Title: A draft fur seal genome provides insights into factors affecting SNP validation and how to mitigate them

Humble, E.[1, 3], Martinez-Barrio, A.[2], Forcada, J.[3], Trathan, P.N.[3], Thorne, M.A.S.[3], Hoffmann M.[4], Wolf, J.B.W.[5*] & Hoffman, J.I.[1*]

[1]Department of Animal Behaviour, University of Bielefeld, Postfach 100131, 33501 Bielefeld, Germany
[2]Science of Life Laboratories and Department of Cell and Molecular Biology, Uppsala University, Husargatan 3, 75124 Uppsala, Sweden.
[3]British Antarctic Survey, High Cross, Madingley Road, Cambridge, CB3 OET, UK
[4]Max Planck Institute for Developmental Biology, Spemannstrasse 35, 72076 Tübingen, Germany
[5]Science of Life Laboratories and Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden

[*] Joint senior authors

Corresponding author:
Emily Humble
Department of Animal Behaviour
University of Bielefeld
Postfach 100131
33501 Bielefeld
Germany
E-mail: emily.humble@uni-bielefeld.de

Keywords: Antarctic fur seal, *Arctocephalus gazella*, draft genome, single nucleotide polymorphism (SNP), cross-validation, SNP array.

Running title (<45 characters): Improving SNP validation success

Abstract

Custom genotyping arrays provide a flexible and accurate means of genotyping single nucleotide polymorphisms (SNPs) in a large number of individuals of essentially any organism. However, validation rates, defined as the proportion of putative SNPs that are verified to be polymorphic in a population, are often very low. A number of potential causes of assay failure have been identified, but none have been explored systematically. In particular, as SNPs are often developed from transcriptomes, parameters relating to the genomic context are rarely taken into account. Here, we assembled a draft Antarctic fur seal (*Arctocephalus gazella*) genome (assembly size: 2.41Gb; scaffold/contig N_{50}: 3.1Mb/27.5kb). We then used this resource to map the probe sequences of 144 putative SNPs genotyped in 480 individuals. The number of probe-to-genome mappings and alignment length together explained almost a third of the variation in validation success, indicating that sequence uniqueness and proximity to intron-exon boundaries play an important role. The same pattern was found after mapping the probe sequences to the Walrus and Weddell seal genomes, suggesting that the genomes of species divergent by as much as 23 million years can hold information relevant to SNP validation outcomes. Additionally, re-analysis of genotyping data from seven previous studies found the same two variables to be significantly associated with SNP validation success across a variety of taxa. Finally, our study reveals considerable scope for validation rates to be improved, either by simply filtering for SNPs whose flanking sequences align uniquely and completely to a reference genome, or through predictive modeling.

Introduction

Single nucleotide polymorphisms (SNPs) are the most abundant form of genetic variation, with an estimated ten million being present in human populations (Kruglyak & Nickerson 2001). Around four million of these have been validated (Jorgenson & White 2006), meaning that they can be reliably scored and are polymorphic in a given population (Conklin *et al.* 2013, Montes *et al.* 2013). SNPs are suitable for addressing many questions in population genetics given their co-dominant, biallelic nature and well understood mutation processes (Brumfield *et al.* 2003; Morin *et al.* 2004). Furthermore, SNPs provide technical advantages compared to other markers such as microsatellites, including the possibility to genotype them on a large scale (Seeb *et al.* 2011) and with minimal error (Hoffman *et al.* 2012). Large scale SNP genotyping can now be readily applied to non-model species, revolutionising many areas of ecology and evolution. In particular, applications previously limited by marker number such as the construction of linkage maps (Kakawami *et al.* 2014), quantitative trait locus mapping (Schielzeth *et al.* 2011), genome-wide association studies (Slate *et al.* 2008), inference of population demographic history (Shafer *et al.* 2015) and studies of inbreeding depression (Hoffman *et al.* 2014) are increasingly benefiting from the enhanced resolution provided by SNPs. Moreover, SNP genotyping will increasingly be used to assay a large number of individuals and populations with high accuracy and low-cost in candidate genomic regions identified by genome scans from whole genome re-sequencing data.

A common approach for SNP genotyping is to mine a sequence resource for putative SNPs, extract the flanking sequences and then use these to develop locus-specific assays. Several different types of genotyping technology are available, which provide considerable flexibility in terms of the numbers of SNPs and individuals that can be typed. Small to medium throughput technologies include Applied Biosystem's SNPlex^{TM} and TaqMan® SNP genotyping assays, Sequenom's iPlex® assay, Beckman Coulter's SNPstream® and LGC's KASP^{TM} assay. Until recently, Illumina's GoldenGate® assay was also popular, but this has recently been discontinued. At the opposite end of the spectrum are high-density arrays, otherwise known as 'SNP chips', including the Illumina

Infinium iSelect® and Affymetrix Axiom® arrays, which can support several thousands to millions of SNPs. Owing to the ease with which large volumes of data can be generated, these high-density arrays are gaining popularity and have already been applied to species as diverse as house sparrows and polar bears (Hagen *et al.* 2013; Malenfant *et al.* 2014).

In humans, where large numbers of SNPs have been pre-validated, it is usual for somewhere in the order of 90% of SNPs to be polymorphic and reliably scored (Montpetit *et al.* 2005; García-Closas *et al.* 2007). However, validation rates for novel SNPs in non-model organisms tend to be much lower, falling to as little as 12.5% and rarely rising above 40% (Chancerel *et al.* 2011; Helyar *et al.* 2011). High failure rates are undesirable both from a financial perspective and due to the loss of data. Nevertheless, only a handful of studies have explored the causes of assay failure for their datasets (Lepoittevin *et al.* 2010; Van Bers *et al.* 2010; Milano *et al.* 2011) and none to our knowledge have tested for broad patterns across species. Addressing this knowledge gap should allow identification of the most common causes of assay failure and may be helpful for improving validation rates in the future.

Many of the reasons for assay failure in non-model organisms stem from the fact that SNPs are often derived *in silico* from a transcriptome or other *de novo* assembled sequence resource, and are rarely validated *in vitro*. Some studies have shown that SNPs with low *in silico* minor allele frequencies (MAF) are less likely to validate, particularly when sequence depth of coverage is low, implying that sequencing errors can sometimes be misinterpreted as SNPs (Lepoittevin *et al.* 2010; Milano *et al.* 2011). In principle, this problem can be mitigated by filtering SNPs based on MAF and depth of coverage, although this could introduce ascertainment bias. Another known cause of failure relates to the physical characteristics of the probe sequences and whether or not these are suitable for a given hybridisation technology. In this case, the use of proprietary algorithms like the Illumina assay design tool (ADT) can identify SNPs that are more likely to fail based on their flanking sequence characteristics.

Variables relating to the genomic context of a SNP are also expected to have a significant impact on validation success, particularly for transcriptome-derived SNPs. In particular, calling SNPs within contigs assembled from paralogous genes can result in probe sequences with multiple target sites in the genome, while another potentially important cause of failure is designing probes that inadvertently span intron-exon boundaries (Wang *et al.* 2008; Helyar *et al.* 2011; Milano *et al.* 2011; De Wit *et al.* 2015). A handful of studies have used reference genomes to elucidate certain aspects of the genomic context, such as proximity to intron-exon boundaries, in order to identify potentially problematic SNPs (Milano *et al.* 2011; Van Bers *et al.* 2012; Hagen *et al.* 2013). However, it is still rare for studies to take into account the genomic context, despite the increasing availability of related species' genomes and the falling cost of sequencing.

An opportunity to explore factors that influence SNP validation success in a non-model species is provided by a study of Antarctic fur seals (*Arctocephalus gazella*). On Bird Island, South Georgia, a breeding colony of this species has been studied since the 1980s, with genetic samples having been collected and analysed since the mid 1990s. To increase the genetic resolution available for studying reproductive success (Hoffman et al. 2003), mate choice (Hoffman et al. 2007) and heterozygosity-fitness correlations (Forcada & Hoffman 2014) we constructed a *de novo* transcriptome assembly from skin biopsy samples (Hoffman 2011) as well as internal organs collected at necropsy (Hoffman *et al.* 2013b). In a pilot study, we then genotyped 144 putative transcriptomic SNPs in 480 individuals using the GoldenGate assay (Hoffman *et al.* 2012). The validation rate was around 70% and, apart from a weak correlation between *in silico* MAF and validation success, most of the deviance in SNP validation could not be explained.

In this study, we present a draft fur seal genome, the first from within the pinniped family Otariidae, which we used to elucidate the genomic context of each of the GoldenGate probe sequences. Our working hypothesis was that information that can be extracted from a reference genome should account for a substantial proportion of the unexplained variation in SNP validation success. To take this approach a step further, we also revisited published studies from a variety of different species for which data on SNP validation could be analysed together with a genome sequence. Finally, we focused on a subset of the larger studies and took a predictive approach to test whether knowledge of the variables influencing SNP validation success could be helpful in improving validation rates.

Materials and methods

Draft fur seal genome

Liver tissue was collected from an adult female Antarctic fur seal that was accidentally crushed to death by a territorial bull. Following digestion with Proteinase K, high molecular weight DNA was extracted using the Qiagen Genomic-tip 100/G kit. Five paired-end libraries with insert sizes ranging from 180–230bp were constructed at the National Genomics Infrastructure (NGI) in Uppsala, Sweden following Illumina's standard TruSeq protocol. Libraries were then paired-end sequenced on an Illumina HiSeq 2500 machine with 150bp read lengths resulting in 147 gigabase pairs (Gb) of raw sequence data, 83% of which remained after removing PCR duplicates and filtering for sequences with a Phred score above 30.

We supplemented the data with seven mate-pair libraries ranging from 3–15 kilobases (kb) and one 40kb fosmid library constructed at the National Genomics Infrastructure (NGI) in Uppsalla, Sweden and the Max-Planck Institute for Developmental Biology, Tübingen, Germany. These were prepared using the Illumina Nextera mate-pair protocol (3–15kb) and the Lucigen NxSeq ® 40kb Mate Pair Cloning Kit (40kb) respectively. Libraries were indexed with different barcodes and were multiplexed across different lanes and runs. These 'jumping' libraries yielded an additional 2.26 billion read pairs (451 Gb) providing longer-distance structural information (Table 1).

In total, we fed 598 Gb of data (200x depth of coverage over a ~3 Gb genome) into ALLPATHS-LG version-R50191 with the default parameters, the haploidify option activated (HAPLOIDIFY=True) and a ploidy value set to two. ALLPATHS-LG was run on a machine equipped with 64 nodes and 2TB RAM memory at the computational infrastructure in Uppsala, UPPMAX (http://www.uppmax.uu.se). The assembly program consists of several modules executed consecutively in an automated fashion. All modules except "FixLocal", which rectifies local assembly errors, finished their computations without showing error messages. The "FixLocal" module was accordingly skipped by setting "FIX_LOCAL=False" when re-running the assembler. According to our previous experience with other vertebrate genomes (Poelstra et al. 2014) omission of this module introduces single base pair errors at a rate of less than one per megabase, thus not bearing on the analyses performed here. ALLPATHS-LG accepts raw data without prior adapter removal or trimming and performs its own read correction steps based on read quality and nucleotide content within each read. The sequencing error rate per base was estimated to be 0.0018 (Q = 27.4) and 21.85% of the raw reads were marked as duplicates. After read correction, 8.2% of the raw reads containing errors were rectified which corresponded to an average of 1.3 corrections per read. Finally, in order to identify redundant scaffolds, we used BLAT to search for identical hits of the assembly against itself.

In order to identify and annotate interspersed repeat regions within the genome, we first generated consensus models of putative repeats for the fur seal using RepeatModeler 1.0.8. The genome was then screened against this database and the vertebrate reference repeat database using RepeatMasker 4.0.3 (http://www.repeatmasker.org). To estimate the status of completeness and contiguity of the fur seal genome, we also used the program CEGMA 2.4 (Parra *et al.* 2007, Parra *et al.* 2009), which uses hidden Markov models to compare the genome assembly to a set of 248 ultra-conserved eukaryotic genes.

Variables affecting SNP validation success in fur seals

We aligned the 121bp GoldenGate probe sequences (i.e. the SNP plus 60bp flanking sequence on either side) of all 144 previously genotyped SNPs to the draft Antarctic fur seal genome using BLASTn with an e-value threshold of 1e^{-10}. To identify variables associated with successful SNP validation success, we constructed a generalized linear model (GLM). As the aim of most studies is to generate a panel of polymorphic SNPs, we modeled SNP validation success as a binary response variable coded as 1 = polymorphic and 0 = monomorphic / failed (following Conklin et al. 2013 and Montes et al. 2013). This may be somewhat conservative, as SNPs that are monomorphic in a given sample could potentially be polymorphic in a larger or different sample of individuals. The following predictor variables were fitted: number of mappings to the draft genome, alignment length, percent identity, bit score, gap opening, mismatches, e-value, Illumina ADT score, *in silico* MAF and depth of coverage, and the type of SNP (transition versus transversion). Alignment length was included as a proxy for presence of intron-exon boundaries, as a full and continuous mapping indicates that a SNP and its flanking sequences lie fully within an exon, whereas a truncated alignment to the genome could arise if the probe sequence spans an intron-exon boundary. The minimal adequate model was chosen based on standard deletion testing procedures (Crawley, 2007) where *F*-tests were used to sequentially remove each term unless doing so significantly reduced the amount of deviance explained.

To test whether the genomes of related species could provide similar insights into validation success, we repeated our analysis after blasting the probe sequences to the genomes of the walrus (*Odobenus rosmarus*) (Foote *et al.* 2015), the Weddell seal (*Leptonychotes weddellii*) (by courtesy of the Weddell Seal Genome Consortium) and the dog (*Canis lupus familiaris*) (Lindblad-Toh *et al.* 2005). We also estimated overall percentage sequence divergence directly from the genome sequences. First, we aligned the draft fur seal genome to both the walrus and the Weddell seal using LASTAL (Kielbasa *et al.* 2011). From the resulting maf alignment files we then used MafFilter (Dutheil *et al.* 2014) to calculate divergence (percentage of mismatch).

Variables affecting SNP validation success in other species

To explore the generality of our findings, we modeled validation success for additional species in which SNP assays have previously been developed and for which draft genome sequences are available. To identify these studies, we conducted Google Scholar and ISI Web of Knowledge searches (on 6th June 2015) using the following keywords: transcriptome, SNP, GoldenGate, Illumina and RAD. We retrieved a total of 22 studies, of which SNP flanking sequences, assay outcomes and genome sequences were all available for seven. Where ADT scores were not available, we generated these from the SNP flanking sequences using Illumina's assay design tool. For each study, we took the final list of SNP flanking sequences submitted for assay design and aligned these to their respective genomes using BLASTn (e-value $1e^{-10}$). GLMs were then constructed using the same predictor variables as in the fur seal model, although in most cases data were not available for *in silico* MAF, depth of coverage and the type of SNP.

Predicting SNP validation success

To test whether a subset of SNPs could be used to predict the outcome of a larger genotyping assay, we focused on five of the above studies that had genotyped at least 8,000 putative SNPs. We then took 1,000 random subsamples of 384 SNPs from each dataset. This number was chosen as a standard TaqMan® panel that represents a reasonable balance between affordability and power, although a number of alternative genotyping technologies are available (see Introduction) that can accommodate custom SNP panels of varying sizes. On each subsample, we then performed *k*-fold cross validation (5-fold, 100 times) using the *bestglm* package in R (R Core Team 2014). This approach splits the observations into $k = 5$ non-overlapping subsets of approximately equal size, uses one subset as a validation sample and the remaining four subsets as training data in order to generate the best predictive model. For each species, we took the 1,000 best models from the cross validation exercise and used the *predict* function in R to output the probability of each SNP in the full dataset successfully validating given values of the predictor variables. A given SNP was predicted as validating successfully if its associated probability value was above an arbitrary threshold of 0.7. In order to estimate the improved assay success rate, we took the SNPs that were predicted to successfully validate, and that would therefore be chosen for inclusion on a SNP assay, and determined the proportion of these that actually did.

Results

Draft fur seal genome assembly

The genome assembly (version 1) of the Antarctic fur seal, generated by ALLPATHS-LG, had a total length of 2.3Gbp excluding gaps, similar to the 2.4Gb and 2.2Gb recently assembled for the walrus and Weddell seal respectively (Table 2). The assembly consisted of a total of 144,410 contigs integrated within 8,126 scaffolds such that 50% of the final assembly was contained within the 233 longest scaffolds. Individual heterozygosity was estimated to be 6.4 x 10^{-4}, average GC content was 45.2% and repeats as estimated by k-mer analyses occupied 21.3% of the genome. Explicit repeat annotation estimated 30.2% of the genome to be repetitive with a strong representation of DNA transposons, LTR retrotransposons, LINEs and SINEs (Supplementary Table 1).

Screening the fur seal genome for the presence and integrity of ultra-conserved genes identified 80.7% of a core set of 248 eukaryotic genes as being complete (i.e. with over 70% of the gene aligning) and 94.4% as partially aligning (over at least 30% of the gene). This number compares well with several other carnivore genomes (Supplementary Table 2) and indicates that the assembly is of good quality in terms of gene content.

Variables affecting SNP validation success

To identify variables associated with the propensity of a given SNP to be successfully validated in the fur seal, we mapped the 121bp probe sequences of 144 putative SNPs genotyped in 480 individuals (Hoffman *et al.* 2012) to the draft genome. 141 of these blasted with an e-value threshold of $1e^{-10}$, allowing us to test for associations between various genomic characteristics and SNP validation success. The number of mappings, alignment length and MAF were all retained in the minimum adequate model, which explained 30.8% of the total deviance in SNP validation success (Table 3a). Specifically, we found a strong negative association between the number of mappings and validation success, together with a weaker positive correlation with alignment length and a negative association with MAF (Figure 1).

To test whether the genomes of related species could also be informative about SNP validation outcomes, we blasted the fur seal probe sequences to the draft genomes of the walrus and Weddell seal and to the dog genome. The two species of seal are thought to share a common ancestor with the Antarctic fur seal 18 and 23 MYA respectively (Higdon *et al.* 2007), corresponding to genomic sequence divergence estimates of 2.9 and 5.1% respectively (this study). The dog is thought to have shared a common ancestor with the Antarctic fur seal around 44 MYA (Hoffman *et al.* 2013a). Similar results were obtained for all three species (Table 3b–d), with the number of mappings in all cases being strongly negatively associated with validation success. However, the number of SNPs mapping to the reference genome declined with phylogenetic distance (fur seal = 99%, walrus =97%, Weddell seal = 92%, and dog = 61%).

We extended our approach to include previously published datasets from a variety of different species. Available data were collated for a total of seven species for which empirical data on SNP validation success could be analysed in combination with probe sequences and a reference genome (see Table 4 for details). These studies differ both in the number of SNPs genotyped (from 384–286,021) and in the genotyping chemistry used (GoldenGate, Infinium BeadChip and Affymetrix Axion). Moreover, the SNPs themselves were derived either from transcriptomic resources (two studies), genomic resources including reduced representation libraries (three studies) or from a combination of the two (two studies). Genome BLASTs resulted in an average of 96% of probe sequences mapping to the respective genomes. As in the fur seal, the number of mappings was retained in all of the models and alignment length was retained in all but one of the models (Table 4). There was also a tendency for studies based on larger numbers of SNPs to retain more explanatory variables, such as gap opening and bit score. The explained deviance varied from 0.25% to 9.73% and was significantly higher for studies incorporating transcriptome-derived SNPs (unpaired *t*-test, $t = -2.74$, $p = 0.04$).

Predicting SNP validation success

Finally, we investigated whether a subset of randomly selected SNPs can be effective at predicting the outcome of a larger genotyping assay. From the studies identified above, we selected five that had genotyped at least 8,000 putative SNPs and from these generated predictive models using *k*-fold cross validation based on 1,000 randomly selected subsets of 384 SNPs (see Materials and methods for details). We then used the resulting models to predict the outcome for the full dataset, assuming that SNPs with associated *p*-values greater than 0.7 would successfully validate. To explore whether this approach might be useful for improving overall validation rates, we then compared the proportion of SNPs correctly identified as validating by the model to the empirical validation rate.

For species with high initial validation rates (sunflower = 80%, soybean= 78%, rainbow trout = 86%), only a fraction of the 1,000 best predictive models retained any predictor variables and, as a consequence, selecting SNPs with a high validation probability would only yield an incremental improvement over the empirical validation rate (4%, 2% and 2% respectively, Figure 2, Table 4). Conversely, for the polar bear and salmon, which had much lower validation rates, the majority of predictive models contained at least one predictor variable (71% and 99% respectively). Using these models to select SNPs with a 70% or greater validation probability would improve the overall validation rate by 16.3% and 27% respectively, but reduce the number of SNPs to 2,549 and 2,436 respectively (Figure 2).

For comparison, we also applied a relatively crude filtering approach in which we selected only SNPs with uniquely mapping probes that align fully to the reference genome. The outcome was similar to that of the predictive approach for the trout, sunflower and soybean (Figure 2). However, for the polar bear and salmon, filtering on the basis of uniqueness and alignment length would not improve the validation rate to the same extent as predictive modeling.

Discussion

SNP assays routinely fail to validate for reasons that in general remain poorly understood. We therefore used a draft fur seal genome to explore the genomic characteristics of 144

SNP probe sequences in order to identify variables associated with the observed genotyping outcomes. We found that probes mapping multiple times to the fur seal genome and with incomplete alignments were less likely to be validated, a pattern that holds up across a variety of species. Our analyses also suggest that filtering raw SNPs on the basis of these two factors alone could help to improve validation rates, although predictive modeling based on pilot SNP data may be desirable when the validation rate is expected to be low.

The fur seal genome

An important outcome of this study is a draft Antarctic fur seal genome. This not only provides insights into factors that influence SNP validation, but should also be a useful resource for future studies of this and other pinniped species. The total scaffold length without gaps was 2.3Gb, similar to the walrus and Weddell seal assemblies. This is somewhat shorter than would be expected from the C-value of the closely related California sea lion (3.15 pg, Du & Wang 2006) and is consistent with the notion that genomes assembled using a short-read shotgun approaches lack a significant portion of highly repetitive genomic regions. We estimated a repeat content of approximately 30% for the fur seal, which is slightly lower than in the Weddell seal (40%) and several other carnivore species (30–43%, http://bit.ly/1X9Vw6z). This difference may arise from the usage of non-specific repeat databases, and/or because the Antarctic fur seal genome may lack certain repetitive regions.

The number of scaffolds assembled was intermediate between the walrus and the Weddell seal, while the scaffold N50 was the highest of the three seal species. This probably reflects the inclusion of numerous 3–15kb jumping mate-paired libraries plus the long-jump 40kb library. Unexpectedly, data from the 40kb library contributed little to the final assembly as the assembler found only 2,634 pairs usable (approx. 0.00001% of the total library reads). To investigate this further, we mapped the raw reads from the 40kb library to the fur seal, Weddell seal, walrus, dog and panda genomes using BWA-MEM 0.7.12 (Li 2013). 91.4% of the reads mapped to the fur seal assembly and this proportion decreased with increasing phylogenetic distance (Supplementary Table 3).

This suggests that the 40kb library comprises high quality fur seal sequences, yet contributes little towards further improving an already high scaffolding length from the 3–15kb libraries.

Variables affecting SNP validation success

Although relatively few studies have explored the effects of SNP characteristics on validation success, a number of factors are thought to be important. First, *in silico* parameters such as depth of sequence coverage and MAF can be informative as to whether or not a SNP is genuine (Sánchez *et al.* 2009; De Wit *et al.* 2015). Second, assembling paralogous sequences can lead to the identification of false positive SNPs, particularly for transcriptomic data (Smith *et al.* 2005; Sánchez *et al.* 2009; Cahais *et al.* 2012; Hagen *et al.* 2013; De Wit *et al.* 2015). Third, technical statistics such as the ADT score provide an indication of how likely a given probe sequence is to work in the assay. Finally, variables relating to the genomic context, including sequence uniqueness (Wang *et al.* 2008; Hagen *et al.* 2013) and proximity to intron-exon boundaries (Wang *et al.* 2008; Hoffman *et al.* 2012; Montes *et al.* 2013), are also expected to have a significant impact on validation success. Our approach attempted to elucidate the importance of the latter by essentially modeling probe hybridization to a reference genome.

The results of the fur seal analysis point towards three variables being important: the number of mappings, alignment length and *in silico* MAF. We included MAF in the model as a preliminary analysis found it to be negatively associated with validation success (Hoffman *et al.* 2012). The number of mappings was by far the most important explanatory variable, suggesting that probe sequence uniqueness is a key factor to consider in SNP development. Alignment length explained a smaller proportion of the total deviance but was nonetheless highly significant, a positive relationship with validation success indicating that SNPs with completely mapping probes are more likely to result in clearly interpretable and polymorphic genotyping assays. Both of these variables were also significantly associated with SNP validation success in all but one of the seven additional species examined. By implication, it appears to be commonplace for studies to include SNPs with probe sequences that are not unique or which span intron-exon boundaries.

One reason for this general pattern could be that many of the studies we examined incorporated transcriptomic SNPs. These can be problematic due to *de novo* assembly artefacts (Gayral *et al.* 2011) and because intron-exon boundaries cannot usually be identified without reference to some form of genomic sequence. However, the same two variables were also associated with validation success in the Atlantic salmon and the soybean, species for which SNPs were developed exclusively from genomic resources. Although the exact reason for this remains unclear, it seems probable that many forms of genomic data will also be affected to a certain extent by assembly artefacts. This could be exacerbated by the fact that both the salmon and the soybean have undergone recent increases in genome ploidy (Shoemaker *et al.* 1996; Davidson *et al.* 2010).

Explained deviance

The proportion of deviance explained by our models varied considerably among the seven species, from 0.25 to 9.73%. To explore why, we constructed a GLM of the proportion of deviance explained, fitting as explanatory variables the overall validation rate of the assay, the total number of SNPs, the number of variables retained in each model, and the source of the SNPs (including or excluding transcriptomic resources). We found a weak tendency for studies with larger numbers of SNPs to retain more variables in the minimum adequate model ($\chi^2 = 13.76$, d.f = 1, $p = 0.08$), reflecting the greater power of large datasets to capture relatively subtle effects. In addition, significantly more deviance could be explained for studies that included SNPs developed from transcriptomic resources ($\chi^2 = 32.74$, d.f = 1, $p = 0.02$). Taken at face value, this suggests that particular care should be taken when developing SNPs from transcriptomes. However, direct comparison is made difficult by the fact that no two studies use the same SNP discovery pipeline, and the two purely genomic studies both incorporated pre-validated SNPs.

Predictive power

We used the five largest SNP datasets to explore whether knowledge of the factors that influence SNP validation success could be used to improve overall validation rates. Given that probe uniqueness and alignment length appear to be consistently associated with validation success across species, we first compared the empirical validation rate of the full dataset with that of a dataset filtered to contain only uniquely and completely mapping SNPs. Success rates of the filtered SNPs were consistently higher, suggesting that even relatively crude filtering based on these two variables alone could help to improve validation rates. As expected, the greatest expected improvement was observed for the salmon, which had the lowest empirical validation rate and hence the greatest room for improvement.

Although the number of mappings and alignment length were retained in most of our models, several other parameters were also found to be important, and these varied from species to species. To integrate all of the available information for each species into a predictive framework, we therefore constructed predictive models using a *k*-fold cross-validation approach. To determine the potential for improvement, we then compared the proportion of SNPs correctly identified as validating by these models to the empirical validation rate. For the trout, soybean and sunflower, selecting SNPs with a validation probability of 0.7 had a similar outcome to filtering SNPs for unique and complete probe alignments. In contrast, for the polar bear and the salmon, which experienced lower overall validation rates, the predictive approach could increase the validation rate by up to around 30%.

Which of these two approaches are best for a particular system will depend on several considerations. Our results suggest that filtering a collection of 'raw' SNPs based on the number of mappings and alignment length is likely to improve the validation rate under most circumstances and this requires minimal effort. In contrast, predictive modeling requires an investment in generating a pilot SNP dataset, but offers greater scope for improving the validation rate when this is expected to be low, for instance when many or all of the SNPs are developed from a transcriptome. However, higher validation rates also come at the cost of fewer SNPs being available for genotyping (Figure 2). How this trade-off between SNP quality and quantity is resolved will differ on a case-by-case basis, although raw SNPs can now be generated in such large numbers that their availability will in many cases not be limiting.

Overall, our study reveals considerable differences among species, both in the explanatory power of different variables and in the potential improvement that could be achieved by pre-selecting SNPs based on prior knowledge of how different variables affect SNP validation. As expected, both explanatory and predictive power correlate negatively with the overall validation rate, which in turn appears to depend on whether or not a given study includes transcriptomic SNPs. This suggests that mapping SNPs to a reference genome may bring the greatest practical benefits where efforts are underway to develop SNP arrays primarily from a transcriptome. However, this is a relatively common endeavor, as transcriptomes provide a rapid and inexpensive means of SNP discovery, as well as a convenient route for mining markers within candidate genes.

Caveats

Genome sequences are not always available and are still challenging or in some cases impossible to generate due to the requirement for large amounts of high quality DNA (Ekblom & Wolf 2014). Nevertheless, our results suggest that, when possible, mapping probe sequences to the genome of a related species may provide useful information on the genomic context. We were able to map most of the fur seal probe sequences to the walrus and Weddell seal genomes, which are divergent by 2.9 and 5.1% respectively, generating qualitatively similar model outputs. Thus, with increasing numbers of non-model species having their genomes sequenced and assembled as part of initiatives like the Genome 10k project (Genome 10K Community of Scientists 2009), growing numbers of studies should at least be able to access the genome of a related species. Failing that, genomic data, even if unassembled, can also be informative in some respects. For instance, a recent study mapped genomic shotgun reads to a transcriptome to help identify intron-exon boundaries (Montes *et al.* 2013).

Another point to bear in mind is that the GoldenGate assay, which we used to identify the main factors affecting SNP validation and to populate a predictive model, has recently been phased out. However, this does not negate our main finding that the genomic context of a SNP appears to affect validation success across a range of species. In addition, although we used a pilot GoldenGate dataset to build a predictive model, several alternative technologies are available that allow similar sized custom SNP panels to be genotyped. We have no reason to believe that these alternative technologies could not be used to similar effect, especially given that the predictive approach integrates diverse information about each SNP, including the genomic context and the likely performance with a specific genotyping technology.

Finally, reduced representation approaches such as targeted amplicon resequencing, Restriction Site Associated (RAD) DNA sequencing (Hohenlohe *et al.* 2010; Peterson *et al.* 2012) and genotyping-by-sequencing (Narum *et al.* 2013) provide alternatives to custom SNP arrays. The method of choice for a given study will depend on a number of factors including cost, the number and specificity of markers required and ease of implementation. RAD sequencing is growing in popularity as it can generate tens of thousands of randomly distributed SNPs in virtually any organism without the need for prior genomic information. However, RAD sequencing is arguably less straightforward than custom SNP genotyping due to the technical difficulty and cost of library preparation and the need for extensive post-processing. Moreover, high-density SNP arrays have very low rates of genotyping error, can target specific genomic regions, generate data with high inter-individual concordance, and can be more easily scaled up to sample sizes of many thousands of individuals. For these and other reasons, custom SNP arrays have an important role to play in the future of the field of molecular ecology (Andrew *et al*. 2013) and are likely to remain the method of choice for large-scale, individual-based studies of natural populations for years to come. Having said that, reduced representation sequencing approaches are increasingly being used to discover SNPs for use in custom arrays (Houston *et al.* 2014; Malenfant *et al.* 2014; Palti *et al.* 2014) and our approach has also been applied in this context.

Conclusions

We used the Antarctic fur seal as a case study to show that mapping probe sequences to a draft reference genome can identify variables with a large effect on SNP validation success. We also demonstrate the potential for filtering and predictive approaches to improve genotyping outcomes, particularly when some or all of the markers are derived from a transcriptome.

References

Andrew RL, Bernatchez L, Bonin A, Buerkle CA *et al.* (2013) A road map for molecular ecology. *Molecular Ecology*, **22**, 2605–2626.

Bachlava E, Taylor CA, Tang S *et al.* (2012) SNP Discovery and Development of a High-Density Genotyping Array for Sunflower. *PLoS ONE*, **7**, e29814.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, **18**, 249–256.

Cahais V, Gayral P, Tsagkogeorga G *et al.* (2012) Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Molecular Ecology Resources,* **12**, 834–845.

Chancerel E, Lepoittevin C, Le Provost G *et al.* (2011) Development and implementation of a highly- multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine. *BMC Genomics,* **12**, 368.

Conklin D, Montes I, Albaina A, Estonba A (2013) Improved conversion rates for SNP genotyping of non-model organisms. IWBBIO 2013: International Work-Conference on Bioinformatics and Biomedical Engineering, Granada, Spain. ISBN:GR 489/2013, 127–134.

Crawley MJ (2007) The R Book. Wiley, Chichester.

Davidson WS, Koop BF, Jones SJM *et al.* (2010) Sequencing the genome of the Atlantic salmon (*Salmo salar*). *Genome Biology,* **11,** 1–7.

De Wit P, Pespeni MH, Palumbi SR (2015) SNP genotyping and population genomics from expressed sequences - current advances and future possibilities. *Molecular Ecology*, **24**, 2310–2323.

Du B, Wang D (2006) C-values of Seven Marine Mammal Species Determined by Flow Cytometry. *Zoological Science*, **23**, 1017–1020.

Dutheil JY, Gaillard S, Stukenbrock EH (2014) MafFilter: a highly flexible and extensible multiple genome alignment files processor. *BMC Genomics* 2014, **15,** 53.

Ekblom R, Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications,* **7**, 1026–1042.

Foote AD, Liu Y, Thomas GWC *et al.* (2015) Convergent evolution of the genomes of marine mammals. *Nature Genetics,* **47**, 272–275.

Forcada J, Hoffman JI (2014) Climate change selects for heterozygosity in a declining fur seal population. *Nature,* **511**, 462–465.

García-Closas M, Malats N, Real FX *et al.* (2007) Large-scale evaluation of candidate genes identifies associations between VEGF polymorphisms and bladder cancer risk. *PLoS Genetics*, **3**, e29.

Gayral P, Weinert L, Chiari Y *et al.* (2011) Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. *Molecular Ecology Resources*, **11**, 650–661.

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

Hagen IJ, Billing AM, Rønning B *et al.* (2013) The easy road to genome-wide medium density SNP screening in a non-model species: development and application of a 10 K SNP-chip for the house sparrow (*Passer domesticus*). *Molecular Ecology Resources*, **13**, 429–439.

Helyar SJ, Hemmer-Hansen J, Bekkevold D *et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, **11**, 123–136.

Higdon JW, Bininda-Emonds OR, Beck RM, Ferguson SH (2007) Phylogeny and divergence of the pinnipeds (Carnivora: Mammalia) assessed using a multigene dataset. *BMC Evolutionary Biology*, **7**, 216.

Hoffman JI, Boyd IL, Amos W (2003) Male reproductive strategy and the importance of maternal status in the Antarctic fur seal Arctocephalus gazella. *Evolution*, **57**, 1917–1930

Hoffman JI, Boyd IL, Amos W (2004) Exploring the relationship between parental relatedness and male reproductive success in the Antarctic fur seal Arctocephalus gazella. *Evolution*, **58**, 2087–2099.

Hoffman JI, Forcada J, Amos W (2006) No relationship between microsatellite variation and neonatal fitness in Antarctic fur seals, Arctocephalus gazella. *Molecular Ecology*, **15**, 1995–2005.

Hoffman JI, Forcada J, Trathan PN, Amos W (2007) Female fur seals show active choice for males that are heterozygous and unrelated. *Nature*, **445**, 912–914.

Hoffman JI (2011) Gene discovery in the Antarctic fur seal (Arctocephalus gazella) skin transcriptome. *Molecular Ecology Resources*, **11**, 703–710.

Hoffman JI, Tucker R, Bridgett SJ *et al.* (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. *Molecular Ecology Resources*, **12**, 861–872.

Hoffman JI, Thorne MA, McEwan R, Forcada J, Ogden R (2013a) Cross-amplification and validation of SNPs conserved over 44 million years between seals and dogs. *PLoS ONE*, **8**, e68365.

Hoffman JI, Thorne MA, Trathan PN, Forcada J (2013b) Transcriptome of the dead: characterisation of immune genes and marker development from necropsy samples in a free-ranging marine mammal. *BMC Genomics*, **14**, 1–27.

Hoffman JI, Simpson F, David P *et al.* (2014) High-throughput sequencing reveals inbreeding depression in a natural population. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 3775–3780.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics,* **6**, e1000862.

Houston RD, Taggart JB, Zard TC *et al.* (2014) Development and validation of a high density SNP genotyping array for Atlantic salmon (*Salmo salar*). *BMC Genomics*, **15**, 90.

Jorgenson E, White JS (2006) A gene-centric approach to genome-wide association studies. *Nature Reviews Genetics,* **7**, 885–891.

Kakawami T, Backström N, Burri R *et al.* (2014) Estimation of linkage disequilibrium and interspecific gene flow in *Ficedula* flycatchers by a newly developed 50K single-nucleotide polymorphism array. *Molecular Ecology Resources,* **14**, 1248–1260.

Kiełbasa SM, Wan R, Sato K *et al*. (2011) Adaptive seeds tame genomic sequence comparison. *Genome Resources,* **21**, 487–493.

Kruglyak L, Nickerson DA (2001) Variation is the spice of life. *Nature Genetics*, **27**, 234–236.

Lapègue S, Harrang E, Heurtebise S *et al.* (2014) Development of SNP-genotyping arrays in two shellfish species. *Molecular Ecology Resources*, **14**, 820–830.

Lepoittevin C, Frigerio J-M, Garnier-Géré P *et al.* (2010) In vitro vs in silico detected SNPs for the development of a genotyping array: what can we learn from a non-model species? *PLoS ONE,* **5**, e11034.

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN]

Lindblad-Toh K, Wade CM, Mikkelsen TS *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature,* **438**, 803–819.

Malenfant RM, Coltman DW, Davis CS (2014) Design of a 9K illumina BeadChip for polar bears (*Ursus maritimus*) from RAD and transcriptome sequencing. *Molecular Ecology Resources*, **15,** 587–600.

Milano I, Babbucci M, Panitz F *et al.* (2011) Novel tools for conservation genomics: comparing two high-throughput approaches for SNP discovery in the transcriptome of the European hake. *PLoS ONE,* **6**, e28008.

Montes I, Conklin D, Albaina A *et al.* (2013) SNP discovery in European anchovy (*Engraulis encrasicolus*) by high-throughput transcriptome and genome sequencing. *PLoS ONE*, **8**, e70051.

Montpetit A, Nelis M, Laflamme P *et al.* (2005) An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PLoS Genetics*, **2**, e27

Morin PA, Luikart G, Wayne RK, the SNP workshop group (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, **19**, 208–216.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.

Palti Y, Gao G, Liu S *et al.* (2014) The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. *Molecular Ecology Resources*, **15,** 1755–0998.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes, *Bioinformatics* **23**, 1061–1067.

Parra G, Bradnam K, Ning Z *et al.* (2009) Assessing the gene space in draft genomes. *Nucleic Acids Resources,* **37**, 289–297.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Poelstra JW, Vijay N, Bossu CM *et al.* (2014) The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, **344**, 1405–1410.

R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

Sánchez C, Smith TP, Wiedmann RT *et al.* (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 559.

Schielzeth H, Forstmeier W, Bart K, Ellegren H (2011) QTL linkage mapping of wing length in zebra finch using genome-wide single nucleotide polymorphisms markers. *Molecular Ecology*, **21**, 329–339.

Seeb JE, Carvalho GR, Hauser L *et al.* (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, **11**, 1–8.

Shafer ABA, Gattepaille LM, Stewart REA, Wolf JBW (2015) Demographic inferences using short-read genomic data in an approximate Bayesian computation framework: in silico evaluation of power, biases and proof of concept in Atlantic walrus. *Molecular Ecology,* **24**, 328–345.

Shoemaker RC, Polzin K, Labate J *et al.* (1996) Genome duplication in soybean (*Glycine subgenus soja*). *Gene*, **144**, 329–338.

Slate J, Gratten J, Beraldi D *et al.* (2008) Gene mapping in the wild with SNPs: guidelines and future directions. *Genetica*, **136**, 97–107.

Smith CT, Elfstrom CM, Seeb LW, Seeb JE (2005) Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. *Molecular Ecology*, **14**, 4193–4203.

Van Bers NEM, Santure AW, van Oers K *et al.* (2012) The design and cross-population application of a genome-wide SNP chip for the great tit *Parus major*. *Molecular Ecology Resources*, **12**, 753–770.

Van Bers NEM, van Oers K, Kerstens HHD *et al.* (2010) Genome-wide SNP detection in the great tit *Parus major* using high throughput sequencing. *Molecular Ecology*, **19**, 89–99.

Wang S, Sha Z, Sonstegard TS *et al.* (2008) Quality assessment parameters for EST-derived SNPs from catfish. *BMC Genomics*, **9**, 450.

Figures

Figure 1: Fur seal SNP validation success in relation to the three predictor variables retained in the minimal adequate model: a) number of mappings, b) alignment length and c) *in silico* MAF. Circle size is proportional to frequency and the shaded areas indicate 95% confidence intervals.

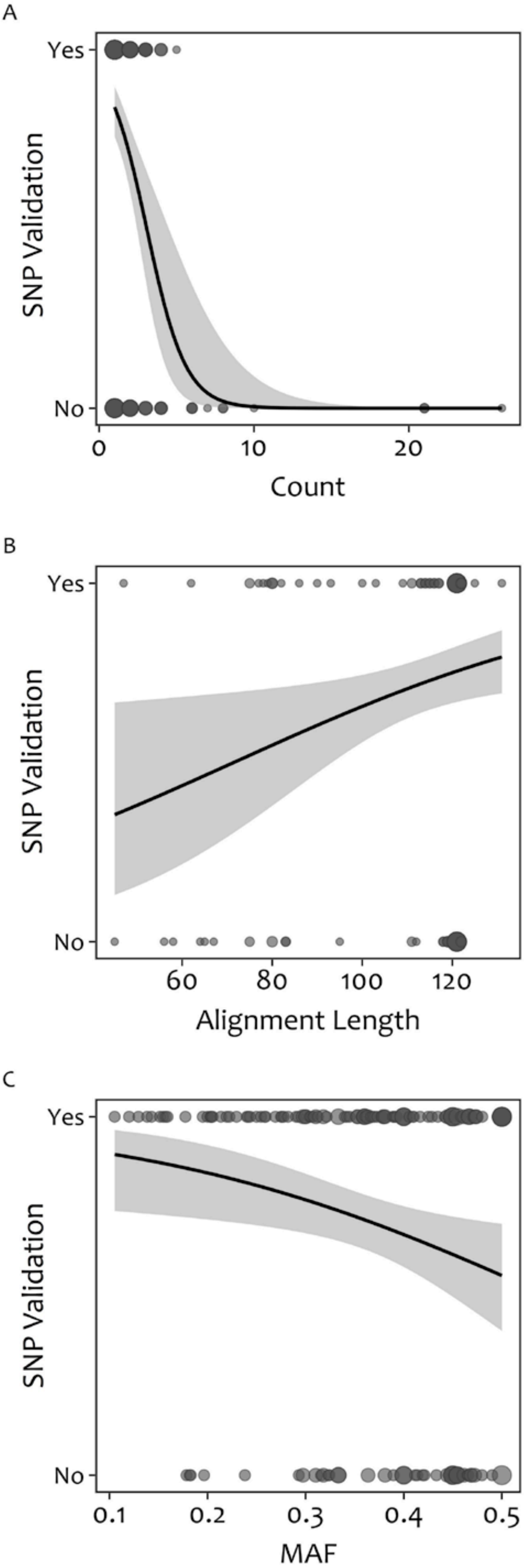
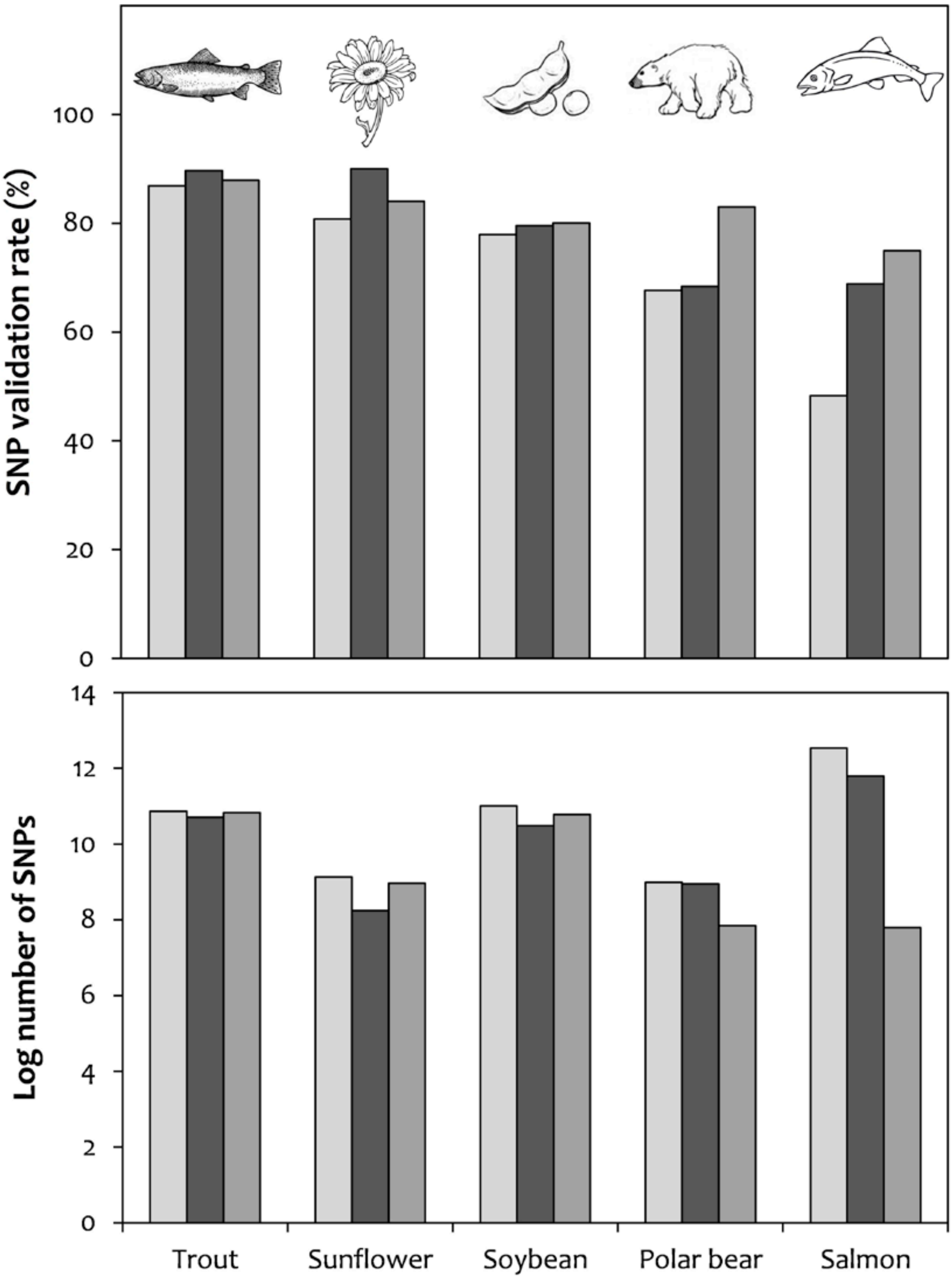

A
Yes
No
SNP Validation
0
10
20
Count
B
Yes
No
SNP Validation
60
80
100
120
Alignment Length
C
Yes
No
SNP Validation
0.1
0.2
0.3
0.4
0.5
MAF

Figure 2: Percent and number of successful SNPs for studies where filtering and predictive modeling approaches were applied (see Materials and methods for details). Light grey bars refer to the observed assay outcomes; dark grey bars refer to assay outcomes following filtering on the basis of the number of mappings and alignment length; medium grey bars indicate the outcomes after selecting SNPs on the basis of predictive models. The studies are ordered from left to right by the observed validation rate.



Acknowledgements

We thank the Broad Institute Genomics Platform, the Weddell Seal Genome Consortium and Kerstin Lindblad-Toh for making the data for *Leptonychotes weddellii* available. We would also like to thank Christa Lanz and Axel Künster from the Tübingen sequencing facility for library preparation and data transfer and the Tübingen Genome Center for fosmid sequencing. We are further indebted to Marc Höppner for running RepeatMasker on the fur seal assembly. We are grateful to Ross Houston and Caird Rexroad for providing additional raw SNP data for the Atlantic salmon and rainbow trout respectively. This work contributes to the Ecosystems project of the British Antarctic Survey, Natural Environment Research Council, and is part of the Polar Science for Planet Earth Programme. This work was supported by a Marie Curie FP7-Reintegration-Grant within the 7th European Community Framework Programme (PCIG-GA-2011-303618 to JIH), core funding from the Natural Environment Research Council to the British Antarctic Survey's Ecosystems Program, the Knut and Alice Wallenberg Foundation to the Wallenberg Advanced Bioinformatics Infrastructure (to JBWW and AMB) and the Swedish Research Council FORMAS (231-2012-450 to JBWW).

Author Contributions

JIH, JBWW, EH conceived and designed the study; JF, PNT, JIH contributed reagents / materials; EH, MAST conducted the analyses; DW, AM-B, JIH, JBWW conducted the genome sequencing and assembly; EH, JIH, AM-B, JBWW wrote the paper. All authors commented on and approved the final manuscript.

Data Accessibility

The Illumina reads have been submitted to the short read archive (http://www.ncbi.nlm.nih.gov/sra) under accession number SRP064853. The draft genome assembly and SNP sequences have been uploaded to Dryad (doi:10.5061/dryad.599f2). The authors declare no competing financial interests. Correspondence should be addressed to E.H (emily.humble@uni-bielefeld.de).

Tables

Table 1. Summary statistics for the sequencing libraries used for the Antarctic fur seal genome assembly.

Library type	**Insert size**	**Read length (bp)**	**Raw data (Gb)**	**Data used (%)**	**Sequence coverage (x)**	**Physical coverage (x)**
paired	180	150	29.20	83.4	10.6	6.6
paired	180	150	27.73	82.1	9.9	6.2
paired	199	150	48.75	82.4	17.5	12.0
paired	200	150	12.11	88.9	4.7	3.2
paired	231	150	29.13	84.4	10.7	8.4
	Total	**--**	**146.92**	**83.4**	**53.5**	**36.5**
jump	3kb	100	151.16	48.2	31.3	313.3
jump	4kb	100	21.45	61.5	5.8	75.7
jump	5kb	100	40.98	46.2	8.3	114.6
jump	6kb	100	101.00	54.7	24.4	473.8
jump	8kb	100	56.63	55.2	13.8	373.6
jump	10kb	100	40.51	61.1	10.9	361.3
jump	15kb	100	13.38	62.5	3.7	19.1
long-jump	40kb*	100	26.42	0.0	0.0	0.0
	Total	**--**	**451.53**	**52.4**	**98.2**	**1731.4**

Further details of the scaffolding with the 40kb library are given in Materials and methods and Results sections.

Table 2. Genome assembly statistics for the *de novo* assembly of the Antarctic fur seal and for two previously assembled pinniped species, the walrus and Weddell seal.

	Fur Seal	**Walrus**	**Weddell Seal**
Total sequence length including gaps	2,405,038,055	2,500,048,309	3,156,902,762
Total sequence length excluding gaps	2,289,802,102	2,400,150,193	2,223,164,129
Number of scaffolds	8,126	3,893	16,711
Scaffold N50	3,169,165	2,616,778	904,031
Number of contigs	144,410	70,655	169,547
Contig N50	27,432	89,951	23,644

Table 3. Logistic regressions of fur seal SNP validation success after blasting to fur seal, Weddell seal, walrus and dog genomes. Predictor variables retained in the minimal adequate models are given together with model estimates, χ^2 values for goodness of fit tests.

(a) Antarctic fur Seal: *n* = 142, total deviance = 170.69, residual deviance = 118.11, explained deviance = 30.80%

Terms fitted in the full model: Number of mappings, percent identity, bit score, gap opening, alignment length, e-value, mismatches, ADT score, MAF, depth & SNP type

	Estimate	χ^2	*d.f*	*p*
Number of mappings	-0.86	40.80	1	1.69e-10 ***
Alignment Length	0.03	6.67	1	0.01 **
MAF	-7.54	9.46	1	0.002 **

(b) Walrus: *n* = 140, total deviance = 169.31, residual deviance = 114.08, explained deviance = 32.62%

Terms fitted in the full model: Number of mappings, percent identity, bit score, gap opening, alignment length, e-value, mismatches, ADT score, MAF, depth & SNP type

	Estimate	χ^2	*d.f*	*p*
Number of mappings	-1.01	43.25	1	4.81e-11 ***
Bit score	0.02	9.83	1	0.0017 **
MAF	-6.86	7.74	1	0.005 **

(c) Weddell Seal: *n* = 133, total deviance = 159.14, residual deviance = 114.50, explained deviance = 28.05%

Terms fitted in the full model: Number of mappings, percent identity, bit score, gap opening, alignment length, e-value, mismatches, ADT score, MAF, depth & SNP type

	Estimate	χ^2	*d.f*	*p*
Number of mappings	-0.95	30.67	1	3.06e-08 ***
Bit score	0.09	6.53	1	0.01 *

Alignment length	-0.14	4.48	1	0.03 *
Mismatches	0.57	5.48	1	0.02 *
MAF	-7.27	9.01	1	0.003 **

(d) Dog: *n* = 88, total deviance = 105.03, residual deviance = 70.34, explained deviance = 33.01%

Terms fitted in the full model: Number of mappings, percent identity, bit score, gap opening, alignment length, e-value, mismatches, ADT score, MAF, depth & SNP type

	Estimate	χ^2	*d.f*	*p*
Number of mappings	-1.17	24.29	1	68.28e-07 ***
Mismatches	0.25	6.87	1	0.009 **
MAF	-9.10	9.17	1	0.002 **

Table 4. Logistic regressions of SNP validation, showing the predictor variables retained in the minimal adequate models together with model estimates, χ^2 values for goodness of fit tests. The terms fitted in each model, the source of the SNPs and genotyping technology are given for each species. Studies are presented in ascending order of the number of SNPs.

(a) Rainbow Trout (Sánchez *et al.* 2009)**:** *n* =347, total deviance = 481.02, residual deviance = 458.16, explained deviance = 4.75%

Terms fitted in the full model: Number of mappings, percent identity, bit score, gap opening, alignment length, e-value, mismatches and ADT score. SNP source: genomic; Genotyping technology: Illumina GoldenGate

Predictor variable	Estimate	χ^2	*d.f*	*p*
Gap Opening	-4.41e-01	9.17	1	0.002 **
Alignment Length	2.55e-02	20.04	1	4.45e-05 ***
E value	2.52	15.10	1	0.0005 ***

(b) Pacific Oyster (Lapègue *et al.* 2014)**:** *n* = 364, total deviance = 488.63, residual deviance = 441.06, explained deviance = 9.73%

Terms fitted in the full model: Number of mappings, percent identity, bit score, gap opening, alignment length, e-value, mismatches and ADT score. SNP source: transcriptomic; Genotyping technology: Illumina GoldenGate

Predictor variable	Estimate	χ^2	*d.f*	*p*
Number of mappings	-2.50e-01	3.60	1	0.05 *

Bit score	1.03e-02	6.71	1	0.01 **
E value	-1.69	21.20		4.14e-06 ***
ADT score	2.52	8.51	1	0.003 **

(c) Polar Bear (Malenfant *et al.* 2014): *n* = 8,033, total deviance = 10,112.20, residual deviance = 9,656.50, explained deviance = 4.50%

Terms fitted in the full model: Number of mappings, percent identity, bit score, gap opening, alignment length, e-value and mismatches and ADT score. SNP source: genomic and transcriptomic; Genotyping technology: Illumina Infinium BeadChip

Predictor variable	Estimate	χ^2	*d.f*	*p*
Number of mappings	-2.62e-05	14.14	1	0.0002 ***
Bit score	-1.24	23.67	1	1.15e-06 ***
Gap opening	-9.64	12.59	1	0.0004 ***
Alignment length	1.82	5.47	1	0.02 *
E value	-1.11	5.28	1	0.02 *
Mismatches	-7.36	32.56	1	1.16e-08 ***

(d) Sunflower (Bachlava *et al.* 2012)**:** *n* = 9,198, total deviance = 9,003.40, residual deviance = 8,520.40, explained deviance = 5.36%

Terms fitted in the full model: Number of mappings, percent identity, bit score, gap opening, alignment length, e-value, mismatches and ADT score. SNP source: transcriptomic; Genotyping technology: Illumina GoldenGate

Predictor variable	Estimate	χ^2	*d.f*	*p*
Number of mappings	-0.01	47.41	1	5.74e-12 ***
Percent identity	0.11	59.78	1	1.06e-14 ***
Alignment length	0.03	391.02	1	< 2.2e-16 ***
ADT score	1.15	4.88	1	0.03 *

(e) Rainbow Trout (Palti *et al.* 2014): *n* = 52,298, total deviance = 40,567.00, residual deviance = 40,336.00, explained deviance = 0.25%.

Terms fitted in the full model: Number of mappings, percent identity, bit score, gap opening, alignment length, e-value, mismatches and ADT score. SNP source: genomic; Genotyping technology: Affymetrix Axion Array

Predictor variable	Estimate	χ^2	*d.f*	*p*
Number of mappings	-2.68e-03	130.95	1	< 2.2e-16 ***
Percent identity	2.72e-01	8.19	1	0.004 **

Bit score	-4.54e-02	3.57	1	0.05 *
Gap opening	-3.97e-01	15.02	1	0.0001 ***
Alignment length	8.70	4.19	1	0.04 *

(f) Soybean (Song *et al.* 2013): *n* = 60,406, total deviance = 63,747.00, residual deviance = 62,954.00, explained deviance = 1.24%

Terms fitted in the full model: Number of mappings, percent identity, bit score, gap opening, alignment length, e-value, mismatches and ADT score. SNP source: genomic; Genotyping technology: Illumina Infinium BeadChip

Predictor variable	Estimate	χ^2	*d.f*	*p*
Number of mappings	-0.0002	16.34	1	5.31e-05 ***
Bit score	-0.09	9.93	1	0.002 **
Gap opening	-1.22	22.64	1	1.95e-06 ***
Alignment length	0.16	10.40	1	0.001 **
Mismatches	-0.60	15.33	1	8.99e-05 ***
ADT score	1.41	617.97	1	< 2.2e-16 ***

(g) Atlantic Salmon (Houston *et al.* 2014): n = 277,363 , total deviance = 384,177, residual deviance = 365,848, explained deviance = 4.77%

Terms fitted in the full model: Number of mappings, percent identity, bit score, gap opening, alignment length, e-value, mismatches and p-convert score. SNP source: genomic and transcriptomic; Genotyping technology: Affymetrix Axiom Array

Predictor variable	Estimate	χ^2	*d.f*	*p*
Number of mappings	-2.50e-03	1038.9	1	< 2.2e-16 ***
Percent Identity	5.29e-01	17.63	1	2.69e-05***
Bit Score	-1.19e-01	20.94	1	4.75e-06 ***
Gap opening	-7.17e-01	34.97	1	3.36e-09 ***
Alignment length	2.81e-01	38.01	1	7.01e-10 ***
E value	5.76	21.88	1	2.89e-06 ***
Mismatches	-2.88e-01	13.10	1	0.00030 ***
P-convert score	2.84	11843	1	< 2.2e-16 ***

Supplementary tables

Supplementary Table 1: Classification of annotated repeats. Proportions were obtained by dividing the total amount in the class by the total genome size without gaps (2,289,802,102 bp).

Class	Number	Total length (bp)	Percentage (%)
Simple_repeat	942,790	41,152,646	1.8
LINE	646,619	396,340,460	17.3
SINE	592,208	109,770,282	4.8
LTR retrotransposon	205,568	88,740,969	3.9
Low_complexity	171,170	8,522,988	0.4
DNA transposon	159,031	43,923,340	1.9
Unknown	19,307	2,406,853	0.1
snRNA	1,507	97,062	<0.01
Satellite	686	187,976	0.01
RC_Helitron	493	125,994	0.01
RNA	401	88,086	<0.01
tRNA	202	13,185	<0.01
rRNA	140	28,821	<0.01
srpRNA	21	5,121	<0.01
scRNA	14	1,322	<0.01
Retroposon	5	417	<0.01
Other	2	178	<0.01
Total	2,739,962	691,405,700	30,20%

Supplementary Table 2: Results of ultra-conserved gene analyses of the Antarctic fur seal and four other carnivore genomes using CEGMA (see Materials and methods for details). Shown are the numbers of ultra-conserved genes aligning completely (>70% aligned) or partially (>30% aligned) together with percentages in parentheses.

	Fur seal	**Walrus**	**Weddell seal**	**Panda**	**Dog**
Complete	200 (80.7)	210 (84.7)	188 (75.8)	202 (81.5)	209 (84.3)
Partial	234 (94.4)	236 (95.2)	241 (97.2)	232 (93.6)	236 (95.2)

Supplementary Table 3: Number of reads mapping uniquely against various carnivore genomes together with percentage (in parentheses), from a total of 264,193,552 raw reads from the 40kb library. Results are shown for when both reads within a pair have mapped and for when only one read within a pair has mapped.

	Fur seal	**Walrus**	**Weddell seal**	**Panda**	**Dog**
Both read pairs	172,566,600 (65.3)	174,555,600 (66.1)	166,934,873 (63.2)	127,695,832 (48.3)	83,347,011 (31.5)
One read pair	137,582,166 (52.1)	140,364,378 (53.1)	132,708,980 (50.2)	84,137,162 (31.9)	42,490,638 (16.1)