Building genomic Infrastructure

Morin, Phillip A.; Alexander, Alana; Blaxter, Mark; Caballero, Susana; Fedrigo, Olivier; Fontaine, Michael C.; Foote, Andrew D.; Kuraku, Shigehiro; Maloney, Brigid; McCarthy, Morgan L.

Published in:
Marine Mammal Science

DOI:
10.1111/mms.12721

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Building genomic infrastructure: Sequencing platinum-standard reference-quality genomes of all cetacean species

In 2001 it was announced that the 3.1 billion base (gigabase, Gb) human genome had been sequenced, but after 13 years of work and US$2.7 billion in cost, it was still considered to be only a draft. The initial assembly was missing over 30% of the genome and was made up of over 100,000 sequence fragments (scaffolds) with an average size of just 81,500 base pairs (bp) (International Human Genome Sequencing Consortium, 2004; Stein, 2004). As technologies improved, the draft human genome assembly has been repeatedly refined and corrected. By the time the genome assembly was published in 2004, the average length of scaffolds had increased to over 38 million bp (megabases, Mb) with only a few hundred gaps in the chromosome-length scaffolds. However, the duplicated and highly repetitive regions of the human genome remained unresolved due to limitations of short-read sequencing technology that requires piecing the genome together from billions of shorter sequences. Over the last decade, as highly parallel, much less expensive, short- and long-read sequencing technologies have revolutionized genomic sequencing, thousands of individual human genomes have been sequenced, further refining the human genome assembly and characterizing its diversity. Together these genome sequences have produced a "reference-quality" human genome assembly that covers 95% of the genome with far fewer and smaller gaps compared to the initial version. Despite this vast improvement, the human genome continues to be updated and refined (v. 39, RefSeq accession GCF_000001405.39).

This example illustrates how all eukaryotic genome assemblies, even those of exemplar quality, are drafts, varying in sequence quality (i.e., error rate), completeness (i.e., how much of the genome is covered), how contiguous DNA sequences within scaffolds are (i.e., how many gaps), and what portions of the genome remain unresolved or incorrect. The "platinum-standard reference genome" that modern genomics strives for is distinguished from other draft assemblies by completeness, low error rates, and a high percentage of the sequences assembled into chromosome-length scaffolds (Anonymous, 2018; Rhie et al., 2020). For the remainder of this note, we use "draft" to refer to the less complete/contiguous "draftier draft" genomes and "reference-quality genomes" to refer to platinum-standard reference genomes as characterized above.

Democratization of genome sequencing has yielded draft genomes across the diversity of life at a rate that was unimaginable just a few years ago. As genome assemblies have become increasingly common, titles of articles often tout "chromosome-level," "complete," "reference-quality," and other adjectives to characterize the quality of a new genome sequence. These terms offer little information about the level of completion or accuracy of genome assemblies, as even chromosome-level genomes may consist of thousands to millions of sequence fragments (e.g., Fan et al., 2019), with significant amounts of missing data, assembly errors, and missing or incomplete genome annotations.

Nevertheless, the utility of draft genomes has been abundantly documented, and there is no doubt that draft genomes provide sufficient data to address many biological questions. For cetaceans, highly fragmented draft genomes have been useful references for mapping data from resequenced individuals, and thus for characterization of variable markers (Morin et al., 2018), phylogenetics and comparative genomics (Arnason, Lammers, Kumar, Nilsson, & Janke, 2018; Fan et al., 2019; Foote et al., 2015; Yim et al., 2014), characterization of intraspecific variability and demographic history (Autenrieth et al., 2018; Foote et al., 2019; Foote et al., 2016; Morin et al., 2015).
Westbury, Petersen, Garde, Heide-Jørgensen, & Lorenzen, 2019; Zhou et al., 2018), molecular evolution of genes and traits (Autenrieth et al., 2018; Fan et al., 2019; Foote et al., 2015; Springer et al., 2016a; Springer, Starrett, Morin, Hayashi, & Gatesy, 2016b; Yim et al., 2014), epigenetic age estimation (Beal, Kiszka, Wells, & Eirin-Lopez, 2019; Polanowski, Robbins, Chandler, & Jarman, 2014), and skin and gut microbiome metagenomics (Hooper et al., 2019; Sanders et al., 2015). The field of conservation genomics has also demonstrated the many applications of genomics that aid in discovery of vulnerable species, identify extinction risks, and implement appropriate management (Garner et al., 2016; Tan et al., 2019).

However, the types of errors common to draft genomes can be misleading (e.g., structural variation; Ho, Urban, & Mills, 2019), and at worst, result in years of lost time and effort characterizing genes and variants that do not exist (Anderson-Trocme et al., 2019; Korlach et al., 2017). In addition, use of a related species reference genome to map sequencing reads (when the new species genome is not available) reduces and biases mapping of the new species reads, compromising estimates of variation (Gopalakrishnan et al., 2017). The completeness and quality of a genome and of its coding and regulatory annotation (e.g., coding regions and identified genes; Scornavacca et al., 2019) affect downstream interpretation of analytic results. Recently, re-analysis of published genomes has shown that appreciable portions of most genome assemblies (e.g., 4.3 Mb of a sperm whale assembly) contain contaminating sequences (including full genes) from parasites and bacteria (Challis, Richards, Rajan, Cochrane, & Blaxter, 2020; Steinegger & Salzberg, 2020).

Recent improvements in sequencing and bioinformatic technologies and a better understanding of the types of errors that can occur and how to minimize them have changed our view of what is possible in genome assembly, such that now it is credible to propose reference-quality genome sequencing for not just a few model taxa of interest, but rather for whole biomes, whole clades and, ultimately all of the planet’s biota. The Earth BioGenome Project (EBP; Lewin et al., 2018) proposes reference genome sequencing of all eukaryotic life on earth. The EBP goals are reflected in local biotic projects, such as the Darwin Tree of Life project (https://darwintreeoflife.org), which aims to sequence all eukaryotic species in Britain and Ireland (including several cetacean species), and clade-focused projects such as the Genome 10K (Genome 10K Community of Scientists, 2009) and its Vertebrate Genomes Project (VGP; https://vertebrategenomesproject.org), which propose sequencing of all Vertebrata. In an effort to establish benchmark quality standards and best practices for reference-quality genome sequencing, the VGP has developed combined sequencing technologies and assembly protocols (Anonymous, 2018) with criteria for evaluation of genomes to meet platinum-quality standards (Rhie et al., 2020). They find that vertebrate genome assemblies that lead to far fewer errors in biological analyses are those that have a contig N50 (without gaps) of 1 Mb or more; chromosomal scaffold N50 of 10 Mb or more; base call accuracy of Q40 or higher (i.e., no more than one nucleotide error per 10,000 bp); paternal and maternal sequences haplotype phased to reduce false gene duplication errors; and manual curation to improve the genome assembly and reduce errors further. These genome assemblies thus far have up to >95% of the genome assembled into chromosomes, with <1,000 gaps/Gb. Both the VGP and the Darwin Tree of Life projects aim to meet these quality standards for all their genome assemblies.

Such reference-quality genomes for each focal cetacean species would offer a platform for analysis that will avoid the types of errors discussed above and obviate the need for cross-species read mapping that is currently the norm. High-quality genomes make correct gene identification possible (Korlach et al., 2017), help phasing of population genomic data (identifying paternal and maternal chromosomes), contribute to identification of population-level structural variation and permit informed analysis of genome architectures (e.g., centromeric and telomeric regions).

As of December 2019, there were 28 cetacean species present in public sequence databases as draft assemblies, but only two species, the vaquita and the blue whale (Table 1, Figure 1), had VGP platinum-standard reference genome assemblies. The vaquita genome, for example, has 99.92% of the assembly assigned to 22 nearly gapless chromosome-level scaffolds (88 gaps/Gb; 0–35 gaps/scaffold), with accuracy Q40.88 (0.8 nucleotide errors per 10,000 bp) (Morin et al., 2020). By contrast, the sperm whale chromosome-level genome assembly (accession GCA_002837175.2; Fan et al., 2019), built from short shotgun reads, 10X Genomics linked reads and Hi-C scaffolding, assigned 95% of the assembly to 21 chromosomes, but contains 51,366 gaps/Gb. The primary reason for the
### TABLE 1  
Cetacean genome assembly information from assemblies in the NCBI Genome Assembly database (https://ncbi.nlm.nih.gov/genome) and DNAzoo (Assembly ID's ending with "HiC"; https://dnazoo.org/assemblies) as of January 2020. The assembly level “scaffold” refers to both unordered contigs and ordered scaffolds. Contig N50 and Scaffold N50 are measures of assembly quality indicating that half of the genome assembly is found in contigs or scaffolds equal to or larger than the N50 size bp. In addition to contig and scaffold N50 metrics, an assessment of whether a genome assembly meets platinum quality standards also relies on other metrics such as genome-wide base-call accuracy level (≥ Q40, or no more than 1 nucleotide error per 10,000 bp), and phased maternal/paternal haplotypes to reduce false gene duplication errors. Rhie et al. (2020) contains additional detail on VGP assembly methods and platinum genome quality standards.

<table>
<thead>
<tr>
<th>Common name</th>
<th>Species name</th>
<th>Assembly ID</th>
<th>Number of contigs</th>
<th>Contig N50 bp</th>
<th>Number of scaffolds</th>
<th>Scaffold N50 bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antarctic minke whale</td>
<td>Balaenoptera bonaerensis</td>
<td>GCA_000978805.1</td>
<td>720,900</td>
<td>8,410</td>
<td>421,444</td>
<td>20,082</td>
</tr>
<tr>
<td>Baiji</td>
<td>Lipotes vexillifer</td>
<td>GCA_000442215.1</td>
<td>155,510</td>
<td>31,902</td>
<td>30,713</td>
<td>2,419,148</td>
</tr>
<tr>
<td>Beluga</td>
<td>Delphinapterus leucas</td>
<td>ASM228892v2_HiC</td>
<td>35,752</td>
<td>156,270</td>
<td>6,972</td>
<td>107,969,763</td>
</tr>
<tr>
<td>Beluga</td>
<td>Delphinapterus leucas</td>
<td>GCA_002288925.3</td>
<td>29,444</td>
<td>196,689</td>
<td>5,905</td>
<td>31,183,418</td>
</tr>
<tr>
<td>Beluga</td>
<td>Delphinapterus leucas</td>
<td>GCA_009917725.1</td>
<td>101,557</td>
<td>76,763</td>
<td>51,177</td>
<td>1,361,507</td>
</tr>
<tr>
<td>Beluga</td>
<td>Delphinapterus leucas</td>
<td>GCA_009917745.1</td>
<td>52,911</td>
<td>141,056</td>
<td>25,931</td>
<td>3,009,037</td>
</tr>
<tr>
<td>Blue whale</td>
<td>Balaenoptera musculus</td>
<td>GCA_009873245.2</td>
<td>1,050</td>
<td>5,963,936</td>
<td>130</td>
<td>110,470,125</td>
</tr>
<tr>
<td>Boto</td>
<td>Inia geoffrensis</td>
<td>GCA_004363515.1</td>
<td>1,218,682</td>
<td>24,570</td>
<td>1,213,610</td>
<td>26,707</td>
</tr>
<tr>
<td>Bowhead whale</td>
<td>Balaena mysticetus</td>
<td>NA</td>
<td>113,673</td>
<td>877,000</td>
<td>7,227</td>
<td>34,800</td>
</tr>
<tr>
<td>Bryde's whale</td>
<td>Balaenoptera edeni</td>
<td>GCA_00177725.1</td>
<td>184,171</td>
<td>71,244</td>
<td>141,314</td>
<td>99,560,599</td>
</tr>
<tr>
<td>Common bottlenose dolphin</td>
<td>Tursiops truncatus</td>
<td>GCA_000151865.3</td>
<td>554,227</td>
<td>11,821</td>
<td>240,557</td>
<td>116,287</td>
</tr>
<tr>
<td>Common bottlenose dolphin</td>
<td>Tursiops truncatus</td>
<td>GCA_001922835.1</td>
<td>116,651</td>
<td>44,299</td>
<td>2,648</td>
<td>26,555,543</td>
</tr>
<tr>
<td>Common bottlenose dolphin</td>
<td>Tursiops truncatus</td>
<td>GCA_00314715.1</td>
<td>139,544</td>
<td>30,985</td>
<td>481</td>
<td>27,166,507</td>
</tr>
<tr>
<td>Common bottlenose dolphin</td>
<td>Tursiops truncatus</td>
<td>GCA_003435595.3</td>
<td>154,206</td>
<td>27,134</td>
<td>42,644</td>
<td>931,081</td>
</tr>
<tr>
<td>Common bottlenose dolphin</td>
<td>Tursiops truncatus</td>
<td>NIST_Tur_tru_v1_HiC</td>
<td>116,947</td>
<td>44,280</td>
<td>2,646</td>
<td>98,188,383</td>
</tr>
<tr>
<td>Common minke whale</td>
<td>Balaenoptera acutorostrata</td>
<td>GCA_000493695.1</td>
<td>184,072</td>
<td>22,690</td>
<td>10,776</td>
<td>12,843,668</td>
</tr>
<tr>
<td>Cuvier's beaked whale</td>
<td>Ziphius cavirostris</td>
<td>GCA_004364475.1</td>
<td>3,761,505</td>
<td>3,606</td>
<td>3,758,276</td>
<td>3,608</td>
</tr>
<tr>
<td>Fin whale</td>
<td>Balaenoptera physalus</td>
<td>GCA_008795845.1</td>
<td>1,270,025</td>
<td>4,493</td>
<td>62,302</td>
<td>871,016</td>
</tr>
<tr>
<td>Finless porpoise</td>
<td>Neophocaena asiacerentalis</td>
<td>GCA_003031525.1</td>
<td>66,346</td>
<td>86,003</td>
<td>13,699</td>
<td>6,341,296</td>
</tr>
<tr>
<td>Franciscana</td>
<td>Pontoporia blainvillii</td>
<td>GCA_004363935.1</td>
<td>1,885,701</td>
<td>2,541</td>
<td>1,885,058</td>
<td>2,541</td>
</tr>
<tr>
<td>Gray whale</td>
<td>Eschrichtius robustus</td>
<td>GCA_002189225.1</td>
<td>375,256</td>
<td>10,066</td>
<td>57,203</td>
<td>187,455</td>
</tr>
<tr>
<td>Gray whale</td>
<td>Eschrichtius robustus</td>
<td>GCA_002738545.1</td>
<td>1,595,257</td>
<td>2,656</td>
<td>1,213,011</td>
<td>10,674</td>
</tr>
<tr>
<td>Gray whale</td>
<td>Eschrichtius robustus</td>
<td>GCA_004363415.1</td>
<td>1,046,770</td>
<td>68,559</td>
<td>1,036,148</td>
<td>94,414</td>
</tr>
<tr>
<td>Harbor porpoise</td>
<td>Phocoena phocoena</td>
<td>GCA_00177725.1</td>
<td>2,347,235</td>
<td>2,773</td>
<td>142,029</td>
<td>27,499,337</td>
</tr>
<tr>
<td>Harbor porpoise</td>
<td>Phocoena phocoena</td>
<td>GCA_004363495.1</td>
<td>1,338,272</td>
<td>89,111</td>
<td>1,331,158</td>
<td>115,969</td>
</tr>
<tr>
<td>Harbor porpoise</td>
<td>Phocoena phocoena</td>
<td>GCA_003071005.1</td>
<td>2,347,235</td>
<td>2,773</td>
<td>142,029</td>
<td>27,499,337</td>
</tr>
<tr>
<td>Harbor porpoise</td>
<td>Phocoena phocoena</td>
<td>GCA_003071005.1</td>
<td>2,347,235</td>
<td>2,773</td>
<td>142,029</td>
<td>27,499,337</td>
</tr>
<tr>
<td>Humpback whale</td>
<td>Megaptera novocoeangiae</td>
<td>GCA_004329385.1</td>
<td>387,694</td>
<td>12,321</td>
<td>2,558</td>
<td>9,138,802</td>
</tr>
<tr>
<td>Indo-Pacific bottlenose dolphin</td>
<td>Tursiops aduncus</td>
<td>ASM322739v1_HiC</td>
<td>58,538</td>
<td>133,491</td>
<td>12,471</td>
<td>111,961,311</td>
</tr>
<tr>
<td>Indo-Pacific bottlenose dolphin</td>
<td>Tursiops aduncus</td>
<td>GCA_003227395.1</td>
<td>44,281</td>
<td>206,065</td>
<td>16,249</td>
<td>1,235,788</td>
</tr>
</tbody>
</table>
difference between the VGP genomes and the sperm whale genome is the use of long-read sequencing to obtain
475× and 140× larger contig N50s (vaquita and blue whale, respectively; Table 1), allowing assembly of all but the most
difficult regions (e.g., some centromeric and telomeric regions). We are aware of whole-genome shotgun (WGS)
sequencing projects underway for most of the 96 recognized cetacean species (Committee on Taxonomy, 2019). Most
of these projects will result in highly fragmented and incomplete draft genome assemblies that may include >90% of
the genes, but are unlikely to resolve chromosome-level scaffolds, let alone full gene or genome structure. A substantial
effort is underway (http://DNAzoo.org) to improve contiguity in new and existing genome assemblies using proximity-
guided assembly methods (Hi-C; Dudchenko et al., 2017; Lieberman-Aiden et al., 2009). This approach generates
chromosome-level scaffolds, and can yield highly contiguous genomes when long reads are used. When used with
short-read data, this approach is very cost-effective and can be used even with somewhat degraded tissue samples.
However, these genome assemblies remain highly fragmented with regions of unresolved structure (e.g., long or com-
plex repeats) and hence do not meet the reference quality standards recommended by the VGP.

### Table 1 (Continued)

<table>
<thead>
<tr>
<th>Common name</th>
<th>Species name</th>
<th>Assembly ID</th>
<th>Number of contigs</th>
<th>Contig N50 bp</th>
<th>Number of scaffolds</th>
<th>Scaffold N50 bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indo-Pacific humpbacked dolphin</td>
<td>Sousa chinensis</td>
<td>GCA_007760645.1</td>
<td>62,803</td>
<td>113,766</td>
<td>23,368</td>
<td>19,436,979</td>
</tr>
<tr>
<td>Indus river dolphin</td>
<td>Platanista minor</td>
<td>GCA_004363435.1</td>
<td>1,110,492</td>
<td>20,879</td>
<td>1,098,790</td>
<td>23,933</td>
</tr>
<tr>
<td>Killer whale</td>
<td>Orcinus Orca</td>
<td>GCA_000331955.2</td>
<td>80,100</td>
<td>70,300</td>
<td>1,668</td>
<td>12,735,091</td>
</tr>
<tr>
<td>Killer whale</td>
<td>Orcinus Orca</td>
<td>Oorc_1.1_HiC</td>
<td>80,502</td>
<td>70,204</td>
<td>1,617</td>
<td>110,405,485</td>
</tr>
<tr>
<td>Long-finned pilot whale</td>
<td>Globicephala melas</td>
<td>ASM654740v1_HiC</td>
<td>21,252</td>
<td>332,801</td>
<td>6,090</td>
<td>106,927,605</td>
</tr>
<tr>
<td>Long-finned pilot whale</td>
<td>Globicephala melas</td>
<td>GCA_006547405.1</td>
<td>21,236</td>
<td>332,801</td>
<td>6,637</td>
<td>18,102,937</td>
</tr>
<tr>
<td>Melon-headed whale</td>
<td>Peponocephala electra</td>
<td>Peponocephala_electra_HiC</td>
<td>84,924</td>
<td>185,978</td>
<td></td>
<td>102,795,557</td>
</tr>
<tr>
<td>Narwhal</td>
<td>Monodon monoceros</td>
<td>GCA_004026685.1</td>
<td>653,473</td>
<td>67,024</td>
<td>644,873</td>
<td>86,766</td>
</tr>
<tr>
<td>Narwhal</td>
<td>Monodon monoceros</td>
<td>GCA_004027045.1</td>
<td>890,705</td>
<td>70,965</td>
<td>882,704</td>
<td>88,921</td>
</tr>
<tr>
<td>Narwhal</td>
<td>Monodon monoceros</td>
<td>GCA_005125345.1</td>
<td>813,468</td>
<td>10,044</td>
<td>21,006</td>
<td>1,483,363</td>
</tr>
<tr>
<td>Narwhal</td>
<td>Monodon monoceros</td>
<td>GCA_005190385.2</td>
<td>25,295</td>
<td>255,327</td>
<td>6,972</td>
<td>107,566,389</td>
</tr>
<tr>
<td>North Atlantic right whale</td>
<td>Eubalaena glacialis</td>
<td></td>
<td>65,924</td>
<td>172,124</td>
<td>101,413,572</td>
<td></td>
</tr>
<tr>
<td>North Pacific right whale</td>
<td>Eubalaena japonica</td>
<td>GCA_004363455.1</td>
<td>1,361,057</td>
<td>34,866</td>
<td>1,353,963</td>
<td>39,813</td>
</tr>
<tr>
<td>Pacific white-sided dolphin</td>
<td>Sagmatias obliquidens</td>
<td>ASM637639v1_HiC</td>
<td>21,805</td>
<td>255,779</td>
<td>5,162</td>
<td>107,447,310</td>
</tr>
<tr>
<td>Pacific white-sided dolphin</td>
<td>Sagmatias obliquidens</td>
<td>GCA_003676395.1</td>
<td>21,793</td>
<td>255,779</td>
<td>5,422</td>
<td>28,371,583</td>
</tr>
<tr>
<td>Pygmy sperm whale</td>
<td>Kogia breviceps</td>
<td>GCA_004363705.1</td>
<td>1,258,125</td>
<td>26,201</td>
<td>1,252,072</td>
<td>28,812</td>
</tr>
<tr>
<td>Sowerby's beaked whale</td>
<td>Mesoplodon bidens</td>
<td>GCA_004027085.1</td>
<td>1,810,317</td>
<td>28,959</td>
<td>1,801,720</td>
<td>33,532</td>
</tr>
<tr>
<td>Sperm whale</td>
<td>Physeter macrocephalus</td>
<td>GCA_000472045.1</td>
<td>110,443</td>
<td>35,258</td>
<td>11,710</td>
<td>427,290</td>
</tr>
<tr>
<td>Sperm whale</td>
<td>Physeter macrocephalus</td>
<td>GCA_002837175.2</td>
<td>143,605</td>
<td>42,542</td>
<td>14,677</td>
<td>122,182,240</td>
</tr>
<tr>
<td>Sperm whale</td>
<td>Physeter macrocephalus</td>
<td>GCA_900411695.1</td>
<td>140,250</td>
<td>43,829</td>
<td>14,676</td>
<td>122,182,240</td>
</tr>
<tr>
<td>Vaquita</td>
<td>Phocoena sinus</td>
<td>GCA_009692025.1</td>
<td>273</td>
<td>20,218,762</td>
<td>65</td>
<td>115,469,292</td>
</tr>
</tbody>
</table>

aVGP platinum-quality reference genomes.
bFrom Keane et al., 2015, Cell Reports, 10, 112–122.
The critical step needed to meet the platinum-level criteria set out by the VGP is long-read sequencing (e.g., Pacific Biosciences or Oxford Nanopore technologies) which generates contiguous raw data tens to hundreds of kilobases in length. Combined with long-range, chromosome-scale scaffolding methods based on Hi-C chromatin contacts and optical mapping (e.g., BioNano), these data allow repetitive regions within scaffolds to be resolved (Figure 2). While this approach is now becoming feasible even on a moderate research budget, the major limitation for many marine mammals is availability of fresh tissues that yield relatively large amounts of ultra-high-quality DNA for long-read sequencing (>40 kb reads) and BioNano approaches (>300 kb reads) (e.g., Mulcahy et al., 2016) and intact chromatin preserving the 3D structure of the DNA in the nuclei for long-range Hi-C linking to build scaffolds. These technologies currently require fresh blood, muscle or organ tissue, or cultured cells, preserved to maintain megabase-length DNA and, preferably, RNA for gene annotation. Although there are rare exceptions, this usually requires rapid freezing and storage at ≤−80°C or culture of live cells, both of which have limited feasibility for

**FIGURE 1** Phylogeny of the extant cetaceans based on phylogenetic analysis of 3191 protein-coding nuclear loci, reproduced from McGowen et al. (2019) and modified to show phylogenetic positions of species with published genome assemblies. Blue triangles mark the species represented by platinum-quality VGP reference genomes (vaquita and blue whale). Orange triangles mark the species for which draft genomes have been published (from Table 1). Parentheses around the triangles indicate that the species is not shown in this phylogeny (but the triangle is placed near congeneric species to indicate approximate position in the phylogeny). Illustrations by Carl Buell.
protected species (due to sampling methods) and in many field conditions (e.g., mass strandings on remote beaches or locations with scarce infrastructure). Skin samples collected by dart biopsy typically yield too little high-quality DNA unless the cells are cultured. Therefore, collection and preservation of appropriate samples is rare for cetaceans.

Given the manifest benefits of reference-quality genome sequencing from at least one specimen of each species, and the extreme logistical difficulty in obtaining appropriate samples for long-read sequencing methods, we propose that a concerted effort should be made to coordinate and facilitate ethical collection of cetacean samples immediately. We estimate that such samples are currently available for about 25% of cetacean species in a few publicly accessible collections that have already contributed samples for cetacean genomics (e.g., the Frozen Zoo tissue culture collection at San Diego Zoo Global’s Institute for Conservation Research and the NOAA National Marine Mammal Tissue Bank). Some of the remaining species may be obtained relatively quickly from captive animals, but the majority will require broad outreach and substantial logistical support to obtain culturable skin biopsies and take advantage of opportunistic sampling (e.g., euthanized animals from beach strandings). This process will take years or decades to complete, but the vast majority of species are likely to be represented within a few years. To accomplish this, we must be cognizant of the existing, and developing, international regulatory systems in place that regulate handling of endangered species sample collection, use and transport (e.g., the Convention on International Trade in Endangered Species of Wild Fauna and Flora; CITES). Recognizing the significant logistical constraints and time commitments needed for permitted international transport of regulated species, VGP has obtained a broad CITES permit for most species, and is currently negotiating expansion to include marine mammals.

**FIGURE 2** Schematic representation of whole genome assembly using short-read or long-read sequencing methods, and combining them with Hi-C scaffolding to link and order contigs into scaffolds. De novo assemblies of short reads result in hundreds of thousands or millions of short, unordered sequence segments. Long read assemblies provide longer, unordered segments that have higher error rates. Combined long and short read assemblies with Hi-C scaffolding orders the contigs to chromosome-length scaffolds, reduces the number of gaps to few per chromosome, resolves most repeat regions or duplicates, and improves sequence accuracy. Black dotted segments represent gaps of unknown length. Blue and black segments within short-reads (e.g., the "yellow" chromosome reads) indicate small differences between highly similar genes in a gene family or repeat region.
The exchange and transport of biological materials should also be underpinned by international legislation such as the Nagoya protocols on Access and Benefit Sharing of the Convention of Biological Diversity (https://www.cbd.int/abs/). In line with this, an important consideration is that sampling (and downstream sequencing) of species sampled from the traditional waters of Indigenous Peoples is only carried out following respectful engagement and collaboration, to ensure appropriate management of downstream data (including implementing “gated access” if desired by Indigenous Peoples), and equitable sharing of benefits and knowledge with these communities (Buck & Hamilton, 2011; Carroll, Rodriguez-Lonebear, & Martinez, 2019; Collier-Robinson, Rayne, Rupene, Thoms, & Steeves, 2019; Gemmell et al., 2019). This requirement also applies to samples collected previously from the waters of Indigenous Peoples, but now currently housed in institutional repositories. As part of this commitment to benefit sharing, we strongly support international capacity building (e.g., conducting all or part of the sequencing in countries with access to endemic species), training and facilitation of genome assembly and data sharing (within international agreements) to provide access to data, benefits and resources, reduce logistical limitations, and serve the regional scientific and conservation communities.

Although genomic sequencing is becoming widespread, expertise in the multiple technologies and complex genome assembly methods required to generate a reference-quality genome discourages most cetacean biologists. The few reference-quality genomes that have been completed have been generated in collaboration with the VGP, an international consortium of genome centers coordinated to optimize and streamline the process. The VGP protocols incorporate existing data where possible, thereby reducing cost and redundancy. The VGP also promotes open access, making raw data and assemblies immediately available as they are completed (https://vgp.github.io/genomemark/ and NCBI BioProject ID PRJNA489243), narrowly embargoed to ensure first publication rights while allowing rapid distribution of data for additional research (https://genome10k.soe.ucsc.edu/data-use-policies/). The Darwin Tree of Life project releases assemblies with fully open access at the time of deposition https://www.darwintreeoflife.org/project-resources/.

With a goal to produce hundreds, and eventually thousands of reference-quality genomes per year, the VGP has been able to substantially reduce costs, currently estimated at less than US$20,000 per mammalian genome, from DNA to curated, annotated assembly. These costs are already 50% lower than they were just 2 years ago and are expected to continue to decline.

For reference-quality genomes to become a reality for all cetacean species, a globally coordinated effort among marine mammalogists is needed to obtain and preserve samples that can yield ultra-high-quality DNA and RNA, as well as the 3D genome structure for Hi-C scaffolding. Furthermore, coordination with genome centers that can perform genome sequencing, assembly, manual curation, and annotation is needed to produce reference-quality genomes and disseminate data rapidly. To begin this process, we have formed the Cetacean Genome Project (CGP) in collaboration with the VGP and Darwin Tree of Life as a coordinated effort to (1) assemble a database of samples available from accessible collections, forge collaborations and solicit appropriate samples from the scientific community; (2) coordinate and disseminate information on best practices for sample collection and preservation (e.g., cell culture, appropriate short-term field preservation methods), with facilitation of sample transportation, storage, and, where appropriate, culture of live cells; (3) coordinate available data (e.g., published short- or long-read data, genome assemblies) to avoid redundancy and reduce costs of completing the reference-quality genomes; and (4) seek funding for individual or groups of species, in coordination with marine mammal researchers with near-term interests in genomic analysis. The CGP will leverage the participation and expertise of the VGP and Darwin Tree of Life project, while providing the focus and expertise necessary to obtain samples and funding, and conduct/facilitate research on reference-quality genomes of all cetacean species. Although we have chosen to focus on a single taxonomic group, cetaceans, the issues, needs, and recommendations discussed here apply to other aquatic mammal species as well.

While we recognize that there is not a one-model approach to accomplishing the CGP goals, the VGP model does provide a streamlined approach to generating the necessary data and releasing the curated reference-quality genome data through recognized genome databases. The interests of scientists, institutions, Indigenous Peoples, and
geopolitical entities will benefit from local involvement in some or all steps of the process, especially as an investment in training and capacity building for scientists and institutions. We foresee multiple approaches to building the platinum-standard set of cetacean genomes, and provide a nexus to coordinate and facilitate the international efforts necessary to reach those goals. Further information is available through the VGP (https://vertebrategenomesproject.org) and CGP (https://www.fisheries.noaa.gov/international/science-data/cetacean-genomes-project).

ACKNOWLEDGMENTS
We are grateful to Kim Parsons and three anonymous reviewers for helpful reviews of a previous version of this Letter. The authors confirm that they have no conflict of interest to declare.

AUTHOR CONTRIBUTIONS
Phillip Morin: Conceptualization; writing-original draft. Alana Alexander: Writing-review and editing. Mark Blaxter: Writing-review and editing. Susana Caballero: Writing-review and editing. Olivier Fedrigo: Writing-review and editing. Michaël Fontaine: Writing-review and editing. Andrew Foote: Conceptualization; writing-review and editing. Shigehiro Kuraku: Writing-review and editing. Brigid Maloney: Writing-review and editing. Morgan McCarthy: Data curation; writing-review and editing. Michael McGowen: Conceptualization; writing-review and editing. Jacquelyn Mountcastle: Writing-review and editing. Mariana Nery: Writing-review and editing. Morten Olsen: Writing-review and editing. Patricia Rosel: Writing-review and editing. Erich Jarvis: Conceptualization; writing-review and editing.

Phillip A. Morin1
Alana Alexander2
Mark Blaxter3
Susana Caballero4
Olivier Fedrigo5
Michaël Fontaine6,7
Andrew D. Foote8
Shigehiro Kuraku9
Brigid Maloney5
Morgan L. McCarthy10
Michael R. McGowen11
Jacquelyn Mountcastle5
Mariana F. Nery12
Morten Tange Olsen10
Patricia E. Rosel13
Erich D. Jarvis5,14

1Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, California
2Department of Anatomy, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand
3Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
4Laboratorio de Ecología Molecular de Vertebrados Acuáticos (LEMVA), Biological Sciences Department, Universidad de los Andes, Colombia
5The Rockefeller University, New York, New York
6Laboratoire MIVEGEC (Université de Montpellier, CNRS, IRD), Centre IRD de Montpellier, Montpellier, France
7Groningen Institute for Evolutionary Life Sciences (GELIFES), University of Groningen, Groningen, The Netherlands
8Department of Natural History, Norwegian University of Science and Technology, Trondheim, Norway
9RIKEN Center for Biosystems Dynamics Research, Kobe, Hyogo, Japan
Evolutionary Genomics, Globe Institute, University of Copenhagen, Copenhagen, Denmark

Department of Vertebrate Zoology, Smithsonian National Museum of Natural History, Washington, DC

Biology Institute, State University of Campinas, Campinas, Brazil

Southeast Fisheries Science Center, National Marine Fisheries Service, NOAA, Louisiana

The Howard Hughes Medical Institute, Chevy Chase, Maryland

Correspondence
Phillip A Morin, Southwest Fisheries Science Center, 8901 La Jolla Shores Drive, La Jolla, CA 92037.
Email: phillip.morin@noaa.gov

ORCID
Phillip A. Morin https://orcid.org/0000-0002-3279-1519
Alana Alexander https://orcid.org/0000-0002-6456-7757
Mark Blaxter https://orcid.org/0000-0003-2861-949X
Susana Caballero https://orcid.org/0000-0002-9285-3873
Olivier Fedrigo https://orcid.org/0000-0002-6450-7551
Michael C. Fontaine https://orcid.org/0000-0003-1156-4154
Andrew D. Foote https://orcid.org/0000-0001-7384-1634
Shigehiro Kuraku https://orcid.org/0000-0003-1464-8388
Brigid Maloney https://orcid.org/0000-0001-8741-8458
Morgan L. McCarthy https://orcid.org/0000-0002-9695-6060
Michael R. McGowen https://orcid.org/0000-0001-9192-3166
Jacquelyn Mountcastle https://orcid.org/0000-0003-1078-4905
Mariana F. Nery https://orcid.org/0000-0001-6501-0486
Morten Tange Olsen https://orcid.org/0000-0001-6716-6345
Patricia E. Rosel https://orcid.org/0000-0002-6481-3989
Erich D. Jarvis https://orcid.org/0000-0001-8931-5049

REFERENCES


Steinegger, M., & Salzberg, S. L. (2020). Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biology*, 21, 115.


