



HAL
open science

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 Mccarthy, et al.

► **To cite this version:**

Phillip A. Morin, Alana Alexander, Mark Blaxter, Susana Caballero, Olivier Fedrigo, et al.. Building genomic infrastructure: Sequencing platinum-standard reference-quality genomes of all cetacean species. *Marine Mammal Science*, 2020, 10.1111/mms.12721 . hal-03020344

HAL Id: hal-03020344

<https://hal.science/hal-03020344v1>

Submitted on 24 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

3
4 Phillip A. Morin¹, Alana Alexander², Mark Blaxter³, Susana Caballero⁴, Olivier Fedrigo⁵,
5 Michael C. Fontaine^{6,7}, Andrew D. Foote⁸, Shigehiro Kuraku⁹, Brigid Maloney⁵, Morgan L.
6 McCarthy¹⁰, Michael R. McGowen¹¹, Jacquelyn Mountcastle⁵, Mariana F. Nery¹², Morten Tange
7 Olsen¹⁰, Patricia E. Rosel¹³, Erich D. Jarvis^{5,14}

8
9
10 1 Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 8901 La Jolla
11 Shores Dr., La Jolla, CA 92037, USA

12 2 Department of Anatomy, School of Biomedical Sciences, University of Otago, Dunedin, New
13 Zealand

14 3 Wellcome Sanger Institute, 215 Euston Road, London, NW1 2BE, UK

15 4 Laboratorio de Ecología Molecular de Vertebrados Acuáticos (LEMVA), Biological Sciences
16 Department, Universidad de los Andes, Carrera 1E # 18 A 10, Bogota, Colombia

17 5 The Rockefeller University, Box 54, 1230 York Avenue, New York, NY 10065, USA

18 6 Laboratoire MIVEGEC (Université de Montpellier, CNRS, IRD), Centre IRD de Montpellier,
19 34394 Montpellier, France

20 7 Groningen Institute for Evolutionary Life Sciences (GELIFES), University of Groningen, PO
21 Box 11103 CC, Groningen, The Netherlands.

22 8 Department of Natural History, Norwegian University of Science and Technology, NO-7491
23 Trondheim, Norway.

24 9 RIKEN Center for Biosystems Dynamics Research, 2-2-3 Minatojima-minamimachi, Chuo-ku,
25 Kobe, Hyogo 650-0047, Japan

26 10 Evolutionary Genomics, Globe Institute, University of Copenhagen, Oester Voldgade 5-7,
27 1350 Copenhagen K, Denmark

28 11 Department of Vertebrate Zoology, Smithsonian National Museum of Natural History, 10th
29 St. & Constitution Ave. NW, Washington, D.C. 20560, USA

30 12 Biology Institute, State University of Campinas, Rua Bertrand Russel, Caixa Postal 6109 –
31 CEP 13083-970 – Campinas, Brazil

32 13 Southeast Fisheries Science Center, National Marine Fisheries Service, NOAA, 646
33 Cajundome Boulevard, Lafayette, Louisiana 70506, U.S.A

34 14 The Howard Hughes Medical Institute, 4000 Jones Bridge Rd, Chevy Chase, MD 20815,
35 USA

36

37

38 In 2001 it was announced that the 3.1 billion base (Gigabase, Gb) human genome had been
39 sequenced, but after 13 years of work and \$2.7 billion, it was still considered to be only a draft,
40 missing over 30% of the genome and made up of over 100,000 sequence fragments (scaffolds)
41 with an average size of just 81,500 base-pairs (bp) (International Human Genome Sequencing
42 Consortium, 2004; Stein, 2004). As technologies improved, the draft human genome assembly
43 has been repeatedly refined and corrected. By the time the genome assembly was published in
44 2004, the average length of scaffolds had increased to over 38 million bp (Megabases, Mb) with
45 only a few hundred gaps in the chromosome-length scaffolds. However, the duplicated and
46 highly repetitive regions of the human genome remained unresolved due to limitations of short-
47 read sequencing technology that required piecing the genome together from billions of shorter
48 sequences. Over the last decade, as highly parallel, much less expensive, short and long-read
49 sequencing technologies have revolutionized genomic sequencing, thousands of individual
50 human genomes have been re-sequenced, further refining the human genome assembly and
51 characterizing its diversity, resulting in a “reference-quality” genome assembly that covers 95%
52 of the genome with far fewer and smaller gaps compared to the initial version. Despite this vast
53 improvement, the human genome continues to be updated and refined (v. 39, RefSeq accession
54 GCF_000001405.39).

55
56 This example illustrates how all eukaryotic genome assemblies, even those of exemplar quality,
57 are drafts, varying in sequence quality (i.e. error rate), completeness (i.e. how much of the
58 genome is covered), how contiguous DNA sequences within scaffolds are, and what portions of
59 the genome remain unresolved or incorrect. The “platinum-standard reference genome” that
60 modern genomics strives for is distinguished from older draft assemblies by completeness, low
61 error rates, and a high percentage of the sequences assembled into chromosome-length scaffolds
62 (Anon., 2018; Rhie et al., in prep). For the remainder of this note, we use ‘draft’ to refer to the
63 less complete/contiguous ‘draftier draft’ genomes and ‘reference-quality genomes’ to refer to
64 “platinum-standard reference genome” draft genomes as characterized above.

65
66 Democratization of genome sequencing has yielded draft genomes across the diversity of life at a
67 rate that was unimaginable just a few years ago. As genome assemblies have become
68 increasingly common, titles of articles often tout “chromosome-level”, “complete”, “reference-

69 quality” and other adjectives to characterize the quality of a new genome sequence. These terms
70 offer little information about the level of completion or accuracy of genome assemblies, as even
71 “chromosome-level” genomes may consist of thousands to millions of sequence fragments (e.g.,
72 Fan et al., 2019), with significant amounts of missing data, assembly errors, and missing or
73 incomplete genome annotations.

74

75 The utility of draft genomes has been abundantly documented, and there is no doubt that draft
76 genomes provide sufficient data to address many biological questions. For cetaceans, highly
77 fragmented draft genomes have been useful references for mapping data from re-sequenced
78 individuals, and thus for characterization of variable markers (Morin et al., 2018), phylogenetics
79 and comparative genomics (Arnason, Lammers, Kumar, Nilsson, & Janke, 2018; Fan et al.,
80 2019; Foote et al., 2015; Yim et al., 2014), characterization of intraspecific variability and
81 demographic history (Autenrieth et al., 2018; Foote et al., 2019; Foote et al., 2016; Morin et al.,
82 2015; Westbury, Petersen, Garde, Heide-Jorgensen, & Lorenzen, 2019; Zhou et al., 2018),
83 molecular evolution of genes and traits (Autenrieth et al., 2018; Fan et al., 2019; Foote et al.,
84 2015; Springer et al., 2016a; Springer, Starrett, Morin, Hayashi, & Gatesy, 2016b; Yim et al.,
85 2014), epigenetic age determination (Beal, Kiszka, Wells, & Eirin-Lopez, 2019; Polanowski,
86 Robbins, Chandler, & Jarman, 2014), and skin and gut microbiome metagenomics (Hooper et al.,
87 2019; Sanders et al., 2015). The field of conservation genomics has also demonstrated the many
88 applications of genomic data that aid in discovery of vulnerable species, identify extinction risks,
89 and implement appropriate management (Garner et al., 2016; Kraus et al., submitted; Tan et al.,
90 2019).

91

92 However, the types of errors common to draft genomes can be at best misleading (e.g., structural
93 variation, Ho, Urban, & Mills, 2019), and at worst may result in years of lost time and effort
94 pursuing genes and variants that do not exist (Anderson-Trocme et al., 2019; Korlach et al.,
95 2017). Use of a related species reference genome to map sequencing reads (when the new
96 species genome is not available) reduces and biases mapping of the new species reads,
97 compromising estimates of variation (e.g., mapping reads to a distantly related species;
98 Gopalakrishnan et al., 2017). The completeness of a genome and of its coding and regulatory
99 annotation (e.g., coding regions and identified genes; Scornavacca et al., 2019) affect

100 downstream interpretation of analytic results. Recently, re-analysis of published genomes has
101 shown that appreciable portions of most genome assemblies (e.g., 4.3 Mb of a sperm whale
102 assembly) contain contaminating sequences (including full genes) from parasites and bacteria
103 (Challis, Richards, Rajan, Cochrane, & Blaxter, 2020; Steinegger & Salzberg, 2020).

104
105 Recent improvements in sequencing and bioinformatic technologies have changed our view of
106 what is possible in genome assembly, such that now it is credible to propose reference quality
107 genome sequencing for not just a few model taxa of interest, but rather for whole biomes, whole
108 clades and, ultimately all of the planet's biota. The Earth BioGenome Project (EBP; Lewin et al.,
109 2018) proposes the reference genome sequencing of all eukaryotic life on earth. The EBP goals
110 are reflected in local biotic projects, such as the Darwin Tree of Life project
111 (darwintreeoflife.org), which aims to sequence all eukaryotic species in Britain and Ireland
112 (including several cetacean species), and clade-focused projects such as the Genome 10k
113 (Genome 10K Community of Scientists, 2009) and its Vertebrate Genomes Project (VGP;
114 vertebratengenomesproject.org), which propose sequencing of all Vertebrata. In an effort to
115 establish benchmark quality standards and best practices for reference-quality genome
116 sequencing, the VGP has developed combined sequencing technologies and assembly protocols
117 (Anon., 2018), and criteria for evaluation of genomes to meet "platinum-quality" standards (Rhie
118 et al., in prep). They find that vertebrate genome assemblies that lead to far fewer errors in
119 biological analyses are those that have a contig N50 (without gaps) of 1 Mb or more;
120 chromosomal scaffold N50 of 10 Mb or more; base call accuracy of Q40 or higher (no more than
121 one nucleotide error per 10,000 bp); paternal and maternal sequences haplotype phased to reduce
122 false gene duplication errors; and manual curation to improve the genome assembly and reduce
123 errors further. These genome assemblies thus far have up to >99% of the genome assembled into
124 chromosomes, with some chromosomes having between 0 to fewer than 20 gaps. Both the VGP
125 and the Darwin Tree of Life projects aim to meet these quality standards for all of their genome
126 assemblies.

127
128 Such reference-quality genomes for each focal cetacean species would offer a platform for
129 analysis that will avoid the types of errors discussed above, obviating the need for cross-species
130 read mapping that is currently the norm. High-quality genomes make correct gene identification

131 possible (Korlach et al., 2017), help phasing of population genomic data (identifying paternal
132 and maternal chromosomes), contribute to identification of population-level structural variation
133 and permit informed analysis of genome architectures (e.g. centromeric and telomeric regions).

134
135 As of December 2019, there were 28 cetacean species present in public sequence databases as
136 draft assemblies, but only two species had VGP platinum-standard reference genome assemblies
137 and they were generated by the VGP: the vaquita and the blue whale (Table 1, Figure 1). The
138 vaquita genome, for example, has 99.92% of the assembly assigned to 22 nearly-gapless (0-35
139 gaps/scaffold) chromosome-level scaffolds, with accuracy Q40.88 (0.8 nucleotide errors per
140 10,000bp) (Morin et al., 2020). By contrast, the sperm whale chromosome-level genome
141 (accession GCA_002837175.2; Fan et al., 2019), assembled from short-read shot-gun, 10X
142 Genomics linked reads and Hi-C scaffolding, assigned 95% of the assembly to 21 chromosomes,
143 but contains 1513-9978 gaps per scaffold. The primary reason for the difference between the
144 VGP genomes and the sperm whale genome is the use of long-read sequencing to obtain 475X
145 and 140X larger contig N50s (vaquita and blue whale, respectively; Table 1), allowing assembly
146 of all but the most difficult regions (e.g., some centromeric and telomeric regions). We are aware
147 of whole-genome shotgun (WGS) sequencing projects underway for most of the 96 recognized
148 cetacean species. Most of these projects will result in highly fragmented and incomplete draft
149 genome assemblies that may include >90% of the genes, but are unlikely to resolve
150 chromosome-level scaffolds, let alone full gene or genome structure. A substantial effort is
151 underway (DNAzoo.org) to improve contiguity in new and existing genome assemblies using
152 proximity-guided assembly methods (Hi-C; Dudchenko et al., 2017; Lieberman-Aiden et al.,
153 2009). This approach generates chromosome-level scaffolds, and can yield highly contiguous
154 genomes when long reads are used. When used with short-read data, this approach is very cost-
155 effective and can be used even with somewhat degraded tissue samples. However, these genome
156 assemblies remain highly fragmented with regions of unresolved structure (e.g., long or complex
157 repeats) and hence do not meet the reference quality standards recommended by the VGP.

158
159 The critical step needed to meet the platinum-level criteria set out by the VGP is long-read
160 sequencing (e.g., Pacific Biosciences or Oxford Nanopore technologies) that generates
161 contiguous raw data tens to hundreds of kilobases in length. Combined with long-range,

162 chromosome-scale scaffolding methods based on Hi-C chromatin contacts and optical mapping
163 (e.g., BioNano), these data allow repetitive regions within scaffolds to be resolved (Figure 2).

164
165 While this approach is now becoming feasible even on a moderate research budget, the major
166 limitation for many marine mammals is availability of fresh tissues that yield relatively large
167 amounts of ultra-high quality DNA for long-read sequencing (>40 Kb) and BioNano approaches
168 (>300 Kb) (e.g., Mulcahy et al., 2016) and intact chromatin preserving the 3D structure in nuclei
169 for long-range Hi-C linking to build scaffolds. These technologies currently require fresh blood,
170 muscle or organ tissue, or cultured cells, preserved to maintain megabase-length DNA and
171 (preferably, for gene annotation) RNA. Although there are rare exceptions, this usually requires
172 rapid freezing and storage at $\leq -80^{\circ}\text{C}$ or culture of live cells, both of which have limited
173 feasibility for protected species (due to sampling methods) and in many field conditions (e.g.,
174 mass strandings on remote beaches or locations with scarce infrastructure). Therefore, collection
175 and preservation of such samples is rare.

176
177 Given the manifest benefits of reference-quality genome sequencing from at least one specimen
178 of each species, and the extreme logistical difficulty in obtaining appropriate samples for long-
179 read sequencing methods, we propose that a concerted effort should be made to coordinate and
180 facilitate ethical collection of cetacean samples immediately. We estimate that such samples are
181 currently available for about 25% of cetacean species in a few publicly accessible collections that
182 have already contributed samples for cetacean genomics (e.g., the Frozen Zoo® tissue culture
183 collection at San Diego Zoo Global's Institute for Conservation Research and the NOAA
184 National Marine Mammal Tissue Bank). Some of the remaining species may be obtained
185 relatively quickly from captive animals, but the majority will require broad outreach and
186 substantial logistical support to obtain culturable skin biopsies and take advantage of
187 opportunistic sampling (e.g., euthanized animals from beach strandings). This process will take
188 years or decades to complete, but the vast majority of species are likely to be represented within
189 a few years. We must be cognizant of the existing, and developing, international regulatory
190 systems in place that regulate handling of endangered species (e.g., the Convention on
191 International Trade in Endangered Species of Wild Fauna and Flora; CITES). Recognizing the
192 significant logistical constraints and time commitments needed for permitted international

193 transport of regulated species, VGP has obtained a broad CITES permit for most species, and is
194 currently negotiating expansion to include marine mammals.

195

196 The exchange and transport of biological materials should also be underpinned by international
197 legislation such as the Nagoya protocols on Access and Benefit Sharing of the Convention of
198 Biological Diversity (<https://www.cbd.int/abs/>). In line with this, an important consideration is
199 that sampling (and downstream sequencing) of species sampled from the traditional waters of
200 Indigenous peoples is only carried out following respectful engagement and collaboration, to
201 ensure appropriate management of downstream data (including implementing ‘gated access’ if
202 desired by indigenous peoples), and equitable sharing of benefits and knowledge with these
203 communities (Buck & Hamilton, 2011; Carroll, Rodriguez-Lonebear, & Martinez, 2019; Collier-
204 Robinson, Rayne, Rupene, Thoms, & Steeves, 2019; Gemmell et al., 2019). This requirement
205 also applies to samples collected previously from the waters of Indigenous peoples, but now
206 currently housed in institutional repositories. As part of this commitment to benefit sharing, we
207 strongly support international capacity building (e.g., conducting all or part of the sequencing in
208 countries with access to endemic species), training and facilitation of genome assembly and data
209 sharing (within international agreements) to provide benefits and resources, reduce logistical
210 limitations, and serve the regional scientific and conservation communities.

211

212 Although genomic sequencing is becoming widespread, expertise in the multiple technologies
213 and complex genome assembly methods required to generate a reference-quality genome
214 discourages most cetacean biologists. The few reference-quality genomes that have been
215 completed have been generated in collaboration with the VGP, an international consortium of
216 genome centers coordinated to optimize and streamline the process. The VGP protocols
217 incorporate existing data where possible, thereby reducing cost and redundancy. The VGP also
218 promotes open access, making raw data and assemblies immediately available as they are
219 completed (<https://vgp.github.io/genomeark/> and NCBI BioProject ID PRJNA489243), narrowly
220 embargoed to ensure first publication rights while allowing rapid distribution of data for
221 additional research (<https://genome10k.soe.ucsc.edu/data-use-policies/>). The Darwin Tree of Life
222 project releases assemblies with fully open access at the time of deposition
223 (<https://www.darwintreeoflife.org/wp-content/uploads/2020/03/DToL-Open-Data-Release->

224 [Policy.pdf](#)). With a goal to produce hundreds, and eventually thousands of reference-quality
225 genomes per year, the VGP has been able to substantially reduce costs, currently estimated at
226 less than US\$20,000 per mammalian genome, from DNA to curated, annotated assembly. These
227 costs are already 50% lower than they were just two years ago, and are expected to continue to
228 decline.

229

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

250

251 While we recognize that there is not a one-model approach to accomplishing the CGP goals, the
252 VGP model does provide a streamlined approach to generating the necessary data and releasing
253 the curated reference-quality genome data through recognized genome databases. The interests
254 of scientists, institutions, states, Indigenous peoples, and geopolitical entities will benefit from
255 local involvement in some or all steps of the process, especially as an investment in training and
256 capacity building for scientists and institutions. We foresee multiple approaches to building the

257 “platinum standard” set of cetacean genomes, and provide a nexus to coordinate and facilitate the
258 international efforts necessary to reach those goals. Further information is available through the
259 VGP (<https://vertebrategenomesproject.org>) and CGP
260 (<https://www.fisheries.noaa.gov/content/cetacean-genomes-research>).

261
262 Literature Cited

263 Anderson-Trocme, L., Farouni, R., Bourgey, M., Kamatani, Y., Higasa, K., Seo, J. S., . . .
264 Gravel, S. (2019). Legacy data confounds genomics studies. *Molecular Biology &*
265 *Evolution*, 37(1), 2-10. doi:10.1093/molbev/msz201

266 Anon. (2018). A reference standard for genome biology. *Nature Biotechnology*, 36(12), 1121.
267 doi:10.1038/nbt.4318

268 Arnason, U., Lammers, F., Kumar, V., Nilsson, M. A., & Janke, A. (2018). Whole-genome
269 sequencing of the blue whale and other rorquals finds signatures for introgressive gene
270 flow. *Science Advances*, 4(4), eaap9873. doi:10.1126/sciadv.aap9873

271 Autenrieth, M., Hartmann, S., Lah, L., Roos, A., Dennis, A. B., & Tiedemann, R. (2018). High-
272 quality whole-genome sequence of an abundant Holarctic odontocete, the harbour
273 porpoise (*Phocoena phocoena*). *Molecular Ecology Resources*, 18(6), 1469-1481.
274 doi:10.1111/1755-0998.12932

275 Beal, A. P., Kiszka, J. J., Wells, R. S., & Eirin-Lopez, J. M. (2019). The bottlenose dolphin
276 epigenetic aging tool (BEAT): A molecular age estimation tool for small cetaceans.
277 *Frontiers in Marine Science*, 6, 561. doi:10.3389/fmars.2019.00561

278 Buck, M., & Hamilton, C. (2011). The Nagoya Protocol on access to genetic resources and the
279 fair and equitable sharing of benefits arising from their utilization to the Convention on
280 Biological Diversity. *Review of European Community & International Environmental*
281 *Law*, 20(1), 47-61. doi:10.1111/j.1467-9388.2011.00703.x

282 Carroll, S. R., Rodriguez-Lonebear, D., & Martinez, A. (2019). Indigenous data gGovernance:
283 Strategies from United States Native Nations. *Data Science Journal*, 18(1), 31.
284 doi:10.5334/dsj-2019-031

285 Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit –
286 Interactive quality assessment of genome assemblies. *G3*, in press. doi:10.1101/844852

287 Collier-Robinson, L., Rayne, A., Rupene, M., Thoms, C., & Steeves, T. (2019). Embedding
288 indigenous principles in genomic research of culturally significant species: a conservation
289 genomics case study. *New Zealand Journal of Ecology*, 43(3).
290 doi:10.20417/nzjecol.43.36

- 291 Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., . . . Aiden,
292 E. L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields
293 chromosome-length scaffolds. *Science*, *356*(6333), 92-95. doi:10.1126/science.aal3327
- 294 Fan, G., Zhang, Y., Liu, X., Wang, J., Sun, Z., Sun, S., . . . Liu, X. (2019). The first
295 chromosome-level genome for a marine mammal as a resource to study ecology and
296 evolution. *Molecular Ecology Resources*, *19*(4), 944-956. doi:10.1111/1755-0998.13003
- 297 Foote, A. D., Liu, Y., Thomas, G. W., Vinar, T., Alfoldi, J., Deng, J., . . . Gibbs, R. A. (2015).
298 Convergent evolution of the genomes of marine mammals. *Nature Genetics*, *47*(3), 272-
299 275. doi:10.1038/ng.3198
- 300 Foote, A. D., Martin, M. D., Louis, M., Pacheco, G., Robertson, K. M., Sinding, M.-H. S., . . .
301 Morin, P. A. (2019). Killer whale genomes reveal a complex history of recurrent
302 admixture and vicariance. *Molecular Ecology*, *28*, 3427-3444. doi:10.1111/mec.15099
- 303 Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., . . .
304 Wolf, J. B. W. (2016). Genome-culture coevolution promotes rapid divergence in the
305 killer whale. *Nature Communications*, *7*, Article No. 11693. doi:10.1038/ncomms11693
- 306 Garner, B. A., Hand, B. K., Amish, S. J., Bernatchez, L., Foster, J. T., Miller, K. M., . . . Luikart,
307 G. (2016). Genomics in conservation: case studies and bridging the gap between data and
308 application. *Trends in Ecology and Evolution*, *31*(2), 81-83.
309 doi:10.1016/j.tree.2015.10.009
- 310 Gemmell, N. J., Rutherford, K., Prost, S., Tollis, M., Winter, D., Macey, J. R., . . . Board, N. T.
311 (2019). The tuatara genome: insights into vertebrate evolution from the sole survivor of
312 an ancient reptilian order. *bioRxiv*. doi:10.1101/867069
- 313 Genome 10K Community of Scientists. (2009). Genome 10K Community of Scientists. Genome
314 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *Journal*
315 *of Heredity*, *100*(6), 659-674. doi:10.1093/jhered/esp086
- 316 Gopalakrishnan, S., Samaniego Castruita, J. A., Sinding, M. S., Kuderna, L. F. K., Raikonen, J.,
317 Petersen, B., . . . Gilbert, M. T. P. (2017). The wolf reference genome sequence (*Canis*
318 *lupus lupus*) and its implications for *Canis* spp. population genomics. *BMC Genomics*,
319 *18*(1), 495. doi:10.1186/s12864-017-3883-3
- 320 Ho, S. S., Urban, A. E., & Mills, R. E. (2019). Structural variation in the sequencing era. *Nature*
321 *Reviews Genetics*. doi:10.1038/s41576-019-0180-9
- 322 Hooper, R., Brealey, J., van der Valk, T., Alberdi, A., Durban, J. W., Fearnbach, H., . . .
323 Guschanski, K. (2019). Host-derived population genomics data provides insights into

- 324 bacterial and diatom composition of the killer whale skin. *Molecular Ecology*, 28(2),
325 484-502. doi:10.1111/mec.14860
- 326 International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic
327 sequence of the human genome. *Nature*, 431(7011), 931-945. doi:10.1038/nature03001
- 328 Korlach, J., Gedman, G., Kingan, S. B., Chin, C. S., Howard, J. T., Audet, J. N., . . . Jarvis, E. D.
329 (2017). De novo PacBio long-read and phased avian genome assemblies correct and add
330 to reference genes generated with intermediate and short reads. *Gigascience*, 6(10), 1-16.
331 doi:10.1093/gigascience/gix085
- 332 Kraus, R. H. S., Paez, S., Ceballos, G., Crawford, A., Fedrigo, O., Finnegan, S., . . . Jarvis, E. D.
333 (submitted). The role of genomics in conserving biodiversity during the sixth mass
334 extinction. *Nature Reviews Genetics*.
- 335 Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., . . .
336 Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life.
337 *Proceedings of the National Academy of Science, USA*, 115(17), 4325-4333.
338 doi:10.1073/pnas.1720115115
- 339 Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragozy, T., Telling, A., .
340 . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding
341 principles of the human genome. *Science*, 326(5950), 289-293.
342 doi:10.1126/science.1181369
- 343 McGowen, M. R., Tsagkogeorga, G., Álvarez-Carretero, S., dos Reis, M., Struebig, M., Deaville,
344 R., . . . Rossiter, S. J. (2019). Phylogenomic resolution of the cetacean tree of life using
345 target sequence capture. *Systematic Biology*, *syz068*. doi:10.1093/sysbio/syz068
- 346 Morin, P. A., Archer, F. I., Balacco, J. R., Bukham, Y. V., Chaisson, M. J. P., Chow, W., . . .
347 Jarvis, E. D. (2020). Reference Genome and demographic history of the most endangered
348 marine mammal, the vaquita. *BioRxiv*.
- 349 Morin, P. A., Foote, A. D., Hill, C. M., Simon-Bouhet, B., Lang, A. R., & Louise, M. (2018).
350 SNP discovery from single and multiplex genome assemblies of non-model organisms. In
351 S. R. Head, P. Ordoukhanian, & D. Salomon (Eds.), *Next-Generation Sequencing*.
352 *Methods in Molecular Biology* (Vol. 1712, pp. 113-144): Humana Press.
- 353 Morin, P. A., Parsons, K. M., Archer, F. I., Ávila-Arcos, M. C., Barrett-Lennard, L. G., Dalla
354 Rosa, L., . . . Foote, A. D. (2015). Geographic and temporal dynamics of a global
355 radiation and diversification in the killer whale. *Molecular Ecology*, 24, 3964-3979.
356 doi:doi: 10.1111/mec.13284

- 357 Mulcahy, D. G., Macdonald, K. S., 3rd, Brady, S. G., Meyer, C., Barker, K. B., & Coddington, J.
358 (2016). Greater than X kb: a quantitative assessment of preservation conditions on
359 genomic DNA quality, and a proposed standard for genome-quality DNA. *PeerJ*, *4*,
360 e2528. doi:10.7717/peerj.2528
- 361 Polanowski, A. M., Robbins, J., Chandler, D., & Jarman, S. N. (2014). Epigenetic estimation of
362 age in humpback whales. *Molecular Ecology Resources*, *14*(5), 976-987.
363 doi:10.1111/1755-0998.12247
- 364 Rhie, A., McCarthy, S., Fedrigo, O., Howe, K., Myers, E. W., Durbin, R., . . . Jarvis, E. D. (in
365 prep). Towards complete and error-free genome assemblies of all vertebrate species.
366 *Nature*.
- 367 Sanders, J. G., Beichman, A. C., Roman, J., Scott, J. J., Emerson, D., McCarthy, J. J., & Girguis,
368 P. R. (2015). Baleen whales host a unique gut microbiome with similarities to both
369 carnivores and herbivores. *Nature Communications*, *6*, 8285. doi:10.1038/ncomms9285
- 370 Scornavacca, C., Belkhir, K., Lopez, J., Dernas, R., Delsuc, F., Douzery, E. J. P., & Ranwez, V.
371 (2019). OrthoMaM v10: Scaling-up orthologous coding sequence and exon alignments
372 with more than one hundred mammalian genomes. *Molecular Biology and Evolution*,
373 *36*(4), 861-862. doi:10.1093/molbev/msz015
- 374 Springer, M. S., Emerling, C. A., Fugate, N., Patel, R., Starrett, J., Morin, P. A., . . . Gatesy, J.
375 (2016a). Inactivation of cone-specific phototransduction genes in rod monochromatic
376 cetaceans. *Frontiers in Ecology and Evolution*, *4*, Article No. 61.
377 doi:10.3389/fevo.2016.00061
- 378 Springer, M. S., Starrett, J., Morin, P. A., Hayashi, C., & Gatesy, J. (2016b). Inactivation of
379 C4orf26 in toothless placental mammals. *Molecular Phylogenetics and Evolution*, *95*, 34-
380 45. doi:10.1016/j.ympev.2015.11.002
- 381 Stein, L. D. (2004). Human genome: end of the beginning. *Nature*, *431*(7011), 915-916.
382 doi:10.1038/431915a
- 383 Steinegger, M., & Salzberg, S. L. (2020). Terminating contamination: large-scale search
384 identifies more than 2,000,000 contaminated entries in GenBank. *bioRxiv doi*
385 *10.1101/2020.01.26.920173*.
- 386 Tan, M. P., Wong, L. L., Razali, S. A., Afiqah-Aleng, N., Mohd Nor, S. A., Sung, Y. Y., . . .
387 Danish-Daniel, M. (2019). Applications of next-generation sequencing technologies and
388 computational tools in molecular evolution and aquatic animals conservation studies: A
389 short review. *Evolutionary Bioinformatics*, *15*, 1-5. doi:10.1177/1176934319892284

- 390 Westbury, M. V., Petersen, B., Garde, E., Heide-Jorgensen, M. P., & Lorenzen, E. D. (2019).
391 Narwhal genome reveals long-term low genetic diversity despite current large abundance
392 size. *iScience*, *15*, 592-599. doi:10.1016/j.isci.2019.03.023
- 393 Yim, H.-S., Cho, Y. S., Guang, X., Kang, S. G., Jeong, J.-Y., Cha, S.-S., . . . Lee, J.-H. (2014).
394 Minke whale genome and aquatic adaptation in cetaceans. *Nature Genetics*, *46*, 88-92.
395 doi:10.1038/ng.2835
- 396 Zhou, X., Guang, X., Sun, D., Xu, S., Li, M., Seim, I., . . . Yang, G. (2018). Population genomics
397 of finless porpoises reveal an incipient cetacean species adapted to freshwater. *Nature*
398 *Communications*, *9*(1), Article number: 1276. doi:10.1038/s41467-018-03722-x
399

400 Table 1. Cetacean genome assembly information from assemblies in NCBI Genome Assembly database (ncbi.nlm.nih.gov/genome)
401 and DNazoo (Assembly ID's ending with "HiC"; dnazoo.org/assemblies) as of January, 2020. The assembly level "scaffold" refers to
402 both unordered contigs and ordered scaffolds. Contig and Scaffold N50 are measures of assembly quality indicating that half of the
403 genome assembly is found in contigs or scaffolds equal to or larger than the N50 size. In addition to Contig and Scaffold N50 metrics,
404 an assessment of whether a genome meets platinum quality standards also relies on other metrics such as genome-wide base-call
405 accuracy level (\geq Q40, or no more than 1 nucleotide error per 10,000bp), and phased maternal/paternal haplotypes to reduce false gene
406 duplication errors. Rhie et al. (2020) contains additional detail on VGP assembly methods and platinum genome quality standards.
407

Species name	Common name	Assembly ID	Number of contigs	Contig N50	Number of scaffolds	Scaffold N50
<i>Balaenoptera bonaerensis</i>	Antarctic Minke Whale	GCA_000978805.1	720,900	8,410	421,444	20,082
<i>Lipotes vexillifer</i>	Baiji	GCA_000442215.1	155,510	31,902	30,713	2,419,148
<i>Delphinapterus leucas</i>	Beluga	ASM228892v2_HiC	35,752	158,270	6,972	107,969,763
<i>Delphinapterus leucas</i>	Beluga	GCA_002288925.3	29,444	196,689	5,905	31,183,418
<i>Delphinapterus leucas</i>	Beluga	GCA_009917725.1	101,557	76,763	51,177	1,361,507
<i>Delphinapterus leucas</i>	Beluga	GCA_009917745.1	52,911	141,056	25,931	3,009,037
<i>Balaenoptera musculus</i>	Blue Whale	GCA_009873245.2*	1,050	5,963,936	130	110,470,125
<i>Inia geoffrensis</i>	Boto	GCA_004363515.1	1,218,682	24,570	1,213,610	26,707
<i>Balaena mysticetus</i>	Bowhead Whale	NA †	113,673	877,000	7,227	34,800
<i>Balaenoptera edeni</i>	Bryde's Whale	Balaenoptera_edeni_HiC	184,171	71,244	141,314	99,560,599
<i>Tursiops truncatus</i>	Common Bottlenose Dolphin	GCA_000151865.3	554,227	11,821	240,557	116,287
<i>Tursiops truncatus</i>	Common Bottlenose Dolphin	GCA_001922835.1	116,651	44,299	2,648	26,555,543
<i>Tursiops truncatus</i>	Common Bottlenose Dolphin	GCA_003314715.1	139,544	30,985	481	27,166,507
<i>Tursiops truncatus</i>	Common Bottlenose Dolphin	GCA_003435595.3	154,206	27,134	42,644	931,081
<i>Tursiops truncatus</i>	Common Bottlenose Dolphin	NIST_Tur_tru_v1_HiC	116,947	44,280	2,646	98,188,383
<i>Balaenoptera acutorostrata</i>	Common Minke Whale	GCA_000493695.1	184,072	22,690	10,776	12,843,668
<i>Ziphius cavirostris</i>	Cuvier's Beaked Whale	GCA_004364475.1	3,761,505	3,606	3,758,276	3,608
<i>Balaenoptera physalus</i>	Fin whale	GCA_008795845.1	1,270,025	4,493	62,302	871,016
<i>Neophocaena asiaeorientalis</i>	Finless Porpoise	GCA_003031525.1	66,346	86,003	13,699	6,341,296
<i>Pontoporia blainvillei</i>	Franciscana	GCA_004363935.1	1,885,701	2,541	1,885,058	2,541
<i>Eschrichtius robustus</i>	Gray Whale	GCA_002189225.1	375,256	10,066	57,203	187,455
<i>Eschrichtius robustus</i>	Gray Whale	GCA_002738545.1	1,595,257	2,656	1,213,011	10,674
<i>Eschrichtius robustus</i>	Gray Whale	GCA_004363415.1	1,046,770	68,559	1,036,148	94,414
<i>Phocoena phocoena</i>	Harbour Porpoise	GCA_003071005.1	2,347,235	2,773	142,029	27,499,337
<i>Phocoena phocoena</i>	Harbour Porpoise	GCA_004363495.1	1,338,272	89,111	1,331,158	115,969
<i>Phocoena phocoena</i>	Harbour Porpoise	Phocoena_phocoena_HiC	610,275	58,076	565,368	97,795,164
<i>Megaptera novaeangliae</i>	Humpback Whale	GCA_004329385.1	387,694	12,321	2,558	9,138,802
<i>Tursiops aduncus</i>	Indo-Pacific Bottlenose Dolphin	ASM322739v1_HiC	58,538	133,491	12,471	111,961,311
<i>Tursiops aduncus</i>	Indo-Pacific Bottlenose Dolphin	GCA_003227395.1	44,281	206,065	16,249	1,235,788
<i>Sousa chinensis</i>	Indo-Pacific Humpbacked Dolphin	GCA_003521335.2	46,900	182,701	20,903	9,008,636
<i>Sousa chinensis</i>	Indo-Pacific Humpbacked Dolphin	GCA_007760645.1	62,803	113,766	23,368	19,436,979
<i>Platanista minor</i>	Indus river dolphin	GCA_004363435.1	1,110,492	20,879	1,098,790	23,933
<i>Orcinus orca</i>	Killer Whale	GCA_000331955.2	80,100	70,300	1,668	12,735,091
<i>Orcinus orca</i>	Killer Whale	Oorc_1.1_HiC	80,502	70,204	1,617	110,405,485
<i>Globicephala melas</i>	Long-Finned Pilot Whale	ASM654740v1_HiC	21,252	332,801	6,090	106,927,605
<i>Globicephala melas</i>	Long-Finned Pilot Whale	GCA_006547405.1	21,236	332,801	6,637	18,102,937

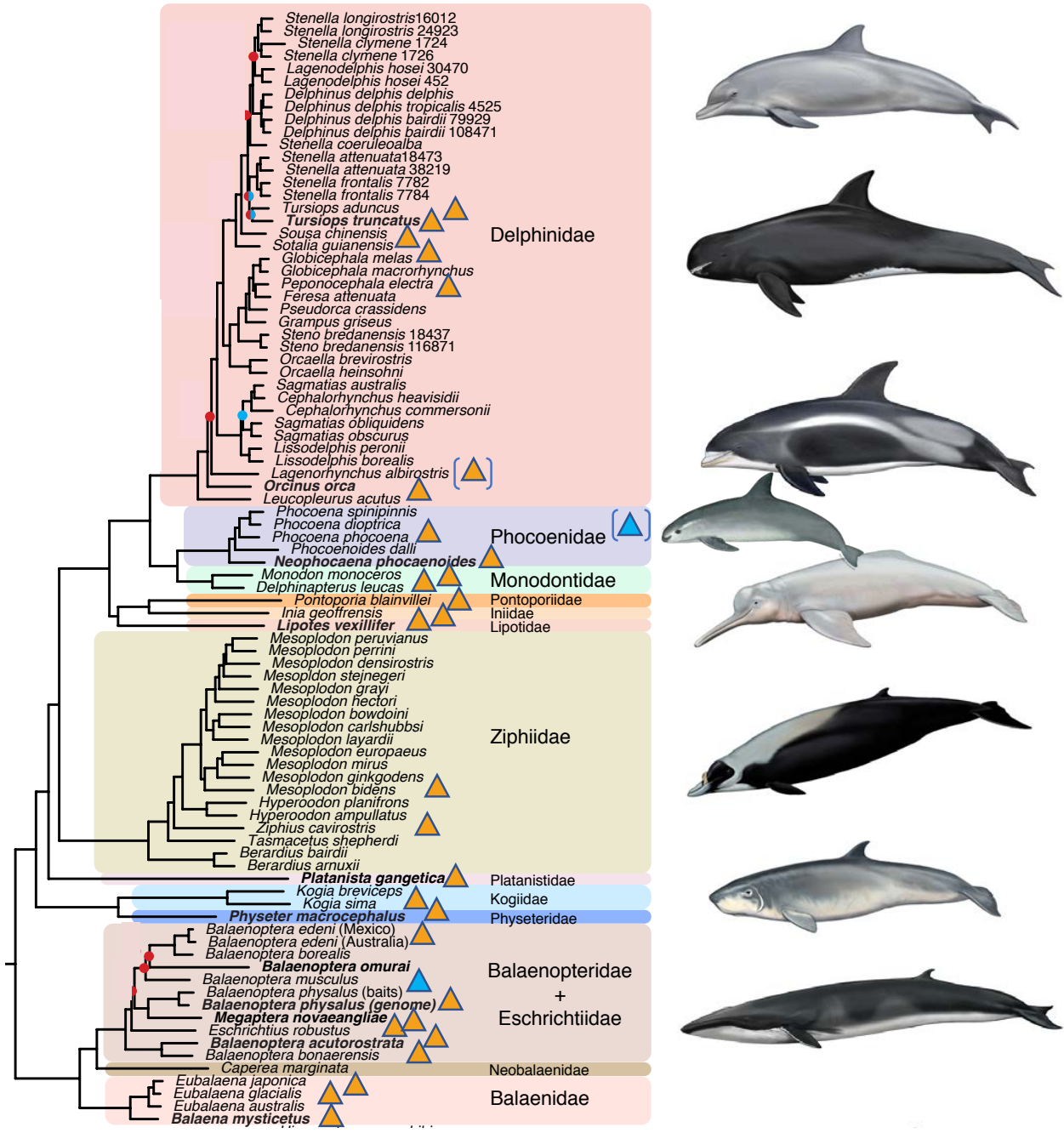
<i>Peponocephala electra</i>	Melon-Headed Whale	Peponocephala_electra_HiC	222,071	84,924	185,978	102,795,557
<i>Monodon monoceros</i>	Narwhal	GCA_004026685.1	653,473	67,024	644,873	86,766
<i>Monodon monoceros</i>	Narwhal	GCA_004027045.1	890,705	70,965	882,704	88,921
<i>Monodon monoceros</i>	Narwhal	GCA_005125345.1	813,468	10,044	21,006	1,483,363
<i>Monodon monoceros</i>	Narwhal	GCA_005190385.2	25,295	255,327	6,972	107,566,389
<i>Eubalaena glacialis</i>	North Atlantic Right Whale	Eubalaena_glacialis_HiC	215,753	65,924	172,124	101,413,572
<i>Eubalaena japonica</i>	North Pacific Right Whale	GCA_004363455.1	1,361,057	34,866	1,353,963	39,813
<i>Lagenorhynchus obliquidens</i>	Pacific White-Sided Dolphin	ASM367639v1_HiC	21,805	255,779	5,162	107,447,310
<i>Lagenorhynchus obliquidens</i>	Pacific White-Sided Dolphin	GCA_003676395.1	21,793	255,779	5,422	28,371,583
<i>Kogia breviceps</i>	Pygmy Sperm Whale	GCA_004363705.1	1,258,125	26,201	1,252,072	28,812
<i>Mesoplodon bidens</i>	Sowerby's Beaked Whale	GCA_004027085.1	1,810,317	28,959	1,801,720	33,532
<i>Physeter macrocephalus</i>	Sperm Whale	GCA_000472045.1	110,443	35,258	11,710	427,290
<i>Physeter macrocephalus</i>	Sperm Whale	GCA_002837175.2	143,605	42,542	14,677	122,182,240
<i>Physeter macrocephalus</i>	Sperm Whale	GCA_900411695.1	140,250	43,829	14,676	122,182,240
<i>Phocoena sinus</i>	Vaquita	GCA_008692025.1*	273	20,218,762	65	115,469,292

408 * VGP “platinum-quality” reference genomes.

409 † from Keane et al., 2015, Cell Reports 10, 112–122, <http://dx.doi.org/10.1016/j.celrep.2014.12.008>

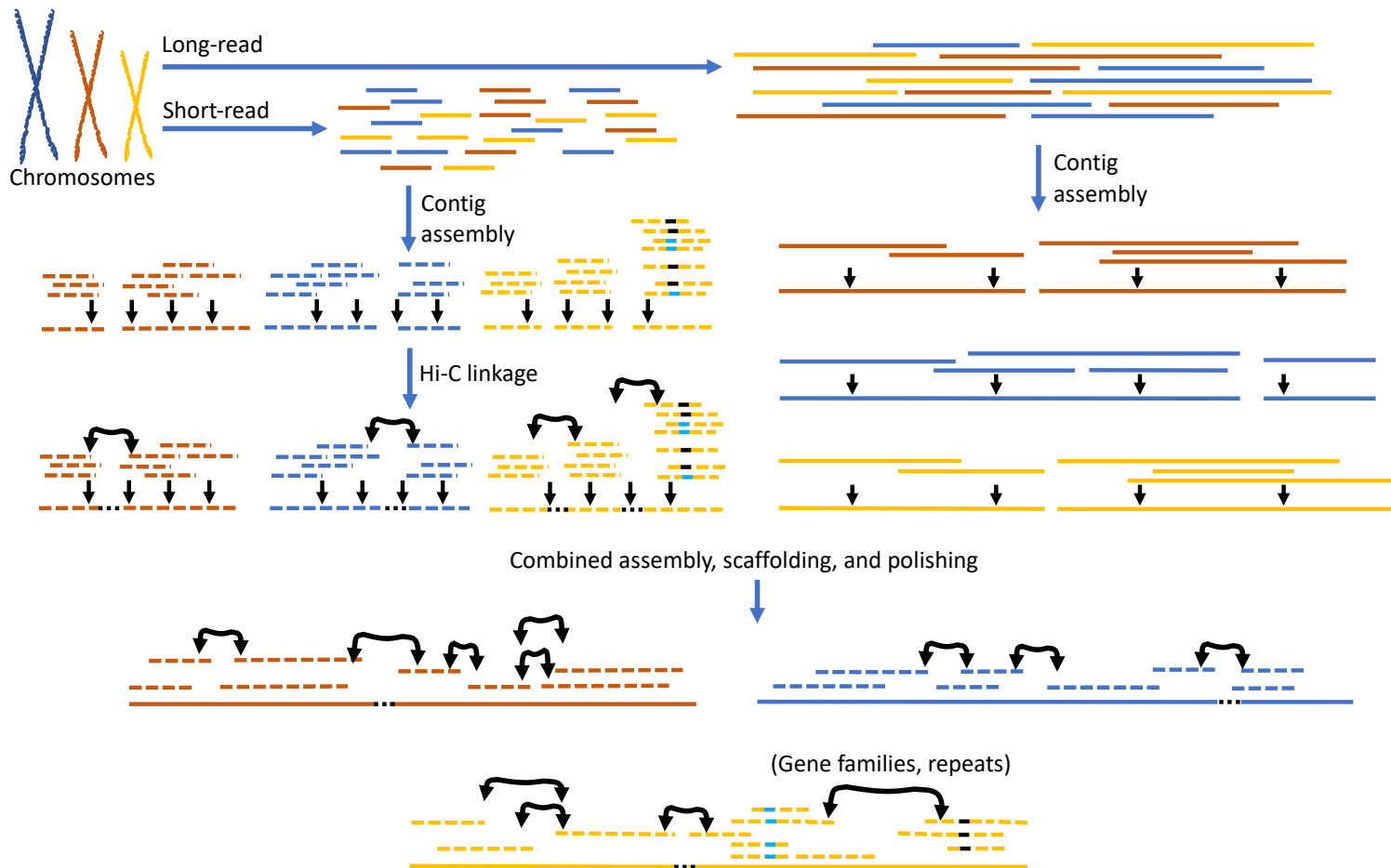
410

411 Figure 1. Phylogeny of the extant cetaceans based on phylogenetic analysis of 3191 protein-
 412 coding nuclear loci, reproduced from McGowen et al. (2019) and modified to show phylogenetic
 413 positions of species with published genome assemblies. Blue triangles mark the species
 414 represented by platinum-quality VGP reference genomes. Orange triangles mark the species for
 415 which draft genomes have been published (from Table 1). Parentheses around the triangles
 416 indicate that the species is not shown in this phylogeny (but the triangle is placed near
 417 congeneric species to indicate approximate position in the phylogeny). Illustrations by Carl
 418 Buell.



419
420

421 Figure 2. Schematic representation of whole genome assembly using short-read or long-read sequencing methods, and combining
 422 them with Hi-C scaffolding to link and order contigs into scaffolds. *De novo* assemblies of short reads result in hundreds of thousands
 423 or millions of short, un-ordered segments. Long read assemblies provide longer, unordered segments that have higher error rates.
 424 Combined long and short read assemblies with Hi-C scaffolding orders the contigs to chromosome-length scaffolds, reduces the
 425 number of gaps to few per chromosome, resolves most repeat regions or duplicates, and improves sequence accuracy. Black dotted
 426 segments represent gaps of unknown length. Blue and black segments within short-reads (e.g., the “yellow” chromosome reads)
 427 indicate small differences between highly similar genes in a gene family or repeat region.
 428



429