Article

Current Biology

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

Highlights

- "Lineage-specific genes" often come from genomes annotated with different methods
- Mixed annotation methods can cause many genes falsely appearing lineage-specific
- A majority of lineage-specific genes came from this effect in many case studies

Authors

Caroline M. Weisman, Andrew W. Murray, Sean R. Eddy

Correspondence

cweisman@princeton.edu

In brief

Some genes are "lineage specific," present only in a few related species. These are often found by comparing species whose gene repertoires have been inferred using different methods. Weisman et al. find that this practice can cause large numbers of spurious lineage-specific genes: often, a majority of the total found in an analysis.

Weisman et al., 2022, Current Biology *32*, 1–8
June 20, 2022
https://doi.org/10.1016/j.cub.2022.04.085

CellPress

Please cite this article in press as: Weisman et al., Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.085

Current Biology

CellPress
OPEN ACCESS

Article

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

Caroline M. Weisman,[1,5,6,*] Andrew W. Murray,[2] and Sean R. Eddy[2,3,4]
[1]Lewis-Sigler Institute for Integrative Genomics, Carl Icahn Laboratory, Princeton University, South Drive, Princeton, NJ 08540, USA
[2]Department of Molecular & Cellular Biology, Harvard University, Divinity Avenue, Cambridge, MA 02138, USA
[3]Howard Hughes Medical Institute, Jones Bridge Road, Chevy Chase, MD 20815, USA
[4]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Oxford Street, Cambridge, MA 02138, USA
[5]Twitter: @WeismanCara
[6]Lead contact
*Correspondence: cweisman@princeton.edu
https://doi.org/10.1016/j.cub.2022.04.085

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 previously hypothesized 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. We refer to this as "annotation heterogeneity."

Annotation heterogeneity is common. Comparative analyses typically use existing annotations rather than producing their own, potentially uniform, ones. Available annotations have been generated by a wide variety of methods. Large consortia, such as bioinformatics institutes or model organism databases, all use different methods, including custom pipelines (NCBI,[12] Ensembl[13]), hand curation (Flybase,[14] Wormbase[15]), and crowd-sourced annotation (VectorBase[16]). Individual research groups also produce annotations, usually using heterogeneous selections of one or a combination of at least 30 available software tools[17] and custom parameters. Annotations from these varied sources can often be downloaded or accessed from large centralized databases (e.g., Refseq/Genbank, Uniprot) but are not homogenized when they are deposited into them: proteomes downloaded from such databases (including NCBI's "non-redundant," or "nr," database, the default in a BLASTP web server search) or searches performed on them are therefore highly heterogeneous. Of 25 lineage-specific studies discussed

Please cite this article in press as: Weisman et al., Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.085

in a prominent 2019 review,[18] 18 (76%) depended on heterogeneous genome annotations, rather than on homogeneous re-annotations or on annotation-independent homology searches with six-frame translations of all open reading frames (ORFs) (Table S1). And of 33 studies published between 2019 and 2022, 21 (64%) depended on heterogeneous annotations, a rate not significantly different ($p = 0.58$, Fisher's exact test) than among the older studies (Table S2).

Here we explore the effect of annotation heterogeneity on inferred numbers of lineage-specific 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. 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. Full details of these annotations, the underlying assemblies, and the methods used to annotate them can be found in Table S3. Each of these four groups is less than approximately 60 million years 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. (This is a limiting case of the "phyletic annotation" described below.) Because the underlying genome sequences are identical, proteins can only appear to be "orphans," unique to one of the annotations, as a result of annotation heterogeneity.

To mimic a typical analysis, for each species' two annotations, we used BLASTP[18] for all proteins in one annotation to see if a significantly similar ($E < 0.001$) homolog was present in the other annotation (Table S4). On average, 1,380 proteins in each annotation lacked a significantly similar sequence in the other. This represented an average of 3% of total proteins, with a range of between 0.6% and 9.7%. Nineteen of the 40 annotations, or nearly half, had over 1,000 proteins without a significant homolog in the other annotation. In an extreme case of the cichlid *Astatotilapia burtonii*, one annotation (Broad Institute) found 4,110 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 "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).

The degree of difference between the annotations of ingroups and outgroups is largest for phyletic annotation, intermediate for semi-phyletic annotation, and smallest for unpatterned annotation. We expected the magnitude of the impact of annotation heterogeneity genes to scale accordingly. We used our four clades to create case studies for each pattern.

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,

Please cite this article in press as: Weisman et al., Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.085

CellPress
OPEN ACCESS

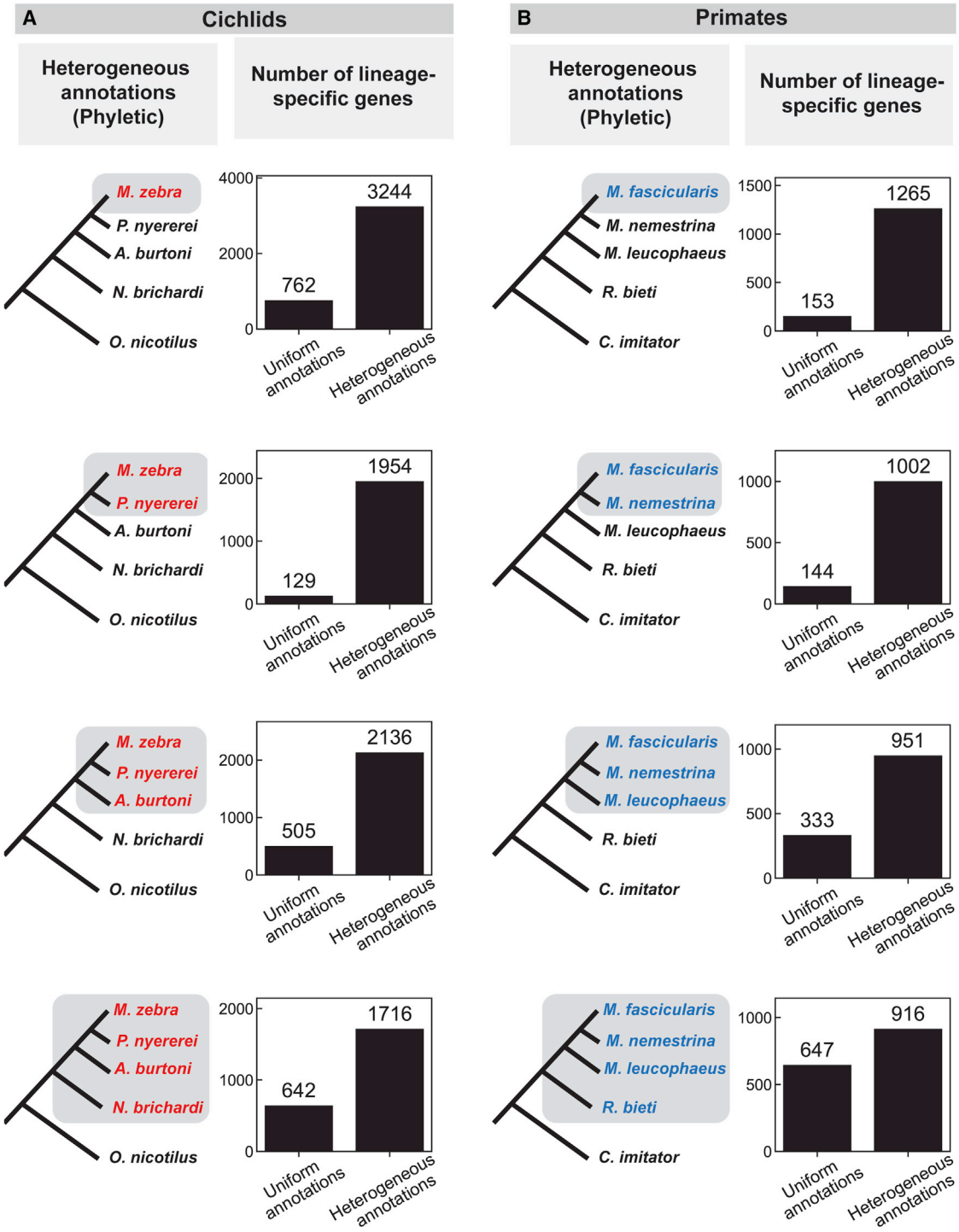

Figure 1. Effect of phyletic annotation heterogeneity

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 (gray shading); different text colors indicate different annotation sources in the heterogeneous annotation analysis (black, NCBI; red, research group at the Broad Institute; blue, Ensembl; see also Tables S3 and S4). 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. See also Table S5 for results of tBLASTx searches in this group.

Please cite this article in press as: Weisman et al., Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.085

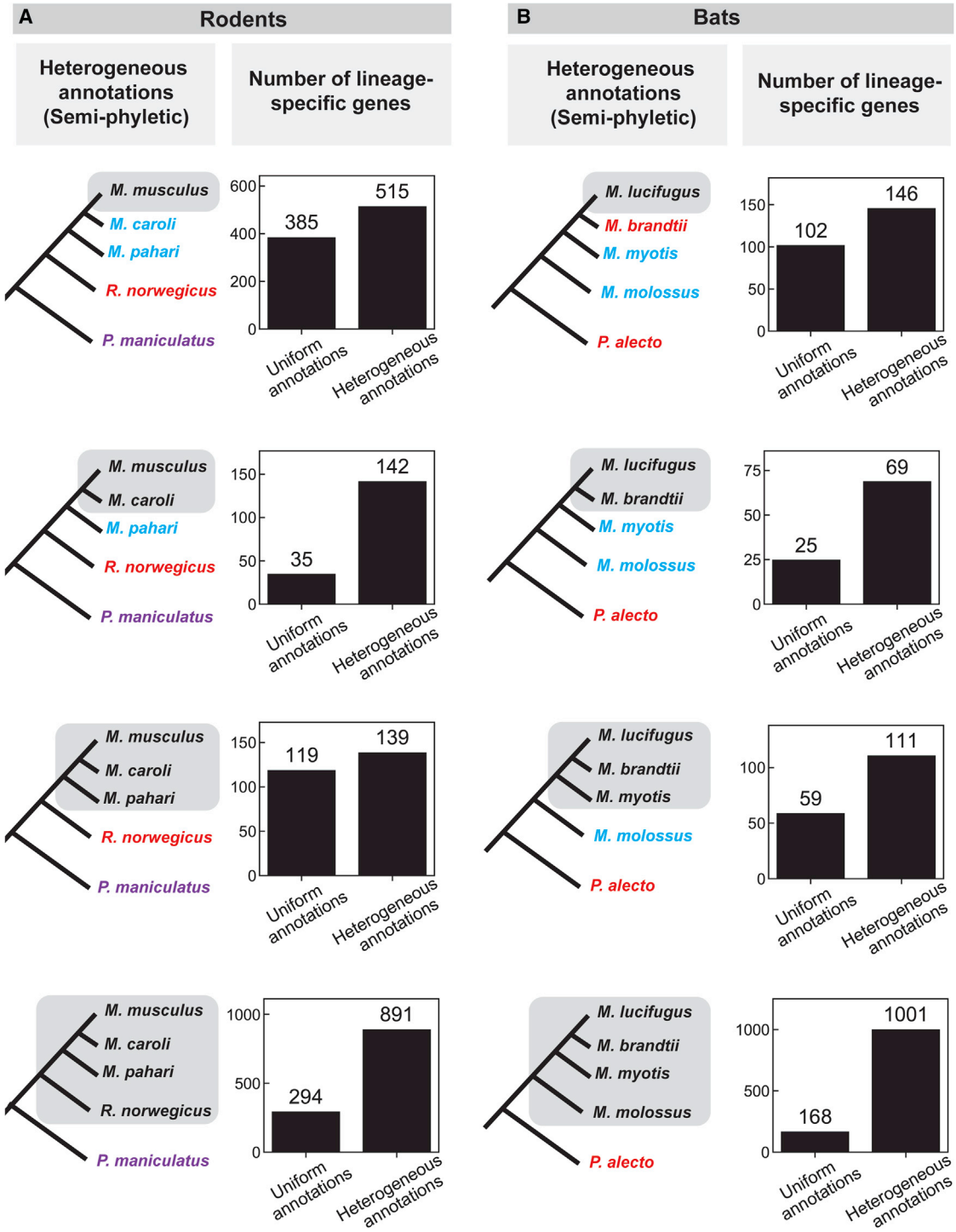

Figure 2. Effect of semi-phyletic annotation heterogeneity

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 (gray 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; see also Tables S3 and S4). 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. See also Table S5 for results of tBLASTx searches in this group.

Please cite this article in press as: Weisman et al., Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.085

CellPress
OPEN ACCESS

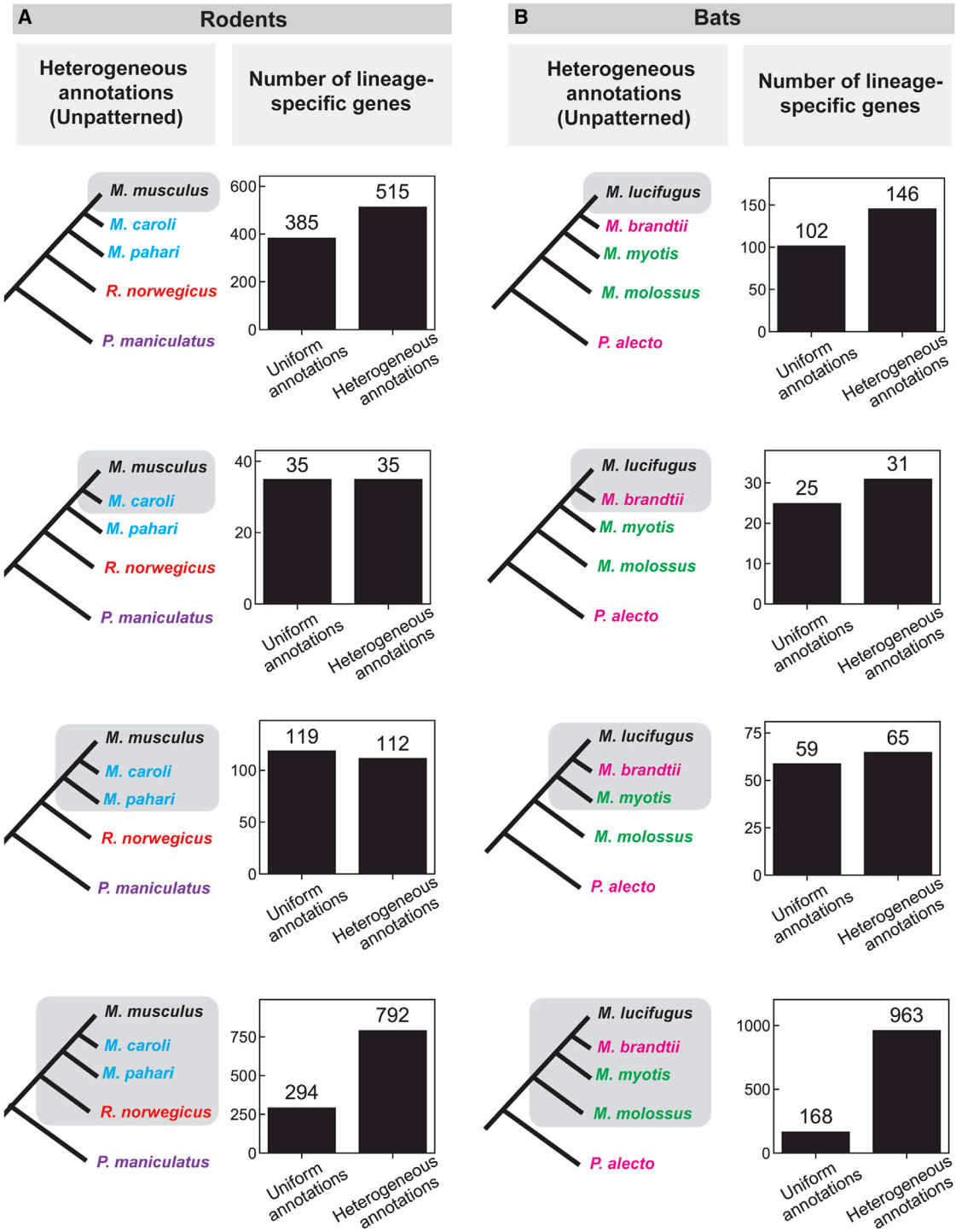

Figure 3. Effect of unpatterned annotation heterogeneity

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 (gray 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; see also Tables S3 and S4). 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. See also Table S5 for results of tBLASTx searches in this group.

Please cite this article in press as: Weisman et al., Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.085

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).

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 (Table S3).

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 (STAR 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

"Semi-phyletic" annotation, where the ingroup is annotated with one method and outgroups with a mixture of methods, was the most common type of annotation heterogeneity in our review of the published literature (Tables S1 and S2). It often occurs in scenarios similar to phyletic annotation, but where outgroup annotations come from a mixture of sources (e.g., a combination of Ensembl and NCBI, or a large heterogeneous database like NCBI's "non-redundant" sequences).

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 (Table S3). 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. It can occur in scenarios similar to semi-phyletic annotation, but when studies use existing annotations for the desired species, which often 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.

Sequence characteristics of genes affected by annotation heterogeneity

We wondered whether certain sequence characteristics of proteins may affect how likely they are to be heterogeneously annotated (included in one annotation of a genome assembly and omitted from another). We therefore classified all proteins in our four focal species as affected or unaffected by annotation heterogeneity. Affected proteins have significantly similar homologs only in one of the two annotations of at least one outgroup species; unaffected proteins either have homologs in both annotations or in neither.

We found that, in all species groups, proteins affected by annotation heterogeneity were shorter than those that were not (*M. musculus*, mean of 377 amino acids versus 691 amino acids, t test, $p = 1.5 \times 10^{-96}$; *M. zebra*, 193 versus 720, $p = 10^{-100}$; *M. fascicularis*, 136 versus 562, $p = 2.8 \times 10^{-208}$; *M. lucufugus*, 299 versus 608, $p = 2.1 \times 10^{-154}$). We also found a significant association between annotation heterogeneity and protein disorder in all groups. However, the direction of this association was inconsistent: heterogeneously annotated proteins were more disordered in two taxa (*M. musculus*, mean IUpred disorder 0.38 versus 0.33, $p = 3.2 \times 10^{-41}$; *M. lucufugus*, 0.35 versus 0.32, $p = 10^{-5}$), and less disordered in two other taxa (*M. zebra*, 0.28 versus 0.35, $p = 6 \times 10^{-154}$; *M. fasciscularis*, $p = 6 \times 10^{-154}$; 0.30 versus 0.31, $p = 10^{-5}$).

We hypothesize that the consistent length effect is due to essentially all annotation methods being more likely to consider longer ORFs as genes. Similarly, the inconsistent disorder effect may be due to different annotation programs weighting disorder, or some other characteristic with which it is correlated, differently in determining whether an ORF is a gene.

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, 2, and 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 100 (Table S5).

DISCUSSION

We used six case studies to ask if varying the annotation method across species in a comparative analysis ("annotation heterogeneity"), common in comparative analyses, 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 was used in both ingroup and outgroup species. Even within types of annotation heterogeneity, however, we find substantial variation in effect size; this likely depends on the details of the particular annotation methods involved, making it difficult to estimate the impact of annotation heterogeneity within any specific study *a priori*. Our results suggest that the numbers of lineage-specific genes found in these studies may be inflated, especially in "phyletic annotation" cases. Annotation heterogeneity may also have consequences that we do not explore here, like producing spurious lineage-specific losses.

We find evidence that sequence characteristics of proteins correlate with their tendency to be heterogeneously annotated. Length and disorder are consistent correlates of heterogeneous annotation, but the direction of the association for disorder varies across our case studies. We speculate that which correlations arise in a particular analysis depends on the particulars of the annotation methods in use. The differences in our results for disorder are reminiscent of conflicting reports of the direction of correlation between apparent gene age and disorder found in previous studies.[8,11,19–23] We speculate that annotation heterogeneity may contribute to these discrepancies, consistent with the previous finding that the direction of the correlation can be reversed when ORFs unlikely to be real genes are carefully excluded.[11]

Our case studies consist of closely related groups of five species. We did not consider larger groups of more distantly related species because we were unable to find ones satisfying our requirement that each species have two available annotations of the same genome assembly, one of which is from the same method for all members of the group. The effects of annotation heterogeneity in larger and more distantly related cases could be less pronounced.

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 independently of annotation errors.[24,25] Previous studies have noted a surprisingly large number of "young" lineage-specific genes found in recently evolved clades,[26] 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 million years 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 non-coding 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. While we do not perform such analyses here, we propose that assessing 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 constrain it.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Identifying lineage-specific proteins
  - Literature review of studies of lineage-specific genes
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Frequency of annotation heterogeneity in published literature
  - Sequence characteristics of genes affected by annotation heterogeneity

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2022.04.085.

ACKNOWLEDGMENTS

This work was primarily funded by a Howard Hughes Medical Institute investigator award to S.R.E. S.R.E. is also supported in part by the NIH (R01-HG009116). A.W.M. is supported by grants from the NIH (RO1-GM43987) and the NSF-Simons Center for the Mathematical and Statistical Analysis of Biology (NSF #1764269, Simons #594596). C.M.W. is supported in part by an NSF-Simons Quantitative Biology PhD Student Fellowship. Computations were done on the Cannon cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

AUTHOR CONTRIBUTIONS

Conceptualization, C.M.W.; formal analysis, C.M.W.; investigation, C.M.W.; writing – original draft, C.M.W.; writing – review & editing, C.M.W., A.W.M., and S.R.E.; supervision, A.W.M. and S.R.E.; funding acquisition, S.R.E.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 4, 2022
Revised: March 17, 2022
Accepted: April 21, 2022
Published: May 18, 2022

REFERENCES

1. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., and Bosch, T.C. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. *25*, 404–413. https://doi.org/10.1016/j.tig.2009.07.006.
2. Tautz, D., and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. Nat. Rev. Genet. *12*, 692–702. https://doi.org/10.1038/nrg3053.
3. Wilson, G.A., Bertrand, N., Patel, Y., Hughes, J.B., Feil, E.J., and Field, D. (2005). Orphans as taxonomically restricted and ecologically important genes. Microbiology *151*, 2499–2501. https://doi.org/10.1099/mic.0.28146-0.
4. McLysaght, A., and Guerzoni, D. (2015). New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philos. Trans. R. Soc. B Biol. Sci. *370*, 20140332. https://doi.org/10.1098/rstb.2014.0332.
5. Tautz, D. (2014). The discovery of de novo gene evolution. Perspect. Biol. Med. *57*, 149–161. https://doi.org/10.1353/pbm.2014.0006.
6. Domazet-Lošo, T., Brajković, J., and Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. *23*, 533–539. https://doi.org/10.1016/j.tig.2007.08.014.
7. McLysaght, A., and Hurst, L.D. (2016). Open questions in the study of de novo genes: what, how and why. Nat. Rev. Genet. *17*, 567–578. https://doi.org/10.1038/nrg.2016.78.
8. Basile, W., and Elofsson, A. (2017). The number of orphans in yeast and fly is drastically reduced by using combining searches in both proteomes and genomes. Preprint at bioRxiv. https://doi.org/10.1101/185983.
9. Casola, C. (2018). From de novo to "de nono": the majority of novel protein-coding genes identified with phylostratigraphy are old genes or recent duplicates. Genome Biol. Evol. *10*, 2906–2918. https://doi.org/10.1093/gbe/evy231.
10. Zile, K., Dessimoz, C., Wurm, Y., and Masel, J. (2020). Only a single taxonomically restricted gene family in the Drosophila melanogaster subgroup can be identified with high confidence. Genome Biol. Evol. *12*, 1355–1366. https://doi.org/10.1093/gbe/evaa127.
11. Wilson, B.A., Foy, S.G., Neme, R., and Masel, J. (2017). Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nat. Ecol. Evol. *1*. 0146-6. https://doi.org/10.1038/s41559-017-0146.
12. Thibaud-Nissen, F., DiCuccio, M., Hlavina, W., Kimchi, A., Kitts, P.A., Murphy, T.D., Pruitt, K.D., and Souvorov, A. (2016). P8008 the NCBI eukaryotic genome annotation pipeline. J. Anim. Sci. *94*, 184.
13. Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., et al. (2021). Ensembl 2021. Nucleic Acids Res. *49*, D884–D891. https://doi.org/10.1093/nar/gkaa942.
14. Larkin, A., Marygold, S.J., Antonazzo, G., Attrill, H., dos Santos, G., Garapati, P.V., Goodman, J.L., Gramates, L.S., Millburn, G., Strelets, V.B., et al. (2008). FlyBase: updates to the Drosophila melanogaster knowledge base. Nucleic Acids Res. *49*, D899–D907.
15. Howe, K., Davis, P., Paulini, M., Tuli, M.A., Williams, G., Yook, K., Durbin, R., Kersey, P., and Sternberg, P.W. (2012). WormBase: annotating many nematode genomes. Worm. *1*, 15–21.
16. Giraldo-Calderón, G.I., Emrich, S.J., MacCallum, R.M., Maslen, G., Dialynas, E., Topalis, P., Ho, N., Gesing, S., VectorBase Consortium, and Madey, G., et al. (2015). VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Res. *43*, D707–D713. https://doi.org/10.1093/nar/gku1117.
17. Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. *13*, 329–342. https://doi.org/10.1038/nrg3174.
18. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
19. James, J.E., Willis, S.M., Nelson, P.G., Weibel, C., Kosinski, L.J., and Masel, J. (2021). Universal and taxon-specific trends in protein sequences as a function of age. eLife *10*, e57347. https://doi.org/10.7554/elife.57347.
20. Foy, S.G., Wilson, B.A., Bertram, J., Cordes, M.H.J., and Masel, J. (2019). A shift in aggregation avoidance strategy marks a long-term direction to protein evolution. Genetics *211*, 1345–1355. https://doi.org/10.1534/genetics.118.301719.
21. Willis, S., and Masel, J. (2018). Gene birth contributes to structural disorder encoded by overlapping genes. Genetics *210*, 303–313. https://doi.org/10.1534/genetics.118.301249.
22. Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., and Santhanam, B. (2012). Proto-genes and de novo gene birth. Nature *487*, 370–374. https://doi.org/10.1038/nature11184.
23. Vakirlis, N., Hebert, A.S., Opulente, D.A., Achaz, G., Hittinger, C.T., Fischer, G., Coon, J.J., and Lafontaine, I. (2018). A molecular portrait of de novo genes in yeasts. Mol. Biol. Evol. *35*, 631–645. https://doi.org/10.1093/molbev/msx315.
24. Weisman, C.M., Murray, A.W., and Eddy, S.R. (2020). Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol. *18*, e3000862. https://doi.org/10.1371/journal.pbio.3000862.
25. Moyers, B.A., and Zhang, J. (2016). Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. Mol. Biol. Evol. *33*, 1245–1256. https://doi.org/10.1093/molbev/msw008.
26. Neme, R., and Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics *14*, 117. https://doi.org/10.1186/1471-2164-14-117.
27. Van Oss, S.B., and Carvunis, A.-R. (2019). De novo gene birth. PLoS Genet. *15*, e1008160. https://doi.org/10.1371/journal.pgen.1008160.

Please cite this article in press as: Weisman et al., Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.085

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Cichlid genome annotations	Broad Institute	ftp://ftp.broadinstitute.org/pub/vgb/cichlids/Annotation/Protein_coding/Peptide_Files/
Cichlid genome annotations	NCBI	GCF_000238955.1, GCF_000239375.1, GCF_000239415.1, GCF_000239395.1, GCF_000188235.2
Primate genome annotations	Ensembl	Release 102, macaca_fascicularis; Release 103, macaca_nemestrina, mandrillus_leucophaeus, rhinopithecus_bieti, cebus_imitator
Primate genome annotations	NCBI	GCF_000364345.1, GCF_000956065.1, GCF_000951045.1, GCF_001698545.1, GCF_001604975.1
Rodent genome annotations	Downloaded from Ensembl; generated by Ensemble and UCSC	Release 101, mus_musculus, mus_caroli, mus_pahari, rattus_norwegicus; Release 104, peromyscus_maniculatus
Rodent genome annotations	NCBI	GCF_000001635.26, GCF_900094665.1, GCF_900095145.1, GCF_000001895.5, GCF_000500345.1
Bat genome annotations	Downloaded from Ensembl and NCBI; generated by Ensembl, BAT1K consortium, Beijing Genomics Institute	GCA_000412655.1, GCA_014108235.1, GCA_014108415.1, GCA_000325575.1, Ensembl release 103, myotis_lucifugus
Bat genome annotations	NCBI	GCF_000147115.1, GCF_000412655.1, GCF_014108235.1, GCF_014108415.1, GCF_000325575.1
Software and algorithms		
Basic Local Alignment Search Tool (BLAST)	Altschul et al.[18]	Version 6.2.0

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Caroline M. Weisman (cweisman@princeton.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All genome annotations on which these analyses were based are publicly available and are listed in the key resources table and Tables S3 and S4. All results summarized in Figures 1, 2, and 3 are publicly available as of the date of publication at https://github.com/caraweisman/Annotation_homology.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Identifying lineage-specific proteins

For each species group, we defined a protein as specific to a particular lineage if a search using BLASTP[18] 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. Results from these searches are summarized in Table S5.

Literature review of studies of lineage-specific genes

We considered each of the papers cited in Table S1 of a recent review paper on *de novo* genes[27] and determined whether or not it used heterogeneously annotated genomes. A list of these papers, whether it uses heterogeneous annotations, and, if so, relevant details about the heterogeneous annotations are shown in Table S1.

We performed a literature search for articles about lineage-specific genes published after 2019. We considered all articles returned by a Google Scholar search for publications between 2019 and present whose titles included the phrases "lineage-specific gene(s)," "orphan gene(s)," "taxonomically-restricted gene(s)," "de novo gene(s)," or "proto-gene(s)." We manually excluded results clearly on unrelated topics (e.g., "lineage-specific genes" in the context of development, where the "lineage" is a cell type and its progenitors). Some of these studies used lineage-specific genes identified in previous studies; in these cases, we considered whether the original study used heterogeneous annotations. A list of these papers, whether it uses heterogeneous annotations, and, if so, relevant details about the heterogeneous annotations are shown in Table S2.

QUANTIFICATION AND STATISTICAL ANALYSIS

Frequency of annotation heterogeneity in published literature

To compare the rates of annotation heterogeneity among the two sets of papers on the subject of lineage-specific genes that we considered, one from pre-2020 and one from post-2020, we used a Fisher's exact test, with N being the total number of papers in the two groups, as described in the text (introduction) and as listed in Tables S1 and S2. No data were excluded.

Sequence characteristics of genes affected by annotation heterogeneity

To assess the statistical association between genes' tendency to be affected by annotation heterogeneity and gene length, GC content, and disorder, we used t tests to compare the distribution of the lengths of genes affected and not affected by annotation heterogeneity. N was the total number of proteins in the annotation of the focal species, given in Table S4, with the results for particular genes available on the Github as described above in data and code availability. No data were excluded.

Current Biology, Volume *32*

Supplemental Information

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

Caroline M. Weisman, Andrew W. Murray, and Sean R. Eddy

	Bats	Cichlids	Rodents	Primates
Orphans	5	7	36	43
Lineage 1	13	2	24	1
Lineage 2	56	6	78	11
Lineage 3	106	18	286	176

Table S5: Results of six-frame translation homology searches. Related to Figures 1-3.
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.

Supplemental References

S1. Wissler, L., Gadau, J., Simola, D.F., Helmkampf, M., and Bornberg-Bauer, E. (2013). Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome biology and evolution *5*, 439-455.

S2. Li, Z.-W., Chen, X., Wu, Q., Hagmann, J., Han, T.-S., Zou, Y.-P., Ge, S., and Guo, Y.-L. (2016). On the origin of de novo genes in Arabidopsis thaliana populations. Genome biology and evolution *8*, 2190-2202.

S3. Sun, W., Zhao, X.-W., and Zhang, Z. (2015). Identification and evolution of the orphan genes in the domestic silkworm, Bombyx mori. FEBS letters *589*, 2731-2738.

S4. Donoghue, M.T., Keshavaiah, C., Swamidatta, S.H., and Spillane, C. (2011). Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. BMC evolutionary biology *11*, 1-23.

S5. Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., and Wang, W. (2008). On the origin of new genes in Drosophila. Genome research *18*, 1446-1455.

S6. Chen, S., Zhang, Y.E., and Long, M. (2010). New genes in Drosophila quickly become essential. science *330*, 1682-1685.

S7. Zhao, L., Saelao, P., Jones, C.D., and Begun, D.J. (2014). Origin and spread of de novo genes in Drosophila melanogaster populations. Science *343*, 769-772.

S8. Heames, B., Schmitz, J., and Bornberg-Bauer, E. (2020). A continuum of evolving de novo genes drives protein-coding novelty in Drosophila. Journal of molecular evolution *88*, 382-398.

S9. Wu, D.-D., Irwin, D.M., and Zhang, Y.-P. (2011). De novo origin of human protein-coding genes. PLoS genetics *7*, e1002379.

S10. Knowles, D.G., and McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. Genome research *19*, 1752-1759.

S11. Dowling, D., Schmitz, J.F., and Bornberg-Bauer, E. (2020). Stochastic gain and loss of novel transcribed open reading frames in the human lineage. Genome biology and evolution *12*, 2183-2195.

S12. Vakirlis, N., Hebert, A.S., Opulente, D.A., Achaz, G., Hittinger, C.T., Fischer, G., Coon, J.J., and Lafontaine, I. (2018). A molecular portrait of de novo genes in yeasts. Molecular Biology and Evolution *35*, 631-645.

S13. Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A.R., Yu, Y., Hou, G., Zi, J., and Zhou, R. (2019). Rapid evolution of protein diversity by de novo origination in Oryza. Nature ecology & evolution *3*, 679-690.

S14. Neme, R., and Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novoevolution. BMC genomics *14*, 1-13.

S15. Schmitz, J.F., Ullrich, K.K., and Bornberg-Bauer, E. (2018). Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. Nature ecology & evolution *2*, 1626-1632.

S16. Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., and Mar Alba, M. (2009). Origin of primate orphan genes: a comparative genomics approach. Molecular biology and evolution *26*, 603-612.

S17. Prabh, N., and Rödelsperger, C. (2019). De novo, divergence, and mixed origin contribute to the emergence of orphan genes in Pristionchus nematodes. G3: Genes, Genomes, Genetics *9*, 2277-2286.
S18. Wilson, B.A., Foy, S.G., Neme, R., and Masel, J. (2017). Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nature ecology & evolution *1*, 1-6.
S19. Ekman, D., and Elofsson, A. (2010). Identifying and quantifying orphan protein sequences in fungi. Journal of Molecular Biology *396*, 396-405.
S20. Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., and Santhanam, B. (2012). Proto-genes and de novo gene birth. Nature *487*, 370.
S21. Wilson, B.A., and Masel, J. (2011). Putatively noncoding transcripts show extensive association with ribosomes. Genome biology and evolution *3*, 1245-1252.
S22. Durand, É., Gagnon-Arsenault, I., Hallin, J., Hatin, I., Dubé, A.K., Nielly-Thibault, L., Namy, O., and Landry, C.R. (2019). Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. Genome research *29*, 932-943.
S23. Lu, T.-C., Leu, J.-Y., and Lin, W.-C. (2017). A comprehensive analysis of transcript-supported de novo genes in Saccharomyces sensu stricto yeasts. Molecular biology and evolution *34*, 2823-2838.
S24. Qi, M., Zheng, W., Zhao, X., Hohenstein, J.D., Kandel, Y., O'Conner, S., Wang, Y., Du, C., Nettleton, D., and MacIntosh, G.C. (2019). QQS orphan gene and its interactor NF-YC 4 reduce susceptibility to pathogens and pests. Plant biotechnology journal *17*, 252-263.
S25. Dossa, K., Zhou, R., Li, D., Liu, A., Qin, L., Mmadi, M.A., Su, R., Zhang, Y., Wang, J., and Gao, Y. (2021). A novel motif in the 5'-UTR of an orphan gene 'Big Root Biomass' modulates root biomass in sesame. Plant biotechnology journal *19*, 1065-1079.
S26. O'Conner, S., and Li, L. (2020). Mitochondrial fostering: the mitochondrial genome may play a role in plant orphan gene evolution. Frontiers in plant science *11*, 1855.
S27. Witt, E., Benjamin, S., Svetec, N., and Zhao, L. (2019). Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in Drosophila. eLife *8*, e47138.
S28. Wang, C., Chen, S., Feng, A., Su, J., Wang, W., Feng, J., Chen, B., Zhang, M., Yang, J., and Zeng, L. (2021). Xa7, a small orphan gene harboring promoter trap for AvrXa7, leads to the durable resistance to Xanthomonas oryzae pv. oryzae. Rice *14*, 1-16.
S29. Lange, A., Patel, P.H., Heames, B., Damry, A.M., Saenger, T., Jackson, C.J., Findlay, G.D., and Bornberg-Bauer, E. (2021). Structural and functional characterization of a putative de novo gene in Drosophila. Nature communications *12*, 1-13.
S30. Yates, T.B., Feng, K., Zhang, J., Singan, V., Jawdy, S.S., Ranjan, P., Abraham, P.E., Barry, K., Lipzen, A., and Pan, C. (2021). The ancient Salicoid genome duplication event: A platform for reconstruction of de Novo gene evolution in Populus trichocarpa. Genome biology and evolution *13*, evab198.
S31. Wang, Y.-W., Hess, J., Slot, J.C., and Pringle, A. (2020). 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 *12*, 2168-2182.

S32. Zhuang, X., and Cheng, C.-H.C. (2021). Propagation of a De Novo Gene under Natural Selection: Antifreeze Glycoprotein Genes and Their Evolutionary History in Codfishes. Genes *12*, 1777.
S33. Yoshioka, Y., Suzuki, G., Zayasu, Y., Yamashita, H., and Shinzato, C. (2021). Comparative Genomics Highlight the Importance of Lineage-Specific Gene Families in Evolutionary Divergence of the Coral Genus, Montipora.
S34. Zelhof, A.C., Mahato, S., Liang, X., Rylee, J., Bergh, E., Feder, L.E., Larsen, M.E., Britt, S.G., and Friedrich, M. (2020). The brachyceran de novo gene PIP82, a phosphorylation target of aPKC, is essential for proper formation and maintenance of the rhabdomeric photoreceptor apical domain in Drosophila. PLoS genetics *16*, e1008890.
S35. Xie, C., Bekpen, C., Künzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K.K., and Tautz, D. (2019). Studying the dawn of de novo gene emergence in mice reveals fast integration of new genes into functional networks.
S36. Lee, B.Y., Kim, J., and Lee, J. (2021). Intraspecific de novo gene birth revealed by presence absence variant genes in Caenorhabditis elegans. bioRxiv.
S37. Delihas, N. (2022). An ancestral genomic sequence that serves as a nucleation site for de novo gene birth. bioRxiv.
S38. Xie, C., Bekpen, C., Künzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K.K., and Tautz, D. (2019). A de novo evolved gene in the house mouse regulates female pregnancy cycles. eLife *8*, e44392.
S39. Rivard, E.L., Ludwig, A.G., Patel, P.H., Grandchamp, A., Arnold, S.E., Berger, A., Scott, E.M., Kelly, B.J., Mascha, G.C., and Bornberg-Bauer, E. (2021). A putative de novo evolved gene required for spermatid chromatin condensation in Drosophila melanogaster. PLoS genetics *17*, e1009787.
S40. McMenamin, A.J., Brutscher, L.M., Daughenbaugh, K.F., and Flenniken, M.L. (2021). The Honey Bee Gene Bee Antiviral Protein-1 Is a Taxonomically Restricted Antiviral Immune Gene. Frontiers in Insect Science, 11.
S41. Ma, D., Ding, Q., Guo, Z., Zhao, Z., Wei, L., Li, Y., Song, S., and Zheng, H.-L. (2021). Identification, characterization and expression analysis of lineage-specific genes within mangrove species Aegiceras corniculatum. Molecular Genetics and Genomics *296*, 1235-1247.
S42. Reinhardt, D., Roux, C., Corradi, N., and Di Pietro, A. (2021). Lineage-specific genes and cryptic sex: parallels and differences between arbuscular mycorrhizal fungi and fungal pathogens. Trends in Plant Science *26*, 111-123.
S43. Jiang, M., Zhan, Z., Li, H., Dong, X., Cheng, F., and Piao, Z. (2020). Brassica rapa orphan genes largely affect soluble sugar metabolism. Horticulture research *7*.
S44. Gori, A., Harrison, O.B., Mlia, E., Nishihara, Y., Chan, J.M., Msefula, J., Mallewa, M., Dube, Q., Swarthout, T.D., and Nobbs, A.H. (2020). Pan-GWAS of Streptococcus agalactiae highlights lineage-specific genes associated with virulence and niche adaptation. MBio *11*, e00728-00720.
S45. Entwistle, S., Li, X., and Yin, Y. (2019). Orphan genes shared by pathogenic genomes are more associated with bacterial pathogenicity. Msystems *4*, e00290-00218.
S46. Jin, G.H., Zhou, Y.L., Yang, H., Hu, Y.T., Shi, Y., Li, L., Siddique, A.N., Liu, C.N., Zhu, A.D., and Zhang, C.J. (2021). Genetic innovations: Transposable element recruitment and de novo formation lead to the birth of orphan genes in the rice genome. Journal of Systematics and Evolution *59*, 341-351.

S47. Li, G., Wu, X., Hu, Y., Muñoz-Amatriaín, M., Luo, J., Zhou, W., Wang, B., Wang, Y., Wu, X., and Huang, L. (2019). Orphan genes are involved in drought adaptations and ecoclimatic-oriented selections in domesticated cowpea. Journal of experimental botany *70*, 3101-3110.
S48. Luna, S.K., and Chain, F.J. (2021). Lineage-Specific Genes and Family Expansions in Dictyostelid Genomes Display Expression Bias and Evolutionary Diversification during Development. Genes *12*, 1628.
S49. Cridland, J.M., Majane, A.C., Zhao, L., and Begun, D.J. (2022). Population biology of accessory gland-expressed de novo genes in Drosophila melanogaster. Genetics *220*, iyab207.
S50. Li, J., Arendsee, Z., Singh, U., and Wurtele, E.S. (2019). Recycling RNA-seq data to identify candidate orphan genes for experimental analysis. BioRxiv, 671263.
S51. LI, T.-p., ZHANG, L.-w., LI, Y.-q., YOU, M.-s., and Qian, Z. (2021). Functional analysis of the orphan genes Tssor-3 and Tssor-4 in male Plutella xylostella. Journal of Integrative Agriculture *20*, 1880-1888.
S52. Ma, S., Yuan, Y., Tao, Y., Jia, H., and Ma, Z. (2020). Identification, characterization and expression analysis of lineage-specific genes within Triticeae. Genomics *112*, 1343-1350.
S53. Gao, Q., Yan, H., Xia, E., Zhang, S., and Li, S. (2019). TOGD: a database of orphan genes in Triticum aestivum. International Journal of Agriculture and Biology *22*, 961-966.
S54. Zhao, Z., and Ma, D. (2021). Genome-Wide Identification, Characterization and Function Analysis of Lineage-Specific Genes in the Tea Plant Camellia sinensis. Frontiers in Genetics *12*, 770570-770570.
S55. Warner, M.R., Qiu, L., Holmes, M.J., Mikheyev, A.S., and Linksvayer, T.A. (2019). Convergent eusocial evolution is based on a shared reproductive groundplan plus lineage-specific plastic genes. Nature communications *10*, 1-11.
S56. Brennan, C.J., Zhou, B., Benbow, H.R., Ajaz, S., Karki, S.J., Hehir, J.G., O'Driscoll, A., Feechan, A., Mullins, E., and Doohan, F.M. (2020). Taxonomically restricted wheat genes interact with small secreted fungal proteins and enhance resistance to Septoria tritici blotch disease. Frontiers in plant science *11*, 433.
S57. Chen, K., Tian, Z., Chen, P., He, H., Jiang, F., and Long, C.-a. (2020). Genome-wide identification, characterization and expression analysis of lineage-specific genes within Hanseniaspora yeasts. FEMS Microbiology Letters *367*, fnaa077.