The Birth of New Genes by RNA- and DNA-mediated Duplication During Mammalian Evolution

Jin Jun¹, Paul Ryvkin², Edward Hemphill³, Ion Mandoiu¹, and Craig Nelson³*

(1) Computer Science & Engineering Department, University of Connecticut, Storrs, CT 06269,
(2) Genomics and Computational Biology Graduate Group, University of Pennsylvania Biomedical Graduate Studies, Philadelphia, PA 19104,
(3) Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269

* To whom all correspondence should be addressed.

Contact information (telephone, fax, and e-mail address)
Jin Jun: (860) 486-1576, (860) 486-1936, jinjun@engr.uconn.edu
Paul Ryvkin: (215) 746-2807, (215) 573-3111, pry@mail.med.upenn.edu
Edward Hemphill: (860) 486-1576, (860) 486-1936, edward.hemphill_iii@uconn.edu
Ion Mandoiu: (860) 486-3784, (860) 486-4817, ion@engr.uconn.edu
Craig Nelson: (860) 486-5617, (860) 486-1936, craig.nelson@uconn.edu

Running title:
Gene duplication during mammalian evolution

KeyWords:
Retrotransposition, RNA-mediated duplication, DNA-mediated duplication, Gene duplication, Segmental duplication
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. In this study we utilize pseudogenes to recreate more complete gene family histories and compare the rates of RNA and DNA-mediated duplication and new functional gene formation in five mammalian genomes. We find that RNA-mediated duplication occurs at a much higher and more variable rate than DNA-mediated duplication, and gives rise to many more duplicated sequences over time. We show that while the chance of RNA-mediated duplicates becoming functional is much lower than that of their DNA-mediated counterparts, the higher rate of retrotransposition leads to nearly equal contributions of new genes by each mechanism. We also find that functional RNA-mediated duplicates are closer to neighboring genes than non-functional RNA-mediated copies, consistent with cooption of regulatory elements at the site of insertion. Overall new genes derived from DNA and RNA-mediated duplication mechanisms are under similar levels of purifying selective pressure, but have broadly different functions. RNA-mediated duplication gives rise to a diversity of genes but is dominated by the highly expressed genes of RNA metabolic pathways. DNA-mediated duplication can copy regulatory material along with the protein coding region of the gene and often gives rise to classes of genes whose function are dependent on complex regulatory information. This mechanistic difference may in part explain why we find that mammalian protein families tend to evolve by either one mechanism or the other, but rarely by both.
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 (Ohno 1970), a number of recent studies have suggested that variation in gene family size may be even more widespread than previously appreciated (Fortna et al. 2004) and that gene copy number variation between individuals may account for differences in disease predisposition within populations (McCarroll and Altshuler 2007; Redon et al. 2006).

Four primary mechanisms of gene duplication have been described (for review see (Hahn in press; Li 1997): whole genome duplication or polyploidization (Hurley et al. 2005; Skrabanek and Wolfe 1998; Van de Peer et al. 2003), unequal crossing over or tandem duplication (Shoja and Zhang 2006), duplicative transposition (Bailey et al. 2003; Friedman and Hughes 2004; Samonte and Eichler 2002), and retrotransposition (Harrison et al. 2005; Zhang et al. 2002). Of these, whole genome duplication, tandem duplication, and duplicative transposition are DNA-mediated duplication events (abbreviated here as DD for DNA Duplication), while retrotransposition is RNA-mediated (abbreviated here as RD for RNA Duplication). Whole genome duplication has been important to the evolution of many lineages (Jaillon et al. 2004; Van de Peer et al. 2003), but it is a relatively rare event. Unlike whole genome duplication events, tandem duplication and duplicative transposition (sometimes collectively referred to as 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 newly retrotransposed gene copies are duplicated by an RNA-mediated mechanism they lack the promoter and other flanking regulatory sequences of the parental gene. For this reason, retroduplication events have long been believed to give rise primarily to non-functional pseudogenes (Petrov and Hartl 1999; Shemesh et al. 2006). 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 (Sakai et al. 2007; Svensson et al. 2006; Vinckenbosch et al. 2006). Very recently retrotransposition has also been shown to contribute siRNA genes to the genome (Tam et al. 2008; Watanabe et al. 2008).

In this study we compare the rates of new gene formation by DNA-mediated duplication (DD)
and RNA-mediated duplication (RD) in five eutherian genomes. We show that, while genes arising from DD events are up to six times more likely to remain functional than those arising from RD events, the number of RD events is nearly ten times that of DD events, resulting in roughly equal quantitative contributions of new genes from each mechanism. Our analysis further shows that (1) duplicate genes generated by each mechanism are under similar levels of constraint on their protein coding regions but belong to distinctly different classes of genes, (2) gene families tend to evolve by one duplication mechanism or the other, but rarely both, (3) silent site substitution profiles of RD duplicate copies are consistent with bursts of retrotransposition during mammalian evolution, while DNA-mediated duplication appears to occur at a more stable rate, and (4) RD copies under stabilizing selective pressure have insertion sites significantly closer to the nearest genes.

2. Methods

2.1 Dataset

Protein sequences for the five species analyzed (human, chimp, mouse, rat and dog) were obtained from Ensembl (release 37) (Hubbard et al. 2005). For genes with multiple alternative transcripts we developed a collapsed gene model that incorporates all predicted exons of the gene. Resulting exon coordinates were used to obtain a representative protein sequence (see Supplementary Data for detail) that was 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 (Zhang et al. 2006) 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 their related pseudogenes.
2.2 Identification of RD and DD 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 (Altschul et al. 1990) 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 \(s_{i,j} = 2\) 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 \(s_{i,j} = 1\) only if the other side had at least two homologous neighbors; otherwise we assigned a synteny level \(s_{i,j} = 0\) (see Supplementary Data for justification of synteny clustering).

Local synteny levels were used in a two-stage clustering algorithm (see Figure 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 DD 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 (also predicted to be orthologs or paralogs from DD events). 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 RD events or DD events that have lost local synteny. 3,018 clusters out of 27,869 clusters from the above two-stage clustering algorithm were ambiguous and were not considered as RD events. 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.
**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:**

$C \leftarrow \emptyset$  // Collection of clusters  
$U \leftarrow \{ g_1, g_2, \ldots, g_N \}$  // Unassigned family members

**Stage 1)** Single-linkage clustering with synteny level 2:

While $U \neq \emptyset$ do

Select an arbitrary member $g_i$ of $U$

$U \leftarrow U \setminus \{ g_i \}$; $C_{open} \leftarrow \{ g_i \}$  // Start new core cluster

While there exists $g_j \in U$ with synteny 2 to a member of $C_{open}$, do

$U \leftarrow U \setminus \{ g_i \}$; $C_{open} \leftarrow C_{open} \cup \{ g_j \}$  // Add $g_j$ to core cluster

$C \leftarrow C \cup C_{open}$  // Add the new core cluster to the collection of clusters

**Stage 2) Merging of clusters with high average pairwise synteny:**

While there is a pair $(C_l, C_m)$ 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 and 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

Figure 1. Two-stage clustering 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.
Within each cluster produced by the above clustering algorithm there may be successive DD events. We used UPGMA (Unweighted Pair Group Method with Arithmetic mean) (Sneath and Sokal 1973) to find these DD 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, $(n_u^i + n_d^i)_k$ and $(n_u^j + n_d^j)_k$, 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 DD 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 (Figure 2). Remaining inner nodes are ambiguous between speciation and DD events followed by gene loss and were disregarded. In total, 39,673 inner tree nodes were classified as speciation events, 2,035 as DD events, and 1,642 were ambiguous.

We distinguish between putative descendants of RD events or DD events that have lost local synteny using intron conservation scores between descendant 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 (Rogozin et al. 2003). An event was identified as an RD duplication if the average intron conservation rate to paralogs outside the cluster was below 1/3 (see Supplementary Data for supporting data).

For RD events that do not have Ensembl gene models we lack Ensembl intron predictions. Accordingly, for these RD duplicates we used PseudoPipe intron predictions (Zhang et al. 2006). We filtered out RD events for which more than half of the descendant PseudoPipe predicted pseudogenes are annotated as probable intron bearing copies. Of 11,193 RD events, we filtered out 1,666 RD events: 1,645 of these have only intron-bearing pseudogene predictions, and 21 events give rise to more intron-bearing pseudogenes than intron-free pseudogenes (or genes). We retained 202 RD events that gave rise to primarily intron-free predictions with only one intron-bearing prediction, however these represent only ~1% of the total number of RD events.
Figure 2. Inferring DD and RD events using local synteny and hierarchical clustering. This example shows how DD and RD 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 RD event. UPGMA builds hierarchical clusters within each syntenic cluster and speciation and DD events are inferred based on species sets.

2.3 Event Assignment to Tree Branches and Evidence of Function

We used parsimony to assign each inferred duplication event to a specific branch of the 5-species tree. We assigned 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-frame stop codons) of the protein coding sequence 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 of genes within the cluster. In order to avoid the disruption by subsequent DD events, we computed average dN, dS, and dN/dS measures for the descendants of a duplication event assigned to branch b considering only the pairs of genes and pseudogenes coming from different lineages rooted at b. Pairwise dN and dS measures were estimated using the YN00 program of PAML (Yang 1997).

2.4 Distance to the Nearest Ensembl Genes from RD Copies

For RD genes and processed pseudogenes located in intergenic regions, the distances from the RD copies to the nearest upstream/downstream Ensembl genes on the same/opposite strand were measured. The distance was defined by the difference between the starting point of predicted coding sequence (intact or interrupted) and the starting point of the coding sequence of neighboring Ensembl genes.

2.5 Gene Ontology Analysis

The correlation between the duplication mechanism and gene families was measured using the GOstat web tool (Beissbarth and Speed 2004) with default parameters. We used the sets of genes from the 10 families most abundant in RD and DD events, and reported the overrepresented GO terms with a p-value below 0.1. For families with one or two duplication events we selected mid-size families having between 7 and 17 Ensembl annotated genes and RD-only or DD-only events (594 RD-only families and 250 DD-only families). We performed the GOstat analysis for biological process terms on this list, using a minimal GO path length of 5.

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 RD events, while events giving rise to clusters of genes with high local synteny to parental genes were classified as DD 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 DD events, 12,507 RD events, and 2,742 ambiguous events. Using parsimony to assign non-ambiguous events to branches of the species tree resulted in 52 DD and 45 RD events on the branch leading to primates and rodents (the in-group), 161 DD and 1,782 RD events on the primate branch leading to humans and chimps, and 88 DD and 522 RD events on the rodent branch leading to mice and rats (Figure 3). Gene duplication events for the root and terminal branches of the tree were also counted, but 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 at the root. A number of 386 DD and 429 RD events could not be reliably assigned to specific branches of the tree using parsimony and were also omitted from further analysis.

Figure 3. Numbers of gene duplication events from DNA-mediated duplication (above the line) and RNA-mediated duplication (below the line). Numbers represent the assigned DD or RD 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 DD events on primate branch. Evolutionary ages are based on (Ureta-Vidal et al. 2003).
Duplication event counts on the three internal branches of the tree reveal an excess of RD events over DD events along all but the deepest branches of the tree, suggesting an average rate of RD copy formation 3-10 times higher than that of DD copy formation (Figure 3). 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 RD events not under purifying selective pressure (Marques et al. 2005).

3.2 Rates of Duplication

Rates of retrotransposition vary significantly over time and bursts of retrotransposition have been reported in several mammalian lineages (Marques et al. 2005; Zhang et al. 2004). The synonymous substitution rate (dS) profiles of the duplicates identified in this study (Figure 4) 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. RD events in all three internal branches show clear peaks in dS (Figure 4A). 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 ~ 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.

DD events show a similar distribution in dS but a more uniform distribution of dS values than RD duplicates (Figure 4B and C), suggesting that DNA-mediated 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 DD events is more uniform than that of the RD duplicates but is not perfectly flat, suggesting that there may be some variation in the rate of DD events over evolutionary time.
3.3 Preservation Rates of Functional Duplicate Copies

In Figure 3, along the in-group branch leading to primates and rodents DD and RD events contribute roughly equal numbers of intact events. However as we move more distally along the tree, intact RD events begin to outnumber intact DD events. While the numbers of intact DD and RD events are roughly equal along the rodent branch, RD intact events outnumber DD intact events almost three fold on the primate branch. Terminal branches of the tree display an even greater excess of RD intact events, ranging from nearly equal numbers in humans and chimps, to a near three-fold excess in mice, and a roughly four-fold excess in rats and dogs. The progressive increase in intact RD events over intact DD events in more recent branches of the tree is likely to be due in part to recent bursts of retrotransposition, and in part due to the persistence of intact protein coding regions in young RD copies that may not be functional, but have not yet experienced inactivating mutations of their protein coding regions.

The progressive influence of purifying selection over time is readily observed in dN vs. dS
plots of DD and RD events from the three internal lineages. In Figure 5, RD events on the primate branch are compressed near the origin, consistent with a recent burst of retrotransposition in this lineage, while DD events display a more even age distribution. Non-intact events are generally interspersed with intact ones, except for the rodent lineage in which the effect of prolonged purifying selection results in the separation of inactivated RD events from the intact RD copies. In the in-group there are very few inactivated duplication events due to the difficulty in finding very old pseudogenes. However, some inactivated events remain interspersed among intact ones, suggesting that the gene duplicates resulting from these events were under purifying selection and may have only recently suffered interruption of their protein coding regions. Alternatively, these genes may encode partial protein products that remain under purifying selection.

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 (Nekrutenko et al. 2002). For example, dN/dS ratio < 0.5 has been used as stringent functionality criteria between retrotransposed genes and their parental genes (Emerson et al. 2004). 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 (Torrents et al. 2003). 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.
Figure 5. Dot-plots of averaged dN and dS over all the pairs of genes and pseudogenes within clusters resulting from DD intact events (black dots in the upper three graphs), DD inactivated events (gray dots in the upper three graphs), RD intact events (black dots in the lower three graphs) and RD inactivated events (gray dots in the lower three graphs) for three internal branches: in-group, primate and rodent branches. The dashed lines on the dot-plots are the line with dN=dS (or dN/dS=1) and the solid lines are the trend lines of the intact events per each panel.

Analysis of the dN/dS ratios of clusters derived from duplication events is quite revealing. Figure 6A compares clusters of RD duplication event descendants with intact protein coding reading frames (intact) and clusters of RD 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 \( \sim 0.5 \). 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 (\(< 0.5\)) are likely to be under
stabilizing selection for protein coding function.

Figure 6. (A) Histograms of average dN/dS ratio over pairs of Ensembl genes for clusters resulting from intact RD events and average dN/dS ratio over pairs of genes and pseudogenes for clusters resulting from inactivated RD events on the rodent lineage. Histograms of average dN/dS ratio over pairs of Ensembl genes for clusters resulting from intact DD events and RD events on the (B) in-group branch leading to primates and rodents, (C) rodent, and (D) primate.

Panels B through D of Figure 6 compare dN/dS values of duplicate clusters derived from RD and DD 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 RD and DD events that gave rise to putatively functional gene copies (aggregate 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 (Figure 5B).

3.4 Relative Position of RD Copies to the Other Genes

In order to address the question of why some RD events are under the stabilizing selective pressure (dN/dS ration \( \leq 0.5 \)), we analyzed the location of RD copies. We categorized RD copies as either genic (mostly intronic) or intergenic, and measured the distance to the nearest Ensembl genes. Three classes of RD copies were defined: intact RD genes with dN/dS \( \leq 0.5 \) (SS-RD), intact RD genes with dN/dS > 0.5 (NI-RD) and processed pseudogenes (PP-RD). Table 1 shows the number and percentage of RT copies on genic and intergenic sequences and corresponding p-values (\( \chi^2 \) test using the proportion of genic area in the whole genome). Although all the three classes of RD copies were less likely to be found inside other genes (all p-values <1E-06), processed pseudogenes (PP-RD) were found inside other genes two or three times more often than intact RD copies (SS-RD and NI-RD).

<table>
<thead>
<tr>
<th>Numbers (%)</th>
<th>SS-RD</th>
<th>NI-RD</th>
<th>PP-RD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genic</td>
<td>23 ( 4.3%)</td>
<td>23 ( 4.5%)</td>
<td>432 (11.2%)</td>
</tr>
<tr>
<td>Intergenic</td>
<td>510 (95.7%)</td>
<td>485 (95.5%)</td>
<td>3,411 (88.8%)</td>
</tr>
<tr>
<td>p-Value*</td>
<td>5.90E-08</td>
<td>3.23E-07</td>
<td>2.58E-07</td>
</tr>
<tr>
<td>Genic</td>
<td>70 (13.1%)</td>
<td>73 (14.4%)</td>
<td>996 (25.9%)</td>
</tr>
<tr>
<td>Intergenic</td>
<td>463 (86.9%)</td>
<td>435 (85.6%)</td>
<td>2,847 (74.1%)</td>
</tr>
<tr>
<td>p-Value*</td>
<td>1.14E-11</td>
<td>7.28E-10</td>
<td>8.55E-07</td>
</tr>
</tbody>
</table>

Table 1. The numbers of retrotransposed insertions on genic vs. intergenic sequence by RD events on three internal branches: in-group, primates and rodents. Three classes of RD copies were used: intact RD genes with dN/dS \( \leq 0.5 \) (SS-RD), intact RD genes with dN/dS > 0.5 (NI-RD) and processed pseudogenes (PP-RD). *: \( \chi^2 \) test p-value using the proportion of genic area in the whole genome.
For intergenic RD copies SS-RD were consistently closer to neighboring genes than PP-RD regardless of strand or orientation (up- or down-stream, all p-values <0.001, Wilcoxon / Kruskal-Wallis test and Student’s t-test, Figure 7). This is consistent with retrotransposed duplicates co-opting preexisting regulatory elements. Not surprisingly, distances to upstream neighboring genes on the opposite strand (in a head-to-head configuration) have the shortest median distance to intact RD copies (95.8KB for SS-RD and 138.7KB for NI-RD) (Figure 7C).

Figure 7. Relative location of RD events to Ensembl genes. Three classes of RD copies were used: intact RD genes with dN/dS ≤ 0.5 (SS-RD), intact RD genes with dN/dS > 0.5 (NI-RD) and processed pseudogenes (PP-RD). (A) Distance to the nearest upstream gene on the same strand, (B) distance to the nearest downstream gene on the same strand, (C) distance to the nearest upstream gene on the opposite strand (head-to-head), (D) distance to the nearest downstream gene on the opposite strand (tail-to-tail). Boxes represent interquartile range with the horizontal line being the median; diamonds span a 95% confidence interval around the mean assuming normality. The vertical lines span the extents of 95% of a normal distribution fit to the data. Three classes have significant different distance measures (p-value<0.0001 for (A), (C) and (D), p-value<0.001 for (B), Wilcoxon / Kruskal-Wallis Test). SS-RD and PP-RD have significantly different mean (p-value<0.0001 for (A) and (D), p-value<0.001 for (B) and (C), Student’s t-test).
3.5 Distribution of Duplication Events within the Mammalian Tree

The total number of RD and DD duplication events detected in this study is illustrated in Figure 3. 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, showing a fraction of the functional events over the total assigned events, e.g. 53/161 for DD events on primate branch. From these numbers it is clear that we detect far more RD events than DD events, but that far fewer of these events give rise to functional protein coding genes than their DD counterparts. Analysis of the internal branches individually reveals small 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 DD events over functional RD 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 DD events, but in primates fewer of these events give rise to functional descendants (Table 2). A decreased rate of functionalization is also apparent in the RD events on the primate lineage. Despite an RD event rate nearly twice that seen in rodents, the number of functional RD events in primates is only ~25% greater than that in rodents.

Table 2. Rates of duplication events for rodent and primate lineages.

<table>
<thead>
<tr>
<th>Events per million yrs</th>
<th>DD events</th>
<th>RD events</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assigned</td>
<td>Intact</td>
</tr>
<tr>
<td>Rodents</td>
<td>1.76</td>
<td>1.56</td>
</tr>
<tr>
<td>Primates</td>
<td>1.87</td>
<td>1.31</td>
</tr>
</tbody>
</table>

3.6 Distribution of Functional Events in Gene Families

Of 8,872 Ensembl families, 262 families have at least one functional duplication event. The distribution of functional events within gene families of different sizes (Figure 8) reveals that the two duplication mechanisms make similar contributions to families of all sizes. However, there is an apparent lack of mixing of the two types of duplication events within families.
Figure 8. Distribution of functional events categorized by RD events in RD only families, DD events in DD only families and RD events and DD events in the families including both events binned by the number of Ensembl genes within family.

Of the 262 families containing at least one functional event, 151 (57.6%) families have exclusively functional RD events and 101 (38.5%) have exclusively functional DD events, suggesting that gene families primarily evolve either by DNA-mediated duplication, or by RNA-mediated duplication (retrotransposition), but rarely by both mechanisms. We tested this idea by looking more closely at the distribution of functional RD and DD events in families with at most 4 functional events. We counted families that contained only DD or only RD functional events and compared these observations to the numbers of families expected based on a binomial distribution of the of two types of events, where the probability of DD functional events is $p=148/335$ (total number of functional DD events / total number of functional events) and the probability of RD functional events is $q=1-p$ (Table 3). Chi-squared tests reveal that the observed mutual exclusivity of these two mechanisms is statistically highly significant.
Table 3. Observed numbers of families having only RD functional events and only DD functional events and $\chi^2$ p-values.

<table>
<thead>
<tr>
<th># Events</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>&gt;4</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD only families</td>
<td>14</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>RD only families</td>
<td>16</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Total families</td>
<td>37</td>
<td>7</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>p-Value</td>
<td>0.007017</td>
<td>0.000655</td>
<td>3.63E-07</td>
<td></td>
</tr>
</tbody>
</table>

3.7 DNA- and RNA-mediated Duplications Give Rise to Different Types of Gene Families

As the mechanisms that give rise to these two types of duplications have very different consequences for the properties of the duplicate copy, it might be reasonable to expect that each mechanism might be biased in the type of new functional genes it creates. We examined the types of genes created by each duplication mechanism by identifying Gene Ontology terms that are overrepresented in families derived from each mechanism. As might be expected the largest gene families generated through RD duplications are dominated by ribosomal proteins (see Supplementary Data for detail). The very high levels of expression of these genes lead to a large number of intact RD events and a correspondingly large number of new gene copies. In contrast, the largest families of DD generated genes include a variety of functional categories including immune function (lipocalins, chemokines, and defensins), and large families of diverse molecules such as the olfactory receptors.

Analysis of the type of genes born through each mechanism in the mid-size families (7-17 members) shows a similar trend. Overrepresented GO terms for RD dominated families again include a number of RNA related categories However, when the contribution of RD genes is analyzed family by family, a much greater diversity of functions is revealed including members of the nuclear pore complex, topoisomerase, DNA binding proteins, cell cycle regulation, apoptosis, and energy metabolism. DD dominated mid-size families are involved in a variety of processes requiring more complex regulation of gene expression including the regulation of development, immune processes and odorant perception. Full lists of the gene families and GO terms can be found in Supplementary Data.
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 miss-assignment of some duplication events to younger branches and a consequent under-estimation of the age of these gene families. But using local synteny to help classify duplication events works 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 the local genome structure surrounding duplicate copies and intron content to differentiate between DNA and RNA based duplications and to order successive DD 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 (see Supplementary Data: Figure D). The very low false-positive rate (0.2%~0.7%) associated with measures of local synteny means that genes that share synteny with paralogs are almost certainly the result of DNA-mediated duplications regardless of intron content. Therefore this method is unlikely to misclassify RD duplicates as DD duplicates. DNA-mediated duplications, however, can lose synteny to their paralogs over time (Huynen and Bork 1998; Rocha 2006), which may result in some DD class being mis-assigned to the RD class. To account for this we use conventional intron content criteria (Vinckenbosch et al. 2006) to further discriminate between non-syntenic DNA based duplications and RD copies. Duplicate pairs that maintain synteny with their paralog are most likely DNA based, while non-syntenic paralogs with significant intron loss are likely RD duplicates. Comparison with other studies identifying RD duplicates in mammalian genomes suggest that using synteny criteria in addition to intron
content improves the reliability of RD duplicate classification and that duplications identified as RD duplicates on the basis of intron content alone may in fact be DD 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 DD events. Gene families generated by rounds of DNA-mediated 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 actually 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 judged to be quite accurate (Consortium 2004; She et al. 2008; Vinckenbosch et al. 2006). 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 (Demuth et al. 2006).

4.2 Rates of duplication

Lineage specific gene duplication, by RNA-mediated or DNA-mediated 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 (Demuth et al. 2006). Retrotransposition, in particular, appears to be peppering the genome with large numbers of duplicate retrocopies that can act as both insertional mutagens (Mills et al. 2007) and new duplicate genes (Vinckenbosch et al. 2006), and siRNA’s (Tam et al. 2008; Watanabe et al. 2008). Analysis of retrotransposon activity during vertebrate evolution shows strong peaks of activity (Ohshima et al. 2003) and it is therefore not surprising that RD duplication of genes
shows similar peaks in birth rates. DNA-mediated 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 DD events identified in our study are more broadly distributed than the RD age profiles, but interestingly, they are not perfectly uniform over time and may indicate of bursts of DNA-mediated duplication activity in the evolutionary history of these genomes (see also Bailey et al. 2002 and She et al. 2006)

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 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 (Figure 5). 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 Figure 6A and C). For the young primate branch the divergence between functional and non-functional descendants is less clear. At virtually all time-points, 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 RD duplicates and DD duplicates would show different
degrees of purifying selective pressure. Interestingly, in age-matched populations of DD and RD
duplicates, there is no dramatic difference in selection pressure on genes born by these two
mechanisms (Figure 6). 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
(150/2304) of RD events give rise to duplicates showing evidence of purifying selection, while
forty percent (101/249) for DD events appear to generate functional descendants (Figure 3). The
very high rate of RD events coupled with the very low rate of functionalization of gene copies
generated by these events, and the lower rate of DD 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.

4.4 Relative Location of Functional RD Copies

Vinckenbosh et al. showed that there are significant more transcribed retrocopies inside other
genes (42.1%) than transcribed intergenic retrocopies (24.8%) (Vinckenbosch et al. 2006). They
found that 29.6% (170 of 575) of intact retrocopies are inside other genes, while 30.8% (929 of
3,015) pseudogenes are inside other genes. These ratios are far higher than our results (Table
1). This difference is likely to be due to the age of the retrocopies identified in these two studies. We
only analyzed RD copies from the three internal branches of our phylogeny, while Vinkenbosh et
al. analyzed _all_ human retrocopies, including those generated since the split from chimpanzee,
which were excluded form our study as too young to be reliably assessed for function. When we
constrain the comparison with the Vinckenbosh data to only those human RD copies with mouse
orthologs, the percentages of the retrocopies and pseudogenes inside other genes decreases to
17.3% and 27.4% respectively, which are comparable to our strand independent results from the
in-group branches in our study (13.1% to 14.4% and 25.9%).

4.5 Family Size and Gene type

While there is little difference between the numbers of new genes contributed to the genome
by these RNA-based and DNA-based duplication mechanisms, we anticipated major differences
in both the size of the gene families created by each mechanism and the types of genes successfully duplicated by each. The first of these expectations proved inaccurate, but revealed a previously unanticipated phenomenon. The size distribution of families generated by RD and DD are virtually identical (see Figure 8). Both very large and very small gene families may be comprised of DD or RD duplicates. Interestingly however, very few families (far fewer than expected by chance) are comprised of duplicates generated by both mechanisms (Table 3). This may be due, in part, to the possibility that some types of genes are more likely to survive duplication in the absence of specific regulatory information. If so, some gene families would be predicted to have no functional descendants of RD events. Why other families might lack functional DD duplicates is less clear.

The possible role of gene function in the tendency for gene families to evolve by one mechanism or the other is supported by an analysis of the types of functions attributed to families derived from duplications of each type. Unsurprisingly, many of the largest and/or most over-represented RD gene families are associated with the ribosome or other aspects of RNA metabolism. The very high levels of expression of ribosomal constituents should lead to correspondingly high levels of cDNA for these genes. During periods of active retro transcription abundant cDNA levels would lead to higher levels of newly inserted retrocopies for these genes than genes expressed at lower levels. In smaller and midsize gene families, however, while ribosomal constituents are still over-represented, many other functions can be found. Metabolic enzymes like dehydrogenases and phosphatases, and a variety of other functions like cyclins, and zinc finger and beta-HLH transcription factors are represented in the minority of functional RD duplicates. DNA-mediated duplications, on the other hand, show over-representation of genes involved in developmental and immune related processes. Large families like the Hox genes, olfactory receptors, T-cell receptors and defensins have all evolved through successive DNA-mediated duplications. Unlike many metabolic proteins, many of these genes require very precise control of gene expression patterns for their function and therefore, in the absence of extremely fortuitous enhancer-trap events would likely require duplication of regulatory DNA for their continued function. Like RD gene families, however, a great diversity of functions is represented in the entire population of DD derived genes. The diversity and abundance of new genes created by both duplication mechanisms illustrates the importance of both mechanisms to genome evolution.
Acknowledgement

A preliminary version of this work (Jun et al. 2008) appeared in the proceedings of the 6th RECOMB Workshop on Comparative Genomics (RECOMB-CG’08). The authors thank the associated editor and the anonymous reviewers for their valuable suggestions.

Disclosure Statement

No competing financial interests exist.

References


### Supplementary Data

#### A. Top-Ten RD and DD abundant families

<table>
<thead>
<tr>
<th>Ensembl family</th>
<th>Family description</th>
<th># Ensembl genes</th>
<th># Intact RD events</th>
<th># Assigned RD events</th>
<th>GO</th>
<th>GO name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSF00000001265</td>
<td>60S RIBOSOMAL L27A</td>
<td>42</td>
<td>30</td>
<td>112</td>
<td>GO:0005842, GO:0005830, GO:0015934</td>
<td>cytosolic large ribosomal subunit (sensu Eukaryota)</td>
</tr>
<tr>
<td>ENSF00000001824</td>
<td>60S RIBOSOMAL L36</td>
<td>39</td>
<td>27</td>
<td>104</td>
<td>GO:0005840, GO:0030529, GO:0006412</td>
<td>Ribosome</td>
</tr>
<tr>
<td>ENSF00000001192</td>
<td>40S RIBOSOMAL SA P40 34/67 KDA LAMININ</td>
<td>44</td>
<td>24</td>
<td>193</td>
<td>GO:0015935, GO:0005055</td>
<td>small ribosomal subunit, laminin receptor activity</td>
</tr>
<tr>
<td>ENSF00000001687</td>
<td>60S RIBOSOMAL L13</td>
<td>30</td>
<td>23</td>
<td>85</td>
<td>GO:0005830, GO:0044445</td>
<td>cytosolic ribosome (sensu Eukaryota)</td>
</tr>
<tr>
<td>ENSF00000001631</td>
<td>40S RIBOSOMAL S10</td>
<td>37</td>
<td>21</td>
<td>90</td>
<td>GO:0016283, GO:0005843</td>
<td>eukaryotic 48S initiation complex</td>
</tr>
<tr>
<td>ENSF00000002263</td>
<td>40S RIBOSOMAL S12</td>
<td>34</td>
<td>20</td>
<td>95</td>
<td>GO:0016283, GO:0005843</td>
<td>eukaryotic 48S initiation complex</td>
</tr>
<tr>
<td>ENSF00000001209</td>
<td>60S ACIDIC RIBOSOMAL P0</td>
<td>29</td>
<td>19</td>
<td>53</td>
<td>GO:0005842, GO:0006414, GO:0005830, GO:0042254, GO:0043284, GO:0015934</td>
<td>cytosolic large ribosomal subunit (sensu Eukaryota)</td>
</tr>
<tr>
<td>ENSF00000001521</td>
<td>60S RIBOSOMAL L12</td>
<td>33</td>
<td>19</td>
<td>97</td>
<td>GO:0005842, GO:0005830, GO:0015934</td>
<td>cytosolic large ribosomal subunit (sensu Eukaryota)</td>
</tr>
<tr>
<td>ENSF00000002778</td>
<td>60S RIBOSOMAL L35A</td>
<td>25</td>
<td>18</td>
<td>169</td>
<td>GO:0000049</td>
<td>tRNA binding</td>
</tr>
<tr>
<td>ENSF00000001964</td>
<td>40S RIBOSOMAL S7</td>
<td>30</td>
<td>18</td>
<td>76</td>
<td>GO:0016283, GO:0005843</td>
<td>eukaryotic 48S initiation complex</td>
</tr>
</tbody>
</table>

Table A-1. Top 10 RD-abundant families ordered in decreasing order of the number of intact RD events. All families here have no DD events and at least 18 RD events.
<table>
<thead>
<tr>
<th>Ensembl family</th>
<th>Family description</th>
<th># Ensembl genes</th>
<th># Intact DD events</th>
<th># Assigned DD events</th>
<th>GO</th>
<th>GO name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSF00000001153</td>
<td>Lipocalin-related protein and Bos/Can/Equ allergen</td>
<td>49</td>
<td>34</td>
<td>44</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>ENSF00000000477</td>
<td>Trace amine associated receptor / tar</td>
<td>42</td>
<td>22</td>
<td>28</td>
<td>GO:0001584, GO:0008227</td>
<td>rhodopsin-like receptor activity, amine receptor activity</td>
</tr>
<tr>
<td>ENSF00000000443</td>
<td>Carboxylesterase precursor EC_3.1.1.1</td>
<td>45</td>
<td>20</td>
<td>28</td>
<td>GO:0004759, GO:0004091</td>
<td>carboxylesterase activity; serine esterase activity</td>
</tr>
<tr>
<td>ENSF00000000958</td>
<td>Small chemokine</td>
<td>34</td>
<td>14</td>
<td>16</td>
<td>GO:0008009, GO:0042379</td>
<td>chemokine receptor binding; chemokine activity</td>
</tr>
<tr>
<td>ENSF00000002170</td>
<td>Proteinase inhibitor I25, cystatin</td>
<td>27</td>
<td>14</td>
<td>15</td>
<td>GO:0004869, GO:0004866</td>
<td>endopeptidase inhibitor activity; cysteine protease inhibitor activity</td>
</tr>
<tr>
<td>ENSF00000000691</td>
<td>Sialic acid binding IG lectin precursor siglec</td>
<td>41</td>
<td>13</td>
<td>22</td>
<td>GO:0016021, GO:0031224</td>
<td>membrane part; intrinsic to membrane</td>
</tr>
<tr>
<td>ENSF00000001948</td>
<td>Apolipoprotein</td>
<td>25</td>
<td>13</td>
<td>18</td>
<td>GO:0042157, GO:0006869</td>
<td>lipoprotein metabolic process, lipid transport</td>
</tr>
<tr>
<td>ENSF00000001951</td>
<td>Cystatin precursor</td>
<td>28</td>
<td>12</td>
<td>16</td>
<td>GO:0004869, GO:0004866</td>
<td>endopeptidase inhibitor activity; cysteine protease inhibitor activity</td>
</tr>
<tr>
<td>ENSF00000002079</td>
<td>Olfactory receptor 8S1</td>
<td>27</td>
<td>12</td>
<td>14</td>
<td>GO:0007608, GO:0007606, GO:0004984</td>
<td>sensory perception of chemical stimulus; sensory perception of smell, olfactory receptor activity</td>
</tr>
<tr>
<td>ENSF00000001990</td>
<td>Defensin related cryptdin precursor</td>
<td>19</td>
<td>11</td>
<td>15</td>
<td>GO:0042742, GO:0009617</td>
<td>response to bacterium; defense response to bacterium</td>
</tr>
</tbody>
</table>

Table A-2. Top 10 DD-abundant families ordered in decreasing order of the number of intact DD event. All families here have more than 10 DD events.
B. GO term analysis for the mid-sized families

The RD/DD abundant families (Supp. A) are from the long tail of the distribution, but the majority of RD/DD events are in the mid sizes of families (Figure 8). We listed the overrepresented (P-Value < 0.1) **biological process** GO terms (Top 10s) from RD event only families with sizes of [7-17] (594 families, 6052 genes, 736 annotated Hs gene names, 767 annotated Mm gene names), with minimal length of GO path as 5. Most of the overrepresented GO terms are overlapped between human and mouse, while all GO terms are ribosomal or RNA related ones.

<table>
<thead>
<tr>
<th>GO Term</th>
<th>P-value (Hs)</th>
<th>P-value (Mm)</th>
<th>GO Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0016071</td>
<td>7.87e-41</td>
<td>4.4e-20</td>
<td>mRNA metabolic process</td>
</tr>
<tr>
<td>GO:0006397</td>
<td>8.57e-40</td>
<td>1.74e-21</td>
<td>mRNA processing</td>
</tr>
<tr>
<td>GO:0022613</td>
<td>7.24e-38</td>
<td>2.08e-11</td>
<td>ribonucleoprotein complex biogenesis and assembly</td>
</tr>
<tr>
<td>GO:0006396</td>
<td>1e-37</td>
<td>3.31e-19</td>
<td>RNA processing</td>
</tr>
<tr>
<td>GO:0006605</td>
<td>3.17e-21</td>
<td></td>
<td>protein targeting</td>
</tr>
<tr>
<td>GO:0006139</td>
<td>3.82e-18</td>
<td>7.26e-14</td>
<td>nucleobase, nucleoside, nucleotide and nucleic acid metabolic process</td>
</tr>
<tr>
<td>GO:0051246</td>
<td>6.12e-18</td>
<td></td>
<td>regulation of protein metabolic process</td>
</tr>
<tr>
<td>GO:0043170</td>
<td>5.39e-16</td>
<td>3.37e-14</td>
<td>macromolecule metabolic process</td>
</tr>
<tr>
<td>GO:0008380</td>
<td>7.37e-16</td>
<td>1.39e-23</td>
<td>RNA splicing</td>
</tr>
<tr>
<td>GO:0044249</td>
<td>1.05e-14</td>
<td></td>
<td>cellular biosynthetic process</td>
</tr>
<tr>
<td>GO:0000377</td>
<td>2.91e-08</td>
<td></td>
<td>RNA splicing, via transesterification reactions with bulged adenosine as nucleophile</td>
</tr>
<tr>
<td>GO:0000398</td>
<td>2.91e-08</td>
<td></td>
<td>nuclear mRNA splicing, via spliceosome</td>
</tr>
<tr>
<td>GO:0000375</td>
<td>2.91e-08</td>
<td></td>
<td>RNA splicing, via transesterification reactions</td>
</tr>
</tbody>
</table>

Table B-1. Top 10 GO terms from GOstat analysis on the human and mouse genes from RD only families with size between 7 and 17.
Overrepresented (P-Value < 0.1) biological process GO terms (Top 10s) from DD event only families with sizes of [7-17] (250 families, 2668 genes, 340 annotated Hs gene names, 379 annotated Mm gene names), with minimal length of GO path as 5, show different result that there is no overlapping between two GO term lists from human genes and mouse genes. The overrepresented GO terms from human genes are mostly development related ones, while mouse genes are associated with sensory ones.

<table>
<thead>
<tr>
<th>GO Term</th>
<th>P-value (Hs)</th>
<th>P-value (Mm)</th>
<th>GO Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006954</td>
<td>0.00042</td>
<td></td>
<td>inflammatory response</td>
</tr>
<tr>
<td>GO:0048731</td>
<td>0.000424</td>
<td></td>
<td>system development</td>
</tr>
<tr>
<td>GO:0009611</td>
<td>0.000654</td>
<td></td>
<td>response to wounding</td>
</tr>
<tr>
<td>GO:0048513</td>
<td>0.000772</td>
<td></td>
<td>organ development</td>
</tr>
<tr>
<td>GO:0048518</td>
<td>0.00473</td>
<td></td>
<td>positive regulation of biological process</td>
</tr>
<tr>
<td>GO:0016998</td>
<td>0.00473</td>
<td></td>
<td>cell wall catabolic process</td>
</tr>
<tr>
<td>GO:0045682</td>
<td>0.0095</td>
<td></td>
<td>regulation of epidermis development</td>
</tr>
<tr>
<td>GO:0000165</td>
<td>0.0104</td>
<td></td>
<td>MAPK KK cascade</td>
</tr>
<tr>
<td>GO:0008366</td>
<td>0.0128</td>
<td></td>
<td>axon ensheathment</td>
</tr>
<tr>
<td>GO:0007272</td>
<td>0.0128</td>
<td></td>
<td>ensheathment of neurons</td>
</tr>
<tr>
<td>GO:0009617</td>
<td>0.00204</td>
<td></td>
<td>response to bacterium</td>
</tr>
<tr>
<td>GO:0050906</td>
<td>0.00204</td>
<td></td>
<td>detection of stimulus during sensory perception</td>
</tr>
<tr>
<td>GO:0001580</td>
<td>0.00209</td>
<td></td>
<td>detection of chemical stimulus during sensory perception of bitter taste</td>
</tr>
<tr>
<td>GO:0050912</td>
<td>0.00209</td>
<td></td>
<td>detection of chemical stimulus during sensory perception of taste</td>
</tr>
<tr>
<td>GO:0050907</td>
<td>0.00244</td>
<td></td>
<td>detection of chemical stimulus during sensory perception</td>
</tr>
<tr>
<td>GO:0050913</td>
<td>0.00276</td>
<td></td>
<td>sensory perception of bitter taste</td>
</tr>
<tr>
<td>GO:0042742</td>
<td>0.00276</td>
<td></td>
<td>defense response to bacterium</td>
</tr>
<tr>
<td>GO:0000188</td>
<td>0.00378</td>
<td></td>
<td>inactivation of MAPK activity</td>
</tr>
<tr>
<td>GO:0009593</td>
<td>0.00576</td>
<td></td>
<td>detection of chemical stimulus</td>
</tr>
<tr>
<td>GO:0009435</td>
<td>0.0119</td>
<td></td>
<td>NAD biosynthetic process</td>
</tr>
</tbody>
</table>

Table B-2. Top 10 GO terms from GOstat analysis on the human and mouse genes from DD only families with size between 7 and 17.
C. The collapsed gene/protein model

One of the features in metazoan genomes is that alternative splicing is observed in many genes. Two approaches can be used to analyze these genes with transcript variation. One method involves choosing a single transcript (usually the longest transcript) to represent the entire gene. However one transcript cannot explain the whole gene ideally. The other method is to create a collapsed gene in which all of the transcripts of the gene are represented. The collapsed gene model assumes that there is a unique set of exons per gene, which incorporates all the transcript variations from that gene.

As shown in Figure C, only the protein-coding genes on annotated chromosomes were taken for the next step. The exons from the multiple transcripts are collapsed into a single set of exons by following rules: 1) take the union set of exons from multiple transcripts, and 2) merge or concatenate two overlapped or adjoining exons if any. From these new collapsed exons the corresponding introns are also created. Some exceptions were considered to make the collapsed exon coordinates, such as the cases that UTRs on inner exons were counted neither as 5' or 3' UTR of corresponding gene, and that no gene coordinates were affected by any overlapped gene ones.

Figure E shows how a collapsed protein is created from a collapsed gene. Since some of the genes in Ensembl database have multiple transcript phases, we used the first phase information in Ensembl database and had to discard any leftover nucleotide letters between two exons during translation. In Figure C (b) an exemplar gene has two transcripts with different phase, 1 for T1 and 0 for T2, for the fourth exon of collapsed gene model. Since T1 is the first transcript in the database we picked 1 for the phase information and discarded 1 leftover letter for the fourth exon in CP. However because T1 doesn’t have exon at the position of the fifth collapsed exon, the phase information of the fifth collapsed exon is from T2, which is 2. Therefore in this collapsed protein CP, the first, second and fourth exons are parts of the protein from T1 transcript and the third and fifth exons are from T2. This phase abnormality is assumed annotation problem or from the lack of information. Since our goal is not related with protein functionality, this collapsing process will provide us enough consensus protein sequence. The collapsed protein sequences were used as input of any following steps that needed protein sequences.
Figure C. The flowcharts and the example of the collapsed gene/protein model. In example, 'T1' and 'T2' are two transcripts of given gene. 'CG' stands for collapsed gene and 'CP' for collapsed protein.

D. Synteny clustering justification

Any gene pair has two numbers of co-linearly ortholog matches on either side, e.g. 2 matches
for upstream and 0 for downstream or (2, 0) in short (we don’t distinguish (2,0) from (0,2)).

From our preliminary experiments, the synteny measure higher than (1, 0), i.e. (m, n) where
1<m+n≤6, was selected as the threshold for DNA-mediated duplicated genes. In addition to this
threshold, we considered any two sided matches of synteny measures, i.e. (m, n) where 0<m, n≤3
are stronger signal of the DD genes than any one sided matches, e.g. (2, 0) or (3, 0). Therefore
we have three synteny levels, ‘0’ for (0, 0) and (1, 0), ‘1’ for (2, 0) and (3,0) and ‘2’ for any two
sided matches. Since any paralog definitions can include either DD duplicates or RD copies we
use ortholog pairs to assess our local synteny measure. The false positive is between 0.2% and
0.7% from our random pairs comparing the true positive values from known 1:1 orthologs,
ranging from 81.4% to 97.8%. Furthermore it is very unlikely to be retrocopied into orthologous
regions.

<table>
<thead>
<tr>
<th>Compared species</th>
<th>Synteny level</th>
<th>Inparanoid 1:1 ortholog</th>
<th>Ensembl 1:1 ortholog</th>
<th>Random 1:1 orthologs</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS vs. MM</td>
<td>0</td>
<td>5.3%</td>
<td>7.0%</td>
<td>99.8%</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>8.0%</td>
<td>8.7%</td>
<td>0.1%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>86.7%</td>
<td>84.3%</td>
<td>0.1%</td>
</tr>
<tr>
<td>HS vs. PT</td>
<td>0</td>
<td>0.6%</td>
<td>1.8%</td>
<td>99.3%</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1.6%</td>
<td>2.3%</td>
<td>0.5%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>97.8%</td>
<td>95.9%</td>
<td>0.2%</td>
</tr>
<tr>
<td>HS vs. CF</td>
<td>0</td>
<td>3.0%</td>
<td>5.2%</td>
<td>99.8%</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5.7%</td>
<td>6.6%</td>
<td>0.1%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>91.3%</td>
<td>88.2%</td>
<td>0.1%</td>
</tr>
<tr>
<td>MM vs. RN</td>
<td>0</td>
<td>6.9%</td>
<td>9.1%</td>
<td>99.7%</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>8.2%</td>
<td>8.7%</td>
<td>0.1%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>84.9%</td>
<td>82.2%</td>
<td>0.2%</td>
</tr>
<tr>
<td>MM vs. CF</td>
<td>0</td>
<td>6.2%</td>
<td>9.4%</td>
<td>99.7%</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>8.2%</td>
<td>9.2%</td>
<td>0.1%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>85.6%</td>
<td>81.4%</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

Table D. The true positive (TP) probability with Inparanoid 1:1, Ensembl 1:1 ortholog pairs compared to the false positive (FP) probability with randomly chosen ortholog pairs for some species pairs.
Correlation between local synteny and other evidence of DNA copy would be another justification of synteny clustering. Figure D shows that there is a significant correlation between local synteny and intron conservation. Non-syntenic pairs have much less intron conservation rate (ICR) than syntenic pairs’ ICR. Also the distribution of the proportions of three ICR bins show that at younger branch (primate branch), the slopes from non-syntenic pairs and syntenic pairs are steeper in opposite direction, and these slopes started to diminish as moved to older branch. Interestingly even syntenic pairs share their introns with their partner less than expected, meaning that there might have been considerable amount of structure evolution/disruption events on DD copies.

Figure D. The histograms showing the correlation between local synteny and intron conservation at three different branches.