Limitations of Pseudogenes in Identifying Gene Losses

James C. Costello[1,2], Mira V. Han[1,2], and Matthew W. Hahn[1,2]

[1] School of Informatics, Indiana University, Bloomington, IN, 47408, USA
[2] Department of Biology, Indiana University, Bloomington, IN, 47405, USA

Abstract. The loss of previously established genes has been proposed as a major force in evolutionary change. While the sequencing of many new species offers the opportunity to identify cases of gene loss, the best method to do this with is unclear. A number of methods to identify gene losses rely on the presence of a pseudogene for each loss. If genes are completely or largely removed from the genome, however, such methods will fail to identify these cases. As the fate of gene losses is still unclear, we attempt to identify losses using nine Drosophila genomes and determine whether these lost genes leave behind pseudogenes in the lineage leading to *D. melanogaster*. We were able to find 109 cases of unambiguous gene loss. Of these, a maximum of 18 have identifiable pseudogenes, while the other 91 do not. We were also able to identify a large number of previously unannotated genes in the *D. melanogaster* genome, most of which also had evidence for transcription. Though our results suggest that pseudogene-based methods for finding gene losses will miss a large proportion of these events, we discuss the dependence of these conclusions on the divergence times among the species considered.

1 Introduction

Comparative genomic approaches to find 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 adaptive natural selection (*e.g.* ref. [1]). Recent extensions to this approach have further considered non-coding sequences and have uncovered several regions involved in human adaptation [2,3]. The availability of high-quality genome sequences has also allowed researchers to discover genes lost during evolution, where sequences are not necessarily shared between species. These changes may also have played important roles in adaptive evolution.

Gene loss is a ubiquitous phenomenon across all sequenced genomes, both eukaryotic and prokaryotic [4,5,6]. Gene loss generally refers to the loss of a functional gene present in a genome, rather than simply the creation of new pseudogenes by gene duplication. In humans, gene loss has been proposed to be an especially important source of adaptive change under the "less is more" hypothesis [7,8]. A number of well-studied examples of human-specific losses are known, including CMAH [9], ELN [10], Siglec-13 [11], and MYH16 [12]. In addition to these individual cases, several groups have conducted computational

C.E. Nelson and S. Vialette (Eds.): RECOMB-CG 2008, LNBI 5267, pp. 14–25, 2008.

searches to identify human- or primate-specific gene losses via comparative genomics [13,14,15]. These searches have collectively discovered over a hundred new gene losses in humans. 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 [14], annotate their own pseudogenes [15], or require there to be an EST for the pseudogene [13]. In each case, a pseudogene is defined as a genomic feature in the focal genome with homology to a functional gene in 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 bias towards deletions in the human genome [16], which may result in the loss of many sequences no longer maintained by selection. Deletion bias is even stronger in Drosophila [17], which may cause methods requiring pseudogene sequences to have extremely high false negative rates when searching for gene losses. However, the publication of 12 Drosophila genomes [18,19] provides a novel comparative genomic dataset that offers the opportunity to identify recent gene losses with unprecedented resolution. Therefore, to determine the extent to which algorithms dependent on pseudogenes may miss gene losses, we conducted an extensive analysis of apparent losses among the genomes within the Sophophora sub-genus of Drosophila (which includes the model organism, *D. melanogaster*). We were able to identify a large number of gene losses along the lineage leading to *D. melanogaster*, only a small fraction of which are present as pseudogenes. Additionally, we examined two *D. melanogaster* genome assemblies and annotations in order to highlight the effect of genome annotation on identifying gene losses. Our results suggest that alternative algorithms may be needed to uncover the full extent of gene loss across species.

2 Data

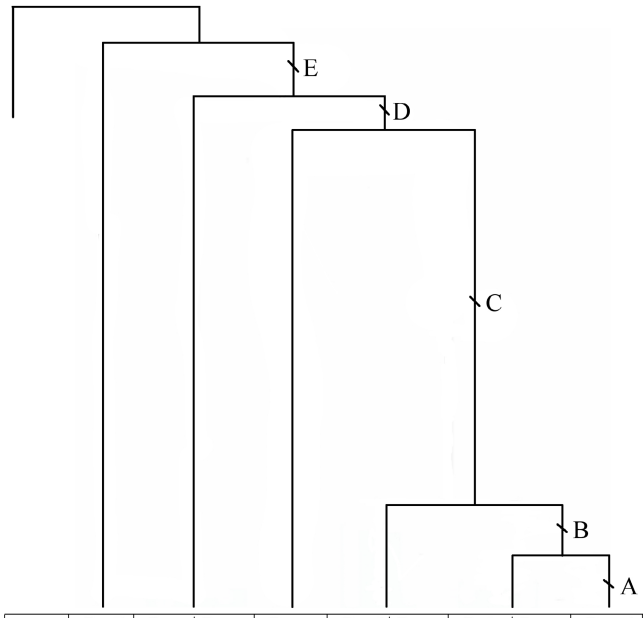
2.1 Drosophila Genomes

The sequences of 12 Drosophila genomes were recently used to compare the complement of protein-coding genes among species [18,19]. In 11 of the 12 species (all except *D. melanogaster*) *de novo* gene prediction was conducted to establish the set of genes in each genome, including in the previously sequenced *D. pseudoobscura* [20]. We used both the reconciled set of predicted genes from the newly sequenced species in the Sophophora sub-genus and the assembly and annotations from *D. melanogaster* v4.3 to initially identify gene losses; these are the same set of genes used for these genomes in the main analyses of ref. [18] and ref. [19]. The genomes in the Sophophora sub-genus are: *D. melanogaster* (*Dmel*), *D. simulans* (*Dsim*), *D. sechellia* (*Dsec*), *D. yakuba* (*Dyak*), *D. erecta* (*Dere*), *D. ananassae* (*Dana*), *D. pseudoobscura* (*Dpse*), *D. persimilis* (*Dper*),

and *D. willistoni* (*Dwil*). The additional 3 sequenced Drosophila genomes are *D. grimshawi* (*Dgri*), *D.virilis* (*Dvir*), and *D. mojavensis* (*Dmoj*).

2.2 Defining Gene Families

Gene families were defined using the Fuzzy Reciprocal BLAST (FRB) method introduced in ref. [19]. FRB compares all proteins in a reciprocal manner between



	Dwil	*Dpse*|*Dper*	*Dana*	*Dyak*|*Dere*	*Dsim*|*Dsec*	*Dmel*	**Count**
A	1	1	1	1	1	0	124(1|7)
B	1	1	1	1	0	0	11(0|2)
C	1	1	1	0	0	0	22(17|4)
D	1	1	0	0	0	0	30(21|4)
E	1	0	0	0	0	0	60(52|1)
Total							**247**(91|18)

Fig. 1. At the top of the figure is the phylogeny for the sub-genus Sophophora. The letters on the phylogeny represent the timing of candidate gene losses. The table below the phylogeny shows the breakdown of all the 247 candidate gene losses considered. A "1" indicates that at least one gene is present in a given gene family and a "0" indicates the absence of a gene defined for a given gene family. The underlined values represent the species from which one gene per gene family was used as a query to the *D. melanogaster* genome. In the case where two species are sister to *D. melanogaster*, genes from the better assembled genomes (*Dpse, Dyak, Dsim*) were taken if possible. The left-most column corresponds to the letters on the phylogeny. The right-most column shows the number of candidates in each category of gene loss, as well as the number of complete losses and pseudogenes in parentheses; displayed as (complete losses|pseudogenes). In total, there are 109 identified gene losses (91 + 18).

all pairs of genomes using BLASTp. Instead of using only the reciprocal best hit, FRB uses a rank-based method to identify potential homologs of each protein. The genes are then clustered based on their reciprocal similarity scores so that the resulting families are maximally connected and disjoint from one another. The method results in families that include both orthologs and paralogs, but has a propensity to break down families into 1:1:1...1 matches across species. This aspect of FRB allows us to easily identify homologs of candidate gene losses. Among the sequenced Drosophila genomes, FRB identified 11,434 families present in the most recent common ancestor of all 12 species, comprising a total of 148,326 genes. By comparing the number of genes within a family across species we were able to identify genes that appear to have been lost along each lineage as shown in Figure 1. See ref. [21] for further details.

2.3 Drosophila Sequences

To verify gene losses in *D. melanogaster* we searched against both the assembly and annotation of this genome used in the initial definition of gene families as well as an updated version. Both v4.3 and v5.3 *D. melanogaster* sequences were downloaded from the FlyBase ftp website.[1] Coordinates for *D. melanogaster* sequences (coding sequences and pseudogenes) were extracted from the fasta headers.

D. melanogaster EST sequences were downloaded from the Berkeley Drosophila Genome Project (BDGP) website.[2] Gene models of the eight non-*melanogaster* Sophophora species were defined by the GLEANR consensus set of the Drosophila Genome Sequencing and Analysis Consortium [19].

3 Results

3.1 Gene Losses

We initially identified potential gene losses along the lineage leading to *D. melanogaster* since the split with *D. willistoni* (Figure 1) by using fuzzy reciprocal BLAST [21]. Because annotated *D. melanogaster* pseudogenes were not used as input to FRB, this method calls genes as absent whether or not a pseudogene can be found. Here we consider only those cases of potential gene losses where a single loss has occurred. This means that the gene family containing the lost gene is required to have at least one intact homolog present in all of the sister branches to the lineage of interest, including at least one homolog in *Dmoj*, *Dvir*, or *Dgri*. For example, for a gene to be considered lost in the *melanogaster* group (*Dmel, Dsec, Dsim, Dyak, Dere, Dana*), there must be at least one gene from the same family present in the obscura group (*Dpse, Dper*), one in the *willistoni* group (*Dwil*), and one among the Drosophila sub-genus species *Dmoj, Dvir,* or *Dgri* (case "D" in Figure 1). All cases involving the parallel loss of genes were therefore not considered. However, because of the low sequence coverage

[1] ftp://www.flybase.net/

[2] http://www.fruitfly.org/sequence/dlcDNA.shtml

of several of the Drosophila genomes, we used the annotations of closely related sister species to eliminate apparent parallel losses that were due to missed predictions in genomes with low sequence coverage. Therefore the following species were treated as individual lineages: *Dsim|Dsec*, *Dyak|Dere*, and *Dpse|Dper*. Figure 1 shows the counts of candidate gene losses in relation to *D. melanogaster*. In total, 247 gene families from the FRB results met the criteria listed above.

For each of the 247 gene families, one gene was selected as a query sequence and used for further analysis. The gene sequence selected was taken from the most closely related species that contained an intact protein-coding gene homologous to the lost gene. Figure 1 identifies the species from which query sequences were taken for each case of gene loss. Since gene families are defined only for protein-coding genes, the coding sequence for a given query gene was used in all subsequent analyses.

As a first step in confirming gene losses along the *D. melanogaster* lineage, the coding sequences of the 247 query genes were searched against the *D. melanogaster* genome using BLASTn. The results from this search constitute the first major division within the candidate gene losses. Of the starting 247 coding sequences, 133 have hits to the v4.3 *D. melanogaster* genome meeting our BLAST criteria (e-value $< 10^{-6}$, percent identity $> 80\%$, and hit length > 40), while 114 do not have a significant hit (Figure 2).

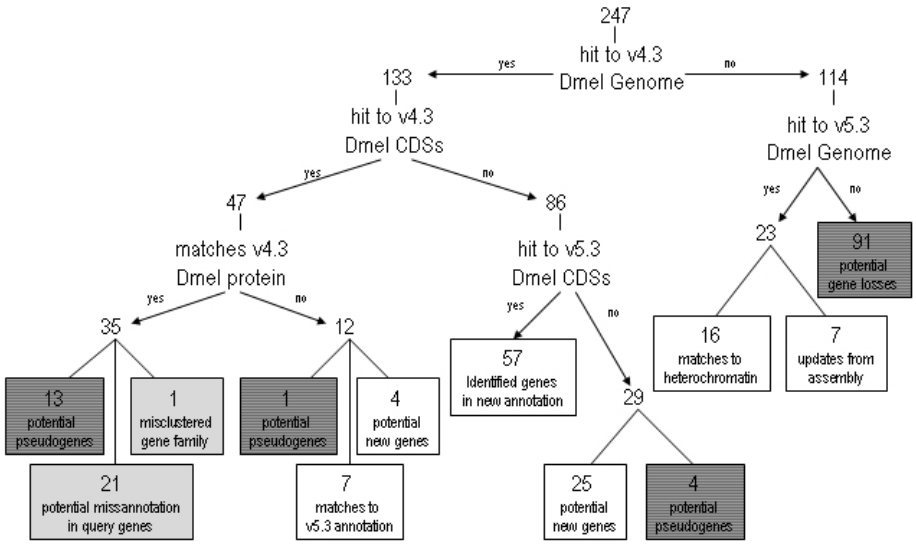


Fig. 2. Results of the gene loss analysis. The boxes shaded with horizontal stripes represent potential gene losses through either pseudogenization or complete removal of a gene from the genome. The white boxes represent genes that were not annotated or improperly annotated in v4.3 of the *D. melanogaster* genome assembly and annotation, but that are called as potential new genes in our analyses. Many of the genes missed in the v4.3 genome are in fact called gene models in the updated v5.3.

Query genes that do not hit the *D. melanogaster* genome. The 114 query genes not hitting the v4.3 *D. melanogaster* genome were first checked against the v5.3 genome assembly and annotation to determine if any of these potential gene losses are simply due to gaps in the v4.3 assembly. The 114 query sequences were searched against the v5.3 *D. melanogaster* genome using BLASTn with the same criteria as before. Interestingly, 23 query genes hit very strongly to genes predicted in the v5.3 genome. Of these 23, 16 mapped to heterochromatin and 7 mapped to euchromatin. The 7 hits to euchromatin are clear examples of gaps in the assembly that have been closed from v4.3 to v5.3 of the *D. melanogaster* genome. The 16 new genes found in heterochromatin are due in large part to recent efforts towards sequencing heterochromatic regions of the *D. melanogaster* genome [22,23].

As an additional verification that these 23 query genes do map to the *D. melanogaster* genome and are not gene losses, these sequences were searched against the *D. melanogaster* EST library using BLASTn with an e-value cutoff of 10^{-6}. Of these, 22 of the 23 query sequences have matches to ESTs, suggesting that they are true genes missed in previous assemblies. The one query sequence that did not map to an EST is dpse_GLEANR_9567, which hits a predicted gene located on an unmapped contig of the *D. melanogaster* genome.

The 91 query genes that do not have a hit to the *D. melanogaster* genome (both v4.3 and v5.3) meeting our requirements are likely losses of genes that were completely removed from the *D. melanogaster* genome. An alternative explanation for not finding these 91 genes is that any remaining remnants of the pseudogenes have been degraded beyond the detectable limits of the given BLAST parameters. To demonstrate that this is not the case, we ignored the percent identity and sequence length cutoffs and also lowered the BLASTn e-value cutoff from 10^{-6} to 10^{-3}. We did not recover a single additional hit to the v4.3 *D. melanogaster* genome using these criteria. A third potential reason for not being able to find these 91 proteins is that the coding regions lie in heterochromatic DNA that was not assembled into either the v4.3 or v5.3 *D. melanogaster* genome. Although this is unlikely given the progress that has been made on recent versions of the *D. melanogaster* genome where great efforts have been taken to fully sequence the heterochromatic regions [22]. We wanted to verify that this was not the case. As mentioned above, the 91 query genes (being a subset of the total 114) were searched against the v5.3 *D. melanogaster* genome with no hits to the heterochromatin; however, potential heterochromatic regions may still exist. Because genes located in heterochromatic regions are assumed to be transcribed, we reasoned that evidence for transcribed sequences could be used to find unassembled genes. In other words, a match to an unmapped EST would suggest a transcription unit that is not assembled into the current *D. melanogaster* genome release. We carried out this check by searching the set of 91 query genes against the *D. melanogaster* EST library with a BLASTn e-value cutoff of 10^{-6}, but no reliable hits were found. These 91 genes therefore represent good cases of gene loss with no identifiable pseudogenes.

Query genes that hit the *D. melanogaster* genome. The 133 sequences that hit the *D. melanogaster* genome were analyzed to determine whether they matched an annotated coding sequence. The physical chromosome coordinates from the v4.3 *D. melanogaster* BLAST results were checked against the physical coordinates of all *D. melanogaster* coding sequences, which resulted in 47 query sequences overlapping at least one *D. melanogaster* coding sequence and 86 not overlapping a *D. melanogaster* coding sequence. These two sets are further explored in the next two sections.

Query genes that do not overlap a *D. melanogaster* coding sequence The set of 86 non-*melanogaster* query genes that hit part of the *D. melanogaster* genome but do not overlap with any *D. melanogaster* coding sequences were first tested against the v5.3 *D. melanogaster* genome to identify missed genes due to poor genome annotation. Coding sequences in v5.3 *D. melanogaster* were searched with the 86 query genes using BLASTn (e-value $< 10^{-6}$, percent identity $> 80\%$, and hit length > 40). Of the 86 query sequences, 57 unambiguously mapped to a putative coding sequence in *D. melanogaster* (v5.3). All of these genes represent annotations that were added from v4.3 to v5.3. To find evidence of gene expression for the 57 query genes we searched the EST sequence database using a BLASTn e-value cutoff of 10^{-6}, resulting in 50 sequences that had EST evidence and only 7 sequences that did not.

Improved annotations can explain 57 of the 86 query sequences, but does not explain the remaining 29 hits to the genome. These 29 cases are suggestive of either pseudogenes or missed annotations (*i.e.* new genes not included in v5.3). To test whether the regions hit by these 29 genes have evidence for transcription, we queried the non-*melanogaster* coding sequences against the *D. melanogaster* EST library. We found good matches to ESTs for 21 genes and no matches for the remaining 8 genes.

To evaluate the potential gene structure of these 29 regions in the *D. melanogaster* genome, we performed gene predictions using GeneWise [24]. We used translated query peptide sequences compared with two different lengths of *D. melanogaster* genomic regions ($\pm$ 2,000 bases or $\pm$ 5,000 bases from the BLAST hit) as input to GeneWise and used the output peptide of longest length as our gene model. Out of the 29 cases, 25 regions were identified as novel genes that are missing even in the v5.3 annotation of *D. melanogaster*. These 25 regions of the *D. melanogaster* genome have valid exon structures that align across the whole query gene without any nonsense or frameshift mutations. Eighteen of these newly predicted genes have independent supporting evidence, such as a perfect match to a third party annotated *D. melanogaster* protein in the nonredundant database of NCBI. Of these, 17 also overlap with new annotations predicted in ref. [18] and have EST evidence.

Finally, the four remaining regions have predicted exons that align only partially to the query gene or have nonsense/frameshift mutations, and are identified as pseudogenes. None of these four pseudogenes have any EST evidence.

Query genes that hit a *D. melanogaster* coding sequence. The set of 47 genes overlapping at least one *D. melanogaster* coding sequence suggests either misannotation or misclustering of the input genes, or requires some other explanation for their high similarity to genes present in *D. melanogaster*. To determine which of these scenarios may have occurred, we conducted further analyses.

In order to verify whether the *D. melanogaster* gene matching the query sequence is indeed a protein homolog, we again used GeneWise to predict exons in the genomic region using the query protein. We then used BLASTp to query the predicted peptide against the v4.3 proteins. We found 12 cases where the peptide matches the genomic nucleotide sequence but does not match an annotated protein in v4.3. Of these, 4 cases appear to be novel genes that overlap already annotated proteins. Because they are overlapping genes present in the current annotation we found significant nucleotide similarity, but no protein similarity. EST evidence was found for all four novel predictions. Another 7 cases match the nucleotide sequence of predicted genes in v4.3 that have since been updated with new predictions in v5.3. In all of these cases, the v5.3 predicted protein is in a different reading frame than the previously annotated gene, and this new protein has significant similarity to the peptide predicted by GeneWise. Our predicted peptides did not have significant protein similarity to the v4.3 annotations. The one remaining predicted peptide does not have a hit to v5.3 and only partially aligns to fragments of the query gene, and therefore is identified as a pseudogene.

The remaining 35 cases do have a matching *D. melanogaster* protein in v4.3, but still fail to cluster together in the same family. We found that 21 of the 35 peptides only partially match the *D. melanogaster* protein in the far 5' or 3' ends of the gene. For all of these cases the query gene is much shorter than the *D. melanogaster* gene it is aligned to. For a few cases the query gene matches a short first exon of the *D.melanogaster* protein that resides more than 10,000 bases upstream of the second exon. We suspect that these are misannotations in the other Drosophila species, where the *de novo* gene prediction program has predicted short exons at either end of long genes as separate genes. It is possible that a gene fusion event has occurred along the *D. melanogaster* lineage [25], though these generally do not occur between initially adjacent genes.

In 1 of the 35 cases, the gene family of the matching *D. melanogaster* gene appears to have one extra member, meaning that the matching *D.melanogaster* gene should have been placed with the query gene in order to explain the gene loss. This is the only case that appears to represent an apparent loss explained by the misclustering of gene families by FRB. For the remaining 13 cases (of the 35) there are one or more genes in the non-*melanogaster* species that are already clustered with the matching *D. melanogaster* gene, and the alignment among those genes is better than the alignment between the query gene and the *D. melanogaster* gene. These cases represent ancient duplications predating the base of the Drosophila tree, for which a gene is lost in one of the paralogous lineages and the query sequence is hitting the other paralog. These represent gene losses, though the high similarity to intact paralogs make it hard to unambiguously say whether a pseudogene is present in the *D. melanogaster* genome.

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 (*e.g.* ref. [26]), 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 (*e.g.* refs. [13,14,15]). Alternatively, true gene losses can be found by identifying annotated genes in other species that do not have significant similarity to genes in the target genome [27,21]. Though this 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 or when there are gaps in the genome sequence.

Here we have used this latter method to determine the utility of algorithms that require the presence of pseudogenes to identify gene losses. While we have not run any of these algorithms on the Drosophila dataset used here, by finding gene losses that do not have pseudogenes we are able to estimate the maximum number of genes that could be identified by such methods. By closely examining a number of cases, we are also able to extend previous results to judge the accuracy of methods based only on the lack of significantly similar genes (*i.e.* ref. [21]).

We initially identified 247 candidate gene losses along the lineage leading to *D. melanogaster*. Note that because we ignored parallel gene losses, these do not represent the full set of losses that have occurred along this lineage since the split with *D. willistoni*. It does mean, however, that we are unambiguously able to assign losses to a specific branch of the tree (Figure 1).

Of the 247 genes we initially identified as candidate gene losses, 109 appear to be unambiguous losses along the lineage leading to *D. melanogaster*. The vast majority of candidates that do not appear to be losses are instead genes that were not annotated in earlier versions of the *D. melanogaster* genome. Some of these were not annotated because of gaps in the genome assembly ($n = 7$), unsequenced heterochromatic regions ($n = 16$), or were simply not found by previous gene-finding algorithms ($n = 86$). The large majority of the annotation updates account for the 124 gene loss candidates between the *Dsim*|*Dsec* and *Dmel* lineages (Figure 1, row A), thus artificially inflating potential gene losses between sister species. We also found a large number of losses on branches D and E relative to C (Figure 1), a result consistent with previous estimates of loss rates along these lineages [21]. The v4.3 *D. melanogaster* genome, though out of date, still represents one of the most high quality assemblies and annotations available, particularly in a metazoan genome. These annotation updates illustrate the large influence that genome assembly and annotation can have on identifying gene losses. Additionally, that this "finished" genome can be missing so many gene annotations attests to the difficulties in identifying eukaryotic protein-coding genes in large genomes. In fact, 29 of the newly predicted proteins from this study are still not included in the v5.3 *D. melanogaster* annotation.

We were only able to identify 5 pseudogenes out of the 109 unambiguous gene losses, though for 13 cases this has not been determined definitively. This result implies that methods depending on the presence of pseudogenes to identify gene losses will find a maximum of 18 losses (5+13) along this lineage. Missing 83% of all gene losses would appear to be a major disadvantage of these methods.

However, the apparent failure of these methods in identifying gene losses masks a more complicated result. In the recent paper by Zhu *et al.* [15] 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 using nine Drosophila species with respect to the *D. melanogaster* lineage. Despite the 91 cases of total gene loss without the presence of a pseudogene, our results appear to at least partly support the Zhu *et al.* [15] supposition: only one of these 91 cases corresponds to the complete removal of a recently lost gene (Figure 1, row A). In other words, most of these losses may indeed have left behind a pseudogene, and only over time have these pseudogenes been degraded beyond recognition. Because there are only a few recent ($<$ 10 million years) losses in *D. melanogaster* among the set considered here, it is hard to determine exactly what proportion initially leave behind a pseudogene as opposed to being completely deleted.

This result also raises the issue of the timeframe over which pseudogene-based methods can be used. For example, the Zhu *et al.* [15] study used the mouse genome to predict gene models of human pseudogenes. Though the divergence time between human and mouse is much greater than even the most distantly related Drosophila, the level of nucleotide divergence is equivalent to approximately the *Dmel-Dyak* split; comparing *D. melanogaster* and *D. willistoni* is equivalent to comparing the human genome to a lizard genome [18]. 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. It should be reiterated, however, that these problems do not result in any false positives, only false negatives.

In contrast to pseudogene-based methods, the clustering method used here identified a large number of gene losses across all time-scales of comparison. While we have not determined how many gene losses potentially identified by pseudogene-based methods were not identified by our clustering method, we expect this number to be small. If a pseudogene were present in the *D. melanogaster* genome, our method should also identify the loss of a homologous gene in the relevant gene family. The clustering method did result in a single false positive due to misclustering of genes into families, but this case was easily identified through follow-up analyses. Finally, the clustering method has the added property of finding a large number of previously unannotated genes initially identified by the lack of homologous proteins in *D. melanogaster* [21,18]; it also found a number of cases of misannotation in the other Drosophila species that can be fixed. These fortuitous results should be of benefit regardless of the divergence times among the genomes considered.

5 Funding

Computing resources provided by the Center for Genomics and Bioinformatics were supported in part by the METACyt Initiative of Indiana University, funded by a major grant from the Lilly Endowment. This research was supported by grants from the National Science Foundation (DBI-0543586) and National Institutes of Health (R01-GM076643A) to MWH.

References

1. Nielsen, R., Bustamante, C., Clark, A., Glanowski, S., Sackton, T., et al.: A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3, e170 (2005)
2. Dermitzakis, E., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., et al.: Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature 420, 578–582 (2002)
3. Pollard, K., Salama, S., Lambert, N., Lambot, M., Coppens, S., et al.: An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443, 167–712 (2006)
4. Aravind, L., Watanabe, H., Lipman, D., Koonin, E.: Lineage-specific loss and divergence of functionally linked genes in eukaryotes. PNAS USA 97, 11319–11324 (2000)
5. Hughes, A., Friedman, R.: Recent mammalian gene duplications: robust search for functionally divergent gene pairs. J. Mol. Evo. 59, 114–120 (2004)
6. Roelofs, J., Van Haastert, P.: Genes lost during evolution. Nature 411, 1013–1014 (2001)
7. Olson, M.: When less is more: gene loss as an engine of evolutionary change, American journal of human genetics. Am. J. Human Genet. 64, 18–23 (1999)
8. Olson, M., Varki, A.: Sequencing the chimpanzee genome: insights into human evolution and disease. Nature Rev. 4, 20–28 (2003)
9. Chou, H., Takematsu, H., Diaz, S., Iber, J., Nickerson, E., et al.: A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. PNAS USA 95, 11751–11756 (1998)
10. Szabo, Z., Levi-Minzi, S., Christiano, A., Struminger, C., Stoneking, M., et al.: Sequential loss of two neighboring exons of the tropoelastin gene during primate evolution. J. Mol. Evo. 49, 664–671 (1999)
11. Angata, T., Margulies, E., Green, E., Varki, A.: Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. PNAS USA 101, 13251–13256 (2004)
12. Stedman, H., Kozyak, B., Nelson, A., Thesier, D., Su, L., et al.: Myosin gene mutation correlates with anatomical changes in the human lineage. Nature 428, 415–418 (2004)
13. Hahn, Y., Lee, B.: Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. Bioinformatics 21(suppl.1), 186–194 (2005)
14. Wang, X., Grus, W., Zhang, J.: Gene losses during human origins. PLoS Biol. 4, 52 (2006)
15. Zhu, J., Sanborn, J., Diekhans, M., Lowe, C., Pringle, T., Haussler, D.: Comparative Genomics Search for Losses of Long-Established Genes on the Human Lineage. PLoS Comput. Biol. 3, 247 (2007)

16. Kvikstad, E., Tyekucheva, S., Chiaromonte, F., Makova, K.: A macaque's-eye view of human insertions and deletions: differences in mechanisms. PLoS Comput. Biol. 3, 1772–1782 (2007)
17. Petrov, D., Hartl, D.: Patterns of nucleotide substitution in Drosophila and mammalian genomes. PNAS USA 96, 1475–1479 (1999)
18. Stark, A., Lin, M., Kheradpour, P., Pedersen, J., Parts, L., et al.: Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. nature 450, 219–232 (2007)
19. Clark, A., Eisen, M., Smith, D., Bergman, C., Oliver, B., et al.: Evolution of genes and genomes on the Drosophila phylogeny. Nature 450, 203–208 (2007)
20. Richards, S., Liu, Y., Bettencourt, B., Hradecky, P., Letovsky, S., et al.: Comparative genome sequencing of Drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution. Genome Res. 15, 1–18 (2005)
21. Hahn, M., Han, M., Han, S.G.: Gene Family Evolution across 12 Drosophila Genomes. PLoS Genet. 3, e197 (2007)
22. Hoskins, R., Carlson, J., Kennedy, C., Acevedo, D., Evans-Holm, M., et al.: Sequence finishing and mapping of Drosophila melanogaster heterochromatin. Science 316, 1625–1628 (2007)
23. Smith, C., Shu, S., Mungall, C., Karpen, G.: The Release 5.1 annotation of Drosophila melanogaster heterochromatin. Science 316, 1586–1591 (2007)
24. Birney, E., Clamp, M., Durbin, R.: GeneWise and Genomewise. Genome Res. 14, 988–995 (2004)
25. Long, M.: A new function evolved from gene fusion. Genome Res. 10, 1655–1657 (2000)
26. Zhang, Z., Gerstein, M.: Large-scale analysis of pseudogenes in the human genome. Curr. Opin. Genet. Dev. 14, 328–335 (2004)
27. Demuth, J., De Bie, T., Stajich, J., Cristianini, N., Hahn, M.: The evolution of mammalian gene families. PLoS One 1, e85 (2006)