MOLECULAR ECOLOGY
RESOURCES

The easy road to genome-wide medium density SNP screening in a non-model species: development and application of a 10K SNP-chip for the house sparrow (Passer domesticus).

Journal:	*Molecular Ecology Resources*
Manuscript ID:	MER-12-0405.R1
Manuscript Type:	Resource Article
Date Submitted by the Author:	11-Jan-2013
Complete List of Authors:	Hagen, Ingerid; Centre for Conservation Biology, Norwegian University of Science and Technology, Dept. of Biology Billing, Anna; Centre for Conservation Biology, Norwegian University of Science and Technology, Dept. of Biology Rønning, Bernt; Centre for Conservation Biology, Norwegian University of Science and Technology, Dept. of Biology Pedersen, Sindre; Norwegian University of Science and Technology, Dept. of Biology Pärn, Henrik; Centre for Conservation Biology, Norwegian University of Science and Technology, Dept. of Biology Slate, Jon; University of Sheffield, Dept. of Animal and Plant Sciences Jensen, Henrik; Centre for Conservation Biology, Norwegian University of Science and Technology, Dept. of Biology
Keywords:	next-generation sequencing, single nucleotide polymorphism, genetic population differentiation, house sparrow, Passer domesticus

Title:

The easy road to genome-wide medium density SNP screening in a non-model species: development and application of a 10K SNP-chip for the house sparrow (*Passer domesticus*).

Ingerid J. Hagen[1], Anna M. Billing[1], Bernt Rønning[1], Sindre A. Pedersen[2], Henrik Pärn[1], Jon Slate[3] and Henrik Jensen[1]

[1] Centre for Conservation Biology, Department of Biology, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway.

[2] Department of Biology, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway.

[3] Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK.

Keywords: next-generation sequencing, single nucleotide polymorphism, genetic population differentiation, house sparrow, *Passer domesticus*

Corresponding authors: Ingerid J. Hagen, Fax: +47 73596100, E-mail: ingerid.hagen@bio.ntnu.no, and Henrik Jensen, Fax: +47 73596100, E-mail: henrik.jensen@bio.ntnu.no

Running title: Medium density SNP screening in sparrows

Abstract

With the advent of next generation sequencing, new avenues have opened to study genomics in wild populations of non-model species. Here, we describe a successful approach to a genome-wide medium density Single Nucleotide Polymorphism (SNP) panel in a non-model species, the house sparrow (*Passer domesticus*), through the development of a 10K Illumina iSelect HD BeadChip. Genomic DNA and cDNA derived from six individuals were sequenced on a 454 GS FLX system and generated a total of 1.2 million sequences, in which SNPs were detected. As no reference genome exists for the house sparrow, we used the zebra finch (*Taeniopygia guttata*) reference genome to determine the most likely position of each SNP. The 10,000 SNPs on the SNP-chip were selected to be distributed evenly across 31 chromosomes, giving on average one SNP per 100,000 bp. The SNP-chip was screened across 1968 individual house sparrows from four island populations. Of the original 10,000 SNPs, 7413 were found to be variable, and 99% of these SNPs were successfully called in at least 93% of all individuals. We used the SNP-chip to demonstrate the ability of such genome-wide marker data to detect population sub-division, and compared these results to similar analyses using microsatellites. The SNP-chip will be used to map Quantitative Trait Loci (QTL) for fitness-related phenotypic traits in natural populations.

Introduction

Single Nucleotide Polymorphisms (SNPs) have over the last decade been established as the most frequently used genetic marker in genome mapping, population genomics and Quantitative Trait Loci (QTL) mapping (Goddard & Hayes 2009; Mackay *et al*. 2009; Allendorf *et al.* 2010; Slate *et al*. 2010). Until recently, SNPs have mainly seen use in studies of the human genome and genetic model species, but - as the cost of next generation sequencing decreases, SNPs are more readily used in studies of wild non-model species, surpassing microsatellite markers as the major genetic marker technology (Luikart *et al*. 2003; Kohn *et al*. 2006; Ekblom & Galindo; 2011). Compared to microsatellites, SNPs offer several advantages including a higher frequency of occurrence in the genome in both coding and non-coding regions (Collins *et al.* 1998; Brumfield *et al.* 2003), and a bi-allelic nature which corresponds more closely with the models of evolution often applied in population genetics (e.g. Hartl & Clark 1997). Additionally, SNPs offer logistical advantages such as a lower rate of genotyping error and greater ease of automating large scale genotyping (Kennedy *et al.* 2003). The ability to efficiently type large numbers of SNP markers across numerous individuals has enabled the study of the genetic architecture of quantitative traits by use of QTL or gene mapping (Slate 2005; 2008; Goddard & Hayes 2009; Stapley *et al.* 2010). Medium and high density genome-wide SNP panels are therefore powerful tools when disentangling the relative roles of selection, genetic drift and gene flow in observed patterns of genetic variation (Luikart *et al.* 2003; Nielsen *et al.* 2005; Stinchcombe & Hoekstra 2008). Historically, QTL mapping has been performed mainly on laboratory populations and outcrossed lines of cultured plants or domesticated animals under controlled conditions (Duncan *et al.* 2007; Goddard & Hayes 2009). QTL mapping studies of natural populations are more challenging than laboratory studies (Slate 2005; Slate *et al*. 2010), but are important

in order to establish the extent to which studies in model organisms can be extrapolated, and to study adaptive genetic processes in systems where forces such as environmental interactions, pleiotropy and epistasis have not been reduced or eliminated (e.g. Roff & Simons 1997; Kroymann & Mitchell-Olds 2005; Ellegren & Sheldon 2008). In order to perform QTL mapping in a natural population, a large number of genetic markers and a detailed pedigree are required. The number of genetic markers has previously been the major limiting factor for QTL mapping in natural populations, but with the advent of applications derived from the now frequently used next generation sequencing this is changing rapidly.

A method for *de-novo* development of a medium density SNP-chip for a non-model species, the great tit (*Parus major*) has been described in Van Bers *et al.* (2012). *De-novo* development of medium to high density SNP chips for non-model species is expected to become more common over the next decade (Seeb *et al.* 2011), and best practice guides that describe successful approaches will likely be important resources for future studies. Additionally, comparisons of results obtained from SNP genotyping and those based on microsatellites - which have a much higher mutation rate than SNPs (Foll & Gaggiotti 2008) - are rare (Helyar *et al.* 2011). Here, we describe the development of a 10K Illumina iSelect HD BeadChip for the house sparrow (*Passer domesticus*), a non-model species for which only a small fraction of the genome has been sequenced but that is used to study many ecological, evolutionary and physiological questions (e.g. Anderson 2006). The house sparrow is globally distributed, either as a result of natural dispersal or from introductions mitigated by humans (Anderson 2006). The ubiquitous nature of the house sparrow and multiple historical bottle-neck events make it attractive for studies into evolution and adaptation (e.g. Summers-Smith 1988; Anderson, 2006). The sparrow breeds in proximity of human habitations, where they nest in accessible cavities in houses, inside barns or within nest boxes. The house sparrow

lends itself as a good study species of adaptation because it has a short generation time (approx. 2 years; Jensen *et al*. 2008), is easy to locate and capture, and can be measured and non-lethally sampled without causing negative effects on its populations. Our study complements that of Van Bers *et al*. (2012) by describing an alternative approach to a successful 10K Illumina iSelect HD BeadChip *de-novo* development and large scale population genotyping. Additionally, we investigated the performance of the SNP panel in tests of population differentiation and compared the results to data derived from microsatellite typing of the same populations.

The 10K Illumina iSelect HD BeadChip developed in this study will be used to establish the first medium density marker map of the house sparrow genome, and will be an important resource for QTL mapping and studies of the genetic architecture of complex traits, genetic drift and adaptive evolution in both natural and manipulated populations.

Materials and Methods

Our data were collected from four islands within a natural house sparrow metapopulation in northern Norway. The islands are separated by at least 20 km, are part of individual-based long-term studies initiated in 1993 (Aldra [66°25'N, 13°04'E] and Hestmannøy [66°33'N, 12°50'E]) or 2001 (Leka [65°06'N, 11°38'E] and and Vega [65°40'N, 11°55'E]) and used as model systems to answer questions in evolutionary biology, ecology and conservation biology (see e.g. Ringsby *et al.* 2002; Jensen *et al.* 2007; 2008; Pärn *et al.* 2009; Holand *et al*. 2011; Billing *et al.* 2012; Pärn *et al.* 2012). Within this metapopulation system, less than 10% of recruits disperse from their natal population (Pärn *et al.* 2009) and populations are morphologically and genetically differentiated (Holand *et al*. 2011; Jensen *et al*. in review). Genetic pedigrees going back > 8 generations are established for all island populations

represented in this study (Jensen *et al.* 2003; 2004; 2008; Billing *et al.* 2012; Rønning *et al*. in prep.) and include approx. 6000 individuals.

Collection of tissues and RNA extraction

In February 2009, three relatively large populations (population size >160 adults) were randomly sampled for one female and one male specimen. The populations included Hestmannøy, Vega and Leka of the coast of central and northern Norway (Fig. 1). The house sparrows were captured in mist nets, after which 47 µl of blood was immediately collected. The sparrows were then quickly euthanized by cervical dislocation following guidelines established and approved by the Norwegian Directorate for Nature Management, and within 10 minutes 80-150 mg of tissues from heart, liver, kidney, lung and brain were dissected. Additionally, testis tissue (10 and 130 mg) was collected from two males, one from Leka and the other from Vega. Blood and tissue samples were transferred to separate 2 ml DNase and RNase free micro tubes (Nunc) containing RNAlater (QIAGEN, 1000 µl for blood samples and 1600 µl for tissue samples) and immediately frozen for later extraction of RNA and DNA in the lab. Total RNA from all tissues from one Hestmannøy male, testis samples from the Leka and Vega Island males, and blood samples from all six individuals were extracted using a RiboPure Blood Kit (Aambion Inc., Austin, TX, USA), with additional DNase treatment following manufacturers' recommendations. For the remaining five individuals, total RNA and DNA from all remaining tissues were extracted using a GeneMole automated nucleic acid extraction system (Mole Genetics AS, Oslo, Norway) and the Total RNA Basic Kit (Mole Genetics) without DNase treatment by Mole Genetics AS (Oslo, Norway). Extracted RNA was stored in <2.5 mM Tris-HCL, pH 7.6.

cDNA library generation and 454-sequencing

For each individual, total RNA from all tissue samples and the blood sample was pooled prior to cDNA library synthesis. The amount of extracted RNA for each tissue and blood was then adjusted to give similar amounts of RNA from each, and 10 µg RNA in total from an individual. One random-primed and normalized cDNA library was synthesized for each individual at Vertis Biotechnologie AG (Friesing-Weihenstephan, Germany), following their in house protocol for cDNA library generation. To prepare the cDNA for 454-Titanium sequencing, normalized cDNA in the size range of 500 – 800 bp was eluted from preparative agarose gels. The six synthesized cDNA libraries contained on average 17.2 ng/µl (range: 9-31 ng/µl) in 20 µl solutions. One run of 454-sequencing was carried out on a GS FLX Titanium (Roche, Switzerland) at the Norwegian High-Throughput Sequencing Centre, University of Oslo, Norway.

SNP-chip development

Read mapping Alignment of the reads was performed in CLC Genomics Workbench 4 (CLC Bio, Aarhus, Denmark). All sequences from the six individuals were mapped together, thus the contigs were allowed to contain sequences from more than one individual. Stringent alignment parameters were used to ensure a high sequence quality from which to detect SNPs: mismatch cost = 3; insertion cost = 3; deletion cost = 3; length fraction = 0.4; similarity = 0.99; ambiguity codes were used; non-specific matches were ignored and minimum contig length set to 150 bp.

SNP detection and selection SNPs were detected in CLC Genomics using the following parameters and Roche Phred scores: window length = 9; maximum number of gaps and mismatches = 1; minimum average quality of surrounding bases = 20; minimum quality of central base = 37; minimum coverage = 3; and minor allele frequency = 5%. Only non-

complex SNPs of the Infinium II design (one probe per SNP) with 60 bp of sequence on either side of the SNP were chosen. The selected 121 bp sequences were searched against the *Taeniopygia guttata* reference genomic sequences database build 1.1 on nucleotide BLAST (http://blast.ncbi.nlm.nih.gov) as follows: optimised for discontinuous megablast; maximum target sequences = 10; expected threshold = 10; word size = 11; match/mismatch scores = 1/-1; gap cost = 2 for existence, 1 for extension and filtered for low complexity regions. Because the 454-reads used to build contigs contained both genomic DNA and cDNA, the search against *T. guttata* reference genomic sequences served to help us determine whether SNPs from cDNA sequences were situated on intron/exon breaks and to inform us of the position of the SNP on the *T. guttata* genome. The retained query sequences that contained intron/exon breaks were then cropped at the intron/exon break and were not allowed to be less than 116 bp long. Query sequences that had hits to more than one position on the *T. guttata* genome were considered repeats and removed. Query sequences with BLAST e-values worse than 2.0E-15 and those with more than six ambiguities were removed (i.e. the SNP itself and maximum 5 other ambiguities within the 116-121 bp sequence was allowed). In order to determine the SNPs that were situated in known *T. guttata* mRNA regions, the remaining sequences were put through BLAST against the *T. guttata* RNA reference genomic database (http://blast.ncbi.nlm.nih.gov), with BLAST parameters as described above. A second BLAST search against the more improved *T. guttata* reference genomic sequences database build 3.2.4.58 was carried out after development of the chip, in order to verify the positions on the genome. The BLAST parameters were identical to previous searches. In total 19,852 SNPs were sent to Illumina for processing by the Illumina Assay Design Tool to generate a score file with a score and a failure code for each SNP that indicated the expected success for designing an assay for the SNP. The SNPs that received failure codes equal to zero and scores

over 0.85 (N = 13,800) were retained. Of these, 9955 unique SNPs were selected to be included on the 10K SNP chip based on their position on the *T. guttata* genome. The SNPs were selected to be evenly distributed across the genome. The minimum distance between two SNPs was set to 675 bp. For positions with more than one SNP to choose from, the one with the highest read depth was selected. See Fig. 2 for an illustration of SNP distribution across the *T. guttata* genome. As positive controls, 45 SNPs were randomly chosen to be typed twice to test for genotyping errors.

The 10,000 SNPs included 10 SNPs located in candidate genes for beak morphology and limb development: Calmodulin (Abzhanov *et al.* 2006; Schneider 2007), FGF8 (Abzhanov & Tabin 2004; Grant *et al.* 2006) and Frizzled (Brugmann *et al.* 2010). These were specifically retained through the selection process despite not always conforming to the criteria described above.

Extraction of genomic DNA for sample screening

The blood samples used for the SNP genotyping were collected from 1968 different individuals on the islands Aldra (N = 406), Hestmannøy (N = 447), Leka (N = 512) and Vega (N = 603) (see Fig. 1) during the years 1993-2010. The six individuals that were sequenced (see above) were included among the 1968 samples. Of the 1968 individuals, 37 were randomly chosen to be genotyped several times as positive controls. These 37 individuals were typed two (N = 31), three (N = 5) or four (N = 1) times, thus the total number of DNA samples screened was 2012. Additionally, four samples containing only ddH_2O were included in the screening as negative controls. In total 2016 SNP chips were used in the genotyping.

Whole blood preserved in 96% ethanol, which had been stored for up to 10 years in room temperature at the time of DNA extraction, was lysed in 60 µl Lairds buffer (Ausubel *et al*. 1989), with 90 µg proteinase K (Sigma Aldrich, St Louis, MO), and incubated at 50ºC for

208 3 hours. Genomic DNA was extracted from the lysate using the ReliaPrep Large Volume HT
209 gDNA Isolation System (Promega, Madison, WI), automated on a Biomek NXp robot
210 (Beckman Coulter, Miami, FL) and following the manufacturer's recommendations; the only
211 exception being elution of DNA in 25 mM Tris HCl (pH 8). The DNA concentrations were
212 measured using a Flurostar Omega scanner (MBG Labtech, Offenburg, Germany). Illumina
213 recommends sample concentrations of 40 – 60 ng/µl. Samples with concentrations above 60
214 ng/µl were normalized to a concentration of 50 ng/µl with 25 mM Tris HCl (pH 8). 760
215 samples had stock DNA concentrations below the recommended 40 ng/µl. For each sample, a
216 four µl aliquot of DNA containing a total of on average 160 ng (SD=56 ng) DNA was stored
217 at -20 ºC until sample screening on the Illumina iSelect HD BeadChip. The SNP screening,
218 clustering and scoring of genotypes were carried out by the Genomics Core Facility,
219 Norwegian University of Science and Technology. The results were checked, filtered and
220 scored using GenomeStudio (Illumina, San Diego) following the guidelines provided by
221 Illumina (Illumina 2010).

222 *Descriptive statistics and analysis of genetic differentiation*
223 Descriptive statistics, tests regarding SNP design principles and SNP typing results were
224 carried out in the statistical software IBM SPSS 19.0 (SPSS Inc., 2010). The dataset used for
225 descriptive statistics was based on all loci for all genotyped samples except the four negative
226 controls. Quality control filtering of the dataset prior to analyses of genetic structure was done
227 in PLINK version 1.07 (http://pngu.mgh.harvard.edu/purcell/plink/; Purcell *et al.* 2007), using
228 a minor allele frequency (MAF) of 0.01 and maximum per-person missing (MIND) of 0.1.
229 Tests for Hardy-Weinberg equilibrium were done separately for each population and with a
230 significance level of 0.05. For the analyses of genetic structure and F_{ST} we used a reduced
231 dataset comprising individuals present on the four islands in 2002 (N = 173), and their

respective genotypes for SNPs that were found on autosomes and that passed the above described quality control filtering (N = 6736). Individuals from 2002 were chosen in order to avoid potential effects of experimental manipulation in two of the populations after 2002 and to avoid inclusion of close relatives across multiple generations. First, a principal component analysis (PCA) was performed using the R-package adegenet version 1.3 (Jombart 2008; Jombart & Ahmed 2011) in the statistical software R version 2.14.2 (R Development Core Team 2012). Second, genetic differentiation between populations was analysed with STRUCTURE 2.3.3 (Pritchard *et al.* 2000) using no prior population information, the admixture model, 10,000 burn-ins and 50,000 iterations, the number of populations from K = 1 to K = 5 and 10 separate runs for each K. The results were processed in STRUCTURE HARVESTER (Earl & von Holdt 2012), which uses the Evanno method to determine the most likely K (Evanno *et al*. 2005). Finally, we performed F_{ST} analyses in the R package HIERFSTAT (Goudet 2005) which estimates F_{ST} with 95% confidence intervals (CI); if the 95% CI do not include zero the estimate is regarded as significantly different from zero at P = 0.05. The results from STRUCTURE and HIERFSTAT were compared to measures of population divergence derived from microsatellite data in Jensen *et al*. (2013).

Results

RNA extraction

The RNA/DNA yield for samples extracted at Mole Genetics was 100 µl eluate containing on average 139.3 ng/µl (SD = 88.0 ng/µl) nucleic acids for liver, kidney, lung and brain tissue samples, and on average 18.1 ng/µl (SD = 4.9 ng/µl) for heart tissue samples. For samples extracted in-house the 100 µl eluate contained on average 450.3 ng/µl (SD = 257.3 ng/µl) total RNA for liver, kidney, lung and brain tissue samples, 27.5 ng/µl for the heart tissue

sample, 401.4 ng/µl (SD = 420.6 ng/µl) for the testis samples, and on average 52.3 ng/µl (SD = 27.6 ng/µl) for the blood samples.

SNP chip development

Sequencing The 454-sequencing generated 1,160,122 reads of mean length 282 bp and 327,536,336 bp in total (Sequence read Archive accession numbers xxxxxxx – xxxxxxx). On average, the number of reads from each of the six individuals was 191,242 (range: 145,889 – 287,470). The five individuals for which tissue samples were extracted at Mole Genetics had length distribution of reads biased towards short reads (mean = 245 bp). This was due to repeats in many of the sequences, which caused premature termination of the 454-sequencing. The likely reason for such repeats was failure to treat the samples with DNase during RNA extraction at Mole Genetics, with the consequence that genomic DNA was present in the samples when cDNA libraries were synthesized. In contrast, for the individual that was extracted in-house, the length distribution was as expected (i.e. a peak in the distribution of read lengths at approx. 450 bp), indicating that sequences from this individual represented RNA.

Contig assembly and SNP detection Stringent alignment of the reads produced 93,351 contigs in which SNPs could be detected. The number of SNPs varied according to the stringency of the search requirements. Using stringent search parameters described above but allowing coverage down to 3 reads per site, we identified a total of 43,198 SNPs. Of these, 37,714 SNPs fulfilled the initial criteria of sequence length and SNP quality and were thus put through BLAST. Of these, 13,800 SNPs satisfied the selection criteria (intron/exon breaks were not allowed to extend more than a total of 5 bp into the 121 bp sequences, Infinium Type

II design, less than 5 ambiguities, a unique position on the *T. guttata* genome, and an Illumina score of over 0.85) and were thus considered for inclusion on the chip.

SNP-chip design characteristics and SNP call rates Among the 10,000 SNPs selected for the SNP-chip, the Illumina score was on average 0.9683, the mean number of ambiguities within the 112-121 bp query sequence was 1.84 (range: 1 - 6, including the SNP itself), the mean coverage was 6.37 (range: 3 - 67), and the mean count of the less frequent allele was 1.69 (range: 1 - 32) (Fig. 3).

Of the 10,000 SNPs on the chip, 8491 were successfully called (dbSNP accession numbers xxxx – xxxx) and 7413 were variable. The Illumina design score, number of ambiguities in the query sequence, coverage and the count of the least frequent allele was similar for the 8491 SNPs that were successfully called and the 1509 SNPs that were not called (Mann-Whitney tests: $P > 0.11$). On the other hand, the successfully called SNPs had on average slightly longer query sequences than non-called SNPs (120.211 [SE = 0.014] vs. 120.123 (SE = 0.036), Mann-Whitney test: $P = 0.015$) and worse e-values (2.0E-18 [SE = 7.7E-19] vs. 5.5E-19 [SE = 2.2E-19], Mann-Whitney test: $P < 0.001$). These results suggest that few of the SNP-chip design criteria affected the probability of a SNP to be successfully called (given that the SNP had fulfilled our detection and selection criteria).

Compared to the 1078 called and non-variable SNPs, the 7413 variable SNPs had on average slightly lower read depth (6.34 [SE = 0.05] vs. 6.67 [SE = 0.13], Mann-Whitney tests: $P = 0.002$) but higher count of the least frequent allele in the sequence data (1.78 [SE = 0.02] vs. 1.14 (SE = 0.01], Table 1, Mann-Whitney tests: $P < 0.001$). This suggests that a SNP was more likely to be a true SNP (i.e. a polymorphic base) if the number of sequences in which its rarest allele was observed was two rather than one.

Assessment of SNP call rate and genotyping error in genotyped samples

Of the 2016 samples that were genotyped (including four negative controls and the six sequenced individuals), SNPs were called for 1999 samples. The 13 samples besides the negative controls that were not called had on average lower DNA concentration (mean: 17.79 ng/µl, SD = 18.20 ng/µl, range: 3.99 - 50 ng/µl) than the 1999 samples that were called (mean: 40.05 ng/µl, SD = 13.83 ng/µl, range: 0.83 - 60.00 ng/µl) (Mann-Whitney test: $P < 0.001$). For samples with a DNA concentration lower than 20 ng/µl, the probability of successful calling of SNPs decreased; no SNPs were called for 10 out of 284 samples (3.5%) with a DNA concentration lower than 20 ng/µl. In contrast, no SNPs were called for only 3 out of 1728 samples (0.2%) with a DNA concentration of 20 ng/µl or more. From the 37 individuals that were typed twice or more, we estimated a SNP typing error rate of 0.0005%. Additionally, 45 SNPs were included twice on the chip: 32 SNPs returned the same genotype for all pairs (typed in between 1995 and 1999 samples), whilst 10 SNPs returned a genotype for one of the assays but failed for the other. For three SNP pairs no individuals were called. Thus there were no conflicts among the 45 SNPs that were run twice, although some assays did not return a genotype.

The mean number of SNPs called for a given sample was 8457 (range: 8169 - 8476). The 7413 variable SNPs were on average typed in 1988 samples (Table 1). The mean minor allele frequency of variable SNPs was 0.2380, and ranged from 0.00025 (i.e. one sample was heterozygous at the SNP and the rest homozygous for the common base) to 0.5 (Table 1). Of the variable SNPs, only 21.4% had a minor allele frequency below 0.1, suggesting that most of the variable SNPs will be valuable in further analyses.

Genetic differentiation of sub-populations

Principal component analysis using SNP data from individuals sampled in 2002 indicated three distinct clusters; the individuals from Aldra constituted one group, as did the individuals from Leka, whilst Vega and Hestmannøy clustered together (but with incomplete overlap) (Fig. 4). In concordance with these results, F_{ST} analyses indicated that Aldra was the most differentiated population, with pair-wise values approximately twice as high as for the other islands (Table 2). This pattern was consistent also in the STRUCTURE analysis, which under the most likely scenario identified two clusters; again with Aldra as a distinct population and the remaining islands clustering together (Fig. 5). Under the less likely scenario of three clusters, the pattern was the same as for the principal component analysis, with Aldra and Leka being separate groups and Hestmannøy and Vega grouping together. Results based on SNP data corresponded closely with results derived from microsatellites, which showed similar levels of F_{ST} between pairs of populations (Table 2) and that - based on analyses in STRUCTURE - the four islands fell into two different clusters; again with Aldra in one cluster and the three remaining islands in a second group (Jensen *et al*. 2013).

Discussion

We have described the development of a 10K Illumina iSelect HD BeadChip for the house sparrow, and have assessed the performance of the genome-wide SNP data to detect and quantify genetic sub-division of a meta-population in central and northern Norway. The available genomic resources derived from our study include an additional 30,000 putative house sparrow SNPs in 93,351 contigs. With a mean length of 475 bp, these contigs cover approximately 44.5 million bp of the house sparrow genome, and feature both coding and non-coding regions. Assuming that the house sparrow genome has the same size as the zebra finch genome (1.2 Gbp; Warren *et al*. 2010), these sequences cover about 3.7% of the total

genome. This resource will be important during future development of SNP-chips for further investigation of specific QTL regions, or for primer design in a candidate gene approach where desired target genes are partly or wholly included in the contigs. Additionally, considering the cross-population success in great tits described in Van Bers *et al*. (2012), we predict that the house sparrow SNP chip will provide reliable results if applied to house sparrow populations outside of Norway, as indicated by microsatellite genetic structure for house sparrow populations distributed globally (Schrey *et al*. 2011). Moreover, successful cross species applications of a 50K *Ovis aries* SNP chip have been described for bighorn and thinhorn sheep, with call rates of 95 and 90% respectively (Miller *et al*. 2011). It is therefore likely that the house sparrow SNP chip could be applied also to other species or sub-species of the European *Passer* genus, which has been found to have multiple hybrid zones and pair-wise F_{ST} estimates that are comparable with those found on house sparrows along the Norwegian coast (Hermansen *et al*. 2011; Jensen *et al*. 2013).

The approach we have described for SNP-chip development proved highly successful. The overall SNP call rate was 85% and approximately 75% of the SNPs were variable and informative for population differentiation and/or marker map development. Of our 2012 samples, 99.35% were called for at least 8169 SNPs. The error rate estimated from duplicate samples and SNPs was very low (0.0005%). For comparison, the call rate expected for application of commercially available SNP-chips developed for humans is approximately 98% (International HapMap Consortium 2010). Other studies on model species have for example reported call rates of 93%, with 89% of the SNPs being polymorphic for a 60K chicken SNP chip (Groenen *et al.* 2011). For non-model species the success rate is generally somewhat lower: when genotyping wild and farmed Atlantic salmon (*Salmo salar*) on a 7K SNP-chip, Karlsson *et al.* (2011) reported that 65% of the SNPs were called and informative,

whilst Van Bers *et al*. (2012) developed a 10K SNP chip for two populations of the great tit and obtained a call rate of 83%, with about 72% of the SNPs being polymorphic. Our custom made 10K SNP-chip thus has a genotyping success rate that is somewhat lower than commercial SNP-chips for humans but has the same proportion of called and variable SNPs as other studies of non-model species.

The inclusion of SNPs located in genomic DNA was un-intentional and caused by failure to treat tissue samples collected from five of the six sequenced individuals with DNase. The original idea was to sequence cDNA to obtain a high read depth in which to detect SNPs, with relatively low sequencing costs, and avoid choosing SNPs with very low minor allele count. In hindsight, it seems that sequencing a mix of cDNA and gDNA did not reduce the quality of our SNP-chip compared to similar studies, despite the fact that we had to choose some SNPs that were detected in regions with only 3x read depth.

A large number of reads at the SNP sites is advantageous, as it allows for greater confidence that the SNP is not an artifact of sequencing error. For their 60K chicken SNP-chip, Groenen *et al.* (2011) used read depth of 12x or more and Van Bers *et al.* (2012) reported a read depth at the SNP site of > 8. The median read depth of the SNPs included on the house sparrow chip was 5 (range 3 – 67), whilst 2569 of the SNPs were situated in regions with a read depth of 3 (see Fig. 3). Of these, 60% (N = 1552) were both successfully called, variable both within and across our sampled populations and with a minor allele frequency ≥ 0.01. Comparably, this is a lower success rate than for SNPs with a higher read depth, but indicates that SNPs detected in low-coverage regions have the potential to be highly useful.

Approximately one third of our samples had a lower than recommended (< 40 ng/µl) DNA concentration, however genotyping success was only marginally affected: there was a 15-20% lower call rate for samples with concentrations below 10 ng/µl, and a 1% lower call

rate for samples with concentrations ranging from 10-20 ng/µl. It therefore appears that the Illumina iSelect HD BeadChip requires less sample material than some sequencing based SNP genotyping techniques (Miller *et al*. 2007, Baird *et al*. 2008). The approach we have developed may therefore be a useful resource for molecular ecology studies on organisms for which only small amounts of DNA can be acquired.

In all analyses of genetic population structure, the island of Aldra was the most divergent, a pattern which is probably explained by the unique history of this population. Aldra is known to have been colonized by four founders in 1998 and has since received few immigrants (Billing *et al.* 2012). The effective population size on this island has been estimated to range from approximately 10 to 30 (Engen *et al*. 2007; Baalsrud *et al*. in review). Accordingly, the level of inbreeding in this island population is significant (Jensen *et al*. 2007; Billing *et al.* 2012), and it is likely that an initial founder effect and subsequent genetic drift can explain the strong genetic divergence of Aldra from the other islands.

The results from STRUCTURE and F_{ST} analyses indicated that our panel of approx. 7000 variable SNPs and a panel of 14 highly variable microsattelites produced very similar results. A denser panel of SNPs is required in order to achieve the same power as microsatellites (Evans & Cardon 2004). For instance Hess *et al.* (2011) found that microsatellites performed better than SNPs when 13 highly variable microsatellite loci were compared to a SNP-panel of 92 loci for fine-scale population identification. Microsatellites are in the process of being replaced by SNPs in a variety of molecular applications, including those within molecular ecology (Allendorf *et al.* 2010; Ekblom & Galindo 2011; Hess *et al.* 2011). However, the costs for development of SNP panels with a power comparable to microsatellites are still significant and may not be feasible for projects focusing on the ecology of non-model species. Provided microsatellites in a panel have low error rates, are

independent and distributed across the genome, there do not seem to be strong reasons to abandon the use of microsatellites for the purpose of investigating population differentiation and assignment analysis. However, in many other cases, such as in QTL mapping, studies of adaptive evolution and effects of genetic drift, a greater number of markers is required and microsatellites fall short compared to high density SNP panels.

In conclusion, we have described an easy and cost effective protocol for successful generation and population scale screening of a 10K medium density SNP chip in a non-model species. We have shown that despite features such as lower than desired read depth (3, for approximately 25% of the SNPs), lower than recommended DNA concentration for some samples, and a large proportion of SNPs situated in genomic DNA, the result was very successful and comparable with other medium density SNP-chip population screens on non-model species. Lastly, our study indicated that for the purpose of population assignment and differentiation, high density SNP data produce results that are comparable with those derived from high quality microsatellite data. In the near future, the SNP data will be used to develop the first marker map for house sparrows, and subsequent analysis into the genetic architecture of quantitative traits in wild populations of house sparrows.

Acknowledgements

We are grateful to the hospitable and friendly inhabitants in the field area who made the study possible. We also thank fieldworkers and laboratory technicians for assistance. We thank the Research Council of Norway (Project no: 191847, Strategic University Program (SUP) in Conservation Biology) and the functional genomics programme at the Norwegian University of Science and Technology (to IJH and AMB) for funding. The SNP microarray service was provided by the Genomics Core Facility, Norwegian University of Science and Technology - a

444 national technology platform supported by the functional genomics program (FUGE) of the
445 Research Council of Norway. Permit to collect tissue samples from 6 individual house
446 sparrows was given from the Norwegian Directorate for Nature Management. The research
447 was carried out in accordance with permits from the Norwegian Directorate for Nature
448 Management and the Bird Ringing Centre at Stavanger Museum, Norway.
449

450 **Data accessibility**

451 All sequence reads have been submitted to the Sequence read Archive (SRA). Accession
452 numbers are xxxxxxx – xxxxxxx. All genotyped SNPs with the 116-121 bp flanking
453 sequences have been submitted to dbSNP. Accession numbers are xxxxxxx – xxxxxxx.
454

References

Abzhanov A, Kuo WP, Hartmann C, *et al.* (2006) The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature* **442**, 563-567.

Abzhanov A, Tabin CJ (2004) Shh and Fgf8 act synergistically to drive cartilage outgrowth during cranial development. *Developmental Biology* **273**, 134-148.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11**, 697-709.

Anderson TR (2006) Biology of the ubiquitous house sparrow: from genes to populations. Oxford University Press, New York.

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K, eds. 1989. Current Protocols in Molecular Biology. John Wiley & Sons, New York.

Baird N, Etter P, Atwood T, *et al*. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PloS one* **3**.

Billing AM, Lee AM, Skjelseth S, *et al.* (2012) Evidence of inbreeding depression but not inbreeding avoidance in a natural house sparrow population. *Molecular Ecology* **21**, 1487-1499.

Brugmann SA, Powder KE, Young NM, *et al.* (2010) Comparative gene expression analysis of avian embryonic facial structures reveals new candidates for human craniofacial disorders. *Human molecular genetics* **19**, 920-930.

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.

Collins FS, Brooks LD, Chakravarti A (1998) A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. *Genome Research* **8**, 1229-1231.

Duncan EJ, Dodds KG, Henry HM, Thompson MP, Phua SH (2007) Cloning, mapping and association studies of the ovine ABCG2 gene with facial eczema disease in sheep. *Animal Genetics* **38**, 126-131.

Earl D, von Holdt B (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359-361.

Ekblom R, Galindo J (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**, 1-15.

Ellegren H, Sheldon BC (2008). Genetic basis of fitness differences in natural populations. *Nature* **452**, 169-175.

Engen S, Ringsby TH, Sæther B-E, *et al*. (2007) Effective size of fluctuating populations with two sexes and overlapping generations. *Evolution* **61**, 1873-1885.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Evans DM, Cardon LR (2004) Guidelines for Genotyping in Genomewide Linkage Studies: Single-Nucleotide–Polymorphism Maps Versus Microsatellite Maps. *The American Journal of Human Genetics* **75**, 687-692.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977-993.

Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* **10**, 381-391.

Goudet J 2005 HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**, 184–186.

503 Grant PR, Grant BR, Abzhanov A (2006) A developing paradigm for the development of bird
504 beaks. *Biological Journal of the Linnean Society* **88**, 17-22.
505 Groenen M, Megens H-J, Zare Y, *et al.* (2011) The development and characterization of a
506 60K SNP chip for chicken. *BMC Genomics* **12**, 274.
507 Hartl DL, Clark AG (1997) Principles of population genetics. 3rd ed. Sinauer Associates,
508 Sunderland.
509 Helyar SJ, Hemmer-Hansen J, Bekkevold D, *et al.* (2011) Application of SNPs for population
510 genetics of nonmodel organisms: new opportunities and challenges. *Molecular*
511 *Ecology Resources* **11**, 123-136.
512 Hermansen JS, Sæther SA, Elgvin TO, *et al.* (2011) Hybrid speciation in sparrows I:
513 phenotypic intermediacy, genetic admixture and barriers to gene flow. *Molecular*
514 *Ecology* **20**, 3812-3822.
515 Hess JE, Matala AP, Narum SR (2011) Comparison of SNPs and microsatellites for fine-scale
516 application of genetic stock identification of Chinook salmon in the Columbia River
517 Basin. *Molecular Ecology Resources* **11**, 137-149.
518 Holand AM, Jensen H, Tufto J, Moe R (2011). Does genetic drift or selection explain
519 geographic differentiation of morphological characters in house sparrows? *Genetics*
520 *Research* **93**, 367-379.
521 Illumina, Inc. (2010) Infinium genotyping data analysis: a guide for analyzing infinium
522 genotyping data using the Illumina GenomeStudio genotyping module. Illumina, Inc,
523 San Diego. Available at:
524 http://www.illumina.com/Documents/products/technotes/technote_infinium_genotypin
525 g_data_analysis.pdf

International HapMap Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58.

Jensen H, Sæther B-E, Ringsby TH, *et al.* (2003) Sexual variation in heritability and genetic correlations of morphological traits in house sparrow (*Passer domesticus*). *Journal of Evolutionary Biology* **16**, 1296-1307.

Jensen H, Sæther B-E, Ringsby TH, *et al.* (2004) Lifetime reproductive success in relation to morphology in the house sparrow Passer domesticus. *Journal of Animal Ecology* **73**, 599-611.

Jensen H, Bremset EM, Ringsby TH, Sæther B-E (2007) Multilocus heterozygosity and inbreeding depression in an insular house sparrow metapopulation. *Molecular Ecology* **16**, 4066-4078.

Jensen H, Steinsland I, Ringsby TH, Sæther B-E (2008) Evolutionary dynamics of a sexual ornament in the house sparrow (*Passer domesticus*): the role of indirect selection within and between sexes. *Evolution* **62**, 1275-1293.

Jensen H, Moe R, Hagen IJ, Holand AM, Kekkonen J, Tufto J, Sæther B-E (2013) Genetic variation and structure of house sparrow populations: is there an island effect? *Molecular Ecology* **xx**, xx-xx.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics (Oxford, England)* **24**, 1403-1405.

Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070-3071.

Karlsson S, Moen T, Lien S, Glover KA, Hindar K (2011) Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP-chip. *Molecular Ecology Resources* **11**, 247-253.

Kennedy GC, Matsuzaki H, Dong S, *et al.* (2003) Large-scale genotyping of complex DNA. *Nature Biotechnology* **21**, 1233-1237.

Kohn MH, Murphy WJ, Ostrander EA, Wayne RK (2006). Genomics and conservation genetics. *Trends in Ecology and Evolution* **21**, 629-637.

Kroymann J, Mitchell-Olds T (2005) Epistasis and balanced polymorphism influencing complex trait variation. *Nature* **435**, 95-98.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* **4**, 981-994.

Mackay T FC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**, 565-577.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2006) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* **17**.

Miller JM, Poissant J, Kijas JW, Coltman DW, the International Sheep Genomics C (2011) A genome-wide set of SNPs detects population substructure and long range linkage disequilibrium in wild sheep. *Molecular Ecology Resources* **11**, 314-322.

Nielsen R, Williamson S, Kim Y, *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Research* **15**, 1566-1575.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Purcell S, Neale B, Todd-Brown K, *et al.* (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Society of Human Genetics* **81**, 559-575.

574 Pärn H, Jensen H, Ringsby TH, Sæther B-E (2009) Sex-specific fitness correlates of dispersal
575 in a house sparrow metapopulation. *Journal of Animal Ecology* **78**, 1216-1225.
576 Pärn H, Ringsby TH, Jensen H, Sæther B-E (2012) Spatial heterogeneity in the effects of
577 climate and density-dependence on dispersal in a house sparrow metapopulation.
578 *Proceedings of the Royal Society B: Biological Sciences* **279**, 144-152.
579 Ringsby TH, Sæther B-E, Tufto J, Jensen H, Solberg EJ (2002) Asynchronous spatiotemporal
580 demography of a house sparrow metapopulation in a correlated environment. *Ecology*
581 **83**, 561-569.
582 R Development Core Team (2012) R: A language and environment for statistical computing.
583 R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL
584 http://www.R-project.org/.
585 Roff DA, Simons AM (1997) The quantitative genetics of wing dimorphism under laboratory
586 and field conditions in the cricket *Gryllus pennsylvanicus*. *Heredity* **78**, 235-240.
587 Schneider RA (2007) How to tweak a beak: molecular techniques for studying the evolution
588 of size and shape in Darwin's finches and other birds. *BioEssays* **29**, 1-6.
589 Schrey AW, Grispo M, Awad M, *et al*. (2011) Broad-scale latitudinal patterns of genetic
590 diversity among native European and introduced house sparrow (*Passer domesticus*)
591 populations. *Molecular Ecology* **20**, 1133-1143.
592 Seeb JE, Carvalho G, Hauser L, *et al.* (2011) Single-nucleotide polymorphism (SNP)
593 discovery and applications of SNP genotyping in nonmodel organisms. *Molecular*
594 *Ecology Resources* **11**, 1-8.
595 Slate J (2005) INVITED REVIEW: Quantitative trait locus mapping in natural populations:
596 progress, caveats and future directions. *Molecular Ecology* **14**, 363-379.

Slate J (2008) Robustness of linkage maps in natural populations: a simulation study. *Proceedings. Biological sciences / The Royal Society* **275**, 695-702.

Slate J, Santure AW, Feulner PGD, *et al*. (2010) Genome mapping in intensively studied vertebrate populations. *Trends in Genetics* **26**, 275-284.

Stapley J, Reger J, Feulner PGD, *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution* **25**, 705-712.

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* **100**, 158-170.

Summers-Smith JD (1988) The sparrows: a study of the genus *Passer*. T & AD Poyser Ltd, Calton.

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**, 753-779.

Warren WC, Clayton DF, Ellegren H, *et al*. (2010) The genome of a songbird. *Nature*, **464**, 757–762.

Tables

Table 1: Characteristics of SNPs on the 10K Illumina custom SNP-chip for house sparrows when typed in 2012 samples.

	Number of SNPs	Minor allele frequency	
		Mean	Median (min - max)
All SNPs	10000	0.197	0.1803 (0 - 0.5)
Non-called SNPs	1509	-	-
Called SNPs	8491	0.208	0.1958 (0 - 0.5)
Non-variable SNPs	1078	0	0 (0 - 0)
Variable SNPs	7413	0.238	0.2286 (0.00025 - 0.5)

Table 2: Pair-wise genetic distances between four island populations of house sparrows in Norway. Values below the diagonal are F_{ST} values based on 14 microsatellite loci (Jensen *et al*. 2013); values above diagonal are F_{ST} values derived from 6736 SNPs. F_{ST} values and their 95% confidence limits (in parentheses) were calculated using the R-package HIERFSTAT (Goudet 2005).

	Hestmannøy	**Aldra**	**Vega**	**Leka**
Hestmannøy	-	0.073 (0.071-0.076)	0.024 (0.022-0.025)	0.029 (0.028-0.031)
Aldra	0.062 (0.044-0.083)	-	0.078 (0.076-0.081)	0.082 (0.079-0.085)
Vega	0.024 (0.016-0.034)	0.069 (0.041-0.101)	-	0.028 (0.027-0.029)
Leka	0.023 (0.013-0.036)	0.074 (0.053-0.096)	0.025 (0.016-0.034)	-

626 **Figure legends**

627 **Figure 1:** Map showing the study area on the Norwegian coast. The four study islands are
628 shown in black.

629

630 **Figure 2:** Distribution of SNPs present on the 10K house sparrow SNP chip when mapped
631 onto the *T. guttata* genome (Tgu chromosomes 1 – 28 and Z). The number of SNPs within
632 500,000 bp intervals is shown.

633

634 **Figure 3:** Plots showing sequencing information for the 10,000 SNPs on the house sparrow
635 SNP chip. Light grey bars (left axis) show the number of SNPs on the chip selected in
636 genomic regions with a given sequence coverage. Dots (right axis) show mean count (± 1SD)
637 of the minor allele for each level of sequence coverage at the SNP location.

638

639 **Figure 4**: Plots of principle component analysis (PCA) of genetic variation between 173 adult
640 individuals present in 2002 in four island house sparrow populations off the coast of Norway.
641 The colours of each individual represent island population: Aldra = red, Hestmannøy = black,
642 Leka = green and Vega = blue. The PCAs are based on 6736 autosomal SNPs.

643

644 **Figure 5**: Structure barplots based on 6736 autosomal SNPs from 173 adult individuals
645 present in 2002 in four island house sparrow populations off the coast of Norway. The upper
646 panel shows results for $K = 2$, whilst the lower panel shows results for $K = 3$.

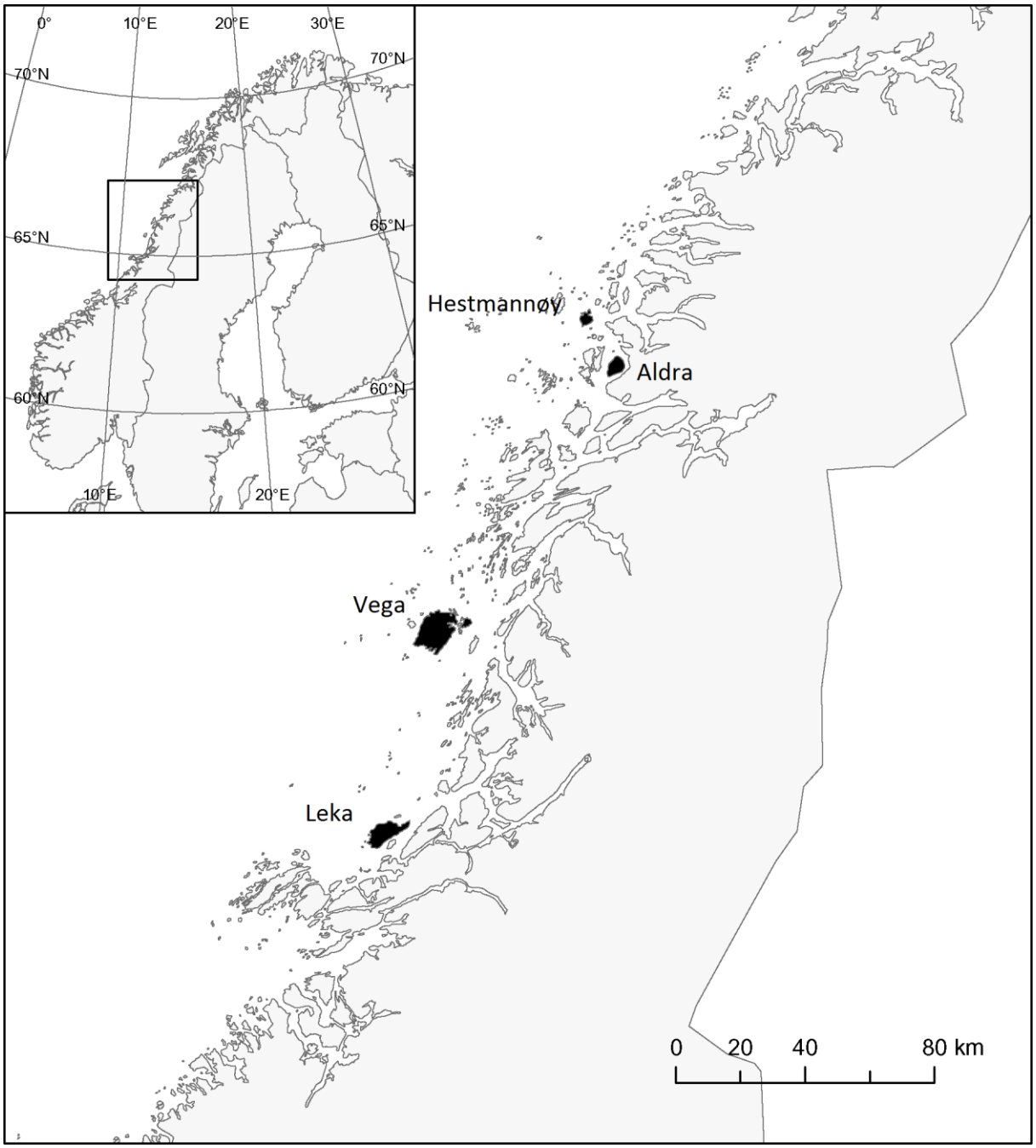
647

648

649 **Figures**

650

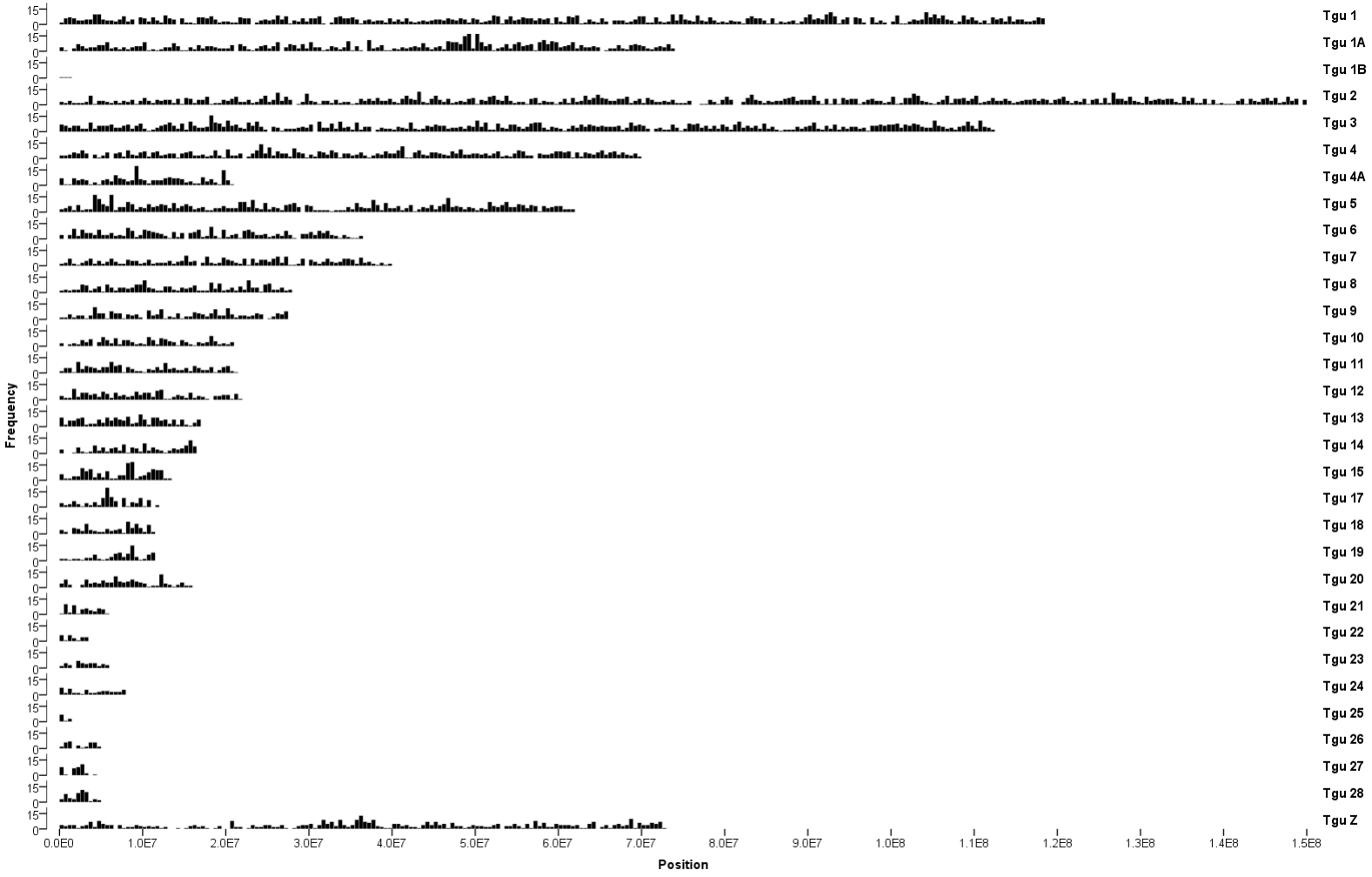
651 **Figure 1**



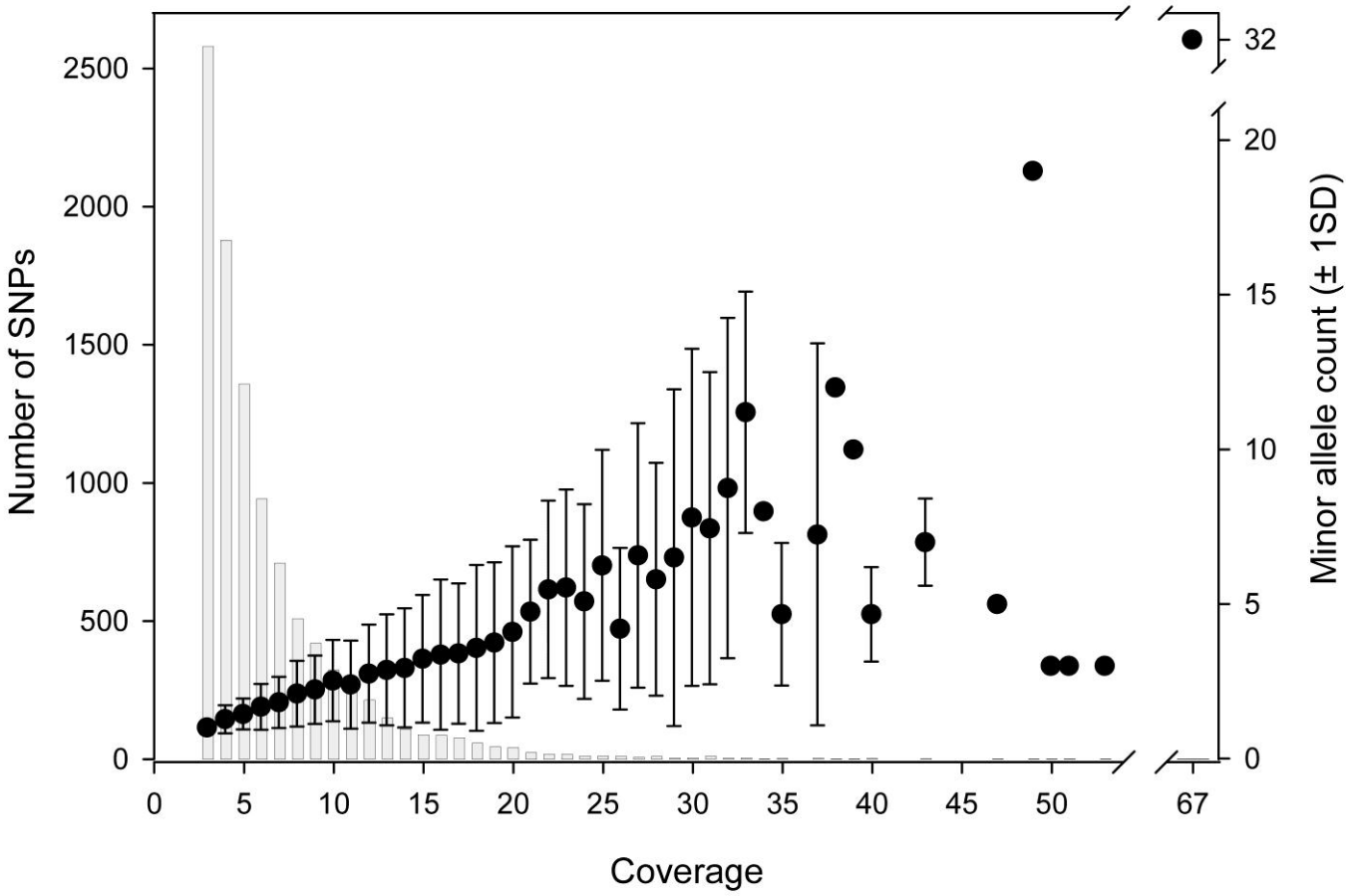
652

653 **Figure 2**

654



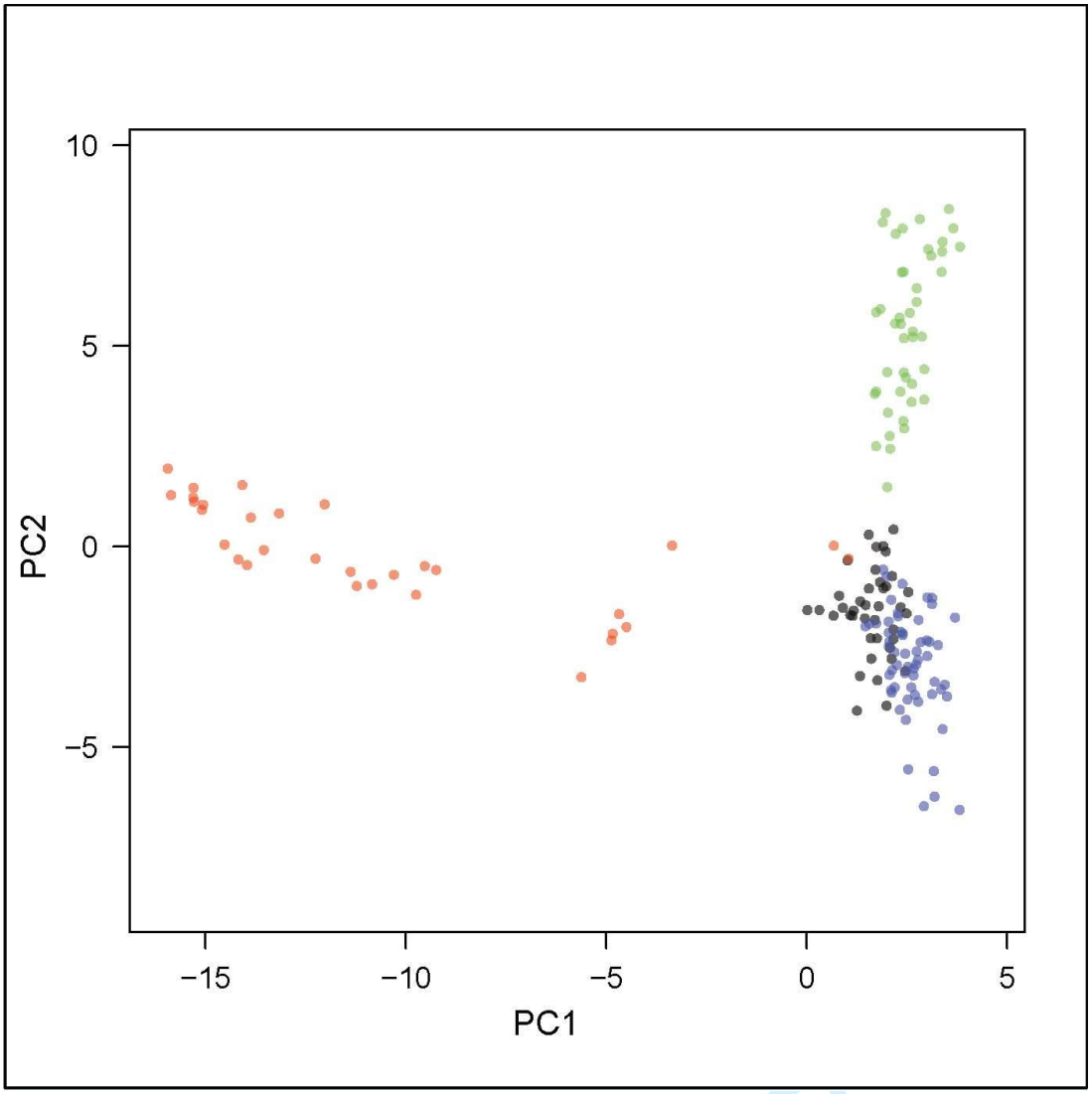
655 **Figure 3**



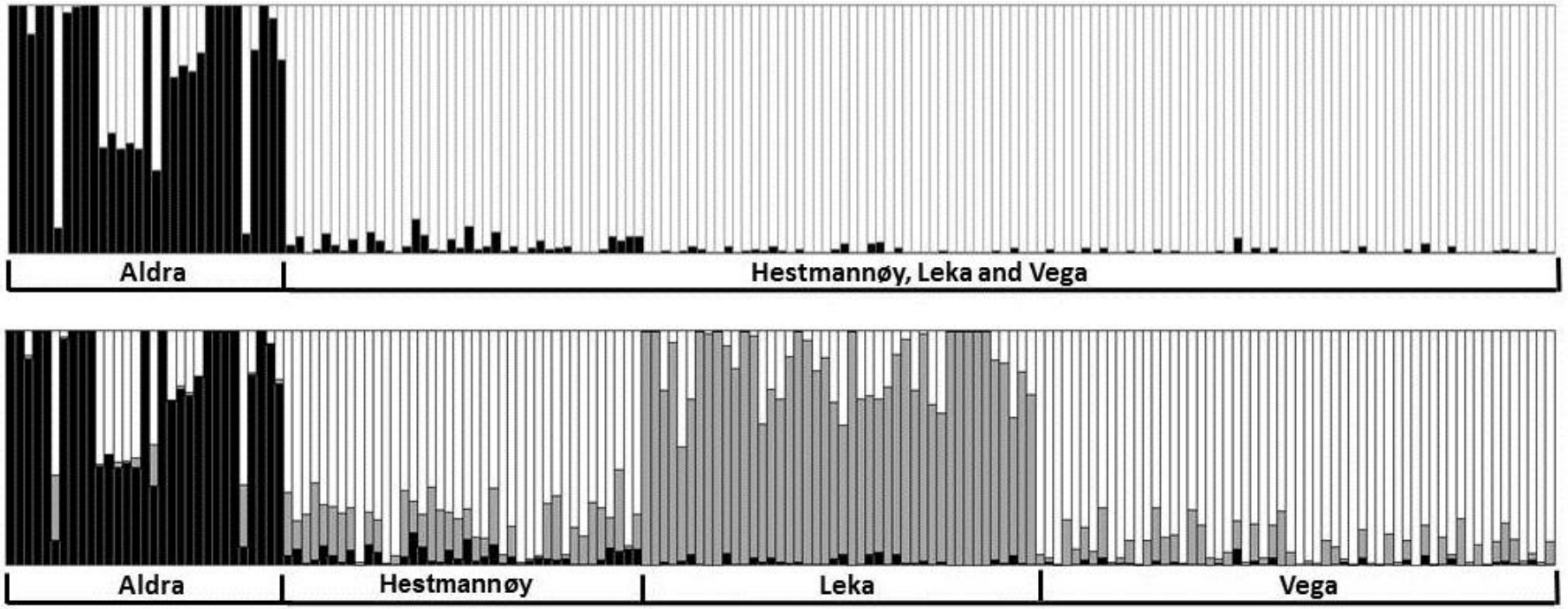
656

657

Figure 4



660 **Figure 5**



661