Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/9/eaaw6671/DC1

Supplementary Materials for

Genes lost during the transition from land to water in cetaceans highlight genomic changes associated with aquatic adaptations

Matthias Huelsmann, Nikolai Hecker, Mark S. Springer, John Gatesy, Virag Sharma, Michael Hiller*

*Corresponding author. Email: hiller@mpi-cbg.de

Published 25 September 2019, *Sci. Adv.* **5**, eaaw6671 (2019)
DOI: 10.1126/sciadv.aaw6671

The PDF file includes:

References (*71–84*)

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/5/9/eaaw6671/DC1)

Table S1 (Microsoft Excel format). Species and genome assemblies used in this study.
Table S2 (Microsoft Excel format). Genes lost in the cetacean stem lineage.
Table S3 (Microsoft Excel format). Validation of all smaller inactivating mutations with raw DNA sequencing reads.
Table S4 (Microsoft Excel format). Analysis of relaxed selection.
Table S5 (Microsoft Excel format). Convergently inactivated genes between (semi-) aquatic mammalian clades, represented by killer whale, manatee, and Pacific walrus.

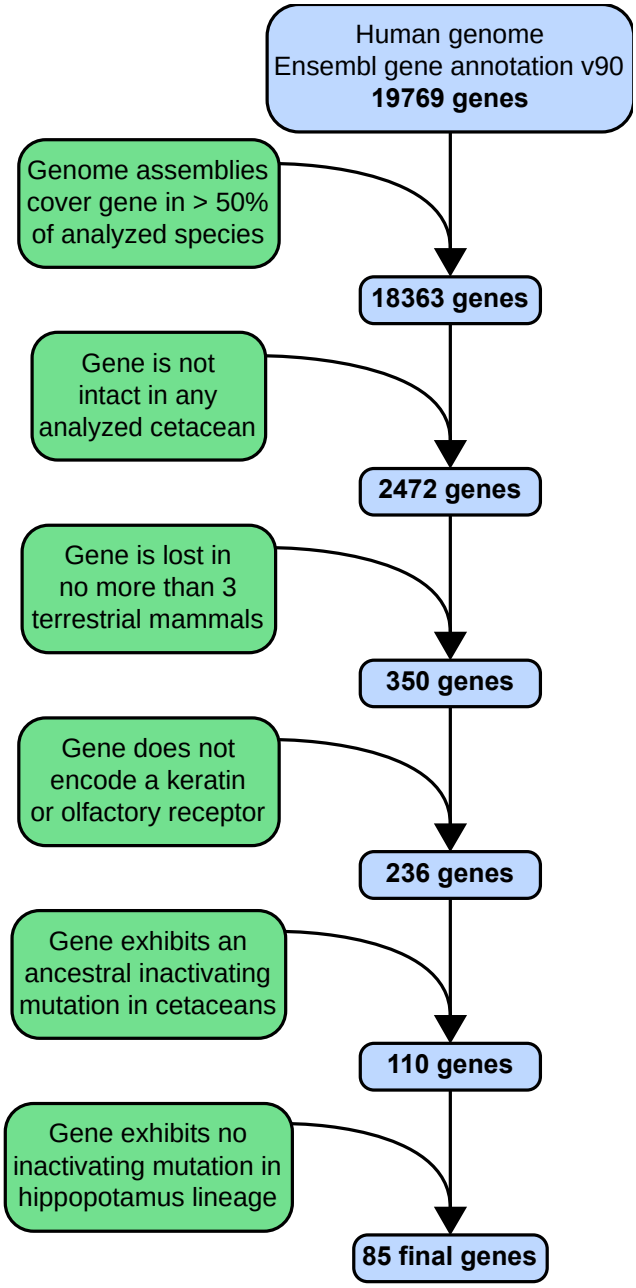


Fig. S1. Workflow for identifying gene losses that are lost in the cetacean stem lineage. All steps are described in detail in the Methods.

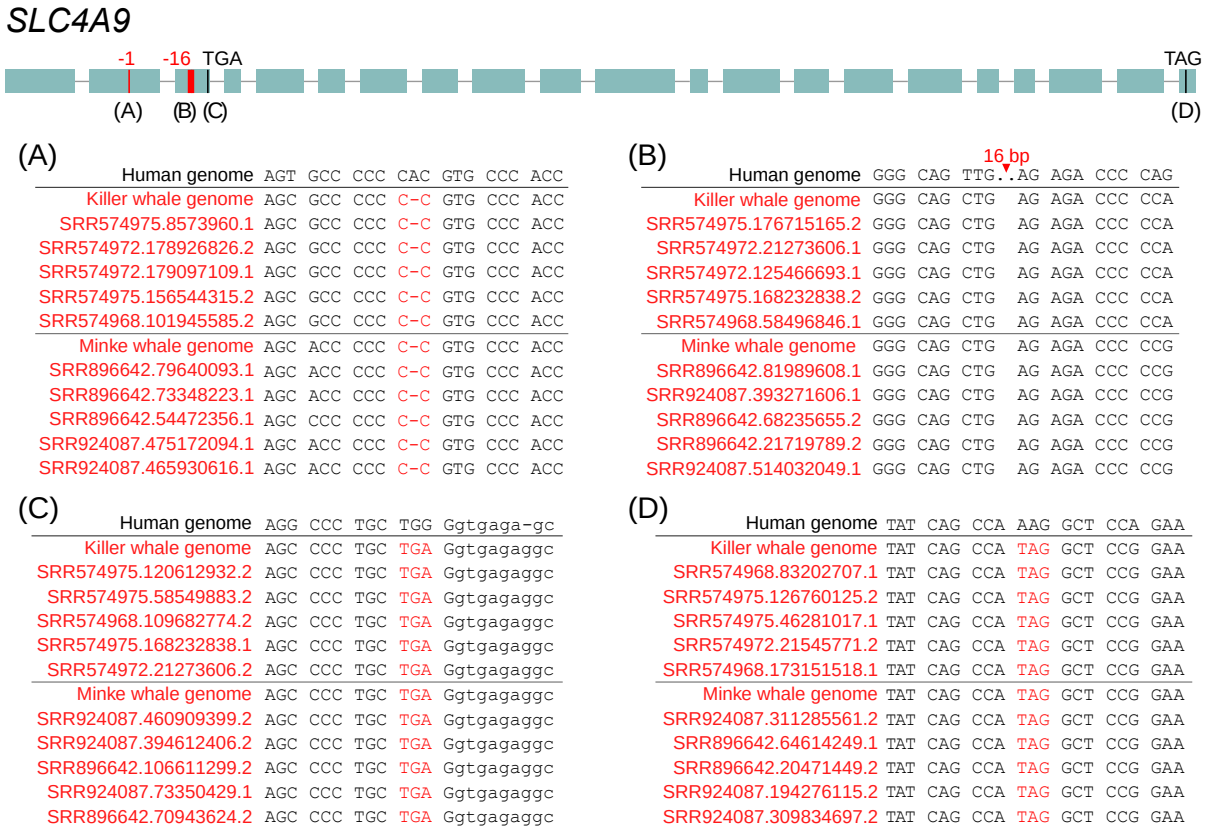
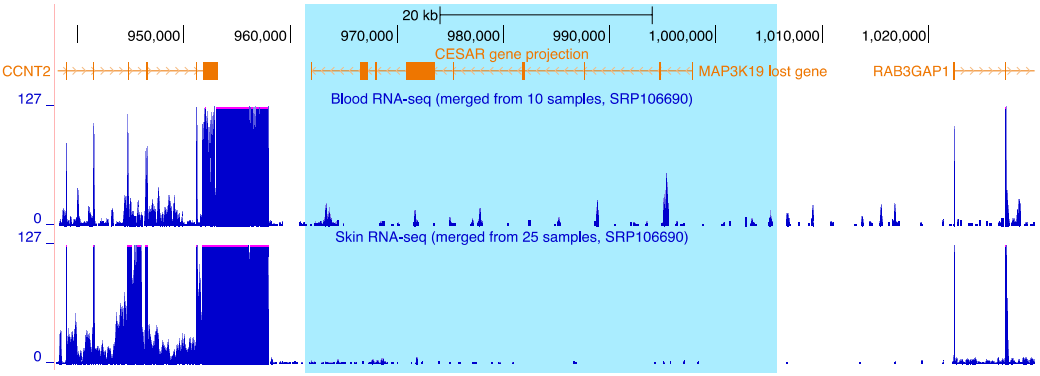


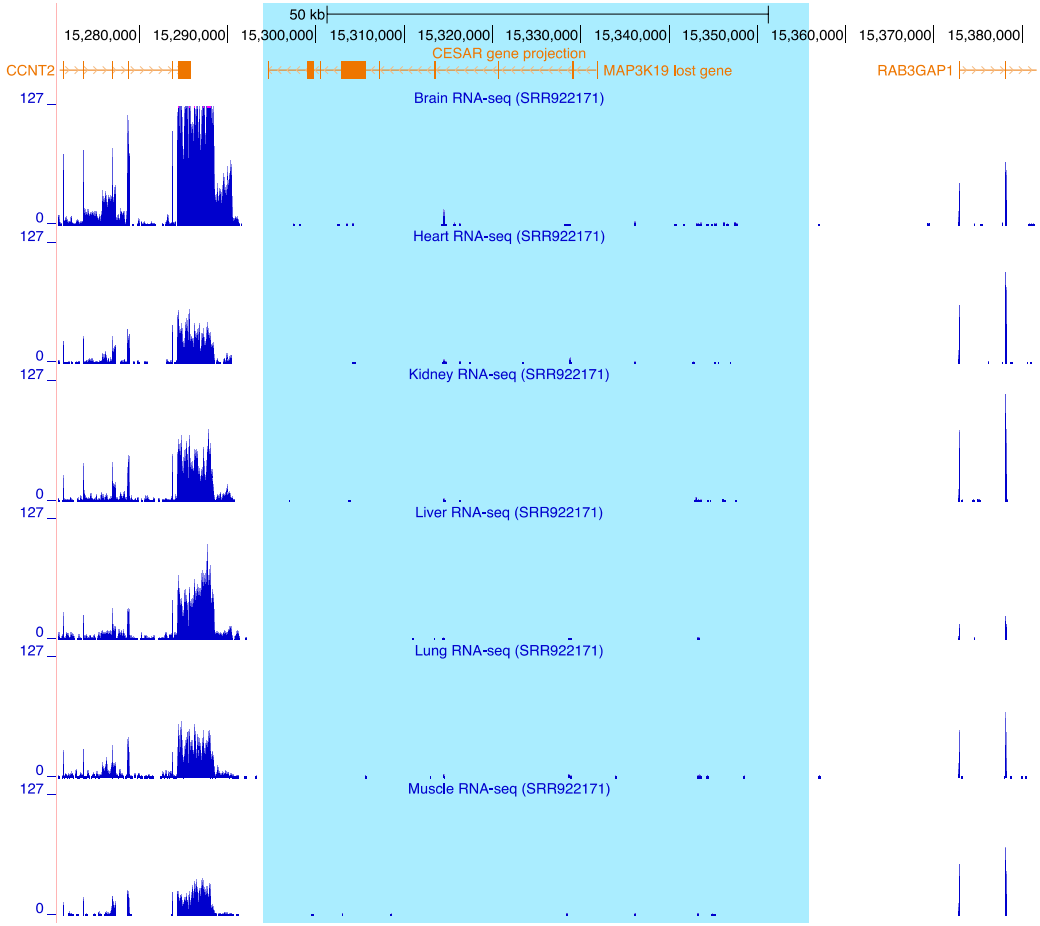
Fig. S2. Raw DNA sequencing read validation of shared inactivating mutations. While individual bases in a single genome assembly may be erroneous due to sequencing or assembly errors, the presence of inactivating mutations shared between several independently sequenced and assembled genomes makes such errors extremely unlikely. To verify this, we used unassembled DNA sequencing reads of the killer whale (an odontocete) and the minke whale (a mysticete) to validate the correctness of all four shared inactivating mutations in *SLC4A9*. To this end, we extracted the genomic sequence up- and downstream of each mutation and aligned these sequences to reads stored in the NCBI sequence read archive (*71*). As shown in panels A-D, each of the four inactivating mutations is supported by multiple sequencing reads for both species. NCBI sequence read archive identifiers are: SRX188930 (killer whale), SRX302112 and SRX316738 (minke whale).

Table S3 lists the results of validating all 251 (shared or lineage-specific) inactivating mutations in four cetaceans and the manatee, of which we could confirm 248 (98.8%).

A Bottlenose dolphin (turTru3 assembly) MRVK01001353:937,998-1,029,986



B Minke whale (balAcu1 assembly) KI538308:15,270,835-15,381,654



C Bottlenose dolphin (turTru3 assembly) MRVK01001161:29,174-60,756

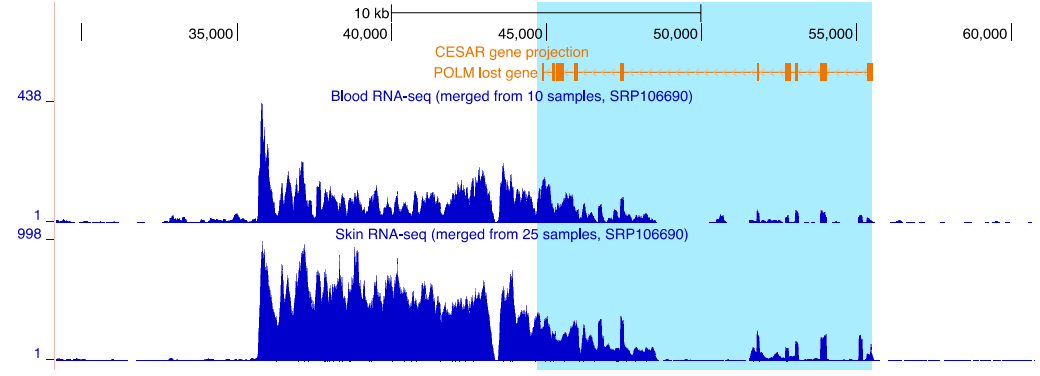


Fig. S3. Expression of the remnants of genes lost in cetaceans. We analyzed available expression data of different tissues of the Bottlenose dolphin and Minke whale. UCSC genome browser (*72*) screenshots show the RNA-seq read coverage in blue and genes projected by CESAR (*69, 73*) in orange. Most of the genes do not appear to have significant expression levels anymore, exemplified by *MAP3K19*, where available RNA-seq data shows no expression of the exons in dolphin (**A**) and Minke whale (**B**). Only for two genes (*POLM* and *F12*), we found evidence of partial gene expression, exemplified for dolphin *POLM* (C), where the upstream gene part may still be correctly spliced in the dolphin. The first exons of *POLM* also show expression in the Minke whale, albeit at an even lower level. However, the mapped read data indicates that both *POLM* and *F12* do not produce full-length and properly spliced transcripts anymore in dolphin and Minke whale.

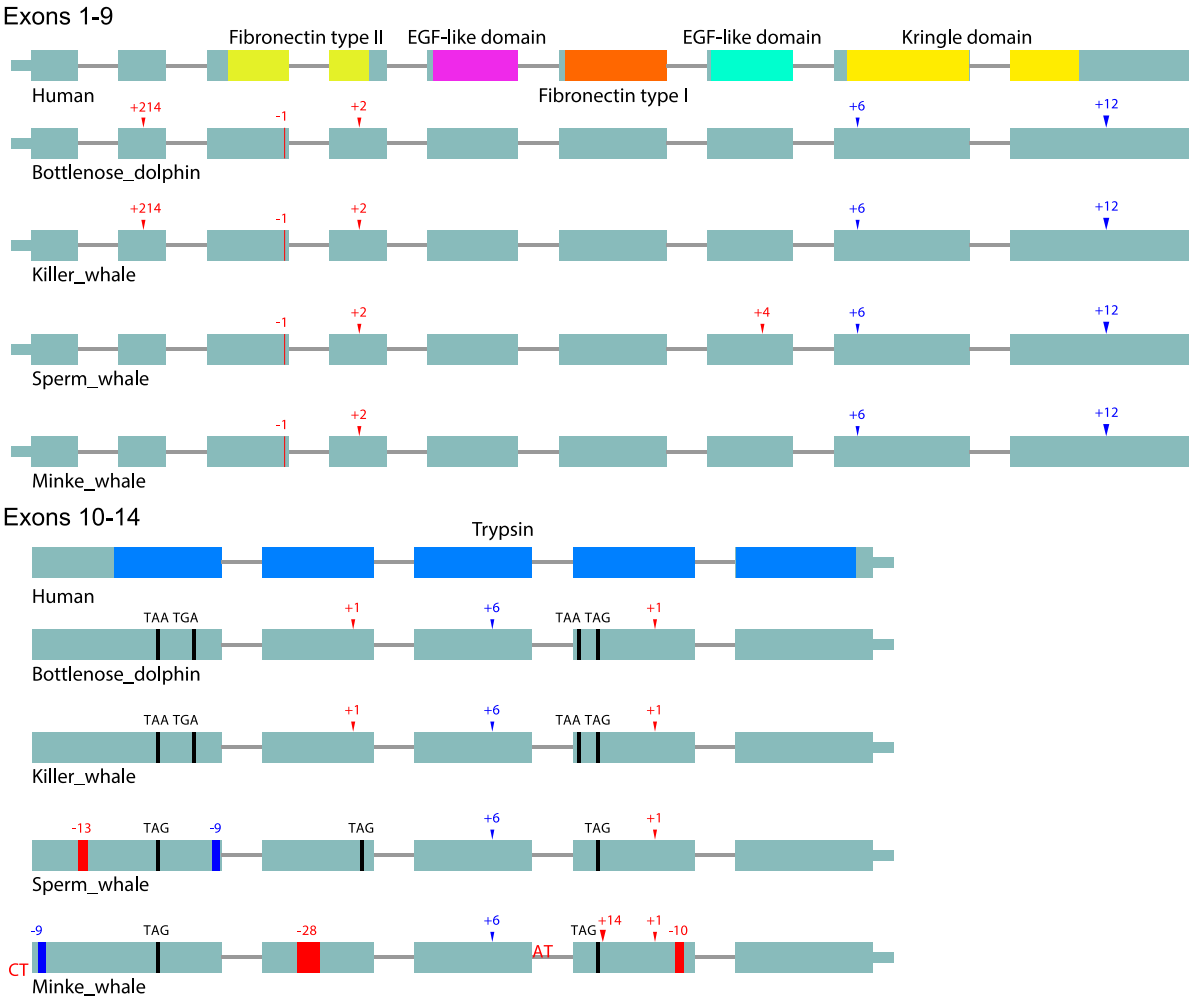


Fig. S4. Inactivating mutations in *F12* in cetaceans. General figure legend for figs. S4-S26: Exons are represented by boxes proportional to their size. Introns are represented by horizontal lines. Small boxes to the left and right indicate the beginning and end of the gene. The exon-intron structure of the human gene is shown as a reference at the top. The location of functional domains in the human protein, downloaded from Ensembl Biomart (*74*), is indicated by colored boxes. The exon-intron structures shown below the human gene visualize the mutations that occurred in the orthologous gene in cetaceans. Here, a filled red box indicates an exon deletion and filled grey box missing genomic sequence. Vertical red lines show frameshifting deletions, whereas vertical blue lines indicate frame preserving deletions. Arrow heads indicate frameshifting (red) or frame preserving (blue) insertions. The size of deletions or insertions is given on top of the mutation. Premature stop codons are indicated by black vertical lines and the corresponding triplet. Splice site mutations are shown by red letters at the end of an exon (donor mutation) or the beginning of an exon (acceptor mutation). Mutated ATG start codons are indicated as ‘noATG’.

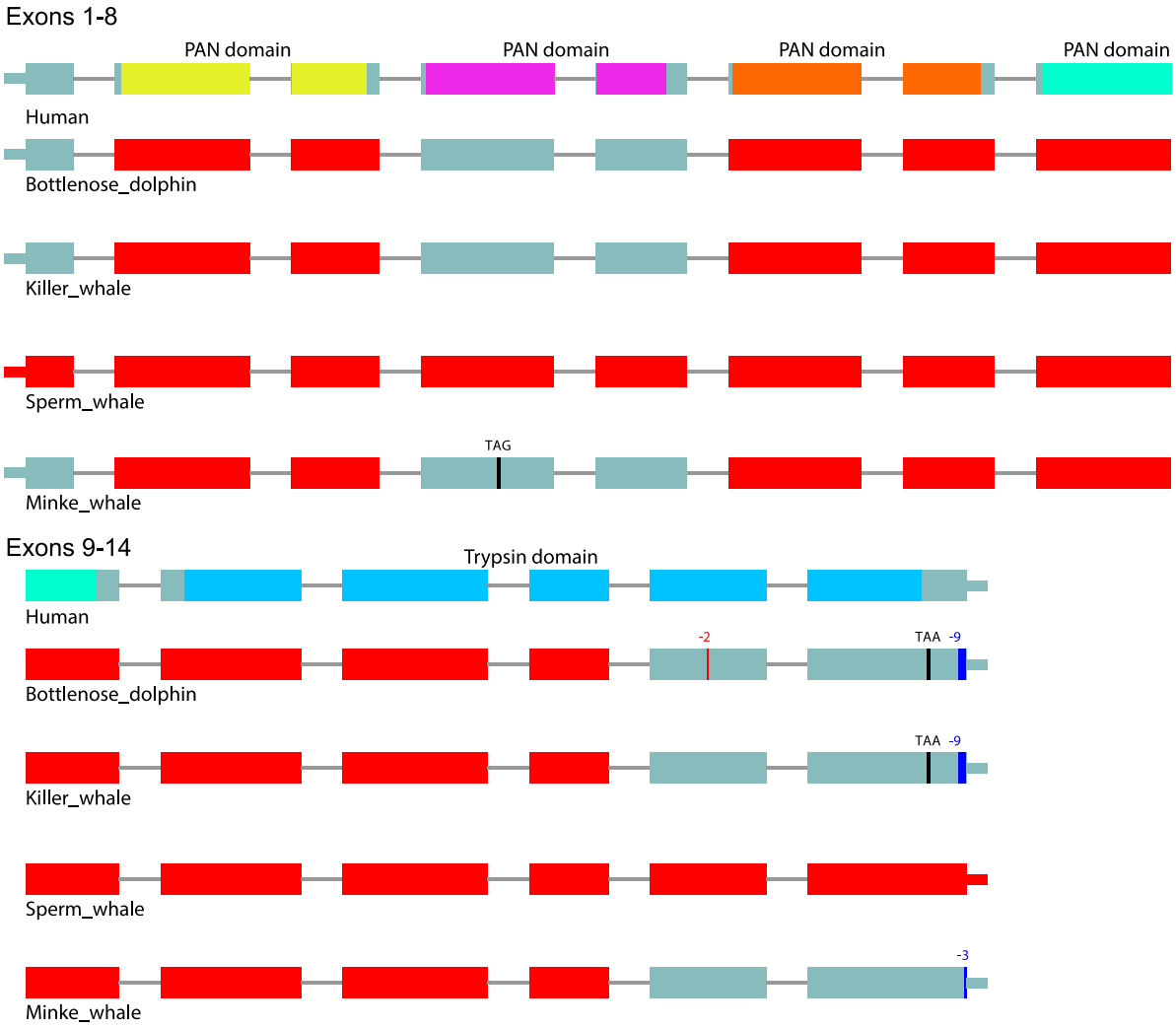


Fig. S5. Inactivating mutations in *KLKB1*. Visualization as in fig. S4. The breakpoints of the exon 6-12 deletion are shared between odontocetes and mysticetes. The entire gene is deleted in the sperm whale. The breakpoints of the exon 2-3 deletion in mysticetes are nested within the breakpoints in odontocetes, which could have arisen by independent (but overlapping) deletions or by a single smaller deletion in the cetacean stem lineage followed by additional deletions in odontocetes.

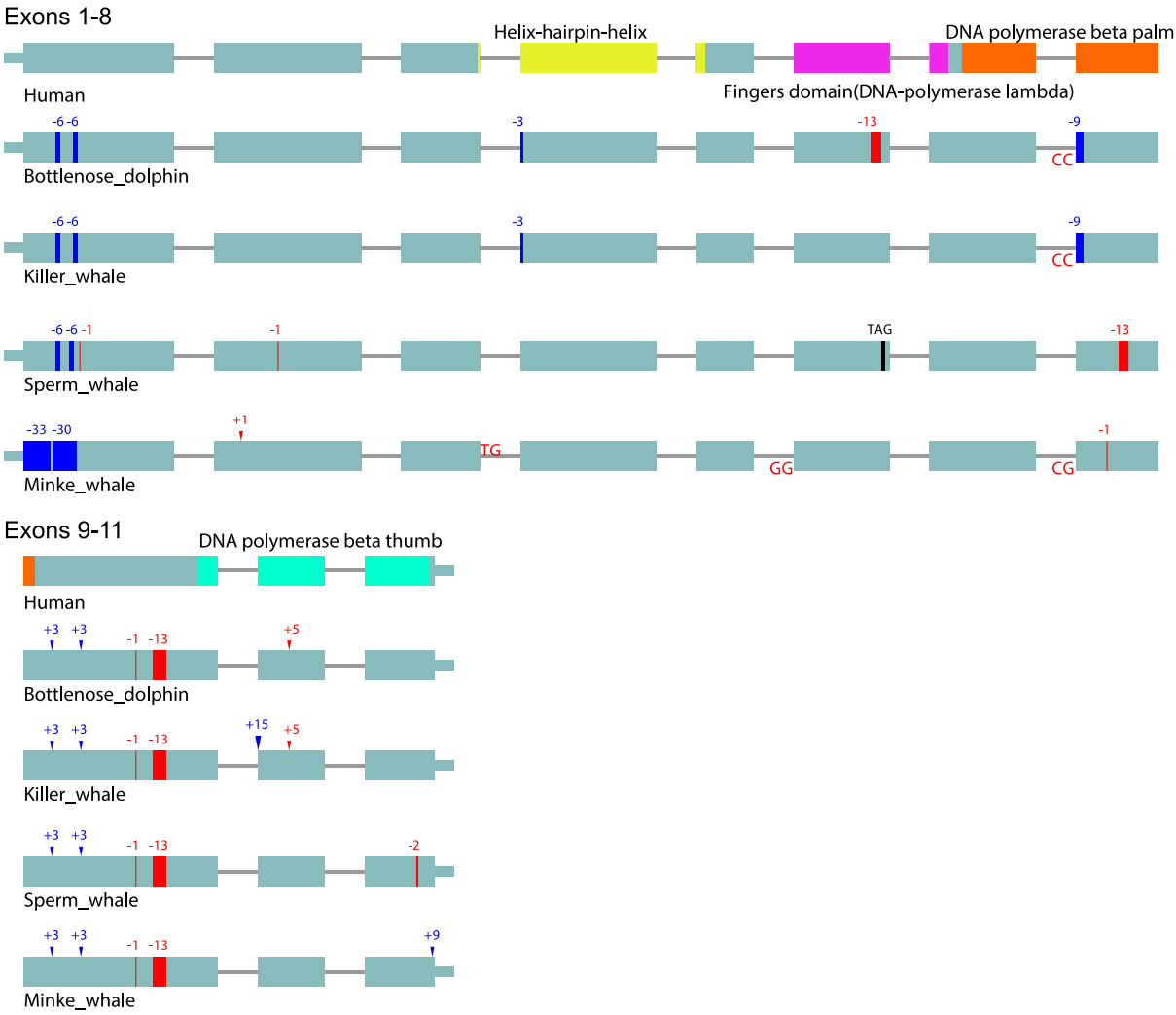


Fig. S6. Inactivating mutations in *POLM*. Visualization as in fig. S4.

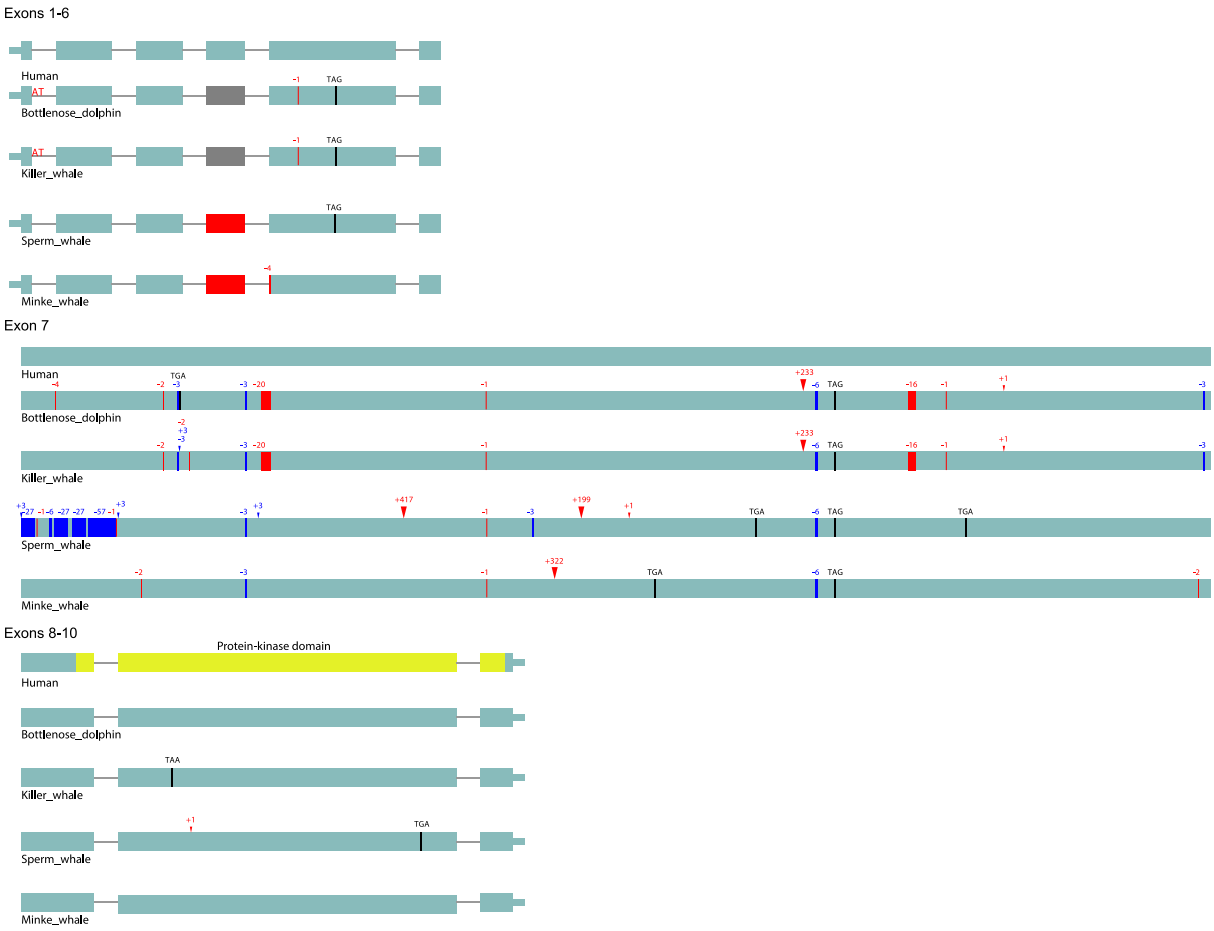


Fig. S7. Inactivating mutations in *MAP3K19*. Visualization as in fig. S4. In addition to the shared frameshift and stop codon mutations in exon 7, the deletion of exon 4 is also shared between all species and has the same breakpoints.

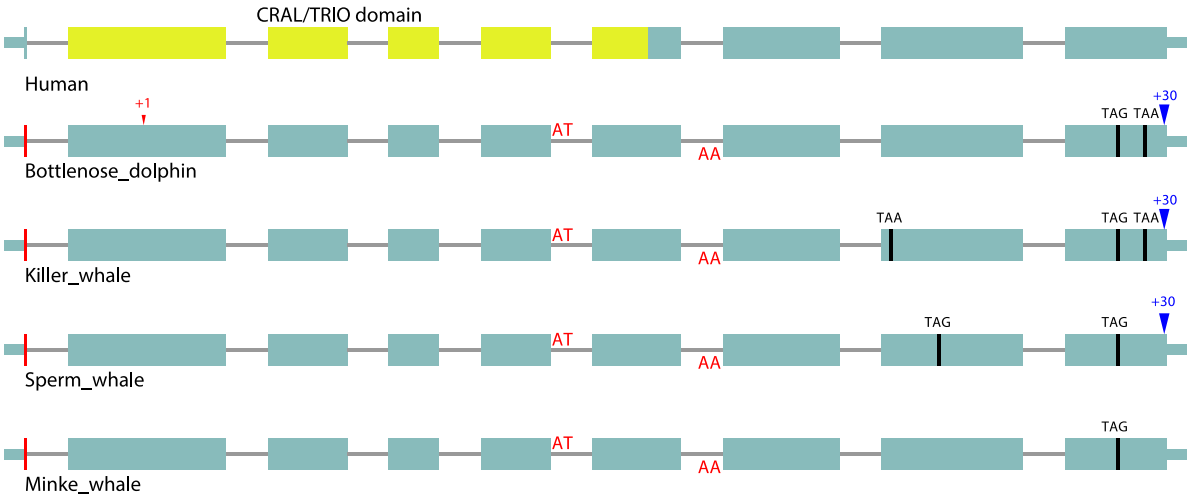


Fig. S8. Inactivating mutations in *SEC14L3*. Visualization as in fig. S4.

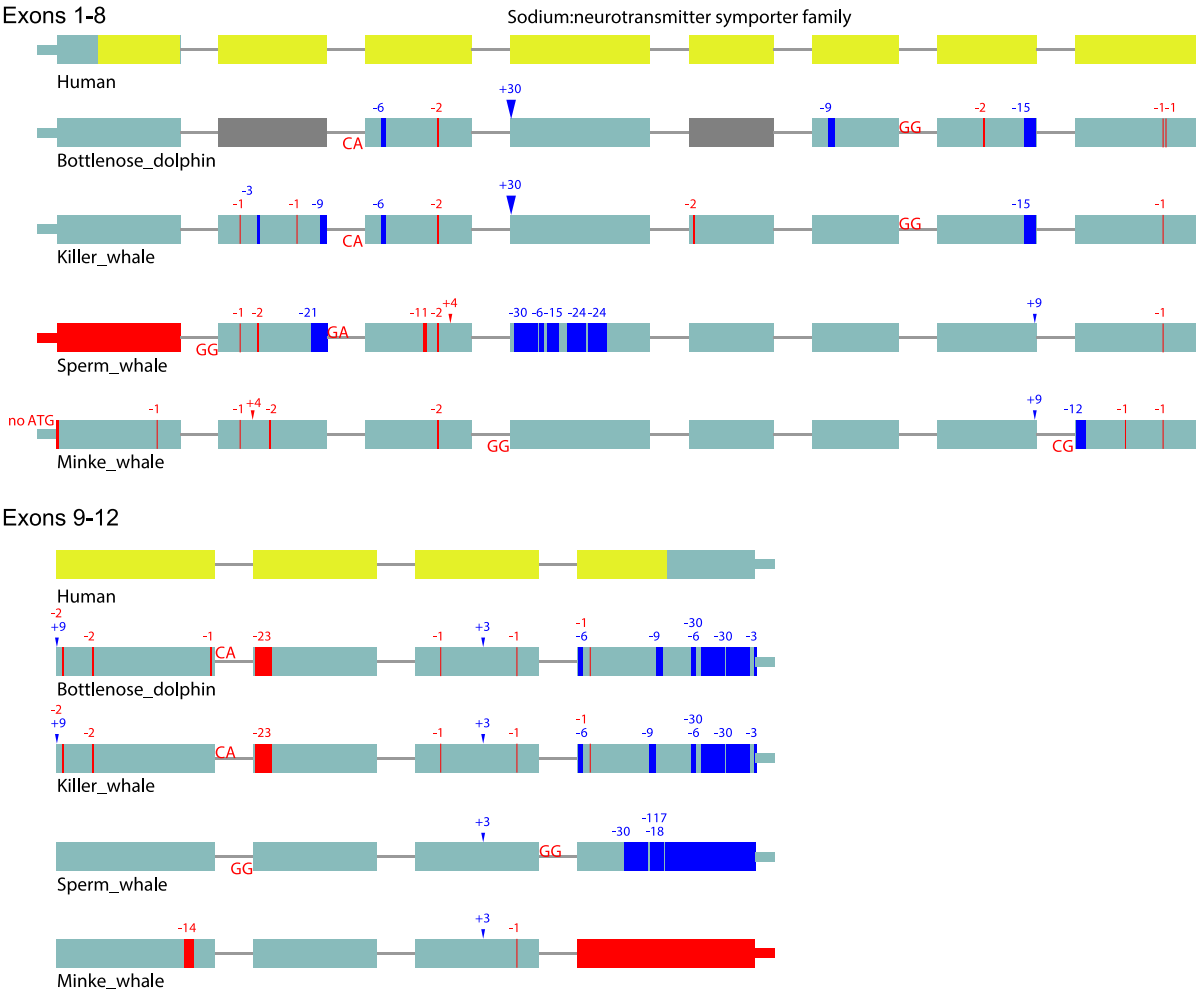


Fig. S9. Inactivating mutations in *SLC6A18*. Visualization as in fig. S4.

Fig. S10. Inactivating mutations in *SLC4A9*. (**A**) Inactivating mutations in cetaceans. (**B**) Inactivating mutations in the manatee. Visualization as in fig. S4.

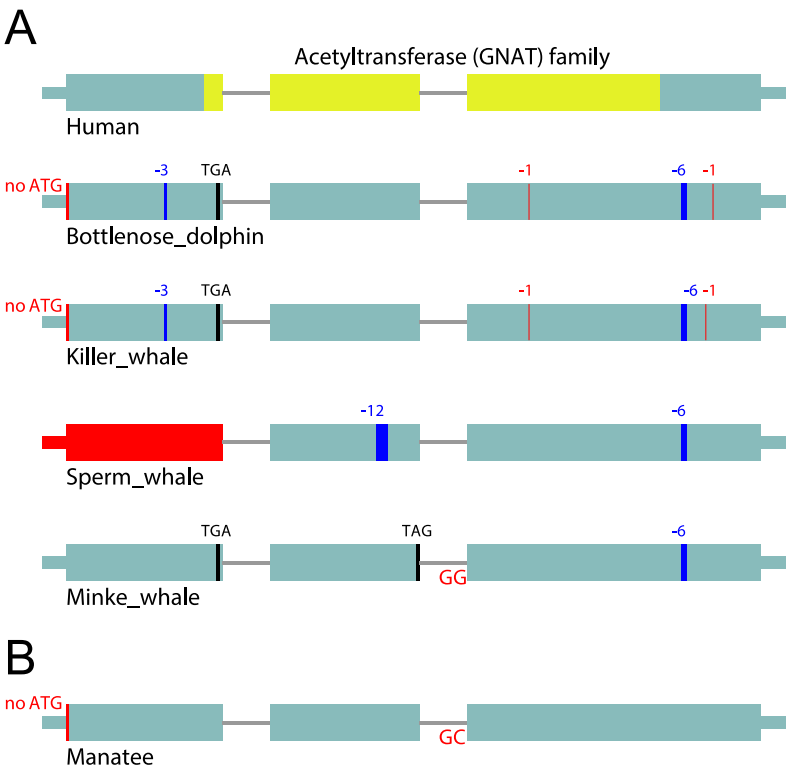


Fig. S11. Inactivating mutations in *AANAT*. (**A**) Inactivating mutations in cetaceans. (**B**) Inactivating mutations in the manatee.
Visualization as in fig. S4.

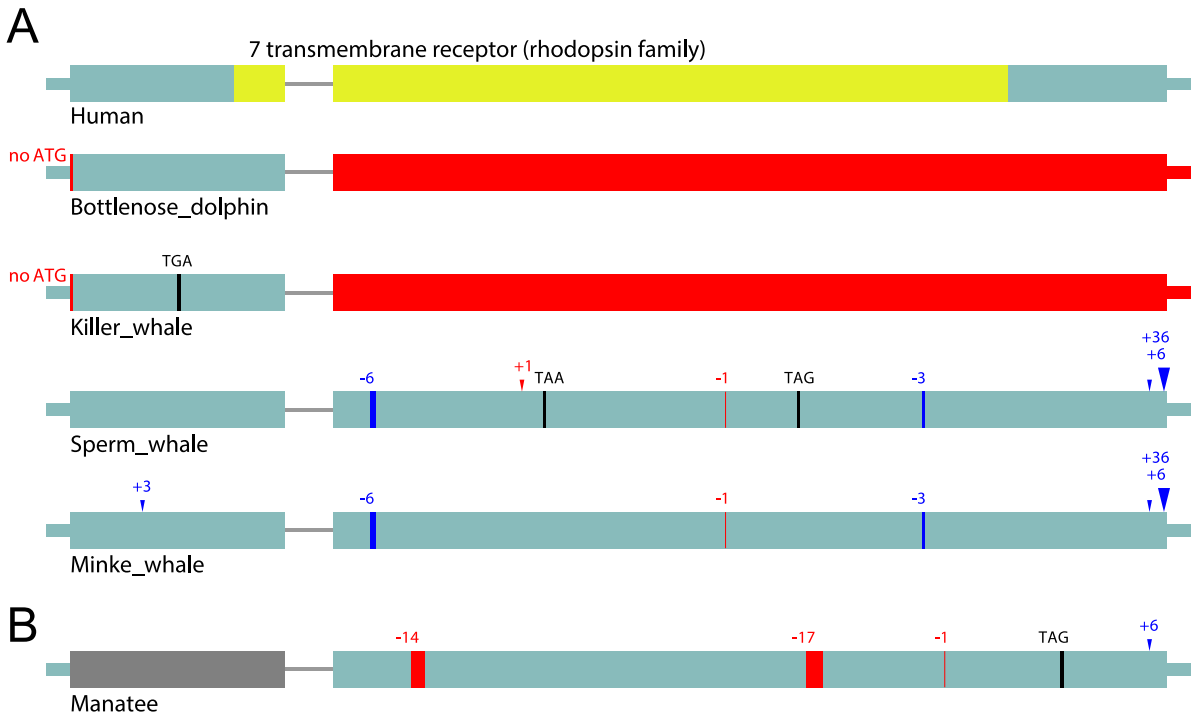


Fig. S12. Inactivating mutations in *MTNR1B*. (**A**) Inactivating mutations in cetaceans. (**B**) Inactivating mutations in the manatee. Visualization as in fig. S4.

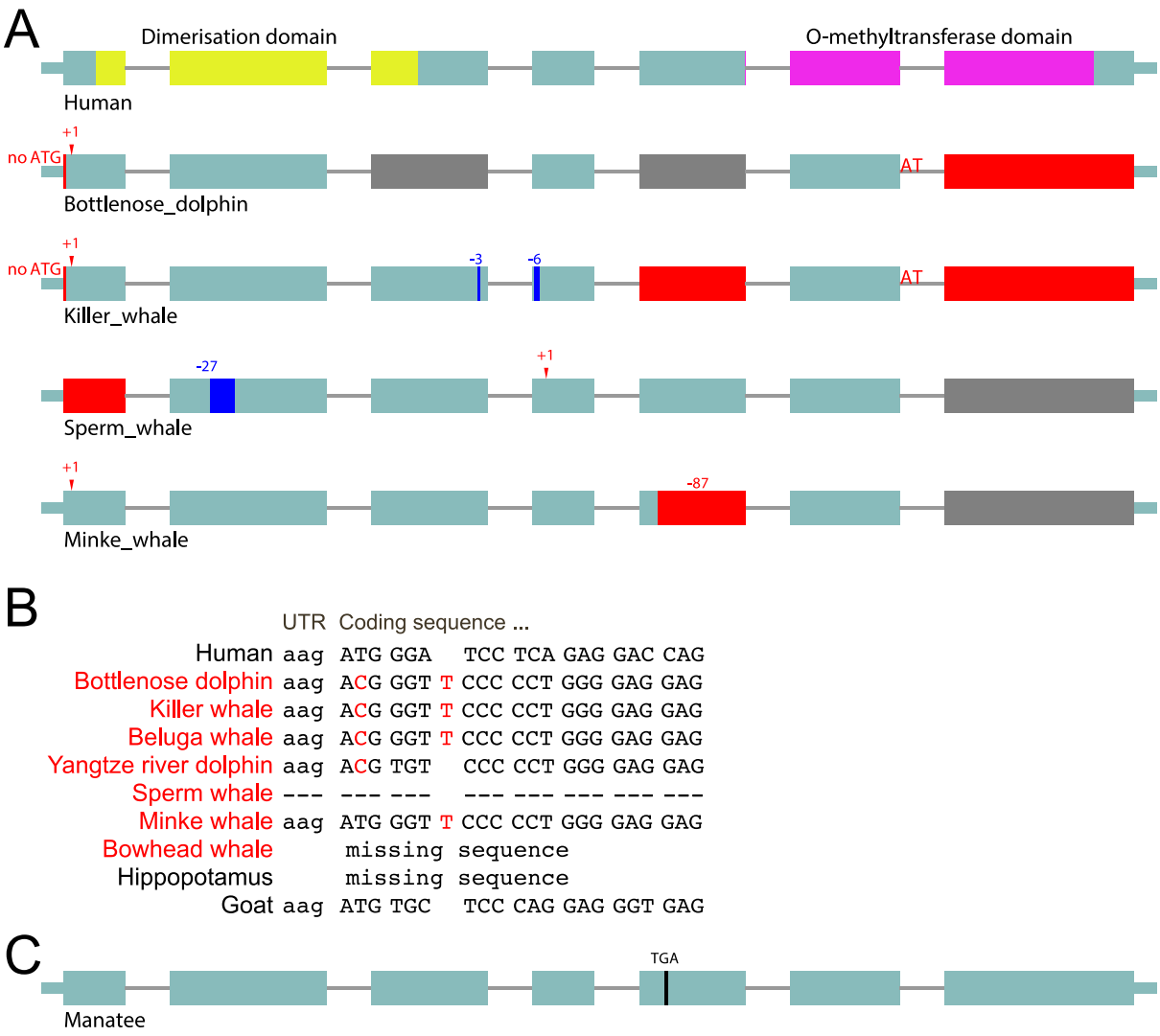


Fig. S13. Inactivating mutations in *ASMT*. (**A**) Inactivating mutations in cetaceans. Visualization as in fig. S4. The last exon is not present in any cetacean. In the sperm and minke whale, the respective genomic locus that is bounded by aligning blocks up- and downstream overlaps an assembly gap. While this does not exclude the possibility that this exon is truly deleted, we conservatively indicate it as missing genomic sequence. The putative upstream deletion breakpoint is very similar between cetaceans; however, the downstream breakpoint appears to vary. (**B**) The 1 bp frameshifting insertion at the beginning of exon 1 is not found in the Yangtze river dolphin. It is possible that this insertion already happened in the cetacean stem lineage and that the inserted base was later deleted in the Yangtze river dolphin. Alternatively, this insertion could have happened independently. (**C**) Inactivating mutations in the manatee. Please note that the single stop codon mutation is heterozygous based on DNA reads stored in the NCBI sequence read archive (see Table S3), indicating that *ASMT* loss likely happened recently in this lineage and is not yet fixed.

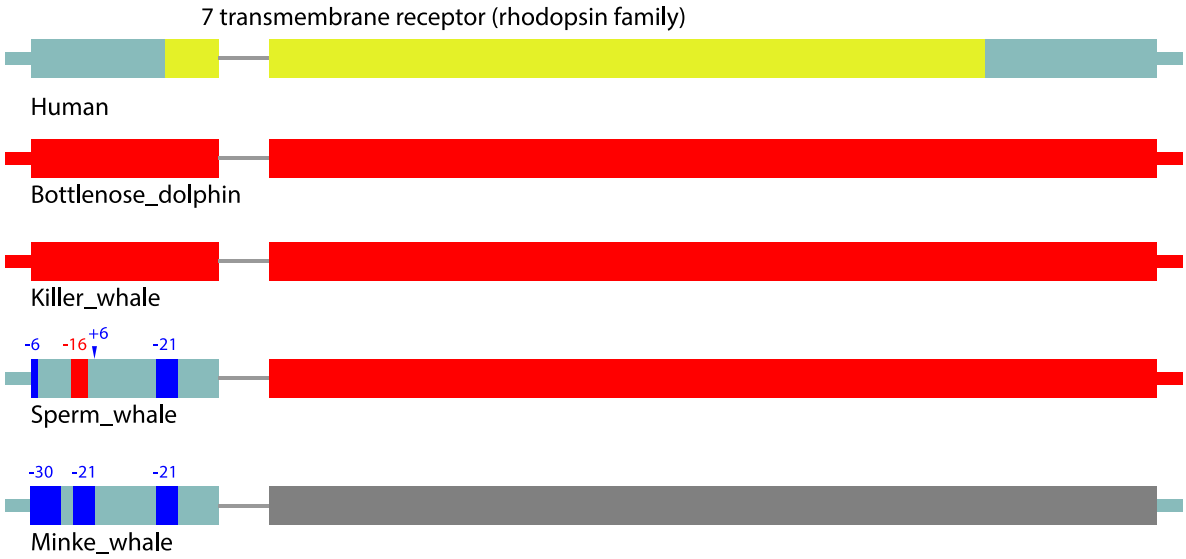


Fig. S14. Inactivating mutations in *MTNR1A*. Visualization as in fig. S4. The deletion breakpoints differ between species, indicating independent gene loss.

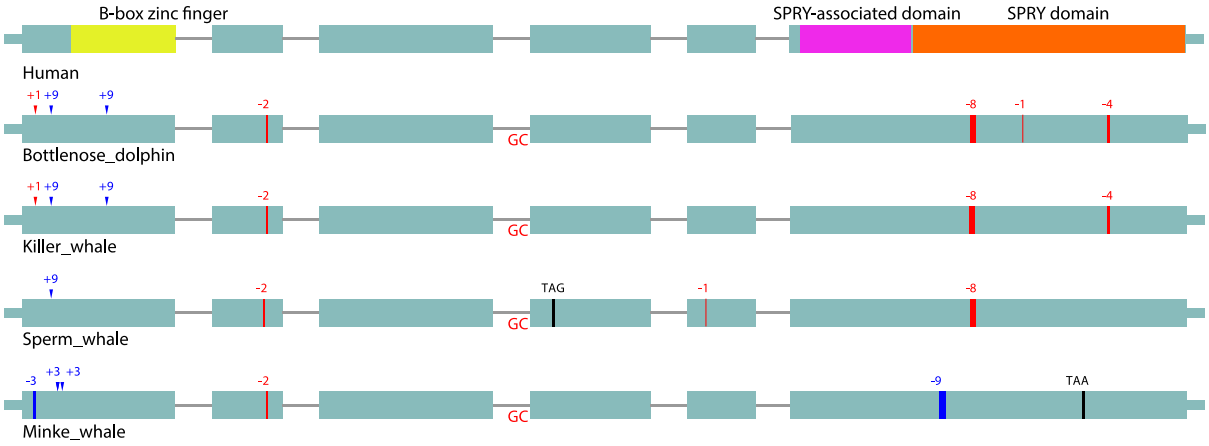


Fig. S15. Inactivating mutations in *TRIM14*. Visualization as in fig. S4.

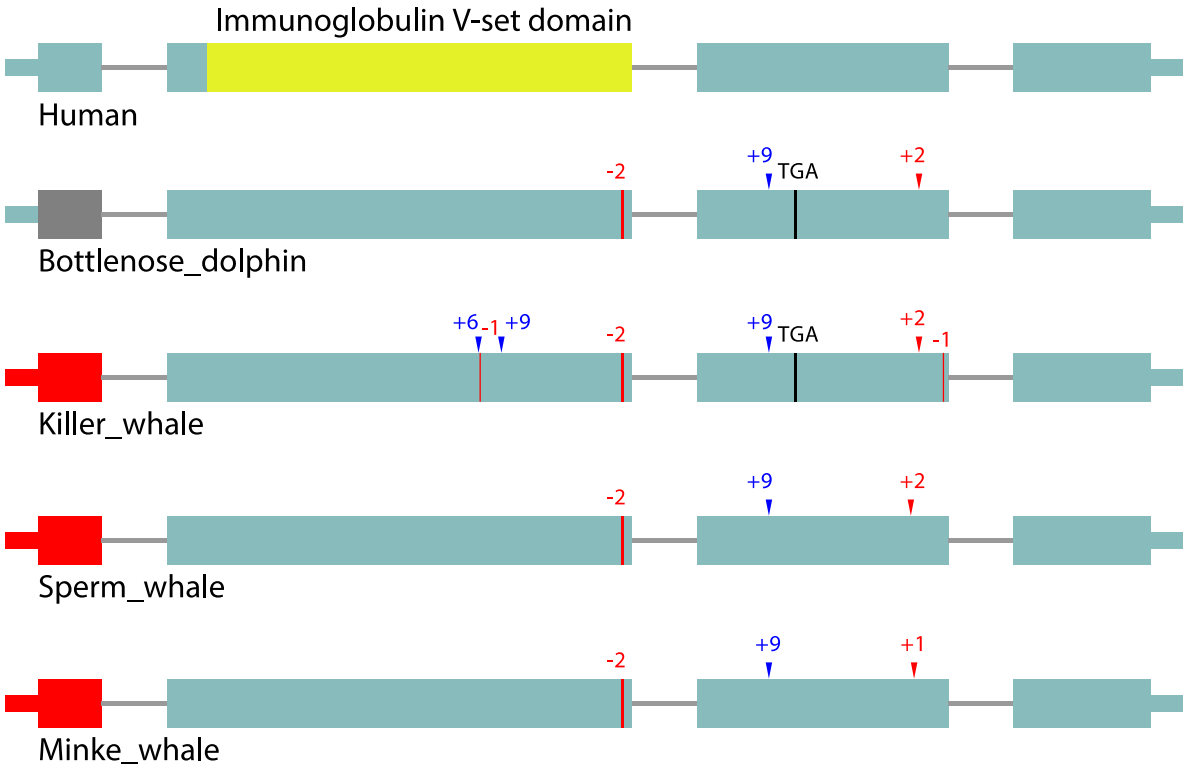


Fig. S16. Inactivating mutations in *TREM1*. Visualization as in fig. S4. In addition to the shared -2 bp frameshifting deletion in exon 2, the deletion of exon 1 is also shared between all species and has the same breakpoints.

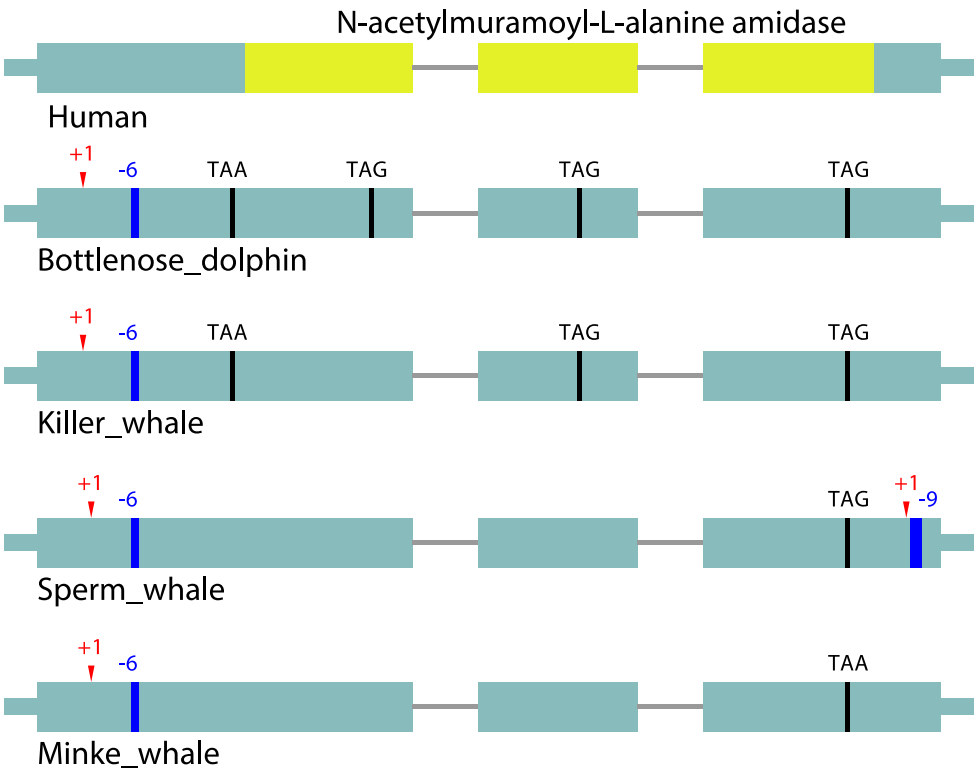


Fig. S17. Inactivating mutations in *PGLYRP1*. Visualization as in fig. S4. *PGLYRP1/3/4* encode peptidoglycan recognition proteins, which are receptors important for antimicrobial function and for maintaining a healthy gut microbiome (*75, 76*).

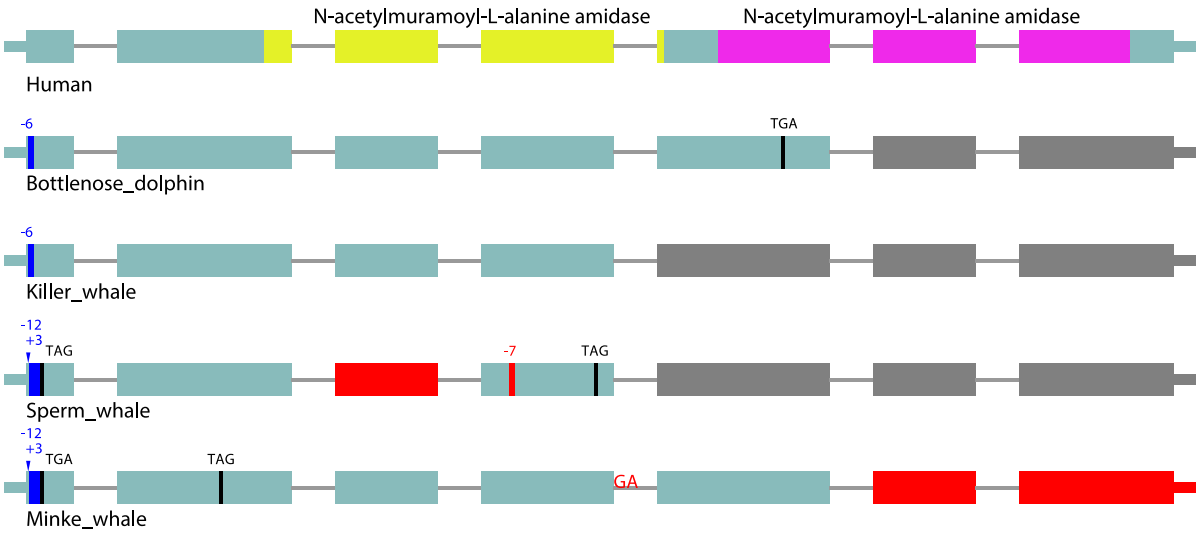


Fig. S18. Inactivating mutations in *PGLYRP3*. Visualization as in fig. S4. The deletion of the last two exons has shared deletion breakpoints in all species (except the sperm whale, where no alignment chain spans this region), consistent with a gene loss in the cetacean stem lineage. We conservatively mark these two exons as missing sequence in the dolphin and killer whale, because there are small assembly gaps in the respective locus. No assembly gap occurs in the minke whale and also the beluga whale (an odontocete), providing strong evidence that the last two exons are lost due to a single deletion that occurred in the cetacean stem lineage.

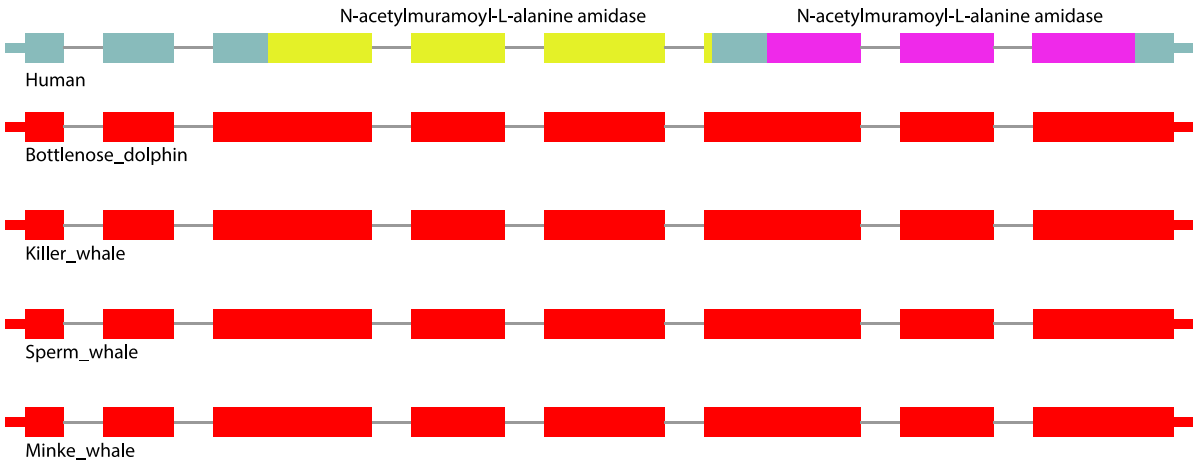


Fig. S19. Inactivating mutations in *PGLYRP4*. Visualization as in fig. S4. This gene is entirely deleted in all analyzed species, with shared deletion breakpoints, indicating an ancestral gene loss.

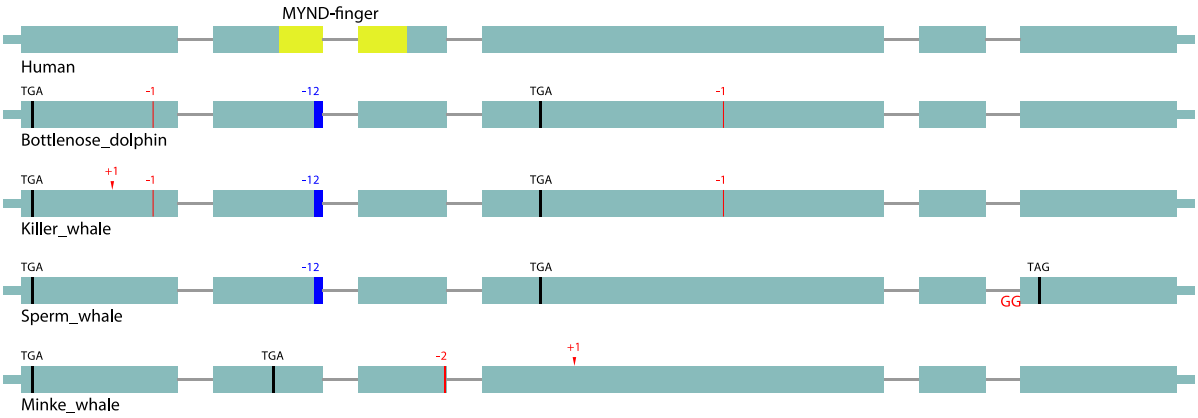


Fig. S20. Inactivating mutations in *MSS51*. Visualization as in fig. S4. Inactivation of *MSS51* in muscle cell lines directs muscle energy metabolism towards beta-oxidation of fatty acids (*77*). Loss of this gene in the cetacean stem lineage suggests that muscle metabolism may be largely fueled by fatty acids, which would be consistent with a high intramuscular lipid content in cetaceans (*78*).

Fig. S21. Inactivating mutations in *ACSM3*. Visualization as in fig. S4. This gene is involved in the oxidation of the short chain fatty acid butyrate (*79*).

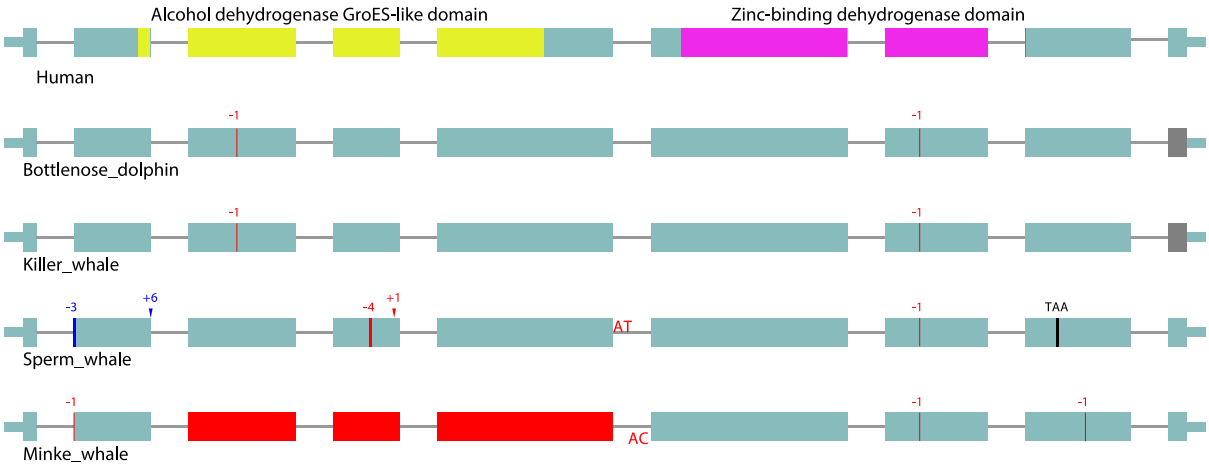


Fig. S22. Inactivating mutations in *ADH4*. Visualization as in fig. S4. *ADH4* metabolizes retinol and other substrates (*80*).

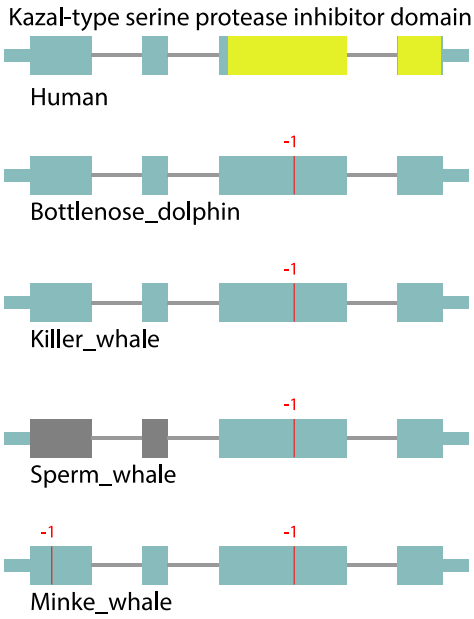


Fig. S23. Inactivating mutations in *SPINK7*. Visualization as in fig. S4. This gene is involved in esophageal epithelium development. Its loss could be linked to the specific ontogeny of the cetacean esophagus, which is homologous to a ruminant's forestomach (*81*).

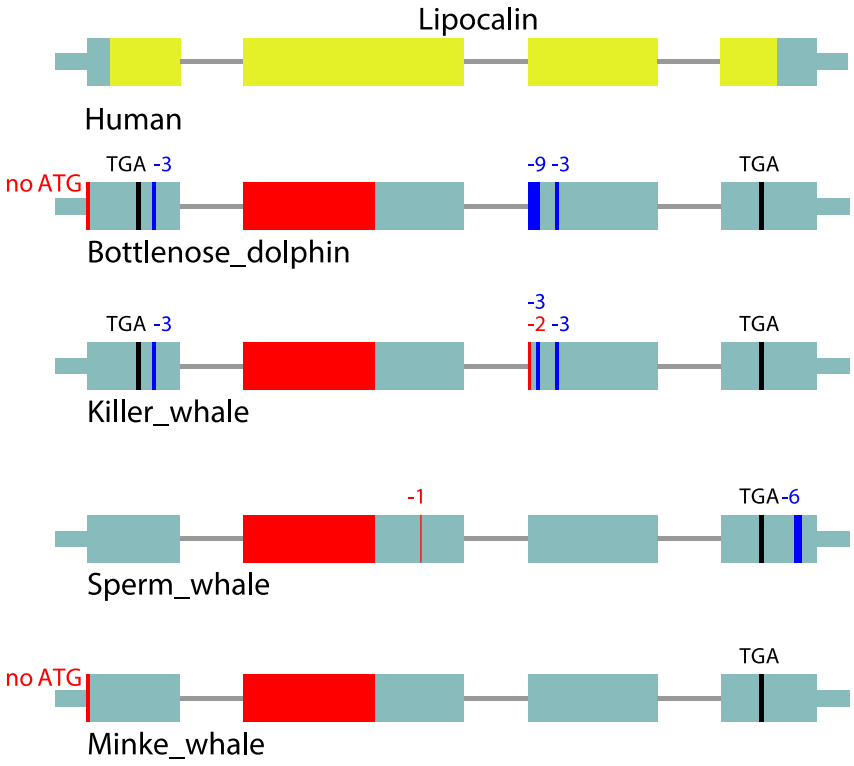


Fig. S24. Inactivating mutations in *FABP12*. Visualization as in fig. S4. In addition to the shared stop codon mutation in exon 4, the partial deletion of exon 2 is also shared between all species and has the same breakpoints. *FABP12* is a member of the fatty acid-binding protein family and is expressed in retina and testis of rats (*82*).

Fig. S25. Inactivating mutations in *ASIC5*. Visualization as in fig. S4. The -1 bp frameshifting deletion in exon 10 is shared between odontocetes and mysticetes and thus likely arose in the cetacean stem lineage. Exons 9 and 10 were likely later deleted in the sperm whale lineage; however, there is a small assembly gap, which is why we conservatively mark both exons as missing sequence. *ASIC5* is an orphan acid-sensing ion channel specifically expressed in interneuron subtypes of the vestibulocerebellum that regulates balance and eye movement (*83*).

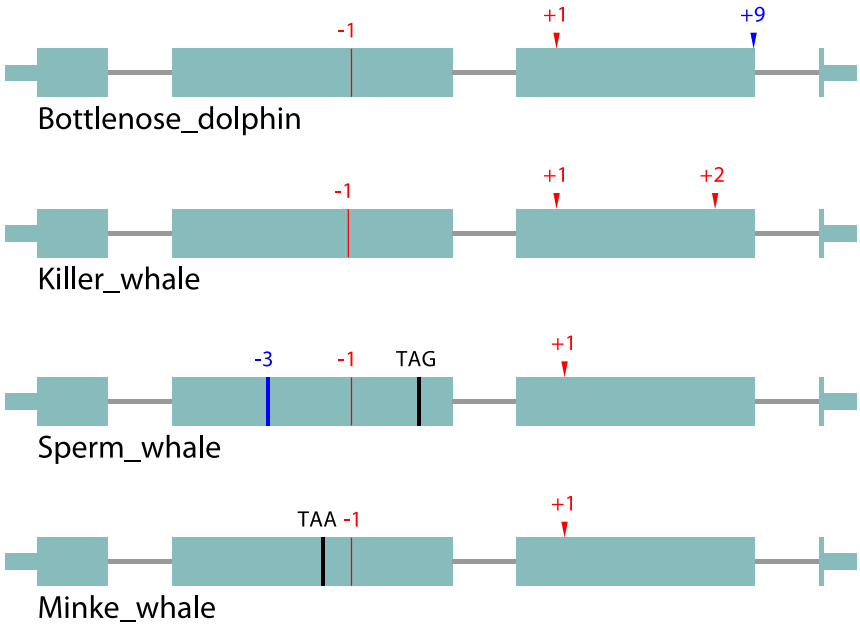


Fig. S26. Inactivating mutations in *C10orf82*. Visualization as in fig. S4. There are no annotated domains in the human protein. The gene is specifically expressed in the human testis (*84*).

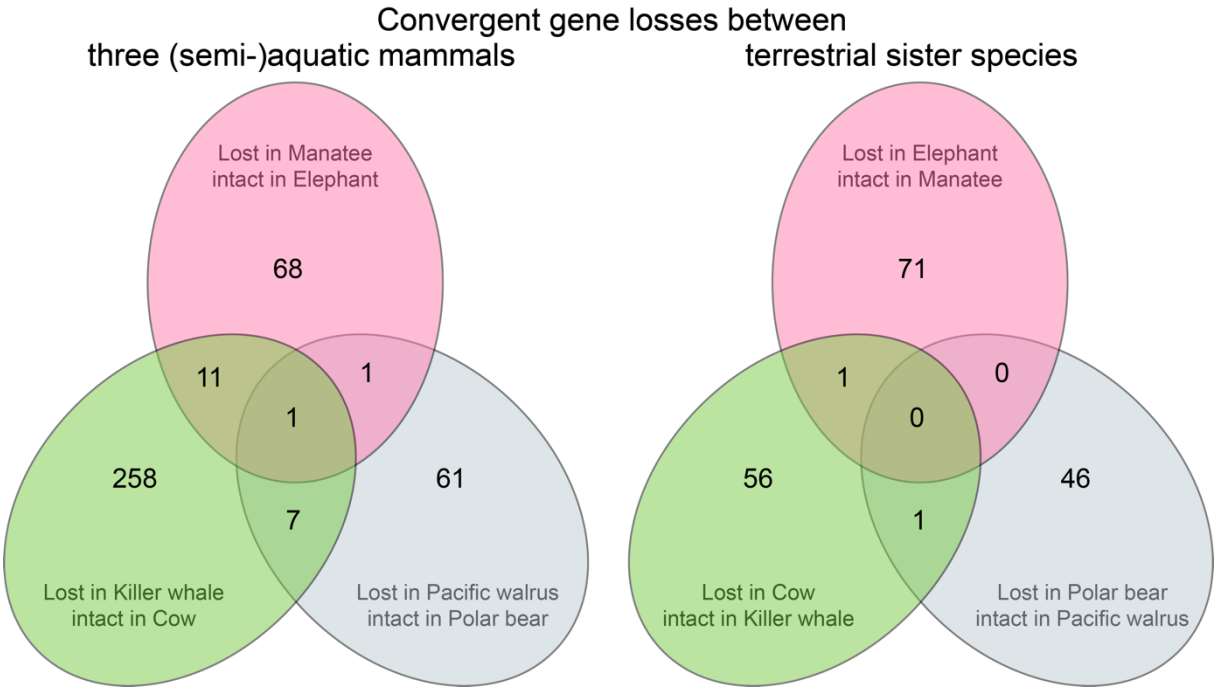


Fig. S27. Convergent gene losses between any of the three aquatic or semi-aquatic mammalian lineages. Left: Venn diagrams visualize genes, not belonging to the olfactory receptor and keratin-associated gene families, that are convergently lost in two or all three aquatic or semi-aquatic lineages, represented by the killer whale, manatee and Pacific walrus. These genes lack inactivating mutations in the respective terrestrial sister species, represented by cow, elephant and polar bear. A total of 20 genes are convergently lost in (semi-)aquatic mammals (Table S5). Right: Venn diagrams show that only two genes are convergently lost between two or three terrestrial mammals that are sister species to the (semi-)aquatic mammals.