Received: 23 June 2021 | Revised: 28 October 2021 | Accepted: 2 November 2021

DOI: 10.1111/1755-0998.13551

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

An accurate assignment test for extremely low-coverage whole-genome sequence data

Giada Ferrari[1] | Lane M. Atmore[1] | Sissel Jentoft[1] | Kjetill S. Jakobsen[1] | Daniel Makowiecki[2] | James H. Barrett[3,4] | Bastiaan Star[1]

[1]Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway

[2]Department of Environmental Archaeology and Human Paleoecology, Institute of Archaeology, Nicolaus Copernicus University, Torun, Poland

[3]McDonald Institute for Archaeological Research, Department of Archaeology, University of Cambridge, Cambridge, UK

[4]Department of Archaeology and Cultural History, NTNU University Museum, Trondheim, Norway

Correspondence
Giada Ferrari, Lane M. Atmore and Bastiaan Star, Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway.
Emails: giada.ferrari@ibv.uio.no; lane@palaeome.org; bastiaan.star@ibv.uio.no

Funding information
The Aqua Genome Project, Grant/Award Number: 221734/O30; Catching the Past, Grant/Award Number: 262777; Marie Skłodowska-Curie, Grant/Award Number: 813383

Abstract

Genomic assignment tests can provide important diagnostic biological characteristics, such as population of origin or ecotype. Yet, assignment tests often rely on moderate- to high-coverage sequence data that can be difficult to obtain for fields such as molecular ecology and ancient DNA. We have developed a novel approach that efficiently assigns biologically relevant information (i.e., population identity or structural variants such as inversions) in extremely low-coverage sequence data. First, we generate databases from existing reference data using a subset of diagnostic single nucleotide polymorphisms (SNPs) associated with a biological characteristic. Low-coverage alignment files are subsequently compared to these databases to ascertain allelic state, yielding a joint probability for each association. To assess the efficacy of this approach, we assigned haplotypes and population identity in *Heliconius* butterflies, Atlantic herring, and Atlantic cod using chromosomal inversion sites and whole-genome data. We scored both modern and ancient specimens, including the first whole-genome sequence data recovered from ancient Atlantic herring bones. The method accurately assigns biological characteristics, including population membership, using extremely low-coverage data (as low as 0.0001x) based on genome-wide SNPs. This approach will therefore increase the number of samples in evolutionary, ecological and archaeological research for which relevant biological information can be obtained.

KEYWORDS
chromosomal inversion, ecotype, genome skimming, haplotype, population assignment

1 | INTRODUCTION

Continuous advances in sequencing technology have resulted in an exponential increase in coverage and genomic resources for many species (Reuter et al., 2015; Wetterstrand, 2021). Nonetheless, although high-coverage whole-genome data allow a plethora of detailed bioinformatic analyses, low-coverage data remains common in fields that cannot always generate large amounts of high-coverage sequences, such as molecular ecology and ancient DNA (Bohmann et al., 2020; Malé et al., 2014; Peterson et al., 2012; Ripma et al.,

Giada Ferrari and Lane M. Atmore contributed equally.

2014; Suchan et al., 2016). Such low coverage data may arise for various reasons. First, whole genome population-level analyses in ecological research may remain cost-prohibitive, leading researchers to turn to techniques such as reduced-representation sequencing (e.g., Dodsworth, 2015; Marcus, 2021; Nevill et al., 2020; Zeng et al., 2018). Second, recent strategies targeting organelle (mitochondrial or chloroplast) reference databases increasingly use genome skimming sequencing approaches (Bohmann et al., 2020). Finally, DNA preservation in subfossil, archaeological, historical, or degraded biological material remains variable and is often context-specific (Ferrari et al., 2021; Keighley et al., 2021; Tin et al., 2014). In order to account for such unpredictability, ancient DNA sequencing studies typically screen many specimens, from which a subset with the best DNA preservation is selected for deeper sequencing (e.g., Martínez-García et al., 2021; Star et al., 2018; van der Valk et al., 2021). These practices result in a proliferation of specimens for which (extremely) sparse genome-wide data is obtained.

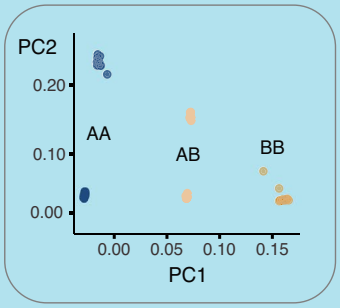
Low-coverage sequence data are difficult to jointly analyse with specimens that have obtained higher coverage without introducing various types of statistical bias (e.g., François & Jay, 2020; Lee et al.,2010; Patterson et al., 2006; Skoglund et al., 2014). Researchers who use low-coverage sequence data are therefore restricted in the analyses they can conduct and are often forced to discard specimens that do not yield sufficient coverage. For any investigation within the field of molecular ecology, the discarding of specimens due to low coverage results in wasted resources, and a reduced number of specimens for downstream analyses. Moreover, for ancient DNA studies this leads to the destruction of limited and unique zooarchaeological material for which no meaningful information is obtained. Efforts to obtain as much relevant information as possible from such specimens are therefore particularly important, from a biological and an ethical perspective (Pálsdóttir et al., 2019).

Genomic assignment tests can provide important biological information in ecological research and in several cases, low-coverage data have been effectively used for the determination of a range of basic biological characteristics (e.g., Grossen et al., 2018). For instance, the genetic sex of ancient mammals can easily be assigned from sparse sequencing data due to its association with extensive genomic differentiation on a chromosomal scale. Sexing has been applied to ancient low-coverage sequences to infer burial practices (Fages et al., 2020; Nistelberger et al., 2019), the impact of historic hunting (Barrett et al., 2020), and the behaviour of extinct species (Pečnerová et al., 2017). Aside from sex determination, other relevant biological characteristics may also be associated with large-scale genomic differentiation. In particular, structural variants (e.g., chromosomal inversions, haploblocks, or supergenes) have been increasingly identified as major drivers of evolutionary and ecological processes (Wellenreuther & Bernatchez, 2018), playing important roles in population structure and evolution. For instance, inversions are involved in the evolution of sex chromosomes (Hughes et al., 2010; Lemaitre et al., 2009) and speciation (Noor et al., 2001), and are critical for within-species adaptation to local environments (Ayala et al., 2013; Barth et al., 2017; Berg et al., 2016; Jones et al., 2012; Leitwein et al., 2017; Lowry & Willis, 2010; Morales et al., 2019; Nadeau et al., 2016; Pettersson et al., 2019; Todesco et al., 2020; Twyford & Friedman, 2015). Chromosomal inversions can affect megabase-sized genomic regions (e.g., Berg et al., 2017; Fang et al., 2012; Twyford & Friedman, 2015), and are often characterized by high levels of linkage disequilibrium (LD; Hoffmann & Rieseberg, 2008) due to inhibited recombination between noncollinear inversion haplotypes. Genotyping of such haplotypes using a subset of segregating genetic markers is feasible using whole genome sequencing data (Donnelly et al., 2010; Salm et al., 2012). Therefore, low-coverage data can retrieve relevant biological characteristics, potentially yielding useful insights on population continuity, species migration and distributions, hunting, historic trade, and burial practices, depending on archaeological context or ecological setting.

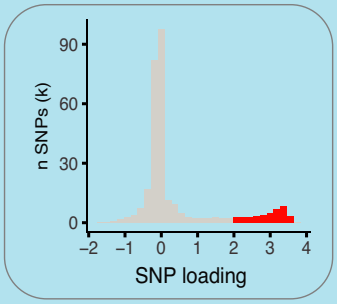
Several methods have been developed for assigning inversion haplotypes in order to facilitate GWAS analysis for SNPs within inversions in the human genome (*scoreInvHap*, Ruiz-Arenas et al., 2019; *pfido*, Salm et al., 2012; *InvClust*, Cáceres & González, 2015; *inveRsion*, Cáceres et al., 2012; and methods proposed by Bansal et al., 2007; Sindi & Raphael, 2010). These methods rely on LD break-points and structural variation (e.g., *InvClust*, *inveRsion*, and *scoreInvHap*, as well as the methods proposed by Bansal et al. and Sindi et al.) or haplotype tagging (*inveRsion*) to identify inversion sites and then conduct various types of SNP calling within those sites. All of these methods are specifically developed for identifying inversions in human genomes (e.g., Ma & Amos, 2012) and their use in disease- and other phenotype-association studies (Ruiz-Arenas et al., 2019; Salm et al., 2012). They have not been tested with sparse genomic data and are specific to use with inversions; indeed, *pfido* was designed for just one inversion in the human genome (Salm et al., 2012). Because of their reliance on signatures of structural variation, they cannot be applied to other types of variation, such as genome-wide population differentiation. There is currently no approach specifically designed to classify extremely low-coverage data with a broad applicability to score different types of large-scale genomic differentiation in a range of species.

FIGURE 1 The BAMscorer pipeline. The BAMscorer pipeline has two main modules—reference database creation and alignment file scoring. (a) Sequence data must be pre-processed and input into the pipeline as a VCF file. smartPCA (Patterson et al., 2006; Price et al., 2006) is used to generate eigenvalues and SNP loading weights, which are then used to assign population groups or inversion types in the reference database and create a database of highly-divergent loci in a given region of interest. (b) These positions are called from the alignment files to be scored. The positions are then compared to the database for allelic similarity. The likelihood of a given allele at a locus belonging to a haplotype is coded as the frequency of that allele at the locus in each database. AB allele frequencies are calculated as the average of frequencies present in AA and BB haplotypes. A joint probability is estimated for each alignment file belonging to each of the three haplotypes (for genome-wide assignment only AA and BB are used) and these values are scaled to one, outputting a probability index of genomic assignment for each individual

(a)



VCF



Reference SNP Database Created

		Major Alleles				Minor Alleles			
Chrom	Position	AA Genotype	AF AA	BB Genotype	AF BB	AA Genotype	AF AA	BB Genotype	AF BB
1	568	A	0.95	G	0.98	C	0.05	A	0.02
1	1289	T	0.85	A	0.97	A	0.15	C	0.03
1	2350	T	0.94	G	0.99	G	0.06	T	0.01
1	4560	G	0.99	C	0.91	C	0.01	G	0.09

(b)

Reference SNP positions read from BAM

Chrom	Position	Reads	Consensus
1	568	A,A	A
1	1289	A	A
1	2350	G	G
1	4560	C,G,C	C

BAM scored against Reference SNP Database

Chrom	Position	Genotype	Database Match	Likelihood AA	Likelihood BB	Likelihood AB
1	568	A	AA Major, BB Minor	0.95	0.02	(0.95 + 0.02) / 2
1	1289	A	AA Minor, BB Major	0.15	0.97	(0.15 + 0.97) / 2
1	2350	T	AA Major, BB Minor	0.94	0.01	(0.94 + 0.01) / 2
1	4560	G	AA Major, BB Minor	0.99	0.09	(0.99 + 0.09) / 2

Joint Probability Estimation

$$p_{i,g} = \prod_{l=1}^{n} f_l$$

Scaled Probability Scores

Individual	AA	BB	AB	SNPs
Ind_1	1.0	0.0	0.0	450
Ind_2	0.0	0.0	1.0	315
Ind_3	1.0	0.0	0.0	678
Ind_4	0.1	0.9	0.0	589

Here, we developed a new method that allows efficient assignment of different biological characteristics using extremely low-coverage sequence data. First, a database is created that contains the allele frequency association of individual SNP loci with a specific biological characteristic (e.g., an inversion type or population membership). These databases are based on moderate- to high-coverage sequences of a subset of specimens (Figure 1a). Second, sequence alignment data of (ancient) specimens are compared to this database and a joint probability (e.g., see Star et al., 2017) is calculated based on the binomial distribution of their frequency association (Figure 1b). This two-step approach is analogous to *scoreInvHap* (Ruiz-Arenas et al., 2019), yet there are some key differences. Importantly, in contrast to earlier approaches, this probability calculation does not make any assumptions regarding specific signatures of structural variation and can therefore be applied to different types of genetic differentiation. For instance, our approach includes differentiation between inversion haplotypes or genome-wide differences associated with ecotype or population structure. Our program depends solely on freely available, commonly used software and file formats, and is freely available for download at: https://github.com/laneatmore/BAMscorer.

We investigated the efficiency of our approach in assigning haplotypes for three chromosomal inversions in species that differ in their availability of reference specimens (P3 on Chr15, *Heliconius numata*, $n = 20$; Chr12, *Clupea harengus*, $n = 19$; and LG01, *Gadus morhua*, $n = 276$). These inversions display clinal distributions that are associated with biological characters such as wing pattern phenotypes (Joron et al., 2006, 2011; Nadeau, 2016), adaptation to water temperature and salinity (Pettersson et al., 2019), and migratory behaviour (Berg et al., 2016). Finally, we investigated the accuracy of this approach for the genome-wide population assignment of Atlantic cod specimens (Barth et al., 2019; Pinsky et al., 2021). To assess the program, we first built reference databases for each species and then used these databases to identify biological characteristics in nonreference alignment files. We used both ancient and modern sequences for scoring, including the first ancient whole-genome sequences recovered from Atlantic herring bones.

2 | MATERIALS AND METHODS

2.1 | Reference and scoring databases

For each species investigated, two different data sets were used. The first data set was used to create the reference SNPs database (hereafter referred to as the 'reference database') and the second data set (hereafter referred to as the 'scoring database'), containing individuals not found in the reference database that were scored utilizing the BAMscorer program. Reference databases were processed and filtered as described below to be input to the BAMscorer program as VCF files. Scoring databases were aligned to appropriate reference genomes and left as unfiltered BAM files for scoring. All the data used in this manuscript—including the newly generated archaeological Atlantic herring data—are publicly available. Below we describe each data set in terms of composition, sample size and biological characteristics.

2.1.1 | Heliconius butterflies

Heliconius numata butterflies are known to exhibit distinct wing-pattern morphs that are associated with different genomic haplotypes (Joron et al., 2006). These wing-pattern morphs are generally thought to be mimicry adaptations to predation and signalling between butterflies, therefore are probably under strong selective pressures (Chouteau et al., 2017; Joron et al., 2011). Research has suggested a supergene on chromosome 15 that determines the wing pattern morph of a particular butterfly (Joron et al., 2006, 2011). There are three inversion sites within the *P* supergene—P1, P2, and P3—which are associated with *Heliconius* wing-pattern morphs (Jay et al., 2018; Joron et al., 2011). We chose to focus on P3, which can be used to discriminate between the major types of wing-pattern morphs (Jay et al., 2021). Two databases were obtained for *Heliconius* butterflies, one as a reference database and one as a scoring database.

The reference database was obtained from Nadeau et al. (2016), which contains 20 individual genomes from various *H. numata* subspecies. This reference database encompasses several different wing morphs, which are associated with inversions at the *P* supergene (Chouteau et al., 2017; Jay et al., 2021). The scoring database consisted of 40 individual genomes obtained from Jay et al. (2021). This database, which has no overlap with the reference database from Nadeau et al. (2016), contains several different *H. numata* subspecies and encompasses various wing pattern morphs, therefore different P3 inversion types. The reference database was well-balanced between the three possible inversion types, with seven individuals belonging to types AA and BB and six individuals belonging to the AB heterozygous type.

Both the reference and scoring databases were aligned to the *Heliconius melpomene* Hmel2.5 reference assembly (http://ensembl.lepbase.org) using PALEOMIX v.1.2.13 (Schubert et al., 2014) with BWA-mem. The ~1.1 Mb P3 inversion is found on reference scaffold Hmel215003o (between 2,000,001 and 3,100,000 bp). Genotypes for the reference database were called using the GATK4 pipeline (Van der Auwera & O'Connor, 2020) and the following filtering parameters: FS<60.0 && SOR<4 && MQ>30.0 && QD >2.0 && INFO/DP<5500, SnpGap 10, minGQ 15 minDP 3, maf 0.001, with indels removed and biallelic variants selected. The reference database, therefore, is contained in a single VCF file, whereas the scoring database is a collection of unfiltered alignment files.

2.1.2 | Atlantic herring

Atlantic herring show a high degree of genomic structure that is probably associated with environmental characteristics such as

salinity and sea surface temperature and behaviours, such as spawning season (Lamichhaney et al., 2017; Martinez Barrio et al., 2016). Several inversion sites have been identified that are thought to be linked with some of these adaptations (Han et al., 2020; Pettersson et al., 2019). These sites show structure between various populations of Atlantic herring as well as between the true Atlantic herring (*Clupea harengus*) and the Baltic herring (*Clupea harengus membras*), a slightly smaller subspecies living in the Baltic Sea (Han et al., 2020).

For Atlantic herring, an ~8Mb inversion on chromosome 12 was investigated, which may be associated with a 'supergene' denoting different Atlantic herring ecotypes associated with salinity (Han et al., 2020; Pettersson et al., 2019). The inversion is located at chr12:17,900,000–25,600,000 bp and is linked to salinity adaptation for autumn-spawning herring populations (Han et al., 2020). A modern reference database was obtained from Han et al. (2020), which consisted of 20 individual genome sequences. These sequences encompassed all major populations of Atlantic and Baltic herring identified by Pettersson et al. (2019) and Han et al. (2020), with the exception of the spring-spawning Baltic herring. There were 6 Baltic individuals and 14 Atlantic individuals in the reference database.

To assess the applicability of BAMscorer for conducting assignment tests on ancient genomes, the scoring database was created from nine newly sequenced ancient Atlantic herring genomes (see following section for DNA extraction and sequencing methodology). Both modern and ancient herring reads were aligned to the Atlantic herring reference genome (GCA_900700415.1, Pettersson et al., 2019). The modern reads were aligned as described above for the *Heliconius* butterflies. Ancient herring reads were aligned as described in Ferrari et al. (2021), using BWA-aln, which is commonly held to be the most appropriate mode for aligning ancient genome sequences (Schubert et al., 2012).

Genotypes for the reference database were called and filtered following the same protocol as for the *Heliconius*, while the scoring database was not processed further. Two individual outliers in the reference database were observed and subsequently checked for relatedness using KING (Manichaikul et al., 2010; Note S1). These individuals appeared to be duplicates and were removed from the database to ensure accuracy of metadata.

In the scoring database of ancient individuals, most specimens have excellent DNA preservation (Table S1) and all show the typical fragmentation and misincorporation patterns of authentic ancient DNA data (Jónsson et al., 2013; Figure S1). The aligned sequences were down-sampled to 100,000 reads and are now available on ENA (accession number PRJEB45393). Similar as for *Heliconius*, the reference database for Atlantic herring thus consisted of a single VCF file and the scoring database of a collection of alignment files.

2.1.3 | Atlantic cod

The Atlantic cod reference database was created using 276 Atlantic cod individuals representing most major geographical locations (western Atlantic, eastern Atlantic, and Baltic Sea) in the species' range (Barth et al., 2019; Pinsky et al., 2021). In Atlantic cod, four large chromosomal inversions have been identified that are associated with differences in migratory behaviour and environmental adaptations (Berg et al., 2016; Kirubakaran et al., 2016; Sodeland et al., 2016). These chromosomal inversions originated independently, between 0.40 and 1.66 million years ago (Matschiner et al., 2021). A data set of 15 unrelated archaeological specimens were obtained from Star et al. (2017) to use for the scoring. Modern and ancient reads were aligned to the gadMor2 reference genome (Star et al., 2011; Tørresen et al., 2017) as above for *Heliconius* and Atlantic herring. SNP calling and filtering for the reference was performed as described in Barth et al. (2019) using the GATK haplotype caller v.3.4.46 (McKenna et al., 2010), bcftools v.1.3 (Li, 2011), VCFtools v.0.1.14 (Danecek et al., 2011), again mirroring methods used for Heliconius and Atlantic herring. For Atlantic cod, we investigated both an inversion locus, as well as genome-wide patterns of divergence. First, we targeted an ~16 Mb double chromosomal inversion on LG01 which is associated with differences in migratory behaviour (Berg et al., 2016; Kirubakaran et al., 2016; Sodeland et al., 2016). This inversion is located at LG01:9,100,000–26,200,000 bp, and the reference database contained 217 nonmigratory and 30 migratory specimens. Second, we analysed genome-wide data separating 24 western Atlantic from 252 eastern Atlantic cod specimens and genome-wide data separating 23 Baltic from 229 eastern Atlantic cod specimens. For the whole genome analyses, we excluded the location of four major inversions (on LG01, LG02, LG07, and LG12, as described in Berg et al., 2016, 2017) following the coordinates used in Star et al. (2017).

2.2 | Ancient DNA extraction and sequencing

Nine Atlantic herring bones from two Polish sites, dated between the 9th and 15th century CE (Domagała & Franczuk, 1992; Iwaszkiewicz, 1991; Makowiecki, 2003; Makowiecki et al., 2016; Table S1), were UV-treated for 10 min per side and cleaned with ultra-pure water. DNA was extracted including a predigestion step, following Damgaard et al. (2015). Then, 10–40 mg of bone were pulverized with micro pestles in the digestion buffer (1 ml 0.5 M EDTA, 0.5 mg/ml proteinase K, and 0.5% N-Lauryl sarcosine). Following overnight digestion, DNA was extracted with nine volumes of a 3:2 mixture of QG buffer (QIAGEN) and isopropanol. MinElute purification was carried out using the QIAvac 24 Plus vacuum manifold system (Qiagen) in a final elution volume of 65 µl. Parallel nontemplate controls were included. Single-indexed blunt-end sequencing libraries were built from 16 µl of DNA extract or nontemplate extraction blank, following the single-tube (BEST) protocol (Carøe et al., 2018) with the modifications described in Mak et al. (2017). All laboratory protocols up to indexing of sequencing libraries were carried out in a dedicated ancient DNA clean laboratory at the University of Oslo following standard anti-contamination and authentication protocols (Cooper & Poinar, 2000; Gilbert et al., 2005; Llamas et al., 2017). Library quality and concentration were inspected with a

High Sensitivity DNA Assay on the Bioanalyser 2100 (Agilent) and sequenced on an Illumina HiSeq 2500 platform at the Norwegian Sequencing Centre with paired-end 125 bp reads, demultiplexed allowing zero mismatches in the index tag.

2.3 | The BAMscorer Pipeline

2.3.1 | Module 1: creation of SNP reference databases

The initial step of the BAMscorer pipeline is to create a reference database of divergent SNPs associated with each haplotype or population in a set of focal individuals (Figure 1a). These divergent SNPs are referred to as belonging to 'AA', 'BB', or 'AB' genotypes (or groups). The BAMscorer program does not conduct SNP calling, but takes a preprepared VCF file as input. SNP calling methods and filtering parameters are therefore at the discretion of the user and can be done using a reference genome as we have done with our species or with de novo SNP calling techniques as is often used for reduced-representation sequencing. As long as the reference data is input to BAMscorer as a VCF file, a reference database can be created.

Reference SNP databases are created as follows:

1. The VCF file is first prepared with VCFtools v.0.1.16 (Danecek et al., 2011) and PLINK v.1.9 (Purcell et al., 2007), selecting only those regions of interest (i.e., where inversions are located, or genome-wide).
2. A Principal component analysis (PCA) is run as implemented in smartPCA (EIGENSOFT v.7.2.1; Patterson et al., 2006; Price et al., 2006) to calculate axes of differentiation and individual SNP loadings between homozygote inversion haplotypes or populations. As a default, the BAMscorer pipeline selects diagnostic loci in the top and bottom 5% of the SNP loading distribution, although the optimal SNP loading cutoff value should be determined by the user. Visualization of the SNP loading profile can help decide such cutoffs (see further below). The BAMscorer program is capable of filtering SNPs in the reference database based on both symmetrical and asymmetrical distribution cutoffs.
3. SNPs that pass cut-off filters form the divergent SNPs database for each haplotype or population. To assist the user in the selection of individuals to represent each haplotype, heterozygosity is calculated per individual based on SNPs in the divergent database.
4. Individuals from the reference database VCF file are scored for PC1 and heterozygosity values, and manually classified into types: when whole genome data are investigated, individuals are separated into groups called AA and BB; when inversions are investigated, individuals are separated into three clusters, representing genotypes that comprise homozygous AA and BB and heterozygous AB haplotypes. Inversion genotypes are known to fall into specific clusters in PCA analysis (see Figure 1a), which allow for easy identification using separation on PC1 and assessing levels of individual heterozygosity.
5. For individuals in the homozygote AA and BB haplotypes/groups, allele frequencies of the divergent SNPs are calculated. Two databases are created, containing the allelic state (i.e., A, C, G, T) and allele frequencies of the major (first database) and minor (second database) alleles in the AA and BB haplotypes. Databases containing few individuals often contain fixed alleles due to limited sampling. The uncertainty associated with sampling fixed alleles is addressed in the BAMscorer program by calculating a minimum expected frequency of $(1/((2*N)+1))$ for the minor allele, where N is the number of individuals in the reference database for fixed alleles in the region of interest. When scoring inversions, the sample probability of obtaining alleles from heterozygous AB genotypes are calculated by averaging the observed allele frequencies in the AA and BB haplotypes. We note that due to the nature of inversions, it is highly likely that a heterozygous genotype will fall directly in between the homozygous genotypes (see Figure 2). This may not be the case for all genomic regions and is certainly not always the case for whole-genome variation. We therefore recommend running BAMscorer with the—wg flag for genomic regions that do not follow such a pattern and/or whole-genome analysis, as this will provide an assignment of either type AA or type BB without attempting to estimate the average allele frequencies between the two.

Once optimal database parameters have been identified (a full list of parameters can be found at https://github.com/laneatmore/BAMscorer), the SNP database can be reused for BAM scoring on many different data sets of the same species.

2.3.2 | Module 2: BAM scoring

The reference database can then be used to assess the scoring database (Figure 1b). The scoring database consists of unfiltered alignment files in BAM format. BAM files can be generated at the user's discretion as long as the coordinate system matches that of the reference database. Rather than conducting SNP calling on the scoring database, BAMscorer takes the position of each SNP in the reference database and pulls these loci from each BAM file in the scoring database. A consensus read for each locus is determined via a simplified genotype calling process, and then the alleles in the alignment files are compared to the reference database to determine the probability of a given individual belonging to each genotype or genetic cluster.

A detailed overview of the BAM scoring process is as follows:

1. The divergent SNPs databases are used to score alignment files (BAM format) for a given set of (low-coverage) individuals. For each locus in the divergent SNPs database, matching reads are pulled from the BAM file using the python module pysam (https://github.com/pysam-developers/pysam). The 'consensus' read is determined based on the most highly-represented allele in all reads for each position. In the event that there are equal numbers of reads for multiple alleles at a given locus,

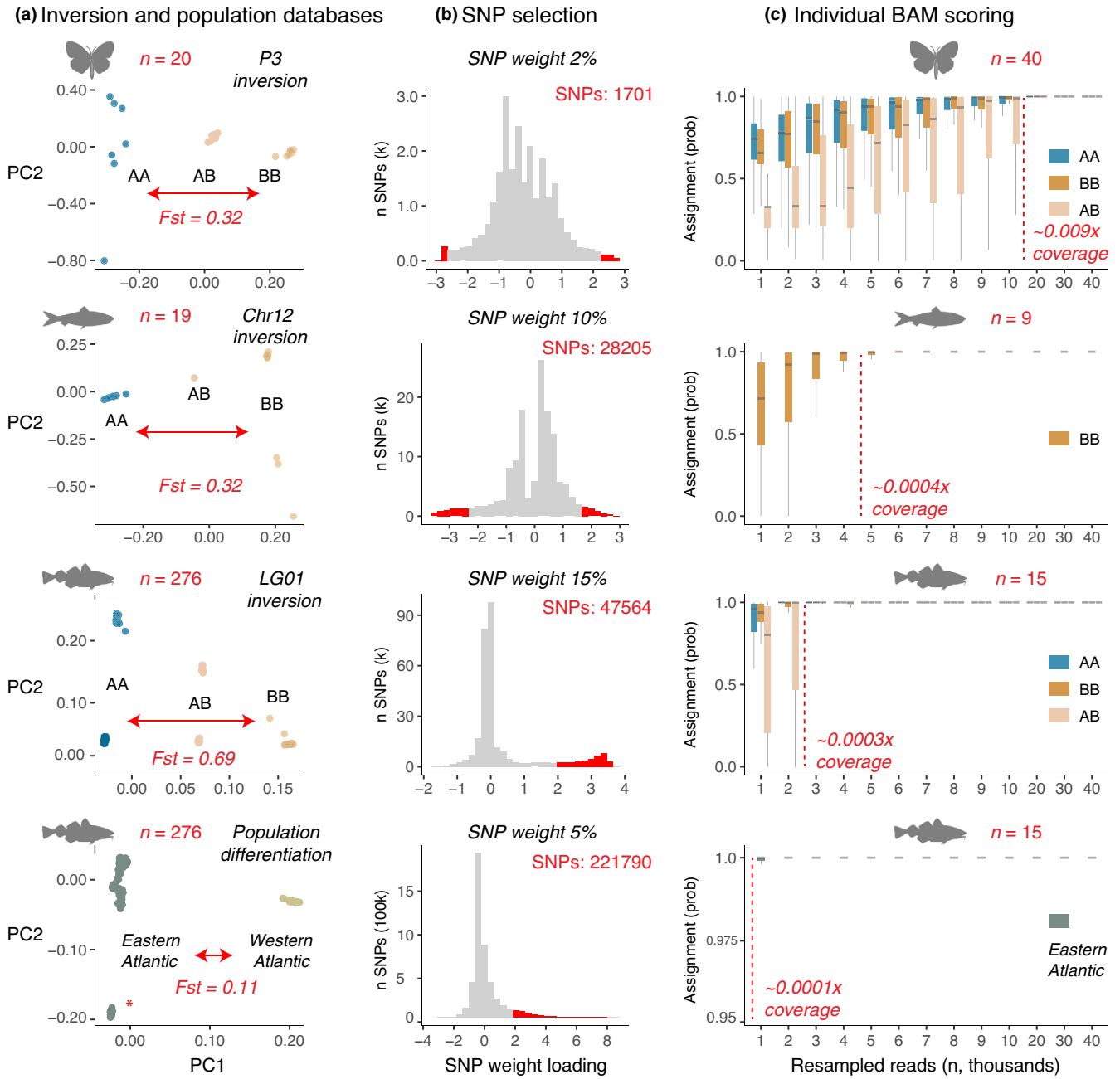


FIGURE 2 Inversion and population assignment for *H. numata* (P3 inversion), *C. harengus* (Chr12 inversion) and *G. morhua* (LG01 inversion, population differentiation) using extremely low-coverage data. (a) Inversion and population PCA plots generated for the three species (silhouettes) using smartPCA (Patterson et al., 2006; Price et al., 2006). Haplotype clusters (AA; blue, AB; beige, BB; sepia) and main population clusters (eastern Atlantic; dark green, western Atlantic; light green) are named. The number of individuals (red) and the weighted F_{ST} differentiation between inversion-loci (between AA and BB) or genome-wide (red arrow) is indicated. The Baltic population is indicated by a red '*' amongst the eastern Atlantic populations. (b) SNPs most associated with either inversion (A or B) haplotype or large-scale population differentiation (western or eastern Atlantic) are selected based on their SNP weight loading distribution along PC1. Those with lowest and highest loadings are most associated with differentiation along PC1. SNP weight indicates the percentage of SNPs selected from the most extreme end(s) of the distribution (red). (c) Assignment probability for individual specimens generated by downsampling BAM files 1000–40,000 reads. At each interval, and for each individual, the downsampling is iterated 20 times in order to generate box plots. Probabilities are calculated based on the joint binomial distribution of observing divergent SNPs associated with either haplotype group or population. Also indicated is the number of individuals scored (red, note these are not the same individuals used to create the original databases) and fold coverage (red dotted line, *x* coverage) at which more than 0.99 median assignment probability is obtained

one allele is then chosen at random, although these instances were extremely rare in the low-coverage data analysed here. This process provides a subset of observed alleles at divergent loci in each inversion or population for each individual BAM file. Using pysam for genotype calling directly from BAM files is the most accurate method when dealing with low-coverage data (Ros-Freixedes et al., 2018).

2. The probability of observing a variant associated with a specific haplotype is calculated using the allele frequencies of matching positions in the reference databases. For example, if the position in the BAM file matches the dominant allele in haplotype group AA, the probability for that locus belonging to the AA genotype is coded as the allele frequency of the dominant allele in haplotype group AA. If the allele also matches the major or minor allele in haplotype group BB, the probability for that locus belonging to the BB genotype is coded as the allele frequency of that allele in the BB reference set. This allows a proxy calculation for heterozygous sites in the BAM file without requiring the extensive computational requirements it would take to determine genotype likelihood directly for each position. Both the dominant and minor allele frequencies for each genotype in the reference database for alleles in homozygote AA and BB haplotypes are used, thereby providing a likelihood estimation that the consensus read pulled from the BAM file is any one of the following: AA dominant, AA minor, BB dominant, BB minor. For inversions, three probabilities are recorded for each position—the frequency of that allele in haplotype groups AA, BB, and AB (only AA and BB for genome-wide analysis).
3. Joint probabilities of all observed alleles belonging to a particular haplotype group or population are calculated for each individual using the following equation:

$$p_{i,g} = \prod_{l=1}^{n} f_l$$

Whereby the probability (p) of the scored individual (i) and genotype (g) is the product of allele frequencies (f) of the number (n) of observed SNP loci (l) in each database. Finally, the joint probability scores for all genotypes are scaled to one to provide a final probability estimate of an individual belonging to a certain haplotype or population. We also provide the number of SNPs in the reference database that were recovered from each individual BAM file to inform on how well scored a specific individual is.

2.4 | Analyses

We ran the above pipeline on each of the four databases outlined above: *Heliconius* P3 inversion, Atlantic herring chr12 inversion, Atlantic cod LG01 inversion, and the whole-genome data set for Atlantic cod (in two different scenarios). For each of these data sets, we also tested program parameters to assess the impact of noise and filtering in the reference database on BAMscorer accuracy. We created separate reference databases for each inversion using SNP loading cutoff values between 1% and 25% and an additional set of reference databases for the genome-wide analysis of Atlantic cod with SNP loading cutoff weights between 1% and 5%. We further assessed the impact of SNP filtering in reference database creation by limiting analysis to asymmetrical tails of the SNP loading distribution (e.g., taking only SNPs in the top 5% of loading weights).

2.5 | Assessing scoring certainty

To investigate the sensitivity of the BAMscorer pipeline, we downsampled each BAM file in the five databases (P3 from *Heliconius*, Jay et al., 2021; chr12 from Atlantic herring; and LG01 and two whole-genome population comparisons from Atlantic cod, Star et al., 2017). Following an approach described in Nistelberger et al. (2019), BAM files containing whole-genome shotgun data were downsampled to contain between 1000 and 40,000 reads (in most instances this is a mere fraction of the available data). At each read interval, and for each individual, the downsampling was randomly iterated 20 times. We compared accuracy of the scoring results of the extremely downsampled *Heliconius* data using BAMscorer by comparing these results to a separate PCA analysis using all data for the individuals in both databases. All of the individuals from the scoring database clustered with one of the three inversions in the reference database (Figure S2) and this clustering fully agreed with the assessment using BAMscorer. For Atlantic herring and Atlantic cod, accuracy of results was confirmed by prior knowledge of the inversion types or geographic origin of specimens. We assume that the ancient Polish herring have a Baltic origin given the archaeological context and age, and should therefore more closely match with the chr12 BB haplotype group, which is associated with the Baltic individuals in our reference database.

3 | RESULTS

We investigated three chromosomal inversions and one genome-wide analysis using BAMscorer. The *Heliconius* P3 inversion is the smallest (1.1 Mb) inversion, followed by the Atlantic herring Chr12 (8 Mb) and Atlantic cod LG01 inversion (16 Mb, Table 1). Principal component analysis (PCA) as implemented through BAMscorer *select_snps* separates the three main inversion genotypes along PC1 for the *Heliconius* P3, Atlantic herring Chr12, and Atlantic cod LG01 databases (Figure 2a), reproducing earlier observations (Barth et al., 2019; Han et al., 2020; Nadeau et al., 2016; Pinsky et al., 2021). Similarly, the whole genome analysis separates western from eastern Atlantic cod specimens along PC1 (Figure 2a, Pinsky et al., 2021) and Baltic from other eastern Atlantic cod (Figure S3, Barth et al., 2019). For the data analysed here, BAMscorer *select_snps* typically runs within 15 min when using the test scoring database provided on a MacBook Pro (running MacOS Catalina with a 1.4 GHz Quad-Core

TABLE 1 Inversion and genome characteristics of *Heliconius*, Atlantic herring, and Atlantic cod. Each comparison differs in terms of size of inversion, overall genome size and relative size of inversion in regards to species-specific genome size, as well as in terms of the optimum number of divergent SNPs (see methods) and individuals used for the reference databases and scoring

Species	Inversion name, location	Inversion size (Mbp)	Genome size (Mbp)	Relative size (%)	Divergent SNPs (*n*)	Database individuals (*n*)	Scored individuals (*n*)
H. numata (Heliconius)	P3, Chr15	1.1	273	0.4	1701	20	40
C. harengus (Atlantic Herring)	Chr12	8	726	1.1	28,205	19	9
G. morhua (Atlantic cod)	LG01	16	643	2.5	47,564	276	15
G. morhua (Atlantic cod)	Whole genome	na	643	na	221,790	276	15

Intel Core i5 and 16GB RAM). The SNP weight loading distribution underlying genetic divergence between inversion haplotypes of populations is either approximately symmetrical (*Heliconius* and Atlantic herring) or asymmetrical (Atlantic cod, Figure 2b). SNP weights are proportional to the correlation (across samples) between each SNP and each PC (Patterson et al., 2006; Price et al., 2006). SNPs that are strongly associated with divergence will have the highest SNP weight loading values and are therefore biologically informative.

An important consideration of our approach therefore lies in the selection of loci based on their SNP loading distribution patterns. In order to maximize the probability of observing loci in low-coverage sequencing data, as many loci as possible should be included in the database. Yet, adding those loci that are not significantly associated with either inversion haplotype or specific population will add noise and uncertainty. We therefore tested the accuracy of our approach using a range of SNP loading filtering parameters. For inversions, databases were created using cutoff values between 1% and 25%, depending on the species under investigation. For our genome-wide analyses, we set the SNP loading cut-off weights between 1% and 5%. The default parameter in the BAMscorer pipeline is to take symmetrical portions from each side of the SNP loading distribution (the 5% cutoff value takes the top and bottom 5% of SNPs), yet we also noticed asymmetrical SNP loading distribution values. We therefore also investigated the effect of selecting SNPs from either the top or bottom of the SNP loading distribution.

For the *Heliconius* P3 inversion, the ability to confidently score heterozygous individuals (Jay et al., 2021) erodes with increasing SNP weight values (Figure 3), and the optimal cutoff to simultaneously score all possible genotypes lies at 2% and 1701 SNPs. For Atlantic herring Chr12, not all inversion types are observed in the ancient read data, yet no major increase in ability of scoring is obtained after a SNP weight of 10% and 28,205 SNPs (Figure S4). For Atlantic cod, best separation of ancient data (Star et al., 2017) was obtained by selecting SNPs from the single, most extreme end of the SNP weight loading distribution (Figure 2b, Figures S5–S7). For Atlantic cod LG01, SNP selection is similar to the *Heliconius* P3, in that the optimal cutoff is a trade-off in scoring homozygotes and heterozygotes, which for cod lies at 15% and 47,564 SNPs (Figure S6). Finally, best population separation for Atlantic cod using whole genome data is obtained at 5% and 221,790 SNPs for the trans-Atlantic separation (Figure S7), and 5% and 217,328 SNPs for the Baltic separation (Figure S3).

After deciding the best-possible cut-off values, several observations can be made regarding the scoring accuracy of BAMscorer *score_bams* depending on the number of reads for each of the comparisons. First, accurate scoring is obtained in extremely low-coverage data for all comparisons (Figure 2c). For *Heliconius*, accurate haplotype determination is obtained with 20,000 reads and 0.009x nuclear coverage. For all other comparisons, even less reads—by an order of magnitude—are required. Second, the scoring accuracy of heterozygous genotypes requires more reads compared to homozygous genotypes (see *Heliconius* P3 and Atlantic cod LG01, Figure 2c). Thus, different levels of accuracy are obtained depending on each sample's haplotype or population of origin. Third, an increase in scoring accuracy at lower numbers of reads is observed for those comparisons for which more SNPs could be obtained (Table1, Figure 2c). Best scoring accuracy is obtained for the population comparison of Atlantic cod, for which population of origin can be determined with 1000 reads or less than 0.0001x nuclear coverage (Figure 2c). The Baltic cod population (separated from other eastern Atlantic populations on PC2 indicated with red '*', Figure 2a), can be identified by iteratively investigating a smaller subset containing the eastern Atlantic specimens only (Figure S3). Finally, BAMscorer *score_bams* takes—on average—approximately 5 min to complete each comparison using a single Intel Xeon-Gold 6138 2.0 GHz CPU with 10 Gb of ram. It takes a similar amount of time to run the test scoring database on the MacBook Pro system as described above.

4 | DISCUSSION

The BAMscorer program allows genomic assignment on extremely low-coverage sequence data, thereby increasing the capacity for conducting population genomics analysis on sparse genome-wide data. Sequence data with low coverage is often discarded as there is little usable information that can be reliably recovered in comparative analyses with higher coverage specimens (e.g., François & Jay, 2020; Lee et al., 2010; Patterson et al., 2006; Skoglund et al., 2014). Applying our method will allow samples with sparse genome-wide data to be used. The method is, additionally, fast and can be applied to large quantities of data at one time, providing an efficient overview of the biological characteristics of a large data set. This approach will therefore expand the amount of information that can be gleaned from sparse genome-wide data (Bohmann et al., 2020), and

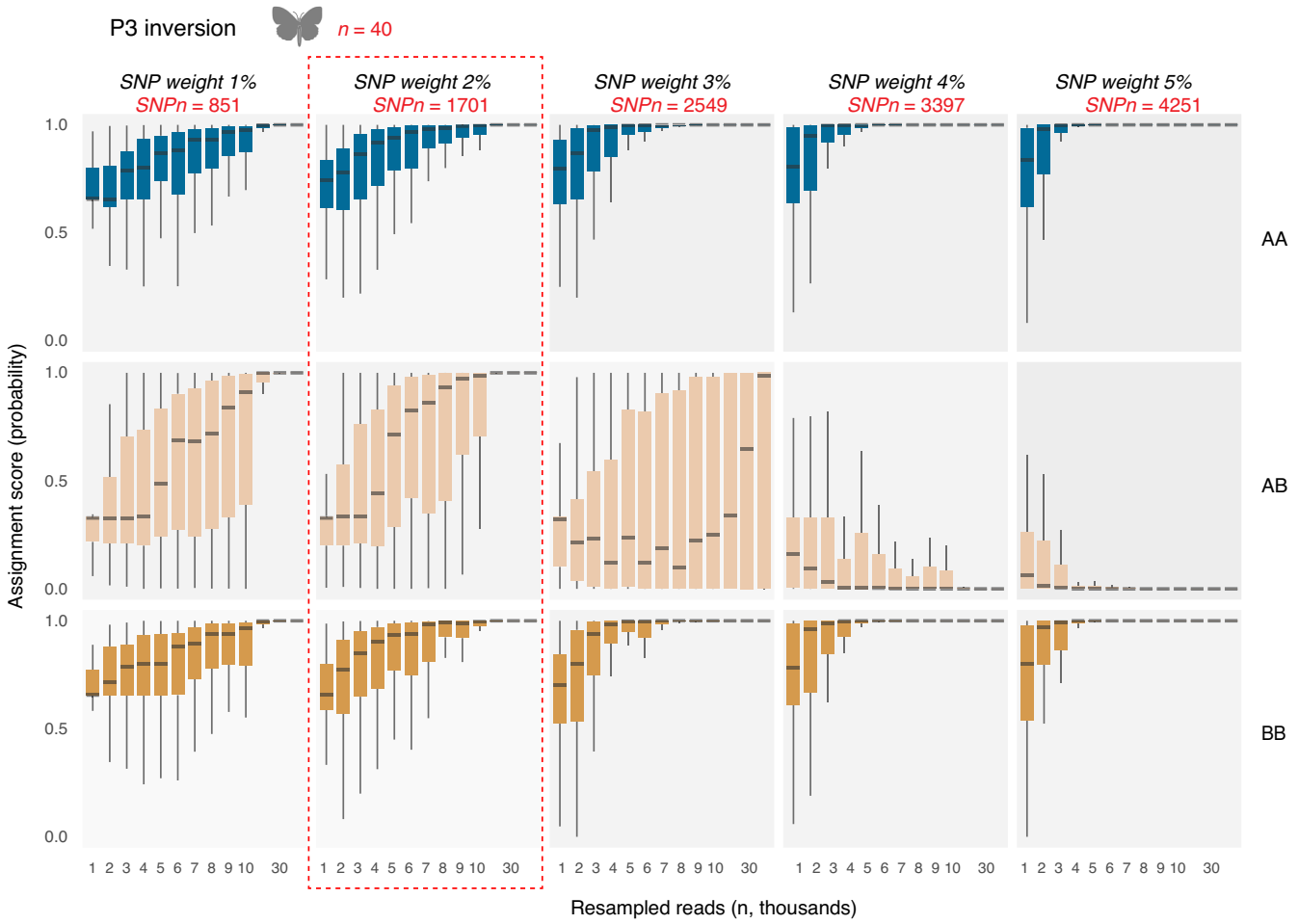


FIGURE 3 SNP selection by varying SNP weight in *H. numata* (P3 inversion). SNP weight is here defined as the percentage of SNPs with the most extreme values at both sides of the SNP loading distribution. Confidence in probability assignment is obtained by down-sampling BAM files 1000–40,000 reads. At each interval, and for each individual, the downsampling is iterated 20 times in order to generate box plots. Probabilities are calculated based on the joint binomial distribution of observing divergent SNPs associated with either haplotype group. Also indicated is the number of individuals (*n*, red) and number of SNPs (SNP*n*, red) and the chosen cutoff value (red dotted lines) at which all three haplotype groups can be efficiently separated

reduce sample dropout in the ancient DNA analyses pipeline where destructive sampling is wide-spread (Pálsdóttir et al., 2019).

We applied our method to three biological examples that have different levels of genomic differentiation. In *Heliconius* butterflies, inversions have not been found to be sympatric barriers to interspecific gene flow (Davey et al., 2017), and there is a high degree of interbreeding between the seven different *H. numata* wing pattern morphs (Chouteau et al., 2017). Within the wing pattern morph supergene on chromosome 15, there is incomplete genetic segregation between several of the *P* locus inversion types, including the P3 inversion site (Jay et al., 2021). This could suggest incomplete lineage sorting—a phenomenon known to be highly prevalent in mimetic species such as *H. numata* (Kozak et al., 2015; Savage & Mullen, 2009)—and/or introgression among *H. numata* morphs. Although BAMscorer exhibited less power in distinguishing this complex pattern of genomic divergence between *H. numata* morphs than between fish ecotypes, the approach still showed a high degree of efficiency. Within the *Heliconius* we were able to correctly identify all morphs without error at just 0.009x coverage, including heterozygous individuals, for which identification becomes significantly more challenging at low sequencing effort.

Both Atlantic herring and Atlantic cod exhibit temporally and geographically isolated spawning reproductive behaviour, although overall levels of genetic differentiation are relatively low (e.g., F_{ST} ~0.1 or less; Barth et al., 2017; Berg et al., 2016, 2017; Martinez Barrio et al., 2016; van Damme et al., 2009). Nonetheless, the divergence time of the LG01 inversion haplotypes of Atlantic cod is estimated to be around 600,000 years ago (Matschiner et al., 2021), driving divergence in many thousands of SNPs, which explains the success of BAMscorer at extremely low-coverage for the LG01 inversion. Similarly, strong selective pressures on Baltic herring have driven differentiation between these populations and their Atlantic conspecifics, as herring populations entering the Baltic Sea ~8000 years ago would have had to adapt to new, brackish conditions. Even though divergence time between Atlantic and Baltic herring cannot be older than 8000 years, herring exhibit distinct signals

of adaptation to low salinity conditions at several thousand loci (Han et al., 2020; Lamichhaney et al., 2012). Our results therefore indicate a high degree of accuracy in determining inversion types even for subspecies and haplotypes with a range of divergence times.

We obtain the highest power in scoring accuracy for the whole genome analyses of Atlantic cod. Both comparisons investigate populations that have diverged relatively recently: the western and eastern Atlantic populations around 65,000 years ago (Matschiner et al., 2021), and have genome-wide nuclear genetic divergence measured at F_{ST} ~0.11 (Pinsky et al., 2021). The Baltic was colonized between 8000 and 6000 years ago (Berg et al., 2015) and has genome-wide nuclear genetic divergence at F_{ST} ~0.04. Neither of these populations or areas exhibits fixed mitogenomic differentiation (Martínez-García et al., 2021). The high scoring accuracy that is obtained at a low number of reads in this whole-genome analysis—despite low overall genetic divergence—suggests a wide range of applicability in different biological settings.

Additionally, the accuracy of our results indicate that the method is robust to deamination damage, a common feature of ancient genomic sequences. By sampling ancient alignment files from different places in the genome, the scoring is robust to the noise created by deamination damage, which typically occurs at the ends of reads. Our ancient sequences for cod and herring exhibited up to 17% deamination damage (Figure S1), yet a high scoring accuracy was obtained despite the presence of such damage.

There are several practical considerations and limitations to take into account while using the BAMscorer program. First, each example provided here is associated with different levels of genomic divergence and size of genomic regions under investigation. We find that there are no optimal program settings that apply to all cases. Each of our three species required different filtering parameters, such as optimum SNP loading weight cut-off value, and required different minimum numbers of reads to obtain high scoring accuracy. Similarly, differences in these parameters were also required within species, such as when analysing the inversion on LG01 as compared to the genome-wide data in Atlantic cod. It is thus recommended that users explore the filtering parameters as we have done above to ascertain the appropriate parameters, as an understanding of the biological system in question is important for assessing the efficacy of BAMscorer.

Second, the whole-genome (population) application of BAMscorer is currently limited to assigning two clusters or populations simultaneously. We are in the process of developing a more generic approach that will allow scoring of an undefined number of populations and aim to make this available in future versions of the BAMscorer. The current version of BAMscorer, however, can be applied iteratively to sequentially score finer scales of genomic differentiation within data sets containing multiple clusters (see Figure S3). Moreover, BAMscorer is reliant on existing reference data to create the database from which alignment files are scored. This is a limiting factor in any assignment test, and probably unavoidable. However, we found that even with a relatively low number of reference genomes (our reference databases ranged from 19 to 276 individuals), we were still able to efficiently identify haplotypes in low-coverage data. The requirements for the reference databases are therefore not especially demanding in order for BAMscorer to be used efficiently.

Finally, an assignment test only evaluates the scenario as given by the user. It is therefore important to use BAMscorer with an understanding of the biological system in question and with these limitations in mind. We further recommend that users assess the impact of filtering parameters for creating the BAMscorer reference SNP database for each biological system. To provide an understanding of how much data BAMscorer is actually getting from each BAM file, we provide a read-out of the number of SNPs in the reference database that were read from the BAM file in question.

We have here introduced a novel software program that can be used to increase the information gleaned from extremely low-coverage sequence data. We have found that biological characteristics and genomic assignment can be recovered from sequences with as little as 1000 aligned reads (at ~0.0001x coverage in the case of Atlantic cod). The method is flexible and can be used on various types of genomic data. Because all SNP calling and alignment takes place prior to using the BAMscorer program, the program itself is not dependent on a reference genome and can therefore be used with SNPs calling generated de novo, as is often the case for reduced-representation sequencing. The program is further scalable for BAM files from 1000 to 50 M reads and can handle up to hundreds of thousands of SNPs without sacrificing computational efficiency.

We have shown that BAMscorer can differentiate between subspecies, populations, ecotypes, and genomic inversions and can successfully recover relevant biological information from extremely low-coverage data. As the underlying methodology is general in design, it can be applied to any low-quality samples and reduced representation sequence data such as ddRAD (Peterson et al., 2012) or hyRAD (Suchan et al., 2016), two common methods for cost-efficient sequencing used in ecology and evolution studies for modern and historic specimens. We expect that the capacity to quickly identify population of origin, determine between domestic and wild types (e.g., farmed vs. wild salmon; Glover et al., 2013, 2017), assess ecotype distribution, and to identify hybrids will be a useful additional tool in the fields of wildlife forensics and conservation genomics.

ACKNOWLEDGEMENTS

We thank A. Gondek-Wyrozemska for processing the ancient Atlantic herring specimens. We are grateful for the computational resources provided by the Saga computing cluster from Uninett-Sigma2 through allocations to the Centre for Ecological and Evolutionary Synthesis at the University of Oslo. We also thank M. Skage, S. Kollias, M. S. Hansen, and A. Tooming-Klunderud and the Norwegian Sequencing Centre for sequencing and processing of samples. The project benefited from early access to extensive genomic resources developed by the RCN-funded project 'The Aqua Genome Project' (221734/O30). Finally, this project received funding from RCN project 'Catching the Past' (262777) and the European

Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813383. The European Research Agency is not responsible for any use that may be made of the information it contains. We thank three reviewers whose comments and suggestions helped improve this manuscript. The first authors (G.F. and L.M.A.) contributed equally and reserve the right to place themselves first on their CV.

AUTHOR CONTRIBUTIONS

Giada Ferrari, Lane M. Atmore, and Bastiaan Star wrote the manuscript. Bastiaan Star conceived the project. Giada Ferrari, Lane M. Atmore, and Bastiaan Star developed the method and statistical framework. Lane M. Atmore wrote the code for the software program. Giada Ferrari, Lane M. Atmore, and Bastiaan Star conducted data analysis and visualization. James H. Barrett and Daniel Makowiecki provided archaeological material for sequencing. Sissel Jentoft and Kjetill S. Jakobsen provided early access to genomic sequence data. All authors have read and approved the manuscript.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://github.com/laneatmore/BAMscorer.

DATA AVAILABILITY STATEMENT

Reference data for all species have been publicly released earlier and are available from the European Nucleotide Archive (ENA) with the following accession numbers: *Heliconius*; PRJEB12740 and PRJEB40136, Atlantic herring; PRJNA642736, Atlantic cod; PRJEB29231 and PRJEB41431. The nine ancient Atlantic herring sequences are available at ENA under accession number PRJEB45393. The full software package is available for download at: https://github.com/laneatmore/BAMscorer.

ORCID

Giada Ferrari https://orcid.org/0000-0002-0850-1518
Lane M. Atmore https://orcid.org/0000-0002-8903-8149
Sissel Jentoft https://orcid.org/0000-0001-8707-531X
Kjetill S. Jakobsen https://orcid.org/0000-0002-8861-5397
Daniel Makowiecki https://orcid.org/0000-0002-1821-4627
James H. Barrett https://orcid.org/0000-0002-6683-9891
Bastiaan Star https://orcid.org/0000-0003-0235-9810

REFERENCES

Ayala, D., Guerrero, R. F., & Kirkpatrick, M. (2013). Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. *Evolution; International Journal of Organic Evolution*, *67*(4), 946–958. https://doi.org/10.1111/j.1558-5646.2012.01836.x

Bansal, V., Bashir, A., & Bafna, V. (2007). Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Research*, *17*(2), 219–230. https://doi.org/10.1101/gr.5774507

Barrett, J. H., Boessenkool, S., Kneale, C. J., O'Connell, T. C., & Star, B. (2020). Ecological globalisation, serial depletion and the medieval trade of walrus rostra. *Quaternary Science Reviews*, *229*, 106122. https://doi.org/10.1016/j.quascirev.2019.106122

Barth, J. M. I., Berg, P. R., Jonsson, P. R., Bonanomi, S., Corell, H., Hemmer-Hansen, J., & André, C. (2017). Genome architecture enables local adaptation of Atlantic cod despite high connectivity. *Molecular Ecology*, *26*(17), 4452–4466. https://doi.org/10.1111/mec.14207

Barth, J. M. I., Villegas-Ríos, D., Freitas, C., Moland, E., Star, B., André, C., & Jentoft, S. (2019). Disentangling structural genomic and behavioural barriers in a sea of connectivity. *Molecular Ecology*, *28*(6), 1394–1411. https://doi.org/10.1111/mec.15010

Berg, P. R., Jentoft, S., Star, B., Ring, K. H., Knutsen, H., Lien, S., & André, C. (2015). Adaptation to low salinity promotes genomic divergence in Atlantic cod (*Gadus morhua* L.). *Genome Biology and Evolution*, *7*(6), 1644–1663. https://doi.org/10.1093/gbe/evv093

Berg, P. R., Star, B., Pampoulie, C., Bradbury, I. R., Bentzen, P., Hutchings, J. A., & Jakobsen, K. S. (2017). Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity*, *119*(6), 418–428. https://doi.org/10.1038/hdy.2017.54

Berg, P. R., Star, B., Pampoulie, C., Sodeland, M., Barth, J. M. I., Knutsen, H., & Jentoft, S. (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Scientific Reports*, *6*, 23246. https://doi.org/10.1038/srep23246

Bohmann, K., Mirarab, S., Bafna, V., & Gilbert, M. T. P. (2020). Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. *Molecular Ecology*, *29*(14), 2521–2534. https://doi.org/10.1111/mec.15507

Cáceres, A., & González, J. R. (2015). Following the footprints of polymorphic inversions on SNP data: From detection to association tests. *Nucleic Acids Research*, *43*(8), e53. https://doi.org/10.1093/nar/gkv073

Cáceres, A., Sindi, S. S., Raphael, B. J., Cáceres, M., & González, J. R. (2012). Identification of polymorphic inversions from genotypes. *BMC Bioinformatics*, *13*(1). https://doi.org/10.1186/1471-2105-13-28

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M. H. S., Samaniego, J. A., Wales, N., Sicheritz-Pontén, T., & Gilbert, M. T. P. (2018). Single-tube library preparation for degraded DNA. *Methods in Ecology and Evolution*, *9*(2), 410–419. https://doi.org/10.1111/2041-210x.12871

Chouteau, M., Llaurens, V., Piron-Prunier, F., & Joron, M. (2017). Polymorphism at a mimicry supergene maintained by opposing frequency-dependent selection pressures. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(31), 8325–8329. https://doi.org/10.1073/pnas.1702482114

Cooper, A., & Poinar, H. N. (2000). Ancient DNA: Do It Right or Not at All. *Science*, *289*(5482), 1139–1139. https://doi.org/10.1126/science.289.5482.1139b

Damgaard, P. B., Margaryan, A., Schroeder, H., Orlando, L., Willerslev, E., & Allentoft, M. E. (2015). Improving access to endogenous DNA in ancient bones and teeth. *Scientific Reports*, *5*, 11184. https://doi.org/10.1038/srep11184

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Davey, J. W., Barker, S. L., Rastas, P. M., Pinharanda, A., Martin, S. H., Durbin, R., & Jiggins, C. D. (2017). No evidence for maintenance of a sympatric *Heliconius* species barrier by chromosomal inversions. *Evolution Letters*, *1*(3), 138–154. https://doi.org/10.1002/evl3.12

Dodsworth, S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science*, *20*(9), 525–527. https://doi.org/10.1016/j.tplants.2015.06.012

Domagała, R., & Franczuk, R. (1992). Wyniki badań archeologiczno-architektonicznych na zamku w Małej Nieszawce. *Rocznik Muzeum W Toruniu*, *9*, 41–53.

Donnelly, M. P., Paschou, P., Grigorenko, E., Gurwitz, D., Mehdi, S. Q., Kajuna, S. L. B., & Kidd, K. K. (2010). The distribution and most recent common ancestor of the 17q21 inversion in humans. *American Journal of Human Genetics*, *86*(2), 161–171. https://doi.org/10.1016/j.ajhg.2010.01.007

Fages, A., Seguin-Orlando, A., Germonpré, M., & Orlando, L. (2020). Horse males became over-represented in archaeological assemblages during the Bronze Age. *Journal of Archaeological Science: Reports*, *31*, 102364. https://doi.org/10.1016/j.jasrep.2020.102364

Fang, Z., Pyhäjärvi, T., Weber, A. L., Dawe, R. K., Glaubitz, J. C., de González, J. S., & Ross-Ibarra, J. (2012). Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics*, *191*(3), 883–894. https://doi.org/10.1534/genetics.112.138578

Ferrari, G., Cuevas, A., Gondek-Wyrozemska, A. T., Ballantyne, R., Kersten, O., Pálsdóttir, A. H., & Star, B. (2021). The preservation of ancient DNA in archaeological fish bone. *Journal of Archaeological Science*, *126*, 105317. https://doi.org/10.1016/j.jas.2020.105317

François, O., & Jay, F. (2020). Factor analysis of ancient population genomic samples. *Nature Communications*, *11*(1), 4661. https://doi.org/10.1038/s41467-020-18335-6

Gilbert, M. T. P., Bandelt, H.-J., Hofreiter, M., & Barnes, I. (2005). Assessing ancient DNA studies. *Trends in Ecology & Evolution*, *20*(10), 541–544. https://doi.org/10.1016/j.tree.2005.07.005

Glover, K. A., Pertoldi, C., Besnier, F., Wennevik, V., Kent, M., & Skaala, Ø. (2013). Atlantic salmon populations invaded by farmed escapees: Quantifying genetic introgression with a Bayesian approach and SNPs. *BMC Genetics*, *14*, 74. https://doi.org/10.1186/1471-2156-14-74

Glover, K. A., Solberg, M. F., McGinnity, P., Hindar, K., Verspoor, E., Coulson, M. W., & Svåsand, T. (2017). Half a century of genetic interaction between farmed and wild Atlantic salmon: Status of knowledge and unanswered questions. *Fish and Fisheries*, *18*(5), 890–927. https://doi.org/10.1111/faf.12214

Grossen, C., Biebach, I., Angelone-Alasaad, S., Keller, L. F., & Croll, D. (2018). Population genomics analyses of European ibex species show lower diversity and higher inbreeding in reintroduced populations. *Evolutionary Applications*, *11*(2), 123–139. https://doi.org/10.1111/eva.12490

Han, F., Jamsandekar, M., Pettersson, M. E., Su, L., Fuentes-Pardo, A. P., Davis, B. W., & Andersson, L. (2020). Ecological adaptation in Atlantic herring is associated with large shifts in allele frequencies at hundreds of loci. *eLife*, *9*, e61076. https://doi.org/10.7554/eLife.61076

Hoffmann, A. A., & Rieseberg, L. H. (2008). Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annual Review of Ecology, Evolution, and Systematics*, *39*, 21–42. https://doi.org/10.1146/annurev.ecolsys.39.110707.173532

Hughes, J. F., Skaletsky, H., Pyntikova, T., Graves, T. A., van Daalen, S. K. M., Minx, P. J., & Page, D. C. (2010). Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature*, *463*(7280), 536–539. https://doi.org/10.1038/nature08700

Iwaszkiewicz, M. (1991). Szczątki ryb z zamku krzyżackiego w Małej Nieszawce (woj. toruńskie), Roczniki Akademii Rolniczej w Poznaniu 227. *Archeozoologia*, *16*, 3–5.

Jay, P., Chouteau, M., Whibley, A., Bastide, H., Parrinello, H., Llaurens, V., & Joron, M. (2021). Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. *Nature Genetics*, *53*(3), 288–293. https://doi.org/10.1038/s41588-020-00771-1

Jay, P., Whibley, A., Frézal, L., Rodríguez de Cara, M. Á., Nowell, R. W., Mallet, J., & Joron, M. (2018). Supergene evolution triggered by the introgression of a chromosomal inversion. *Current Biology: CB*, *28*(11), 1839–1845.e3. https://doi.org/10.1016/j.cub.2018.04.072

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., & Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., & Orlando, L. (2013). mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, *29*(13), 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., & ffrench-Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, *477*(7363), 203–206. https://doi.org/10.1038/nature10341

Joron, M., Papa, R., Beltrán, M., Chamberlain, N., Mavárez, J., Baxter, S., & Jiggins, C. D. (2006). A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biology*, *4*(10), e303. https://doi.org/10.1371/journal.pbio.0040303

Keighley, X., Bro-Jørgensen, M. H., Ahlgren, H., Szpak, P., Ciucani, M. M., Sánchez Barreiro, F., & Olsen, M. T. (2021). Predicting sample success for large-scale ancient DNA studies on marine mammals. *Molecular Ecology Resources*, *21*(4), 1149–1166. https://doi.org/10.1111/1755-0998.13331

Kirubakaran, T. G., Grove, H., Kent, M. P., Sandve, S. R., Baranski, M., Nome, T., & Andersen, Ø. (2016). Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*, *25*(10), 2130–2143. https://doi.org/10.1111/mec.13592

Kozak, K. M., Wahlberg, N., Neild, A. F. E., Dasmahapatra, K. K., Mallet, J., & Jiggins, C. D. (2015). Multilocus species trees show the recent adaptive radiation of the mimetic *Heliconius* butterflies. *Systematic Biology*, *64*(3), 505–524. https://doi.org/10.1093/sysbio/syv007

Lamichhaney, S., Fuentes-Pardo, A. P., Rafati, N., Ryman, N., McCracken, G. R., Bourne, C., & Andersson, L. (2017). Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(17), E3452–E3461. https://doi.org/10.1073/pnas.1617728114

Lamichhaney, S., Martinez Barrio, A., Rafati, N., Sundström, G., Rubin, C.-J., Gilbert, E. R., & Andersson, L. (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*(47), 19345–19350. https://doi.org/10.1073/pnas.1216128109

Lee, S., Zou, F., & Wright, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Annals of Statistics*, *38*(6), 3605–3629. https://doi.org/10.1214/10-AOS821

Leitwein, M., Garza, J. C., & Pearse, D. E. (2017). Ancestry and adaptive evolution of anadromous, resident, and adfluvial rainbow trout (Oncorhynchus mykiss) in the San Francisco bay area: application of adaptive genomic variation to conservation in a highly impacted landscape. *Evolutionary Applications*, *10*(1), 56–67. https://doi.org/10.1111/eva.12416

Lemaitre, C., Braga, M. D. V., Gautier, C., Sagot, M.-F., Tannier, E., & Marais, G. A. B. (2009). Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. *Genome Biology and Evolution*, *1*, 56–66. https://doi.org/10.1093/gbe/evp006

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Llamas, B., Valverde, G., Fehren-Schmitz, L., Weyrich, L. S., Cooper, A., & Haak, W. (2017). From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. *STAR: Science & Technology of Archaeological Research*, *3*(1), 1–14. https://doi.org/10.1080/20548923.2016.1258824

Lowry, D. B., & Willis, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biology*, *8*(9), e1000500. https://doi.org/10.1371/journal.pbio.1000500

Ma, J., & Amos, C. I. (2012). Investigation of inversion polymorphisms in the human genome using principal components analysis. *PLoS One*, *7*(7), e40224. https://doi.org/10.1371/journal.pone.0040224

Mak, S. S. T., Gopalakrishnan, S., Carøe, C., Geng, C., Liu, S., Sinding, M.-H.- S., & Gilbert, M. T. P. (2017). Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *GigaScience*, *6*(8), 1–13. https://doi.org/10.1093/gigascience/gix049

Makowiecki, D. (2003). *Historia ryb i rybołówstwa w holocenie na Niżu Polskim w świetle badań archeoichtiologicznych*. Institute of Archaeology and Ethnology, Polish Academy of Sciences.

Makowiecki, D., Orton, D. C., & Barrett, J. H. (2016). Cod and herring in medieval Poland. In J. H. Barrett, & D. Orton (Eds.), *Cod & herring: The archaeology & history of medieval sea fishing* (pp. 117–132). Oxbow Books.

Malé, P.-J.- G., Bardon, L., Besnard, G., Coissac, E., Delsuc, F., Engel, J., & Chave, J. (2014). Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Molecular Ecology Resources*, *14*(5), 966–975. https://doi.org/10.1111/1755-0998.12246

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Marcus, J. M. (2021). Our love-hate relationship with DNA barcodes, the Y2K problem, and the search for next generation barcodes. *AIMS Genetics*, *05*(01), 001–023. https://doi.org/10.3934/genet.2018.1.1

Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., & Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife*, *5*, e12081. https://doi.org/10.7554/eLife.12081

Martínez-García, L., Ferrari, G., Oosting, T., Ballantyne, R., van der Jagt, I., Ystgaard, I., & Star, B. (2021). Historical demographic processes dominate genetic variation in ancient Atlantic cod mitogenomes. *Frontiers in Ecology and Evolution*, *9*, 342. https://doi.org/10.3389/fevo.2021.671281

Matschiner, M., Barth, J. M. I., Tørresen, O. K., Star, B., Baalsrud, H. T., Brieuc, M. S. O., & Jentoft, S. (2021). Origin and fate of supergenes in Atlantic cod. *BioRxiv*. https://doi.org/10.1101/2021.02.28.433253

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Morales, H. E., Faria, R., Johannesson, K., Larsson, T., Panova, M., Westram, A. M., & Butlin, R. K. (2019). Genomic architecture of parallel ecological divergence: Beyond a single environmental contrast. *Science Advances*, *5*(12), eaav9963. https://doi.org/10.1126/sciadv.aav9963

Nadeau, N. J. (2016). Genes controlling mimetic colour pattern variation in butterflies. *Current Opinion in Insect Science*, *17*, 24–31. https://doi.org/10.1016/j.cois.2016.05.013

Nadeau, N. J., Pardo-Diaz, C., Whibley, A., Supple, M. A., Saenko, S. V., Wallbank, R. W. R., & Jiggins, C. D. (2016). The gene cortex controls mimicry and crypsis in butterflies and moths. *Nature*, *534*(7605), 106–110. https://doi.org/10.1038/nature17961

Nevill, P. G., Zhong, X., Tonti-Filippini, J., Byrne, M., Hislop, M., Thiele, K., & Small, I. (2020). Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. *Plant Methods*, *16*, 1. https://doi.org/10.1186/s13007-019-0534-5

Nistelberger, H. M., Pálsdóttir, A. H., Star, B., Leifsson, R., Gondek, A. T., Orlando, L., & Boessenkool, S. (2019). Sexing Viking Age horses from burial and non-burial sites in Iceland using ancient DNA. *Journal of Archaeological Science*, *101*, 115–122. https://doi.org/10.1016/j.jas.2018.11.007

Noor, M. A. F., Grams, K. L., Bertucci, L. A., & Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(21), 12084–12088. https://doi.org/10.1073/pnas.221274498

Pálsdóttir, A. H., Bläuer, A., Rannamäe, E., Boessenkool, S., & Hallsson, J. H. (2019). Not a limitless resource: Ethics and guidelines for destructive sampling of archaeofaunal remains. *Royal Society Open Science*, *6*(10), 191059. https://doi.org/10.1098/rsos.191059

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Pečnerová, P., Díez-Del-Molino, D., Dussex, N., Feuerborn, T., von Seth, J., van der Plicht, J., & Dalén, L. (2017). Genome-based sexing provides clues about behavior and social structure in the woolly mammoth. *Current Biology: CB*, *27*(22), 3505–3510.e3. https://doi.org/10.1016/j.cub.2017.09.064

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Pettersson, M. E., Rochus, C. M., Han, F., Chen, J., Hill, J., Wallerman, O., & Andersson, L. (2019). A chromosome-level assembly of the Atlantic herring genome—Detection of a supergene and other signals of selection. *Genome Research*, *29*(11), 1919–1928. https://doi.org/10.1101/gr.253435.119

Pinsky, M. L., Eikeset, A. M., Helmerson, C., Bradbury, I. R., Bentzen, P., Morris, C., & Star, B. (2021). Genomic stability through time despite decades of exploitation in cod on both sides of the Atlantic. *Proceedings of the National Academy of Sciences*, *118*(15), e2025453118. https://doi.org/10.1073/pnas.2025453118

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell*, *58*(4), 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

Ripma, L. A., Simpson, M. G., & Hasenstab-Lehman, K. (2014). Geneious! Simplified genome skimming methods for phylogenetic systematic studies: A case study in Oreocarya (Boraginaceae). *Applications in Plant Sciences*, *2*(12), 1400062. https://doi.org/10.3732/apps.1400062

Ros-Freixedes, R., Battagin, M., Johnsson, M., Gorjanc, G., Mileham, A. J., Rounsley, S. D., & Hickey, J. M. (2018). Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. *Genetics, Selection, Evolution: GSE*, *50*(1), 1–14. https://doi.org/10.1186/s12711-018-0436-4

Ruiz-Arenas, C., Cáceres, A., López-Sánchez, M., Tolosana, I., Pérez-Jurado, L., & González, J. R. (2019). scoreInvHap: Inversion genotyping for genome-wide association studies. *PLoS Genetics*, *15*(7), e1008203. https://doi.org/10.1371/journal.pgen.1008203

Salm, M. P. A., Horswell, S. D., Hutchison, C. E., Speedy, H. E., Yang, X., Liang, L., & Shoulders, C. C. (2012). The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome Research*, *22*(6), 1144–1153. https://doi.org/10.1101/gr.126037.111

Savage, W. K., & Mullen, S. P. (2009). A single origin of Batesian mimicry among hybridizing populations of admiral butterflies (Limenitis arthemis) rejects an evolutionary reversion to the ancestral phenotype. *Proceedings. Biological Sciences/The Royal Society*, *276*(1667), 2557–2565. https://doi.org/10.1098/rspb.2009.0256

Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., & Orlando, L. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols*, *9*(5), 1056–1082. https://doi.org/10.1038/nprot.2014.063

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., Al-Rasheid, K. A. S., Willerslev, E., & Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, *13*, 178. https://doi.org/10.1186/1471-2164-13-178

Sindi, S. S., & Raphael, B. J. (2010). Identification and frequency estimation of inversion polymorphisms from haplotype data. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *17*(3), 517–531. https://doi.org/10.1089/cmb.2009.0185

Skoglund, P., Sjödin, P., Skoglund, T., Lascoux, M., & Jakobsson, M. (2014). Investigating population history using temporal genetic differentiation. *Molecular Biology and Evolution*, *31*(9), 2516–2527. https://doi.org/10.1093/molbev/msu192

Sodeland, M., Jorde, P. E., Lien, S., Jentoft, S., Berg, P. R., Grove, H., & Knutsen, H. (2016). "Islands of divergence" in the Atlantic cod genome represent polymorphic chromosomal rearrangements. *Genome Biology and Evolution*, *8*(4), 1012–1022. https://doi.org/10.1093/gbe/evw057

Star, B., Barrett, J. H., Gondek, A. T., & Boessenkool, S. (2018). Ancient DNA reveals the chronology of walrus ivory trade from Norse Greenland. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1884), 20180978. https://doi.org/10.1098/rspb.2018.0978

Star, B., Boessenkool, S., Gondek, A. T., Nikulina, E. A., Hufthammer, A. K., Pampoulie, C., & Barrett, J. H. (2017). Ancient DNA reveals the arctic origin of viking age cod from Haithabu, Germany. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(34), 9152–9157. https://doi.org/10.1073/pnas.1710186114

Star, B., Nederbragt, A. J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T. F., & Jakobsen, K. S. (2011). The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, *477*(7363), 207–210. https://doi.org/10.1038/nature10342

Suchan, T., Pitteloud, C., Gerasimova, N. S., Kostikova, A., Schmid, S., Arrigo, N., & Alvarez, N. (2016). Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS One*, *11*(3), e0151651. https://doi.org/10.1371/journal.pone.0151651

Tin, M.- M.-Y., Economo, E. P., & Mikheyev, A. S. (2014). Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics. *PLoS One*, *9*(5), e96793. https://doi.org/10.1371/journal.pone.0096793

Todesco, M., Owens, G. L., Bercovich, N., Légaré, J.-S., Soudi, S., Burge, D. O., & Rieseberg, L. H. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, *584*(7822), 602–607. https://doi.org/10.1038/s41586-020-2467-6

Tørresen, O. K., Star, B., Jentoft, S., Reinar, W. B., Grove, H., Miller, J. R., & Nederbragt, A. J. (2017). An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics*, *18*(1), 95. https://doi.org/10.1186/s12864-016-3448-x

Twyford, A. D., & Friedman, J. (2015). Adaptive divergence in the monkey flower Mimulus guttatus is maintained by a chromosomal inversion. *Evolution; International Journal of Organic Evolution*, *69*(6), 1476–1486. https://doi.org/10.1111/evo.12663

van Damme, C. J. G., Dickey-Collas, M., Rijnsdorp, A. D., & Kjesbu, O. S. (2009). Fecundity, atresia, and spawning strategies of Atlantic herring (Clupea harengus). *Canadian Journal of Fisheries and Aquatic Sciences*, *66*(12), 2130–2141. https://doi.org/10.1139/f09-153

Van der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media.

van der Valk, T., Pečnerová, P., Díez-Del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., & Dalén, L. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, *591*(7849), 265–269. https://doi.org/10.1038/s41586-021-03224-9

Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, *33*(6), 427–440. https://doi.org/10.1016/j.tree.2018.04.002

Wetterstrand, K. A. (2021). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: https://www.genome.gov/sequencingcostsdata. Accessed 19 September 2021.

Zeng, C.-X., Hollingsworth, P. M., Yang, J., He, Z.-S., Zhang, Z.-R., Li, D.-Z., & Yang, J.-B. (2018). Genome skimming herbarium specimens for DNA barcoding and phylogenomics. *Plant Methods*, *14*, 43. https://doi.org/10.1186/s13007-018-0300-0

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Ferrari, G., Atmore, L. M., Jentoft, S., Jakobsen, K. S., Makowiecki, D., Barrett, J. H., & Star, B. (2021). An accurate assignment test for extremely low-coverage whole-genome sequence data. *Molecular Ecology Resources*, 00, 1–15. https://doi.org/10.1111/1755-0998.13551