TOGA integrates gene annotation with orthology inference at scale

Bogdan M. Kirilenko[1,2,3,4,5,6], Chetan Munegowda[1,2,3], Ekaterina Osipova[1,2,3,4,5,6], David Jebb[1,2,3], Virag Sharma[1,2,3], Moritz Blumer[1,2,3], Ariadna Morales[4,5,6], Alexis-Walid Ahmed[4,5,6], Dimitrios-Georgios Kontopoulos[4,5,6], Leon Hilgers[4,5,6], Zoonomia Consortium[7], and Michael Hiller[1,2,3,4,5,6*]

[1] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
[2] Max Planck Institute for the Physics of Complex Systems, Dresden, Germany
[3] Center for Systems Biology Dresden, Germany
[4] LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt, Germany
[5] Senckenberg Research Institute, Senckenberganlage 25, 60325 Frankfurt, Germany
[6] Goethe-University, Faculty of Biosciences, Max-von-Laue-Str. 9, 60438 Frankfurt, Germany
[7] Zoonomia Consortium Members are listed at the end of the document

*To whom correspondence should be addressed:
Michael Hiller
LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt, Germany
Tel: +49 69 7542-1398
E-Mail: Michael.Hiller@senckenberg.de

Running title: TOGA annotates orthologous genes at scale

Keywords: comparative genomics, orthology inference, gene annotation, genome alignment, gene loss, machine learning

1 sentence Summary: A scalable gene annotation approach using a novel paradigm to detect orthologous loci provides comparative data for hundreds of mammals and birds.

Abstract

Annotating coding genes and inferring orthologs are two classical challenges in genomics and evolutionary biology that have traditionally been approached separately, which limits scalability. We present TOGA, the first method that integrates gene annotation and orthology inference. TOGA implements a novel paradigm to infer orthologous genes, improves ortholog detection and annotation completeness compared to state-of-the-art methods, and handles even highly-fragmented assemblies. TOGA scales to hundreds of genomes, which we demonstrate by applying it to 488 placental mammal and 308 bird assemblies, creating the largest comparative gene resources so far. Additionally, TOGA detects gene losses, enables selection screens, and automatically provides a superior measure of mammalian genome quality. Together, TOGA is a powerful and scalable method to annotate and compare genes in the genomic era.

Introduction

Distinguishing homologs -- genes with a common ancestry -- into orthologs and paralogs is a fundamental problem in evolutionary and molecular biology. Orthology and paralogy are defined for a pair of homologous genes that originated by either speciation (ortholog) or gene duplication (paralog) (*1*). Inferring orthologous genes is a prerequisite for many genomic analyses, including reconstructing phylogenetic trees from molecular data, predicting gene function, investigating molecular and genome evolution, and discovering differences in genes that underlie phenotypes of the sequenced species (*2-7*).

Current methods for orthology inference are either based on graph or tree approaches or a combination of both (*8*). Graph-based methods cluster genes into pairs or groups of orthologs based on pairwise sequence similarity such as (reciprocal) best alignment hits (*9-18*). Tree-based methods determine whether the evolutionary lineages of a pair of genes coalesce in a speciation or a duplication node in the gene tree (*19-26*). Importantly, the input for these approaches is a set of annotated genes with their coding or protein sequences for each to-be-considered species. This is why gene identification and annotation until now has preceded orthology inference, resulting in two limitations. First, gene annotation quality has a large influence on the accuracy of orthology inference (*27*). Second, since generating a high-quality annotation is time-consuming and typically requires comprehensive transcriptomics (gene expression) data, there is a growing gap between genome sequencing and genome annotation including orthology inference.

Here, we present TOGA (Tool to infer Orthologs from Genome Alignments), a new method that provides several key innovations. First, TOGA uses a new paradigm to accurately infer orthologous genes that largely relies on alignments of intronic and intergenic sequences instead

of alignments of only coding sequences. Second, TOGA is the first method that integrates orthology detection with comparative gene annotation, making it applicable to un-annotated genome assemblies. Third, TOGA explicitly investigates whether orthologs likely encode an intact protein, have missing exonic sequence, or have gene-inactivating mutations (e.g. frameshifts or premature stop codons), which is important for distinguishing functional from inactivated orthologous genes. We show that TOGA accurately detects orthologs and generates comprehensive gene annotations at a quality similar to, or better, than state-of-the-art methods. TOGA's ability to join fragments of orthologous genes facilitates the use of less contiguous assemblies in comparative gene analyses. We also show that TOGA provides a superior benchmark for mammalian genome quality. Finally, we demonstrate that TOGA scales to the hundreds of already sequenced genomes by annotating genes and inferring orthologs for 488 placental mammals and 308 birds, creating the largest comparative gene datasets for both groups.

Results

A novel paradigm for orthology detection

The principle used implicitly or explicitly by all orthology detection methods is that orthologous sequences are generally more similar to each other than to paralogous sequences (*1*). Existing methods focus on similarity between coding sequences that typically evolve under purifying selection. However, this principle also extends to non-exonic regions (introns, intergenic regions) that largely evolve neutrally. The key innovation implemented in TOGA is that intronic and flanking intergenic regions of orthologous gene loci will also be more similar to each other (Figure 1A), provided that the evolutionary distance between the species is sufficiently short such that neutrally evolving regions still partially align. This is given for placental mammals that shared a common ancestor up to ~100 Mya (*28*), since the evolutionary distance between human and other placental mammals is at most 0.55 substitutions per neutral site (Figure S1, Tables S1, S2). Similarly, the evolutionary distance between chicken and other birds that shared a common ancestor up to ~100 Mya (*29*) is at most 0.51 substitutions per neutral site (Table S2). This explains why orthologous introns and intergenic regions retain enough sequence similarity that they partially align between species within these clades (Figures 1A, S2). In contrast, the evolutionary distance between paralogs that duplicated before the speciation event is often much larger and exceeds 1 substitution per neutral site (Figure 1C). At such distances, introns and intergenic regions of paralogous genes have been largely randomized, and alignments can only be detected for coding sequences that generally evolve slowly due to purifying selection (Figures 1A, S2). TOGA exploits this principle by (i) taking a well annotated genome such as human, mouse or chicken as the reference, (ii) inferring all (co-)orthologous loci for all genes from a genome alignment between the reference and a query species (e.g. other placental mammals or birds), and (iii) annotating and classifying these genes (Figure 1B).

The TOGA annotation and orthology detection pipeline

TOGA implements a multi-step pipeline, comprising the detection of orthologous loci, annotation and classification, and orthology type determination. In the first step, TOGA uses machine learning to distinguish orthologous from paralogous genomic loci or loci containing processed pseudogenes, largely relying on alignments of intronic and intergenic regions around the gene of interest. To this end, TOGA uses a whole genome alignment between an annotated reference species and an aligned query species, exemplified by human and mouse in Figure 1A. A powerful method to compute and visualize a pairwise genome alignment are chains of co-linear local alignments that capture both orthologous as well as paralogous genes or processed pseudogene loci (*30*). To distinguish between them, TOGA computes for each gene and each overlapping chain four characteristic features that capture the amount of intronic and intergenic alignments (Figure S3). Additionally, TOGA uses synteny (conserved gene order) as another feature, which can help to distinguish orthologs from paralogs (*24, 31-33*).

We trained a machine learning classifier using known orthologous genes between human (reference) and mouse (query) from Ensembl Compara (*24*) (Figure S4). We then tested the classifier on several independent query species (rat, dog, armadillo) from different placental mammalian orders. We obtained a near perfect orthologous chain classification for both multi- and single-exon genes (Figure 1D, Table S3). The features capturing intronic/intergenic alignments are most important for the classification performance (Figures 1E,F). In contrast, synteny is the least important feature, likely reflecting our training data sets that we deliberately enriched with translocated orthologs (Figures S5). Using synteny as an auxiliary but not determining feature enables TOGA to also accurately detect orthologs that underwent rearrangements such as translocations or inversions and therefore lack conserved gene order (Figure 1D), as exemplified in Figure S6.

For the human-rat test dataset, we manually investigated discrepancies between TOGA's classifications and Ensembl. We found that chains classified as false positives mostly represent partial or full gene duplications in rat (Figure S7), indicating that TOGA is able to detect lineage-specific gene duplications and actually correctly classified these chains as co-orthologous loci. A limitation of our approach is exemplified by the 12 false negative chain classifications in the test set. These exhibited both exceptional intron divergence and lacked intergenic alignments due to rearrangements, resulting in alignment chains that resemble paralogous loci (Figure S8). Interestingly, 7 of the 12 false negatives are X-chromosome linked genes, indicating that faster X chromosome evolution (*34*) could be involved in the exceptional divergence of neutrally evolving regions of these loci. It should be noted that TOGA still annotated these genes, but labeled them as putative paralogs (Figure S8).

In a second step, TOGA uses CESAR 2.0 (*35, 36*) to determine the positions and boundaries of all coding exons for each (co-)orthologous query locus of the gene (Figures 1B, S9, S10). Since

orthology between genomic loci, as determined in the first step, does not imply that the gene encodes a functional protein, TOGA subsequently assesses for each transcript and each orthologous locus whether it preserves the intact reading frame (Figure S11). To this end, TOGA identifies gene-inactivating mutations (frameshifting, stop codon or splice site mutations, exon or gene deletions) by implementing an improved version of our gene loss detection approach (*6*) (Figures 1B, S12-S17). We only classified a gene as lost, if all transcripts at all (co-)orthologous loci are classified as lost. We benchmarked this approach on a large set of 11,161 conserved genes that are annotated as 1:1 orthologs by Ensembl in mouse, rat, cow and dog. Only 21, 22, 12 and 21 genes are misclassified as inactivated for the four species, indicating a very high specificity of 99.80 to 99.89% (Table S4). Manual inspection showed that the few mis-classified cases include highly-diverged genes, genes that evolved drastic changes in exon-intron structure or protein length, and a lost gene that is compensated by a processed pseudogene copy, which highlights cases of less certain gene conservation (Figures S18-S22).

An interesting example demonstrating the importance of detecting all orthologous loci and determining reading frame intactness is the *STRC* and *CKMT1B* gene locus. This locus was duplicated four times in the lineage leading to guinea pig, and TOGA recognizes all co-orthologous loci with high probabilities (Figure 1G). However, despite the quadruplication, only one copy of each gene encodes an intact reading frame. In case of *STRC*, the gene encoded by the ancestral locus became inactivated, while one of the new copies maintained an intact reading frame. TOGA correctly classifies and annotates both genes as 1:1 orthologs, but also annotates exons of the remnants of the otherwise inactivated gene copies in the guinea pig genome (Figure 1G).

In the third step, TOGA determines the orthology type by considering all reference genes and all orthologous query loci that encode an intact reading frame (Figure 1B, S23). Finally, TOGA uses an orthology graph approach to resolve weakly-supported orthology relationships among many:many orthologs (Figures 1B, S24).

TOGA improves ortholog detection

To assess the performance of TOGA's orthology detection pipeline, we compared it against Ensembl Compara, which integrates graph- and tree-based methods and provides high-quality ortholog gene sets (*24*). Using orthologs between human and three representative mammals (rat, cow, elephant), we found that TOGA detected 97.6%, 98.9% and 96.5% of the orthologs provided by Ensembl (Figure 2A, Table S5), showing a good agreement. Furthermore, for >90% of these commonly-detected orthologs, TOGA inferred the same orthology type as Ensembl (1:1, 1:many, many:1, many:many) (Figure 2C). A quarter of the discrepancies are cases where TOGA infers 1:1 and Ensembl 1:many. In several of these cases, Ensembl annotates a processed pseudogene copy as a second ortholog (Figure S25).

For the orthologs detected only by Ensembl, TOGA did identify an orthologous locus in >93% of the cases, but detected either inactivating mutations indicating gene loss or that large parts of the gene overlap assembly gaps (classified as a missing gene) (Figures 2D, S26, S27). Consistent with these cases including more questionable orthologs, parameters measuring alignment identity (mean 51%), alignment coverage (mean 44%) and orthology confidence (mean 32%) are substantially lower compared to orthologs detected by both methods (means 81%, 94%, 91%) (Figure 2B).

TOGA predicted for the three species 1,532 (rat), 1,711 (cow) and 2,174 (elephant) additional orthologs that are not listed in Ensembl (Figure 2A). For rat, this includes *PAX1*, an important developmental transcription factor that was potentially missed by Ensembl because of a mis-annotated N-terminus (Figure S28). About half of these genes belong to large families such as zinc finger genes, olfactory receptors or keratin associated proteins (Figure 2C). While establishing orthology between genes in large families is generally more challenging, these genes exhibit alignment identity (mean 70%), alignment coverage (mean 83%) and orthology confidence (mean 94%) values that are more similar to the orthologs detected by both methods (means 82%, 94%, 99%) (Figure 2B), supporting that these genes are undetected orthologs.

TOGA improves gene annotation completeness

To assess the completeness of annotations generated by TOGA, we performed a direct comparison to annotations generated by Ensembl and by the NCBI Eukaryotic Genome Annotation Pipeline (*37, 38*), two state-of-the-art methods that integrate transcriptomics, homology-based data (transcripts and proteins from RefSeq and GenBank) and *ab initio* gene predictions. To this end, we applied TOGA using human as the reference to genomes of 70 / 118 placental mammals that have Ensembl / NCBI annotations. Using BUSCO (Benchmarking Universal Single-Copy Orthologs), a widely used tool to assess the completeness of protein-coding gene annotations (*39*), we surprisingly found that TOGA annotations have a higher completeness score for the mammalian BUSCO odb10 gene set for 97% (Ensembl) and 91.5% (NCBI) of the species (Figure 3A, B, Tables S6, S7). On average, TOGA's annotations have a 4.1% (Ensembl) and 0.7% (NCBI) higher completeness, which corresponds to ~377 and ~64 genes in the set of 9,226 BUSCO genes.

To show that this performance is not specific to the use of human as the reference, we compared Ensembl and NCBI to TOGA annotations obtained by using mouse (mm10 assembly) as the reference. Like human, mouse also provides a high-quality gene annotation, which is important for reference-based methods like TOGA. Using mouse, we found that TOGA annotations have a higher BUSCO completeness for 98.5% (Ensembl) and 64% (NCBI) of the species (Figure 3A, B, Tables S6, S7). While reference-based methods cannot annotate orthologs of genes not contained in the reference annotation, this downside can be counteracted by combining multiple references. Indeed, combining the human- and mouse-based TOGA annotations achieves a higher

completeness for almost all (>98%) of the assemblies with an average increase of 4.5% (Ensembl) and 0.97% (NCBI) (Figure 3A, B). These tests show that the BUSCO gene completeness of TOGA's comparative annotations are often higher than those produced by state-of-the-art annotation pipelines that include transcriptomics data.

TOGA improves annotation completeness even if transcriptomics data are available

Transcriptomics data is undoubtedly very useful for gene annotation, as it provides direct evidence of transcripts expressed in the sampled tissues. Therefore, we next tested whether TOGA can increase annotation completeness, even if transcriptomics data and other gene evidence are already available. To this end, we used six high-quality bat genomes (*7*) and first annotated genes by integrating available transcriptomics data (both RNA-seq and Iso-seq), *ab initio* gene predictions (Augustus (*40*)), aligned proteins from closely related bats, and comparative gene predictions (Augustus-CGP applied to a multiple genome alignment (*41*)). For the six bats, these annotations contained 87.7% to 95.4% of the genes in the mammalian BUSCO odb10 set (Figure 3C, Table S8). Adding TOGA with human as the reference as an additional evidence consistently increased the annotation completeness by 3.9% to 11.4%, reaching a BUSCO completeness score of 98.8% to 99.3%. This shows that even if a comprehensive set of gene evidence including transcriptomics data are available, annotation completeness can still be improved by TOGA.

TOGA joins split genes in fragmented assemblies

Genes that are split between different scaffolds are currently either missed or annotated as fragments, hampering downstream analyses. Although current genome projects aim to generate highly-complete, chromosome-level assemblies (*7, 42*), even such assemblies can contain a few fragmented genes (Figure S29). Furthermore, many currently available mammal or bird assemblies exhibit fragmentation (*43, 44*) and are therefore more difficult to annotate. To improve comparative annotation and orthology inference of fragmented genes, we leveraged TOGA's ability to detect orthologous loci of partial genes. We implemented a gene joining procedure that recognizes orthologous parts of 1:1 orthologous genes, joins them together, and generates an annotation and codon/protein alignments for the full gene. Figure 4A illustrates this procedure for a gene split into six parts in the fragmented pygmy sperm whale assembly (*43*).

To evaluate the accuracy of this step, we utilized the fact that orthologous but not paralogous genes from two closely related species are expected to be highly similar. We used assemblies of two sperm whale species: *Kogia breviceps* with a low scaffold N50 of 29 kb (*43*) and *Physeter macrocephalus* with a high scaffold N50 of 122 Mb (*45*). Orthologous genes, for which no joining is necessary as they are contained on a single scaffold in both species, have a high median coding exon identity of 98.28% (mean 98.70%) (Figure 4B), which serves as a positive control. In contrast, paralogous genes, which we used as a negative control, have a lower median coding exon identity of 77% (mean 75.18%). Consequently, if TOGA's gene joining procedure was misidentifying paralogous as orthologous fragments, we would expect a decreased nucleotide identity compared

to orthologs located on a single scaffold. However, we observed an equally high identity for orthologous genes joined from two, three and at least four fragments (Figures 4B, S30), indicating a high accuracy.

Demonstrating the effectiveness of this fragment joining procedure, the median length of the coding sequence of split *Kogia* genes after joining orthologous fragments is 100% (mean 97%) of the length of the orthologous human gene. This is a substantial improvement over the largest orthologous fragment in the assembly (median 58%, mean 59%) (Figure 4C, Table S9). We obtained similar improvements for other highly-fragmented assemblies. Even for an assembly of the extinct Steller's sea cow with a scaffold N50 value of 1.4 kb (*46*), TOGA improved relative coding sequence length from 28% to 70% (Figure 4C, Table S9).

TOGA scales to hundreds of genomes

Given the wealth of genomes that are generated in the current era, there is a strong need for annotation and orthology inference methods that are able to handle hundreds or thousands of genomes. Unlike previous graph- or tree-based methods that often perform all-against-all comparisons that scale quadratically with the number of species, TOGA considers a pair of reference-query species and thus scales linearly with the number of query species. To demonstrate this, we applied TOGA with human as the reference to a large set of placental mammals, comprising 488 different assemblies of 427 distinct species (Figure 5A, Table S1). As expected, with an average of 19,144 (median 19,192) genes, TOGA annotates more genes in the six Hominoidea (apes) species that are closely related to human. Importantly, for the remaining 482 assemblies, TOGA also annotated on average 17,779 (median 18,049) genes, indicating that TOGA is an effective annotation method across placental mammals.

Fitting generalized linear models shows that the number of annotated orthologs is influenced by several factors. These include assembly quality metrics (contig and scaffold N50), which are both positively correlated with the number of detected orthologs, and the evolutionary distance (substitutions per neutral site) and divergence time (millions of years) to human, which are both negatively correlated (Figure S31, Table S10). Evolutionary distance has a stronger influence than divergence time. This is exemplified for Perissodactyla, where TOGA consistently annotates more genes than in many rodents, despite the fact that the rodent lineage split from human more recently.

To explore the influence of the reference genome, we next applied TOGA to the same 488 placental mammal assemblies using mouse as the reference (Figure 5B, Table S1). Corroborating the influence of evolutionary distance and divergence time, TOGA annotated more genes for the 20 closely related Muridae assemblies (mean 20,597, median 20,918) than for the remaining 466 assemblies (mean 17,852, median 18,115). Overall, the number of annotated genes is similar to the human-based annotations.

TOGA provides a superior approach for assessing mammalian assembly quality

In addition to annotation and orthology inference, TOGA's gene classification also provides a powerful benchmark to measure assembly completeness and quality. To this end, we first compiled a comprehensive set of 18,430 ancestral placental mammal genes, defined as human coding genes that have an intact reading frame in the basal placental clades Afrotheria and Xenarthra (Table S11). For each of the 488 placental mammal assemblies, we then used TOGA's gene classification to determine which percent of these ancestral genes have an intact reading frame without missing sequence. We found that this completeness measure is significantly correlated with the completeness value computed by BUSCO (Pearson r = 0.73, $P=10^{-81}$) (Figure 6A). However, BUSCO's values saturate at ~97% for highly complete assemblies, while TOGA's completeness values exhibit a larger dynamic range (Figure 6B), which is important to distinguish highly- from less-contiguous assemblies. This is exemplified by two closely related bats: a high-quality assembly of *Rhinolophus ferrumequinum* (contig N50 21.7 Mb) and a less-contiguous assembly of *R. sinicus* (contig N50 38 kb) that have a very similar BUSCO completeness (96.4% vs. 96.3% complete genes) but are separated markedly by TOGA's completeness value (94.4 vs. 88.2%) (Figure 6C).

BUSCO's fragmented or missing gene classification indicates how much of the gene was detected, but does not distinguish between the two major underlying reasons: assembly gaps that result in missing gene sequence vs. assembly base errors that destroy the reading frame. TOGA's gene classification explicitly distinguishes between these two different assembly issues, which provides valuable information on assembly quality. For example, TOGA detects a substantially higher percentage of genes exhibiting inactivating mutations in the *Bos gaurus* (gaur, 14.2%) compared to the *Bos taurus* (cow, 4.3%) assembly, indicating that the *B. gaurus* assembly has an elevated base error rate, whereas both assemblies are indistinguishable in terms of BUSCO completeness scores (95.8 vs. 95.5%) (Figure 6D). Similarly, TOGA shows that the dog canFam5 assembly exhibits an elevated base error rate compared to dog canFam4 or the dingo, whereas all three assemblies have highly similar BUSCO scores (Figure 6E). An informative example illustrating that assemblies can suffer from different issues are two assemblies of the spotted hyena: the NCBI GCA_008692635.1 assembly has less missing sequence, but a noticeably higher base error rate compared to the DNAzoo assembly of the same species (Figure 6E). Finally, illustrating extreme cases among seal assemblies, TOGA reveals that 56% of the genes in the Antarctic fur seal have inactivating mutations and that 31% of the genes in the Weddell seal have missing exonic sequence (Figure 6F).

In summary, TOGA automatically provides a measure for mammalian genome completeness with two advantages. High sensitivity provides the resolution to detect smaller differences in gene completeness of high-quality assemblies and the ability to distinguish between assembly incompleteness and base error rate provides insight into these two distinct assembly challenges.

TOGA facilitates more accurate codon alignments

Codon or protein alignments are important to screen for selection patterns and to reconstruct phylogenetic trees, but alignment error can substantially impact the outcome (*28, 47*). TOGA implements two features that help to avoid errors when aligning coding sequences. First, TOGA masks all gene-inactivating mutations such as frameshifts, which can otherwise result in misalignments (Figure S32). Second, whereas existing methods consider entire orthologous coding or protein sequences, TOGA is aware of orthology at the exon level. This enables a new "exon by exon" procedure that generates codon or protein alignments by first aligning orthologous exons and then joining exon alignments together with potential split codons at exon boundaries. Figure S33 illustrates that this procedure avoids alignment errors in case of insertions or deletions that occurred at exon boundaries.

Applying TOGA to 308 bird as well as other non-mammalian genomes

To further demonstrate TOGA's ability to scale to many genomes, we used chicken (galGal6) as the reference and applied TOGA to a large set of bird genomes generated by the B10K project and many individual laboratories (*29, 44, 48*). The set comprises 308 different assemblies of 298 distinct species. Across all assemblies, TOGA annotated on average 13,994 (median: 14,058) orthologous genes (Figure 5C, Table S12).

We also explored whether TOGA can be applied to species other than mammals and birds. Our tests with turtles, fish, and sea urchins provide encouraging results (Figure 5D) that may be further improved by adjusting the method to these clades.

Comprehensive resources for comparative genomics

For the 488 placental mammal and 308 bird assemblies, we provide comparative gene annotations, ortholog sets, lists of inactivated genes and multiple codon alignments generated with MACSE v2 (*49*) for download at http://genome.senckenberg.de/download/TOGA/. To our knowledge, these comprise the largest comparative genomics datasets for both clades so far. To facilitate visualizing and analyzing these data, we further implemented a TOGA annotation track as part of the UCSC genome browser (*50*) (Figure S34). Our UCSC browser mirror at https://genome.senckenberg.de/ provides these annotation tracks for all analyzed mammal and bird assemblies.

Discussion

TOGA is an integrative pipeline that jointly addresses two fundamental problems in genomics and evolutionary biology: gene annotation and orthology inference. We show that alignments between non-coding sequences in introns and intergenic regions enable an accurate detection of orthologous gene loci, establishing a novel paradigm for inference of orthologous genes. Comparisons with state-of-the-art methods show that TOGA often improves gene annotation completeness, even if transcriptomics data are available. Here, TOGA benefits from great efforts that generated high-quality annotations for human and mouse (*38, 51*), and provides an approach to effectively utilize these to annotate other placental mammals. Furthermore, by joining split genes in fragmented assemblies, TOGA increases the utility of such genomes for comparative analyses. In addition to generating annotations, TOGA detects inactivated genes and provides orthologous sequences for codon alignments. These enable phylogenomic analyses as well as screens for selection patterns and gene losses that are linked to relevant phenotypes, as previously demonstrated in the Bat1K and other projects (*7, 52-54*). TOGA's gene annotations and classifications can also be used to assess assembly quality, featuring an increased sensitivity and the ability to distinguish assembly incompleteness from assembly base errors, which are both important as more and more highly complete and accurate assemblies are being produced (*7, 42, 55, 56*). Finally, TOGA's reference-based methodology scales linearly, handling hundreds and -- when available in the clades of interest -- even thousands of genomes.

TOGA's application range comprises species with "alignable" genomes, which we define in our context as genomes where orthologous neutrally evolving regions partially align. In general, this holds for evolutionary distances of ~0.6 substitutions per neutral site, which from a human or mouse point of view includes other placental mammals. At larger evolutionary distances, neutrally evolving intronic and intergenic regions are too diverged to be of use for TOGA's orthologous locus detection approach (Figure S35A). Interestingly, applying TOGA with human as the reference to 18 marsupial and two monotreme species reveals that TOGA is still able to annotate on average 13,096 orthologs (Figure 5A,B), largely because gene order is often conserved (Figure S35B). Nevertheless, human is obviously not a powerful reference for these more distant clades. Instead, a marsupial and a monotreme species should be used as the reference to annotate genes and infer orthologs in these clades.

With the tree of life becoming more densely populated with genomes thanks to great efforts of large-scale projects (*42-44, 57*), TOGA provides a general strategy to cope with the annotation and orthology inference bottleneck. For every "alignable" clade of interest, one can select one (or a few) species to be used as the reference for others in the clade. The resulting annotations can be enriched with transcriptomics data of the query species (when available) to detect novel lineage-specific genes or novel splice variants that are expressed in the sampled tissues. Genome and annotation of the reference(s) should ideally be highly complete, since the quality of the input

impacts the quality of the output. References can be defined for different taxonomic ranks, from the class to the family or genus level. For example, in the Bat1K project (*58*), we aim at generating a high-quality assembly and gene annotation for representatives of all bat families to serve as references for dozens or hundreds of other bats in these families.

Data and code availability

The TOGA source code, and all scripts to run TOGA, create training and test data sets and browser tracks are available at https://github.com/hillerlab/TOGA. TOGA is also available in a singularity container environment. All data generated in this manuscript are available for download at http://genome.senckenberg.de/download/TOGA/ and for browsing in our UCSC genome browser mirror at https://genome.senckenberg.de.

Competing interests

The authors have no competing interests.

Acknowledgment

We thank the genomics community for sequencing and assembling the genomes and the UCSC genome browser group for providing software, and Ensembl for genome annotations. We also thank Ingo Ebersberger and Kerstin Lindblad-Toh for helpful comments on the manuscript, Franziska Friedrich for help with the TOGA logo, and the Computer Service Facilities of the MPI-CBG and MPI-PKS and Christoph Sinai for their excellent technical support. This work was supported by the Max Planck Society and the LOEWE-Centre for Translational Biodiversity Genomics (TBG) funded by the Hessen State Ministry of Higher Education, Research and the Arts (HMWK).

Supplementary Materials

Materials and Methods
Tables S1 – S13
Figures S1 – S37

References

1. T. Gabaldon, E. V. Koonin, Functional and evolutionary implications of gene orthology. *Nature reviews. Genetics* **14**, 360-366 (2013).
2. P. Kapli, Z. Yang, M. J. Telford, Phylogenetic tree building in the genomic age. *Nature reviews. Genetics* **21**, 428-444 (2020).
3. A. Meyer *et al.*, Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature* **590**, 284-289 (2021).
4. A. M. Altenhoff, R. A. Studer, M. Robinson-Rechavi, C. Dessimoz, Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS computational biology* **8**, e1002514 (2012).
5. J. Huerta-Cepas *et al.*, Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular biology and evolution* **34**, 2115-2122 (2017).
6. V. Sharma *et al.*, A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nature communications* **9**, 1215 (2018).
7. D. Jebb *et al.*, Six reference-quality genomes reveal evolution of bat adaptations. *Nature* **583**, 578-584 (2020).
8. A. M. Altenhoff, N. M. Glover, C. Dessimoz, Inferring Orthology and Paralogy. *Methods in molecular biology* **1910**, 149-175 (2019).
9. M. Remm, C. E. Storm, E. L. Sonnhammer, Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of molecular biology* **314**, 1041-1052 (2001).
10. L. Li, C. J. Stoeckert, Jr., D. S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189 (2003).
11. C. Dessimoz *et al.* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2005), pp. 61-72.
12. A. C. Roth, G. H. Gonnet, C. Dessimoz, Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9**, 518 (2008).
13. C. M. Train, N. M. Glover, G. H. Gonnet, A. M. Altenhoff, C. Dessimoz, Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. *Bioinformatics* **33**, i75-i82 (2017).
14. E. V. Kriventseva, N. Rahman, O. Espinosa, E. M. Zdobnov, OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res* **36**, D271-275 (2008).
15. E. M. Zdobnov *et al.*, OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* **45**, D744-D749 (2017).
16. L. J. Jensen *et al.*, eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* **36**, D250-254 (2008).
17. B. Linard, J. D. Thompson, O. Poch, O. Lecompte, OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics* **12**, 11 (2011).
18. D. M. Emms, S. Kelly, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**, 157 (2015).
19. C. M. Zmasek, S. R. Eddy, A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* **17**, 821-828 (2001).
20. H. Li *et al.*, TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* **34**, D572-580 (2006).

21. J. Huerta-Cepas, H. Dopazo, J. Dopazo, T. Gabaldon, The human phylome. *Genome Biol* **8**, R109 (2007).
22. J. Huerta-Cepas, S. Capella-Gutierrez, L. P. Pryszcz, M. Marcet-Houben, T. Gabaldon, PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res* **42**, D897-902 (2014).
23. R. T. van der Heijden, B. Snel, V. van Noort, M. A. Huynen, Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* **8**, 83 (2007).
24. A. J. Vilella *et al.*, EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**, 327-335 (2009).
25. J. Herrero *et al.*, Ensembl comparative genomics resources. *Database : the journal of biological databases and curation* **2016**, (2016).
26. D. M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).
27. K. Trachana *et al.*, Orthology prediction methods: a quality assessment using curated protein families. *Bioessays* **33**, 769-780 (2011).
28. W. J. Murphy, N. M. Foley, K. R. Bredemeyer, J. Gatesy, M. S. Springer, Phylogenomics and the Genetic Architecture of the Placental Mammal Radiation. *Annu Rev Anim Biosci*, (2020).
29. E. D. Jarvis *et al.*, Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320-1331 (2014).
30. W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, D. Haussler, Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 11484-11489 (2003).
31. J. Lehmann, P. F. Stadler, S. J. Prohaska, SynBlast: assisting the analysis of conserved synteny information. *BMC Bioinformatics* **9**, 351 (2008).
32. J. Jun, Mandoiu, II, C. E. Nelson, Identification of mammalian orthologs using local synteny. *BMC Genomics* **10**, 630 (2009).
33. S. Jahangiri-Tazehkand, L. Wong, C. Eslahchi, OrthoGNC: A Software for Accurate Identification of Orthologs Based on Gene Neighborhood Conservation. *Genomics Proteomics Bioinformatics* **15**, 361-370 (2017).
34. R. P. Meisel, T. Connallon, The faster-X effect: integrating theory and data. *Trends in Genetics* **29**, 537-544 (2013).
35. V. Sharma, P. Schwede, M. Hiller, CESAR 2.0 substantially improves speed and accuracy of comparative gene annotation. *Bioinformatics* **33**, 3985-3987 (2017).
36. V. Sharma, A. Elghafari, M. Hiller, Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. *Nucleic Acids Res* **44**, e103 (2016).
37. F. Thibaud-Nissen *et al.*, P8008 The NCBI Eukaryotic Genome Annotation Pipeline. *Journal of Animal Science* **94**, 184-184 (2016).
38. N. A. O'Leary *et al.*, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745 (2016).
39. M. Manni, M. R. Berkeley, M. Seppey, F. A. Simao, E. M. Zdobnov, BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular biology and evolution* **38**, 4647-4654 (2021).

40. M. Stanke, O. Schoffmann, B. Morgenstern, S. Waack, Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
41. S. Konig, L. W. Romoth, L. Gerischer, M. Stanke, Simultaneous gene finding in multiple genomes. *Bioinformatics* **32**, 3388-3395 (2016).
42. A. Rhie *et al.*, Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737-746 (2021).
43. C. Zoonomia, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240-245 (2020).
44. S. Feng *et al.*, Dense sampling of bird diversity increases power of comparative genomics. *Nature* **587**, 252-257 (2020).
45. G. Fan *et al.*, The first chromosome-level genome for a marine mammal as a resource to study ecology and evolution. *Mol Ecol Resour* **19**, 944-956 (2019).
46. F. S. Sharko *et al.*, Steller's sea cow genome suggests this species began going extinct before the arrival of Paleolithic humans. *Nature communications* **12**, 2215 (2021).
47. G. Jordan, N. Goldman, The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular biology and evolution* **29**, 1125-1139 (2012).
48. T. B. Sackton *et al.*, Convergent regulatory evolution and loss of flight in paleognathous birds. *Science* **364**, 74-78 (2019).
49. V. Ranwez, E. J. P. Douzery, C. Cambon, N. Chantret, F. Delsuc, MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Molecular biology and evolution* **35**, 2582-2584 (2018).
50. B. T. Lee *et al.*, The UCSC Genome Browser database: 2022 update. *Nucleic Acids Res*, (2021).
51. A. Frankish *et al.*, Gencode 2021. *Nucleic Acids Res*, (2020).
52. J. Damas *et al.*, Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 22311-22322 (2020).
53. D. E. Gordon *et al.*, Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science* **370**, (2020).
54. M. Blumer *et al.*, Gene losses in the common vampire bat illuminate molecular adaptations to blood feeding. *bioRxiv*, 2021.2010.2018.462363 (2021).
55. K. H. Miga *et al.*, Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79-84 (2020).
56. A. M. M. Cartney *et al.*, Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *bioRxiv*, 2021.2007.2002.450803 (2021).
57. H. A. Lewin *et al.*, Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 4325-4333 (2018).
58. E. Teeling *et al.*, Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for all Living Bat Species. *Annu Rev Anim Biosci*, (2017).

Figures

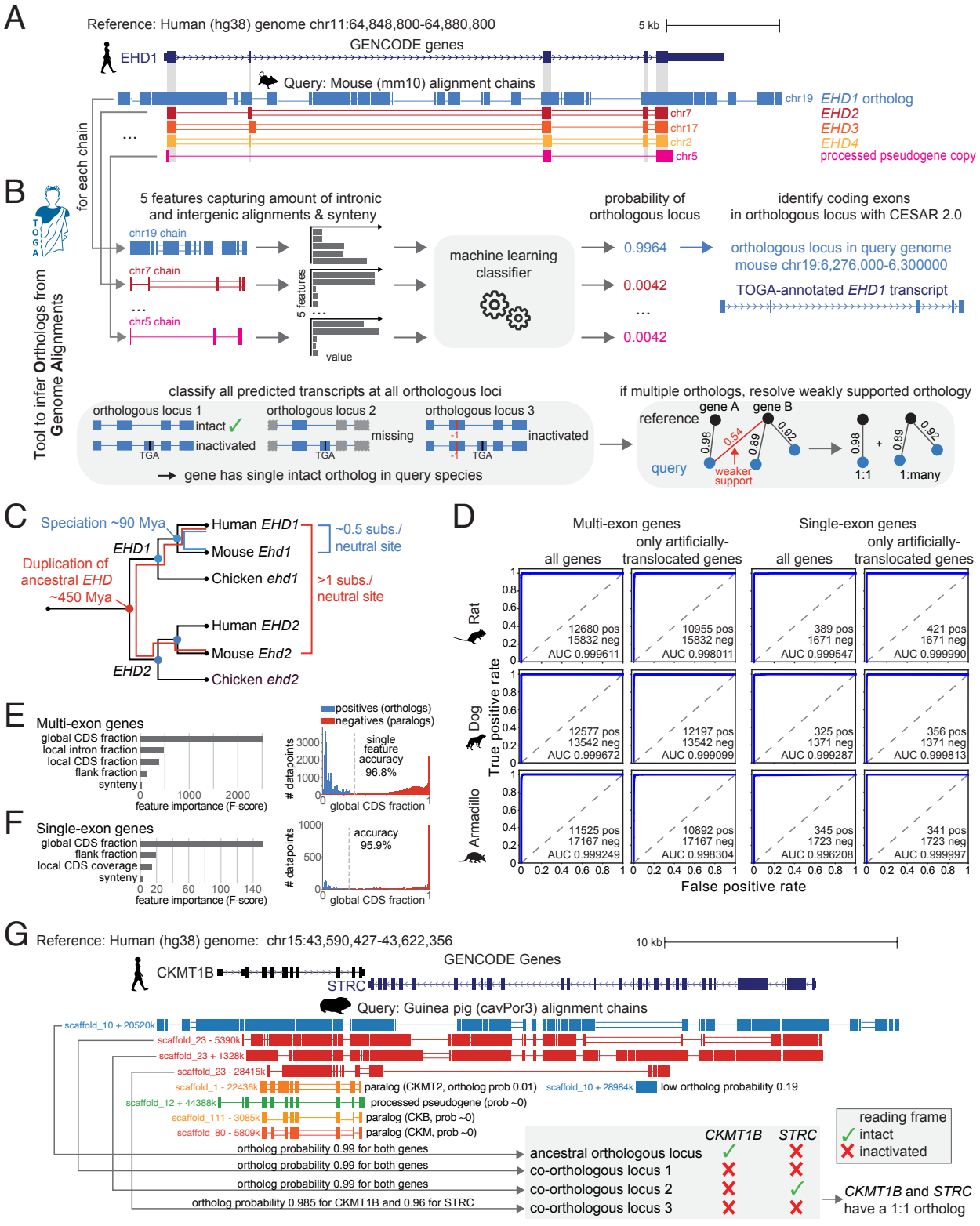

Figure 1: TOGA utilizes a novel methodology to detect orthologous genes.

(A) Illustration of TOGA and the key principle that orthologous genes have intronic and intergenic alignments. A UCSC genome browser view of the human *EHD1* gene locus shows five alignment chains (boxes represent local alignments that occur in a co-linear order, single lines represent deletions and double lines unaligning sequence) to the query species mouse, indicating that five mouse loci (chr19, 7, 17, 2, 5) have sequence similarity to coding exons of *EHD1*. The chr19 locus that encodes the mouse *Ehd1* ortholog aligns both exons and parts of introns and flanking intergenic regions, whereas the other loci that encode paralogs or processed pseudogene copies align only coding exons. Alignments for other placental mammals show similar properties (Figure S2).

(B) Illustration of TOGA. For each gene of interest and each alignment chain, we compute characteristic alignment features and use machine learning to obtain a probability that the chain alignment represents an orthologous locus. For each orthologous locus in the query, coding exons are inferred for every reference transcript of this gene. TOGA then determines for each transcript at each orthologous locus whether it encodes an intact reading frame, taking assembly incompleteness and real inactivating mutations into account. Finally, for many:many orthologs, an orthology graph is used to resolve potential weak orthology connections.

(C) The principle exploited in TOGA. The difference in the number of substitutions separating aligning orthologous and paralogous loci explains the characteristic difference that only orthologous loci show partial intronic and intergenic alignments.

(D) Orthology detection performance. Receiver Operating Characteristics curves show the true positive rate for a given false positive rate in blue. Dashed lines indicate a random classifier. The areas under these curves are close to 1 for three independent test species (rat, dog, armadillo), indicating a very high accuracy in distinguishing orthologous from non-orthologous loci. This holds both for single- and multi-exon genes as well as for genes that lack synteny due to artificial translocations that we introduced.

(E/F) Importance of the features used by TOGA to detect orthologous multi-exon (E) and single-exon (F) genes (left side). The distribution of the single most important feature (global CDS fraction, which measures the proportion of coding exon alignments in all aligning blocks of a chain) shows a clear difference between orthologous and non-orthologous chains (blue and red) for the human-rat comparison (right side). Indeed, this feature alone has high predictive power, resulting in a classification accuracy of >95%.

(G) Importance of detecting all orthologous loci and determining reading frame intactness. UCSC genome browser view shows the human genomic locus comprising *STRC* and *CKMT1B*, which is quadruplicated in the guinea pig (top four alignment chains). TOGA correctly recognizes the four co-orthologous loci with a high probability (>0.96) and distinguishes them from non-orthologous alignment chains representing paralogs and a processed pseudogene copy (probabilities <0.01). Analyzing reading frame intactness of both genes, TOGA finds that only one of the four co-orthologous loci encodes an intact reading frame (green checkmark), and correctly infers a 1:1 orthology relationship.

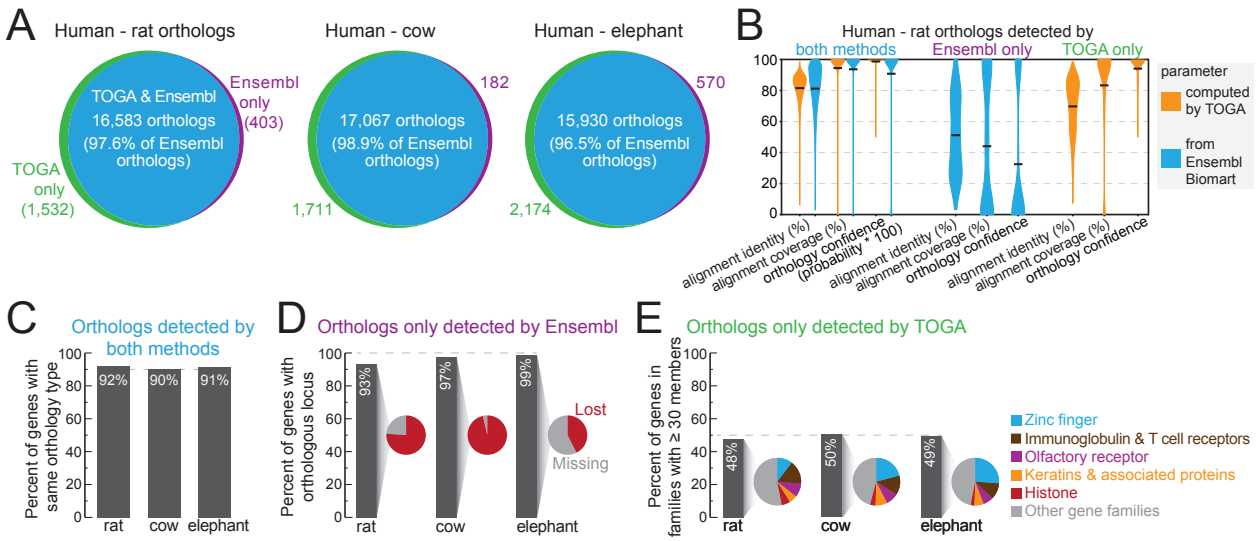

Figure 2: TOGA improves ortholog detection.

(A) Overlap of orthologs provided by Ensembl Compara and detected by TOGA for three representative placental mammals.

(B) Violin plots compare the identity and coverage of the coding region alignment and the orthology confidence probability for human-rat orthologs, detected by both or either Ensembl and TOGA. Horizontal black lines represent the mean. Note that for orthologs only detected by TOGA, these features are not available on Ensembl Biomart, and vice versa.

(C) Percent of orthologs detected by both methods, for which Ensembl and TOGA infer the same orthology type (1:1, 1:many, many:1, many:many).

(D) Percent of orthologs only detected by Ensembl, for which TOGA detects an orthologous locus (bar chart) but classifies the gene as lost (inactivated reading frame) or missing (more than half of the coding region overlaps assembly gaps), as shown by the pie charts.

(E) Percent of orthologs only detected by TOGA that belong to gene families with at least 30 members (bar chart). Pie charts show the proportion of the most frequent gene families.

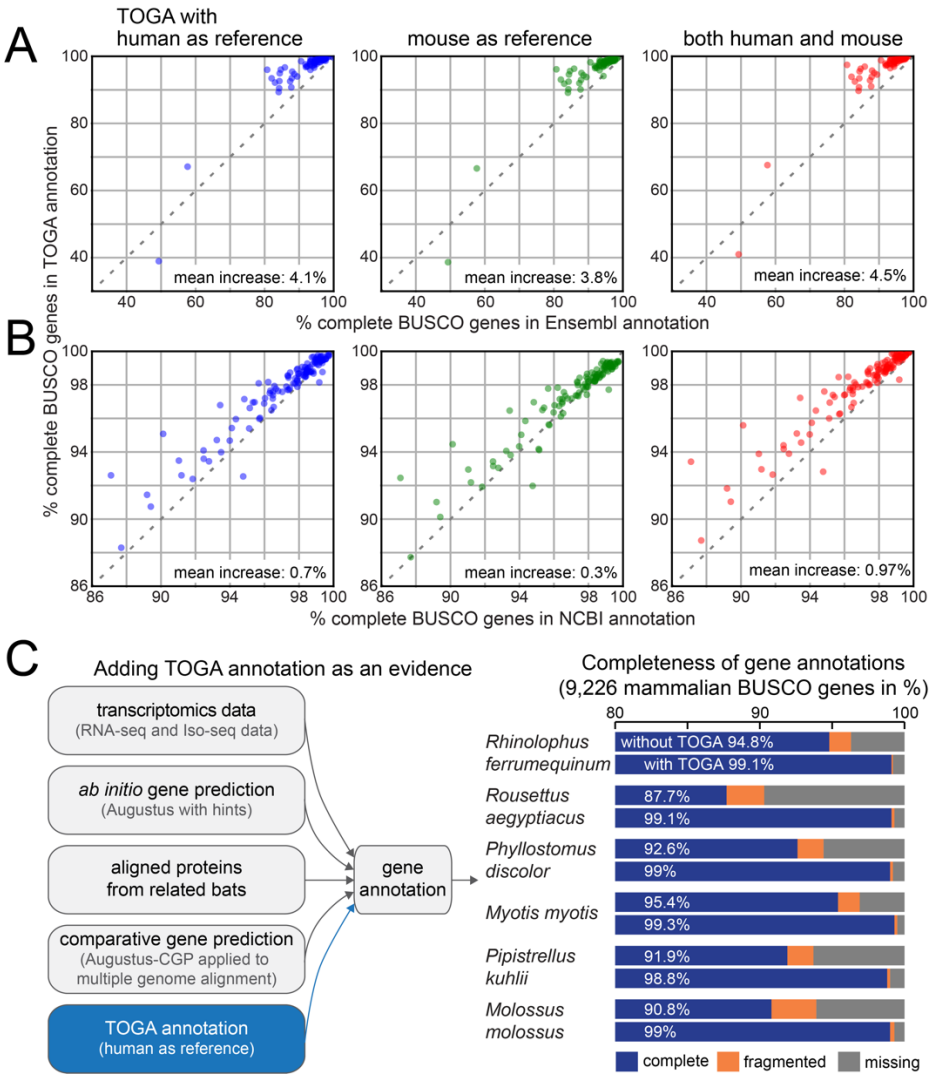

Figure 3: TOGA improves gene annotation completeness.

(A,B) Comparison of the completeness of annotations generated by TOGA and Ensembl (panel A, 70 placental mammals) and the NCBI Eukaryotic Genome Annotation Pipeline (panel B, 118 placental mammals). For most species, TOGA annotations have a higher annotation completeness according to the percent of completely detected mammalian BUSCO genes. Note that the set of species in A and B overlaps but is not identical.

(C) List of gene evidence that was integrated to annotate six bat species, once without TOGA and once with TOGA. Bar charts compare annotation completeness as a percentage of detected mammalian BUSCO genes. Adding TOGA as evidence increases annotation completeness by 3.9% to 11.4%.

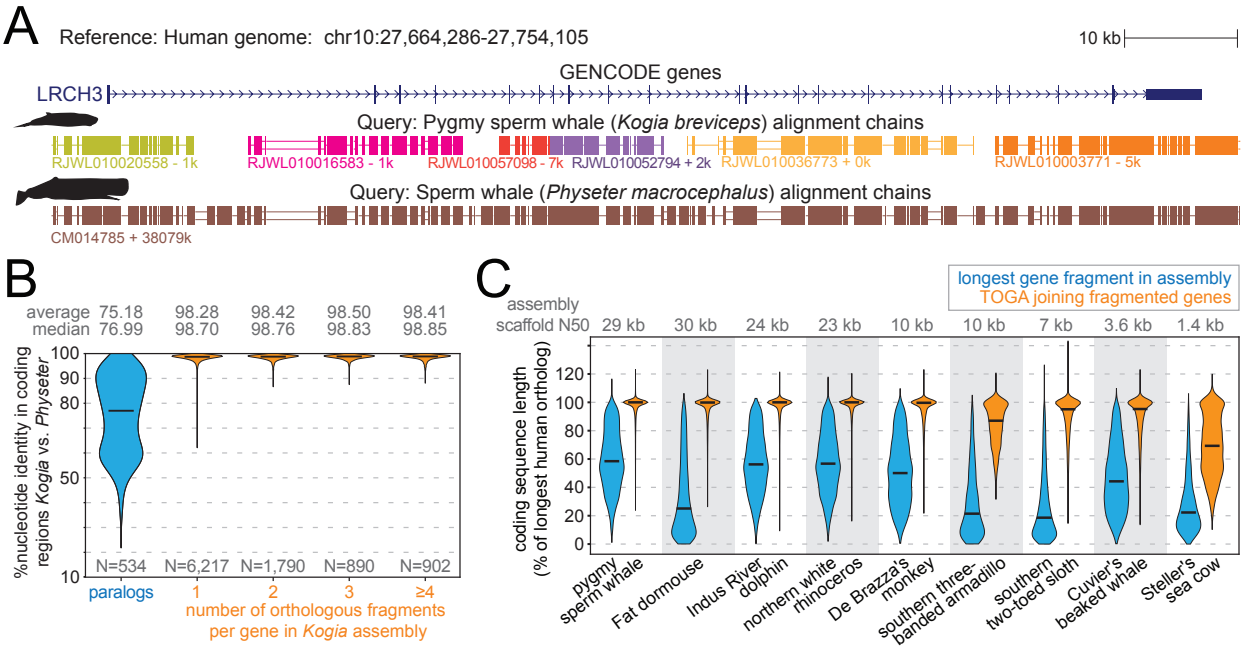

Figure 4: TOGA detects and joins genes split in fragmented genome assemblies.

(A) The ortholog of human *LRCH3* is split into six parts in the fragmented pygmy sperm whale (*Kogia breviceps*) assembly that comprises 1.2 million scaffolds (*43*). Different chain colors represent different scaffolds. Despite some chains aligning as little as one or two coding exons, TOGA correctly detects and joins all six orthologous chains to obtain the complete gene. For comparison, *LRCH3* is located on a single scaffold (thus on a single chain) in the closely related sperm whale (*Physeter macrocephalus*), which highlights the highly-similar alignment block structure.

(B) Violin plots show the coding exon identity between *Kogia breviceps* and *Physeter macrocephalus*. Horizontal black lines represent the median. Supporting the high accuracy of TOGA's fragmented gene joining procedure, genes that are present as two or more fragments in the *Kogia* assembly have a highly-similar identity distribution compared to genes for which no joining was necessary as they are already present on a single scaffold.

(C) Effectiveness of joining fragmented genes. Violin plots show the length of the coding sequence for the largest genomic fragment of split genes (blue) and after joining orthologous fragments (orange). Length is relative to the longest transcript of the orthologous human gene. In case of codon insertions, the relative length can be >100%.

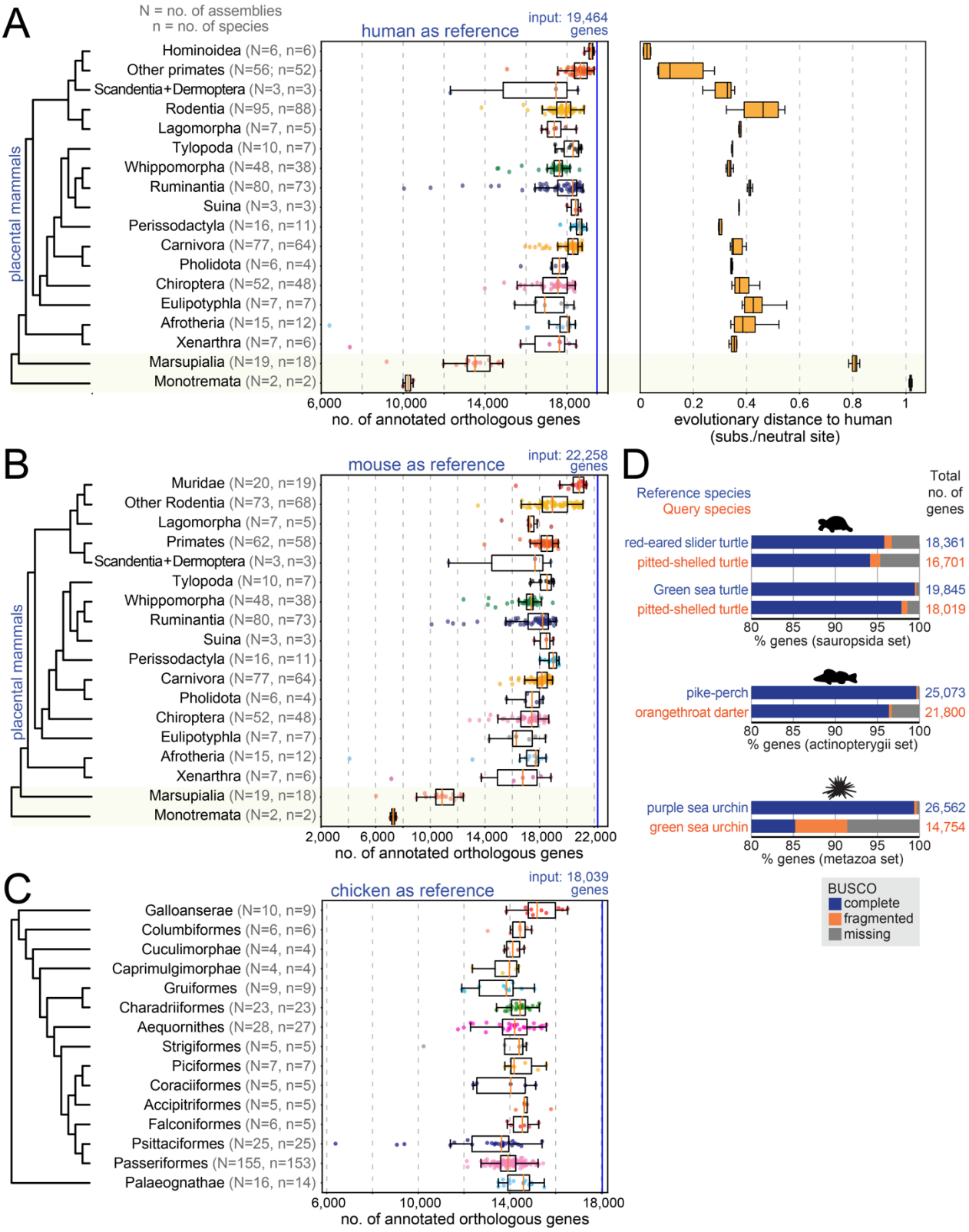

Figure 5: Large-scale application of TOGA to hundreds of genomes.

(A) TOGA with human as the reference. Left: Phylogenetic tree of mammal orders (7). Box plots with overlaid data points show the number of annotated orthologs. Hominoidea are shown as a

685 separate group. Non-placental mammals (marsupials and monotremes) are highlighted with a
686 yellow background. Right: Box plots showing the distributions of evolutionary distances to human
687 (Table S2).
688 (B) TOGA with mouse as the reference. Muridae are shown as a separate group.
689 (C) TOGA with chicken as the reference, applied to 308 bird assemblies.
690 (D) Using TOGA with other reference species (blue) to annotate related query species (orange).
691 The bar charts compare the BUSCO gene completeness of the input (reference) annotation, which
692 provides an upper bound, and the query annotation generated by TOGA. It should be noted that
693 the two sea urchins split ~200 Mya.
694
695

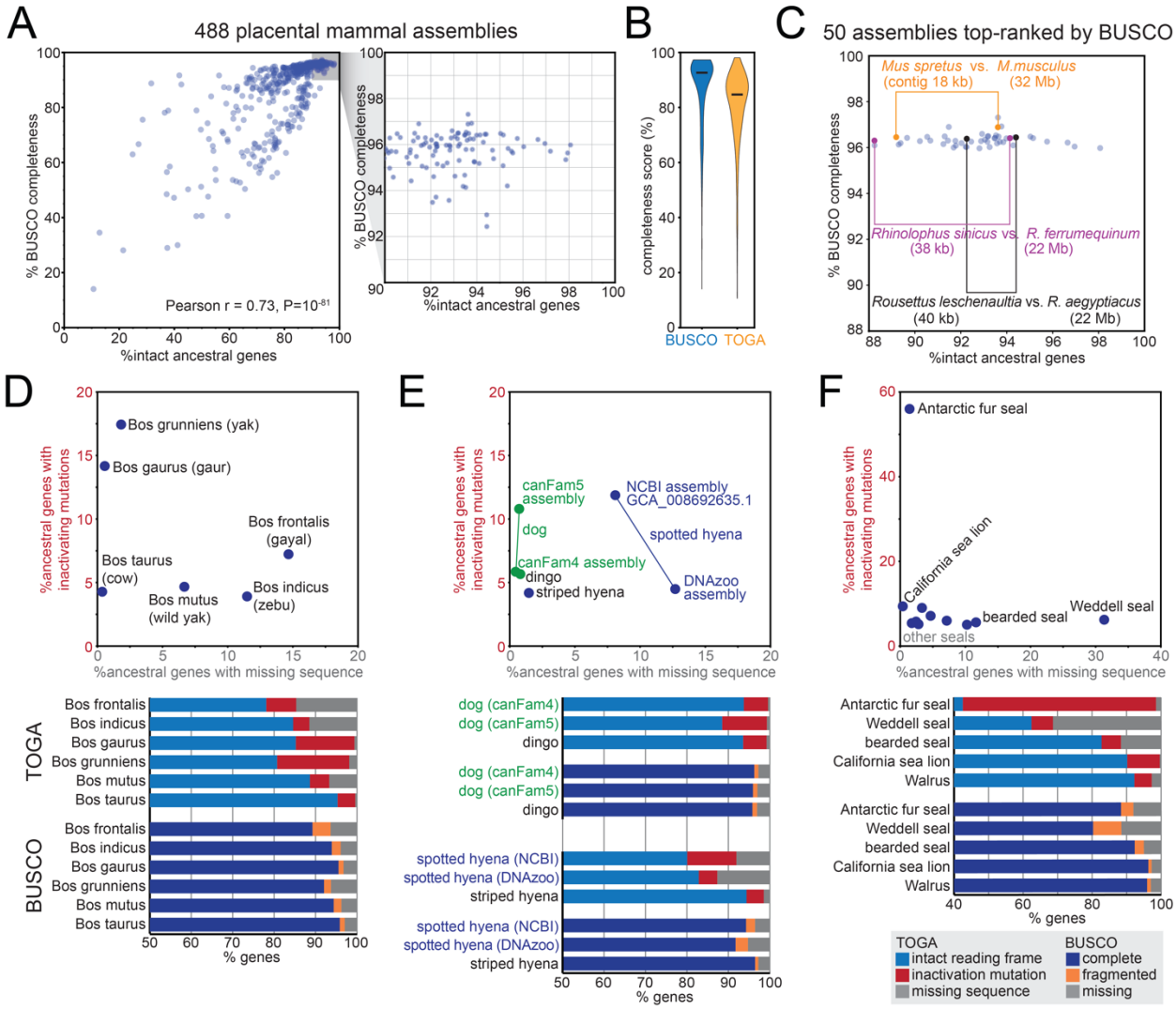

Figure 6: TOGA provides a superior measure of mammalian genome quality.

(A) Comparison of the percent complete BUSCO genes (Y-axis) and TOGA's percent of intact ancestral genes (X-axis) for 488 placental mammal assemblies. The inset shows that BUSCO's completeness values saturate at a maximum of 97.3%, whereas TOGA's value offers a larger dynamic range.

(B) Violin plots of BUSCO's and TOGA's completeness values. Horizontal black lines represent the median.

(C) BUSCO and TOGA values for the 50 assemblies that are top-ranked by BUSCO. Three pairs of closely related species are highlighted that have substantially different assembly contiguity (contig N50) values and are distinguishable in terms of gene completeness by TOGA but not by BUSCO.

(D-F) TOGA determines the percent of ancestral genes that have missing sequence and that have inactivating mutations (X and Y-axis in the dot plots at the top). Bar charts compare the TOGA gene classification with the percent of complete, fragmented and missing genes computed by BUSCO. The three panels highlight assemblies with a higher incompleteness or base error rate

712 (inferred from an increased percentage of genes with inactivating mutations) that is often not
713 detectable by the BUSCO metrics.
714
715
716
717

Zoonomia Consortium:

Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R Bredemeyer[5], Ana M Breit[6], Matthew J Christmas[3], Joana Damas[7], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[8], Kaili Fan[1], Cornelia Fanter[9], Nicole M. Foley[5], Karin Forsberg-Nilsson[10], Carlos J. Garcia[11], John Gatesy[12], Steven Gazal[13], Diane P. Genereux[4], Daniel Goodman[14], Linda Goodman[15], Jenna Grimshaw[11], Michaela K. Halsey[11], Andrew J Harris[5], Glenn Hickey[16], Michael Hiller[17,51,52], Allyson G. Hindle[9], Robert M. Hubley[18], Laura Huckins[53], Graham M. Hughes[19], Jeremy Johnson[4], David Juan[20], Irene M. Kaplow[21,22], Elinor K. Karlsson[1,4], Kathleen C. Keough[23,24], Bogdan Kirilenko[17,51,52], Klaus-Pieter Koepfli[54], Jennifer M. Korstian[11], Sergey V. Kozyrev[3], Alyssa J. Lawler[25], Colleen Lawless[19], Danielle L. Levesque[6], Harris A. Lewin [7,26,27], Xue Li[1,4] , Yun Li 47, Abigail Lind[23,24], Kerstin Lindblad-Toh[3,4], Voichita D. Marinescu[3], Tomas Marques-Bonet[20,28,29,30], Victor C Mason[31], Jennifer R. S. Meadows[3], Jill E, Moore[1], Diana D. Moreno-Santillan[11], Kathleen M. Morrill[1,4], Gerard Muntané[20], William J Murphy[5], Arcadi Navarro[20,32,33,34], Martin Nweeia[35,36,37,38], Austin Osmanski[11], Benedict Paten[2], Nicole S. Paulat[11], Eric Pederson[3], Andreas R. Pfenning[21,22], BaDoi N. Phan[21], Katherine S. Pollard[23,24,39], Kavya Prasad[21], Henry Pratt[1], David A. Ray[11], Jeb Rosen[18], Irina Ruf [40], Louise Ryan[19], Oliver A. Ryder[41,42], Daniel Schäffer[21], Aitor Serres[20], Beth Shapiro[43,44,] Arian F. A. Smit[18], Mark Springer[45], Chaitanya Srinivasan[21], Cynthia Steiner[46], Jessica M. Storer[18], Patrick F. Sullivan[47,48], Kevin A. M. Sullivan[11], Quan Sun[47], Elisabeth Sundström[3], Megan A Supple[44], Ross Swofford[4], Jin Szatkiewicz[47], Joy-El Talbot[49], Emma Teeling[19], Jason Turner-Maier[4], Alejandro Valenzuela[20], Franziska Wagner[50], Ola Wallerman[3], Chao Wang[3], Juehan Wang[13], Jia Wen 47, Zhiping Weng[1], Aryn P. Wilder[41], Morgan E. Wirthlin[21,22], Shuyang Yao[48], Xiaomeng Zhang[21]

1 Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA
2 Genomics Institute, UC Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA
3 Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, 751 32, Sweden
4 Broad Institute of MIT and Harvard, Cambridge MA 02139, USA
5 Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA
6 School of Biology and Ecology, University of Maine, Orono, Maine 04469, USA
7 The Genome Center, University of California Davis, Davis, CA 95616, USA
8 School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, 90619-900, Brazil
9 School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA
10 Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, 751 85, Sweden
11 Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA
12 Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA
13 Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA
14 University of California San Francisco, San Francisco, CA 94143 USA
15 Fauna Bio Inc., Emeryville, CA 94608, USA
16 Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA
17 LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt, Germany
18 Institute for Systems Biology, Seattle, WA 98109, USA
19 School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland

20 Institute of Evolutionary Biology (UPF-CSIC), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, 08003, Spain
21 Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
22 Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
23 Gladstone Institutes, San Francisco, CA 94158, USA
24 Department of Epidemiology & Biostatistics, University of California, San Francisco, CA 94158, USA
25 Department of Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA
26 Department of Evolution and Ecology, University of California, Davis, CA 95616, USA
27 John Muir Institute for the Environment, University of California, Davis, CA 95616, USA
28 Catalan Institution of Research and Advanced Studies (ICREA), 08010, Barcelona, Spain
29 CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), 08036, Barcelona, Spain
30 Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193, Cerdanyola del Vallès, Barcelona, Spain
31 Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland
32 Catalan Institution of Research and Advanced Studies (ICREA), 08010, Barcelona, Spain
33 CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), 08036, Barcelona, Spain
34 BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, Barcelona, 08005 Spain
35 Narwhal Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA
36 Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA
37 Department of Vertebrate Zoology, Smithsonian Institution, Washington, DC 20002, USA
38 Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, Ontario K2P 2R1, Canada
39 Chan Zuckerberg Biohub, San Francisco, CA 94158, USA
40 Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany
41 Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA
42 Department of Evolution, Behavior and Ecology, Division of Biology, University of California, San Diego, La Jolla, CA 92039 USA
43 Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA
44 Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA
45 Department of Evolution, Ecology and Organismal Biology, University of California, Riverside, CA 92521, USA
46 Conservation Science Wildlife Health, San Diego Zoo Wildlife Alliance, Escondido CA 92027, USA
47 Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA
48 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
49 Iris Data Solutions, LLC, Orono, ME 04473, USA
50 Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany
51 Senckenberg Research Institute, Senckenberganlage 25, 60325 Frankfurt, Germany
52 Goethe-University, Faculty of Biosciences, Max-von-Laue-Str. 9, 60438 Frankfurt, Germany
53 Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
54 Center for Species Survival, Smithsonian Conservation Biology Institute, National Zoological Park, Washington, DC, USA