neofunctionalization proposed by Susumu Ohno (Ohno 1970): After duplication, one of the copies continues to carry out the ancestral gene function, while the other is free to evolve a new function. We find that the gene copy in the original genomic location appears to be strongly constrained in function, while the copy transposed to a new genomic context is freed from constraint. The new genomic and epigenetic environment experienced by the daughter paralog makes it unlikely that it will be expressed in the full range of tissues necessary for ancestral function. So it is more likely that the fully capable, original copy will be constrained to continue with the original roles, while the new copy is allowed to gain a new function via adaptive natural selection.

Conclusion

The molecular changes responsible for organismal adaptation are generally identified by comparing orthologous regions between species. In humans, such methods have identified adaptive changes in protein-coding genes (Enard et al. 2002), regulatory regions (Rockman et al. 2005; Haygood et al. 2007), and even noncoding genes (Pollard et al. 2006). Our results suggest that, by including analyses of duplicated genes, we can recover substantial information on the adaptive evolution of lineage-specific phenotypes, especially human-specific phenotypes. In fact, these data show that the frequency of adaptive evolution is much higher among duplicated genes than among orthologs. This implies that we cannot use the proportion of positively selected genes from studies of orthologs to parameterize models of duplicate gene evolution or even models of organismal adaptation. Considering the conservative nature of the tests for positive selection used here, our results strongly support a common role for adaptive evolution in the maintenance of duplicate copies.

Methods

Data

We used the gene models for human, chimpanzee, macaque, mouse, and rat defined in Ensembl v41 as the main data set. Each of these genes is placed into a gene family by Ensembl based on an all-against-all BLASTP sequence similarity search, followed by clustering of similar proteins using the MCL algorithm (Enright et al. 2002). The corresponding protein and cDNA sequences for each gene from each family were downloaded for construction of the gene trees and analyses of positive selection.

Identification of lineage-specific duplicates

We built gene trees for all 10,204 families with PHYLIP (Felsenstein 1989) using neighbor-joining and JTT protein distances with 100 bootstrap replicates. Duplications in the tree were identified by reconciliation with the known species tree using NOTUNG 2.5 (Chen et al. 2000); only duplication nodes with bootstrap support of >90% were used. Duplicated nodes specific to each lineage (i.e., in mouse after the split with rat, in rat after the split with mouse, in human after the split with macaque, and in macaque after the split with human) were extracted, ignoring the chimpanzee lineage for the main data set ("Reconciliation" in Table 1).

Since duplication events can incorrectly appear to be lineage-specific when multiple copies are lost in sister species, we further filtered the duplicates based on branch lengths. We used the genetic distance (d_S) corresponding to the time of speciation between each lineage and its sister species, and required the distance between any two duplicate genes in the trees to be less than twice the distance since speciation. The average d_S values for each of the four lineages were taken from the genomic average of one-to-one orthologs (Wang et al. 2007). Figure 1 shows the d_S cutoffs for each lineage, which may differ between sister lineages because of unequal rates of nucleotide substitution. The cutoff for human-specific genes relative to chimpanzee was $d_S = 0.0075$. As an example, because the average distance along the human lineage back to the human–macaque ancestor is $d_S = 0.032$, for a duplication to be specific to humans two paralogs could not be diverged more than $d_S = 0.064$. In addition, we identified paralogs in the same family whose sequences did not overlap each other in alignment, possibly due to gene fission or fusion events; these families were split or the genes removed. These steps resulted in 2791 lineage-specific subtrees containing 7271 paralogs that went

through 4480 duplication events in the four separate lineages ("d_S filtering" in Table 1). The number of new duplicates in each lineage is also shown in Table 1; this number represents the additional genes that have been added to each genome and can be calculated by subtracting the number of trees from the total number of genes contained in these trees.

Test for positive selection

We used the program PAML 3.15 (Yang 2007) to test for positive selection. We used two different likelihood ratio tests to identify genes under positive selection: the "site" models M1a vs. M2a (Yang et al. 2005) and the "branch-site" models included in test 2 (Zhang et al. 2005). Test 2 compares a model in which the branch under consideration is evolving without constraint ($d_N/d_S = 1$) to a model in which this branch has some proportion of sites evolving under positive selection ($d_N/d_S > 1$). Sequences that have few substitutions between them offer little to no power to detect selection for both tests. We therefore removed the subtrees whose sequences had fewer than three mutations or whose alignments were <50 bases. We were left with 6315 genes in 2377 trees ("PAML" in Table 1; Fig. 1). Simulated data sets of sequences of length 1500 bases were generated using the EVOLVER program of the PAML 4 (Yang 2007) package.

The extracted proteins were realigned using ClustalW (Thompson et al. 1994) and the corresponding cDNA sequences were aligned according to the protein alignments. In order to avoid false signals of selection due to poor alignment, regions with low quality alignments were masked out. The method to mask poor alignments was essentially a method to find regions that contained codons with >1 substitution or that were aligned to a gap. The exact criteria for filtering the alignment was as follows: We required that in a window of five codons, for all three overlapping windows of three contiguous codons, there must be at least two codons in each that have no more than one base substitution. The following equations specify when the codon is masked, i.e., when mask is 1:

$$m_i : \text{number of mismatches for codon } i$$

$$I_i = \begin{cases} 1 \text{ if } mi < 2 \\ 0 \text{ otherwise} \end{cases}$$

$$nmask_i = \begin{cases} 1 \text{ if } I_i + I_{i+1} + I_{i+2} < 2 \\ 0 \text{ otherwise} \end{cases}$$

$$mask_i = \begin{cases} 1 \text{ if } (I_{i-1} \wedge \neg mask_{i-1}) + I_i + (I_{i+1} \wedge \neg nmask_i) < 2 \\ 0 \text{ otherwise} \end{cases}$$

Each family underwent at least two to at most 10 PAML runs to ensure that likelihood values converged. If the maximum likelihood score and the next best score differed by >0.001, or if the likelihood ratio was a negative value, we assumed failed convergence and removed the family from the final data set. The likelihood ratio test between models M1a and M2a was conducted with a critical value of $\chi^2 = 5.99$ (i.e., P-value = 0.05, $d.f.$ = 2), as suggested by Yang (2007). This cutoff can be very conservative (Wong et al. 2004). False discovery rates (FDR) were calculated using Q-value (Storey 2002). For the branch-site test 2, we included as an outgroup the closest protein sequences from the closest sister species based on the reconciled gene trees in order to polarize lineage-specific branches. The critical value used for branch-site tests was $\chi^2 = 3.84$ (P-value = 0.05, $d.f.$ = 1), also as recommended by Yang (2007). FDRs for test 2 were calculated by correcting for the number of lineage-specific branches tested.

Validation

Orthologous pairs of genes are from Gibbs et al. (2007), with alignments kindly provided by T. Vinar and A. Siepel. Positive selection among the orthologs was detected using the likelihood ratio test between models M1a and M2a in PAML, as above. To assess any bias due to misannotations of paralogs, all the proteins were queried against the more recent version 48 of Ensembl. The number of trees with evidence for positive selection was recalculated excluding the proteins modified in version 48. Pairwise Nei–Gojobori d_N/d_S values were calculated among all proteins in the tree by PAML and the average value for the family was used. Gene conversion was detected on all duplicates using GENECONV v1.81 (Sawyer 1989; McGrath et al. 2009), and we again tested for positive selection after removing trees with evidence of gene conversion at $P < 0.05$. All results from these validation steps are presented in Supplemental Table 3.

Parent–daughter classification

In order to classify the duplicates into parents and daughters, we looked for conserved synteny in the flanking region of the genes. Assuming that the parent gene will maintain a longer stretch of conserved synteny with the outgroup gene, while the daughter copy will have a shorter syntenic block that only comprises the duplicated segment, we can determine the parental relationship based on the length of the conserved synteny. In order to estimate the conserved synteny we used a probabilistic model that takes into account the length of the conserved block, and the probability of observing homologous genes within the sytenic block versus outside the sytenic block. More details on the model and estimation are described in Han and Hahn (2009). For each lineage-specific gene, the identity of flanking genes was noted for 10-Mb regions both upstream and downstream. After the length of the conserved syntenic block was estimated, we clustered the genes into two clusters, parent vs. daughter, based on the length. In order to test for an association between positive selection and synteny, we only counted the cases where we could identify a one-to-one relationship between a single parental gene and a translocated daughter gene. All families containing tandem duplicates were excluded from this analysis, even when parent–daughter relationships could be determined. We used the UCSC Genome Browser (Karolchik et al. 2004), and the Galaxy server (Giardine et al. 2005) to calculate GC content, recombination rate, and the number of conserved elements predicted by phastCons (Siepel et al. 2005) for the flanking region (±1 Mb) of each gene. Exons were removed from the phastCons elements analysis to avoid any bias due to possible gene enrichment in the selected regions.

Acknowledgments

We thank M. Lynch for comments and discussion. This work was funded by a grant from the National Science Foundation (DBI-0543586) to M.W.H.

References

Anisimova, M., Nielsen, R., and Yang, Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164:** 1229–1236.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003–1007.

Bakewell, M.A., Shi, P., and Zhang, J.Z. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc. Natl. Acad. Sci.* **104:** 7489–7494.

Betran, E. and Long, M. 2003. *Dntf-2r*, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. *Genetics* **164:** 977–988.
Birtle, Z., Goodstadt, L., and Ponting, C.P. 2005. Duplication and positive selection among hominin-specific *PRAME* genes. *BMC Genomics* **6:** 120. doi: 10.1186/1471-2164-6-120.
Casola, C. and Hahn, M.W. 2009. Gene conversion among paralogs results in moderate false detection of positive selection using likelihood methods. *J. Mol. Evol.* (in press).
Chen, K., Durand, D., and Farach-Colton, M. 2000. NOTUNG: A program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.* **7:** 429–447.
Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R.K., Paabo, S., et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437:** 88–93.
Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D., Kejariwal, A., Todd, M.A., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B., et al. 2003. Inferring non-neutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302:** 1960–1963.
Conant, G.C. and Wagner, A. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Res.* **13:** 2052–2058.
Cusack, B.P. and Wolfe, K.H. 2007. Not born equal: Increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol. Biol. Evol.* **24:** 679–686.
Demuth, J.P., De Bie, T., Stajich, J.E., Cristianini, N., and Hahn, M.W. 2006. The evolution of mammalian gene families. *PLoS One* **1:** e85. doi: 10.1371/journal.pone.0000085.
Dykhuizen, D. and Hartl, D.L. 1980. Selective neutrality of 6pgd allozymes in *E. coli* and the effects of genetic background. *Genetics* **96:** 801–817.
Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S.L., Wiebe, V., Kitano, T., Monaco, A.P., and Paabo, S. 2002. Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* **418:** 869–872.
Enright, A.J., Van Dongen, S., and Ouzounis, C.A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30:** 1575–1584.
Felsenstein, J. 1989. PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164–166.
Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.-l., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.
Galiegue, S., Jbilo, O., Combes, T., Bribes, E., Carayon, P., Le Fur, G., and Casellas, P. 1999. Cloning and characterization of PRAX-1. *J. Biol. Chem.* **274:** 2938–2952.
Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. 2005. Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.* **15:** 1451–1455.
Gibbs, R., Rogers, J., Katze, M., Bumgarner, R., Weinstock, G., Mardis, E., Remington, K., Strausberg, R., Venter, J., Wilson, R., et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316:** 222–234.
Hahn, M.W. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J. Hered.* (in press).
Hahn, M.W., Demuth, J.P., and Han, S.-G. 2007. Accelerated rate of gene gain and loss in primates. *Genetics* **177:** 1941–1949.
Han, M.V. and Hahn, M.W. 2009. Identifying parent-daughter relationships among duplicated genes. *Pacific Symposium on Biocomputing.* **14:** 114–125.
Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.-D., and Wray, G.A. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat. Genet.* **39:** 1140–1144.
Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. B. Biol. Sci.* **256:** 119–124.
Johnson, M.E., Viggiano, L., Bailey, J.A., Abdul-Rauf, M., Goodwin, G., Rocchi, M., and Eichler, E.E. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413:** 514–519.
Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32:** D493–D496.
Kondrashov, F., Rogozin, I., Wolf, Y. and Koonin, E. 2002. Selection in the evolution of gene duplications. *Genome Biol.* **3:** research0008.0001–research0008.0009. doi: 10.1186/gb-2002-3-2-research0008.
Kosiol, C., Vinař, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R., and Siepel, A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4:** e1000144. doi: 10.1371/journal.pgen.1000144.
Lahn, B.T. and Page, D.C. 2000. A human sex-chromosomal gene family expressed in male germ cells and encoding variably charged proteins. *Hum. Mol. Genet.* **9:** 311–319.
Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.
Maston, G.A. and Ruvolo, M. 2002. Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. *Mol. Biol. Evol.* **19:** 320–335.
McGrath, C.L., Casola, C., and Hahn, M.W. 2009. Minimal effect of ectopic gene conversion among recent duplicates in four mammalian genomes. *Genetics* (in press). doi: 10.1534/genetics.109.101428.
Merritt, T.J.S. and Quattro, J.M. 2001. Evidence for a period of directional selection following gene duplication in a neurally expressed locus of triosephosphate isomerase. *Genetics* **159:** 689–697.
Mittelstaedt, T. and Schoch, S. 2007. Structure and evolution of RIM-BP genes: Identification of a novel family member. *Gene* **403:** 70–79.
Moore, R.C. and Purugganan, M.D. 2003. The early stages of duplicate gene evolution. *Proc. Natl. Acad. Sci.* **100:** 15682–15687.
Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3:** e170. doi: 10.1371/journal.pbio.0030170.
Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin, Germany.
Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443:** 167–172.
Popesco, M.C., MacLaren, E.J., Hopkins, J., Dumas, L., Cox, M., Meltesen, L., McGavran, L., Wyckoff, G.J., and Sikela, J.M. 2006. Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. *Science* **313:** 1304–1307.
Rockman, M.V., Hahn, M.W., Soranzo, N., Zimprich, F., Goldstein, D.B., and Wray, G.A. 2005. Ancient and recent positive selection transformed opioid *cis*-regulation in humans. *PLoS Biol.* **3:** e387. doi: 10.1371/journal.pbio.0030387.
Rodriguez-Trelles, F., Tarrio, R., and Ayala, F.J. 2003. Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the *xanthine dehydrogenase* gene. *Proc. Natl. Acad. Sci.* **100:** 13413–13417.
Sapolsky, R.M. 2005. The influence of social hierarchy on primate health. *Science* **308:** 648–652.
Sawyer, S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6:** 526–538.
Seixas, S., Suriano, G., Carvalho, F., Seruca, R., Rocha, J., and Di Rienzo, A. 2007. Sequence diversity at the proximal 14q32.1 SERPIN subcluster: Evidence for natural selection favoring the pseudogenization of SERPINA2. *Mol. Biol. Evol.* **24:** 587–598.
She, X., Liu, G., Ventura, M., Zhao, S.Y., Misceo, D., Roberto, R., Cardone, M.F., Rocchi, M., Program, N.C.S., Green, E.D., et al. 2006. A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome Res.* **16:** 576–583.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
Spofford, J.B. 1969. Heterosis and evolution of duplications. *Am. Nat.* **103:** 407–432.
Stoltzfus, A. 1999. On the possibility of constructive neutral evolution. *J. Mol. Evol.* **49:** 169–181.
Storey, J.D. 2002. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64:** 479–498.
Studer, R.A., Penel, S., Duret, L., and Robinson-Rechavi, M. 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res.* **18:** 1393–1402.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Thorn, C.F. and Whitehead, A.S. 2002. Differential glucocorticoid enhancement of the cytokine-driven transcriptional activation of the human acute phase serum amyloid A genes, *SAA1* and *SAA2*. *J. Immunol.* **169:** 399–406.
Thornton, K. and Long, M. 2005. Excess of amino acid substitutions relative to polymorphism between X-linked duplications in *Drosophila melanogaster*. *Mol. Biol. Evol.* **22:** 273–284.
Vandepoele, K., Van Roy, N., Staes, K., Speleman, F., and van Roy, F. 2005. A novel gene family NBPF: Intricate structure generated by gene duplications during primate evolution. *Mol. Biol. Evol.* **22:** 2265–2274.

Van Esch, H., Hollanders, K., Badisco, L., Melotte, C., Van Hummelen, P., Vermeesch, J.R., Devriendt, K., Fryns, J.-P., Marynen, P., and Froyen, G. 2005. Deletion of VCX-A due to NAHR plays a major role in the occurrence of mental retardation in patients with X-linked ichthyosis. *Hum. Mol. Genet.* **14:** 1795–1803.

Walsh, J.B. 1995. How often do duplicated genes evolve new functions? *Genetics* **139:** 421–428.

Wang, H.-Y., Chien, H.-C., Osada, N., Hashimoto, K., Sugano, S., Gojobori, T., Chou, C.-K., Tsai, S.-F., Wu, C.-I., and Shen, C.K.J. 2007. Rate of evolution in brain-expressed genes in humans and other primates. *PLoS Biol.* **5:** e13. doi: 10.1371/journal.pbio.0050013.

Wong, W.S.W., Yang, Z., Goldman, N., and Nielsen, R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168:** 1041–1051.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24:** 1586–1591.

Yang, Z., Wong, W.S.W., and Nielsen, R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22:** 1107–1118.

Zhang, Z. and Kishino, H. 2004. Genomic background predicts the fate of duplicated genes: Evidence from the yeast genome. *Genetics* **166:** 1995–1999.

Zhang, J.Z., Rosenberg, H.F., and Nei, M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci.* **95:** 3708–3713.

Zhang, L., Vision, T.J., and Gaut, B.S. 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **19:** 1464–1473.

Zhang, P., Gu, Z.L., and Li, W.H. 2003. Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol.* **4:** R56. doi: 10.1186/gb-2003-4-9-r56.

Zhang, J.Z., Nielsen, R., and Yang, Z.H. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22:** 2472–2479.

Zheng, D. 2008. Asymmetric histone modifications between the original and derived loci of human segmental duplications. *Genome Biol.* **9:** R105. doi: 10.1186/gb-2008-9-7-r105.

Received September 1, 2008; accepted in revised form February 9, 2009.