MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2012) **12**, 861–872 doi: 10.1111/j.1755-0998.2012.03158.x

Rates of assay success and genotyping error when single nucleotide polymorphism genotyping in non-model organisms: a case study in the Antarctic fur seal

J. I. HOFFMAN,* R. TUCKER,† S. J. BRIDGETT,‡ M. S. CLARK,§ J. FORCADA§ and J. SLATE†
*Department of Animal Behaviour, University of Bielefeld, Postfach 100131, 33501 Bielefeld, Germany, †Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK, ‡The GenePool, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK, §British Antarctic Survey, Natural Environment Research Council, High Cross, Madingley Road, Cambridge CB3 0ET, UK

Abstract

Although single nucleotide polymorphisms (SNPs) are increasingly being recognized as powerful molecular markers, their application to non-model organisms can bring significant challenges. Among these are imperfect conversion rates of assays designed from *in silico* resources and the enhanced potential for genotyping error relative to pre-validated, highly optimized human SNPs. To explore these issues, we used Illumina's GoldenGate assay to genotype 480 Antarctic fur seal (*Arctocephalus gazella*) individuals at 144 putative SNPs derived from a 454 transcriptome assembly. One hundred and thirty-five polymorphic SNPs (93.8%) were automatically validated by the program GenomeStudio, and the initial genotyping error rate, estimated from nine replicate samples, was 0.004 per reaction. However, an almost tenfold further reduction in the error rate was achieved by excluding 31 loci (21.5%) that exhibited unclear clustering patterns, manually editing clusters to allow rescoring of ambiguous or incorrect genotypes, and excluding 18 samples (3.8%) with unreliable genotypes. After stringent quality filtering, we also found a counter-intuitive negative relationship between *in silico* minor allele frequency and the conversion rate, suggesting that some of our assays may have been designed from paralogous loci. Nevertheless, we obtained over 45 000 individual SNP genotypes with a final error rate of 0.0005, indicating that the GoldenGate assay is eminently capable of generating large, high-quality data sets for non-model organisms. This has positive implications for future studies of the evolutionary, behavioural and conservation genetics of natural populations.

Keywords: 454 sequencing, Antarctic fur seal, *Arctocephalus gazella*, BeadXpress, conversion rate, genotyping error, GoldenGate assay, heterozygosity, Illumina, marine mammal, pinniped, single nucleotide polymorphism, transcriptome, VeraCode

Received 24 February 2012; revision received 30 March 2012; accepted 4 April 2012

Introduction

Single nucleotide polymorphisms (SNPs) are rapidly becoming the marker of choice for many molecular ecological studies. They are the most abundant source of genetic variation in most if not all genomes, with around 4 million SNPs having been validated in humans (Sobrino *et al.* 2005) and the total number probably exceeding 10 million (Kruglyak & Nickerson 2001). SNPs have a low enough mutation rate to effectively preclude recurrent mutations, making them largely bi-allelic (Krawczak 1999). This not only facilitates high-throughput genotyping, but also makes them analytically highly tractable (Brumfield *et al.* 2003). Other desirable properties including co-dominant inheritance, ease of calibration across laboratories and the ability to target SNPs within specific regions of the genome (Morin *et al.* 2004) suggest that these markers will become increasingly prominent in evolutionary, behavioural and conservation genetic studies.

Classical approaches to SNP discovery in non-model organisms include Sanger sequencing random genomic DNA fragments (Bensch *et al.* 2002; Primmer *et al.* 2002; Seddon *et al.* 2005) and targeting conserved regions of orthologous sequences from closely related species to amplify the intervening introns (Aitken *et al.* 2004). However, these approaches typically yield modest numbers of markers for a large experimental effort. An alternative is to exploit emerging 'next-generation' sequencing methods to generate large *in silico* resources such as transcriptome assemblies (Vera *et al.* 2008). From these it is usually possible to identify many thousands of

Correspondence: Joseph I. Hoffman, Fax: +49 (0)521 1062998; E-mail: joseph.hoffman@uni-bielefeld.de

putative SNPs, depending on the depth of sequence coverage obtained, the representation of different transcripts and the genetic diversity of the study species. Moreover, SNPs derived from transcribed sequences are associated with functional genes, allowing direct links to be established between functional genetic variation and phenotypic traits (Wang *et al.* 2008; Hemmer-Hansen *et al.* 2011).

A wide variety of technologies are available for genotyping SNPs, the optimal selection of which depends on numerous factors including sensitivity, accuracy, reliability, cost (both overall and per genotype), multiplexing capacity, throughput, ease of assay development and the need for specialized equipment (De la Vega *et al.* 2005; Sobrino *et al.* 2005; Syvanen 2005; Ragoussis 2009; Slate *et al.* 2009). Small to medium throughput technologies include Applied Biosystem's SNPlex™ and TaqMan® SNP genotyping assays, Sequenom's iPlex® assay, Beckman Coulter's SNPstream® and Illumina's GoldenGate assay. The latter is arguably the most flexible in terms of multiplexing, allowing 48–384 loci to be genotyped in a single reaction on the BeadXpress platform and 96–1536 loci to be typed on the BeadArray platform. This has made it popular for Quantitative Trait Locus discovery, genetic diversity assessment, association mapping and marker-assisted selection in a variety of commercially important species (Rostoks *et al.* 2006; Hyten *et al.* 2008; Wang *et al.* 2008; Akhunov *et al.* 2009; Yan *et al.* 2010).

Although these technologies also show great promise for studying natural populations, SNP genotyping in non-model organisms is not always straightforward (Lepoittevin *et al.* 2010). One reason for this is that the majority of SNP genotyping technologies were initially developed for use in humans, which benefit from the availability of vast numbers of pre-validated, optimized SNPs (Fan *et al.* 2003). In contrast, several factors can hinder the development of SNPs in species with partially characterized genomes. For example, sequencing errors can lead to the identification of false-positive SNPs, particularly when the depth of sequence coverage and/or the *in silico* Minor Allele Frequency (MAF) are low (Wang *et al.* 2008). Moreover, poor-quality flanking sequences including the presence of undetected SNPs, repetitive elements or exon-intron boundaries can also result in assays failing to convert into scoreable polymorphic SNPs (Wang *et al.* 2008; Grattapaglia *et al.* 2011). Finally, in species with large and complex genomes, there is the additional risk of interpreting paralogous loci containing fixed differences as SNPs (Smith *et al.* 2005; Sanchez *et al.* 2009). Thus, while human studies typically report GoldenGate conversion rates above 92% (Montpetit *et al.* 2006; Garcia-Closas *et al.* 2007), equivalent rates for non-model organisms are almost invariably lower. These range from 12.5% to 40.6% for SNPs that have not been validated using *in vitro* approaches (Wang *et al.* 2008; Chancerel *et al.* 2011) to 89.1% for pre-validated markers (Hyten *et al.* 2008).

Another area for concern relates to the enhanced potential for genotyping error in custom assays. Moderate rates of error can be tolerated for the assessment of population structure, but even low rates can have a strong detrimental effect on parentage analysis (Hoffman & Amos 2005a;b), the estimation of population size (Waits & Leberg 2003) and linkage and association studies (Douglas *et al.* 2000; Abecasis *et al.* 2001; Lamina *et al.* 2010). In theory, far lower error rates should be achievable for SNPs than other markers such as microsatellites because of their biallelic nature and the increased potential for automated genotyping and scoring. However, some authors have argued that error rates could actually increase 'as laboratories rush to implement high-throughput SNP methods' (Sobel *et al.* 2002). In practice, the GoldenGate assay has been reported to be highly accurate in humans, with error rates in the order of 0.3–0.4% (Oliphant *et al.* 2002; Fan *et al.* 2003). However, far more variable error rates, ranging from zero to around 4% per reaction, have been reported for non-model organisms (Akhunov *et al.* 2009; Lepoittevin *et al.* 2010; Yan *et al.* 2010; Chancerel *et al.* 2011; Grattapaglia *et al.* 2011) and the exact causes of this variation are unclear.

An ideal opportunity for exploring the prevalence and causes of GoldenGate genotyping error in a non-model species is provided by the Antarctic fur seal (*Arctocephalus gazella*). On Bird Island, South Georgia, a breeding colony of this species has been studied since the 1970s, with an aerial walkway providing unprecedented ease of access for tissue sampling and the collection of detailed behavioural observations, which started in 1994. Genetic analysis using nine hypervariable microsatellite loci has shown that most if not all pups are conceived ashore (Hoffman *et al.* 2003) and hence that lifetime reproductive profiles can be constructed for virtually every territory-holding male. Heterozygosity at the same panel of markers has also been shown to correlate with virtually every fitness trait measured to date, from male reproductive success through body size to attractiveness to females (Hoffman *et al.* 2004, 2007, 2010a). However, because the markers used are both anonymous and few in number, the mechanisms underlying these associations remain unclear (Hoffman *et al.* 2010b). To circumvent this problem, a transcriptome assembly was recently generated (Hoffman 2011) from which we have begun to develop functionally annotated genetic markers (Hoffman & Nichols 2011).

The aim of this study was to develop a high-quality, genome-wide distributed panel of SNPs for the Antarctic fur seal using the GoldenGate assay. A total of 144 putative SNPs identified from the transcriptome assembly were genotyped in 480 individuals. Replicate individuals were included, allowing us to quantify genotyping error

rates and to evaluate the efficacy of three main approaches for driving down the genotyping error rate for the resulting data set. After having stringently filtered out unreliable genotypes, we also explored relationships between several relevant parameters and assay conversion success.

Materials and methods

Tissue sampling and DNA extraction

Skin biopsy samples were collected from 440 fur seal individuals at a designated study colony on Bird Island, South Georgia (54º 00′ S, 38º 02′ W) during the austral summers of 2000/2001–2008/2009. Sampling procedures are described in detail by Hoffman *et al.* (2003). Skin samples were transferred to Dimethyl Sulphoxide saturated with salt and stored individually at −20 °C. Total genomic DNA was extracted using an adapted Chelex 100 protocol (Walsh *et al.* 1991) followed by phenol-chloroform purification (Sambrook *et al.* 1989). Each sample was then quantified using PicoGreen (Invitrogen) fluorometry. DNA concentrations averaged 96.8 ng/μL and ranged from 3.4 to 243.1 ng/μL.

Transcriptome assembly and SNP discovery

A fur seal transcriptome assembly was previously generated using protocols described by Hoffman (2011). Briefly, a normalized cDNA library derived from skin samples collected from twelve individuals was sequenced on a Roche GS-FLX Titanium DNA sequencer (Roche Diagnostic). A total of 1 443 397 reads with a mean length of 286 bp were generated. These were assembled *de novo* using Roche Newbler assembler version 2.3 into 23 025 isotigs, which in turn clustered into 18 576 isogroups (different isotigs from a given isogroup can be inferred as alternative splice-variants). The mean isotig length was 854 bp and the average depth of coverage was 19.4×. Basic Local Alignment Search Tool (BLASTX) sequence similarity searches to the non-redundant (nr) database with an e-value threshold of $1e^{-4}$ produced matches for 10 825 isotig sequences (47.0%), with 76.9% of the top matches being to mammals and these most frequently comprising the dog. Restricting the BLAST search set to canine sequences, the majority of isotigs (n = 22 541, 97.9%) were also mapped to unique locations within the dog genome. A final set of BLAST searches against a subset of sequences with known Gene Ontology (GO) annotations recovered a total of 111 446 annotation terms.

Single nucleotide polymorphism detection was conducted using the Swap454 pipeline (Brockman *et al.* 2008) which incorporates a *phred*-based quality score into the SNP-calling algorithm to reduce the false-positive rate. The phred score is an estimate of the probability that the corresponding base-call is correct, based on the image intensities recorded during sequencing. The Swap454 program first maps the raw sequence reads back to the assembled isotigs, ignoring alignments having <80% identity. It then scores these alignments by adding indels plus mismatches and declares a read ambiguous if the score for its best alignment exceeds a one-fourth of the score of the second-best. We specified an 11-base Neighbourhood Quality Standard (NQS) of 20/15 (i.e. for the parameter 'MIN_QUAL', we specified at least quality '20' at the central base and a window of five bases on each side with a neighbourhood quality 'NQ' of at least '15'). No more than two mismatches and zero indels were allowed in this window. Using only these mapped reads, and taking into account an error model for the 454 data, Swap454 then determines which positions are called as SNPs according to two user-specified thresholds. The first of these thresholds, 'MIN_RATIO' corresponds to the fraction of reads that differ from the reference sequence at a given position and the second, 'MIN_READS' to the number of copies present of the minor allele. To minimize the possibility of false positives arising from sequencing error, we applied stringent SNP discovery criteria, setting MIN_RATIO to 0.1 (meaning at least 10% of the reads at this position must differ from the reference sequence to call a SNP) and MIN_READS to 6. The parameter 'NEED_RC' was set to 'True' meaning that reads need to be seen aligned in both directions to call a SNP. This identified a total of 1599 putative SNPs located within 1004 different isotigs. These were then further filtered using a relational database to include only those that were functionally annotated with respect to the nr database and which mapped to known locations in the dog genome. This reduced the total number of putative SNPs to 1101, which were located within 613 different isotigs.

Single nucleotide polymorphism selection and GoldenGate assay design

Individual SNPs were selected to provide the best possible balance between two main criteria: genomic distribution and quality. To achieve our primary aim of developing a genome-wide panel of SNPs in the Antarctic fur seal, we selected SNPs distributed evenly across chromosomes, as inferred using BLASTN to map our isotigs to the closest matching sequences within the dog (*Canis familaris*) genome. Additionally, we also evaluated a subset of SNPs located within candidate genes relating to immunity and growth (Hoffman & Nichols 2011). This involved filtering all of the SNP-containing isotigs for a subset with GO annotation terms containing the strings 'immune' or 'growth', which recovered 41 and 107 SNPs respectively. The most promising of these, which met the

additional criteria outlined below, were selected for further development. These comprised seven immune-related and twelve growth-related SNPs (See Table S1, Supporting information for details). The former included two SNPs residing within isotigs revealing sequence similarity to Major Histocompatibility Complex (MHC) class II genes.

Our second major criterion was SNP quality. To maximize the likelihood of SNPs successfully converting into polymorphic assays, the following steps were taken: (i) we selected only SNPs with at least 60 bases of flanking sequence on either side to allow the design of allele and locus-specific oligonucleotides; (ii) because the presence of additional SNPs within the flanking sequences can have a strong detrimental effect on assay performance (Wang *et al.* 2008), we visually inspected all of the sequences within the alignment viewing program Tablet v1 (Milne *et al.* 2010) and discarded loci showing evidence of flanking SNPs; (iii) isotigs carrying high SNP densities were also avoided because of the possibility of these having been assembled from paralogous loci (Sanchez *et al.* 2009); and (iv) as a final guard against false-positive SNPs and to maximize the informativeness of our panel, we also opted not to develop assays for SNPs with *in silico* Minor Allele Frequencies (MAFs) below 10%.

We initially submitted sequences for 200 putative SNPs to Illumina for processing by the Assay Design Tool (ADT). This software generates a score for each SNP that takes into account the sequence conformation around the SNP, the presence of repetitive elements and, in the case of model organisms, sequence redundancy against the available database. The resulting score varies between 0 and 1, with values of 0.6 or above indicating a high probability of conversion into a successful genotyping assay. ADT scores for the 200 SNPs ranged from 0.35 to 1 and averaged 0.86. Sequences that gave scores below 0.6 were discarded, and a subset of 144 SNPs that met the above criteria was subsequently selected to populate the custom Oligo Pool Array (OPA). The ADT scores for these SNPs ranged from 0.78 to 1 with a mean of 0.92 (see Table S1, Supporting information for details).

Single nucleotide polymorphism genotyping

Highly multiplexed SNP genotyping was conducted using Veracode GoldenGate genotyping on a BeadXpress Reader at the Sheffield University Molecular Ecology Laboratory following the manufacturer's protocols. Genotyping was carried out using 5 μL of each DNA sample normalized to a concentration of ≤ 50 ng/μL. The template was first subjected to oligonucleotide hybridization involving two allele-specific primers and one locus-specific primer for each of the 144 loci (the OPA comprised a total of 432 custom oligonucleotides). Allele-specific primers were then extended across the SNP site and ligated to the locus-specific primer to create a polymerase chain reaction (PCR) template. This step was followed by universal PCR for all 144 loci using complementary primers, with the allele-specific oligonucleotides being labelled with either *Cy3* or *Cy5* to distinguish each of the alleles. The fluorescent products were then hybridized onto Sentrix Array Matrices (SAMs) containing beads coated with oligonucleotides complementary to address sequences within each of the locus-specific primers. Finally, the bead array signal was read using an Illumina Beadstation 500 GX (Illumina, San Diego, SA). Homozygous genotypes are expected to return predominantly *Cy3* or *Cy5* signals, whereas heterozygotes display a signal of roughly equal strength in both channels.

Automated allele calling was implemented using the software GenomeStudio 2010.1 (Genotyping module 1.7.4; Illumina). This program normalized the intensity data for each of the loci and then assigned each sample a cluster position. The resulting genotype output was then produced with two quality measures, the GenTrain and GenCall scores (Fan *et al.* 2003). The GenTrain score provides a locus-specific measure that takes into account the quality and shape of the genotype clusters and their relative distances from one another. The GenCall score, estimated for each individual at each SNP, allows individuals or loci to be ranked. Genotypes with lower GenCall scores are located further away from the centre of clusters and are therefore considered less reliable. We only accepted loci with a GenTrain score ≥0.25 and only called individual genotypes with GenCall scores ≥0.25. These scores represent stringent thresholds previously applied in studies of humans (Fan *et al.* 2003) and other species (Namroud *et al.* 2008; Sanchez *et al.* 2009; Lepoittevin *et al.* 2010). After calling the data automatically (dataset1), we then checked each of the loci manually within GenomeStudio and excluded from further analysis any SNPs that did not show clear patterns of cluster separation (see Fig. 1c,d for examples). This process yielded data set 2. Following Yan *et al.* (2010), we then generated a third data set in which minor manual adjustments were made to the clustering to allow the re-scoring of any genotypes that we believed to be either ambiguous or incorrect. Finally we removed any samples from the analysis that had call rates <0.9, as we suspected these samples may be prone to error at those loci for which they were called (data set 4).

Estimation of genotyping error rates and multilocus heterozygosity

Nine individuals were genotyped in duplicate following Hoffman & Amos (2005b) to estimate overall reproducibility. The error rate per reaction was calculated as the

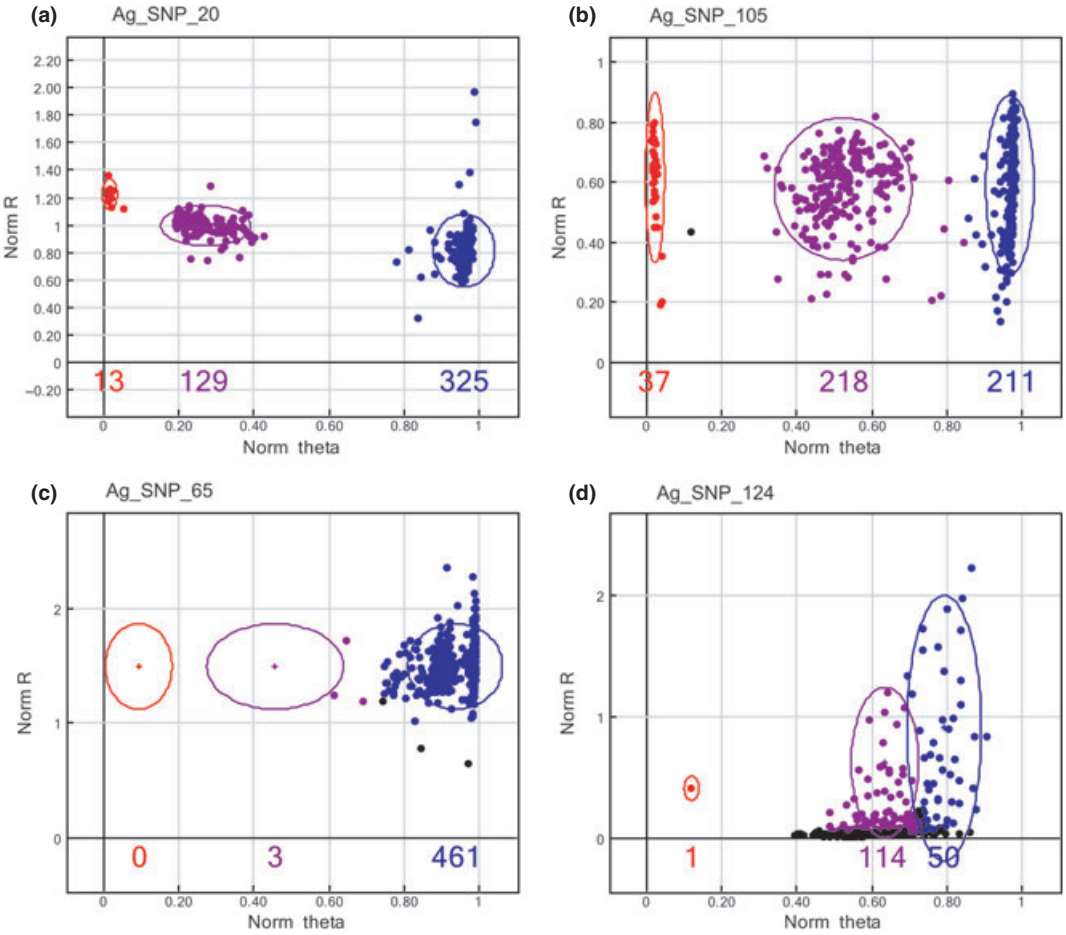

Fig. 1 Examples of clustering results obtained for the Antarctic fur seal single nucleotide polymorphism (SNP) array. Each point represents the mean normalized intensity derived from a population of beads for a single sample. 'Norm R' (*y*-axis) is the normalized sum of the intensities of the two channels (*Cy3* and *Cy5*). 'Norm Theta' (*x*-axis) is $((2/\pi)\mathrm{Tan}^{-1}$ (*Cy5/Cy3*)) where a value near 0 represents a homozygote for allele A (denoted by red points) and a value near 1 represents a homozygote for allele B (denoted by blue points). Heterozygotes fall approximately mid-way between these values and are denoted by purple points. Samples not scored at a given locus due to their not having passed the GenCall threshold of 0.25 are denoted by black points. The numbers of samples called by GenomeStudio for each of the three possible genotypes are shown below the *x*-axis (thirteen failed samples were removed from the data set leaving a total of 467). (a) Classical three-cluster pattern for a SNP considered successful and polymorphic; (b) A second polymorphic SNP showing greater scatter; (c) A locus showing 'cluster compression', in which the clusters cannot be clearly distinguished because of their being closer to one another than expected; (d) A locus showing ambiguous clustering. We classified c and d as genotyping failures.

number of genotypes that differed divided by the total number of comparisons made (Bonin *et al.* 2004). We also calculated multilocus standardized heterozygosity (SH) for each individual using the program Rhh (Alho *et al.* 2010).

Data analyses

Genepop (Raymond & Rousset 1995) was used to calculate observed and expected heterozygosities for each of the loci and to test for deviations from Hardy–Weinberg equilibrium and for linkage disequilibrium among pairs of loci. To explore factors potentially influencing whether or not a given assay successfully converted into a high-quality polymorphic SNP, we constructed a Generalized Linear Model (GLM) within R (R Development Team 2005). Conversion was modelled as a binary response variable (coded as 0 = failed and 1 = successful) using a binomial error structure. The following predictor variables were fitted: class of SNP (as a factor with transition = 0, transversion = 1), depth of sequence coverage at the SNP (accepted reads only), MAF (accepted reads only) and ADT score plus all second-order interactions. Using standard deletion-testing procedures (Crawley 2002), each term was then progressively dropped from models unless doing so significantly reduced the amount

of deviance explained. The change in deviance between full and reduced models was distributed as chi-square with degrees of freedom equal to the difference in degrees of freedom between the models with and without the term in question. For all models, distributions of standardized residuals about regressions were inspected to verify that they were approximately normally distributed.

Results

Automated single nucleotide polymorphism calling (dataset 1)

Fully automated analysis within GenomeStudio using a GenTrain threshold of 0.25 resulted in all 144 putative SNPs being validated. Most of the loci showed clear clustering patterns (see Fig. 1, panels a and b for examples) and received correspondingly high GenTrain scores (mean = 0.72). Nine loci were monomorphic in all of the samples tested and were excluded from further analysis, leaving 135 (93.8%) polymorphic SNPs. Applying a GenCall threshold of 0.25 to the scoring of individual genotypes, 13 individuals (2.7%) failed to generate data at any of the 144 loci. After excluding these individuals from the data set, the call rate averaged over all individuals and loci was 94.6%. The genotyping error rate, estimated by repeat-genotyping nine individuals, was estimated at 0.0040 per reaction (nine genotypes were called differently out of 2255 reactions compared).

Removal of poor-quality loci (dataset 2)

We next used GenomeStudio to visually inspect the clustering results obtained for each of the 135 polymorphic assays. A further 31 loci were found to exhibit unclear clustering patterns, either because of poor cluster separation (n = 19, Fig. 1c) or because of a combination of poor clustering and highly variable signal intensity (n = 12, Fig. 1d). We considered these assays to be genotyping failures despite their having passed the GenTrain threshold. Retaining only clear, polymorphic SNPs (n = 104, 72.2%), the average GenTrain score increased to 0.77 and the call rate increased to 99.4%. The genotyping error rate for the reduced data set was estimated at 0.0016 (three genotypes were called differently out of 1867 reactions compared).

Manually adjusted clustering (data set 3)

We next revisited the data set of 104 high-quality loci and visually inspected the automated scoring of each individual within GenomeStudio. Manual adjustments were made to the clustering to allow the program to exclude any ambiguous genotypes and to rescore any genotypes that we believed to be incorrect. This resulted in one of the SNPs no longer being scored as polymorphic, leaving a total of 103 variable loci remaining. The resulting data set had a marginally lower call rate than before (98.8%) but the average GenTrain score increased to 0.80 indicating an improved fit of the manually adjusted clusters to the data. The genotyping error rate was also further reduced to 0.0005 (1 genotype was scored differently out of 1839 reactions compared).

Removal of poor-quality samples (data set 4)

It is common practice when using microsatellites to discard the multilocus genotypes of any individuals that fail to generate interpretable PCR products at or above a given number of loci. Although the exact threshold is often arbitrarily defined, this approach makes good sense because poor-quality samples tend not only to have higher genotyping failure rates but also to suffer from allelic dropout, which can produce 'false homozygotes' (Taberlet *et al.* 1999). To explore whether a similar bias could exist within our SNP data set, we pooled data from all 103 SNPs and calculated SH for each of the individuals. Comparing this with the number of SNPs scored, a highly significant correlation was obtained (Fig. 2, $\chi^2 = 99.66$, d.f. = 1, $P < 0.0001$), although contrary to expectations, the direction of the trend was negative. By implication, samples excluded from being scored at multiple loci may show a bias towards being preferentially assigned to heterozygous clusters at the remaining loci.

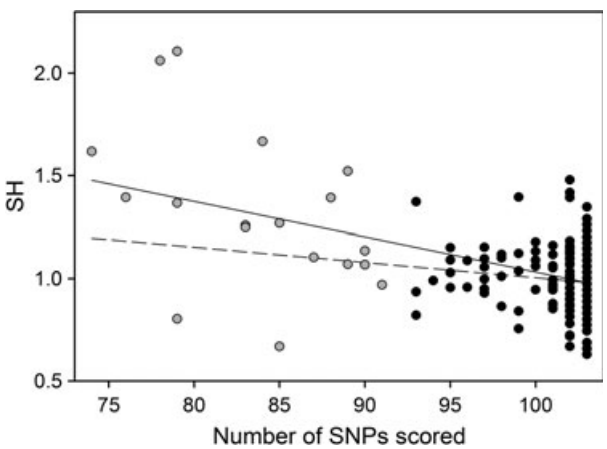

Fig. 2 Relationship between the number of single nucleotide polymorphisms scored and individual multilocus standardized heterozygosity. Grey circles depict samples that failed to score at 10% or more of loci. The solid line represents the regression fitted to the entire data set (y = −0.0171x + 2.742, r^2 = 0.175), while the broken line represents the regression after excluding the grey data points (y = −0.007x + 1.744, r^2 = 0.002).

Fortunately, the trend became no longer significant after we excluded 18 samples that failed to score at 10% or more of the loci ($\chi^2 = 3.67$, d.f. = 1, $P > 0.05$), although this also resulted in an additional locus being discarded as monomorphic, leaving a total of 102 polymorphic SNPs. The call rate increased to 99.5% but the overall genotyping error rate was not further affected.

Descriptive statistics

Using the final, stringently filtered data set, we tested for conformity to Hardy–Weinberg equilibrium (HWE) and calculated a variety of summary statistics (see Table S2, Supporting information for details). Three of the SNPs were found to deviate significantly from HWE at $P < 0.05$, although only one (Ag_SNP_1) remained significant following table-wide correction for the false discovery rate (Benjamini & Hochberg 1995) implemented within the program Q-value (Storey & Tibshirani 2003). At this locus, all of the individuals were called as homozygotes but for different alleles, a pattern attributable to our having inadvertently developed this marker within the mitochondrial NADH dehydrogenase gene. Tests for linkage disequilibrium (LD) among the remaining 101 polymorphic SNPs yielded only nine *P*-values that were robust to table-wide false discovery rate correction, consistent with the broad inferred genomic distribution of the marker panel (Fig. 3, Tables S1 and S2, Supporting information). However, in five of nine instances where significant LD was inferred, both loci mapped to adjacent positions on the same chromosome in the dog (Ag_SNP_10 and Ag_SNP_11 both mapped to chromosome 1, Ag_SNP_6 and Ag_SNP_103 to chromosome 10, Ag_SNP_34 and Ag_SNP_43 to chromosome 36, Ag_SNP_61 and Ag_SNP_103 to chromosome 10, and Ag_SNP_58 and Ag_SNP_138 to chromosome 24) consistent with their being physically linked.

Candidate gene markers

There is a growing interest in developing genetic markers targeted towards specific classes of functionally relevant gene (i.e. 'candidate genes'). Consequently, we explored the feasibility of developing SNPs within a subset of 19 isotigs with functional annotations relating to immunity and growth (see Materials and methods for details). Thirteen of these converted into high-quality polymorphic SNPs, giving a success rate that did not differ significantly from that of non-candidate SNPs (13/19 versus 89/125, binomial proportions test, $P > 0.05$). Moreover, all but one of the immune-related SNPs (85.7%) successfully converted, including two located within isotigs revealing homology to MHC class II genes.

Assay conversion rates

Finally, to explore factors influencing the propensity of GoldenGate assays to successfully convert into clearly interpretable polymorphic SNPs, we constructed a GLM (see Materials and Methods for details). Only *in silico*

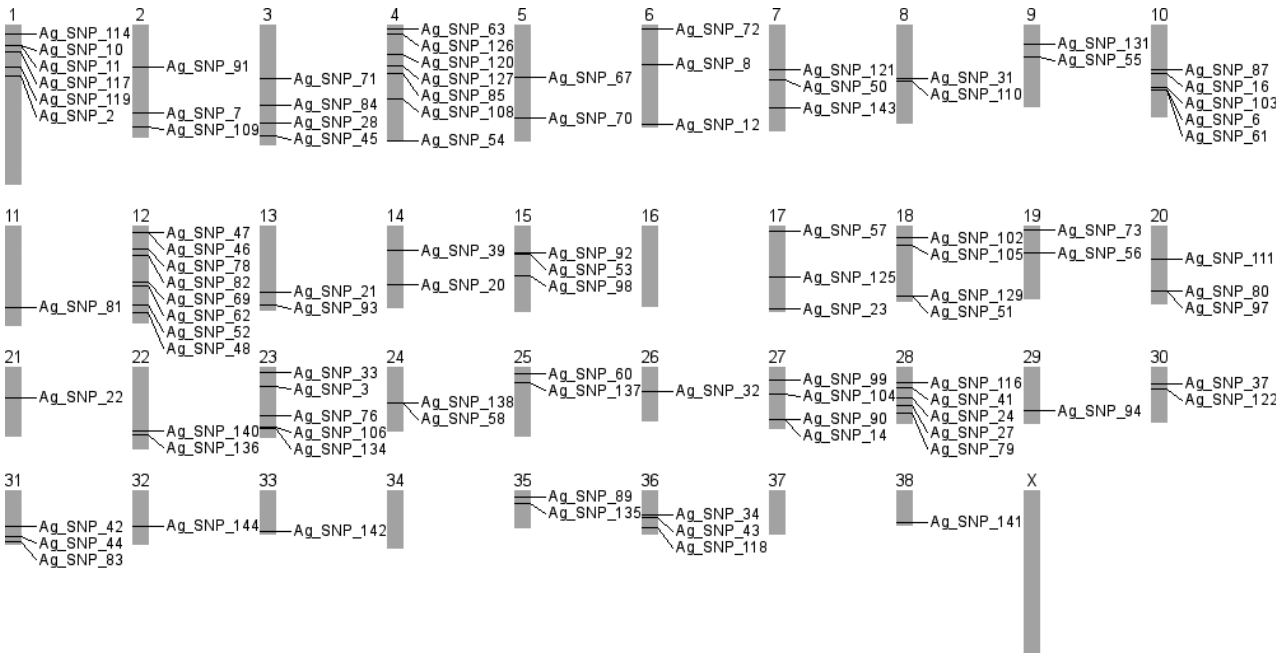

Fig. 3 Genomic distribution of 101 polymorphic nuclear single nucleotide polymorphisms inferred by mapping isotigs to the dog (*Canis familiaris*) genome.

MAF was found to explain significant deviance in conversion success ($\chi^2 = 7.32$, d.f. = 1, $P = 0.007$). Unexpectedly, the slope of the relationship was negative, implying that SNPs with low to intermediate *in silico* MAFs were more likely to convert into successful assays than those with MAFs approaching 0.5.

Discussion

Although SNPs are rapidly emerging as a marker of choice for many population genetic applications, they have not yet been widely applied to natural populations of non-model organisms. Consequently, examining the efficacy of assay design and measuring and reducing the genotyping error rate remain important challenges, particularly given that many SNP genotyping technologies were initially developed for use in humans. Here, we successfully developed a genome-wide panel of SNPs for a marine mammal, the Antarctic fur seal, while also exploring several approaches for reducing the genotyping error rate. We demonstrate highly repeatable SNP-calling after a combination of removing poor-quality loci, manually editing clusters within GenomeStudio and excluding unreliable samples.

Single nucleotide polymorphism genotyping errors

Theoretically, enhanced automation of genotyping and scoring should allow SNP genotyping error rates to be driven far lower than is achievable for microsatellites. In practice, however, highly variable GoldenGate genotyping error rates have been reported for non-model organisms, ranging from zero to around 4% per reaction (Akhunov *et al.* 2009; Lepoittevin *et al.* 2010; Yan *et al.* 2010; Chancerel *et al.* 2011; Grattapaglia *et al.* 2011). Our initial error rate of 0.4% per reaction falls towards the lower end of this range, but we were nevertheless able to identify several sources of error that could have a bearing on why rates vary so greatly from system to system.

The first main source of error we identified in our SNP data set was locus-specific. Upon visual inspection, over 20% of loci ($n = 31$) validated by GenomeStudio were found to exhibit unclear clustering patterns, either due to poor cluster separation or variable to low signal intensities. Removal of these loci resulted in a 2.5-fold reduction in the genotyping error rate, consistent with a previous study reporting a ninefold reduction after omitting just four out of 188 SNPs with the highest error rates (Lepoittevin *et al.* 2010). An alternative to identifying poor-quality loci by eye is to apply stringent *ad hoc* reliability thresholds. For example, Grattapaglia *et al.* (2011) only scored SNPs with median GenCall scores greater than or equal to 0.4 and with a call rate of 95% or above. However, without visual inspection and depending on the exact thresholds applied, this approach risks discarding data of adequate quality.

The second source of error we observed relates to the automated scoring of genotypes within GenomeStudio. Following the approach of Yan *et al.* (2010), we made minor manual adjustments to the clustering of each locus to allow the program to re-score any genotypes that were either ambiguous or likely to be incorrect. However, in contrast to Yan *et al.* (2010) who did not directly quantify the effect of their adjustments on the quality of the resulting data, we were able to demonstrate a greater than 3-fold reduction in the genotyping error rate, even after having previously filtered poor-quality loci from the data set. Thus, although manual cluster adjustment is clearly undesirable for large-scale projects (Yan *et al.* 2010), it may prove a valuable way of minimizing errors in small- to medium-scale studies.

A third source of error documented in this study was sample-specific, with a highly significant negative correlation being found between SH and the number of SNPs scored. This indicates a tendency for samples that perform poorly in the assay to be preferentially assigned to heterozygous rather than homozygous clusters and is probably related to the fact that heterozygous clusters are usually more scattered on the *x*-axis (i.e. they have a larger variance in normalized theta) than homozygous ones (see Fig. 1a,b for examples). Fortunately, the relationship became no longer significant after removing a small number of samples that failed to score at 10% or more of loci, suggesting a relatively straightforward means of mitigating the problem. Nevertheless, the effect on individual samples was in some cases extreme, with two samples in particular yielding SH values over twice that of the mean SH of the full data set (2.11 and 2.06 vs. 1.00) and well outside the interquartile range (0.89–1.09). If overlooked, this could have led to erroneous conclusions being drawn about the distribution of individual heterozygosity in the population, particularly if the inclusion of unreliable genotypes were to generate an artefactual correlation in heterozygosity among loci that could be interpreted as evidence for inbreeding depression (Balloux *et al.* 2004).

An alternative to filtering samples on the basis of the proportion of SNPs scored would be to use another quality criterion, such as p50GC, the 50th percentile (or mode) of the distribution of the GenCall scores for a given individual. Genotypes with lower average GenCall scores tend to be located further away from the centre of clusters and are therefore considered to be less reliable. However, GenCall scores provide an imperfect measure of cluster separation because they are based on the degree to which the two homozygote clusters are separated from the heterozygote cluster rather than the degree of separation of the two homozygote clusters (Hyten *et al.* 2008).

Moreover, because heterozygous clusters tend to be more diffuse than homozygous clusters *per se*, individuals who are heterozygous for many loci will also tend to have higher average GenCall values. This leads to the prediction that SH and p50GC could, under certain circumstances, be negatively correlated even in the absence of genotyping error.

Assay conversion rates

Even after having stringently filtered our data set, resulting in a substantial proportion of loci being discarded as either low quality or monomorphic, our rate of conversion into polymorphic SNPs still compares favourably with similar studies based exclusively on *in silico* resources (e.g. 12.5–19.5% in Maritime pine (Chancerel *et al.* 2011) and 40.6% in Catfish (Wang *et al.* 2008)). There are several potential reasons for this. Most obviously, sequencing errors can lead to the identification of false-positive SNPs, especially where sequence coverage is low (Wang *et al.* 2008). However, we guarded against this by applying stringent SNP-calling thresholds. For example, our requirement of a minimum 6× depth of coverage of the minor allele led to assays being designed for isotigs with an average of 57× total coverage (range = 12–433×), substantially higher than reported for most previous studies.

A second factor known to impact rates of assay conversion is flanking sequence quality, including the possible presence of additional SNPs or intron-exon boundaries close to the targeted marker (Wang *et al.* 2008; Grattapaglia *et al.* 2011). Although the latter is difficult to guard against in the absence of a reference genome, we conducted a thorough *in silico* inspection of each SNP and its flanking regions, allowing us to discard any loci that were closely flanked by other polymorphisms, areas of low sequence coverage or homopolymer tracts. We were also careful to select assays with ADT scores well in excess of Illumina's recommended threshold of 0.6 (see Materials and methods).

Another potential pitfall relates to the assembly of paralogous sequences, which can lead to the identification of false-positive SNPs (Smith *et al.* 2005; Sanchez *et al.* 2009). This can be particularly problematic for species that have undergone recent genome duplications, such as many commercial crop species and some vertebrates including salmonids. Usually, assays targeted towards these regions will tend either to result in no signal (Smith *et al.* 2005; Sanchez *et al.* 2009) or to report all samples as being heterozygous (Hosking *et al.* 2004; Sanchez *et al.* 2009). However, cases of 'cluster compression', where the two homozygous clusters are shifted together, have also been reported for SNPs residing within near-identical paralogues in species with highly duplicated genomes (Hyten *et al.* 2008; Yan *et al.* 2010). Moreover, it has also been cautioned that it may prove difficult generally to distinguish between true heterozygote clusters and those resulting from targeted sequence redundancy (Yan *et al.* 2010), although the latter should in many instances lead to violation of HWE (Lee *et al.* 2008). Fortunately, pinnipeds do not have particularly large or complex genomes (Du & Wang 2006) and our use of Swap454 may also partially compensate for this problem because the program only calls SNPs on the basis of reads that assemble reliably to a single isotig (Brockman *et al.* 2008). However, we took the additional precaution of avoiding designing SNPs within isotigs carrying qualitatively high SNP densities, as these are more likely to originate from paralagous loci (Sanchez *et al.* 2009). This appears to have been largely successful, because only six of our failed assays exhibited clustering patterns indicative of cluster compression (for an example, see Fig. 1c). Moreover, only a single SNP was found to deviate significantly from HWE after correction for the false discovery rate, and this was because of the locus in question having inadvertently been designed within a mitochondrial gene.

To test whether additional factors might also have contributed to assay success or failure, we constructed a GLM. In contrast to several previous studies (Wang *et al.* 2008; Lepoittevin *et al.* 2010; Chancerel *et al.* 2011), we found no effect of the type of SNP (transition vs. transversion), depth of sequence coverage or Illumina's quality score, the latter combining information about flanking sequence complexity and context. One reason for this could be that our relatively small sample size of loci, combined with an above-average conversion rate, gave us little power to dissect apart the underlying causes of assay failure. Alternatively, many of these variables might not have come into play because of our stringent *in silico* filtering criteria. For example, we only evaluated assays with ADT scores above the arbitrary threshold of 0.6. In contrast, Lepoittevin *et al.* (2010) reported a significant effect of ADT score, but this was driven primarily by poorly converting assays with ADT scores in the range 0.4–0.6. We nevertheless detected a significant effect of *in silico* MAF on the conversion rate, although contrary to our initial expectations, the slope of the relationship was negative. The exact reasons for this are unclear, although if paralogous loci were sequenced at roughly equal depth, any fixed differences residing within them should be manifest as false-positive SNPs with *in silico* MAFs close to 0.5.

Overall performance of the assay in Antarctic fur seals

Our findings also add to a growing body of evidence suggesting that, despite imperfect conversion rates, GoldenGate genotyping can perform remarkably well for SNPs

derived *in silico* from non-model organisms. Thus, even after excluding poor-quality loci and samples, we obtained over 45 000 high-quality individual SNP genotypes at the cost of <€10 000 and 2 weeks spent in the laboratory. This compares favourably with a similar sized Antarctic fur seal microsatellite data set (5000 individuals genotyped at 9–76 loci) that has taken over a decade to amass using manual techniques (J. Hoffman, unpublished data). Moreover, our final SNP genotyping error rate (0.0005 per reaction) is roughly an order of magnitude lower than that previously estimated for our microsatellite data set (Hoffman & Amos 2005b). Finally, our attempts at targeting SNPs within candidate genes were surprisingly successful, 13 of 19 assays converting into polymorphic SNPs, including two residing within MHC class II like-genes.

Conclusion

We have demonstrated the feasibility of developing and genotyping a panel of genome-wide distributed SNPs in a marine mammal species. Our approach, based on the development of GoldenGate assays from a skin transcriptome, should be particularly useful for marine mammal taxa as well as other species that can only be remotely biopsy sampled.

Acknowledgements

We are grateful to D. Briggs, M. Jessop, K. Reid, R. Taylor, T. Walker, N. Warren, S Robinson, D. Malone and E Edwards for tissue sampling and logistical support. We also acknowledge cDNA synthesis performed by Evrogen, Russia, and library preparation and 454 sequencing performed by Anna Montazam and Denis Cleven of the GenePool Facility, Edinburgh. This work contributes to the British Antarctic Survey (BAS) Ecosystems (Polar Science for Planet Earth) programme. Fieldwork was approved by BAS, and samples were collected and retained under permits issued by the Department for Environment, Food and Rural Affairs (DEFRA) and in accordance with the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). The molecular work was funded by a Marie Curie Career Integration Grant (PCIG-GA-2011-303618) awarded to J. Hoffman and NERC core funding to the British Antarctic Survey Ecosystems Programme.

References

Abecasis GR, Cherny SS, Cardon LR (2001) The impact of genotyping error on family-based analysis of quantitative traits. *European Journal of Human Genetics*, **9**, 130–134.

Aitken N, Smith S, Schwarz C, Morin PA (2004) Single nucleotide polymorphism (SNP) discovery in mammals: a targeted gene approach. *Molecular Ecology*, **13**, 1423–1431.

Akhunov E, Nicolet C, Dvorak J (2009) Single nucleotide polgymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theoretical and Applied Genetics*, **119**, 507–517.

Alho J, Valimaki K, Merila J (2010) Rhh: an R extension for estimating multilocus heterozygosity and heterozygosity-heterozygosity correlation. *Molecular Ecology Resources*, **10**, 720–722.

Balloux F, Amos W, Coulson T (2004) Does heterozygosity estimate inbreeding in real populations? *Molecular Ecology*, **13**, 3021–3031.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300.

Bensch S, Akesson S, Irwin DE (2002) The use of AFLP to find an informative SNP: genetic differences across a migratory divide in willow warblers. *Molecular Ecology*, **11**, 2359–2366.

Bonin A, Bellemain E, Bronken Eidesen P *et al.* (2004) How to track and assess genotyping errors in population genetic studies. *Molecular Ecology*, **13**, 3261–3273.

Brockman W, Alvarez P, Young S *et al.* (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research*, **18**, 763–770.

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

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.

Crawley MJ (2002) *Statistical Computing, an Introduction to Data Analysis Using S-plus*. John Wiley and Sons Ltd, Chichester.

De la Vega FM, Lazaruk KD, Rhodes MD, Wenz MH (2005) Assessment of two flexible and compatible SNP genotyping platforms: Taqman SNP genotyping assays and the SNPlex genotyping system. *Mutation Research*, **573**, 111–135.

Douglas JA, Boehnke M, Lange K (2000) A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *American Journal of Human Genetics*, **66**, 1287–1297.

Du B, Wang D (2006) C-values of seven marine mammal species determined by flow cytometry. *Zoological Science*, **23**, 1017–1020.

Fan JB, Oliphant A, Shen R *et al.* (2003) Highly parallel SNP genotyping. *Cold Spring Harbor Symposia on Quantitative Biology*, **68**, 69–78.

Garcia-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.

Grattapaglia D, Silva-Junior OB, Kirst M *et al.* (2011) High-throughput SNP genotyping in the highly heterozygous genome of *Eucalyptus*: assay success, polymorphism and transferability across species. *BMC Plant Biology*, **11**, 65.

Hemmer-Hansen J, Nielsen EE, Meldrup D, Mittelholzer C (2011) Identification of single nucleotide polymorphisms in candidate genes for growth and reproduction in a nonmodel organism; the Atlantic cod, *Gadus morhua*. *Molecular Ecology Resources*, **11**, 71–80.

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

Hoffman JI, Amos W (2005a) Does kin selection influence fostering behaviour in Antarctic fur seals (*Arctocephalus gazella*)? *Proceedings of the Royal Society of London Series B-Biological Sciences*, **272**, 2017–2022.

Hoffman JI, Amos W (2005b) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, **14**, 599–612.

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

Hoffman JI, Forcada J, Amos W (2010a) Getting long in the tooth: a strong positive correlation between canine size and heterozygosity in the

Antarctic fur seal *Arctocephalus gazella*. *Journal of Heredity*, **101**, 527–538.

Hoffman JI, Forcada J, Amos W (2010b) Exploring the mechanisms underlying a heterozygosity-fitnesss correlation for canine size in the Antarctic fur seal *Arctocephalus gazella*. *Journal of Heredity*, **101**, 539–552.

Hoffman JI, Nichols HJ (2011) A novel approach for mining polymorphic microsatellite markers *in silico*. *PLoS ONE*, **6**, e23283.

Hosking L, Lumsden S, Lewis K *et al*. (2004) Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European Journal of Human Genetics*, **12**, 395–399.

Hyten DL, Song Q, Choi I-Y *et al*. (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. *Theoretical and Applied Genetics*, **116**, 945–952.

Krawczak M (1999) Informatiivity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis*, **28**, 1676–1681.

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

Lamina C, Kuchenhoff H, Chang-Claude J *et al*. (2010) Haplotype misclassification resulting from statistical reconstruction and genotype error, and its impact on association estimates. *Annals of Human Genetics*, **74**, 452–462.

Lee S, Kasif S, Weng Z, Cantor CR (2008) Quantitative analysis of single nucelotide polymorphisms within copy number variation. *PLoS ONE*, **3**, e3906.

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.

Milne I, Bayer M, Cardle L *et al*. (2010) Tablet-next generation sequence assembly visualization. *Bioinformatics*, **26**, 401–402.

Montpetit A, Nelis M, Laflamme P *et al*. (2006) An evaluation of the performance of tag SNPs derived from HapMap in a caucasian population. *PLoS Genetics*, **2**, e27.

Morin PA, Luikart G, Wayne R, The SNP Workshop Group (2004) SNP's in ecology, evolution and conservation. *Trends in Ecology and Evolution*, **19**, 208–216.

Namroud M-C, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology*, **17**, 3599–3613.

Oliphant A, Barker DL, Stuelpnagel JR, Chee MS (2002) BeadArrayTM technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *BioTechniques*, **32**, S56–S61.

Primmer CR, Borge T, Lindell J, Saetre GP (2002) Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. *Molecular Ecology*, **11**, 603–612.

R Development Team (2005) *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ragoussis J (2009) Genotyping technologies for genetic research. *Annual Review of Genomics and Human Genetics*, **10**, 117–133.

Raymond M, Rousset F (1995) Genepop (Version 1.2) – population genetics software for exact tests of ecumenicism. *Journal of Heredity*, **86**, 248–249.

Rostoks N, Ramsay L, MacKenzie K *et al*. (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *Proceedings of the National Acadamy of Sciences*, **103**, 18656–18661.

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Spring Harbour Laboratory Press, New York.

Sanchez CC, Smith TPL, Wiedmann RT *et al*. (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 599.

Seddon JM, Parker HG, Ostrander EA, Ellegren H (2005) SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. *Molecular Ecology*, **14**, 503–511.

Slate J, Gratten J, Beraldi D *et al*. (2009) 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.

Sobel E, Papp JC, Lange K (2002) Detection and integration of genotyping errors in statistical genetics. *American Journal of Human Genetics*, **70**, 496–508.

Sobrino B, Brion M, Carracedo A (2005) SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Science International*, **154**, 181–194.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440–9445.

Syvanen A-C (2005) Toward genome-wide SNP genotyping. *Nature Genetics*, **37**, S5–S10.

Taberlet P, Waits LP, Luikart G (1999) Noninvasive genetic sampling: look before you leap. *Trends in Ecology and Evolution*, **14**, 323–327.

Vera JC, Wheat CW, Fescemyer HW *et al*. (2008) Rapid transcriptome characterization for a non-model organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Waits JL, Leberg PL (2003) Biases associated with population estimation using molecular tagging. *Animal Conservation*, **3**, 191–199.

Walsh PS, Metzger DA, Higuchi R (1991) Chelex100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *BioTechniques*, **10**, 506–513.

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

Yan J, Yang X, Shah T *et al*. (2010) High-throughput SNP genotyping with the GoldenGate assay in maize. *Molecular Breeding*, **25**, 441–451.

J.I.H. conceived and designed the study. J.I.H., and R.T. designed SNP assays and genotyped the samples. J.I.H., S.J.B., and J.S. analyzed the data. J.I.H., M.S.C., J.F., and J.S. contributed reagents / materials. J.I.H., and J.S. wrote the paper. All authors approved the final manuscript.

Data Accessibility

DNA sequence assembly, previously uploaded to Dryad by Hoffman (2011): doi: 10.5061/dryad.8268. Single nucleotide polymorphism data: dbSNP accession numbers Ag_SNP_1 511224113, Ag_SNP_2 511224116, Ag_SNP_3 511224119, Ag_SNP_6 511224122, Ag_SNP_7 511224124, Ag_SNP_8 511224126, Ag_SNP_10 511224129, Ag_SNP_11 511224131, Ag_SNP_12 511224134, Ag_SNP_14 511224136, Ag_SNP_16 511224138, Ag_SNP_20 511224141, Ag_SNP_21 511224144, Ag_SNP_22 511224146, Ag_SNP_23 511224148, Ag_SNP_24 511224151, Ag_SNP_27 511224154, Ag_SNP_28 511224157, Ag_SNP_31 511224159, Ag_SNP_32 511224161, Ag_SNP_33 511224163, Ag_SNP_34 511224166, Ag_SNP_37 511224167, Ag_SNP_39 511224170, Ag_SNP_41 511224173, Ag_SNP_42 511224175, Ag_SNP_43 511224177, Ag_SNP_44 511224180, Ag_SNP_45 511224183, Ag_SNP_46 511224186, Ag_SNP_47 511224187, Ag_SNP_48 511224190, Ag_SNP_50 511224193, Ag_SNP_51 511224196, Ag_SNP_52 511224197, Ag_SNP_53 511224200, Ag_SNP_54 511224203, Ag_SNP_55 511224206, Ag_SNP_56 511224208, Ag_SNP_57 511224210, Ag_SNP_58 511224213, Ag_SNP_60 511224216, Ag_SNP_61 511224218,

Ag_SNP_62 511224220, Ag_SNP_63 511224223, Ag_SNP_67 511224226, Ag_SNP_69 511224229, Ag_SNP_70 511224231, Ag_SNP_71 511224233, Ag_SNP_72 511224236, Ag_SNP_73 511224239, Ag_SNP_76 511224240, Ag_SNP_78 511224243, Ag_SNP_79 511224246, Ag_SNP_80 511224248, Ag_SNP_81 511224250, Ag_SNP_82 511224253, Ag_SNP_83 511224256, Ag_SNP_84 511224259, Ag_SNP_85 511224260, Ag_SNP_87 511224263, Ag_SNP_89 511224266, Ag_SNP_90 511224268, Ag_SNP_91 511224271, Ag_SNP_92 511224273, Ag_SNP_93 511224276, Ag_SNP_94 511224278, Ag_SNP_97 511224280, Ag_SNP_98 511224283, Ag_SNP_99 511224286, Ag_SNP_102 511224288, Ag_SNP_103 511224290, Ag_SNP_104 511224293, Ag_SNP_105 511224296, Ag_SNP_106 511224299, Ag_SNP_108 511224301, Ag_SNP_109 511224303, Ag_SNP_110 511224306, Ag_SNP_111 511224308, Ag_SNP_114 511224310, Ag_SNP_116 511224312, Ag_SNP_117 511224315, Ag_SNP_118 511224318, Ag_SNP_119 511224319, Ag_SNP_120 511224322, Ag_SNP_121 511224325, Ag_SNP_122 511224327, Ag_SNP_125 511224330, Ag_SNP_126 511224332, Ag_SNP_127 511224335, Ag_SNP_129 511224338, Ag_SNP_131 511224340, Ag_SNP_134 511224342, Ag_SNP_135 511224345, Ag_SNP_136 511224348, Ag_SNP_137 511224351, Ag_SNP_138 511224353, Ag_SNP_140 511224355, Ag_SNP_141 511224358, Ag_SNP_142 511224361, Ag_SNP_143 511224364, Ag_SNP_144 511224366.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Details of 144 GoldenGate SNP assays developed from the Antarctic fur seal transcriptome assembly (see Materials and methods for details). 'Chromosome in the dog' refers to the genomic location inferred by mapping each isotig to the dog (*Canis familaris*) genome. Basic Local Alignment Search Tool (BLAST) results indicate the top match of each isotig to the non-redundant (nr) database. Gene Ontology codes are given only for isotigs that recovered functional annotations relating to growth or immunity. *In silico* Minor Allele Frequency and depth of coverage refer to the exact site of each SNP and are given only for reads accepted as unambiguous by the program Swap454 (Brockman *et al.* 2008). Assay Design Tool scores vary between 0 and 1, with values of 0.6 or above indicating a high probability of conversion into a successful genotyping assay.

Table S2 Polymorphism characteristics of 102 polymorphic SNP assays in 440 Antarctic fur seals individuals (see Materials and methods for details). The GenTrain score takes into account the quality, shape and degree of separation of the genotype clusters, with higher values indicating improved clustering (Fan *et al.* 2003).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.