Building genomic infrastructure: Sequencing platinum-standard reference-quality genomes of all cetacean species

Phillip A. Morin, Alana Alexander, Mark Blaxter, Susana Caballero, Olivier Fedrigo, Michael C. Fontaine, Andrew D. Foote, Shigehiro Kuraku, Brigid Maloney, Morgan L. McCarthy, Michael R. McGowen, Jacquelyn Mountcastle, Mariana F. Nery, Morten Tange Olsen, Patricia E. Rosel, Erich D. Jarvis

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 genomic data 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 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 complex repeats) and hence do not meet the reference quality standards recommended by the VGP.

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 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.

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/genomeark/ 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.

REFERENCES

Anderson-Trocme, L., Farouni, R., Bourgey, M., Kamatani, Y., Higasa, K., Seo, J. S., … Gravel, S. (2019). Legacy data confounds genomics studies. *Molecular Biology & Evolution*, 37, 2–10.

Anonymous. (2018). A reference standard for genome biology. *Nature Biotechnology*, 36(12), 1121.

Arnason, U., Lammers, F., Kumar, V., Nilsson, M. A., & Janke, A. (2018). Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. *Science Advances*, 4(4), eaap9873.

Autenrieth, M., Hartmann, S., Lah, L., Roos, A., Dennis, A. B., & Tiedemann, R. (2018). High-quality whole-genome sequence of an abundant Holarctic odontocete, the harbour porpoise (*Phocoena phocoena*). *Molecular Ecology Resources*, 18, 1469– 1481.

Beal, A. P., Kiszka, J. J., Wells, R. S., & Eirin-Lopez, J. M. (2019). The bottlenose dolphin epigenetic aging tool (BEAT): A molecular age estimation tool for small cetaceans. *Frontiers in Marine Science*, 6, 561.

Buck, M., & Hamilton, C. (2011). The Nagoya Protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the Convention on Biological Diversity. *Review of European Community & International Environmental Law*, 20, 47– 61.

Carroll, S. R., Rodriguez-Lonebear, D., & Martinez, A. (2019). Indigenous data governance: Strategies from United States Native Nations. *Data Science Journal*, 18(1), 31.

Challis, R., Richards, E., Rajan, J., Cochrane, *G*., & Blaxter, M. (2020). BlobToolKit – Interactive quality assessment of genome assemblies. *G3*, 20, 1361– 1374.

Collier-Robinson, L., Rayne, A., Rupene, M., Thoms, C., & Steeves, T. (2019). Embedding indigenous principles in genomic research of culturally significant species: A conservation genomics case study. *New Zealand Journal of Ecology*, 43(3), 3389.

Committee on Taxonomy. (2019). List of marine mammal species and subspecies. *Society for Marine Mammalogy*. Retrieved from https://www.marinemammalscience.org/species-information/list-marine-mammal-species-subspecies/. Consulted December, 2019.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., … Aiden, E. L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333), 92– 95.

Fan, G., Zhang, Y., Liu, X., Wang, J., Sun, Z., Sun, S., … Liu, X. (2019). The first chromosome-level genome for a marine mammal as a resource to study ecology and evolution. *Molecular Ecology Resources*, 19, 944– 956.

Foote, A. D., Liu, Y., Thomas, G. W., Vinar, T., Alfoldi, J., Deng, J., … Gibbs, R. A. (2015). Convergent evolution of the genomes of marine mammals. *Nature Genetics*, 47, 272– 275.

Foote, A. D., Martin, M. D., Louis, M., Pacheco, G., Robertson, K. M., Sinding, M.-H. S., … Morin, P. A. (2019). Killer whale genomes reveal a complex history of recurrent admixture and vicariance. *Molecular Ecology*, 28, 3427– 3444.

Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., … Wolf, J. B. W. (2016). Genome-culture coevolution promotes rapid divergence in the killer whale. *Nature Communications*, 7, Article No. 11693.

Garner, B. A., Hand, B. K., Amish, S. J., Bernatchez, L., Foster, J. T., Miller, K. M., … Luikart, G. (2016). Genomics in conservation: Case studies and bridging the gap between data and application. *Trends in Ecology and Evolution*, 31(2), 81– 83.

Gemmell, N. J., Rutherford, K., Prost, S., Tollis, M., Winter, D., Macey, J. R., … Board, N. T. (2019). The tuatara genome: Insights into vertebrate evolution from the sole survivor of an ancient reptilian order. *bioRxiv*. Retrieved from. https://www.biorxiv.org/content/10.1101/867069v1.

Genome 10K Community of Scientists. (2009). Genome 10K: A proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, 100, 659– 674.

Gopalakrishnan, S., Samaniego Castruita, J. A., Sinding, M. S., Kuderna, L. F. K., Raikkonen, J., Petersen, B., … Gilbert, M. T. P. (2017). The wolf reference genome sequence (*Canis lupus lupus*) and its implications for *Canis* spp. population genomics. *BMC Genomics*, 18(1), 495.

Ho, S. S., Urban, A. E., & Mills, R. E. (2019). Structural variation in the sequencing era. *Nature Reviews Genetics*, 21, 171– 189.

Hooper, R., Brealey, J., van der Valk, T., Alberdi, A., Durban, J. W., Fearnbach, H., … Guschanski, K. (2019). Host-derived population genomics data provides insights into bacterial and diatom composition of the killer whale skin. *Molecular Ecology*, 28, 484– 502.

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931– 945.

Korlach, J., Gedman, G., Kingan, S. B., Chin, C. S., Howard, J. T., Audet, J. N., … Jarvis, E. D. (2017). De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *Gigascience*, 6(10), 1– 16.

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Science of the United States of America*, 115, 4325–4333.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., … Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950), 289– 293.

McGowen, M. R., Tsagkogeorga, G., Álvarez-Carretero, S., dos Reis, M., Struebig, M., Deaville, R., … Rossiter, S. J. (2019). Phylogenomic resolution of the cetacean tree of life using target sequence capture. *Systematic Biology*, 69, 479– 501.

Morin, P. A., Archer, F. I., Avila, C. D., Balacco, J. R., Bukham, Y. V., Chow, W., … Jarvis, E. D. (2020). Reference genome and demographic history of the most endangered marine mammal, the vaquita. *BioRxiv.* Retrieved from. https://www.biorxiv.org/content/10.1101/2020.05.27.098582v1.

Morin, P. A., Foote, A. D., Hill, C. M., Simon-Bouhet, B., Lang, A. R., & Louise, M. (2018). SNP discovery from single and multiplex genome assemblies of non-model organisms. In S. R. Head, P. Ordoukhanian, & D. Salomon (Eds.), Next-generation sequencing: Methods and protocols (pp. 113– 144). Totowa, NJ: Humana Press.

Morin, P. A., Parsons, K. M., Archer, F. I., Ávila-Arcos, M. C., Barrett-Lennard, L. G., Dalla Rosa, L., … Foote, A. D. (2015). Geographic and temporal dynamics of a global radiation and diversification in the killer whale. *Molecular Ecology*, 24, 3964– 3979.

Mulcahy, D. G., Macdonald, K. S., III, Brady, S. G., Meyer, C., Barker, K. B., & Coddington, J. (2016). Greater than X kb: A quantitative assessment of preservation conditions on genomic DNA quality, and a proposed standard for genome-quality DNA. *PeerJ*, 4, e2528.

Polanowski, A. M., Robbins, J., Chandler, D., & Jarman, S. N. (2014). Epigenetic estimation of age in humpback whales. *Molecular Ecology Resources*, 14, 976– 987.

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., … Jarvis, E. D. (2020). Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv*. Retrieved from. Google Scholar

Sanders, J. G., Beichman, A. C., Roman, J., Scott, J. J., Emerson, D., McCarthy, J. J., & Girguis, P. R. (2015). Baleen whales host a unique gut microbiome with similarities to both carnivores and herbivores. *Nature Communications*, 6, 8285.

Scornavacca, C., Belkhir, K., Lopez, J., Dernat, R., Delsuc, F., Douzery, E. J. P., & Ranwez, V. (2019). OrthoMaM v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Molecular Biology and Evolution*, 36, 861–862.

Springer, M. S., Emerling, C. A., Fugate, N., Patel, R., Starrett, J., Morin, P. A., … Gatesy, J. (2016a). Inactivation of cone-specific phototransduction genes in rod monochromatic cetaceans. *Frontiers in Ecology and Evolution*, 4, Article No. 61.

Springer, M. S., Starrett, J., Morin, P. A., Hayashi, C., & Gatesy, J. (2016b). Inactivation of C4orf26 in toothless placental mammals. *Molecular Phylogenetics and Evolution*, 95, 34–45.

Stein, L. D. (2004). Human genome: End of the beginning. *Nature*, 431(7011), 915– 916.

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.

Tan, M. P., Wong, L. L., Razali, S. A., Afiqah-Aleng, N., Mohd Nor, S. A., Sung, Y. Y., … Danish-Daniel, M. (2019). Applications of next-generation sequencing technologies and computational tools in molecular evolution and aquatic animals conservation studies: A short review. *Evolutionary Bioinformatics*, 15, 1– 5.

Westbury, M. V., Petersen, B., Garde, E., Heide-Jørgensen, M. P., & Lorenzen, E. D. (2019). Narwhal genome reveals long-term low genetic diversity despite current large abundance size. *iScience*, 15, 592– 599.

Yim, H.-S., Cho, Y. S., Guang, X., Kang, S. G., Jeong, J.-Y., Cha, S.-S., … Lee, J.-H. (2014). Minke whale genome and aquatic adaptation in cetaceans. *Nature Genetics*, 46, 88– 92.

Zhou, X., Guang, X., Sun, D., Xu, S., Li, M., Seim, I., … Yang, G. (2018). Population genomics of finless porpoises reveal an incipient cetacean species adapted to freshwater. *Nature Communications*, 9(1), 1276.

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.

Common name	**Species name**	**Assembly ID**	**Number of contigs**	**Contig N50 bp**	**Number of scaffolds**	**Scaffold N50 bp**
Antarctic minke whale	*Balaenoptera bonaerensis*	GCA_000978805.1	720,900	8,410	421,444	20,082
Baiji	*Lipotes vexillifer*	GCA_000442215.1	155,510	31,902	30,713	2,419,148
Beluga	*Delphinapterus leucas*	ASM228892v2_HiC	35,752	158,270	6,972	107,969,763
Beluga	*Delphinapterus leucas*	GCA_002288925.3	29,444	196,689	5,905	31,183,418
Beluga	*Delphinapterus leucas*	GCA_009917725.1	101,557	76,763	51,177	1,361,507
Beluga	*Delphinapterus leucas*	GCA_009917745.1	52,911	141,056	25,931	3,009,037
Blue whale	*Balaenoptera musculus*	GCA_009873245.2[a]	1,050	5,963,936	130	110,470,125
Boto	*Inia geoffrensis*	GCA_004363515.1	1,218,682	24,570	1,213,610	26,707
Bowhead whale	*Balaena mysticetus*	NA[b]	113,673	877,000	7,227	34,800
Bryde's whale	*Balaenoptera edeni*	Balaenoptera_edeni_HiC	184,171	71,244	141,314	99,560,599
Common bottlenose dolphin	*Tursiops truncatus*	GCA_000151865.3	554,227	11,821	240,557	116,287
Common bottlenose dolphin	*Tursiops truncatus*	GCA_001922835.1	116,651	44,299	2,648	26,555,543
Common bottlenose dolphin	*Tursiops truncatus*	GCA_003314715.1	139,544	30,985	481	27,166,507
Common bottlenose dolphin	*Tursiops truncatus*	GCA_003435595.3	154,206	27,134	42,644	931,081

Common name	Species name	Assembly ID	Number of contigs	Contig N50 bp	Number of scaffolds	Scaffold N50 bp
Common bottlenose dolphin	*Tursiops truncatus*	NIST_Tur_tru_v1_HiC	116,947	44,280	2,646	98,188,383
Common minke whale	*Balaenoptera acutorostrata*	GCA_000493695.1	184,072	22,690	10,776	12,843,668
Cuvier's beaked whale	*Ziphius cavirostris*	GCA_004364475.1	3,761,505	3,606	3,758,276	3,608
Fin whale	*Balaenoptera physalus*	GCA_008795845.1	1,270,025	4,493	62,302	871,016
Finless porpoise	*Neophocaena asiaeorientalis*	GCA_003031525.1	66,346	86,003	13,699	6,341,296
Franciscana	*Pontoporia blainvillei*	GCA_004363935.1	1,885,701	2,541	1,885,058	2,541
Gray whale	*Eschrichtius robustus*	GCA_002189225.1	375,256	10,066	57,203	187,455
Gray whale	*Eschrichtius robustus*	GCA_002738545.1	1,595,257	2,656	1,213,011	10,674
Gray whale	*Eschrichtius robustus*	GCA_004363415.1	1,046,770	68,559	1,036,148	94,414
Harbor porpoise	*Phocoena phocoena*	GCA_003071005.1	2,347,235	2,773	142,029	27,499,337
Harbor porpoise	*Phocoena phocoena*	GCA_004363495.1	1,338,272	89,111	1,331,158	115,969
Harbor porpoise	*Phocoena phocoena*	Phocoena_phocoena_HiC	610,275	58,076	565,368	97,795,164
Humpback whale	*Megaptera novaeangliae*	GCA_004329385.1	387,694	12,321	2,558	9,138,802
Indo-Pacific bottlenose dolphin	*Tursiops aduncus*	ASM322739v1_HiC	58,538	133,491	12,471	111,961,311
Indo-Pacific bottlenose dolphin	*Tursiops aduncus*	GCA_003227395.1	44,281	206,065	16,249	1,235,788

Common name	Species name	Assembly ID	Number of contigs	Contig N50 bp	Number of scaffolds	Scaffold N50 bp
Indo-Pacific humpbacked dolphin	*Sousa chinensis*	GCA_003521335.2	46,900	182,701	20,903	9,008,636
Indo-Pacific humpbacked dolphin	*Sousa chinensis*	GCA_007760645.1	62,803	113,766	23,368	19,436,979
Indus river dolphin	*Platanista minor*	GCA_004363435.1	1,110,492	20,879	1,098,790	23,933
Killer whale	*Orcinus orca*	GCA_000331955.2	80,100	70,300	1,668	12,735,091
Killer whale	*Orcinus orca*	Oorc_1.1_HiC	80,502	70,204	1,617	110,405,485
Long-finned pilot whale	*Globicephala melas*	ASM654740v1_HiC	21,252	332,801	6,090	106,927,605
Long-finned pilot whale	*Globicephala melas*	GCA_006547405.1	21,236	332,801	6,637	18,102,937
Melon-headed whale	*Peponocephala electra*	Peponocephala_electra_HiC	222,071	84,924	185,978	102,795,557
Narwhal	*Monodon monoceros*	GCA_004026685.1	653,473	67,024	644,873	86,766
Narwhal	*Monodon monoceros*	GCA_004027045.1	890,705	70,965	882,704	88,921
Narwhal	*Monodon monoceros*	GCA_005125345.1	813,468	10,044	21,006	1,483,363
Narwhal	*Monodon monoceros*	GCA_005190385.2	25,295	255,327	6,972	107,566,389
North Atlantic right whale	*Eubalaena glacialis*	Eubalaena_glacialis_HiC	215,753	65,924	172,124	101,413,572
North Pacific right whale	*Eubalaena japonica*	GCA_004363455.1	1,361,057	34,866	1,353,963	39,813

Common name	Species name	Assembly ID	Number of contigs	Contig N50 bp	Number of scaffolds	Scaffold N50 bp
Pacific white-sided dolphin	*Sagmatias obliquidens*	ASM367639v1_HiC	21,805	255,779	5,162	107,447,310
Pacific white-sided dolphin	*Sagmatias obliquidens*	GCA_003676395.1	21,793	255,779	5,422	28,371,583
Pygmy sperm whale	*Kogia breviceps*	GCA_004363705.1	1,258,125	26,201	1,252,072	28,812
Sowerby's beaked whale	*Mesoplodon bidens*	GCA_004027085.1	1,810,317	28,959	1,801,720	33,532
Sperm whale	*Physeter macrocephalus*	GCA_000472045.1	110,443	35,258	11,710	427,290
Sperm whale	*Physeter macrocephalus*	GCA_002837175.2	143,605	42,542	14,677	122,182,240
Sperm whale	*Physeter macrocephalus*	GCA_900411695.1	140,250	43,829	14,676	122,182,240
Vaquita	*Phocoena sinus*	GCA_008692025.1[a]	273	20,218,762	65	115,469,292

- [a] VGP platinum-quality reference genomes.
- [b] From Keane et al., 2015, *Cell Reports*, *10*, 112–122.

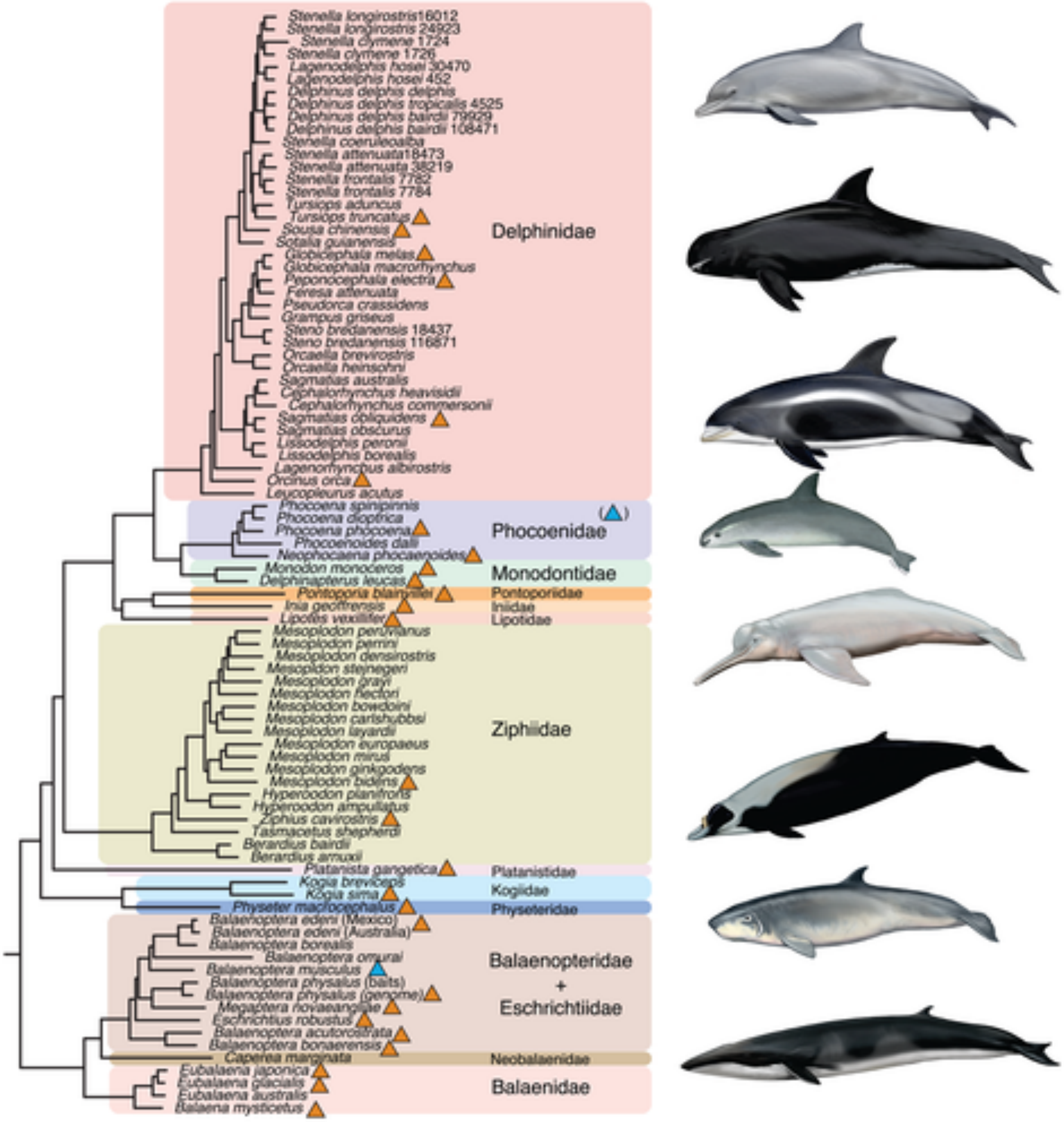

FIGURE 1 Phylogeny of the extant cetaceans based on phylogenetic analysis of 3191 protein-coding nuclear loci, reproduced from McGowen et al. (2019) and modified to show phylogenetic positions of species with published genome assemblies. Blue triangles mark the species represented by platinum-quality 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.

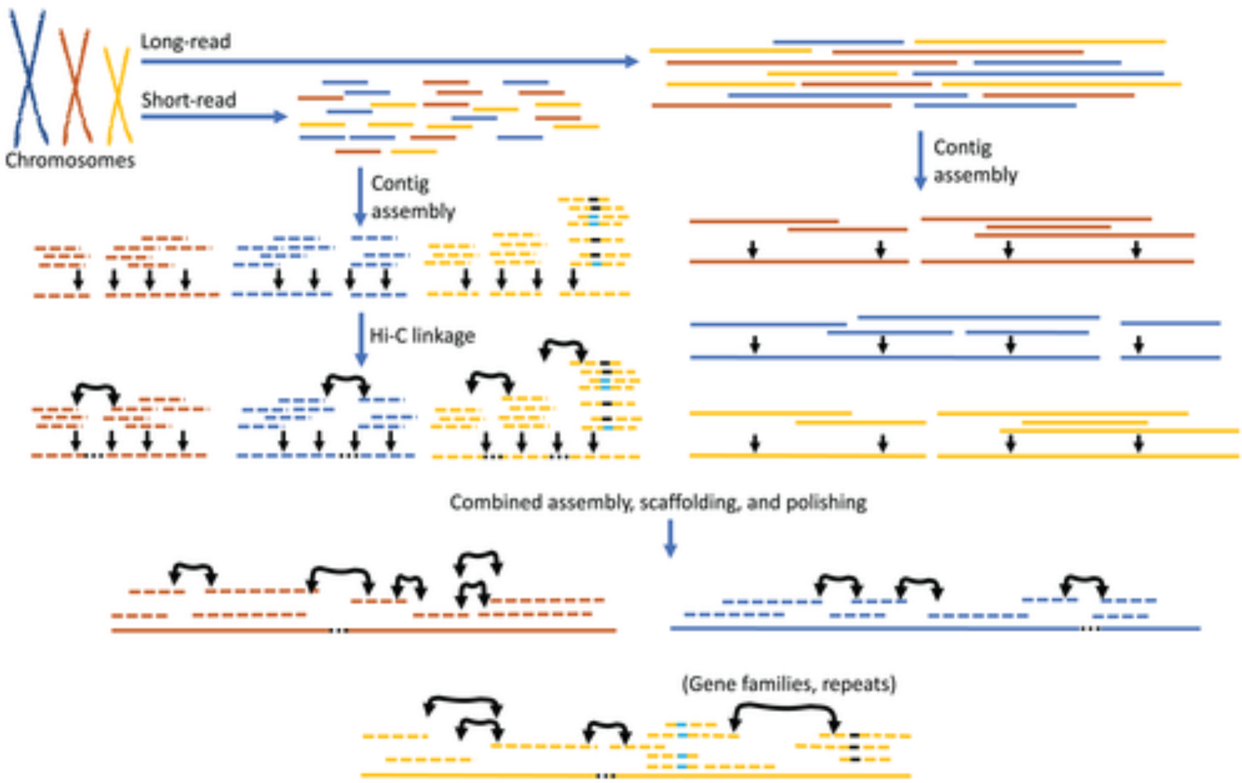

FIGURE 2 Schematic representation of whole genome assembly using short-read or long-read sequencing methods, and combining them with Hi-C scaffolding to link and order contigs into scaffolds. De novo assemblies of short reads result in hundreds of thousands or millions of short, unordered sequence segments. 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.