Received Date : 14-Mar-2016

Revised Date : 27-Jun-2016

Accepted Date : 02-Jul-2016

Article type : Resource Article

Practical low-coverage genome-wide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in non-model species

Nina Overgaard Therkildsen[1,2] and Stephen R. Palumbi

Hopkins Marine Station, Department of Biology, Stanford University, 120 Oceanview Blvd., CA-93950 Pacific Grove, USA

[1]Current address: Department of Natural Resources, Cornell University, 208 Fernow Hall, NY-14853 Ithaca, USA

[2]Corresponding author. Email: nt246@cornell.edu. Fax: 607-255-0349

Keywords: Low-coverage sequencing, library preparation, non-model population genomics, exome capture, mitogenomics

Running title: Practical low-coverage exome sequencing

ABSTRACT

Today most population genomic studies of non-model organisms either sequence a subset of the genome deeply in each individual or sequence pools of unlabeled individuals. With a step-by-step workflow, we illustrate how low-coverage whole genome sequencing of hundreds of individually barcoded samples is now a practical alternative strategy for obtaining genome-wide data on a population scale. We used a highly efficient protocol to generate high-quality libraries for ~6.5 USD from each of 876 Atlantic silversides (a teleost fish with a genome size ~730Mb) that we sequenced to 1-4x genome coverage. In the absence of a reference genome, we developed a bioinformatic pipeline for mapping the genomic reads to a *de novo* assembled reference transcriptome. This provides an 'in silico' method for exome capture that avoids the complexities and expenses of using wet chemistry for target isolation. Using novel tools for analysis of low-coverage data, we extracted population allele frequencies, individual genotype likelihoods and polymorphism data for 2,504,335 SNPs across the exome for the 876 fish. To illustrate the use of the resulting data, we present a preliminary analysis of geographic patterns in the exome data and a comparison of complete mitochondrial genome sequences for each individual (constructed from the low-coverage data) that show population colonization patterns along the US East coast. With a total cost per sample of less than 50 USD (including sequencing) and ability to prepare 96 libraries in only five hours, our approach adds a viable new option to the population genomics toolbox.

INTRODUCTION

DNA sequencing costs have decreased more than 300,000-fold since the turn of the century (Wetterstand 2015). Yet, despite the increasing affordability of sequencing data, researchers remain faced with decisions about how to distribute sequencing effort between breadth of genomic coverage, depth of coverage per individual, and the total number of individuals analyzed (Sims *et al.* 2014).

Studies focusing on intra-specific genetic variation often need large sample sizes to accurately estimate population allele frequencies. This need has spurred the development of reduced-representation techniques that focus sequencing on a small fraction of the genome, which can then be sequenced to a sufficient depth for reliable genotype calls across many individuals. Approaches targeting the sequence flanking restriction cut sites (such as RAD-seq (Davey *et al.* 2011)) have become the method of choice for many studies on non-model organisms because they provide a relatively cheap and fast way to generate genome-wide population genomic data and do not require prior knowledge about the genome sequence, enabling important insights in a diversity of systems (e.g. Davey *et al.* 2011; Narum *et al.* 2013; Andrews et al. 2016).

However, a large proportion of RAD markers are typically located outside protein-coding sequence, precluding functional analysis of polymorphism patterns if markers cannot be anchored to an annotated reference genome (Jones & Good 2016). For studies focusing on selection and adaptation, it may be preferable to target protein-coding or promotor regions of the genome (Pespeni *et al.* 2012; De Wit *et al.* 2015). RNA-seq (sequencing of expressed transcripts (Wang *et al.* 2009; De Wit *et al.* 2012)) offers an efficient way to target sequencing to protein-coding regions and obtain transcriptome-wide polymorphism data, but requires high-quality RNA (De Wit *et al.* 2012; 2015). A more versatile alternative is targeted sequence capture with hybridization probes (Grover *et al.* 2012; Jones & Good 2016). This method requires prior knowledge about the target sequence for probe design, but several effective workflows have been developed for non-model species (e.g. Cosart *et al.* 2011; Bi *et al.* 2012). The key drawback is the high cost of synthesizing custom capture probes, as well as logistics surrounding probe design and laboratory protocols.

As an alternative to sequencing a subset of the genome deeply enough to reliably call individual genotypes, an increasingly popular option is to spread the sequencing over greater parts of the genome or across more individuals. Simulation studies have demonstrated that sampling many individuals at low read depth provides more precise estimates of population parameters than higher read depth for fewer individuals (Fumagalli 2013; Buerkle & Gompert 2013; Nevado *et al.* 2014). In fact, these studies have suggested that spreading sequencing depth to 1-2 reads per locus and individual (1-2x coverage or less) maximizes the information gained about a population. At such low read depth, individual genotype calls are highly uncertain, so this sequencing design is not suitable for analysis requiring accurate individual genotypes. However, even though most population genetic software packages currently require individual genotypes as input, there is no inherent need for genotypes to be accurately known for many types of analysis of selection, population structure or demographic history. In fact, many commonly used software packages simply collapse genotype data into allele counts for populations (Buerkle & Gompert 2013). New methods and software packages that estimate population-level statistics directly from genotype likelihoods without calling genotypes (and without genotype imputation) offer powerful opportunities for taking advantage of the full information contained in low-coverage sequence data for more accurate population parameter estimates (e.g. Li 2011; Nielsen *et al.* 2011; Buerkle & Gompert 2013; Fumagalli *et al.* 2014). Such methods can also estimate individual-based parameters like admixture coefficients and parentage probabilities by combining information across loci to compensate for low coverage at individual loci (e.g. Buerkle & Gompert 2013; Fumagalli *et al.* 2014; Lindtke *et al.* 2014; Korneliussen and Moltke 2015).

With decreasing sequencing costs and improved analytical tools, whole genome resequencing of population samples is becoming a viable strategy for many studies (e.g. Jones *et al.* 2012; Liu *et al.* 2014; Xia *et al.* 2015). Yet so far, most population-level resequencing has been conducted on pools of unlabeled individuals (Pool-seq), rather than on individually barcoded samples (reviewed by Schlötterer *et al.* 2014). Undoubtedly, Pool-seq offers the most cost-effective approach for estimating population allele frequencies across the entire genome. Both theoretical and empirical studies have demonstrated that it can generate reliable estimates when the number of individuals in a pool and the sequencing depth are sufficiently high (Futschik & Schlötterer 2010; Zhu *et al.* 2012; Schlötterer *et al.* 2014). The obvious downside is that all information about individuals is lost, making it difficult to control for uneven contribution to the pool and precluding any individual-level analysis. Although barcoded sequencing adapters make it possible to multiplex hundreds of labeled samples in a sequencing run, Pool-seq has become popular because with traditional methods, it would be very labor-intensive and costly to prepare separate libraries for hundreds of individuals (the cost could easily exceed the costs of sequencing).

Here we demonstrate a workflow that overcomes these limitations with a library preparation protocol that is sufficiently rapid and inexpensive to allow high-throughput individual sample processing, eliminating the need for pooling un-barcoded samples in many studies. The library preparation protocol was originally developed for small microbial genomes (<15 Mb; Kryazhimskiy *et al.* 2014; Baym *et al.* 2015). Our contribution is to demonstrate that it also produces high-quality libraries for a teleost fish with a genome size several orders of magnitude larger (~730 Mb), and that the library fragment lengths can easily be tuned to suit different sequencing applications. With a sequencing depth of only 1-2x per individual, we are able to recover the complete mitochondrial genome sequence for each individual and obtain individual genotype likelihoods across the nuclear genome.

The libraries can be used for whole genome resequencing analysis when a relevant reference is available for the species under study. However, even though we can now sequence the full genome of any organism, assembling a high-quality reference genome still remains a non-trivial, costly, and time-consuming task, especially for highly polymorphic species (Voskoboynik *et al.* 2013; Ellegren 2014). We therefore also demonstrate an 'in silico' exome capture approach that is a more straight-forward starting point for organisms with no prior genomic resources. The idea is to take advantage of our simple sample preparation procedure for shotgun whole genome sequencing, but then focus the analysis on reads that map to a reference transcriptome, which is easier to assemble *de novo* than a full genome (see also Lamichhaney *et al.* 2012 for a Pool-seq version of this approach).

In this paper, we describe our step-by-step workflow (Fig. 1) for using this approach to generate exome-wide data for the Atlantic silverside, a small estuarine fish (730Mb estimated genome size). With a total of cost of 50 USD per sample (including both library preparation and 1x genome sequencing) and the ability to process 96 samples in only five hours, our study illustrates that this method will be a viable strategy for obtaining individual-level genome-wide data for many organisms.

MATERIALS AND METHODS

Study organism and samples

Our target species is the Atlantic silverside *Menidia menidia*, a small estuarine fish with almost no prior genomic resources. The most closely related species with fully annotated reference genomes in the Ensembl browser are the medaka (*Oryzias latipes*), the platyfish (*Xiphophorus maculatus*) and the tilapia (*Oreochromis niloticus*) with estimated genome sizes ranging from 730 to 927 Mb (golden path length in Ensembl release 82; Cunningham *et al.* 2015). These species all diverged from the silverside more than 100 million years ago (Setiamarga *et al.* 2009; Near *et al.* 2012), so are unlikely to be good references for mapping genomic silverside reads, hence the need for a species-specific reference.

Our target samples had been stored whole in a -20°C freezer for between 8 and 17 years. These fish were either collected directly from the wild at one of four different locations along the East coast of North America (from Hice *et al.* 2012) or were generation F1-F11 from a selection experiment started with wild-caught parents collected in New York (from Conover & Munch 2002). Because RNA was unrecoverable from these archived individuals, we collected eight fresh specimens in New York for RNA-seq to generate a reference transcriptome.

Generating a reference transcriptome

To capture a broad diversity of transcripts expressed at different life stages and in different tissue types, we prepared cDNA libraries with RNA from 5 whole silverside larvae and multiple tissues from 3 adults, sequenced them in one Illumina HiSeq lane with 100 bp paired-end reads, and merged *de novo* assemblies created from the combined read set with two different programs (CLC Genomics Workbench (http://www.clcbio.com) and Trinity (Haas *et al.* 2013)). To reduce redundancy in our merged assembly, we meta-assembled transcripts based on sequence similarity, and used a sequential reciprocal best-hit-blast approach to select a single best representative for each non-redundant protein from related fish species with available reference genomes (platyfish, medaka and tilapia). We also added transcripts that contained full open reading frames but had low similarity to the reciprocal-best-BLAST hit contigs.

DNA extraction and library preparation

For the genomic analysis, we used the Qiagen DNeasy Blood and Tissue kit to extract DNA from muscle tissue. We evaluated the degradation level of each extract through 1.5%

agarose gel electrophoresis: only samples that showed clear high molecular weight bands and limited smearing were retained for library preparation. To ensure DNA integrity in the retained samples, we removed fragments shorter than ~1000 bp from each extract using Agencourt AMPure XP beads in a 0.4:1 AMPure to sample ratio, and eluted the DNA in 10mM Tris-Cl, pH 8.5. DNA yield and degradation varied between samples, but we were able to obtain good quality DNA extracts from a total of 876 individuals (at least 50 individuals from each population sample).

We measured DNA concentrations with a Quant-iT high sensitivity assay (Invitrogen) and prepared a separate barcoded library for each individual with Illumina's Nextera kit according to the protocol developed by Kryazhimskiy *et al.* (2014) (slightly modified by Baym *et al.* 2015). Briefly, the tagmentation reaction, which simultaneously fragments the DNA and incorporates partial adapters, was carried out in a 2.5 µl volume with 1.6 – 7.9 ng of input DNA for each library (this is 1/20 of the reaction volume and DNA input suggested by the manufacturer). We then used a two-step PCR procedure with a total of 12 cycles (8+4) to add the remaining Illumina adapter sequence with dual index barcodes and amplify the libraries. The PCR was conducted with the KAPA Library Amplification Kit and the Illumina Nextera index kit with primers N501-N508 + S511 and N701-N712 + N714. As a final step, we purified and size-selected the amplification products with Agencourt AMPure XP beads and quantified the concentration of the final libraries with the Invitrogen Quant-iT high sensitivity assay (see Table S1 for reagent catalogue numbers). We also examined the fragment size distribution of multiple libraries from each plate on an Agilent BioAnalyzer instrument. Similar to what was shown in the original implementation by Baym et al. (2015), the entire library preparation protocol can be completed for 96 samples in less than five hours for about 6.50 USD per sample (Table S1), reducing the cost more than 10-fold compared to the regular protocol for Illumina's Nextera kit.

In addition to slightly increasing the amount of input DNA for some samples, our only modifications to the Kryazhimskiy *et al.* (2014) protocol was to extend the elongation step to 3 minutes in the initial PCR and to 2 minutes in the reconditioning PCR (to promote amplification of longer fragments) and to change the size selectivity in the final library purification step. We made these changes because the majority of fragments in our first batch of 76 libraries prepared with the original protocol were shorter than the combined length of the paired reads and the adapter sequence (2x125 bp read length + 138 bp adapter sequence = 388 bp, see results). By using a 0.6:1 AMPure XP beads to library ratio (rather than the 1:1 ratio used in the original protocol), we increased average fragment lengths by almost 100 bp.

We combined equimolar amounts of 56-76 libraries into separate pools for sequencing in 13.5 lanes of paired-end 125bp reads on an Illumina HiSeq 2000 (v4 chemistry) at the University of Utah's Bioinformatics Core Facility. To even out the data yield among samples, we re-pooled libraries that initially had obtained the lowest read output for supplementary sequencing in 4.5 additional HiSeq lanes.

To assess the effect of our preparation method on library bias and complexity, we compared our results to two libraries that we had previously prepared with Illumina's TruSeq DNA PCR-Free Sample Preparation Kit. Each of these libraries were prepared according to the manufacturer's instructions for the 550bp insert size workflow (except using a Branson Sonifier Cell Disruptor 200 for fragmentation instead of a Covaris instrument) with a total of 2 µg input DNA pooled in equimolar amounts from 50 of the same silverside individuals included in the low-coverage set. The two pooled PCR-free libraries were sequenced in 1.5 lanes of paired-end 125bp reads on an Illumina HiSeq instrument.

Data filtering

Prior to any downstream analysis, we filtered the raw reads to remove potential artifacts and low quality data. We first removed exact duplicate read pairs (likely caused by PCR amplification) with the program Fastuniq v1.1 (Xu *et al.* 2012), and then used Trimmomatic v0.32 (Bolger *et al.* 2014) run in both palindrome and simple mode to trim off adapter sequence. We also used the Trimmomatic sliding window approach to trim off the rest of the read if the average sequence quality over any four bases fell below 15. We used the program FLASH v. 1.2.9 (Magoč & Salzberg 2011) to merge overlapping paired-end reads into single consensus reads and removed sequence that mapped to potential contaminant sources including human, bacterial and viral genomes (between 0.4 and 1.4% of reads per library, Supplementary note 1).

Mapping

Through a benchmarking test with the program Teaser (Smolka 2015), we identified Bowtie2 v2.2.3 (Langmead & Salzberg 2012) in preset mode --very-sensitive-local as one of the best-performing mappers for our dataset (in a comparison of five commonly used mappers). We therefore used this program and preset mode to map all the genomic reads to the reference transcriptome. We used samtools v1.2 (Li *et al.* 2009*)* to filter the alignment files, retaining both unpaired, orphaned, and concordantly paired reads with a mapping quality > 20 but discarding discordantly mapped pairs. The inferred fragment lengths (based on mapping position) revealed a residual presence of overlapping read ends not merged by FLASH for 4-16% of the mapped pairs. To avoid double-counting the sequencing support during SNP calling, we used the clipOverlap program in the bamUtil package v1.0.14 (Breese & Liu 2013) to soft clip overlapping read ends (maintaining only the read with the highest quality score in overlapping regions). We removed duplicate reads with the MarkDuplicates module of Picard Tools v1.139 (http://broadinstitute.github.io/picard/) and summarized the coverage and mapping depth across the reference transcriptome for each individual with the BEDTools coverageBed program v2.19.1 (Quinlan & Hall 2010) and the samtools mpileup module (Li *et al.* 2009). The outputs from coverageBed and mpileup were processed with custom scripts to compute average mapping depths after excluding positions with >4x the mean mapping depth (likely repetitive sequence). We estimated the genome size based on the relationship between the observed mapping depth and the amount of quality-filtered sequence for each sample. The extent of GC-bias was evaluated by computing the average depth of coverage for non-overlapping 200 bp windows along the entire transcriptome with a script by R.V. Panday (available at http://www.popoolation.at/mauritiana_genome/). We compared patterns between the merged read set from 50 Nextera-style libraries and the pooled PCR-free libraries, both downsampled to 50 million mapped reads.

SNP calling and posterior genotype probabilities

Prior to calling variants, we realigned reads around indels with the GATK IndelRealigner (McKenna *et al.* 2010). We then used the program ANGSD v0.910 (Korneliussen *et al.* 2014) to call single nucleotide polymorphisms (SNPs) at sites with a probability <1e-6 of being monomorphic based on the mapped reads for all 876 individuals (excluding sites with a total read depth <300 and >3028 (mean depth +2 standard deviations) and bases with a quality score <20). We also used ANGSD to estimate allele frequencies in each population sample of 50 individuals and computed posterior genotype probabilities, which incorporate the uncertainty about true genotypes for each individual at each SNP. The posterior genotype probabilities for all SNPs with a global minor allele frequency >1% were used to estimate the covariance matrix among individuals with ngsTools (Fumagalli *et al.* 2014), and we used the base package in R for eigen-decomposition to summarize the covariance patterns with a principle components analysis.

Mitochondrial genome reconstruction

We mapped the genomic reads to the mitochondrial genome sequence for *M. menidia* (Genbank Acc. GI: 197311199; Setiamarga *et al.* 2008) with the same procedure and filtering criteria as described above. We then used freebayes v0.9.21-5-g018c661 (Garrison & Marth 2012) in haploid mode to call variants for each sample, requiring at least three observations of alternative alleles and retaining only non-reference genotype calls with phred-scale quality >20. Based on the haploid genotype calls, we extracted a consensus mitochondrial genome sequence for each individual with the vcf2fasta tools from the vcflib package (https://github.com/ekg/vcflib). We generated a multiple sequence alignment of these sequences with Clustal Omega (Sievers *et al.* 2011), and used the R-package haplotypes (Aktas 2015) to compute a pairwise mismatch distribution and quantify the number of unique mitochondrial haplotypes observed within each population. To illustrate the applicability to species without a pre-existing mitochondrial genome reference, we in parallel *de novo* assembled the mitochondrial genome for our 200 samples collected along the geographical cline with the baiting and iterative mapping approach implemented in the program MITObim v.1.8 (Hahn *et al.* 2013). We used only the commonly available COI barcoding gene sequence as a starting seed and used both the "denovo" and the "mapping" mode to iteratively extend to the full sequence based on matching reads from the quality-filtered read pool from each individual.

RESULTS

Reference transcriptome

The final nuclear transcriptome assembly contained 20,998 contigs with a minimum length of 200 bp, an N50 of 3,347 bp, and a combined length of 53.3 Mb. The average GC-content was 49.7% (compared to ~40% across all the genomic reads). Overall, 74-78% of RNA-seq reads mapped uniquely onto the assembly and only 1-2% of reads mapped to multiple locations, indicating a low level of redundancy. The contig set shows significant BLASTx hits to 84% of gene models in the related platyfish genome. For 74% of these genes, the top HSP covers >90% of the total length of the reference protein, indicating complete or nearly complete transcripts.

Genomic sequence data quality

The first round of sequencing yielded 0.17 – 4.16 Gb of raw sequence per library (757 Gb in total). Our follow-up sequencing added another 229 Gb, resulting in an average of 1.13 Gb raw sequence data (~9 million read pairs) per individual (see Supplementary Note 2 and Fig. S1 for details on how we evened out the sequencing effort across individuals). Quality trimming removed on average only 3% of bases, except from one lane with notably lower quality scores, which had an average of >9% of bases removed in the filtering.

The two size-selection protocols had substantial impact on the sequence usability. With the original protocol (Kryazhimskiy *et al.* 2014), the two ends overlapped for 74% of read pairs, with an average overlap of 58 bp, making 22% of the read data redundant after merging. The short length of our inserts also caused adapter read-through in the shortest constructs, so another 8% of the raw data were adapter sequence. The modified protocol substantially increased the mean insert length in our libraries, resulting in a much lower proportion of paired reads overlapping (11-64%, average of 25.4%). This reduced overlap redundancy to 6.3% and the level of discarded adapter sequence to 1.9% (Supplementary Fig. S2), so that on average 87% of bases were retained for mapping after all quality filtering. One aspect of our results that bodes well for future use of this approach is that the amount of input DNA did not appear to systematically affect the mean insert length or the total amount of quality-filtered data generated per library (Supplementary Fig. S3-5)

Mapping

On average, 14% of the reads mapped uniquely to the reference transcriptome with a mapping quality >20. Almost half of these (46%) were mapped in concordant pairs (both ends mapped to the same contig in the expected orientation and with an inferred insert length <1500bp). Less than 3% of the mapped reads were in discordant pairs with the two ends mapped to different contigs, and these were filtered out. Because we mapped genomic reads to a transcriptome reference that spans intron-exon boundaries, ~37% of all the mapped reads were orphans from pairs where only one end mapped (the other end likely originated from intronic or intergenic sequence that is ignored for the present analysis). Mapping depth tended to only decrease moderately around intron-exon boundaries, however, because the applied mapping algorithm allows local mapping of reads that only match the reference for part of their length (Fig. 2). Indeed, distinct break positions where a large number of reads either start or end their local alignment can be observed in the alignment files (Fig, 2). As also highlighted by Montes *et al.* 2013, localizing these break positions represents a novel way to infer the location of exon-intron boundaries within the transcriptome reference.

Library complexity, mapping depth and GC-bias

Ideal sequencing libraries should contain a large number of unique DNA fragments so that most molecules are sequenced only once, and these fragments should be randomly sampled from the genome for homogenous sequencing coverage. The MarkDuplicates algorithm flagged 3-14% (on average 6%) of our mapped reads as potential duplicates based on identical mapping positions. However, the duplication rates estimated for single-end reads (for which only the mapping position of one end of a fragment is known) were up to 28x greater than the duplication rates estimated for paired reads (for which the mapping position of both ends of a fragment can be taken into account, Fig. 3), indicating that the overall estimates are inflated by false positives (see Supplementary Note 3 and Fig. S6). As expected, the total amount of sequence generated for each library correlated strongly with the duplication rate estimates (Fig. 3). Yet, even in the libraries for which we had 5x genome coverage, the highest paired-read duplication rate estimated was only 2.4% (the inflated estimate including single-end reads was 12%), indicating a high level of complexity and very limited PCR duplication in the libraries.

About 1.6% of positions in the transcriptome sequence showed excessive mapping depth (greater than 4 times the mean depth) - likely because of repetitive sequence or assembly errors - and these positions were excluded from all further analysis. For the rest of the transcriptome, the average mapping depth was 1.3x per individual (Supplementary Fig. S7). Due to the stochastic sampling process involved in sequencing, the reads did not cover the entire reference sequence for each individual. However, within individuals, an average of 66% of positions in the transcriptome were covered by at least one read. In samples with greater than 3.5x coverage, >90% of positions were covered (Fig. 4). The number of years (8-17) a sample had been stored frozen did not seem to affect the average mapping depth or transcriptome coverage, indicating robustness to variable levels of degradation. Across samples, we observed the highest depth of coverage in regions with GC-content between 30 and 50% and reduced coverage in regions with extreme GC-content. Because an almost equally strong bias was seen for the PCR-free libraries (Fig. 5b), a large part of this bias most likely arose during sequencing and not in library preparation. Within the range of GC-content observed across 99% of the transcriptome (26.5-73.6% GC, Fig 5a-b), the average coverage only varied two-fold. The fraction of the reference not covered by any reads was identical for the PCR-free and the Nextera library sets (2.1%), and although the slightly higher GC-bias resulted in a slightly less even distribution of reads than for the PCR-free libraries (Fig. 5c), we consistently saw an average coverage >60x across the population groups of 50 individuals with >82% of bases covered by reads from at least 25 of the 50 different individuals (Supplementary Fig. S8).

Genome size estimation and per sample sequencing costs

We did not sequence any samples deeply enough to robustly estimate the silverside genome size with k-mer based approaches (Marçais & Kingsford 2011; Chikhi & Medvedev 2014). However, comparisons of the amount of cleaned sequence per individual and the average mapping depth achieved across the reference transcriptome (excluding highly repetitive sites) showed that we needed about 730 Mb of cleaned sequence reads to get an average of 1x mapping depth, indicating that this is the approximate size of the silverside genome (similar to genome sizes reported for related species). Given that our conservative quality filtering discarded about 13% of the raw data, we would need about 840 Mb of raw reads per individual to achieve 1x genome coverage, which means that 65 individuals with a silverside-sized genome can be pooled in a single Illumina lane (assuming 54Gb of raw sequence output, as we saw on average). This brings the current cost of sequencing to ~40USD / individual (for 1x genome coverage; see Table S2 for a comparison of sequencing costs for organisms with different genome sizes).

SNP calling and posterior genotype probabilities

ANGSD detected a total of 2,504,335 SNPs with a minor allele frequency >1% across all samples. The population-specific allele frequencies for each of these SNPs estimated from the genotype likelihoods can be used for a suite of downstream applications including scans for signatures of selection and reconstruction of demographic history (Therkildsen et al.*, in prep.*). Despite the low certainty about individual genotypes for each individual SNP (given the only 1-2x coverage), the PCA that integrated genotype probability information across the entire exome clearly grouped individuals into distinct clusters that largely corresponded to geographic origin (Fig. 6). The first principle component separates the geographical samples along a North-South continuum, while PC 2 shows the distinctness of silversides from the Gulf of St. Lawrence, as has been reported from D-loop sequence data (Mach *et al.* 2010).

In addition, the PCA also reveals that a few individuals cluster with a different group than most other individuals collected at the same locations, indicating that they may be migrants. The effects of such divergent individuals would be impossible to quantify in pooled sequencing designs.

Mitochondrial genome reconstruction

On average 0.2% of the quality-filtered reads mapped to the mitochondrial genome with a mapping quality >20, resulting in an average mapping depth of 79x along the mitochondrial genome for each individual (range 2x-740x). The ratio of mitochondrial to transcriptome mapping depths depended strongly on the amount of time a sample had been stored frozen prior to sequencing (Fig. 7). Samples that had been stored for 17 years showed only about 6 times greater mapping depth for the mitochondrial genome compared to the transcriptome, while the enrichment was on average 99 times for samples stored for only 8 years. This large and consistent difference cannot be attributed to library preparation or sequencing lane effects because samples had been randomized between batches. As mentioned above, we also did not see a similar effect of sample age on the mapping rate for the transcriptome. These results therefore indicate differential rates of degradation of mitochondrial and nuclear DNA in storage.

Among our samples, 784 (98%) of the samples stored for fewer than 17 years had a mapping depth of at least 5x throughout >95% of positions in the mitochondrial genome, allowing high-confidence haploid genotype calls. The multiple alignment of all consensus sequences revealed that 1,287 of the 16,458 positions in the mitochondrial genome were variable (651 variants were singletons, i.e. the alternative allele was only observed in a single individual). All individuals differed by at least four positions from the reference sequence, but many individuals had completely identical sequences, resulting in only 272 unique haplotypes among the samples. The unique sequences differed by on average 40.7 positions providing strong resolution for tracking of individual lineages.

Among the F1 offspring from wild parents, 89% of haplotypes were unique. By contrast, only 13% of haplotypes were unique in F6 samples after five generations of rearing in the lab, and 7% were unique in samples collected after 10 generations of lab rearing, illustrating strong reduction of maternal lineages and potential inbreeding.

For the field-collected individuals from the four geographic locations, pairwise mismatch distributions show that the southernmost samples have the most divergent mitochondrial genomes (largest number of polymorphisms in pairwise comparisons), while the mitochondrial genomes sampled at northern locations differ at fewer sites (Fig. 8). This pattern is consistent with northward migration from southern refugia since the last glaciation, as has been previously suggested for this species based on D-loop sequences (Mach *et al.* 2010). The full-length mitochondrial genomes provide a much richer data set for detailed reconstruction of demographic history of this species.

We also tested whether we could reconstruct mitochondrial genomes without a full reference sequence. Using only COI as a seed and MITObim's default parameters, we were able to *de novo* assemble the full mitochondrial genome in a single contig for 166 of the 200 geographical samples over 14-157 iterations. Parameter optimization or using a mitochondrial genome sequence from a related species could probably improve the success rate substantially, showing that the low-coverage data make it possible to recover full mitochondrial genomes even without a pre-assembled reference.

DISCUSSION

This study demonstrates that whole genome sequencing of hundreds of individuals now represents a practical approach for obtaining population genomic data from model and non-model species alike. We illustrate an efficient workflow for preparing high-quality individually barcoded libraries that enable more robust and versatile data analysis than pooled sample sequencing, and we show how 'in silico' capture of genomic reads can be used to generate exome-wide polymorphism data for organisms without a reference genome sequence. With a total cost per sample of less than 50 USD (including both library preparation and 1x genome sequencing), our approach offers a viable alternative to RAD-seq, physical target enrichment, and Pool-seq for many studies.

High-quality libraries despite low DNA input amounts

The Nextera-based libraries showed only slightly greater GC-bias and a slightly less even genomic distribution of reads than previous libraries generated from a subset of the same samples with a PCR-free method, indicating that the cost reduction and simplified workflow do not substantially compromise library quality. We also saw that DNA input amounts as low as 2 ng did not significantly reduce the diversity of sequenced molecules in libraries prepared from the ~730 Mb silverside genome, as we consistently saw duplication rates <2.4%, even in libraries sequenced to >4x genome coverage. The ability to generate high-complexity libraries with ng amounts of input DNA makes whole genome sequencing available for a broad set of samples that may not yield sufficient DNA for alternative methods that require as much as 1-2 μg of DNA.

Robustness to DNA degradation

The efficiency with which DNA is incorporated into the library probably depends to some degree on DNA quality. For nuclear DNA, we saw no consistent differences in the quality of libraries generated with samples stored in a minus 20°C freezer for 8 to 17 years. We did, however, only use DNA extracts that contained high molecular weight DNA, and used a pre-cleaning procedure to remove fragments shorter than 1000 bp prior to DNA quantification. The notable decrease in enrichment for mitochondrial DNA with sample age suggests faster decay of mitochondrial vs. nuclear DNA under our sample storage conditions (whole fish frozen at -20°C). This faster decay is surprising considering that mitochondrial fragments are often the only recoverable DNA from ancient samples and have been reported to decay at slower rates than nuclear DNA over long time scales (hundreds to thousands of years; Schwarz *et al.* 2009; Allentoft *et al.* 2012). Several other studies have reported faster decay of mitochondrial than nuclear DNA in experimental comparisons, however, indicating that different processes may affect short- and long-term DNA preservation (Foran 2006; Higgins *et al.* 2015).

Efficient workflow with easy tuning for different sequencing applications

The entire library preparation protocol can be completed for 96 samples in less than 5 hours total time, and without expensive specialized laboratory equipment. We found that simply altering the ratio of AMPure XP beads to library volume caused consistent shifts in the size distribution of fragments in the final library, thereby allowing flexible tuning to different sequencing applications. To minimize redundancy from sequence overlap and adapter read-through, library fragments should be longer than the total read length (the sum of both ends for paired reads) plus the adapters. However, fragments larger than 600 bp tend to cluster much less efficiently than fragments in the 250-500 bp range on Illumina flow cells (Bronner *et al.* 2013). This resulted in lower sequence yields for libraries with longer fragments, so the optimal balance between data loss and data yield was reached for libraries with inserts between 250 and 350 bp (Supplementary Fig. S5).

Contrary to several previous studies (Adey *et al.* 2010; Baym *et al.* 2015) and Illumina's Nextera Technical Note, we did not find that variation in the amount of input DNA (in the range 1.6-7.9 ng) predictably affected the fragment length distribution of the libraries. A similar result has been reported in another study optimizing the Nextera protocol for a range of different organisms (Lamble *et al.* 2013), suggesting that exact normalization prior to library preparation - currently the most time-consuming part of the protocol - may not always be necessary and can perhaps be replaced with a bead normalization (Lamble *et al.* 2013). However, since the effect of input concentration is likely to vary between organisms (depending e.g. on genome composition, GC-content, and sample quality), we echo the recommendation of Baym *et al.* (2015) to always calibrate input amounts and size selection conditions on dilution series of a few representative samples before scaling up library preparation for novel DNA sources.

Individual barcoding vs. pooled sequencing

Our study shows that the price difference between pooling unlabeled DNA into a single library and preparing separate barcoded libraries for each individual is diminishing. Our library preparations cost ~6.5 USD per sample, so the total reagent cost for individual barcoding of 50 individuals is about ~325 USD, a modest portion of an overall NGS sequencing budget. Considering furthermore that we can prepare 96 individual libraries in the same amount of time it takes to make a pooled library, we believe that our approach is practical for broad adoption. The slightly higher cost compared to Pool-seq should for many studies be well worth the much increased versatility of individual-level low-coverage data compared to an anonymous pool of sequences.

A key advantage of individually barcoded data is the ability to account for uneven sequencing depth among individuals. This is particularly important when degradation levels vary between samples, as was the case for our silversides. Perhaps more important, however, is the suite of additional individual-level analysis made possible. Novel tools that integrate signals from low-confidence genotype likelihoods across thousands of nuclear loci now allow accurate inference about relatedness and genetic similarities among individuals even at 1-2x coverage (Fumagalli 2013; Gompert *et al.* 2014; Korneliussen *et al.* 2014; 2015; Snyder-Mackler et al. 2016; Vieira *et al.* 2015). As illustrated here with the PCA, this individual-level analysis can help detect cryptic differentiation within a population sample. When undetected in Pool-seq data, such heterogeneity among individuals can dramatically bias population allele frequency estimates. Estimation of individual admixture proportions also make certain types of individual clustering possible with low-coverage data (Skotte *et al.* 2013), so that analyses of population structure are not limited to pre-defined sample groupings. In a similar vein, individually barcoded individuals are not limited to a single pool. Samples can be re-grouped in many different configurations, for example according to different phenotypic traits for association studies (Kim *et al.* 2011; Skotte *et al.* 2012). This allows for flexible analysis of both allele frequencies and patterns of linkage disequilibrium across the genome (Li 2011; Maruki and Lynch 2015). We foresee a rapid development in tools for analyzing sequencing data in a probabilistic framework without calling genotypes over the next years, which should enable even more powerful analysis from low-coverage sequencing.

Another “added bonus” of barcoding individuals is the high-confidence full mitochondrial genome sequence we were able to recover for each individual, allowing powerful inference about maternal ancestry and evolutionary history. As utilized in genome skimming projects (Straub *et al.* 2012; Dodsworth 2015), high copy numbers per cell typically ensures deep sequencing of the mitochondrial genome even when the nuclear genome-wide coverage is shallow. The average of 79x coverage we observed across the mitochondrial genome for each individual allowed for both *de novo* assembly of the full mitochondrial sequence and polymorphism detection based on mapping to a pre-existing reference. The resulting full-length mitochondrial genome sequences, representing non-recombining, maternally inherited markers, provide valuable complementary information to nuclear data at a resolution not achievable with earlier techniques. A previous study that examined a 340 bp fragment of the D-loop in 1,029 silversides found the same haplotype in 66% of their samples and most other haplotypes differing only by a single substitution (Mach *et al.* 2010). The full mitochondrial genomes of our 200 individuals collected over the same geographical cline differed on average by >40 nucleotides (Fig. 8). At a divergence rate of 2% per million years, two 16,500 bp mitochondrial genomes will differ by 1 change per 3000 years. As a result, full mtDNA data sets have the possibility of capturing demographic events that have occurred during the last several glacial cycles. Such changes may be important in temperate species affected by swings in latitudinal range. In our data set, there is a strong peak in northern populations at an average of 6 bp differences, representing perhaps invasion of these northern populations over the last 18,000 years as glaciers receded. Peaks in southern populations at 18, 32 and 66 differences perhaps suggest lineage divergence that occurred 50,000, 90,000 and 200,000 years ago. More detailed recent historical demography might be mined from these kinds of data sets.

Expanding the tool box for both model and non-model species genomics

As discussed above, dividing a given sequencing effort over individually barcoded samples should provide substantially more robust and versatile data than sequencing a pooled sample to the same combined depth. Our cheap and efficient library preparation method should therefore be useful for many population whole genome resequencing studies of species with full reference genome sequences.

Our “brute force” approach to obtaining exome and mitochondrial sequence for a species without a reference genome may seem inefficient since we only end up using ~8% of the data. However, the cost of non-target sequencing must be weighed against the cost and effort associated with isolating the target parts of the genome prior to sequencing. Although several workflows have been developed for designing capture probes for non-model species (reviewed by Gasc *et al.* 2016 and Jones and Good 2016) and methodological improvements along with new commercial providers have made capture technology more affordable, the cost of custom-synthesizing probes the entire transcriptome for hundreds of individuals is still relatively high. Although it is currently somewhat more expensive to capture ‘in silico’ rather than with physical baits for many applications, the benefits include a much simpler and faster high-throughput sample preparation method that introduces less bias, and, importantly, the availability of all the non-target genomic reads for later analysis. Because our approach is based on shotgun genome sequencing, the sequencing cost per sample will scale with the genome size of the study organism, so the relative advantages of ‘in silico’ vs. physical capture will depend strongly on the genome-size of the organism under study (see Supplementary Table S2 for an comparison of the per individual cost of our approach for organisms with different genome sizes). However, as sequencing costs further decrease, the set of organisms for which our approach is cost-effective will expand.

For some time to come, our method will be more expensive than RAD-seq for many studies. However, the ability to focus on specific parts of the genome adds a valuable functional dimension to data sets. Our approach also does not suffer from the technical artifacts associated with polymorphisms in restriction cut sites that in RAD-seq can reduce overlap in the loci sequenced in different individuals and cause allele dropout that may complicate and bias inferences, especially when the genetic diversity within species is high (Gautier *et al.* 2012; Arnold *et al.* 2013). We therefore believe that ‘in silico’ capture adds a new promising tool to the non-model species population genomics toolbox. We have shown here for the silverside how it can be incorporated into a workflow for species with no prior genomic resources and can powerfully address both individual-level and population-level questions. For applications where high-confidence individual genotypes are required, however, other methods will be more appropriate, so method selection should always depend on particular study objectives.

Future potential

We here demonstrate a workflow focused on 'in silico' exome capture with a *de novo* assembled transcriptome reference as a tractable and straightforward starting point for population genomic analysis. Since the raw data cover the whole genome, however, many other regions of interest can be targeted (e.g. microRNAs, transcription factors or transposable elements). An alternative strategy could also be to *de novo* assemble the genomic reads to the extent possible with only low-coverage sequencing of polymorphic individuals, potentially using the transcriptome reference contigs as seeds for assembling genomic regions flanking the exons (Lamichhaney *et al.* 2012; Ruby *et al.* 2013). The resulting longer contigs can then be used as a more extensive reference for mapping, covering a larger part of the genome (but also complicating the analysis pipeline). Reference-free variant calling methods also offer promising prospects (Iqbal *et al.* 2012; Leggett & MacLean 2014; Uricaru *et al.* 2015), and with the increasing availability of long sequence reads and methods for improving assembly contiguity, it will become easier and faster to develop full genome reference sequences for new species, making low-coverage whole genome re-sequencing data even more powerful. Our exome-focused approach allows a short-cut to start exploring genomic patterns in versatile sequence data that will have a long-lasting value and enduring applicability as reference sequences improve.

Acknowledgements

We would like to thank Sergey Kryazhimskiy for advice on protocol modifications, Beth Sheets for assistance in the laboratory, and Steve Munch, David Conover, and Hannes Baumann for supplying samples. Matt Hare, the non-model genomics discussion group at Cornell University, three anonymous reviewers and the editor provided helpful comments on earlier versions of the manuscript. The study was funded through a Villum Foundation postdoctoral fellowship to NOT and an NSF grant to SRP (OCE-1434325).

Literature cited

Adey A, Morrison HG, Asan *et al.* (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biology*, **11**, R119.

Aktas C. 2015. haplotypes: Haplotype inference and statistical analysis of genetic variation. R package version 1.0. Available at https://cran.r-project.org/web/packages/haplotypes/index.html

Allentoft ME, Collins M, Harker D *et al.* (2012) The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **279**, 4724–4733.

Andrews, KR, Hohenlohe PA, Miller MR, Hand B, Seeb JE, and Luikart, G. 2014. Trade-offs and utility of alternative RADseq methods. *Molecular Ecology*, **23**, 5943–5946

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

Baym M, Kryazhimskiy S, Lieberman TD *et al.* (2015) Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS ONE*, **10**, e0128036.

Bi K, Vanderpool D, Singhal S *et al.* (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, **13**, 403.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Breese MR, Liu Y (2013) NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics*, **29**, 494–496.

Bronner IF, Quail MA, Turner DJ (2013) Improved protocols for illumina sequencing. *Current Protocols in Human Genetics*, **79**, 18.2.1–18.2.42.

Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.

Chikhi R, Medvedev P (2014) Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, **30**, 31–37.

Conover DO, Munch SB (2002) Sustaining fisheries yields over evolutionary time scales. *Science*, **297**, 94–96.

Cosart T, Beja-Pereira A, Chen S *et al.* (2011) Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics*, **12**, 347.

Cunningham F, Amode MR, Barrell D *et al.* (2015) Ensembl 2015. *Nucleic Acids Research*, **43**, D662–9.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

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

De Wit P, Pespeni MH, Ladner JT *et al.* (2012) The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, **12**, 1058–1067.

Dodsworth S (2015) Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science*, **20**, 525–527.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.

Foran DR (2006) Relative degradation of nuclear and mitochondrial DNA: an experimental approach. *Journal of Forensic Sciences*, **51**, 766–770.

Fumagalli M (2013) Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS ONE*, **8**, e79667.

Fumagalli M, Vieira FG, Linderoth T, Nielsen R (2014) ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, **30**, 1486–1487.

Futschik A, Schlötterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *Preprint at arXiv:1207.3907v2 [q-bio.GN]*.

Gasc C, Peyretaillade E, Peyret P. 2016. Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Research*, **44**, 4504–4518.

Gautier M, Gharbi K, Cezard T *et al.* (2012) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165–3178.

Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany*, **99**, 312–319.

Gompert Z, Lucas L, Buerkle, CA *et al.* (2014) Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. Molecular Ecology, **23**, 4555–4573.

Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.

Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - a baiting and iterative mapping approach. *Nucleic Acids Research*, **41**, e129–e129.

Hice LA, Duffy TA, Munch SB, Conover DO (2012) Spatial scale and divergent patterns of variation in adapted traits in the ocean. *Ecology Letters*, **15**, 568–575.

Higgins D, Rohrlach AB, Kaidonis J, Townsend G, Austin JJ (2015) Differential nuclear and mitochondrial DNA preservation in post-mortem teeth with implications for forensic and ancient DNA studies. *PLoS ONE*, **10**, e0126935.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, **44**, 226–232.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185–202.

Kim S, Lohmueller KE, Albrechtsen A *et al.* (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, **12**, 231.

Korneliussen T, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, **15**, 356.

Korneliussen TS, Moltke I. 2015. NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics*, **31**, 4009-4011.

Kryazhimskiy S, Rice DP, Jerison ER, Desai MM (2014) Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*, **344**, 1519–1522.

Lamble S, Batty E, Attar M *et al.* (2013) Improved workflows for high throughput library preparation using the transposome-based Nextera system. *BMC Biotechnology*, **13**, 1–10.

Lamichhaney S, Martinez Barrio A, Rafati N *et al.* (2012) Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 19345–19350.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Leggett RM, MacLean D (2014) Reference-free SNP detection: dealing with the data deluge. *BMC Genomics*, **15 Suppl 4**, S10.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lindtke D, Gompert Z, Lexer C, Buerkle CA (2014) Unexpected ancestry of *Populus* seedlings from a hybrid zone implies a large role for postzygotic selection in the maintenance of species. *Molecular Ecology*, **23**, 4316–4330.

Liu S, Lorenzen ED, Fumagalli M *et al.* (2014) Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, **157**, 785–794.

Mach ME, Sbrocco EJ, Hice LA *et al.* (2010) Regional differentiation and post-glacial expansion of the Atlantic silverside, *Menidia menidia*, an annual fish with high dispersal potential. *Marine Biology*, **158**, 515–530.

Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.

Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

Maruki T, Lynch M (2014) Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. *Genetics*, **197**,1303–1313.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

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

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

Near TJ, Eytan RI, Dornburg A *et al.* (2012) Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 13698–13703.

Nevado B, Ramos-Onsins SE, Perez-Enciso M (2014) Resequencing studies of nonmodel organisms using closely related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics. *Molecular Ecology*, **23**, 1764–1779.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Pespeni MH, Garfield DA, Manier MK, Palumbi SR (2012) Genome-wide polymorphisms show unexpected targets of natural selection. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **279**, 1412–1420.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Ruby JG, Bellare P, DeRisi JL (2013) PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda, Md.)*, **3**, 865–880.

Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, **15**, 749–763.

Schwarz C, Debruyne R, Kuch M *et al.* (2009) New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. *Nucleic Acids Research*, **37**, 3215–3229.

Setiamarga DHE, Miya M, Yamanoue Y *et al.* (2008) Interrelationships of Atherinomorpha (medakas, flyingfishes, killifishes, silversides, and their relatives): The first evidence based on whole mitogenome sequences. *Molecular Phylogenetics and Evolution*, **49**, 598–605.

Setiamarga DHE, Miya M, Yamanoue Y *et al.* (2009) Divergence time of the two regional medaka populations in Japan as a new time scale for comparative genomics of vertebrates. *Biology Letters*, **5**, 812–816.

Sievers F, Wilm A, Dineen D *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, **7**, 539.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, **15**, 121–132.

Skotte L, Korneliussen TS, Albrechtsen A (2012) Association testing for next-generation sequencing data using score statistics. *Genetic Epidemiology*, **36**, 430–437.

Skotte L, Korneliussen TS, Albrechtsen A (2013) Estimating individual admixture proportions from next generation sequencing data. *Genetics*, **195**, 693–702.

Smolka M (2015) Teaser: Individualized benchmarking and optimizationof read mapping results for NGS data. *Genome Biology*, **16**, 235

Snyder-Mackler N, Majoros WH, Yuan ML *et al.* (2016) Efficient genome-wide sequencing and low-coverage pedigree analysis from noninvasively collected samples. *Genetics*, **203**, 699-714

Straub SCK, Parks M, Weitemier K *et al.* (2012) Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany*, **99**, 349–364.

Thorvaldsdóttir H, Robinson JT, and Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, **14**, 178–192.

Uricaru R, Rizk G, Lacroix V *et al.* (2015) Reference-free detection of isolated SNPs. *Nucleic Acids Research*, **43**, e11.

Voskoboynik A, Neff NF, Sahoo D *et al.* (2013) The genome sequence of the colonial chordate, *Botryllus schlosseri*. *eLife*, **2**, e00569.

Vieira FG, Lassalle F, Korneliussen TS, Fumagalli M. 2015. Improving the estimation of genetic distances from Next-Generation Sequencing data. *Biological Journal of the Linnean Society*, **117**, 139–149.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57–63.

Wetterstrand, K. A. 2015. DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed November 5 2015.

Xia JH, Bai Z, Meng Z *et al.* (2015) Signatures of selection in tilapia revealed by whole genome resequencing. *Scientific Reports*, **5**, 14168.

Xu H, Luo X, Qian J *et al.* (2012) FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS ONE*, **7**, e52249.

Zhu Y, Bergland AO, González J, Petrov DA (2012) Empirical validation of pooled whole genome population re-sequencing in *Drosophila melanogaster*. *PLoS ONE*, **7**, e40637.

Data accessibility

The Atlantic silverside transcriptome is deposited in the NCBI GenBank Transcriptome Shotgun Assembly Sequence Database (TSA) under accession GEVY00000000.

A table with detailed statistics on all steps in the quality filtering and mapping procedures is available from the Dryad Digital Repository: doi:10.5061/dryad.ft596. This repository also includes a list of the command line arguments used for running each step in the pipeline.

Author contributions

The authors conceived of and designed the study together. NOT performed the experiments and analyzed the data with input from SRP, and the authors wrote the manuscript together.

Fig. 1. Diagram illustrating our workflow. The sample input material is marked in red, the procedure for getting from RNA and DNA to the final datasets is marked in blue, and examples of the types of information that can be extracted from the different datasets are shown in green (applications demonstrated in this paper are highlighted in darker green).

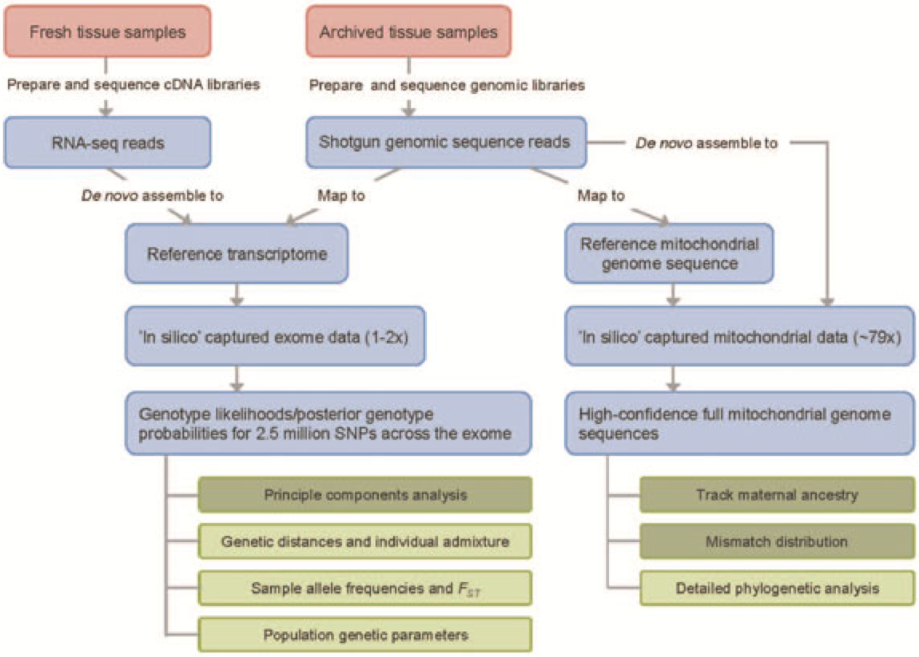

Fresh tissue samples
Archived tissue samples
Prepare and sequence cDNA libraries
Prepare and sequence genomic libraries
RNA-seq reads
Shotgun genomic sequence reads
De novo assemble to
De novo assemble to
Map to
Map to
Reference transcriptome
Reference mitochondrial genome sequence
'In silico' captured exome data (1-2x)
'In silico' captured mitochondrial data (~79x)
Genotype likelihoods/posterior genotype probabilities for 2.5 million SNPs across the exome
High-confidence full mitochondrial genome sequences
Principle components analysis
Genetic distances and individual admixture
Sample allele frequencies and F_{ST}
Population genetic parameters
Track maternal ancestry
Mismatch distribution
Detailed phylogenetic analysis

Fig. 2. Example illustration of how the genomic reads from 50 individuals map to a transcriptome contig from a protein-coding gene. The barplot in the top panel shows the total sequencing depth along the 710 bp contig and each horizontal grey bar underneath shows the mapped portion of a sequencing read. Discontinuities in the mapping boundaries indicate positions where the reads contain data from introns that do not align to the transcriptome contig, thereby denoting intron-exon boundaries. Figure generated with the Integrative Genomics Viewer (Thorvaldsdottír *et al.* 2013)

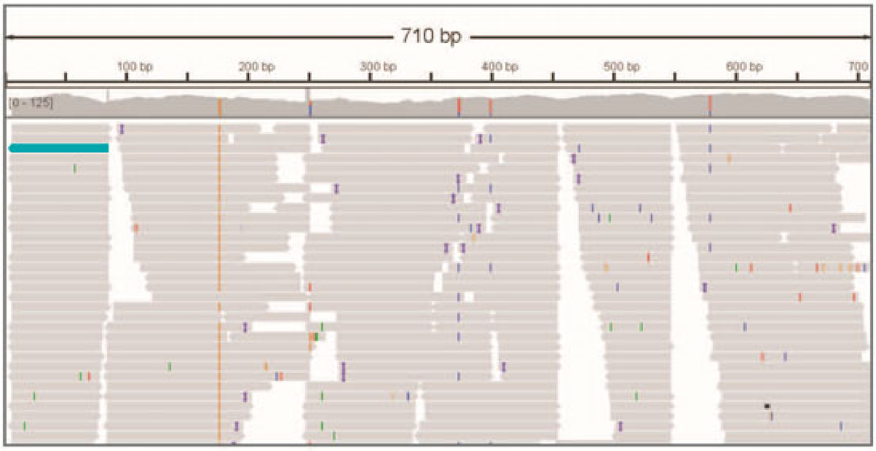


Fig. 3. Scatter plot showing estimated levels of duplication and the total amount of raw sequence generated for each individual. The dark red crosses show the overall duplication rate estimates from MarkDuplicates (paired and single-end reads), the pink crosses show MarkDuplicate estimates based on single-end reads only, the light purple dots show estimates of identical sequence duplicates (from FastUniq), and the dark purple dots show the sum of exact sequence duplicates (FastUniq) and the paired duplicates flagged by MarkDuplicates. We believe that the latter category (dark purple) provides the most reliable estimates of true duplication rates (see text).

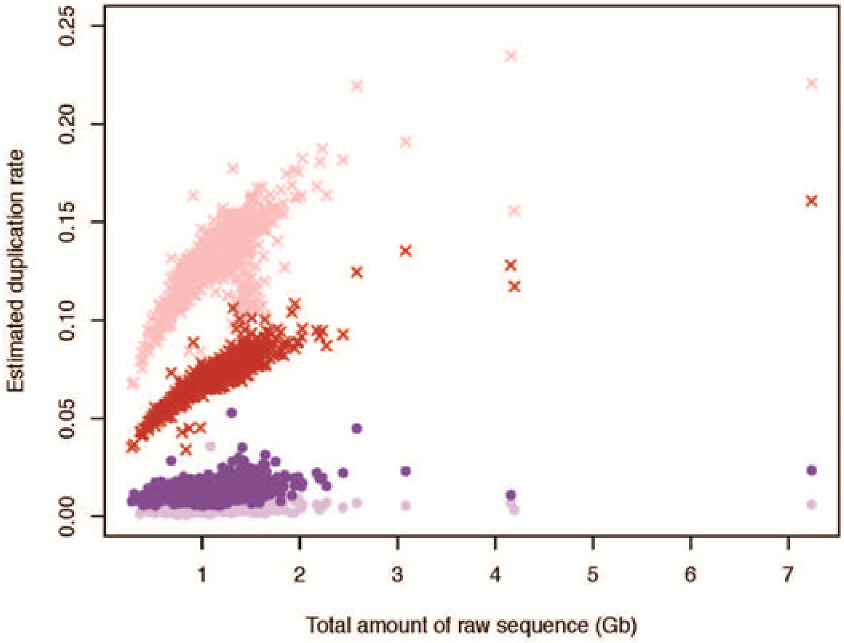


Fig. 4. Scatter plot showing the percentage of the transcriptome reference covered by at least one read as a function of the average mapped depth for each individual. The marginal histograms show the overall distributions of mapped depth per individual (x-axis) and the percentage of the reference covered by each individual (y-axis).

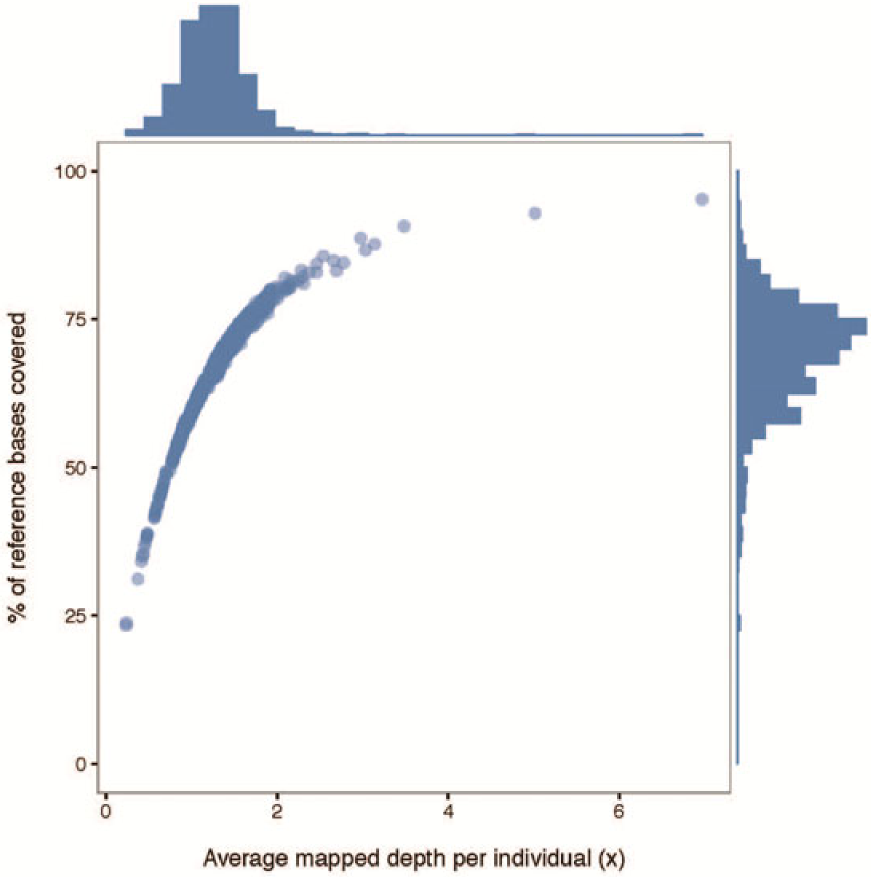


Fig. 5. Comparison of mapping patterns in the PCR-free (red) and the Nextera-based (blue) libaries downsampled to 50 million mapped reads. Panel a) shows a density plot of GC-content in 200 bp windows across the transcriptome reference. Panel b) shows the mean depth of coverage plotted against GC-content in 200 bp windows across the transcriptome for the two library types. The shading represents the 25th and 75th inner quartile regions for each library type and the vertical black lines delimit the range of GC-content observed across 99% of the transcriptome. Panel c) shows density plots of the depth of coverage per position in the transcriptome for the two kinds of libraries.

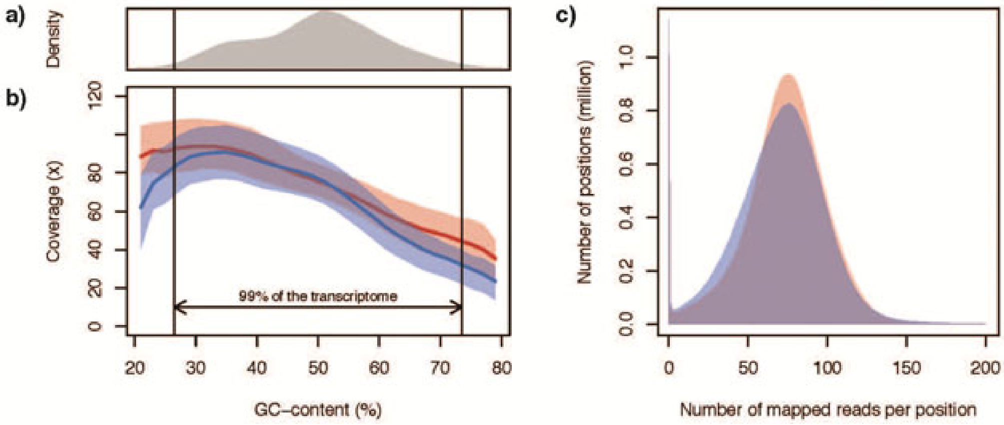


Fig. 6. Clustering of individuals from different sampling locations along the first two principle components (generated from the exome SNP data). Each transparent dot represents an individual sampled from the color-coded locations indicated on the inset map.

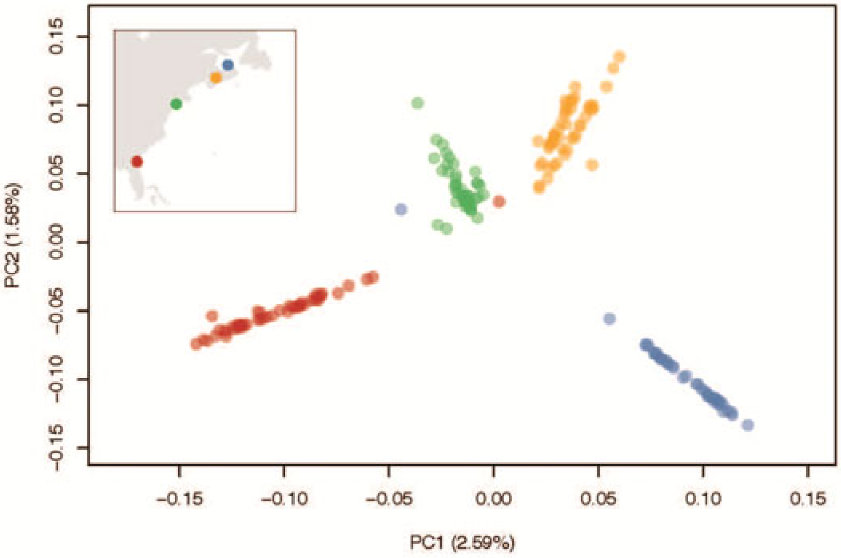


Fig. 7. Boxplot summarizing the ratio between mapping depth at the mitochondrial genome and mapping depth at the transcriptome for samples stored frozen for different lengths of time. The horizontal band in each box represents the median, the top and bottom of the boxes represent the 25th and 75th percentiles, and the error bars define the 5th and the 95th percentiles (data points falling outside these-percentiles are marked by dots).

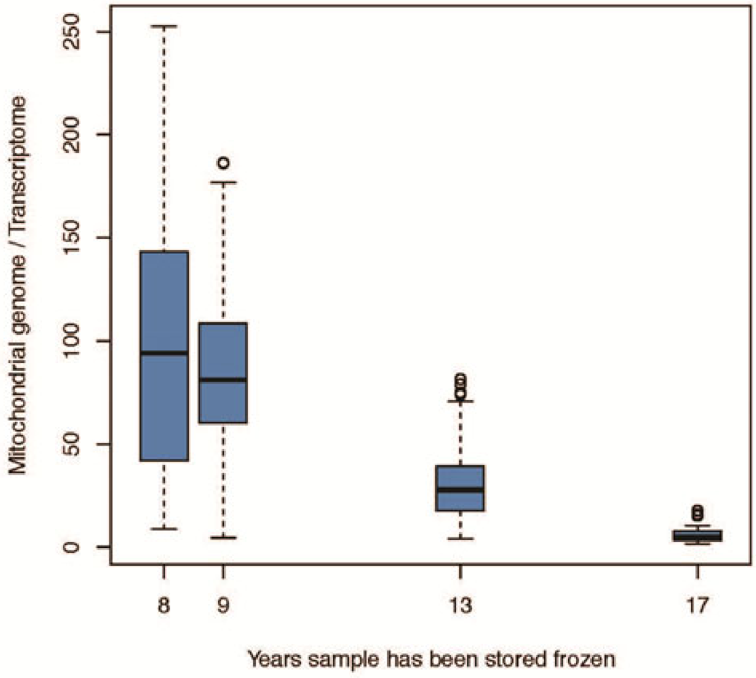
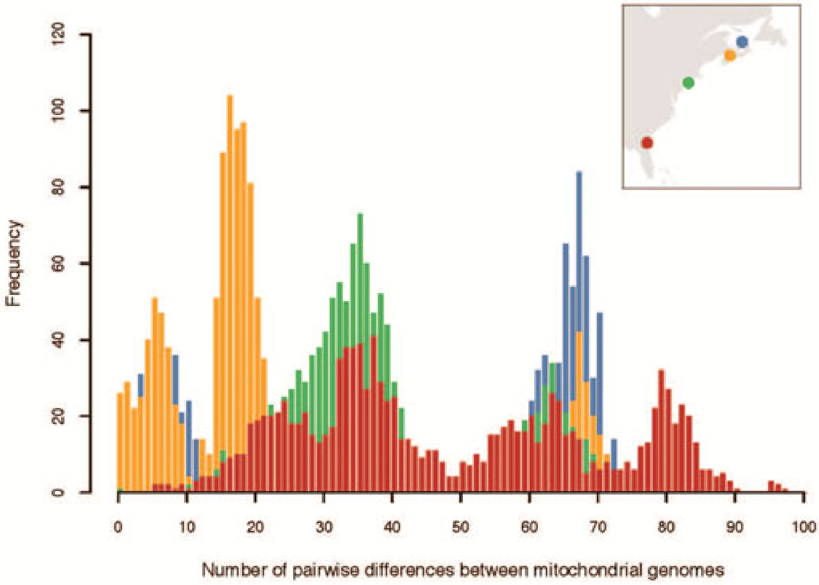


Fig. 8. Mismatch distributions showing the number of pairwise nucleotide differences between the full mitochondrial genome sequences of silversides samples from four different locations along North America's east coast (the color-coded sampling locations are shown on the inset map).


Frequency
Number of pairwise differences between mitochondrial genomes