Supplementary Information for

Transition to an aquatic habitat permitted the repeated loss of the pleiotropic KLK8 gene in mammals

Nikolai Hecker[1,2], Virag Sharma[1,2] and Michael Hiller[1,2*]

[1] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany.
[2] Max Planck Institute for the Physics of Complex Systems, Dresden, Germany.

*To whom correspondence should be addressed:
Michael Hiller
Computational Biology and Evolutionary Genomics, Max Planck Institute of Molecular Cell Biology and Genetics & Max Planck Institute for the Physics of Complex Systems, Dresden, Germany.
Tel: +49 351 210 2781
Fax: +49 351 210 1209
Email: hiller@mpi-cbg.de

The Supplementary Material contains

- Figures 1-5
- Tables 1-11

A *KLK8* coding exon 2

Genome alignment

```
Human        TGGGTCCTTACAGCTGCCCACTGTAAAAAACCgtgagtggatgatgggggcagaggtcagc
Dolphin      TGTATCCTCACAGCAGCCCACTGT--------------ggattccgggggcagaggtgggc
Killer whale TGTGTCCTCACAGCAGCCCACTGTAAAAAACTatgagtggattccgggagcagaggtgggc
```

CESAR exon alignment

```
Human   TGGGTCCTTACAGCTGCCCACTGTAAA---------AAACC
Dolphin TGTATCCTCACAGCAGCCCACTGTGGATTCCGGGGGCAGAGgtgggc
```

B *KLK8* coding exon 5

Genome alignment

```
Human    CTGGACTGGATCAAGAAGATCATAGGCAGCAAGGGCTGAttctag
Manatee  CTGGACTGGATCAAGAAGACCATAGGTAAC--GGGTTGAtcctca
```

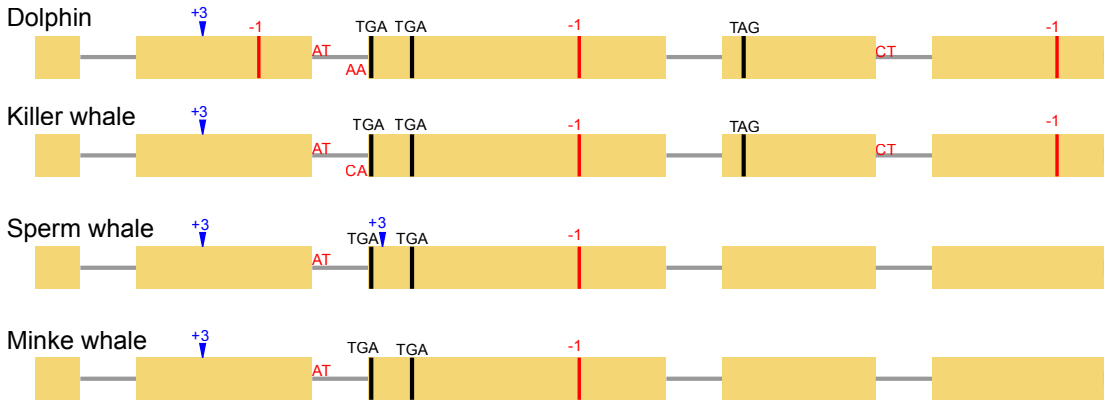
CESAR exon alignment

```
Human    CTGGACTGGATCAAGAAGATCATAGGCAGCAAGGGCTGA
Manatee  CTGGACTGGATCAAGAAGACCATAGGTAAC---GGGTTGATCCTCAAAGAGTCCTAG
```

Supplementary Figure 1: Application of CESAR results in a conservative number of gene-inactivating mutations.

(A) The genome alignment shows that dolphin (but not killer whale) exon 2 exhibits a deletion that removes the last 8 bp of the exon and the first 6 bp of the intron (confirmed by SRA reads), which could be counted as a frameshift and a splice site mutation. However, CESAR searches for other potential splice site candidates that occur in the same reading frame, and found a consensus splice site candidate (red font) in dolphin. As a conservative approach, we do not show inactivating mutations but the resulting 9 bp insertion at the end of the exon in Figures 1 and 3.

(B) The genome alignment shows a 2 bp frameshifting deletion close to the stop codon in *KLK8* coding exon 5. CESAR reports an alignment with a frame-preserving 3 bp deletion (shown in Figures 1 and 3) and a reading frame that terminates 6 codons downstream.



Supplementary Figure 2: *KLK11* is lost in cetaceans.
KLK11 is a downstream target of KLK8, since the KLK11 pro-enzyme was shown to be activated by KLK8 *in vitro* (Eissa, et al. 2011). *KLK11* is also expressed in the epidermis (Komatsu, et al. 2005; Komatsu, et al. 2006) and hippocampus (Mitsui, et al. 2000); however, its functional role is unknown. Based on our observation that *KLK11* is not expressed in the dolphin skin (Figure 4), we inspected the coding sequence of this gene. As shown here, we found a number of inactivating mutations in the four cetacean species included in our genome alignment. Several mutations are shared between all four species, suggesting a common loss in the cetacean ancestor. All mutations are also supported by sequencing reads. *KLK11* has no inactivating mutations in manatee.

A

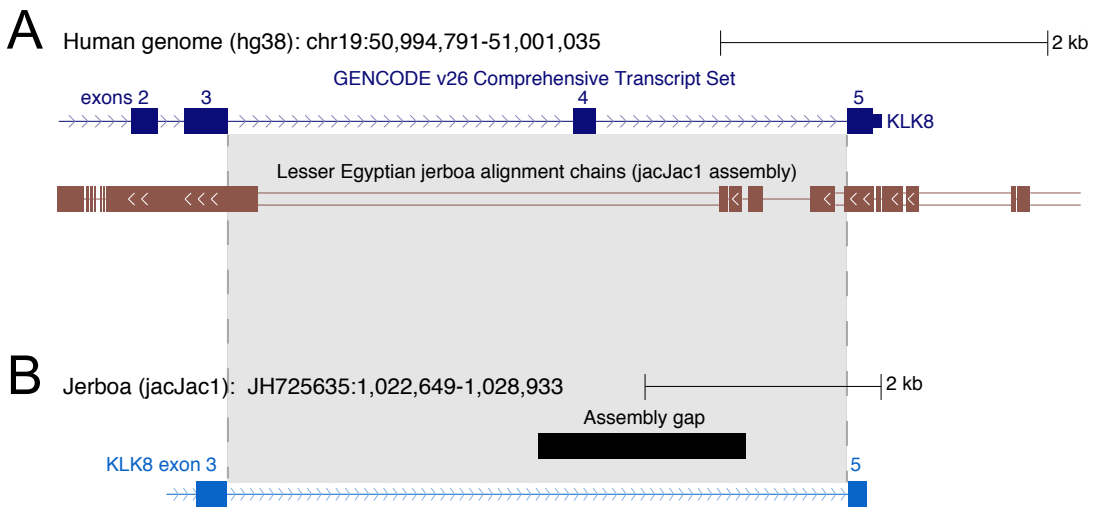
```
Genome(Human)       CCCAGGATGTGATGCCCTGGAGTGCACCATCACA
Genome(Armadillo)   CCCAGGATG-GATGCCC-GGAG-ACACCACCACA
ti:1949622794       CCCAGGATGTGATGCCCTGGAGAACACCACCACA
ti:553388016        CCCAGGATGTGATGCCCTGGAGAACACCACCACA
                    ********* ******* ****  ***** ****
```

B

```
            Genome (Human)    GAGGATGCTT-ACCCGGGG
          Genome (Marmoset)   GAGGATGCCTTACCCAGAG
SRA:DRR036757.218821465.2     GAGGATGCCT-ACCCAGAG
SRA:DRR036757.217876685.1     GAGGATGCCT-ACCCAGAG
SRA:DRR036757.208735343.1     GAGGATGCCT-ACCCAGAG
SRA:DRR036757.190750292.2     GAGGATGCCT-ACCCAGAG
                              ********* *********
```

Supplementary Figure 3: Sequencing errors in *KLK8* exons in armadillo and marmoset.

Unassembled sequencing reads show that putative gene-inactivating mutations (frameshifts) in armadillo exon 5 (A) and marmoset exon 4 (B) are sequencing errors as the putative mutation that is present in the genome assembly is not supported by sequencing reads from the trace archive or the SRA. Note that there is no armadillo or marmoset read that confirms the inactivating mutation.



Supplementary Figure 4: *KLK8* exon 4 is not contained in the current jerboa genome assembly.

(A) UCSC genome browser screenshot of the human genome locus comprising *KLK8* coding exons 2-5 and the alignment chains to the jerboa genome. Blocks in the alignment chain indicate aligning regions, the double-line overlapping exon 4 shows that this exon does not align.

(B) The corresponding locus in the jerboa genome (grey box) contains a 1,759 bp assembly gap that overlaps the locus where exon 4 should reside, making it likely that the exon 4 sequence was not added to the genome assembly.

A Alignment of human *KLK8* coding exon 3 (top) to pangolin (manPen1, bottom)

Genome alignment

```
        gAAATACACAGTACGCCTGGGAGACCACAGCCTACAGAATAAAGATGGCCCAGAGCAAGAAATACC
ttacaggAAGTACACAGTATGCCTGGGAGAGCACAGCCTGAAAAACAAGGAAGGATCAGAGCAAGAAATGGC
      *** ********* ********** ********  * ** ** ** **  *************  *

TGTGGTTCAGTCCATCCCACACCCCTGCTACAACAGCAGCGAT                    GTGGAGGACCA
TGTGGCTCAATCCATCCCACACCCCTGCTATGACAGCAGCAGT  Insert(1979bp)   ----AAGACCA
***** *** ********************  ********  *                        * *****

CAACCATGATCTGATGCTTCTTCAACTGCGTGACCAGGCATCCCTGGGGTCCAAAGTGAAGCCCATCAGCCT
CAGACATGATCTGATGCTTATTCGACTACATGGTTGGGCATCCTTGGGGTCCAAAGTGAAGCCCATCAACCT
**  *************** *** *** * **    ******* ************************ ***

GGCAGATCATTGCACCCAGCCTGGCCAGAAGTGCACCGTCTCAGGCTGGGGCACTGTCACCAGTCCCCGAg
GACAGATCACTGCCCTCGATCTGGCCAGAAGTGCACCATCTCGGGGTGGGGCACAGTCACCAGCCCTCAAggtagt
* ******* *** * *   ***************** **** ** ******** ******** ** * **
```

CESAR alignment

```
        gAAATACACAGTACGCCTGGGAGACCACAGCCTACAGAATAAAGATGGCCCAGAGCAAGAAATACC
ttacaggAAGTACACAGTATGCCTGGGAGAGCACAGCCTGAAAAACAAGGAAGGATCAGAGCAAGAAATGGC
      *** ********* ********** ********  * ** ** ** **  *************  *

TGTGGTTCAGTCCATCCCACACCCCTGCTACAACAGCAGCGATGTGGAGGACCACAACCATGATCTGATGCT
TGTGGCTCAATCCATCCCACACCCCTGCTATGACAGCAGCAGT---CAAGACCACAGACATGATCTGATGCT
***** *** ********************  ********  *    * *******  **************

TCTTCAACTGCGTGACCAGGCATCCCTGGGGTCCAAAGTGAAGCCCATCAGCCTGGCAGATCATTGCACCCA
TATTCGACTACATGGTTGGGCATCCTTGGGGTCCAAAGTGAAGCCCATCAACCTGACAGATCACTGCCCTCG
* *** *** * **    ******* ************************ **** ******* *** * *

GCCTGGCCAGAAGTGCACCGTCTCAGGCTGGGGCACTGTCACCAGTCCCCGAg
ATCTGGCCAGAAGTGCACCATCTCGGGGTGGGGCACAGTCACCAGCCCTCAAggtagt
  ***************** **** ** ******** ******** ** * **
```

B Alignment of human *KLK8* coding exon 1 (top) to naked mole rat (hetGla2, bottom)

Genome alignment

```
        ATGGGACGCCCC----CGACCTCG----TGCGGCCAAGACGTGGATGTTCCTGCTCTTGCTGGGG-
gcccccCTGGGATGCCCCGCCCCCGCCCCGGCCGTGCAGGCCTGATGTGGACGGTCCTGCTGCTGCTGCTGC
       ***** *****    *  ** **    *** * *  ** ***** * *******  *****  *

-----------GGAGCCTGGGCAg
TGCTGCTGCTGGCCTCTTGGCCAggtgag
           *   * *** ***
```

CESAR alignment

```
        ATGGGACGCCCCCGACCT---CGTGCGGCCAAGACGTGG------------ATGTTCCTGCTCTTG
cctgggATGCCCCGCCCCCGCCCCGGCCGTGCAGGCCTGATGTGGACGGTCCTGCTGCTGCTGCTGCTGCTG
      ***   ******** **    ***** * *  ** ****             ** * *****  **

CTGGGGGGAGCCTGGGCAg
CTGCTGGCCTCTTGGCCAggtgag
***  **   * *** ***
```

Supplementary Figure 5: CESAR reveals an intact exon alignment for several species with putative gene-inactivating mutations in the genome alignment.

(A) The genome alignment of *KLK8* coding exon 3 to the pangolin contains a 1.9 kb insertion in the middle of the exon, which would inactivate this exon due to the presence of in-frame stop codons. This insertion is the result of a

duplication in the pangolin manPen1 genome that copied the second half of the exon ~1.9 kb downstream. In contrast to the genome alignment, CESAR does not align the second half of the exon to the downstream duplication but aligns the exon as one block, which results in an alignment with an intact reading frame and consensus splice site, as shown here.
(B) The genome alignment of *KLK8* coding exon 1 to the naked mole rat contains two frameshifting 4 bp insertions and lacks the ATG start codon. In contrast, CESAR finds an intact exon alignment without frameshifts and an ATG, and two frame-preserving insertions.
Note that the identity of the genome and CESAR alignment is very similar in both cases, showing that both represent plausible nucleotide alignments. However, in contrast to CESAR, the genome alignment is not aware of the exon's reading frame, splice sites and the start codon in exon 1.

Species	Genome assembly	Species	Genome assembly
Human	hg38	Bactrian camel	camFer1
Chimp	panTro4	Dolphin	turTru2
Bonobo	panPan1	Killer whale	orcOrc1
Gorilla	gorGor3	Sperm whale	phyCat1
Orangutan	ponAbe2	Minke whale	balAcu1
Gibbon	nomLeu3	Tibetan antelope	panHod1
Rhesus	rheMac3	Cow	bosTau8
Crab-eating macaque	macFas5	Bison	bisBis1
Baboon	papAnu2	Sheep	oviAri3
Green monkey	chlSab2	Domestic goat	capHir1
Proboscis monkey	nasLar1	Horse	equCab2
Golden snub-nosed monkey	rhiRox1	Rhinoceros	cerSim1
Marmoset	calJac3	Cat	felCat8
Squirrel monkey	saiBol1	Dog	canFam3
Tarsier	tarSyr2	Ferret	musFur1
Bushbaby	otoGar3	Panda	ailMel1
Mouse lemur	micMur2	Polar bear	ursMar1
Chinese tree shrew	tupChi1	Pacific walrus	odoRosDiv1
Squirrel	speTri2	Weddell seal	lepWed1
Lesser Egyptian jerboa	jacJac1	Chinese pangolin	manPen1
Prairie vole	micOch1	Black flying-fox	pteAle1
Prairie deer mouse	perManBai1	Megabat	pteVam1
Chinese hamster	criGri1	Big brown bat	eptFus1
Golden hamster	mesAur1	Davids myotis bat	myoDav1
Mouse	mm10	Microbat	myoLuc2
Rat	rn6	Hedgehog	eriEur2
Upper Galilee mountains blind mole rat	nanGal1	Shrew	sorAra2
Naked mole-rat	hetGla2	Star-nosed mole	conCri1
Guinea pig	cavPor3	Elephant	loxAfr3
Chinchilla	chiLan1	Cape elephant shrew	eleEdw1
Brush-tailed rat	octDeg1	Manatee	triMan1
Rabbit	oryCun2	Cape golden mole	chrAsi1
Pika	ochPri3	Tenrec	echTel2
Pig	susScr3	Aardvark	oryAfe1
Alpaca	vicPac2	Armadillo	dasNov3

Supplementary Table 1: Species and their genome assemblies for which we analyzed the *KLK8* coding sequence.

Species	SRA accession
Dolphin	SRX1136398, SRX1136399
Killer whale	SRX188930, SRX188933
Sperm whale	SRX2447272, SRX2447273
Minke whale	SRX872217, SRX872216
Manatee	SRX091948, SRX091947
Marmoset	DRX032998, DRX033000

Supplementary Table 2: SRA identifiers that we used to validate the putative gene-inactivating mutations in *KLK8*. For dolphin, we also confirmed inactivating mutations and inferred intact versions of sequences with reads from the NCBI Trace Archive (Trace\Tursiops_truncatus_WGS).

Dolphin sequence spanning breakpoint
CCTTGGATTTGTCCCCTTCATGGAAGCCCCGCCCCCTTTGCTTATGATTGGTCCCTTAGAGTCAGGGACAGCTGCTGCAG CACAGCAAGATACCAGTTCTCA
Killer whale sequence spanning breakpoint
TCATGGAAGCCCCGCCCCCTTTGCTTATGATTGGTCCCTTAGAGTAAGGGACAGCCGCTGCAGCACAGCAAGATACCAGT TCTCAGGTGTGATCATGACTGGA

Supplementary Table 3: Genomic sequences spanning the breakpoint of the deletion that resulted in the loss of exon 5 in dolphin and killer whale. These genomic sequences are confirmed by several unassembled reads from the NCBI Trace archive and the SRA.

Killer whale exon1 stop codon inactivating
ATGGGACATCCCACAGCTGCTGCAGAGTCAATCTAGATGTTCCCGCTCTTGCTGTTGGAATCCCGGGCAG
Killer whale exon 1 stop codon ancestral state
ATGGGACATCCCACAGCTGCTGCAGAGTCAATCTGGATGTTCCCGCTCTTGCTGTTGGAATCCCGGGCAG
Killer whale exon 2 splice site inactivating
GTTCTCATAGATGACCACTGTGTCCTCACAGCAGCCCACTGTAAAAAACTATGAGTGGATTCCGGGAGCAGAGGTGGGCT GGAGCCTGGGGCAAGAGGGGGC
Killer whale exon 2 splice site ancestral state
GTTCTCATAGATGACCACTGTGTCCTCACAGCAGCCCACTGTAAAAAACTGTGAGTGGATTCCGGGAGCAGAGGTGGGCT GGAGCCTGGGGCAAGAGGGGGC
Killer whale exon 3 stop codon inactivating
CCTGCTACAACAGCAGCAACAAGGACCACAACCATGATCTGATGCTCATTTGACCATGTGAATGGGCATCCCTGGGGCTC AAAGTGAAGCCCATCAACCTGGC
Killer whale exon 3 stop codon ancestral state
CCTGCTACAACAGCAGCAACAAGGACCACAACCATGATCTGATGCTCATTCAACCATGTGAATGGGCATCCCTGGGGCTC AAAGTGAAGCCCATCAACCTGGC

Supplementary Table 4: Killer whale query sequences comprising the inactivating mutations that are present in its genome (red) and putative intact of versions of the sequences for which we replaced the inactivating mutations with the ancestral state (blue) inferred from the aligned cow sequences. While all sequences with inactivating mutations are confirmed by unassembled reads from the SRA and from sequencing data of distinct killer whale ecotypes (Supplementary Table 8), none of the putative intact sequences are confirmed.

Dolphin exon 1 frame shift inactivating
ATGGGACATCCCACAGCTGCTGCAGAGTCTGGATCTAGATGTTCCCGCTCTTGCCGTTGGAATCCCGGGCAG
Dolphin exon 1 frame shift ancestral state
ATGGGACATCCCACAGCTGCTGCAGTCTGGATCTAGATGTTCCCGCTCTTGCCGTTGGAATCCCGGGCAG
Dolphin exon 3 stop codon inactivating
CCTGCTACAACAGCAGCAACAAGGACCACAACCATGATCTGCTGCTCATTTGACCATGTGAACGGGCATCCCTGGGGCCC AAAGTGAAGCCCATCAACCTGGC
Dolphin exon 3 stop codon ancestral state
CCTGCTACAACAGCAGCAACAAGGACCACAACCATGATCTGCTGCTCATTCAACCATGTGAACGGGCATCCCTGGGGCCC AAAGTGAAGCCCATCAACCTGGC

Supplementary Table 5: Dolphin query sequences comprising the inactivating mutations that are present in its genome (red) and putative intact of versions of the sequences for which we replaced the inactivating mutations with the ancestral state (blue) inferred from the aligned cow sequences. While all sequences with inactivating mutations are confirmed by unassembled reads from the NCBI Trace Archive (Trace\Tursiops_truncatus_WGS), none of the putative intact sequences are confirmed.

Minke whale exon 1 frame shift inactivating
GTGGGACATCCTACACCTGCTGCAGAGTCTGGATCTGGATGTTCCAGCTCTTGCTGTTGGAATCCCGGGCCG
Minke whale exon 1 frame shift ancestral state
GTGGGACATCCTACACCTGCTGCAGTCTGGATCTGGATGTTCCAGCTCTTGCTGTTGGAATCCCGGGCCG
Minke whale exon 2 stop codon inactivating
TGAGAGCACAGGAGACCAAGGTGCTGGAGGGCCAGGAGTGCAGGCCCCATTAGCAGCCTTGGCAGACGGCCTTGTTCCAG GGTGTCCGGCTAATATGTGGGGA
Minke whale exon 2 stop codon ancestral state
TGAGAGCACAGGAGACCAAGGTGCTGGAGGGCCAGGAGTGCAGGCCCCATTCTCAGCCTTGGCAGACGGCCTTGTTCCAG GGTGTCCGGCTAATATGTGGGGA
Minke whale exon 4 splice site inactivating
CTGTCTCTGGGAACTTGCAACCTCTGCCCCCTCGGAGAATTTTCCTGACACCCTCAGCTGCGCAGAA
Minke whale exon 4 splice site ancestral state
CTGTCTCTGGGAACTTGCAACCTCTGCCCCCTCAGAGAATTTTCCTGACACCCTCAGCTGCGCAGAA

Supplementary Table 6: Minke whale query sequences comprising the inactivating mutations that are present in its genome (red) and putative intact of versions of the sequences for which we replaced the inactivating mutations with the ancestral state (blue) inferred from the aligned cow sequences. All inactivating mutations could be confirmed by unassembled reads from the SRA. The sequence covering the ancestral state of the splice site mutation in exon 4 retrieved a single matching read (SRA:SRR1802585.28974202.1 SRX872217); however, given that the same SRA datasets provide overwhelming support (20 matching reads) for the sequence with the inactivated splice site, we can conclude that this read has a sequencing error at the mutated position.

Manatee exon 1 splice site inactivating
TTCCTGCTTTTGCTGTTGGAAGCCTGGGCAGGGAAGGGGTTTGGGGAAGGGGCTGGAACACA
Manatee exon 1 splice site ancestral state
TTCCTGCTTTTGCTGTTGGAAGCCTGGGCAGTGAAGGGGTTTGGGGAAGGGGCTGGAACACA
Manatee exon 3 stop codon 1 inactivating
TACACAGTTTGCCTGGGAGATCACAGCCTGTAGAGTAAGGATGGGCTGGAGCAAGAAATGGCT
Manatee exon 3 stop codon 1 ancestral state
TACACAGTTTGCCTGGGAGATCACAGCCTGCAGAGTAAGGATGGGCTGGAGCAAGAAATGGCT
Manatee exon 3 stop codon 2 inactivating
AGTTGGTAGATCGCTGCCCCCAGGCTGGCCAGCTGTGCACCATCTCTGGCTGAGGCACTGTCACCAGCCCCCAAGTTACT GGGCTTGGCCAGCACT
Manatee exon 3 stop codon 2 ancestral state
AGTTGGTAGATCGCTGCCCCCAGGCTGGCCAGCTGTGCACCATCTCTGGCTGGGGCACTGTCACCAGCCCCCAAGTTACT GGGCTTGGCCAGCACT
Manatee exon 3 splice site inactivating
GTGCACCATCTCTGGCTGAGGCACTGTCACCAGCCCCCAAGTTACTGGGCTTGGCCAGCACTGTGAGAGAGAGGAGGAGT TGGTGGGCCTAGTGAACTCCAA
Manatee exon 3 splice site ancestral state
GTGCACCATCTCTGGCTGAGGCACTGTCACCAGCCCCCAAGTTACTGGGCGTGGCCAGCACTGTGAGAGAGAGGAGGAGT TGGTGGGCCTAGTGAACTCCAA

Supplementary Table 7: Manatee query sequences comprising the inactivating mutations that are present in its genome (red) and putative intact of versions of the sequences for which we replaced the inactivating mutations with the ancestral state (blue) inferred from the aligned elephant sequences. All inactivating mutations could be confirmed by unassembled reads from the SRA. Similarly to the minke whale case above (Supplementary Table 6), the “exon 3 stop codon 2” sequence with the ancestral state retrieved a single matching read (SRA:SRR331137.123171751.3 SRX091948); however, as in the minke whale case, the same SRA datasets provide 38 matching reads that support the gene-inactivating stop codon mutation. Thus, a A->G sequencing error (TGA->TGG) explains this single read.

	exon 1 stop codon		exon 2 splice site		exon 3 stop codon	
Accession	**inactiv.**	**ancestral**	**inactiv.**	**ancestral**	**inactiv.**	**ancestral**
ERR637306	4	0	0	0	0	0
ERR637307	5	0	7	0	12	0
ERR637308	6	0	0	0	1	0
ERR637309	1	0	1	0	3	0
ERR637310	5	0	2	0	2	0
ERR637311	0	0	0	0	0	0
ERR637312	3	0	2	0	1	0
ERR637313	0	0	6	0	4	0
ERR637314	3	0	1	0	3	0
ERR637315	5	0	6	0	2	0
ERR637316	3	0	2	0	3	0
ERR637317	2	0	0	0	3	0
ERR637318	3	0	2	0	3	0
ERR637319	2	0	1	0	1	0
ERR637320	1	0	0	0	0	0
ERR637321	1	0	1	0	3	0
ERR637322	3	0	4	0	4	0
ERR637323	2	0	1	0	0	0
ERR637324	5	0	2	0	2	0
ERR637325	0	0	2	0	4	0
ERR637326	2	0	4	0	1	0
ERR637327	3	0	5	0	2	0
ERR637328	1	0	3	0	5	0
ERR637329	4	0	1	0	2	0
ERR637330	2	0	6	0	5	0
ERR637331	1	0	2	0	3	0
ERR637332	5	0	0	0	3	0
ERR637333	2	0	1	0	1	0
ERR637334	2	0	0	0	0	0
ERR637335	0	0	0	0	1	0
ERR637336	2	0	1	0	4	0
ERR637337	3	0	2	0	5	0
ERR637338	0	0	2	0	5	0
ERR637339	0	0	1	0	3	0
ERR637340	0	0	4	0	1	0
ERR637341	7	0	8	0	6	0
ERR637342	1	0	1	0	1	0
ERR637343	5	0	5	0	8	0
ERR637344	1	0	3	0	4	0
ERR637345	1	0	0	0	0	0
ERR637346	2	0	4	0	3	0
ERR637347	3	0	3	0	4	0
ERR637348	2	0	2	0	3	0
ERR637349	0	0	1	0	0	0
ERR637350	2	0	2	0	5	0
ERR637351	1	0	1	0	1	0
ERR637352	1	0	2	0	3	0
ERR637353	0	0	2	0	2	0
SUM	**107**	**0**	**106**	**0**	**132**	**0**

Supplementary Table 8: Sequencing data of 48 killer whale individuals from five distinct ecotypes confirm that *KLK8* is lost throughout different killer whale populations.

The number of sequencing reads that confirm inactivating mutations (inactiv.) and intact ancestral state sequences (Supplementary Table 4) is shown for 48 killer whale individuals (Foote, et al. 2016). The respective European Nucleotide Archive (Cochrane, et al. 2013) accessions are listed.

Gene	$r_{mean:exon/intron}$	$r_{med:exon/intron}$	$e_{mean:exon}$	$e_{mean:intron}$	$e_{med:exon}$	$e_{med:intron}$
KLK9	11.4	14.4	5818	510	7219	503
KLK8	1.0	1.4	256	249	246	171
KLK7	32.0	34.9	52517	1640	56669	1623
KLK6	14.4	15.2	22281	1548	23269	1532
KLK4	1.3	3.3	76	57	90	27

Supplementary Table 9: Exon and intron expression levels in the *KLK* gene locus based on dolphin RNA-seq data. For the genes on the same scaffold as *KLK8* (JH479757), we calculated the mean and median read coverage for each exon and intron using reads from all 116 samples of the dolphin skin RNA-seq data set (GEO-accession: GSE90941). For each gene, we then averaged the exon and intron values to obtain a per-gene exon and intron expression level ($e_{mean:exon}$ and $e_{mean:intron}$). Similarly, we computed the per-exon and per-intron median of all 116 samples and then computed the median of the exon/intron values for each gene, obtaining $e_{med:exon}$ and $e_{med:intron}$. $r_{mean:exon/intron}$ or $r_{med:exon/intron}$ refer to the ratio between mean exon and mean intron expression or the ratio between median exon and median intron expression, respectively. An expressed multi-exon gene is expected to exhibit much higher exon than intron expression levels (large ratios), which we observed for *KLK6/7/9* that are clearly expressed in these RNA-seq datasets (Figure 4). Genes which are not expressed in these dolphin skin samples are expected to have (i) ratios close to 1, indicating background expression levels, and (ii) overall lower exon expression levels compared to expressed genes. This was observed for the inactivated *KLK8* and for *KLK4*, which is expressed in the prostate and developing teeth. These results corroborate the data shown in Figure 4 that *KLK8* is not expressed in dolphin skin.

SRA accession	Gene	Exon 1	Exon 2	Exon 3	Exon 4	Exon 5
SRX220351	*KLK8*	29	57	57	0	0
SRX220351	*KLK7*	6616	>20000	>20000	>20000	>20000

Supplementary Table 10: *KLK8* is not expressed in the sperm whale skin. The table shows the read counts, obtained with a Megablast search for the five coding exons of sperm whale *KLK7* and *KLK8*. For *KLK7*, which is located next to *KLK8*, we found thousands of reads for all exons, showing that this gene is clearly expressed. In contrast, we found not a single read for *KLK8* exon 4 and 5, and only a small number of reads for the first three coding exons. The expression ratio between *KLK7* and *KLK8* is at least 1:228, suggesting that the expression of *KLK8* is negligibly small. SRX220351 is part of a study of five sperm whale samples (SRA accession: SRP016870). For the other samples (SRX220352-SRX220358), Megablast did not find any reads aligning to *KLK8* exons.

		Test set			Background		
Test set	p-value	ω_1	ω_2	ω_3	ω_1	ω_2	ω_3
Sperm whale	$< 10^{-4}$	1 (1,7%)	1 (20.5%)	1 (77.8%)	0.08 (77.8%)	0.81 (20.5%)	8.49 (1.7%)
Fully-aquatic	$< 10^{-5}$	1 (87.8%)	1 (10.1%)	1 (2.1%)	0.13 (87.8%)	1 (10.1%)	7.46 (2.1%)

Supplementary Table 11: Neutral evolution of *KLK8* in the sperm whale and in the fully-aquatic lineages.

Relaxation of selection rates computed with RELAX (Wertheim, et al. 2015) for the sperm whale branch and all branches associated with fully-aquatic species (cetaceans, manatee), compared with all branches associated with non-fully-aquatic species as a background set. ω_1, ω_2 and ω_3 refer to the estimated dN/dS values and parenthesis indicate the proportion of codons per ω class.

References

Cochrane G, Alako B, Amid C, Bower L, Cerdeno-Tarraga A, Cleland I, Gibson R, Goodgame N, Jang M, Kay S, et al. 2013. Facing growth in the European Nucleotide Archive. Nucleic acids research 41:D30-35.

Eissa A, Amodeo V, Smith CR, Diamandis EP. 2011. Kallikrein-related peptidase-8 (KLK8) is an active serine protease in human epidermis and sweat and is involved in a skin barrier proteolytic cascade. The Journal of biological chemistry 286:687-706.

Foote AD, Vijay N, Avila-Arcos MC, Baird RW, Durban JW, Fumagalli M, Gibbs RA, Hanson MB, Korneliussen TS, Martin MD, et al. 2016. Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. Nat Commun 7:11693.

Komatsu N, Saijoh K, Toyama T, Ohka R, Otsuki N, Hussack G, Takehara K, Diamandis EP. 2005. Multiple tissue kallikrein mRNA and protein expression in normal skin and skin diseases. Br J Dermatol 153:274-281.

Komatsu N, Tsai B, Sidiropoulos M, Saijoh K, Levesque MA, Takehara K, Diamandis EP. 2006. Quantification of eight tissue kallikreins in the stratum corneum and sweat. J Invest Dermatol 126:925-929.

Mitsui S, Yamada T, Okui A, Kominami K, Uemura H, Yamaguchi N. 2000. A novel isoform of a kallikrein-like protease, TLSP/hippostasin, (PRSS20), is expressed in the human brain and prostate. Biochem Biophys Res Commun 272:205-211.

Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. 2015. RELAX: detecting relaxed selection in a phylogenetic framework. Mol Biol Evol 32:820-832.