**Supplemental Methods**

**Genome assembly and annotation**

Our high-coverage reference sperm whale (*Physeter macrocephalus*)genome, was obtained from an adult female in the northern Gulf of Mexico, with DNA extracted with a DNAeasy kit (Qiagen) according to manufacturer’s protocol. Our collection of libraries built for sequencing on the Illumina HiSeq2000 consisted of short-insert (200 bp), and mate-pair libraries (insert sizes: 3, 8, and 40 kbp) all sequenced using 100 bp paired-end reads. The combined sequence reads were assembled with the ALLPATHS software (Gnerre et al. 2011) using default parameter settings**.** This draft assembly was gap-filled with a version of Image (Tsai et al. 2010) that was modified for large genomes, and cleaned of contaminated contigs by performing a MegaBLAST (Zhang et al. 2000) of the contigs against bacterial and vertebrate genome databases. Contigs that displayed the best alignment over 50% of length with a different species were removed. Using a genome size estimate of 2.8 Gbp, the total raw sequence depth of Illumina reads was >90×. Our final sperm whale assembly was repeat-masked using WindowMasker prior to gene annotation using the NCBI (National Center for Biotechnology Information) pipeline described here: (<http://www.ncbi.nlm.nih.gov/books/NBK169439/>. We obtained RNAseq data for gene annotation as described below. The relative completeness in terms of expected gene content of the cetacean genomes and their annotated gene sets was assessed using the Benchmarking Universal Single-Copy Ortholog (BUSCO) assessment tool (Simão et al. 2015) with the laurasiatheria\_odb9 lineage dataset that contains 6253 BUSCOs. Software versions were: BUSCO v3.0.0, Augustus v3.2.3, and HMMER v3.1b1. Gene sets were first filtered to select the longest protein-coding transcript per gene. DNA for the four additional samples re-sequenced to medium depth (**Supplementary Table 1)** were extracted from ‘Voyage of the *Odyssey* samples’ using a high-salt procedure as described in Godard et al. (2003). These samples were then sequenced to medium depth (~20-30×) by sequencing paired-end short insert libraries (~300bp) to 125bp length on the Illumina HiSeq X10 instrument. Sex was determined for all samples based on PCR of the SRY gene (Richard et al. 1994). Whole genome resequencing data for all samples are archived in the NCBI SRA as detailed in **Supplementary Table 1**.

**RNA sequencing**

Total RNA was extracted from skin tissues (n=5) collected from Gulf of Mexico sperm whales using Trizol reagent (Invitrogen) according to the manufacturer’s protocol. RNA quality was assessed by electrophoresis using an Agilent 2100 Bioanalyzer (Santa Rosa, CA). RNAseq paired-end data (100 bp length) were generated from Illumina TruSeq cDNA (stranded) libraries using the HiSeq2000 instrument. All RNAseq data are available through the NCBI SRA with accession numbers SRS387918-387922.

**Selection analyses**

Annotated gene sets from NCBI for the sperm whale, common minke whale (*Balaenoptera acutorostrata*), Yangtze river dolphin (*Lipotes vexillifer*), killer whale (*Orcinus orca*), and bottlenose dolphin (*Tursiops truncatus*), were mapped to OrthoDB v8 Eutheria and Cetartiodactyla orthologous groups (a hierarchical catalog of orthologs) (Kriventseva et al. 2015). In addition, OrthoDB v8 already included orthologs from human (*Homo sapiens*), pig (*Sus scrofa*) and cow (*Bos taurus*), which we also used in our analyses (**see Supplemental Table 4**). For tests of natural selection on orthologous gene pairs we followed previously published methods (Foote et al. 2015). In brief, we identified 5,938 sets of 1:1 orthologs across these eight species from NCBI or ENSEMBL 81 (Cunningham et al. 2015). We aligned all coding nucleotide sequences based on translated amino acid sequences using Transalign (Bininda-Emonds 2005) and clustalW (Thompson et al. 1994). Protein sequence alignments were evaluated with the software Gblocks (Talavera and Castresana 2007) to remove poorly aligned regions. Model testing and likelihood ratio tests (LRT) were performed using PAML4*.*0 (Yang 2007) using the following Newick tree as input: (((((((OO, TT), LV), PM),BA),BT),SS), HS); OO *Orcinus or*c*a*, TT *Tursiops truncatus*, LV *Lipotes vexillifer*, PM *Physter macrocephalus*, BA *Balaenoptera acutorostrata*, BT *Bos taurus*, SS *Sus scrofa*, HS *Homo sapiens*. For tests of positive selection we contrasted paired models for branch tests (free ratio vs one ratio model; two ratio vs one ratio model) and branch-site tests (model 1a: nearly neutral vs. model 2a: positive selection; model 7: gamma vs. model 8: gamma & ω). Branch-specific tests were used to identify accelerated rates of positive selection across genes from specific branches of the evolutionary tree. Site-specific tests were used to detect natural selection acting on specific amino acid sites of proteins on a given branch. Significance of LRT results employed a threshold of *p*<0.01.

The potential impact of amino acid site substitutions on protein structure was tested with Provean (Choi and Chan 2015). We assessed canonical pathway enrichment of gene clusters under positive selection using WebGeStalt (Wang et al. 2013) or Reactome (Fabregat et al. 2016). Entrez Gene IDs were input as gene symbols, with the organism of interest set to human using the genome as the reference set. Significant WikiPathways (Kelder et al. 2012) and KEGG Pathways (Kanehisa et al. 2016) were reported using a hypergeometric test, and the significance level was set at *p*<0.05. We implemented the Benjamini and Hochberg (1995) multiple test adjustment to control for false discovery.

**Proteolytic gene annotation**

We mined the sperm whale genome for protease genes, using the BATI (Blast, Annotate, Tune, Iterate) algorithm (http://degradome.uniovi.es/downloads.html). Briefly, a collection of curated human protease sequences (http://degradome.uniovi.es/dindex.html) was compared to the sperm whale genome with the TBLASTN algorithm, using the tbex script. Then, the resulting files were used to predict the locations of sperm whale protease genes with bsniffer and to set the precise exon/intron boundaries with genetuner. Possible novel proteases were predicted by merging all of the TBLASTN hits in one sorted file, which was visually inspected, with bgmix. This last script provides a summary of all the input TBLASTN files. If several hits overlap in a region, only the best one, as assessed by the expect value, is kept. Therefore, visual inspection of this file can also be used to set orthology relationships, and, importantly, to find any support for the existence of a gene. In this regard, to report a gene loss no supporting hit can be found in the bgmix file. To further ensure that assembly artifacts are not mistaken for gene losses, we used the Trace Archive Nucleotide BLAST to search inside RNA-Seq sets SRX220358, SRX220357, SRX220356, SRX220355, SRX220354, SRX220353, SRX220352, SRX220351 and SRX220350. The query in each case was a part of the orthologous sequence of the putatively lost gene in the closest available organism. This strategy was also employed to gather support for important variants affecting proteases by querying the databases with the corresponding flanking sequences in the *P. macrocephalus* genome. However, complete gene losses must be interpreted with care, as under some circumstances these procedures cannot filter all false negatives.

**Gene duplication among segmental duplications genome-wide**

We generated genome wide segmental duplication maps based on a read depth approach as previously described (Alkan et al. 2009). To this end, we produced an extensively repeat masked assembly by masking out all repeats as identified by repeat masker (Smit et al. 1996) and tandem repeat finder (Benson 1999). Additionally, to mask out any potential cryptic repeats not identified by the aforementioned tools, we fragmented the assembly into kmers of size 36 with an offset of 5 bp, and mapped them against the reference by using gem (Marco-Sola et al. 2012), allowing an edit distance of up 2. We retained all possible placements of a given kmer, and masked those with more than 20 placements along the genome. We then fragmented the reads into non-overlapping kmers of 36 bp, and mapped them against the masked reference, retaining all placements up to an edit distance of 2. We then went on call absolute copy number from read depth (RD) in non-overlapping windows of 1 kb of unmasked sequence (i.e. the effective window size may be larger than 1kb). To avoid biases in RD drop off at the boundaries of masked sequence where reads cannot be properly mapped, we also introduced a padding of 36 bp around any masked region. Finally, we calculated a genome wide RD distribution by iteratively excluding windows with extreme RD, and retaining the remaining ones as ‘control regions’. The absolute copy number of a window was then determined using mrCaNaVar (Alkan et al. 2009) as the GC-corrected RD normalized by the median read depth of control regions, and centered to 2, given the ploidity of the species. Due to asymmetrical read depth distributions, we filtered out the individuals from the Seychelles. We then conservatively defined segmental duplications requiring at least 5 consecutive windows with RD above the mean RD in control regions plus 3 standard deviations, allowing for one internal window to be only above 2 standard deviations, and a minimum length of 10 kb. To check for genes falling within segmental duplications, we intersected the union of segmental duplications of all individuals with the gene annotation, requiring at least 60% of the feature to overlap the duplication.

**Inferring global diversity from whole genome resequencing**

For our high-coverage Gulf of Mexico reference genome, and re-sequenced medium coverage genomes, we performed adaptor and quality trimming of the raw sequencing data (short-insert library only, for the Gulf of Mexico sample) using Trimmomatic (v. 0.36) (Bolger et al. 2014) in paired-end mode and using the following parameters: ILLUMINACLIP:<adaptors.fa>:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:20

We mapped the resulting clipped reads onto the reference genome (Physeter\_macrocephalus-2.0.2) using bwa mem (v. 0.7.7-r441) (Li and Durbin 2009; Li 2013) with standard parameters, and removed duplicates using the bammarkduplicates utility from the biobambam package (Tischler and Leonard 2014). The resulting mappings were used to call SNPs with freebayes (v. 0.9.20) (Garrison and Marth 2012) using the following parameters: -m 30 -q 20 -R 0 -S 0 -m 30 -q 20 --no-population-priors -p 2 --report-genotype-likelihood-max. We then generated a callability mask for each individual using the GATK CallableLoci (v. 3.7) (McKenna et al. 2010) utility with default parameters. We conservatively kept only biallelic SNPs within the intersection of callable regions of all individuals, with a minimum quality of 30 and the DP tag set to a value above 4. Additionally, we filtered out regions with variations in depth of coverage >100 fold across the genomes. The total size of the callable genome was thus calculated to be 1,617,616,407 sites, or about 70% of the assembled genome.

To screen for potential contaminants in the sequencing data, we calculated the allelic imbalance (AI) for each heterozygous position in each individual, defined as the proportion of reads that support the alternative genotype observation. Given normal sampling of the genome, and no contamination, the resulting distribution should be roughly symmetrical and centered at 0.5. **Supplementary Figure 1** shows the distribution of AI for each sample. We do not observe major peaks at low frequency of alternative observations or significant deviations of AI per sample that would imply high levels of contamination. We calculated heterozygosity on a per-individual basis using VCFtools (Danecek et al. 2011).

Effective population size was reconstructed using PSMC (Li and Durbin 2011: Version: 0.6.5-r67, with options -N25 -t15 -r5 -p "4+25\*2+4+6"). To run PSMC, we constructed personal diploid genome references for all individuals with samtools mpileup and bcftools call. We retained only loci with a minimum genotype quality of 30, and depths between 7-100× to reduce false positives. Upon examination of the fastqcf file for sample SEY 420021031-063-1, we found some evidence for low quality or possible contamination (some runs show a bimodal GC distribution, typically contamination). This could, in theory, lead to higher heterozygous rates and in consequence a higher *Ne* estimate, therefore we excluded this sample. We performed 100 bootstrap replicates for each individual to check the consistency of the PSMC. Given the unknown mutation rate of sperm whales we used a prior rate calculated for minke whale (Yim et al. 2013). We used an estimate of the sperm whale generation time of 32 years (Taylor et al. 2007). Although using different mutation rates (1.05–2.5 × 10−8 mutations/site/generation) varied estimates of *N*e and timing of population history events, overall demographic history patterns remained similar as seen in other studies (Nadachowska-Brzyska et al. 2015).

**Supplemental Results**

**Genome assembly and annotation**

The high coverage Gulf of Mexico female sperm whale genome was generated from a mix of short-insert and mate pair libraries sequenced with Illumina 2000 100 bp paired end reads, and assembled into 11,711 scaffolds totaling 2.28 Gbp in length, using ALLPATHS. The size of the sperm whale assembly is consistent with that observed for all previously published cetacean genomes (**Supplementary Table 2**) (Lindblad-Toh et al. 2011; Yim et al. 2013; Zhou et al. 2013; Foote et al. 2015; Keane et al. 2015; Kishida et al. 2015). The lower-than-expected (2.80 Gbp) assembly size is likely due to the presence of highly repetitive regions as was noted in the assembly of the bowhead whale (Keane et al. 2015). We estimated that 40.4% of our assembly was comprised of interspersed repeats, similar to previously published cetacean genomes (**Supplementary Table 2**). Utilizing the NCBI gene annotation pipeline (Brown et al. 2015), we built 18,686 protein coding sperm whale gene models, similar to the killer and minke whale counts of 18,183 and 18,470, respectively (Yim et al. 2013; Foote et al. 2015) (**Supplementary Table 5**). A total of 1,241 non-coding RNAs and 4,062 pseudogenes were also annotated. We assessed the genic representation of our assembly by aligning the core eukaryotic genes of the Laurasiatheria lineage (BUSCO; Simão et al. 2015) and Cetartiodactyla known RefSeq transcripts (n=19,608) to the sperm whale reference assembly (**Supplementary Tables 3 and 5**). The sperm whale had comparable transcript and protein coverage to other cetaceans annotated through the NCBI gene annotation pipeline (**Supplementary Table 5**).

**Selection analyses**

Based on OrthoDB v8 the majority of sperm whale genes had orthologs in both human and other cetartiodactyls (**Supplementary Table 4).** Of the 8,108 sperm whale genes with orthologs, 5,938 genes found across all 8 taxa (pig, cow, bottlenose dolphin, human, Yangtze River dolphin, common minke whale, sperm whale and killer whale) were examined for signatures of positive selection. We found 197 genes that passed the significance threshold (<.01) for the codeml branch test, suggesting positive selection within those genes along the sperm whale lineage. In comparison, a total of 279 genes passed the likelihood significance threshold along the minke whale branch (Yim et al. 2013). Out of the 197 genes putatively under positive selection on the sperm whale branch, 45 passed our stringent functional impact tests (<2.5 cutoff) using Provean (**Supplemental File 1**)**.** Some genes (24 in total) were shown to be also under positive selection along the Cetacean branch (Foote et al. 2015), such as the *ANPEP* involved in glutathione metabolism. However, many were unique to the sperm whale. Of the 197 total genes under selection several significant enrichments for genes associated with canonical pathways (KEGG or Wiki), human diseases, or gene ontology (GO) were repeatedly detected, e.g. genes that protect skin barriers and maintain cellular structure (**Table 1**). One example was desmosome genes – genes that are crucial to tissues that experience mechanical stress, such as skin, where multiple gene members were found to be enriched (*PPL, EVPL, DSP* and *DSG3:* **Table 1**).

We next characterized the novel evolution of sperm whale proteolytic enzymes. Overall, most of the predicted losses and gains of protease genes were similar to those already described in minke and bowhead (*Balaena mysticetus*)whales (Yim et al. 2013; Keane et al. 2015). We found multiple instances of loss-of-function in proteases associated with skin function. For instance, the cysteine-protease *CAPN12* has been lost through different, non-overlapping premature stop codons and frameshifts in the sperm whale, bottlenose dolphin, bowhead and minke whale. This might be related to the lack of hair in cetaceans since this protease is preferentially expressed at the cortex of the hair follicle (Dear et al. 2000). Likewise, the serine-protease *KLK8* appears to be absent in all cetaceans through different mechanisms. In sperm whales, this kallikrein displays a complete open-reading frame, but its catalytic site is mutated to a theoretically inactive protease. Finally, another kallikrein, *KLK7*, seems to have been specifically lost in the common ancestor of mysticetes, but not in odontocetes. Both *KLK7* and *KLK8* have been related to skin homeostasis (Kishibe et al. 2007), and some experiments suggest that *KLK8* is involved in skin desquamation in mammals (Kuwae et al. 2002).

We found that several proteases related to the immune system, including *MMP12* and *CASP12*, were lost early during the evolution of cetaceans. Other proteases, such as *TPSAB1* and *MASP2* (**Figure 1**), seem to have been lost in several cetaceans independently. The apoptotic cysteine-protease *CASP3* (McIlwain et al. 2013) has been specifically duplicated in sperm whales in a retrotranscription-involving event. In addition to its putative role in cancer progression, *CASP3* has also been involved in brain physiology as the predominant caspase involved in the cleavage of amyloid-beta 4A precursor protein. Further studies will be required to ascertain the putative sperm whale-specific consequences of this duplication.

The aquatic environment, where cetaceans experience hydrostatic pressure and lack of net weight, must prompt compensatory mechanisms in the control of blood pressure and coagulation to avoid hemostatic accidents. Consistent with this, we have confirmed the lack of both *F12* and *KLKB1*, two serine proteases which participate in the kinin-kallikrein system, in all cetaceans. Furthermore, the related serine proteases *F7*, *TMPRSS11B* and *TMPRSS11F* have been lost in mysticetes, probably related to their role in coagulation. In our data set, *F7* appears to be a functional gene in sperm whales, whereas both *TMPRSS11F* and *TMPRSS11B* have been lost through premature stop codons. Finally, *AMZ2*, *DPEP2* and *DPEP3*, seem to have been lost in sperm whales specifically. These peptidases are expected to participate in the maturation and inactivation of vasoactive peptides, including angiotensin and leukotriene D4. Therefore, blood homeostasis in sperm whales seems to feature different regulation levels compared to other cetaceans, potentially linked to the deeper pressures they encounter on foraging dives in comparison with mysticetes. Together, these changes suggest that the mammalian potential for clotting and blood pressure are excessive in an aquatic environment, and these systems had to be modulated through the loss of proteases implicated in related proteolytic cascades.

**Inferring global diversity from whole genome resequencing**

To better understand patterns of genetic diversity among sperm whales from different ocean basins, we carried out medium coverage (~20-28×) resequencing of two individuals from the Pacific Ocean (Gulf of California), and two individuals from the Indian Ocean (Seychelles). We identified 5.08 million SNPs across all individuals, with values ranging from 1.74-2.66 million SNPs per individual genome, including the reference individual (**Supplementary Table** **1**). In addition to the lower heterozygosity, the Gulf of Mexico sample also had a higher transition/transversion ratio (1.86) than the other samples (≤1.84). Levels of genetic diversity appeared similar between individuals sampled in the Pacific and Indian Oceans.

**Gene duplication**

Among all filtered segmental duplications we found a total of 350 genes, with only 28 having gene ontology for various enrichment analyses. After evaluating the 28 genes for overrepresentation significance among canonical pathways, gene ontology or disease phenotypes we found none reached an FDR of <.7. We observe several genes annotated as pseudogenized. A full list of genes falling within segmental duplications can be found in Supplementary Table 6.

**Supplementary Table 1.** Summary of genome heterozygosity by ocean. Statistics were generated across 1,617,616,407 sites categorized as callable (see Supplementary Methods). “Total SNPs” refers to sites where an individual had at least one allele that differed with the reference genome. “Non-reference homozygote sites” is the subset of total SNPs where both alleles in an individual genome were the alternate alleles. Heterozygotes is the subset of total SNPs where one copy of the reference and one copy of the alternate alleles were present. Sequence depth gives the average sequence depth across all 1,617,616,407 sites. Singletons refers to the subset of total SNPs unique to that genome. Heterozygosity was calculated at a per-base level.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Location/Sample** | **Total SNPs** | **Non-Reference homozygote sites** | **Heterozygotes** | **Sequence depth** | **Singletons** | **Ts/Tv**  **Ratio** | **Heterozygosity** |
| **Atlantic** |  |  |  |  |  |  |  |
| GMX-SRS387925  (SRX220366,  SRX220367) | 1,735,624 | 25,860 | 1,709,764 | 45.6 | 243,439 | 1.86 | 0.00106 |
| **Pacific** |  |  |  |  |  |  |  |
| SC991024-177-1  (SRX2447269, SRX2447272) | 2,597,535 | 816,310 | 1,781,225 | 21.8 | 446,350 | 1.84 | 0.00110 |
| SC991025-181-1  (SRX2447271, SRX2447274) | 2,591,704 | 813,729 | 1,777,975 | 28.4 | 437,971 | 1.84 | 0.00110 |
| Pacific 95% Lower CI | 2,588,905 | 812,490 | 1,776,415 | 18.6 | 433,949 | 1.84 | 0.00110 |
| Pacific 95% Upper CI | 2,600,334 | 817,549 | 1,782,785 | 31.6 | 450,372 | 1.84 | 0.00110 |
| **Indian** |  |  |  |  |  |  |  |
| SEY220012029-053-1  (SRX2447270, SRX2447273) | 2,609,090 | 810,119 | 1,798,971 | 24.8 | 448,938 | 1.84 | 0.00111 |
| SEY420021031 -063-1  (SRX2447268, SRX2447275) | 2,660,142 | 791,470 | 1,868,672 | 20.8 | 462,405 | 1.83 | 0.00116 |
| Indian 95% Lower CI | 2,584,585 | 782,518 | 1,765,515 | 18.9 | 442,474 | 1.83 | 0.00109 |
| Indian 95% Upper CI | 2,684,647 | 819,071 | 1,902,128 | 26.7 | 468,869 | 1.84 | 0.00118 |
| TOTAL 95% Lower CI | 2,574,350 | 784,718 | 1,751,924 | 18.4 | 436,876 | 1.83 | 0.00108 |
| TOTAL 95% Upper CI | 2,642,633 | 817,296 | 1,844,769 | 26.5 | 457,723 | 1.84 | 0.00114 |

**Supplementary Table 2.** Genome assembly contiguity metrics for all cetaceans deposited with NCBI (and accession numbers). Species are ranked by their N50 scaffold length. The sperm whale assembly generated in this current study is highlighted in blue.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Species** | **Assembled size** | **N50 contig length (kbp)** | **N50 scaffold length**  **(Mbp)** | **Total scaffolds** | **Total contigs** | **% Masked with RepeatMasker** |
| *Tursiops truncatus*  (GCA\_001922835.1) | 2.1 | 44 | 26 | 2,648 | 116,651 | 43.6 |
| *Orcinus orca*  (GCA\_000331955.2) | 2.3 | 70 | 12 | 1,667 | 80,099 | 41.0 |
| *Balaenoptera acutorostrata*  (GCA\_000493695.1) | 2.4 | 22 | 12 | 10,776 | 184,072 | 39.0 |
| *Lipotes vexillifer*  (GCA\_000442215.1) | 2.4 | 31 | 2.4 | 30,713 | 155,510 | 42.8 |
| *Physeter macrocephalus*  (GCA\_000472045.1) | 2.3 | 35 | 0.43 | 11,711 | 110,444 | 40.4 |
| *Tursiops truncatus*  (GCA\_000151865.3) | 2.6 | 12 | 0.12 | 240,557 | 554,227 | -- |
| *Balaenoptera bonaerensis*  (GCA\_000978805.1) | 2.2 | 8 | 0.02 | 421,444 | 720,900 | -- |

**Supplementary Table 3.** A summary of gene representation utilizing core eukaryotic genes of the Laurasiatheria lineage aligned against each species reference assembly. See Supplementary Methods for details.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Species** | **Genome or Gene Set** | **Size: Mbps or No. Genes** | **% Complete** | **% Complete Single-Copy** | **% Complete Duplicated** | **% Fragmented** | **% Missing** |
| *Balaenoptera*  *acutorostrata* | GCA\_000493695.1  Bacut-1.0 | 2431.7 | 91.3 | 90.2 | 1.1 | 4.6 | 4.1 |
| NCBI 100 | 18,413 | 98.4 | 97.4 | 1.0 | 1.3 | 0.3 |
| *Lipotes*  *vexillifer* | GCA\_000442215.1  Lvexi-v1 | 2429.2 | 88.9 | 87.6 | 1.3 | 4.1 | 7.0 |
| NCBI 100 | 18,877 | 98.6 | 97.8 | 0.8 | 1.1 | 0.3 |
| *Orcinus*  *orca* | GCA\_000331955.2  Oorca-1.1 | 2372.9 | 91.4 | 90.4 | 1.0 | 3.3 | 5.3 |
| NCBI 101 | 17,558 | 98.8 | 97.8 | 1.0 | 0.7 | 0.5 |
| *Physeter*  *macrocephalus* | GCA\_000472045.1  Pmacr-2.0.2 | 2280.7 | 89.6 | 88.6 | 1.0 | 4.7 | 5.7 |
| NCBI 100 | 18,639 | 94.7 | 93.8 | 0.9 | 4.9 | 0.4 |
| *Tursiops*  *truncatus* | GCA\_000151865.3  Ttrun-1.4 | 2551.4 | 77.3 | 76.6 | 0.7 | 9.2 | 13.5 |
| NCBI 101 | 19,550 | 77.0 | 76.0 | 1.0 | 16.8 | 6.2 |

**Supplementary Table 4**. Counts of orthologs among five cetaceans, four artiodactyls, and human: (a) Number of genes with orthologs in human and at least seven other species, single-copy; (b) Number of genes with orthologs in human and at least seven other species, multi-copy; (c) Number of genes with orthologs in human and fewer than seven other species; (d) Number of genes with orthologs in at least one artiodactyl and one cetacean but without human orthologs; (e) Number of genes with only artiodactyl or cetacean orthologs. The sperm whale sequenced and assembled in this study is highlighted in blue.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Primates** |  |  |  | **Artiodactyla (even-toed ungulates)** |  |  |  |  |  | **Cetacea (whales, dolphins and porpoises)** |  |
|  | Human |  | Alpaca | Pig | Sheep | Cow |  | Minke whale | Sperm whale | Yangtze river dolphin | Bottlenose dolphin | Killer whale |
|  | *(Homo sapiens)* |  | *(Vicugna pacos)* | *(Sus scrofa)* | *(Ovis aries)* | *(Bos taurus)* |  | *(Balaenoptera acutorostrata)* | *(Physeter macrocephalus)* | *(Lipotes vexillifer)* | *(Tursiops truncatus)* | *(Orcinus orca)* |
| (a) | 13,235 |  | 9,112 | 10,879 | 12,650 | 13,024 |  | 13,003 | 12,717 | 13,063 | 11,977 | 13,151 |
| (b) | 3,578 |  | 1,269 | 6,918 | 4,312 | 4,039 |  | 3,608 | 3,954 | 3,475 | 4,792 | 3,235 |
| (c) | 1,752 |  | 549 | 1,478 | 1,185 | 1,584 |  | 673 | 514 | 595 | 397 | 529 |
| (d) | *NA* |  | 136 | 501 | 563 | 428 |  | 336 | 342 | 280 | 357 | 255 |
| (e) | *NA* |  | 59 | 426 | 544 | 559 |  | 17 | 24 | 14 | 25 | 15 |

**Supplementary Table 5**. Gene annotation summary for cetaceans with annotations available through NCBI (with annotation release ID in parentheses). ‘% Transcript coverage’ and ‘% Protein coverage’ refer to the average % coverage of Cetartiodactyla known RefSeq (NM\_/NR\_\_) for transcript, and Cetartiodactyla GenBank for protein alignments, respectively. Species are ranked by % Protein coverage. The sperm whale assembly generated in this current study is highlighted in blue.

|  |  |  |  |
| --- | --- | --- | --- |
| **Species** | **Protein coding genes** | **% Transcript coverage** | **% Protein coverage** |
| *Orcinus orca* (101) | 18,183 | 97.0 | 87.1 |
| *Balaenoptera acutorostrata* (100) | 18,470 | 95.2 | 84.0 |
| *Lipotes vexillifer* (100) | 18,906 | 96.0 | 85.5 |
| *Physeter macrocephalus* (100) | 18,686 | 91.9 | 83.2 |
| *Tursiops truncatus* (101) | 17,096 | 93.5 | 81.8 |

**Supplementary Table 6.**  Genes intersecting at least 60% with large segmental duplications (>10 kb).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ACOT2 | LOC102978094 | LOC102984931 | LOC102991179 | LOC102995950 |
| ADI1 | LOC102978121 | LOC102985023 | LOC102991300 | LOC102996014 |
| BMP8B | LOC102978153 | LOC102985122 | LOC102991329 | LOC102996082 |
| CCDC150 | LOC102978207 | LOC102985123 | LOC102991386 | LOC102996245 |
| CEACAM1 | LOC102978338 | LOC102985267 | LOC102991652 | LOC102996255 |
| CEP57 | LOC102978444 | LOC102985445 | LOC102991670 | LOC102996292 |
| CXCL17 | LOC102978797 | LOC102985710 | LOC102991822 | LOC102996397 |
| EIF2S3 | LOC102978889 | LOC102985929 | LOC102991850 | LOC102996428 |
| FAM120C | LOC102978921 | LOC102986123 | LOC102991889 | LOC102996499 |
| FAM205A | LOC102979301 | LOC102986197 | LOC102991913 | LOC102996522 |
| FAM206A | LOC102979333 | LOC102986255 | LOC102991999 | LOC102996531 |
| FMO5 | LOC102979474 | LOC102986328 | LOC102992008 | LOC102996727 |
| KDM4D | LOC102979489 | LOC102986363 | LOC102992068 | LOC102997006 |
| LOC102973143 | LOC102979574 | LOC102986413 | LOC102992163 | MKRN1 |
| LOC102973212 | LOC102979721 | LOC102986538 | LOC102992175 | NDUFA9 |
| LOC102973252 | LOC102979755 | LOC102986615 | LOC102992178 | NKAP |
| LOC102973680 | LOC102979802 | LOC102986717 | LOC102992250 | NXF2B |
| LOC102973702 | LOC102979889 | LOC102986833 | LOC102992265 | OXCT2 |
| LOC102973704 | LOC102979958 | LOC102986909 | LOC102992273 | PCDHGB5 |
| LOC102973711 | LOC102980074 | LOC102986945 | LOC102992277 | POLD3 |
| LOC102973802 | LOC102980293 | LOC102987070 | LOC102992454 | PPAP2C |
| LOC102973807 | LOC102980499 | LOC102987236 | LOC102992462 | PPP2CB |
| LOC102974172 | LOC102980543 | LOC102987304 | LOC102992730 | PSMD5 |
| LOC102974179 | LOC102980564 | LOC102987374 | LOC102992814 | REST |
| LOC102974281 | LOC102980607 | LOC102987474 | LOC102992834 | RNF214 |
| LOC102974644 | LOC102980745 | LOC102987511 | LOC102992888 | SLC25A32 |
| LOC102974756 | LOC102980761 | LOC102987554 | LOC102992931 | SNRPC |
| LOC102974993 | LOC102980783 | LOC102987569 | LOC102992994 | TCAIM |
| LOC102975029 | LOC102980813 | LOC102987685 | LOC102993197 | TCEAL3 |
| LOC102975326 | LOC102980882 | LOC102987789 | LOC102993290 | TDRD1 |
| LOC102975371 | LOC102980899 | LOC102987792 | LOC102993368 | TMEM42 |
| LOC102975401 | LOC102981104 | LOC102987846 | LOC102993439 | TRNAA-UGC |
| LOC102975441 | LOC102981531 | LOC102987859 | LOC102993446 | TRNAC-GCA |
| LOC102975523 | LOC102981796 | LOC102987969 | LOC102993544 | TRNAC-GCA |
| LOC102975557 | LOC102981924 | LOC102988008 | LOC102993656 | TRNAE-UUC |
| LOC102975588 | LOC102981927 | LOC102988236 | LOC102993711 | TRNAE-UUC |
| LOC102975615 | LOC102982014 | LOC102988404 | LOC102993723 | TRNAE-UUC |
| LOC102975617 | LOC102982319 | LOC102988426 | LOC102993753 | TRNAE-UUC |
| LOC102975669 | LOC102982383 | LOC102988439 | LOC102993801 | TRNAE-UUC |
| LOC102975812 | LOC102982400 | LOC102988479 | LOC102993821 | TRNAE-UUC |
| LOC102975883 | LOC102982402 | LOC102988527 | LOC102993898 | TRNAE-UUC |
| LOC102975940 | LOC102982657 | LOC102988598 | LOC102993929 | TRNAE-UUC |
| LOC102975965 | LOC102982702 | LOC102988792 | LOC102993932 | TRNAE-UUC |
| LOC102975966 | LOC102982874 | LOC102988842 | LOC102993987 | TRNAE-UUC |
| LOC102975999 | LOC102983088 | LOC102988987 | LOC102993988 | TRNAE-UUC |
| LOC102976077 | LOC102983095 | LOC102988994 | LOC102993992 | TRNAE-UUC |
| LOC102976092 | LOC102983376 | LOC102989027 | LOC102994028 | TRNAE-UUC |
| LOC102976175 | LOC102983468 | LOC102989103 | LOC102994074 | TRNAE-UUC |
| LOC102976364 | LOC102983668 | LOC102989177 | LOC102994088 | TRNAE-UUC |
| LOC102976378 | LOC102983730 | LOC102989326 | LOC102994116 | TRNAG-CCC |
| LOC102976392 | LOC102983762 | LOC102989375 | LOC102994138 | TRNAG-CCC |
| LOC102976418 | LOC102983823 | LOC102989384 | LOC102994268 | TRNAG-CCC |
| LOC102976464 | LOC102983994 | LOC102989513 | LOC102994356 | TRNAG-CCC |
| LOC102976490 | LOC102984038 | LOC102989709 | LOC102994405 | TRNAG-UCC |
| LOC102976547 | LOC102984170 | LOC102989806 | LOC102994440 | TRNAG-UCC |
| LOC102976672 | LOC102984321 | LOC102989821 | LOC102994506 | TRNAG-UCC |
| LOC102976752 | LOC102984410 | LOC102989934 | LOC102994556 | TRNAG-UCC |
| LOC102976765 | LOC102984414 | LOC102989951 | LOC102994592 | TRNAQ-CUG |
| LOC102976913 | LOC102984442 | LOC102990068 | LOC102994636 | TRNAR-CCU |
| LOC102976986 | LOC102984447 | LOC102990088 | LOC102994674 | TRNAR-CCU |
| LOC102977009 | LOC102984450 | LOC102990102 | LOC102994841 | TRNAV-AAC |
| LOC102977132 | LOC102984473 | LOC102990200 | LOC102994882 | TRNAV-AAC |
| LOC102977216 | LOC102984544 | LOC102990220 | LOC102994887 | TRNAV-UAC |
| LOC102977259 | LOC102984565 | LOC102990450 | LOC102994921 | TRNAW-CCA |
| LOC102977324 | LOC102984611 | LOC102990547 | LOC102995165 | TRNAW-CCA |
| LOC102977416 | LOC102984682 | LOC102990725 | LOC102995462 | VWA2 |
| LOC102977542 | LOC102984705 | LOC102990738 | LOC102995478 | ZNF211 |
| LOC102977687 | LOC102984723 | LOC102990847 | LOC102995492 | ZNF304 |
| LOC102977705 | LOC102984838 | LOC102991040 | LOC102995519 | ZNF548 |
| LOC102977747 | LOC102984839 | LOC102991110 | LOC102995740 | ZNF77 |

**Suppementary Figure 1**: Distribution of allelic imbalance for all heterozygous calls, defined as proportion of reads containing the alternative observation. Samples with the PMMB-SC prefix originate from the Pacific (Gulf of California), and samples with the PMBB-SEY prefix originate from the Indian Ocean (Seychelles). The sample with the SRS prefix is the Gulf of Mexico sample.

Macintosh HD:Users:lukas:PROJECTS:SPERM_WHALE:allelic_imbalance_incRef.pdf

**Supplementary Material References**

Alkan C., Kidd J.M., Marques-Bonet T., Aksay G., Antonacci F., Hormozdiari F., Kitzman J.O., Baker C., Malig M., Mutlu O., Sahinalp S.C., Gibbs R.A., Eichler E.E. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 41(10):1061-1070.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57:289–300. Wiley for the Royal Statistical Society.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999 27(2):573-80.

Bininda-Emonds, O. R. P. 2005. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. BMC Bioinformatics 6:156.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Brown, G. R., V. Hem, K. S. Katz, M. Ovetsky, C. Wallin, O. Ermolaeva, I. Tolstoy, T. Tatusova, K. D. Pruitt, D. R. Maglott, and T. D. Murphy. 2015. Gene: A gene-centered information resource at NCBI. Nucleic Acids Res. 43:D36–D42.

Choi, Y., and A. P. Chan. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31:btv195.

Croft, D., A. Mundo, R. Haw, and M. Milacic. 2014. The Reactome pathway knowledgebase. Nucleic acids 42:D472–D477.

Cunningham, F., M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, B. L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. J. Searle, G. Spudich, S. J. Trevanion, A. Yates, D. R. Zerbino, and P. Flicek. 2015. Ensembl 2015. Nucleic Acids Res. 43:D662–D669.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Dear, T. N., N. T. Meier, M. Hunn, and T. Boehm. 2000. Gene structure, chromosomal localization, and expression pattern of Capn12, a new member of the calpain large subunit gene family. Genomics 68:152–160.

Fabregat, A., K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, L. Matthews, B. May, M. Milacic, K. Rothfels, V. Shamovsky, M. Webber, J. Weiser, M. Williams, G. Wu, L. Stein, H. Hermjakob, and P. D’Eustachio. 2016. The reactome pathway knowledgebase. Nucleic Acids Res. 44:D481–D487.

Foote, A. D., Y. Liu, G. W. C. Thomas, T. Vinař, J. Alföldi, J. Deng, S. Dugan, C. E. van Elk, M. E. Hunter, V. Joshi, Z. Khan, C. Kovar, S. L. Lee, K. Lindblad-Toh, A. Mancia, R. Nielsen, X. Qin, J. Qu, B. J. Raney, N. Vijay, J. B. W. Wolf, M. W. Hahn, D. M. Muzny, K. C. Worley, M. T. P. Gilbert, and R. A. Gibbs. 2015. Convergent evolution of the genomes of marine mammals. Nat. Genet. 47:272–5.

Garrison, E., and G. Marth. 2012. Haplotype-based variant detection from short-read sequencing. arXiv Prepr. arXiv1207.3907 9.

Gnerre, S., I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U. S. A. 108:1513–8.

Godard, C., R. Clark, I. Kerr, P. T. Madsen, and R. Payne. 2003. Preliminary report on the sperm whale data collected during the Voyage of the Odyssey. Pap. SC/55/O17 Present. to IWC Sci. Committee, May 2003, Berlin 8pp.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44:D457–D462.

Keane, M., J. Semeiks, A. E. Webb, Y. I. Li, V. Quesada, T. Craig, L. B. Madsen, S. van Dam, D. Brawand, P. I. Marques, P. Michalak, L. Kang, J. Bhak, H.-S. Yim, N. V. Grishin, N. H. Nielsen, M. P. Heide-Jørgensen, E. M. Oziolor, C. W. Matson, G. M. Church, G. W. Stuart, J. C. Patton, J. C. George, R. Suydam, K. Larsen, C. López-Otín, M. J. O’Connell, J. W. Bickham, B. Thomsen, and J. P. De Magalhães. 2015. Insights into the Evolution of Longevity from the Bowhead Whale Genome. Cell Rep. 10:112–122.

Kelder, T., M. P. Van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico. 2012. WikiPathways: Building research communities on biological pathways. Nucleic Acids Res. 40.

Kishibe, M., Y. Bando, R. Terayama, K. Namikawa, H. Takahashi, Y. Hashimoto, A. Ishida-Yamamoto, Y. P. Jiang, B. Mitrovic, D. Perez, H. Iizuka, and S. Yoshida. 2007. Kallikrein 8 is involved in skin desquamation in cooperation with other kallikreins. J. Biol. Chem. 282:5834–5841.

Kishida, T., J. G. M. Thewissen, T. Hayakawa, H. Imai, and K. Agata. 2015. Aquatic adaptation and the evolution of smell and taste in whales. Zool. Lett. 1:9.

Kriventseva, E. V., F. Tegenfeldt, T. J. Petty, R. M. Waterhouse, F. A. Simão, I. A. Pozdnyakov, P. Ioannidis, and E. M. Zdobnov. 2015. OrthoDB v8: Update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. 43:D250–D256.

Kuwae, K., K. Matsumoto-Miyai, S. Yoshida, T. Sadayama, K. Yoshikawa, K. Hosokawa, and S. Shiosaka. 2002. Epidermal expression of serine protease, neuropsin (KLK8) in normal and pathological skin samples. Mol. Pathol. 55:235–41.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr. arXiv 1303.3997.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Li, H., and R. Durbin. 2011. Inference of human population history from individual whole-genome sequences. Nature 475:493–496.

Lindblad-Toh, K., M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, and others. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478:476–482.

Marco-Sola S., Sammeth M., Guigó R., Ribeca P. 2012. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods. 9(12):1185-88.

McIlwain, D. R., T. Berger, and T. W. Mak. 2013. Caspase functions in cell death and disease.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Nadachowska-Brzyska, K., C. Li, L. Smeds, G. Zhang, and H. Ellegren. 2015. Temporal dynamics of avian populations during pleistocene revealed by whole-genome sequences. Curr. Biol. 25:1375–1380.

Richard, K. R., S. W. McCarrey, and J. M. Wright. 1994. DNA sequence from the SRY gene of the sperm whale (*Physeter macrocephalus*) for use in molecular sexing. Can. J. Zool. 72:873–877.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31(19):3210-2.

Smit, A., Hubley, R., Green, P. 1996. RepeatMasker Open 3.0.

Talavera, G., and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564–577.

Taylor, B. L., S. J. Chivers, J. Larese, and W. F. Perrin. 2007. Generation length and percent mature estimates for IUCN assessments of cetaceans. Adm. Rep. LJ-07-01, Southwest Fish. Sci. Center, 8604 La Jolla Shores Blvd., La Jolla, CA 92038, USA 24 pp.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Tischler, G., and S. Leonard. 2014. biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol. Med. 9:13.

Tsai, I. J., T. D. Otto, and M. Berriman. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 11:R41.

Wang, J., D. Duncan, Z. Shi, and B. Zhang. 2013. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res 41:W77-83.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yim, H.-S., Y. S. Cho, X. Guang, S. G. Kang, J.-Y. Jeong, S.-S. Cha, H.-M. Oh, J.-H. J.-H. Lee, E. C. Yang, K. K. Kwon, Y. J. Kim, T. W. H. Kim, W. Kim, J. H. Jeon, S. S.-J. Kim, D. H. Choi, S. Jho, H.-M. H. H.-S. Kim, J. Ko, H.-M. H. H.-S. Kim, Y.-A. Shin, H.-J. Jung, Y. Zheng, Z. Wang, Y. Chen, M. Chen, A. Jiang, E. Li, S. Zhang, H. Hou, T. W. H. Kim, L. Yu, S. Liu, K. Ahn, J. Cooper, S.-G. Park, C. P. Hong, W. Jin, H.-M. H. H.-S. Kim, C. Park, K. Lee, S. Chun, P. A. Morin, S. J. O’Brien, H. H. S. Lee, J. Kimura, D. Y. Moon, A. Manica, J. Edwards, B. C. Kim, S. S.-J. Kim, J. Wang, J. Bhak, H. H. S. Lee, and J.-H. J.-H. Lee. 2013. Minke whale genome and aquatic adaptation in cetaceans. Nat. Genet. 46:88–92.

Zhang, Z., S. Schwartz, L. Wagner, and W. Miller. 2000. A greedy algorithm for aligning DNA sequences. J. Comput. Biol. 7:203–214.

Zhou, X., F. Sun, S. Xu, G. Fan, K. Zhu, X. Liu, Y. Chen, C. Shi, Y. Yang, Z. Huang, J. Chen, H. Hou, X. Guo, W. Chen, Y. Chen, X. Wang, T. Lv, D. Yang, J. Zhou, B. Huang, Z. Wang, W. Zhao, R. Tian, Z. Xiong, J. Xu, X. Liang, B. Chen, W. Liu, J. Wang, S. Pan, X. Fang, M. Li, F. Wei, X. Xu, K. Zhou, J. Wang, and G. Yang. 2013. Baiji genomes reveal low genetic variability and new insights into secondary aquatic adaptations. Nat. Commun. 4:2708.