# Comparative Genomics Search for Losses of Long-Established Genes on the Human Lineage

Jingchun Zhu[1], J. Zachary Sanborn[1], Mark Diekhans[1], Craig B. Lowe[1], Tom H. Pringle[1], David Haussler[1,2*]

1 Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California, United States of America, 2 Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, California, United States of America

**Taking advantage of the complete genome sequences of several mammals, we developed a novel method to detect losses of well-established genes in the human genome through syntenic mapping of gene structures between the human, mouse, and dog genomes. Unlike most previous genomic methods for pseudogene identification, this analysis is able to differentiate losses of well-established genes from pseudogenes formed shortly after segmental duplication or generated via retrotransposition. Therefore, it enables us to find genes that were inactivated long after their birth, which were likely to have evolved nonredundant biological functions before being inactivated. The method was used to look for gene losses along the human lineage during the approximately 75 million years (My) since the common ancestor of primates and rodents (the euarchontoglire crown group). We identified 26 losses of well-established genes in the human genome that were all lost at least 50 My after their birth. Many of them were previously characterized pseudogenes in the human genome, such as *GULO* and *UOX*. Our methodology is highly effective at identifying losses of single-copy genes of ancient origin, allowing us to find a few well-known pseudogenes in the human genome missed by previous high-throughput genome-wide studies. In addition to confirming previously known gene losses, we identified 16 previously uncharacterized human pseudogenes that are definitive losses of long-established genes. Among them is *ACYL3*, an ancient enzyme present in archaea, bacteria, and eukaryotes, but lost approximately 6 to 8 Mya in the ancestor of humans and chimps. Although losses of well-established genes do not equate to adaptive gene losses, they are a useful proxy to use when searching for such genetic changes. This is especially true for adaptive losses that occurred more than 250,000 years ago, since any genetic evidence of the selective sweep indicative of such an event has been erased.**

Citation: Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, et al. (2007) Comparative genomics search for losses of long-established genes on the human lineage. PLoS Comput Biol 3(12): e247. doi:10.1371/journal.pcbi.0030247

## Introduction

It is intuitive to think that changes leading to increased complexity, adaptation, and intelligence are achieved by the gain and improvement of genetic components such as genes and regulatory elements. However, in certain scenarios, a loss of function can also bring a selective advantage. The best-known examples are losses of cell surface receptors to confer pathogenic resistance, such as the inactivation of the *DUFFY* gene contributing to malaria resistance [1] and homozygosity for a null allele of chemokine receptor *CCR5* conveying resistance to infection by various pathogens, including HIV [2]. In addition, the loss of an existing biological component can open new developmental opportunities. The human-specific loss of a myosin heavy chain isoform expressed in the masticatory muscles has been linked to the weakening of human jaw muscles, possibly allowing the increase of cranial capacity in humans, although this is still quite speculative [3]. Adaptive gene loss is the type of genetic change that leads to better fitness for an organism by inactivating a functional gene. As argued by the "less-is-more" hypothesis, gene losses may be an important engine of evolutionary innovation [4].

In addition to adaptive evolution, gene losses can play an important role in human diseases where conditionally advantageous mutations improve fitness in a particular environment. For example, deleterious mutations affecting hemoglobin and other red blood cell proteins are common in many human populations due to a heterozygote advantage in malaria epidemic environments. This improved fitness comes at a cost for those born with deleterious mutations on both alleles, since the homozygous state causes anemia including sickle cell disease [5–7]. Other human diseases such as glucose-6-phosphate dehydrogenase deficiency [7] and cystic fibrosis [8,9] have also been associated with the heterozygote advantage.

Despite the apparent importance of adaptive gene loss, we know surprisingly little about its contribution and significance at the genomic level and over a broad time scale. Most research on adaptive evolution in mammals focuses on new

**Editor:** Gary Stormo, Washington University, United States of America

**Received** July 31, 2007; **Accepted** October 30, 2007; **Published** December 14, 2007

A previous version of this article appeared as an Early Online Release on October 30, 2007 (doi:10.1371/journal.pcbi.0030247.eor).



**Abbreviations:** My, million years; Mya, million years ago; OR, olfactory receptor

* To whom correspondence should be addressed. E-mail: haussler@soe.ucsc.edu

## Author Summary

One of the most important questions in biology is to identify the genetic changes underlying evolution, especially those along the lineage leading to the modern human. Although counterintuitive, losing a gene might actually bring a selective advantage to the organism. This type of gene loss is called adaptive gene loss. Although a few cases have been characterized in the literature, this is the first study to address adaptive gene losses on a scale of the whole human genome and a time period of up to 75 million years. The difficulty of identifying adaptive gene losses is in part the large number of pseudogenes in the human genome. To circumvent this problem, we used two methods to enrich the process for the adaptive candidates. The first is a novel approach for pseudogene detection that is highly sensitive in identifying single-copy pseudogenes that bear no apparent sequence homology to any functional human genes. Second, we used the length of time a gene is functional before loss as a proxy for biological importance, which allows us to differentiate losses of long-established genes from mere losses due to functional redundancy after gene duplication.


genes or regulatory elements as well as on modifications to known genes, such as amino acid substitutions [10,11]. With the complete genomes of human and several other mammals including chimp, rhesus, mouse, rat, and dog [12–16], it is now feasible to systematically identify adaptive gene losses in the human lineage through the course of mammalian evolution.

A claim for adaptive genetic change typically requires evidence of DNA signatures indicating directional selection, and is accompanied by the identification of selective pressures acting on the organisms that are consistent with DNA, fossil, or historical evidence. Methods for detecting amino acid or DNA signatures left by natural selection are not generally applicable for identifying adaptive gene loss [17,18]. An inactivated gene is no longer maintained through the forces of natural selection, and secondary mutations begin to accumulate at the neutral rate. Therefore, methods based on sequence conservation or ratio of synonymous versus nonsynonymous mutations are not suitable to detect adaptive gene losses [11,19–21]. Recent adaptive losses can be detected by the distinct DNA signatures left by positive selection; however, those signatures only persist for a narrow evolutionary window of at most 250,000 years [22–24]. To detect adaptive gene losses further back into the evolutionary past, it is reasonable to assume that a nonredundant gene that was functional for a long time and then inactivated is a good candidate for adaptive gene loss. While not every loss of a well-established gene is adaptive, searching for those candidates can be used to enrich for adaptive gene losses.

Gene loss normally leaves behind a pseudogene. However, the vast majority of pseudogenes in a genome did not bring a selective advantage to the organism. Most pseudogenes arise through a gene copying operation of either retrotransposition (reverse-transcribing a processed mRNA back to DNA, which is reinserted in the genome at a different location) [25], or by segmental or tandem duplication of a genomic region [26]. These are called processed or unprocessed pseudogenes, respectively. While processed pseudogenes in general have a single exon and a polyadenine tail, unprocessed pseudogenes typically have multiple exons and preserve the intron–exon structures of the parental gene. The vast majority of processed pseudogenes are "dead on arrival," due to the lack of complete coding regions or necessary transcription and translation signals in the new genomic location. Even when a functional gene is formed by segmental duplication, one copy often becomes silenced by degenerative mutations due to functional redundancy [27]. In contrast, adaptive gene losses arise from degradation of genes with well-established functions, which often do not have close homologs in the genome. Taking advantage of the genomic signatures left behind by retrotransposition or gene duplication, several genomic surveys identified tens of thousands of pseudogenes in the human genome using sequence homology to a functional parental gene [28–32]. However, because they lack close homologs, many losses of well-established genes were missed by these studies. More importantly, these analyses focused on cataloging pseudogenes in the human genome, but not on addressing whether the pseudogenizations played a role in evolution.

This study identified losses of well-established protein-coding genes in the human lineage since the common ancestor of euarchontoglires (primates, lemurs, tree shrews, rodents, and lagomorphs such as rabbits). We applied a novel comparative genomic method to identify pseudogenes by syntenic mapping of gene structures between the human–mouse–dog trio of genomes. This approach is able to systematically detect the sequence signature left by losses of well-established genes, distinguishing true losses from mere loss of redundant genes following duplication or retrotransposition. Our analysis was able to differentiate the losses of well-established genes from the large background of human pseudogenes. Twenty six losses of well-established genes were identified in the human lineage since the common ancestor of euarchontoglires, approximately 75 million years ago (Mya). Sixteen of those were previously uncharacterized gene losses in the human genome, such as the loss of acyltransferase 3 during great ape evolution.

## Results

### Detecting Gene Losses Using Gene Structure Conservation

After a mutation inactivates a functional gene, the signature of the intron–exon structure can still be detected for some time before neutral decay erases it from the genome. Based on the observation that mammalian gene structures are typically conserved between species, a gene prediction program called TransMap was developed that exploits the large-scale conservation of gene order and orientation on mammalian chromosomes to map gene structures between genomes. TransMap is essentially a cross-species mRNA alignment grogram (Text S1) that relies upon the "syntenic" alignments produced by the BLASTZ program [33]. TransMap is highly sensitive in detecting gene structures for both genes and pseudogenes (Table S2). Unlike most existing pseudogene detection methods [30,31,34–36], TransMap does not rely on sequence homology to a parental gene from the same genome; therefore, it is well-suited for detecting losses of well-established genes whose functional precursor has not been recently duplicated. To identify gene losses, the mapped coding region is conceptually translated and scanned for ORF-disrupting mutations. A TransMap

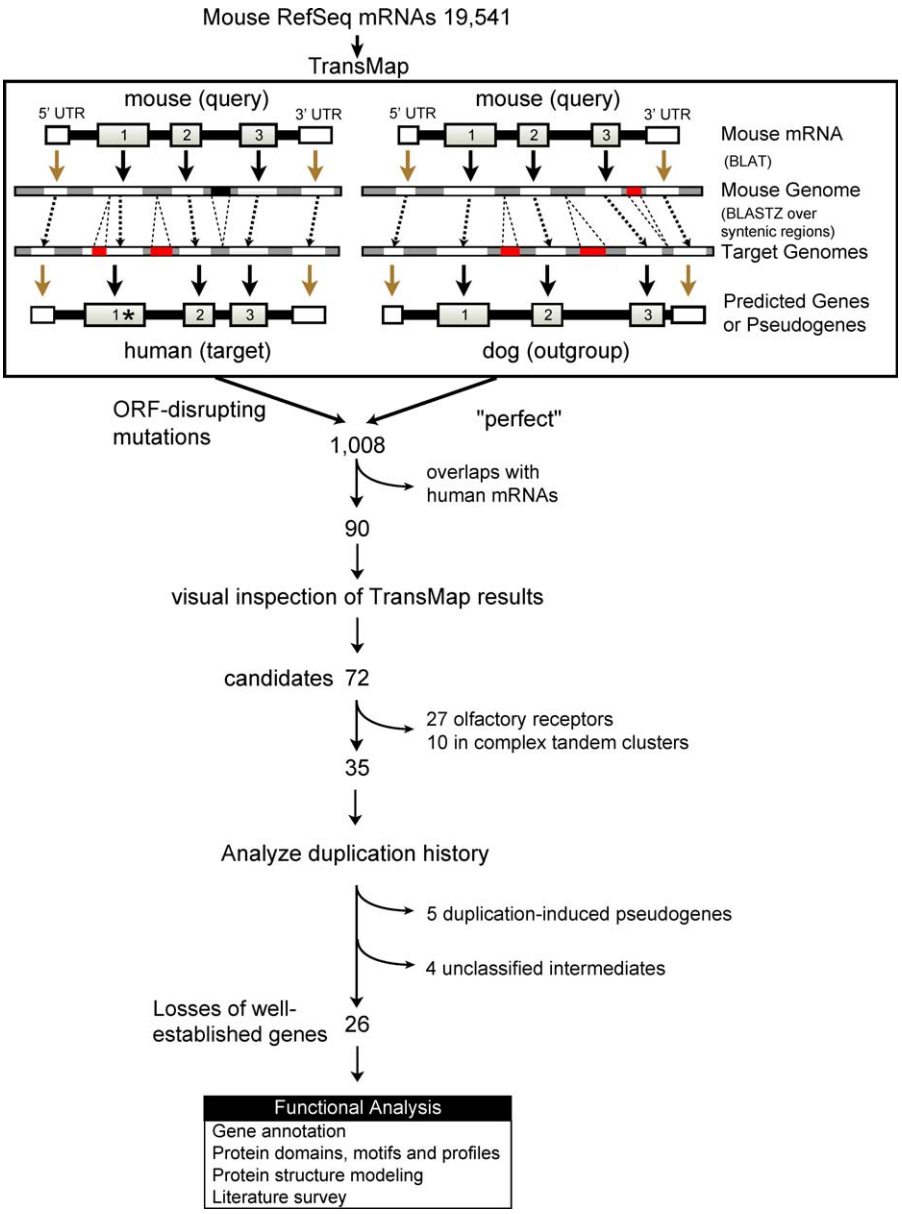


**Figure 1.** Algorithm for Identifying Losses of Well-Established Genes in the Human Lineage Since the Common Ancestor of Euarchontoglires

TransMap predicts (pseudo)genes in the human and dog genomes by syntenically mapping mouse mRNA gene structures to the target genomes through genome alignments and then transferring over the corresponding genomic coordinates. The predicted coding regions are conceptually translated and scanned for ORF-disrupting mutations. In this example, a stop codon (labeled with an "*") is detected in the first coding exon mapped to the human genome, which has also experienced an insertion. Genomic insertions or deletions are shown as red rectangles. Of 19,541 mouse RefSeq genes, 1,008 are identified as initial candidate gene losses in the human lineage based on the differential mutation status in the TransMap results. The list is narrowed down to 72 after eliminating those overlapping with human transcription evidence and filtered out by a manual inspection. Twenty-six are identified as losses of well-established genes in the human lineage after analyzing their duplication histories.

doi:10.1371/journal.pcbi.0030247.g001

prediction is a candidate for gene loss if any of the following ORF-disrupting mutations are detected in the mapped coding regions: stop codons, frameshifts, a splice junction that is not GT-AG, or when less than 50% of the coding region can be mapped to the target (Figure S1). To decrease the number of false positives, a valid conceptual translation is required in an outgroup genome. For example, if a mouse mRNA TransMaps from mouse to the dog genome with a valid conceptual translation, but fails to do so from mouse to the human genome, the gene is a candidate for a human gene loss (Figure 1).

Using mouse mRNAs as queries and the dog genome as an outgroup, we identified gene losses in the human lineage since the common ancestor of euarchontoglires. We chose the RefSeq database because it is one of the most comprehensive collections of mouse transcripts [37]. Based on the differential prediction status in the dog and human genomes, 1,008 of the 19,541 mouse RefSeq transcripts were identified as potential human losses (Table S3). The candidates were reduced to 90 after filtering out overlaps with GenBank human mRNAs in order to identify pseudogenes where no transcriptional activity in the form of mature mRNA has been

observed in humans, and to remove the false positive pseudogenes predicted by TransMap (Text S1). A visual examination of an alignment of the mouse mRNA sequence, mouse, human, and dog genomic sequences and their three-frame translations further reduced the number of candidates to 72, eliminating 18 that we are not confident represent true pseudogenes in the human genome. The pipeline is illustrated in Figure 1.

## Gene Losses in the Human Lineage

Of the 72 candidates, 27 are predicted to be olfactory receptors (ORs) and ten are members of large gene families, such as keratins. These gene families are organized as tandem gene clusters that have experienced copy number changes and/or complex local rearrangements since the common ancestor of euarchontoglires. The dynamics of gene clusters make it difficult to unambiguously discern ortholog/paralog relationships among species and analyze the evolutionary history of the lost genes. Therefore, we focused on the remaining 35 (non-OR, non-cluster) candidates that were confirmed by visual inspection, which we referred to as the "definite" losses in the human lineage. Among them, 21 gene losses are annotated with some biological functions in mouse and 14 have not been characterized functionally (Table 1).

The large number of ORs found by this method is consistent with observations that human OR genes experienced a rapid acceleration of pseudogene formation [38,39]. Previous studies have shown evidence that human genes involved in olfaction have a significant tendency to be under positive selection, indicating ORs have undergone directional selection in humans, including by pseudogenization [18,39]. Many previously identified non-olfactory gene losses were confirmed as well. These include gulonolactone (L-) oxidase (*GULO*), a vitamin C biosynthesis enzyme that is the genetic basis for scurvy [40], and urate oxidase (*UOX*), an enzyme converting uric acid to allantoin [41]. In addition, a human-specific nonsense mutation was confirmed in an orphan chemoattractant G protein–coupled receptor 33 (*GPR33*). Additionally, confirmed gene losses included the human-specific loss of cardiotrophin-2 (*CTF2*) due to a 8 bp deletion [42], cytochrome c oxidase subunit VIIIb (*COX8B*) with only 40% mapping to the human genome yielding an ORF of eight amino acids [43], and others listed in Table 1, note a.

Our analysis also identified 23 previously uncharacterized losses, of which 21 belong to the definite loss group and two are members of complex gene clusters. One example of a previously unknown definite loss is acyltransferase 3 (*ACYL3*; NM__177028), identified by the Riken mouse cDNA project [37,44] and whose function in mammals has not been characterized experimentally. We conducted protein profile analysis and determined that this gene has a highly conserved acyltransferase 3 domain (hence we annotated the gene *ACYL3*) [45]. Further structural modeling (SAM, TMHMM, SignalP) revealed mouse Acyl3 is a multipass transmembrane protein with its C-terminal domain forming a helix bundle. The N-terminal is extracellular and hydrophilic with conserved cysteine residues able to form disulfide bonds [46–49]. The extracellular domain might be involved in cellular response to external signals. Thus, Acyl3 might be a membrane protein with acyltransferase activity or a multipass transporter to pass molecules across the membrane upon external signals [50,51]. Phylogenetically, *ACYL3* is ancient and conserved in archaea, bacteria, fungi, worms, flies, and mammals. While numerous copies of *ACYL3* are encoded in fly and worm genomes, mammalian genomes have only one copy. A nonsense mutation (TGG to TGA) located in one of the transmembrane helices is shared by the human and chimpanzee genomes. However, the ancestral TGG codon is present in two orangutan trace sequences, and a valid conceptual translation is present in the rhesus genome. To narrow down the timing of the inactivation, we sequenced a PCR product amplified from the corresponding region in a gorilla DNA sample. The sequencing result showed the TGG (W) codon is present in the gorilla genome. Based on this evidence, it appears that the nonsense mutation inactivated *ACYL3* after the divergence of gorillas from the human lineage, and before the divergence of humans and chimpanzees (Figure 2). It is intriguing that the last copy of such an ancient enzyme as *ACYL3* was lost during the evolution of great apes. Although we do not know the precise evolutionary impact of the loss, its expression pattern in the mouse pituitary gland and developmental abnormalities observed in *Drosophila* null mutants suggests the loss might be related to development or hormonal regulation [52–54].

As shown in Table 1, other previously unknown gene losses include *CETN4*, a mammalian centrin expressed in ciliated cells including those present in the cerebellum [55], *NEPN* (nephrocan), an inhibitor of TGF-β signaling pathway [56], and *NRADD*, a death domain containing membrane protein involved in mediating apoptosis in response to ER stress [57]. It is worth noting that ER-mediated apoptosis triggers a cascade leading to the activation of Caspase 12 (*CASP12*), a gene that is also lost in humans. The loss of *CASP12* is still polymorphic in humans and has been shown to have experienced a recent selective sweep [58,59].

## Timing of the Gene Losses

The timing of the gene losses was determined by finding the branch interval that encloses the earliest shared ORF-disrupting mutations between humans and other mammals on a phylogenetic tree. The branch intervals on the human lineage for the 35 definite losses are illustrated on a mammalian phylogeny (Figure 3).

Using complete genome sequences of human, chimp, rhesus, mouse, and dog, six genes were determined to be human-specific losses, i.e., lost after the divergence of humans and chimpanzees. Ten genes were found to be lost during the period between the human–chimp split and the divergence of old world monkeys from the human lineage. Among the ten genes, seven were observed to have independent ORF-disrupting mutations in the rhesus lineage that are not shared with humans or chimps. Seventeen genes were determined to be lost prior to the divergence of old world monkeys from the human lineage and after the common ancestor of euarchontoglires. Due to insufficient sequence information in the rhesus genome, two genes could only be determined to be lost at some point during the 70 million years (My) prior to the human–chimp split and after the common ancestor of euarchontoglires. To refine the timing of the gene losses, we extracted trace sequences from several additional primates (orangutan, marmoset, tarsier, galago) and tree shrew. Using these trace sequences, we were able to narrow down 50% of gene losses to a much more precise branch on a phylogenetic tree. For example, the timing of

**Table 1.** Gene Losses in the Human Lineage since the Common Ancestor of Euarchontoglires

| Row | Mouse Gene Name | Annotated Function in Mouse | ORF-Disrupting Mutation | Note |
|---|---|---|---|---|
| 1 | Ctf2 | Cardiotrophin 2 | 8 bp indel | a |
| 2 | Cyp2g1 | Cytochrome P450, family 2, subfamily g | Nonsense | a |
| 3** | Gpr33 | G protein-coupled receptor 33 | Nonsense | a,b |
| 4 | 1700013G24Rik | Adult testis expressed cDNA, a helix-turn-helix domain | Nonsense,indel | |
| 5** | Sord | Sorbitol dehydrogenase 1 | Indel | a |
| 6** | S100a15 | S100 calcium binding protein A15 | 46% missing | a |
| 7 | Acyl3 | Acyltransferase 3 | Nonsense | |
| 8 | Crygf | Crystallin, gamma F | Indel (indel) | a, b |
| 9 | Taar4 | Trace amine-associated receptor 4 | Indel | a, b |
| 10 | Uox | Urate oxidase | Nonsense | a, b |
| 11 | Nradd | Death domain containing membrane protein | Nonsense (indel) | |
| 12 | Gsta4 | Glutathione S-transferase, alpha 4 isozyme | Indel (indel) | a |
| 13 | Sult1d1 | Sulfotransferase family 1D, member 1 | Nonsense (indel) | a, b |
| 14* | Pfpl | Similar to pore forming protein | Indel (indel) | b |
| 15 | 4933429E10Rik | Multi antimicrobial extrusion protein (LOC380701) | Indel (indel) | |
| 16* | LOC433492 | Vomeromodulin-like protein | 40% missing | |
| 17 | Tex21 | Contains leucine zipper motif, expressed in adult testis | 35% missing | |
| 18 | 0610012H03Rik | Thioesterase superfamily (LOC74088) | Indel | |
| 19 | Gucy2d | GC guanylate cyclase 2d receptor | Nonsense | |
| 20 | Nepn | Nephrocan | Nonsense,indel | b |
| 21* | BC018465 | Contains a lipid-binding serum glycoprotein domain | Indel | b |
| 22 | Gm766 | Contains a sulfotransferase domain | Indel | |
| 23 | Gulo | Gulonolactone (L-) oxidase | 54% missing,indel,nonsense | a |
| 24* | Abca14 | ATP-binding cassette, sub-family A | 87% missing | a, b |
| 25 | Cox8b | Cytochrome 8, subunit VIIIb | 60% missing (indel) | a |
| 26 | 4921517D21Rik | Actin-like domain containing protein | SINE insertion | |
| 27 | BC048502 | LOC223927 | SINE insertion,indel | |
| 28 | 2700097O09Rik | Contains methyltransferases structure (LOC72658) | 52% missing | |
| 29 | 2610318N02Rik | Contains proline-rich profile (LOC70458) | 62% missing | |
| 30 | D730039F16Rik | Contains CutA1 domain (LOC77996) | 68% missing | |
| 31 | B430306N03Rik | Immunoglobulin-like protein | 80% missing | |
| 32** | Cxcl7 | Pro-platelet basic protein | Indel | b |
| 33** | Unc93a | Unc-93 homolog A | 88% missing | b |
| 34 | Cetn4 | Centrin 4 | Indel | |
| 35 | Slc7a15 | Aromatic-preferring amino acid transporter | Nonsense | a |
| +27 olfactory receptors | | | | a, c |
| +2 UDP glucuronosyltransferase | | | | a, c |
| +4 keratin proteins | | | | a, c |
| +1 cytochrome P450 | | | | a, c |
| +1 retinol dehydrogenase | | | | c |
| +1 tryptase | | | | a, c |
| +1 hypothetical protein similar to putative membrane protein Re9 (2310042E22Rik) | | | | c |

Among the 72 candidates, after excluding 27 ORs, 35 (numbers 1–35) have been confirmed by visual inspection to be definite losses in the human lineage. The remaining ten genes belong to large gene families that cluster on the human genome in tandem and had experienced copy number change and/or complex genomic rearrangements since the common ancestors of euarchontoglires, making it difficult to confirm gene losses by visual inspection. Of the 35 definite losses, 26 are classified as losses of well-established genes (no * label), five are duplication-induced pseudogenes (labeled with **), and the remaining cannot be determined with the current data set (labeled with *). ORF-disrupting mutations shown in column four are the earliest shared mutations identified between human and other mammals. Mutations that give rise to an independent gene loss in rhesus are shown in parentheses.

a Previously characterized human gene losses.

b Genes with >60% overlap with pseudogenes from VEGA or Yale pseudogenes [64,65].

c Duplication history as well as ORF-disrupting mutations were undetermined for OR genes and the aforementioned ten genes that belong to tandem gene clusters.

doi:10.1371/journal.pcbi.0030247.t001

gene losses for *GUCY2D*, *NEPN*, and others (number 19 to 22) was narrowed down to the period of approximately 25 to 40 Mya, between the dates when old world and new world monkeys split off from the human lineage (Figure 3).

In addition to identifying previously unknown gene losses in the human genome, our analysis refined the timing of several previously known gene losses. For example, *SULT1D1*, a sulfotransferase, and *GSTA4*, a glutathione S-transferase alpha 4 isozyme are known to be pseudogenes in human while their mouse orthologs remain functional [60,61]; however, it was unclear when the inactivation occurred. Our analysis discovered that the pseudogenization of *GSTA4* and *SULT1D1* occurred approximately 14 to 25 Mya, between the dates when orangutans and old world monkeys split off from the human lineage. *GULO* is known to be inactive in primates with the inactivation dating to some time prior to the separation of apes and old world monkeys (>25 Mya) [62]. Frameshift indels, nonsense mutations, and genomic deletions are observed in the human *GULO* sequence, indicating an older pseudogene that has experienced numerous secondary mutations. The shared mutation analysis has shown that human, chimp, rhesus, and marmoset share a nonsense mutation, while galago, mouse, and dog share the TGG tryptophan codon. Therefore, the inactivation of *GULO*

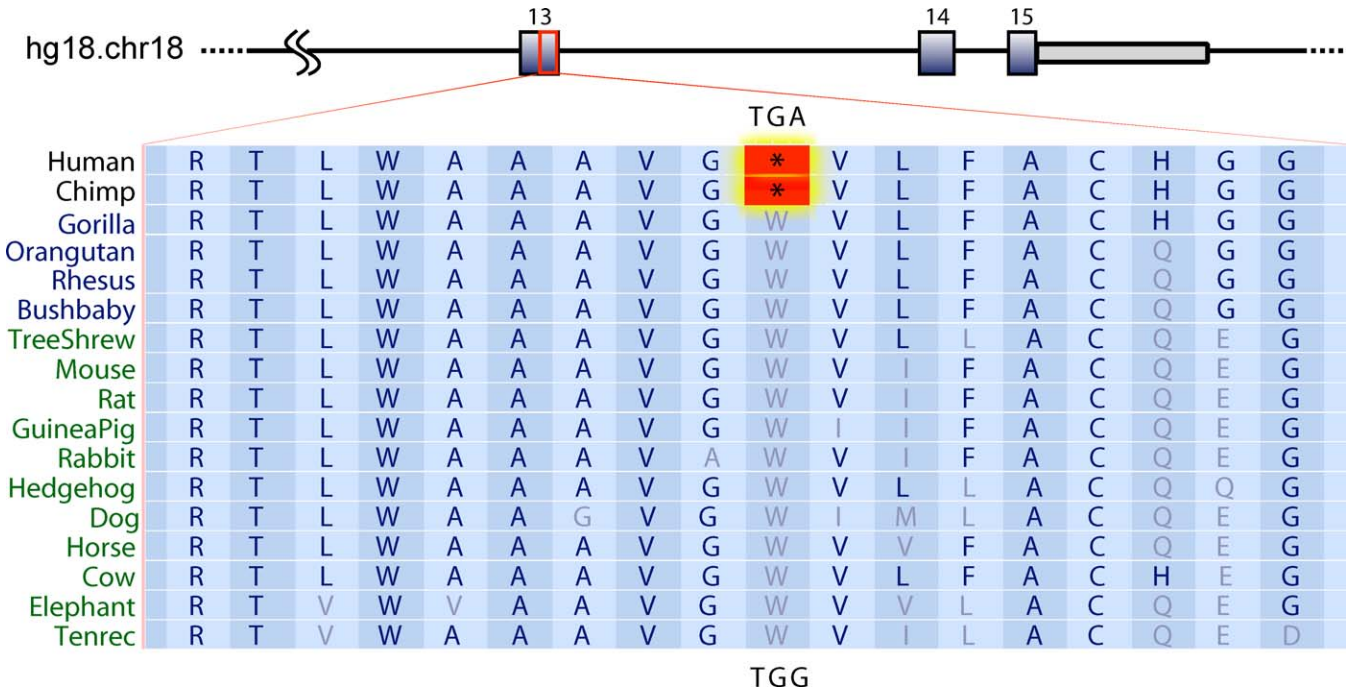


**Figure 2.** A Nonsense Mutation Inactivated *ACYL3* During Great Ape Evolution

It occurred after the divergence of gorillas from the human lineage and before the human–chimp split. The nonsense mutation is located in exon 13 of *ACYL3*. A multispecies syntenic alignment showing the nonsense mutation ("*") lies in a highly conserved protein coding region. The stop codon mutation (TGA) is present in the human and chimp genomes, but a TGG tryptophan (W) codon is present in the rhesus, mouse, rat, dog, and other mammalian genomes. The region maps to human Chromosome 18 at location 54,881,070–54,881,124. We sequenced the genomic region in a gorilla DNA sample to show that the codon TGG (W) codon is present in the gorilla genomes.

doi:10.1371/journal.pcbi.0030247.g002

occurred before the separation of new world and old world monkeys (>40 Mya).

## Losses of Well-Established Genes versus Duplication-Induced Pseudogenes

A significant contribution of this analysis is to differentiate losses of well-established genes from the large background of pseudogenes caused by retrotransposition or formed shortly after segmental or tandem duplication. The method of syntenically mapping gene structures to both a target and outgroup genome is likely to filter out almost all processed pseudogenes. However, TransMap does not fully eliminate those genes that were silenced soon after duplication, which we referred to as duplication-induced pseudogenes. To identify duplication-induced gene losses, we need to determine when the duplication occurred.

Recent segmental duplications can be detected by within-genome sequence homology. If there is a self-alignment chain in the UCSC human genome browser [63] enclosing the gene loss region, the region is determined to have been recently duplicated. We determined when the duplications had occurred by tracing along the human lineage through a seven-species syntenic alignment (human, chimp, rhesus, mouse, rat, dog, opossum) to determine the origin of each duplicate in the best self-alignment (the one with the highest alignment score recorded in the UCSC genome browser). If a gene and its duplicate trace back to a single region in an outgroup genome, the duplication was determined to have occurred on the branch immediately after the outgroup split off from the human lineage. If the gene and its duplicate consistently traced back to different regions through the series of outgroups all the way back to opossum, the duplication was determined to occur prior to the common ancestor of human and opossum. Figure 4A is a schematic illustration of this procedure. The branch of gene duplication is the branch of gene birth (by duplication).

In many cases, there are no detectable self-alignments, indicating an ancient duplication had formed the functional precursor to the pseudogene. We presume the functional precursor of the pseudogene existed prior to the earliest common ancestor of human and the species whose genomic sequence can be aligned to the human exons, therefore providing a lower bound timing of the gene birth event. To narrow down the timing of gene birth on the long branch between dog and opossum, we included scaffold assemblies of the elephant, tenrec, and armadillo genomes. To infer gene birth events that occurred further back in the evolutionary past, we included the chicken genome in the analysis (Figure 4B). Using the above method, the gene birth by duplication branch for the 35 non-OR, non-cluster definite gene losses was determined and is shown in Table 2 (Gene Birth Branch).

Subsequently, we estimated the length of time (in My) a gene remained functional before its pseudogenization using the separation of the gene birth and death branches. Since the timings of both events are estimated as branch intervals, an upper and lower bound estimation of the separation was obtained. Using a 50 My threshold, we classified the candidates based on their functional time lengths as losses of well-established genes, duplication-induced pseudogenes, or undetermined. If the lower estimation of functional time

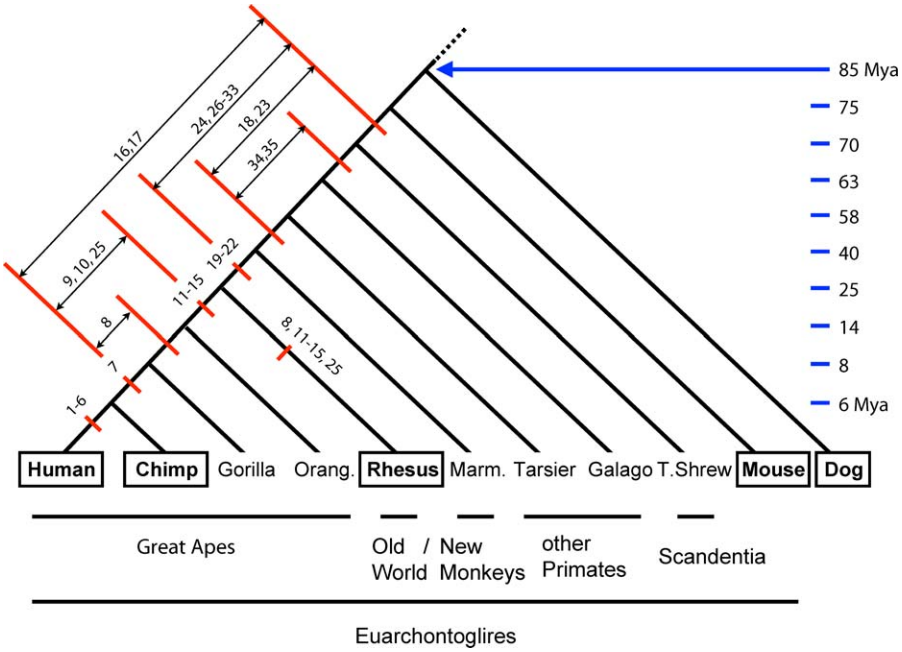


**Figure 3.** Timing of the Gene Losses in the Human Lineage Since the Common Ancestor of Euarchontoglires, Estimated Based on Shared Mutation Analysis

Branch intervals enclosing the earliest ORF-disrupting mutations shared between human and other mammals are illustrated on the human lineage of a mammalian species tree. Genes are represented by numbers, which correspond to their row numbers in Tables 1 and 2. Marks on the rhesus lineage represent independent ORF-disrupting mutations that are not shared with the ones in the human lineage. Approximate time when the species diverged from the human lineage is shown in Mya. Species with complete genome sequences are enclosed by rectangles, while others only have trace sequences available for analysis. Orang.: orangutan; Marm.: marmoset; T.shrew: tree shrew.

doi:10.1371/journal.pcbi.0030247.g003

length is greater than 50 My, the candidate is classified as a loss of a well-established gene. If the upper estimation of functional time length is smaller than 50 My, the gene loss is classified as a duplication-induced pseudogene. The gene loss is classified as undetermined if its functional time length overlaps the 50 My threshold.

Table 2 gives the estimated functional time length for the 35 definite losses. Among them, 26 are classified to be losses of well-established genes, which accounted for the majority (74%) of the definite losses. Five are classified as duplication-induced losses—*CYP2G1*, *SORD*, *S100a15*, *CXCL7*, and *UNC93a*— (labeled "**" in Tables 1 and 2). The remaining four are undetermined. Of these 26 losses of well-established genes, 16 have not been previously characterized as human pseudogenes in the literature. Among these 16, four have been functionally characterized in mouse, which are *NRADD*, *NEPN*, *CETN4*, and *GUCY2D*. Table S5 describes various subsets constructed using the 35 "definite" losses. All four candidates do not have detectable homologs in the human genome. Most strikingly, *NRADD*, *NEPN*, and *CETN4* remained functional for more than 300 My before being inactivated.

## Discussion

This study presents the first attempt to systematically identify adaptive gene losses in the human genome since the common ancestor of euarchontoglires, approximately 75 Mya. Using losses of well-established genes as the proxy for adaptive gene losses, we focused on identifying a class of pseudogenes that were once functional and retained this function through tens of millions of years of evolution. We confidently identified 26 losses of well-established genes, including 16 that were not previously known in the literature. The highlight of this analysis is the ability to automatically detect losses of genes bearing no significant homology to any functional paralog in the human genome. Their functional precursors had an ancient origin, but enough evolutionary time has elapsed to erase any significant homology with other genes in the human genome. These genes were functioning for hundreds of millions of years and silenced recently within the past 75 My.

It has been proposed that the majority of pseudogenes are either dead-on-arrival [58] or inactivated quickly after duplication [27]. Therefore, it is not surprising that we have identified a much smaller number of pseudogenes as compared to the thousands identified by previous whole genome analysis that aimed to catalog the human genome for unprocessed pseudogenes [30,31,36,64]. We overlapped our results with two well-known pseudogene databases, Yale pseudogene database, composed of mostly various computational predictions [64], and VEGA pseudogene collection, compiled by manual curation [65]. We found limited overlap between the losses identified in Table 1 with both pseudogene sets (Table S4). Only two out of 31 annotated, zero out of 14 hypothetical, and five out of 27 ORs were found by all three analyses. Neither database has *GULO*, Cardiotropin 2, or many others listed in Table 1 (see note b). A recent genome scan identified 67 human-specific gene losses, including 36 ORs [58]. Excluding ORs, only one out of the six human-specific gene losses identified in Table 1, Gpr33, was also discovered in that study [58]. Another possible overlap is

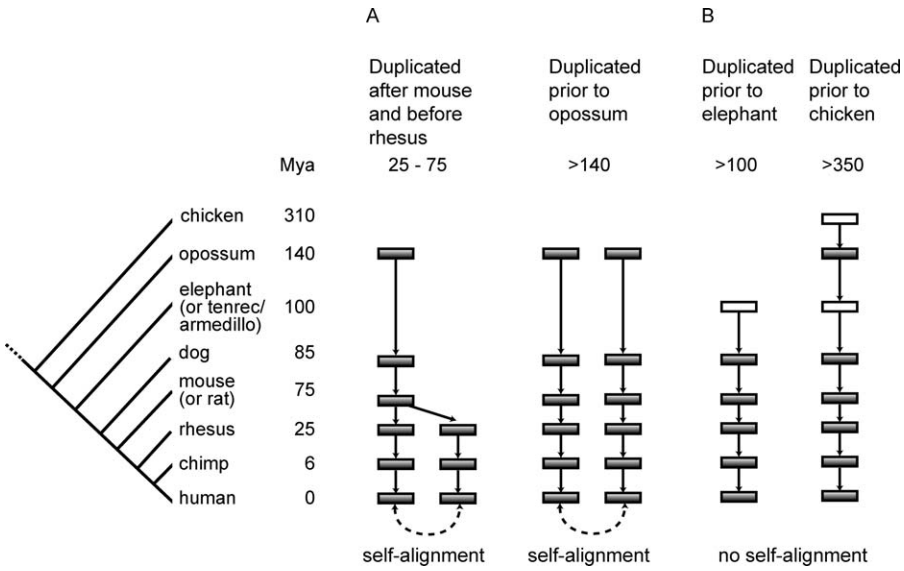


**Figure 4.** Timing of Gene Birth Is Estimated by Determining Duplication Histories of Genomic Regions Surrounding the Gene Losses

For the subset of gene losses with detectable human self-alignment, the duplication branch is determined by tracing each duplicate of the best human self-alignment through a seven-species syntenic genomic alignment. For those without detectable self-alignments, the duplication branch is determined by the seven-species syntenic alignments plus alignments with the chicken genome, and the elephant, tenrec, and armadillo scaffolds when available. Filled rectangles represent syntenic alignment to the human genome, and open rectangles represent genomes or scaffolds aligned to the human genome without the syntenic constraint. Approximate time when the species diverged from the human lineage is shown in Mya.

doi:10.1371/journal.pcbi.0030247.g004

Ugt2b1, which belongs to a tandem cluster of Ugt2B genes on Chromosome 4. The limited overlap in part reflects the difference in methodology used to identify the pseudogenes, but also makes apparent that none of these methods in their present state are able to form the complete set of losses of genes with ancient origins. It also confirms that we have identified some unprocessed pseudogenes derived from functional precursors of ancient origin, where evolution has erased any significant homology to their current functional paralogs.

The gene loss candidates shown in Table 1 are by no means a complete list of losses of well-established genes in the human lineage during the past 75 My. TransMap gene model prediction methodology is not perfect, many factors can introduce prediction errors including uncertainties in sequence alignments, errors generated by the gene model prediction and evaluation procedures, and evolutionary changes of the gene structures across mammalian species (Text S1). For example, the well-known human specific loss of *CMAH* (a CMP-sialic acid hydroxylase) [66] was not found by this analysis due to the strictness of TransMap gene model predictions, causing a valid *CMAH* gene model in the dog genome to be excluded because it featured a noncanonical GC-AG splice junction. However, the use of an outgroup genome and the mRNA filter makes the analysis far more likely to produce false negatives than false positives. Several other factors also contribute to this incompleteness. First, our method using human–mouse–dog comparison relied upon well-defined mouse genes to seed the search and valid dog predictions for outgroup confirmation. Problems in either one will return a false negative result. Our analysis missed *MYH16* because it is not in mouse RefSeq, which could be due to an independent loss or a misannotation. We further investigated its absence and found that the *MYH16* syntenic region is not present in the mouse genome, indicating an independent loss in mouse via genomic deletion. Our analysis required a valid conceptual translation in the dog genome, which may fail to occur due to TransMap prediction errors, sequencing gaps, or an independent loss in dog. However, the chance of producing a valid mapping increases if multiple outgroups, such as the opossum genome [67] or a computationally reconstructed ancestral genome [68], were used and the resultant gene loss predictions were combined. For example, the previously documented human specific loss of Htr5b [69] can be identified using a reconstructed boreoeutherian genome as the outgroup (Haussler lab, unpublished data). Our analysis can also be improved by extending our seed mRNAs to include those from other species and by using multiple outgroup genomes. For example, using chimpanzee *MYH16* mRNA as a seed could have found this pseudogene in human.

Our analysis may not identify human polymorphic gene losses. For example, the human-specific loss of *CASP12* [58,59] was not identified by our analysis because the latest human genome assembly (NCBI release 36) has the functional allele. Several other human polymorphic losses were also missed by our analysis for the same reason [70,71]. These polymorphic null alleles are potentially crucial to human diseases, e.g., *CASP12* in sepsis and *CCR5* in HIV infection. Incorporating human EST and mRNA information, as was done by Hahn et al. [70,71], or the human SNP dataset [72], could help our method identify human polymorphic gene losses. Overlapping those alleles with human disease loci, such as those documented in OMIM database [73] or identified by genetic association studies, might lead to the identification of new human disease associated genes. Another factor that may cause the method to overlook gene losses is related to segmental duplication. After a gene is duplicated, both the

**Table 2.** Losses of Well-Established Genes in the Human Genome, Classified Using Estimated Functional Time Length

| Row | Gene Birth Branch | Gene Death Branch | Functional Time Length (My) | Mouse Gene Name |
|---|---|---|---|---|
| 1 | >OP | H | >134 | Ctf2 |
| 2** | H | H | 0–6 | Cyp2g1 |
| 3 | >OP | H | >134 | Gpr33 |
| 4 | >elephant / tenrec | H | >94 | 1700013G24Rik |
| 5** | H | H | 0–6 | Sord |
| 6** | C-R | H | 0–6 | S100a15 |
| 7 | >OP | C–G | >132 | Acyl3 |
| 8 | >OP | C–O | >126 | Crygf |
| 9 | >OP | C–R | >115 | Taar4 |
| 10 | >OP | C–R | >115 | Uox |
| 11 | >chicken | O–R | >300 | Nradd |
| 12 | >elephant | O–R | >75 | Gsta4 |
| 13 | >elephant | O–R | >75 | Sult1d1 |
| 14* | M-D | O–R | 30–71 | Pfpl |
| 15 | >chicken | O–R | >300 | 4933429E10Rik |
| 16* | >D | R–M | >10 | LOC433492 |
| 17 | >OP | C–M | >115 | Tex21 |
| 18 | >chicken | Marm–M | >300 | 0610012H03Rik |
| 19 | >elephant./ armadillo | R–Marm | >60 | Gucy2d |
| 20 | >chicken | R–Marm | >300 | Nepn |
| 21* | >D | R–Marm | >35 | BC018465 |
| 22 | >chicken | R–Marm | >300 | Gm766 |
| 23 | >chicken | Marm–M | >300 | Gulo |
| 24* | D-OP | R–M | 10–115 | Abca14 |
| 25 | >chicken | C–R | >300 | Cox8b |
| 26 | >tenrec / armadillo | R–M | >60 | 4921517D21Rik |
| 27 | >elephant | R–M | >60 | BC048502 |
| 28 | >chicken | R–M | >300 | 2700097O09Rik |
| 29 | >tenrec | R–M | >60 | 2610318N02Rik |
| 30 | >chicken | R–M | >300 | D730039F16Rik |
| 31 | >armadillo | R–M | >60 | B430306N03Rik |
| 32** | R–M | R–M | 0–50 | Cxcl7 |
| 33** | R–M | R–M | 0–50 | Unc93a |
| 34 | >chicken | Marm–T.shrew | >300 | Cetn4 |
| 35 | >chicken | Marm–T.shrew | >300 | Slc7a15 |
| +27 olfactory receptors | | | | |
| +2 UDP glucuronosyltransferase | | | | |
| +4 keratin proteins | | | | |
| +1 cytochrome P450 (Cyp4f16) | | | | |
| +1 retinol dehydrogenase (Rdh7) | | | | |
| +1 tryptase (Prss32) | | | | |
| +1 hypothetical protein similar to putative membrane protein Re9 (2310042E22Rik) | | | | |

The functional time length (in My) of a gene is estimated using the separation between the gene birth branch (duplication branch) and the gene death branch (pseudogenization). Twenty-six gene losses have lower-bound estimations of their functional time length greater than 50 My and are classified as losses of well-established genes (those not marked with asterisks), five were alive for less than 50 My and are classified as duplication-induced pseudogenes (labeled with "**"). Four are undetermined because their estimated functional time length overlaps with the 50 My threshold (labeled with "*"). H: human; C: chimpanzee; G: gorilla; O: orangutan; R: rhesus; Marm: marmoset; T.shrew: tree shrew; M: mouse; D: dog; OP: opossum.

doi:10.1371/journal.pcbi.0030247.t002

ancestral copy (the copy in the original genomic context) and the daughter copy (the copy duplicated in the new genomic context) are equally subject to degenerative mutations. Since our analysis evaluates based on the status of the ancestral copy, if evolution silences the daughter copy, it will not be identified by our method. However, this type of false negative is quite limited in our results because it only applies when a segmental duplication occurred after the boreoeutherian common ancestor. Treating the daughter copy in the same way as the ancestral copy will solve this problem, except in the case of a tandem segmental duplication, where it is difficult to distinguish the ancestral copy from the daughter copy.

Among the 26 losses of well-established genes, six were identified to be lost independently in the human and old world monkey lineages (numbers 8, 11, 12, 13, 15, 25 in Table 1). This can be interpreted as a confirmation for adaptive evolution, if we believe that a common selection pressure forced these genes to be lost in separate clades. Other known independent losses such as Caspase15 and Gpr33 seem to confirm this hypothesis [74,75]. An alternative interpretation is that the gene function is no longer needed, such as the loss of *GULO* in guinea pigs and humans [40]. However, it is also quite probable that the original loss did not occur independently on different lineages, but rather a common mutation that was missed by the analysis might have occurred earlier on a shared ancestor to inactivate the gene. This might have been a mutation in a noncoding region, or a mutation that

was erased by secondary mutations such as genomic deletions. For example, a prior, noncoding mutation in any of the six cases we found could have disrupted the transcription, translation, or regulatory signals of the gene in the common ancestor of old world monkeys and apes, rendering the gene effectively inactive at the time that these lineages split. Since the gene is no longer under selective pressure to maintain its integrity, secondary ORF-disrupting mutations could follow, occurring independently in the separate lineages, as observed by our analysis.

To identify genes that are truly lost, we have focused on regions lacking any reported mRNA evidence, including in cell lines derived from cancer cells. A large number of candidates with differential mutational status in the human and dog gene predictions (918 out of 1,008) were filtered out because they overlap with some mRNA evidence in humans. The majority of these are likely to be TransMap prediction errors (Text S1, Table S1). However, some pseudogenes still generate transcripts if the transcription signal is intact, and these would be overlooked by our method. An example of a transcribed pseudogene in the human genome that appears on this list is *CATSPER2* (chr15: 41815434–41825788), represented by GenBank mRNA BC066967, and BC047442. The mammalian gene collection annotates it as a transcribed pseudogene. If a pseudogene is transcribed and spliced, its mRNA transcript with ORF-disrupting mutations (i.e., premature stop codon) is targeted and degraded by the cell's RNA surveillance pathway of nonsense mediated decay [76], although this process may not be complete. Only with time will these pseudogenes will be completely silenced at the level of transcription. In addition, studies have shown that occasionally a pseudogene, like Makorin1, not only transcribes but also plays a vital biological role in stabilizing the mRNA of its homologous coding gene [77]. Thus it is difficult to prove that a transcribed pseudogene is completely nonfunctional.

Theories of molecular evolution suggest three outcomes for new genes arising from gene duplication: degeneration due to functional redundancy, evolution into a new function, or function sharing by both copies [27]. The expected time that elapses before a gene is inactivated is thought to be relatively short [27]. Lynch and Conery estimated the half-life of a new duplicate to be around 16 My in the human lineage [27,78,79]. Using this estimate, after our cutoff of 50 My, 11% of redundant genes caused by duplications are expected to be intact by chance. After 60 My (the shortest estimation that passes the cutoff in Table 2), only 7.5% will be left. Twenty-six candidates in Table 2 are classified as losses of established genes using the 50 My cutoff, and many have an estimated functional period after duplication that is much longer than 50 My. This suggests that they are likely to have evolved independent functions before pseudogenization and thus likely to be true losses of well-established genes. In addition, our method used the lower-bound estimation for the functional time length for this classification. Although the higher-bound estimations for four candidates (*PFPL*, *ABCA14*, *LOC344492*, *BC018465* in Table 2) satisfy the 50 My cutoff, their low estimations do not. As complete genome sequences for additional mammals become available in the future, the timing of duplication and pseudogenization can be greatly refined, potentially classifying some of these four candidates as losses of established genes as well.

It is nontrivial to determine whether these losses we have found were truly adaptive. It is very likely that neutral losses at dispensable loci account for a subset of our results. For example, *GULO*, a vitamin C biosynthesis gene, is thought to have been lost in primates because primates have ample dietary supply of ascorbic acid, reducing or removing the selective pressure that maintains this gene. In general, it is difficult to differentiate between neutral loss due to removal of selective pressure, as proposed by the "use it or lose it hypothesis," and positively selected adaptive loss, as by the "less is more" hypothesis, without knowing the gene's precise biological functions. Given our current knowledge of human genes, identifying the losses of established genes seem to be the best strategy in the search for more ancient (before 250 Mya) adaptive gene losses on a genomic scale. The resulting list is a much more enriched set of candidates.

In summary, our analysis identified a set of losses that are highly enriched for well-established genes in the human genome against a large background of pseudogenes. Expanding these results to include genes and genomes from the entire mammalian clade will generate a more accurate and comprehensive picture of adaptive gene losses in human evolution. From a theoretical standpoint, it will provide insight into the role that loss of functional genes plays in evolutionary adaptation [4]. The method presented here can also be generalized to discover gene losses in other organisms on a genomic scale.

## Materials and Methods

**Mapping mouse mRNAs to the human and dog genomes.** BLAT [80] alignments of mouse mRNAs sequences from Genbank [81] and RefSeq [37] to the cognate genome were obtained from the UCSC Genome Browser Database [82]. These alignments, along with the coding sequence annotations associated with the mRNAs, provide annotations of gene structure in the genome. BLASTZ [33] syntenic chained alignments [63] of cognate genome (mouse) to the target genome (human or dog) are used to project the mouse mRNA alignments to the target genome. This algorithm, known as TransMap and illustrated in Figure 1, results in predictions of orthologous gene structures in the target organism. The TransMap prediction methodology is described in detail in Text S1.

**Sources of sequences, mRNAs, and pseudogene databases.** Genomic sequences used in this study were obtained from the UCSC genome browser [82]. Sequence release of the following species are human (NCBI release 36.1; UCSC hg18 March 2006), chimp (Chimpanzee Sequencing and Analysis Consortium Build 2 version 1; UCSC March 2006), rhesus macaque (BCM HGSC version 1.0, Mmul_051212; UCSC Jan 2006), mouse (NCBI release 36; UCSC Feb 2006), dog (Broad Institute assembly version 2.0; UCSC May 2005), rat (Baylor Human Genome Sequencing Center HGSC version 3.4; UCSC Nov 2004), opossum (draft assembly produced by the Broad Institute; UCSC Jan 2006), chicken (version 2.1 draft assembly produced by the Genome Sequencing Center at the Washington University; UCSC May 2006), elephant (Broad Institute version 1.0; UCSC May 2005), tenrec (Broad Institute echTel 1.0; UCSC Jul 2005), and armadillo (Broad Institute version 1.0; UCSC May 2005). Trace sequences of orangutan, marmoset, tarsier, galago, and tree shrew were downloaded from NCBI Trace Archive. Mouse RefSeq genes were obtained from the UCSC mouse genome browser, which is consistent with the NCBI mouse genome build 36.1. Yale and VEGA pseudogene datasets were also obtained from the UCSC human genome browser track "Yale Pseudo" and track "Vega Pseudogenes." Human mRNAs filter is the GenBank human mRNAs, obtained from the UCSC human genome browser track "Human mRNAs". The identifiers of the two orangutan trace sequences used to confirm the TGG (W) codon is present in the orangutan genome are ti865941905 and ti1012155976 in the NCBI trace archive.

***ACYL3* domain analysis and structural modeling.** The mouse gene *ACYL3* (NM_177028) is predicted to contain an acyltransferase 3 (acyl3, IPR002656) and a nose resistant to fluoxetine-4 (NRF, IPR006621) domain in the InterPro database [45]. We predicted

*ACYL3* to have eight or nine transmembrane helixes in its C-terminal sequence using TMHMM v2.0 [48] and to have a 20 amino acid signal peptide in its N-terminal sequence by SignalP 3.0 [49]. We performed protein structure modeling using SAM-T05 [46]. SAM (dssp-eh12 model) predicted *ACYL3* to be a multi-pass membrane protein with four conserved cysteine residues in its N-terminal sequence and 12 helixes in its C-terminal sequence. The nonsense stop codon shared by humans and chimpanzees is located in the tenth helix of the structural prediction.

**DNA amplification and sequencing of ACYL3 region in gorilla.** Genomic DNA surrounding the *ACYL3* TGG (W) codon was PCR-amplified from a gorilla DNA sample. Degenerate PCR primers were designed based on the conservation in human, chimp, and rhesus genomic sequences. The forward primer is 5′-GGTCACCC-TATTTGCGGTGGCCGCTCTGGCATACA-3′ and the reverse primer is 5′-TGGGCTGGGTCCTCTTTGCGTGCCACNGAGGATATG-GAGGTATGGA-3′. The PCR product was sequenced in both forward and reverse directions, and results were combined to generate a 162 bp gorilla sequence.

**Determining earliest shared mutations.** The earliest shared mutations were determined by examining an alignment of genomic sequences of human, chimp, rhesus, mouse, and dog, plus 200 bp trace sequences surrounding the mutation site, when available in the NCBI trace archive, from orangutan, marmoset, tarsier, galago, and tree shrew. Trace sequence analysis was limited to candidates with only point mutations (i.e., stop codons, frameshift indels, or noncanonical splice sites). Candidates with coverage problems (>50% missing) were only analyzed to check whether the limited coverage is also shared by the chimp or rhesus genomes, and their point mutations were not analyzed except for a stop codon mutation in *GULO*.

## Supporting Information

**Figure S1.** A Histogram of the Number of TransMap Gene Models Predicted in the Human Genome against the Fraction of the Length of the Mouse mRNA Coding Sequence That Can Be Aligned

Found at doi:10.1371/journal.pcbi.0030247.sg001 (21 KB PDF).

**Table S1.** Evaluation of the TransMap Gene Model Classifications against ENCODE Gene Annotations

Found at doi:10.1371/journal.pcbi.0030247.st001 (52 KB DOC).

**Table S2.** Comparison of Sequence Coverage by TransMap and Other Sequence Aligners

Found at doi:10.1371/journal.pcbi.0030247.st002 (22 KB DOC).

**Table S3.** The Percentage of Alignments or Gene Models with a "Valid" Code Assignment

Found at doi:10.1371/journal.pcbi.0030247.st003 (20 KB DOC).

**Table S4.** Genomic Coordinates of the Lost Genes and Their Overlap with YALE and VEGA Pseudogenes

Found at doi:10.1371/journal.pcbi.0030247.st004 (58 KB DOC).

**Table S5.** Description of the Various Subsets Constructed Using the 35 "Definite Losses"

Found at doi:10.1371/journal.pcbi.0030247.st005 (27 KB DOC).

**Text S1.** Description of the TransMap Gene Prediction Methodology

Found at doi:10.1371/journal.pcbi.0030247.sd001 (49 KB DOC).

## Acknowledgments

We thank Ting Wang for critical reading of the manuscript and scientific discussions. We thank Sofie Salama for helping to carry out the Acyl3 sequencing experiment. We thank Webb Miller for the BLASTZ alignments.

**Author contributions.** JZ and DH conceived and designed the experiments. JZ, JZS, MD, and CBL performed the experiments. JZ, JZS, and THP analyzed the data. JZ, JZS, and MD contributed reagents/materials/analysis tools. JZ, JZS, and DH wrote the paper.

**Funding.** This work is supported by the National Human Genome Research Institute (JZ, CBL), the NIH Training Grant T32 GM070386 (JZS), the National Cancer Institute NO1-CO-12400 (MD), and the Howard Hughes Medical Institute (DH).

**Competing interests.** The authors have declared that no competing interests exist.

**References**

1. Tournamille C, Colin Y, Cartron JP, Le Van Kim C (1995) Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat Genet 10: 224–228.
2. Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, et al. (1996) Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. Science 273: 1856–1862.
3. Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, et al. (2004) Myosin gene mutation correlates with anatomical changes in the human lineage. Nature 428: 415–418.
4. Olson MV (1999) When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet 64: 18–23.
5. Ringelhann B, Hathorn MK, Jilly P, Grant F, Parniczky G (1976) A new look at the protection of hemoglobin AS and AC genotypes against plasmodium falciparum infection: a census tract approach. Am J Hum Genet 28: 270–279.
6. Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, et al. (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293: 455–462.
7. Ruwende C, Khoo SC, Snow RW, Yates SN, Kwiatkowski D, et al. (1995) Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature 376: 246–249.
8. Schroeder SA, Gaughan DM, Swift M (1995) Protection against bronchial asthma by CFTR delta F508 mutation: a heterozygote advantage in cystic fibrosis. Nat Med 1: 703–705.
9. Gabriel SE, Brigman KN, Koller BH, Boucher RC, Stutts MJ (1994) Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. Science 266: 107–109.
10. Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413: 519–523.
11. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, et al. (2006) An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443: 167–172.
12. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69–87.
13. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316: 222–234.
14. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.
15. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803–819.
16. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428: 493–521.
17. Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39: 197–218.
18. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, et al. (2003) Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. Science 302: 1960–1963.
19. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. (2004) Ultraconserved elements in the human genome. Science 304: 1321–1325.
20. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17: 368–376.
21. Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19: 908–917.
22. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive natural selection in the human lineage. Science 312: 1614–1620.
23. Kreitman M (2000) Methods to detect selection in populations with applications to the human. Annu Rev Genomics Hum Genet 1: 539–559.
24. Bustamante CD, Wakeley J, Sawyer S, Hartl DL (2001) Directional selection and the site-frequency spectrum. Genetics 159: 1779–1788.
25. Vanin EF (1985) Processed pseudogenes: characteristics and evolution. Annu Rev Genet 19: 253–272.
26. Mighell AJ, Smith NR, Robinson PA, Markham AF (2000) Vertebrate pseudogenes. FEBS Lett 468: 109–114.
27. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.
28. Zhang Z, Harrison PM, Liu Y, Gerstein M (2003) Millions of years of

evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res 13: 2541–2558.
29. Zhang Z, Carriero N, Gerstein M (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet 20: 62–67.
30. Suyama M, Harrington E, Bork P, Torrents D (2006) Identification and analysis of genes and pseudogenes within duplicated regions in the human and mouse genomes. PLoS Comput Biol 2: e76. doi:10.1371/journal.pcbi.0020076
31. Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. Genome Res 13: 2559–2567.
32. Khelifi A, Duret L, Mouchiroud D (2005) HOPPSIGEN: a database of human and mouse processed pseudogenes. Nucleic Acids Res 33: D59–D66.
33. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, et al. (2003) Human–mouse alignments with BLASTZ. Genome Res 13: 103–107.
34. Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, et al. (2006) PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics 22: 1437–1439.
35. Zheng D, Gerstein MB (2006) A computational approach for identifying pseudogenes in the ENCODE regions. Genome Biol 7 (Supplement 1): S13, 11–10.
36. Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, et al. (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res 12: 272–280.
37. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61–D65.
38. Gilad Y, Man O, Paabo S, Lancet D (2003) Human specific loss of olfactory receptor genes. Proc Natl Acad Sci U S A 100: 3324–3327.
39. Gilad Y, Bustamante CD, Lancet D, Paabo S (2003) Natural selection on the olfactory receptor gene family in humans and chimpanzees. Am J Hum Genet 73: 489–501.
40. Nishikimi M, Fukuyama R, Minoshima S, Shimizu N, Yagi K (1994) Cloning and chromosomal mapping of the human nonfunctional gene for L-gulono-gamma-lactone oxidase, the enzyme for L-ascorbic acid biosynthesis missing in man. J Biol Chem 269: 13685–13688.
41. Wu XW, Lee CC, Muzny DM, Caskey CT (1989) Urate oxidase: primary structure and evolutionary implications. Proc Natl Acad Sci U S A 86: 9412–9416.
42. Derouet D, Rousseau F, Alfonsi F, Froger J, Hermann J, et al. (2004) Neuropoietin, a new IL-6-related cytokine signaling through the ciliary neurotrophic factor receptor. Proc Natl Acad Sci U S A 101: 4827–4832.
43. Goldberg A, Wildman DE, Schmidt TR, Huttemann M, Goodman M, et al. (2003) Adaptive evolution of cytochrome c oxidase subunit VIII in anthropoid primates. Proc Natl Acad Sci U S A 100: 5873–5878.
44. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420: 563–573.
45. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, et al. (2000) InterPro—an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16: 1145–1150.
46. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D (1994) Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 235: 1501–1531.
47. Karchin R, Hughey R (1998) Weighting hidden Markov models for maximum discrimination. Bioinformatics 14: 772–782.
48. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567–580.
49. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783–795.
50. Choy RK, Thomas JH (1999) Fluoxetine-resistant mutants in C. elegans define a novel family of transmembrane proteins. Mol Cell 4: 143–152.
51. Choy RK, Kemner JM, Thomas JH (2006) Fluoxetine-resistance genes in Caenorhabditis elegans function in the intestine and may act in drug transport. Genetics 172: 885–892.
52. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, et al. (2007) Genome-wide atlas of gene expression in the adult mouse brain. Nature 445: 168–176.
53. Dzitoyeva S, Dimitrijevic N, Manev H (2003) Identification of a novel Drosophila gene, beltless, using injectable embryonic and adult RNA interference (RNAi). BMC Genomics 4: 33.
54. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062–6067.
55. Gavet O, Alvarez C, Gaspar P, Bornens M (2003) Centrin4p, a novel mammalian centrin specifically expressed in ciliated cells. Mol Biol Cell 14: 1818–1834.
56. Mochida Y, Parisuthiman D, Kaku M, Hanai J, Sukhatme VP, et al. (2006) Nephrocan, a novel member of the small leucine-rich repeat protein family, is an inhibitor of transforming growth factor-beta signaling. J Biol Chem 281: 36044–36051.
57. Wang X, Shao Z, Zetoune FS, Zeidler MG, Gowrishankar K, et al. (2003) NRADD, a novel membrane protein with a death domain involved in mediating apoptosis in response to ER stress. Cell Death Differ 10: 580–591.
58. Wang X, Grus WE, Zhang J (2006) Gene losses during human origins. PLoS Biol 4: e52. doi:10.1371/journal.pbio.0040052
59. Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, et al. (2006) Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am J Hum Genet 78: 659–670.
60. Meinl W, Glatt H (2001) Structure and localization of the human SULT1B1 gene: neighborhood to SULT1E1 and a SULT1D pseudogene. Biochem Biophys Res Commun 288: 855–862.
61. Morel F, Rauch C, Coles B, Le Ferrec E, Guillouzo A (2002) The human glutathione transferase alpha locus: genomic organization of the gene cluster and functional characterization of the genetic polymorphism in the hGSTA1 promoter. Pharmacogenetics 12: 277–286.
62. Ohta Y, Nishikimi M (1999) Random nucleotide substitutions in primate nonfunctional gene for L-gulono-gamma-lactone oxidase, the missing enzyme in L-ascorbic acid biosynthesis. Biochim Biophys Acta 1472: 408–411.
63. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A 100: 11484–11489.
64. Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, et al. (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res 35: D55–D60.
65. Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, et al. (2005) The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Res 33: D459–D465.
66. Chou HH, Hayakawa T, Diaz S, Krings M, Indriati E, et al. (2002) Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. Proc Natl Acad Sci U S A 99: 11736–11741.
67. Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, et al. (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447: 167–177.
68. Blanchette M, Green ED, Miller W, Haussler D (2004) Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res 14: 2412–2423.
69. Grailhe R, Grabtree GW, Hen R (2001) Human 5-HT(5) receptors: the 5-HT(5A) receptor is functional but the 5-HT(5B) receptor was lost during mammalian evolution. Eur J Pharmacol 418: 157–167.
70. Hahn Y, Lee B (2006) Human-specific nonsense mutations identified by genome sequence comparisons. Hum Genet 119: 169–178.
71. Hahn Y, Lee B (2005) Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. Bioinformatics 21 (Supplement 1): i186–194.
72. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308–311.
73. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33: D514–D517.
74. Rompler H, Schulz A, Pitra C, Coop G, Przeworski M, et al. (2005) The rise and fall of the chemoattractant receptor GPR33. J Biol Chem 280: 31068–31075.
75. Eckhart L, Uthman A, Sipos W, Tschachler E (2006) Genome sequence comparison reveals independent inactivation of the caspase-15 gene in different evolutionary lineages of mammals. Mol Biol Evol 23: 2081–2089.
76. Lewis BP, Green RE, Brenner SE (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A 100: 189–192.
77. Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, et al. (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature 423: 91–96.
78. Zhang L, Gaut BS, Vision TJ (2001) Gene duplication and evolution. Science 293: 1551.
79. Long M, Thornton K (2001) Gene duplication and evolution. Science 293: 1551.
80. Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12: 656–664.
81. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2004) GenBank: update. Nucleic Acids Res 32: D23–D26.
82. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. (2003) The UCSC Genome Browser Database. Nucleic Acids Res 31: 51–54.