How low can you go: sex identification from sequencing data of species lacking assembled sex chromosomes

Andrea A. Cabrera[1], Alba Rey-Iglesia[1], Marie Louis[1,2], Mikkel Skovrind[1], Michael V Westbury[1]*, Eline D Lorenzen[1]*

[1] GLOBE Institute, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark
[2] Greenland Institute of Natural Resources, Kivioq 2, Nuuk 3900, Greenland
* authors contributed equally

Corresponding author: EDL (elinelorenzen@sund.ku.dk), MVW (m.westbury@sund.ku.dk) and AAC (andrea_ca_gt@yahoo.com)

Abstract

Accurate sex identification is crucial for elucidating the biology of a species. Here, we present SeXY, a sex-identification pipeline, for very low-coverage shotgun sequencing data from a single individual. The method does not require a conspecific sex-chromosome assembly as reference. SeXY was specifically designed to utilise low-effort screening data for sexing, but can also be applied to samples of higher-effort sequencing. We assess the accuracy of our pipeline to data quantity by downsampling sequencing data from 100,000 to 1,000 mapped reads, and mapping to a variety of reference genomes of various quality and phylogenetic distance. We show that when mapping to a high-quality (highly contiguous N50 > 30 Mb in our case, or chromosome-level) conspecific genome, our method is 100% accurate even down to 1,000 mapped reads. For lower-quality reference assemblies (N50 < 30 Mb), our method is 100% accurate with 50,000 mapped reads, regardless of reference assembly quality or phylogenetic distance. The SeXY pipeline provides several advantages over previously implemented methods; SeXY (i) requires data from only a single individual, (ii) does not require assembled conspecific sex-chromosomes, or even a conspecific reference assembly, (iii) takes into account variation in coverage across the genome, and (iv) is accurate with only 1,000 mapped reads in many cases. SeXY is broadly applicable to any target species with a heterogametic sex, including birds, mammals, and certain reptiles, fish, and insects.

Introduction

Accurate sex identification is critical for elucidating the life history, behaviour, social structure, and demography of a species. It is particularly important for taxa where females and males differ in prey preference (e.g., Louis et al. 2021), social interactions and mating behaviour (e.g., Amos, Schlötterer, and Tautz 1993; Pečnerová et al. 2017), and seasonal movements and dispersal (e.g., Dobson and Stephen Dobson 1982; Greenwood 1980;

Gower et al. 2019). Reliable sex identification may also elucidate the impacts of past and present anthropogenic activities on wildlife, including prehistoric hunting or domestication practices (e.g., Nistelberger et al. 2019), and the identification of the sex of and sex biases in ongoing wildlife poaching (e.g., Malisa et al. 2005).

In the absence of directly observable sexual characteristics, such as morphology or behaviour (Fairbairn, Blanckenhorn, and Székely 2008), sex identification of wild fauna in the field remains challenging, if not impossible. An additional challenge for research based on museum or palaeontological specimens is the sex identification of skeletal remains. In most cases, such as in the (sub-)fossil record, only small skeletal fragments are available. Osteological sex determination may be further limited by the degree of preservation, the age of the individual, or access to appropriate reference material with which to compare (Buonasera et al. 2020).

Molecular sexing can be used as an alternative to morphological sexing; it only requires a small tissue sample, and may even be applied to environmental samples (Hrovatin and Kunej 2018). Many molecular sexing techniques utilise information regarding the homogametic and heterogametic sexes. In mammals, and in many fishes, females are homogametic and males are heterogametic with XX and XY chromosomes, respectively (Ellegren 2000; Moore 1925; Í Kongsstovu et al. 2020). In birds and certain reptiles, the pattern is reversed, with females having ZW and males having ZZ chromosomes.

For tissue samples with high-quality DNA, molecular sex identification is relatively fast, inexpensive, and straightforward. Methods for mammals include PCR-based techniques that (i) amplify the SRY gene of the Y chromosome (Bryja and Konečný 2003; Pomp et al. 1995), or (ii) target specific regions of the ZFX and ZFY genes found on the X and Y chromosomes, respectively (e.g., Bérubé and Palsbøll 1996; Aasen and Medrano 1990; Curtis, Stewart, and Karl 2007). However, these approaches require specific laboratory work targeting loci on sex chromosomes, and are not suitable for samples with highly fragmented and/or degraded DNA, such as material not specifically sampled and preserved for DNA analysis (including skeletal remains, wildlife products, and museum specimens). PCR failure in method (i) and a biased amplification of the ZFX over the ZFY region (Sinding et al. 2016) in method (ii) may cause males to be misidentified as females.

The analysis of shotgun sequencing data offers a more robust approach to identify the sex of an individual; endogenous shotgun data can be retrieved from samples with low-quality DNA, with no additional laboratory procedures required to specifically target loci on sex chromosomes. Sex-identification pipelines for DNA data with a low number of target reads were originally developed for human ancient DNA data, and were based on either the ratio of number of reads aligning to the X and Y chromosomes (Skoglund et al. 2013), or on the ratio of number of reads aligning to the X chromosome *versus* the autosomes (Mittnik et al. 2016). This last method has recently been utilised on elephants and other mammalian taxa for which the X chromosome of either a conspecific or a related reference genome is available (de Flamingh et al. 2020; Bro-Jørgensen et al. 2021). Although this approach has been shown to be efficient down to ~10,000 mapped sequencing reads, it requires either a conspecific chromosome-level assembly with known sex chromosomes, or mapping to a more distantly related chromosome-level assembly, with decreased mapping efficiency as a result.

Reference genome assemblies from non-model vertebrate species with assembled sex chromosomes are relatively scarce. Available mammalian genome assemblies with at least one sex chromosome (most commonly the X chromosome) include humans, several domesticates such as cat (*Felis catus*), cow (*Bos taurus*), dog (*Canis familiaris*), horse (*Equus caballus*), sheep (*Ovis aries*), and wild species such as blue whale (*Balaenoptera musculus*), bottlenose dolphin (*Tursiops truncatus*), greater horseshoe bat (*Rhinolophus ferrumequinum*), gorilla (*Gorilla gorilla*), meerkat (*Suricata suricatta*), orangutan (*Pongo pygmaeus*), and vaquita (*Phocoena sinus*) (de Flamingh et al. 2020; Cabrera et al. 2021). In the absence of a conspecific chromosome-level assembly, alternative approaches can be used to identify scaffolds originating from sex chromosomes. Approaches include synteny-based, whole-genome alignments (e.g., Grabherr et al. 2010), and the estimation of relative coverage of each scaffold using data from known females and males of the target species (reviewed in Palmer et al. 2019). Sex identification using synteny or coverage approaches has been applied in some studies using ancient (e.g., Kirch et al. 2021) or degraded DNA (e.g., Skovrind et al. 2019). However, the pipelines have been developed for specific species and datasets, and an assessment of the minimum level of required sequencing data and of the impact of reference genome assembly choice is lacking.

Methods exist that circumvent the need to *a priori* identify sex-linked scaffolds. For example, a recent fast and automated method “Sex Assignment Through Coverage” uses principal component analysis to identify sex-related scaffolds, and the sex of an individual (Nursyifa et al. 2021). This approach holds promise for studies that include a relatively large number of samples, as the method requires a set of both male and female samples. However, these sample requirements may not always be met.

Here, we present a sex-identification method (SeXY) for taxa lacking a conspecific chromosome-level assembly. The method can be applied to shotgun sequencing data from any species with a heterogametic sex (e.g, mammals, birds, and some reptiles, fish, and insects). We use a synteny-based approach to identify putative X-linked scaffolds in the reference assembly, and determine sex using the expectation that males (in mammals) have half the amount of X-chromosome genetic material compared to females. We assessed the robustness of this method using raw shotgun sequencing data from two target marine mammal species: beluga whale (*Delphinapterus leucas*) and polar bear (*Ursus maritimus*). The read data were subsampled and mapped to reference assemblies of various quality and phylogenetic distance. We show our approach to be highly accurate (i) with as few as 1,000 mapped reads when mapping to a high-quality (chromosome level) reference genome assembly, or as few as 50,000 mapped reads when mapping to a lower-quality reference genome assembly (N50 < 30 million base pairs (Mb)); (ii) also when using a phylogenetically distant reference genome assembly; and (iii) without known sex chromosomes.

Materials and Methods

The SeXY method requires (i) raw shotgun sequencing reads of a target individual; (ii) an assembled genome from either a conspecific or related species (RefGEN); and (iii) assembled X and Y chromosomes (RefX and RefY, respectively), which can be either from the same or another species than the RefGEN.

We assessed the applicability of SeXY using data from two target species: beluga and polar bear. We also assessed the impact of reference assembly using four RefGEN of varying quality and phylogenetic distance to each target species, and two reference sex chromosome assemblies (each comprising RefX and RefY) from species of varying phylogenetic distance. To ascertain the applicability of our method to specimens with low DNA yield, we additionally tested the impact of the number of mapped reads on the sex determination using various downsamplings ranging from 100,000 to 1,000 mapped reads.

1. Target species data and reference assemblies

We used publicly available Illumina shotgun sequencing reads from ten beluga and ten polar bear individuals (Supplementary Table 1). Each species dataset comprised five females and five males. As we were interested in results produced with <=100,000 mapped reads only, all read files were randomly downsampled to one million reads using the sample option in seqtk v1.3 (https://github.com/lh3/seqtk), to reduce computational time during the mapping step.

To evaluate the impact of reference genome assembly, we used four reference assemblies (RefGEN) for each target species (beluga, polar bear): two conspecific RefGEN of differing assembly quality, and two RefGEN from more divergent species (Figure 2; Supplementary Table 2). All scaffolds shorter than 10 kilobase (kb) were removed from the RefGEN files and excluded from downstream analyses using reformat.sh from the BBmap toolsuite (Bushnell 2014).

For beluga, we included two beluga reference assemblies: one of lower quality (Beluga v1, N50 161 kb (Jones et al. 2017)) and one highly contiguous (Beluga v3, N50 31 Mb (Dudchenko et al. 2018, 2017)). We also included a killer whale (*Orcinus orca*) assembly (Orca, N50 13 Mb (Foote et al. 2015)) and a chromosome-level cow assembly (Cow, N50 103 Mb (Zimin et al. 2009)). Assuming a divergence time between the beluga and killer whale of ~19 million years ago (Ma) (McGowen et al. 2020) and a yearly mutation rate for beluga of 5.16 x 10^{-10} (Westbury et al. 2019), the divergence between the beluga and killer whale genomes is estimated at ~2%. The divergence between the beluga and cow genomes is estimated at ~6.8% assuming a divergence time of ~66 Ma (McGowen et al. 2020), and above-mentioned beluga mutation rate.

For polar bear, we included two polar bear reference assemblies: the lower quality Polar bear v1, N50 16 Mb (Liu et al. 2014) and the chromosome-level Polar bear v1 HiC, N50 71 Mb (Dudchenko et al. 2018, 2017). We also included a chromosome-level panda (*Ailuropoda melanoleuca*) assembly (Panda, N50 129 Mb (Fan et al. 2019)), and a chromosome-level dog assembly (Dog, N50 64 Mb (Lindblad-Toh et al. 2005)). The estimated divergence between the polar bear and dog genomes is ~6.4%, assuming a divergence time of ~19.5 Ma (Hu et al. 2017) and a mutation rate for polar bear of 1.6 x 10^{-9} (Liu et al. 2014). The divergence between the polar bear and dog genomes is estimated at ~17%, assuming a divergence time of ~52 Ma (Hu et al. 2017) and above-mentioned polar bear mutation rate.

2. Identification of putative sex-linked and autosomal scaffolds

We identified scaffolds putatively originating from sex chromosomes (both X and Y) from all RefGEN lacking assembled sex chromosomes as well as from Cow and Dog, which include assembled sex chromosomes. We did this by aligning each RefGEN with a designated pair of RefX and RefY assemblies, using satsuma synteny v2.1 (Grabherr et al. 2010) with default parameters (Figure 1). Although our method relies on comparing X chromosome and autosomal coverage (which we term X:A ratio), we included the Y chromosome to remove possible biases due to pseudoautosomal regions (homologous regions between the X and Y chromosomes) (Helena Mangs and Morris 2007). To reduce this bias, we removed any overlapping coordinates between the X- and Y-linked scaffold bed files using bedtools v.2.29.0 intersect (Quinlan and Hall 2010). We identified putative autosomal scaffolds by removing the previously identified putative sex-linked scaffolds from each RefGEN.

We selected three RefX and RefY combinations: (i) HumanX and HumanY, (ii) CowX and HumanY, and (iii) DogX and DogY (Supplementary Table 3). The human sex chromosome assemblies were selected as they are the most well-assembled mammalian sex chromosomes available. We selected the cow and dog sex-chromosome assemblies, as they each represent the highest-quality, chromosome-level assemblies with defined sex chromosomes within the same phylogenetic order as each of our target species: beluga (Artiodactyla) and polar bear (Carnivora). We used the three RefX and RefY combinations to assess the influence of phylogenetic distance to the target species on downstream sex determination. For the cetacean/cow RefGEN dataset used for beluga, combinations (i) and (ii) were used (Figure 1a). For the bear/dog RefGEN dataset used for polar bear, combinations (i) and (iii) were used (Figure 1b) (Supplementary Table 3). For the Cow and Dog RefGENs only one combination of RefX and RefY was tested (CowX and HumanY for the former, and DogX and DogY for the latter).

3. Mapping and downsampling of mapped reads

Processing and mapping of raw beluga and polar bear sequencing reads to each designated RefGEN (Figure 1A) was performed using the Paleomix pipeline v.1.3.2 (Schubert et al. 2014). Adapter sequences were trimmed from the raw reads with AdapterRemoval v.2.3.1 (Schubert et al. 2014; Schubert, Lindgreen, and Orlando 2016) using default settings and a minimum read length of 30 bp. Trimmed reads were mapped with BWA-MEM v.0.7.17 (Heng Li 2013) to each RefGEN. Mapped reads with mapping quality < 30 were removed using SAMtools v1.9 (H. Li et al. 2009). Duplicates were removed using Picard MarkDuplicates (http://broadinstitute.github.io/picard).

To evaluate the impact of number of mapped reads on genetic sex determination, we randomly downsampled the bam files to 100,000; 50,000; 10,000; 5,000; 2,500 and 1,000 mapped reads (Figure 2) using BBMap (Bushnell 2014). We evaluated the differences in the mapping efficiency to each RefGEN, measured as the number of raw reads required to obtain a specific number of mapped reads (Figure 2, Supplementary Figure 1, Supplementary Table 4).

4. Sex determination

The sex of each individual was estimated based on the X chromosome:autosome coverage ratio (X:A ratio). We calculated the read depth of all sites from the X-linked scaffolds and from the autosomal scaffolds using SAMtools depth v.1.9 (H. Li et al. 2009), specifying minimum base and mapping qualities of 25. To take into account variation across genomic regions, we randomly selected 10 million sites from both X-linked and autosomal scaffolds independently, calculated the average coverage for those sites, and calculated the X:A ratio from the average coverages. This step was repeated ten times (Supplementary Table 5). As female mammals have two copies of the X chromosome, and males carry only one copy, we expected X:A ratios of ~1 and ~0.5 for females and males, respectively. We determined a female as correctly identified if the mean X:A ratio of the ten replicates was >=0.8 and a male if the mean X:A ratio of the ten replicates was <= 0.7. We considered a X:A ratio of 0.7 - 0.8 as 'undetermined' sex .

When interpreting the accuracy of the method, we considered (i) correctly determined sex; (ii) 'undetermined' sex, (iii) incorrectly determined sex (Supplementary Table 6). We did this to indicate whether accuracy below 100% was due to individuals with undetermined sex (with a X:A ratio of 0.7-0.8), or due to with incorrectly determined sex, as the latter is more detrimental to biological inference than simply the inability to determine sex.

Results

1. Mapping

In agreement with previous results (Prasad, Lorenzen, and Westbury 2021), we found a decline in mapping efficiency as phylogenetic distance to the RefGEN increased (Supplementary Table 4). For the beluga dataset, the average percentage of raw reads successfully mapping and passing filters were as follows: Beluga v1 - 81%, Beluga v3 - 82%, Orca - 75%, and Cow - 25%. For the polar bear dataset, the average percentage was: Polar bear v1 - 91%, Polar bear v1 HiC - 91%, Panda - 80%, and Dog - 24%.

2. Sex determination

We found the sexing approach implemented in SeXY provided 100% accuracy in sex determination across all combinations of reference genome assembly (RefGEN) and reference sex-chromosome assemby (RefX, RefY), when 100,000 and 50,000 mapped reads were available (Figure 2, Table 1, Supplementary Figure1, Supplementary Table 6). Moreover, 100% accuracy was observed for most trials involving lower numbers of mapped reads; 10,000 and 5,000. Clear exceptions could be seen when using Beluga v1 (N50 161 kb) and Orca (N50 13 Mb) as RefGEN in the beluga data set. Inaccuracies were especially prevalent when the low-quality Beluga v1 RefGEN (N50 161 kb) was used; we found a marked decline in accuracy when using ≤10,000 mapped reads, with sex determination accuracy in some cases equivalent to random chance (down to 50%) (Table 1).

Taken together, our results showed scaffold contiguity of the RefGEN influences the accuracy of sex determination more than phylogenetic distance. Across all trials, we found the highest percentage of correctly identified sex was obtained with highly contiguous (Beluga v3) or chromosome-level (Polar bear v1 HiC, Panda, Dog) RefGEN, regardless of whether the RefGEN was from a conspecific or a more divergent species (Table 1, Figure 2).

For the beluga dataset and CowX and HumanY RefXY (table 1 and figure 1), we found 100% accuracy in sex determination down to 10,000 mapped reads when using the higher-quality Beluga v3 (N50 31 Mb) and Cow (N50 103 Mb) RefGENs (Table 1). When we decreased the number of mapped reads below 5,000, we obtained a 10%-20% decrease in accuracy, which resulted in some undetermined individuals. However, for the trials where we were able to determine sex, the sex was determiend with 100% accuracy down to 1,000 and 2,500 mapped reads with Beluga v3 and Cow as RefGEN, respectively.

When analysing the polar bear dataset and DogX and DogY RefXY (table 1 and figure 1), we found 100% accuracy in sex determination down to 5,000 mapped reads for all RefGEN. Both polar bear RefGENs (Polar bear v1, Polar bear v1 HiC) produced similar sex determination accuracies (Table 1), with 100% accuracy down to 2,500 mapped reads. However, when we decreased the number of mapped reads to 1,000, mapping to the less contiguous Polar bear v1 correctly determined the sex in 70% of individuals (30% were incorrect), while the chromosome-level Polar bear v1 HiC correctly determined the sex with 100% accuracy. When using the Dog assembly as RefGEN, we found 100% accuracy regardless of the number of mapped reads.

We also tested whether the two combinations of RefX and RefY used in each species data set (CowX/HumanY *vs* HumanX/HumanY for beluga; DogX/DogY *vs* HumanX/HumanY for polar bear) provided the same results. We observed a small fraction of contradictions in sex identification, where an individual was identified as a female when using one RefX/Y set, and as a male in the other RefX/Y set, despite the RefGEN and number of mapped reads being identical (Supplementary Table 5-6). When comparing sex identifications produced using identical RefGEN and number of mapped reads, but different RefX/Y combinations, results were identical in 94% of the pairwise comparisons (337 out of 360 comparisons, including both beluga and polar bear data sets). The inability to designate the sex of an individual with both combinations of RefX/Y and RefY was only observed in two comparison. In the remaining 6% of comparisons, 2% (eight comparisons) yielded contradicting sex identifications. In six of the comparisons the more distant HumanXY RefX/Y produced the correction results, in one comparison the DogXY gave the correct result (polar bear dataset) and, in the remaining comparison the CowXHumanY gave the correct result (beluga dataset) The last 4% (15 comparisons) comprised one determined sex (male or female) and one undetermined sex (X:A ratio of 0.7-0.8). We obtained contradicting sex determination only in comparisons using relatively few reads and with the low-quality Beluga v1 RefGEN (using 5,000 and 2,500 mapped reads), and with Beluga v3 and Polar bear v1 RefGEN (using 1,000 mapped reads).

Discussion

Many biological specimens for which sex cannot be identified using morphology or other traditional approaches, such as faecal, environmental, and archaeological or palaeontological material, are also likely to contain highly contaminated and/or degraded DNA (Hrovatin and Kunej 2018). Therefore, by assessing the reliability of SeXY to various levels of sequencing effort, we evaluate its applicability to such samples. Although our results differed between reference genomes, we show that less than 5,000 mapped reads can be used to accurately identify the biological sex of an individual, depending on the quality of the mapping reference. This finding opens a world of possibility for studies that employ low-effort shotgun sequencing approaches to identify specimens of sufficient preservation for deeper sequencing, but which discard any data/specimens not deemed of sufficient quality. By utilising our method, sequence information that would previously have been discarded can now be used to obtain sex-related evolutionary and biological insights. Although this has been done on several taxa (e.g., Gower et al. 2019; Pečnerová et al. 2017), our method, which does not require *a priori* sex-chromosome information or a reference panel of known females and males, will hopefully enable such analyses from a much wider range of species. Although only tested with up to 100,000 mapped reads, the increasing accuracy as the number of mapped reads increased means this method is also suitable for well-preserved specimens with more available sequencing data. In such cases, data could even be downsampled to increase computational speed.

SeXY identifies sex-linked scaffolds using a synteny approach (Grabherr et al. 2010), where the reference sex-chromosome assemblies (RefX and RefY) of a chromosome-level assembly from a closely related species is used to identify sequence similarities on the reference genome assembly (RefGEN). Although this method may have limitations due to computational time or the lacking identification of new (neo)-sex chromosomes (Marshall Graves 2008; Nursyifa et al. 2021), our results show that SeXY could accurately determine the sex of the beluga and polar bear individuals, even with a relatively distant sex-chromosome assembly (in our case, human). In addition, the identification of sex-linked scaffolds is performed only once per reference genome assembly used, and hence computation time will not increase with the number of samples.

Our finding of 100% accurate sex identification when mapping polar bear reads to the dog as RefGEN, even with only 1,000 mapped reads, was somewhat unexpected, as we anticipated a decline in mapping efficiency with increasing phylogenetic distance (Prasad, Lorenzen, and Westbury 2021). However, these results become less surprising when considering the mapping efficiency to each RefGEN. Although sex determination was 100% accurate down to 1,000 mapped reads when using these two species with ~17% divergence, approximately four times as many raw reads are required to reach the target number of mapped reads, relative to when mapping to a conspecific RefGEN (Figure 2, Supplementary Table 4). Therefore, when < 5,000 endogenous reads are available, it is important to weigh the number of mapped reads *versus* the number of raw reads, to evaluate whether mapping to a conspecific reference genome or a phylogenetically distant reference genome is more beneficial. Although not tested here, alterations in mapping quality filters may facilitate the recovery of more mapped reads and thereby more accurate sex identification. However, decreased mapping quality may also result in misalignments, biasing results. Such low endogenous read counts are unlikely to arise when sequencing DNA from well-preserved samples, but it is much more common when considering highly degraded samples such as faecal, environmental, or subfossil material.

When comparing results produced by mapping beluga reads to the more fragmented Beluga v1 *versus* the more contiguous Beluga v3, we show the quality of the reference genome assembly can significantly impact the accuracy of sex determination. The two beluga assemblies are vastly different in quality, with scaffold N50s of 161 kb and 31 Mb, respectively. When considering < 50,000 mapped reads, the more fragmented Beluga v1 assembly could not be used to accurately determine sex. A fragmented reference genome assembly of lower quality, as with Beluga v1, may lead to difficulties in accurately identifying the sex-linked scaffolds, which our method is reliant on. Therefore, although not comprehensively investigated here, it is advisable to rather use a high-quality reference genome assembly from a phylogenetically more distant species, than a low-quality conspecific assembly. However, the accuracy of the X:A ratio using Beluga v1 as mapping reference provided 100% accuracy at 50,000 and above mapped reads. Therefore, we show that SeXY can still be used to accurately identify sex even if only a highly fragmented assembly is available, if the number of mapped reads is sufficiently high. This holds promise for the applicability of our method moving forward, as there are an increasing number of high-quality reference genomes available, and initiatives such as the Vertebrate Genome project aim to generate near error-free reference genome assemblies of many vertebrate species in the near future (Rhie et al. 2021).

Phylogenetic distance of the mapping reference genome assembly also appears to play a role. In the case of beluga mapped to the Orca RefGEN, comparisons using < 10,000 mapped reads were unable to accurately identify an individual's sex. However, this finding may reflect the more fragmented assembly of the Orca (N50 = 13 Mb) relative to the other mapping references, as we were able to identify sex with 80% accuracy (89% excluding undetermined sex) using Cow as RefGEN down to 1,000 mapped reads. Furthermore, while Panda as RefGEN produced less consistent results for the polar bear than the two conspecific reference genome assemblies, the Panda results were far more consistent than when Orca was used as RefGEN for beluga, perhaps owing to the higher assembly quality of the Panda (N50 = 129 Mb). Thus, our results suggest that the quality of the reference genome assembly is far more important than phylogenetic distance between the species of interest and the mapping reference.

In conclusion, we demonstrate the method implemented in SeXY can accurately determine the sex of individuals based on very low sequencing effort, and when no conspecific chromosome-level assembly is available. The SeXY pipeline provides several advantages over previously implemented methods: SeXY (i) requires data from only a single individual (a mix of male and female individuals is not required), (ii) does not require assembled conspecific sex-chromosomes, or even a conspecific reference assembly, (ii) takes into account variation in coverage across the genome when calculating the X:A ratio, and (iv) can work on very low-coverage shotgun data, down to 1,000 mapped reads in many cases. Although we assessed the method based on XY sex chromosomes (as in mammals), the method can in theory be applied to any species with a heterogametic and a homogametic sex (e.g, birds, and some reptiles, fish, and insects).

Acknowledgements

This work was supported by the Carlsberg Foundation Distinguished Associate Professor Fellowship (grant no CF16-0202) and the Villum Fonden Young Investigator program (grant no 13151) to E.D.L. A.A.C. was supported by a fellowship from Rubicon-NWO (project 019.183EN.005) and M.L. was supported by a fellowship from the Greenland Research Council.

References

Aasen, E., and J. F. Medrano. 1990. “Amplification of the ZFY and ZFX Genes for Sex Identification in Humans, Cattle, Sheep and Goats.” *Bio/technology* 8 (12): 1279–81.

Amos, B., C. Schlötterer, and D. Tautz. 1993. “Social-Structure of Pilot Whales Revealed by Analytical DNA Profiling.” *Science* 260 (5108): 670–72.

Bérubé, M., and P. Palsbøll. 1996. “Identification of Sex in Cetaceans by Multiplexing with Three ZFX and ZFY Specific Primers.” *Molecular Ecology* 5 (2): 283–87.

Bro-Jørgensen, Maiken Hemme, Xénia Keighley, Hans Ahlgren, Camilla Hjorth Scharff-Olsen, Aqqalu Rosing-Asvid, Rune Dietz, Steven H. Ferguson, et al. 2021. “Genomic Sex Identification of Ancient Pinnipeds Using the Dog Genome.” *Journal of Archaeological Science* 127 (March): 105321.

Bryja, J., and A. Konečný. 2003. “Fast Sex Identification in Wild Mammals Using PCR Amplification of the Sry Gene.” *Folia Zoo* 52 (3): 269–74.

Buonasera, Tammy, Jelmer Eerkens, Alida de Flamingh, Laurel Engbring, Julia Yip, Hongjie Li, Randall Haas, et al. 2020. “A Comparison of Proteomic, Genomic, and Osteological Methods of Archaeological Sex Estimation.” *Scientific Reports* 10 (1): 11897.

Bushnell, Brian. 2014. “BBMap: A Fast, Accurate, Splice-Aware Aligner.” In *No. LBNL-7065E*. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://www.semanticscholar.org/paper/f64dd54444a724574deb7710888091350eebb2b9.

Cabrera, Andrea A., Martine Bérubé, Xênia M. Lopes, Marie Louis, Tom Oosting, Alba Rey-Iglesia, Vania E. Rivera-León, Dóra Székely, Eline D. Lorenzen, and Per J. Palsbøll. 2021. “A Genetic Perspective on Cetacean Evolution.” *Annual Review of Ecology, Evolution, and Systematics* 52: 131–51.

Curtis, C., B. S. Stewart, and S. A. Karl. 2007. “Sexing Pinnipeds with ZFX and ZFY Loci.” *The Journal of Heredity* 98 (3): 280–85.

Dobson, F. Stephen, and F. Stephen Dobson. 1982. “Competition for Mates and Predominant Juvenile Male Dispersal in Mammals.” *Animal Behaviour*. https://doi.org/10.1016/s0003-3472(82)80209-1.

Dudchenko, Olga, Sanjit S. Batra, Arina D. Omer, Sarah K. Nyquist, Marie Hoeger, Neva C. Durand, Muhammad S. Shamim, et al. 2017. “De Novo Assembly of the Aedes Aegypti Genome Using Hi-C Yields Chromosome-Length Scaffolds.” *Science* 356 (6333): 92–95.

Dudchenko, Olga, Muhammad S. Shamim, Sanjit S. Batra, Neva C. Durand, Nathaniel T. Musial, Ragib Mostofa, Melanie Pham, et al. 2018. “The Juicebox Assembly Tools Module Facilitates de Novo Assembly of Mammalian Genomes with Chromosome-Length Scaffolds for under $1000.” *bioRxiv*. https://doi.org/10.1101/254797.

Ellegren, Hans. 2000. “Evolution of the Avian Sex Chromosomes and Their Role in Sex Determination.” *Trends in Ecology & Evolution*. https://doi.org/10.1016/s0169-5347(00)01821-8.

Fairbairn, Daphne J., Wolf U. Blanckenhorn, and Tamás Székely. 2008. *Sex, Size and Gender Roles: Evolutionary Studies of Sexual Size Dimorphism*. OUP Oxford.

Fan, Huizhong, Qi Wu, Fuwen Wei, Fengtang Yang, Bee Ling Ng, and Yibo Hu. 2019.

"Chromosome-Level Genome Assembly for Giant Panda Provides Novel Insights into Carnivora Chromosome Evolution." *Genome Biology* 20 (1): 267.

Flamingh, Alida de, Ashley Coutu, Alfred L. Roca, and Ripan S. Malhi. 2020. "Accurate Sex Identification of Ancient Elephant and Other Animal Remains Using Low-Coverage DNA Shotgun Sequencing Data." *G3* 10 (4): 1427–32.

Foote, Andrew D., Yue Liu, Gregg W. C. Thomas, Tomáš Vinař, Jessica Alföldi, Jixin Deng, Shannon Dugan, et al. 2015. "Convergent Evolution of the Genomes of Marine Mammals." *Nature Genetics* 47 (3): 272–75.

Gower, Graham, Lindsey E. Fenderson, Alexander T. Salis, Kristofer M. Helgen, Ayla L. van Loenen, Holly Heiniger, Emilia Hofman-Kamińska, et al. 2019. "Widespread Male Sex Bias in Mammal Fossil and Museum Collections." *Proceedings of the National Academy of Sciences of the United States of America* 116 (38): 19019–24.

Grabherr, Manfred G., Pamela Russell, Miriah Meyer, Evan Mauceli, Jessica Alföldi, Federica Di Palma, and Kerstin Lindblad-Toh. 2010. "Genome-Wide Synteny through Highly Sensitive Sequence Alignment: Satsuma." *Bioinformatics* 26 (9): 1145–51.

Greenwood, Paul J. 1980. "Mating Systems, Philopatry and Dispersal in Birds and Mammals." *Animal Behaviour*. https://doi.org/10.1016/s0003-3472(80)80103-5.

Hrovatin, Karin, and Tanja Kunej. 2018. "Genetic Sex Determination Assays in 53 Mammalian Species: Literature Analysis and Guidelines for Reporting Standardization." *Ecology and Evolution* 8 (2): 1009–18.

Hu, Yibo, Qi Wu, Shuai Ma, Tianxiao Ma, Lei Shan, Xiao Wang, Yonggang Nie, et al. 2017. "Comparative Genomics Reveals Convergent Evolution between the Bamboo-Eating Giant and Red Pandas." *Proceedings of the National Academy of Sciences of the United States of America* 114 (5): 1081–86.

Í Kongsstovu, Sunnvør, Hans Atli Dahl, Hannes Gislason, Eydna Homrum, Jan Arge Jacobsen, Paul Flicek, and Svein-Ole Mikalsen. 2020. "Identification of Male Heterogametic Sex-Determining Regions on the Atlantic Herring Clupea Harengus Genome." *Journal of Fish Biology* 97 (1): 190–201.

Jones, Steven J. M., Gregory A. Taylor, Simon Chan, René L. Warren, S. Austin Hammond, Steven Bilobram, Gideon Mordecai, et al. 2017. "The Genome of the Beluga Whale (*Delphinapterus Leucas*)." *Genes* 8 (12). https://doi.org/10.3390/genes8120378.

Kirch, Melanie, Anders Romundset, M. Thomas P. Gilbert, Felicity C. Jones, and Andrew D. Foote. 2021. "Ancient and Modern Stickleback Genomes Reveal the Demographic Constraints on Adaptation." *Current Biology: CB* 31 (9): 2027–36.e8.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *arXiv E-Prints*, March, arXiv:1303.3997.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp352.

Lindblad-Toh, Kerstin, Claire M. Wade, Tarjei S. Mikkelsen, Elinor K. Karlsson, David B. Jaffe, Michael Kamal, Michele Clamp, et al. 2005. "Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog." *Nature* 438 (7069): 803–19.

Liu, Shiping, Eline D. Lorenzen, Matteo Fumagalli, Bo Li, Kelley Harris, Zijun Xiong, Long Zhou, et al. 2014. "Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears." *Cell* 157 (4): 785–94.

Louis, Marie, Mikkel Skovrind, Eva Garde, Mads Peter Heide-Jørgensen, Paul Szpak, and Eline D. Lorenzen. 2021. "Population-Specific Sex and Size Variation in Long-Term Foraging Ecology of Belugas and Narwhals." *Royal Society Open Science* 8 (2): 202226.

Malisa, A., P. Gwakisa, S. Balthazary, S. Wasser, and B. Mutayoba. 2005. "Species and Gender Differentiation between and among Domestic and Wild Animals Using Mitochondrial and Sex-Linked DNA Markers." *African Journal of Biotechnology* 4 (11). https://doi.org/10.4314/ajb.v4i11.71391.

Marshall Graves, Jennifer A. 2008. "Weird Animal Genomes and the Evolution of Vertebrate

Sex and Sex Chromosomes.” *Annual Review of Genetics* 42: 565–86.

McGowen, Michael R., Georgia Tsagkogeorga, Sandra Álvarez-Carretero, Mario Dos Reis, Monika Struebig, Robert Deaville, Paul D. Jepson, et al. 2020. “Phylogenomic Resolution of the Cetacean Tree of Life Using Target Sequence Capture.” *Systematic Biology* 69 (3): 479–501.

Mittnik, Alissa, Chuan-Chao Wang, Jiří Svoboda, and Johannes Krause. 2016. “A Molecular Approach to the Sexing of the Triple Burial at the Upper Paleolithic Site of Dolní Věstonice.” *PloS One* 11 (10): e0163019.

Moore, Carl R. 1925. “Sex Determination and Sex Differentiation in Birds and Mammals.” *The American Naturalist*. https://doi.org/10.1086/280026.

Nistelberger, Heidi M., Albína Hulda Pálsdóttir, Bastiaan Star, Rúnar Leifsson, Agata T. Gondek, Ludovic Orlando, James H. Barrett, Jón Hallsteinn Hallsson, and Sanne Boessenkool. 2019. “Sexing Viking Age Horses from Burial and Non-Burial Sites in Iceland Using Ancient DNA.” *Journal of Archaeological Science*. https://doi.org/10.1016/j.jas.2018.11.007.

Nursyifa, Casia, Anna Brüniche-Olsen, Genis Garcia Erill, Rasmus Heller, and Anders Albrechtsen. 2021. “Joint Identification of Sex and Sex-Linked Scaffolds in Non-Model Organisms Using Low Depth Sequencing Data.” *Molecular Ecology Resources* 00 (August): 1–10.

Palmer, Daniela H., Thea F. Rogers, Rebecca Dean, and Alison E. Wright. 2019. “How to Identify Sex Chromosomes and Their Turnover.” *Molecular Ecology* 28 (21): 4709–24.

Pečnerová, Patrícia, David Díez-Del-Molino, Nicolas Dussex, Tatiana Feuerborn, Johanna von Seth, Johannes van der Plicht, Pavel Nikolskiy, Alexei Tikhonov, Sergey Vartanyan, and Love Dalén. 2017. “Genome-Based Sexing Provides Clues about Behavior and Social Structure in the Woolly Mammoth.” *Current Biology: CB* 27 (22): 3505–10.e3.

Pomp, D., B. A. Good, R. D. Geisert, C. J. Corbin, and A. J. Conley. 1995. “Sex Identification in Mammals with Polymerase Chain Reaction and Its Use to Examine Sex Effects on Diameter of Day-10 or -11 Pig Embryos.” *Journal of Animal Science* 73 (5): 1408–15.

Prasad, Aparna, Eline D. Lorenzen, and Michael V. Westbury. 2021. “Evaluating the Role of Reference-Genome Phylogenetic Distance on Evolutionary Inference.” *Molecular Ecology Resources*, June. https://doi.org/10.1111/1755-0998.13457.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq033.

Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. “Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species.” *Nature* 592 (7856): 737–46.

Schubert, Mikkel, Luca Ermini, Clio Der Sarkissian, Hákon Jónsson, Aurélien Ginolhac, Robert Schaefer, Michael D. Martin, et al. 2014. “Characterization of Ancient and Modern Genomes by SNP Detection and Phylogenomic and Metagenomic Analysis Using PALEOMIX.” *Nature Protocols* 9 (5): 1056–82.

Schubert, Mikkel, Stinus Lindgreen, and Ludovic Orlando. 2016. “AdapterRemoval v2: Rapid Adapter Trimming, Identification, and Read Merging.” *BMC Research Notes* 9 (February): 88.

Sinding, Mikkel-Holger S., Outi M. Tervo, Bjarne Grønnow, Hans Christian Gulløv, Peter A. Toft, Lutz Bachmann, Katharina Fietz, et al. 2016. “Sex Determination of Baleen Whale Artefacts: Implications for Ancient DNA Use in Zooarchaeology.” *Journal of Archaeological Science: Reports* 10 (December): 345–49.

Skoglund, Pontus, Jan Storå, Anders Götherström, and Mattias Jakobsson. 2013. “Accurate Sex Identification of Ancient Human Remains Using DNA Shotgun Sequencing.” *Journal of Archaeological Science* 40 (12): 4477–82.

Skovrind, Mikkel, Jose Alfredo Samaniego Castruita, James Haile, Eve C. Treadaway, Shyam Gopalakrishnan, Michael V. Westbury, Mads Peter Heide-Jørgensen, Paul Szpak, and Eline D. Lorenzen. 2019. “Hybridization between Two High Arctic Cetaceans Confirmed by Genomic Analysis.” *Scientific Reports*.

https://doi.org/10.1038/s41598-019-44038-0.

Westbury, Michael V., Bent Petersen, Eva Garde, Mads Peter Heide-Jørgensen, and Eline D. Lorenzen. 2019. “Narwhal Genome Reveals Long-Term Low Genetic Diversity despite Current Large Abundance Size.” *iScience*, May. https://doi.org/10.1016/j.isci.2019.03.023.

Zimin, Aleksey V., Arthur L. Delcher, Liliana Florea, David R. Kelley, Michael C. Schatz, Daniela Puiu, Finnian Hanrahan, et al. 2009. “A Whole-Genome Assembly of the Domestic Cow, Bos Taurus.” *Genome Biology* 10 (4): R42.

Data Availability Statement

All raw data used in this paper are publicly available in GenBank, NCBI and DNA Zoo. Accession codes can be found within the Materials and Methods and Supplementary Table 1-3. SeXY pipeline is available on Github: https://github.com/andreidae/SeXY

Author Contributions

A.A.C.: Conceptualization, Formal analysis, Writing -Original Draft, Visualization; A.R-I: Conceptualization, Formal analysis, Writing - Review & Editing; M.L.: Conceptualization, Formal analysis, Writing - Review & Editing; M.S.: Conceptualization, Formal analysis, Writing - Review & Editing; M.V.W.: Conceptualization, Methodology, Writing - Review & Editing, Supervision; E.D.L.: Conceptualization, Resources, Writing - Review & Editing, Supervision, Funding acquisition.

Tables and Figures (with captions)

Tables

Table 1: Summary table showing percentage of correct sex determination across tested combinations of reference genome assembly (RefGEN), reference sex-chromosome assembly (RefX and RefY), and number of mapped reads. Results are shown for the beluga data and the cetacean/cow RefGEN assemblies tested (left columns), and for the polar bear data and the bear/dog RefGEN assemblies tested (right columns). The value below each RefGEN indicates the assembly N50. For cells with two estimates, the left value indicates estimates including undetermined sex, and the right value indicates estimates excluding undetermined sex. Only one value is included if both estimates were the same. Percentages in each cell are based on ten sample individuals; five females and five males. Sex determination for each indvidual was calculated using the average value of ten replicates, assuming a threshold of <= 0.7: male; 0.7-0.8: undetermined sex; >= 0.8: female. Corresponding summary table for tests using HumanX and HumanY as RefX and RefY, respectively, is provided in Supplementary Table 6.

	Beluga				Polar bear			
Number of mapped reads	**Beluga v1**	**Beluga v3**	**Orca**	**Cow**	**Polar bear v1**	**Polar bear v1 HiC**	**Panda**	**Dog**
	161 kb	**31 Mb**	**13 Mb**	**103 Mb**	**16 Mb**	**71 Mb**	**129 Mb**	**64 Mb**
	CowX and HumanY				**DogX and DogY**			
100,000	100	100	100	100	100	100	100	100
50,000	100	100	100	100	100	100	100	100
10,000	50	100	90/100	100	100	100	100	100
5,000	50/56	90/100	80/89	100	100	100	100	100
2,500	100	100	50/63	80/100	100	100	90/100	100
1,000	50	80/100	60	80/89	70	100	80/89	100

Figures

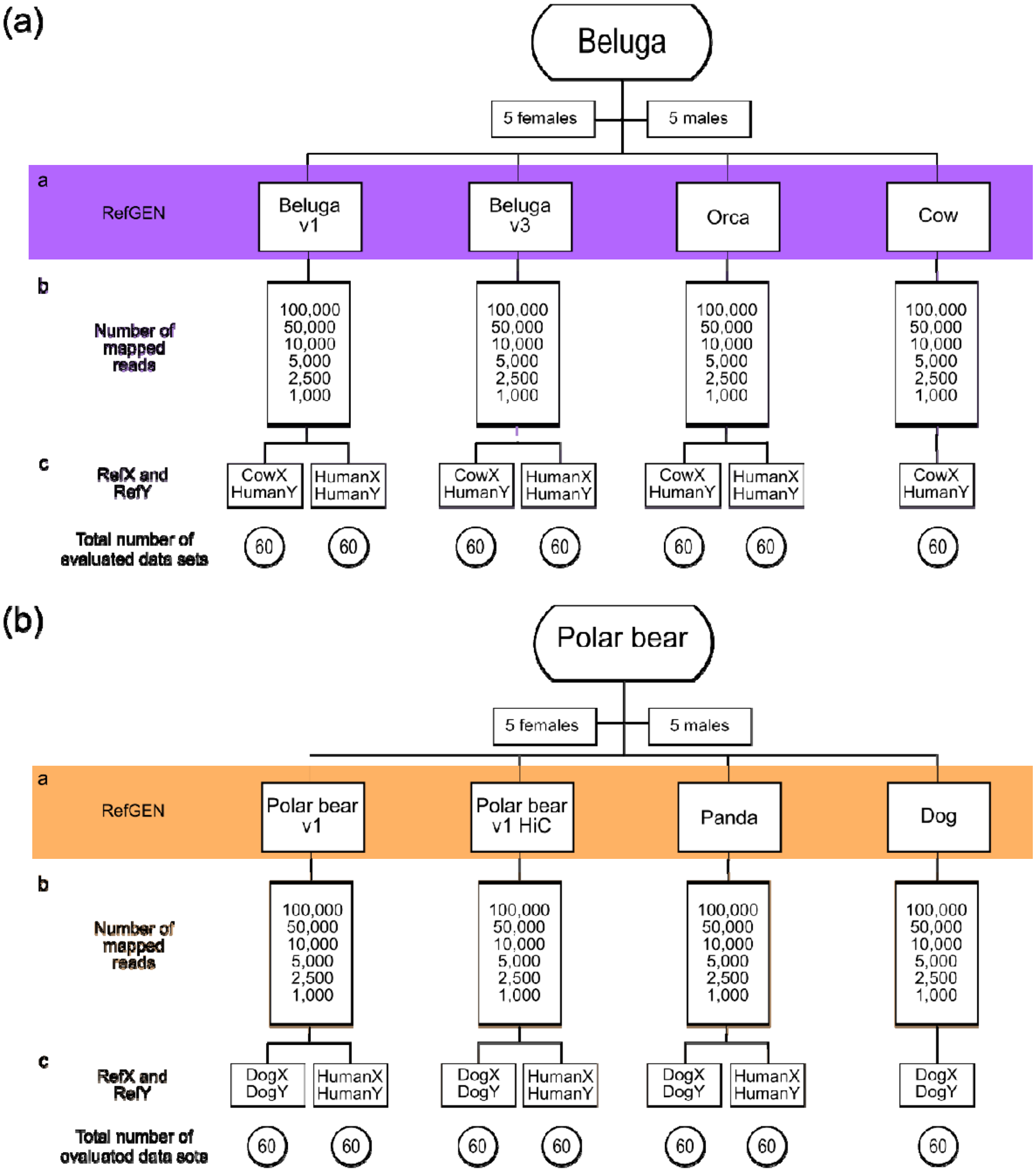

Figure 1. Schematic representation of the data sets and reference assemblies (RefGEN, RefX, RefY) analyzed for the two target species: beluga and polar bear. Each branch of the flowchart shows the evaluated combination of (a) reference genome assembly (RefGEN) used as mapping reference for the raw reads of each target species, (b) number of mapped reads of the target species (representing six independent data sets), and (c) reference sex-chromosome assembly (RefX and RefY) used to localize the sex-linked scaffolds (synteny). Total number of evaluated data sets per branch of the flow chart is shown at the bottom of the figure.

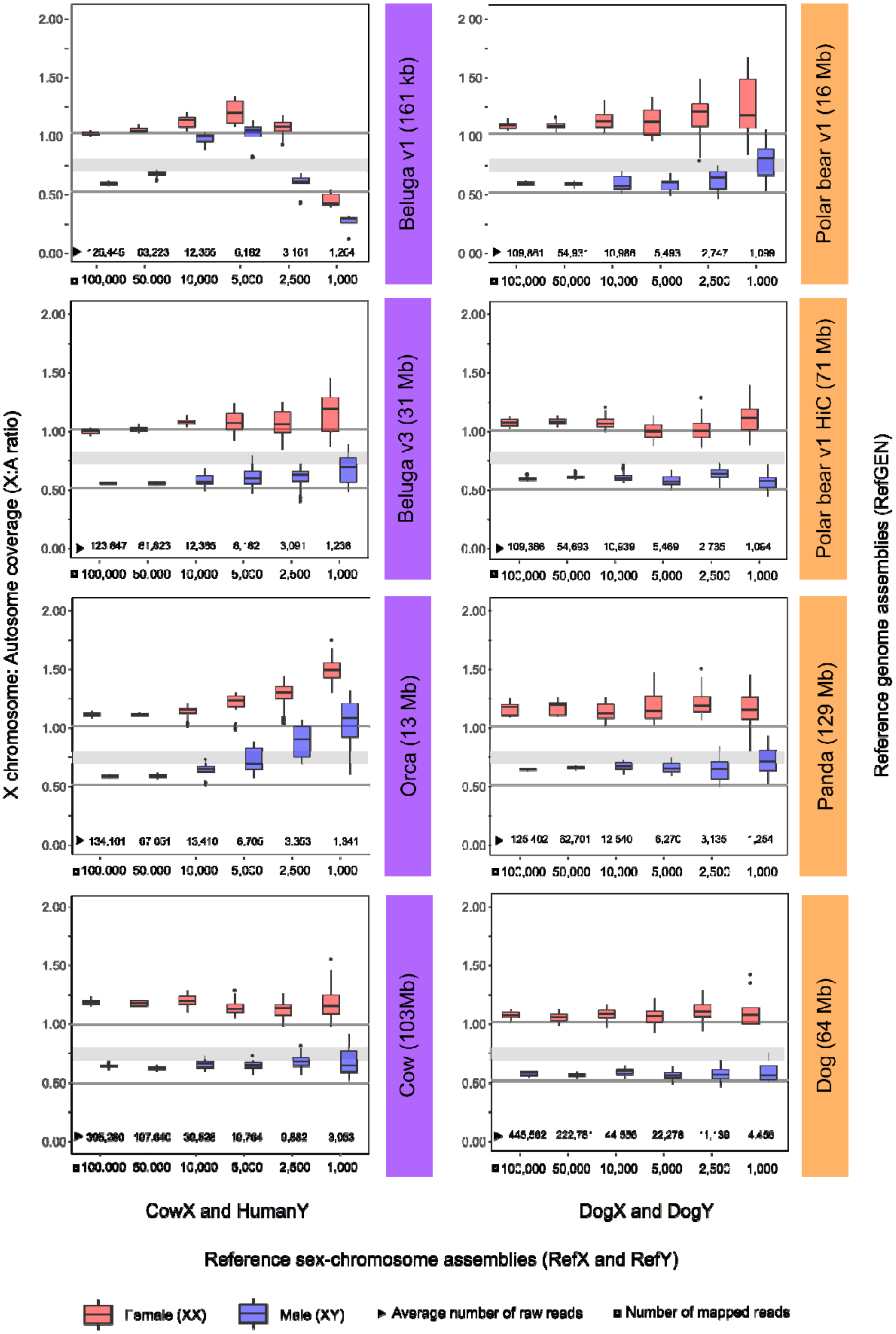
X chromosome: Autosome coverage (X:A ratio)
2.00
1.50
1.00
0.50
0.00
Beluga v1 (161 kb)
▶ 126,445 63,223 12,335 6,162 3 161 1,264
■ 100,000 50.000 10,000 5,000 2,500 1,000
Polar bear v1 (16 Mb)
▶ 109,861 54,931 10,986 5,493 2,747 1,099
■ 100,000 50,000 10,000 5,000 2,500 1.000
Beluga v3 (31 Mb)
▶ 123 647 61,823 12,365 6,182 3,091 1,238
■ 100,000 50.000 10,000 5,000 2,500 1,000
Polar bear v1 HiC (71 Mb)
▶ 109,386 54,693 10,939 5,469 2 735 1,064
■ 100,000 50,000 10,000 5,000 2,500 1,000
Orca (13 Mb)
▶ 134,101 67 051 13,410 6,705 3,353 1,341
■ 100.000 50.000 10,000 5,000 2,500 1,000
Panda (129 Mb)
▶ 125 402 62,701 12 540 6,270 3,135 1,254
■ 100,000 50,000 10,000 5,000 2,500 1,000
Cow (103Mb)
▶ 305,260 167,640 30,526 16,764 6,882 3,053
■ 100.000 50.000 10,000 5,000 2,500 1,000
Dog (64 Mb)
▶ 445,562 222,781 44 556 22,278 11,138 4,456
■ 100,000 50,000 10,000 5,000 2,500 1,000
Reference genome assemblies (RefGEN)
CowX and HumanY
DogX and DogY
Reference sex-chromosome assemblies (RefX and RefY)
Female (XX)
Male (XY)
▶ Average number of raw reads
■ Number of mapped reads

Figure 2. Sex determination of beluga and polar bear individuals using four reference genome assemblies (RefGEN), one combination of reference sex-chromosome assembly (RefX and RefY) for each target species, and various numbers of mapped reads. The ten beluga and ten polar bear individuals tested both comprised five females (red) and five males (blue). X axis shows number of mapped reads (square) and average number of raw reads necessary to obtain the required number of mapped reads (triangle). Y axis shows comparison of X chromosome and autosome coverage (X:A ratio) for each combination of RefGEN, RefX and RefY (CowX and HumanY, DogX and DogY), and number of mapped reads. Individuals were determined as females if their X:A ratio was >=0.8, and as males if their X:A ratio was <=0.7. Grey shaded horizontal bars indicate an X:A ratio of 0.7-0.8, which we interpreted as undetermined sex.