JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 10, 2009

Pp. 1419–1427
DOI: 10.1089/cmb.2009.0085

Research Article

All Human-Specific Gene Losses Are Present in the Genome as Pseudogenes

DANIEL R. SCHRIDER, JAMES C. COSTELLO, and MATTHEW W. HAHN

ABSTRACT

The loss of previously established genes has been proposed as a major force in evolutionary change. While genome sequencing of many new species offers the opportunity to identify cases of gene loss, it is unclear which algorithms offer the greatest accuracy or sensitivity. A number of methods to identify gene losses rely on the presence of a pseudogene for each loss. If genes are deleted when lost, however, such methods will fail to identify these cases. As the fate of gene losses is still unclear, we identified gene losses through a method that does not require pseudogenes to identify human-specific gene losses. Of the several hundred probable gene losses initially identified, we were unable to find a single case of unambiguous gene loss via deletion. We were also able to identify a large number of previously unannotated genes in the human genome, some of which also had evidence for transcription. Though our results suggest that pseudogene-based methods for finding gene losses in humans will not miss many events, we discuss the dependence of these conclusions on the divergence times among the species considered. Supplementary Material is provided (see online Supplementary Material at www.liebertonline.com).

Key words: algorithms, genomics.

1. INTRODUCTION

Comparative genomic approaches to finding evolutionarily important genes have traditionally involved comparisons between orthologous protein-coding sequences. Such comparisons can identify rapidly evolving genes whose high rate of evolution may indicate the action of adaptive natural selection (Nielsen et al., 2005). Recent extensions to this approach have further considered non-coding sequences and have uncovered several regions involved in human adaptation (Dermitzakis et al., 2002; Pollard et al., 2006). The availability of high-quality genome sequences has allowed researchers to discover genes lost during evolution, where sequences are not necessarily shared between species. These changes may have also played important roles in adaptive evolution.

Gene loss is a ubiquitous phenomenon across all sequenced genomes, both eukaryotic and prokaryotic (Aravind et al., 2000; Roelofs and Van Haastert, 2001; Hughes and Friedman, 2004). Gene loss generally refers to the loss of a functional gene present in a genome for at least several million years, rather than simply the creation of new pseudogenes by gene duplication. In humans, gene loss has been proposed to be

Department of Biology and School of Informatics and Computing, Indiana University, Bloomington, Indiana.

an especially important source of adaptive change under the ''less is more'' hypothesis (Olson, 1999; Olson and Varki, 2003). A number of well-studied examples of human-specific losses are known (reviewed in Kehrer-Sawatzki and Cooper, 2007), including *CMAH* (Chou et al., 1998), *ELN* (Szabo et al., 1999), *Siglec-13* (Angata et al., 2004), and *MYH16* (Stedman et al., 2004).

In addition to these individual cases, several groups have conducted computational searches to identify human- or primate-specific gene losses via comparative genomics (IHGSC, 2004; Hahn and Lee, 2005; Wang et al., 2006; Zhu et al., 2007). These searches have discovered 6 (IHGSC, 2004), 9 (Hahn and Lee, 2005), 67 (Wang et al., 2006), and 3 (Zhu et al., 2007) gene losses specific to humans (i.e., since the split from chimpanzees). Though the methods introduced in these papers differ in their details, they have one important thing in common: they all initialize their search for gene losses using sequences currently present in the focal (i.e., human) genome. This means that they use either previously annotated pseudogenes (Wang et al., 2006), annotate their own pseudogenes (Zhu et al., 2007), or require there to be an EST for the pseudogene (Hahn and Lee, 2005). In each case, a pseudogene is defined as a genomic feature in the focal genome with homology to a functional gene in the other species, but that has lost its ability to code for a protein. Any gene loss resulting from a complete or near-complete deletion of a gene, or any sequence that has been deleted since becoming a pseudogene is therefore missed.

It is currently unknown how many gene losses have gone undiscovered because of the limitations of these algorithms. There is a slight bias towards deletions in the human genome (Kvikstad et al., 2007), which may result in the loss of many sequences no longer maintained by selection. Deletion bias is even stronger in *Drosophila* (Petrov and Hartl, 1999), which may cause methods requiring pseudogene sequences to have extremely high false negative rates when searching for gene losses. We previously examined gene loss among 12 *Drosophila* genomes and found that a large number of gene losses were completely deleted from the *D. melanogaster* genome (Costello et al., 2008). However, because of the differences in deletion biases, these results may not hold in humans. Therefore, to determine the extent to which algorithms dependent on pseudogenes may miss gene losses, we conducted an extensive analysis of apparent losses in the human genome. We were able to identify a number of human-specific gene losses, all of which we found present as pseudogenes. Our results therefore suggest that alternative algorithms may not be needed to uncover the full extent of recent gene loss.

2. DATA

2.1. Mammalian genomes

Exon sequences were acquired for all protein-coding genes in human, chimpanzee, orangutan, Rhesus macaque, mouse, rat, and dog from Ensembl v49 (Flicek et al., 2008). The genome assemblies corresponding to these genomes are as follows: NCBI36 for human, CHIMP2.1 for chimpanzee, PPYG2 for Orangutan, MMUL1 for macaque, NCBIm37 for mouse, RGSC3.4 for rat, and CanFam2.0 for dog. For each gene, any overlapping exon sequences from alternative transcripts were merged, and exons were then concatenated. When UTR sequences were available for the 5′- and 3′-most exons of the gene, they were removed from the concatenated sequences, resulting in the full set of protein-coding DNA for each annotated gene. In cases where multiple UTRs were present for the same exon, only the smallest UTR sequence was removed (i.e., sequence that was protein-coding in any transcript was included).

2.2. Defining gene families

Gene families were defined using the MCL clustering algorithm (Enright et al., 2002). Each of the concatenated exonic sequences constituting a single gene from all genomes (150,127 genes total) were BLASTed against every other gene in all species (BLASTn). A weighted undirected graph was then created, where genes are represented as nodes. Gene pairs where the average BLAST E-values were 10^{-2} were connected by an edge, with the weight of the edge equal to the negative log of their average E-value. MCL was then run on this graph using an inflation parameter of 2.3 (Demuth et al., 2006); this resulted in 25,777 gene families. As discussed in Section 3.2, this method of defining gene families is somewhat error-prone: the gene families identified are sensitive to changes in the inflation parameter (Demuth et al., 2006). However, manual verification of all gene losses detected (also discussed in Section 3.2) ensured that our final set of losses contained no false positives. Families with members only in primates, only in rodents, or

only in dog were then removed from the set of gene families to avoid the problem of inferring ancestral states for families that are not as old as the ancestor of all the mammals considered here. This left 15,960 mammalian gene families in our analysis.

3. RESULTS

3.1. Identifying human-specific gene losses

We initially identified potential gene losses along the lineage leading to humans since the split with chimpanzees (Fig. 1) by using the gene families defined by MCL as input into the program CAFE (De Bie et al., 2006). CAFE uses a birth-death stochastic process to infer the most likely number of genes present in the ancestral nodes of the species tree for each family (Hahn et al., 2005). Families for which both the chimpanzee-human most recent common ancestor (MRCA) and the chimpanzee genome were inferred to have more genes than the number of human genes in the family were inferred to have human-specific gene losses. Because annotated human pseudogenes were not included in the input to MCL, this method calls genes as absent whether or not a pseudogene can be found. For each family, the difference between the number of genes inferred to be present in the chimpanzee-human MRCA and the number of genes in human was taken as the number of human-specific gene losses. In total, we inferred a total of 234 losses in 210 gene families. For the following analyses, we only consider those families with fewer than 10 members in humans, as it was too difficult to unambiguously verify inferred losses in families this large. In total, there were only 4 cases of inferred genes losses found in families of greater than 10 members and exclusion of these cases did not affect the conclusions of our analyis.

For each family determined to have at least one human-specific loss by CAFE, a representative gene was chosen to search the human genome for similar genes that may have been missed by the Ensembl gene-calling process and therefore omitted from our gene set. These query genes were chosen by calculating, for each gene, the average E-value (using BLASTn) for each gene against all other genes in the family. The concatenated coding sequence of the gene with the lowest average E-value was then used to search the human genome. In all subsequent analyses this gene was used as the exemplar gene model for the family.

As a first step in confirming gene losses along the human lineage, the coding sequences of the 210 query genes were used to search against the human genome assembly using BLASTn. To ensure that all genes similar to the query gene were detected, all hits with E-values below the relatively high threshold of 10^{6} were recorded (with sequences too dissimilar from the query gene filtered out as described in the last paragraph of this section). Overlapping hits were then merged. To prevent exons from being counted as individual genes (and then later filtered out for being too short), hits within 500 kilobases (kb) of one another were merged. (This relatively high distance threshold allows for the possibility that tandem duplicates would be collapsed. However, any spurious losses detected due to this error would be corrected by the manual verification described in the next section.) Of the starting 210 query sequences, all of them have hits to the human genome meeting our BLAST criteria (Fig. 2). This result is in stark contrast to previous

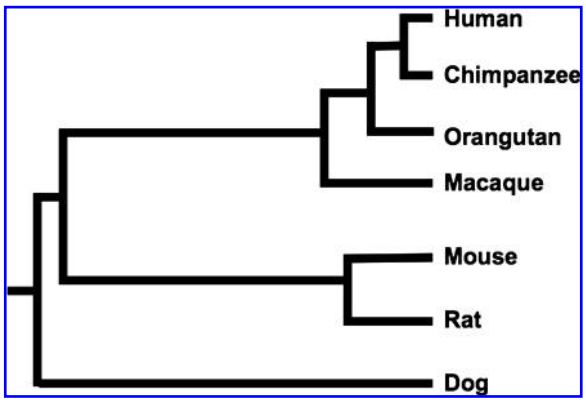

FIG. 1. Phylogeny of mammal species included in the analysis. Branch lengths approximate current estimates of divergence times. Human-specific gene losses (occurring after the human-chimpanzee split) were identified in this analysis.

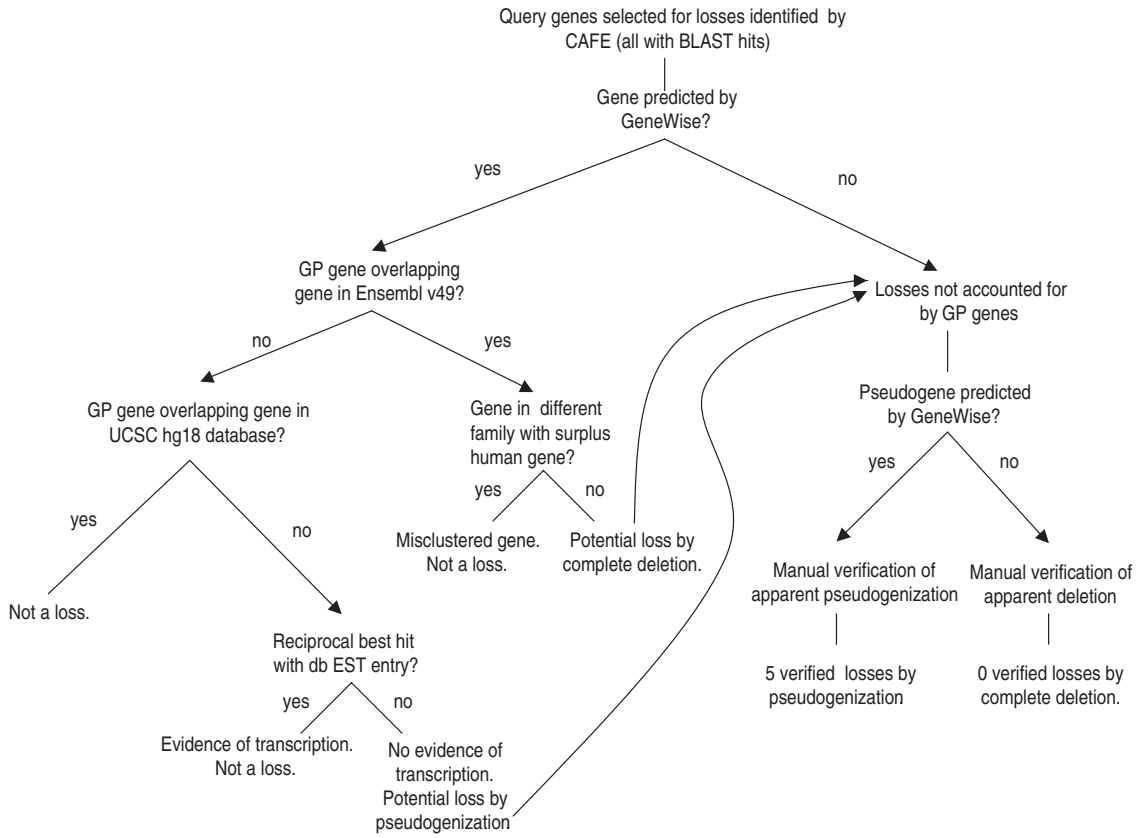

FIG. 2. Tree representing the analysis and verification of gene losses identified by CAFE. This flowchart illustrates the steps taken to detect and verify human-specific gene losses. In this figure, "GP genes" refer to protein coding genes predicted by GeneWise.

analyses from Drosophila, where homologs of almost half of all the genes lost did not hit the *D. melanogaster* genome using the same BLAST thresholds (Costello et al., 2008). Possible reasons for this difference are enumerated in the Discussion.

As the human genome is many times larger than any of the 12 sequenced Drosophila genomes (3500 megabases vs. ~150 megabases) and is filled with repetitive elements, the fact that there are weak BLAST hits for all of the query genes does not necessarily mean that none have been completely deleted. In addition—and unlike in our previous analysis of Drosophila—we have included families that have lost individual genes but may still have some representatives present in the genome (e.g., families that have contracted from two copies to one copy); query genes from these families will always hit their remaining paralogs. Therefore, to determine whether these hits represent previously known homologs, previously unknown and unannotated genes, pseudogenes, or are simply spurious similarities, we conducted further analyses. We ran GeneWise (Birney et al., 2004) to determine if the regions identified by BLAST could be translated into peptides similar to those coded for by the query gene. GeneWise was run on all merged blast hits with an additional 2 kb added upstream and downstream of the identified genomic region. For each GeneWise run, the Ensembl peptide sequence of the query gene was used as the template peptide. (In cases where the query genes had multiple ENSEMBL peptides due to alternative splicing, GeneWise was run once with each Ensembl peptide.) The pseudo option was used to filter out genes with internal stop codons and/or frameshifts relative to the peptide template. These filtered GeneWise predicted genes were then labeled as pseudogenes. The results from this analysis constitute the first major division within the candidate gene losses.

The output from GeneWise was filtered to remove sequences not likely to be genes, genes apparently dissimilar to the query gene, and redundant genes. This was done using ClustalW (Thompson et al., 1994) to construct pairwise alignments between each peptide associated with the Ensembl query gene and the peptide predicted by GeneWise. The gene predicted by GeneWise was categorized as a potential protein-

coding gene in the query gene's family if its predicted peptide was at least half the length of, and had at least 30% identity with, any of the Ensembl peptides translated from the query gene. If the GeneWise predicted gene's peptide sequence was too short relative to all peptides from the query gene, or was labeled a pseudogene by GeneWise, it was recorded as a pseudogene. Genes predicted by GeneWise that were long enough but too dissimilar to the query gene were ignored. Each potential gene was also compared to the release 33 list of annotated human pseudogenes from www.pseudogene.org (Karro et al., 2007). Any GeneWise predicted genes overlapping an entry in this database were categorized as a pseudogene. Of the starting 210 query genes, potential genes in the human genome were identified for 194. The remaining 16 gene families had a predicted pseudogene loss or no gene/pseudogene was predicted (Fig. 2).

3.2. Query genes with a potential homolog in the human genome

For the 194 query genes with a potential homolog predicted by GeneWise, we first checked whether the predicted peptide represented a known Ensembl gene included in our dataset. If the predicted gene matched a known Ensembl gene in the same family as the query (e.g., the one remaining human gene in a family that contracted from two copies to one copy), this hit was ignored as it does not account for any losses. However, some genes predicted by GeneWise match known Ensembl genes in other families, which may represent either ancient paralogs or a misclustering of genes into families by MCL. To examine possible misclustering of genes, we asked whether these known genes were in families that had more human than chimpanzee members—in these cases the apparent loss in one family is most likely explained by a gain in the other family. We found that in 18 cases genes were initially called as losses due to gene family misclustering, and therefore do not represent true losses (Fig. 2). Our previous analysis of Drosophila genomes revealed only 1 apparent loss due to misclustering (Costello et al., 2008), though a different method was used to group genes into families. These results therefore indicate that the previously used method—Fuzzy Reciprocal BLAST—may be a more accurate way of constructing gene family relationships than the MCL clustering method.

All other genes predicted by GeneWise that were categorized as potential genes and did not hit a known Ensembl gene were then compared to all known UCSC genes from the hg18 human genome release from the UCSC genome browser (Karolchik et al., 2004). A GeneWise predicted gene matching an UCSC annotated gene was considered to be a previously known gene not annotated in Ensembl v49; therefore, losses inferred in families of genes similar to these genes are not actual losses (Fig. 2). The remaining GeneWise genes are suggestive of either pseudogenes or missed annotations (i.e., new genes not included in the Ensembl v49 or UCSC version hg18 annotation of the human genome). To test whether the genes predicted by GeneWise have evidence for transcription we used them to search against dbEST (Boguski et al., 1993) using BLASTn. For each gene that hit an EST, the sequence of the best hit was then acquired from GenBank and BLASTed back against the human genome. Each GeneWise gene that was the reciprocal best hit of an EST was categorized as a new gene with evidence for transcription; there were 65 predicted genes in this category. All 198 remaining GeneWise genes may represent either new genes without evidence for transcription or pseudogenes without obvious coding-region alterations but that nonetheless no longer function. As dbEST contains over 8 million ESTs from humans, it seems unlikely that the remaining open reading frames are truly functional genes. Therefore, these genes were treated as pseudogenes for the remainder of our analysis.

The new (non-Ensembl) genes found in the preceding analyses still do not account for all the losses in families with multiple human-specific losses. Therefore, for each gene family initially considered to have lost a gene, the number of non- Ensembl genes (with EST evidence or present in the UCSC hg18 annotation) was subtracted from the inferred number of human-specific losses in the family. Any family in which there was still a difference between the inferred chimp-human ancestral family size and the updated human family size was considered to have a loss by either deletion or inactivation, and were examined further along with those families that had no predicted genes by GeneWise.

3.3. Query genes that do not have a predicted homolog in the human genome

For 16 gene families, GeneWise did not predict a peptide at least half as long or at least 30% similar to the query sequence, and for 116 more families, at least one human-specific loss remained even after adding the GeneWise-predicted genes to the family (top-right of Fig. 2). To determine whether these cases represent pseudogenes still present in the genome or complete gene losses via deletion—or are simply

annotation errors—we further examined the results provided by GeneWise and conducted manual comparisons of genome alignments between the human and chimpanzee genomes.

Of the genes predicted by GeneWise using the 210 query genes as templates, 1348 were called as pseudogenes by GeneWise. Adding these instances to the pseudogenes found by comparison to pseudogene.org, cases for which GeneWise called a peptide that was <50% the length of the query gene, and genes predicted by GeneWise absent from the UCSC and Ensembl databases and lacking evidence of transcription, there were 92 total losses in 89 families apparently represented by a pseudogene (Fig. 2). (This was determined by subtracting the number of remaining losses for each family by the number of pseudogenes found for that family.) Note that since some families had more pseudogenes found than they had losses inferred by CAFE, many pseudogenes found did not account for gene losses. The excess pseudogenes may be ''dead on arrival'' duplications. We manually examined these pseudogenes and their orthologous sequence in the chimpanzee using an updated version of the Ensembl database (v52). We found that many of the chimpanzee protein-coding genes initially called as functional in earlier annotations had either been ''retired'' in v52 (i.e., they were removed from the gene set) or were clearly misidentified as protein-coding genes in v49 of Ensembl (e.g., genes with a 1-bp intron: ENSPTRG00000024286). In total, we were only able to confirm 5 of these pseudogenes as representing true human-specific losses (see Discussion).

For the remaining 47 losses in 43 families, no pseudogene or remnant of a protein-coding gene was found by GeneWise; these cases could therefore represent gene loss via deletion. In order to confirm that these were deletions, we again manually examined an alignment of the orthologous regions between the human and chimpanzee genomes. We found that none of these apparent losses were due to deletion, and in fact, none represent losses at all. The most common cause of our misidentification was due to a human gene that was split into two separate genes in the chimpanzee annotation (and in many other mammals). Both of these chimp genes would cluster with the single human gene, but of course no pseudogene would ever be found to account for this apparent ''loss.'' It is also possible that the human gene is an incorrect fusion of two distinct genes, though in either case there is no human-specific loss of a gene. In summary, we did not find a single gene loss in humans that occurred via deletion of the entire ancestral copy.

4. DISCUSSION

Identifying cases where previously functional genes maintained by natural selection are lost is one of the novel and important challenges posed by comparative genomics. Though a large number of pseudogenes have been identified in many genomes (Harrison et al., 2005), the vast majority of pseudogenes identified are duplicated genes that were never maintained by selection. A number of new methods have been used to find true gene losses, but they require the remnants of lost genes to be identified in the target genome (IHGSC, 2004; Hahn and Lee, 2005; Wang et al., 2006; Zhu et al., 2007). Alternatively, true gene losses can be found by identifying genes in other species that do not have significant similarity to annotated genes in the target genome (Demuth et al., 2006; Hahn et al., 2007). Though this clustering method does not require the presence of pseudogenes, it may misidentify gene losses when genes present in the target genome are not clustered with their homologous genes. Additionally, this method can be used to detect any pseudogene with an orthologous functional gene, regardless of whether or not they have a functional paralog. Such genes may not be detected by within-genome searches for pseudogenes paralogous to functional genes (Zheng et al., 2007).

Here we have used this clustering method to determine whether any human-specific gene losses would be missed by methods that require the presence of a pseudogene. Briefly, we obtained exon sequences (excluding UTRs) for protein-coding genes in seven mammalian species, clustered them into families using MCL, and inferred human-specific gene losses via CAFE (De Bie et al., 2006) (Fig. 2). We then verified losses by searching for un-annotated genes similar to those in gene families inferred to have lost members, and verified all losses remaining after this step by examining genomic alignments. Note that because we did not examine gene losses in large families (>10 paralogs) or in families restricted only to other primate species, these do not represent the full set of losses that have occurred along this lineage (but see below). It does mean, however, that we are unambiguously able to assign a mechanism of loss to each case.

Of the 234 candidate gene losses we originally identified, only 6 appear to be unambiguous gene losses (by pseudogenization) along the lineage leading to modern humans (Table S1) (see online Supplementary

Material at www.liebertonline.com). For 51 of the candidate gene losses we were able to identify a human gene annotated in the UCSC genome database, and for another 55 cases we predicted a new human gene based on both gene structure and the presence of ESTs for the gene. A further 18 of these losses were simply due to the misclustering of genes into families by MCL. Finally, a large number of candidates that are not losses are instead errors in the Ensembl annotation of either the human or chimpanzee genomes. These annotation errors occur for a variety of reasons; namely, short open-reading frames in chimpanzee that have subsequently become unannotated or are clearly not genes (because they have 1-bp introns), or single genes that are annoated as being split apart in chimpanzee and other mammals (or fused in humans).

The small number of losses found using these methods was surprising, as previous analyses had counted either 86 (Demuth et al., 2006) or 80 (Wang et al., 2006) human-specific gene losses. Though 8 losses inferred by CAFE in 4 large gene families were not considered here, we noticed that our results did not include many known pseudogenes in humans. To check our results, we first compared our human-specific gene losses against cases of well-known, experimentally verified losses. It appears that many of these known pseudogenes were mistakenly annotated as protein-coding genes in Ensembl v49 (Table S1) (see online Supplementary Material at www.liebertonline.com). For instance, the human-specific loss of *MYH16* has been associated with major structural changes to human cranial morphology (Stedman et al., 2004). However, it was annotated as a protein-coding gene in Ensembl v49, which was the source of our gene set. The word ''pseudogene'' was subsequently added to the gene description in v51 (November 2008), though it was still considered a ''known protein coding gene,'' and in v52 Ensembl (December 2008) it was annotated as a ''known pseudogene.'' The *MYH16* pseudogene has an inactivating frameshift mutation in exon 18. Since a large ORF still remains (>600 amino acids), it is not surprising that automated methods like GeneWise misclassified such cases as functional genes. Similarly, some of our predicted functional genes may also in fact be pseudogenes, as we have only required that a gene be at least 50% the length of its homolog to be considered functional. These cases may explain at least some of the discrepancy between our results and previous results on the absolute number of human-specific losses. In order to further incorporate updated annotations of the human genome, we compiled a list of protein-coding genes that were retired from the Ensembl annotation between versions 49 and 52, and where the chimpanzee homolog(s) was not retired (so that they represent human-specific losses). This analysis provided an additional 22 gene losses in humans, all of which were present as pseudogenes.

Our analyses revealed no examples of human genes that were lost via complete deletion of the protein-coding sequence. All losses either had nearly full-length pseudogenes or large, identifiable pieces present in the human genome. Even for a loss that was previously described—informally—as a deletion, we were able to find a significant portion of the remaining pseudogene. The lost gene, *Siglec-13*, has been called a ''gene deletion'' by Angata et al. (2004), but by aligning the chimpanzee gene to the human genome we were able to find remnants of all but the final exon. We were surprised by the complete absence of full gene deletions for a number of reasons. First, our analyses in Drosophila had previously revealed that a large proportion of gene losses were due to deletions (Costello et al., 2008). While there are several important differences between the two studies, we expected to identify at least a small number of deleted genes. Second, most molecular mechanisms responsible for the duplication of genes also result in the deletion of genes. Any unequal crossing-over event between tandemly arranged sequences or non-allelic homologous recombination event between dispersed sequences will result in both a chromosome with one extra sequence element (i.e., a duplication) and a chromosome with one fewer (i.e., a deletion). Only duplication via retrotransposition does not also produce an allele with a deleted gene. As there are hundreds of human-specific gene duplications (Demuth et al., 2006; Hahn et al., 2007) it is reasonable to expect a commensurate number of deletions. Our results instead indicate that there may be strong selection against the fixation of alleles containing completely deleted genes.

Our results do come with several caveats. First, we have not accounted for the mechanisms of gene loss in very large gene families because of the difficulty involved in assigning specific causes to each loss event. It is possible that some losses in these families are due to complete deletions, though we know of no reason why it should be more common in families with more than 10 members. Second, we did not include primate-specific families in our analyses, and it is possible that human-specific losses in these families (i.e., families with genes present only in chimpanzee, orangutan, and Rhesus macaque) were caused by deletions. However, further analyses of the four human-specific losses we found in these families revealed pseudogenes present for each one (data not shown). Finally, it is formally possible that even losses for which we were able to find a pseudogene were actually losses due to deletions. This scenario could come

about if the pseudogene we have identified in the human genome had been produced shortly before the original gene was completely deleted. In the unlikely case that this scenario occurred, we would have accounted for the loss with the pseudogene.

Our results suggest that pseudogene-based methods for identifying gene losses will be successful because few to no genes are lost via deletion. However, this conclusion masks a more complicated result. In the recent article by Zhu et al. (2007), the authors state that, ''gene loss normally leaves behind a pseudogene.'' Motivated to determine the accuracy of this statement, we have examined the pattern of gene loss in the human genome. While our results support this claim, they do not specify for how long the pseudogene will remain in the genome. In other words, most of these losses may indeed have left behind a pseudogene, but over time these pseudogenes may be degraded beyond recognition. So while a comparison of the human and chimpanzee genomes can depend on finding pseudogenes for all losses, it is not clear that a human-mouse comparison would be as reliable.

This result raises the issue of the timeframe over which pseudogene-based methods can be used. It is obvious that pseudogene-based methods cannot be used beyond the limits of our ability to identify the homologs of pseudogenes, and it may simply be that they are inappropriate or less useful in rapidly evolving lineages or over long timescales. Our previous analyses of gene loss along the lineage leading to *D. melanogaster* revealed many lost genes that had been completely deleted (Costello et al., 2008), but most of these genes had been lost many millions of years ago. Only one gene was completely deleted between the sister species *D. melanogaster* and *D. simulans* (Costello et al., 2008), which split 2–5 million years ago (Clark et al., 2007). This implies that the results from humans and Drosophila are quite comparable given the appropriate divergence time, as there is little evidence for the complete deletion of a gene as the causal mutation in gene loss. It is only over longer periods of time that pseudogenes accumulate either point mutations or deletions that make them unidentifiable. Future studies aiming to identify lineage-specific gene losses should take the divergence times between species into account, as pseudogene-based methods are likely only accurate for detecting relatively recent losses, while family-based methods can be used to detect more ancient losses.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (grant DBI-0543586 to M.W.H.).

DISCLOSURE STATEMENT

No competing financial interest exists.

REFERENCES

Angata, T., Margulies, E.H., et al. 2004. Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. *Proc. Natl. Acad. Sci. USA* 101, 13251–13256.

Aravind, L., Watanabe, H., et al. 2000. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl. Acad. Sci. USA* 97: 11319–11324.

Birney, E., Clamp, M., et al. 2004. GeneWise and Genomewise. *Genome Res.* 14, 988–995.

Boguski, M.S., Lowe, T.M., et al. 1993. dbEST—database for ''expressed sequence tags.'' *Nat. Genet.* 4, 332–333.

Chou, H.H., Takematsu, H., et al. 1998. A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. *Proc. Natl. Acad. Sci. USA* 95, 11751–11756.

Clark, A.G., Eisen, M.B., et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450, 203–218.

Costello, J.C., Han, M.V., et al. 2008. Limitations of pseudogenes in identifying gene losses. *RECOMB–CG 2008* 14–25.

De Bie, T., Demuth, J.P., et al. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271.

Demuth, J.P., De Bie, T., et al. 2006. The evolution of mammalian gene families. *PLoS ONE* 1, e85.

Dermitzakis, E.T., Reymond, A., et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 420, 578–582.

Enright, A.J., Van Dongen, S., et al. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584.

Flicek, P., Aken, B.L., et al. 2008. Ensembl 2008. *Nucleic Acids Res.* 36(Database issue): D707–D714.

Hahn, M.W., De Bie, T., et al. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.* 15: 1153–1160.

Hahn, M.W., Demuth, J.P., et al. 2007. Accelerated rate of gene gain and loss in primates. *Genetics* 177, 1941–1949.

Hahn, M.W., Han, M.V., et al. 2007. Gene family evolution across 12 Drosophila genomes. *PLoS Genet.* 3, e197.

Hahn, Y., and Lee B., 2005. Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. *Bioinformatics* 21(Supp 1), i186–i194.

Harrison, P.M., Zheng, D., et al. 2005. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res.* 33, 2374–2383.

Hughes, A.L., and Friedman R., 2004. Recent mammalian gene duplications: robust search for functionally divergent gene pairs. *J. Mol. Evol.* 59, 114–120.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945.

Karolchik, D., Hinrichs, A.S., et al. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32(Database issue), D493–D496.

Karro, J. E., Yan, Y., et al. (2007). Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.* 35, D55–D60.

Kehrer-Sawatzki, H., and Cooper D.N., 2007. Understanding the recent evolution of the human genome: insights from human-chimpanzee genome comparisons. *Hum. Mutat.* 28, 99–130.

Kvikstad, E.M., Tyekucheva, S., et al. 2007. A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLoS Comput. Biol.* 3, 1772–1782.

Nielsen, R., Bustamante, C., et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 3, e170.

Olson, M.V. 1999. When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* 64, 18–23.

Olson, M.V., and Varki A., 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nat. Rev. Genet.* 4, 20–28.

Petrov, D.A., and Hartl. D.L., 1999. Patterns of nucleotide substitution in Drosophila and mammalian genomes. *Proc. Natl. Acad. Sci. USA* 96, 1475–1479.

Pollard, K.S., Salama, S.R., et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443, 167–172.

Roelofs, J., and Van Haastert P.J., 2001. Genes lost during evolution. *Nature* 411, 1013–1014.

Stedman, H.H., Kozyak, B.W., et al. 2004. Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* 428, 415–418.

Szabo, Z., Levi-Minzi, S.A., et al. 1999. Sequential loss of two neighboring exons of the tropoelastin gene during primate evolution. *J. Mol. Evol.* 49, 664–671.

Thompson, J.D., Higgins, D.G., et al. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680.

Wang, X., Grus, W.E., et al. 2006. Gene losses during human origins. *PLoS Biol.* 4, e52.

Zhu, J., Sanborn, J.Z., et al. 2007. Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput. Biol.* 3, e247.

Address correspondence to:
Dr. Matthew W. Hahn
Department of Biology
Indiana University
Bloomington, IN 47405

E-mail: mwh@indiana.edu