Estimating the Relative Contributions of New Genes from Retrotransposition and Segmental Duplication Events During Mammalian Evolution

Jin Jun[1], Paul Ryvkin[2], Edward Hemphill[3], Ion Măndoiu[1], and Craig Nelson[3]

[1] Computer Science & Engineering Department, University of Connecticut, Storrs, CT, 06269, USA. {jinjun,ion}@engr.uconn.edu
[2] Genomics & Computational Biology Graduate Group, University of Pennsylvania, Philadelphia, PA 19104, USA. pry@mail.med.upenn.edu
[3] Genetics & Genomics Program, Department of Molecular & Cell Biology, University of Connecticut, Storrs, CT 06269, USA. {edward.hemphill_iii,craig.nelson}@uconn.edu

Abstract. Gene duplication has long been recognized as a major force in genome evolution and has recently been recognized as an important source of individual variation. For many years the origin of functional gene duplicates was assumed to be whole or partial genome duplication events, but recently retrotransposition has also been shown to contribute new functional protein coding genes and siRNA's. Here we present a method for the identification and classification of retrotransposed and segmentally duplicated genes and pseudogenes based on local synteny. Using the results of this approach we compare the rates of segmental duplication and retrotransposition in five mammalian genomes and estimate the rate of new functional protein coding gene formation by each mechanism. We find that retrotransposition occurs at a much higher and temporally more variable rate than segmental duplication, and gives rise to many more duplicated sequences over time. While the chance that retrotransposed copies become functional is much lower than that of their segmentally duplicated counterparts, the higher rate of retrotransposition events leads to nearly equal contributions of new genes by each mechanism.

1 Introduction

The impact of changes in gene copy number on both evolution and human health are under increasing scrutiny. While the creation of new genes and the modulation of gene copy-number via duplication has long been recognized as an important mechanism for the evolution of lineage-specific traits [14], a number of recent studies have suggested that variation in gene family size may be even more widespread than previously appreciated [7] and that gene copy number variation between individuals may account for differences in disease predisposition within populations [18].

Three primary mechanisms of gene duplication have been described: whole genome duplication [9, 31], segmental duplication [3, 23], and retrotransposition [11, 35]. Whole genome duplication has been important to the evolution of many lineages [31], but it is a relatively rare event. Unlike whole genome duplication events, segmental duplications occur continuously and have contributed significantly to the divergence of gene content between mammalian genomes. Duplication by retrotransposition also occurs quite frequently, but because these new retrotransposed gene copies lack the flanking regulatory material of the parental gene, they have long been believed to give rise primarily to non-functional pseudogenes [16, 25]. Recent studies however, have indicated the presence of many apparently functional retrocopies in various mammalian genomes, challenging traditional perspectives on the relevance of this event to genome evolution [17, 21, 27, 32]. Very recently retrotransposition has also been shown to contribute siRNA's [28, 33].

In this study we compare the rates of new gene formation by segmental duplication (SD) and retrotransposition (RT) in five eutherian genomes. We show that, while genes arising from SD events are up to six times more likely to remain functional than those arising from RT events, the number of RT events is nearly ten times that of SD events, resulting in roughly equal quantitative contributions of new genes by each duplication mechanism. Our analysis further shows that duplicate genes generated by each mechanism are under similar levels of constraint on their protein coding regions and that silent site substitution profiles of RT duplicate copies are consistent with bursts of retrotransposition during mammalian evolution, while segmental duplication appears to occur at a more stable rate.

2 Methods

2.1 Dataset

Protein sequences for the five species analyzed (human, chimp, mouse, rat and dog) were obtained from Ensembl (release 37) [8]. For genes with multiple alternative transcripts we developed a collapsed gene model that incorporates all potential exons of that gene. Resulting exon coordinates were used to obtain a representative protein sequence used for subsequent homology assignment and dN/dS computations. Ensembl protein family annotations served as a starting point for our analysis. Over all five species, there were 17,341 Ensembl families comprising 113,543 genes. Excluding families with members on unassembled contigs (no reliable synteny information) and families with more than 50 Ensembl genes (due to the excessive computation time required to generate multiple alignments) resulted in 8,872 gene families containing 53,733 genes.

Pseudogenes were identified using Pseudopipe [36] seeded with known transcripts from Ensembl release 37. Over all five species, 17,226 pseudogenes (14,189 processed pseudogenes and 3,037 non-processed pseudogenes) were detected. Each pseudogene was added to one of the 8,872 Ensembl gene families. This

process resulted in super-families consisting of both protein coding genes and related pseudogenes.

2.2 Identification of RT and SD events

Within each super-family a local synteny level was computed for all pairwise combinations of super-family members. Local synteny is defined as homology of upstream and downstream neighboring genes. For each pair, we checked homology between the 3 nearest up- and downstream neighboring Ensembl annotated genes. Homology between neighbors was defined by a BlastP [1] score of 50 or more and sequence similarity over 80% of corresponding protein sequences. After this analysis, for every pair (g_i, g_j) of family members we obtained two numbers $0 \leq n_u^{ij}, n_d^{ij} \leq 3$ representing the homology upstream and downstream neighbors. A synteny level $s_{i,j}$ of **2** was assigned to every pair of genes or pseudogenes that had homologous neighbors on both sides, up and down (i.e., whenever $n_u^{ij}, n_d^{ij} \geq 1$). When one side lacked homologous neighbors, we assigned a synteny level $s_{i,j}$ of **1** only if the other side had at least two homologous neighbors; otherwise (i.e., when $n_u^{ij} + n_d^{ij} \leq 1$) we assigned a synteny level $s_{i,j}$ of **0**.

Local synteny levels were used in a two-stage clustering algorithm (see Algorithm 1) to identify syntenic ortholog/paralog clusters. In our algorithm, for a set X of genes and pseudogenes, $Sp(X)$ denotes the set of species represented in X. For a set S of species, $LCA(S)$ denotes the last common ancestor in the phylogenetic tree. In the first stage, we used a single-linkage clustering algorithm to obtain core clusters by merging pairs of genes and pseudogenes with local synteny level of 2, predicted to be either orthologs or paralogs resulting from SD events which preserve up and downstream neighbors. In the second stage, we merged pairs of core clusters if every member of one cluster had synteny level of 1 to every member of the other cluster. Any two non-overlapping clusters from this two-stage clustering algorithm are mutually non-syntenic. Second stage clusters spanning a phylogenetically contiguous subset of the species represented in larger clusters from the same super-family represent putative descendants of RT events or SD events that have lost local synteny. Since retrotransposed gene copies generally lack introns due to their RNA-intermediate nature, we distinguish between these possibilities using intron content conservation scores as described below.

Within each cluster produced by the two-stage clustering algorithm there may be successive segmental duplication events. We use UPGMA (Unweighted Pair Group Method with Arithmetic mean) [26] to find these successive SD events. For input to UPGMA we compute the distance between two members g_i and g_j as the Pearson's correlation coefficient between the two vectors, $(n_u^{ik} + n_d^{ik})_k$ and $(n_u^{jk} + n_d^{jk})_k$, i.e. sums of upstream and downstream homologous neighbors with remaining genes g_k in the cluster. Given the UPGMA gene trees, we counted the inner nodes as SD events when two subtrees from such an inner node are in a species-subset relationship. If two subtrees from an inner node had disjoint species sets, this node was considered as a speciation event (Fig. 1).

We distinguish between putative descendants of RT events or SD events that have lost local synteny using intron conservation scores between descendant

Algorithm 1 Two-Stage Clustering Algorithm

Input: Family of genes and pseudogenes $F = \{g_1, g_2, \ldots, g_N\}$ with species information and pairwise synteny levels $s_{i,j}$

Initialization:
$C \leftarrow \emptyset$
$U \leftarrow \{g_1, g_2, \ldots, g_N\}$
(Stage1) Single-linkage clustering with synteny level 2:
While $U \neq \emptyset$ **do**
Select an arbitrary member g_i of U
$U \leftarrow U \setminus \{g_i\}; C_{open} \leftarrow \{g_i\}$
While there exists $g_j \in U$ with synteny 2 to a member of C_{open}, **do**
$U \leftarrow U \setminus \{g_i\}; C_{open} \leftarrow C_{open} \cup \{g_i\}$ // Add g_j to core cluster
$C \leftarrow C \cup C_{open}$
(Stage2) Merging of clusters with high average pairwise synteny:
While there is a (C_l, C_m) where SYNTENIC_TEST (C_l, C_m) is true, **do**
$C \leftarrow C \setminus \{C_l, C_m\}$
$C \leftarrow C \cup \{C_l \cup C_m\}$
Return C

SYNTENIC_TEST(A, B)
If $Sp(A)$ and $Sp(B)$ are subsets of different lineages, i.e.
$LCA(Sp(A)) \neq LCA(Sp(A \cup B))$ and $LCA(Sp(B)) \neq LCA(Sp(A \cup B))$, **then**
If $s_{i,j} = 1$ for every pair $g_i \in A$, $g_j \in B$ **then return true**
Else, if $LCA(Sp(A)) = LCA(Sp(A \cup B))$ **then**
$A' \leftarrow$ set of genes/pseudogenes of A of species descending from $LCA(Sp(B))$
If $s_{i,j} = 1$ for every pair $g_i \in A'$, $g_j \in B$ **then return true**
Else, return false

genes and pseudogenes. The intron conservation rate between two paralogous genes was calculated as the ratio of the number of shared introns divided by the total number of intron positions from the protein/intron alignment between two genes (based upon the method of [20]. An event was identified as an RT duplication if the average intron conservation rate to paralogs outside the cluster was below 1/3.

2.3 Event assignment to tree branches and evidence of function

We use parsimony to assign each inferred duplication event to a specific branch of the 5-species tree. We assign each event to the tree branch corresponding to the exact set of species spanned by the descendant genes of the detected duplication event, which we refer to as *assigned* events. *Intact* events are defined as those duplication events that have no apparent disruption (e.g. in stop codons) of the protein coding reading frame and an Ensembl annotated gene in each of the species spanned by the cluster. *Functional* events are defined by the clusters of putative protein coding genes with average dN/dS ratio below 0.5 over all pairs

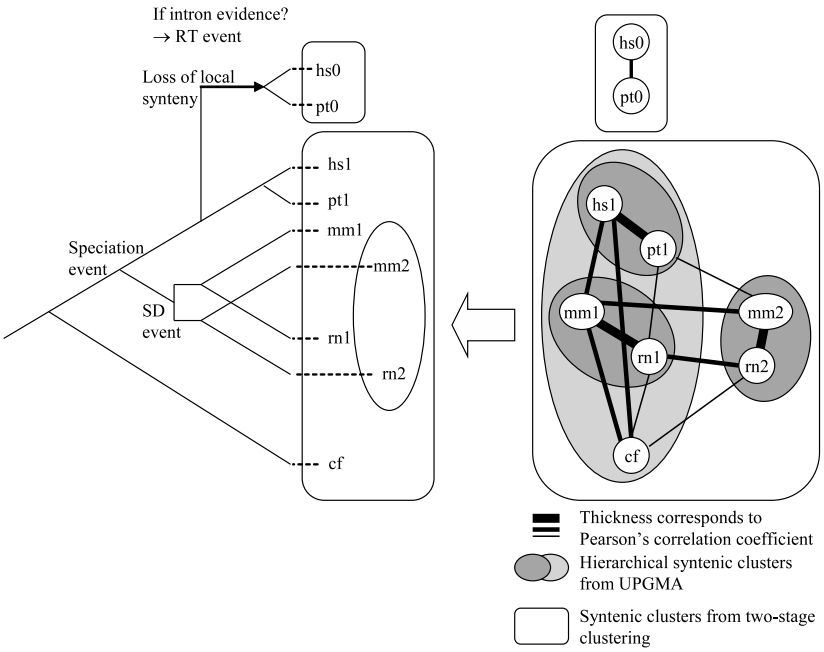

Fig. 1. Inferring SD and RT events using local synteny and hierarchical clustering. This example shows how SD and RT events are inferred from a super-family having 9 members: 2 members per each species except for dog, from the results of our clustering algorithms (on the right side) to corresponding events (on the left side). By using two-stage clustering algorithm, two syntenic clusters are formed, shown as hollow rounded rectangles. Loss of introns in one cluster suggests that the loss of synteny was due to an RT event. UPGMA builds hierarchical clusters within each syntenic cluster and speciation and SD events are inferred based on species sets.

of genes within the cluster. Pairwise dN and dS measures were estimated using the `YN00` program of PAML [34].

3 Results

3.1 Lineage distribution of duplication events

Events giving rise to clusters of genes with no conservation of synteny relative to "parental" genes and low inter-cluster intron conservation rates were classified as *RT events*, while events giving rise to clusters of genes with high local synteny to parental genes were classified as *SD events*. Events corresponding to gene clusters with indeterminate intron conservation or local synteny to parental genes were classified as *ambiguous*. This analysis resulted in the classification of a total of 2,035 SD events, 12,507 RT events, and 2,742 ambiguous events. Using parsimony to assign non-ambiguous events to branches of the species tree resulted in 52 SD

	3 internal branches	Whole tree
SD **functional** / assigned events	**148** / 301 = 49.17%	1,649
RT **functional** / assigned events	**187** / 2,349 = 7.96%	12,078

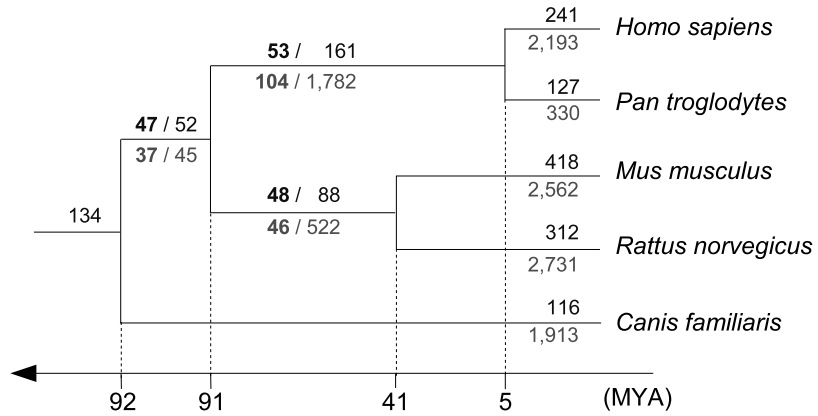

Fig. 2. Numbers of gene duplication events from segmental duplication (above the line) and retrotransposition (below the line). Numbers represent the assigned SD or RT events on each branch. Numbers typeset in bold on three internal branches are counts of functional events, defined in this study as intact events that yield clusters with average dN/dS ratio below 0.5 over pairs of homologous Ensembl genes. For three internal branches, fractions of the functional events over the total assigned events are shown, e.g. **53**/161 for SD events on primate branch. Evolutionary ages are based on [30].

and 45 RT events on the branch leading to primates and rodents (the in-group), 161 SD and 1,782 RT events on the primate branch leading to humans and chimps, and 88 SD and 522 RT events on the rodent branch leading to mice and rats (Fig. 2). Gene duplication events for the root and terminal branches of the tree were also counted, but were not used for further analysis due to the difficulty in estimating the degree of purifying selection on very recent duplication on the terminal branches and the age of duplications on the root. 386 SD and 429 RT events could not be reliably assigned to specific branches of the tree using parsimony and were also omitted from further analysis.

Duplication event counts on the three internal branches of the tree reveal an excess of RT events over SD events along all but the deepest branches of the tree, suggesting an average rate of RT copy formation 3-10 times higher than that of SD copy formation (Fig. 2). Deviation from this ratio along the in-group branch may be the result of a period of relative inactivity of retrotransposition compounded with the difficulty of detecting the products of old RT events not under purifying selective pressure [11].

3.2 Rates of duplication

Rates of retrotransposition vary significantly over time and bursts of retrotransposition have been reported in several mammalian lineages [11, 35]. The synonymous substitution rate (dS) profiles of the duplicates identified in this study (Fig. 3) are shaped by the rate of generation of new duplicates, the mutation rates along each lineage, the age of the genes identified in each interval, and our ability to identify genes uniformly along each lineage. Pseudogenes, for instance, become increasingly difficult to identify as they get older and diverge from their original sequence. RT events in all three internal branches show clear peaks in dS (Fig. 3A). For duplications occurring on the primate branch this peak occurs around dS=0.1, while in rodents it occurs around dS=0.3 and in in-groups around dS=0.6 $\sim$ 0.8. This pattern is consistent with bursts of retrotransposition in each of these lineages, a high mutation rate in the rodent lineage, and the 36Myr gap between the speciation events leading to rodent and primate lineages. Duplications occurring prior to the rodent/primate split display a dS distribution significantly shifted toward higher dS values, consistent with the greater age of these duplicates.

Segmental duplications show similar patterns in dS but a more uniform distribution of dS values than RT duplicates (Fig. 3B and C), suggesting that segmental duplication is a more uniform process that occurs at less variable rates than retrotransposition. It is interesting to note that the inferred age distribution of segmental duplication events is more uniform than that of the RT duplicates but is not perfectly flat, suggesting that there may be some variation in the rate of segmental duplication over evolutionary time.

3.3 Functional preservation rates

It is probable that young duplicate genes may escape inactivation for some time despite lacking any apparent function. Since Ensembl gene predictions rely upon the presence of an intact coding region rather than any evidence of selection pressure upon the sequence, the gene clusters resulting from intact duplication events should be comprised of both functional genes and duplicates that are not functional, but have escaped inactivation. Evidence of purifying selection is often used as evidence for function, and the ratio of synonymous to non-synonymous changes (dN/dS) in the protein-coding region of a gene is a convenient way of estimating this selective pressure [13]. For example, dN/dS ratio < 0.5 has been used as stringent functionality criteria between retrotransposed genes and their parental genes [6]. Also Torrents et al. showed that there is a clear discrimination between dN/dS ratios of pseudogenes and those of functional genes, supporting the use of dN/dS ratios as evidence of function [29]. Here we compute dN/dS ratios between all pairs of descendants from each duplication event. This pairwise approach is computationally rapid, is independent of precise reconstruction of the entire gene tree, and allows for the detection of functionalized descendant clusters of a duplication event that are not constrained relative to the parental genes.

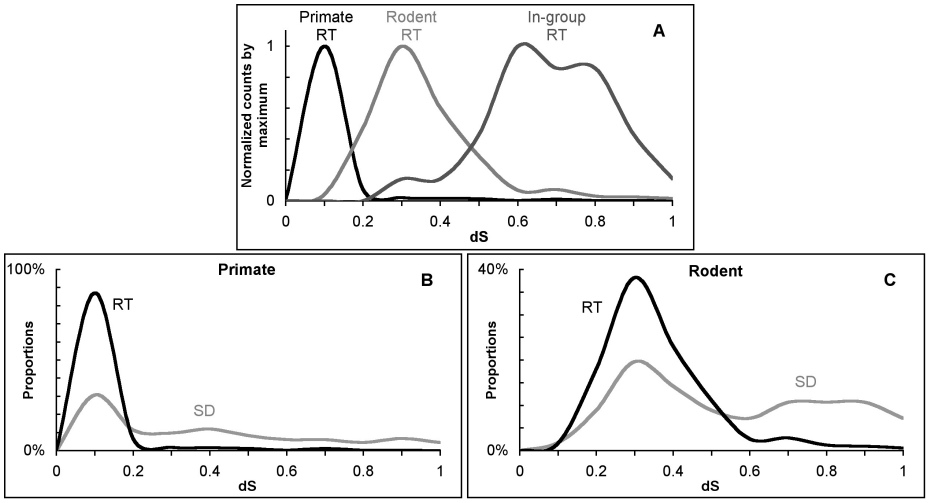

Fig. 3. Histograms of average dS over pairs of Ensembl genes and pseudogenes. (A) for clusters resulting from RT events on the primate, rodent, and the in-group branch leading to primates and rodents, (B) for clusters resulting from SD events and RT events on the primate lineages and (C) on the rodent lineages.

Analysis of the dN/dS ratios of clusters derived from duplication events is quite revealing. Fig. 4A compares clusters of RT duplication event descendants with intact protein coding reading frames (intact) and clusters of RT duplicates with inactivated reading frames (inactivated). Aggregate dN/dS values of a significant portion of intact clusters overlap with the dN/dS values of inactivated clusters in the region of the graph where dN/dS is greater than ~ 0.5. Assuming that the vast majority of inactivated clusters (clusters whose members have inactivating mutations in their protein coding regions) are not under purifying selection for protein coding function, those intact clusters that fall into this range are unlikely to encode functional proteins, despite lacking any clearly inactivating mutation. By inference, those clusters that display significantly lower aggregate dN/dS values (< 0.5) than inactivated clusters are likely to be under stabilizing selection for protein coding function.

Panels B through D of Fig. 4 compare dN/dS values of duplicate clusters derived from RT and SD events on each of the three internal branches of the mammalian tree. In the oldest internal branch of the tree (in-group) very few clusters generated by either duplication mechanism can be detected that are not under some degree of purifying selection pressure. This is probably due to the difficulty in identifying very old non-functional sequences. Such sequences are expected to drift away from their parental sequence making identification increasingly difficult with advanced age. Clusters derived from duplication events along the rodent branch have a bimodal distribution of dN/dS ratio resulting from RT and SD events that gave rise to putatively functional gene copies (ag-

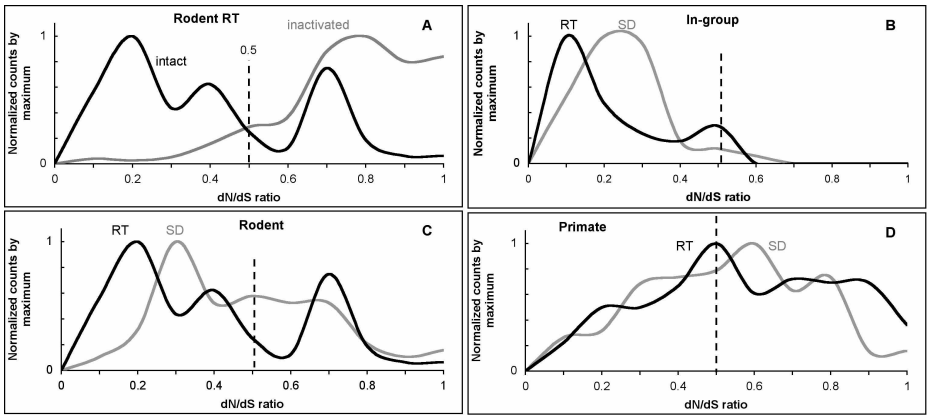

Fig. 4. (A) Histograms of average dN/dS ratio over pairs of Ensembl genes for clusters resulting from intact RT events and average dN/dS ratio over pairs of genes and pseudogenes for clusters resulting from inactivated RT events on the rodent lineage. Histograms of average dN/dS ratio over pairs of Ensembl genes for clusters resulting from intact SD events and RT events on the (B) in-group branch leading to primates and rodents, (C) rodent, and (D) primate.

gregate dN/dS values < 0.5), and clusters with no clear evidence of stabilizing selective pressure. Duplication events along the primate branch gave rise to clusters with more uniformly distributed aggregate dN/dS values spanning the entire range of measurements. This is likely to be a reflection of the relatively short period of time these new genes have been under purifying selection and is consistent with the relatively low dS values of duplicates detected along this branch (Fig. 3B).

3.4 Distribution of duplication events within the mammalian tree

The total number of RT and SD duplication events detected in this study is illustrated in Fig. 2. Along each branch the number of events giving rise to clusters with evidence of purifying selective pressure on their protein coding regions is in bold typeset, while the total number of events detected is in denominators. From these numbers it is clear that that we detect far more RT events than SD events, but that far fewer of these events give rise to functional protein coding genes than their SD counterparts. Analysis of the internal branches individually reveals possible differences in the relative probability of these events giving rise to functional genes in different lineages. In the most basal branch shared by rodents and primates there is a slight excess of functional SD events over functional RT events, while the two mechanisms appear to contribute equal numbers of functional events in the rodent lineage. The primate and rodent branches show similar rates of assigned SD events, but in primates fewer of these events give rise to functional descendants (Table 1). A decreased rate of functionaliza-

tion is apparent in the RT events on the primate lineage. Despite an RT event rate nearly twice that seen in rodents, the number of functional RT events in primates is only $\sim 25\%$ greater than that in rodents.

Table 1. Rates of duplication events for rodent and primate lineages.

Events per	SD events			RT evnets		
million yrs	Assigned	Intact	Functional	Assigned	Intact	Functional
Rodents	1.76	1.56	0.96	10.4	1.42	0.92
Primates	1.87	1.31	0.62	20.7	3.41	1.21

4 Discussion

4.1 Identification and characterization of gene duplications during mammalian evolution

Identifying gene duplication events and placing them in a phylogenetic framework depends upon sensitive identification of duplicate copies, reliable clustering of orthologs, and differentiating between lineage specific gene loss events and more recent duplications. To identify groups of duplicated sequences we combine Ensembl gene predictions with Pseudopipe pseudogene identification. Combining predicted genes and pseudogenes in our gene families significantly reduces the complexity of placing duplication events on the phylogenetic tree; gene loss events are represented by pseudogenes and need not be inferred. Of course, this approach is less effective as pseudogenes age and become more difficult to detect deep in the tree. Undetected gene loss events deeper in the tree may lead to misassignment of some duplication events to younger branches and a consequent underestimation of the age of these gene families. But using local synteny to help classify duplication events appears to work relatively well for the species analyzed in this study.

Once duplicated genes have been identified and assigned to large gene families, clusters of orthologs within those families must be constructed to infer the time of the duplication event that gave rise to each cluster. Our clustering algorithm uses both the protein-coding information embedded in the Ensembl gene family assignments, and the local genome structure surrounding duplicate copies, to differentiate between DNA and RNA based duplications and to order successive segmental duplication events. This method is effective because random insertion of a retrocopied cDNA into the genome is very unlikely to recreate any significant synteny with orthologs or paralogs (data not shown). The very low false-positive rate associated with measures of local synteny means that genes that share synteny with paralogs are almost certainly the result of segmental duplications regardless of intron content. Therefore this method is

unlikely to misclassify RT duplicates as segmental duplications. Segmental duplications, however, can lose synteny to their paralogs over time [10, 19], which may result in some segmental duplications being mis-assigned to the RT class. To account for this we use conventional intron content criteria [32] to further discriminate between non-syntenic DNA based duplications and RT duplicate copies. Duplicate pairs that maintain synteny with their paralog are most likely DNA based, while non-syntenic paralogs with significant intron loss are likely RT duplicates. Comparison with other studies identifying RT duplicates in mammalian genomes suggest that using synteny criteria in addition to intron based criteria improves the reliability of RT duplicate classification and that duplications characterized as RT duplicates on the basis of intron content alone may in fact be SD duplicates.

While the gradual degradation of synteny can create problems for placing duplication events on a phylogenetic tree, it conversely enables the differentiation of successive segmental duplication events. Gene families generated by rounds of segmental duplication can be difficult to classify into definitive orthologous groups using protein-coding sequences alone. By examining flanking gene content, however, orthologous groups of paralogs can often be clearly resolved and iterative DNA based duplications placed on the phylogenetic tree. As a result we can see that while synteny decays over time, dN/dS values may also decrease, reflecting the prolonged influence of stabilizing selection.

Detection of duplicate genes will always depend strongly on the depth and quality of genome annotation. This fact is reflected in our results in the highest number of duplicates detected in the two most well annotated genomes in the study, human and mouse. While it is difficult to predict how many duplicates have been missed in current genome annotations, estimates of duplication rates from the most well-annotated genomes are now judged to be quite accurate [4, 22, 32]. The consistency of these estimates across the tree suggests that the number of duplications events is not highly variable between these species, but definitive demonstration of that finding must await further annotation (see also [5]).

4.2 Rates of duplication

Lineage specific gene duplication, by retrotransposition or segmental duplication, is a major force in the evolution of differences between genomes. Thousands of new genes have been born over the course of mammalian evolution, and while not all of these new genes live, they provide significant quantities of raw material for species-specific evolution and account for many of the known differences between closely related mammalian genomes [5]. Retrotransposition, in particular, appears to be peppering the genome with large numbers of duplicate retrocopies that can act as insertional mutagens [12], new duplicate genes [32], and siRNA's [28, 33]. Analysis of retrotransposon activity during vertebrate evolution shows strong peaks of activity [15] and it is therefore not surprising that RT duplication of genes shows similar peaks in birth rates. Segmental duplications, however, are

not expected to be dependant on retrotransposon duplication machinery and appear to occur at a more stable rate. Consistent with these expectations, the age profiles of the segmental duplications identified in our study are more broadly distributed than the RT age profiles, but interestingly, they are not perfectly uniform over time and may indicate of bursts of segmental duplication activity in the evolutionary history of these genomes (see also [2, 24]).

4.3 The fate of newly duplicated genes

At the moment a newly duplicated gene is born it is presumed to be an exact copy of the duplicated portion of the parental gene (cDNA for retrocopies; and introns, exons, and flanking material for segmental duplicates). Over time, however, mutation, coupled with selection, leads to the divergence of the new copy's sequence from its parent/paralog. The progressive aging of a duplicate is revealed in its dS profile, as we move deeper on the tree, dS values between duplicate pairs become progressively larger, reflecting the age of the duplications. If a new duplicate is functional, purifying selection will serve to remove deleterious non-synonymous mutations from the population, and the ratio of non-synonymous to synonymous changes (dN/dS) will diverge from that of non-functional copies. Full resolution of the degree of purifying selective pressure however, takes time, and estimating this pressure on young duplicates can be difficult. Indeed, we find significant separation between putative functional and non-functional descendants of a duplication event in populations of genes that have had sufficient time for this difference to become apparent (see the rodent branch Fig. 4A and C). For the young primate branch the divergence between functional and non-functional descendants is less clear. At virtually all timepoints, however, there are duplicates that have not yet been inactivated, but also show no evidence of purifying selection on their protein coding sequence. Whether this is the result of copies evading inactivation simply due to chance, or the reflection of some other phenomenon is unknown. We also observe the converse phenomenon, old copies that appear to have dN/dS ratios consistent with purifying selection, but inactivating mutations in their protein-coding region. This could be the result of recent inactivating mutations after long periods of purifying selection, or the result of purifying selection acting on fragments of the original protein coding sequence.

While the general effects of time, mutation, and selective pressure discussed above apply to all new duplicates, we wondered if RT duplicates and SD duplicates would show different degrees of purifying selective pressure. Interestingly, in age-matched populations of segmental and retrotransposed duplicates, there is no dramatic difference in selection pressure on genes born by these two mechanisms (Fig. 4). What is most clearly different between these two populations is the *proportion* of copies that show evidence of purifying selective pressure. Of the duplication events assigned to the branches leading to primates and rodents, only about six percent (150/2,304) of RT events give rise to duplicates showing evidence of purifying selection, while forty percent (101/249) of SD events appear to generate functional descendants (Fig. 2). The very high rate of RT

events coupled with the very low rate of functionalization of gene copies generated by these events, and the lower rate of SD events with much higher rate of descendant gene functionalization, results in nearly equal contributions of new genes to eutherian genomes by each of these two mechanisms.

References

1. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.
2. J. A. Bailey and E. E. Eichler. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*, 7(7):552–64, 2006.
3. J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler. Recent segmental duplications in the human genome. *Science*, 297(5583):1003–7, 2002.
4. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, 2004.
5. J. P. Demuth, T. De Bie, J. E. Stajich, N. Cristianini, and M. W. Hahn. The evolution of mammalian gene families. *PLoS ONE*, 1:e85, 2006.
6. J. J. Emerson, H. Kaessmann, E. Betran, and M. Long. Extensive gene traffic on the mammalian x chromosome. *Science*, 303(5657):537–40, 2004.
7. A. Fortna, Y. Kim, E. MacLaren, K. Marshall, G. Hahn, L. Meltesen, M. Brenton, R. Hink, S. Burgers, T. Hernandez-Boussard, A. Karimpour-Fard, D. Glueck, L. McGavran, R. Berry, J. Pollack, and J. M. Sikela. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol*, 2(7):E207, 2004.
8. T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. M. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, and E. Birney. Ensembl 2005. *Nucleic Acids Res*, 33(Database issue):D447–53, 2005.
9. I. Hurley, M. E. Hale, and V. E. Prince. Duplication events and the evolution of segmental identity. *Evol Dev*, 7(6):556–67, 2005.
10. M. A. Huynen and P. Bork. Measuring genome evolution. *Proc Natl Acad Sci U S A*, 95(11):5849–56, 1998.
11. A. C. Marques, I. Dupanloup, N. Vinckenbosch, A. Reymond, and H. Kaessmann. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol*, 3(11):e357, 2005.
12. R. E. Mills, E. A. Bennett, R. C. Iskow, and S. E. Devine. Which transposable elements are active in the human genome? *Trends Genet*, 23(4):183–91, 2007.
13. A. Nekrutenko, K. D. Makova, and W. H. Li. The k(a)/k(s) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res*, 12(1):198–202, 2002.
14. S Ohno. *Evolution by gene duplication*. Allen and Unwin, London, United Kingdom, 1970.
15. K. Ohshima, M. Hattori, T. Yada, T. Gojobori, Y. Sakaki, and N. Okada. Whole-genome screening indicates a possible burst of formation of processed pseudogenes

and alu repeats by particular l1 subfamilies in ancestral primates. *Genome Biol*, 4(11):R74, 2003.

16. D. A. Petrov and D. L. Hartl. Patterns of nucleotide substitution in drosophila and mammalian genomes. *Proc Natl Acad Sci U S A*, 96(4):1475–9, 1999.
17. L. Potrzebowski, N. Vinckenbosch, A. C. Marques, F. Chalme, B. Jègou, and H. Kaessmann. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol*, 6(4):e80, 2008.
18. R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–54, 2006.
19. E. P. Rocha. Inference and analysis of the relative stability of bacterial chromosomes. *Mol Biol Evol*, 23(3):513–22, 2006.
20. I. B. Rogozin, Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, and E. V. Koonin. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*, 13(17):1512–7, 2003.
21. H. Sakai, K. O. Koyanagi, T. Imanishi, T. Itoh, and T. Gojobori. Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. *Gene*, 389(2):196–203, 2007.
22. X. She, Z. Cheng, S. Zollner, D. M. Church, and E. E. Eichler. Mouse segmental duplication and copy number variation. *Nat Genet*, 2008.
23. X. She, Z. Jiang, R. A. Clark, G. Liu, Z. Cheng, E. Tuzun, D. M. Church, G. Sutton, A. L. Halpern, and E. E. Eichler. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, 431(7011):927–30, 2004.
24. X. She, G. Liu, M. Ventura, S. Zhao, D. Misceo, R. Roberto, M. F. Cardone, M. Rocchi, E. D. Green, N. Archidiacano, and E. E. Eichler. A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome Res*, 16(5):576–83, 2006.
25. R. Shemesh, A. Novik, S. Edelheit, and R. Sorek. Genomic fossils as a snapshot of the human transcriptome. *Proc Natl Acad Sci U S A*, 103(5):1364–9, 2006.
26. P. H. A. Sneath and R. R. Sokal. *Numerical Taxonomy*. W.H. Freeman and Company, San Francisco, 1973.
27. O. Svensson, L. Arvestad, and J. Lagergren. Genome-wide survey for biologically functional pseudogenes. *PLoS Comput Biol*, 2(5):e46, 2006.
28. O. H. Tam, A. A. Aravin, P. Stein, A. Girard, E. P. Murchison, S. Cheloufi, E. Hodges, M. Anger, R. Sachidanandam, R. M. Schultz, and G. J. Hannon. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453(7194):534–8, 2008.
29. D. Torrents, M. Suyama, E. Zdobnov, and P. Bork. A genome-wide survey of human pseudogenes. *Genome Res*, 13(12):2559–67, 2003.
30. A. Ureta-Vidal, L. Ettwiller, and E. Birney. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet*, 4(4):251–62, 2003.
31. Y. Van de Peer, J. S. Taylor, and A. Meyer. Are all fishes ancient polyploids? *J Struct Funct Genomics*, 3(1-4):65–73, 2003.

32. N. Vinckenbosch, I. Dupanloup, and H. Kaessmann. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A*, 103(9):3220–5, 2006.
33. T. Watanabe, Y. Totoki, A. Toyoda, M. Kaneda, S. Kuramochi-Miyagawa, Y. Obata, H. Chiba, Y. Kohara, T. Kono, T. Nakano, M. A. Surani, Y. Sakaki, and H. Sasaki. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, 453(7194):539–43, 2008.
34. Z. Yang. Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13(5):555–6, 1997.
35. Z. Zhang, N. Carriero, and M. Gerstein. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet*, 20(2):62–7, 2004.
36. Z. Zhang, N. Carriero, D. Zheng, J. Karro, P. M. Harrison, and M. Gerstein. Pseudopipe: an automated pseudogene identification pipeline. *Bioinformatics*, 22(12):1437–9, 2006.