Genome Analysis

Reconstructing ancestral genome content based on symmetrical best alignments and Dollo parsimony

Onur Sakarya1,2, Kenneth S. Kosik2,3 and Todd H. Oakley4

1Department of Computer Science, 2Neuroscience Research Institute, 3Department of Molecular, Cellular and Developmental Biology and 4Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, Santa Barbara, California 93106, USA

ABSTRACT

Motivation: Gene duplications and losses (GDLs) are important events in genome evolution. They result in expansion or contraction of gene families, with a likely role in phenotypic evolution. As more genomes become available and their annotations are improved, software programs capable of rapidly and accurately identifying the content of ancestral genomes and the timings of GDLs become necessary to understand the unique evolution of each lineage.

Results: We report EvolMAP, a new algorithm and software that utilizes a species-tree based gene clustering method to join all-to-all symmetrical similarity comparisons of multiple gene sets in order to infer the gene composition of multiple ancestral genomes. The algorithm further uses Dollo parsimony-based comparison of the inferred ancestral genes to pinpoint the timings of GDLs onto evolutionary intervals marked by speciation events. Using EvolMAP, first we analyzed the expansion of four families of G-protein Coupled Receptors (GPCRs) within animal lineages. Additional to demonstrating the unique expansion tree for each family, results also show that the ancestral eumetazoan genome contained many fewer GPCRs than modern animals, and these families expanded through concurrent lineage-specific duplications. Second, we analyzed the history of GDLs in mammalian genomes by comparing seven proteomes. In agreement with previous studies, we report that the mammalian gene family sizes have changed drastically through their evolution. Interestingly, although we identified a potential source of duplication for 75% of the gained genes, remaining 25% did not have clear-cut sources, revealing thousands of genes that have likely gained their distinct sequence identities within the descent of mammals.

Availability: Query server, source code and executable are available from http://kosik-web.mcdb.ucsb.edu/evolmap/index.htm

Contact: kosik@lifesci.ucsb.edu, oakley@lifesci.ucsb.edu

1 INTRODUCTION

A robust understanding of gene family evolution can leverage knowledge of gene function in one species to allow predictions in other species. Gene family history may also indicate specific mutational events responsible for phenotypic changes during evolution. At the heart of gene function prediction is distinguishing orthologous and paralogous gene pairs. A pair of orthologous genes are derived from a speciation event (Fitch, 1970; Ingram, 1961) whereas a pair of paralogous genes are derived from a mutational event that copied a gene within an ancestral genome (a gene duplication event). Orthologous genes often retain similar functions, whereas paralogous genes are more likely to diverge in function (Koonin, 2005). Gains or losses in gene copy number in genomes may be associated with the evolution of phenotypic traits (Ohno, 1970; Plachetzk and Oakley, 2007). Primarily for these reasons, gene family evolution has received considerable attention in bioinformatics, where numerous methods have been developed.

Most methods to understand gene history can be classified as similarity-based or tree based methods. “Similarity-based” methods primarily rely on the relative similarity between pairs of genes, without explicitly utilizing information from a phylogeny of the species containing the genes of interest. For example, a common approach to understanding gene family history uses the symmetrical best scores from all-to-all BLAST alignments, which have been effectively used for detection of orthologs from two given genomes (Li, et al., 2003; O’Brien, et al., 2005; Wall, et al., 2003). When only two species are involved, their species-level relationships are known. Once genes from more than two species are analyzed, reliable orthology detection becomes more challenging because GDLs are common during evolution, and gene similarity is correlated with, but does not always reflect common ancestry. Different similarity-based algorithms, including COG/KOG (Tatusov, et al., 2003), InParanoid/ MultiParanoid (Alexeyenko, et al., 2006; Remm, et al., 2001), OrthoMCL (Li, et al., 2003) and Roundup (Dela, et al., 2006) deal with these challenges in different ways. Other ortholog identification methods use “tree-based” approaches, including RIO (Zmasek and Eddy, 2002), Orthostrapper (Storm and Sonnhammer, 2002), TreeFam (Li, et al., 2006), and OrthologID (Chiu, et al., 2006). These methods explicitly consider both gene and species history, in some cases simultaneously, to estimate reconciled trees (Durand, et al., 2006). A challenge for tree-based methods is that they rely on correct estimation of both gene tree and species tree and require extensive amounts of computer time, often making genome scale analyses impractical or impossible.

Two other recent approaches, implemented in CAFE and SYNERGY use elements of both similarity-based and tree-based
approaches (De Bie, et al., 2006; Wapinski, et al., 2007). CAFE primarily uses information about the numbers of genes in pre-computed orthologous families, and does not simultaneously consider gene family phylogeny or similarity among genes. SYNERGY employs a novel algorithm to first find families of sufficiently similar genes and then to count GDLs by considering their gene family phylogeny trees.

In this contribution, we present EvolMAP software which implements a new algorithm to estimate the gene content of hypothetical ancestral species and to find the timings of gene duplication and loss events, relative to evolutionary transitions marked by speciation events. Unlike De Bie, et al. (2006) but like the tree-based approaches mentioned above, our algorithm simultaneously considers information on species history and gene history. However, unlike many of the tree-based methods above, our approach does not perform the computationally costly calculations required to explicitly estimate both species phylogeny and gene phylogeny.

Below, we explain our approach in detail, report availability of software for making the calculations, and analyze multiple data sets to allow comparison with previous methods. First, we analyze the history of G-protein coupled receptors (GPCRs) in animals. GPCRs constitute a diverse superfamily of proteins involved in signal transduction pathways and are found in many copies in animal genomes. An important question is whether those genes existed in high copy numbers in the early animal ancestors or were expanded subsequently in separate lineages. Through our analyses, we demonstrate the unique expansion tree for each GPCR family and find that eumetazoan and bilaterian ancestors had much fewer GPCRs than modern animals. Second, we reanalyze the history of mammalian GDLs which was studied previously under a gene family framework (Demuth, et al., 2006). In addition to reaching a higher resolution of gene history by calculating and comparing genome composition of multiple ancestral genomes, we include two more species in the analysis and use more recently annotated gene sets. With that, we obtain results that concur qualitatively with Demuth et al. (2006) but differ quantitatively. Like them, we also report that mammalian gene family sizes have changed drastically during evolution, including numerous changes in gene number since the split between humans and chimps. Finally, we compare our results from five genome-wide analyses against those of SYNERGY, InParanoid and MultiParanoid.

2 METHODS

For a given set of species-specific sequences and a species tree, we describe an all-to-all similarity clustering algorithm to reconstruct the ancestral genes of each ancestral node of the tree and pinpoint the timings of GDLs. In the first section, we describe a new similarity metric which is based on normalized, global optimal pair-wise alignments. In the second section, we explain how the orthologous groups are detected for ancestral nodes, which is essentially based on post-order traversal of the tree (from descendants to the root) while performing all-to-all similarity comparisons of each node’s nearest descendants (NDs). In the third section, we explain how the tree is traversed back from root to descendants to find Dollo parsimonious timings of gene gains and losses. In the final sections, we present two new species-topology based trees that are generated as result, and summarize the implementation of the algorithm.

2.1 Normalized Similarity Scoring

[Deleted 2 sentences] Many ortholog detection algorithms use very fast scoring schemes. For example, all-against-all BLASTP (Blat) searches offer high-speed detection of homologous sequence segments (Altschul, et al., 1997). However, BLAST often generates multiple, overlapping hits between a pair of genes. Any one of these overlapping hits may often represent inaccurate estimates of overall similarity and joining them together is complicated. In contrast to very rapid similarity scoring, dynamic programming algorithms - like Needleman-Wunsch (NW) optimal global alignments - offers a highly accurate scoring metric (Needleman and Wunsch, 1970), but is slow in comparison to BLAST. As a compromise between speed and accuracy, we first find the most similar gene pairs using BLAST, then generate more accurate similarity scores of those most similar genes by pairwise NW alignments assuming a BLOSUM62 scoring matrix (Henikoff and Henikoff, 1992). Gap extensions are minimally penalized to treat alignments of genes with different splice variants or with lost or recombined domains similarly to longer but evenly homologous alignments. Following standard NW alignments, similarity scores are normalized using:

\[
NSS(i, j) = \frac{NW_S(i, j)}{GAL(i, j)} \times \frac{\text{Min}(NW_S(i, j), NW_S(j, j))}{\text{Min}(GAL(i, j), GAL(j, j))}
\]

where \(NSS(i, j)\) is the normalized similarity score between sequences \(i\) and \(j\); \(NW_S\) are NW alignment scores; and \(GAL\) are gap-free alignment lengths where \(GAL(i, j)\) and \(GAL(j, i)\) are equal to length of the sequences \(i\) and \(j\) respectively. If \(NSS(i, j)\) returns a negative score, it is set to 0. With that, it defines globally comparable similarity scores that range from 0 to 1 (1 being most similar). In the implementation, we multiply the final score by 1000 to have a more memory efficient, integer representation of the score.

2.2 Detecting ancestral genes using symmetrical best alignments

Orthologous genes are commonly identified between a pair of genomes as symmetrical best alignments that score above a minimum threshold (sym-bet) (Remm, et al., 2001). Sequences within the same genome that align better to a sym-bet gene than its ortholog are defined as ‘in-paralogs’ (Sonnhammer and Koonin, 2002). In this study, we employ these concepts and use the terms ‘ancestral gene’ and ‘orthologous group’ interchangeably to refer to a sym-bet and its in-paralogs. Therefore, once identified, an orthologous group is assumed to include the detectable descendants of a single, hypothetical ancestral gene of a given ancestral node (Fig. 1 A).

In the first phase of the algorithm, the species tree is traversed in post-order (from descendants to root). For each ancestral node, sequences of its two NDs are compared in an all-to-all fashion using the NSS matrix to find sym-bets and their in-paralogs. In cases where the NDs are also ancestral nodes, similarity scores between their ancestral genes are calculated as the average similarity of members of one ancestral gene to those of the other. Genes that are neither a sym-bet nor an in-paralog of another sym-bet is defined as a ‘singular’ gene of the ancestral node. Those singular genes are considered as ambiguous ancestral genes at this stage where there is no evidence from the pairwise comparisons that the gene was present in that ancestral node. Through the second phase of the algorithm, earlier ancestral nodes are rigorously scanned for any such evidence, and the ambiguity is further resolved.
Reconstruction of ancestral genome content

2.3 Detecting gene gains and losses using Dollo parsimony

In the second phase of the algorithm, the species tree is traversed in pre-order (root to descendants) to better estimate the composition of the ancestral genomes and detect the timings of gene gains and losses. If a gene is ambiguously present (singular) in a given ancestral node and present (member of an orthologous group) in one of its NDs, then it was either present in this ancestral node and lost in the branch towards nearest descendant (BTND) missing the gene (Fig. 1 B), or it was not present as an individual loci in this node and was duplicated in the other BTND and diverged quickly and lost its in-paralog signal (Fig. 1 C). Comparing the corresponding genes of earlier ancestors may help clarify this ambiguity. Since the loss of an existing gene is more likely than a parallel gain of the same gene in another genome (ignoring horizontal gene transfers), a Dollo parsimony model (Farris, 1977) is fast and appropriate for inferring if and when a gene was gained or lost. For each singular, orthologous groups of all earlier ancestral nodes are searched to see if it was already inferred ‘present’ in any of those earlier ancestors. If such a ‘present’ orthologous group is found, and it does not include other paralogs from the ND missing the gene (i.e. it is not a larger group containing multiple paralogs), then the gene is inferred to be present in this ancestor and lost in the BTND missing the gene by the Dollo parsimony criterion (Fig. 2 A). If none of the ancestors have such an orthologous group, then according to the Dollo criterion the gene is inferred ‘not present’ in this ancestor and is gained in the BTND containing the gene (Fig. 2 B).

Gained genes are further categorized into one of the three groups of in-paralogs, diverged in-paralogs or ambiguous gains. In-paralogs are detected as genes that are inferred present in the ancestor as descendants of a single gene, but split into at least two genes in the descendant. They are essentially paralogous genes that were conserved better than the orthologs after the duplication event. Diverged in-paralogs are defined as singular genes of an ancestor, which are present in the descendant, and they have significant sequence similarity above a user defined statistical threshold to another present gene of the descendant. They are essentially genes that have likely duplicated in the given branch and diverged under relaxed selection. Ambiguous gains are defined as singular genes of the ancestor which are inferred present in the descendant and do not have significant sequence similarity above the user defined threshold to other genes of the descendant. Such genes might have experienced very high rates of mutation that erased their history, horizontally transferred from another genome, lost in all earlier lineages, or could just be dataset artifacts.

2.4 Average ortholog divergence and gene expansion trees

As a summary of results, two species-topology based trees are generated. A first tree, called the average ortholog divergence (AOD) tree, displays the average distance of orthologs of ancestral genes as the length of each branch of the species tree. A second tree, called the ‘gene expansion’ tree,
displays branch lengths as absolute number of genes gained along each branch. In this case, absolute gene gain is defined as total number of gene gains minus losses. \( AOD \) for each ancestral node \( A \) of the tree is derived by the following formula:

\[
AOD(A) = 1 - \left( \frac{\sum_{\text{for each sym-bet }(i,j) \text{ of } A} NSS(i,j)}{\text{number of sym-bets } (A)} \right)
\]

(2)

For species nodes, \( AOD \) is defined as 0. From this equation, lengths for each branch (A→D) of the \( AOD \) tree are calculated using the equation:

\[
AOD(A \rightarrow D) = \frac{AOD(A) - AOD(D)}{2}
\]

(3)

Number of present genes (\( NPG \)) of ancestral nodes is used to find the branch lengths of the gene expansion tree where branch lengths are derived by the following formula (with negative branch lengths set to 0):

\[
GE(A \rightarrow D) = NPG(D) - NPG(A)
\]

(4)

2.5 Implementation and Data Collection

EvolMAP is implemented using Java 1.6.0 in a package called EvolMAP. It will work under most operating systems running a Java Virtual Machine. Source codes and an executable (JAR) file are available. The JAR file may be used from command-line with a simple options file and a graphical user interface (GUI) is also provided. Basic input to EvolMAP is a binary newick format species tree and fasta files containing sequences from the corresponding species. Once an analysis is complete, inferred ancestral newick format species tree and fasta files containing sequences from the investigated protostome genomes (fly and worm) did not experience massive expansions of this family independently, while the investigated protostome genomes (fly and worm) did not (Fig. 3 A). The analyses of \( 7tm_2 \) and \( 7tm_3 \) containing-genes displayed slower and otherwise different profiles of expansion (Fig. 3 B, C). The \( 7tm_4 \) domain, which has very high copy number in the \( C.elegans \) genome, is not highly represented in other genomes, including the sea anemone. A phylogenetic tree with branch lengths proportional to gene numbers containing this domain shows a massive lineage-specific expansion on the branch leading to \( C.elegans \) (Fig. 3 D).

![Fig. 3. Expansion of GPCR-coding genes.](image)

**Fig. 3. Expansion of GPCR-coding genes.** Gene expansion trees were generated using EvolMAP with all-to-all NSS matrix for genes containing selected GPCR domains of (A) \( 7tm_1 \) (Opsin-rhodopsin family) (B) \( 7tm_2 \) (Secretin family) (C) \( 7tm_3 \) (metabotropic glutamate receptor family) and (D) \( 7tm_4 \) (chemoreceptor family). For each analysis, all-to-all NSS score matrices were used to find ancestral genes with the minimum ortholog threshold set as 250. Scale bar corresponds to hundred gene gains.
3.2 Evolution of the mammalian protein-coding genes

To scrutinize the gene complexity of ancestral mammalian genomes and identify the timings of GDLs during their descent, we analyzed seven mammalian proteomes using EvolMAP (Fig. 4). Dog and opossum genomes were used as outgroups to support the values for the inner clades. Gene expansion, gene gain and gene loss trees are displayed in Supplementary Fig. 1. Number of modern descendants of the inferred ancestral genes is summarized in Supplementary Fig. 2. On average, gene gains were detected to be 3.3 times more frequent than losses over all branches suggesting a steady expansion of genome size. An exception was the lineage leading to chimp after their split with human where the number of losses was more than the gains. *In-paralogs, diverged in-paralogs* and *ambiguous gains* constituted 27%, 48% and 25% of total gene gains respectively. This suggests that most gained genes (75%) have recognizable sources of duplication and yet most of them (48%) diverged more than the orthologs, undetectable as proper *in-paralogs*. Ambiguously gained genes represent perhaps the most interesting gene gains as they have significantly diverged from their original sources and may be important in species differences. Although some may be annotation errors, many are likely not. For example, the 544 and 1370 genes estimated to be ambiguously gained in the primate and human-chimp ancestral lineages, respectively, are detectable symmetrical best hits within their respective clades, but are missing any significant homolog in the outgroup genomes. Those are unlikely to be due to annotation errors since it would require the same annotation error in multiple genomes. In any event, it is expected that the numbers presented in this study will improve as the genomes are better annotated and more species added to the analysis.

3.3 Comparisons to previous methods

EvolMAP was first compared against SYNERGY on a previously analyzed set of nine yeast genomes from Wapinski, *et al.* (2007). *AOD* tree was generated and *GDLs* were mapped onto its branches (Supplementary Fig. 3). The orthologous assignments were observed to be in general agreement (avg. 91%) (Table 1). Furthermore, the numbers of genes estimated for 7 out of 9 ancestors were on average only 3% different. But for 2 remaining ancestors, EvolMAP predicted on average 15% fewer genes.

Then, we compared EvolMAP against InParanoid and MultiParanoid. For the InParanoid comparison, we observed almost complete overlap for the closely related species comparison of human and chimp (avg. 99%). The fly and worm comparison yielded somewhat less overlap (avg. 89%) and the biggest differences were observed for the comparison of human and Ciona (avg. 81%). Comparison to Multiparanoid analysis of four species from Alexeyenko, *et al.* (2006) display similar differences with the comparison to InParanoid on human and Ciona dataset (avg. 81%). EvolMAP results tried against different cut-off settings of MultiParanoid (0, 25, 50, 75 and 100) gave the most ortholog overlap with the cutoff set at 25.

![Fig. 4. Gene duplication and loss history of the mammalian proteome.](image)

EvolMAP was run with BLAST first to retrieve 300 top-scoring gene pairs for which NSS scores were generated where minimum ortholog threshold was set at 250. Scale bar corresponds to AOD (average ortholog divergence). Numbers within boxes show estimated number of present genes for that ancestral genome. Gene gains (+) and losses (−) are printed for each branch. Number of *in-paralogs, diverged in-paralogs* and *ambiguous gains* are printed respectively below each gene gain with a ‘+’ separator. Number of genes from the species is shown within parenthesis next to the species name where the second number following the ‘+’ symbol shows the number of genes omitted from the study. Those genes were omitted since they did not score a BLAST hit (above minimum threshold) to other genes at all.
Table 1. Orthologous gene family comparison with other methods.

<table>
<thead>
<tr>
<th>Matching groups</th>
<th>Matching genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYNERGY (9 yeasts)</td>
<td>9224/9224 (100%)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>9008/9062 (100%)</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>45287/48654 (88%)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>45401/48768 (93%)</td>
</tr>
<tr>
<td>InParanoid (HS – PT)</td>
<td>19306/19603 (98%)</td>
</tr>
<tr>
<td>Ciona intestinalis</td>
<td>19343/19594 (99%)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>39101/39361 (99%)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>39066/39396 (99%)</td>
</tr>
<tr>
<td>InParanoid (DM – CE)</td>
<td>3984/4333 (92%)</td>
</tr>
<tr>
<td>Ciona intestinalis</td>
<td>3985/4265 (93%)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>9305/9823 (93%)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>9254/11139 (93%)</td>
</tr>
<tr>
<td>InParanoid (HS – CI)</td>
<td>4373/3590 (78%)</td>
</tr>
<tr>
<td>Ciona intestinalis</td>
<td>4339/4512 (96%)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>10297/12395 (83%)</td>
</tr>
<tr>
<td>Protozoa</td>
<td>10345/13140 (83%)</td>
</tr>
<tr>
<td>MultiParanoid (HS/CI vs DM/CE)</td>
<td>5431/6281 (87%)</td>
</tr>
<tr>
<td>Ciona intestinalis</td>
<td>5463/5650 (97%)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>23194/26625 (87%)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>23302/31253 (75%)</td>
</tr>
</tbody>
</table>

First column for each comparison displays the comparison of the program from left versus EvolMAP and second column displays EvolMAP versus the other program. For each gene of a matching group, the one with most overlapping members with the original group was picked for comparison. For each gene of a matching group, we counted whether it is also contained in the corresponding group (Matching genes). Abbreviations used are HS, Homo sapiens; PT, Pan troglodytes; CI, Ciona intestinalis; DM, Drosophila melanogaster; CE, Caenorhabditis elegans. For all analyses, EvolMAP was run with BLAST first to retrieve 250 top-scoring genes for which NSS scores were generated, and orthologous groups were identified with the minimum ortholog threshold set as 300.

4 DISCUSSION

4.1 Advantages and limitations of EvolMAP and comparisons to other methods

The first step of EvolMAP’s algorithm is based on consecutive all-to-all NSS comparisons of two genomes to find evolutionary relationships of orthology and paralogy. The comparison of EvolMAP and InParanoid, a previous method that generates all-to-all comparisons of a pair of genomes based on BLAST scores, shows that even though closely related species comparisons, like that of human and chimp, are in very high agreement (99%), the agreement drops for evolutionarily more distant comparisons, such as that of human and Ciona (81%). The differences are likely due to different scoring schemes used by the two programs. As explained in Section 2.1, we reason that NSS is more reliable than BLAST scores for predicting orthology between evolutionarily more diverged sequences.

Species-tree based clustering of the pairwise all-to-all comparisons enables re-construction of ancestral genes of multiple hypothetical ancestors. In contrast to MultiParanoid which clusters results from separate InParanoid analyses to find orthologous groups of a single ancestral genome, we estimate contents of multiple ancestral genomes. When multiple ancestral genomes are available, comparing ancestral gene families with the descendant gene families can help detect the timings of GDLs. We applied the Dollo parsimony model to detect those events rapidly.

SYNERGY is a gene phylogeny based algorithm developed for the same purpose of ancestral gene content estimation. In that method, both gene and synteny similarity are utilized. Synteny similarity, which corresponds to the similarity of order of genes on a chromosome, might be advantageous for finding orthologs of closely related species. Even though EvolMAP does not consider synteny, orthologous assignments of the two programs for the nine yeast genomes were highly overlapping (91% same). A further challenge for synteny-based ortholog detection is the near absence of synteny for evolutionarily more distant species comparisons, such as that of a protostome and a deuterostome.

When we compared ancestral genes detected by SYNERGY and EvolMAP for a reconstructed yeast species tree, we observed only 3% average difference for 7 of 9 ancestral nodes (Supplementary Fig. 3). However, for the remaining two nodes EvolMAP predicted 15% fewer genes on average. Those nodes were the ancestral nodes immediately following the branch where the proposed yeast whole genome duplication (WGD) occurred. In fact, in the SYNERGY analysis, the branch of the WGD was assigned a priori a much higher duplication rate than other branches (0.5 for WGD-branch vs. 0.05 for all other branches). That assignment led to differences between the two results, since EvolMAP finds the branches where the GDLs occurred based on the most Dollo parsimonious scenario. Fewer stabilized genes predicted by EvolMAP for those branches may be due to massive concurrent gene losses at after-WGD lineages, and retention of different paralogs in lineages that branched shortly after the WGD. We suggest that using the option of SYNERGY to set different rates of duplications and losses for different branches may generate ‘circular’ results supporting the original hypothesis, whether or not it is fully supported by gene/synteny similarity data.

4.2 Analyses of specific gene families and complete proteomes using EvolMAP

Restricting EvolMAP analyses on selected gene families is informative for discerning clade-specific expansions of certain genes. In this study, we focused on the evolutionary history of genes containing GPCR domain families of 7tm_1 to 4. These domains are not found in plant or yeast genomes by HMMER searches. Further, no 7tm_1 or 7tm_4 domains and only one 7tm_2 and 7tm_3 domains are found in the genome of the choanoflagellate Monosiga brevicollis. This suggests that two of these domains originated in metazoan ancestors and all of them expanded only in metazoans. For such animal-specific expansions, the recently sequenced sea anemone genome lies in a highly informative phylogenetic position (Putnam, et al., 2007). Of the four analyzed GPCR domain families, the Opsin/Rhodopsin (7tm_1) domain containing-family is the largest animal-specific GPCR in size (Wistrand, et al., 2006). HMMER identified 916 genes containing that domain in the sea anemone. This number surprisingly exceeds that of worm (133) and fly (64), and is comparable to that of human (713) and zebrafish (755), all of which are dwarfed by that of mouse (1426). However, EvolMAP estimated only 24 genes containing 7tm_1 domains that were present in the eumetazoan ancestor suggesting that both vertebrate and the sea anemone genomes experienced massive parallel expansions of this family, whereas the two investigated protostome lineages did not undergo such a large expansion.

Using EvolMAP, we studied the history of GDLs of seven mammalian genomes. Our results concur with previous studies that used CAFE (Demuth, et al., 2006) to demonstrate numerous GDLs during evolution of five mammalian genomes (our study added newly available monkey and opossum genomes). Yet our results differ in detail from Demuth et al. (2006) and indicate that GDLs may be even more common than previously suspected. The biggest difference between results from EvolMAP and CAFE was
the number of gene gains estimated along the chimpspecific lineage (26 with CAFE vs. 493 with EvolMAP). We do not attribute this difference to changes in annotation with our use of a new dataset (Ensembl v46) because analyzing the same dataset used by Demuth et al. (2006) (5 genomes, Ensembl v41) with EvolMAP also yields much higher numbers of gene gains along the chimp lineage (694).

Instead, we suggest that the differences between EvolMAP and CAFE may be caused by highly conservative estimates made by CAFE in cases where genes radiated rapidly and were differentially lost. More specifically, after a rapid gene radiation, different subsets of paralogs may be retained in descendant species (Supplementary Fig. 4). In such cases, CAFE infers those paralogs to share a direct ancestry because the method only counts the number of genes in closely related gene families without referencing the relative similarity between them. To test for such occurrences, we performed homology tests between genes inferred by EvolMAP to be gained along the human and chimp lineages. We found 16% of human-specific and 61% of chimpspecific sequences to be significantly similar to a gene inferred as gained along the other species’ lineage (i.e. they would be in the same orthologous group in CAFE analysis). Therefore, we were able to attribute some of the differences between the two results to rapid gene duplications. The results of Demuth et al. (2006) were conservative with respect to their provocative claim of a genomic “revolving door”, where genes are gained and lost routinely. Our results indicate that GDLs occur at an even faster rate than previously estimated.

But perhaps the most interesting genes categorized by EvolMAP are ambiguous gains, which are highly novel sequences that are clearly lineage-specific. Here there is no concern for being an artifact of rapid gene radiations. Although these genes are unlikely to be due to annotation errors, those that are not may contribute unique functions to the biology of their species. Such genes may originate by mechanisms such as exon-domain shuffling, bursts of positive selection or horizontal gene transfers.

FUNDING
Support for this project derived from the W. M. Keck Foundation and a gift from Harvey Karp.

ACKNOWLEDGEMENTS
We like to thank Tsunglin Liu, Firat Kart and Kawther Abu-Elneel for comments and discussions. Thanks also to I. Wapinski for help with SYNERGY data and analyses.

REFERENCES