A genome and single-nucleus cerebral cortex transcriptome atlas of the short-finned pilot whale *Globicephala macrorhynchus*

Kang, Hui; Liu, Qun; Seim, Inge; Zhang, Wenwei; Li, Hanbo; Gao, Haiyu; Lin, Wenzhi; Lin, Mingli; Zhang, Peijun; Zhang, Yaolei; Gao, Haoyang; Wang, Yang; Qin, Yating; Liu, Mingming; Dong, Lijun; Yang, Zixin; Zhang, Yingying; Han, Lei; Fan, Guangyi; Li, Songhai

Published in:
Molecular Ecology Resources

DOI:
10.1111/1755-0998.13775

Publication date:
2023

Document version
Peer reviewed version

Citation for published version (APA):
A genome and single-nucleus cerebral cortex transcriptome atlas of the short-finned pilot whale *Globicephala macrorhynchus*

Hui Kang1,2,*, Qun Liu3,4,5,6,*, Inge Seim7,*, Wenwei Zhang8,*, Hanbo Li3,4, Haiyu Gao1,2, Wenzhi Lin1, Mingli Lin1, Peijun Zhang1, Yaolei Zhang3, Haoyang Gao3, Yang Wang3, Yating Qin3, Mingming Liu1, Lijun Dong1, Zixin Yang1, Yingying Zhang3, Lei Han8, Guangyi Fan3,8,†, Songhai Li1,9,†

1Marine Mammal and Marine Bioacoustics Laboratory, Institute of Deep-sea Science and Engineering, Chinese Academy of Sciences, Sanya 572000, China
2University of Chinese Academy of Sciences, Beijing 100049, China
3BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China
4Lars Bolund Institute of Regenerative Medicine, Qingdao-Europe Advanced Institute for Life Sciences, BGI-Qingdao, Qingdao 266555, China
5Department of Biology, University of Copenhagen, Copenhagen 2100, Denmark
6Qingdao Key Laboratory of Marine Genomics, BGI-Qingdao, Qingdao 266555, China
7Integrative Biology Laboratory, College of Life Sciences, Nanjing Normal University, Nanjing 210046, China
8State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China
9Lead Contact
†Correspondence:
lish@idsse.ac.cn (S.L.),
fanguangyi@genomics.cn (GF.)

*These authors contributed equally to this work.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.13775
ABSTRACT

Cetaceans (dolphins, whales, and porpoises) have large and anatomically sophisticated brains. To expand our understanding of the cellular makeup of cetacean brains and the similarities and divergence between the brains of cetaceans and terrestrial mammals, we here report a short-finned pilot whale (*Globicephala macrorhynchus*) single-nucleus transcriptome atlas. To achieve this goal, we assembled a chromosome-scale reference genome spanning 2.25 Gb on 22 chromosomes and profiled the gene expression of five major anatomical cortical regions of the short-finned pilot whale by single-nucleus RNA-sequencing (snRNA-seq). We identified six major cell lineages in the cerebral cortex (excitatory neurons, inhibitory neurons, oligodendrocytes, oligodendrocyte precursor cells, astrocytes, and endothelial cells), eight molecularly distinct subclusters of excitatory neurons, and four subclusters of inhibitory neurons. Finally, a comparison of snRNA-seq data from the short-finned pilot whale, human, and rhesus macaque revealed a broadly conserved cellular makeup of brain cell types. Our study provides genomic resources and molecular insights into cetacean brain evolution.

KEYWORDS: *Globicephala macrorhynchus*; cetacean; genome assembly; cerebral cortex; single-nucleus RNA-sequencing; cross-species analysis
1 INTRODUCTION

The first cetaceans emerged during the early Paleocene around 55.5 million years ago (Yuan et al., 2021). Their terrestrial-to-aquatic transition is considered a seminal macroevolutionary event. This transition is manifested by changes to the cetacean brain, as illustrated by several histological, radiological, and morphological appraisals (Oelschläger et al., 2008; Sacchini et al., 2022; Smaers et al., 2021). The cetacean brain is a blend of early mammalian conservative features and uniquely derived characteristics (Glezer et al., 1988). Idiosyncratic changes include a reduction of the olfactory system and adaptations to the vision and auditory systems (Ridgway, 1990). Cetaceans also show apparent convergence with primates in terms of neurobiological features, such as social cognition and interaction (Franchini, 2021; Marino, 2002). Cetaceans and primates diverged from a common mammal ancestor around 90-95 million years ago (Bromham et al., 1999). Given the apparent similarity in brain structure and function, juxtaposing cetaceans and primates is of great interest. To date, such studies have relied on phenotypic (fossil and morphological features) or single-gene data (Marino, 2002; McGowen et al., 2011).

Recent RNA-seq studies have assessed the expression of genes in the cetacean brain, either in brain regions (i.e., traditional bulk-level methods) (Krüger et al., 2020) or by isolation of particular cell types (e.g., by laser-capture microdissection) (Geßner et al., 2022). However, methods for direct localization of gene expression are desired for complex tissue such as the brain. High-throughput sequencing techniques now allow the profiling of cell transcriptomes and provide powerful, complementary approaches to anatomical studies (Bakken et al., 2021; Hodge et al., 2019; Tabula Sapiens Consortium, 2022). One such technique, single-nucleus RNA sequencing (snRNA-seq), allows the characterization of distinct cell populations at single-nucleus resolution from frozen
tissue (Bakken et al., 2021; Hodge et al., 2019; Krienen et al., 2020). Transcriptome-defined cell types can also be compared across species (Bakken et al., 2021; Hodge et al., 2019; Sjöstedt et al., 2020). Recent large-scale snRNA-seq studies have investigated cell type classifications, functions, and differentiation in brains of human and non-human primates (Hodge et al., 2019; Zhu et al., 2018).

The cetacean brain is composed of multiple cell types that form distinct functional regions (Hof et al., 2005; Hof & Gucht, 2007; Spocter et al., 2017). Systematic efforts to identify cell types by marker genes provide a fundamental step in unraveling the molecular and cellular network underpinning the cetacean brain. To address this challenge, here we generated a chromosome-level genome assembly and profiled single-nuclei brain transcriptomes of the short-finned pilot whale (*Globicephala macrorhynchus*), a highly social species with a pan-tropical and pan-temperate distribution (Mahaffy et al., 2015). As a member of the toothed whale superfamily Delphinoidea, the short-finned pilot whale possesses a large and complex brain (Smaers et al., 2021). Although transcriptome analysis of cetacean brains may be an arduous task due to the difficulty of collecting and preserving such rare and large specimens from stranded individuals, we were able to create a high-integrity single-nucleus transcriptome atlas of the *G. macrorhynchus* cerebral cortex. We also compared its snRNA-seq profile with the human and rhesus macaque (*Macaca mulatta*) (see (Lake et al., 2017; Zhu et al., 2018)) to illuminate conserved and divergent cell type features across a large evolutionary distance.
2 MATERIALS AND METHODS

2.1 Ethical statement

Muscle tissues and brain tissues were collected after the death of a female short-finned pilot whale stranded in Sanya, Hainan, in June 2019. Sample collection and use protocols were approved by the Institute of Deep-sea Science and Engineering, Chinese Academy of Sciences (ethical statement number: IDSSE-SYLL-MMMBL-01). All experiments in this study were approved by the Institutional Review Board on the Ethics Committee of BGI.

2.2 Genomic DNA preparation and genome sequencing

The muscle tissue of short-finned pilot whale was stored and transported on ice until frozen storage or tissue dissociation to minimize degradation. DNA was extracted using the cetyltrimethylammonium bromide (CTAB) (Thomas et al., 1997). DNA quality control was performed by using Pulsed-Field Gel Electrophoresis Systems (CHEF MAPPER Bio-Rad). DNA concentration and integrity were verified using an Agilent Genomic DNA ScreenTape (Waldbronn, Germany). High molecular weight genomic DNA was used to prepare a single tube Long Fragment Read (stLFR) library (Wang et al., 2019) with an MGIEasy stLFR Library Prep kit (PN: 1000005622). The resulting library was sequenced on the BGISEQ-500 platform. To allow the generation of a chromosome-level assembly, we prepared a Hi-C library (Leberman-Aiden et al., 2009) using the standard protocol and obtained 100-bp reads on the BGISEQ-500 platform.

2.3 Genome assembly

The stLFR data were split into two parts: paired-end 100 bp short reads and the corresponding
barcode information. Next, stLFR barcode information was parsed to generate barcodes compatible with the 10X Genomics format (see https://github.com/BGIQingdao/stLFR2supernova_pipeline) to allow the use of the linked-read assembler Supernova v2.11 (Weisenfeld et al., 2017). Redundans (Pryszcz & Gabaldón, 2016) (parameters: –noscaffolding –nogapclosing) was used to remove possible redundant heterozygous assembled sequences, after which stLFR_GapCloser (https://github.com/BGI-Qingdao/stLFR_GapCloser) (Luo et al., 2012) was used to fill in gaps. For raw data generated using Hi-C libraries, the same HiC-Pro pipeline in the previous study was used (Yuan et al., 2021).

2.4 Genome annotation

Gene annotations were carried out following the procedure described in the previous study (Yuan et al., 2021). Repetitive sequences and protein-coding genes were identified and annotated by homology-based prediction and de novo methods. A final, non-redundant reference gene set was produced by integrating all the above-annotated gene evidence using EVidenceModeler v1.1.1 with default parameters (Haas et al., 2008). Functional annotations of the best alignments in each database were used as the final consensus gene annotation result (Bairoch & Apweiler, 2000; Bru et al., 2005; El-Gebali et al., 2019; Hulo et al., 2006; Kanehisa & Goto, 2000; Mi et al., 2005; Mitchell et al., 2019; Ponting et al., 1999; The Gene Ontology Consortium, 2000; The UniProt Consortium, 2012; Zdobnov & Apweiler, 2001). Pseudochromosomes and genome annotations (DNA TE density, RNA TE density, GC content, and gene density) were visualized using Circos (Krzywinski et al., 2009).
2.5 Genome assembly assessment

In addition to assessing standard assembly statistics (e.g., N50 values), we employed Benchmarking Universal Single-Copy Orthologs (BUSCO) v2.0 (Simão et al., 2015) with the 4,104-gene mammalia_odb9 dataset (default parameters). We also compared the length of mRNAs, coding sequences (CDS), exons, and introns of the short-finned pilot whale with 11 other mammals. To gauge the accuracy of *G. macrorhynchus* genome, the assembly was aligned to the long-finned pilot whale using LASTZ v1.1 (Harris, 2007) (parameters: “T = 2 C = 2 H = 2000 Y = 3,400 L = 6,000 K = 2,200”). After filtering out aligned blocks shorter than 5 kb in length, we plotted the LASTZ data using RectChr v0.69 (https://github.com/BGI-shenzhen/RectChr).

2.6 Phylogenetic analysis

To infer the phylogenetic position of short-finned pilot whale, we obtained the genomes of an outgroup species, Ornithorhynchus anatinus (the platypus), six cetaceans (*Globicephala melas, Peponocephala electra, Grampus griseus, Lagenorhynchus obliquidens, Physeter microcephalus, and Balaenoptera musculus*), a semi-aquatic mammal (*Hippopotamus amphibius*), and three terrestrial mammals (*Bos taurus, Homo sapiens, and Macaca mulatta*) (see Table S1 for details). The longest protein-coding transcript was denoted as the gene representative for each species. We filtered out sequences with coding sequences less than 50 amino acids, premature termination codon(s), and a sequence length not divisible by three (i.e., out of frame). All-versus-all protein similarities were obtained using BLASTP v2.6.0+ (Altschul et al., 1990; Camacho et al., 2009). TreeFam v1.1 (Li et al., 2006) was used to identify gene families (default parameters). We constructed a 12-species phylogenetic tree from 6,078 single-copy orthologous genes using MrBayes v3.2.7a (Huelsenbeck &
Ronquist, 2001) and PhyML v3.0 (Guindon et al., 2009).

2.7 Nuclei extraction from brain tissue

We performed a tissue-level sampling of major functional regions from two hemispheres of the short-finned pilot whale. These included the prefrontal cortex (PFC), the motor cortex (MC), the visual cortex (VC), the auditory cortex (AC), and temporal lobe (TL). Tissues were snap-frozen in liquid nitrogen on-site and later transferred to -80 °C. We extracted the nuclei from brain tissue as previously described (Bakken et al., 2018), with minor modifications. Briefly, frozen tissues were placed in a 2 ml KIMBLE Dounce tissue grinder set (Sigma) with homogenization buffer consisting of 20 mM Tris pH 8.0 (Thermo Fisher Scientific), 500 mM sucrose (Sigma), 50 mM KCl (Thermo Fisher Scientific), 10 mM MgCl₂ (Thermo Fisher Scientific), 0.1% NP-40 (Roche), 0.2U/μl RNase inhibitor (MGI), 1× protease inhibitor cocktail (Roche), 1% nuclease-free BSA, and 0.1 mM DTT (Sigma). Tissues were then homogenized by ten strokes with a loose Dounce pestle. The homogenate was filtered through a 70μm cell strainer (Falcon), ground with a tight Dounce pestle for ten strokes to release the nuclei, filtered through a 30μm cell strainer (Sysmex), and centrifuged at 500 g for 5 minutes at 4 °C to pellet nuclei. Nuclei were resuspended with a blocking buffer containing 1× PBS supplemented with 1% BSA and 0.2 U/μl RNase inhibitor (BGI), then centrifuged at 500 g for 5 minutes, and resuspended with cell resuspension buffer (MGI). The cells were incubated with DAPI (Beyotime, China) for ten min at room temperature, and observed under a fluorescent microscope, and images were obtained and quantified using IMAGEJ software. We counted the nuclei and adjusted the final concentration to 1,000 nuclei/μl for snRNA-seq library construction.
2.8 Single-nucleus RNA-sequencing library construction and sequencing

Single nucleus capturing and cDNA library generation were performed using a 10X Chromium 3’ library construction kit v2 and C4 scRNA-seq kit (MGI). The libraries were pooled and sequenced on the MGISEQ-2000 sequencing platform (Huang et al., 2017) with the following read lengths: 47 bp for read 1, 100 bp for read 2.

2.9 Data processing

Raw sequencing reads were filtered using PISA v1.10.2 (http://github.com/shiquan/PISA) (parameters: “-f -q 4 -dropN”) and aligned to the short-finned pilot whale genome with STAR v2.7.9a (parameters: “--outSAMmultNmax 1”) (Dobin et al., 2013). Cell-gene matrices were generated using PISA.

2.10 Cell type clustering with Seurat

Cell-gene matrices were imported into the R toolkit Seurat v4.0.3 (Stuart et al., 2019) and combined into a Seurat object. Cells were filtered to retain genes expressed in at least three cells from each sample. Cells were further filtered to retain cells with a comparable number of features (> 200) and unique molecular identifiers (UMIs). We performed all downstream analyses in Seurat, including normalization, shared nearest neighbor graph-based clustering, differential expression analysis, and visualization. Doublet cells were classified by DoubletFinder v2.0.3 with default parameters (Mcginnis et al., 2019) and excluded from further analysis. In order to augment the identification of common clusters, we integrated datasets (i.e., performed batch effect correction) using Harmony (https://github.com/immuniogenomics/harmony) (Korsunsky et al., 2019). Clusters were identified
using a community identification algorithm implemented in the Seurat “FindClusters” function (parameters: “resolution=0.2”). Cell clusters were visualized by applying the unsupervised dimensionality reduction method UMAP (uniform manifold approximation and projection) (McInnes et al., 2018) based on plots generated by the Seurat “RunUMAP” function (parameters: “dims = 1:30”). Clusters were classified and annotated based on the expression of markers identified by Seurat “FindAllMarkers” (parameters: “min.pct = 0.25, logfc.threshold = 0.25”). After initial cell type classification, we focused on re-clustering neurons to further delineate excitatory neuron subtypes and inhibitory neuron subtypes. We used t-distributed stochastic neighbor embedding (t-SNE) (Maaten & Hinton, 2008) as a dimensional reduction method to visualize excitatory neuron subclusters since this method may better separate highly similar cell types (Liu et al., 2020).

2.11 DEGs identification and functional enrichment

Differentially expressed genes for each cluster (i.e., cell type) were identified in one-vs-all comparisons using the Seurat “FindAllMarkers” function. DEGs were filtered by log2 (fold-change) and a min.pct > 0.25 cut-off. Gene-set enrichment analysis of the top 100 DEGs was performed using the Metascape webtool (www.metascape.org) (Zhou et al., 2019), which allows statistical analysis and visualization of functional profiles for gene sets. The following datasets were interrogated: Gene Ontology (GO) terms, Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways (Kanehisa & Goto, 2000), Reactome (Jassal et al., 2020), the Pathway Interaction Database (PID) (Schaefer et al., 2009), WikiPathways (Slentner et al., 2018), and NABA Matrisome (genes encoding extracellular matrix components) (Naba et al., 2016; Naba et al., 2012).
2.12 Cross-species cell atlas comparisons

Single-nucleus transcriptomes from human and rhesus macaque were obtained from two recent studies (Lake et al., 2017; Zhu et al., 2018). One-to-one orthologous protein-coding gene pairs with the short-finned pilot whale were identified by aligning genes to EggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) v5.0 with default parameters (Huerta-Cepas et al., 2016). Next, we calculated a PCA matrix with 50 components using these genes and then feed this PCA matrix into the “HarmonyMatrix” function implemented in Harmony with default parameters (Korsunsky et al., 2019) to correct for batch effect differences (here, three since each species dataset come from a single batch). The result was visualized using UMAP (McInnes et al., 2018). To further ensure that the cortical cell types in our snRNA-seq datasets were true “cell-type homologs”, we used the R pipeline TOME (transcriptional similarity using trajectories of mammalian embryogenesis) (Qiu et al., 2022) to perform all possible pairwise cross-species transcriptome comparisons by nonnegative least-squares regression. The inferred relationships between the transcriptomes of major cell types (nodes) and links (species) generated by TOME were visualized using Sankey diagrams generated by the ‘networkD3’ R package v0.4 (Allaire et al., 2017).

To assess the cell-type identity of the six major cortical cell types and their subtypes across the three species datasets, we performed an unsupervised MetaNeighbor analysis (parameters: “pseudocell_size = 50”) (Crow et al., 2018; Fischer et al., 2021). Briefly, MetaNeighbor employs a neighbor voting framework to match cell types from a training dataset to a target dataset, where the matching strength (i.e., similar gene expression profiles) is quantified as Mean Area under the Receiver Operating Characteristic score (AUROC). AUROCs were here calculated based on 620 highly variable one-to-one orthologous genes (out of a total 9971 genes) recurrent across the three
species datasets (short-finned pilot whale vs. human, 912 genes; short-finned pilot whale vs. rhesus macaque, 2042 genes; 1058 human vs. rhesus macaque genes). Meanwhile, the heatmap of the Euclidean distances of cell types and subtypes for three species was generated using the Hclust function within the vegan package of R (default parameters) (Oksanen et al., 2015).

Cell-type level differentially expressed genes (DEGs) that are conserved or divergent across species were identified as follows: scaled, normalized gene count data in the species-integrated Seurat object was interrogated using the Seurat “FindAllMarkers” function (same ‘min.pct’ cut-off as above). For each species, we compared a cell type to the other cell types (e.g., astrocyte-vs-all). DEGs shared (i.e., conserved) or unique (i.e., divergent) to species were illustrated by Venn diagrams generated using the R package ‘VennDiagram’. The expression of DEGs in each major cell type was visualized on heatmaps by plotting scaled, normalized counts using the Seurat “DoHeatmap” function. In addition, we analyzed the average expression of cell identity genes in all three species using scaled data in the Seurat object, including transcription factors (n=1053 genes) and neurotransmitters (n=63 genes) obtained from a recent study (Sjöstedt et al., 2020).

3 RESULTS

3.1 The first assembled and annotated genome of short-finned pilot whale

We sampled muscle tissue from a female short-finned pilot whale stranded in Sanya, China on June 10, 2019. Single tube long fragment read (stLFR) sequencing (Wang et al., 2019) was used to generate 285.84 Gb sequence data (~127 \times depth) and assemble a 2.25 Gb genome with a scaffold N50 of 26.83 Mb (Table S2). We obtained 154.52 Gb of Hi-C sequencing data and anchored 98.73% of the 2.25 Gb assembly onto 22 chromosomes (scaffold N50 of 107.35 Mb) (Figure 1a, b, Table S2)
The 22 chromosomes ranged from 35.48 to 181.39 Mb (Table S3) and were highly homologous to the related long-finned pilot whale (G. melas) (Figure 1c, Table S1) (ASM654740v1; https://www.dnazoo.org/assemblies/Globicephala_melas) (Culibrk et al., 2019). The genome assembly obtained a BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simão et al., 2015) genome completeness score of 90.0%, indicating high completeness and accuracy (Table S4). We predicted 19,587 protein-coding genes, 18,772 of which were supported by homologs in other species (five cetaceans and cow) or public protein databases (Tables S5-S6). Gene features, such as the average length of coding sequences (CDS), were similar to other mammals (Figure 1d, Tables S3, S7). Transposable elements (TEs) accounted for 37.63% of the short-finned pilot whale genome and were universally distributed across the genome (Figure 1b, Table S8). Long interspersed nuclear elements (LINEs) were the most predominant TE subtype (34.23%) (Table S8). Phylogenomic analysis supported that the short-finned pilot whale along with the long-finned pilot whale is closely related to the melon-headed whale (Peponocephala electra), consistent with a recent large-scale phylogenetic analysis of cetaceans (Cabrera et al., 2021) (Figure 1d). Taken together, we generated a genome assembly suitable for single-nucleus transcriptome analyses.

### 3.2 Single-nucleus transcriptome profiling of short-finned pilot whale cortical tissue

Single-nucleus RNA-seq was performed on cortical tissue from the stranded short-finned pilot whale employed for genome sequencing. Sequencing libraries generated from the prefrontal cortex (PFC), temporal lobe (TL), motor cortex (MC), visual cortex (VC), and auditory cortex (AC) (Figures 2a and S1a) were sequenced separately on the 10X Genomics Chromium and MGI DNBelab C4 platforms. After stringent filtering at the cell and gene levels to eliminate potential doublets, we
captured a total of 21,971 nuclei. To visualize cell clusters in our single-nucleus data set, we first applied the unsupervised dimensionality reduction method UMAP (uniform manifold approximation and projection) (McInnes et al., 2018). UMAP revealed six major clusters attributed to canonical cell lineage-specific markers (Figures 2b and c). These were astrocytes (Ast; e.g., SLC1A2 and GFAP) (Lake et al., 2017; Zhong et al., 2018), endothelial cells (End; e.g., IGFBP7 and FLT1) (Darmanis et al., 2015; Fan et al., 2018), excitatory neurons (Ex; e.g., SATB2 and NEFM) (Lake et al., 2016; Rowell et al., 2010), inhibitory neurons (In; e.g., GAD2 and BTBD11) (Bygrave et al., 2021; Shi et al., 2021), oligodendrocytes (Oli; e.g., MBP, OPALIN, MOG, MOBP, and ENPP2) (Darmanis et al., 2015; Hodge et al., 2019; Mizrak et al., 2019; Welch et al., 2019), and oligodendrocyte precursor cells (Opc; e.g., VCAN and TNR) (Lake et al., 2017; Mizrak et al., 2019). The number of unique molecular identifiers (UMIs) and genes in each cell type showed that this dataset passed stringent high-quality filtering and was suitable for snRNA-seq analysis (Figure S1b). We performed gene set enrichment analysis on the top 100 differentially expressed genes of each cell type using Metascape (Zhou et al., 2019), revealing terms associated with expected biological processes (Figure S2a, Tables S9-S10). For example, oligodendrocyte-specific genes were enriched in categories such as “glial cell development” (GO:0021782), “gliogenesis” (GO:0042063), and “myelination” (GO:0042552), and endothelial cell-specific genes were enriched for “blood vessel development” (GO:0001568) and “regulation of endothelial cell migration” (GO:0010594). We next focused on neurons to further delineate excitatory and inhibitory neuron subtypes with a more detailed analysis. The excitatory neurons were further divided into eight subclusters and the inhibitory neurons into four subclusters (Figure S2b). The majority of our defined neuron subtypes were evenly distributed among five cortical regions (Figure S2c-2d). In addition, we employed t-
SNE (t-distributed stochastic neighbor embedding) (Maaten & Hinton, 2008) to visualize eight discrete excitatory neuron cell subtypes (Figure 2d). While considered largely complementary methods, UMAP has been reported to display global relationships while t-SNE better separates highly similar cell types (Liu et al., 2020). We performed a gene enrichment analysis of the top 100 differentially expressed genes in each t-SNE cluster identified (Tables S11-S12) (Satija et al., 2015). While excitatory neuron subtypes in general, as expected, showed enrichment for “neuron projection guidance (GO:0097485)” and “axon development (GO:0061564)” (also see Figure S2a, Table S10), further stratification was observed (Figure 2e). For example, clusters 3 and 6 demonstrate enrichment for the gene ontology term “muscle contraction” (GO: 0006936) – specifically genes hinting at excitatory neuron regulation of action potential, rapid membrane depolarization events (i.e., ryanodine receptors, RyRs, and the sodium-calcium exchanger NCX1 encoded by SLC8A1) (Table S12). For four inhibitory neuron subtypes, we also performed a gene enrichment analysis of the top 100 differentially expressed genes in each subtype, revealing terms associated with expected biological processes (Figure S2b, Tables S13-S14). Notable, cluster 3 shows enrichment of the gene ontology terms "ear development" (GO:0043583) and "visual system development" (GO:0150063), which may be related to the sensory system development of the short-finned pilot whale. These data reveal high-resolution cell populations, extending previous findings on the neuroanatomy of pilot whales (e.g., see (Mortensen et al., 2014)).

3.3 Globally similar cellular diversity in the cerebral cortex of short-finned pilot whales, rhesus macaques, and humans

We next integrated our dataset with single-nucleus cerebral cortex transcriptomes from human and
rhesus macaque (Lake et al., 2017; Zhu et al., 2018). We found that the cell type profile of the short-finned pilot whale is consistent with pooled cell populations from the human and rhesus macaque (Figure 3a, b, and S3a). Among short-finned pilot whale neuronal cells (excitatory neurons and inhibitory neurons) accounted for 61% of the cells in five cortical regions, whereas glial cells (oligodendrocyte progenitor cells, oligodendrocytes, and astrocytes) accounted for 38% (Figure 3b). The cell distribution is similar to the human and the rhesus macaque, although the proportion of glial cells is relatively high in the whale. Previous studies have suggested that cell type similarity in orthologous gene expression overrides species differences (Han et al., 2020). In agreement, we observed that cell types with the same annotation names matched and were highly conserved within the same category (Figure 3c). MetaNeighbor (Crow et al., 2018) employing a shared set of 620 highly variable one-to-one orthologous genes was used to provide a measure of cell subtype and cluster replicability across species. Again, a global similarity was observed (Figure 3d, S3b, and S3c).

As expected, more consistent clustering was observed in the pairwise MetaNeighbor assessment of primates than between the short-finned pilot whale and primates. There were also notable species differences (Figure 3d, S3b, and S3c). For example, the excitatory neuron subtypes Ex7 of short-finned pilot whale and ExN9 of rhesus macaque could be readily differentiated from their closely related subtypes (the former, as observed in the previous within-species analysis) and had a robust (>0.90) pairwise AUROC value, indicative of high cell type similarity. In contrast, these subclusters did not show high similarity to any human excitatory neuron subtype. A heatmap of the Euclidean distances of cell types and subtypes for three species also showed that Ex7 of the short-finned pilot whale has a matching type in rhesus macaque (ExN9), which is similar to the results from MetaNeighbor (Figure S3d).
Because we observed a globally similar cellular diversity, we next focused on identifying genes that may manifest conserved and divergent features of the six molecularly defined cortical cell types (astrocytes, endothelial cells, excitatory neurons, oligodendrocytes, and oligodendrocyte precursor cells) in whales and primates. In other words, we wished to identify gene expression that may underly conservative features and uniquely derived characteristics of the mammalian cerebral cortex. To this end, we considered cell type-specific differentially expressed genes (DEGs) and 1116 genes associated with brain cell type identity (1053 transcription factors (TFs) and 63 neurotransmitters) (Deneris & Hobert, 2014; Sjöstedt et al., 2020).

A small number of genes were conserved and differentially expressed in a particular cell type in all species (Figure 4a-d, Figure S4a, b, Table S15). Excitatory neurons had the lowest number of conserved DEGs (Table S15), in agreement with the above MetaNeighbour analysis. Interestingly, we observed that the conserved DEGs exhibited an ostensibly more cell-type-specific expression pattern (Figure 4b). These conserved DEGs may thus reflect cell types of the cerebral cortex with highly similar physiological properties – the basic mammalian brain architecture. Conserved astrocyte DEGs included the homeobox transcription factor MEIS2 (myeloid ecotropic insertion site 2) (Figure 4c) and the neurotransmitters SLC1A2 and SLC1A3 (Figure 4d). MEIS2 has roles in embryonic and adult neurogenesis and its high expression in astrocytes agrees with recent transcriptome studies on rats (Zhou et al., 2022) and mice (Herrero-Navarro et al., 2021). Conserved DEGs in oligodendrocyte precursor cells included transcription factor SOX6 (Figure 4c), a gene critical for myelin-forming oligodendrocyte development (Wittstatt et al., 2019). Conserved DEGs in inhibitory neurons of all three species included the GABAergic neuron marker glutamate decarboxylase 2 (GAD2), in agreement with recent snRNA-seq studies (Di Bella et al., 2021; Hodge...
et al., 2019) (Figure 4d). DEGs unique to the short-finned pilot whale may underly unique neuroanatomical features of cetaceans and are candidates for adaptations to a fully aquatic environment. For example, the cetacean vascular system, including the cranial vascular system, is highly complex (Lillie et al., 2018; Piscitelli et al., 2010; Racicot et al., 2018). Cetacean-specific endothelial cell DEGs were enriched for gene enrichment categories associated with vasculature such as “vasculature development (GO:0001944),” “regulation of vasculature development (GO:1901342),” and “regulation of angiogenesis (GO:0045765)” (Table S16).

4 DISCUSSION

This report presents a chromosome-level genome assembly of the short-finned pilot whale. The high-quality genome allowed the generation of the first cetacean single-nucleus cerebral cortex transcriptome atlas.

Previous appraisals on brain structure and function illustrate that cell types of the nervous system work together efficiently to maintain processes such as cognition and the motor, auditory, and visual systems (Masland, 2004; Zeng & Sanes, 2017). For example, glial cells provide neuronal and immunological support (Simons & Trajkovic, 2006). Transcriptomics at the cellular level (in this context, both single-cell and -nucleus RNA-sequencing) enables systematic characterization of cellular diversity in the brain, allowing a paradigm shift in neuroscience from the historical emphasis on cellular anatomy to the molecular classification of cell types (Yao et al., 2021; Zeisel et al., 2015; Zhu et al., 2021).

Echoing anatomical studies, profiling the short-finned pilot whale cerebral cortex using snRNA-seq demonstrated great cellular diversity. We transcriptomically defined six major cortical cell types and found that each showed gene expression that likely reflects distinct physiology and
morphology (Table S10). We observed broad cortical cell type conservation between cetaceans and primates (Figure 3). Broad cell type conservation is a phenomenon also observed by recent cellular level cross-species transcriptomics studies on primary motor cortex in human, marmoset, and mouse (Bakken et al., 2021).

Conservation across species, including at the level of a cell type transcriptome, is evidence for functional relevance under evolutionary constraints, while divergence indicates adaptations or drift (Arendt et al., 2016). To narrow down cell-type-specific genes, we considered 1053 transcription factors and 63 neurotransmitters associated with brain cell type identity (Deneris & Hobert, 2014; Sjöstedt et al., 2020). A small number of the 1116 genes assessed were differentially expressed in the six cortical cell types (Figure 4c, d, Figure S4a, b), many known cell type markers. Endothelial cells line the vasculature. Our analysis shows that the genes encoding the oxygen-sensitive α subunits of hypoxia-inducible factors 1 (HIF1A) and 2 (EPAS1), master regulators of the cellular response to hypoxia (Kunej, 2021; Majmundar et al., 2010), are endothelial cell DEGs in the cetacean cortex (Figure 4c). While EPAS1 is an endothelial cell-specific DEG in all species examined here, HIF1A is only a DEG in the short-finned pilot whale. Several studies have shown that gene-level changes of HIF α-subunits mediate hypoxia tolerance in mammals (Allen & Vázquez-Medina, 2019), including cetaceans (Bi et al., 2015). Taken together, we propose that the co-expression of HIF1A and EPAS1 in cetacean cortical endothelial cells underlies the remarkable hypoxia tolerance of this group of mammals. We also observed enrichment for categories such as “myelination (GO:0042552)”, “gliogenesis (GO:0042063)”, and “glial cell differentiation (GO:0010001)” by oligodendrocyte-specific DEGs of the short-finned pilot whale (Table S16). This enrichment may be associated with the fast-conducting axons of cetaceans since myelination affects conduction times in vertebrate
axons (Chinea, 2017). Thus, these genes may help explain studies showing that the auditory evoked potentials in the central nervous system of dolphins have shorter latencies compared to the human brain (Ridgway et al., 1981) and a faster cognitive processing and voluntary acoustic response in dolphins compared to humans (Ridgway, 2011).

As described above, we observed that most transcription factors and neurotransmitters associated with brain cell type identity showed evidence of divergence. One cause of this observation may be that the three datasets employed here are from independent investigators. Thus, sample processing and sequencing differences may mask bona fide gene expression changes (Shekhar & Sanes, 2021). While such experimental noise could result in an underestimation of conserved cell-type-specific gene expression, we contend that our results are broadly robust. Instead, we argue that our results illustrate the large phylogenetic distance (90-95 million years) and likely the core transcriptional program of the mammalian brain (i.e., cell-type gene expression present in the last common ancestor of the three species). While primates and cetaceans have remarkably complex brains, and there is apparent convergence of anatomical structure and neurobiological features, there are equally dramatic differences between these groups of mammals (Franchini, 2021).

The remarkable evolutionary transition from a terrestrial ancestor ~55.5 million years ago (Yuan et al., 2021) to the fully-aquatic lifestyle of today most certainly reshaped the cetacean brain. Larger-scale comparative studies of shared and unique neuroanatomy of primates, cetaceans, and elephants are required to assess the degree to which mammalian brain evolution (in particular, intelligence) is manifested as molecular convergence [e.g., at the gene or gene pathway level. See (Wu et al., 2020) for a recent review]. We also appreciate that the cerebral cortex (and indeed the entire brain) is anatomically complex. While single-cell and -nucleus RNA-seq have provided a wealth of insights...
on gene expression at the cellular level, emerging technologies may further reconcile cross-species differences and similarities between cetaceans and other mammals. For example, SpaTial Enhanced REsolution Omics-sequencing (Stereo-seq) (Chen et al., 2022) can resolve the cerebral cortex at near single-cell spatial resolution (Maynard et al., 2021; Ortiz et al., 2020).

In summary, we have generated a short-finned pilot whale genome assembly and gene set, which is, to our knowledge, the first public available genome resource for the short-finned pilot whale. This reference genome lays the foundation for deepening our understanding of the evolutionary and comparative genomics of the short-finned pilot whale. The single-nucleus cerebral cortex transcriptome resource presented provides an important step toward gaining insights into the molecular underpinnings of the cetacean brain.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

ACKNOWLEDGEMENTS
We are very grateful to Luka Culibrk and Dr. Steven Jones from Canada's Michael Smith Genome Sciences Centre for providing the long-finned pilot whale assembly, and all colleagues and students for their assistance during sampling. We thank Fujian Tan for providing technical assistance and useful discussions. This work was financially supported by the Major Science and Technology Project of Hainan Province (No. ZDKJ2019011 to S.L.), the National Natural Science Foundation of China (Nos. 42225604 and 41422604 to S.L.), the “One Belt and One Road” Science and Technology Cooperation Special Program of the International Partnership Program of the Chinese
Academy of Sciences (No. 183446KYSB20200016 to S.L.), the Key Deployment Project of Center for Ocean Mega-Science of the Chinese Academy of Sciences (No. COMS2020Q15 to S.L.), the Specially-Appointed Professor Program of Jiangsu Province (to I.S.), the Jiangsu Foreign Expert Bureau (to I.S.), and the Jiangsu Provincial Department of Technology (grant JSSCTD202142, to I.S.).

AUTHOR CONTRIBUTIONS

Songhai Li conceived and led the study. Haiyu Gao, Wenzhi Lin, Mingli Lin, Lijun Dong, and Zixin Yang performed sample collection. Yang Wang and Yating Qin performed sequencing libraries construction. Hui Kang, Qun Liu, and Haoyang Gao performed data analysis and manuscript writing with input of the other authors. Inge Seim, Guangyi Fan, Wenwei Zhang, Hanbo Li, Yaolei Zhang, Wenzhi Lin, Peijun Zhang, and Lei Han performed manuscript amending. All authors read and approved the manuscript.

DATA ACCESSIBILITY STATEMENT

The data including whole genome sequencing reads, Hi-C data, genome assembly, repetitive sequence and gene annotation, snRNA-seq reads, cell-gene matrix, and cell metadata (cell types, cluster labels for single nuclei, and the coordinates of dimensional reduction results) in this study are public and available in the China National GeneBank Nucleotide Sequence Archive (CNSA: https://db.cngb.org/cnsa; accession number CNP0003161).
Figure 1 Assembly and annotation statistics of Short-finned pilot whale Genomes. (a) Hi-C heatmap based on the chromosome-scale assembly of the short-finned pilot whale genome and length statistics of each chromosome. (b) Circles from the outside inwards: (I) pseudochromosomes, (II) the number of DNA TEs, (III) the number of RNA TEs, (IV) the GC content, and (V) gene density. These density metrics were calculated with 1-Mb sliding windows. (c) Genomic syntenic blocks between short-finned pilot whales and long-finned pilot whales are shown (SP: short-finned pilot whale, LP: long-finned pilot whale). (d) Left, a phylogenetic tree constructed from 6,078 single-copy genes of 12 species, with Ornithorhynchus anatinus (platypus) as an outgroup. Middle, the percentage of mRNAs with different length ranges for each species. Right, the average lengths of the CDSs, exons, and introns of the species.

Figure 2 Overview of the experimental approach and the short-finned pilot whale snRNA-seq cerebral cortex dataset. (a) Anatomical localization of analyzed brain regions within the short-finned pilot whale cerebral cortex. (b) Uniform Manifold Approximation and Projection (UMAP) of cortical cell types. Cells are color-coded by cluster cell type. Ast, astrocyte; End, endothelial cell; Ex, excitatory neuron; Oli, oligodendrocyte; In, inhibitory neuron; Opc, oligodendrocyte precursor cell. (c) Gene expression dot plot showing the relative expression of cell-type marker genes (y-axis) across all clusters (x-axis). The color intensity of dots represents the average expression level (standardized to a Z-score), the size of dots the proportion (%) of cells expressing a gene. (d) (Left) t-distributed stochastic neighbor embedding (t-SNE) projection of excitatory neuron cell types. (Right) The relative proportions of excitatory neuron subtypes. (e) Heatmap of excitatory neuron subtypes gene expression and their corresponding Metascape GO biological processes (BP) enrichment. Left, scaled, normalized gene counts of the top 100 differentially expressed genes in one-vs-all comparisons. Right, GO BP enrichment. Benjamin-Hochberg-corrected P-values derived by a hypergeometric test are shown.

Figure 3 Integration of single-nucleus transcriptome data of short-finned pilot whale, human, and rhesus macaque reveals extensive conservation of cell types. (a) UMAP of integrated short-finned pilot whale, human, and rhesus macaque cortical cells colored by species. (b) UMAP of integrated short-finned pilot whale, human, and rhesus macaque cortical cells colored by annotated cell types. The bar chart displays the cell ratio of each cell type in the three species. (c) Sankey diagram showing the connectivity between cortical cell types from the short-finned pilot whale (left), rhesus macaque (middle), and human (right). Cell types as in (B). Box heights and line widths are relative to the number of cells in each cell type and connection, respectively. (d) Heatmap showing the similarity between short-finned pilot whale and human cortical cell types. Cell-type similarities are defined as MetaNeighbor AUROC scores, which range between 0 and 1 (a transition of colors from blue to red), where 0 indicates low similarity and 1 high similarity. Ast, astrocyte; End, endothelial cell; Ex, excitatory neuron; In, inhibitory neuron; Oli, oligodendrocyte; Opc, oligodendrocyte precursor cell.

Figure 4 Identification of differentially expressed genes of cortical cell types conserved and divergent across cetaceans and primates. (a) Venn diagrams of cell type-specific differentially expressed genes (DEGs) in three species. For each species, DEGs were identified separately for each of the six cell types in all-vs-all comparisons. HM denotes human; RM, rhesus macaque; SP, short-
finned pilot whale. A separate Venn diagram is shown for each cell type. Ast, astrocyte; End, endothelial cell; Ex, excitatory neuron; Oli, oligodendrocyte; In, inhibitory neuron; Opc, oligodendrocyte precursor cell. (b) Heatmaps of conserved and divergent DEGs from (a). Top, short-finned pilot whale; middle, human; bottom, rhesus macaque. In each heatmap, species are ordered by cell type (rows). (c) Gene expression heatmap of selected conserved and divergent transcription factor DEGs from (a). The expression range between 0 and 1 (a transition of colors from white to red), where 0 indicates low expression and 1 high expression. (d) Gene expression heatmap of selected conserved and divergent neurotransmitter DEGs from (a). Annotated as in (c).
REFERENCES


Mizrak, D., Levitin, H. M., Delgado, A. C., Crotet, V., Yuan, J., Chaker, Z., Silva-Vargas, V., Sims, P. A.,...


layer structure in the ferret. *Journal of Comparative Neurology*, 518(16), 3272-3289. https://doi.org/10.1002/cne.22399


The UniProt Consortium. (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Research, 40(Database issue), D71-D75. https://doi.org/10.1093/nar/gkr981


Figure 1

(a) [Image of a graph showing contact frequency across chromosomes with a color gradient.
(b) [Image of a circular genome view with chromosomes labeled and different colored bands indicating gene density.
(c) [Image of a horizontal view of chromosomes with annotations and colored bands.
(d) [Image of a phylogenetic tree with species names and colored bars representing length in base pairs.]
Figure 3

(a) Human, Rhesus macaque, Short-finned pilot whale

(b) Bar chart showing AUROC values for different categories.

(c) Heatmap and dendrogram showing relationships between categories.

(d) Heatmap with color scale indicating AUROC values.
Figure 4

(a) 

(b) 

Conserved DEGs

Species-specific DEGs

(c) 

(d)