Science
AAAS

Supplementary Materials for

Integrating gene annotation with orthology inference at scale

Bogdan M. Kirilenko[1,2,3,4,5,6], Chetan Munegowda[1,2,3,4,5,6], Ekaterina Osipova[1,2,3,4,5,6], David Jebb[1,2,3], Virag Sharma[1,2,3,#], Moritz Blumer[1,2,3], Ariadna E. Morales[4,5,6], Alexis-Walid Ahmed[4,5,6], Dimitrios-Georgios Kontopoulos[4,5,6], Leon Hilgers[4,5,6], Kerstin Lindblad-Toh[7,8], Elinor K. Karlsson[8,9,10], Zoonomia Consortium[11], and Michael Hiller[1,2,3,4,5,6*]

[1] Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307 Dresden, Germany.

[2] Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Str. 38, 01187 Dresden, Germany.

[3] Center for Systems Biology Dresden, Pfotenhauerstr. 108, 01307 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 Frankfurt, Faculty of Biosciences, Max-von-Laue-Str. 9, 60438 Frankfurt, Germany.

[7] Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University; Uppsala, 751 32, Sweden.

[8] Broad Institute of MIT and Harvard; Cambridge, MA 02139, USA.

[9] Program in Bioinformatics and Integrative Biology, UMass Chan Medical School; Worcester, MA 01605, USA.

[10] Program in Molecular Medicine, UMass Chan Medical School; Worcester, MA 01605, USA

[11] Zoonomia Consortium Members are listed at the end of the document

[#] Current affiliation: Department of Chemical Sciences, School of Natural Sciences, University of Limerick, Limerick, V94 T9PX, Ireland

Correspondence to: michael.hiller@senckenberg.de

This PDF file includes:

Figs. S1 to S45

Captions for Tables S1-S15

Other Supplementary Materials for this manuscript include the following:

Tables S1-S15 as sheets in an Excel file

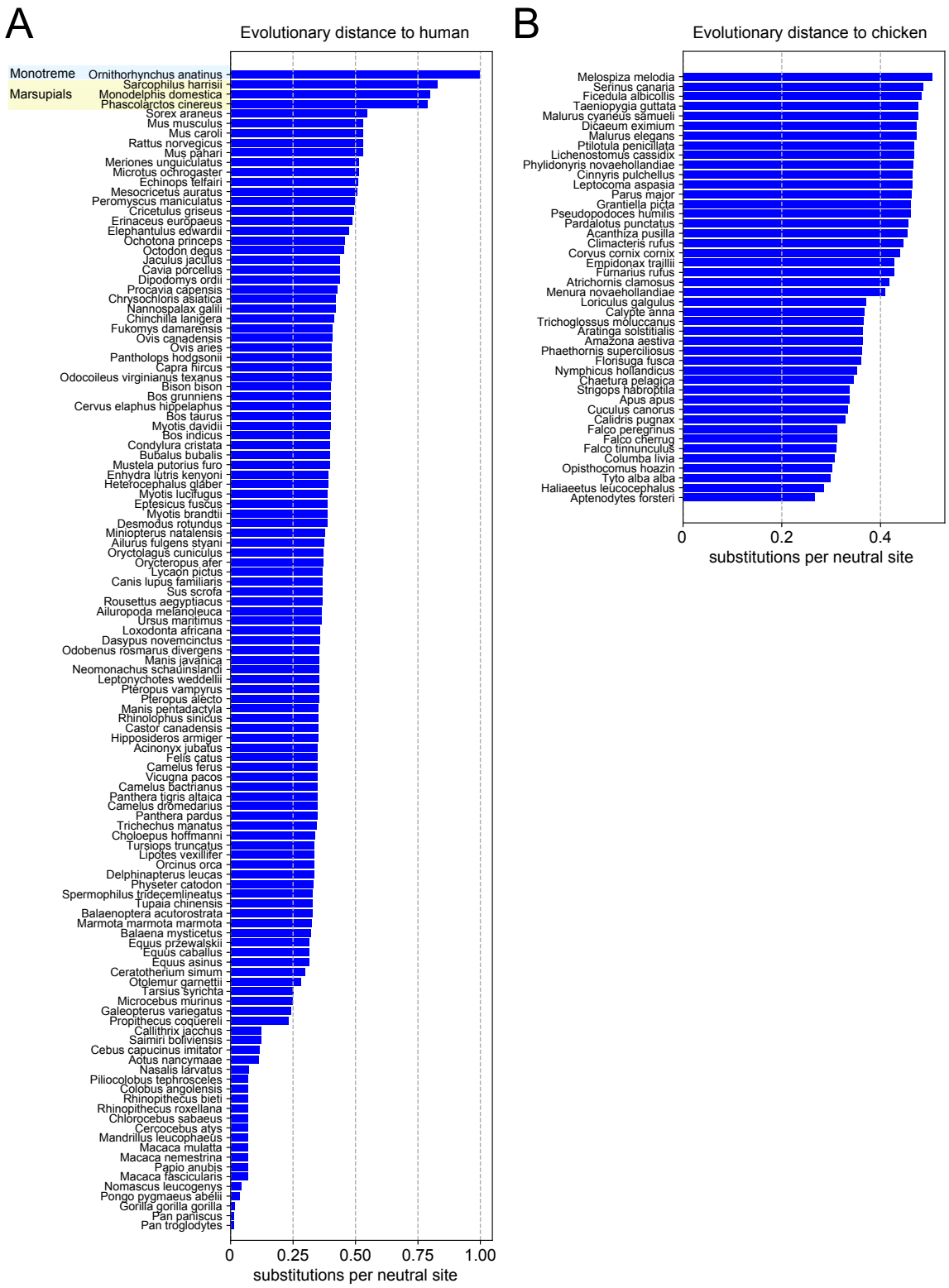


Fig. S1: Evolutionary distance between human and other mammals (A), and chicken and other birds (B).

Evolutionary distance is measured as the number of substitutions per neutral site (Y-axis). In (A), three marsupial and one monotreme species are highlighted. Placental mammals shared a common ancestor up to ~100 Mya (*31*). Birds also shared a common ancestor up to ~100 Mya (*32*).

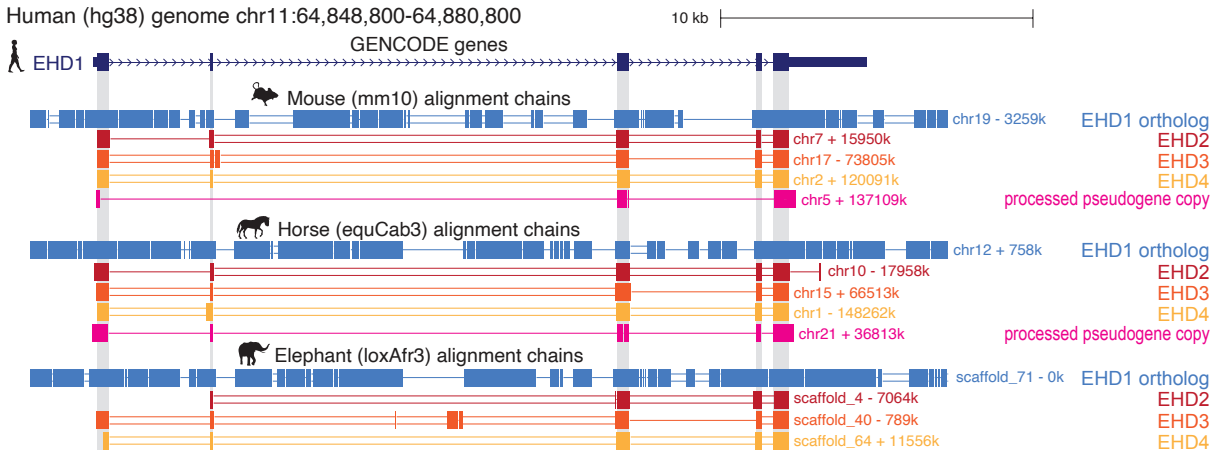


Fig. S2: Characteristic features of orthologous alignment chains.
UCSC genome browser view of the human *EHD1* gene locus and alignment chains to three representative placental mammal species: mouse (recapitulated from Fig. 1A for comparison), horse and elephant. For all species, several chains align coding exons of *EHD1*; however, only the chain (blue) that represents the orthologous *EHD1* locus in other species aligns both exons and parts of introns and flanking intergenic regions. The other chains lack intronic and intergenic alignments and represent paralogs or, in case of horse and mouse, a processed pseudogene copy.

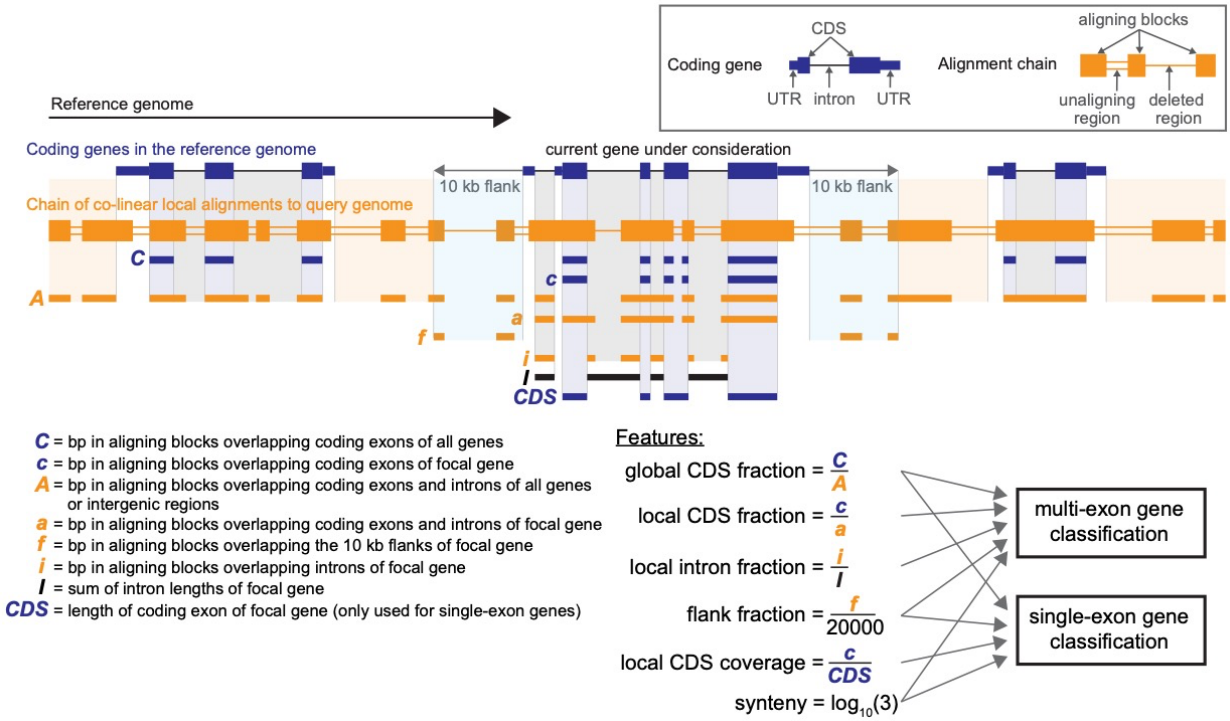


Fig. S3: Illustration of the features TOGA extracts from each candidate chain for orthology classification.

The X-axis represents a part of the reference genome. Coding and untranslated exonic parts are shown as large and small dark-blue boxes. The alignment chain to the query genome is shown in orange with boxes representing local alignments, double lines regions that do not align between reference and query, and single lines representing regions deleted in the query or inserted in the reference genome. The example chain overlaps three coding genes. As illustrated, the "global" variables *C* and *A* are computed for all genes, while the "local" variables *c*, *a*, *f*, *i*, *I*, and *CDS* are computed only for the focal gene. All variables are computed using reference coordinates. Note that UTR regions (white background) are ignored. Since features measuring intronic alignments cannot be computed for single-exon genes, the features used for orthology classification of single- or multi-exon genes differ slightly.

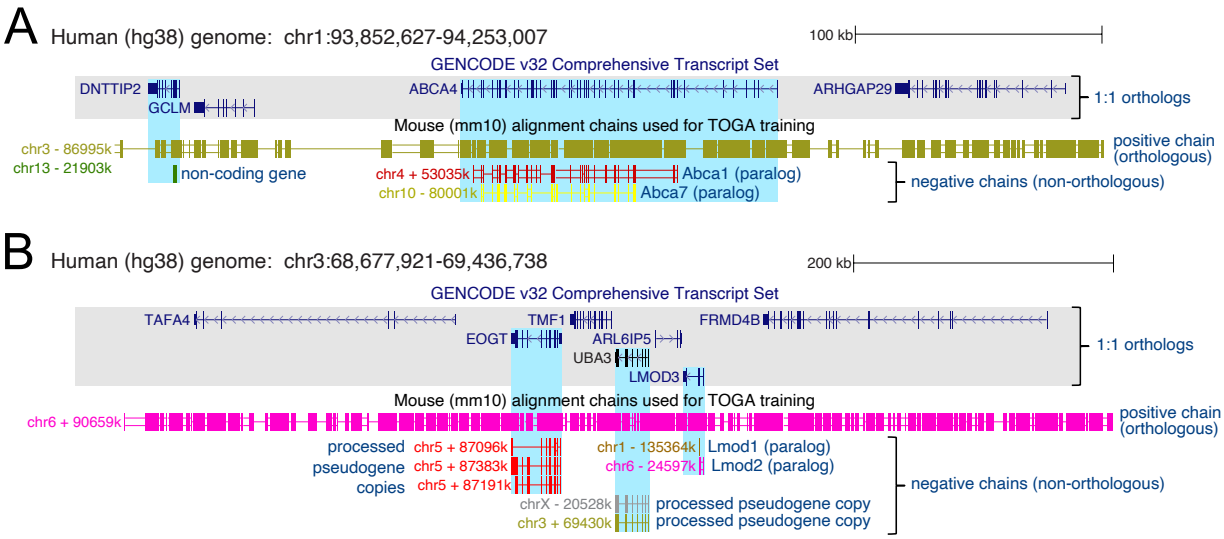


Fig. S4: Generating positive (1:1 orthologous) and negative (non-orthologous) training data for human-mouse genome alignments.
(A) and (B) show UCSC genome browser screenshots of two exemplary human loci and chains of co-linear local alignments to mouse (mm10 assembly). The genes shown are classified as 1:1 orthologs between human and mouse by Ensembl. For each human gene, the Ensembl-annotated mouse ortholog is located at the coordinates provided by the top-level chr3 (A) and chr6 (B) chain. Thus, these two chains represent the orthologous chain for the respective genes and we used both chains as positive training data. A 1:1 orthology relationship implies that other exon-overlapping chains cannot represent co-orthologs. Indeed, these chains represent alignments to paralogous or processed pseudogenes, as annotated in blue font. Hence, we used these chains as negative (non-orthologous) training data.

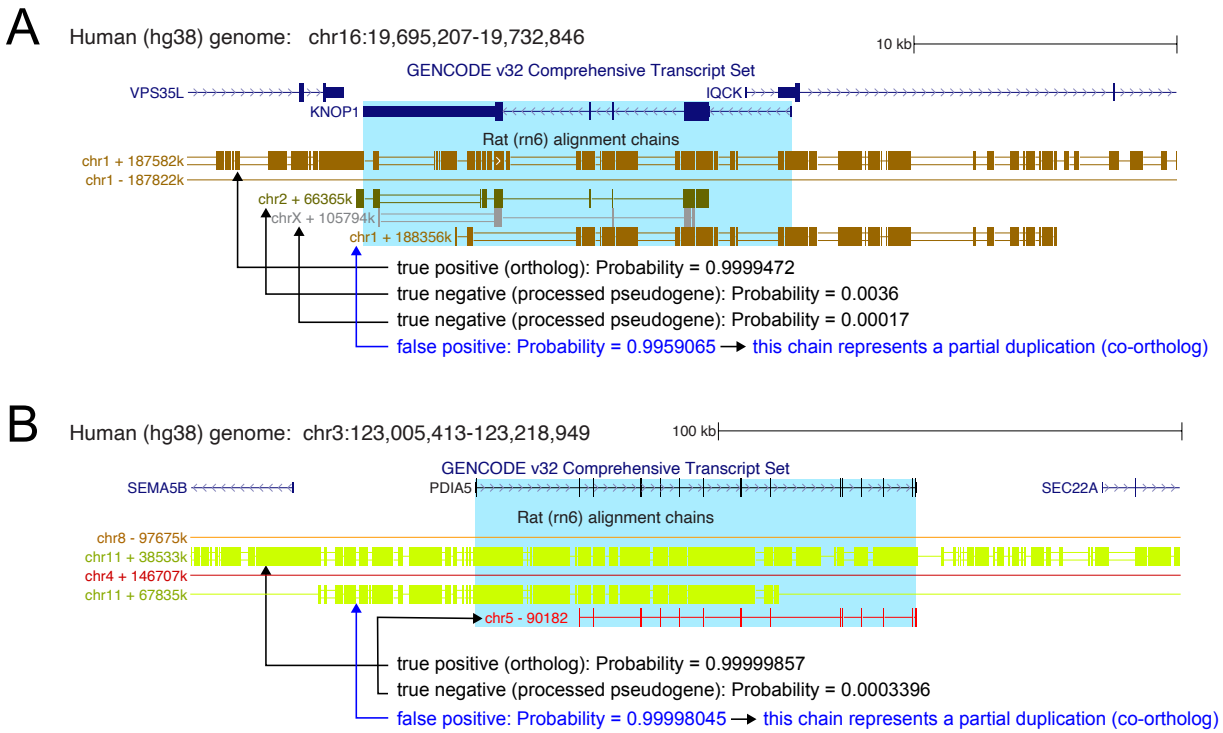


Fig. S5: False positive chain classifications in the human-rat test dataset often represent partial co-orthologs (gene duplications).

(A) UCSC genome browser screenshot showing the human *KNOP1* gene locus. Ensembl classifies *KNOP1* as a 1:1 ortholog between human and rat. The top level chain represents the Ensembl-annotated ortholog of *KNOP1* and was correctly classified as such by TOGA (probability >0.99). Two other chains (green and grey) are processed pseudogene copies of *KNOP1* and were also correctly classified as non-orthologous (probabilities <0.004). Importantly, the fourth chain (brown) shows that parts of *KNOP1* and the neighboring *IQCK* gene were duplicated in rat (note the near identical chain block structure). The duplication is specific to the rat rn6 genome and the duplicated region is located ~200 kb upstream of the original *KNOP1* locus in rat. Since Ensembl annotated *KNOP1* as a 1:1 ortholog between human and rat, we considered all other exon-overlapping chains including this fourth chain as negatives (paralogs or processed pseudogenes) in the test data. However, this chain is actually a co-ortholog (lineage-specific duplication) and TOGA indeed estimates a high probability of 0.996 that this chain represents an orthologous locus. Supporting this, Ensembl (ENSRNOT00000075020.1) does annotate a shorter 380 amino acid-comprising *KNOP1* gene at the duplicated locus; however, Ensembl does not classify this gene as a co-ortholog to human *KNOP1*. In summary, the fourth chain was incorrectly labeled as negative, resulting in a TOGA classification that we conservatively consider as a false positive, even though TOGA actually correctly classified this chain as one orthologous locus.

(B) UCSC genome browser screenshot showing the human *PDIA5* gene locus. Ensembl classifies *PDIA5* as a 1:1 ortholog between human and rat, and TOGA correctly classifies the respective top chr11 chain (green) as orthologous (probability >0.999). The second chr11 chain shows that the upstream part of *PDIA5* is duplicated in the rat rn6 genome. As in (A), relying on the Ensembl ortholog annotation of *PDIA5*, we considered this chain as negative in our test data, even though this chain actually represents a co-orthologous (lineage-specific duplicated) locus and TOGA correctly estimates a high probability (>0.999).

In summary, these findings indicate that TOGA is able to detect lineage-specific gene duplications and actually correctly classified these chains as co-orthologous loci.

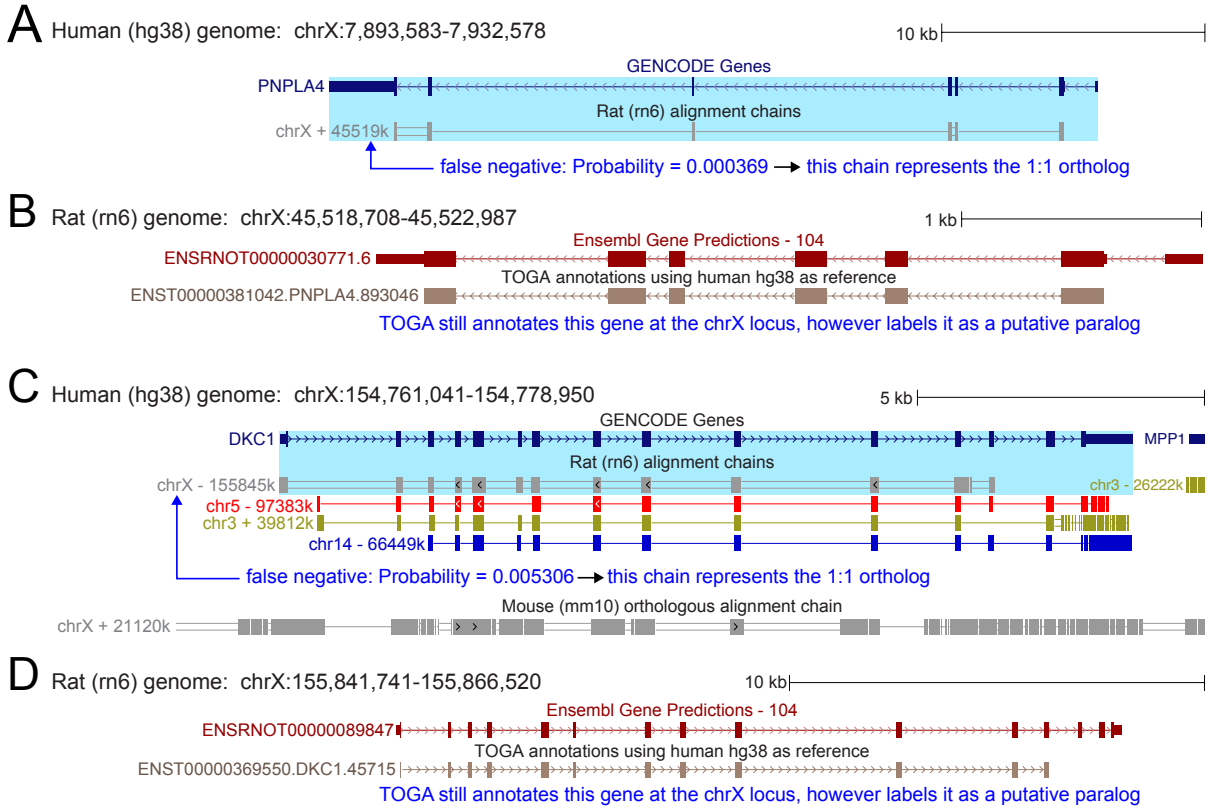


Fig. S6: False negative chain classifications in the human-rat test dataset are characterized by exceptional intron divergence and lack of conserved gene order, which can partially be explained by faster X chromosome evolution.

(A) UCSC genome browser screenshot showing the human *PNPLA4* gene locus, which is correctly annotated as a 1:1 ortholog between human and rat by Ensembl. The chrX chain lacks intronic and intergenic alignments. While this chain appears to be a typical paralogous chain, it actually represents the orthologous locus of *PNPLA4*. Showing a limitation of our approach, TOGA incorrectly classifies the chain as a non-orthologous locus.

(B) While TOGA incorrectly infers a paralogous locus, TOGA does annotate a gene at the rat chrX locus, since this locus receives no annotation via an orthologous chain. However, the annotation is labeled as a paralogous projection as the chain used for annotation was classified as non-orthologous.

(C) UCSC genome browser screenshot showing the human *DKC1* gene locus that Ensembl correctly annotated as a 1:1 ortholog between human and rat. As in (A), the chrX chain represents the true orthologous locus; however, this chain contains barely any intronic and intergenic alignments and was therefore incorrectly classified as a non-orthologous locus by TOGA. For comparison, we show the orthologous mouse chain, which does not exhibit this exceptional divergence. The other rat chains (red, green, blue color) are non-orthologous and likely represent processed pseudogene copies.

(D) As in (B), TOGA annotates a gene at the rat chrX locus, but labels it as a putative paralog. The reason underlying the exceptional sequence divergence of all neutrally evolving intronic regions in these loci are not known, but faster X chromosome evolution (*17*) could be one factor, since 58% (7 of 12) of the false negatives for human-rat are X-chromosome linked genes.

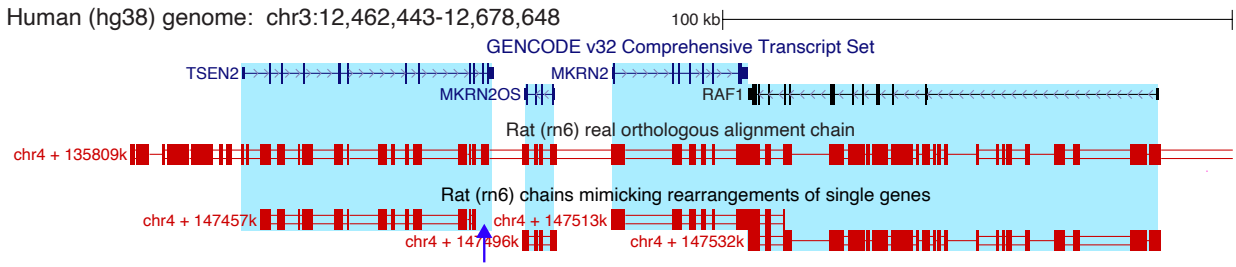


Fig. S7: Augmenting the positive training and test dataset of TOGA by single gene rearrangements.

UCSC genome browser screenshot showing the human genome and the top-level alignment chain to rat, which represents the orthologs of all four genes shown in this 216 kb locus. The entire chain spans 12.5 Mb of the human genome at the beginning of chromosome 3 and represents dozens of genes, thus the synteny feature of this chain is high. Since real inversions or translocations are rarely observed in our data and since our goal was that TOGA is also able to accurately detect inverted or translocated orthologs, we created artificial rearrangements of individual genes by trimming the long syntenic chain. The resulting single gene-covering chains, shown at the bottom, mimic real rearrangements of a single gene and were added as real orthologs to the positive training data (mouse chains) and test data (rat chains, shown here). Trimmed chains were allowed to extend at most 10 kb upstream or downstream of the CDS. Please note that a trimmed chain was allowed to miss up to 20% of the CDS of the gene, as shown for *TSEN2*, where the alignments of the last two coding exons were trimmed (blue arrow). Thus, a trimmed chain can be shorter than the gene, mimicking a partial rearrangement.

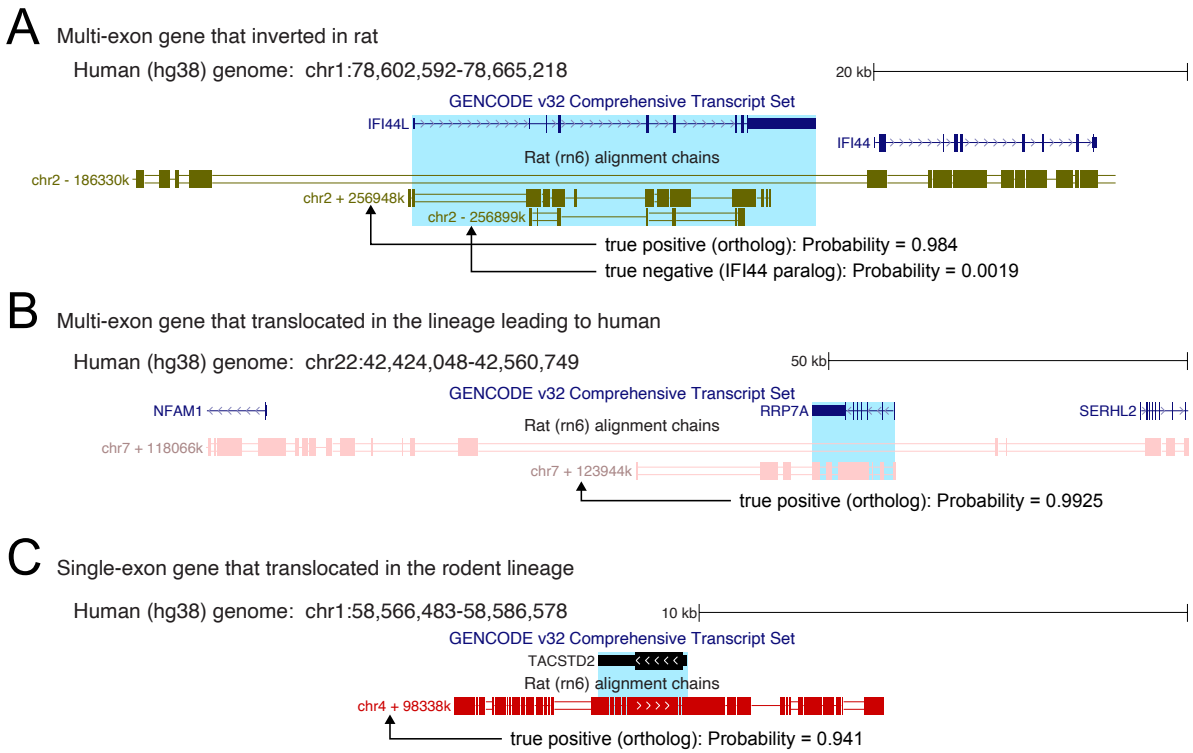


Fig. S8: Examples of real single gene rearrangements that were correctly-classified as orthologous by TOGA.

(A) UCSC genome browser screenshot showing human locus comprising the *IFI44L* and *IFI44* genes. In contrast to mouse and other mammals, the *IFI44L* gene is inverted in rat, as shown by a second level chr2 + strand chain, which represents the rat ortholog. This local inversion breaks co-linearity with the surrounding alignments, resulting in a chain that covers only this gene. Consequently, the synteny feature of this chain has a low value of $\log_{10}(1)$. Nevertheless, due to training TOGA on a dataset augmented with such single gene rearrangements and using other features in addition to synteny, TOGA correctly classifies this chain as the ortholog with a high probability of 0.98. The third-level chain represents the alignment between human *IFI44L* and the rat paralog *IFI44*, and this chain is correctly classified as non-orthologous.

(B) UCSC genome browser screenshot showing the human *RRP7A* gene locus. The second level alignment chain for rat represents the ortholog and shows that this gene was translocated. Importantly, inspecting chains of other mammals (for clarity not shown here) revealed that this rearrangement happened along the primate lineage, before the split of the great apes. Thus, the rearrangement occurred in the lineage leading to the reference species (here human). As in (A), the chain covers only *RRP7A*. Nevertheless, due to intronic and gene-flanking alignments, TOGA is able to correctly classify this chain as the ortholog with a high probability of 0.99.

(C) UCSC genome browser screenshot showing the single exon *TACSTD2* gene. The chain representing the ortholog in rat covers only this gene with upstream and downstream flanks, because *TACSTD2* translocated to a different locus in rat (this translocation is shared with mouse, not shown here). TOGA computes features quantifying the amount of gene-flanking alignments, enabling us to correctly classify this chain as orthologous with probability 0.94, despite the lack of conserved gene order.

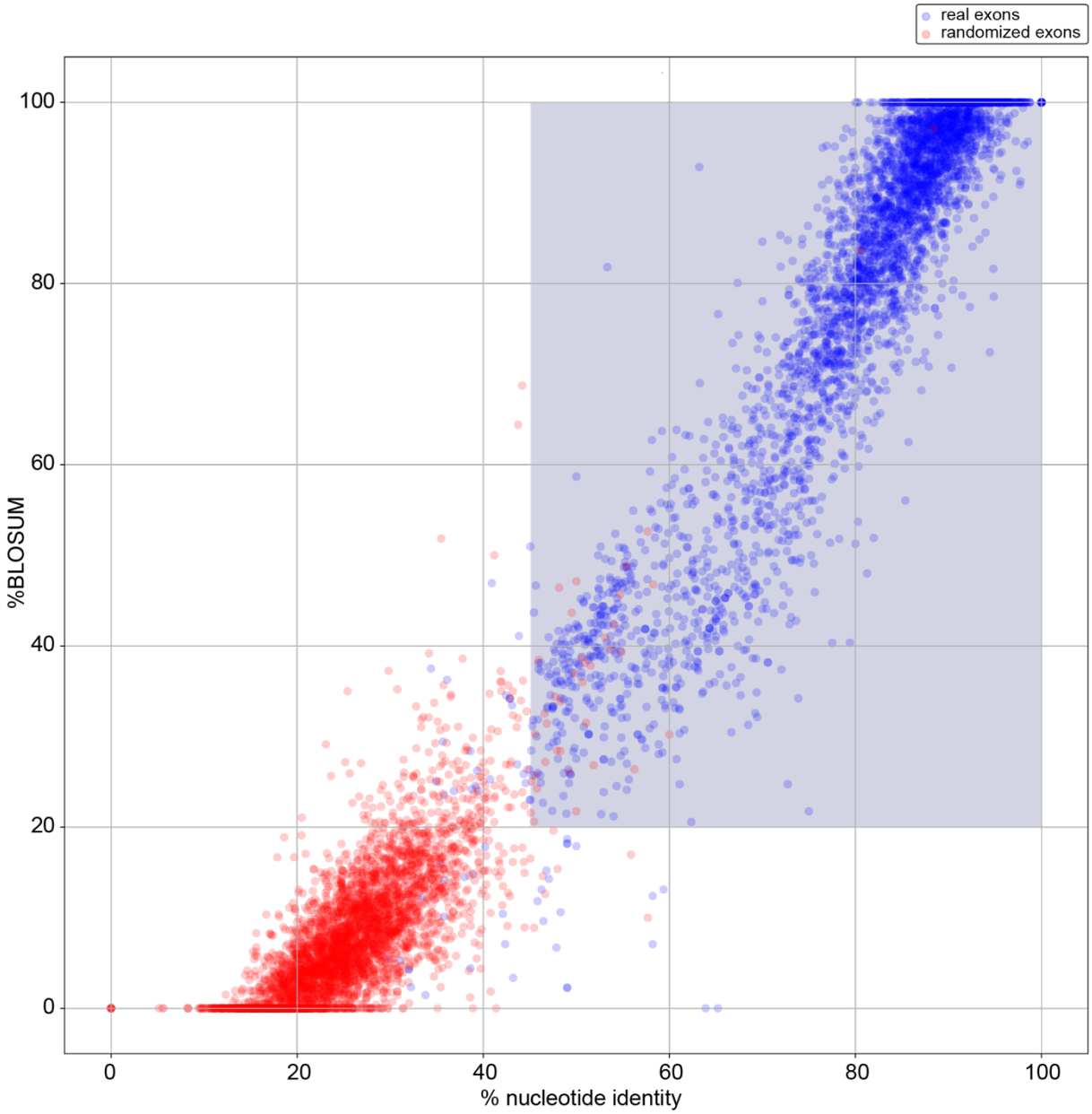


Fig. S9: Percent nucleotide identity and BLOSUM thresholds obtained from alignments between real and randomized exons.

We extracted 137,935 exons belonging to human-mouse 1:1 orthologous genes, where the TOGA-annotated exon overlaps an Ensembl-annotated exon (real exons, blue color). For each exon, we computed %nucleotide identity (X-axis) and %BLOSUM (Y-axis), two metrics that quantify how well this exon aligns at the nucleotide and protein level. To compare %nucleotide identity and %BLOSUM between real and random exon alignments, we aligned the 137,935 reference exons to randomized exons, obtained by reversing the query sequence (red color). Based on this data, we defined a threshold (blue background) as %nucleotide identity $\geq$ 45% and %BLOSUM $\geq$ 20%. This threshold has a sensitivity of 0.9808 and a precision of 0.99075, and is used by TOGA to distinguish real from random exon alignments. For visualization clarity, only 4,000 randomly-selected real and randomized exons are plotted.

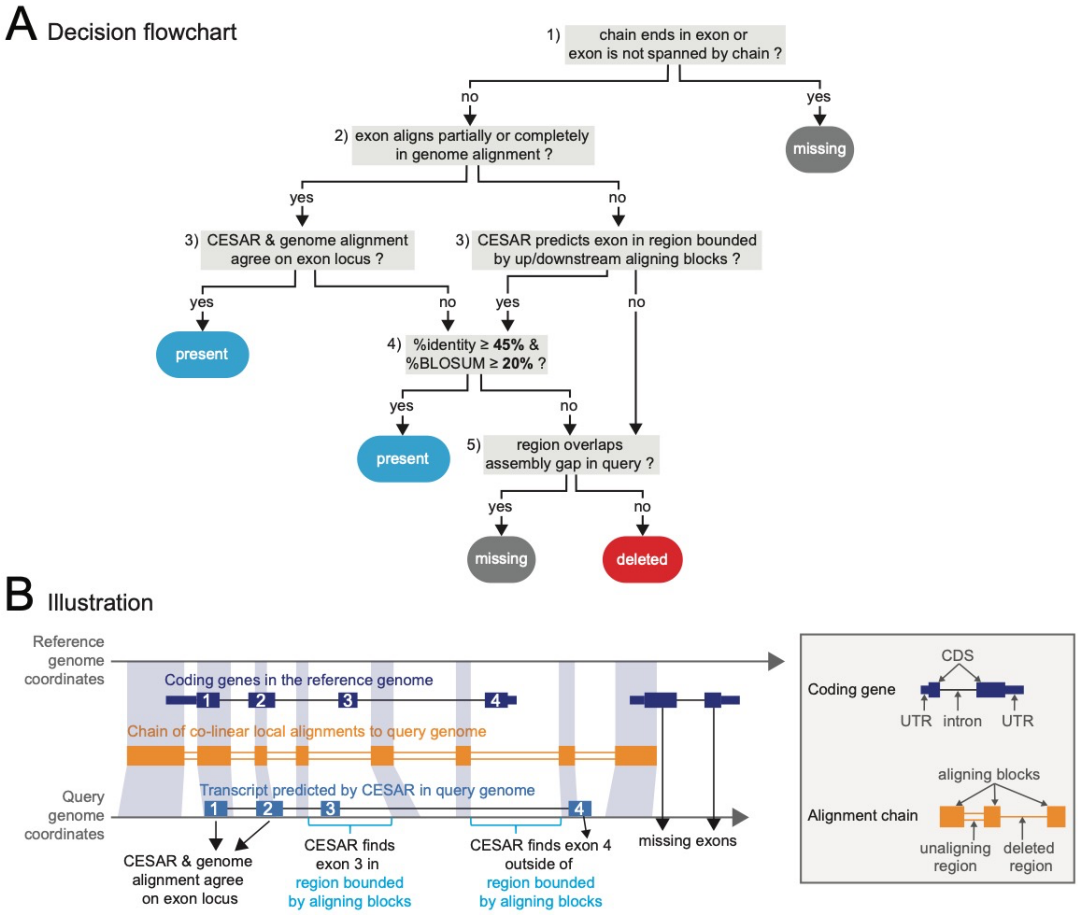


Fig. S10: Exon classification implemented in TOGA.

(A) Flowchart of the decisions used to classify exons as present, missing or deleted. For a given exon, TOGA assesses whether the orthologous chain ends inside this exon or whether the exon is not spanned by an orthologous chain (1). These exons are classified as missing. For all other exons, TOGA determines an 'expected region', which is a region in the query genome that should overlap or contain the exon according to the genome alignment chain. For exons that partially or completely align in the genome alignment, this expected region is defined by the coordinates of exon-overlapping alignment blocks (2). TOGA then assesses whether CESAR 2.0 in multi-exon mode (aligning all coding exons of the reference transcript to the orthologous query gene sequence that comprises exons and introns) finds an exon candidate in the expected region (3), which shows that genome alignment and CESAR agree on the exon location in the query. Such exons are classified as present. Exons, for which genome alignment and CESAR disagree on the query location (3), are classified as present if the CESAR exon alignment is better than randomized exon alignments, which TOGA determines using a threshold of %nucleotide identity ≥ 45% and %BLOSUM ≥ 20% (4). For exons that do not align at all in the genome alignment (2), the 'expected region' is defined as the query region bounded by the nearest up- and downstream alignment block in the chain (light blue curly bracket in panel B). Exons for which the CESAR exon candidate is found in the expected region (3) and that align better than randomized exons (4) are also classified as present. For all other exons, TOGA determines whether the expected region overlaps an assembly gap in the query genome (5). If so, the exon is classified missing, otherwise as deleted.

(B) Illustration of decisions (1), (2) and (3).

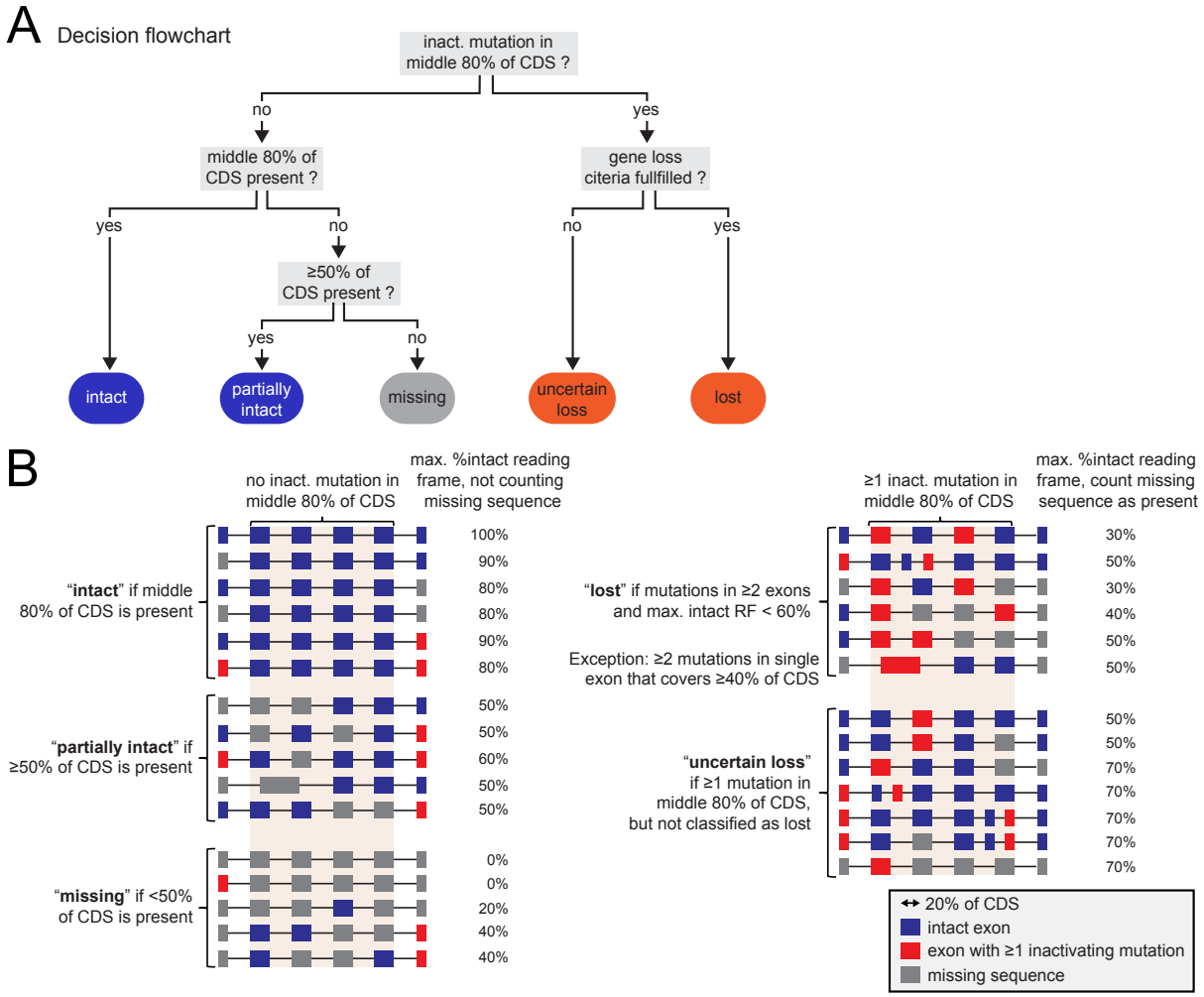


Fig. S11: Transcript classification implemented in TOGA.
(A) Flowchart implemented in TOGA to classify transcripts as "intact", "partially intact", "missing", "uncertain loss" and "lost". The first criterion determines whether a transcript exhibits inactivating mutations in the middle 80% of the CDS. This is based on observations that frameshift and stop codon mutations in truly conserved genes largely occur in the first or last 10% of the CDS (*19*) (Fig. S12). "Present" refers to part of the CDS that consists of exons classified as present.
(B) Illustration of different transcript classifications, together with the maximum percent of the reading frame that is intact in the query species. The example transcripts consist of 5-7 coding exons that make up 10%, 20% or 40% of the total CDS (boxes of different sizes). Blue and red boxes represent coding exons with and without inactivating mutations, grey boxes represent coding exons that are missing in the query (often due to assembly gaps). The criteria to classify a transcript as "lost" are detailed in the Methods section. Note that for transcripts exhibiting mutations in the middle 80% of the CDS, we conservatively count missing sequence parts as sequence that is present and lacks inactivating mutations, when determining the maximum percent of the intact reading frame (see also Fig. S13).

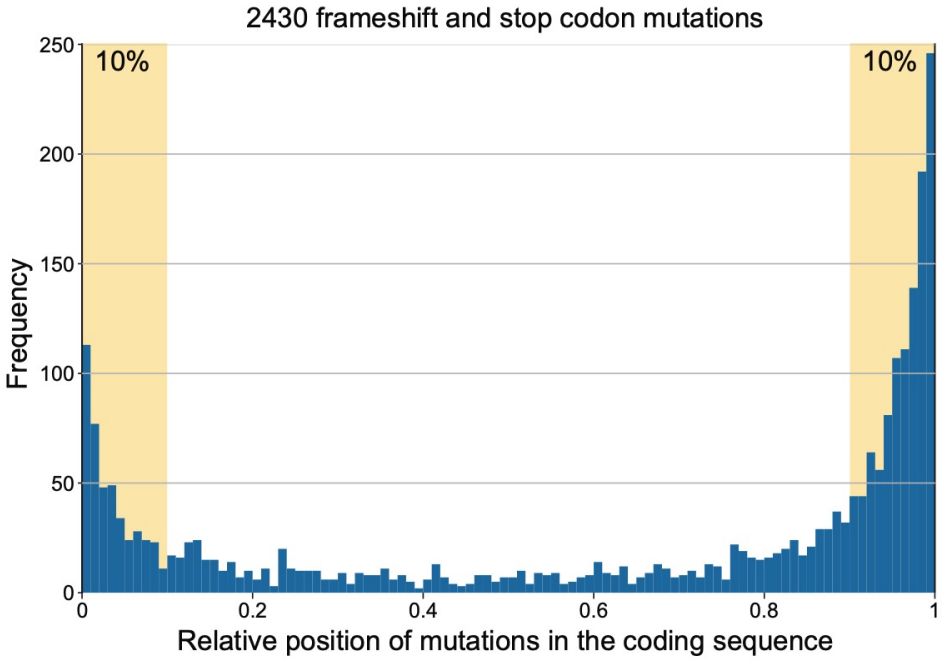


Fig. S12: Relative position of a total of 2,430 frameshift and stop codon mutations that TOGA detected in the coding sequence of human-mouse conserved genes.
Only 1:1 human-mouse orthologous genes according to the Ensembl were analyzed. The histogram shows that the relative position of 2,016 frameshifts and 414 premature stop codons (detected in 1,684 exons) is heavily biased toward the N-terminal and in particular to the C-terminal end of the coding region. In total, 62% of all frameshifts and stop codons occurred in the first or last 10% of CDS.

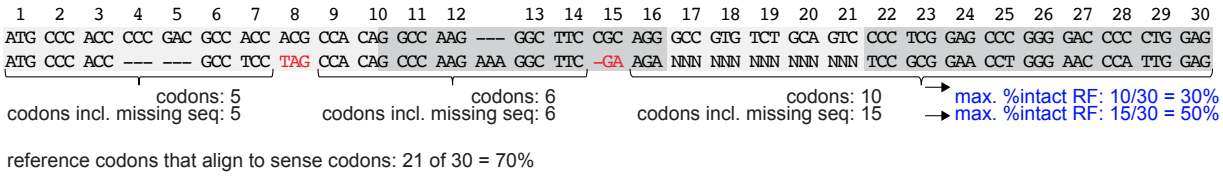


Fig. S13: Illustration of the computation of the maximum percent intact reading frame of a transcript.

The transcript has 30 codons in the reference and four exons (alternating shades of grey). NNN represents missing sequence, e.g. due to assembly gaps. Two inactivating mutations (stop codon and frameshift, red font) determine the boundaries of three individual parts of the reading frame that remain intact in the query species (curly brackets). For each part, we determine the number of codons that align, ignoring deleted and inserted codons. To distinguish between intact, partially intact and missing transcripts, we ignore missing sequence. In this example, the maximum percent of reading frame that remains intact is the third part with 10 of 30 codons aligning (30%). To distinguish between uncertain loss and lost transcripts, we count missing sequence as aligning codons, making the conservative assumption that NNN codons correspond to sense codons in the unknown query sequence. Then, the maximum percent of reading frame that remains intact covers 15 of 30 total codons (50%).

The figure also illustrates the calculation of the percentage of reference codons that align to sense codons in the query, which is 70% (21 of 30 reference codons).

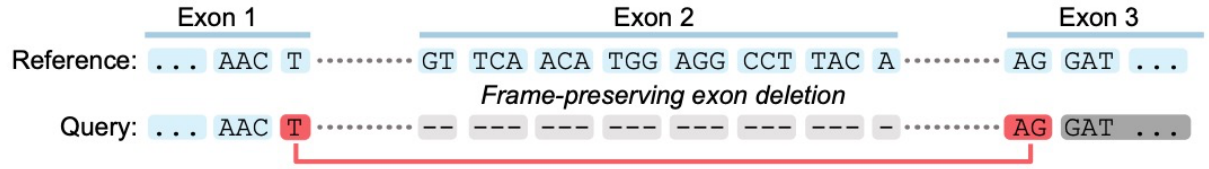


Fig. S14: Exon deletion(s) can introduce new premature stop codons.

Hypothetical example illustrating that exon deletions can result in stop codons which are "assembled" at the new exon-exon boundaries. While TOGA does not consider the frame-preserving deletion of exon 2 as gene-inactivating, it detects the stop codon that arises at the boundaries of exons 1 (T-) and 3 (-AG) as an inactivating mutation.

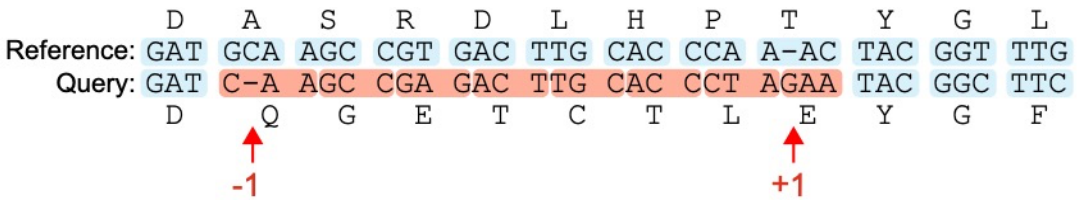


Fig. S15: Compensated frameshifts are not counted as gene-inactivating mutations.
The figure illustrates two consecutive -1 and +1 frameshifts that compensate each other in a hypothetical sequence. The sequence between these frameshifts can be translated in an alternative reading frame (red background) without stop codons, thus both mutations are not considered to be gene-inactivating. The sequence downstream of the second frameshift is translated in the ancestral reading frame.

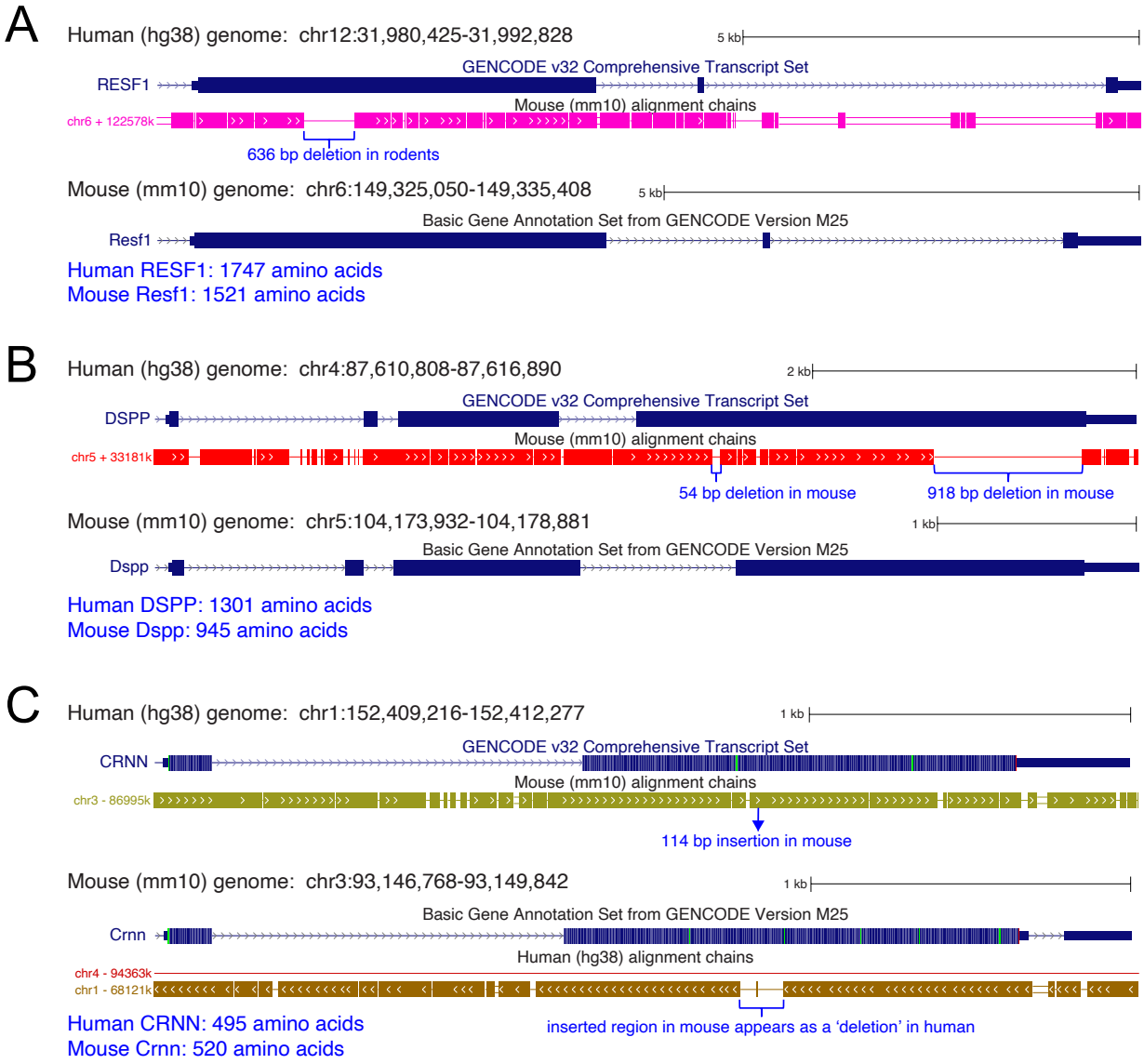
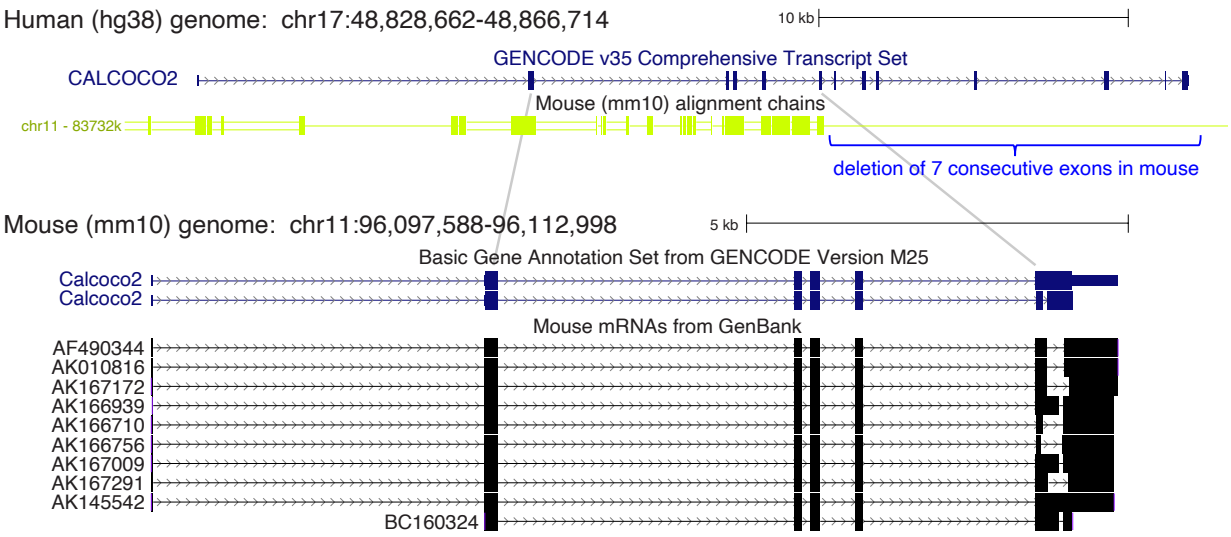


Fig. S16. Large frame-preserving deletions in large exons may not result in gene loss.
(A-C) UCSC genome browser screenshots show three human genes with large (>1 kb) exons. Alignment chains show that these exons exhibit large frame-preserving deletions in mouse. The length of the human and shorter mouse protein is given in blue font. Mouse genome browser screenshots show that the genes are annotated as coding genes, showing that these large frame-preserving deletions do not result in gene loss.
(A) A 636 bp deletion in a ~5 kb long coding exon of *RESF1* occurred in mouse and other rodents.
(B) A ~2.8 kb long coding exon of *DSPP* exhibit large size differences between human and mouse.
(C) A ~1.3 kb long coding exon of *CRNN* exhibits a large insertion in mouse and related rodents.



Human *CALCOCO2*: 12 coding exons; protein length 446 amino acids
Mouse *Calcoco2*: 5 coding exons; protein length 331 amino acids

Fig. S17. Deletions of several frame-preserving exons may not result in gene loss. *CALCOCO2* exhibits drastic exon structure changes between human and mouse. The human gene consists of 12 coding exons that are conserved in placental mammals and encodes a protein of 446 amino acids. In mouse, the last 7 consecutive exons are deleted, as shown in the UCSC genome browser screenshot by the human to mouse alignment chain. Despite these exon deletions, the mouse gene has an intact reading frame, encodes a shorter 331 amino acid long protein, and is expressed as shown by the mouse mRNA track in the UCSC genome browser.

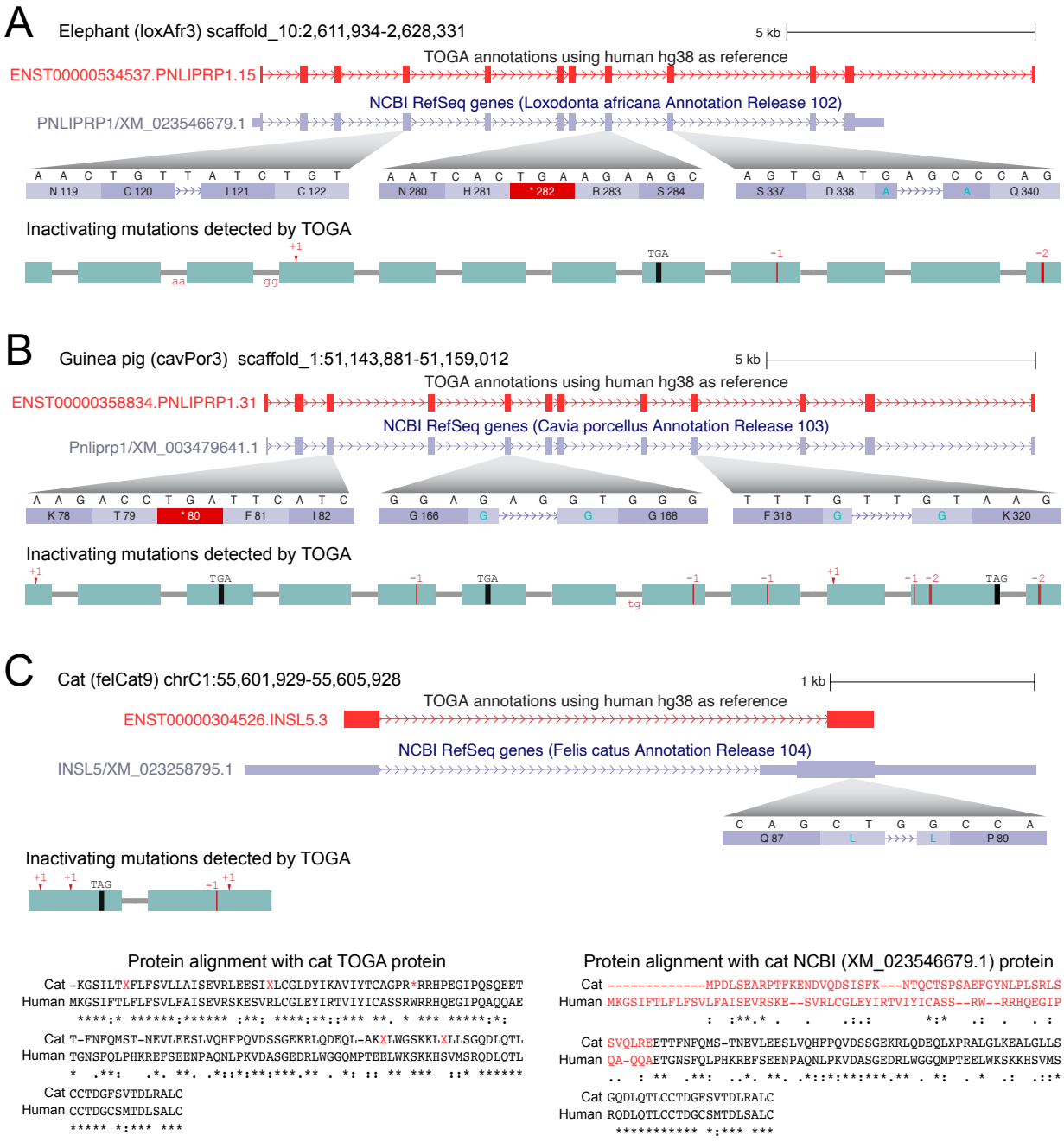


Fig. S18: TOGA does not attempt to fix putative assembly base errors.
(A-C) Three examples of genes lost in elephant (A), guinea pig (B) and cat (C). These examples represent genuine gene losses and not assembly base errors since the inactivating mutations are validated by unassembled DNA sequencing reads and are shared with independent assemblies of related sister species (*62*). The loss of *PNLIPRP1* is likely beneficial for herbivores by increasing the capacity for digesting dietary triglycerides and the loss of *INSL5* in cat and other carnivores is related to a carnivorous diet (*62*). Whereas TOGA classifies the three genes as lost, NCBI creates micro-exons around frameshifts and skips over stop codons in an attempt to correct putative base errors in the assembly (fig. S27 shows a similar case for Ensembl). In case of *INLS5*, NCBI also extends the 5' end of the second exon and annotates a non-homologous N-terminal protein sequence (red font).
TOGA's strategy is to classify genes based on the mutations present in the assembly. This strategy has the downside that some gene loss candidates are due to base errors (fig. S19), but it also provides two advantages. First, determining the number of genes with inactivating

mutations provides a benchmark for assembly quality, as assemblies with a higher base error rate are detectable by an increased percentage of genes with inactivating mutations (illustrated in our Fig. 6D-F). Second, TOGA detects putative gene losses, allowing users to validate mutations with unassembled DNA sequencing reads or assemblies of sister species (*6, 37, 62-64*).

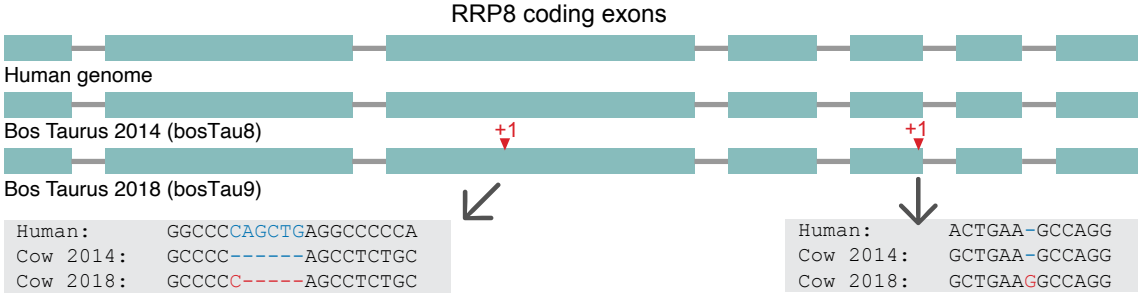
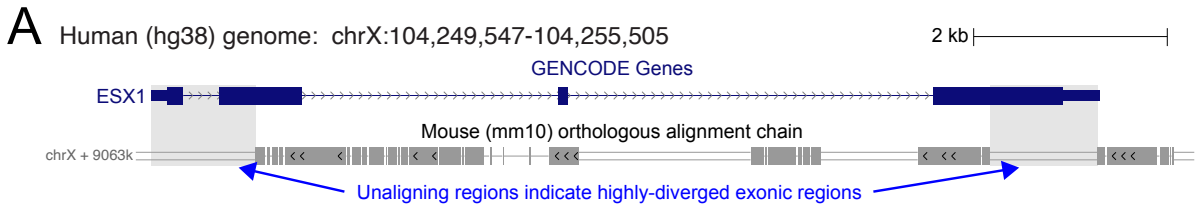
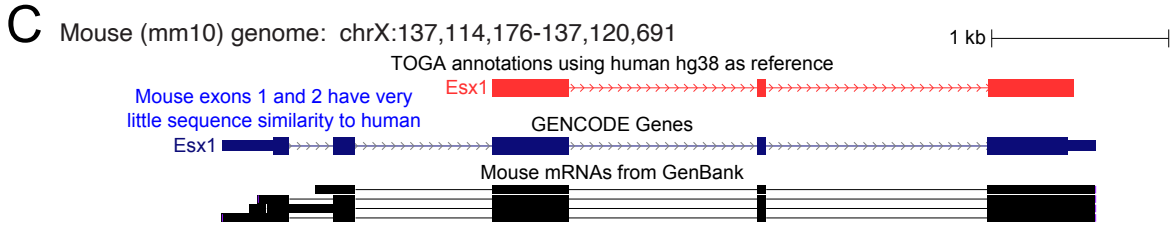


Fig. S19: Assembly base errors can lead to false gene losses.
The plot shows the conserved intron-exon structure of the *RRP8* gene in human and two distinct cow genome assemblies, published in 2014 (GCA_000003055.4, bosTau8) and 2018 (GCA_002263795.2, bosTau9). While TOGA classifies the *RRP8* in the newer bosTau9 assembly (bottom) as lost, because it exhibits two frameshifting mutations, *RRP8* encodes an intact reading frame in the older bosTau8 assembly. Thus, the frameshifting mutations that led TOGA to classify *RRP8* as lost in the bosTau9 assembly are most likely base errors.





D *ESX1* exhibits low sequence similarity between human-mouse (data from Ensembl)

Species	Gene ID	Pept. length	%identity(cDNA)
Human	ENSG00000123576	406 aa	37%
Mouse	ENSMUSG00000023443	382 aa	39%

E Muscle alignment shows high divergence between human and mouse ESX1, even at the protein level

```
ENSMUST00000074698      MEShKkcpccyctdlKTfvgavkeEtLqDpQpSLSStLleGaDyQENaeSEttFysdfga
ENST00000372588         MESlRgythsdi-gyRSlavgediEeVnDeKlTVTSlMarGgEdEENtrSKpeY-----g

ENSMUST00000074698      SpkEeeVemEnSLfNsmflEEaNyQepeGfEPsRgEaaaPVaEapqawngnenlgggfle
ENST00000372588         TeaEnnVgtEgSVpS----DDqDrEgggGhEPeQqQeepPLtK-----------------

ENSMUST00000074698      snaqlgeadaaPvRQslmrPLMQpVaQsspQPlPanpLQaPQqpEeqEeeEeeQPgEeqp
ENST00000372588         -----------PeQQqeepPLLE-LkQeqeEP-PqttVEgPQpaEgpQtaEgpQPpE---

ENSMUST00000074698      qqepKpRRyRicFTpiQLQELEaFFqrvQYPDLfARveLArRLgLpEpRVQVWFQNRRAK
ENST00000372588         ----RkRRrRtaFTqfQLQELEnFFdesQYPDVvARerLAaRLnLtEdRVQVWFQNRRAK

ENSMUST00000074698      WRRlRRaqafRNmvpvaMSpPVgVYLddhYgpiPiVEviwkcY-----------PMvPrP
ENST00000372588         WKRnQRvlmlRNtatadLAhPLdMFLggaYyaaPaLDpalcvHlvpqlprppvlPVpPmP

ENSMUST00000074698      mhPqMMPLPPRP-----------PPGfRM---PPpfRppPLPPF--------PWPPV----
ENST00000372588         prPpMVPMPPRPpiapmppmapvPPGsRMapvPPgpRmaPVPPWppmapvpPWPPMapvp

ENSMUST00000074698      -----------PPdAhIPnaareynpfPfPfpfPfPfpnPfpnpnPnPnpnPnPnpnP--
ENST00000372588         tgppmapvppgPPmArVPpgppmarvpPgPpmaPlPpgpPmaplpPgPpmaPlPpgpPma

ENSMUST00000074698      ------------------------NqnFAgPrYry
ENST00000372588         plpprshvphtglapvhitwapviNsyYAcPfF--
```

Fig. S20: The highly-diverged *ESX1* gene is misclassified as lost in mouse.
(A) UCSC genome browser screenshot of the *ESX1* gene locus in human. The orthologous mouse alignment chain indicates that exon 1 and parts of exon 2 and 4 are so diverged that no alignment at the nucleotide level can be detected, despite our sensitive alignment parameters.
(B) Exon visualization indicating the absence of exon 1 (red) and showing a frameshifting insertion in exon 4. TOGA classifies this gene as lost.
(C) UCSC genome browser screenshot of the mouse *Esx1* locus. While TOGA (using the DNA to codon aligner CESAR) can only detect three exons, GENCODE annotates a five exon gene, supported by mRNAs. However, exons 1 and 2 have very little sequence similarity and exon 2 differs in size, raising the question of whether both exons are truly orthologous.
(D/E) The high divergence between human and mouse ESX1 is shown by weak alignment identity in Ensembl Biomart and the protein sequence alignment (generated with Muscle).

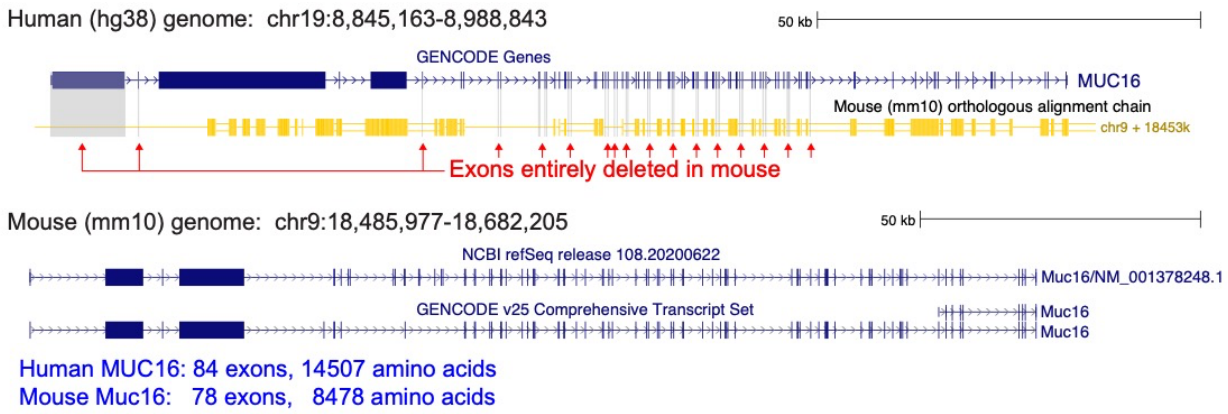


Fig. S21: Potential cases of mis-classified gene losses can be due to drastic exon-intron structure changes.

(A) UCSC genome browser screenshot of the human *MUC16* gene, which has 84 exons and encodes a 14,507 amino acid-long protein. There is a clear orthologous alignment chain to mouse. However, as shown by the grey boxes, several exons are deleted in mouse. Therefore, TOGA classifies this gene as lost in mouse.

(B) Orthologous locus in mouse. Both NCBI and GENCODE annotate an intact *Muc16* ortholog; however the annotated gene has only 78 exons and encodes a substantially shorter protein (8,478 amino acids, 58% of the human protein length and thus below our default threshold of 60% for the maximum percent intact reading frame).

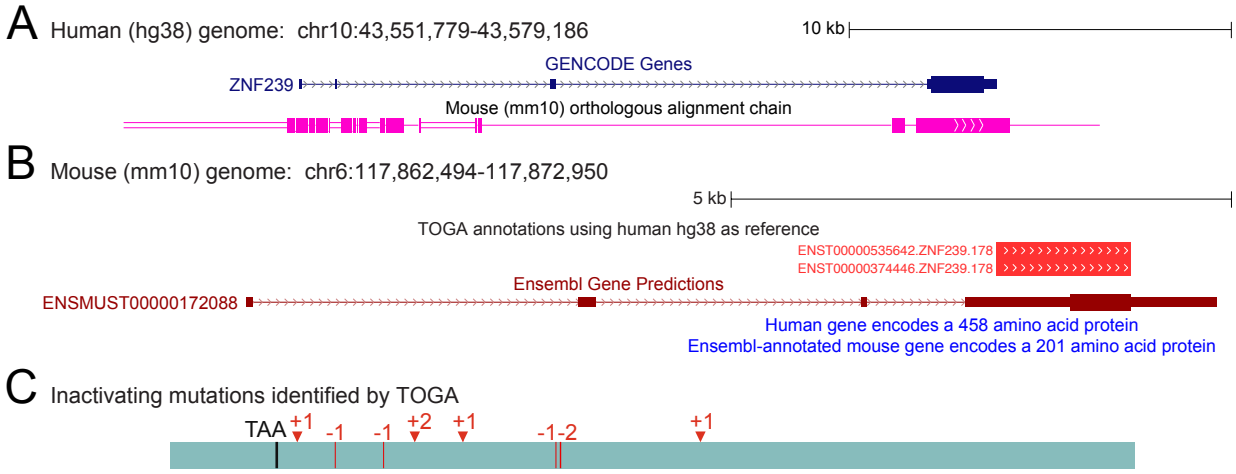


Fig. S22: Potential cases of misclassified gene losses can be due to drastic gene length changes. (A) UCSC genome browser screenshot of the *ZNF239* gene locus in human together with the orthologous mouse chain.
(B) While TOGA detects several inactivating mutations in the single codon exon of the *ZNF239* ortholog in mouse and classifies this gene as lost (indicated by the red color), Ensembl annotates a gene whose coding region starts downstream of the inactivating mutations (see also panel C). However, while the human gene encodes a 458 amino acid protein, the gene annotated by Ensembl is substantially shorter with only 201 amino acids. If the shorter mouse gene is functional, the drastic length change may be an indication that the function is not conserved between human and mouse.
(C) Inactivating mutations that TOGA detected in the upstream half of the gene.

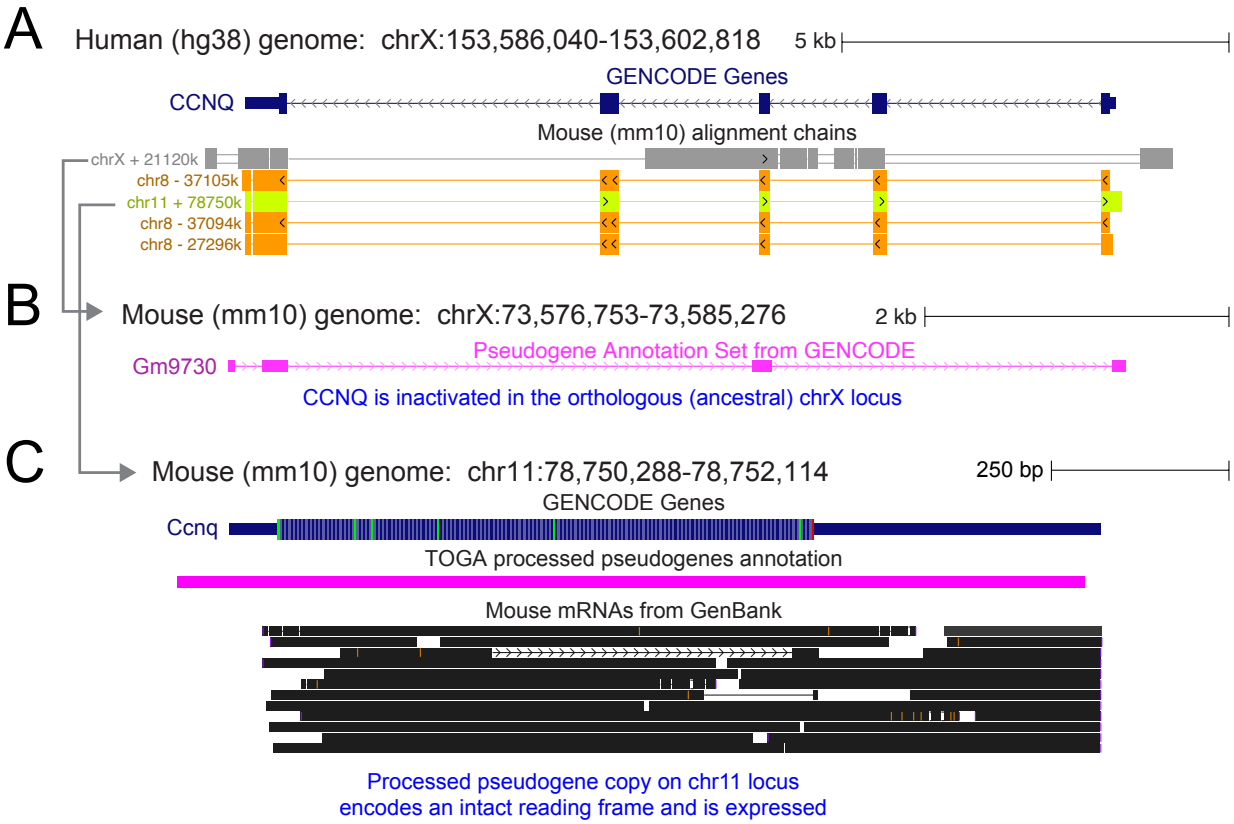


Fig. S23: A real gene loss is compensated by a processed pseudogene copy.
(A) UCSC genome browser screenshot of the human *CCNQ* gene, which is contained in our set of conserved genes that Ensembl annotates as 1:1 orthologs in mouse. The browser screenshot shows several mouse alignment chains. The chrX chain (grey) represents the orthologous locus, but shows that two coding exons are deleted. The other chains represent processed pseudogene copies.
(B) In the orthologous mouse locus, both GENCODE (Ensembl) and TOGA annotate an inactivated *CCNQ* ortholog (labeled Gm9730).
(C) TOGA annotates a processed pseudogene copy in the mouse chr11 locus, which is in agreement with the single-exon gene annotated by GENCODE. However, this processed pseudogene copy encodes an intact reading frame and GENCODE annotates this gene as the ortholog of human *CCNQ*. Mouse mRNAs further show that the gene is expressed, raising the possibility that the loss of *CCNQ* in the ancestral chrX locus is compensated by this processed pseudogene. Since processed pseudogenes lack the promoter and distal regulatory elements, it remains to be determined whether tissue expression and function is conserved.

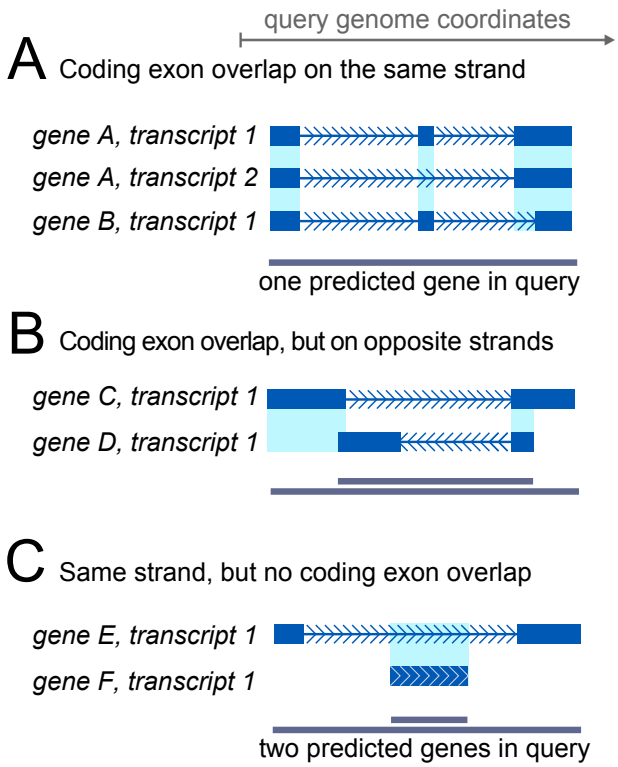


Fig. S24: TOGA infers predicted genes in the query by same-strand coding exon overlap. The three panels illustrate three different query loci, where more than one reference gene was annotated.
(A) The coding exons of transcripts from both genes A and B overlap on the same strand, thus TOGA predicts that there is a single gene in this query locus. If A and B have no other predicted orthologous loci in the query, TOGA infers a many:1 orthology relationship.
(B) Both transcripts have coding exon overlap but on the opposite strand. TOGA infers that two genes exist in the query locus (in antisense orientation).
(C) Both transcripts are projected to the same strand but coding exons do not overlap. TOGA infers that two (nested) genes exist in this query locus.

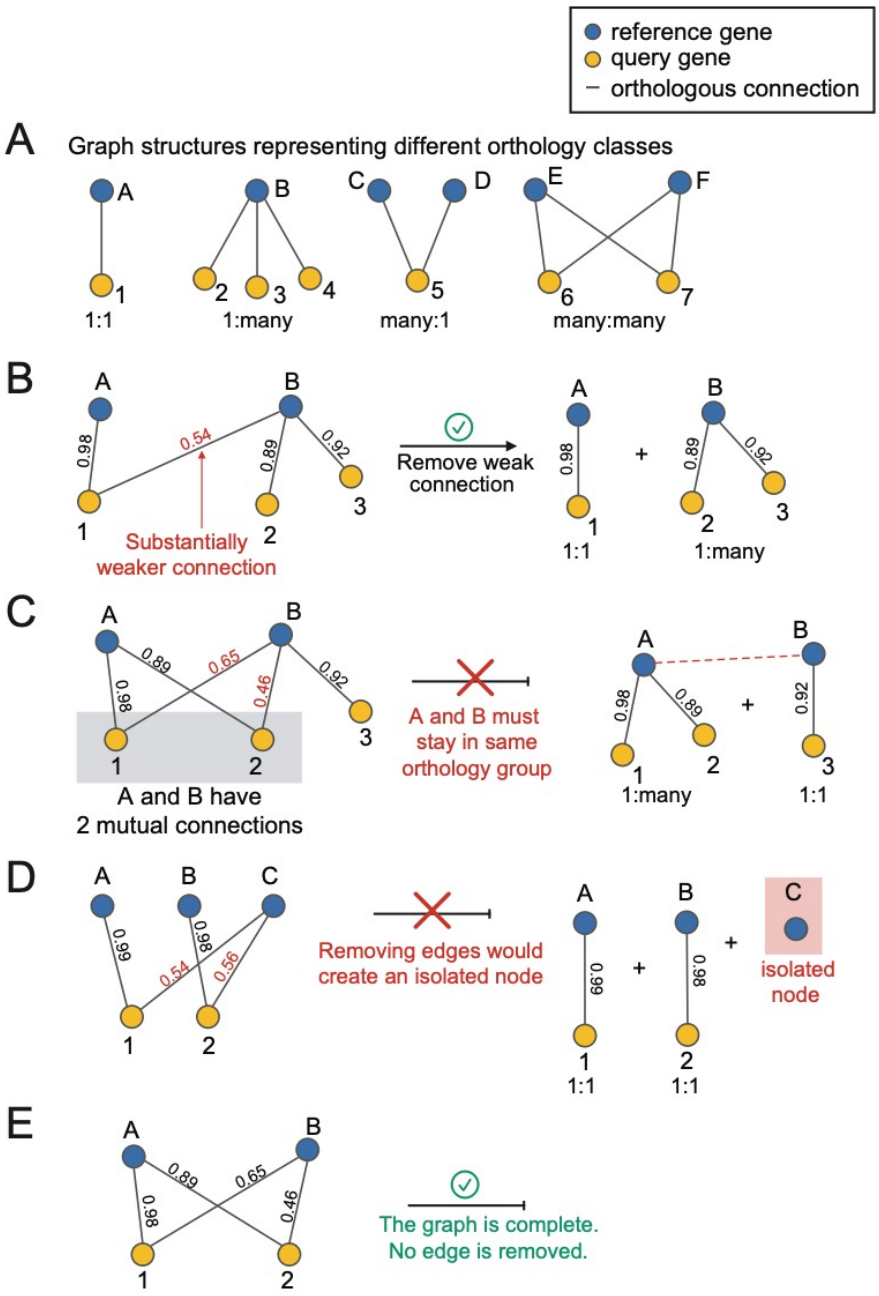


Fig. S25: Resolving weakly supported orthology relationships in a bipartite orthology graph.
(A) TOGA builds an orthology graph that represents genes of the reference (capital letters) and query (numbers) as nodes and orthology relationships as edges. Edges are weighted by the orthology score of the chain (shown in B-D). The panel illustrates graph structures that represent different orthology classes.
(B) The weakly supported B-1 edge is removed, resulting in two orthology groups with higher support: 1:1 (A-1) and 1:many (B-2-3).
(C) Since reference genes A and B have more than one mutual orthology connections, TOGA does not remove edges that result in separating A and B into different orthology groups.
(D) Removal of weakly supported edges would result in isolating gene C, which is not permitted.
(E) In a complete bipartite graph, TOGA does not remove any edges.

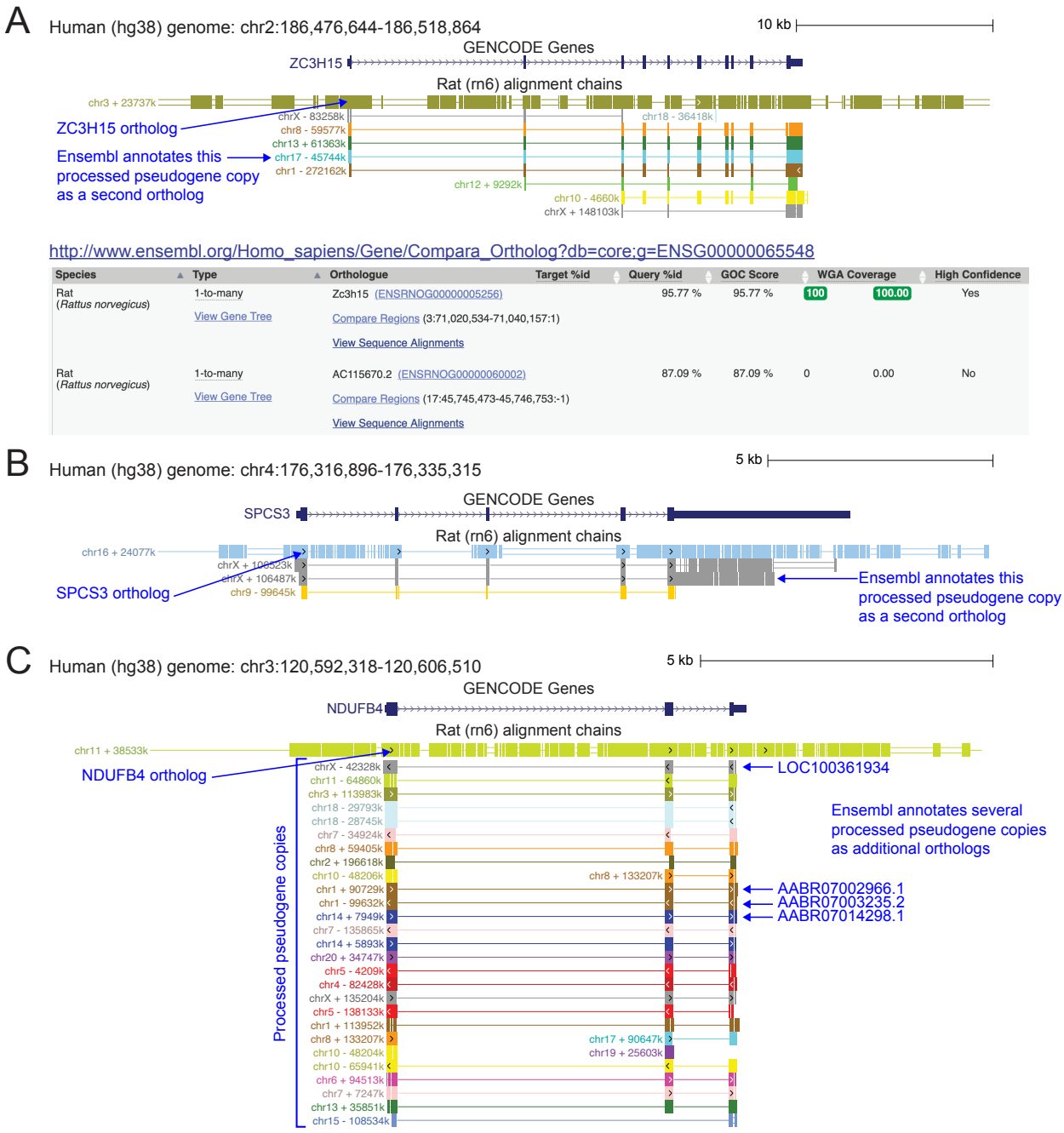


Fig. S26: Orthology type differences between Ensembl and TOGA.
Of the 1,312 genes, where both Ensembl and TOGA predict human-rat orthology but infer a different orthology type, 24.5% are cases where TOGA infers 1:1 and Ensembl 1:many. Manual inspection of these revealed several cases where Ensembl annotates processed pseudogene copies as orthologs. Panel A-C present three such examples.
(A) Human-rat alignment chains show a single orthologous locus for *ZC3H15* on rat chr3, which TOGA annotates as a 1:1 ortholog. The chains also show additional processed pseudogene copies that lack intronic sequences (single lines connecting chain blocks refer to "deletions" in the query species). Ensembl annotates the pseudogene located on chr17 as a second ortholog, resulting in a 1:many (1:2) orthology relationship.
(B) As in (A), alignment chains reveal a single orthologous locus for *SPCS3* on rat chr16 as well as additional processed pseudogene chains. While TOGA infers a 1:1 orthology

relationship, Ensembl annotates one of the chrX pseudogenes as an additional ortholog (1:many).
(C) TOGA infers a single *NDUFB4* ortholog (1:1), located on rat chr11. Ensembl additionally annotates four of the processed pseudogene copies as orthologs, resulting in a 1:many relationship.

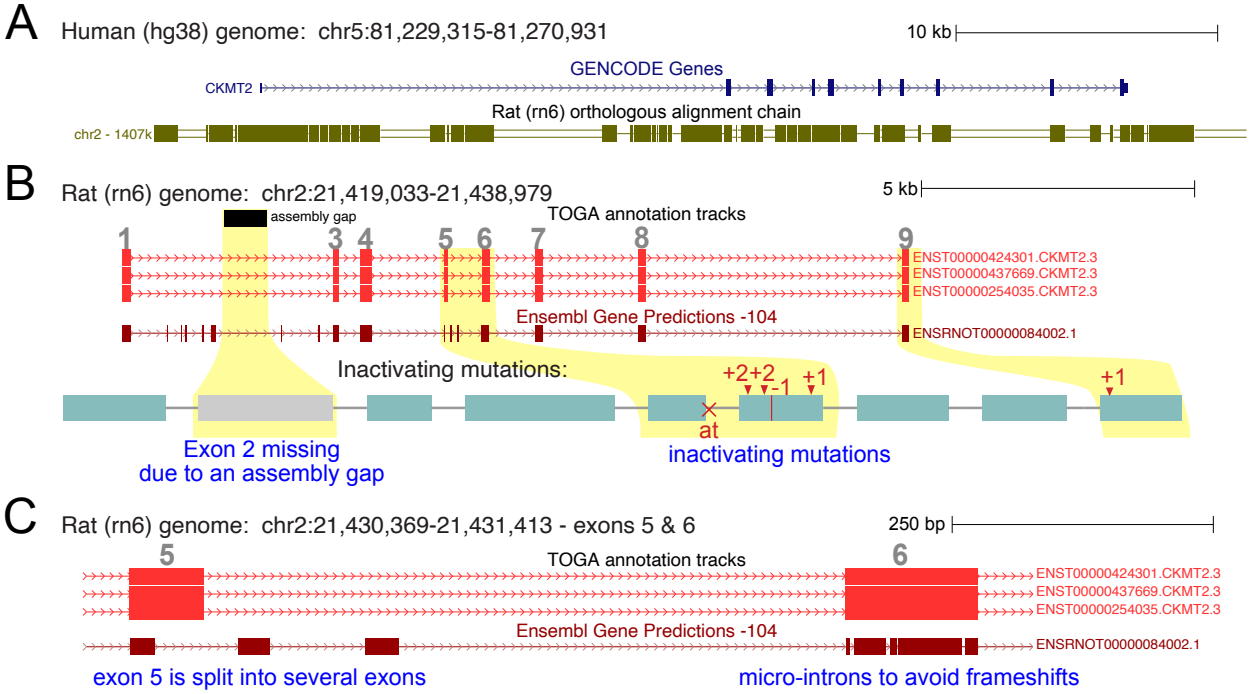


Fig. S27: Micro-introns explain why some inactivated genes are classified as orthologous by Ensembl.
(A) UCSC genome browser screenshot of the human *CKMT2* locus and the orthologous alignment chain to rat.
(B) At the orthologous rat locus, TOGA identifies numerous inactivating mutations and classifies the orthologous rat gene as lost (note that the missing exon 2 is not counted as an inactivating mutation). In contrast, Ensembl annotates a different exon intron structure for *CKMT2* that comprises many additional 'micro-introns', some being as short as two or four bases. These micro-introns attempt to avoid frameshift and splice site mutations in order to maintain a reading frame.
(C) Close up of the rat exon 5-6 region highlights the erroneous exon-intron structure and the micro-introns in exon 6.

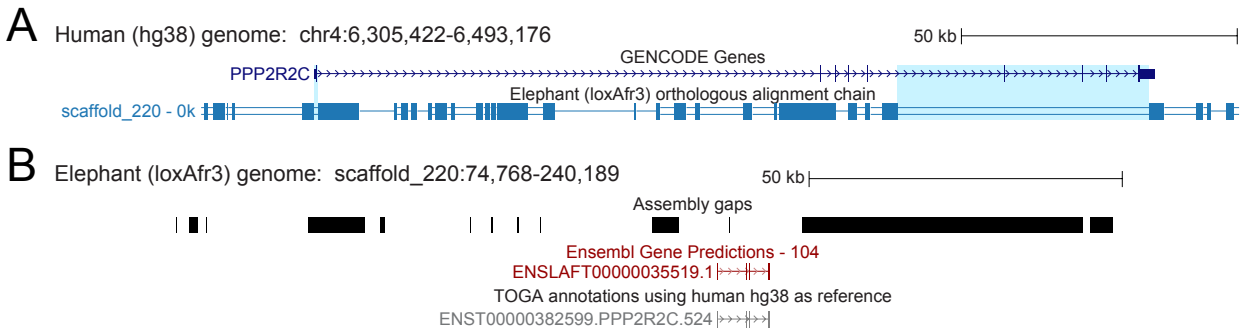


Fig. S28: Assembly incompleteness explains why some genes are classified as orthologous by Ensembl and as missing by TOGA.

(A) UCSC genome browser screenshot of the human *PPP2R2C* locus and the orthologous alignment chain to elephant. As shown by the blue highlights, large parts of the gene do not align to elephant.

(B) The orthologous elephant locus contains numerous small and large assembly gaps, explaining why the respective human gene parts (panel A) do not align. Both TOGA and Ensembl annotate the same highly-truncated gene, which encodes a 447 amino acid protein in human, but only a 188 amino acid protein in elephant. While Ensembl classifies it as a 1:1 ortholog, TOGA classifies this ortholog as missing in elephant, as more than 50% of the coding region overlaps assembly gaps.

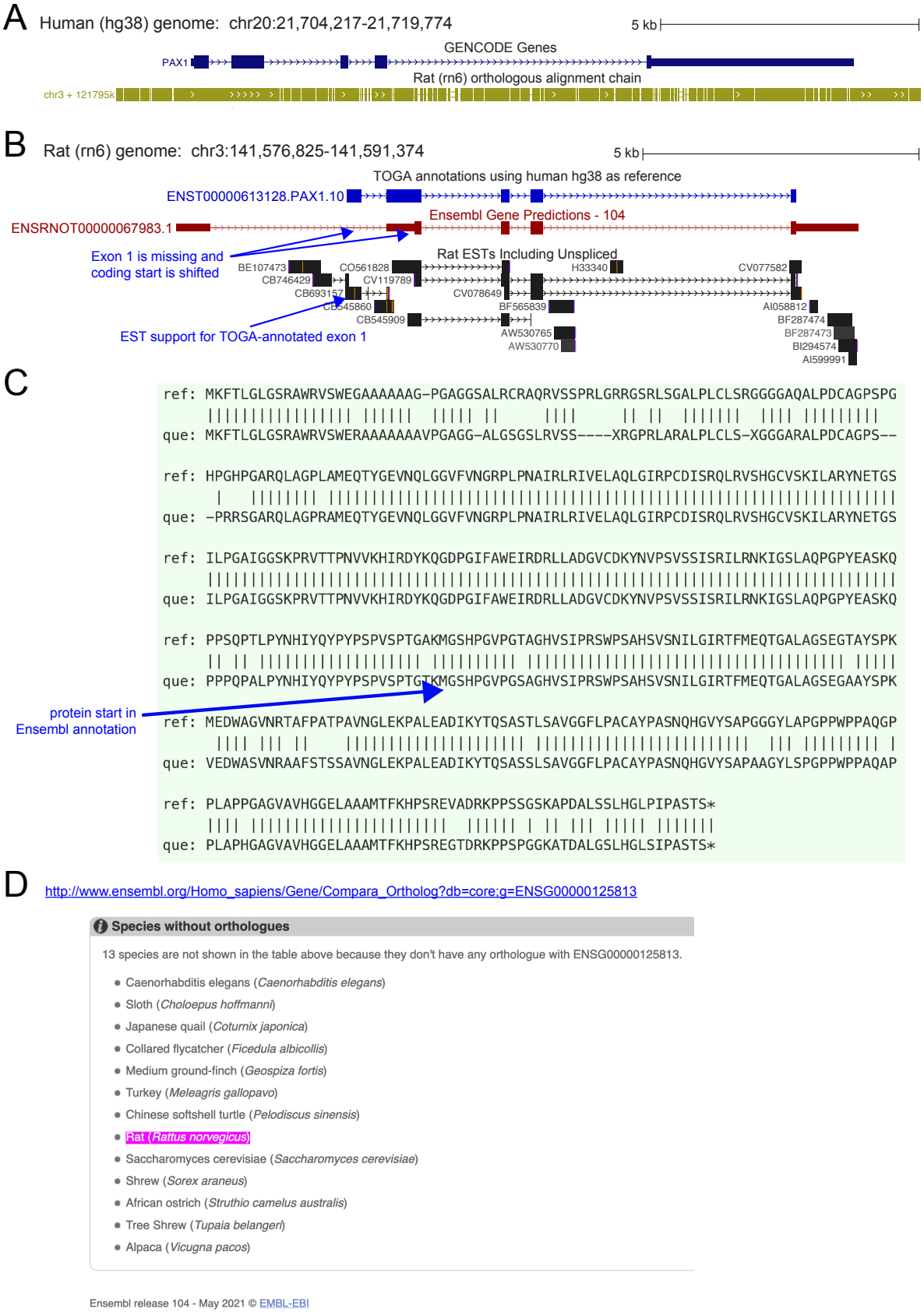


Fig. S29: TOGA but not Ensembl detects a human-rat ortholog of *PAX1*.
(A) UCSC genome browser screenshot of the locus encoding the human *PAX1* gene, encoding a key developmental transcription factor. As indicated by the alignment chain, there is an orthologous locus on rat chr3.

(B) While TOGA annotates all five coding exons of *PAX1*, the Ensembl annotation misses the coding exon 1 and shows a shifted coding start. ESTs support the location and splice site of the TOGA-annotated coding exon 1.
(C) Protein alignment of the TOGA annotated PAX1 protein (ref = human, que = rat) shows that the N-terminus is highly-conserved. The arrow indicates the start codon annotated by Ensembl.
(D) Ensembl Compara does not identify a rat ortholog for the human *PAX1* (ENSG00000125813), potentially because of the missing N-terminus.

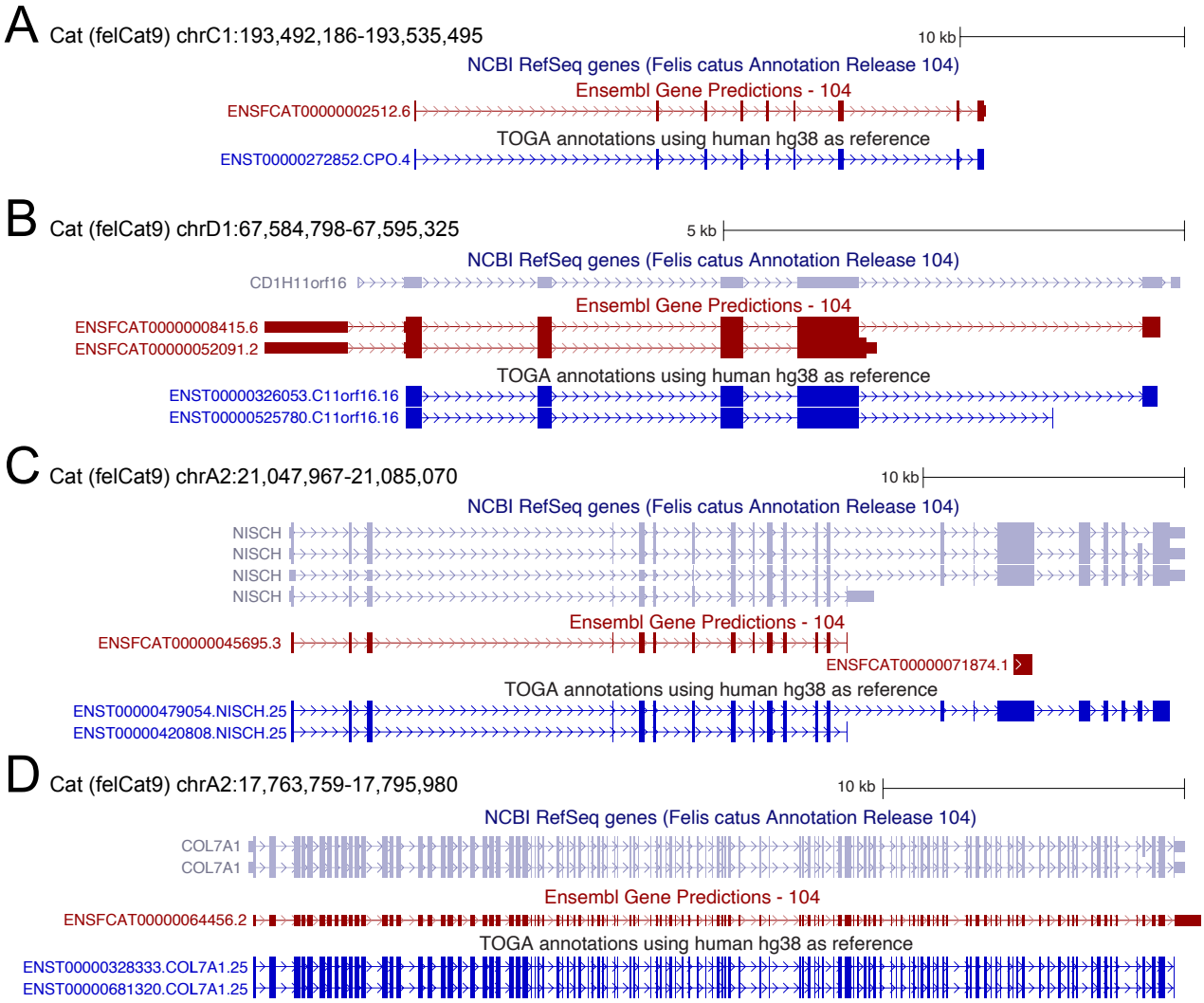


Fig. S30: BUSCO genes detected in cat (felCat9 assembly) by TOGA but missing in the cat annotation generated by NCBI or Ensembl.

(A) *CPO* is annotated by TOGA and Ensembl but NCBI does not annotate a gene in this locus.

(B) *C11orf16* is annotated as a coding gene by TOGA and Ensembl. NCBI annotates a similar exon-intron structure but labels the transcript as non-coding.

(C) A full length coding annotation of *NISCH* is provided by TOGA and NCBI. Ensembl annotates only the first part of the gene.

(D) *COL7A1* is annotated as a coding gene by TOGA and NCBI, while Ensembl annotates a non-coding transcript.

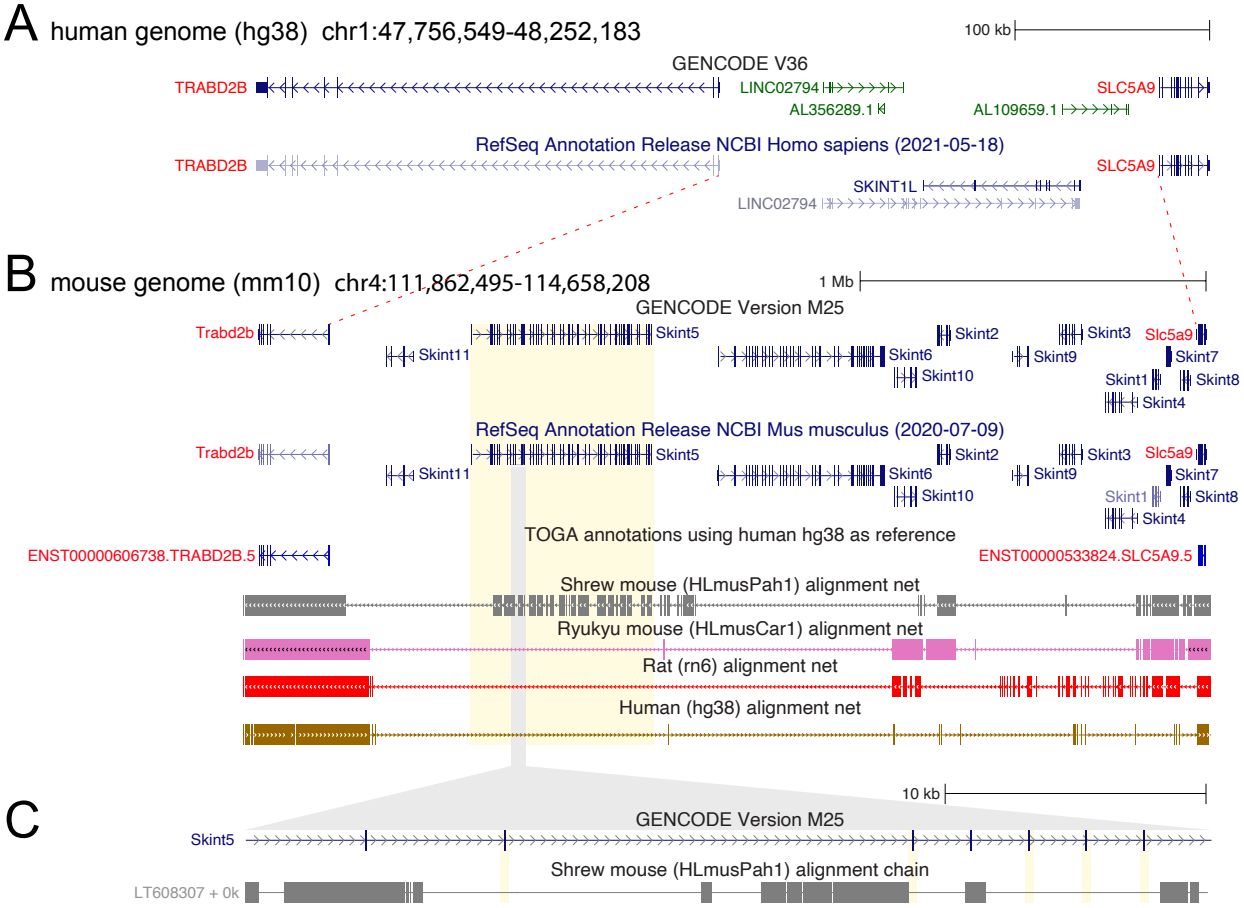


Fig. S31: Species-specific genes cannot be annotated with homology-based methods like TOGA.

Genome browser screenshots of the Skint gene locus in human (A) and house mouse (B,C). The *TRABD2B* and *SLC5A9* genes flank the Skint locus and are shown in red font. Skint genes are a rapidly evolving immunoglobulin gene family that are required for specific T cell subtypes in the thymus and skin (*65, 66*).

(A) The Human GENCODE and RefSeq gene annotation show that the Skint gene locus lacks a protein-coding Skint gene. The annotated *SKINT1L* is a Skint gene whose reading frame is inactivated by multiple stop codon mutations (*67*). TOGA with human as the reference detects the flanking *TRABD2B* and *SLC5A9* genes in the house mouse, but does not annotate a single Skint gene (B), because the reference (human) annotation lacks a protein-coding Skint gene.

(B,C) The corresponding house mouse (*Mus musculus*) genomic locus shows 11 Skint genes that are annotated by GENCODE and RefSeq (B). Genome alignments show that many Skint genes do not align to other rodents, exemplified for *Skint5* (yellow background). While *Skint5* aligns to the closely-related shrew mouse (*Mus pahari*), a magnified view shows that many of the repetitive *Skint5* exons are not present in the shrew mouse (C), indicating the *Skint5* is a *Mus musculus* specific gene. Indeed TOGA classifies the shrew mouse *Skint5* as 'uncertain loss' and does not find an intact *Skint5* gene in any other placental mammal. This is in agreement with Ensembl gene trees (www.ensembl.org/Mus_musculus/Gene/Compara_Tree?db=core;g=ENSMUSG00000078598) that infer *Skint5* to be a mouse-specific gene.

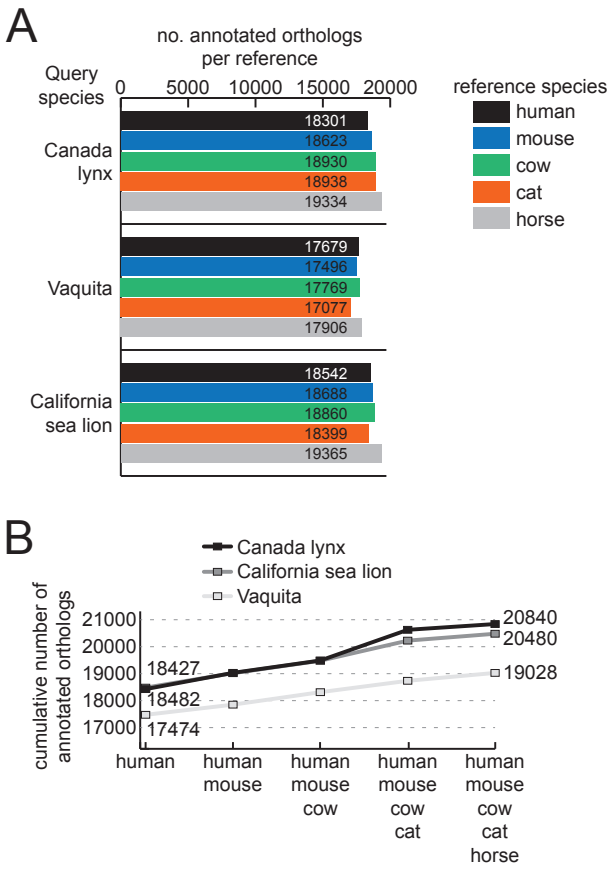


Fig. S32: Multiple references reduce reference bias and increase the number of annotated genes. We used three reference-quality assemblies generated by the Vertebrate Genomes Project (*26, 68*) as query species and applied TOGA with five different reference species using GENCODE (human, mouse) or NCBI RefSeq (cow, cat, horse) annotations (*21, 22*).
(A) Number of annotated orthologs with each individual reference.
(B) Cumulative number of annotated orthologs when adding multiple references.

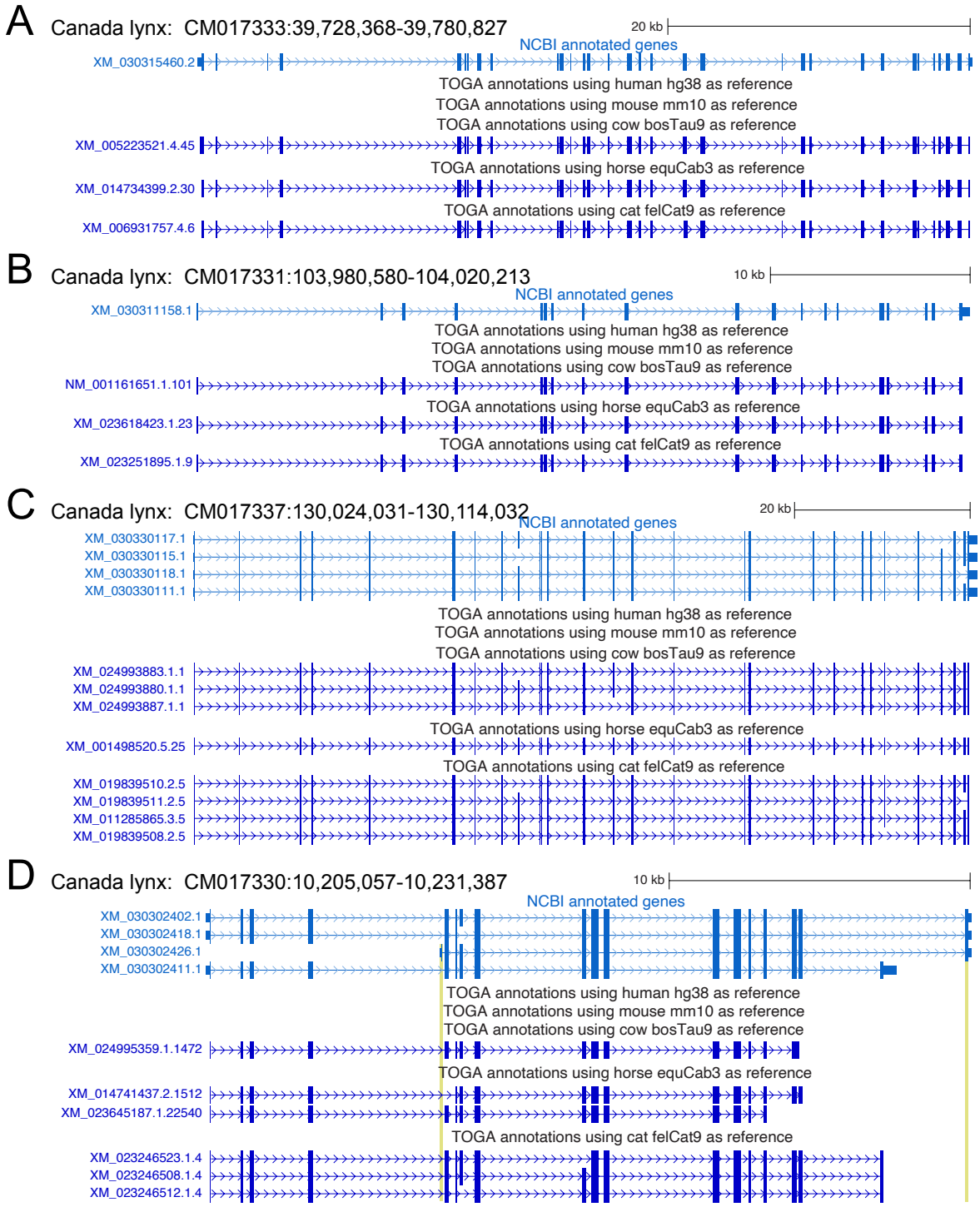


Fig. S33: Multiple references improve TOGA's gene annotation in the Canada lynx genome. UCSC genome browser screenshots showing the NCBI RefSeq annotation of the Canada lynx assembly (*26*) and TOGA annotations generated with different references.
(A-C) Examples of three larger multi-exon genes that TOGA did not detect with human or mouse as the reference, but that were annotated with references (cow, cat, horse) that are evolutionarily closer to the lynx.
(D) Example illustrating that TOGA with closer references annotates the gene, but misses putative lineage-specific alternative exons at the 5' and 3' end of the gene (highlighted in yellow).

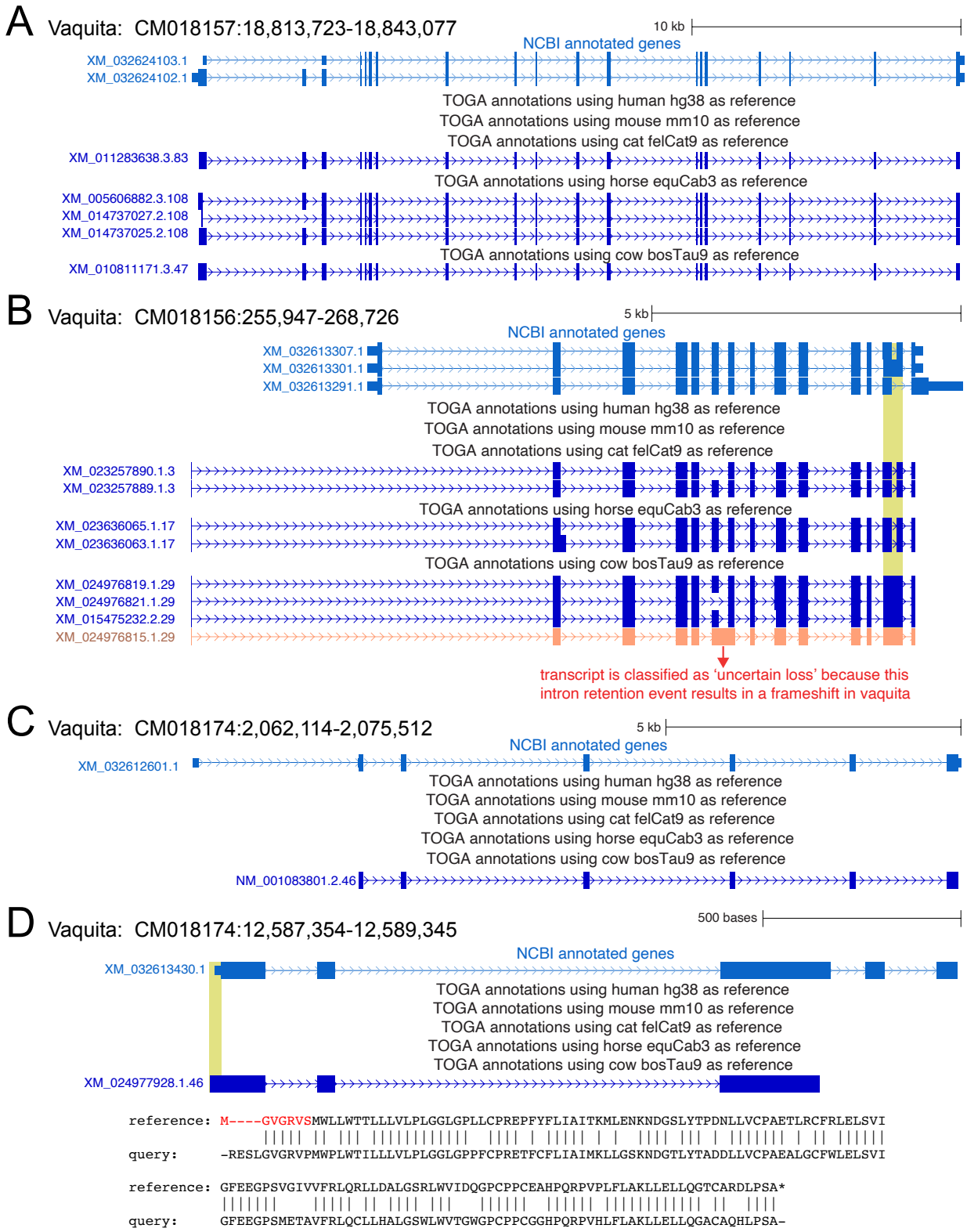


Fig. S34: Multiple references improve TOGA's gene annotation in the Vaquita genome. UCSC genome browser screenshots showing the NCBI RefSeq annotation of the Vaquita genome (*68*) and TOGA annotations generated with different references.
(A) A multi-exon gene that TOGA only annotates with reference species (cow, cat, horse) that are evolutionarily closer to the vaquita.
(B) TOGA with cat, horse and cow (but not human or mouse) as the reference annotates the gene, but predicts a different first coding exon. The last intron can be spliced or retained in

some NCBI-annotated transcripts (yellow highlight). Using cat/horse and cow as reference that provide both the intron splicing and intron retention splice variant, TOGA is able to annotate both transcripts in vaquita. The orange transcript shows a TOGA prediction classified as uncertain loss, because the retention of intron 6 results in a frameshift in vaquita.
(C) Example of a gene that TOGA only annotates with cow but not with any of the other four reference species. This example also illustrates that TOGA annotations only contain coding exons, which match the coding exons annotated by NCBI. To annotate untranslated regions and exons, transcriptomics data is required.
(D) Example of a partial gene annotation. Using cow as the reference, TOGA annotates a gene in this locus, but predicts a wrong N-terminus (yellow highlight) and misses the last two exons. The underlying reason is a difference in gene structure between cow and vaquita. The annotated cow gene has an extended N-terminus (7 amino acids longer, red font) and a stop codon in the third exon. Vaquita lacks this stop and has a longer C-terminus.

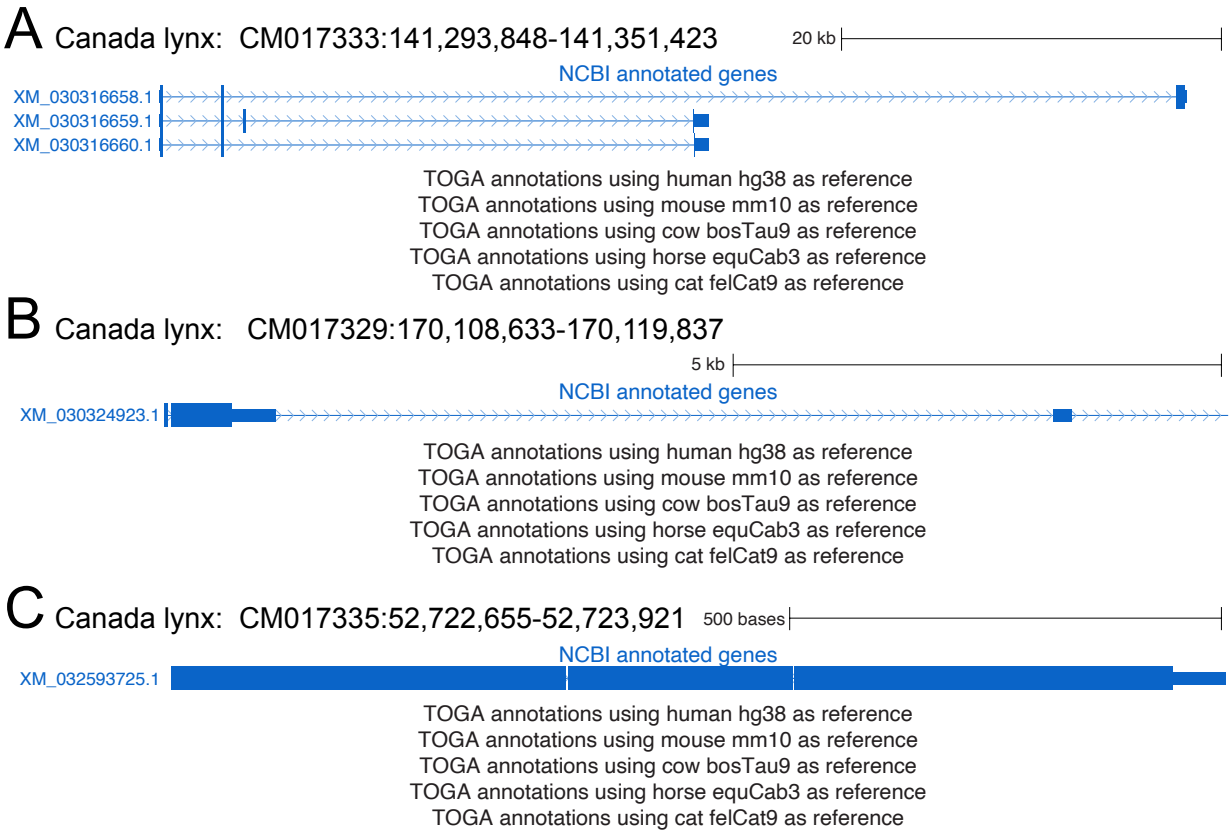


Fig. S35: Examples of genes not annotated by TOGA illustrate that transcriptomics data and *ab initio* gene predictions are required to comprehensively annotate lineage-restricted genes. (A-C) UCSC genome browser screenshots of three genes that NCBI annotated in the Canada lynx genome (*26*). TOGA misses these genes, even when using multiple and closely-related reference species.

It should be noted that the gene in (B) is a target for nonsense mediated decay as it has two introns in the 3' UTR and that NCBI corrected three frameshifts with micro-introns in the gene model shown in (C).

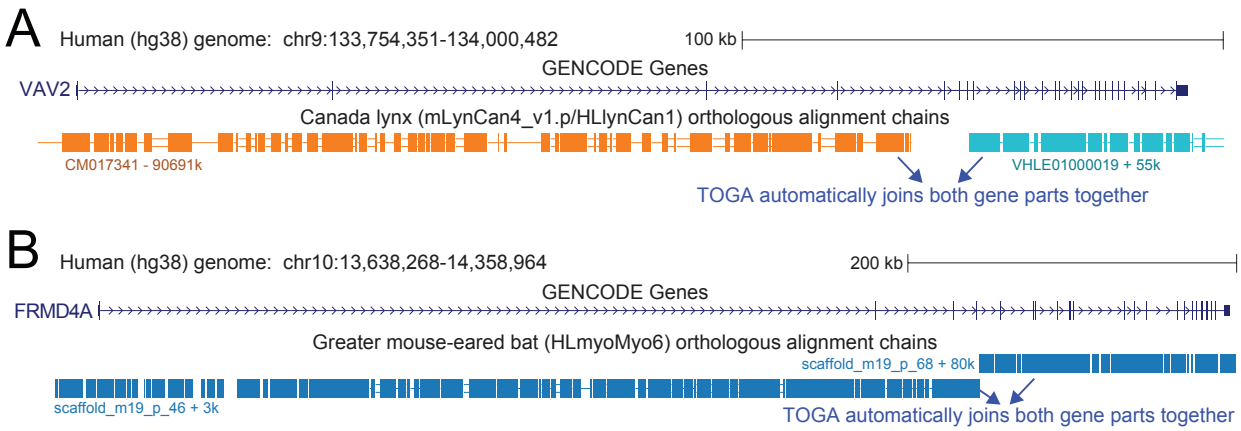


Fig. S36: TOGA detects and joins split genes also in chromosome-scale assemblies.
(A) UCSC genome browser screenshot showing that the ortholog of the human *VAV2* gene is fragmented in the highly contiguous (contig N50 of 7.5 Mb), chromosome-level Canada lynx genome assembly generated by the Vertebrate Genome Project (*26*).
(B) The ortholog of the human *FRMD4A* gene is fragmented in the greater mouse-eared bat (*Myotis myotis*), despite the fact that this assembly is highly contiguous (contig N50 of 11.8 Mb) and has chromosome-scale scaffolds (*6*).
In both cases, TOGA recognizes the two gene parts and joins them together.

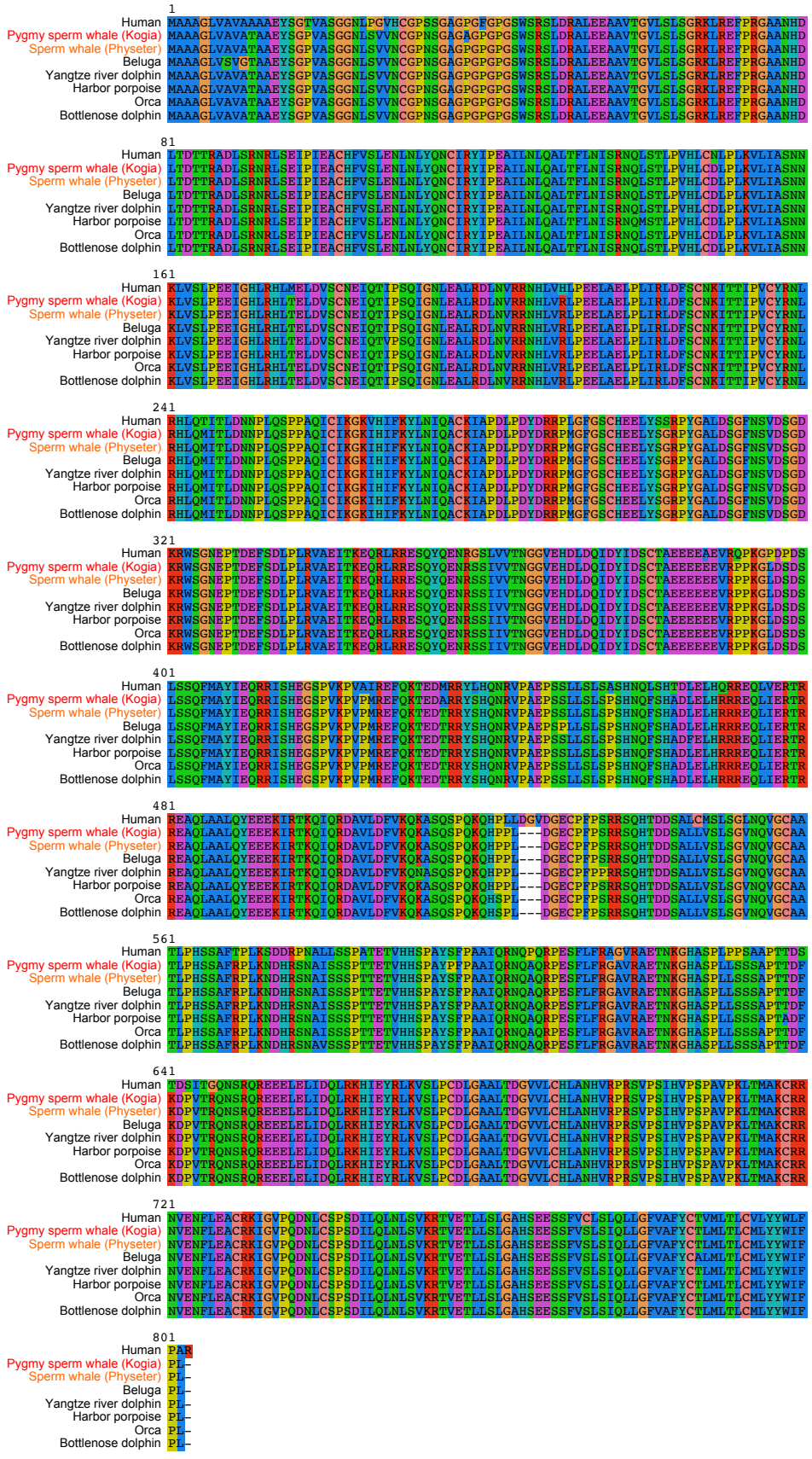


Fig. S37: Protein sequence alignment of the *Kogia breviceps* LRCH3 ortholog.
As shown in Fig. 4A (main text), this gene is split in six different scaffolds in the pygmy sperm whale (*Kogia breviceps*) assembly. TOGA recognizes and joins the six parts together, obtaining a full length protein. The sequence alignment with human (reference genome) and other

cetaceans shows a highly-conserved protein sequence, indicating that TOGA detected the correct orthologous genome fragments. In particular, the *Kogia* sequence is virtually identical to the sequence of the other sperm whale species (*Physeter macrocephalus*).

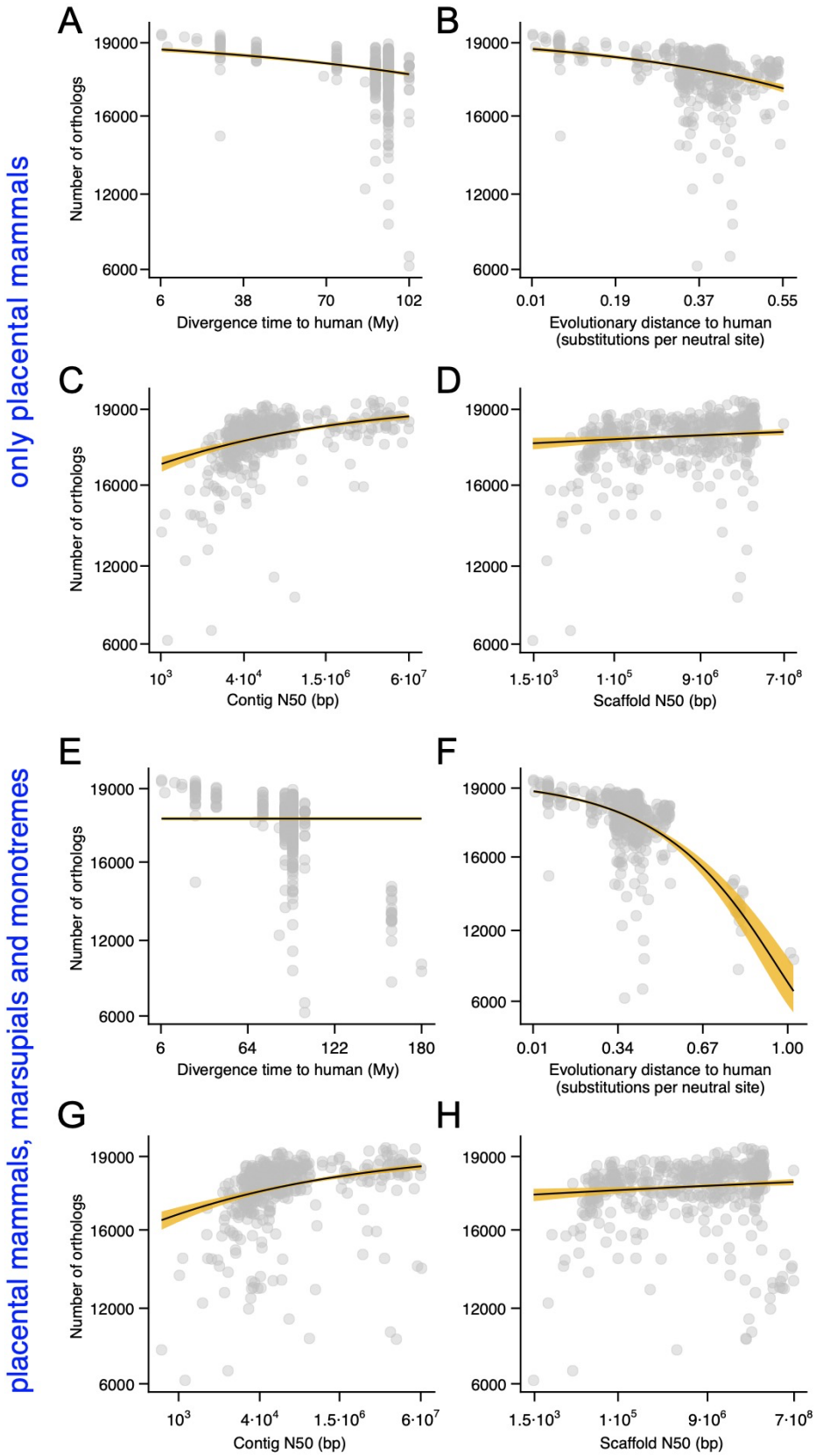


Fig. S38: The influence of evolutionary distance, divergence time and assembly metrics on the number of orthologs detected by TOGA using human as the reference.
In each panel (A-H), the variable on the X-axis is allowed to vary in the best fitting model, whereas the three remaining variables are kept fixed. Yellow areas represent the uncertainty (±1.96 standard errors) around the fit.

(A-D) Considering only placental mammals, all four predictor variables have a systematic (statistically significant) influence on the number of detected orthologs. The contig N50 value has the strongest impact, followed by the evolutionary distance to human, the divergence time to human, and the scaffold N50 value.
(E-H) Considering all mammals (placental mammals, marsupials and monotremes), the best fitting model according to AIC includes only three predictor variables: evolutionary distance to human (the most influential variable), contig and scaffold N50 value. In contrast, divergence time to human does not have a systematic effect in this model, because including marsupials and monotremes strengthens the correlation between evolutionary distance and divergence time to human from 0.82 to 0.90, making divergence time a redundant predictor. Nevertheless, a model that includes all four predictor variables is the second best fitting model with a ΔAIC of 3.44 (Table S12).

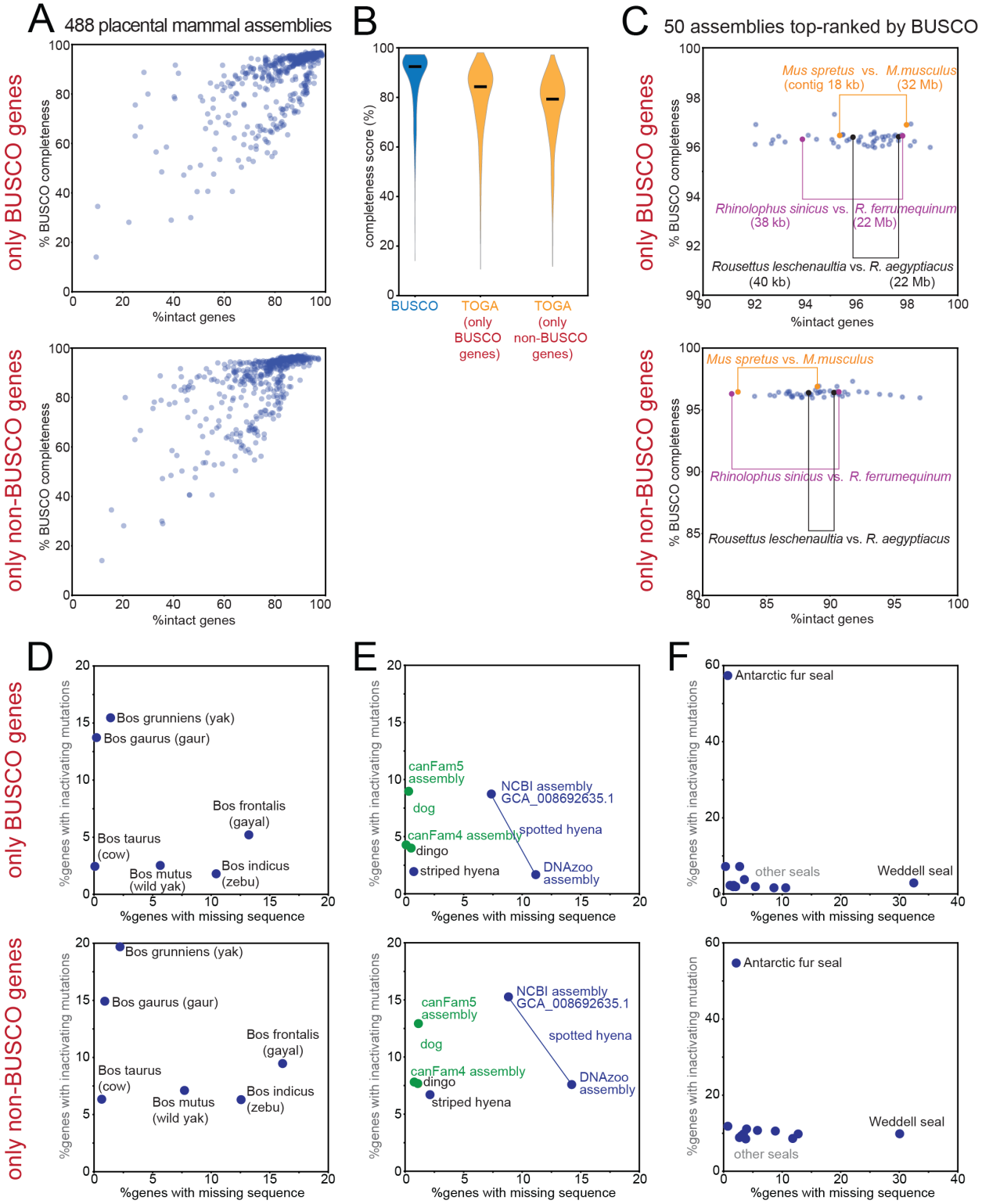


Fig. S39: TOGA provides a superior measure of mammalian genome quality, both with BUSCO and with non-BUSCO genes.

In Fig. 6, we compared BUSCO using the set of 9,226 mammalian odb10 genes with the TOGA status of 18,430 ancestral placental mammal genes. We found that TOGA provides a higher resolution to detect smaller differences in gene completeness of high-quality assemblies and that TOGA is able to distinguish between assembly incompleteness and base error rate, which represent two distinct assembly challenges. To investigate whether these results are driven by the TOGA methodology or the twofold increased gene number (9,226 vs. 18,430 genes), we

repeated the TOGA analyses, once using only the BUSCO genes and once using only non-BUSCO genes.
(A,B) Comparison of the percent complete BUSCO genes and TOGA's percent of intact genes for 488 placental mammal assemblies, shown as dot plots (A) and violin plots (B; horizontal black lines represent the median). TOGA's completeness values have a larger dynamic range for both BUSCO and non-BUSCO genes.
(C) BUSCO and TOGA completeness values for 50 assemblies that are top-ranked by BUSCO. Highlighted are three pairs of closely related species with different contig N50 values. TOGA's percent intact gene measure distinguishes low-contiguity assemblies from high-contiguity assemblies, both with BUSCO and with non-BUSCO genes.
(D-F) Dot plots compare the percent of genes with missing sequence and percent of genes with inactivating mutations. By definition, BUSCO genes are more conserved than non-BUSCO genes, which is reflected by non-BUSCO genes exhibiting inactivating mutations more frequently. Nevertheless, both using BUSCO and non-BUSCO genes, TOGA consistently highlights assemblies with a higher incompleteness or an elevated base error rate.

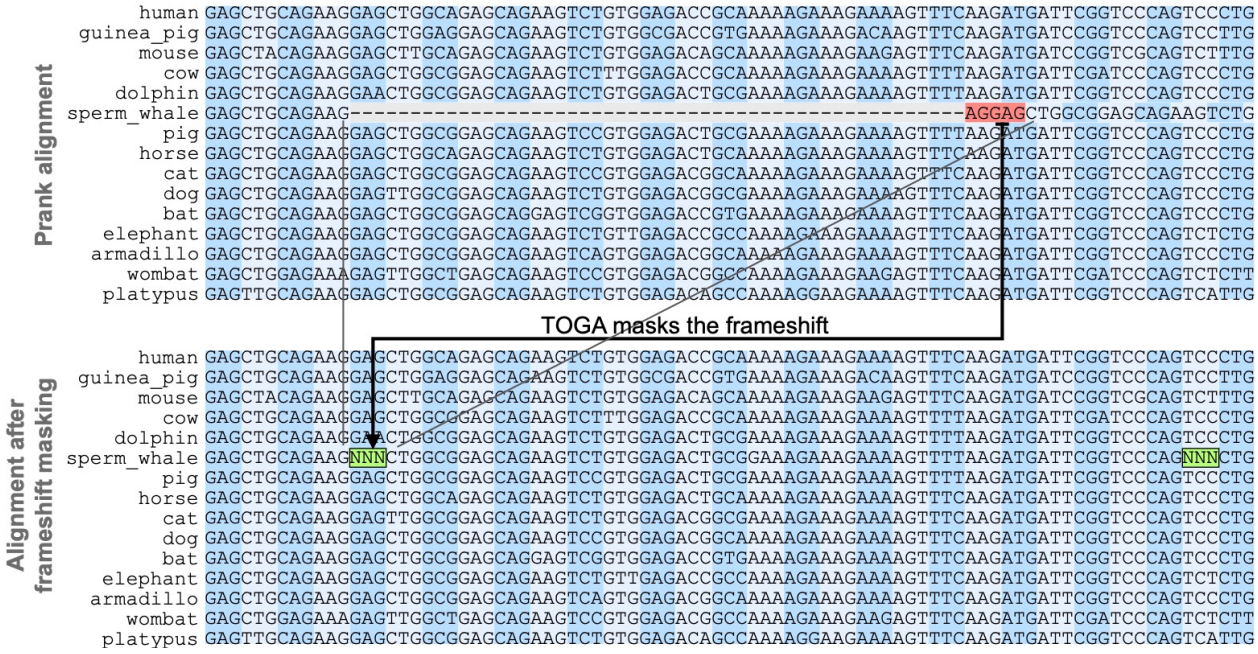


Fig. S40: TOGA enables more accurate codon alignments by masking inactivating mutations.
Top: Codon alignment of *AMPD3* generated by Prank (*69*). Red background highlights a frameshifting mutation in the sperm whale GCA_002837175.2 assembly.
Bottom: Codon alignment generated by Prank after masking this frameshift with TOGA (NNN). Grey lines connect the sequence up- and downstream of the frameshift in both alignments.

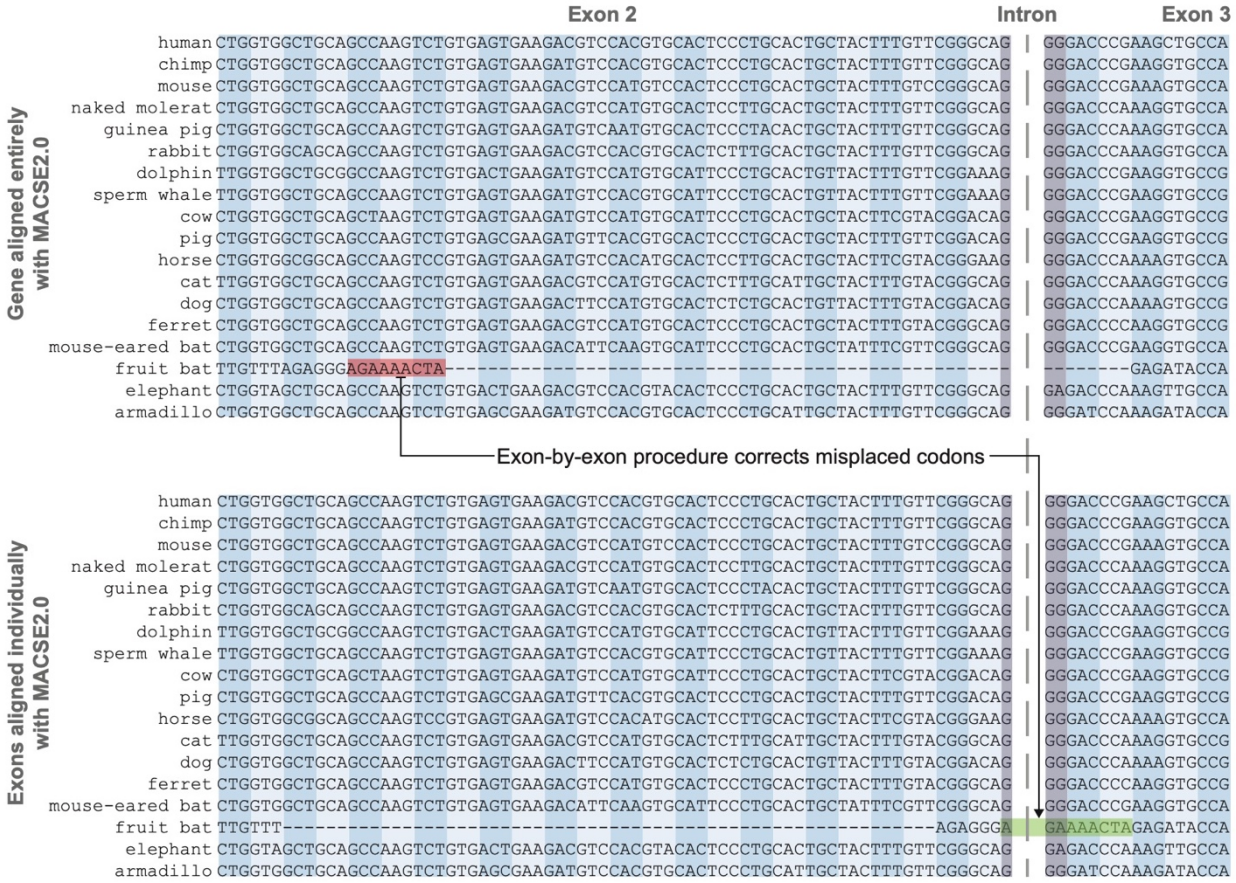


Fig. S41: TOGA enables an exon-by-exon alignment procedure that can avoid alignment errors in case of insertions or deletions that occurred at exon boundaries.
Top: Codon alignment of *ACOT8* generated by MACSE v2 (*33*). The Egyptian rousette has a larger deletion in exon 2. The red background highlights codons that belong to the Egyptian rousette exon 3, but are incorrectly aligned to the exon 2 sequence of other mammals when aligning the entire coding sequence.
Bottom: Applying MACSE v2 to individual exons avoids this misalignment.

Fig. S42: TOGA annotation track visualization in a UCSC genome browser mirror.
The figure shows a screenshot of our browser mirror (https://genome.senckenberg.de), visualizing the TOGA annotation with human as the reference. Clicking on a transcript annotation (here the orca *PITX1* ortholog) opens a page providing detailed information, a visualization of the exon-intron structure, and protein and exon alignments.

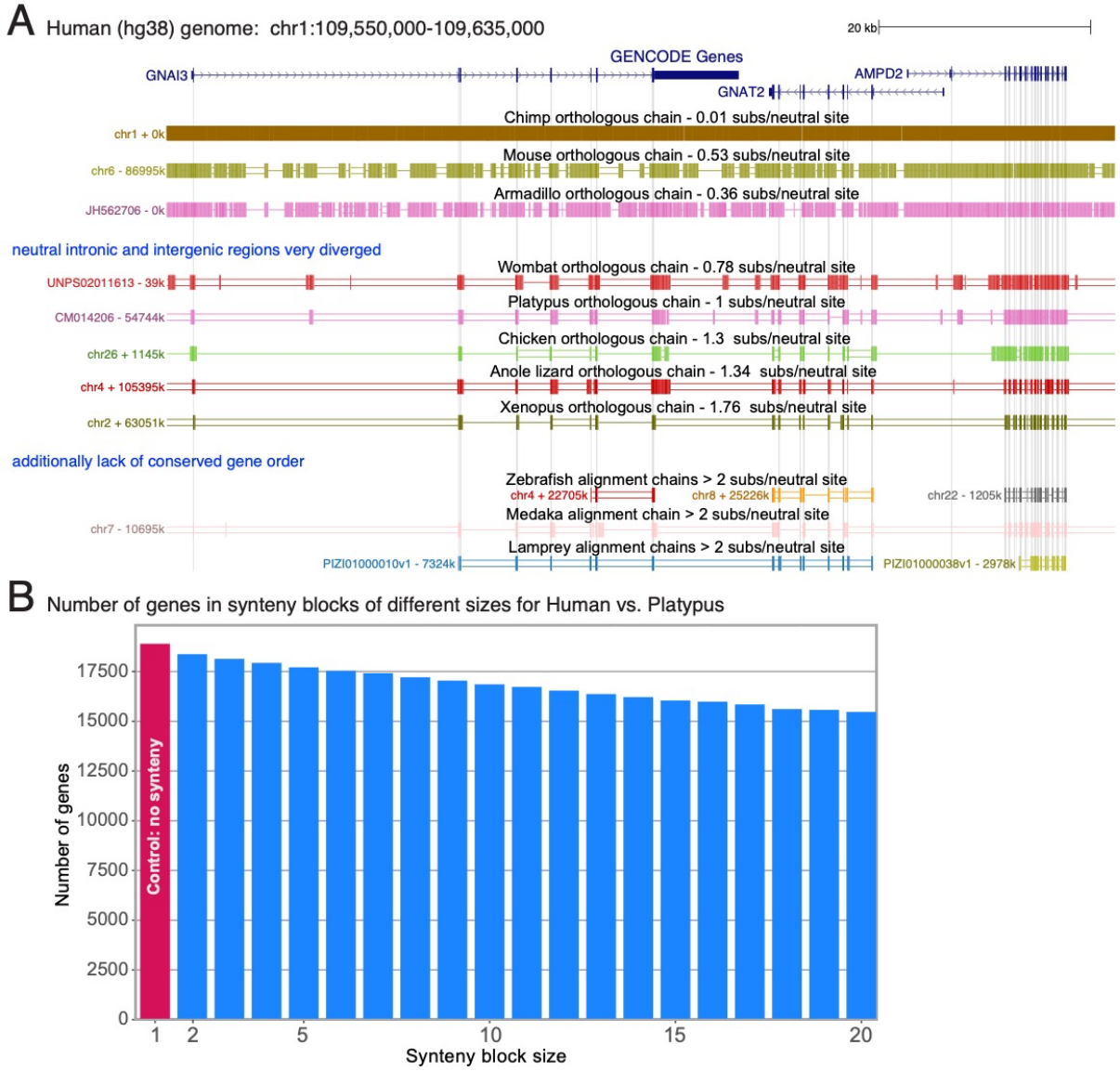


Fig. S43: Neutrally evolving regions diverge over greater evolutionary distances but conserved gene order is sometimes maintained.

(A) UCSC genome browser screenshot of three human genes and alignment chains to various vertebrates, ordered by the evolutionary distance to human (in substitutions per neutral site). For species exceeding an evolutionary distance of ~0.6 (marsupials, monotremes, reptiles etc), the divergence of intronic and intergenic regions is very high, making it hard for TOGA to use intronic/intergenic alignments to distinguish orthologous from non-orthologous genomic loci. For more distant species like fish or lamprey, conserved gene order (synteny) is also sometimes not preserved.

(B) More distant species can show a high extent of conserved gene order, which can be used for ortholog detection. For human vs. platypus, the bar chart shows how many genes occur in synteny blocks that have a minimum number of genes, ranging from at least 1 to at least 20. A synteny block size of 1 corresponds to all genes that align between both species. There are more than 17,500 genes that occur in synteny blocks with at least 5 genes and more than 15,000 genes in synteny blocks with at least 20 genes.

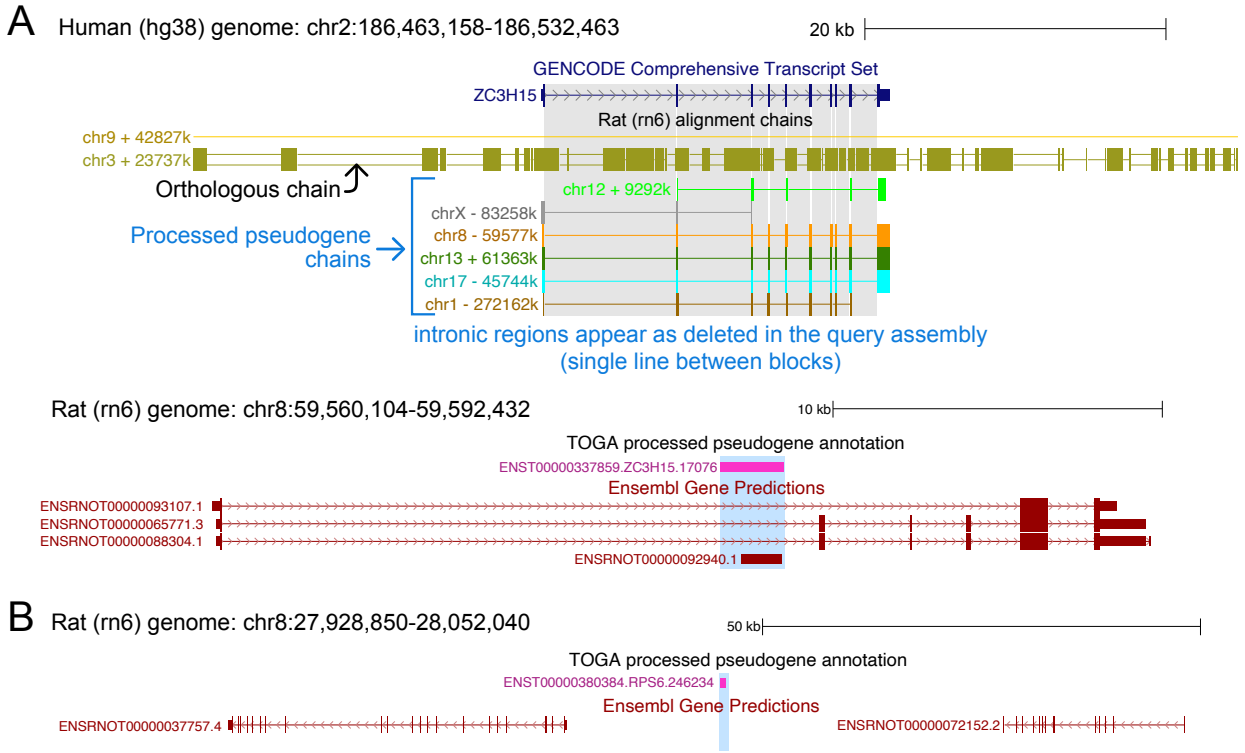
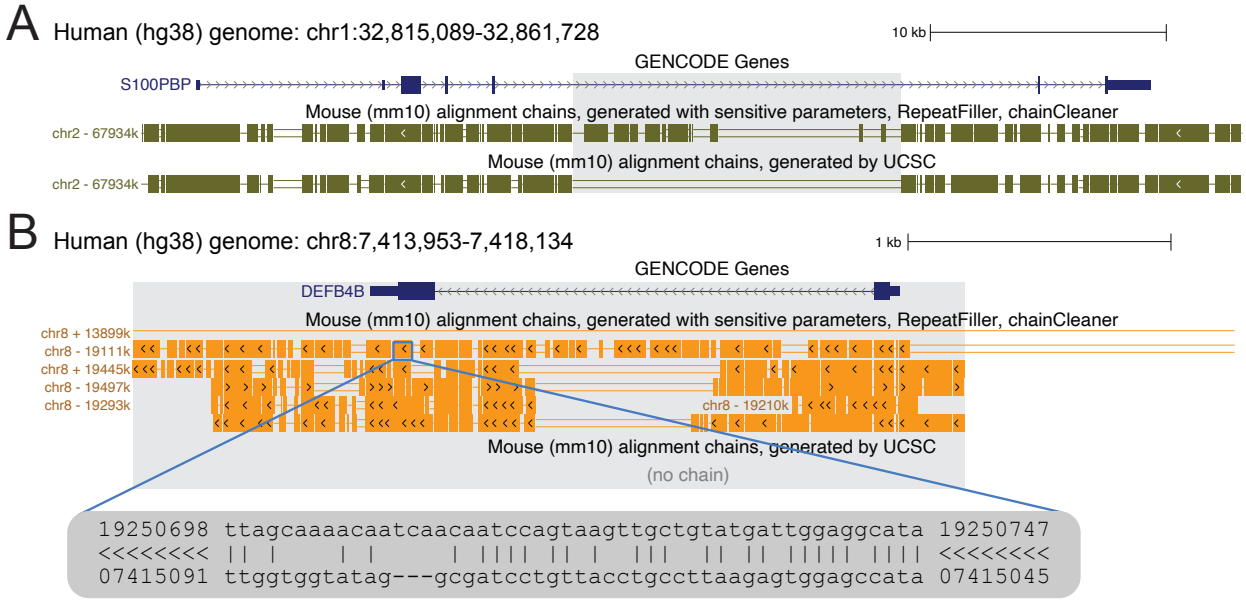


Fig. S44: TOGA detects chains aligning processed pseudogene copies and annotates these loci in the query genome.

(A) UCSC genome browser screenshot showing the human genome locus of the *ZC3H15* gene. The olive chain represents the orthologous locus in rat, while the other chains represent alignments of processed pseudogene copies in rat. Since processed pseudogenes are reverse-transcribed spliced mRNAs that have been inserted back into the genome, intronic regions appear as deleted in the query loci. Consequently, the span of the chain in query coordinates is similar to the summed size of the aligning blocks. TOGA uses this property to distinguish between paralogous and processed pseudogene chains. As one example, the rat locus corresponding to the chr8 pseudogene chain is visualized below, showing that TOGA annotates the locus as a processed pseudogene (pink). Ensembl also annotates a processed pseudogene in the same locus.

(B) TOGA annotates a processed *RPS6* pseudogene on chromosome 8 in the rat genome that is missing in the Ensembl annotation.



C Number of orthologous genes predicted for our more sensitive alignment chains and UCSC chains

Query species	sensitive chains	UCSC chains	Difference
Mouse (mm10)	18268	18045	223
Cow (bosTau9)	18779	18659	120
Dog (canFam3)	18368	18254	114

Fig. S45: Importance of the sensitivity of genome alignment chains.
(A) UCSC genome browser screenshot of the human *S100PBP* gene locus and the orthologous alignment chain to mouse. The chain at the top differs from the chain computed by UCSC (bottom) in more sensitive lastz parameters, and application of RepeatFiller and chainCleaner (*50, 51*). The grey box highlights a region where local alignments between intronic regions were only detected in the more sensitive chain.
(B) UCSC genome browser screenshot of the human *DEFB4B* locus showing that the more sensitive chains align several orthologous loci in mouse. These loci are missing in the chains computed by UCSC. The inset shows a local alignment of the more sensitive chains.
(C) Comparison of the number of orthologs that TOGA annotates for the same assemblies of mouse, cow and dog, using once the more sensitive chains and once the UCSC chains. With the more sensitive chains, TOGA is able to annotate more genes.

Captions for Tables S1-S15

Table S1: Mammalian species and assemblies, together with TOGA annotation statistics and genomic BUSCO completeness scores

Table S2: Evolutionary distances between human and other placental mammals (A) and between chicken and other birds (B)

Table S3: Test dataset for gradient boosting model evaluation

Table S4: Evaluation of the specificity of the gene loss detection pipeline integrated into TOGA

Table S5: Comparison of orthologs detected by TOGA and Ensembl

Table S6: Comparison between TOGA and Ensembl annotation completeness

Table S7: Comparison between TOGA and NCBI annotation completeness

Table S8: Mammalian BUSCO odb10 genes that are not present in human or mouse.

Table S9: Using TOGA as an additional source of gene evidence

Table S10: Relative length of genes, for which TOGA joined orthologous parts in fragmented assemblies

Table S11: Mammalian and bird assemblies generated by the Vertebrate Genomes Project for which TOGA annotations are available

Table S12: Generalized linear models to explore the influence of different variables on the number of detected orthologs

Table S13: Ancestral placental mammalian gene set

Table S14: Bird species and assemblies, together with number of TOGA-annotated orthologs

Table S15: TOGA for species other than mammals and birds