Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes

Caroline M. Weisman[1] *, Andrew W. Murray[2], Sean R. Eddy[2,3,4]

[1] Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA [2] Department of Molecular & Cellular Biology, Harvard University, Cambridge MA, USA, [3] Howard Hughes Medical Institute, [4]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge MA, USA

* Correspondence: cweisman@princeton.edu, @WeismanCara

Summary

Comparisons of genomes of different species are used to identify lineage-specific genes, those genes that appear unique to one species or clade. Lineage-specific genes are often thought to represent genetic novelty that underlies unique adaptations. Identification of these genes depends not only on genome sequences, but also on inferred gene annotations. Comparative analyses typically use available genomes that have been annotated using different methods, increasing the risk that orthologous DNA sequences may be erroneously annotated as a gene in one species but not another, appearing lineage-specific as a result. To evaluate the impact of such "annotation heterogeneity," we identified four clades of species with sequenced genomes with more than one publicly available gene annotation, allowing us to compare the number of lineage-specific genes inferred when differing annotation methods are used to those resulting when annotation method is uniform across the clade. In these case studies, annotation heterogeneity increases the apparent number of lineage-specific genes by up to 15-fold, suggesting that annotation heterogeneity is a substantial source of potential artifact.

Introduction

Comparing the genome sequences of different organisms can yield inferences about the genetic basis of the biological differences between them. One such analysis aims to identify genes unique to a particular monophyletic group. Such genes, called "orphan genes" when restricted to one species and "lineage-specific" or "taxonomically-restricted" when restricted to a clade of several species, are interesting from the perspective of genetic and evolutionary novelty. For example, they have been proposed to underlie lineage-specific structural and functional innovations, and to be novel genes that have emerged from noncoding DNA [1-5].

Lineage-specific genes are typically identified by searching for homologs in outgroup species: genes for which homologs cannot be found are considered lineage-specific. Such analyses typically begin not with raw genome sequences, but with particular "annotations" of them: inferences about what genes they encode. Often, only genes included in these annotations are considered in the homology search [2, 6, 7].

Previous work has recognized two ways in which errors in genome annotations could produce spurious lineage-specific genes. A real gene could be annotated in the focal lineage, but its homologs incorrectly unannotated in outgroups [8-10]. Conversely, a non-genic sequence could be incorrectly annotated as a gene in the lineage, but correctly omitted in outgroups [11]. Such errors could occur even when all genomes in an analysis are consistently annotated by the same annotation methodology, but the potential for error is expected to increase if genomes are annotated by different methods, which use different criteria in determining which sequences are genic. Because comparative analyses typically depend on publicly available genomes whose annotations come from different authors and sources, such "annotation heterogeneity" is common [12-16]. Many gene annotation methods are in wide use, including custom pipelines at large bioinformatics data providers (NCBI [17], Ensembl [18]), hand-curated model organism annotation (Flybase [19], Wormbase [20]), crowd-sourced annotation (VectorBase [21]), and various software packages (Maker [22], PASA [23]), used independently or in combination, with custom parameters chosen by individual researchers.

Here we evaluate the impact of annotation heterogeneity on inferred numbers of lineage-specific protein-coding genes. We identify four clades of species with available genome sequences for which multiple different annotations are publicly available. These enable us to conduct case studies in which we compare the number of lineage-specific genes when all species are annotated with the same method ("uniform annotations") to when they are annotated with different methods ("heterogeneous annotation"). We find that annotation heterogeneity consistently and substantially increases the inferred number of lineage-specific genes. This effect is strongest when all species within the lineage are annotated with one method and all outgroup species with a different one. Our results suggest that annotation heterogeneity can produce many spurious lineage-specific genes, potentially a majority of those found in a study.

Results

Identification of clades of sequenced genomes with annotations from two methods

To directly compare lineage-specific genes found using uniform annotations and heterogeneous annotations, we manually searched the literature and bioinformatic databases for species groups in which all species were annotated with the same method, and, additionally, the same assembly of each species had been independently annotated with some other method. We used existing annotations from a variety of standard sources instead of generating our own to make results maximally representative of real studies. We identified four groups of five species: cichlids, primates, bats, and rodents (Table 1, Supplemental Table 1). For cichlids and primates, all five species were annotated with the same two methods, whereas for bats and rodents, one method was applied to all five species, and the other available annotation was from three different methods, with each species being annotated by one of the three. Each of these four groups is less than approximately 60 My old.

Different annotations of the same genome have many proteins unique to each method

Spurious lineage-specific genes may result from annotation heterogeneity when different annotation methods differentially annotate homologous sequences. Spurious lineage-specific genes may also result from such erroneous differential annotation even when a single annotation method is used, as sequence differences between the species may alter a given method's determination regarding genic status. To get a sense of how many spurious lineage-specific protein-coding genes annotation heterogeneity *per se* can produce, we compared two protein annotations of the same species to identify proteins appearing to be unique to one of the annotations. Because the underlying genome sequences are identical, any such apparently unique proteins must be spurious, due only to annotation heterogeneity.

To mimic a typical analysis, for each species' two annotations, we used BLASTP [24] for all proteins in one annotation to see if a significantly similar ($E<0.001$) homolog was present in the other annotation. Between 0.6% and 9.7% of proteins in one annotation had no significantly similar sequence in the other (Table 1). Of the 40 (20 pairs) annotations, 19 had over 1000 proteins without a significant homolog in the other annotation. In an extreme case of the cichlid *Astatotilapia burtonii*, one annotation (Broad Institute) found 4110 genes that had no significant similarities in the other (NCBI eukaryotic annotation pipeline), and 799 proteins in the NCBI annotation lacked significant similarities in the Broad annotation. These substantial differences between two annotations of one genome illustrate the potential for spurious lineage-specific genes in comparisons of different genomes.

Different patterns of annotation heterogeneity may differently affect the inferred number of lineage-specific genes

When different annotation methods are used for species within an analysis, different patterns in which those methods are arranged on the species topology are possible. These different patterns may differently affect the number of spurious lineage-specific genes produced by annotation heterogeneity. In particular, because a gene is called as "lineage-specific" if no significant homologs are found in any species outside the lineage, we expected that the number of spurious lineage-specific genes would be positively related to the overall degree of difference between the lineage and outgroup annotations.

We considered three such patterns. In the first, one annotation method is used for all ingroup species (in the lineage, the gray boxes in the figures), and a different method for all outgroup species (outside the lineage); we refer to this as "phyletic" annotation (Figure 1). In the second, one method is used for all ingroup species, but a mixture of methods is used for the outgroup species; we refer to this as "semi-phyletic" annotation (Figure 2). In the third, a mixture of methods is used for both the ingroup species and the outgroup species; we refer to this as "unpatterned" annotation (Figure 3). We used our four clades to create case studies for each pattern.

The differences in annotation methods between ingroups and outgroups are largest for phyletic annotation, intermediate for semi-phyletic annotation, and smallest for unpatterned annotation; we expected the number of spurious lineage-specific genes to scale accordingly.

Annotating a lineage with one method and outgroups with a different method greatly increases the apparent number of lineage-specific genes

Phyletic annotation occurs in at least two scenarios. Studies that newly sequence a lineage often use their own method to annotate that lineage, and may then compare it to outgroup annotations from another single source (e.g. Ensembl). Additionally, studies using existing annotations may encounter a correlation between taxon and annotation method because genome sequencing groups (with their annotation teams) often select species taxonomically (e.g. studies of particular taxa, sequencing consortia/database initiatives for particular taxa) [25, 26].

We tested the impact of phyletic annotation on the apparent number of lineage-specific genes on two groups of species, where the same genome assembly for every species had been annotated by the same two methods: five cichlids, annotated both by the Broad Institute and NCBI; and five primates, annotated both by Ensembl and NCBI (Supplemental Table 1).

For each tree of five species, we exploited the ladder-like topology (Figure 1) of the tree to perform four analyses, comparing each of the four monophyletic groups including the focal species to the remaining outgroups. For each lineage that included the focal species, we conducted a typical analysis of lineage-specific genes by identifying genes in the focal species that have a significantly similar homolog in the deepest rooted member of the ingroup (and thus are "present" in that clade), but lack significant similarity to any protein in any outgroup species in a BLASTP search (Methods). We compared the number of lineage-specific genes found when all species (both ingroups and outgroups) were annotated with the same method to the number found when the annotations for all outgroup species were switched to the other method in a "phyletic" annotation pattern (Figure 1).

Heterogeneous annotation consistently caused a large increase of hundreds to thousands of apparent lineage-specific genes, typically about a 4-fold (ranging from 1.4-fold to 15-fold) difference relative to uniform annotation. In all but one of the eight cases in Figure 1, the increase is more than 2-fold, suggesting that the majority of lineage-specific genes inferred in heterogeneous annotations are artifacts of the heterogeneity.

Annotating a lineage with one method and outgroups with a mixture of other methods increases the apparent number of lineage-specific genes

Examples of what we call "semi-phyletic" annotation, where the ingroup is annotated with one method and outgroups with a mixture of methods, are common in the literature on lineage-specific genes [12, 13, 26-31]. This can occur in scenarios similar to phyletic annotation, but where outgroup annotations are available from a mixture of sources (e.g. a combination of Ensembl and NCBI).We created case studies of semi-phyletic annotation using groups of species for which every species had been annotated both by the same method and by one of a mix of other methods: five rodents and five bats (Supplemental Table 1). We repeated the procedure described for phyletic annotation above to compare the number of lineage-specific genes in semi-phyletic annotations to those in uniform annotations (Figure 2).

Semi-phyletic annotation heterogeneity caused a smaller but still substantial increase in the number of apparent lineage-specific genes in all lineages in both groups (Figure 2). The magnitude of this effect ranged from 20 to 833 additional lineage-specific genes, corresponding to 1.2-fold to 6-fold increases.

Annotating species with a mixture of methods without taxonomic bias increases the apparent number of lineage-specific genes

Examples of what we call "unpatterned" annotation, where the annotation method varies within the ingroup as well as the outgroup, are also common in the literature [15, 16, 27, 32-35]. This occurs when studies use existing available annotations for the desired species, which may come from a variety of sources. We created case studies of unpatterned annotation using the same rodent and bat species we used for semi-phyletic annotation (Figure 2), with the difference that we always compared the uniform annotations to the full set of mixed annotations (Figure 3) to produce unpatterned annotation heterogeneity.

Unpatterned annotation heterogeneity usually caused an increase in apparent lineage-specific genes (Figure 3), though the effect was smaller than for phyletic or semi-phyletic annotations. Two cases showed equal numbers or slight decreases, and the other six cases showed increases of 1.1-fold to 5.7-fold; the largest increases were in the cases with a single outgroup species.

As expected, six-frame translation homology searches dramatically reduce the apparent number of lineage-specific genes

A homology search in which the query protein is compared directly to a six-frame translation of the target genome does not rely on an annotation of the target species, and so should reduce this source of spurious lineage-specific genes. Such translated searches have previously been shown to reduce the inferred number of lineage-specific genes [8, 9]. In agreement with these expectations, we find that, for all of the lineages described above (depicted in Figures 1-3), a search for the focal species' proteins against six-frame translations of all comparator species genomes dramatically reduces the number of lineage-specific genes: to below the number inferred with uniform annotations, and often to less than one hundred (Supplemental Table 2).

Discussion

We used six case studies to ask if varying the annotation method across species in a comparative analysis ("annotation heterogeneity") alters the apparent number of lineage-specific genes. We found that switching from uniform to heterogeneous annotations consistently increased the number of genes that were classified as lineage-specific, with increases ranging from tens to thousands of genes, corresponding to increases of up to 15-fold. The largest increases were seen when one annotation method was used for all the ingroup species and another was used for all the outgroup species ("phyletic annotation"). The smallest increases were seen when a mixture of annotation methods were used in both ingroup and outgroup species. Our case studies consist of trees of five species; mixtures of annotations in larger numbers of outgroup and ingroup species may reduce the artifact.

Annotation heterogeneity is common in comparative studies. Our results suggest that the numbers of lineage-specific genes found in these studies may be inflated, especially in "phyletic annotation" cases, and where the number of species compared is small. Annotation heterogeneity may also have consequences that we do not explore here, like producing spurious lineage-specific losses.

Recent work from us and others has shown that homology detection failure, in which homology searches fail to detect homologs that are actually present in outgroups, can also produce spurious lineage-specific genes [36, 37]. Previous studies have noted a surprisingly large number of "young" lineage-specific genes found in recently evolved clades [15], which, compared to older lineage-specific genes, are less readily explained by homology detection failure, which is minimized at short evolutionary distances. The results here are all for young (<60 My old) clades, showing that annotation heterogeneity can be a significant source of spurious lineage-specific genes in young clades.

In accordance with previous results, we show that annotation heterogeneity artifacts can be reduced by performing homology searches of six-frame translated genomic DNA sequence in search of unannotated homologs in target species. This approach has caveats. At short evolutionary distances, a sequence may be sufficiently similar for successful detection in such a search without having the same coding status as the query; for example, a truly de novo originated gene is expected to have significant nucleotide similarity to a homologous noncoding locus in close outgroup species. This approach also still relies on an accurate annotation of the focal species.

When annotation methods disagree, which is correct? Our results do not address this, only demonstrating a consequence of this disagreement. Even homogeneous annotations are imperfect. Of particular concern, methods in general rely on features (homology to known genes, length, expression level, codon optimization) that seem likely to be absent or weaker in newly evolved (*de novo*) genes, and so may fail to identify these genes. We consider annotation accuracy primarily accountable to experimental data. Testing transcription, translation, and function in all species in question is of ultimate importance in accurately identifying lineage-specific genes. In light of our results, we suggest more emphasis on these metrics. In the meantime, the true number of lineage-specific genes remains difficult to ascertain, but better understanding sources of spurious ones helps us constrain it.

Author Contributions

Conceptualization, C.W.; Formal Analysis, C.W.; Investigation, C.W.; Writing – Original Draft: C.W.; Writing – Review & Editing: C.W., A.W.M, S.R.E; Supervision, A.W.M, S.R.E; Funding Acquisition, S.R.E.

Declaration of interests

The authors declare no competing interests.

Methods

Identifying lineage-specific proteins

For each species group, we defined a protein as specific to a particular lineage if a search using BLASTP [24] version 6.2.0 had no similar protein at a significance threshold of E=0.001

in the annotation of any species that was an outgroup to that lineage. We did not require that a protein be present in all members of the lineage to be specific to that lineage: a protein was defined as specific to a lineage based on the most distant species in which it was detected. For example, if a protein in *M. musculus* was detected only in *R. norvegicus*, it was defined as specific to that lineage; if a gene in *M. musculus* was detected in *M. caroli, M. pahari, and R. norvegicus,* it was also defined as specific to that same lineage. If a protein was found in the earliest-branching member of the species group, it was considered "conserved" and so not counted as any kind of lineage-specific gene. This way of classifying lineage-specificity coheres with standard practice [6].

For the six-frame translated searches, we first generated a six-frame translation of the genome assembly of each species using the 'esl-translate' command in the hmmer easel package, and then used it as the target database in a BLASTP search, as described in the previous paragraph.

Supplemental Information

Supplemental Table 1: Sources, brief descriptions, and links to protein annotations and genome assemblies used in this study.

Supplemental Table 2: Results of six-frame translation homology searches. Numbers in the table indicate the inferred number of genes specific to the indicated lineage (corresponding to the four lineages depicted in Figures 1-3) in each of the described taxa.

Data availability

All raw results summarized in Figures 1-3 are available at https://github.com/caraweisman/Annotation_homology.

Tables

Species	Annotation 1 source	No. proteins in annotation 1	Number/percent of proteins in annotation 1 w/ no homologs in annotation 2	Annotation 2 source	No. proteins in annotation 2	Number/percent of proteins in annotation 2 w/ no homologs in annotation 1
Cichlid fish						
Metraclimia zebra	Broad Institute	51772	3592/6.9%	NCBI Eukaryotic annotation pipeline	40043	706/1.8%
Pundamilia nyererei	Broad Institute	42152	3276/7.7%	NCBI Eukaryotic annotation pipeline	38583	668/1.7%
Astatotilapia burtoni	Broad Institute	52845	4110/7.8%	NCBI Eukaryotic annotation pipeline	44653	799/1.8%
Neolamprologus brichardi	Broad Institute	36873	3568/9.7%	NCBI Eukaryotic annotation pipeline	31372	755/2.4%
Oreochromis niloticus	Broad Institute	66482	4143/6.2%	NCBI Eukaryotic annotation pipeline	47713	765/1.6%
Primates						
Macaca fascicularis	Ensembl "full genebuild"	46148	1510/3.3%	NCBI Eukaryotic annotation pipeline	62672	1044/1.7%
Macaca nemestrina	Ensembl "full genebuild"	46238	1295/2.8%	NCBI Eukaryotic annotation pipeline	66484	1623/2.4%
Mandrillus leucophaeus	Ensembl "full genebuild"	40903	1406/3.4%	NCBI Eukaryotic annotation pipeline	38336	693/1.8%
Rhinopithecus bieti	Ensembl "full genebuild"	43730	1233/2.8%	NCBI Eukaryotic annotation pipeline	49595	1476/3.0%
Cebus imitator	Ensembl "full genebuild"	40677	926/1.7%	NCBI Eukaryotic	55885	602/1.5%

				annotation pipeline		
Rodents						
Mus musculus	Ensembl "full genebuild"	68381	1523/2.2%	NCBI Eukaryotic annotation pipeline	84985	471/0.6%
Mus caroli	UCSC	50492	1496/3.0%	NCBI Eukaryotic annotation pipeline	47409	590/1.2%
Mus pahari	UCSC	50002	1596/3.2%	NCBI Eukaryotic annotation pipeline	42226	413/1.0%
Rattus norwegicus	Ensembl "mixed genebuild"	29107	458/1.6%	NCBI Eukaryotic annotation pipeline	56110	716/1.3%
Peromyscus maniculatus	Ensembl "full genebuild"	28866	267/0.9%	NCBI Eukaryotic annotation pipeline	45588	517/1.1%
Bats						
Myotis lucifugus	Ensembl "full genebuild"	20719	197/1.0%	NCBI Eukaryotic annotation pipeline	43106	622/1.4%
Myotis brandtii	Beijing Genomics Institute	19484	867/4.5%	NCBI Eukaryotic annotation pipeline	40808	1370/3.4%
Myotis myotis	Bat1K	46057	3448/7.5%	NCBI Eukaryotic annotation pipeline	61156	704/1.2%
Molossus molossus	Bat1K	53797	3107/5.8%	NCBI Eukaryotic annotation pipeline	42753	486/1.1%
Pteropus alecto	Beijing Genomics Institute	19619	1338/6.8%	NCBI Eukaryotic annotation pipeline	39706	955/2.4%

Table 1: Genome annotations used in this study. Brief description of annotation source, number of genes in the annotation, and the number and percentage of genes in each annotation with no significant homologs found by a BLASTP search in the other annotation for the given species are listed. Note that, where large differences in the number of proteins included in a pair of annotations occurs, this is often due in part to one annotation including a larger number of different isoforms of the same locus, all or many of which may have significant similarity to the same protein(s) in the other annotation.

Figures

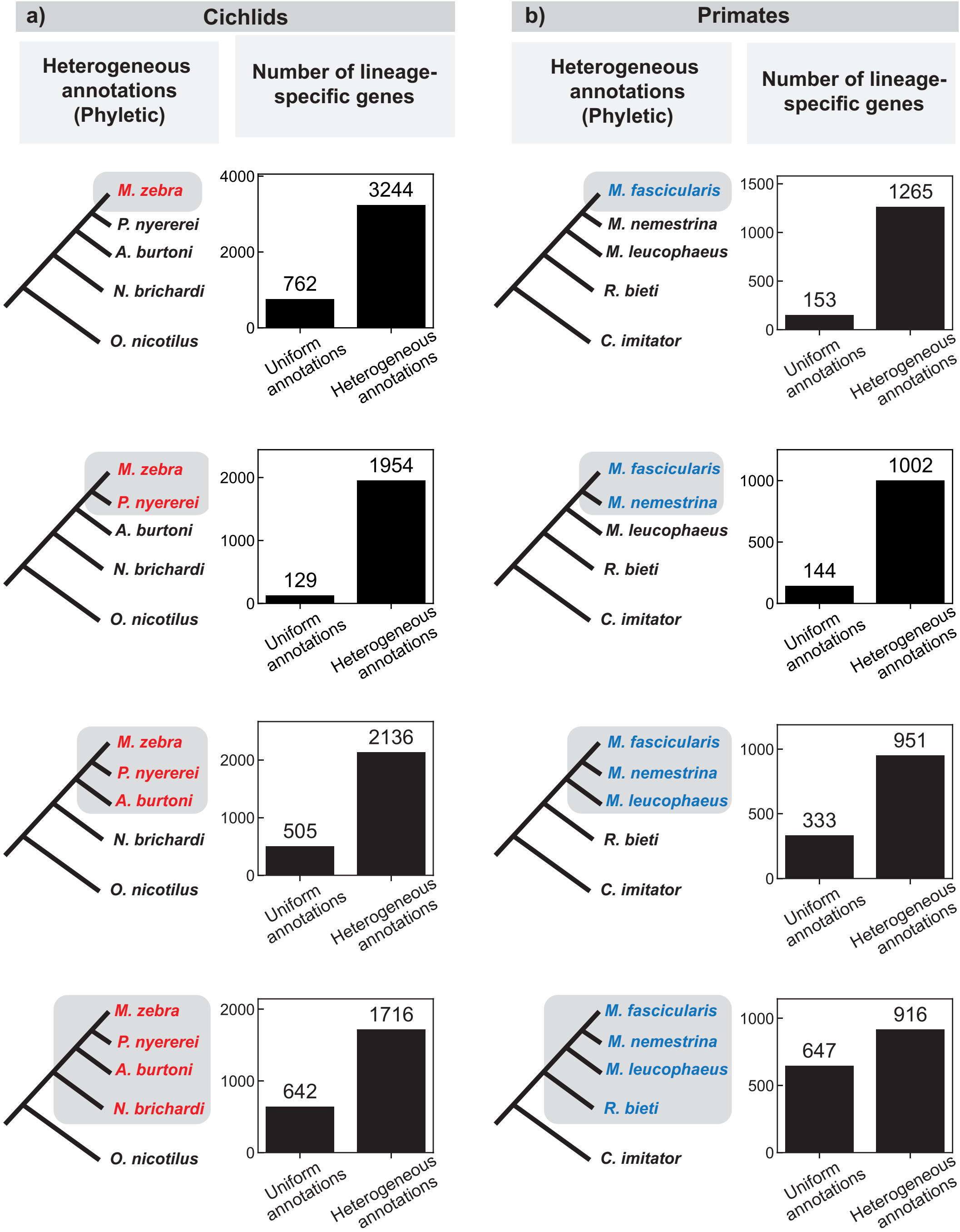
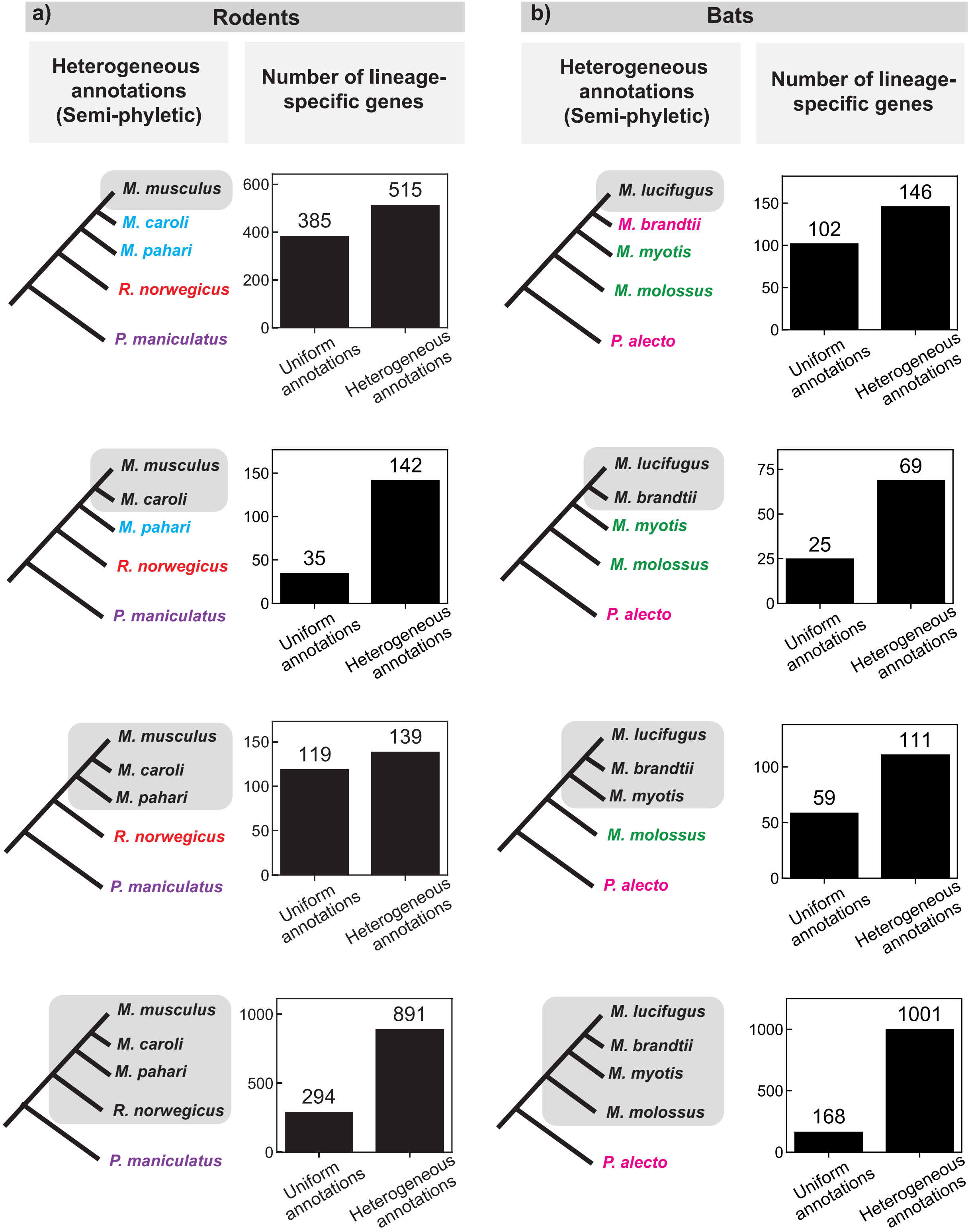
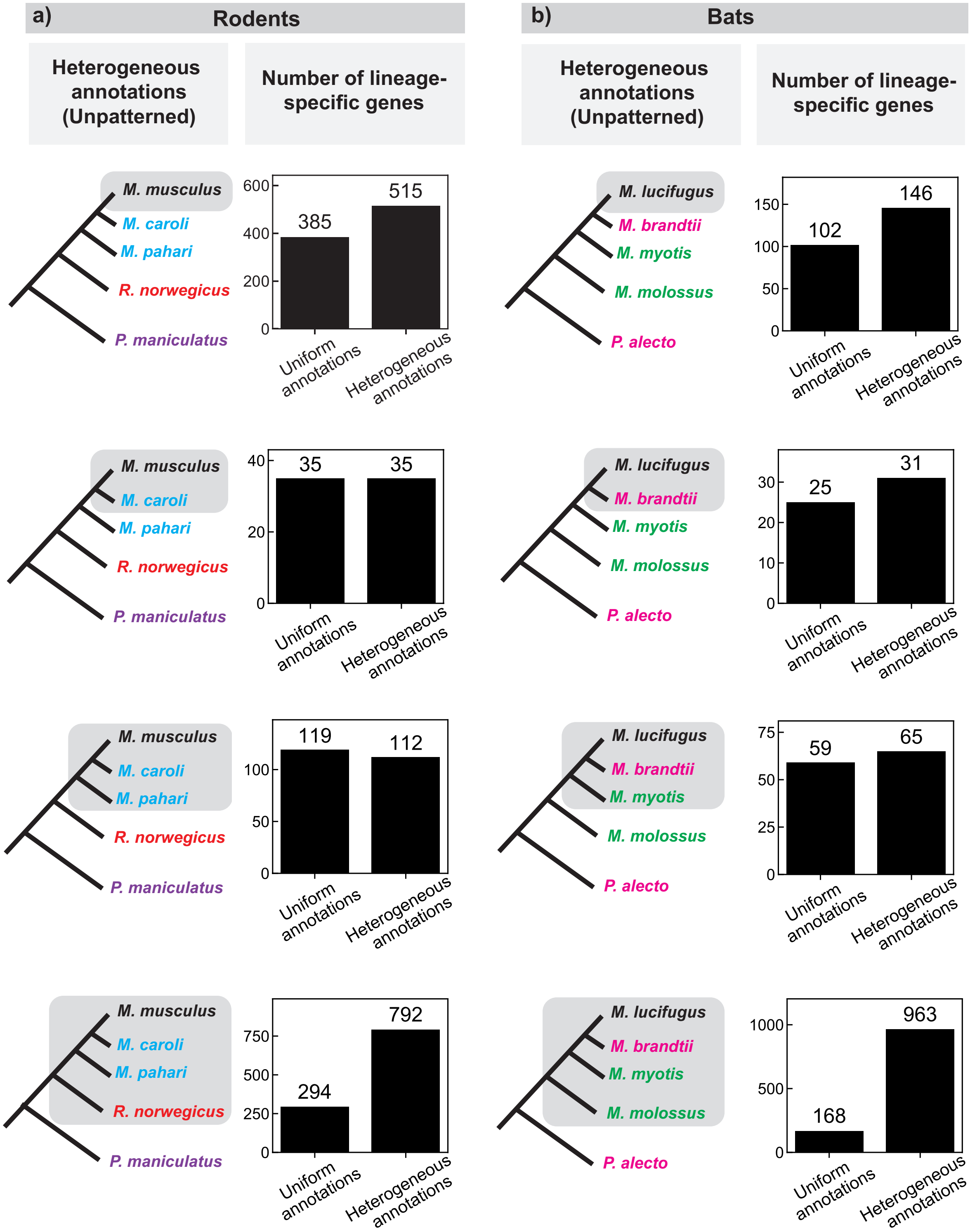

a)
Cichlids
Heterogeneous annotations (Phyletic)
Number of lineage-specific genes
M. zebra
P. nyererei
A. burtoni
N. brichardi
O. nicotilus
4000
2000
0
762
3244
Uniform annotations
Heterogeneous annotations
2000
1000
0
129
1954
2000
1000
0
505
2136
2000
1000
0
642
1716
b)
Primates
Heterogeneous annotations (Phyletic)
Number of lineage-specific genes
M. fascicularis
M. nemestrina
M. leucophaeus
R. bieti
C. imitator
1500
1000
500
0
153
1265
1000
500
0
144
1002
1000
500
0
333
951
1000
500
0
647
916

Figure 1: Comparison of the number of lineage-specific genes found using uniform and heterogeneous (phyletic) annotations in a) cichlids and b) primates. The species tree on the left indicates the lineage under consideration (grey shading); different text colors indicate different annotation sources in the heterogeneous annotation analysis (black, NCBI; red, research group at the Broad Institute; blue, Ensembl). A depiction of the uniform annotation pattern, in which all annotations are from NCBI (black), is not shown. Bar graphs indicate the number of genes that appear specific to the lineage shaded on the species tree to the left using either uniform or heterogeneous annotations.


a)
Rodents
Heterogeneous annotations (Semi-phyletic)
Number of lineage-specific genes
M. musculus
M. caroli
M. pahari
R. norwegicus
P. maniculatus
600
400
200
0
385
515
Uniform annotations
Heterogeneous annotations
150
100
50
0
35
142
119
139
1000
500
294
891
b)
Bats
Heterogeneous annotations (Semi-phyletic)
Number of lineage-specific genes
M. lucifugus
M. brandtii
M. myotis
M. molossus
P. alecto
150
100
50
0
102
146
75
50
25
0
25
69
100
50
0
59
111
1000
500
0
168
1001

327

Figure 2: Comparison of the number of lineage-specific genes found using uniform and heterogeneous (semi-phyletic) annotations in a) rodents and b) bats. The species tree on the left indicates the lineage under consideration (grey shading); different text colors indicate different annotation sources in the heterogeneous annotation analysis (black, NCBI; blue, UCSC; red, Ensembl “mixed genebuild”; purple, Ensembl “full genebuild”; green, Bat1k; pink, Beijing Genomics Institute). A depiction of the uniform annotation pattern, in which all annotations are from NCBI (black), is not shown. Bar graphs indicate the number of genes that appear specific to the lineage shaded on the species tree to the left using either uniform or heterogeneous annotations.


a)
Rodents
Heterogeneous annotations (Unpatterned)
Number of lineage-specific genes
b)
Bats
Heterogeneous annotations (Unpatterned)
Number of lineage-specific genes
M. musculus
M. caroli
M. pahari
R. norwegicus
P. maniculatus
M. lucifugus
M. brandtii
M. myotis
M. molossus
P. alecto
385
515
0
200
400
600
102
146
0
50
100
150
35
35
0
20
40
25
31
0
10
20
30
119
112
0
50
100
59
65
0
25
50
75
294
792
0
250
500
750
168
963
0
500
1000
Uniform annotations
Heterogeneous annotations

342
343
344

Figure 3: Comparison of the number of lineage-specific genes found using uniform and heterogeneous (unpatterned) annotations in a) rodents and b) bats. The species tree on the left indicates the lineage under consideration (grey shading); different text colors indicate different annotation sources in the heterogeneous annotation analysis (black, NCBI; blue, UCSC; red, Ensembl "mixed genebuild"; purple, Ensembl "full genebuild"; green, Bat1k; pink, Beijing Genomics Institute). A depiction of the uniform annotation pattern, in which all annotations are from NCBI (black), is not shown. Bar graphs indicate the number of genes that appear specific to the lineage shaded on the species tree to the left using either uniform or heterogeneous annotations.

References

1. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. More than just orphans: are taxonomically-restricted genes important in evolution? Trends in Genetics. 2009;25:404-13.

2. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nature Reviews Genetics. 2011;12:692.

3. Wilson G, Bertrand N, Patel Y, Hughes J, Feil E, Field D. Orphans as taxonomically restricted and ecologically important genes. Microbiology. 2005;151:2499-501.

4. McLysaght A, Guerzoni D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philosophical Transactions of the Royal Society B: Biological Sciences. 2015;370:20140332.

5. Tautz D. The discovery of de novo gene evolution. Perspectives in biology and medicine. 2014;57:149-61.

6. Domazet-Lošo T, Brajković J, Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends in Genetics. 2007;23:533-9.

7. McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. Nature Reviews Genetics. 2016;17:567.

8. Basile W, Elofsson A. The number of orphans in yeast and fly is drastically reduced by using combining searches in both proteomes and genomes. bioRxiv. 2017:185983.

9. Casola C. From de novo to “de nono”: the majority of novel protein-coding genes identified with phylostratigraphy are old genes or recent duplicates. Genome Biology and Evolution. 2018;10:2906-18.

10. Zile K, Dessimoz C, Wurm Y, Masel J. Only a single taxonomically restricted gene family in the Drosophila melanogaster subgroup can be identified with high confidence. Genome Biology and Evolution. 2020.

11. Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nature ecology & evolution. 2017;1:1-6.

12. Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and de novo gene birth. Nature. 2012;487:370.

13. Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, et al. A molecular portrait of de novo genes in yeasts. Molecular Biology and Evolution. 2017;35:631-45.

14. Bowles AM, Bechtold U, Paps J. The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty. Current Biology. 2020.

15. Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics. 2013;14:117.

16. Paps J, Holland PW. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nature communications. 2018;9:1-8.

17. Thibaud-Nissen F, DiCuccio M, Hlavina W, Kimchi A, Kitts P, Murphy T, et al. P8008 the NCBI eukaryotic genome annotation pipeline. Journal of Animal Science. 2016;94:184-.

18. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic acids research. 2021;49:D884-D91.

19. Drysdale R, Consortium F. FlyBase. Drosophila. 2008:45-59.

20. Howe K, Davis P, Paulini M, Tuli MA, Williams G, Yook K, et al., editors. WormBase: annotating many nematode genomes. Worm; 2012: Taylor & Francis.

21. Giraldo-Calderón GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, et al. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic acids research. 2015;43:D707-D13.

22. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome research. 2008;18:188-96.

23. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology. 2008;9:1-22.

24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215:403-10.

25. Schmitz JF, Chain FJ, Bornberg-Bauer E. Evolution of novel genes in three-spined stickleback populations. Heredity. 2020;125:50-9.

26. Prabh N, Rödelsperger C. De novo, divergence, and mixed origin contribute to the emergence of orphan genes in pristionchus nematodes. G3: Genes, Genomes, Genetics. 2019;9:2277-86.

27. Li Z-W, Chen X, Wu Q, Hagmann J, Han T-S, Zou Y-P, et al. On the origin of de novo genes in Arabidopsis thaliana populations. Genome biology and evolution. 2016;8:2190-202.

28. Heames B, Schmitz J, Bornberg-Bauer E. A continuum of evolving de novo genes drives protein-coding novelty in Drosophila. Journal of molecular evolution. 2020;88:382-98.

29. Aguilera F, McDougall C, Degnan BM. Co-option and de novo gene evolution underlie molluscan shell diversity. Molecular Biology and Evolution. 2017;34:779-92.

30. Huang J, Chen J, Fang G, Pang L, Zhou S, Zhou Y, et al. Two novel venom proteins underlie divergent parasitic strategies between a generalist and a specialist parasite. Nature communications. 2021;12:1-16.

31. Wang Y-W, Hess J, Slot JC, Pringle A. De Novo Gene Birth, Horizontal Gene Transfer, and Gene Duplication as Sources of New Gene Families Associated with the Origin of Symbiosis in Amanita. Genome biology and evolution. 2020;12:2168-82.

32. Papanicolaou A, Schetelig MF, Arensburger P, Atkinson PW, Benoit JB, Bourtzis K, et al. The whole genome sequence of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. Genome biology. 2016;17:1-31.

33. Forêt S, Knack B, Houliston E, Momose T, Manuel M, Quéinnec E, et al. New tricks with old genes: the genetic bases of novel cnidarian traits. Trends in Genetics. 2010;26:154-8.

34. Johnson BR, Tsutsui ND. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. BMC Genomics. 2011;12:164.

35. Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. BMC evolutionary biology. 2011;11:1-23.

36. Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS biology. 2020;18:e3000862.

37. Moyers BA, Zhang J. Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. Molecular biology and evolution. 2016;33:1245-56.