



THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

### **Genomic methods take the plunge: recent advances in high-throughput sequencing of marine mammals**

**Citation for published version:**

Cammen, KM, Andrews, KR, Carroll, EL, Foote, AD, Humble, E, Khudyakov, JI, Louis, M, McGowen, MR, Olsen, MT & Van Cise, AM 2016, 'Genomic methods take the plunge: recent advances in high-throughput sequencing of marine mammals' *Journal of Heredity*, vol. 107, no. 6, pp. 481-495. DOI: 10.1093/jhered/esw044

**Digital Object Identifier (DOI):**

[10.1093/jhered/esw044](https://doi.org/10.1093/jhered/esw044)

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Peer reviewed version

**Published In:**

*Journal of Heredity*

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



## Genomic methods take the plunge: recent advances in high-throughput sequencing of marine mammals

Journal:	<i>Journal of Heredity</i>
Manuscript ID	JOH-2016-093.R2
Manuscript Type:	Invited Review
Date Submitted by the Author:	n/a
Complete List of Authors:	<p>Cammen, Kristina; University of Maine, School of Marine Sciences          Andrews, Kim; University of Idaho, Department of Fish and Wildlife Sciences          Carroll, Emma; University of St Andrews, Scottish Oceans Institute          Foote, Andy; University of Bern, Institute of Ecology and Evolution          Humble, Emily; University of Bielefeld, Department of Animal Behaviour;          British Antarctic Survey          Khudyakov, Jane; Sonoma State University, Department of Biology          Louis, Marie; University of St Andrews, Scottish Oceans Institute          McGowen, Michael; Queen Mary University of London, School of Biological          and Chemical Sciences          Olsen, Morten; University of Copenhagen, Natural History Museum of          Denmark          Van Cise, Amy; University of California San Diego, Scripps Institute of          Oceanography</p>
Subject Area:	Genomics and gene mapping
Keywords:	RADseq, target sequence capture, whole genome sequencing, RNAseq, SNP array, non-model organisms

SCHOLARONE™  
Manuscripts

1  
2  
3 **1 Genomic methods take the plunge: recent advances in high-throughput sequencing of**  
4 **2 marine mammals**  
5  
6

7 4 KRISTINA M. CAMMEN<sup>1\*</sup>, KIMBERLY R. ANDREWS<sup>2</sup>, EMMA L. CARROLL<sup>3</sup>, ANDREW D.  
8 5 FOOTE<sup>4</sup>, EMILY HUMBLE<sup>5,6</sup>, JANE I. KHUDYAKOV<sup>7</sup>, MARIE LOUIS<sup>3</sup>, MICHAEL R.  
9 6 MCGOWEN<sup>8</sup>, MORTEN TANGE OLSEN<sup>9</sup>, AND AMY M. VAN CISE<sup>10</sup>  
10 7

11 8 <sup>1</sup>School of Marine Sciences, University of Maine, Orono, Maine 04469, USA

12 9 <sup>2</sup>Department of Fish and Wildlife Sciences, University of Idaho, 875 Perimeter Drive MS 1136,  
13 10 Moscow, Idaho 83844-1136, USA

14 11 <sup>3</sup>Scottish Oceans Institute, University of St Andrews, East Sands, St Andrews, Fife KY16 8LB,  
15 12 UK

16 13 <sup>4</sup>Computational and Molecular Population Genetics Lab, Institute of Ecology and Evolution,  
17 14 University of Bern, Bern CH-3012, Switzerland

18 15 <sup>5</sup>Department of Animal Behaviour, University of Bielefeld, Postfach 100131, 33501 Bielefeld,  
19 16 Germany

20 17 <sup>6</sup>British Antarctic Survey, High Cross, Madingley Road, Cambridge CB3 0ET, UK

21 18 <sup>7</sup>Department of Biology, Sonoma State University, Rohnert Park, California 94928, USA

22 19 <sup>8</sup>School of Biological and Chemical Sciences, Queen Mary University of London,  
23 20 Mile End Road, London E1 4NS, UK

24 21 <sup>9</sup>Evolutionary Genomics Section, Natural History Museum of Denmark, University of  
25 22 Copenhagen, DK-1353 Copenhagen K, Denmark

26 23 <sup>10</sup>Scripps Institution of Oceanography, 8622 Kennel Way, La Jolla, California 92037, USA  
27 24

28 25 \*Corresponding author: kristina.cammen@maine.edu  
29 26  
30 27

31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
Running title: Marine mammal genomics

1  
2  
3 **Abstract**  
4

5 29 The dramatic increase in the application of genomic techniques to non-model organisms over the  
6  
7 30 past decade has yielded numerous valuable contributions to evolutionary biology and ecology,  
8  
9 31 many of which would not have been possible with traditional genetic markers. We review this  
10  
11 32 recent progression with a particular focus on genomic studies of marine mammals, a group of  
12  
13 33 taxa that represent key macroevolutionary transitions from terrestrial to marine environments and  
14  
15 34 for which available genomic resources have recently undergone notable rapid growth. Genomic  
16  
17 35 studies of non-model organisms utilize an expanding range of approaches, including whole  
18  
19 36 genome sequencing, restriction site-associated DNA sequencing, array-based sequencing of  
20  
21 37 single nucleotide polymorphisms and target sequence probes (e.g., exomes), and transcriptome  
22  
23 38 sequencing. These approaches generate different types and quantities of data, and many can be  
24  
25 39 applied with limited or no prior genomic resources, thus overcoming one traditional limitation of  
26  
27 40 research on non-model organisms. Within marine mammals, such studies have thus far yielded  
28  
29 41 significant contributions to the fields of phylogenomics and comparative genomics, as well as  
30  
31 42 enabled investigations of fitness, demography, and population structure. Here we review the  
32  
33 43 primary options for generating genomic data, introduce several emerging techniques, and discuss  
34  
35 44 the suitability of each approach for different applications in the study of non-model organisms.  
36  
37 45

38  
39  
40  
41  
42  
43  
44  
45  
46 **Keywords:** RADseq, SNP array, target sequence capture, whole genome sequencing, RNAseq,  
47 non-model organisms  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 48 **Introduction**

49 Recent advances in sequencing technologies, coincident with dramatic declines in cost, have  
50 increasingly enabled the application of genomic sequencing in non-model systems (Ekblom and  
51 Galindo 2011; Ellegren 2014). These advances in molecular technologies have in many ways  
52 begun to blur the distinction between model and non-model organisms (Armengaud et al. 2014).  
53 Non-model organisms (NMOs) have traditionally been defined as those for which whole-  
54 organism experimental manipulation is rarely, if ever, possible due to logistical and/or ethical  
55 constraints (Ankeny and Leonelli 2011). Further, NMOs have typically been characterized by  
56 limited genomic resources, but this is becoming increasingly less so as the number of NMO  
57 reference genomes grows rapidly, for example through efforts like the Genome 10K Project  
58 (Koepfli et al. 2015). In fact, in some taxonomic orders, we are approaching the point at which  
59 all species have at least one representative reference genome available for a closely related  
60 species (Fig 1).

61  
62 Despite the limitations of working with NMOs, including potentially small sample sizes, low  
63 DNA quantity, and limited information on gene function, genetic and genomic investigations of  
64 NMOs have yielded numerous valuable contributions to understanding their evolutionary  
65 biology and ecology. For the past several decades, traditional genetic markers such as  
66 microsatellites and short fragments of mitochondrial DNA (e.g., the control region) have been  
67 extensively used in molecular ecology. These markers, which typically evolve under neutral  
68 expectations, have proven useful for identifying population structure and reconstructing  
69 population demographic history (Hedrick 2000). However, the power of such studies is limited  
70 by the number of markers that can feasibly be evaluated using traditional approaches. The advent  
71 of low-cost high-throughput sequencing has led to dramatic increases in the number of neutral  
72 markers that can be evaluated, in many cases improving our power to resolve fine-scale or  
73 cryptic population structure in species with high dispersal capability (e.g., Corander et al. 2013)  
74 and improving the accuracy of estimating some (though not all) demographic parameters (Li and  
75 Jakobsson 2012; Shafer et al. 2015). Importantly, high-throughput sequencing has also further  
76 enabled genomic studies of non-neutral processes in NMOs, for example, characterizing both  
77 deleterious and adaptive variation within and across species (Stinchcombe and Hoekstra 2008;

1  
2  
3 78 Künstner et al. 2010). It is increasingly evident that genomic analyses of NMOs can and have  
4  
5 79 provided important insights that could not be identified with traditional genetic markers.  
6  
7 80

8  
9 81 Many molecular ecologists now face the challenge of deciding which of the broad range of  
10  
11 82 genomic approaches to apply to their study systems. Here we review the primary options for  
12  
13 83 generating genomic data and their relative suitability for different applications in the study of  
14  
15 84 NMOs. We focus on marine mammals, which represent several mammalian clades with notably  
16  
17 85 rapid growth in available genomic resources in recent years. This growth is clearly evident in  
18  
19 86 both publication rate (Fig 2) and the rise in number and size of genomic sequences deposited in  
20  
21 87 public resources (Fig 3). We comprehensively review the literature on marine mammal  
22  
23 88 genomics, highlighting recent trends in methodology and applications, and then describe in detail  
24  
25 89 the molecular approaches that are most commonly applied to studies of NMO genomics. Our  
26  
27 90 hope is that this review will highlight the promise of genomics for NMOs and offer guidance to  
28  
29 91 researchers considering the application of genomic techniques in their non-model study system  
30  
31 92 of choice.  
32

### 33 94 **Why study marine mammal genomics?**

34 95 Marine mammals represent key macroevolutionary transitions from terrestrial to marine  
35  
36 96 environments (McGowen et al. 2014) and accordingly are an exemplary system for investigating  
37  
38 97 the evolution of several morphological and physiological adaptations (Foote et al. 2015)  
39  
40 98 associated with locomotion (Shen et al. 2012), sight (Meredith et al. 2013), echolocation (Parker  
41  
42 99 et al. 2013; Zou and Zhang 2015), deep diving (Mirceta et al. 2013), osmoregulation (Ruan et al.  
43  
44 100 2015), and cognition (McGowen et al. 2012). Furthermore, studies of marine mammal evolution  
45  
46 101 to date have characterized several unique aspects of their genome evolution that merit further  
47  
48 102 investigation, including low genomic diversity and a relatively slow molecular clock, especially  
49  
50 103 in cetaceans (Jackson et al. 2009; McGowen et al. 2012; Zhou et al. 2013). As many cetacean  
51  
52 104 species are highly mobile with no obvious physical geographic barriers to dispersal, they provide  
53  
54 105 a unique opportunity to study the role of behavior and culture in shaping population structure and  
55  
56 106 genetic diversity (Riesch et al. 2012; Carroll et al. 2015; Alexander et al. 2016). Though highly  
57  
58 107 mobile, many marine mammals exhibit evidence of local adaptation; for example, several species  
59  
60 108 show parallel divergent morphological and behavioral adaptations to coastal and pelagic

1  
2  
3 109 environments (Moura et al. 2013; Louis et al. 2014; Viricel and Rosel 2014). These species may  
4  
5 110 be studied across ocean basins as emerging examples of ecological adaptation and speciation  
6  
7 111 (Morin et al. 2010a).  
8  
9 112

10 113 Beyond their value as systems of evolutionary study, many marine mammals are also of broader  
11 114 interest relating to their historical and present conservation status. Many marine mammal  
12 115 populations share histories of dramatic decline due to hunting and other human impacts.  
13 116 Genomics provides a promising tool with which to expand our insights into these historical  
14 117 population changes, which so far primarily have been derived from archival review and  
15 118 traditional genetic approaches (Ruegg et al. 2013; Sremba et al. 2015). More recently, since the  
16 119 implementation of national and international protections, many marine mammal populations  
17 120 have partially or fully recovered (Magera et al. 2013), yet the conservation status of certain  
18 121 marine mammal populations remains of concern. Such vulnerable populations could benefit  
19 122 greatly from an improved understanding of their genetic diversity and evolution, especially in  
20 123 ways that can inform predictions of adaptive capacity to anthropogenic pressures and expand the  
21 124 toolkit for conservation policy (Garner et al. 2016; Taylor and Gemzell 2016).  
22  
23 125

### 24 126 **Recent trends in marine mammal genomics**

25 127 We conducted a meta-analysis of the peer-reviewed marine mammal genomics literature to  
26 128 evaluate trends in publication rates across research methodologies and aims. A search of the Web  
27 129 of Science database using the term “genom\*” and one of the following terms indicating study  
28 130 species - “marine mammal”, “pinniped”, “seal”, “sea lion”, “sea otter”, “whale”, “dolphin”,  
29 131 “polar bear”, “manatee” - identified 825 records on December 11, 2015. We excluded 77% of the  
30 132 search results that were not directly related to genomic studies in marine mammal systems. The  
31 133 remaining 101 articles that were relevant to marine mammal genomics were further categorized  
32 134 by primary research methodology and general research aim. A subset of these articles is  
33 135 described briefly in Supplemental Table 1.  
34  
35 136

36 137 From the early 1990s through 2015, published literature in the field shifted from an early focus  
37 138 on mitogenome sequencing to more sequence-intensive approaches, such as transcriptome and  
38 139 whole genome sequencing (Figs 2 and 4). This trajectory closely follows trends in sequencing  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 140 technologies, from Sanger sequencing of short- and long-range PCR products for mitogenome  
4  
5 141 sequencing (Arnason et al. 1991) and SNP discovery (Olsen et al. 2011), to high-throughput  
6  
7 142 sequencing of reduced-representation genomic libraries (RRLs) that consist of selected subsets  
8  
9 143 of the genome (e.g., Viricel et al. 2014), to high-throughput sequencing of whole genomes with  
10  
11 144 varying levels of depth, coverage, and contiguity. Today, high-throughput sequencing can be  
12  
13 145 used both to generate high-quality reference genome assemblies (Yim et al. 2014; Foote et al.  
14  
15 146 2015; Humble et al. 2016) and to re-sequence whole genomes at a population scale (Liu et al.  
16  
17 147 2014a; Foote et al. 2016). Similarly, the scale of gene expression studies has increased from  
18  
19 148 quantitative real-time PCR of candidate genes (Tabuchi et al. 2006) to microarrays containing  
20  
21 149 hundreds to thousands of genes (Mancia et al. 2007) and high-throughput RNAseq that evaluates  
22  
23 150 hundreds of thousands of contigs across the genome (Khudyakov et al. 2015b). As the cost of  
24  
25 151 high-throughput sequencing continues to decline, we anticipate an increase in studies that  
26  
27 152 sequence RRLs, whole genomes, and transcriptomes in NMOs at a population scale.

28  
29 154 Marine mammal genomic studies thus far have primarily contributed to the fields of  
30  
31 155 phylogenomics and comparative genomics (Fig 2, Table S1). Several of these comparative  
32  
33 156 genomics studies have aimed to improve our understanding of the mammalian transition to an  
34  
35 157 aquatic lifestyle and describe the evolutionary relationships within and among marine mammals  
36  
37 158 and their terrestrial relatives (McGowen et al. 2014; Foote et al. 2015). Whereas such studies  
38  
39 159 require only a single representative genome per species, an emerging class of studies applying  
40  
41 160 genomic techniques at a population scale enables further investigations of fitness, demography,  
42  
43 161 and population structure within species (Table S1). However, expanding the scale of genomic  
44  
45 162 studies requires careful selection of an appropriate method for data generation and analysis from  
46  
47 163 a growing number of approaches that are becoming available to non-model systems.

#### 48 **Data generation**

49  
50 166 Our review of marine mammal genomics highlights an increasing number of options for the  
51  
52 167 generation and analysis of genomic data. Choosing which of these sequencing strategies to apply  
53  
54 168 is a key step in any genomics study. Here we describe approaches that have been used  
55  
56 169 successfully in order to help guide future studies of ecological, physiological, and evolutionary  
57  
58 170 genomics in NMOs. Across data generation methods, we highlight approaches that can be used  
59  
60



1  
2  
3 171 with limited or no prior genomic resources, overcoming one traditional challenge of genomic  
4  
5 172 studies of NMOs (the need for a reference genome to which sequencing reads can be mapped).  
6  
7 173 These methods produce a range in quantity and type of data output, from hundreds of SNPs to  
8  
9 174 whole genome sequences, and from single individuals to population samples, reflecting the  
10  
11 175 trade-off between number of samples and amount of data generated per sample.  
12  
13 176

#### 14 177 Sample collection, storage and extraction

15 178 Prior to starting a genomic study, researchers must recognize that many recent methods for high-  
16  
17 179 throughput sequencing require genetic material of much higher quality and quantity than  
18  
19 180 techniques used to characterize traditional genetic markers. These more stringent sample  
20  
21 181 requirements necessitate new standards for tissue sampling, storage, and DNA/RNA extraction.  
22  
23 182 Ideally, samples should be collected from live or newly deceased individuals and stored at -80°C,  
24  
25 183 or when this is not possible at -20°C in RNAlater, Trizol, ethanol, salt-saturated DMSO, or dry,  
26  
27 184 depending on the intended application. Given the sensitivity of new sequencing methods, great  
28  
29 185 care should be taken to minimize cross-contamination during sampling, as even minute amounts  
30  
31 186 of genetic material from another individual can bias downstream analyses, for example variant  
32  
33 187 genotyping and gene expression profiles. Choice of extraction method varies with sample type  
34  
35 188 and study aim, but typically genomic methods require cleanup and treatment with RNase to yield  
36  
37 189 pure extracts, whereas RNAseq methods require rigorous DNase treatment to remove genomic  
38  
39 190 contamination that can bias expression results. Depending on the genomic methodology, target  
40  
41 191 quantities for a final sample may range from as low as 50 ng of DNA for some RRL sequencing  
42  
43 192 methods (Andrews et al. 2016) up to ~1 mg for sequencing the full set of libraries (of different  
44  
45 193 insert sizes) necessary for high-quality genome assemblies (Ekblom and Wolf 2014). Most  
46  
47 194 commercial RNAseq library preparation services require at least 500-1,000 ng of pure total RNA  
48  
49 195 that shows minimal degradation as measured by capillary gel electrophoresis (RNA Integrity  
50  
51 196 Number (RIN)  $\geq$  8). Samples should ideally consist of high molecular weight genetic material  
52  
53 197 (with little shearing), though continuing molecular advances enable genomic sequencing even of  
54  
55 198 low quantity or poor quality starting material. Extreme examples of the latter include  
56  
57 199 successfully sequenced whole genomes from ancient material (e.g., Rasmussen et al. 2010;  
58  
59 200 Meyer et al. 2012; Allentoft et al. 2015), including a more than 500,000-year-old horse (Orlando  
60  
201 et al. 2013).

1  
2  
3 202  
4  
5 203 Reduced-representation genome sequencing  
6  
7 204 *i. RADseq*  
8  
9 205 Reduced-representation sequencing methods evaluate only a small portion of the genome,  
10 206 allowing researchers to sequence samples from a larger number of individuals within a given  
11 207 budget in comparison to sequencing whole genomes. Restriction site-associated DNA  
12 208 sequencing (RADseq) is currently the most widely used RRL sequencing method for NMOs  
13 209 (Davey et al. 2011; Narum et al. 2013; Andrews et al. 2016). RADseq generates sequence data  
14 210 from short regions adjacent to restriction cut sites and therefore targets markers that are  
15 211 distributed relatively randomly across the genome and occur primarily in non-coding regions.  
16 212 This method allows simultaneous discovery and genotyping of thousands of genetic markers for  
17 213 virtually any species, regardless of availability of prior genomic resources. Of greatest interest  
18 214 are variable markers, characterized either as single SNPs or phased alleles that can be resolved  
19 215 from the identification of several variants within a single locus.  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29

30 217 The large number of markers generated by RADseq dramatically increases genomic resolution  
31 218 and statistical power for addressing many ecological and evolutionary questions when compared  
32 219 to studies using traditional markers (Table S1). For example, heterozygosity fitness correlations  
33 220 in harbor seals (*Phoca vitulina*) were nearly fivefold higher when using 14,585 RADseq SNPs  
34 221 than when using 27 microsatellite loci (Hoffman et al. 2014). A recent study on the Atlantic  
35 222 walrus (*Odobenus rosmarus rosmarus*) using 4,854 RADseq SNPs to model demographic  
36 223 changes in connectivity and effective population size associated with the Last Glacial Maximum  
37 224 (Shafer et al. 2015) both supported and extended inferences from previous studies using  
38 225 traditional markers (Shafer et al. 2010; Shafer et al. 2014).  
39  
40  
41  
42  
43  
44  
45  
46  
47

48 227 Furthermore, RADseq can provide sufficient numbers of markers across the genome to identify  
49 228 genomic regions influenced by natural selection. These analyses require large numbers  
50 229 (thousands to tens of thousands) of markers to ensure that some markers will be in linkage  
51 230 disequilibrium with genomic regions under selection and to minimize false positives, particularly  
52 231 under non-equilibrium demographic scenarios (Narum and Hess 2011; De Mita et al. 2013;  
53 232 Lotterhos and Whitlock 2014). Extreme demographic shifts, as experienced by many marine  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 233 mammal populations (e.g., killer whales, Foote et al. 2016), can drive shifts in allele frequencies  
4  
5 234 that confound the distinction of drift and selection and make it difficult to detect genomic  
6  
7 235 signatures of selection (Poh et al. 2014). Proof of concept of the application of RADseq for  
8  
9 236 identifying genomic signatures of selection in wild populations was demonstrated in three-spined  
10  
11 237 sticklebacks (*Gasterosteus aculeatus*), for which analyses of over 45,000 SNPs (Hohenlohe et al.  
12  
13 238 2010) identified genomic regions of known evolutionary importance associated with differences  
14  
15 239 between marine and freshwater forms (Colosimo et al. 2005; Barrett et al. 2008). RADseq  
16  
17 240 studies with similar aims in marine mammals have resulted in comparatively sparser sampling of  
18  
19 241 SNPs (<10,000), likely due to both methodological differences and generally low genetic  
20  
21 242 diversity particularly among cetaceans. Nonetheless, genomic regions associated with resistance  
22  
23 243 to harmful algal blooms in common bottlenose dolphins (*Tursiops truncatus*) were identified  
24  
25 244 across multiple pairwise comparisons using 7,431 RADseq SNPs (Cammen et al. 2015), and  
26  
27 245 genomic regions associated with habitat use and resource specialization in killer whales (*Orcinus*  
28  
29 246 *orca*) were identified using 3,281 RADseq SNPs (Moura et al. 2014a). Some of these RADseq  
30  
31 247 SNPs associated with diet in killer whales were later also confirmed as occurring in genomic  
32  
33 248 regions of high differentiation and reduced diversity consistent with a signature of selection  
34  
35 249 identified in a study utilizing whole genome re-sequencing (Foote et al. 2016). It will remain  
36  
37 250 important for further studies of genomic signatures of selection in NMOs to carefully consider  
38  
39 251 which approach will generate a sufficiently large number of SNPs to accurately identify the  
40  
41 252 range of putatively neutral  $F_{ST}$  values (and thus outliers) given the demographic history of the  
42  
43 253 population (Lotterhos and Whitlock 2014).

44  
45 254  
46  
47 255 Numerous laboratory methods have been developed for generating RADseq data (reviewed in  
48  
49 256 Andrews et al. 2016), with the most popular library preparation methods currently being the  
50  
51 257 original RAD (Miller et al. 2007; Baird et al. 2008), Genotyping by Sequencing (GBS, Elshire et  
52  
53 258 al. 2011; Poland et al. 2012), and double digest RAD (ddRAD, Peterson et al. 2012). All  
54  
55 259 RADseq methods share the common goal of sequencing regions adjacent to restriction cut sites  
56  
57 260 across the genome, but differ in technical details, such as the number and type of restriction  
58  
59 261 enzymes used, the mechanisms for reducing genomic DNA fragment sizes, and the strategies for  
60  
262 attaching sequencing adapters to the target DNA fragments. For example, both the original RAD  
263  
method and GBS use a single enzyme digest, but the original RAD method uses a rare-cutting

1  
2  
3 264 enzyme and mechanical shearing to reduce DNA fragment size (Baird et al. 2008), whereas GBS  
4  
5 265 uses a more frequent-cutting enzyme and relies on preferential PCR amplification of shorter  
6  
7 266 fragments for indirect size selection (Elshire et al. 2011). These modifications lead to differences  
8  
9 267 across methods in the time and cost of library preparation, the number and lengths of loci  
10  
11 268 produced, and the types of error and bias present in the resulting data. Different RADseq  
12  
13 269 methods will be better suited to different research questions, study species, and research budgets,  
14  
15 270 and therefore researchers embarking on a RADseq study should carefully consider the suitability  
16  
17 271 of each method for their individual projects. Further details on the advantages and disadvantages  
18  
19 272 of each method are described in Andrews et al. (2016).

20 273

21 274 *ii. SNP arrays*

22  
23 275 An alternative high-throughput reduced-representation genotyping approach involves the use of  
24  
25 276 custom arrays designed to capture and sequence targeted regions of the genome. Such array-  
26  
27 277 based approaches may provide certain advantages over RADseq, including the ability to easily  
28  
29 278 estimate genotyping error rates, scalability to thousands of samples, lower requirements for DNA  
30  
31 279 quantity/quality and technical effort, greater comparability of markers across studies, and the  
32  
33 280 ability to genotype SNPs within candidate genomic regions. However, unlike RADseq, array-  
34  
35 281 based techniques require prior knowledge of the study system's genome or the genome of a  
36  
37 282 closely related species, which remains unavailable for some NMOs. Furthermore, SNP arrays  
38  
39 283 must take into account the potential for ascertainment bias (e.g., Malenfant et al. 2015), whereas  
40  
41 284 RADseq avoids ascertainment bias by simultaneously discovering and genotyping markers.

42 285

43 286 To identify SNPs for NMO array development, researchers must rely on existing genomic  
44  
45 287 resources or generate new reference sequences, in the form of whole or reduced-representation  
46  
47 288 genomes or transcriptomes (Hoffman et al. 2012; Malenfant et al. 2015). When a whole genome  
48  
49 289 reference assembly is available for the target species or a related species, multiplex shotgun  
50  
51 290 sequencing can facilitate the rapid discovery of hundreds of thousands of SNPs for array  
52  
53 291 development. This SNP discovery approach involves high-throughput sequencing of sheared  
54  
55 292 genomic DNA that can be sequenced at a low depth of coverage (i.e., low mean read depth  
56  
57 293 across the genome) if suitable genotype likelihood-based methods (O'Rawe et al. 2015) are used  
58  
59 294 to identify polymorphic sites. Thus, this approach is less restrictive in terms of DNA quality. For  
60

1  
2  
3 295 example, shotgun sequencing of 33 Northeast Atlantic common bottlenose dolphins, which  
4  
5 296 included degraded DNA collected from stranded specimens, on one Illumina HiSeq2000 lane of  
6  
7 297 100 bp single-end sequencing identified 440,718 high-quality SNPs (M. Louis unpublished data).  
8  
9 298 Such dense sampling of SNPs is essential for studies of population genomics that require a large  
10  
11 299 number of markers, such as for inferences of demographic history (Gutenkunst et al. 2009;  
12  
13 300 Excoffier et al. 2013; Liu and Fun 2015) and selective sweeps (Chen et al. 2010). Once a set of  
14  
15 301 putative markers has been identified, hybridization probes can be designed from their flanking  
16  
17 302 sequences and printed onto a SNP array. The two principal SNP genotyping platforms supporting  
18  
19 303 thousands to millions of SNPs are the Illumina Infinium iSelect® and Affymetrix Axiom®  
20  
21 304 arrays.

22  
23 306 The use of SNP arrays in NMOs has thus far been somewhat limited, potentially due to low SNP  
24  
25 307 validation rates (Chancerel et al. 2011; Helyar et al. 2011), issues of ascertainment bias  
26  
27 308 (Albrechtsen et al. 2010; McTavish and Hillis 2015), and cost of SNP discovery. However, using  
28  
29 309 both SNP data and whole genome sequence from the Antarctic fur seal (*Arctocephalus gazella*),  
30  
31 310 Humble et al. (2016) recently demonstrated that careful filtering based on SNP genomic context  
32  
33 311 prior to array development has the potential to substantially increase assay success rates. Further,  
34  
35 312 ascertainment bias can be reduced by selecting samples for SNP discovery that span the  
36  
37 313 geographic range of populations that will be target sequenced (Morin et al. 2004). By accounting  
38  
39 314 for ascertainment bias, Malenfant et al. (2015) were able to demonstrate population structure in  
40  
41 315 Canadian polar bears (*Ursus maritimus*) more clearly using a 9K SNP array than 24  
42  
43 316 microsatellite markers.

44 317  
45 318 *iii. Target sequence capture*

46 319 Target sequence capture (TSC, also called target enrichment, direct selection, or Hyb-seq) has  
47  
48 320 many of the same advantages and disadvantages as the array-based SNP approaches described  
49  
50 321 above, but differs in library preparation, sequencing platform, and resulting sequence data. While  
51  
52 322 SNP arrays genotype single variable positions, TSC can be used to sequence selected short  
53  
54 323 fragments. With TSC, researchers can amplify and sequence up to a million target probes on  
55  
56 324 solid-state arrays, and even more if in-solution arrays are used. This gives the user the ability to  
57  
58 325 choose to sequence many samples in parallel (Cummings et al. 2010), as many as 100-150 per  
59  
60

1  
2  
3 326 Illumina HiSeq lane, or to sequence many regions per individual. Recent advances in target  
4  
5 327 enrichment, such as genotyping in thousands (Campbell et al. 2015), anchored hybrid enrichment  
6  
7 328 (Lemmon et al. 2012), and target capture of ultraconserved elements (UCEs, Faircloth et al.  
8  
9 329 2012; McCormack et al. 2012), have further increased the number of regions and individuals that  
10  
11 330 can be sampled in a single lane. In addition, UCEs overcome the need for a reference genome,  
12  
13 331 enabling their wide application across many NMOs (though designing custom probe sets from  
14  
15 332 closely related species will remain preferable in many cases (Hancock-Hanser et al. 2013)).  
16  
17 333 Although a number of methodological variants have been developed and optimized (Bashiardes  
18  
19 334 et al. 2005; Noonan et al. 2006; Hodges et al. 2009; Cummings et al. 2010; Mamanova et al.  
20  
21 335 2010; Hancock-Hanser et al. 2013), TSC generally relies on hybridization and amplification of  
22  
23 336 specially prepared libraries consisting of fragmented genomic DNA. Many companies offer kits  
24  
25 337 for TSC, such as Agilent (SureSelect) and MYcroarray (MYbaits), with MYcroarray specifically  
26  
27 338 marketing their kits for use with NMOs.

28  
29 340 The most common use of TSC has been the capture of whole exomes in model organisms,  
30  
31 341 including humans (Ng et al. 2009). However, as costs have plummeted, TSC is increasingly  
32  
33 342 being used in investigations of NMOs. TSC is particularly useful in sequencing ancient DNA,  
34  
35 343 where it can enrich the sample for endogenous DNA content relative to exogenous DNA (i.e.,  
36  
37 344 contamination) and thereby increase the relative DNA yield (Ávila-Arcos et al. 2011; Enk et al.  
38  
39 345 2014). For example, TSC has been used to generate mitogenome sequences from subfossil killer  
40  
41 346 whale specimens originating from the mid-Holocene for comparison with modern lineages  
42  
43 347 (Foote et al. 2013). TSC was also recently utilized to compare >30 kb of exonic sequence from  
44  
45 348 museum specimens of the extinct Steller's sea cow (*Hydrodamalis gigas*) and a modern dugong  
46  
47 349 (*Dugong dugon*) specimen to investigate evolution within Sirenia (Springer et al. 2015). Springer  
48  
49 350 et al. (2016) further used TSC to examine gene evolution related to dentition across edentulous  
50  
51 351 mammals, including mysticetes. Finally, TSC of both exonic and intronic regions has been used  
52  
53 352 to assess genetic divergence across cetacean species (Hancock-Hanser et al. 2013; Morin et al.  
54  
55 353 2015). These studies show the potential use of TSC across evolutionary timescales for population  
56  
57 354 genomics, phylogenomics, and studies of selection and gene loss across divergent lineages  
58  
59 355 (Table S1).

60 356

1  
2  
3 357 Whole genome sequencing

4  
5 358 Beyond advances enabled by the reduced-representation methods presented above, our power  
6  
7 359 and resolution to elucidate evolutionary processes, including selection and demographic shifts,  
8  
9 360 can be further increased by sequencing whole genomes.

10 361

11  
12 362 *i. Reference genome sequencing*

13  
14 363 At the time of publication, there are 12 publicly available<sup>1</sup> whole (or near-whole) marine  
15  
16 364 mammal genomes of varying quality representing 10 families, including 7 cetaceans (Fig 1A), 3  
17  
18 365 pinnipeds (Fig 1B), the West Indian manatee (*Trichechus manatus*), and the polar bear. The first  
19  
20 366 sequenced marine mammal genome was that of the common bottlenose dolphin, which was  
21  
22 367 originally sequenced to ~2.5x depth of coverage using Sanger sequencing (Lindblad-Toh et al.  
23  
24 368 2011). This genome was later improved upon by adding both 454 and Illumina HiSeq data  
25  
26 369 (Foote et al. 2015). Other subsequent marine mammal genomes were produced solely using  
27  
28 370 Illumina sequencing and mate-paired or paired-end libraries with varied insert sizes (Miller et al.  
29  
30 371 2012; Zhou et al. 2013; Yim et al. 2014; Foote et al. 2015; Keane et al. 2015; Kishida et al. 2015;  
31  
32 372 Humble et al. 2016).

33  
34 374 Whole genome sequencing has been used to address many issues in marine mammal genome  
35  
36 375 evolution, usually by comparison with other existing mammalian genomes. Biological insights  
37  
38 376 discussed in the genome papers listed above include the evolution of transposons and repeat  
39  
40 377 elements, gene evolution and positive selection, predicted population structure through time,  
41  
42 378 SNP validation, molecular clock rates, and convergent molecular evolution (Table S1). For  
43  
44 379 example, analyses of the Yangtze river dolphin (*Lipotes vexillifer*) genome confirmed that a  
45  
46 380 bottleneck occurred in this species during the last period of deglaciation (Zhou et al. 2013). In  
47  
48 381 addition, following upon earlier smaller-scale studies (e.g., Deméré et al. 2008; McGowen et al.  
49  
50 382 2008; Hayden et al. 2010), genomic analyses have confirmed the widespread decay of gene  
51  
52 383 families involved in olfaction, gustation, enamelogenesis, and hair growth in some cetaceans  
53  
54 384 (Yim et al. 2014; Kishida et al. 2015). Perhaps the most widespread use of whole genome studies

54  
55  
56  
57  
58  
59  
60  

---

<sup>1</sup> These genomes are available on NCBI's online genome database or Dryad, but they have not all been published. As agreed upon in the Fort Lauderdale Convention, the community standard regarding such unpublished genomic resources is to respect the data generators' right to publish with these data first.

1  
2  
3 385 has been the use of models of selection to detect protein-coding genes that show evidence of  
4  
5 386 natural selection in specific lineages. A recent study by Foote et al. (2015) extended this  
6  
7 387 approach to investigate convergent positive selection among cetaceans, pinnipeds, and sirenians.  
8  
9 388 This study exemplifies a trend in recent genomic studies that sequence multiple genomes to  
10  
11 389 address a predetermined evolutionary question, in this case, the molecular signature of aquatic  
12  
13 390 adaptation.

14 391  
15  
16 392 In addition to these evolutionary insights that typically stem from a comparative genomics  
17  
18 393 approach, the development of high-quality reference genome assemblies provide an important  
19  
20 394 resource that facilitates mapping of reduced-representation genomic data (see previous section)  
21  
22 395 as well as short-read sequencing data with relatively low depth of coverage (see following  
23  
24 396 section). These data types can be generated at relatively low cost on larger sample sizes enabling  
25  
26 397 population-scale genomic studies. In many cases, genome assemblies from closely related  
27  
28 398 species are sufficient for use as a reference. Particularly among marine mammals, given their  
29  
30 400 generally slow rate of nucleotide divergence, it is therefore likely unnecessary to sequence a  
31  
32 401 high-quality reference genome assembly for every species. Instead, resources could be allocated  
33  
34 402 toward population-scale studies, including genome re-sequencing efforts.

35 403 *ii. Population-level genome re-sequencing*  
36  
37 404 In contrast to reference genome sequencing that today often exceeds 100x mean read depth and  
38  
39 405 typically combines long- and short-insert libraries to generate high-quality assemblies for one to  
40  
41 406 a few individuals, genome re-sequencing studies aim to achieve only  $\geq 2x$  mean read depth on  
42  
43 407 tens to hundreds of individuals from short-insert libraries whose reads are anchored to existing  
44  
45 408 reference assemblies. Despite the inherent trade-offs between cost, read depth, coverage, and  
46  
47 409 sample size, genome re-sequencing of large numbers of individuals for population-level  
48  
49 410 inference can be conducted at a relatively low cost. In the past five years, several influential  
50  
51 411 studies have used genome re-sequencing to advance our understanding of the genomic  
52  
53 412 underpinnings of different biological questions in model systems. For example, population  
54  
55 413 genomics of *Heliconius* butterflies highlighted the exchange of genes between species that  
56  
57 414 exhibit convergent wing patterns (The *Heliconius* Genome Consortium 2012); whole genome re-  
58  
59 415 sequencing of three-spined sticklebacks highlighted the re-use of alleles in replicated  
60



1  
2  
3 416 divergences associated with ecological speciation and local adaptation (Jones et al. 2012); and  
4  
5 417 combined population genomics and phylogenomics have identified regions of the genome  
6  
7 418 associated with variation in beak shape and size in Darwin's finches (Lamichhaney et al. 2015).  
8  
9 419

10 420 To date only two marine mammal population genomics studies using whole genome re-  
11  
12 421 sequencing have been published. These studies involved re-sequencing the genomes of 79  
13  
14 422 individuals from three populations of polar bears (Liu et al. 2014a) and 48 individuals from five  
15  
16 423 evolutionarily divergent ecotypes of killer whale (Foote et al. 2016). The findings of Foote et al.  
17  
18 424 (2016) confirmed results of population differentiation that had previously been established using  
19  
20 425 traditional genetic markers (Morin et al. 2010a). However, the study also provided new insights  
21  
22 426 into the demographic history, patterns of selection associated with ecological niche, and evidence  
23  
24 427 of episodic ancestral admixture that could not have been obtained using traditional markers.  
25 428

26 429 Several new resources have made such population genomic studies economically possible for a  
27  
28 430 greater number of NMOs, including the availability of reference genome assemblies (see section  
29  
30 431 above), relatively low-cost high-throughput sequencing (further increases in throughput expected  
31  
32 432 with the new Illumina HiSeq X Ten (van Dijk et al. 2014)), and crucially, the development of  
33  
34 433 likelihood-based methods that allow estimation of population genetic metrics from re-sequencing  
35  
36 434 data (Fumagalli et al. 2013; O'Rawe et al. 2015). One last consideration is the ease of laboratory  
37  
38 435 methods necessary to generate whole genome re-sequencing data when compared to other  
39  
40 436 methods such as RADseq or TSC. DNA simply needs to be extracted from the samples and,  
41  
42 437 using proprietary kits, built into individually index-amplified libraries that are equimolarly  
43  
44 438 pooled and submitted for sequencing.  
45 439

46 440 Many population genomic analyses are based on the coalescent model that gains most  
47  
48 441 information from the number of independent genetic markers, not the number of individuals  
49  
50 442 sampled. Sample sizes of ~10 individuals are usually considered sufficient (Robinson et al.  
51  
52 443 2014) and have been standard in many genome-wide studies in the eco-evolutionary sciences  
53  
54 444 (Ellegren et al. 2012; Jones et al. 2012). Thus, sampling fewer individuals by whole genome re-  
55  
56 445 sequencing is a salient approach that allows us to consider many more gene trees, whilst  
57  
58 446 continuing to provide robust estimates of per-site genetic metrics (e.g.,  $F_{ST}$ ). The robustness of  
59  
60

1  
2  
3 447 inference from data with low mean read depth across the genome was recently confirmed using a  
4  
5 448 comparison of per-site  $F_{ST}$  estimates for the same sites from high-depth ( $\geq 20x$ ) RADseq data and  
6  
7 449 low-depth ( $\approx 2x$ ) whole genome re-sequencing data in pairwise comparisons between the same  
8  
9 450 two killer whale ecotypes (Foote et al. 2016).

10 451  
11  
12 452 Beyond the increased power afforded by sequencing more polymorphic sites, whole genome re-  
13  
14 453 sequencing also allows inference of demographic history from the genome of even just a single  
15  
16 454 individual by identifying Identical By Descent (IBD) segments and runs of homozygosity (Li  
17  
18 455 and Durbin 2011; Harris and Nielsen 2013). For example, Liu et al. (2014a) found evidence for  
19  
20 456 ongoing gene flow from polar bears into brown bears after the two species initially diverged.  
21  
22 457 Genome re-sequencing of sufficient numbers of individuals also facilitates haplotype phasing,  
23  
24 458 which has many applications, including the detection of ongoing selective sweeps (Ferrer-  
25  
26 459 Admetlla et al. 2014) and the inference of demographic history of multiple populations based on  
27  
28 460 coalescence of pairs of haplotypes in different individuals (Schiffels and Durbin 2014).  
29  
30 461 However, haplotype phasing typically requires genomic data with higher mean read depth ( $\sim 20x$ )  
31  
32 462 from tens of individuals (though recent advances in genotype imputation suggest success with  
33  
34 463 data of lower mean read depth (VanRaden et al. 2015)). Thus far, phasing has been restricted to  
35  
36 464 relatively few NMO studies, and no marine mammal studies to the best of our knowledge.

37 465  
38 466 Transcriptome sequencing  
39 467 In comparison with the DNA-based genomic approaches described above, RNA-based genomic  
40  
41 468 approaches are a relatively new and emerging application in NMOs such as marine mammals.  
42  
43 469 Transcriptomics by RNA sequencing (RNAseq) can rapidly generate vast amounts of  
44  
45 470 information regarding genes and gene expression without any prior genomic resources. This  
46  
47 471 approach can resolve differences in global gene expression patterns between populations,  
48  
49 472 individuals, tissues, cells, and physiological or environmental conditions, and can yield insights  
50  
51 473 into the molecular basis of environmental adaptation and speciation in wild animals (Wolf 2013;  
52  
53 474 Alvarez et al. 2015). Furthermore, RNAseq is a valuable tool for resource development, for  
54  
55 475 example as a precursor to designing SNP and TSC arrays (e.g., Hoffman et al. 2012). However,  
56  
57 476 applying RNAseq to NMOs requires several unique considerations in comparison to the DNA-  
58  
59 477 based methods described above. Most importantly, the labile nature of gene transcription and  
60

1  
2  
3 478 high detection sensitivity of RNAseq have the potential to amplify transcriptional “noise” and  
4  
5 479 are thus extremely sensitive to experimental design.  
6  
7 480

8  
9 481 If the experimental goal is to capture a comprehensive transcriptome profile for a study  
10 482 organism, multiple tissues from individuals of varied life history stages should be sampled.  
11 483 However, if the aim is to characterize transcriptional responses to physiological or environmental  
12 484 stimuli, efforts should focus on minimizing variability in individuals and sampling conditions  
13 485 (Wolf 2013). For differential expression analyses, pairwise comparisons should be made within  
14 486 the same individual if at all possible (e.g., before and after treatment, between two  
15 487 developmental stages). As RNAseq only captures a ‘snapshot’ of gene expression in time,  
16 488 repeated sampling or time-course studies are necessary to obtain a more complete picture of  
17 489 cellular responses to the condition(s) in question (Spies and Ciaudo 2015). Sampling and  
18 490 sequencing depth requirements will depend on the study design. Simulation studies have shown  
19 491 that a minimum of 5-6 biological replicates sequenced at a depth of 10-20 million reads per  
20 492 sample is necessary for differential expression analysis (Liu et al. 2014b; Schurch et al. 2015).  
21 493 RNAseq can also be used for biomarker development to expand molecular toolkits for NMOs  
22 494 without sequenced genomes (Hoffman et al. 2013). In this case, higher sequencing depths of 30-  
23 495 60 million reads per sample are recommended for SNP discovery and genotyping (De Wit et al.  
24 496 2015).  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37 497

38  
39 498 Following sequence generation, transcript annotation remains a challenge for NMOs without  
40 499 reference transcriptomes or genomes. *De novo* transcriptomes can be annotated through detection  
41 500 of assembled orthologs of highly conserved proteins, but these analyses remain limited by the  
42 501 quality of reference databases. As a result, NMO transcriptomes are biased in favor of highly  
43 502 conserved terrestrial mammal genes and therefore provide an incomplete understanding of  
44 503 animal adaptations to natural environments (Evans 2015). For example, while 70.0% of northern  
45 504 elephant seal (*Mirounga angustirostris*) skeletal muscle transcripts had BLASTx hits to mouse  
46 505 genes, only 54.1% of blubber transcripts could be annotated due to poor representation of this  
47 506 tissue in terrestrial mammal reference proteomes (Khudyakov et al. 2015b).  
48  
49  
50  
51  
52  
53  
54  
55 507

1  
2  
3 508 To date, RNAseq has been used for gene discovery and phylogenomics analyses in Antarctic fur  
4 seal (Hoffman 2011; Hoffman et al. 2013), polar bear (Miller et al. 2012), Indo-Pacific  
5 509 humpback dolphin (*Sousa chinensis* (Gui et al. 2013)), spotted seal (*Phoca largha* (Gao et al.  
6 510 2013)), bowhead whale (*Balaena mysticetus* (Seim et al. 2014)), narrow-ridged finless porpoise  
7 511 (*Neophocaena asiaeorientalis* (Ruan et al. 2015)), and humpback whale (*Megaptera*  
8 512 *novaeangliae* (Tsagkogeorga et al. 2015)) (Table S1). Due to the challenges of repeated  
9 513 sampling of wild marine mammals, few studies have examined cetacean or pinniped  
10 514 transcriptome responses to environmental or experimental stimuli. The majority of such  
11 515 functional gene expression studies have used microarrays (Mancia et al. 2008; Mancia et al.  
12 516 2012; Mancia et al. 2015); however, RNAseq has been employed to profile sperm whale  
13 517 (*Physeter macrocephalus*) skin cell response to hexavalent chromium (Pabuwal et al. 2013) and  
14 518 free-ranging northern elephant seal skeletal muscle response to an acute stress challenge  
15 519 (Khudyakov et al. 2015a; Khudyakov et al. 2015b). With decreasing sequencing costs and  
16 520 improvements in bioinformatics tools, RNAseq has the potential to accelerate molecular  
17 521 discoveries in marine mammal study systems and supplement existing functional genomics  
18 522 approaches.  
19 523  
20 524

### 21 525 Emerging techniques

22 526 In addition to the relatively proven NMO genomic data generation techniques described above, a  
23 527 suite of emerging techniques is entering the field, with exciting promise for exploration of  
24 528 existing and new research areas. For example, high-throughput shotgun sequencing is  
25 529 increasingly being used to identify genetic material from multiple species in a single sample  
26 530 (metagenomics and metatranscriptomics), rather than focus on characterizing variation in a  
27 531 single target individual. These multi-species approaches can be used, for example, to  
28 532 characterize diet from fecal samples (Deagle et al. 2009) and to investigate microbiomes (Nelson  
29 533 et al. 2015), objectives with implications for improving our understanding of both basic ecology  
30 534 and health in natural populations of NMOs. Furthermore, high-throughput sequencing of  
31 535 environmental DNA dramatically increases the throughput of NMO detection in environmental  
32 536 (e.g., seawater) samples (Thomsen et al. 2012), using degenerate primers for multi-species  
33 537 detection rather than requiring the design and implementation of numerous single-species  
34 538 protocols (Foote et al. 2012).  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

539  
540 A second broad area of emerging interest moves beyond the study of variation at the DNA and  
541 RNA levels to examine epigenetic effects of histone modification on gene regulation and  
542 evolution. Epigenomic studies often examine changes in DNA methylation in association with  
543 processes such as cancer and ageing. Such approaches, from targeted gene to genome-wide, have  
544 only very recently and not yet frequently been applied in NMOs. Polanowski et al. (2014) used a  
545 targeted gene approach to examine changes in DNA methylation in age-associated genes,  
546 previously identified in humans and mice, in humpback whales of known age. The most  
547 informative markers were able to estimate humpback whale ages with standard deviations of  
548 approximately 3-5 years, demonstrating the potential transferability of these approaches from  
549 model to non-model organism. Villar et al. (2015) utilized a genome-wide approach – chromatin  
550 immunoprecipitation followed by high-throughput sequencing (ChIPseq) – to examine gene  
551 regulatory element evolution across mammals, including four species of cetaceans. This study  
552 identified highly conserved gene regulatory elements based on their histone modifications  
553 (H3K27ac and H3K4me3), showed that recently evolved enhancers were associated with genes  
554 under positive selection in marine mammals, and identified unique *Delphinus*-specific enhancers.  
555 Finally, reduced-representation epigenomic approaches have also been developed (Gu et al.  
556 2011), and although they have not yet been used in marine mammals to our knowledge, these  
557 techniques could facilitate future studies of how changes in DNA methylation patterns affect  
558 other biological processes, such as stress levels or pregnancy.

559

### 560 **Data analysis**

561 Following the generation of genomic data, researchers must select the most appropriate genomic  
562 analysis (i.e., bioinformatics) pipelines, which often differ significantly from those used in  
563 traditional genetic studies of NMOs. The choice of analysis pipeline will depend on multiple  
564 factors including the availability of a reference genome, the level of diversity within the dataset  
565 (e.g., single- or multi-species), the type of data generated (e.g., single- or paired-end), and the  
566 computing resources available. The computational needs, both in terms of hardware and  
567 competency in computer science, for analysis of genomic data typically far exceed those  
568 necessary for traditional genetic markers. On the smaller end of the spectrum, one lane of 50 bp  
569 single-end sequencing on an Illumina HiSeq 2500 can produce tens of gigabytes of data, while

1  
2  
3 570 data files associated with a single high-quality vertebrate genome may reach hundreds of  
4  
5 571 gigabytes in size (Ekblom and Wolf 2014). Computing resources necessary for the analysis of  
6  
7 572 these genomic datasets can range from ~10 gigabytes for a pilot study using a reduced-  
8  
9 573 representation sequencing approach to over a terabyte for whole genome sequence assembly  
10  
11 574 (Ekblom and Wolf 2014). Fortunately, university computing clusters, cloud-based (Stein 2010)  
12  
13 575 and high-performance computing clusters (e.g., XSEDE; Towns et al. 2014), and open web-  
14  
15 576 based platforms for genomic research (e.g., Galaxy; Goecks et al. 2010) are becoming  
16  
17 577 increasingly accessible. Furthermore, new pipelines are continuously being developed and  
18  
19 578 improved, and there are a growing number of resources aimed at training molecular ecologists  
20  
21 579 and evolutionary biologists in computational large-scale data analysis (Andrews and Luikart  
22  
23 580 2014; Belcaid and Toonen 2015; Benestan et al. 2016). We provide an indicative list of the  
24  
25 581 current, most commonly used analysis pipelines that are specific to each data generation method  
26  
27 582 in Table 1. Here we briefly summarize current genomic data analysis pipelines and discuss  
28  
29 583 considerations that are likely to be similar across multiple data generation methods.

30 584  
31 585 Genomic data analysis often involves multiple steps, and the choice of analysis tool for each step  
32  
33 586 can greatly affect the outcome, with different tools producing different (though usually  
34  
35 587 overlapping) sets of results (e.g., Schurch et al. 2015). All analyses begin by evaluating data  
36  
37 588 quality, trimming sequences if necessary to remove erroneous nucleotides (MacManes 2014),  
38  
39 589 and implementing appropriate data quality filters (e.g., phred scores, read length, and/or read  
40  
41 590 depth). Raw reads also need to be demultiplexed based on unique barcodes if pools of  
42  
43 591 individuals were sequenced in a single lane. Analyses then usually proceed in a *de novo* or  
44  
45 592 genome-enabled manner, depending on available resources. Briefly, sequences can be compared  
46  
47 593 (e.g., to identify variants) by mapping all reads to a reference genome or *de novo* assembling  
48  
49 594 stacks of sequences putatively derived from the same locus based on sequence similarity. *De*  
50  
51 595 *nov*o methods are sensitive to sequencing error, as well as true genetic variation, and therefore  
52  
53 596 can erroneously assemble polymorphic sequences as separate loci or transcripts, requiring further  
54  
55 597 filtering to remove redundancy. The opposite problem can also occur in both *de novo* and  
56  
57 598 reference mapping approaches, where two distinct loci (e.g., paralogous loci) may assemble as a  
58  
59 599 single locus or map to the same reference location. Researchers should therefore recognize the  
60

1  
2  
3 600 inherent trade-offs when carefully selecting their thresholds for acceptable levels of variation  
4  
5 601 within and among loci.  
6

7 602  
8  
9 603 Considerations relevant to the selection of subsequent downstream analyses are specific to the  
10 604 type of data generated and the research objective. For example, RADseq analysis pipelines differ  
11 605 in the algorithms used to genotype variants (Table 1). Similarly, there are several gene  
12 606 expression analysis pipelines for RNAseq data that compare transcript abundance between  
13  
14 607 samples (Table 1). Analysis of TSC data usually uses standard *de novo* assemblers (e.g., Trinity,  
15  
16 608 Velvet); these assemblers can be run using packages such as PHYLUCE (Faircloth 2015), which  
17  
18 609 is designed specifically for use with ultraconserved elements. Unfortunately, for most analyses,  
19  
20 610 there are no unifying recommendations currently available and researchers must evaluate several  
21  
22 611 approaches, each with their own advantages and disadvantages, in order to select the most  
23  
24 612 appropriate tool for their particular experiment and system. Furthermore, we can expect that the  
25  
26 613 recommendations for analysis tools will continue to evolve as new programs become available in  
27  
28 614 the future.  
29

30 615

### 31 32 616 Guidelines for data quality control and sharing

33 617 With rapid growth in sequencing platforms and bioinformatics analysis pipelines comes the need  
34  
35 618 to extend existing principles (e.g., Bonin et al. 2004) on quality control, analysis, and  
36  
37 619 transparency. General recommendations for sample and data handling, library preparation, and  
38  
39 620 sequencing have been discussed elsewhere (Paszkiwicz et al. 2014). We therefore focus on the  
40  
41 621 need to produce guidelines on data quality evaluation and reporting for genomic data (e.g.,  
42  
43 622 Morin et al. 2010b). A primary challenge in this area is that quality metrics vary widely across  
44  
45 623 sequencing technologies. Yet, regardless of sequencing platform, the quality of sequencing reads  
46  
47 624 must be evaluated (e.g., using FastQC; Andrews 2010) and reported.  
48

49 625  
50 626 Best practices guidelines for reference genome sequencing and RNAseq data generation,  
51 627 analysis, and reporting are available from the human-centric ENCODE consortium  
52  
53 628 ([www.encodeproject.org](http://www.encodeproject.org)). These include minimum depth of sequencing and number and  
54  
55 629 reproducibility of biological replicates. For RNAseq experiments, evaluation of *de novo*  
56  
57 630 assembly quality remains a challenge. Suggested quality metrics include percentage of raw reads  
58  
59  
60

1  
2  
3 631 mapping back to the assembly and number of assembled transcripts with homology to known  
4  
5 632 proteins (MacManes 2016). Emerging tools such as Transrate (Smith-Unna et al. 2015) attempt  
6  
7 633 to integrate these and other metrics into a comprehensive assembly quality score.  
8  
9 634

10 635 In contrast, there is not yet any standard way to estimate or report error rates with RADseq or  
11  
12 636 genome re-sequencing methods (but see Mastretta-Yanes et al. 2015; Fountain et al. 2016).  
13  
14 637 Recommendations to improve confidence in genotyping include using methods that account for  
15  
16 638 population-level allele frequencies when calling individual genotypes, mapping reads to  
17  
18 639 reference genomes rather than *de novo* assembly (Nadeau et al. 2014; Fountain et al. 2016),  
19  
20 640 filtering out PCR duplicates (Andrews et al. 2014), identifying and removing markers in possible  
21  
22 641 repeat regions, and filtering data to include only those with high read depth (>10-20x per locus  
23  
24 642 per individual) (Nielsen et al. 2011). Other analysis methods, such as robust Bayesian methods  
25  
26 643 and likelihood-based approaches that account for read quality in calculations of posterior  
27  
28 644 probabilities of genotypes and per-site allele frequencies utilizing the sample mean site  
29  
30 645 frequency spectrum as a prior (Fumagalli et al. 2013), can account for uncertainty and/or error in  
31  
32 646 the data, and are therefore suitable for use with low to moderate read depths (2-20x per locus;  
33  
34 647 e.g., Han et al. 2015; O'Rawe et al. 2015).  
35  
36 648

37 649 Due to the large number of analysis tools that are available, data quality and reproducibility  
38  
39 650 ultimately depend on methods and data transparency. All raw sequencing reads should be  
40  
41 651 publicly archived, for example deposited in the NCBI Sequence Read Archive. Many journals,  
42  
43 652 including the *Journal of Heredity* (Baker 2013), now also require that primary data supporting  
44  
45 653 the published results and conclusions (e.g., SNP genotypes, assemblies) be publicly archived in  
46  
47 654 online data repositories (e.g., Dryad). We further recommend making public the analysis  
48  
49 655 pipelines, scripts (e.g., using GitHub), and additional outputs, as appropriate, in order for  
50  
51 656 analyses to be fully reproducible and transparent, which is the cornerstone of the scientific  
52  
53 657 method (Nosek et al. 2015).  
54  
55 658

### 56 659 **Future directions**

57 660 As demonstrated here for one group of mammalian taxa, the rapid growth of the field of non-  
58  
59 661 model genomics has been both impressive and empowering. As we approach a point of relative  
60



1  
2  
3 662 saturation in reference genomes, we anticipate an increase in population-scale genomic studies  
4  
5 663 that produce lower depth or coverage datasets per individual but across larger sample sizes. In  
6  
7 664 addition (or alternatively), we hope to see increasing efforts to sequence reference transcriptomes  
8  
9 665 and improve NMO genome annotation in ways beyond the inherently limited approach of  
10  
11 666 comparison to gene lists from a few model organisms. Population-scale genomic studies will  
12  
13 667 facilitate greater ecological understanding of natural populations, while efforts to improve  
14  
15 668 annotation will address persistent limitations in our understanding of gene function for NMOs.  
16  
17 669 Ultimately, improving our understanding of local adaptation, adaptive potential, and  
18  
19 670 demographic history through the use of genomic toolkits such as those described here is likely to  
20  
21 671 have important implications for the future conservation of these populations.  
22

23 673 Advances in sequencing technologies and analytical tools will no doubt continue, in some cases  
24  
25 674 drawing on established techniques in model organisms, posing both new opportunities and new  
26  
27 675 challenges for researchers in NMO genomics. Likely the most persistent challenge will remain  
28  
29 676 selecting the data generation and experimental design that is most appropriate for the respective  
30  
31 677 research objective. Our review identified few cases that exhibit relative dominance of a single  
32  
33 678 methodology and analytical pipeline (e.g., RADseq and STACKS, RNAseq and Trinity); rather,  
34  
35 679 more often we found a diversity of approaches even within each category of data generation. In  
36  
37 680 fact, such diversity of approaches has its benefits, with each approach promoting its own  
38  
39 681 advantages (and limitations). Overall, our reflections on lessons learned from the past decade of  
40  
41 682 NMO genomics in one well-studied group of mammalian taxa clearly demonstrate the value,  
42  
43 683 increased ease, and future promise of applying genomic techniques across a wide range of non-  
44  
45 684 model species to gain previously unavailable insights into evolution, population biology, and  
46  
47 685 physiology on a genome-wide scale.  
48

#### 48 687 **Acknowledgements**

49 688 This review paper is the outcome of two international workshops held in 2013 and 2015 on  
50  
51 689 marine mammal genomics. The workshops were organized by KMC, AF, and C. Scott Baker and  
52  
53 690 hosted by the Society for Marine Mammalogy, with support from a Special Event Award from  
54  
55 691 the American Genetic Association. We sincerely thank all the workshop participants for their  
56  
57 692 contributions to inspiring discussions on marine mammal genomics. We would also like to thank  
58  
59  
60

693 two anonymous reviewers and C. Scott Baker for their helpful feedback on an earlier version of  
694 this manuscript. Illustrations are by C. Buell with permission for use granted by J. Gatesy.

## 696 **Funding**

697 The authors involved in this work were supported by a National Science Foundation Postdoctoral  
698 Research Fellowship in Biology (Grant No. 1523568) to KMC; an Office of Naval Research  
699 Award (No. N00014-15-1-2773) to JIK; a Marie Slodowska Curie Fellowship to ELC  
700 (Behaviour-Connect) funded by the EU Horizon2020 program; Royal Society Newton  
701 International Fellowships to ELC and MRM; a Deutsche Forschungsgemeinschaft studentship to  
702 EH; a Fyssen Foundation postdoctoral fellowship to ML; postdoctoral funding from the  
703 University of Idaho College of Natural Resources to KRA; a short visit grant from the European  
704 Science Foundation-Research Networking Programme ConGenOmics to ADF; and a Swiss  
705 National Science Foundation grant (31003A-143393) to L. Excoffier that further supported ADF.  
706 The first marine mammal genomics workshop we held to begin discussions towards this review  
707 was supported by a Special Event Award from the American Genetic Association.

## 709 **References**

- 710 Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures  
711 of population divergence. *Mol Biol Evol.* 27:2534-2547.
- 712 Alexander A, Steel D, Hoekzema K, Mesnick S, Engelhaupt D, Kerr I, Payne R, Baker CS.  
713 2016. What influences the worldwide genetic structure of sperm whales (*Physeter*  
714 *macrocephalus*)? *Mol Ecol.*
- 715 Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB,  
716 Schroeder H, Ahlstrom T, Vinner L, *et al.* 2015. Population genomics of Bronze Age  
717 Eurasia. *Nature.* 522:167-172.
- 718 Alvarez M, Schrey AW, Richards CL. 2015. Ten years of transcriptomics in wild populations:  
719 what have we learned about their ecology and evolution? *Mol Ecol.* 24:710-725.
- 720 Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome*  
721 *Biol.* 11:R106.
- 722 Andrews K, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of  
723 RADseq for ecological and evolutionary genomics. *Nat Rev Genet.* 17:81-92.
- 724 Andrews KR, Hohenlohe PA, Miller MR, Hand BK, Seeb JE, Luikart G. 2014. Trade-offs and  
725 utility of alternative RADseq methods: Reply to Puritz *et al.* 2014. *Mol Ecol.* 23:5943-  
726 5946.
- 727 Andrews KR, Luikart G. 2014. Recent novel approaches for population genomics data analysis.  
728 *Mol Ecol.* 23:1661-1667.
- 729 Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available  
730 online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>

- 1  
2  
3 731 Ankeny RA, Leonelli S. 2011. What's so special about model organisms? *Studies in History and*  
4 732 *Philosophy of Science*. 42:313-323.
- 5 733 Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM. 2014. Non-model  
6 734 organisms, a species endangered by proteogenomics. *J Proteomics*. 105:5-18.
- 7 735 Arnason U, Adegoke JA, Bodin K, Born EW, Esa YB, Gullberg A, Nilsson M, Short RV, Xu X,  
8 736 Janke A. 2002. Mammalian mitogenomic relationships and the root of the eutherian tree.  
9 737 *Proc Natl Acad Sci USA*. 99:8151-8156.
- 10 738 Arnason U, Gullberg A, Widegren B. 1991. The complete nucleotide sequence of the  
11 739 mitochondrial DNA of the fin whale, *Balaenoptera physalus*. *J Mol Evol*. 33:556-568.
- 12 740 Ávila-Arcos M, Cappellini E, Romero-Navarro JA, Wales N, Moreno-Mayar JV, Rasmussen M,  
13 741 Fordyce SL, Montiel R, Vielle-Calzada J-P, Willerslev E, *et al*. 2011. Application and  
14 742 comparison of large-scale solution-based DNA capture-enrichment methods on ancient  
15 743 DNA. *Sci Rep*. 1:74.
- 16 744 Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA,  
17 745 Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD  
18 746 markers. *PLoS One*. 3:e3376.
- 19 747 Baker CS. 2013. *Journal of Heredity* adopts Joint Data Archiving Policy. *J Hered*. 104:1.
- 20 748 Barrett RDH, Rogers SM, Schluter D. 2008. Natural selection on a major armor gene in  
21 749 threespine stickleback. *Science*. 322:255-257.
- 22 750 Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. 2005. Direct genomic  
23 751 selection. *Nat Methods*. 2:63-69.
- 24 752 Belcaid M, Toonen RJ. 2015. Demystifying computer science for molecular ecologists. *Mol*  
25 753 *Ecol*. 24:2619-2640.
- 26 754 Benestan LM, Ferchaud A-L, Hohenlohe PA, Garner BA, Naylor GJP, Baums IB, Schwartz MK,  
27 755 Kelley JL, Luikart G. 2016. Conservation genomics of natural and managed populations:  
28 756 building a conceptual and practical framework. *Mol Ecol*.
- 29 757 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina Sequence  
30 758 Data. *Bioinformatics*. 30:2114-2120.
- 31 759 Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C, Taberlet P. 2004. How  
32 760 to track and assess genotyping errors in population genetics studies. *Mol Ecol*. 13:3261-  
33 761 3273.
- 34 762 Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. 2012. A reference-free algorithm for  
35 763 computational normalization of shotgun sequencing data. *arXiv*. 1203:4802.
- 36 764 Cammen KM, Schultz TF, Rosel PE, Wells RS, Read AJ. 2015. Genomewide investigation of  
37 765 adaptation to harmful algal blooms in common bottlenose dolphins (*Tursiops truncatus*).  
38 766 *Mol Ecol*. 24:4697-4710.
- 39 767 Campbell NR, Harmon SA, Narum SR. 2015. Genotyping-in-Thousands by sequencing (GT-  
40 768 seq): a cost effective SNP genotyping method based on custom amplicon sequencing.  
41 769 *Mol Ecol Resour*. 15:855-867.
- 42 770 Carroll EL, Baker CS, Watson M, Alderman R, Bannister J, Gaggiotti OE, Gröcke DR,  
43 771 Patenaude N, Harcourt R. 2015. Cultural traditions across a migratory network shape the  
44 772 genetic structure of southern right whales around Australia and New Zealand. *Sci Rep*.  
45 773 5:16182.
- 46 774 Catchen JM, Amores A, Hohenlohe PA, Cresko WA, Postlethwait JH. 2011. *Stacks*: building  
47 775 and genotyping loci *de novo* from short-read sequences. *G3*. 1:171-182.

- 1  
2  
3 776 Catchen JM, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool  
4 777 set for population genomics. *Mol Ecol.* 22:3124-2140.
- 5 778 Chancerel E, Lepoittevin C, Le Provost G, Lin Y-C, Jaramillo-Correa JP, Eckert AJ, Wegrzyn  
6 779 JL, Zelenika D, Boland A, Frigerio J-M, *et al.* 2011. Development and implementation of  
7 780 a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative  
8 781 mapping with loblolly pine. *BMC Genomics.* 12:368.
- 9 782 Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps.  
10 783 *Genome Res.* 20:393-402.
- 11 784 Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Grimwood J, Schmutz J, Myers RM,  
12 785 Schluter D, Kingsley DM. 2005. Widespread parallel evolution in sticklebacks by  
13 786 repeated fixation of ectodysplasin alleles. *Science.* 307:1928-1933.
- 14 787 Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal  
15 788 tool for annotation, visualization and analysis in functional genomics research.  
16 789 *Bioinformatics.* 21:3674-3676.
- 17 790 Corander J, Majander KK, Cheng L, Merilä J. 2013. High degree of cryptic population  
18 791 differentiation in the Baltic Sea herring *Clupea harengus*. *Mol Ecol.* 22:2931-2940.
- 19 792 Cummings N, King R, Rickers A, Kaspi A, Lunke S, Haviv I, Jowett JBM. 2010. Combining  
20 793 target enrichment with barcode multiplexing for high throughput SNP discovery. *BMC*  
21 794 *Genomics.* 11:641.
- 22 795 Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide  
23 796 genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev*  
24 797 *Genet.* 12:499-510.
- 25 798 De Mita S, Thuillet A-C, Gay L, Ahmadi N, Manel S, Ronfort J, Vigouroux Y. 2013. Detecting  
26 799 selection along environmental gradients: analysis of eight methods and their effectiveness  
27 800 for outbreeding and selfing populations. *Mol Ecol.* 22:1383-1399.
- 28 801 De Wit P, Pespeni MH, Palumbi SR. 2015. SNP genotyping and population genomics from  
29 802 expressed sequences - current advances and future possibilities. *Mol Ecol.* 24:2310-2323.
- 30 803 Deagle BE, Kirkwood R, Jarman SN. 2009. Analysis of Australian fur seal diet by  
31 804 pyrosequencing prey DNA in faeces. *Mol Ecol.* 18:2022-2038.
- 32 805 Deméré TA, McGowen MR, Berta A, Gatesy J. 2008. Morphological and molecular evidence for  
33 806 a stepwise evolutionary transition from teeth to baleen in mysticete whales. *Syst Biol.*  
34 807 57:15-37.
- 35 808 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del  
36 809 Angel G, Rivas MA, Hanna M, *et al.* 2011. A framework for variation discovery and  
37 810 genotyping using next-generation DNA sequencing data. *Nat Genet.* 43:491-498.
- 38 811 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras  
39 812 TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 29:15-21.
- 40 813 Eaton DAR. 2014. PyRAD: assembly of *de novo* RADseq loci for phylogenetic analysis.  
41 814 *Bioinformatics.* 30:1844-1849.
- 42 815 Ekblom R, Galindo J. 2011. Applications of next generation sequencing in molecular ecology of  
43 816 non-model organisms. *Heredity.* 107:1-15.
- 44 817 Ekblom R, Wolf JBW. 2014. A field guide to whole-genome sequencing, assembly and  
45 818 annotation. *Evolutionary Applications.* 7:1026-1042.
- 46 819 Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms.  
47 820 *Trends Ecol Evol.* 29:51-63.
- 48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 821 Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Künstner A, Mäkinen H,  
4 822 Nadachowska-Brzyska K, Qvarnström A, *et al.* 2012. The genomic landscape of species  
5 823 divergence in *Ficedula* flycatchers. *Nature*. 491:756-760.
- 6 824 Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A  
7 825 robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.  
8 826 *PLoS One*. 6:e19379.
- 9 827 Enk J, Devault A, Kuch M, Murgha Y, Rouillard J-M, Poinar H. 2014. Ancient whole genome  
10 828 enrichment using baits built from modern DNA. *Mol Biol Evol*. 31:1292-1294.
- 11 829 Evans TG. 2015. Considerations for the use of transcriptomics in identifying the 'genes that  
12 830 matter' for environmental adaptation. *J Exp Biol*. 218:1925-1935.
- 13 831 Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic  
14 832 inference from genomic and SNP data. *PLoS Genetics*. 9:e1003905.
- 15 833 Faircloth BC. 2015. PHYLUCE is a software package for the analysis of conserved genomic  
16 834 loci. *Bioinformatics*. 32:786-788.
- 17 835 Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012.  
18 836 Ultraconserved elements anchor thousands of genetic markers spanning multiple  
19 837 evolutionary timescales. *Syst Biol*. 61:717-726.
- 20 838 Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On detecting incomplete soft or  
21 839 hard selective sweeps using haplotype structure. *Mol Biol Evol*. 31:1275-1291.
- 22 840 Flicek P, Birney E. 2009. Sense from sequence reads: methods for alignment and assembly. *Nat*  
23 841 *Methods*. 6:S6-S12.
- 24 842 Foote AD, Liu Y, Thomas GWC, Vinař Ts, Alföldi J, Deng J, Dugan S, van Elk CE, Hunter ME,  
25 843 Joshi V, *et al.* 2015. Convergent evolution of the genomes of marine mammals. *Nat*  
26 844 *Genet*. 47:272-275.
- 27 845 Foote AD, Newton J, Ávila-Arcos MC, Kampmann M-L, Samaniego JA, Post K, Rosing-Asvid  
28 846 A, Sinding M-HS, Gilbert MTP. 2013. Tracking niche variation over millennial  
29 847 timescales in sympatric killer whale lineages. *Proc R Soc Lond B Biol Sci*. 280:20131481.
- 30 848 Foote AD, Thomsen PF, Sveegaard S, Wahlberg M, Kielgast J, Kyhn LA, Salling AB, Galatius  
31 849 A, Orlando L, Gilbert MTP. 2012. Investigating the potential use of environmental DNA  
32 850 (eDNA) for genetic monitoring of marine mammals. *PLoS One*. 7:e41781.
- 33 851 Foote AD, Vijay N, Ávila-Arcos M, Baird RW, Durban JW, Fumagalli M, Gibbs RA, Hanson  
34 852 MB, Korneliussen TS, Martin MD, *et al.* 2016. Genome-culture coevolution promotes  
35 853 rapid divergence of killer whale ecotypes. *Nat Commun*. 7:11693.
- 36 854 Fountain ED, Pauli JN, Reid BN, Palsbøll PJ, Peery MZ. 2016. Finding the right coverage: the  
37 855 impact of coverage and sequence quality on single nucleotide polymorphism genotyping  
38 856 error rates. *Mol Ecol Resour*.
- 39 857 Fumagalli M, Vieira FG, Korneliussen TS, Linderöth T, Huerta-Sánchez E, Albrechtsen A,  
40 858 Nielsen R. 2013. Quantifying population genetic differentiation from next-generation  
41 859 sequencing data. *Genetics*. 195:979-992.
- 42 860 Fumagalli M, Vieira FG, Linderöth T, Nielsen R. 2014. *ngsTools*: methods for population  
43 861 genetics analyses from Next-Generation Sequencing data. *Bioinformatics*. 30:1486-1487.
- 44 862 Gao X, Han J, Lu Z, Li Y, He C. 2013. *De novo* assembly and characterization of spotted seal  
45 863 *Phoca largha* transcriptome using Illumina paired-end sequencing. *Comp Biochem*  
46 864 *Physiol D Genom Proteom*. 8:103-110.
- 47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 865 Garner BA, Hand BK, Amish SJ, Bernatchez L, Foster JT, Miller KM, Morin PA, Narum SR,  
4 866 O'Brien SJ, Roffler G, *et al.* 2016. Genomics in conservation: case studies and bridging  
5 867 the gap between data and application. *Trends Ecol Evol.* 31:81-83.
- 6  
7 868 Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. 2014. TASSEL-  
8 869 GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One.* 9:e90346.
- 9  
10 870 Gnerre S, MacCallum I, Przbylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea  
11 871 TP, Sykes S, *et al.* 2011. High-quality draft assemblies of mammalian genomes from  
12 872 massively parallel sequence data. *Proc Natl Acad Sci USA.* 108:1513-1518.
- 13 873 Goecks J, Nekrutenko A, Taylor J, The Galaxy Team. 2010. Galaxy: a comprehensive approach  
14 874 for supporting accessible, reproducible, and transparent computational research in the life  
15 875 sciences. *Genome Biol.* 11:R86.
- 16 876 Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. 2011. Preparation of reduced  
17 877 representation bisulfite sequencing libraries for genome-scale DNA methylation  
18 878 profiling. *Nat Protoc.* 6:468-481.
- 19 879 Gui D, Jia K, Xia J, Yang L, Chen J, Wu Y, Yi M. 2013. *De novo* assembly of the Indo-Pacific  
20 880 humpback dolphin leucocyte transcriptome to identify putative genes involved in the  
21 881 aquatic adaptation and immune response. *PLoS One.* 8:e72417.
- 22 882 Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint  
23 883 demographic history of multiple populations from multidimensional SNP frequency data.  
24 884 *PLoS Genetics.* 5:e1000695.
- 25 885 Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D,  
26 886 Li B, Lieber M, *et al.* 2013. *De novo* transcript sequence reconstruction from RNA-seq  
27 887 using the Trinity platform for reference generation and analysis. *Nat Protoc.* 8:1494-  
28 888 1512.
- 29 889 Han E, Sinsheimer JS, Novembre J. 2015. Fast and accurate site frequency spectrum estimation  
30 890 from low coverage sequence data. *Bioinformatics.* 31:720-727.
- 31 891 Hancock-Hanser BL, Frey A, Leslie MS, Dutton PH, Archer FI, Morin PA. 2013. Targeted  
32 892 multiplex next-generation sequencing: advances in techniques of mitochondrial and  
33 893 nuclear DNA sequencing for population genomics. *Mol Ecol Resour.* 13:254-268.
- 34 894 Harris K, Nielsen R. 2013. Inferring demographic history from a spectrum of shared haplotype  
35 895 lengths. *PLoS Genetics.* 9:e1003521.
- 36 896 Hayden S, Bekaert M, Crider TA, Mariani S, Murphy WJ, Teeling EC. 2010. Ecological  
37 897 adaptation determines functional mammalian olfactory subgenomes. *Genome Res.* 20:1-  
38 898 9.
- 39 899 Hedrick PW. 2000 *Genetics of Populations.* Jones and Bartlett Publishers, Sudbury, MA.
- 40 900 Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, Cariani A,  
41 901 Maes GE, Diopere E, Carvalho GR, *et al.* 2011. Application of SNPs for population  
42 902 genetics of nonmodel organisms: new opportunities and challenges. *Mol Ecol Resour.*  
43 903 11:123-136.
- 44 904 Higdon JW, Bininda-Emonds ORP, Beck RMD, Ferguson SH. 2007. Phylogeny and divergence  
45 905 of the pinnipeds (Carnivora: Mammalia) assessed using a multigene dataset. *BMC Evol*  
46 906 *Biol.* 7:216.
- 47 907 Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Gordon DB, Brizuela L, McCombie WR,  
48 908 Hannon GJ. 2009. Hybrid selection of discrete genomic intervals on custom-designed  
49 909 microarrays for massively parallel sequencing. *Nat Protoc.* 4:960-974.

- 1  
2  
3 910 Hoffman JI. 2011. Gene discovery in the Antarctic fur seal (*Arctocephalus gazella*) skin  
4 911 transcriptome. *Mol Ecol Resour.* 11:703-710.  
5 912 Hoffman JI, Nicholas HJ. 2011. A novel approach for mining polymorphic microsatellite  
6 913 markers *in silico*. *PLoS One.* 6:e23283.  
7 914 Hoffman JI, Simpson F, David P, Rijks JM, Kuiken T, Thorne MAS, Lacy RC, Dasmahapatra  
8 915 KK. 2014. High-throughput sequencing reveals inbreeding depression in a natural  
9 916 population. *Proc Natl Acad Sci USA.* 111:3775-3780.  
10 917 Hoffman JI, Thorne MAS, Trathan PN, Forcada J. 2013. Transcriptome of the dead:  
11 918 characterisation of immune genes and marker development from necropsy samples in a  
12 919 free-ranging marine mammal. *BMC Genomics.* 14:52.  
13 920 Hoffman JI, Tucker R, Bridgett SJ, Clark MS, Forcada J, Slate J. 2012. Rates of assay success  
14 921 and genotyping error when single nucleotide polymorphism genotyping in non-model  
15 922 organisms: a case study in the Antarctic fur seal. *Mol Ecol Resour.* 12:861-872.  
16 923 Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population  
17 924 genomics of parallel adaptation in threespine stickleback using sequenced RAD tags.  
18 925 *PLoS Genet.* 6:e1000862.  
19 926 Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management  
20 927 tool for second-generation genome projects. *BMC Bioinformatics.* 12:491.  
21 928 Humble E, Martinez-Barrio A, Forcada J, Trathan PN, Thorne MAS, Hoffmann M, Wolf JBW,  
22 929 Hoffman JI. 2016. A draft fur seal genome provides insights into factors affecting SNP  
23 930 validation and how to mitigate them. *Mol Ecol Resour.*  
24 931 Jackson JA, Baker CS, Vant M, Steel DJ, Medrano-González L, Palumbi SR. 2009. Big and  
25 932 slow: phylogenetic estimates of molecular evolution in baleen whales (suborder  
26 933 Mysticeti). *Mol Biol Evol.* 26:2427-2440.  
27 934 Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody  
28 935 MC, White S, *et al.* 2012. The genomic basis of adaptive evolution in threespine  
29 936 sticklebacks. *Nature.* 484:55-61.  
30 937 Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M,  
31 938 Nagayasu E, Maruyama H, *et al.* 2014. Efficient *de novo* assembly of highly  
32 939 heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24:1384-  
33 940 1395.  
34 941 Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, Madsen LB, van Dam S, Brawand  
35 942 D, Marques PI, *et al.* 2015. Insights into the evolution of longevity from the bowhead  
36 943 whale genome. *Cell Reports.* 10:112-122.  
37 944 Khudyakov JI, Champagne CD, Preeyanon L, Ortiz RM, Crocker DE. 2015a. Muscle  
38 945 transcriptome response to ACTH administration in a free-ranging marine mammal.  
39 946 *Physiol Genomics.* 47:318-330.  
40 947 Khudyakov JI, Preeyanon L, Champagne CD, Ortiz RM, Crocker DE. 2015b. Transcriptome  
41 948 analysis of northern elephant seal (*Mirounga angustirostris*) muscle tissue provides a  
42 949 novel molecular resource and physiological insights. *BMC Genomics.* 16:64.  
43 950 Kishida T, Thewissen JGM, Hayakawa T, Imai H, Agata K. 2015. Aquatic adaptation and the  
44 951 evolution of smell and taste in whales. *Zoolog Lett.* 1:9.  
45 952 Koepfli K-P, Paten B, Genome 10K Community of Scientists, O'Brien SJ. 2015. The Genome  
46 953 10K Project: a way forward. *Annu Rev Anim Biosci.* 3:57-111.  
47 954 Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation  
48 955 Sequencing Data. *BMC Bioinformatics.* 15:356.  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 956 K $\ddot{u}$ stner A, Wolf JBW, Backstr $\ddot{o}$ m N, Whitney O, Balakrishnan CN, Day L, Edwards SV, Janes  
4 957 DE, Schlinger BA, Wilson RK, *et al.* 2010. Comparative genomics based on massive  
5 958 parallel transcriptome sequencing reveals patterns of substitution and selection across 10  
6 959 bird species. *Mol Ecol.* 19:266-276.
- 8 960 Lamichhaney S, Berglund J, Alm $\acute{e}$ n MS, Maqbool K, Grabherr M, Martinez-Barrio A,  
9 961 Promerov $\acute{a}$  M, Rubin C-J, Wang C, Zamani N, *et al.* 2015. Evolution of Darwin's finches  
10 962 and their beaks revealed by genome sequencing. *Nature.* 518:371-375.
- 12 963 Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment  
13 964 of short DNA sequences to the human genome. *Genome Biol.* 10:R25.
- 14 965 Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-  
15 966 throughput phylogenomics. *Syst Biol.* 61:727-744.
- 16 967 Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or  
17 968 without a reference genome. *BMC Bioinformatics.* 12:323.
- 19 969 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.  
20 970 *Bioinformatics.* 25:1754-1760.
- 21 971 Li H, Durbin R. 2011. Inference of human population history from individual whole-genome  
22 972 sequences. *Nature.* 475:493-496.
- 23 973 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,  
24 974 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map  
25 975 form and SAMtools. *Bioinformatics.* 25:2078-2079.
- 27 976 Li S, Jakobsson M. 2012. Estimating demographic parameters from large-scale population  
28 977 genomic data using Approximate Bayesian Computation. *BMC Genet.* 13:22.
- 29 978 Li Y, Hu Y, Bolund L, Wang J. 2010. State of the art *de novo* assembly of human genomes from  
30 979 massively parallel sequencing data. *Human Genomics* 4:271-277.
- 32 980 Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J,  
33 981 Jordan G, Mauceli E, *et al.* 2011. A high-resolution map of human evolutionary  
34 982 constraint using 29 mammals. *Nature.* 478:476-482.
- 35 983 Lindqvist C, Schuster SC, San Y, Talbot SL, Qi J, Ratan A, Tomsho LP, Kasson L, Zeyl E, Aars  
36 984 J, *et al.* 2010. Complete mitochondrial genome of a Pleistocene jawbone unveils the  
37 985 origin of polar bear. *Proc Natl Acad Sci USA.* 107:5053-5057.
- 39 986 Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, Xiong Z, Zhou L, Korneliussen TS, Somel M,  
40 987 Babbitt C, *et al.* 2014a. Population genomics reveal recent speciation and rapid  
41 988 evolutionary adaptation in polar bears. *Cell.* 157:785-794.
- 42 989 Liu X, Fun Y-X. 2015. Exploring population size changes using SNP frequency spectra. *Nat*  
43 990 *Genet.* 47:555-559.
- 44 991 Liu Y, Zhou J, White KP. 2014b. RNA-seq differential expression studies: more sequence or  
45 992 more replication? *Bioinformatics.* 30:301-304.
- 47 993 Lotterhos KE, Whitlock MC. 2014. Evaluation of demographic history and neutral  
48 994 parameterization on the performance of  $F_{ST}$  outlier tests. *Mol Ecol.* 23:2178-2192.
- 49 995 Louis M, Viricel A, Lucas T, Peltier H, Alfonsi E, Berrow S, Brownlow A, Covelo P, Dabin W,  
50 996 Deaville R, *et al.* 2014. Habitat-driven population structure of bottlenose dolphins,  
51 997 *Tursiops truncatus*, in the North-east Atlantic. *Mol Ecol.* 23:857-874.
- 53 998 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for  
54 999 RNA-seq data with DESeq2. *Genome Biol.* 15:550.
- 55 1000 MacManes MD. 2014. On the optimal trimming of high-throughput mRNA sequence data. *Front*  
56 1001 *Genet.* 5:13.
- 57  
58  
59  
60



- 1  
2  
3 1002 MacManes MD. 2016. Establishing evidence-based best practice for the *de novo* assembly and  
4 1003 evaluation of transcriptomes from non-model organisms. *bioRxiv*. doi:  
5 1004 <http://dx.doi.org/10.1101/035642>.  
6  
7 1005 Magera AM, Mills Flemming JE, Kaschner K, Christensen LB, Lotze HK. 2013. Recovery  
8 1006 trends in marine mammal populations. *PLoS One*. 8:e77908.  
9 1007 Malenfant RM, Coltman DW, Davis CS. 2015. Design of a 9K Illumina BeadChip for polar  
10 1008 bears (*Ursus maritimus*) from RAD and transcriptome sequencing. *Mol Ecol Resour*.  
11 1009 15:587-600.  
12  
13 1010 Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J,  
14 1011 Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. *Nat*  
15 1012 *Methods*. 7:111-118.  
16  
17 1013 Mancía A, Abelli L, Kucklick JR, Rowles TK, Wells RS, Balmer BC, Hohn AA, Baatz JE, Ryan  
18 1014 JC. 2015. Microarray applications to understand the impact of exposure to environmental  
19 1015 contaminants in wild dolphins (*Tursiops truncatus*). *Mar Genomics*. 19:47-57.  
20 1016 Mancía A, Lundqvist ML, Romano TA, Peden-Adams MM, Fair PA, Kindy MS, Ellis BC,  
21 1017 Gattoni-Celli S, McKillen DJ, Trent HF, *et al.* 2007. A dolphin peripheral blood  
22 1018 leukocyte cDNA microarray for studies of immune function and stress reactions. *Dev*  
23 1019 *Comp Immunol*. 31:520-529.  
24  
25 1020 Mancía A, Ryan JC, Chapman RW, Wu Q, Warr GW, Gulland FMD, Van Dolah FM. 2012.  
26 1021 Health status, infection and disease in California sea lions (*Zalophus californianus*)  
27 1022 studied using a canine microarray platform and machine-learning approaches. *Dev Comp*  
28 1023 *Immunol*. 36:629-637.  
29  
30 1024 Mancía A, Warr GW, Chapman RW. 2008. A transcriptomic analysis of the stress induced by  
31 1025 capture-release health assessment studies in wild dolphins (*Tursiops truncatus*). *Mol*  
32 1026 *Ecol*. 17:2581-2589.  
33  
34 1027 Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC. 2015.  
35 1028 Restriction site-associated DNA sequencing, genotyping error estimation and *de novo*  
36 1029 assembly optimization for population genetic inference. *Mol Ecol Resour*. 15:28-41.  
37  
38 1030 McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. 2012.  
39 1031 Ultraconserved elements are novel phylogenomic markers that resolve placental mammal  
40 1032 phylogeny when combined with species-tree analysis. *Genome Res*. 22:746-754.  
41 1033 McGowen MR. 2011. Toward the resolution of an explosive radiation - a multilocus phylogeny  
42 1034 of oceanic dolphins (Delphinidae). *Mol Phylogenet Evol*. 60:345-357.  
43 1035 McGowen MR, Clark C, Gatesy J. 2008. The vestigial olfactory receptor subgenome of  
44 1036 odontocete whales: phylogenetic congruence between gene-tree reconciliation and  
45 1037 supermatrix methods. *Syst Biol*. 57:574-590.  
46 1038 McGowen MR, Gatesy J, Wildman DE. 2014. Molecular evolution tracks macroevolutionary  
47 1039 transitions in Cetacea. *Trends Ecol Evol*. 29:336-346.  
48 1040 McGowen MR, Grossman LI, Wildman DE. 2012. Dolphin genome provides evidence for  
49 1041 adaptive evolution of nervous system genes and a molecular rate slowdown. *Proc R Soc*  
50 1042 *Lond B Biol Sci*. 279:3643-3651.  
51  
52 1043 McGowen MR, Spaulding M, Gatesy J. 2009. Divergence date estimation and a comprehensive  
53 1044 molecular tree of extant cetaceans. *Mol Phylogenet Evol*. 53:891-906.  
54 1045 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,  
55 1046 Altshuler D, Gabriel S, Daly M, *et al.* 2010. The Genome Analysis Toolkit: A  
56  
57  
58  
59  
60

- 1  
2  
3 1047 MapReduce framework for analyzing next-generation DNA sequencing data. *Genome*  
4 1048 *Res.* 20:1297-1303.
- 5 1049 McTavish EJ, Hillis DM. 2015. How do SNP ascertainment schemes and population  
6 1050 demographics affect inferences about population history? *BMC Genomics.* 16:266.
- 7 1051 Meredith RW, Gatesy J, Emerling CA, York VM, Springer MS. 2013. Rod monochromacy and  
8 1052 the coevolution of cetacean retinal opsins. *PLoS Genetics.* 9:e1003432.
- 9 1053 Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K,  
10 1054 de Filippo C, *et al.* 2012. A high-coverage genome sequence from an archaic Denisovan  
11 1055 individual. *Science.* 338:222-226.
- 12 1056 Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. 2007. Rapid and cost-effective  
13 1057 polymorphism identification and genotyping using restriction site associated DNA  
14 1058 (RAD) markers. *Genome Res.* 17:240-248.
- 15 1059 Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, Zhao F, Kim HL, Burhans RC,  
16 1060 Drautz DI, Wittekindt NE, *et al.* 2012. Polar and brown bear genomes reveal ancient  
17 1061 admixture and demographic footprints of past climate change. *Proc Natl Acad Sci USA.*  
18 1062 109:E2382-E2390.
- 19 1063 Mirceta S, Signore AV, Burns JM, Cossins AR, Campbell KL, Berenbrink M. 2013. Evolution  
20 1064 of mammalian diving capacity traced by myoglobin net surface charge. *Science.*  
21 1065 340:1234192.
- 22 1066 Morin PA, Archer FI, Foote AD, Vilstrup J, Allen EE, Wade P, Durban JW, Parsons K, Pitman  
23 1067 R, Li L, *et al.* 2010a. Complete mitochondrial genome phylogeographic analysis of killer  
24 1068 whales (*Orcinus orca*) indicates multiple species. *Genome Res.* 20:908-916.
- 25 1069 Morin PA, Luikart G, Wayne RK, SNP workshop group. 2004. SNPs in ecology, evolution and  
26 1070 conservation. *Trends Ecol Evol.* 19:208-216.
- 27 1071 Morin PA, Martien KK, Archer FI, Cipriano F, Steel D, Jackson J, Taylor BL. 2010b. Applied  
28 1072 conservation genetics and the need for quality control and reporting of genetic data used  
29 1073 in fisheries and wildlife management. *J Hered.* 101:1-10.
- 30 1074 Morin PA, Parsons KM, Archer FI, Ávila-Arcos M, Barrett-Lennard LG, Dalla Rosa L, Duchêne  
31 1075 S, Durban JW, Ellis GM, Ferguson SH, *et al.* 2015. Geographic and temporal dynamics  
32 1076 of a global radiation and diversification in the killer whale. *Mol Ecol.* 24:3964-3979.
- 33 1077 Moura AE, Kenny JG, Chaudhuri R, Hughes MA, Welch AJ, Reisinger RR, de Bruyn PJN,  
34 1078 Dahlheim ME, Hall N, Hoelzel AR. 2014a. Population genomics of the killer whale  
35 1079 indicates ecotype evolution in sympatry involving both selection and drift. *Mol Ecol.*  
36 1080 23:5179-5192.
- 37 1081 Moura AE, Nielsen SCA, Vilstrup JT, Moreno-Mayar JV, Gilbert MTP, Gray HWI, Natoli A,  
38 1082 Möller L, Hoelzel AR. 2013. Recent diversification of a marine genus (*Tursiops* spp.)  
39 1083 tracks habitat preference and environmental change. *Syst Biol.* 62:865-877.
- 40 1084 Moura AE, van Rensburg CJ, Pilot M, Tehrani A, Best PB, Thornton M, Plön S, de Bruyn PJN,  
41 1085 Worley KC, Gibbs RA, *et al.* 2014b. Killer whale nuclear genome and mtDNA reveal  
42 1086 widespread population bottleneck during the last glacial maximum. *Mol Biol Evol.*  
43 1087 31:1121-1131.
- 44 1088 Nadeau NJ, Ruiz M, Salazar P, Counterman B, Alejandro Medina J, Ortiz-Zuazaga H, Morrison  
45 1089 A, McMillan WO, Jiggins CD, Papa R. 2014. Population genomics of parallel hybrid  
46 1090 zones in the mimetic butterflies, *H. melpomene* and *H. erato*. *Genome Res.* 24:1316-  
47 1091 1333.

- 1  
2  
3 1092 Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA. 2013. Genotyping-by-  
4 1093 sequencing in ecological and conservation genomics. *Mol Ecol.* 22:2841-2847.  
5 1094 Narum SR, Hess JE. 2011. Comparison of  $F_{ST}$  outlier tests for SNP loci under selection. *Mol*  
6 1095 *Ecol Resour.* 11:184-194.  
7 1096 Nelson TM, Apprill A, Mann J, Rogers TL, Brown MV. 2015. The marine mammal microbiome:  
8 1097 current knowledge and future directions. *Microbiology Australia.* 36:8-13.  
9 1098 Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M,  
10 1099 Bhattacharjee A, Eichler EE, *et al.* 2009. Targeted capture and massively parallel  
11 1100 sequencing of twelve human exomes. *Nature.* 461:272-276.  
12 1101 Nielsen R, Paul JS, Anders A, Song YS. 2011. Genotype and SNP calling from next-generation  
13 1102 sequencing data. *Nat Rev Genet.* 12:433-451.  
14 1103 Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S,  
15 1104 Pritchard JK, *et al.* 2006. Sequencing and analysis of Neanderthal genomic DNA.  
16 1105 *Science.* 314:1113-1118.  
17 1106 Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD,  
18 1107 Chin G, Christensen G, *et al.* 2015. Promoting an open research culture: Author  
19 1108 guidelines for journals could help to promote transparency, openness, and reproducibility.  
20 1109 *Science.* 348:1422-1425.  
21 1110 O'Rawe JA, Ferson S, Lyon GJ. 2015. Accounting for uncertainty in DNA sequencing data.  
22 1111 *Trends Genet.* 31:61-66.  
23 1112 Olsen MT, Volny VH, Bérubé M, Dietz R, Lydersen C, Kovacs KM, Dodd RS, Palsbøll PJ.  
24 1113 2011. A simple route to single-nucleotide polymorphisms in a nonmodel species:  
25 1114 identification and characterization of SNPs in the Arctic ringed seal (*Pusa hispida*  
26 1115 *hispida*). *Mol Ecol Resour.* 11:9-19.  
27 1116 Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E,  
28 1117 Petersen B, Moltke I, *et al.* 2013. Recalibrating *Equus* evolution using the genome  
29 1118 sequence of an early Middle Pleistocene horse. *Nature.* 499:74-78.  
30 1119 Pabuwal V, Boswell M, Pasquali A, Wise SS, Kumar S, Shen Y, Garcia T, Lacerte C, Wise JP,  
31 1120 Jr., Wise JP, Sr., *et al.* 2013. Transcriptomic analysis of cultured whale skin cells exposed  
32 1121 to hexavalent chromium [Cr(VI)]. *Aquat Toxicol.* 134-135:74-81.  
33 1122 Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-  
34 1123 wide signatures of convergent evolution in echolocating mammals. *Nature.* 502:228-231.  
35 1124 Paszkiewicz KH, Farbox A, O'Neill P, Moore K. 2014. Quality control on the frontier. *Front*  
36 1125 *Genet.* 5:157.  
37 1126 Patro R, Duggal G, Kingsford C. 2015. Accurate, fast, and model-aware transcript expression  
38 1127 quantification with Salmon. *bioRxiv.* doi: <http://dx.doi.org/10.1101/021592>.  
39 1128 Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: an  
40 1129 inexpensive method for *de novo* SNP discovery and genotyping in model and non-model  
41 1130 species. *PLoS One.* 7:e37135.  
42 1131 Poh Y-P, Domingues VS, Hoekstra HE, Jensen JD. 2014. On the prospect of identifying adaptive  
43 1132 loci in recently bottlenecked populations. *PLoS One.* 9:e110579.  
44 1133 Poland JA, Brown PJ, Sorrells ME, Jannink J-L. 2012. Development of high-density genetic  
45 1134 maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing  
46 1135 approach. *PLoS One.* 7:e32253.  
47 1136 Polanowski AM, Robbins J, Chandler D, Jarman SN. 2014. Epigenetic estimation of age in  
48 1137 humpback whales. *Mol Ecol Resour.* 14:976-987.

- 1  
2  
3 1138 Puritz JB, Hollenbeck CM, Gold JR. 2014. *dDocent*: a RADseq, variant-calling pipeline  
4 1139 designed for population genomics of non-model organisms. *PeerJ*. 2:e431.  
5 1140 Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu  
6 1141 E, Kivisild T, Gupta R, *et al.* 2010. Ancient human genome sequence of an extinct  
7 1142 Palaeo-Eskimo. *Nature*. 463:757-762.  
8 1143 Riesch R, Barrett-Lennard LG, Ellis GM, Ford JKB, Deecke VB. 2012. Cultural traditions and  
9 1144 the evolution of reproductive isolation: ecological speciation in killer whales? *Biol J Linn*  
10 1145 *Soc Lond*. 2012:1-17.  
11 1146 Robinson JD, Coffman AJ, Hickerson MJ, Gutenkunst RN. 2014. Sampling strategies for  
12 1147 frequency spectrum-based population genomic inference. *BMC Evol Biol*. 14:254.  
13 1148 Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential  
14 1149 expression analysis of digital gene expression data. *Bioinformatics*. 26:139-140.  
15 1150 Ruan R, Guo A-H, Hao Y-J, Zheng J-S, Wang D. 2015. *De novo* assembly and characterization  
16 1151 of narrow-ridged finless porpoise renal transcriptome and identification of candidate  
17 1152 genes involved in osmoregulation. *Int J Mol Sci*. 16:2220-2238.  
18 1153 Ruegg K, Rosenbaum HC, Anderson EC, Engel M, Rothschild A, Baker CS, Palumbi SR. 2013.  
19 1154 Long-term population size of the North Atlantic humpback whale within the context of  
20 1155 worldwide population structure. *Cons Gen*. 14:103-114.  
21 1156 Schiffels S, Durbin R. 2014. Inferring human population size and separation history from  
22 1157 multiple genome sequences. *Nat Genet*. 46:919-925.  
23 1158 Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming,  
24 1159 identification, and read merging. *BMC Res Notes*. 9:88.  
25 1160 Schurch NJ, Schofield P, Gierlinski M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K,  
26 1161 Simpson GG, Owen-Hughes T, *et al.* 2015. Evaluation of tools for differential gene  
27 1162 expression analysis by RNA-seq on a 48 biological replicate experiment. *arXive*.  
28 1163 1505:02017.  
29 1164 Seim I, Ma S, Zhou X, Gerashchenko MV, Lee SG, Suydam R, George JC, Bickham JW,  
30 1165 Gladyshev VN. 2014. The transcriptome of the bowhead whale *Balaena mysticetus*  
31 1166 reveals adaptations of the longest-lived mammal. *Aging*. 6:879-899.  
32 1167 Shafer ABA, Cullingham CI, Côté SD, Coltman DW. 2010. Of glaciers and refugia: a decade of  
33 1168 study sheds new light on the phylogeographic patterns of northwestern North America.  
34 1169 *Mol Ecol*. 19:4589-4621.  
35 1170 Shafer ABA, Davis CS, Coltman DW, Stewart REA. 2014. Microsatellite assessment of walrus  
36 1171 (*Odobenus rosmarus rosmarus*) stocks in Canada. *NAMMCO Scientific Publications*. 9.  
37 1172 Shafer ABA, Gattepaille LM, Stewart REA, Wolf JBW. 2015. Demographic inferences using  
38 1173 short-read genomic data in an approximate Bayesian computation framework: *in silico*  
39 1174 evaluation of power, biases and proof of concept in Atlantic walrus. *Mol Ecol*. 24:328-  
40 1175 345.  
41 1176 Shen Y-Y, Zhou W-P, Zhou T-C, Zeng Y-N, Li G-M, Irwin DM, Zhang Y-P. 2012. Genome-  
42 1177 wide scan for bats and dolphin to detect their genetic basis for new locomotive styles.  
43 1178 *PLoS One*. 7:e46455.  
44 1179 Smith-Unna RD, Boursnell C, Patro R, Hibberd JM, Kelly S. 2015. TransRate: reference free  
45 1180 quality assessment of *de-novo* transcriptome assemblies. *bioRxiv*.  
46 1181 Spies D, Ciaudo C. 2015. Dynamics in transcriptomics: advancements in RNA-seq time course  
47 1182 and downstream analysis. *Comput Struct Biotechnol J*. 13:469-477.  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 1183 Springer MS, Signore AV, Paijmans JLA, Vélez-Juarbe J, Domning DP, Bauer CE, He K, Crerar  
4 1184 L, Campos PF, Murphy WJ, *et al.* 2015. Interordinal gene capture, the phylogenetic  
5 1185 position of Steller's sea cow based on molecular and morphological data, and the  
6 1186 macroevolutionary history of Sirenia. *Mol Phylogenet Evol.* 91:178-193.
- 7 1187 Springer MS, Starrett J, Morin PA, Lanzetti A, Hayashi C, Gatesy J. 2016. Inactivation of  
8 1188 *C4orf26* in toothless placental mammals. *Mol Phylogenet Evol.* 95:34-45.
- 9 1189 Sremba AL, Martin AR, Baker CS. 2015. Species identification and likely catch time period of  
10 1190 whale bones from South Georgia. *Mar Mamm Sci.* 31:122-132.
- 11 1191 Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: *ab initio*  
12 1192 prediction of alternative transcripts. *Nucleic Acids Res.* 34:W435-W439.
- 13 1193 Stein LD. 2010. The case for cloud computing in genome informatics. *Genome Biol.* 11:207.
- 14 1194 Stinchcombe JR, Hoekstra HE. 2008. Combining population genomics and quantitative genetics:  
15 1195 finding genes underlying ecologically important traits. *Heredity.* 100:158-170.
- 16 1196 Tabuchi M, Veldhoen N, Dangerfield N, Jeffries S, Helbing CC, Ross PS. 2006. PCB-related  
17 1197 alteration of thyroid hormones and thyroid hormone receptor gene expression in free-  
18 1198 ranging harbor seals (*Phoca vitulina*). *Environ Health Perspect.* 114:1024-1031.
- 19 1199 Taylor BL, Gemmell NJ. 2016. Emerging technologies to conserve biodiversity: further  
20 1200 opportunities via genomics. Response to Pimm *et al.* *Trends Ecol Evol.* 31:171-172.
- 21 1201 The *Heliconius* Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of  
22 1202 mimicry adaptations among species. *Nature.* 487:94-98.
- 23 1203 Thomsen PF, Kielgast J, Iversen LL, Møller PR, Rasmussen M, Willerslev E. 2012. Detection of  
24 1204 a diverse marine fish fauna using environmental DNA from seawater samples. *PLoS One.*  
25 1205 7:e41732.
- 26 1206 Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, Hazlewood V, Lathrop S,  
27 1207 Lifka D, Peterson GD, *et al.* 2014. XSEDE: accelerating scientific discovery. *Computing*  
28 1208 *in Science and Engineering.* 16:62-74.
- 29 1209 Tsagkogeorga G, McGowen MR, Davies KT, Jarman S, Polanowski A, Bertelsen MF, Rossiter  
30 1210 SJ. 2015. A phylogenomic analysis of the role and timing of molecular adaptation in the  
31 1211 aquatic transition of cetartiodactyl mammals. *R Soc Open Sci.* 2:150156.
- 32 1212 van Dijk EL, Auger H, Jaszczyn Y, Thermes C. 2014. Ten years of next-generation  
33 1213 sequencing technology. *Trends Genet.* 30:418-426.
- 34 1214 VanRaden PM, Sun C, O'Connell JR. 2015. Fast imputation using medium or low-coverage  
35 1215 sequence data. *BMC Genet.* 16:82.
- 36 1216 Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R,  
37 1217 Erichsen JT, Jasinska AJ, *et al.* 2015. Enhancer evolution across 20 mammalian species.  
38 1218 *Cell.* 160:554-566.
- 39 1219 Viricel A, Pante E, Dabin W, Simon-Bouhet B. 2014. Applicability of RAD-tag genotyping for  
40 1220 interfamilial comparisons: empirical data from two cetaceans. *Mol Ecol Resour.* 14:597-  
41 1221 605.
- 42 1222 Viricel A, Rosel PE. 2014. Hierarchical population structure and habitat differences in a highly  
43 1223 mobile marine species: the Atlantic spotted dolphin. *Mol Ecol.* 23:5018-5035.
- 44 1224 Wolf JB. 2013. Principles of transcriptome analysis and gene expression quantification: an RNA-  
45 1225 seq tutorial. *Mol Ecol Resour.* 13:559-572.
- 46 1226 Xiong Y, Brandley MC, Xu S, Zhou K, Yang G. 2009. Seven new dolphin mitochondrial  
47 1227 genomes and a time-calibrated phylogeny of whales. *BMC Evol Biol.* 9:20.

- 1  
2  
3 1228 Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet.*  
4 1229 13:329-342.  
5 1230 Yeh R-F, Lim LP, Burge CB. 2001. Computational inference of homologous gene structures in  
6 1231 the human genome. *Genome Res.* 11:803-816.  
7 1232 Yim H-S, Cho YS, Guang X, Kang SG, Jeong J-Y, Cha S-S, Oh H-M, Lee J-H, Yang EC, Kwon  
8 1233 KK, *et al.* 2014. Minke whale genome and aquatic adaptation in cetaceans. *Nat Genet.*  
9 1234 46:88-92.  
10 1235 Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X, Hao P. 2011. Optimizing *de novo* transcriptome  
11 1236 assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics.*  
12 1237 12:S2.  
13 1238 Zhou X, Sun F, Xu S, Fan G, Zhu K, Liu X, Chen Y, Shi C, Yang Y, Huang Z, *et al.* 2013. Baiji  
14 1239 genomes reveal low genetic variability and new insights into secondary aquatic  
15 1240 adaptations. *Nat Commun.* 4:2708.  
16 1241 Zou Z, Zhang J. 2015. No genome-wide protein sequence convergence for echolocation. *Mol*  
17 1242 *Biol Evol.* 32:1237-1241.  
18 1243

1244 Table 1. Current and commonly used tools for analysis of genomic data generated in non-model organisms. Please note that this list is  
 1245 not exhaustive and new computational tools are continuously being developed.  
 1246

Computational Tool	Purpose	Strengths/Weaknesses	Reference
<i>RADseq*</i>			
STACKS	quality filtering, <i>de novo</i> assembly or reference-aligned read mapping, variant genotyping	scalable (new data can be compared against existing locus catalog); flexible filtering and export options; recently implemented a gapped alignment algorithm to process insertion-deletion (indel) mutations; secondary algorithm adjusts SNP calls using population-level allele frequencies; compatible with input data from multiple RADseq methods	Catchen et al. (2011; 2013), <a href="http://catchenlab.life.illinois.edu/stacks/">http://catchenlab.life.illinois.edu/stacks/</a>
PyRAD	quality filtering, <i>de novo</i> assembly, read mapping, variant genotyping	efficiently processes indel mutations, thus optimal for analysis of highly divergent species; high speed and quality of paired-end library assemblies; compatible with input data from multiple RADseq methods	Eaton (2014)
TASSEL-GBS	quality filtering, reference-aligned read mapping, variant genotyping	optimized for single-end data from large sample sizes (tens of thousands of individuals) with a reference genome; performs genome-wide association studies	Glaubitz et al. (2014)
dDocent	quality trimming, <i>de novo</i> assembly, read mapping, variant genotyping	beneficial in analysis of paired-end data; identifies both SNP and indel variants; most appropriate for ezRAD and ddRAD data	Puritz et al. (2014)
AfrRAD	quality filtering, <i>de novo</i> assembly, read mapping, variant genotyping	identifies both SNP and indel variants; computationally faster than STACKS and PyRAD	Sovic et al. (2015)
<i>Array-based high-throughput sequencing</i>			
Affymetrix Axiom™ Analysis Suite, Illumina® GenomeStudio	genotype scoring	visualization of genotype clusters; quality scores assigned to genotype calls allow user-specific filtering; manual editing possible	
<i>Whole genome sequencing</i>			
AdapterRemoval v2, Trimmomatic	trim raw sequences	remove adapter sequences and low-quality bases prior to assembly	Bolger et al. (2014), Schubert et al. (2016)
ALLPATHS-LG, PLATANUS, SOAPdenovo	<i>de novo</i> genome assembly	designed for short-read sequences of large heterozygous genomes	Li et al. (2010), Gnerre et al. (2011), Kajitani et al. (2014)
AUGUSTUS, GenomeScan, MAKER2	gene annotation	highly accurate evidence-driven or BLASTX-guided gene prediction (Yandell and Ence 2012)	Yeh et al. (2001), Stanke et al. (2006), Holt and Yandell (2011)

1				
2				
3	Bowtie, bwa	read mapping	rapid short-read alignment with compressed reference genome index, but limited number of acceptable mismatches per alignment (Flicek and Birney 2009)	Langmead et al. (2009), Li and Durbin (2009)
4				
5				
6	SAMtools	data processing, variant calling	multi-purpose tool that conducts file conversion, alignment sorting, PCR duplicate removal, and variant (SNP and indel) calling for SAM/BAM/CRAM files	Li et al. (2009)
7				
8				
9				
10	GATK	data processing and quality control, variant calling	suitable for data with low to high mean read depth across the genome; initially optimized for large human datasets, then modified for use with non-model organisms	McKenna et al. (2010), DePristo et al. (2011)
11				
12				
13				
14	ANGSD/NGStools	data processing, variant calling, estimation of diversity metrics, population genomic analyses	suitable for data with low mean read depth, including palaeogenomic data; allow downstream analyses such as D-statistics and SFS estimation	Fumagalli et al. (2014), Korneliusen et al. (2014)
15				
16				
17				
18	<i>RNAseq</i>			
19	Fastx Toolkit, Trimmomatic	trim raw sequences	remove erroneous nucleotides from reads prior to assembly	MacManes (2014)
20				
21	khmer diginorm, Trinity normalization	<i>in silico</i> read normalization	reduce memory requirements for assembly, but can result in fragmented assemblies and collapse heterozygosity	Brown et al. (2012); Haas et al. (2013)
22				
23				
24	Trinity	<i>de novo</i> and genome-guided transcriptome assembly	accurate assembly across conditions, but requires long runtime if normalization is not used (Zhao et al. 2011)	Haas et al. (2013)
25				
26	bowtie, bowtie2, STAR	read alignment to genome or transcriptome assembly	required for many downstream analyses, but bowtie is computationally intensive and all produce very large output BAM files	Langmead et al. (2009), Dobin et al. (2013)
27				
28				
29				
30	eXpress, kallisto, RSEM, Sailfish, Salmon	estimation of transcript abundance	RSEM requires computationally intensive read mapping back to the assembly; the others are faster streaming alignment, quasi-alignment, or alignment-free algorithms	Li and Dewey (2011), Patro et al. (2015)
31				
32				
33				
34	DESeq, DESeq2, edgeR	differential expression analysis	exhibit highest true positive and lowest false positive rates in experiments with smaller sample sizes (Schurch et al. 2015)	Anders and Huber (2010), Robinson et al. (2010), Love et al. (2014)
35				
36				
37				
38	blast2GO, Trinotate	functional annotation of assembled transcripts	complete annotation pipelines including gene ontology and pathway enrichment analyses	Conesa et al. (2005), Haas et al. (2013)
39				

1247 \* This is a non-exhaustive list of software that focuses on *de novo* loci assembly and genotype calling for RADseq data, as many practitioners working on NMOs  
 1248 will not have access to a reference genome. Other programs (e.g., GATK and ANGSD) that undertake genotype calling using reference-aligned loci are described  
 1249 in the whole genome sequencing section.



Figure 1

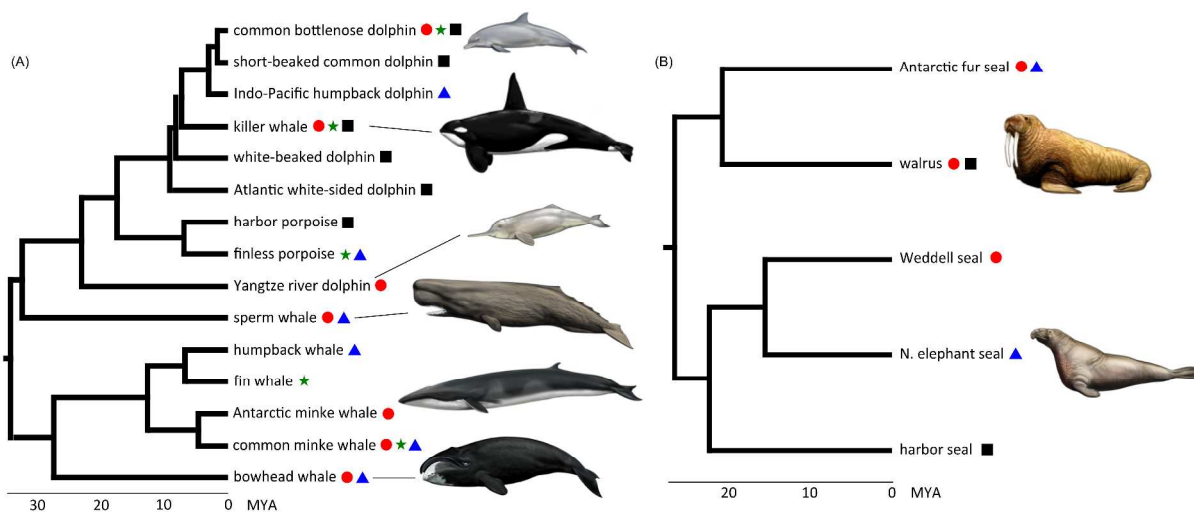


Figure 1. Phylogenetic tree showing current genomic resources available for (A) cetaceans and (B) pinnipeds; relationships and branch lengths are based on molecular dating estimates from McGowen et al. (2009), McGowen (2011), and Higdon et al. (2007). Scale is in millions of years ago (MYA). Red circles indicate species with high-quality reference genomes; green stars indicate whole genome re-sequencing data; blue triangles indicate transcriptomes (generated by microarray or RNAseq); and black squares indicate RADseq data.

Figure 2

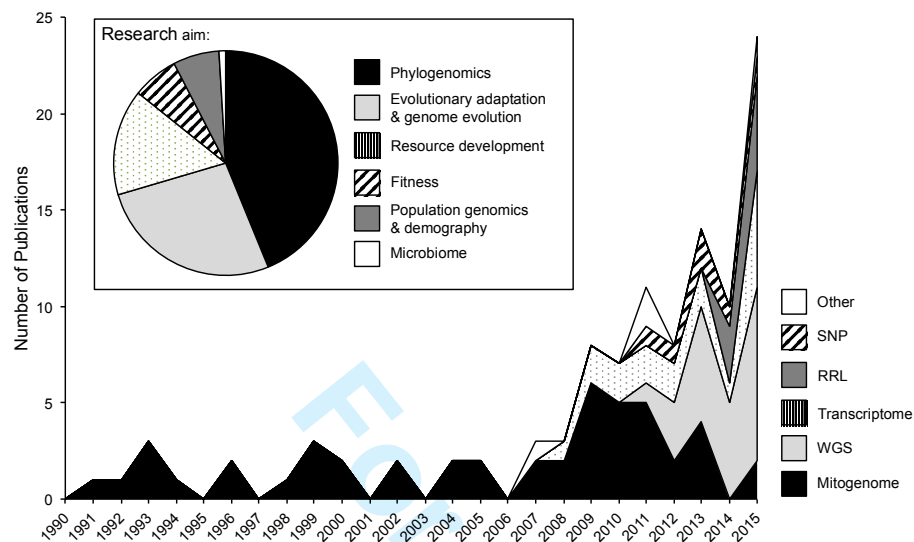


Figure 2. Number of marine mammal genomics publications from 1990 to 2015, categorized by primary methodology and research aim. Genomic methodologies include high-throughput single nucleotide polymorphism (SNP) genotyping and sequencing of mitogenomes, whole genomes (WGS), transcriptomes (generated by microarray or RNAseq), and reduced-representation genomic libraries (RRL). The “Other” category includes studies of microbiomes, BAC libraries, and large (~100) gene sets.

Figure 3

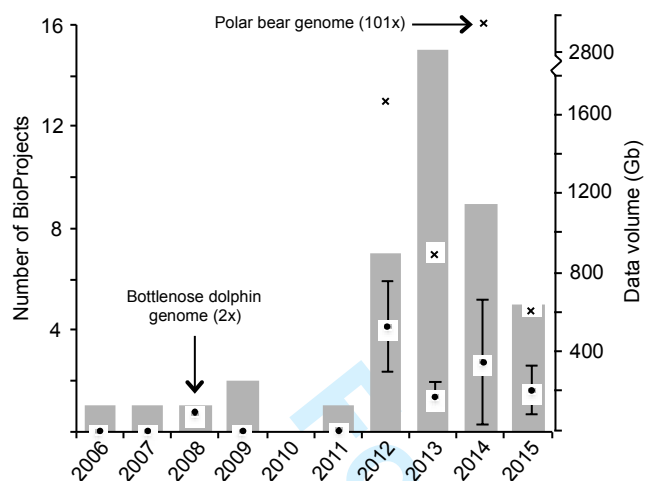


Figure 3. Number of BioProjects (gray bars) related to marine mammal genomics submitted from 2006 to 2015 to an online public database maintained by NCBI. Early BioProjects were largely microarray datasets. The number of projects created each year, as well as the yearly average (black dots  $\pm$  SE) and maximum ( $\times$ ) size of data submitted in each BioProject, increased dramatically after 2011, reflecting advances in high-throughput sequencing technologies that facilitated their use in non-model systems.

Figure 4

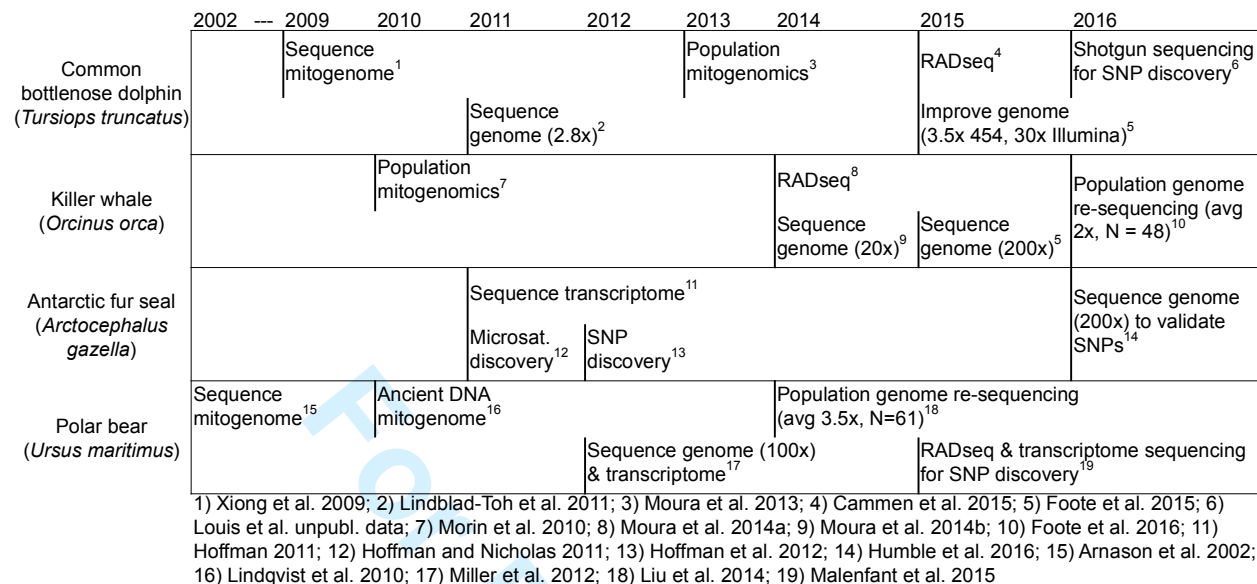
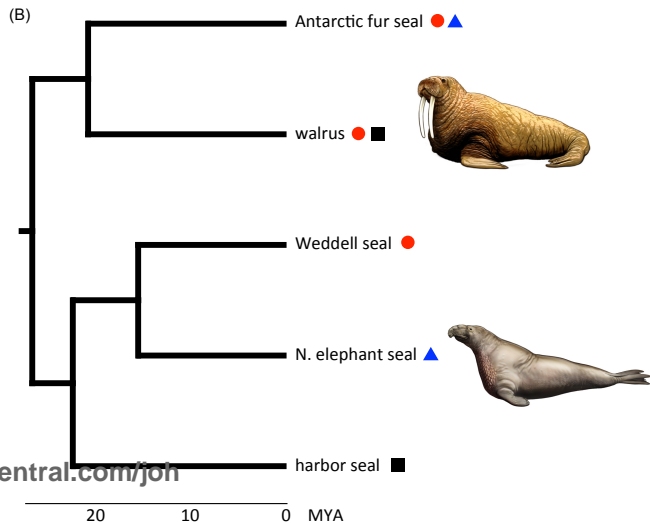
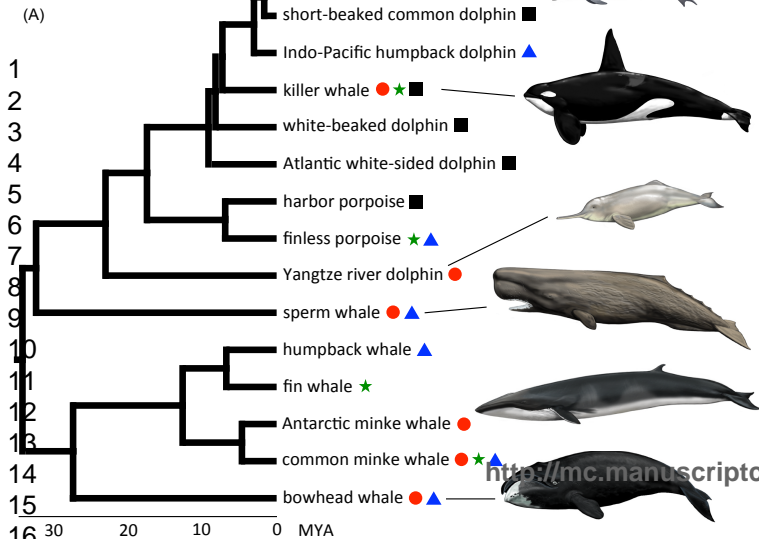
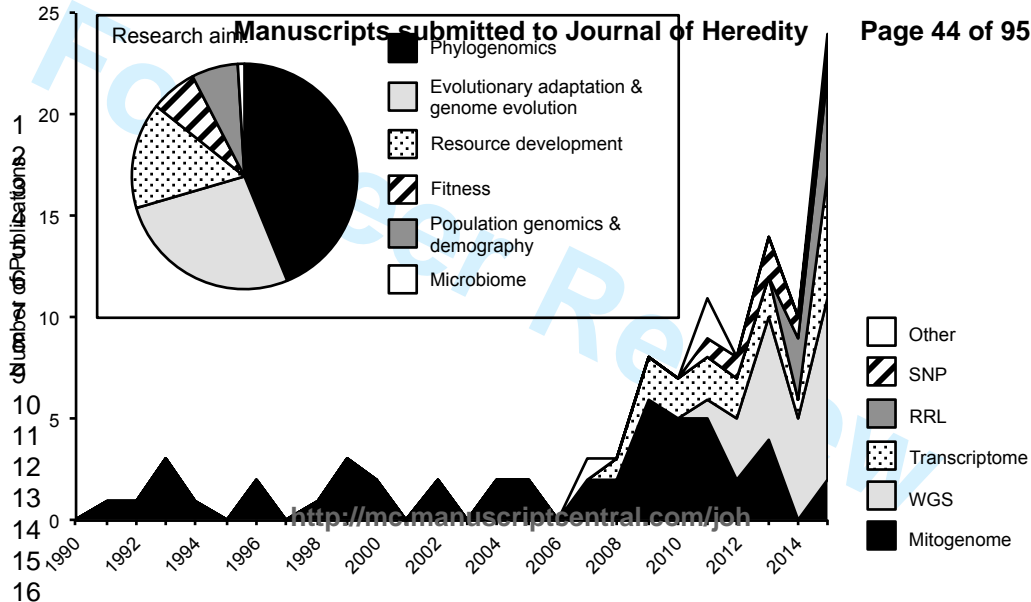


Figure 4. Timelines depicting the independent progression of genomic studies for four representative marine mammal species. Trajectories show the common progression for non-model species from mitogenome sequencing to whole genome sequencing, as well as from sequencing reference specimens to population-scale genomic sequencing. In addition, the timelines reveal the utility of genomic and transcriptomic sequencing for subsequent genetic marker development.



<http://mc.manuscriptcentral.com/joh>



Number of BioProjects

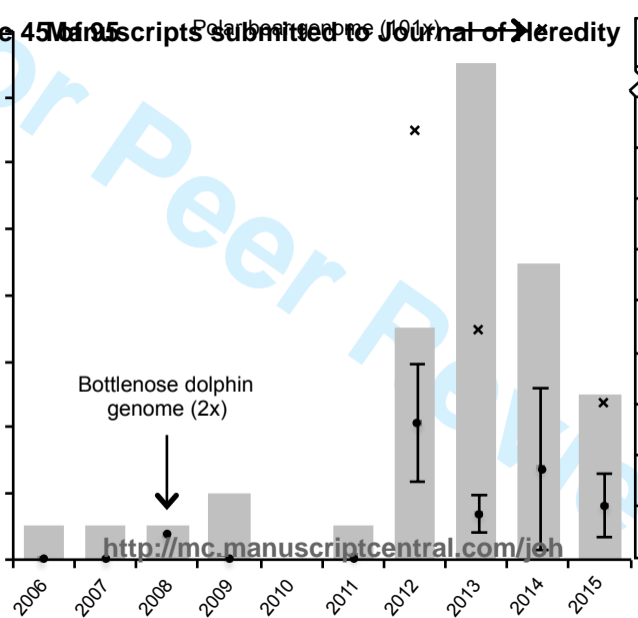
1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17

Bottlenose dolphin genome (2x)

<http://mc.manuscriptcentral.com/jeh>

Data volume (Gb)

3000  
1600  
1200  
800  
400



Page 46 of 95

	2002	2009	2010	2011	2012	2013	2014	2015	2016
Common bottlenose dolphin ( <i>Tursiops truncatus</i> )		Sequence mitogenome <sup>1</sup>		Sequence genome (2.8x) <sup>2</sup>		Population mitogenomics <sup>3</sup>		RADseq <sup>4</sup>	Shotgun sequencing for SNP discovery <sup>6</sup>
Killer whale ( <i>Orcinus orca</i> )			Population mitogenomics <sup>7</sup>				RADseq <sup>8</sup>	Sequence genome (20x) <sup>9</sup>	Sequence genome (200x) <sup>5</sup>
Antarctic fur seal ( <i>Arctocephalus gazella</i> )				Sequence transcriptome <sup>11</sup>	Microsat. discovery <sup>12</sup>	SNP discovery <sup>13</sup>			Population genome re-sequencing (avg 2x, N = 48) <sup>10</sup>
Polar bear ( <i>Ursus maritimus</i> )	Sequence mitogenome <sup>15</sup>		Ancient DNA mitogenome <sup>16</sup>				Population genome re-sequencing (avg 3.5x, N=61) <sup>18</sup>		Sequence genome (200x) to validate SNPs <sup>14</sup>
						Sequence genome (100x) & transcriptome <sup>17</sup>		RADseq & transcriptome sequencing for SNP discovery <sup>19</sup>	

1) Xiong et al. 2009; 2) Lindblad-Toh et al. 2011; 3) Moura et al. 2013; 4) Cammen et al. 2015; 5) Foote et al. 2015; 6) Louis et al. unpubl. data; 7) Monn et al. 2010; 8) Miquel et al. 2014; 9) Moura et al. 2014b; 10) Foote et al. 2016; 11) Hoffman 2011; 12) Hoffman and Nicholas 2011; 13) Hoffman et al. 2012; 14) Humble et al. 2016; 15) Arnason et al. 2002; 16) Lindqvist et al. 2010; 17) Miller et al. 2012; 18) Liu et al. 2014; 19) Malenfant et al. 2015



Table 1. Current and commonly used tools for analysis of genomic data generated in non-model organisms. Please note that this list is not exhaustive and new computational tools are continuously being developed.

Computational Tool	Purpose	Strengths/Weaknesses	Reference
<i>RADseq*</i>			
STACKS	quality filtering, <i>de novo</i> assembly or reference-aligned read mapping, variant genotyping	scalable (new data can be compared against existing locus catalog); flexible filtering and export options; recently implemented a gapped alignment algorithm to process insertion-deletion (indel) mutations; secondary algorithm adjusts SNP calls using population-level allele frequencies; compatible with input data from multiple RADseq methods	Catchen et al. (2011; 2013), <a href="http://catchenlab.life.illinois.edu/stacks/">http://catchenlab.life.illinois.edu/stacks/</a>
PyRAD	quality filtering, <i>de novo</i> assembly, read mapping, variant genotyping	efficiently processes indel mutations, thus optimal for analysis of highly divergent species; high speed and quality of paired-end library assemblies; compatible with input data from multiple RADseq methods	Eaton (2014)
TASSEL-GBS	quality filtering, reference-aligned read mapping, variant genotyping	optimized for single-end data from large sample sizes (tens of thousands of individuals) with a reference genome; performs genome-wide association studies	Glaubitz et al. (2014)
dDocent	quality trimming, <i>de novo</i> assembly, read mapping, variant genotyping	beneficial in analysis of paired-end data; identifies both SNP and indel variants; most appropriate for ezRAD and ddRAD data	Puritz et al. (2014)
AfrRAD	quality filtering, <i>de novo</i> assembly, read mapping, variant genotyping	identifies both SNP and indel variants; computationally faster than STACKS and PyRAD	Sovic et al. (2015)
<i>Array-based high-throughput sequencing</i>			
Affymetrix Axiom™ Analysis Suite, Illumina® GenomeStudio	genotype scoring	visualization of genotype clusters; quality scores assigned to genotype calls allow user-specific filtering; manual editing possible	
<i>Whole genome sequencing</i>			
AdapterRemoval v2, Trimmomatic	trim raw sequences	remove adapter sequences and low-quality bases prior to assembly	Bolger et al. (2014), Schubert et al. (2016)
ALLPATHS-LG, PLATANUS, SOAPdenovo	<i>de novo</i> genome assembly	designed for short-read sequences of large heterozygous genomes	Li et al. (2010), Gnerre et al. (2011), Kajitani et al. (2014)
AUGUSTUS, GenomeScan, MAKER2	gene annotation	highly accurate evidence-driven or BLASTX-guided gene prediction (Yandell and Ence 2012)	Yeh et al. (2001), Stanke et al. (2006), Holt and Yandell (2011)

1				
2				
3	Bowtie, bwa	read mapping	rapid short-read alignment with compressed reference genome index, but limited number of acceptable mismatches per alignment (Flicek and Birney 2009)	Langmead et al. (2009), Li and Durbin (2009)
4				
5				
6	SAMtools	data processing, variant calling	multi-purpose tool that conducts file conversion, alignment sorting, PCR duplicate removal, and variant (SNP and indel) calling for SAM/BAM/CRAM files	Li et al. (2009)
7				
8				
9				
10	GATK	data processing and quality control, variant calling	suitable for data with low to high mean read depth across the genome; initially optimized for large human datasets, then modified for use with non-model organisms	McKenna et al. (2010), DePristo et al. (2011)
11				
12				
13				
14	ANGSD/NGStools	data processing, variant calling, estimation of diversity metrics, population genomic analyses	suitable for data with low mean read depth, including palaeogenomic data; allow downstream analyses such as D-statistics and SFS estimation	Fumagalli et al. (2014), Korneliusen et al. (2014)
15				
16				
17				
18	<i>RNAseq</i>			
19	Fastx Toolkit, Trimmomatic	trim raw sequences	remove erroneous nucleotides from reads prior to assembly	MacManes (2014)
20				
21	khmer diginorm, Trinity normalization	<i>in silico</i> read normalization	reduce memory requirements for assembly, but can result in fragmented assemblies and collapse heterozygosity	Brown et al. (2012); Haas et al. (2013)
22				
23				
24	Trinity	<i>de novo</i> and genome-guided transcriptome assembly	accurate assembly across conditions, but requires long runtime if normalization is not used (Zhao et al. 2011)	Haas et al. (2013)
25				
26	bowtie, bowtie2, STAR	read alignment to genome or transcriptome assembly	required for many downstream analyses, but bowtie is computationally intensive and all produce very large output BAM files	Langmead et al. (2009), Dobin et al. (2013)
27				
28				
29				
30	eXpress, kallisto, RSEM, Sailfish, Salmon	estimation of transcript abundance	RSEM requires computationally intensive read mapping back to the assembly; the others are faster streaming alignment, quasi-alignment, or alignment-free algorithms	Li and Dewey (2011), Patro et al. (2015)
31				
32				
33				
34	DESeq, DESeq2, edgeR	differential expression analysis	exhibit highest true positive and lowest false positive rates in experiments with smaller sample sizes (Schurch et al. 2015)	Anders and Huber (2010), Robinson et al. (2010), Love et al. (2014)
35				
36				
37				
38	blast2GO, Trinotate	functional annotation of assembled transcripts	complete annotation pipelines including gene ontology and pathway enrichment analyses	Conesa et al. (2005), Haas et al. (2013)
39				

\* This is a non-exhaustive list of software that focuses on *de novo* loci assembly and genotype calling for RADseq data, as many practitioners working on NMOs will not have access to a reference genome. Other programs (e.g., GATK and ANGSD) that undertake genotype calling using reference-aligned loci are described in the whole genome sequencing section.

Cammen\_SupMat\_TableS1 - Marine mammal genomics - *JHered*

Table S1. Broad applications of genomic tools in studies of non-model organisms are provided with concrete examples of research areas drawn from the field of marine mammal genomics. The number of loci used in each study provides an estimate of the scope of the respective genomic tools and study, but represents the outcome of several filtering steps from raw sequence data that vary across studies. Further details of each method can be found in the listed references. Please note that this is not an exhaustive list. GBS: Genotyping by Sequencing; RADseq: restriction site-associated DNA sequencing; SNP: single nucleotide polymorphism; TSC: target sequence capture; WGS: whole genome sequencing.

Method	# loci	Research area	Reference
<i>Evolutionary genomics: describe evolutionary history and adaptation</i>			
Mitogenome sequencing	Mitogenome	Cetacean phylogenomics	McGowen et al. (2009)
TSC	Mitogenome	Comparison of sub-fossil and modern killer whales	Foote et al. (2013)
TSC	>30kb coding sequence	Evolution of Sirenia	Springer et al. (2015)
WGS	Whole genome	Yangtze river dolphin genome analysis	Zhou et al. (2013)
WGS	Whole genome	Minke whale genome analysis	Yim et al. (2014)
WGS	Whole genome	Bowhead whale genome analysis	Keane et al. (2015)
WGS	Whole genome	Analysis of convergent evolution in marine mammal lineages	Foote et al. (2015)
WGS	10,025 coding sequences	Positive selection in common bottlenose dolphin genome	McGowen et al. (2012)
WGS	Sensory genes	Analysis of gene loss in olfaction and taste in Antarctic minke whale	Kishida et al. (2015)
Genome re-seq	Whole genome	Speciation and adaptation in brown and polar bears	Liu et al. (2014)
Transcriptomics	9,395 genes	Evolution of longevity in bowhead whales	Seim et al. (2014)
Transcriptomics	103,077 unigenes	Osmoregulatory divergence in narrow-ridged finless porpoise	Ruan et al. (2015)
<i>Population genomics: characterize population structure and investigate demography</i>			
RADseq	3,281 SNPs	Killer whale ecotype divergence	Moura et al. (2014)
RADseq (GBS)	24,996 loci; 4,854 SNPs	Historical demography in Atlantic walrus	Shafer et al. (2015)
TSC	Mitogenome and 43-118 nuclear loci	Phylogeography and population genomics of cetaceans	Hancock-Hanser et al. (2013); Morin et al. (2015)
Genome re-seq	Whole genome	Demographic history, population differentiation, and ecotype divergence in killer whales	Foote et al. (2016)

Cammen\_SupMat\_TableS1 - Marine mammal genomics - *JHered*

<i>Adaptation genomics: describe relationships between genomic variation and fitness</i>			
RADseq	83,148 loci; 14,585 SNPs	Effect of inbreeding depression on parasite infection in harbor seals	Hoffman et al. (2014)
RADseq	129,494 loci; 7,431 SNPs	Common bottlenose dolphin adaptation to harmful algal blooms	Cammen et al. (2015)
Transcriptomics	11,286 contigs	Sperm whale skin cell response to hexavalent chromium	Pabuwal et al. (2013)
Transcriptomics	164,966 contigs	Physiological stress response in northern elephant seals	Khudyakov et al. (2015a; 2015b)
<i>Develop molecular resources</i>			
RADseq	3,595 loci	Comparison of short-beaked common dolphin and harbor porpoise	Viricel et al. (2014)
Shotgun sequencing	440,718 SNPs	SNP discovery in Northeast Atlantic common bottlenose dolphins	M. Louis (unpubl. data)
WGS	144 SNPs	SNP validation in Antarctic fur seal	Humble et al. (2016)
Transcriptomics	23,096 contigs; 144 SNPs	Gene and SNP discovery in Antarctic fur seal	Hoffman et al. (2011; 2012; 2013)
Transcriptomics & RADseq	9,000 SNPs	Development of SNP array for polar bear and demonstration of utility in population genomics	Malenfant et al. (2015)

Cammen\_SupMat\_TableS1 - Marine mammal genomics - *JHered*

## References

- Cammen KM, Schultz TF, Rosel PE, Wells RS, Read AJ. 2015. Genomewide investigation of adaptation to harmful algal blooms in common bottlenose dolphins (*Tursiops truncatus*). *Mol Ecol*. 24:4697-4710.
- Footo AD, Liu Y, Thomas GWC, Vinař Ts, Alföldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V, *et al.* 2015. Convergent evolution of the genomes of marine mammals. *Nat Genet*. 47:272-275.
- Footo AD, Newton J, Ávila-Arcos MC, Kampmann M-L, Samaniego JA, Post K, Rosing-Asvid A, Sinding M-HS, Gilbert MTP. 2013. Tracking niche variation over millennial timescales in sympatric killer whale lineages. *Proc R Soc Lond B Biol Sci*. 280:20131481.
- Footo AD, Vijay N, Ávila-Arcos M, Baird RW, Durban JW, Fumagalli M, Gibbs RA, Hanson MB, Korneliussen TS, Martin MD, *et al.* 2016. Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nat Commun*. 7:11693.
- Hancock-Hanser BL, Frey A, Leslie MS, Dutton PH, Archer FI, Morin PA. 2013. Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Mol Ecol Resour*. 13:254-268.
- Hoffman JI. 2011. Gene discovery in the Antarctic fur seal (*Arctocephalus gazella*) skin transcriptome. *Mol Ecol Resour*. 11:703-710.
- Hoffman JI, Simpson F, David P, Rijks JM, Kuiken T, Thorne MAS, Lacy RC, Dasmahapatra KK. 2014. High-throughput sequencing reveals inbreeding depression in a natural population. *Proc Natl Acad Sci USA*. 111:3775-3780.
- Hoffman JI, Thorne MAS, Trathan PN, Forcada J. 2013. Transcriptome of the dead: characterisation of immune genes and marker development from necropsy samples in a free-ranging marine mammal. *BMC Genomics*. 14:52.
- Hoffman JI, Tucker R, Bridgett SJ, Clark MS, Forcada J, Slate J. 2012. Rates of assay success and genotyping error when single nucleotide polymorphism genotyping in non-model organisms: a case study in the Antarctic fur seal. *Mol Ecol Resour*. 12:861-872.
- Humble E, Martinez-Barrio A, Forcada J, Trathan PN, Thorne MAS, Hoffmann M, Wolf JBW, Hoffman JI. 2016. A draft fur seal genome provides insights into factors affecting SNP validation and how to mitigate them. *Mol Ecol Resour*.
- Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, Madsen LB, van Dam S, Brawand D, Marques PI, *et al.* 2015. Insights into the evolution of longevity from the bowhead whale genome. *Cell Reports*. 10:112-122.
- Khudyakov JI, Champagne CD, Preeyanon L, Ortiz RM, Crocker DE. 2015a. Muscle transcriptome response to ACTH administration in a free-ranging marine mammal. *Physiol Genomics*. 47:318-330.
- Khudyakov JI, Preeyanon L, Champagne CD, Ortiz RM, Crocker DE. 2015b. Transcriptome analysis of northern elephant seal (*Mirounga angustirostris*) muscle tissue provides a novel molecular resource and physiological insights. *BMC Genomics*. 16:64.
- Kishida T, Thewissen JGM, Hayakawa T, Imai H, Agata K. 2015. Aquatic adaptation and the evolution of smell and taste in whales. *Zool J Linn Soc*. 176:1-9.
- Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, Xiong Z, Zhou L, Korneliussen TS, Somel M, Babbitt C, *et al.* 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*. 157:785-794.

Cammen\_SupMat\_TableS1 - Marine mammal genomics - *JHered*

- 1  
2  
3 Malenfant RM, Coltman DW, Davis CS. 2015. Design of a 9K Illumina BeadChip for polar bears (*Ursus maritimus*) from RAD and  
4 transcriptome sequencing. *Mol Ecol Resour.* 15:587-600.
- 5  
6 McGowen MR, Grossman LI, Wildman DE. 2012. Dolphin genome provides evidence for adaptive evolution of nervous system genes  
7 and a molecular rate slowdown. *Proc R Soc Lond B Biol Sci.* 279:3643-3651.
- 8  
9 McGowen MR, Spaulding M, Gatesy J. 2009. Divergence date estimation and a comprehensive molecular tree of extant cetaceans.  
10 *Mol Phylogenet Evol.* 53:891-906.
- 11  
12 Morin PA, Parsons KM, Archer FI, Ávila-Arcos M, Barrett-Lennard LG, Dalla Rosa L, Duchêne S, Durban JW, Ellis GM, Ferguson  
13 SH, *et al.* 2015. Geographic and temporal dynamics of a global radiation and diversification in the killer whale. *Mol Ecol.*  
14 24:3964-3979.
- 15  
16 Moura AE, Kenny JG, Chaudhuri R, Hughes MA, Welch AJ, Reisinger RR, de Bruyn PJN, Dahlheim ME, Hall N, Hoelzel AR. 2014.  
17 Population genomics of the killer whale indicates ecotype evolution in sympatry involving both selection and drift. *Mol Ecol.*  
18 23:5179-5192.
- 19  
20 Pabuwal V, Boswell M, Pasquali A, Wise SS, Kumar S, Shen Y, Garcia T, Lacerte C, Wise JP, Jr., Wise JP, Sr., *et al.* 2013.  
21 Transcriptomic analysis of cultured whale skin cells exposed to hexavalent chromium [Cr(VI)]. *Aquat Toxicol.* 134-135:74-81.
- 22  
23 Ruan R, Guo A-H, Hao Y-J, Zheng J-S, Wang D. 2015. *De novo* assembly and characterization of narrow-ridged finless porpoise  
24 renal transcriptome and identification of candidate genes involved in osmoregulation. *Int J Mol Sci.* 16:2220-2238.
- 25  
26 Seim I, Ma S, Zhou X, Gerashchenko MV, Lee S-G, Suydam R, George JC, Bickham JW, Gladyshev VN. 2014. The transcriptome of  
27 the bowhead whale *Balaena mysticetus* reveals adaptations of the longest-lived mammal. *Aging.* 6:879-899.
- 28  
29 Shafer ABA, Gattepaille LM, Stewart REA, Wolf JBW. 2015. Demographic inferences using short-read genomic data in an  
30 approximate Bayesian computation framework: *in silico* evaluation of power, biases and proof of concept in Atlantic walrus.  
31 *Mol Ecol.* 24:328-345.
- 32  
33 Springer MS, Signore AV, Paijmans JLA, Vélez-Juarbe J, Domning DP, Bauer CE, He K, Crerar L, Campos PF, Murphy WJ, *et al.*  
34 2015. Interordinal gene capture, the phylogenetic position of Steller's sea cow based on molecular and morphological data, and  
35 the macroevolutionary history of Sirenia. *Mol Phylogenet Evol.* 91:178-193.
- 36  
37 Viricel A, Pante E, Dabin W, Simon-Bouhet B. 2014. Applicability of RAD-tag genotyping for interfamilial comparisons: empirical  
38 data from two cetaceans. *Mol Ecol Resour.* 14:597-605.
- 39  
40 Yim H-S, Cho YS, Guang X, Kang SG, Jeong J-Y, Cha S-S, Oh H-M, Lee J-H, Yang EC, Kwon KK, *et al.* 2014. Minke whale  
41 genome and aquatic adaptation in cetaceans. *Nat Genet.* 46:88-92.
- 42  
43 Zhou X, Sun F, Xu S, Fan G, Zhu K, Liu X, Chen Y, Shi C, Yang Y, Huang Z, *et al.* 2013. Baiji genomes reveal low genetic  
44 variability and new insights into secondary aquatic adaptations. *Nat Commun.* 4:2708.
- 45  
46  
47  
48  
49

1  
2  
3 1 **Genomic methods take the plunge: recent advances in high-throughput sequencing of**  
4 2 **marine mammals**

5 3  
6 4 KRISTINA M. CAMMEN<sup>1\*</sup>, KIMBERLY R. ANDREWS<sup>2</sup>, EMMA L. CARROLL<sup>3</sup>, ANDREW D.  
7 5 FOOTE<sup>4</sup>, EMILY HUMBLE<sup>5,6</sup>, JANE I. KHUDYAKOV<sup>7</sup>, MARIE LOUIS<sup>3</sup>, MICHAEL R.  
8 6 MCGOWEN<sup>8</sup>, MORTEN TANGE OLSEN<sup>9</sup>, AND AMY M. VAN CISE<sup>10</sup>  
9 7

10 8 <sup>1</sup>School of Marine Sciences, University of Maine, Orono, Maine 04469, USA

11 9 <sup>2</sup>Department of Fish and Wildlife Sciences, University of Idaho, 875 Perimeter Drive MS 1136,  
12 10 Moscow, Idaho 83844-1136, USA

13 11 <sup>3</sup>Scottish Oceans Institute, University of St Andrews, East Sands, St Andrews, Fife KY16 8LB,  
14 12 UK

15 13 | <sup>4</sup>Computational and Molecular Population Genetics **CMPG-L**lab, Institute of Ecology and  
16 14 Evolution, University of Bern, Bern CH-3012, Switzerland

17 15 <sup>5</sup>Department of Animal Behaviour, University of Bielefeld, Postfach 100131, 33501 Bielefeld,  
18 16 Germany

19 17 <sup>6</sup>British Antarctic Survey, High Cross, Madingley Road, Cambridge CB3 0ET, UK

20 18 <sup>7</sup>Department of Biology, Sonoma State University, Rohnert Park, California 94928, USA

21 19 <sup>8</sup>School of Biological and Chemical Sciences, Queen Mary University of London,  
22 20 Mile End Road, London E1 4NS, UK

23 21 <sup>9</sup>Evolutionary Genomics Section, Natural History Museum of Denmark, University of  
24 22 Copenhagen, DK-1353 Copenhagen K, Denmark

25 23 <sup>10</sup>Scripps Institution of Oceanography, 8622 Kennel Way, La Jolla, California 92037, USA  
26 24

27 25 \*Corresponding author: kristina.cammen@maine.edu  
28 26

29 27 Running title: Marine mammal genomics  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## Abstract

The dramatic increase in the application of genomic techniques to non-model organisms over the past decade has yielded numerous valuable contributions to evolutionary biology and ecology, many of which would not have been possible with traditional genetic markers. We review this recent progression with a particular focus on genomic studies of marine mammals, a group of taxa that represent key macroevolutionary transitions from terrestrial to marine environments and for which available genomic resources have recently undergone notable rapid growth. Genomic studies of non-model organisms utilize an expanding range of approaches, including ~~low- and high-coverage~~ whole genome sequencing, restriction site-associated DNA sequencing, array-based ~~high-throughput~~ sequencing of single nucleotide polymorphisms and target sequence probes (e.g., exomes), and transcriptome sequencing. These approaches generate different types and quantities of data, and many can be applied with limited or no prior genomic resources, thus overcoming one traditional limitations of research on non-model organisms. Within marine mammals, such studies have thus far yielded significant contributions to the fields of phylogenomics and comparative genomics, as well as enabled investigations of fitness, demography, and population structure ~~in natural populations~~. Here, we review the primary options for generating genomic data, introduce several emerging techniques, and discuss the suitability of each approach for different applications in the study of non-model organisms.

**Keywords:** RADseq, SNP array, target sequence capture, whole genome sequencing, RNAseq, non-model organisms



## 49 **Introduction**

50 Recent advances in sequencing technologies, coincident with dramatic declines in cost, have  
51 increasingly enabled the application of genomic sequencing in non-model systems (Ekblom and  
52 Galindo 2011; Ellegren 2014). These advances in molecular technologies have in many ways  
53 begun to blur the distinction between model and non-model organisms (Armengaud et al. 2014).  
54 Non-model organisms (NMOs) have traditionally been defined as those for which whole-  
55 organism experimental manipulation is rarely, if ever, possible due to logistical and/or ethical  
56 constraints (Ankeny and Leonelli 2011). Further, NMOs have typically been characterized by  
57 limited genomic resources, but this is becoming increasingly less so as the number of NMO  
58 reference genomes grows rapidly, for example through efforts like the Genome 10K Project  
59 (Koepfli et al. 2015). In fact, in some taxonomic orders, we are approaching the point at which  
60 all species have at least one representative reference genome available for a closely related  
61 species (Fig 1).

62  
63 Despite the limitations of working with NMOs, including potentially small sample sizes, low  
64 DNA quantity, and limited information on gene function, genetic and genomic investigations of  
65 NMOs have yielded numerous valuable contributions to understanding their evolutionary  
66 biology and ecology. For the past several decades, traditional genetic markers such as  
67 microsatellites and short fragments of mitochondrial DNA (e.g., the control region) have been  
68 extensively used in molecular ecology. These markers, which typically evolve under neutral  
69 expectations, have proven useful for identifying population structure and reconstructing  
70 population demographic history (Hedrick 2000). However, the power of such studies is limited  
71 by the number of markers that can feasibly be evaluated using traditional approaches. The advent  
72 of low-cost high-throughput sequencing has led to dramatic increases in the number of neutral  
73 markers that can be evaluated, in many cases improving our power to resolve fine-scale or  
74 cryptic population structure in species with high dispersal capability (e.g., Corander et al. 2013)  
75 and improving the accuracy of estimating some (though not all) demographic parameters (Li and  
76 Jakobsson 2012; Shafer et al. 2015). Importantly, high-throughput sequencing has also further  
77 enabled genomic studies of non-neutral processes in NMOs, for example, characterizing both  
78 deleterious and adaptive variation within and across species (Stinchcombe and Hoekstra 2008;

1  
2  
3 79 K nstner et al. 2010). It is increasingly evident that genomic analyses of NMOs can and have  
4  
5 80 provided important insights that could not be identified with traditional genetic markers.  
6  
7 81

8  
9 82 Many molecular ecologists now face the challenge of deciding which of the broad range of  
10  
11 83 genomic approaches to apply to their study systems. Here we review the primary options for  
12  
13 84 generating genomic data and their relative suitability for different applications in the study of  
14  
15 85 NMOs. We focus on marine mammals, which represent several mammalian clades with notably  
16  
17 86 rapid growth in available genomic resources in recent years. This growth is clearly evident in  
18  
19 87 both publication rate (Fig 2) and the rise in number and size of genomic sequences deposited in  
20  
21 88 public resources (Fig 3). We comprehensively review the literature on marine mammal  
22  
23 89 genomics, highlighting recent trends in methodology and applications, and then describe in detail  
24  
25 90 the molecular approaches that are most commonly applied to studies of ~~non-model~~NMO  
26  
27 91 genomics. Our hope is that this review will highlight the promise of genomics for NMOs and  
28  
29 92 offer guidance to researchers considering the application of genomic techniques in their non-  
30  
31 93 model study system of choice.  
32

### 33 94

### 34 95 **Why study marine mammal genomics?**

35 96 Marine mammals represent key macroevolutionary transitions from terrestrial to marine  
36  
37 97 environments (McGowen et al. 2014) and accordingly are an exemplary system for investigating  
38  
39 98 the evolution of several morphological and physiological adaptations (Foote et al. 2015)  
40  
41 99 associated with locomotion (Shen et al. 2012), sight (Meredith et al. 2013), echolocation (Parker  
42  
43 100 et al. 2013; Zou and Zhang 2015), deep diving (Mirceta et al. 2013), osmoregulation (Ruan et al.  
44  
45 101 2015), and cognition (McGowen et al. 2012). Furthermore, studies of marine mammal evolution  
46  
47 102 to date have characterized several unique aspects of their genome evolution that merit further  
48  
49 103 investigation, including low genomic diversity and a relatively slow molecular clock, especially  
50  
51 104 in cetaceans (Jackson et al. 2009; McGowen et al. 2012; Zhou et al. 2013). As many cetacean  
52  
53 105 species are highly mobile with no obvious physical geographic barriers to dispersal, they provide  
54  
55 106 a unique opportunity to study the role of behavior and culture in shaping population structure and  
56  
57 107 genetic diversity (Riesch et al. 2012; Carroll et al. 2015; Alexander et al. 2016). ~~Finally, I~~though  
58  
59 108 highly mobile, many marine mammals exhibit evidence of local adaptation; for example, several  
60  
109 species show parallel divergent morphological and behavioral adaptations to coastal and pelagic

1  
2  
3 110 environments (Moura et al. 2013; Louis et al. 2014; Viricel and Rosel 2014). These species may  
4  
5 111 be studied across ocean basins as emerging examples of ecological adaptation and speciation  
6  
7 112 (Morin et al. 2010a).  
8  
9 113

10 114 Beyond their value as systems of evolutionary study, many marine mammals are also of broader  
11 115 interest relating to their historical and present conservation status. Many marine mammal  
12 116 populations share histories of dramatic decline due to hunting and other human impacts.  
13 117 Genomics provides a promising tool with which to expand our insights into these historical  
14 118 population changes, which so far primarily have been derived from archival review and  
15 119 traditional genetic approaches (Ruegg et al. 2013; Sremba et al. 2015). More recently, since the  
16 120 implementation of national and international protections, many marine mammal populations  
17 121 have partially or fully recovered (Magera et al. 2013), yet the conservation status of certain  
18 122 marine mammal populations remains of concern. Such vulnerable populations could benefit  
19 123 greatly from an improved understanding of their genetic diversity and evolution, especially in  
20 124 ways that can inform predictions of adaptive capacity to anthropogenic pressures and expand the  
21 125 toolkit for conservation policy (Garner et al. 2016; Taylor and Gemmill 2016).  
22  
23 126

### 24 127 **Recent trends in marine mammal genomics**

25 128 We conducted a meta-analysis of the peer-reviewed marine mammal genomics literature to  
26 129 evaluate trends in publication rates across research methodologies and aims. A search of the Web  
27 130 of Science database using the term “genom\*” and one of the following terms indicating study  
28 131 species - “marine mammal”, “pinniped”, “seal”, “sea lion”, “sea otter”, “whale”, “dolphin”,  
29 132 “polar bear”, “manatee” - identified 825 records on December 11, 2015. We excluded 77% of the  
30 133 search results that were not directly related to genomic studies in marine mammal systems. The  
31 134 remaining 101 articles that were relevant to marine mammal genomics were further categorized  
32 135 by primary research methodology and general research aim. A subset of these articles is  
33 136 described briefly in Supplemental Table 1.  
34  
35 137

36 138 | From the early 1990s through 2015, published literature in the field ~~has~~ shifted from an early  
37 139 focus on mitogenome sequencing to more sequence-intensive approaches, such as transcriptome  
38 140 and whole genome sequencing (Figs 2 and 4). This trajectory closely follows trends in  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 141 sequencing technologies, from Sanger sequencing of short- and long-range PCR products for  
4  
5 142 mitogenome sequencing (Arnason et al. 1991) and SNP discovery (Olsen et al. 2011), to high-  
6  
7 143 throughput sequencing of reduced-representation genomic libraries (RRLs) that consist of  
8  
9 144 selected subsets of the genome (e.g., Viricel et al. 2014), to high-throughput sequencing of whole  
10  
11 145 genomes with varying levels of depth, ~~of~~ coverage, and contiguity. Today, high-throughput  
12  
13 146 sequencing can be used both to generate high-quality reference genome assemblies (Yim et al.  
14  
15 147 2014; Foote et al. 2015; Humble et al. 2016) and to re-sequence whole genomes at a population  
16  
17 148 scale (Liu et al. 2014a; Foote et al. 2016). Similarly, the scale of gene expression studies has  
18  
19 149 increased from quantitative real-time PCR of candidate genes (Tabuchi et al. 2006) to  
20  
21 150 microarrays containing hundreds to thousands of genes (Mancia et al. 2007) and high-throughput  
22  
23 151 RNAseq that evaluates hundreds of thousands of contigs across the genome (Khudyakov et al.  
24  
25 152 2015b). As the cost of high-throughput sequencing continues to decline, we anticipate an  
26  
27 153 increase in studies that sequence RRLs, whole genomes, and transcriptomes in NMOs at a  
28  
29 154 population scale.

30  
31 156 Marine mammal genomic studies thus far have primarily contributed to the fields of  
32  
33 157 phylogenomics and comparative genomics (Fig 2, Table S1). Several of these comparative  
34  
35 158 genomics studies have aimed to improve our understanding of the mammalian transition to an  
36  
37 159 aquatic lifestyle and describe the evolutionary relationships within and among marine mammals  
38  
39 160 and their terrestrial relatives (McGowen et al. 2014; Foote et al. 2015). Whereas such studies  
40  
41 161 require only a single representative genome per species, an emerging class of studies applying  
42  
43 162 genomic techniques at a population scale enables further investigations of fitness, demography,  
44  
45 163 and population structure within ~~a~~-species (Table S1). However, expanding the scale of genomic  
46  
47 164 studies requires careful selection of an appropriate method for data generation and analysis; from  
48  
49 165 a growing number of approaches that are becoming available to non-model systems.

### 167 **Data generation**

50  
51 168 Our review of marine mammal genomics highlights an increasing number of options for the  
52  
53 169 generation and analysis of genomic data. Choosing which of these sequencing strategies to apply  
54  
55 170 is a key step in any genomics study. Here, we describe approaches that have been used  
56  
57 171 successfully in order to help guide future studies of ecological, physiological, and evolutionary

1  
2  
3 172 genomics in NMOs. Across data generation methods, we highlight approaches that can be used  
4  
5 173 with limited or no prior genomic resources, overcoming one traditional challenge of genomic  
6  
7 174 studies of NMOs (the need for a reference genome to which sequencing reads can be mapped).  
8  
9 175 These methods produce a range in quantity and type of data output, from hundreds of SNPs to  
10  
11 176 whole genome sequences, and from single individuals to population samples, reflecting the  
12  
13 177 trade-off between number of samples and amount of data generated per sample.  
14

178

#### 15 16 179 Sample collection, storage and extraction

17  
18 180 Prior to starting a genomic study, researchers must recognize that many recent methods for high-  
19  
20 181 throughput sequencing require genetic material of much higher quality and quantity than  
21  
22 182 techniques used to characterize traditional genetic markers. These more stringent sample  
23  
24 183 requirements necessitate new standards for tissue sampling, storage, and DNA/RNA extraction.  
25  
26 184 Ideally, samples should be collected from live or newly deceased individuals and stored at -80°C,  
27  
28 185 or when this is not possible at -20°C in RNAlater, Trizol, ethanol, salt-saturated DMSO, or dry,  
29  
30 186 depending on the intended application. Given the sensitivity of new sequencing methods, great  
31  
32 187 care should be taken to minimize cross-contamination during sampling, as even minute amounts  
33  
34 188 of genetic material from another individual can bias downstream analyses, for example variant  
35  
36 189 genotyping and gene expression profiles. Choice of extraction method varies with sample type  
37  
38 190 and study aim, but typically genomic methods require cleanup and treatment with RNase to yield  
39  
40 191 pure extracts, whereas RNAseq methods require rigorous DNase treatment to remove genomic  
41  
42 192 contamination that can bias expression results. Depending on the genomic methodology, target  
43  
44 193 quantities for a final sample may range from as low as 50 ng of DNA for some RRL sequencing  
45  
46 194 methods (Andrews et al. 2016) up to ~1 mg for sequencing the full set of libraries (of different  
47  
48 195 insert sizes) necessary for high-quality genome assemblies (Ekblom and Wolf 2014). Most  
49  
50 196 commercial RNAseq library preparation services require at least 500-1,000 ng of pure total RNA  
51  
52 197 that shows minimal degradation as measured by capillary gel electrophoresis (RNA Integrity  
53  
54 198 Number (RIN)  $\geq 8$ ). Samples should ideally consist of high molecular weight genetic material  
55  
56 199 (with little shearing), though continuing molecular advances enable genomic sequencing even of  
57  
58 200 low quantity or poor quality starting material. Extreme examples of the latter include  
59  
60 201 successfully sequenced whole genomes from ancient material (e.g., Rasmussen et al. 2010;

202 Meyer et al. 2012; Allentoft et al. 2015), including a more than 500,000-year-old horse (Orlando  
203 et al. 2013).

## 204 205 Reduced-representation genome sequencing

### 206 *i. RADseq*

207 Reduced-representation sequencing methods evaluate only a small portion of the genome,  
208 allowing researchers to sequence samples from a larger number of individuals within a given  
209 budget in comparison to sequencing whole genomes. Restriction site-associated DNA  
210 sequencing (RADseq) is currently the most widely-used RRL sequencing method for NMOs  
211 (Davey et al. 2011; Narum et al. 2013; Andrews et al. 2016). RADseq generates sequence data  
212 from short regions adjacent to restriction cut sites and therefore targets markers that are  
213 distributed relatively randomly across the genome and occur primarily in non-coding regions.  
214 This method allows simultaneous discovery and genotyping of thousands of genetic markers for  
215 virtually any species, regardless of availability of prior genomic resources. Of greatest interest  
216 are variable markers, characterized either as single SNPs or phased alleles that can be resolved  
217 from the identification of several [SNPs-variants](#) within a single locus.

218  
219 The large number of markers generated by RADseq dramatically increases genomic resolution  
220 and statistical power for addressing many ecological and evolutionary questions when compared  
221 to studies using traditional markers (Table S1). For example, heterozygosity-fitness [associations](#)  
222 [correlations](#) in harbor seals (*Phoca vitulina*) were nearly fivefold higher when using 14,585  
223 RADseq SNPs than when using 27 microsatellite loci (Hoffman et al. 2014). A recent study on  
224 the Atlantic walrus (*Odobenus rosmarus rosmarus*) using 4,854 RADseq SNPs to model  
225 demographic changes in connectivity and effective population size associated with the Last  
226 Glacial Maximum (Shafer et al. 2015) both supported and extended inferences from previous  
227 studies using traditional markers (Shafer et al. 2010; Shafer et al. 2014).

228  
229 Furthermore, RADseq can provide sufficient numbers of markers across the genome to identify  
230 genomic regions influenced by natural selection ~~in some cases~~. These analyses require large  
231 numbers (thousands to tens of thousands) of markers to ensure that some markers will be in  
232 linkage disequilibrium with genomic regions under selection and to minimize false positives,

1  
2  
3 233 particularly under non-equilibrium demographic scenarios (Narum and Hess 2011; De Mita et al.  
4 234 2013; Lotterhos and Whitlock 2014). Extreme demographic shifts, as experienced by many  
5 235 marine mammal populations (e.g., killer whales, Foote et al. 2016), can drive shifts in allele  
6 236 frequencies that confound the distinction of drift and selection and make it difficult to detect  
7 237 genomic signatures of selection (Poh et al. 2014). Proof of concept of the application of RADseq  
8 238 for identifying genomic signatures of selection in wild populations was demonstrated in three-  
9 239 spined sticklebacks (*Gasterosteus aculeatus*), for which analyses of over 45,000 SNPs  
10 240 (Hohenlohe et al. 2010) identified genomic regions of known evolutionary importance associated  
11 241 with differences between marine and freshwater forms (Colosimo et al. 2005; Barrett et al.  
12 242 2008). RADseq studies with similar aims in marine mammals have resulted in comparatively  
13 243 sparser sampling of SNPs (<10,000), likely due to both methodological differences and generally  
14 244 low genetic diversity particularly among cetaceans. Nonetheless, genomic regions associated  
15 245 with resistance to harmful algal blooms in common bottlenose dolphins (*Tursiops truncatus*)  
16 246 were identified across multiple pairwise comparisons using 7,431 RADseq SNPs (Cammen et al.  
17 247 2015), and genomic regions associated with habitat use and resource specialization in killer  
18 248 whales (*Orcinus orca*) were identified using 3,281 RADseq SNPs (Moura et al. 2014a). Some of  
19 249 these RADseq SNPs associated with diet in killer whales were later also confirmed as occurring  
20 250 in genomic regions of high differentiation and reduced diversity consistent with a signature of  
21 251 selection identified in a study utilizing ~~low-coverage~~ whole genome re-sequencing (Foote et al.  
22 252 2016). It will remain important for further studies of genomic signatures of selection in NMOs to  
23 253 carefully consider which approaches will generate a sufficiently large number of SNPs to  
24 254 accurately identify the range of putatively neutral  $F_{ST}$  values (and thus outliers) given the  
25 255 demographic history of the population (Lotterhos and Whitlock 2014).

256  
257 Numerous laboratory methods have been developed for generating RADseq data (reviewed in  
258 Andrews et al. 2016), with the most popular library preparation methods currently being the  
259 original RAD (Miller et al. 2007; Baird et al. 2008), Genotyping by Sequencing (GBS, Elshire et  
260 al. 2011; Poland et al. 2012), and double digest RAD (ddRAD, Peterson et al. 2012). All  
261 RADseq methods share the common goal of sequencing regions adjacent to restriction cut sites  
262 across the genome, but differ in technical details, such as the number and type of restriction  
263 enzymes used, the mechanisms for reducing genomic DNA fragment sizes, and the strategies for

1  
2  
3 264 attaching sequencing adapters to the target DNA fragments. For example, both the original RAD  
4  
5 265 method and GBS use a single enzyme digest, but the original RAD ~~proteool-method~~ uses a rare-  
6  
7 266 cutting enzyme and mechanical shearing to reduce DNA fragment size (Baird et al. 2008),  
8  
9 267 whereas GBS uses a more frequent-cutting enzyme and relies on preferential PCR amplification  
10  
11 268 of shorter fragments for indirect size selection (Elshire et al. 2011). These ~~types-of~~  
12 269 ~~variationmodifications~~ lead to differences across methods in the time and cost of library  
13  
14 270 preparation, the number and lengths of loci produced, and the types of error and bias present in  
15  
16 271 the resulting data. Different RADseq methods will be better suited to different research  
17  
18 272 questions, study species, and research budgets, and therefore researchers embarking on a  
19  
20 273 RADseq study should carefully consider the suitability of each method for their individual  
21  
22 274 projects. Further details on the advantages and disadvantages of each method are described in  
23  
24 275 Andrews et al. (2016).

25 276

26 277 *ii. SNP arrays*

28 278 An alternative high-throughput reduced-representation genotyping approach involves the use of  
29  
30 279 custom arrays designed to capture and sequence targeted regions of the genome. Such array-  
31  
32 280 based approaches may provide certain advantages over RADseq, including the ability to easily  
33  
34 281 estimate genotyping error rates, scalability to thousands of samples, lower requirements for DNA  
35  
36 282 quantity/quality and technical effort, greater comparability of markers across studies, and the  
37  
38 283 ability to genotype SNPs within candidate genomic regions. However, unlike RADseq, array-  
39  
40 284 based techniques require prior knowledge of the study system's genome or the genome of a  
41  
42 285 closely related species, which remains unavailable for some NMOs. Furthermore, SNP arrays  
43  
44 286 must take into account the potential for ascertainment bias (e.g., Malenfant et al. 2015), whereas  
45  
46 287 RADseq avoids ascertainment bias by simultaneously discovering and genotyping markers.

47 288

48 289 To identify SNPs for NMO array development, researchers must rely on existing genomic  
49  
50 290 resources or generate new reference sequences, in the form of whole or reduced-representation  
51  
52 291 genomes or transcriptomes (Hoffman et al. 2012; Malenfant et al. 2015). When a whole genome  
53  
54 292 reference assembly is available for the target species or a related species, multiplex shotgun  
55  
56 293 sequencing can facilitate the rapid discovery of hundreds of thousands of SNPs for array  
57  
58 294 development. This SNP discovery approach involves high-throughput sequencing of sheared  
59  
60



1  
2  
3 295 genomic DNA, which that can be sequenced at a low depth of coverage (i.e., low mean read  
4 depth across the genome) if suitable genotype likelihood-based methods (O'Rawe et al. 2015) are  
5 296  
6  
7 297 used to identify polymorphic sites. Thus, this approach is less restrictive in terms of DNA  
8  
9 298 quality. For example, shotgun sequencing of 33 Northeast Atlantic common bottlenose dolphins,  
10  
11 299 which included degraded DNA collected from stranded specimens, on one Illumina HiSeq2000  
12  
13 300 lane of 100\_-bp single-end sequencing identified 440,718 high-quality SNPs (M. Louis  
14 301 unpublished data). Such dense sampling of SNPs is essential for studies of population genomics  
15  
16 302 that require a large number of markers, such as for inferences of demographic history  
17  
18 303 (Gutenkunst et al. 2009; Excoffier et al. 2013; Liu and Fun 2015) and selective sweeps (Chen et  
19  
20 304 al. 2010). Once a set of putative markers has been identified, hybridization probes can be  
21  
22 305 designed from their flanking sequences and printed onto a SNP array. The two principal SNP  
23  
24 306 genotyping platforms supporting thousands to millions of SNPs are the Illumina Infinium  
25 307 iSelect® and Affymetrix Axiom® arrays.

26 308  
27  
28 309 The use of SNP arrays in NMOs has thus far been somewhat limited, potentially due to low SNP  
29  
30 310 validation rates (Chancerel et al. 2011; Helyar et al. 2011), issues of ascertainment bias  
31  
32 311 (Albrechtsen et al. 2010; McTavish and Hillis 2015), and cost of SNP discovery. However, using  
33  
34 312 both SNP data and whole genome sequence from the Antarctic fur seal (*Arctocephalus gazella*),  
35  
36 313 Humble et al. (2016) recently demonstrated that careful filtering based on SNP genomic context  
37  
38 314 prior to array development has the potential to substantially increase assay success rates. Further,  
39  
40 315 ascertainment bias can be reduced by selecting samples for SNP discovery that span the  
41  
42 316 geographic range of populations that will be target\_-sequenced (Morin et al. 2004). By  
43  
44 317 accounting for ascertainment bias, Malenfant et al. (2015) were able to demonstrate population  
45  
46 318 structure in Canadian polar bears (*Ursus maritimus*) more clearly using a 9K SNP array than 24  
47  
48 319 microsatellite markers.

### 49 321 *iii. Target sequence capture*

50 322 Target sequence capture (TSC, also called target enrichment, direct selection, or Hyb-seq) has  
51  
52 323 many of the same advantages and disadvantages as the array-based SNP approaches described  
53  
54 324 above, but differs in library preparation, sequencing platform, and resulting sequence data. While  
55  
56 325 SNP arrays genotype single variable positions, TSC can be used to sequence selected short  
57  
58  
59  
60

1  
2  
3 326 fragments. With TSC, researchers can amplify and sequence up to a million target probes on  
4  
5 327 solid-state arrays, and even more if in-solution arrays are used. This gives the user the ability to  
6  
7 328 choose to sequence many samples in parallel (Cummings et al. 2010), as many as 100-150 per  
8  
9 329 Illumina HiSeq lane, or to sequence many regions per individual. Recent advances in target  
10  
11 330 enrichment, such as genotyping in thousands (Campbell et al. 2015), anchored hybrid enrichment  
12  
13 331 (Lemmon et al. 2012), and target capture of ultra-conserved elements (UCEs, Faircloth et al.  
14  
15 332 2012; McCormack et al. 2012), have further increased the number of regions and individuals that  
16  
17 333 can be sampled in a single lane. In addition, UCEs overcome the need for a reference genome,  
18  
19 334 enabling their wide application across many NMOs (though designing custom probe sets from  
20  
21 335 closely related species will remain preferable in many cases (Hancock-Hanser et al. 2013)).  
22  
23 336 Although a number of methodological variants have been developed and optimized (Bashiardes  
24  
25 337 et al. 2005; Noonan et al. 2006; Hodges et al. 2009; Cummings et al. 2010; Mamanova et al.  
26  
27 338 2010; Hancock-Hanser et al. 2013), TSC generally relies on hybridization and amplification of  
28  
29 339 specially prepared libraries consisting of fragmented genomic DNA. Many companies offer kits  
30  
31 340 for TSC, such as Agilent (SureSelect) and MYcroarray (MYbaits), with MYcroarray specifically  
32  
33 341 marketing their kits for use with NMOs.

34  
35 343 The most common use of TSC has been the capture of whole exomes in model organisms,  
36  
37 344 including humans (Ng et al. 2009). However, as costs have plummeted, TSC is increasingly  
38  
39 345 being used in investigations of NMOs. TSC is particularly useful in sequencing ancient DNA,  
40  
41 346 where it can enrich the sample for endogenous DNA content relative to exogenous DNA (i.e.,  
42  
43 347 contamination) and thereby increase the relative DNA yield (Ávila-Arcos et al. 2011; Enk et al.  
44  
45 348 2014). For example, TSC has been used to generate mitogenome sequences from subfossil killer  
46  
47 349 whale specimens originating from the mid-Holocene; for comparison with modern lineages  
48  
49 350 (Foote et al. 2013). TSC was also recently utilized to compare >30 kb of exonic sequence from  
50  
51 351 museum specimens of the extinct Steller's sea cow (*Hydrodamalis gigas*) and a modern dugong  
52  
53 352 (*Dugong dugon*) specimen to investigate evolution within Sirenia (Springer et al. 2015). Springer  
54  
55 353 et al. (2016) further used TSC to examine gene evolution related to dentition across edentulous  
56  
57 354 mammals, including mysticetes. Finally, TSC of both exonic and intronic regions has been used  
58  
59 355 to assess genetic divergence across cetacean species (Hancock-Hanser et al. 2013; Morin et al.  
60  
356 2015). These studies show the potential use of TSC across evolutionary time-scales for

1  
2  
3 357 population genomics, phylogenomics, and studies of selection and gene loss across divergent  
4 358 lineages (Table S1).

5  
6  
7 359

8  
9 360 Whole genome sequencing

10 361 Beyond advances enabled by the reduced-representation methods presented above, our power  
11 362 and resolution to elucidate evolutionary processes, including selection and demographic shifts,  
12 363 can be further increased by sequencing whole genomes.

13  
14  
15 364

16  
17 365 *i. High-coverage Reference genome sequencing*

18 366 At the time of publication, there exist-are 12 publicly available<sup>1</sup> whole (or near-whole) marine  
19 367 mammal genomes of varying quality representing 10 families, including 7 cetaceans (Fig 1A), 3  
20 368 pinnipeds (Fig 1B), the West Indian manatee (*Trichechus manatus*), and the polar bear. The first  
21 369 sequenced marine mammal genome was that of the common bottlenose dolphin, which was  
22 370 originally sequenced to ~2.5x depth of coverage using Sanger sequencing (Lindblad-Toh et al.  
23 371 2011). This genome was later improved upon by adding both 454 and Illumina HiSeq data  
24 372 (Foote et al. 2015). Other subsequent marine mammal genomes were produced solely using  
25 373 Illumina sequencing and mate-paired or paired-end libraries with varied insert sizes (Miller et al.  
26 374 2012; Zhou et al. 2013; Yim et al. 2014; Foote et al. 2015; Keane et al. 2015; Kishida et al. 2015;  
27 375 Humble et al. 2016).

28  
29  
30 376

31 377 Whole genome sequencing has been used to address many issues in marine mammal genome  
32 378 evolution, usually by comparison with other existing mammalian genomes. Biological insights  
33 379 discussed in the genome papers listed above include the evolution of transposons and repeat  
34 380 elements, gene evolution and positive selection, predicted population structure through time,  
35 381 SNP validation, molecular clock rates, and convergent molecular evolution (Table S1). For  
36 382 example, analyses of the Yangtze river dolphin (*Lipotes vexillifer*) genome confirmed that a  
37 383 bottleneck occurred in this species during the last period of deglaciation (Zhou et al. 2013). In  
38 384 addition, following upon earlier smaller-scale studies (e.g., Deméré et al. 2008; McGowen et al.

39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  

---

<sup>1</sup> These genomes are available on NCBI's online genome database or Dryad, but they have not all been published. As agreed upon in the Fort Lauderdale Convention, the community standard regarding such unpublished genomic resources is to respect the data generators' right to publish with these data first.

1  
2  
3 385 2008; Hayden et al. 2010), genomic analyses have confirmed the widespread decay of gene  
4  
5 386 families involved in olfaction, gustation, enamelogenesis, and hair growth in some cetaceans  
6  
7 387 (Yim et al. 2014; Kishida et al. 2015). Perhaps the most widespread use of whole genome studies  
8  
9 388 has been the use of models of selection to detect protein-coding genes that show evidence of  
10  
11 389 natural selection in specific lineages. A recent study by Foote et al. (2015) ~~has~~ extended this  
12  
13 390 approach to investigate convergent positive selection among cetaceans, pinnipeds, and sirenians.  
14  
15 391 This study exemplifies a trend in recent genomic studies, ~~which that~~ sequence multiple genomes  
16  
17 392 to address a predetermined evolutionary question, in this case, the molecular signature of aquatic  
18  
19 393 adaptation.

20  
21 394  
22 395 In addition to these evolutionary insights that typically stem from a comparative genomics  
23  
24 396 approach, the development of high-quality reference genome assemblies provide an important  
25  
26 397 resource that facilitates mapping of reduced-representation genomic data (see previous section)  
27  
28 398 as well as ~~relatively low coverage~~ short-read sequencing data with relatively low depth of  
29  
30 399 coverage (see following section). These data types can be generated at relatively low cost on  
31  
32 400 larger sample sizes enabling population-scale genomic studies. In many cases, genome  
33  
34 401 assemblies from closely related species are sufficient for use as a reference. Particularly among  
35  
36 402 marine mammals, given their generally slow rate of nucleotide divergence, it is therefore likely  
37  
38 403 unnecessary to sequence a high-quality reference genome assembly for every species. Instead,  
39  
40 404 resources could be allocated toward population-scale studies, including ~~low coverage~~ genome  
41  
42 405 re-sequencing efforts.

43 406  
44 407 *ii. Population-level ~~low coverage~~ genome re-sequencing*

45 408 In contrast to ~~high coverage~~ reference genome sequencing that today often exceeds 100x mean  
46  
47 409 read coverage depth and typically combines long- and short-insert libraries to generate high-  
48  
49 410 quality assemblies for one to a few individuals, ~~low coverage~~ genome re-sequencing studies  
50  
51 411 ~~capitalize on existing reference assemblies and aim~~ to achieve only  $\geq 2x$  coverage mean read  
52  
53 412 depth on tens to hundreds of individuals from short-insert libraries ~~which that are then whose~~  
54  
55 413 reads are anchored to ~~the existing~~ reference assemblies. ~~Given Despite~~ the inherent trade-offs  
56  
57 414 between cost, read depth, coverage, and sample size, ~~low coverage~~ genome re-sequencing of  
58  
59 415 large numbers of individuals for population-level inference can be conducted at a relatively low  
60

1  
2  
3 416 cost. In the past five years, several influential studies have used genome re-sequencing to  
4  
5 417 advance our understanding of the genomic underpinnings of different biological questions in  
6  
7 418 model systems. For example, population genomics of *Heliconius* butterflies highlighted the  
8  
9 419 exchange of genes between species that exhibit convergent wing patterns (The *Heliconius*  
10  
11 420 Genome Consortium 2012); whole genome re-sequencing of three-spined sticklebacks  
12  
13 421 highlighted the re-use of alleles in replicated divergences associated with ecological speciation  
14  
15 422 and local adaptation (Jones et al. 2012); and combined population genomics and phylogenomics  
16  
17 423 have identified regions of the genome associated with variation in beak shape and size in  
18  
19 424 Darwin's finches (Lamichhaney et al. 2015).

20  
21 425  
22 426 To date only two marine mammal population genomics studies using whole genome re-  
23  
24 427 sequencing have been published. These studies involved re-sequencing the genomes of 79  
25  
26 428 individuals from three populations of polar bears (Liu et al. 2014a) and 48 individuals from five  
27  
28 429 evolutionarily divergent ecotypes of killer whale (Foote et al. 2016). The findings of Foote et al.  
29  
30 430 (2016) confirmed results of population differentiation that had previously been established using  
31  
32 431 traditional genetic markers (Morin et al. 2010a). However, the study also provided new insights  
33  
34 432 into the demographic history, patterns of selection associated with ecological niche, and evidence  
35  
36 433 of episodic ancestral admixture that could not have been obtained using traditional markers.

37 434  
38 435 Several new resources have made such population genomic studies economically possible for a  
39  
40 436 greater number of NMOs, including the availability of ~~a~~ reference genome assemblies (see  
41  
42 437 section above), relatively low-cost high-throughput sequencing (further increases in throughput  
43  
44 438 expected with the new Illumina HiSeq X Ten (van Dijk et al. 2014)), and crucially, the  
45  
46 439 development of likelihood-based methods that allow estimation of population genetic metrics  
47  
48 440 from ~~low-coverage~~ re-sequencing data (Fumagalli et al. 2013; O'Rawe et al. 2015). One last  
49  
50 441 consideration is the ease of laboratory methods necessary to generate whole genome re-  
51  
52 442 sequencing data when compared to other methods such as RADseq or TSC. DNA simply needs  
53  
54 443 to be extracted from the samples and, using proprietary kits, built into individually index-  
55  
56 444 amplified libraries ~~using proprietary kits, which that~~ are ~~then~~ equimolarly pooled and submitted  
57  
58 445 for sequencing.

59  
60 446

1  
2  
3 447 Many population genomic analyses are based on the coalescent model that gains most  
4  
5 448 information from the number of independent genetic markers, not the number of individuals  
6  
7 449 sampled. Sample sizes of ~10 individuals are usually considered sufficient (Robinson et al.  
8  
9 450 2014) and have been standard in many genome-wide studies in the eco-evolutionary sciences  
10  
11 451 (Ellegren et al. 2012; Jones et al. 2012). Thus, sampling fewer individuals ~~at lower coverage but~~  
12 452 ~~for orders of magnitude more data~~ by whole genome re-sequencing is a salient approach, ~~which~~  
13  
14 453 that allows us to consider many more gene trees, whilst continuing to provide robust estimates of  
15  
16 454 per-site genetic metrics (e.g.,  $F_{ST}$ ). The robustness of inference from ~~low coverage~~ data with low  
17 455 mean read depth across the genome was recently confirmed using a comparison of per-site  $F_{ST}$   
18 456 estimates for the same sites from high-~~coverage~~ depth ( $\geq 20x$ ) RADseq data and low-  
19 457 ~~coverage~~ depth ( $\approx 2x$ ) whole genome re-sequencing data in pairwise comparisons between the  
20  
21 458 same two killer whale ecotypes (Foote et al. 2016).

22  
23  
24  
25 459  
26 460 Beyond the increased power afforded by sequencing more polymorphic sites, whole genome re-  
27  
28 461 sequencing also allows inference of demographic history from the genome of even just a single  
29  
30 462 individual by identifying Identical By Descent (IBD) segments and runs of homozygosity (Li  
31  
32 463 and Durbin 2011; Harris and Nielsen 2013). For example, Liu et al. (2014a) found evidence for  
33  
34 464 ongoing gene flow from polar bears into brown bears after the two species initially diverged.  
35  
36 465 Genome re-sequencing of sufficient numbers of individuals also facilitates haplotype phasing,  
37  
38 466 which has many applications, including the detection of ongoing selective sweeps (Ferrer-  
39  
40 467 Admetlla et al. 2014) and the inference of demographic history of multiple populations based on  
41  
42 468 coalescence of pairs of haplotypes in different individuals (Schiffels and Durbin 2014).

43 469 However, haplotype phasing ~~has~~ typically requires ~~sd genomic higher coverage~~ data with higher  
44 470 mean read depth ( $\sim 20x$ ) from tens of individuals (though recent advances in genotype imputation  
45  
46 471 suggest success with ~~lower coverage~~ data of lower mean read depth (VanRaden et al. 2015)).  
47  
48 472 Thus far, phasing has been restricted to relatively few NMO studies, and no marine mammal  
49  
50 473 studies to the best of our knowledge.

#### 51 474 52 53 475 Transcriptome sequencing

54 476 In comparison with the DNA-based genomic approaches described above, RNA-based genomic  
55  
56 477 approaches are a relatively new and emerging application in NMOs such as marine mammals.

1  
2  
3 478 Transcriptomics by RNA sequencing (RNAseq) can rapidly generate vast amounts of  
4  
5 479 information regarding genes and gene expression without any prior genomic resources. This  
6  
7 480 approach can resolve differences in global gene expression patterns between populations,  
8  
9 481 individuals, tissues, cells, and physiological or environmental conditions, and can yield insights  
10  
11 482 into the molecular basis of environmental adaptation and speciation in wild animals (Wolf 2013;  
12  
13 483 Alvarez et al. 2015). Furthermore, RNAseq is a valuable tool for resource development, for  
14  
15 484 example as a precursor to designing SNP and TSC arrays (e.g., Hoffman et al. 2012). However,  
16  
17 485 applying RNAseq to NMOs requires several unique considerations in comparison to the DNA-  
18  
19 486 based methods described above. Most importantly, the labile nature of gene transcription and  
20  
21 487 high detection sensitivity of RNAseq have the potential to amplify transcriptional “noise” and  
22  
23 488 are thus extremely sensitive to experimental design.

24  
25 490 If the experimental goal is to capture a comprehensive transcriptome profile for a study  
26  
27 491 organism, multiple tissues from individuals of varied life history stages should be sampled.  
28  
29 492 However, if the aim is to characterize transcriptional responses to physiological or environmental  
30  
31 493 stimuli, efforts should focus on minimizing variability in individuals and sampling conditions  
32  
33 494 (Wolf 2013). For differential expression analyses, pairwise comparisons should be made within  
34  
35 495 the same individual if at all possible (e.g., before and after treatment, between two  
36  
37 496 developmental stages). As RNAseq only captures a ‘snapshot’ of gene expression in time,  
38  
39 497 repeated sampling or time-course studies are necessary to obtain a more complete picture of  
40  
41 498 cellular responses to the condition(s) in question (Spies and Ciaudo 2015). Sampling and  
42  
43 499 sequencing depth requirements will depend on the study design. Simulation studies have shown  
44  
45 500 that a minimum of 5-6 biological replicates sequenced at a depth of 10-20 million reads per  
46  
47 501 sample is necessary for differential expression analysis (Liu et al. 2014b; Schurch et al. 2015).  
48  
49 502 RNAseq can also be used for biomarker development to expand molecular toolkits for NMOs  
50  
51 503 without sequenced genomes (Hoffman et al. 2013). In this case, higher sequencing depths of 30-  
52  
53 504 60 million reads per sample are recommended for SNP discovery and genotyping (De Wit et al.  
54  
55 505 2015).

56  
57 506  
58  
59 507 Following sequence generation, transcript annotation remains a challenge for NMOs without  
60  
508 reference transcriptomes or genomes. *De novo* transcriptomes can be annotated through detection

1  
2  
3 509 of assembled orthologs of highly conserved proteins, but these analyses remain limited by the  
4  
5 510 quality of reference databases. As a result, NMO transcriptomes are biased in favor of highly  
6  
7 511 conserved terrestrial mammal genes and therefore provide an incomplete understanding of  
8  
9 512 animal adaptations to natural environments (Evans 2015). For example, while 70.0% of northern  
10  
11 513 elephant seal (*Mirounga angustirostris*) skeletal muscle transcripts had BLASTx hits to mouse  
12  
13 514 genes, only 54.1% of blubber transcripts could be annotated due to poor representation of this  
14  
15 515 tissue in terrestrial mammal reference proteomes (Khudyakov et al. 2015b).  
16

17  
18 517 To date, RNAseq has been used for gene discovery and phylogenomics analyses in Antarctic fur  
19  
20 518 seal (Hoffman 2011; Hoffman et al. 2013), polar bear (Miller et al. 2012), Indo-Pacific  
21  
22 519 humpback dolphin (*Sousa chinensis* (Gui et al. 2013)), spotted seal (*Phoca largha* (Gao et al.  
23  
24 520 2013)), bowhead whale (*Balaena mysticetus* (Seim et al. 2014)), narrow-ridged finless porpoise  
25  
26 521 (*Neophocaena asiaorientalis* (Ruan et al. 2015)), and humpback whale (*Megaptera*  
27 522 | *novaeangliae* (Tsagkogeorga et al. 2015)) (Table S1). Due to the challenges of repeated  
28  
29 523 sampling of wild marine mammals, few studies have examined cetacean or pinniped  
30  
31 524 transcriptome responses to environmental or experimental stimuli. The majority of such  
32  
33 525 functional gene expression studies have used microarrays (Mancia et al. 2008; Mancia et al.  
34  
35 526 2012; Mancia et al. 2015); however, RNAseq has been employed to profile sperm whale  
36  
37 527 (*Physeter macrocephalus*) skin cell response to hexavalent chromium (Pabuwal et al. 2013) and  
38  
39 528 free-ranging northern elephant seal skeletal muscle response to an acute stress challenge  
40  
41 530 (Khudyakov et al. 2015a; Khudyakov et al. 2015b). With decreasing sequencing costs and  
42  
43 531 improvements in bioinformatics tools, RNAseq has the potential to accelerate molecular  
44  
45 532 discoveries in marine mammal study systems and supplement existing functional genomics  
46  
47 533 approaches.

#### 48 534 Emerging techniques

49 535 In addition to the relatively proven NMO genomic data generation techniques described above, a  
50  
51 536 suite of emerging techniques is entering the field, with exciting promise for exploration of  
52  
53 537 existing and new research areas. For example, high-throughput shotgun sequencing is  
54  
55 538 increasingly being used to identify genetic material from multiple species in a single sample  
56  
57 539 (metagenomics and metatranscriptomics), rather than focus on characterizing variation in a  
58  
59  
60



1  
2  
3 540 single target individual. These multi-species approaches can be used, for example, to  
4  
5 541 characterize diet from fecal samples (Deagle et al. 2009) and to investigate microbiomes (Nelson  
6  
7 542 et al. 2015), objectives with implications for improving our understanding of both basic ecology  
8  
9 543 and health in natural populations of NMOs. Furthermore, high-throughput sequencing of  
10  
11 544 environmental DNA dramatically increases the throughput of NMO detection in environmental  
12  
13 545 (e.g., seawater) samples (Thomsen et al. 2012), using degenerate primers for multi-species  
14  
15 546 detection rather than requiring the design and implementation of numerous single-species  
16  
17 547 protocols (Foote et al. 2012).

18 548  
19 549 A second broad area of emerging interest moves beyond the study of variation at the DNA and  
20  
21 550 RNA levels to examine epigenetic effects of histone modification on gene regulation and  
22  
23 551 evolution. Epigenomic studies often examine changes in DNA methylation in association with  
24  
25 552 processes such as cancer and ageing. Such approaches, from targeted gene to genome-wide, have  
26  
27 553 only very recently and not yet frequently been applied in NMOs. Polanowski et al. (2014) used a  
28  
29 554 targeted gene approach to examine changes in DNA methylation in age-associated genes,  
30  
31 555 previously identified in humans and mice, in humpback whales of known age. The most  
32  
33 556 informative markers were able to estimate humpback whale ages with standard deviations of  
34  
35 557 approximately 3-5 years, demonstrating the potential transferability of these approaches from  
36  
37 558 model to non-model organism. Villar et al. (2015) utilized a genome-wide approach – chromatin  
38  
39 559 immunoprecipitation followed by high-throughput sequencing (ChIPseq) – to examine gene-  
40  
41 560 regulatory element evolution across mammals, including four species of cetaceans. This study  
42  
43 561 identified highly conserved gene-regulatory elements based on their histone modifications  
44  
45 562 (H3K27ac and H3K4me3), showed that recently evolved enhancers were associated with genes  
46  
47 563 under positive selection in marine mammals, and identified unique *Delphinus*-specific enhancers.  
48  
49 564 Finally, reduced-representation epigenomic approaches have also been developed (Gu et al.  
50  
51 565 2011), and although they have not yet been used in marine mammals to our knowledge, these  
52  
53 566 techniques could facilitate future studies of how changes in DNA methylation patterns affect  
54  
55 567 other biological processes, such as stress levels or pregnancy.

## 568 569 **Data analysis**

57  
58  
59  
60

1  
2  
3 570 Following the generation of genomic data, researchers must select the most appropriate genomic  
4  
5 571 analysis (i.e., bioinformatics) pipelines, which often differ significantly from those used in  
6  
7 572 traditional genetic studies of NMOs. The choice of analysis pipeline will depend on multiple  
8  
9 573 factors including the availability of a reference genome, the level of diversity within the dataset  
10  
11 574 (e.g., single- or multi-~~ple~~ species), the type of data generated (e.g., single-~~end vs. or~~ paired-  
12  
13 575 end), and the computing resources available. The computational needs, both in terms of hardware  
14  
15 576 and competency in computer science, for analysis of genomic data typically far exceed those  
16  
17 577 necessary for traditional genetic markers. On the smaller end of the spectrum, one lane of 50 bp  
18  
19 578 single-end sequencing on an Illumina HiSeq 2500 can produce tens of gigabytes of data, while  
20  
21 579 data files associated with a single high-~~coverage-quality~~ vertebrate genome may reach hundreds  
22  
23 580 of gigabytes in size (Ekblom and Wolf 2014). Computing resources necessary for the analysis of  
24  
25 581 these genomic datasets can range from ~10 gigabytes for a pilot study using a reduced-  
26  
27 582 representation sequencing approach to over a terabyte for whole-~~genome~~ sequence assembly  
28  
29 583 (Ekblom and Wolf 2014). Fortunately, university computing clusters, cloud-based (Stein 2010)  
30  
31 584 and high-performance computing clusters (e.g., XSEDE; Towns et al. 2014), and open web-  
32  
33 585 based platforms for genomic research (e.g., Galaxy; Goecks et al. 2010) are becoming  
34  
35 586 increasingly accessible. Furthermore, new pipelines are continuously being developed and  
36  
37 587 improved, and there are a growing number of resources aimed at training molecular ecologists  
38  
39 588 and evolutionary biologists in computational large-scale data analysis (Andrews and Luikart  
40  
41 589 2014; Belcaid and Toonen 2015; Benestan et al. 2016). We provide ~~a limited~~ an indicative list of  
42  
43 590 the current, most commonly used analysis pipelines that are specific to each data generation  
44  
45 591 method in ~~Supplemental~~ Table 12. Here, we briefly summarize current genomic data analysis  
46  
47 592 pipelines and discuss considerations that are likely to be similar across multiple data generation  
48  
49 593 methods.

50  
51 594  
52  
53 595 Genomic data analysis often involves multiple steps, and the choice of analysis tool for each step  
54  
55 596 can greatly affect the outcome, with different tools producing different (though usually  
56  
57 597 overlapping) sets of results (e.g., Schurch et al. 2015). All analyses begin by evaluating data  
58  
59 598 quality, trimming sequences if necessary to remove erroneous nucleotides (MacManes 2014),  
60  
61 599 and implementing appropriate data quality filters (e.g., phred scores, read length, and/or read  
62  
63 600 depth). Raw reads also need to be demultiplexed based on unique barcodes if pools of

1  
2  
3 601 individuals were sequenced in a single lane. Analyses then usually proceed in a *de novo* or  
4  
5 602 genome-enabled manner, depending on available resources. Briefly, sequences can be compared  
6  
7 603 (e.g., to identify variants) by mapping all reads to a reference genome or *de novo* assembling  
8  
9 604 stacks of sequences putatively derived from the same locus; based on sequence similarity. *De*  
10  
11 605 *novo* methods are sensitive to sequencing error, as well as true genetic variation, and therefore  
12  
13 606 can erroneously assemble polymorphic sequences as separate loci or transcripts, requiring further  
14  
15 607 filtering to remove redundancy. The opposite problem can also occur in both *de novo* and  
16  
17 608 reference mapping approaches, where two distinct loci (e.g., paralogous loci) may assemble as a  
18  
19 609 single locus or map to the same reference location. Researchers should therefore recognize the  
20  
21 610 inherent trade-offs when carefully selecting their thresholds for acceptable levels of variation  
22  
23 611 within and among loci.

24  
25 613 Considerations relevant to the selection of subsequent downstream analyses are specific to the  
26  
27 614 type of data generated and the research objective. For example, RADseq analysis pipelines differ  
28  
29 615 in the algorithms used to genotype variants (Table 1S2). Similarly, there are several gene  
30  
31 616 expression analysis pipelines for RNAseq data that compare transcript abundance between  
32  
33 617 samples (Table 1S2). Analysis of TSC data usually uses standard *de novo* assemblers (e.g.,  
34  
35 618 Trinity, Velvet); these assemblers can be run using packages such as PHYLUCE (Faircloth  
36  
37 619 2015), which is designed specifically for use with ultraconserved elements. Unfortunately, for  
38  
39 620 most analyses, there are no unifying recommendations currently available and researchers must  
40  
41 621 evaluate several approaches, each with their own advantages and disadvantages, in order to  
42  
43 622 select the most appropriate tool for their particular experiment and system. Furthermore, we can  
44  
45 623 expect that the recommendations for analysis tools will continue to evolve as new programs  
46  
47 624 become available in the future.

#### 625 626 Guidelines for data quality control and sharing

627 With rapid growth in sequencing platforms and bioinformatics analysis pipelines comes the need  
628  
629 to extend existing principles (e.g., Bonin et al. 2004) on quality control, analysis, and  
630  
631 transparency. General recommendations for sample and data handling, library preparation, and  
sequencing have been discussed elsewhere (Paszkiwicz et al. 2014). We therefore focus on the  
need to produce guidelines on data quality evaluation and reporting for genomic data (e.g.,

1  
2  
3 632 Morin et al. 2010b). A primary challenge in this area is that quality metrics vary widely across  
4  
5 633 sequencing technologies. Yet, regardless of sequencing platform, the quality of sequencing reads  
6  
7 634 must be evaluated (e.g., using FastQC; Andrews 2010) and reported.  
8  
9 635

10 636 | Best practices guidelines for ~~high-coverage-wholereference~~ genome sequencing and RNAseq  
11  
12 637 | data generation, analysis, and reporting are available from the human-centric ENCODE  
13  
14 638 consortium ([www.encodeproject.org](http://www.encodeproject.org)). These include minimum depth of sequencing and number  
15  
16 639 and reproducibility of biological replicates. For RNAseq experiments, evaluation of *de novo*  
17  
18 640 assembly quality remains a challenge. Suggested quality metrics include percentage of raw reads  
19  
20 641 mapping back to the assembly and number of assembled transcripts with homology to known  
21  
22 642 proteins (MacManes 2016). Emerging tools such as Transrate (Smith-Unna et al. 2015) attempt  
23  
24 643 to integrate these and other metrics into a comprehensive assembly quality score.  
25  
26 644

27 645 In contrast, there is not yet any standard way to estimate or report error rates with RADseq or  
28  
29 646 | ~~low-coverage~~ genome re-sequencing methods (but see Mastretta-Yanes et al. 2015; Fountain et  
30  
31 647 al. 2016). Recommendations to improve confidence in genotyping include using methods that  
32  
33 648 | account for population--level allele frequencies when calling individual genotypes, mapping  
34  
35 649 reads to reference genomes rather than *de novo* assembly (Nadeau et al. 2014; Fountain et al.  
36  
37 650 2016), filtering out PCR duplicates (Andrews et al. 2014), identifying and removing markers in  
38  
39 651 possible repeat regions, and filtering data to include only those with high read depth (>10-20x  
40  
41 652 per locus per individual) (Nielsen et al. 2011). Other analysis methods, such as robust Bayesian  
42  
43 653 | methods and likelihood--based approaches that account for read quality in calculations of  
44  
45 654 posterior probabilities of genotypes and per-site allele frequencies utilizing the sample mean site  
46  
47 655 frequency spectrum as a prior (Fumagalli et al. 2013), can account for uncertainty and/or error in  
48  
49 656 the data, and are therefore suitable for use with low to moderate read depths (2-20x per locus;  
50  
51 657 e.g., Han et al. 2015; O'Rawe et al. 2015).  
52  
53 658

54 659 Due to the large number of analysis tools that are available, data quality and reproducibility  
55  
56 660 ultimately depend on methods and data transparency. All raw sequencing reads should be  
57  
58 661 | ~~publicly archived, for example~~ deposited in the NCBI Sequence Read Archive. Many journals,  
59  
60 662 including the *Journal of Heredity* (Baker 2013), now also require that primary data supporting

1  
2  
3 663 the published results and conclusions (e.g., SNP genotypes, assemblies) be publicly archived in  
4  
5 664 online data repositories (e.g., Dryad). We further recommend making public the analysis  
6  
7 665 pipelines, scripts (e.g., using GitHub), and additional outputs, as appropriate, in order for  
8  
9 666 analyses to be fully reproducible and transparent, which is the cornerstone of the scientific  
10  
11 667 method (Nosek et al. 2015).

12 668

### 14 669 **Future directions**

15 670 As demonstrated here for one group of mammalian taxa, the rapid growth of the field of non-  
16  
17 671 model genomics has been both impressive and empowering. As we approach a point of relative  
18  
19 672 saturation in reference genomes, we anticipate an increase in population-scale genomic studies  
20  
21 673 that produce lower depth or coverage datasets per individual but across larger sample sizes  
22  
23 674 relative to high coverage sequencing of a few individuals of each species. In addition (or  
24  
25 675 alternatively), we hope to see increasing efforts to sequence reference transcriptomes and  
26  
27 676 improve NMO genome annotation in ways beyond the inherently limited approach of  
28  
29 677 comparison to gene lists from a few model organisms. Population-scale genomic studies will  
30  
31 678 facilitate greater ecological understanding of natural populations, while efforts to improve  
32  
33 679 annotation will address persistent limitations in our understanding of gene function for NMOs.  
34  
35 680 Ultimately, improving our understanding of local adaptation, adaptive potential, and  
36  
37 681 demographic history through the use of genomic toolkits such as those described here is likely to  
38  
39 682 have important implications for the future conservation of these populations.

40 683

41 684 Advances in sequencing technologies and analytical tools will no doubt continue, in some cases  
42  
43 685 drawing on established techniques in model organisms, posing both new opportunities and new  
44  
45 686 challenges for researchers in NMO genomics. Likely the most persistent challenge will remain  
46  
47 687 selecting the data generation and experimental design that is most appropriate for the respective  
48  
49 688 research objective. Our review identified few cases that exhibit relative dominance of a single  
50  
51 689 methodology and analytical pipeline (e.g., RADseq and STACKS, RNAseq and Trinity); rather,  
52  
53 690 more often we found a diversity of approaches even within each category of data generation. In  
54  
55 691 fact, such diversity of approaches has its benefits, with each approach promoting its own  
56  
57 692 advantages (and limitations). Overall, our reflections on lessons learned from the past decade of  
58  
59 693 NMO genomics in one well-studied group of mammalian taxa clearly demonstrate the value,

1  
2  
3 694 increased ease, and future promise of applying genomic techniques across a wide range of non-  
4  
5 695 model species to gain previously unavailable insights into evolution, population biology, and  
6  
7 696 physiology on a genome-wide scale.  
8

9 697

## 10 698 **Acknowledgements**

11 699 This review paper is the outcome of two international workshops held in 2013 and 2015 on  
12  
13 700 marine mammal genomics. The workshops were organized by KMC, AF, and C. Scott Baker and  
14  
15 701 hosted by the Society for Marine Mammalogy, [with support from a Special Event Award from](#)  
16  
17 702 [the American Genetic Association](#). We sincerely thank all the workshop participants for their  
18  
19 703 contributions to inspiring discussions on marine mammal genomics. We would also like to thank  
20  
21 704 two anonymous reviewers and C. Scott Baker for their helpful feedback on an earlier version of  
22  
23 705 this manuscript. Illustrations are by C. Buell with permission for use granted by J. Gatesy.  
24

25 706

## 26 707 **Funding**

27  
28 708 The authors involved in this work were supported by a National Science Foundation Postdoctoral  
29  
30 709 Research Fellowship in Biology (Grant No. 1523568) to KMC; an Office of Naval Research  
31  
32 710 Award (No. N00014-15-1-2773) to JIK; a Marie Skłodowska Curie Fellowship to ELC  
33  
34 711 (Behaviour-Connect) funded by the EU Horizon2020 program; Royal Society Newton  
35  
36 712 International Fellowships to ELC and MRM; a Deutsche Forschungsgemeinschaft studentship to  
37  
38 713 EH; a Fyssen Foundation postdoctoral fellowship to ML; postdoctoral funding from the  
39  
40 714 University of Idaho College of Natural Resources to KRA; a short visit grant from the European  
41  
42 715 Science Foundation-Research Networking Programme ConGenOmics to ADF; and a Swiss  
43  
44 716 National Science Foundation grant (31003A-143393) to L. Excoffier that further supported ADF.  
45  
46 717 The first marine mammal genomics workshop we held to begin discussions towards this review  
47  
48 718 was supported by a Special Event Award from the American Genetic Association.  
49

50 719

## 51 720 **References**

52 721 Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures  
53 722 of population divergence. *Mol Biol Evol.* 27:2534-2547.  
54 723 Alexander A, Steel D, Hoekzema K, Mesnick S, Engelhaupt D, Kerr I, Payne R, Baker CS.  
55 724 2016. What influences the worldwide genetic structure of sperm whales (*Physeter*  
56 725 *macrocephalus*)? *Mol Ecol.*  
57  
58  
59  
60

- 1  
2  
3 726 Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB,  
4 727 Schroeder H, Ahlstrom T, Vinner L, *et al.* 2015. Population genomics of Bronze Age  
5 728 Eurasia. *Nature*. 522:167-172.
- 7 729 Alvarez M, Schrey AW, Richards CL. 2015. Ten years of transcriptomics in wild populations:  
8 730 what have we learned about their ecology and evolution? *Mol Ecol*. 24:710-725.
- 9 731 Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome*  
10 732 *Biol*. 11:R106.
- 11 733 Andrews K, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of  
12 734 RADseq for ecological and evolutionary genomics. *Nat Rev Genet*. 17:81-92.
- 13 735 Andrews KR, Hohenlohe PA, Miller MR, Hand BK, Seeb JE, Luikart G. 2014. Trade-offs and  
14 736 utility of alternative RADseq methods: Reply to Puritz *et al.* 2014. *Mol Ecol*. 23:5943-  
15 737 5946.
- 17 738 Andrews KR, Luikart G. 2014. Recent novel approaches for population genomics data analysis.  
18 739 *Mol Ecol*. 23:1661-1667.
- 20 740 Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available  
21 741 online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- 22 742 Ankeny RA, Leonelli S. 2011. What's so special about model organisms? *Studies in History and*  
23 743 *Philosophy of Science*. 42:313-323.
- 24 744 Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM. 2014. Non-model  
25 745 organisms, a species endangered by proteogenomics. *J Proteomics*. 105:5-18.
- 27 746 Arnason U, Adegoke JA, Bodin K, Born EW, Esa YB, Gullberg A, Nilsson M, Short RV, Xu X,  
28 747 Janke A. 2002. Mammalian mitogenomic relationships and the root of the eutherian tree.  
29 748 *Proc Natl Acad Sci USA*. 99:8151-8156.
- 30 749 Arnason U, Gullberg A, Widegren B. 1991. The complete nucleotide sequence of the  
31 750 mitochondrial DNA of the fin whale, *Balaenoptera physalus*. *J Mol Evol*. 33:556-568.
- 33 751 Ávila-Arcos M, Cappellini E, Romero-Navarro JA, Wales N, Moreno-Mayar JV, Rasmussen M,  
34 752 Fordyce SL, Montiel R, Vielle-Calzada J-P, Willerslev E, *et al.* 2011. Application and  
35 753 comparison of large-scale solution-based DNA capture-enrichment methods on ancient  
36 754 DNA. *Sci Rep*. 1:74.
- 37 755 Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA,  
38 756 Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD  
39 757 markers. *PLoS One*. 3:e3376.
- 41 758 Baker CS. 2013. *Journal of Heredity* adopts Joint Data Archiving Policy. *J Hered*. 104:1.
- 42 759 Barrett RDH, Rogers SM, Schluter D. 2008. Natural selection on a major armor gene in  
43 760 threespine stickleback. *Science*. 322:255-257.
- 44 761 Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. 2005. Direct genomic  
45 762 selection. *Nat Methods*. 2:63-69.
- 47 763 Belcaid M, Toonen RJ. 2015. Demystifying computer science for molecular ecologists. *Mol*  
48 764 *Ecol*. 24:2619-2640.
- 49 765 Benestan LM, Ferchaud A-L, Hohenlohe PA, Garner BA, Naylor GJP, Baums IB, Schwartz MK,  
50 766 Kelley JL, Luikart G. 2016. Conservation genomics of natural and managed populations:  
51 767 building a conceptual and practical framework. *Mol Ecol*.
- 52 768 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina Sequence  
53 769 Data. *Bioinformatics*. 30:2114-2120.
- 54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 770 Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C, Taberlet P. 2004. How  
4 771 to track and assess genotyping errors in population genetics studies. *Mol Ecol.* 13:3261-  
5 772 3273.
- 7 773 Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. 2012. A reference-free algorithm for  
8 774 computational normalization of shotgun sequencing data. *arXive.* 1203:4802.
- 9 775 Cammen KM, Schultz TF, Rosel PE, Wells RS, Read AJ. 2015. Genomewide investigation of  
10 776 adaptation to harmful algal blooms in common bottlenose dolphins (*Tursiops truncatus*).  
11 777 *Mol Ecol.* 24:4697-4710.
- 13 778 Campbell NR, Harmon SA, Narum SR. 2015. Genotyping-in-Thousands by sequencing (GT-  
14 779 seq): a cost effective SNP genotyping method based on custom amplicon sequencing.  
15 780 *Mol Ecol Resour.* 15:855-867.
- 16 781 Carroll EL, Baker CS, Watson M, Alderman R, Bannister J, Gaggiotti OE, Gröcke DR,  
17 782 Patenaude N, Harcourt R. 2015. Cultural traditions across a migratory network shape the  
18 783 genetic structure of southern right whales around Australia and New Zealand. *Sci Rep.*  
20 784 5:16182.
- 21 785 Catchen JM, Amores A, Hohenlohe PA, Cresko WA, Postlethwait JH. 2011. *Stacks*: building  
22 786 and genotyping loci *de novo* from short-read sequences. *G3.* 1:171-182.
- 23 787 Catchen JM, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. *Stacks*: an analysis tool  
24 788 set for population genomics. *Mol Ecol.* 22:3124-2140.
- 26 789 Chancerel E, Lepoittevin C, Le Provost G, Lin Y-C, Jaramillo-Correa JP, Eckert AJ, Wegrzyn  
27 790 JL, Zelenika D, Boland A, Frigerio J-M, *et al.* 2011. Development and implementation of  
28 791 a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative  
29 792 mapping with loblolly pine. *BMC Genomics.* 12:368.
- 31 793 Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps.  
32 794 *Genome Res.* 20:393-402.
- 33 795 Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Grimwood J, Schmutz J, Myers RM,  
34 796 Schluter D, Kingsley DM. 2005. Widespread parallel evolution in sticklebacks by  
35 797 repeated fixation of ectodysplasin alleles. *Science.* 307:1928-1933.
- 36 798 Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal  
37 799 tool for annotation, visualization and analysis in functional genomics research.  
38 800 *Bioinformatics.* 21:3674-3676.
- 40 801 Corander J, Majander KK, Cheng L, Merilä J. 2013. High degree of cryptic population  
41 802 differentiation in the Baltic Sea herring *Clupea harengus*. *Mol Ecol.* 22:2931-2940.
- 42 803 Cummings N, King R, Rickers A, Kaspi A, Lunke S, Haviv I, Jowett JBM. 2010. Combining  
43 804 target enrichment with barcode multiplexing for high throughput SNP discovery. *BMC*  
44 805 *Genomics.* 11:641.
- 46 806 Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide  
47 807 genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev*  
48 808 *Genet.* 12:499-510.
- 49 809 De Mita S, Thuillet A-C, Gay L, Ahmadi N, Manel S, Ronfort J, Vigouroux Y. 2013. Detecting  
50 810 selection along environmental gradients: analysis of eight methods and their effectiveness  
51 811 for outbreeding and selfing populations. *Mol Ecol.* 22:1383-1399.
- 53 812 De Wit P, Pespeni MH, Palumbi SR. 2015. SNP genotyping and population genomics from  
54 813 expressed sequences - current advances and future possibilities. *Mol Ecol.* 24:2310-2323.
- 55 814 Deagle BE, Kirkwood R, Jarman SN. 2009. Analysis of Australian fur seal diet by  
56 815 pyrosequencing prey DNA in faeces. *Mol Ecol.* 18:2022-2038.



- 1  
2  
3 816 Deméré TA, McGowen MR, Berta A, Gatesy J. 2008. Morphological and molecular evidence for  
4 817 a stepwise evolutionary transition from teeth to baleen in mysticete whales. *Syst Biol.*  
5 818 57:15-37.
- 7 819 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del  
8 820 Angel G, Rivas MA, Hanna M, *et al.* 2011. A framework for variation discovery and  
9 821 genotyping using next-generation DNA sequencing data. *Nat Genet.* 43:491-498.
- 10 822 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras  
11 823 TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 29:15-21.
- 12 824 Eaton DAR. 2014. PyRAD: assembly of *de novo* RADseq loci for phylogenetic analysis.  
13 825 *Bioinformatics.* 30:1844-1849.
- 15 826 Ekblom R, Galindo J. 2011. Applications of next generation sequencing in molecular ecology of  
16 827 non-model organisms. *Heredity.* 107:1-15.
- 17 828 Ekblom R, Wolf JBW. 2014. A field guide to whole-genome sequencing, assembly and  
18 829 annotation. *Evolutionary Applications.* 7:1026-1042.
- 20 830 Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms.  
21 831 *Trends Ecol Evol.* 29:51-63.
- 22 832 Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Künstner A, Mäkinen H,  
23 833 Nadachowska-Brzyska K, Qvarnström A, *et al.* 2012. The genomic landscape of species  
24 834 divergence in *Ficedula* flycatchers. *Nature.* 491:756-760.
- 26 835 Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A  
27 836 robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.  
28 837 *PLoS One.* 6:e19379.
- 29 838 Enk J, Devault A, Kuch M, Murgha Y, Rouillard J-M, Poinar H. 2014. Ancient whole genome  
30 839 enrichment using baits built from modern DNA. *Mol Biol Evol.* 31:1292-1294.
- 32 840 Evans TG. 2015. Considerations for the use of transcriptomics in identifying the 'genes that  
33 841 matter' for environmental adaptation. *J Exp Biol.* 218:1925-1935.
- 34 842 Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic  
35 843 inference from genomic and SNP data. *PLoS Genetics.* 9:e1003905.
- 36 844 Faircloth BC. 2015. PHYLUCE is a software package for the analysis of conserved genomic  
37 845 loci. *Bioinformatics.* 32:786-788.
- 39 846 Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012.  
40 847 Ultraconserved elements anchor thousands of genetic markers spanning multiple  
41 848 evolutionary timescales. *Syst Biol.* 61:717-726.
- 42 849 Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On detecting incomplete soft or  
43 850 hard selective sweeps using haplotype structure. *Mol Biol Evol.* 31:1275-1291.
- 45 851 Flicek P, Birney E. 2009. Sense from sequence reads: methods for alignment and assembly. *Nat*  
46 852 *Methods.* 6:S6-S12.
- 47 853 Foote AD, Liu Y, Thomas GWC, Vinař Ts, Alföldi J, Deng J, Dugan S, van Elk CE, Hunter ME,  
48 854 Joshi V, *et al.* 2015. Convergent evolution of the genomes of marine mammals. *Nat*  
49 855 *Genet.* 47:272-275.
- 51 856 Foote AD, Newton J, Ávila-Arcos MC, Kampmann M-L, Samaniego JA, Post K, Rosing-Asvid  
52 857 A, Sinding M-HS, Gilbert MTP. 2013. Tracking niche variation over millennial  
53 858 timescales in sympatric killer whale lineages. *Proc R Soc Lond B Biol Sci.* 280:20131481.
- 54 859 Foote AD, Thomsen PF, Sveegaard S, Wahlberg M, Kielgast J, Kyhn LA, Salling AB, Galatius  
55 860 A, Orlando L, Gilbert MTP. 2012. Investigating the potential use of environmental DNA  
56 861 (eDNA) for genetic monitoring of marine mammals. *PLoS One.* 7:e41781.

- 1  
2  
3 862 Foote AD, Vijay N, Ávila-Arcos M, Baird RW, Durban JW, Fumagalli M, Gibbs RA, Hanson  
4 863 MB, Korneliussen TS, Martin MD, *et al.* 2016. Genome-culture coevolution promotes  
5 864 rapid divergence of killer whale ecotypes. *Nat Commun.* 7:11693.
- 7 865 Fountain ED, Pauli JN, Reid BN, Palsbøll PJ, Peery MZ. 2016. Finding the right coverage: the  
8 866 impact of coverage and sequence quality on single nucleotide polymorphism genotyping  
9 867 error rates. *Mol Ecol Resour.*
- 10 868 Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A,  
11 869 Nielsen R. 2013. Quantifying population genetic differentiation from next-generation  
12 870 sequencing data. *Genetics.* 195:979-992.
- 14 871 Fumagalli M, Vieira FG, Linderoth T, Nielsen R. 2014. *ngsTools*: methods for population  
15 872 genetics analyses from Next-Generation Sequencing data. *Bioinformatics.* 30:1486-1487.
- 16 873 Gao X, Han J, Lu Z, Li Y, He C. 2013. *De novo* assembly and characterization of spotted seal  
17 874 *Phoca largha* transcriptome using Illumina paired-end sequencing. *Comp Biochem*  
18 875 *Physiol D Genom Proteom.* 8:103-110.
- 20 876 Garner BA, Hand BK, Amish SJ, Bernatchez L, Foster JT, Miller KM, Morin PA, Narum SR,  
21 877 O'Brien SJ, Roffler G, *et al.* 2016. Genomics in conservation: case studies and bridging  
22 878 the gap between data and application. *Trends Ecol Evol.* 31:81-83.
- 23 879 Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. 2014. TASSEL-  
24 880 GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One.* 9:e90346.
- 26 881 Gnerre S, MacCallum I, Przbylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea  
27 882 TP, Sykes S, *et al.* 2011. High-quality draft assemblies of mammalian genomes from  
28 883 massively parallel sequence data. *Proc Natl Acad Sci USA.* 108:1513-1518.
- 29 884 Goecks J, Nekrutenko A, Taylor J, The Galaxy Team. 2010. Galaxy: a comprehensive approach  
30 885 for supporting accessible, reproducible, and transparent computational research in the life  
31 886 sciences. *Genome Biol.* 11:R86.
- 33 887 Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. 2011. Preparation of reduced  
34 888 representation bisulfite sequencing libraries for genome-scale DNA methylation  
35 889 profiling. *Nat Protoc.* 6:468-481.
- 36 890 Gui D, Jia K, Xia J, Yang L, Chen J, Wu Y, Yi M. 2013. *De novo* assembly of the Indo-Pacific  
37 891 humpback dolphin leucocyte transcriptome to identify putative genes involved in the  
38 892 aquatic adaptation and immune response. *PLoS One.* 8:e72417.
- 40 893 Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint  
41 894 demographic history of multiple populations from multidimensional SNP frequency data.  
42 895 *PLoS Genetics.* 5:e1000695.
- 43 896 Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D,  
44 897 Li B, Lieber M, *et al.* 2013. *De novo* transcript sequence reconstruction from RNA-seq  
45 898 using the Trinity platform for reference generation and analysis. *Nat Protoc.* 8:1494-  
46 899 1512.
- 48 900 Han E, Sinsheimer JS, Novembre J. 2015. Fast and accurate site frequency spectrum estimation  
49 901 from low coverage sequence data. *Bioinformatics.* 31:720-727.
- 50 902 Hancock-Hanser BL, Frey A, Leslie MS, Dutton PH, Archer FI, Morin PA. 2013. Targeted  
51 903 multiplex next-generation sequencing: advances in techniques of mitochondrial and  
52 904 nuclear DNA sequencing for population genomics. *Mol Ecol Resour.* 13:254-268.
- 54 905 Harris K, Nielsen R. 2013. Inferring demographic history from a spectrum of shared haplotype  
55 906 lengths. *PLoS Genetics.* 9:e1003521.

- 1  
2  
3 907 Hayden S, Bekaert M, Crider TA, Mariani S, Murphy WJ, Teeling EC. 2010. Ecological  
4 908 adaptation determines functional mammalian olfactory subgenomes. *Genome Res.* 20:1-  
5 909 9.
- 6  
7 910 Hedrick PW. 2000 *Genetics of Populations*. Jones and Bartlett Publishers, Sudbury, MA.
- 8 911 Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, Cariani A,  
9 912 Maes GE, Diopere E, Carvalho GR, *et al.* 2011. Application of SNPs for population  
10 913 genetics of nonmodel organisms: new opportunities and challenges. *Mol Ecol Resour.*  
11 914 11:123-136.
- 12  
13 915 Higdon JW, Bininda-Emonds ORP, Beck RMD, Ferguson SH. 2007. Phylogeny and divergence  
14 916 of the pinnipeds (Carnivora: Mammalia) assessed using a multigene dataset. *BMC Evol*  
15 917 *Biol.* 7:216.
- 16  
17 918 Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Gordon DB, Brizuela L, McCombie WR,  
18 919 Hannon GJ. 2009. Hybrid selection of discrete genomic intervals on custom-designed  
19 920 microarrays for massively parallel sequencing. *Nat Protoc.* 4:960-974.
- 20 921 Hoffman JI. 2011. Gene discovery in the Antarctic fur seal (*Arctocephalus gazella*) skin  
21 922 transcriptome. *Mol Ecol Resour.* 11:703-710.
- 22 923 Hoffman JI, Nicholas HJ. 2011. A novel approach for mining polymorphic microsatellite  
23 924 markers *in silico*. *PLoS One.* 6:e23283.
- 24 925 Hoffman JI, Simpson F, David P, Rijks JM, Kuiken T, Thorne MAS, Lacy RC, Dasmahapatra  
25 926 KK. 2014. High-throughput sequencing reveals inbreeding depression in a natural  
26 927 population. *Proc Natl Acad Sci USA.* 111:3775-3780.
- 27 928 Hoffman JI, Thorne MAS, Trathan PN, Forcada J. 2013. Transcriptome of the dead:  
28 929 characterisation of immune genes and marker development from necropsy samples in a  
29 930 free-ranging marine mammal. *BMC Genomics.* 14:52.
- 30 931 Hoffman JI, Tucker R, Bridgett SJ, Clark MS, Forcada J, Slate J. 2012. Rates of assay success  
31 932 and genotyping error when single nucleotide polymorphism genotyping in non-model  
32 933 organisms: a case study in the Antarctic fur seal. *Mol Ecol Resour.* 12:861-872.
- 33 934 Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population  
34 935 genomics of parallel adaptation in threespine stickleback using sequenced RAD tags.  
35 936 *PLoS Genet.* 6:e1000862.
- 36 937 Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management  
37 938 tool for second-generation genome projects. *BMC Bioinformatics.* 12:491.
- 38 939 Humble E, Martinez-Barrio A, Forcada J, Trathan PN, Thorne MAS, Hoffmann M, Wolf JBW,  
39 940 Hoffman JI. 2016. A draft fur seal genome provides insights into factors affecting SNP  
40 941 validation and how to mitigate them. *Mol Ecol Resour.*
- 41 942 Jackson JA, Baker CS, Vant M, Steel DJ, Medrano-González L, Palumbi SR. 2009. Big and  
42 943 slow: phylogenetic estimates of molecular evolution in baleen whales (suborder  
43 944 Mysticeti). *Mol Biol Evol.* 26:2427-2440.
- 44 945 Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody  
45 946 MC, White S, *et al.* 2012. The genomic basis of adaptive evolution in threespine  
46 947 sticklebacks. *Nature.* 484:55-61.
- 47 948 Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M,  
48 949 Nagayasu E, Maruyama H, *et al.* 2014. Efficient *de novo* assembly of highly  
49 950 heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24:1384-  
50 951 1395.
- 51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 952 Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, Madsen LB, van Dam S, Brawand  
4 953 D, Marques PI, *et al.* 2015. Insights into the evolution of longevity from the bowhead  
5 954 whale genome. *Cell Reports*. 10:112-122.
- 7 955 Khudyakov JI, Champagne CD, Preeyanon L, Ortiz RM, Crocker DE. 2015a. Muscle  
8 956 transcriptome response to ACTH administration in a free-ranging marine mammal.  
9 957 *Physiol Genomics*. 47:318-330.
- 10 958 Khudyakov JI, Preeyanon L, Champagne CD, Ortiz RM, Crocker DE. 2015b. Transcriptome  
11 959 analysis of northern elephant seal (*Mirounga angustirostris*) muscle tissue provides a  
12 960 novel molecular resource and physiological insights. *BMC Genomics*. 16:64.
- 14 961 Kishida T, Thewissen JGM, Hayakawa T, Imai H, Agata K. 2015. Aquatic adaptation and the  
15 962 evolution of smell and taste in whales. *Zoolog Lett*. 1:9.
- 16 963 Koepfli K-P, Paten B, Genome 10K Community of Scientists, O'Brien SJ. 2015. The Genome  
17 964 10K Project: a way forward. *Annu Rev Anim Biosci*. 3:57-111.
- 18 965 Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation  
19 966 Sequencing Data. *BMC Bioinformatics*. 15:356.
- 21 967 Künstner A, Wolf JBW, Backström N, Whitney O, Balakrishnan CN, Day L, Edwards SV, Janes  
22 968 DE, Schlinger BA, Wilson RK, *et al.* 2010. Comparative genomics based on massive  
23 969 parallel transcriptome sequencing reveals patterns of substitution and selection across 10  
24 970 bird species. *Mol Ecol*. 19:266-276.
- 26 971 Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A,  
27 972 Promerová M, Rubin C-J, Wang C, Zamani N, *et al.* 2015. Evolution of Darwin's finches  
28 973 and their beaks revealed by genome sequencing. *Nature*. 518:371-375.
- 29 974 Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment  
30 975 of short DNA sequences to the human genome. *Genome Biol*. 10:R25.
- 32 976 Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-  
33 977 throughput phylogenomics. *Syst Biol*. 61:727-744.
- 34 978 Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or  
35 979 without a reference genome. *BMC Bioinformatics*. 12:323.
- 36 980 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.  
37 981 *Bioinformatics*. 25:1754-1760.
- 39 982 Li H, Durbin R. 2011. Inference of human population history from individual whole-genome  
40 983 sequences. *Nature*. 475:493-496.
- 41 984 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,  
42 985 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map  
43 986 form and SAMtools. *Bioinformatics*. 25:2078-2079.
- 44 987 Li S, Jakobsson M. 2012. Estimating demographic parameters from large-scale population  
45 988 genomic data using Approximate Bayesian Computation. *BMC Genet*. 13:22.
- 47 989 Li Y, Hu Y, Bolund L, Wang J. 2010. State of the art *de novo* assembly of human genomes from  
48 990 massively parallel sequencing data. *Human Genomics* 4:271-277.
- 49 991 Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J,  
50 992 Jordan G, Mauceli E, *et al.* 2011. A high-resolution map of human evolutionary  
51 993 constraint using 29 mammals. *Nature*. 478:476-482.
- 53 994 Lindqvist C, Schuster SC, San Y, Talbot SL, Qi J, Ratan A, Tomsho LP, Kasson L, Zeyl E, Aars  
54 995 J, *et al.* 2010. Complete mitochondrial genome of a Pleistocene jawbown unveils the  
55 996 origin of polar bear. *Proc Natl Acad Sci USA*. 107:5053-5057.

- 1  
2  
3 997 Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, Xiong Z, Zhou L, Korneliussen TS, Somel M,  
4 998 Babbitt C, *et al.* 2014a. Population genomics reveal recent speciation and rapid  
5 999 evolutionary adaptation in polar bears. *Cell*. 157:785-794.
- 7 1000 Liu X, Fun Y-X. 2015. Exploring population size changes using SNP frequency spectra. *Nat*  
8 1001 *Genet*. 47:555-559.
- 9 1002 Liu Y, Zhou J, White KP. 2014b. RNA-seq differential expression studies: more sequence or  
10 1003 more replication? *Bioinformatics*. 30:301-304.
- 12 1004 Lotterhos KE, Whitlock MC. 2014. Evaluation of demographic history and neutral  
13 1005 parameterization on the performance of  $F_{ST}$  outlier tests. *Mol Ecol*. 23:2178-2192.
- 14 1006 Louis M, Viricel A, Lucas T, Peltier H, Alfonsi E, Berrow S, Brownlow A, Covelo P, Dabin W,  
15 1007 Deaville R, *et al.* 2014. Habitat-driven population structure of bottlenose dolphins,  
16 1008 *Tursiops truncatus*, in the North-east Atlantic. *Mol Ecol*. 23:857-874.
- 17 1009 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for  
18 1010 RNA-seq data with DESeq2. *Genome Biol*. 15:550.
- 20 1011 MacManes MD. 2014. On the optimal trimming of high-throughput mRNA sequence data. *Front*  
21 1012 *Genet*. 5:13.
- 22 1013 MacManes MD. 2016. Establishing evidence-based best practice for the *de novo* assembly and  
23 1014 evaluation of transcriptomes from non-model organisms. *bioRxiv*. doi:  
24 1015 <http://dx.doi.org/10.1101/035642>.
- 26 1016 Magera AM, Mills Flemming JE, Kaschner K, Christensen LB, Lotze HK. 2013. Recovery  
27 1017 trends in marine mammal populations. *PLoS One*. 8:e77908.
- 28 1018 Malenfant RM, Coltman DW, Davis CS. 2015. Design of a 9K Illumina BeadChip for polar  
29 1019 bears (*Ursus maritimus*) from RAD and transcriptome sequencing. *Mol Ecol Resour*.  
30 1020 15:587-600.
- 32 1021 Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J,  
33 1022 Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. *Nat*  
34 1023 *Methods*. 7:111-118.
- 35 1024 Mancía A, Abelli L, Kucklick JR, Rowles TK, Wells RS, Balmer BC, Hohn AA, Baatz JE, Ryan  
36 1025 JC. 2015. Microarray applications to understand the impact of exposure to environmental  
37 1026 contaminants in wild dolphins (*Tursiops truncatus*). *Mar Genomics*. 19:47-57.
- 39 1027 Mancía A, Lundqvist ML, Romano TA, Peden-Adams MM, Fair PA, Kindy MS, Ellis BC,  
40 1028 Gattoni-Celli S, McKillen DJ, Trent HF, *et al.* 2007. A dolphin peripheral blood  
41 1029 leukocyte cDNA microarray for studies of immune function and stress reactions. *Dev*  
42 1030 *Comp Immunol*. 31:520-529.
- 44 1031 Mancía A, Ryan JC, Chapman RW, Wu Q, Warr GW, Gulland FMD, Van Dolah FM. 2012.  
45 1032 Health status, infection and disease in California sea lions (*Zalophus californianus*)  
46 1033 studied using a canine microarray platform and machine-learning approaches. *Dev Comp*  
47 1034 *Immunol*. 36:629-637.
- 48 1035 Mancía A, Warr GW, Chapman RW. 2008. A transcriptomic analysis of the stress induced by  
49 1036 capture-release health assessment studies in wild dolphins (*Tursiops truncatus*). *Mol*  
50 1037 *Ecol*. 17:2581-2589.
- 52 1038 Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC. 2015.  
53 1039 Restriction site-associated DNA sequencing, genotyping error estimation and *de novo*  
54 1040 assembly optimization for population genetic inference. *Mol Ecol Resour*. 15:28-41.

- 1  
2  
3 1041 McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. 2012.  
4 1042 Ultraconserved elements are novel phylogenomic markers that resolve placental mammal  
5 1043 phylogeny when combined with species-tree analysis. *Genome Res.* 22:746-754.  
6 1044 McGowen MR. 2011. Toward the resolution of an explosive radiation - a multilocus phylogeny  
7 1045 of oceanic dolphins (Delphinidae). *Mol Phylogenet Evol.* 60:345-357.  
8 1046 McGowen MR, Clark C, Gatesy J. 2008. The vestigial olfactory receptor subgenome of  
9 1047 odontocete whales: phylogenetic congruence between gene-tree reconciliation and  
10 1048 supermatrix methods. *Syst Biol.* 57:574-590.  
11 1049 McGowen MR, Gatesy J, Wildman DE. 2014. Molecular evolution tracks macroevolutionary  
12 1050 transitions in Cetacea. *Trends Ecol Evol.* 29:336-346.  
13 1051 McGowen MR, Grossman LI, Wildman DE. 2012. Dolphin genome provides evidence for  
14 1052 adaptive evolution of nervous system genes and a molecular rate slowdown. *Proc R Soc*  
15 1053 *Lond B Biol Sci.* 279:3643-3651.  
16 1054 McGowen MR, Spaulding M, Gatesy J. 2009. Divergence date estimation and a comprehensive  
17 1055 molecular tree of extant cetaceans. *Mol Phylogenet Evol.* 53:891-906.  
18 1056 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,  
19 1057 Altshuler D, Gabriel S, Daly M, *et al.* 2010. The Genome Analysis Toolkit: A  
20 1058 MapReduce framework for analyzing next-generation DNA sequencing data. *Genome*  
21 1059 *Res.* 20:1297-1303.  
22 1060 McTavish EJ, Hillis DM. 2015. How do SNP ascertainment schemes and population  
23 1061 demographics affect inferences about population history? *BMC Genomics.* 16:266.  
24 1062 Meredith RW, Gatesy J, Emerling CA, York VM, Springer MS. 2013. Rod monochromacy and  
25 1063 the coevolution of cetacean retinal opsins. *PLoS Genetics.* 9:e1003432.  
26 1064 Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K,  
27 1065 de Filippo C, *et al.* 2012. A high-coverage genome sequence from an archaic Denisovan  
28 1066 individual. *Science.* 338:222-226.  
29 1067 Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. 2007. Rapid and cost-effective  
30 1068 polymorphism identification and genotyping using restriction site associated DNA  
31 1069 (RAD) markers. *Genome Res.* 17:240-248.  
32 1070 Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, Zhao F, Kim HL, Burhans RC,  
33 1071 Drautz DI, Wittekindt NE, *et al.* 2012. Polar and brown bear genomes reveal ancient  
34 1072 admixture and demographic footprints of past climate change. *Proc Natl Acad Sci USA.*  
35 1073 109:E2382-E2390.  
36 1074 Mirceta S, Signore AV, Burns JM, Cossins AR, Campbell KL, Berenbrink M. 2013. Evolution  
37 1075 of mammalian diving capacity traced by myoglobin net surface charge. *Science.*  
38 1076 340:1234192.  
39 1077 Morin PA, Archer FI, Foote AD, Vilstrup J, Allen EE, Wade P, Durban JW, Parsons K, Pitman  
40 1078 R, Li L, *et al.* 2010a. Complete mitochondrial genome phylogeographic analysis of killer  
41 1079 whales (*Orcinus orca*) indicates multiple species. *Genome Res.* 20:908-916.  
42 1080 Morin PA, Luikart G, Wayne RK, SNP workshop group. 2004. SNPs in ecology, evolution and  
43 1081 conservation. *Trends Ecol Evol.* 19:208-216.  
44 1082 Morin PA, Martien KK, Archer FI, Cipriano F, Steel D, Jackson J, Taylor BL. 2010b. Applied  
45 1083 conservation genetics and the need for quality control and reporting of genetic data used  
46 1084 in fisheries and wildlife management. *J Hered.* 101:1-10.

- 1  
2  
3 1085 Morin PA, Parsons KM, Archer FI, Ávila-Arcos M, Barrett-Lennard LG, Dalla Rosa L, Duchêne  
4 1086 S, Durban JW, Ellis GM, Ferguson SH, *et al.* 2015. Geographic and temporal dynamics  
5 1087 of a global radiation and diversification in the killer whale. *Mol Ecol.* 24:3964-3979.  
6 1088 Moura AE, Kenny JG, Chaudhuri R, Hughes MA, Welch AJ, Reisinger RR, de Bruyn PJN,  
7 1089 Dahlheim ME, Hall N, Hoelzel AR. 2014a. Population genomics of the killer whale  
8 1090 indicates ecotype evolution in sympatry involving both selection and drift. *Mol Ecol.*  
9 1091 23:5179-5192.  
10 1092 Moura AE, Nielsen SCA, Vilstrup JT, Moreno-Mayar JV, Gilbert MTP, Gray HWI, Natoli A,  
11 1093 Möller L, Hoelzel AR. 2013. Recent diversification of a marine genus (*Tursiops* spp.)  
12 1094 tracks habitat preference and environmental change. *Syst Biol.* 62:865-877.  
13 1095 Moura AE, van Rensburg CJ, Pilot M, Tehrani A, Best PB, Thornton M, Plön S, de Bruyn PJN,  
14 1096 Worley KC, Gibbs RA, *et al.* 2014b. Killer whale nuclear genome and mtDNA reveal  
15 1097 widespread population bottleneck during the last glacial maximum. *Mol Biol Evol.*  
16 1098 31:1121-1131.  
17 1099 Nadeau NJ, Ruiz M, Salazar P, Counterman B, Alejandro Medina J, Ortiz-Zuazaga H, Morrison  
18 1100 A, McMillan WO, Jiggins CD, Papa R. 2014. Population genomics of parallel hybrid  
19 1101 zones in the mimetic butterflies, *H. melpomene* and *H. erato*. *Genome Res.* 24:1316-  
20 1102 1333.  
21 1103 Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA. 2013. Genotyping-by-  
22 1104 sequencing in ecological and conservation genomics. *Mol Ecol.* 22:2841-2847.  
23 1105 Narum SR, Hess JE. 2011. Comparison of  $F_{ST}$  outlier tests for SNP loci under selection. *Mol*  
24 1106 *Ecol Resour.* 11:184-194.  
25 1107 Nelson TM, Apprill A, Mann J, Rogers TL, Brown MV. 2015. The marine mammal microbiome:  
26 1108 current knowledge and future directions. *Microbiology Australia.* 36:8-13.  
27 1109 Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M,  
28 1110 Bhattacharjee A, Eichler EE, *et al.* 2009. Targeted capture and massively parallel  
29 1111 sequencing of twelve human exomes. *Nature.* 461:272-276.  
30 1112 Nielsen R, Paul JS, Anders A, Song YS. 2011. Genotype and SNP calling from next-generation  
31 1113 sequencing data. *Nat Rev Genet.* 12:433-451.  
32 1114 Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S,  
33 1115 Pritchard JK, *et al.* 2006. Sequencing and analysis of Neanderthal genomic DNA.  
34 1116 *Science.* 314:1113-1118.  
35 1117 Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD,  
36 1118 Chin G, Christensen G, *et al.* 2015. Promoting an open research culture: Author  
37 1119 guidelines for journals could help to promote transparency, openness, and reproducibility.  
38 1120 *Science.* 348:1422-1425.  
39 1121 O'Rawe JA, Ferson S, Lyon GJ. 2015. Accounting for uncertainty in DNA sequencing data.  
40 1122 *Trends Genet.* 31:61-66.  
41 1123 Olsen MT, Volny VH, Bérubé M, Dietz R, Lydersen C, Kovacs KM, Dodd RS, Palsbøll PJ.  
42 1124 2011. A simple route to single-nucleotide polymorphisms in a nonmodel species:  
43 1125 identification and characterization of SNPs in the Arctic ringed seal (*Pusa hispida*  
44 1126 *hispida*). *Mol Ecol Resour.* 11:9-19.  
45 1127 Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E,  
46 1128 Petersen B, Moltke I, *et al.* 2013. Recalibrating *Equus* evolution using the genome  
47 1129 sequence of an early Middle Pleistocene horse. *Nature.* 499:74-78.  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 1130 Pabuwal V, Boswell M, Pasquali A, Wise SS, Kumar S, Shen Y, Garcia T, Lacerte C, Wise JP,  
4 1131 Jr., Wise JP, Sr., *et al.* 2013. Transcriptomic analysis of cultured whale skin cells exposed  
5 1132 to hexavalent chromium [Cr(VI)]. *Aquat Toxicol.* 134-135:74-81.
- 7 1133 Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-  
8 1134 wide signatures of convergent evolution in echolocating mammals. *Nature.* 502:228-231.
- 9 1135 Paszkiewicz KH, Farbox A, O'Neill P, Moore K. 2014. Quality control on the frontier. *Front*  
10 1136 *Genet.* 5:157.
- 11 1137 Patro R, Duggal G, Kingsford C. 2015. Accurate, fast, and model-aware transcript expression  
12 1138 quantification with Salmon. *bioRxiv.* doi: <http://dx.doi.org/10.1101/021592>.
- 14 1139 Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: an  
15 1140 inexpensive method for *de novo* SNP discovery and genotyping in model and non-model  
16 1141 species. *PLoS One.* 7:e37135.
- 18 1142 Poh Y-P, Domingues VS, Hoekstra HE, Jensen JD. 2014. On the prospect of identifying adaptive  
19 1143 loci in recently bottlenecked populations. *PLoS One.* 9:e110579.
- 20 1144 Poland JA, Brown PJ, Sorrells ME, Jannink J-L. 2012. Development of high-density genetic  
21 1145 maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing  
22 1146 approach. *PLoS One.* 7:e32253.
- 24 1147 Polanowski AM, Robbins J, Chandler D, Jarman SN. 2014. Epigenetic estimation of age in  
25 1148 humpback whales. *Mol Ecol Resour.* 14:976-987.
- 26 1149 Puritz JB, Hollenbeck CM, Gold JR. 2014. *dDocent*: a RADseq, variant-calling pipeline  
27 1150 designed for population genomics of non-model organisms. *PeerJ.* 2:e431.
- 28 1151 Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu  
29 1152 E, Kivisild T, Gupta R, *et al.* 2010. Ancient human genome sequence of an extinct  
30 1153 Palaeo-Eskimo. *Nature.* 463:757-762.
- 32 1154 Riesch R, Barrett-Lennard LG, Ellis GM, Ford JKB, Deecke VB. 2012. Cultural traditions and  
33 1155 the evolution of reproductive isolation: ecological speciation in killer whales? *Biol J Linn*  
34 1156 *Soc Lond.* 2012:1-17.
- 35 1157 Robinson JD, Coffman AJ, Hickerson MJ, Gutenkunst RN. 2014. Sampling strategies for  
36 1158 frequency spectrum-based population genomic inference. *BMC Evol Biol.* 14:254.
- 38 1159 Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential  
39 1160 expression analysis of digital gene expression data. *Bioinformatics.* 26:139-140.
- 40 1161 Ruan R, Guo A-H, Hao Y-J, Zheng J-S, Wang D. 2015. *De novo* assembly and characterization  
41 1162 of narrow-ridged finless porpoise renal transcriptome and identification of candidate  
42 1163 genes involved in osmoregulation. *Int J Mol Sci.* 16:2220-2238.
- 44 1164 Ruegg K, Rosenbaum HC, Anderson EC, Engel M, Rothschild A, Baker CS, Palumbi SR. 2013.  
45 1165 Long-term population size of the North Atlantic humpback whale within the context of  
46 1166 worldwide population structure. *Cons Gen.* 14:103-114.
- 47 1167 Schiffels S, Durbin R. 2014. Inferring human population size and separation history from  
48 1168 multiple genome sequences. *Nat Genet.* 46:919-925.
- 49 1169 Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming,  
50 1170 identification, and read merging. *BMC Res Notes.* 9:88.
- 52 1171 Schurch NJ, Schofield P, Gierlinski M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K,  
53 1172 Simpson GG, Owen-Hughes T, *et al.* 2015. Evaluation of tools for differential gene  
54 1173 expression analysis by RNA-seq on a 48 biological replicate experiment. *arXive.*  
55 1174 1505:02017.



- 1  
2  
3 1175 Seim I, Ma S, Zhou X, Gerashchenko MV, Lee SG, Suydam R, George JC, Bickham JW,  
4 1176 Gladyshev VN. 2014. The transcriptome of the bowhead whale *Balaena mysticetus*  
5 1177 reveals adaptations of the longest-lived mammal. *Aging*. 6:879-899.
- 7 1178 Shafer ABA, Cullingham CI, Côté SD, Coltman DW. 2010. Of glaciers and refugia: a decade of  
8 1179 study sheds new light on the phylogeographic patterns of northwestern North America.  
9 1180 *Mol Ecol*. 19:4589-4621.
- 10 1181 Shafer ABA, Davis CS, Coltman DW, Stewart REA. 2014. Microsatellite assessment of walrus  
11 1182 (*Odobenus rosmarus rosmarus*) stocks in Canada. *NAMMCO Scientific Publications*. 9.
- 12 1183 Shafer ABA, Gattepaille LM, Stewart REA, Wolf JBW. 2015. Demographic inferences using  
13 1184 short-read genomic data in an approximate Bayesian computation framework: *in silico*  
14 1185 evaluation of power, biases and proof of concept in Atlantic walrus. *Mol Ecol*. 24:328-  
15 1186 345.
- 17 1187 Shen Y-Y, Zhou W-P, Zhou T-C, Zeng Y-N, Li G-M, Irwin DM, Zhang Y-P. 2012. Genome-  
18 1188 wide scan for bats and dolphin to detect their genetic basis for new locomotive styles.  
19 1189 *PLoS One*. 7:e46455.
- 21 1190 Smith-Unna RD, Boursnell C, Patro R, Hibberd JM, Kelly S. 2015. TransRate: reference free  
22 1191 quality assessment of *de-novo* transcriptome assemblies. *bioRxiv*.
- 23 1192 Spies D, Ciaudo C. 2015. Dynamics in transcriptomics: advancements in RNA-seq time course  
24 1193 and downstream analysis. *Comput Struct Biotechnol J*. 13:469-477.
- 26 1194 Springer MS, Signore AV, Paijmans JLA, Vélez-Juarbe J, Domning DP, Bauer CE, He K, Crerar  
27 1195 L, Campos PF, Murphy WJ, *et al*. 2015. Interordinal gene capture, the phylogenetic  
28 1196 position of Steller's sea cow based on molecular and morphological data, and the  
29 1197 macroevolutionary history of Sirenia. *Mol Phylogenet Evol*. 91:178-193.
- 30 1198 Springer MS, Starrett J, Morin PA, Lanzetti A, Hayashi C, Gatesy J. 2016. Inactivation of  
31 1199 *C4orf26* in toothless placental mammals. *Mol Phylogenet Evol*. 95:34-45.
- 33 1200 Sremba AL, Martin AR, Baker CS. 2015. Species identification and likely catch time period of  
34 1201 whale bones from South Georgia. *Mar Mamm Sci*. 31:122-132.
- 35 1202 Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: *ab initio*  
36 1203 prediction of alternative transcripts. *Nucleic Acids Res*. 34:W435-W439.
- 37 1204 Stein LD. 2010. The case for cloud computing in genome informatics. *Genome Biol*. 11:207.
- 39 1205 Stinchcombe JR, Hoekstra HE. 2008. Combining population genomics and quantitative genetics:  
40 1206 finding genes underlying ecologically important traits. *Heredity*. 100:158-170.
- 41 1207 Tabuchi M, Veldhoen N, Dangerfield N, Jeffries S, Helbing CC, Ross PS. 2006. PCB-related  
42 1208 alteration of thyroid hormones and thyroid hormone receptor gene expression in free-  
43 1209 ranging harbor seals (*Phoca vitulina*). *Environ Health Perspect*. 114:1024-1031.
- 45 1210 Taylor BL, Gemmel NJ. 2016. Emerging technologies to conserve biodiversity: further  
46 1211 opportunities via genomics. Response to Pimm *et al*. *Trends Ecol Evol*. 31:171-172.
- 47 1212 The *Heliconius* Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of  
48 1213 mimicry adaptations among species. *Nature*. 487:94-98.
- 49 1214 Thomsen PF, Kielgast J, Iversen LL, Møller PR, Rasmussen M, Willerslev E. 2012. Detection of  
50 1215 a diverse marine fish fauna using environmental DNA from seawater samples. *PLoS One*.  
51 1216 7:e41732.
- 53 1217 Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, Hazlewood V, Lathrop S,  
54 1218 Lifka D, Peterson GD, *et al*. 2014. XSEDE: accelerating scientific discovery. *Computing*  
55 1219 *in Science and Engineering*. 16:62-74.

- 1  
2  
3 1220 Tsagkogeorga G, McGowen MR, Davies KT, Jarman S, Polanowski A, Bertelsen MF, Rossiter  
4 1221 SJ. 2015. A phylogenomic analysis of the role and timing of molecular adaptation in the  
5 1222 aquatic transition of cetartiodactyl mammals. *R Soc Open Sci.* 2:150156.  
6 1223 van Dijk EL, Auger H, Jaszczyn Y, Thermes C. 2014. Ten years of next-generation  
7 1224 sequencing technology. *Trends Genet.* 30:418-426.  
8 1225 VanRaden PM, Sun C, O'Connell JR. 2015. Fast imputation using medium or low-coverage  
9 1226 sequence data. *BMC Genet.* 16:82.  
10 1227 Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R,  
11 1228 Erichsen JT, Jasinska AJ, *et al.* 2015. Enhancer evolution across 20 mammalian species.  
12 1229 *Cell.* 160:554-566.  
13 1230 Viricel A, Pante E, Dabin W, Simon-Bouhet B. 2014. Applicability of RAD-tag genotyping for  
14 1231 interfamilial comparisons: empirical data from two cetaceans. *Mol Ecol Resour.* 14:597-  
15 1232 605.  
16 1233 Viricel A, Rosel PE. 2014. Hierarchical population structure and habitat differences in a highly  
17 1234 mobile marine species: the Atlantic spotted dolphin. *Mol Ecol.* 23:5018-5035.  
18 1235 Wolf JB. 2013. Principles of transcriptome analysis and gene expression quantification: an RNA-  
19 1236 seq tutorial. *Mol Ecol Resour.* 13:559-572.  
20 1237 Xiong Y, Brandley MC, Xu S, Zhou K, Yang G. 2009. Seven new dolphin mitochondrial  
21 1238 genomes and a time-calibrated phylogeny of whales. *BMC Evol Biol.* 9:20.  
22 1239 Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet.*  
23 1240 13:329-342.  
24 1241 Yeh R-F, Lim LP, Burge CB. 2001. Computational inference of homologous gene structures in  
25 1242 the human genome. *Genome Res.* 11:803-816.  
26 1243 Yim H-S, Cho YS, Guang X, Kang SG, Jeong J-Y, Cha S-S, Oh H-M, Lee J-H, Yang EC, Kwon  
27 1244 KK, *et al.* 2014. Minke whale genome and aquatic adaptation in cetaceans. *Nat Genet.*  
28 1245 46:88-92.  
29 1246 Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X, Hao P. 2011. Optimizing *de novo* transcriptome  
30 1247 assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics.*  
31 1248 12:S2.  
32 1249 Zhou X, Sun F, Xu S, Fan G, Zhu K, Liu X, Chen Y, Shi C, Yang Y, Huang Z, *et al.* 2013. Baiji  
33 1250 genomes reveal low genetic variability and new insights into secondary aquatic  
34 1251 adaptations. *Nat Commun.* 4:2708.  
35 1252 Zou Z, Zhang J. 2015. No genome-wide protein sequence convergence for echolocation. *Mol*  
36 1253 *Biol Evol.* 32:1237-1241.  
37 1254 |

1255 Table 1. Current and commonly used tools for analysis of genomic data generated in non-model organisms. Please note that this list is  
 1256 not exhaustive and new computational tools are continuously being developed.  
 1257

<b>Computational Tool</b>	<b>Purpose</b>	<b>Strengths/Weaknesses</b>	<b>Reference</b>
<i>RADseq*</i>			
STACKS	quality filtering, <i>de novo</i> assembly or reference-aligned read mapping, variant genotyping	scalable (new data can be compared against existing locus catalog); flexible filtering and export options; recently implemented a gapped alignment algorithm to process insertion-deletion (indel) mutations; secondary algorithm adjusts SNP calls using population-level allele frequencies; compatible with input data from multiple RADseq methods	Catchen et al. (2011; 2013), <a href="http://catchenlab.life.illinois.edu/stacks/">http://catchenlab.life.illinois.edu/stacks/</a>
PyRAD	quality filtering, <i>de novo</i> assembly, read mapping, variant genotyping	efficiently processes indel mutations, <u>thus</u> optimal for analysis of highly divergent species; high speed and quality of paired-end library assemblies; compatible with input data from multiple RADseq methods	Eaton (2014)
TASSEL-GBS	quality filtering, reference-aligned read mapping, variant genotyping	optimized for single-end data from large sample <u>sizes</u> (tens of thousands of individuals) with a reference genome; performs genome-wide association studies	Glaubitz et al. (2014)
dDocent	quality trimming, <i>de novo</i> assembly, read mapping, variant genotyping	beneficial in analysis of paired-end data; identifies both SNP and indel variants; most appropriate for ezRAD and ddRAD data	Puritz et al. (2014)
AfrRAD	quality filtering, <i>de novo</i> assembly, read mapping, variant genotyping	identifies both SNP and indel variants; computationally faster than STACKS and PyRAD	Sovic et al. (2015)
<i>Array-based high-throughput sequencing</i>			
Affymetrix Axiom™ Analysis Suite, Illumina® GenomeStudio	genotype scoring	visualization of genotype clusters; quality scores assigned to genotype calls allow user-specific filtering; manual editing possible	
<i>Whole genome sequencing</i>			
AdapterRemoval v2, Trimmomatic	trim raw sequences	remove adapter sequences and low-quality bases prior to assembly	Bolger et al. (2014), Schubert et al. (2016)
ALLPATHS-LG, PLATANUS, SOAPdenovo	<i>de novo</i> genome assembly	designed for short-read sequences of large heterozygous genomes	Li et al. (2010), Gnerre et al. (2011), Kajitani et al. (2014)
AUGUSTUS, GenomeScan, MAKER2	gene annotation	highly accurate evidence-driven or BLASTX-guided gene prediction (Yandell and Ence 2012)	Yeh et al. (2001), Stanke et al. (2006), Holt and Yandell (2011)

1				
2				
3	Bowtie, bwa	read mapping	rapid short-read alignment with compressed reference genome index, but limited number of acceptable mismatches per alignment (Flicek and Birney 2009)	Langmead et al. (2009), Li and Durbin (2009)
4				
5				
6				
7	SAMtools	data processing, variant calling (SNP and indel discovery)	multi-purpose tool that conducts file conversion, alignment sorting, PCR duplicate removal, and variant (SNP and indel) calling for SAM/BAM/CRAM files	Li et al. (2009)
8				
9				
10	GATK	data processing and quality control, variant calling	suitable for <del>processing and analyses of data with</del> low to high <del>mean read depth across the genome</del> <del>coverage</del> <del>data</del> ; initially optimized for large human datasets, then modified for use with non-model organisms	McKenna et al. (2010), DePristo et al. (2011)
11				
12				
13				
14	ANGSD/NGStools	data processing, variant calling, estimation of diversity metrics, population genomic analyses	suitable for <del>processing and analyses of data with</del> low <del>mean read depth, including coverage and</del> palaeogenomic data; allow downstream analyses such as D-statistics and SFS estimation	Fumagalli et al. (2014), Korneliusen et al. (2014)
15				
16				
17				
18				
19	<i>RNAseq</i>			
20	Fastx Toolkit, Trimmomatic	trim raw sequences	remove erroneous nucleotides from reads prior to assembly	MacManes (2014)
21				
22	khmer diginorm, Trinity normalization	<i>in silico</i> read normalization	reduces memory requirements for assembly, but can result in fragmented assemblies and collapse heterozygosity	Brown et al. (2012); Haas et al. (2013)
23				
24				
25	Trinity	<i>de novo</i> and genome-guided transcriptome assembly	accurate assembly across conditions, but requires long runtime if normalization is not used (Zhao et al. 2011)	Haas et al. (2013)
26				
27				
28	bowtie, bowtie2, STAR	read alignment to genome or transcriptome assembly	required for many downstream analyses, but bowtie is computationally intensive and all produce very large output BAM files	Langmead et al. (2009), Dobin et al. (2013)
29				
30				
31	eXpress, kallisto, RSEM, Sailfish, Salmon	estimation of transcript abundance	RSEM requires computationally intensive read mapping back to the assembly; the others are faster streaming alignment, quasi-alignment, or alignment-free algorithms	Li and Dewey (2011), Patro et al. (2015)
32				
33				
34				
35	DESeq, DESeq2, edgeR	differential expression analysis	exhibit highest true positive and lowest false positive rates in experiments with smaller sample sizes (Schurch et al. 2015)	Anders and Huber (2010), Robinson et al. (2010), Love et al. (2014)
36				
37				
38				
39	blast2GO, Trinotate	functional annotation of assembled transcripts	complete annotation pipelines including gene ontology and pathway enrichment analyses	Conesa et al. (2005), Haas et al. (2013)
40				
41				
42				
43				
44				
45				
46				
47				
48				
49				

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49

1258 | \* This is a non-exhaustive list of software that ~~include~~ focuses on *de novo* loci assembly and genotype calling for RADseq data, as many practitioners working on  
1259 | NMOs will not have access to a reference genome. Other programs (e.g., GATK and ANGSD) that undertake genotype calling using reference-aligned loci ~~only~~  
1260 | are described in the whole genome sequencing section.

For Peer Review

Figure 1

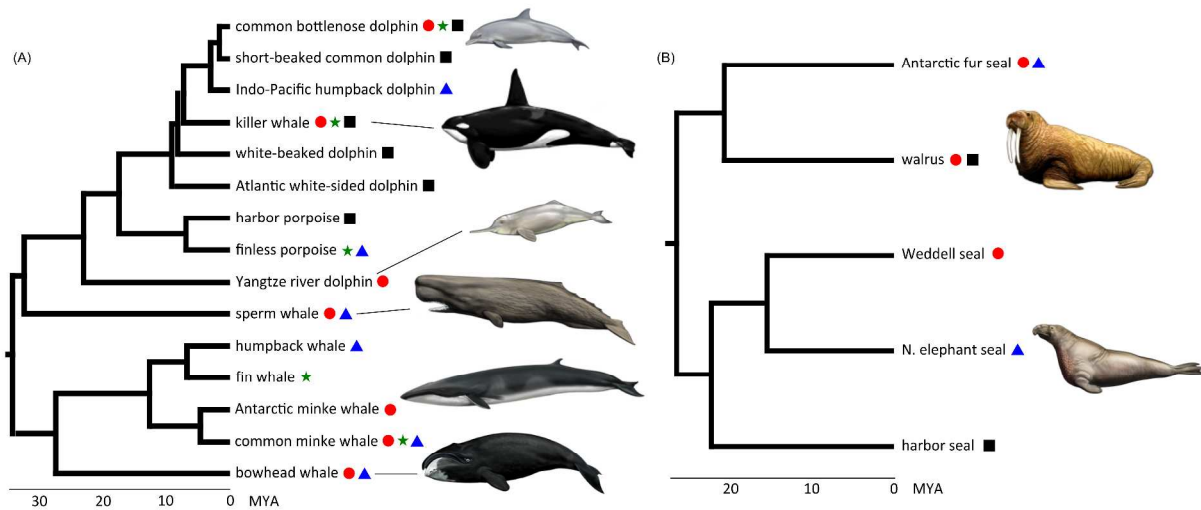


Figure 1. Phylogenetic tree showing current genomic resources available for (A) cetaceans and (B) pinnipeds; relationships and branch lengths are based on molecular dating estimates from McGowen et al. (2009), McGowen (2011), and Higdon et al. (2007). Scale is in millions of years ago (MYA). Red circles indicate species with high-coverage-quality whole-reference genomes; green stars indicate low-coverage whole genome re-sequencing data; blue triangles indicate transcriptomes (generated by microarray or RNAseq); and black squares indicate RADseq data.

Figure 2

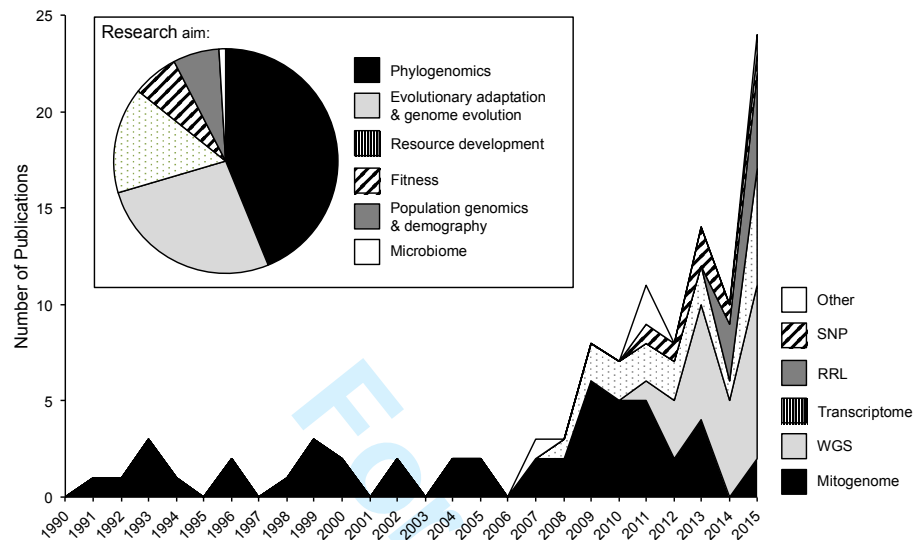


Figure 2. Number of marine mammal genomics publications from 1990 to 2015, categorized by primary methodology and research aim. Genomic methodologies include high-throughput single nucleotide polymorphism (SNP) genotyping and sequencing of mitogenomes, whole genomes (WGS), transcriptomes (generated by microarray or RNAseq), and reduced-representation genomic libraries (RRL). The “Other” category includes studies of microbiomes, BAC libraries, and large (~100) gene sets.

Figure 3

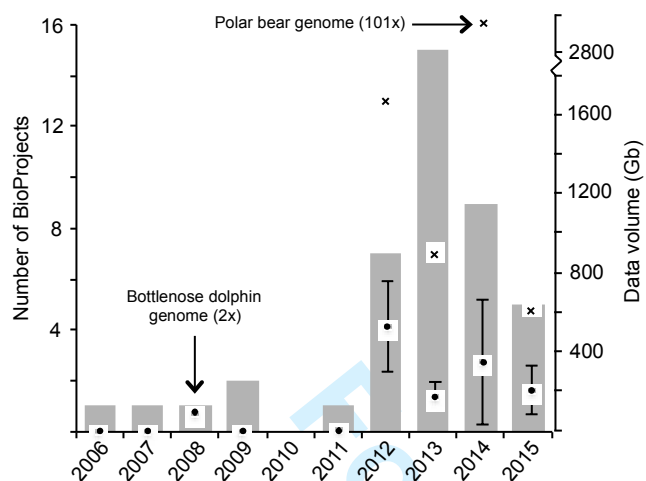


Figure 3. Number of BioProjects (shaded gray bars) related to marine mammal genomics submitted from 2006 to 2015 to an online public database maintained by NCBI. Early BioProjects were largely microarray datasets. The number of projects created each year, as well as the yearly average (black dots  $\pm$  SE) and maximum ( $\times$ ) size of data submitted in each BioProject, increased dramatically after 2011, reflecting advances in high-throughput sequencing technologies that facilitated their use in non-model systems.



Figure 4

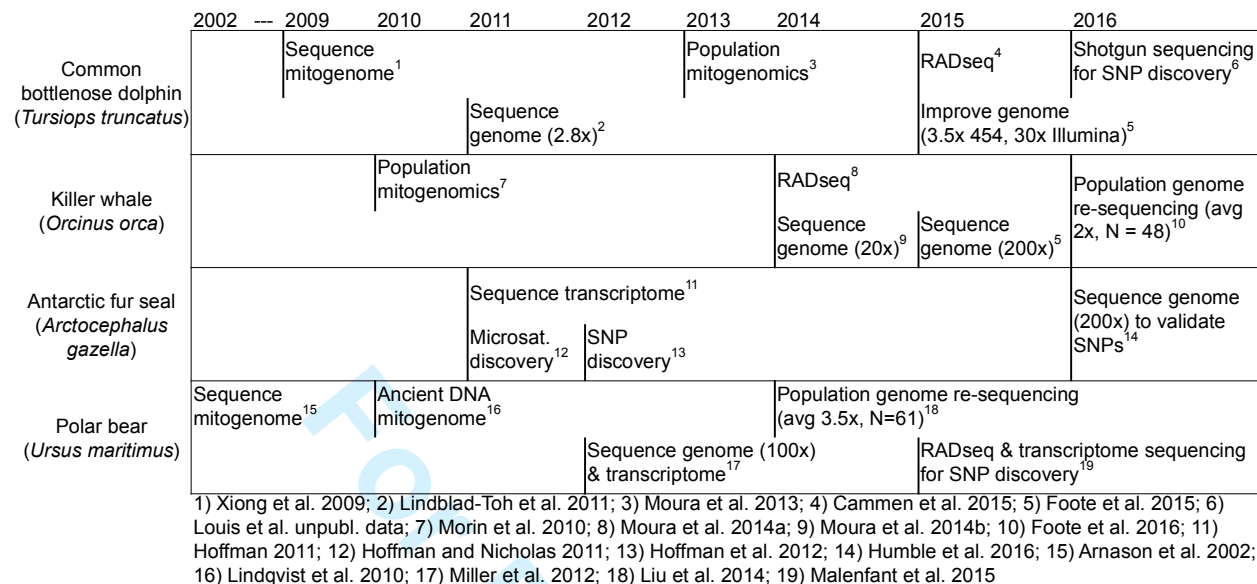


Figure 4. Timelines depicting the independent progression of genomic studies for four representative marine mammal species. Trajectories show the common progression for non-model species from mitogenome sequencing to whole genome sequencing, as well as from sequencing reference specimens to population-scale genomic sequencing. In addition, the timelines reveal the utility of genomic and transcriptomic sequencing for subsequent genetic marker development.