Received: 19 April 2017 | Revised: 23 June 2017 | Accepted: 28 June 2017

DOI: 10.1111/mec.14264

INVITED REVIEWS AND SYNTHESES

WILEY MOLECULAR ECOLOGY

Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations

Angela P. Fuentes-Pardo | Daniel E. Ruzzante

Department of Biology, Dalhousie University, Halifax, NS, Canada

Correspondence
Angela P. Fuentes-Pardo and Daniel E. Ruzzante, Department of Biology, Dalhousie University, Halifax, NS, Canada.
Emails: apfuentesp@gmail.com; daniel.ruzzante@dal.ca

Funding information
Natural Sciences and Engineering Research Council of Canada; Killam Trust Foundation; Vanier Canada Graduate Scholarship; the President's Award of Dalhousie University; Nova Scotia Graduate Scholarship

Abstract

Whole-genome resequencing (WGR) is a powerful method for addressing fundamental evolutionary biology questions that have not been fully resolved using traditional methods. WGR includes four approaches: the sequencing of individuals to a high depth of coverage with either unresolved or resolved haplotypes, the sequencing of population genomes to a high depth by mixing equimolar amounts of unlabelled-individual DNA (Pool-seq) and the sequencing of multiple individuals from a population to a low depth (lcWGR). These techniques require the availability of a reference genome. This, along with the still high cost of shotgun sequencing and the large demand for computing resources and storage, has limited their implementation in nonmodel species with scarce genomic resources and in fields such as conservation biology. Our goal here is to describe the various WGR methods, their pros and cons and potential applications in conservation biology. WGR offers an unprecedented marker density and surveys a wide diversity of genetic variations not limited to single nucleotide polymorphisms (e.g., structural variants and mutations in regulatory elements), increasing their power for the detection of signatures of selection and local adaptation as well as for the identification of the genetic basis of phenotypic traits and diseases. Currently, though, no single WGR approach fulfils all requirements of conservation genetics, and each method has its own limitations and sources of potential bias. We discuss proposed ways to minimize such biases. We envision a not distant future where the analysis of whole genomes becomes a routine task in many nonmodel species and fields including conservation biology.

KEYWORDS
conservation biology, low-coverage sequencing, management, Pool-seq, population genomics, whole-genome sequencing

1 | INTRODUCTION

Over the last 40 years, genetics has emerged as an important tool in the conservation of threatened species. Based on the analysis of genetic variation of individuals and populations, genetics has provided insights on diverse areas in conservation biology including species identification, hybridization, kinship, evolutionary history, effective population size (N_e), population substructure, population connectivity, adaptive genetic variation, local adaptation and inbreeding (Hedrick & Miller, 1992; von der Heyden et al., 2014; Haig et al., 2016).

Genetic variation has traditionally been examined using from a single to a handful (12–24) of molecular markers including allozymes, mitochondrial DNA and microsatellites (see review by Allendorf

(2016)). Most of these markers target a few neutral positions in the genome, limiting the ability to estimate genomewide parameters (Avise, 2010). The development of high-throughput sequencing (HTS) technologies over a decade ago revolutionized the way genetic variation is assessed (Goodwin, McPherson, & McCombie, 2016). These technologies allow the massive sequencing of thousands to millions of loci in a short time for an affordable cost, resulting in a much higher marker density than experienced with past technologies. Today, individual research groups have the option of sequencing the reference genome of their focal species and resequencing genomes of individuals and populations for the detection of both, neutral and adaptive variation (Ellegren, 2014). The extraordinary increase in number of markers available with genomic approaches has sparked much expectation within the conservation community, reflected in several recent review papers on this topic (Primmer, 2009; Avise, 2010; Frankham, 2010; Ouborg, Pertoldi, Loeschcke, Bijlsma, & Hedrick, 2010; Allendorf, Hohenlohe, & Luikart, 2010; Angeloni, Wagemaker, Vergeer, & Ouborg, 2012; Funk, McKay, Hohenlohe, & Allendorf, 2012; Steiner, Putnam, Hoeck, & Ryder, 2013; McMahon, Teeling, & Höglund, 2014; Shafer et al. 2015; Garner et al., 2016; Benestan et al., 2016). The hype is a reflection of the promise of increased statistical power in population genetics tests, but most importantly, of the possibility of addressing long-standing questions in conservation biology not fully resolved with traditional methods. Some of these questions are as follows: What is the phylogenetic relationship between unresolved taxa? What are the loci responsible for speciation, for local adaptation, for interactions among species or for inbreeding depression? What is the genetic basis of traits related to fitness? (Ouborg et al., 2010; Allendorf et al., 2010; McMahon et al., 2014).

The advances achieved with HTS promise an exciting time for genomics-based research, although these developments have their own limitations. For example, short-read sequences (~100 base pairs [bp] long), that are commonly obtained with current sequencing technologies, are problematic for genome assembly and detection of large structural variants. The relatively high error rate of existing sequencing platforms makes it necessary to obtain high depth of coverage for the correct identification of variants (Goodwin et al., 2016) and sequencing cost is still high for population studies that require analysing multiple individuals. Currently, some alternatives to overcome the cost limitation are (i) using reduced-representation sequencing (RRS) methods that screen a fraction of the genome (da Fonseca et al., 2016), (ii) obtaining whole-genome resequencing (WGR) data from pooled DNA of individuals per population to a high coverage (known as Pool-seq, Schlötterer, Tobler, Kofler, & Nolte, 2014), or (iii) low-coverage WGR data of individuals from a population (known as lcWGR; Nielsen, Paul, Albrechtsen, & Song, 2011). These approaches have successfully screened multiple loci genomewide in several species and have been instrumental in addressing a variety of questions in molecular ecology (Hohenlohe et al., 2010; Foote et al., 2016; Lamichhaney et al., 2017). These methods, however, have their own restrictions and sources of bias and error that should be minimized for the correct inference of population parameters (Anderson, Skaug, & Barshis, 2014; Lowry et al., 2017a).

Today, when virtually the genome of any species can be sequenced, it is pertinent to ask, when is the analysis of whole-genome data justified in conservation biology? What are the limitations of current WGR methods, and how could they be overcome? These questions are particularly important for three main reasons: (i) traditional molecular methods can solve some of the questions in conservation for a small fraction of the cost and effort relative to genomic approaches (e.g., dozens of polymorphic microsatellites generate acceptable estimates of population structure, gene flow, N_e, kinship) (McMahon et al., 2014; Allendorf, 2016); (ii) RRS methods generate thousands of molecular markers genomewide, increasing the power of statistical tests for a lower cost compared to whole-genome approaches (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016); (iii) current short-read sequence data present some restrictions that limit the kind of analysis that can be performed (Goodwin et al., 2016).

Given the increased interest in the use of genomics in conservation biology, in this review, we first provide a general background on sequencing technologies and whole-genome sequencing (Box 1). We then describe the various WGR approaches used in population genomics (Box 2), discuss their limitations and potential solutions (Box 3 and 4) and compare WGR to RRS methods (Table 1). We also discuss limitations of genome scans for detecting selection and inferring adaptation from genomic data (Box 5). We subsequently present case studies for the areas of conservation biology that can in principle be benefited by WGR analysis (Table 2). Finally, we provide guidelines for choosing between RRS and WGR methods depending on the type of genetic variation of interest and the expected haplotype block size and explore recent innovations that promise overcoming the limitations of current methodologies.

2 | GENOME SEQUENCING TECHNIQUES

The understanding of the complexity of the genome gained in the Human Genome Project (1990–2003) (Human Genome Research Institute (NIH), https://www.genome.gov/12011239/), coupled with advances in molecular techniques and equipment set the stage for the beginning of the “genomic era” two decades ago. The development of sequencing technologies in particular has revolutionized the way we examine and comprehend the genome. Three major sequencing generations have taken place thus far. Sanger sequencing (or “chain-termination method”), considered the first generation, was introduced in 1977 (Sanger, Nicklen, & Coulson, 1977). This sequencing method provides high per-base accuracy (99,999%, Shendure and Ji (2008)) and medium-read length (~1,000 bp), but low throughput and relatively high cost per base. Genome assembly is achieved via sequencing of bacterial artificial chromosome libraries containing pieces of the whole-genome. Using specialized software, the sequence of each fragment is assembled into a contiguous sequence (NIH, https://www.genome.gov/12011239/). The first genome sequences of several model species (e.g., yeast, *Drosophila*

BOX 1 Genome assembly and completeness of genomes sequenced to date

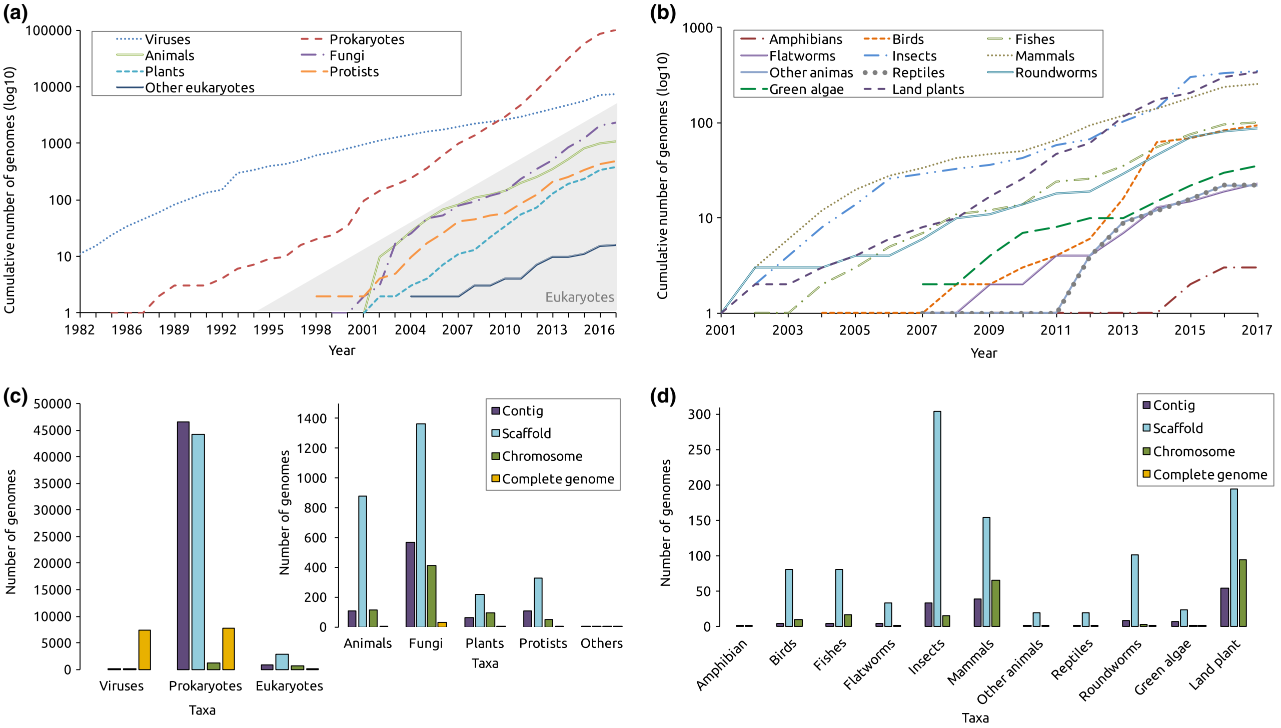


FIGURE B1 State of the art of genomes publicly available in GenBank to date (data retrieved in June 2017). (a) Cumulative number of genomes per year for some major taxonomic groups. The grey shadow indicates eukaryotic groups. (b) Cumulative number of genomes per year for some taxonomic groups within animals and plants. (c) Genome completeness of some major taxonomic groups and within eukaryotes (inset). (d) Genome completeness of some taxonomic groups within animals and plants. Charts c and d were made based on the "assembly level" annotation associated with each genome listed in the Genome browser of GenBank (https://www.ncbi.nlm.nih.gov/genome/browse/#) and were used as proxy of genome completeness. Four levels of assembly were used (from lowest to highest): contigs, scaffolds, chromosomes, and complete genome (https://www.ncbi.nlm.nih.gov/assembly/help/#definition)

The assembly of a genome consists in joining sequences of the DNA of one or several conspecific individuals into a single sequence. DNA is first fragmented to a particular length and sequenced to a certain coverage depending on the sequencing platform. DNA sequences are then assembled using specialized computer algorithms. As per the current assembly algorithms, a haploid sequence is generally obtained, implying species genome diversity and assembly accuracy could be compromised (Baker, 2012; Paten et al., 2017).

A genome project generally has the goal of obtaining a contiguous and complete genome sequence with annotated genes (Veeckman, Ruttink, & Vandepoele, 2016). To achieve this using current sequencing technologies of second- and third generation, it is necessary to get high sequence depth (>50–60×) evenly distributed across the genome, counteracting the relatively high sequencing error rate (Goodwin et al., 2016; Lee et al., 2016; Bleidorn, 2016). As mentioned, repetitive sequences and large structural variants are difficult to assemble using just short reads. Thus, the combined assembly of short and long reads (or long reads only (Chakraborty, Baldwin-Brown, Long, & Emerson, 2016; Bickhart et al., 2017)) is common practice (Ekblom & Wolf, 2014).

The quality of a genome assembly (how complete and accurate it is) has traditionally been assessed using different metrics such as N50 and L50. These two metrics assess contiguity, N50 estimates the contig/scaffold length at which 50% of the total bases fall in a given assembly, and L50 is the number of contigs/scaffolds that are longer than or equal to the N50 length, including 50% of the total bases of a given assembly (https://www.ncbi.nlm.nih.gov/assembly/help/). However, these metrics have several limitations (Salzberg et al., 2012; Gurevich, Saveliev, Vyahhi, & Tesler, 2013) for which new ones have been proposed, for example, NG50 and NA50 (Earl et al., 2011; Bradnam et al., 2013). Guidelines for achieving high-quality de novo genome assembly of nonmodel species are presented in Ekblom and Wolf (2014) and in Koepfli et al. (2015), and recent advances in genome assembly are addressed in a special issue of the journal Genome Research (Phillippy, 2017).

BOX 1 **Continued**

Genomes assembled to date are publicly available online in GenBank of the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/genbank/), the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena) and the DNA DataBank of Japan (DDBJ) (http://www.ddbj.nig.ac.jp/), institutions that constitute the International Nucleotide Sequence Database Collaboration. Ongoing genome projects are listed in GenBank in the Bioprojects webpage (https://www.ncbi.nlm.nih.gov/bioproject/) and in the Genomes Online Database (GOLD) (https://gold.jgi.doe.gov/projects) (Mukherjee et al., 2017).

With the advances in sequencing technology and computation achieved with the Human Genome Project (HGP) (1990–2003) (Human Genome Sequencing Consortium 2004), there has been an exponential growth in the number of genomes published in GenBank per year (Figure B1a). The great majority of genomes correspond to viruses and prokaryotes, and within eukaryotes, fungi, animals and protists follow in representation. Within animals, mammals and insects have the highest number of genomes sequenced, whereas amphibians and reptiles have the lowest. Within plants, land plants have the highest representation (Figure B1b). Contrary to expectations, most prokaryotic and eukaryotic genome sequences to date are incomplete (Figure B1c,d), as they are assembled to the scaffold level. In general, it is only prokaryotes, viruses and a few model eukaryotic species (e.g., yeast, 12.1 megabases—Mb, and fruit fly, 175 Mb) with relatively small, simple or only moderately complex genomes that have their sequence complete (Figure B1c, d). Within animals, mammals, fishes, insects and birds have their genomes assembled to the chromosome level, whereas mammals, insects, flat and round worms followed by fishes and birds have their genomes at the contig level. Land plants have also a great proportion of genomes at the chromosome and contig level.

Within eukaryotes, there is a great diversity in genome size, complexity and proportions of repetitive sequence content [e.g., Atlantic salmon (*Salmo salar*): 2.97 Giga bases (Gb) with ~60% repetitive elements (Lien et al., 2016); loblolly pine tree (*Pinus taeda*): 23.2 Billion bases (Bb; largest genome sequenced to date with 82% repetitive content (Neale et al., 2014). Genome complexity increases with the occurrence of duplicated genes (or paralogs), long repeat sequences, polymorphic genes (e.g., MHC, Trowsdale & Knight, 2013), GC content and ploidy (Treangen & Salzberg, 2011). The complexity of a genome, and especially its repetitive content (Reinert, Langmead, Weese, & Evers, 2015; Goodwin et al., 2016), imposes significant challenges for sequence assembly (Treangen & Salzberg, 2011; Baker, 2012; Ekblom & Wolf, 2014; Ellegren, 2014). This could largely explain the varying degree of completeness observed in the genomes sequenced to date.

How complete and accurate a genome assembly is will determine its suitability for posterior analyses. For example, a very incomplete genome can still be useful for the identification of SNPs but it would fail in the detection of large structural variation (da Fonseca et al., 2016). Despite its utility in some applications, the general consensus is that incomplete draft genomes bring more problems than solutions, especially for the accuracy of SNP calling (Li & Wren, 2014).

melanogaster, *Caenorhabditis elegans*, *Arabidopsis thaliana*) including humans were obtained with this technology, which incidentally, also led to the expansion of genetics research overall (Pettersson, Lundeberg, & Ahmadian, 2009; Goodwin et al., 2016; Heather & Chain, 2016).

The second generation of sequencing technologies appeared between 2005 and 2010 with the development of “sequencing-by-synthesis” and innovative high-throughput systems (i.e., 454-pyrosequencing, Illumina, Ion-Proton). These technologies have a higher error rate (accuracy >99.5%) and produce shorter sequences (75- to 300-bp Illumina, <400-bp Ion-Proton, <700-bp 454-pyrosequencing) than Sanger sequencing, but offer an exceptional high throughput owing to the massive parallel sequencing of DNA fragments. Genome assembly is accomplished with the joining of paired-end short reads (~100 bp) of various DNA libraries with different insert sizes (350 bp–40 kilobases—Kb). The assembly process involves the building of contigs (overlapping sequences with no gaps or runs of more than 10 ambiguous bases [Ns]), that are merged into scaffolds (larger sequences formed by joining nonoverlapping contigs). Scaffolds are linked and ordered with sequence data of long insert size libraries to obtain a consensus sequence. Joining short reads, however, is problematic specially at repetitive sequences, gaps (or portions of DNA with unknown sequence) are thus common in genomes assembled with second-generation sequencing technologies. Similarly, the detection of large structural variants (SVs) and the ability to assign groups of genetic variants that are located in the same chromosome (i.e., “haplotype phasing,” Snyder, Adey, Kitzman, & Shendure, 2015) are limited with short reads due to the small and fragmented nature of these sequences. The main contributions of second-generation technologies have been the substantial drop in sequencing cost and the exponential increase in throughput, unlocking genome-, exome-, transcriptome- and epigenome- sequencing approaches to nonmodel organisms and in doing so, revolutionizing medicine, agriculture and biological research (Pettersson et al., 2009; Goodwin et al., 2016; Heather & Chain, 2016).

The third generation appeared between 2011 and 2014 with sequencing technologies that produce reads of unprecedented length (average ~2–10 Kb). Currently, long reads can be obtained with two methods, “single-molecule real-time” sequencing (SMRT-seq; i.e., Pacific Biosystems [PacBio], Oxford Nanopore Technologies [ONT])

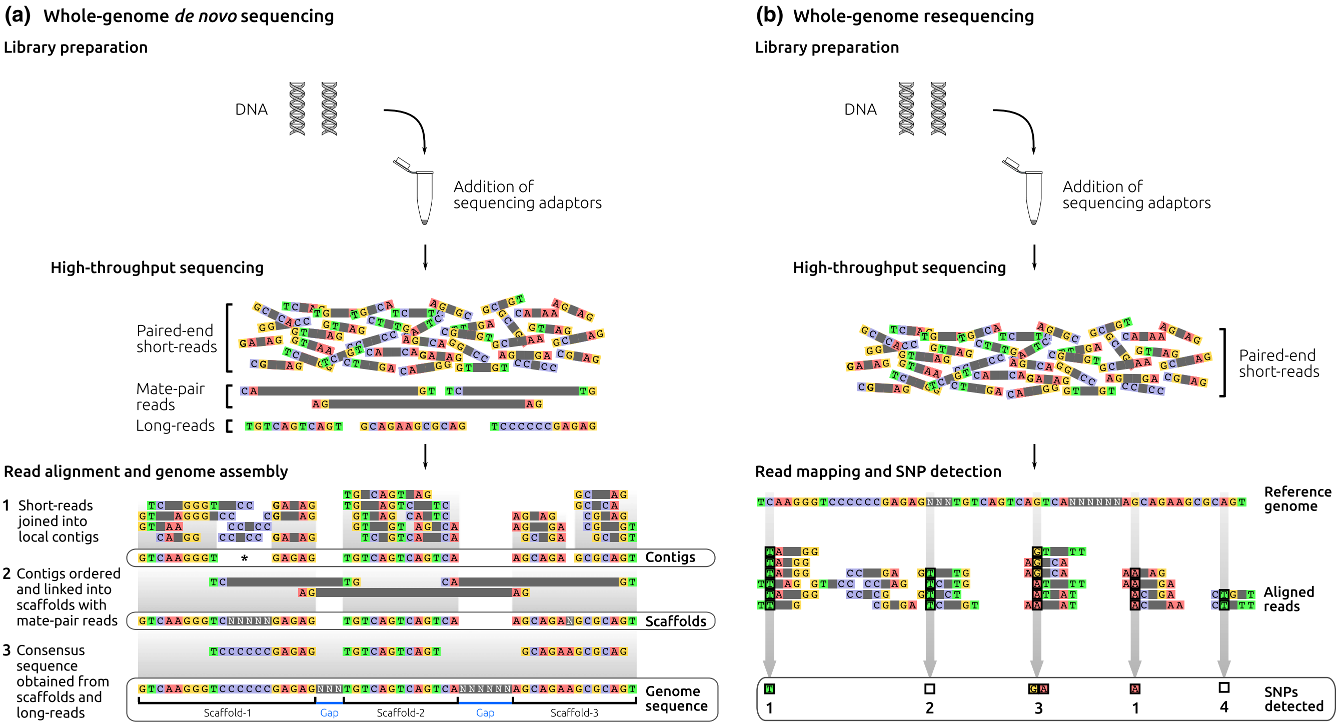

(a) Whole-genome de novo sequencing
Library preparation
DNA
Addition of sequencing adaptors
High-throughput sequencing
Paired-end short-reads
Mate-pair reads
Long-reads
TGTCAGTCAGT
GCAGAAGCGCAG
TCCCCCCGAGAG
Read alignment and genome assembly
1 Short-reads joined into local contigs
GTCAAGGGT
*
GAGAG
TGTCAGTCAGTCA
AGCAGA
GCGCAGT
Contigs
2 Contigs ordered and linked into scaffolds with mate-pair reads
GTCAAGGGTCNNNNNGAGAG
TGTCAGTCAGTCA
AGCAGANGCGCAGT
Scaffolds
3 Consensus sequence obtained from scaffolds and long-reads
TCCCCCCGAGAG
TGTCAGTCAGT
GCAGAAGCGCAG
GTCAAGGGTCCCCCCGAGAGNNNTGTCAGTCAGTCANNNNNNAGCAGAAGCGCAGT
Genome sequence
Scaffold-1
Gap
Scaffold-2
Gap
Scaffold-3
(b) Whole-genome resequencing
Library preparation
DNA
Addition of sequencing adaptors
High-throughput sequencing
Paired-end short-reads
Read mapping and SNP detection
TCAAGGGTCCCCCCGAGAGNNNTGTCAGTCAGTCANNNNNNAGCAGAAGCGCAGT
Reference genome
Aligned reads
1
2
3
1
4
SNPs detected

FIGURE 1 Whole-genome sequencing. (a) De novo whole-genome assembly consists on sequencing and assembling a species complete genome for the first time. First, high-quality genomic DNA is fragmented for library preparation that involves addition of sequencing adaptors to DNA fragments. Paired-end short reads (~100 bp) are obtained using high-throughput sequencing from libraries with different insert sizes to maximize coverage of the genome (standard libraries: ~350–550 bp, mate-pair libraries: ~2–20 kilobases (Kbp), fosmid-end libraries: ~40 Kb, not shown). Long-read sequences (~2–10 Kb long) can also complement the sequence pool. (1) Read alignment starts with the building of local contigs (i.e., sequence formed by overlapping DNA fragments). (*) Repetitive regions are difficult to assemble with short reads. (2) Mate-pair reads can help orient and link contigs, building larger sequence stretches called scaffolds (or supercontigs). Gaps of unknown sequence are denoted with "Ns." (3) Long reads can help in the assembly of repetitive regions. The final product of genome assembly is a consensus sequence often corresponding to a series of contiguous scaffolds separated by gaps. (b) Whole-genome resequencing compares variable sites between the genomes of individuals or populations and requires the species genome sequence for read mapping. This image shows an example for one individual and using short-read sequencing. High-quality genomic DNA of an individual is fragmented for library preparation that adds sequencing adaptors to the DNA fragments and has an insert size of ~350–500 bp. Paired-end short reads (~100 bp) are obtained from the DNA library using high-throughput sequencing. Short reads are mapped onto the species reference genome based on sequence similarity. A SNP is detected when the specific base observed in a position in the reference genome differs from the base observed in the reads. Notice the uneven read coverage for some positions. (1) Variant sites correspond to a base change present only in the subject reads but not in the reference genome, (2) some SNPs may be lost because they are absent in the reference genome, (3) some SNPs may be heterozygous in the subject reads, and (4) others may be lost during filtering, e.g. because of low coverage. The final result is a file that contains the variable sites of the individual. In this image, paired-end reads are represented by a rectangular shape with bases at both ends but not in the middle. The middle part is in grey and corresponds to unknown sequence in between paired reads. Figures 1–5 were created using the free software Inkscape (https://inkscape.org/en/)

and "synthetic long-read" sequencing (SLR-seq; i.e., Illumina synthetic long reads, 10× Genomics). SMRT-seq methods produce long reads of single DNA molecules, whereas the synthetic approaches do not; in the latter, long sequences are computationally assembled from barcoded short reads coming from the same DNA molecule (Goodwin et al., 2016; Lee et al., 2016; Bleidorn, 2016). Throughput, error rate and cost vary between long-read approaches. For example, within SMRT-seq methods, throughput of PacBio is lower than any second-generation technique but nanopore's sequencing competes with Illumina HiSeqX. In contrast, the throughput of SLR-seq methods is the same as in Illumina systems. Error rate is much higher in SMRT-seq methods than in any second-generation technique (15%–20% in PacBio, although 99,99% accuracy can be achieved with ~50× coverage (Berlin et al., 2015); 30%–40% in nanopore-sequencers). In SLR-seq, error rate is the same as in Illumina. In terms of cost, PacBio is pricey (~USD$ 1000/Giga byte [GB]; Goodwin et al., 2016) and nanopore-sequencing promises low cost (~USD$ 20/hr) although it has not yet been disclosed (Bleidorn, 2016). In SLR-seq methods, cost can be high as it includes short-read Illumina sequencing to very high coverage (~1,000× (Lee et al., 2016)) and library preparation that incorporates barcodes. In 10× Genomics additional equipment is required. Genome assembly can be achieved with only long-read data (>50×), or with a combination of long- and short reads. Long reads help resolve complex stretches of the genome (i.e., repetitive sequences and SVs) that are poorly handled by short reads, significantly improving quality of genome assembly. With long reads, it is also possible to sequence entire transcripts, which promise enhancing metabarcoding and metagenomics studies, and perform direct haplotype assignment of genomes (Bleidorn, 2016) that otherwise can only be limitedly inferred from population-level short-read data (Snyder et al., 2015). The composition and extension of haplotypes constitute valuable information for many analyses including, demographic history, linkage disequilibrium, genomewide association studies (GWAS), genealogical tracing of mutations and allele-specific expression, among others (Browning & Browning, 2011; Snyder et al., 2015; Lee et al., 2016). For a more detailed description of sequencing techniques, see Goodwin et al. (2016); Lee et al. (2016); Bleidorn (2016).

2.1 | Comparison of whole-genome resequencing and de novo sequencing

Whole-genome sequencing can be classified in two categories (Figure 1): (i) de novo whole-genome sequencing (WGS); and (ii) whole-genome resequencing (WGR). (i) The goal of WGS is the assembly of a genome sequence for the first time. This can be a demanding task depending on size and genome complexity, desired level of completeness, computing resources and bioinformatics experience. Programming skills and understanding of assembly algorithms are fundamental for optimal results (Ekblom & Wolf, 2014). (ii) The objective of WGR is instead, to compare genomic variability among individuals or populations. This approach requires previous availability of the reference genome for read mapping and variant identification. Sequencing is dedicated to obtaining reads from the genome of individuals or populations to a particular coverage depending on the application.

The absence of the reference genome of the species of interest likely constitutes the main limitation faced by conservation geneticists when implementing a WGR approach, justifying the use of the genome sequence of a closely related species (Lamichhaney et al., 2012; Dennenmoser, Vamosi, Nolte, & Rogers, 2017). Caution is however warranted with this procedure as differences in genomic organization (e.g., copy number variation, structural variants) can exist, even between closely related species (Ekblom & Wolf, 2014). The use of a reference genome of another species restricts the mapping of short reads to conserved regions between the two taxa. The power of WGR could be diminished as potentially informative variation present uniquely in the focal species is likely to be missed following this procedure. Additionally, the genomic differences between taxa can affect the accuracy of both, read mapping and SNP calling (Nevado, Ramos-Onsins, & Perez-Enciso, 2014). Thus, when possible, it is preferable to use the genome of the focal

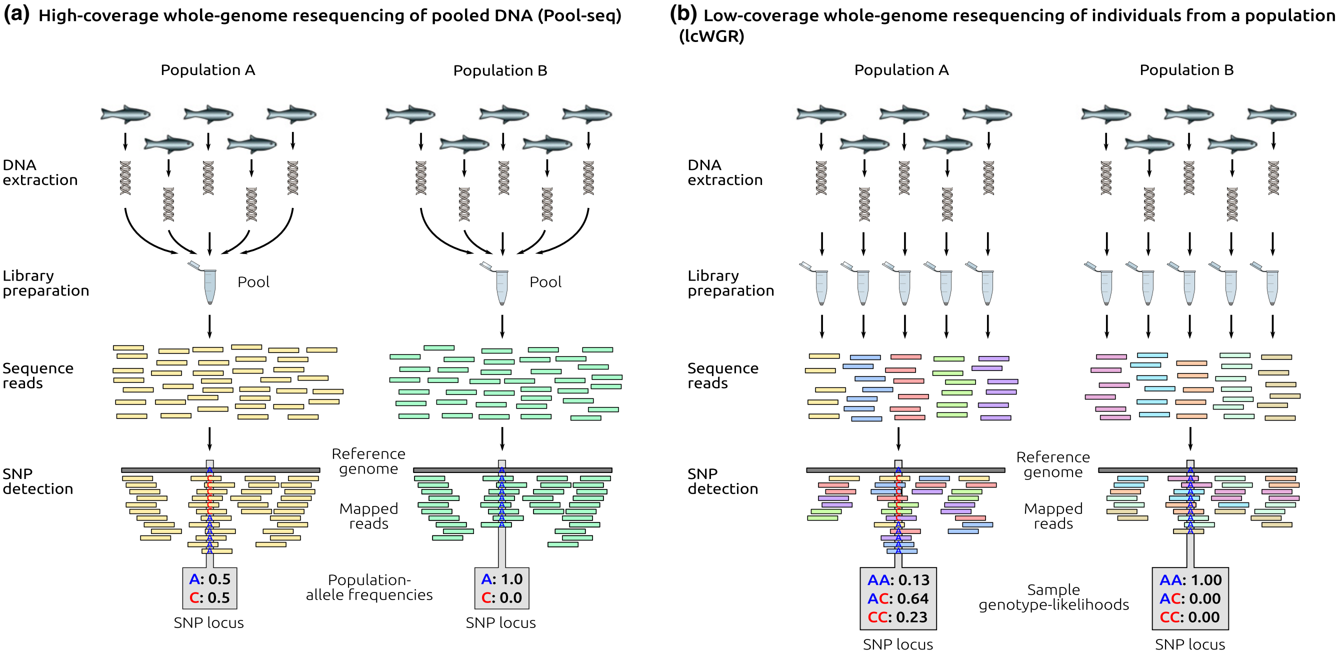

(a) High-coverage whole-genome resequencing of pooled DNA (Pool-seq)
(b) Low-coverage whole-genome resequencing of individuals from a population (lcWGR)
Population A
Population B
DNA extraction
Library preparation
Pool
Sequence reads
SNP detection
Reference genome
Mapped reads
A: 0.5
C: 0.5
A: 1.0
C: 0.0
Population-allele frequencies
SNP locus
AA: 0.13
AC: 0.64
CC: 0.23
AA: 1.00
AC: 0.00
CC: 0.00
Sample genotype-likelihoods

FIGURE 2 Data acquisition in current population-based WGR methods. (a) Pool-seq starts with mixing in a single tube equimolar amount of DNA of several individuals from a population. An aliquot of the DNA pool is used for sequencing library preparation and a single barcode is assigned to each population. Barcodes are represented by different colours in the sequence reads, yellow for population A and green for population B. The pooled-DNA library is sequenced to a high depth of coverage (>50×). SNP detection and population-allele frequency estimation require the mapping of reads to the reference genome and are based on sequence read coverage. Allele frequency differences between populations are then detected from allele read counts for a given polymorphic site. (b) lcWGR, starts with the preparation of a single sequencing library per individual, each with its own barcode (represented as ten different colours of short read). Individual DNA libraries are sequenced to a low depth of coverage (~1–4×). Read mapping to the reference genome is required for SNP detection and sample genotype likelihoods calculation, which are based on the alleles present in the individual reads supporting a variant site

BOX 2 General workflow for whole-genome resequencing data acquisition

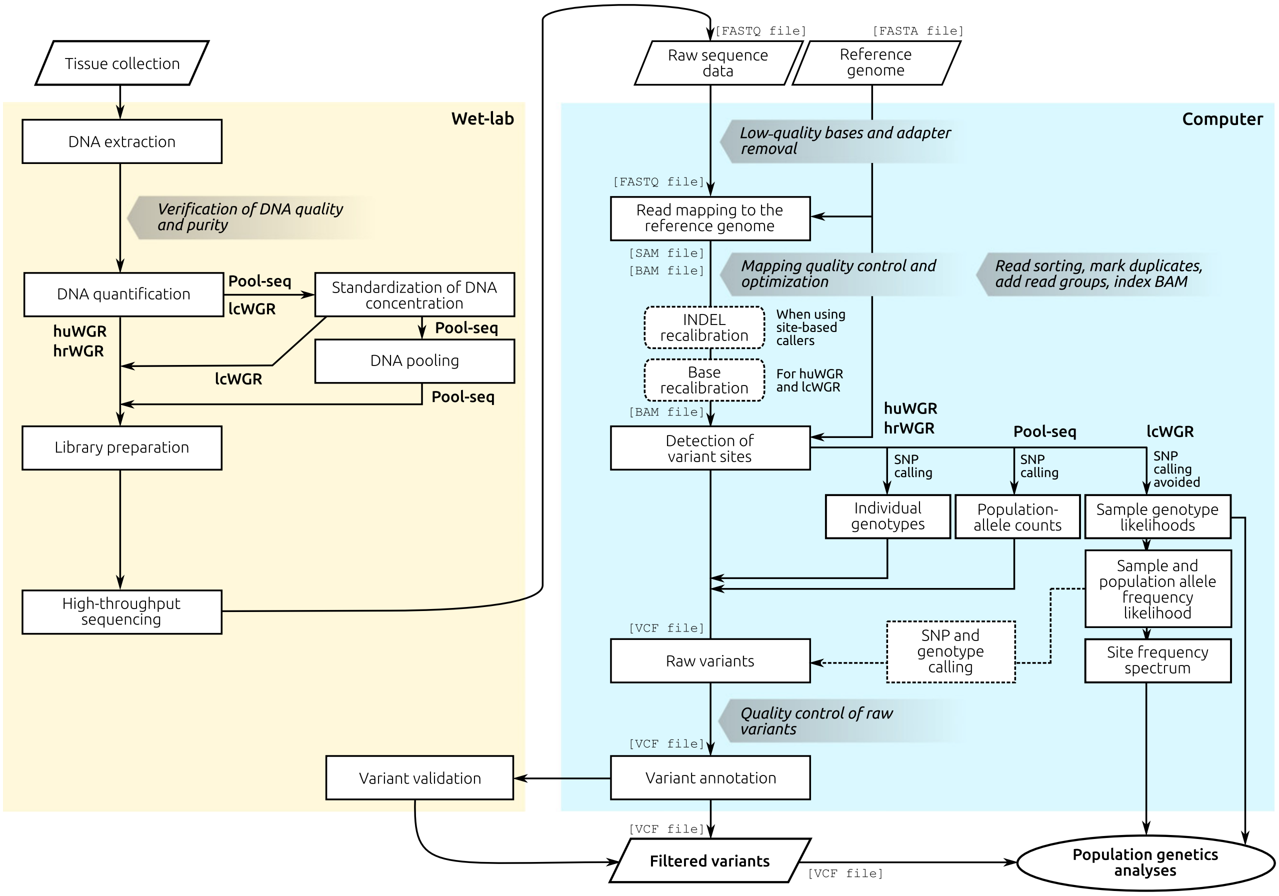


FIGURE B2 Schematic illustration of the general WGR workflow. The four WGR approaches shown are as follows: the sequencing of individuals to a high depth of coverage with either unresolved (huWGR) or resolved haplotypes (hrWGR), the sequencing of population genomes to a high depth by mixing equimolar amounts of unlabelled-individual DNA (Pool-seq) and the sequencing of multiple individuals from a population to a low depth (lcWGR. Dashed lines denote optional steps, shadowed arrows indicate quality check point steps, and file format is shown within brackets.)

Wet-laboratory procedures

Tissue sampling and preservation

Damage of DNA should be avoided. Example of good practices for tissue collection and preservation is presented in Wong et al., (2012).

DNA extraction, quality and quantity

DNA quality is assessed with ~0.8%–1% agarose gel electrophoresis and a 25-Kbp molecular weight ladder. A single high molecular weight band (~23 Kbp) indicates good DNA integrity. High DNA purity is confirmed with a 260/280 nm absorbance ratio of ~1.8–2.0. Highly fragmented DNA should be avoided as it cannot be accurately quantified using fluorometric-based methods (recommended for precise double-strand DNA quantification) (Sedlackova, Repiska, Celec, Szemes, & Minarik, 2013). For Pool-seq this is particularly important as the even contribution of individual DNA in a pool relies on accurate quantification. The amount of starting DNA depends on the library preparation kits' input requirements described in Table 1.

Standardization of DNA concentration across samples (for Pool-seq and lcWGR)

Each DNA sample is diluted or concentrated to a desired standard value (ng/μl). The diluting liquid should stabilize and protect DNA from damage (e.g., lowTE). A liquid handling robot is recommended for this step to eliminate the potential for pipetting error (Figure B2).

BOX 2 Continued

DNA pooling (Pool-seq)

Pooling consists on mixing equimolar amounts of DNA of several individuals from a population. When the interest is to identify the genetic basis of a trait, pools should comprise individuals sharing the same trait (not necessarily from the same population) and extreme trait categories have increased potential to lead to clearer genetic signals. A minimum of 50 individuals is recommended per pool, but including more (>100) (assuming proportional increase in sequencing depth) can help minimize slight unevenness in the representation of few individuals leading to more accurate allele frequency estimates (Gautier et al., 2013; Schlötterer et al., 2014). Individual DNA is then diluted to a standard concentration and verified through a quantification step. Once normalized, the same amount of DNA from individual samples can be pooled into a single tube.

Sequencing library preparation

Several kits for library preparation are available commercially. They differ in cost per sample, need of a sonicator, incorporation of a PCR step and amount of input DNA. For current price and DNA input requirements of Illumina kits, see Table 1. DNA amplification with PCR is convenient when low DNA amounts are available, but PCR can introduce biases (e.g., under-representation of GC-rich fragments, preferential amplification of short fragments and duplicates) that can lead to uneven coverage in some loci. Some of these biases can be minimized by making PCR protocol adjustments (Aird et al., 2011) (e.g., using as few PCR cycles as possible, typically 6–8) and by removing duplicates in silico using Picard tools, http://broadinstitute.github.io/picard, or SAMTOOLS (Li, Handsaker, et al., 2009). Small structural variants (INDELs and CNVs) are detected from short reads of standard libraries (~350–550 bp insert size), whereas the detection of large structural variants (spanning Mbs) may require the use of mate-pair libraries (~2–20 Kb insert size) or long-read data. Additional considerations are discussed in Head et al. (2014).

High-throughput sequencing of DNA libraries

Currently, the most popular technology for short read high-throughput sequencing is Illumina, although new technologies are being developed (Goodwin et al., 2016). Illumina offers an overall accuracy >99.5%, which is high relative to other platforms but still restrictive as it is difficult to distinguish true genetic variation from technical artefacts (Laehnemann, Borkhardt, & McHardy, 2016). The suggested minimum coverage for huWGR is >30×/individual (Sims, Sudbery, Ilott, Heger, & Ponting, 2014), and for Pool-seq, it is >50×/pool (Schlötterer et al., 2014), although a much higher coverage should be targeted (>100–200×) for rare allele detection (Wang, Skoog, et al., 2016) and for lcWGR it is 1–4×/individual (Nielsen et al., 2011; Buerkle & Gompert, 2013). The number of Illumina lanes needed depends on the trade-off between genome size, target coverage per sample/pool and flow-cell yield. Illumina sequencing is potentially prone to lane-to-lane variation (Ross, Russ, & Costello, 2013), a problem that can be minimized by distributing barcoded libraries across multiple lanes (TCAG DNA sequencing facility pers. comm.).

Computer procedures

Quality control of raw sequences

Raw sequence data come from the sequencer in FASTQ format (Cock, Fields, Goto, Heuer, & Rice, 2009). Sequence quality and overrepresented sequences can be assessed with FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Quality control typically consists in the removal of low-quality bases (PHRED quality score <20) and adapter sequences using programs like TRIMMOMATIC (Bolger, Lohse, & Usadel, 2014) or CUTADAPT (Martin, 2011).

Read mapping to a reference genome

High-quality reads are mapped to a genome based on sequence similarity. Multiple algorithms for short-read mapping exist and have been reviewed elsewhere (Fonseca, Rung, Brazma, & Marioni, 2012; Hatem, Bozdağ, & Çatalyürek, 2013; Reinert et al., 2015; Ye, Meehan, Tong, & Hong, 2015). Some of the most commonly used free aligners are BWA (Li & Durbin 2009, 2010; Li, 2013) (Table 2) and Bowtie2 (Langmead & Salzberg, 2012). Alignment artefacts could arise due to multiple factors, including misalignments around INDELs and divergence between the subject reads and the reference genome. It is thus important to understand how the various algorithms work to make informed decisions on how to optimize running parameters (see Box 3). The final product of read mapping is a SAM (Sequence Alignment/Map) file (several Gb in size), format that contains a line for each read and fields with associated information including read position and mapping quality score (MAPQ or MQ) (Li, Ruan, & Durbin, 2008) that can be used for SNP filtering. A BAM file, the compressed light-weighted binary version of the SAM file, is obtained using Picard tools, and it is the format commonly preferred as input file by other programs. Read sorting, marking of duplicates, addition of read groups and indexing are additional steps to prepare the BAM files for variant calling (Van der Auwera et al., 2013).

BOX 2 Continued

Quality control of mapped reads

A visual exploration of the BAM file with the Integrative Genomics Viewer (IGV) (Robinson et al., 2011; Thorvaldsdóttir, Robinson, & Mesirov, 2013) can help identify regions with extremely high or low coverage, strand bias, misalignments around INDELs and repetitive regions, among others. As alignment errors can occur, it is important to verify that reads mapped evenly and correctly to minimize false variant calls. Evaluation of mapping quality can be complemented with summary statistics (e.g., average depth of coverage; insert size distribution, and number of mapped reads, properly paired reads, singletons and ambiguous mappings) that can be obtained with samtools, ea-utils (http://expressionanalysis.github.io/ea-utils/), or qualimap (Okonechnikov, Conesa, & García-Alcalde, 2015).

Indel realignment (optional, depending on SNP caller)

Punctual mapping artefacts around INDELs may not be resolved by optimizing global mapping parameters. Local INDEL realignments are a necessary prerequisite when using a site-based SNP calling algorithm like samtools (Li, Handsaker, et al., 2009) or gatk-unifiedgenotyper (McKenna et al., 2010). This step is not needed when using haplotype-based callers like freebayes (Garrison & Marth, 2012) or the gatk-haplotypecaller (http://gatkforums.broadinstitute.org/gatk/discussion/7847). INDEL realignment can be done with specific functions in gatk (McKenna et al., 2010) (tutorial: https://software.broadinstitute.org/gatk/guide/article?id=7156). A file with known INDELs can help defining targets for realignment (Van der Auwera et al., 2013), but in its absence, INDELs identified during read mapping can be used instead (default mode)(https://software.broadinstitute.org/gatk/events/slides/1504/GATKwr7-X-3-Non_human.pdf).

Base recalibration (optional but recommended)

Per-base quality scores obtained from sequencers often present errors. Because SNP calling and genotype likelihood algorithms consider such quality scores, they should be corrected. This can be achieved using the base quality score recalibration (BQSR) package implemented in gatk (DePristo et al., 2011; Van der Auwera et al., 2013). A known set of variants is required, but in its absence, an iterative bootstrapping approach can be followed instead (Tung, Zhou, Alberts, Stephens, & Gilad, 2015; Snyder-Mackler et al., 2016).

Detection of variant sites

Specific software exists for the detection of the different types of genetic variants (i.e., SNPs and INDELs, SVs, and CNVs). Such algorithms implement particular models of variation and sources of information for the discovery of polymorphisms from short-read data. Variant positions are detected differently in hrWGR, huWGR, Pool-seq and lcWGR data. In the first three, polymorphic site detection is based on per-site read coverage and quality per individual or population, whereas in the latter, it is based on coverage and quality of all the reads covering a site from several individuals in a given sample. SNPs are not called in lcWGR, instead, per-site genotype likelihoods are calculated using software like angsd (Korneliussen et al., 2014). In hrWGR, huWGR and Pool-seq, SNPs are called using software like gatk-haplotypecaller, samtools, or freebayes (Table 2). A comprehensive review of SNP calling using NGS data can be found in Nielsen et al. (2011) and (2012), and for structural variants in Alkan et al. (2011). Each SNP calling algorithm makes a series of assumptions that can lead to different results. Thus, a good practice is to compare the SNPs detected by at least two algorithms (O'Rawe et al., 2013). The product of variant calling is a VCF (Variant Call Format) file containing raw polymorphisms and annotations (Danecek et al., 2011).

The selection of a SNP calling algorithm for Pool-seq data requires consideration of whether it handles ploidies larger than 2. In theory, *Pool ploidy* = *Ploidy per individual* × *Number of individuals*. Assuming 50 diploid individuals are mixed, pool ploidy is 100. Such large ploidies, however, deplete system memory and multiply runtime (in gatk-haplotypecaller https://software.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php, and freebayes https://github.com/ekg/freebayes/commit/576bc703c246035342538a0feeecd1, accessed June 2017).

Using the default ploidy (2) leads the software to call only the 2 most common alleles in a pool, as ploidy assumes 50/50 allele frequency (http://gatkforums.broadinstitute.org/gatk/discussion/6551/what-if-ploidy-is-set-to-2-for-pooled-dna-sequencing-experiment, accessed June 2017). This might not be an issue when calling SNPs among closely related samples as SNPs are considered biallelic, but it would limit the overall number of alleles detected when comparing more distantly related samples. Use of large ploidies is now partially solved by establishing the maximum number of alternative alleles to be considered. In gatk v.3.7 this can be set with the flag *–maxGenotypeCnt* (https://software.broadinstitute.org/gatk/blog?id=8692), and in FreeBayes with *–use-best-n-alleles* and setting a pooled mode (*–pooled-discrete* or *–pooled-continuous*). These settings make the algorithms run faster at the expense of missing low-frequency alleles in multi-allelic loci (https://github.com/ekg/freebayes).

BOX 2 Continued

Quality control of raw variants

SNPs with low support should be removed from the final data set as most likely are false calls. This can be achieved either using variant quality score recalibration (VQSR) or applying hard filters. VQSR is generally preferred as it is an unbiased filtering based on a large number of validated variants that train the algorithm (Van der Auwera et al., 2013). Hard filters are usually applied in the absence of known variants and include SNPs removal based on annotation parameters assigned to each SNP during read mapping and variant calling. Common filters include low complexity, maximum depth, allele balance, double-strand, Fisher strand and quality filter (Van der Auwera et al., 2013; Li & Wren, 2014), as well as mapping quality (MQ) (Li et al., 2008). Each mapping algorithm calculates the MQ score differently (Ruffalo, Koyutürk, Ray, & LaFramboise, 2012) for which scores should not be compared between programs. The application of hard filters, however, can bias the Site Frequency Spectrum by excluding low-frequency variants and is limited by the absence of guidelines to select which annotations or cut-off values should be applied to a given data. The appropriate choice of cut-off values is a function of the data. The recommendation is thus to test different parameter combinations and thresholds to optimize these filters. This forum, https://software.broadinstitute.org/gatk/guide/article?id=6925, can offer some insight on hard filtering using GATK. Additionally, SNPs within low-complexity regions should be removed as these regions are troublesome for read mapping and SNP calling (Li & Wren, 2014). The final VCF file after quality control is then ready for downstream analyses.

Variant annotation

Sequence ontology terms can be annotated to variants in a VCF file using for instance, VCFANNO (Pedersen et al., 2016), ANNOVAR (Yang & Wang, 2015) or SNPEFF (Cingolani et al., 2012) programs.

Variant validation

Variants detected from WGR data should be treated as putative polymorphisms, especially in Pool-seq and lcWGR. SNP genotyping PCR-based methods can be used for SNP validation. PCR amplification and Sanger sequencing can be used for SVs validation.

For additional guidelines on how to obtain high-quality variants from high-throughput sequencing data see Van der Auwera et al. (2013) and Pfeifer (2017).

species for WGR analysis, unless the research question can be addressed examining conserved regions alone. A brief overview of genome assembly guidelines and completeness status of genomes sequenced to date is provided in Box 1.

The steady decrease in sequencing cost promised by new technologies suggests that access to the genome sequence may in the near future no longer be an unsurmountable obstacle for nonmodel species (Goodwin et al., 2016). Proof of this are the multiple international initiatives that are collaboratively sequencing genomes of various taxa including fungi (Grigoriev et al., 2014), invertebrates (GIGA 2014), arthropods (Evans et al., 2013), birds (Zhang, 2015), fishes (Macqueen et al., (2017); Malmstrøm, Matschiner, Tørresen, Jakobsen, & Jentoft, 2017), mammals (Fontanesi et al., 2016), vertebrates (Koepfli, Paten, & O'Brien, 2015), among others.

2.2 | Comparison of whole-genome resequencing approaches for population genomics

A population genomics study can be based on the analysis of individual sequence data (individual-based approaches) or on the analysis of the sequences of a group of individuals as a whole (population-based approaches). In individual-based approaches, the goal is obtaining high-quality individual genotypes, required for analysis on population demographic history and N_e estimation, and genealogical tracing of mutations, among others. There are currently two techniques: (i) high-coverage haplotype-unresolved individual WGR (huWGR) and (ii) high-coverage haplotype-resolved individual WGR (hrWGR). In both techniques, high read depth (>30–50× depth) is targeted for achieving accurate SNP, short INDEL (>50 bp) and genotype calling, as multiple reads (observations) help distinguish true variation from sequencing error (Nagasaki et al., 2015). (i) In huWGR, short-read data per individual results in unphased individual genotypes that are used for subsequent analyses. If haplotype information is required, phasing can be indirectly achieved using statistical methods that rely on genotypes of several related or unrelated individuals. Such methods are then limited by the need for large sample sizes and by the extend of linkage disequilibrium blocks that vary across the genome (Browning & Browning, 2011). (ii) In hrWGR, the goal is to directly obtain haplotype-resolved genomes of single individuals using specific experimental procedures and short- and/or long-read sequencing, implying large sample sizes are not required (reviewed in Snyder et al., 2015).

In population-based approaches, the goal is obtaining population-level genomic data (e.g., allele frequencies or genotype likelihoods) from several individuals analysed as a whole and sequenced to a high or a low depth (Buerkle & Gompert, 2013), to offset the cost of obtaining high-coverage per individual. Such population-level data can be used for inference of population structuring, detection of outlier loci and signatures of selection, among others. Two methods can be identified (Figure 2): (i) Pool-seq, or the sequencing at a high coverage

BOX 3 Pool-seq: Limitations, sources of error and bias and potential solutions

Individual genotypes are missed

Pool-seq has been used for the estimation of genetic differences among populations, the detection of outlier loci, and for mapping genetic variation underlying phenotypic traits (see Table 2). This method, however, prevents the identification of individual reads as only one barcoded library is prepared for the pooled DNA of individuals (Schlötterer et al., 2014). This has four important implications: (i) errors during library preparation or shotgun sequencing that could affect the homogeneous contribution of individual DNA to the final data set cannot be detected; (ii) as individual genotypes are lost, presence of migrants in the sample cannot be evaluated; (iii) individual haplotypes and linkage disequilibrium (LD) cannot be assessed (e.g., the LD method for the estimation of effective population size cannot be used); (iv) only total allele frequencies can be calculated for a given pooled DNA. These factors can bias the population-level allele frequency estimates and limit population genetics analyses, leading to concern about the suitability of the Pool-seq method for population genomic studies (Anderson et al., 2014; Therkildsen & Palumbi, 2017).

Several ways to mitigate these concerns have been proposed. Knowledge of population structure can help avoid accidentally pooling individuals of different origin. Mixed aggregations, however, can also be indirectly detected from Pool-seq data, and conveniently excluded of posterior analysis if required, as they exhibit extremely large branch lengths in a phylogenetic tree based on pairwise genetic distances (see Lamichhaney et al. (2017) Figures 2 and 5: WFB location). In the absence of prior information on population structure, ecological and biological knowledge (e.g., timing of reproduction, maturity stage at time of sampling, size and age) can help infer the local origin of individuals. Uneven representation of individual DNA samples in the final pool can be minimized following stringent laboratory procedures (Box 2). The limitations above concerning the lack of haplotype and hence individual genotypes and linkage disequilibrium information can be overcome by the selection of a subset of informative SNPs identified from the Pool-seq data and the subsequent genotyping of a number of individuals per population at these informative SNPs using PCR-based or NGS-based genotyping methods. This SNP validation step should be performed anyways as part of a good practice protocol for the Pool-seq workflow (Box 2).

Allele frequency estimation is susceptible to multiple factors

A main weakness of Pool-seq is the potential uneven distribution of reads across the genome due to technical and computational artefacts (Anderson et al., 2014). Depth of coverage is the basis for variant detection where read counts are used for the estimation of allele frequencies in individual SNPs. Therefore, uneven coverage not resulting from biological processes can greatly increase false SNP calls or leave out informative SNPs, biasing downstream analysis and interpretation (Sims et al., 2014). Read depth of coverage can be affected by several factors including (i) duplicates and GC bias produced by PCR during library preparation (Sims et al., 2014); (ii) amplification bias of NGS sequencing technologies (Goodwin et al., 2016); (iii) incompleteness of the reference genome especially at repetitive regions that could produce misalignments and false calls (Li & Wren, 2014); (iv) structural variations (e.g., CNVs, chromosomal inversions, transposable elements) can inflate allele frequency estimates (Schlötterer et al., 2014); and (v) poor read mapping can result in read misplacement or missing of divergent reads (Kofler, Orozco-terWengel et al. 2011; Kofler et al. 2016; Schlötterer et al., 2014). Factor (i) can be controlled using PCR-free sequencing library preparation kits when plenty DNA is available; if this is not the case, then PCR duplicates can be easily removed using bioinformatic tools. Factor (ii) is out of researchers' control but to minimize lane-to-lane variation, DNA libraries should be spread in different lanes (Ross et al., 2013). Factor (iii) can be minimized by including an indel realignment step before variant detection when using site-based SNP callers, by using haplotype-based callers that incorporate indel realignment, (Van der Auwera et al., 2013), or by excluding variants detected in and around INDELs (Kofler et al., 2011) and repetitive regions. The effect of factor (iv) can be minimized by excluding structural variants during variant calling (Kofler, Orozco-terWengel et al., 2011), although this type of variation should be included in genome scans for the detection of outlier loci or the characterization of the genetic basis of traits. And factor (v) requires the optimization of read mapping by selecting an adequate algorithm and running parameters that minimize misalignments.

Mapping short reads against a reference sequence is a difficult task considering the large number of reads to process and the computational challenge of determining their exact position in the genome. For instance, repetitive sequences are difficult to map because their sequence can be present in multiple positions, and the genome could also have assembly errors. Moreover, there may be sequence differences between subject reads and the reference genome, because the latter is usually not representative of the species genetic diversity (usually a genome sequence is built from DNA of one or a few individuals) (reviewed by Treangen and Salzberg (2011), Phan, Gao, Tran, and Vo (2014), Laehnemann et al. (2016)). Thus, the optimal choice of a mapping algorithm depends on the sequence data structure (e.g., repetitive content) and degree of divergence between the reads and the reference genome. Multiple mapping algorithms exist which are optimized for different levels of sequence divergence. For example, BWA excels in the mapping of reads with low divergence (<2%) (Li, 2013), however, the running parameters can be modified for mapping divergent reads (Kofler,

BOX 3 Continued

Orozco-terWengel et al., 2011). In contrast, Stampy is better for mapping reads with high-divergence as the user can input a substitution rate value (Lunter & Goodson, 2011). Default settings of mapping algorithms are usually optimized for specific data sets and highly curated genomes (e.g., humans), thus running a mapper with default parameters may not be ideal for all data sets as this can produce multiple spurious outlier loci and false positive SNPs (Kofler, Langmuller, Nouhaud, Otte, & Schlotterer, 2016). Time investment into optimizing read mapping parameters before SNP calling is thus recommended. Unfortunately there is no golden rule applicable to all data sets but, in general, the idea is to experiment with parameters (e.g., mismatch and gap opening penalties), conduct read mapping, compare alignment statistics, and so on, following a multidimensional optimization process. The ideal parameter combination will maximize the number of properly paired-end reads while minimizing presence of discordant mates, singletons and ambiguous mappings. This rationale is explained in detail for RAD-seq data by Jonathan Puritz (https://github.com/jpuritz/Winter.School2017/blob/master/Exercises/Day%201/Mapping%20Exercise.md).

Despite the various factors potentially affecting depth of coverage in Pool-seq data, numerous studies have demonstrated the method can produce reliable population-level allele frequency estimates (Fracassetti et al., 2015; Wang, Skoog, et al., 2016; Martinez Barrio et al., 2016; Lamichhaney et al., 2017).

Rare and low-frequency variants are hard to detect

It is generally assumed that Pool-seq only allows for the discovery of common variants of large effect as different factors can affect read coverage, making it difficult to distinguish low-frequency variants from sequencing error. Recent studies by Wang, Skoog, et al. (2016), however, challenge this idea as they recovered rare variants from high-depth Pool-seq data of Bull Terrier dogs (average 130×) and humans (average 150×). Minor allele frequency errors were evaluated using three variant calling programs, SAMTOOLS, GATK (ploidy setting) and Freebayes (ploidy setting). A good proportion of rare SNPs identified from the pooled data were validated through individual genotyping of several samples using the MassARRAY System (Agena Bioscience, US) and the Illumina SNP array (Illumina, US) systems. Thus, Pool-seq can be a fast and affordable initial approach for the assessment of rare variants in large-scale association studies.

BOX 4 lcWGR: Limitations, sources of error and bias and potential solutions

Genotype likelihood (GL) values can vary

Genotype likelihoods are the foundation of the statistical framework for population genetics inference from low-coverage sequencing data (Nielsen et al. 2011, 2012; Buerkle & Gompert, 2013). GLs and additional analyses for this type of data are implemented in the programs ANGSD (Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012; Korneliussen et al., 2014) and NGSTOOLS (Fumagalli et al., 2014). GLs per polymorphic site are estimated based on the observed sequence reads covering the site from a given sample of individuals, and the reads' PHRED quality scores. The base-error rate estimation method differs among GL models and error rates can be fixed or estimated from the quality scores or the sequence data. The 4 models for GLs and the 2 models for base-error rate calculations implemented in ANGSD are described by Korneliussen et al. (2014). Previous studies indicate the GL models can generate different results in some circumstances (Korneliussen et al., 2014). Thus, the choice of model can potentially introduce bias (http://www.popgen.dk/angsd/index.php/Genotype_Likelihoods) (Korneliussen et al., 2014) but models have not been compared nor, to our knowledge, has the procedure for model selection been discussed in the literature thus far. The base-error rate methods differ in how they model the error structures in the data, which is important in the calculation of GLs and downstream analyses. If the modelling misses error sources; then, the GLs are likely to be biased and verification from mathematical derivations that the proper error structure in the data has been correctly incorporated is not straightforward.

Genotype likelihoods can be affected by (i) accuracy of base-calling and quality score (Fumagalli, Vieira, & Korneliussen, 2013), (ii) read coverage distribution and filtering, (iii) sample size and individuals included in the sample, (iv) how accurately model assumptions are met, including (v) the assumption that markers are diallelic and organisms are diploid.

(i) Sequencing error is extensive in available sequencing platforms (Goodwin et al., 2016), and the per-base quality scores obtained directly from the sequencing machines also have errors. Therefore, lcWGR data should be obtained using sequencing platforms that offer the lowest error possible, and the per-base quality scores need to be recalibrated before data analysis. It is also important to

BOX 4 Continued

consider that the GL calculation assumes independence among reads, which may be violated in the presence of alignment error or PCR artefacts (Nielsen et al., 2011).

(ii) It is currently almost impossible to obtain an even read coverage distribution across the genome and among individuals. This is a result of the sequencing chemistry itself (usually following a Poisson distribution) and of laboratory procedures that can skew the representation of individual DNA samples added to a flow cell (errors in pipetting or DNA quantification, and variability in fragment sizes during library preparation). In addition, a very low sequencing depth per individual (<2×) could also limit the possibility of sequencing both alleles in a diploid organism (Fumagalli, 2013). Similarly, the joint effect of read filtering (including in silico coverage cut-offs across individuals in a sample), reference genome quality and completeness, and read mapping can also bias the individual read representation for a particular locus in the final data set. Thus, a varying proportion of missing data per individual could be expected in lcWGR data sets, which implies polymorphic sites are covered by reads from a varying proportion of individuals in a sample. This could be a problem when reads of a small number of individuals are supporting a particular site because they may not be representative, potentially biasing GL estimates and downstream analyses. An excess of missing data can also bring convergence problems that impact the accuracy of many calculations including the individual admixture analysis implemented in NGSADMIX (Skotte, Korneliussen, & Albrechtsen, 2013). In GWAS in humans, missing genotypes in extremely low-coverage sequencing data (0.1–0.5×/sample for 909 individuals) have been treated using imputation methods that rely on the availability of a set of known haplotypes (Pasaniuc et al., 2012). For nonmodel species, such set is commonly absent, implying that imputation is not an option, unless individuals are highly inbred (Wang, Xu, et al., 2016). In conclusion, lcWGR studies require a relatively even read coverage distribution among individuals (see Therkildsen & Palumbi, 2017). The removal of sites with large amounts of missing data (>80%) (Skotte et al., 2013) and a target read depth that assures both alleles are sequenced in a diploid organism (>2×) are recommended (Fumagalli, 2013).

Similarly, (iii) sample size and the actual individuals included in a sample determine the alleles and sites assessed as well as the population structure and admixture estimates. Simulation studies have shown good accuracy in population genetics parameters using a low depth per individual (~1–2×) as long as the number of individuals in the sample is large (Fumagalli, 2013; Buerkle & Gompert, 2013). A minimum number of individuals per sample or population has, however, not been proposed. Presumably this is because such number depends on several factors including the species genomic architecture (e.g., LD decay, recombination rate, mutation rate) and effective population size, and the funding available, among other variables. However, it would be useful to have at least some reference obtained from real data that can be used in the design of sampling programs. The genetic makeup of individuals composing a sample would determine the alleles and sites evaluated, therefore it is important to ensure individuals are not closely related to avoid inflated estimates of genetic differentiation. When comparing populations, it is fundamental to verify the degree of individual admixture before performing GLs calculation as the presence of mixed individuals will likely bias the alleles represented and thus, subsequent analyses and interpretation. Individual admixture can be verified using the program NGSADMIX (Skotte et al., 2013).

(iv) The accuracy of GLs and subsequent metrics depend on the fulfilment of the assumptions made in the mathematical models, for instance, independence among reads for GLs calculation (Nielsen et al., 2011), independence among sites and Hardy–Weinberg equilibrium for calculation of the likelihood function for the site frequency spectrum (Nielsen et al., 2012), and independence among individuals for the estimation of allele frequencies (Kim et al., 2011).

Finally, (v) the current four GL models were developed for diallelic markers in diploid organisms implying that lcWGR cannot currently be applied to nondiploid species or pooled-DNA data. The method is thus limited to the assessment of genetic variation in SNP loci; INDELs are included in the models but not used for posterior analyses (Korneliussen et al., 2014). Putative structural variants could be detected, however, from high-density SNPs as they facilitate the identification of sweeps to a fixed allele, as observed in a cryptic subgroup of *Anopheles gambiae* s.l. (GOUNDRY) where ~500 SNPs allowed the detection of a putative large inversion (1.67-Mb) on the X chromosome (Crawford et al., 2016).

A few programs accept GLs

Many traditional population genetics software packages require individual genotypes as input data, limiting the possibility of use for the analysis of lcWGR data. To solve this problem, the software ANGSD includes several genotype callers (Korneliussen et al., 2014).

It is a relatively new approach

lcWGR was implemented in the 1,000 Genomes Project (2008–2015) (Auton et al., 2015), where the initial statistical models, file formats and programs were developed. Its use has been restricted mostly to humans and, more recently extended to agricultural and other nonmodel species. However, only few laboratories have used this approach thus far, explaining perhaps the scarcity of software available for this type of data. As this approach gains popularity, new computer packages are likely to be developed.

TABLE 1 Comparison of requirements and different aspects of RAD-seq, Pool-seq and lcWGR approaches for population genomics studies

Aspect	RAD-seq (original protocol)[a]	Pool-seq	lcWGR
Expected percentage of the genome covered	~1%–5%	>70% (depending on reference genome completeness)	>70% (depending on reference genome completeness)
DNA quality	High molecular weight	High molecular weight	High molecular weight
DNA quantity per sample	>200 ng	>1 μg per pool (for TruSeq PCR-free kit) >200 ng per pool (for TruSeq Nano kit)	>50 ng (for Nextera kit), although it could be less[b]
Need of a reference genome	Not indispensable but desirable	Required	Required
Type of library	Usually noncommercial	Commercial	Commercial
Library insert size	~350–550 bp (standard library)	~350–550 bp (standard library). For detection of large structural variants with short-read sequencing, ~2–20 Kb (mate-pair library)	~350–550 bp (standard library)
Cost of library preparation	~USD$ 5–10/individual[c]	~$USD 46/pool[c]	~$USD 6/individual[b,c]
Minimum number of individuals per population	Usually ≥20[d]	≥50[e]	Number not established but usually ≥50
Popular sequencing platform	Illumina MiSeq and HiSeq, Ion-Proton	Illumina HiSeq	Illumina HiSeq
Type of sequence reads	Single-end or paired-end reads, ≥100 bp per read	Paired-end reads, ≥100 bp per read	Single-end or paired-end reads, ≥100 bp per read
Minimum sequencing depth of coverage	High coverage: ≥20× per individual for diploids and higher depth for polyploids[f]	High coverage: ≥50–100× per pool[e]	Low coverage: ~1–4× (per individual) depending on ploidy, for example, 2× recommended for diploid organisms[g]
Minimum computing resources	Desktop computer with multicore processor (≥24) and ≥64 GB RAM	QC and trimming of raw reads, read mapping, duplicate marking and read sorting of small to large size genomes (~1 Gb) can be performed in a desktop computer with multicore processor (≥32) and ≥128–256 GB RAM. Greater computing resources (i.e., computer cluster or computing facilities in the cloud) are required for larger genomes (≥1 GB), specially for SNP calling	
Computer data storage[h]	MBs per sample (depending on genome size and depth)	GBs per pool (depending on genome size and depth). For example, for one population of Atlantic herring (genome size ~900 Mb, 50 individuals per pool, depth 50× per pool): raw data ~50 GB, clean data ~40 GB, SAM file ~150GB, BAM file ~35 GB, gVCF file ~40 GB	GBs per individual (depending on genome size and depth)
Programming skills	Basic-Intermediate	Intermediate	Intermediate
Expected number of SNP loci per sample	Thousands (without reference), hundreds of thousands (with reference genome)	Millions	Millions
Type of variant assessed	Mostly SNPs, inversions when a reference genome is available	SNPs, INDELs, large SVs, CNVs	SNPs, INDELs detected but not used in software, some SVs (depending on genome coverage per individual)
Type of genetic variation screened	Mostly neutral and sometimes functional (depending on marker density)	Neutral and functional	Neutral and functional
Output data obtained per sample	Individual SNP genotypes (based on coverage)	Population-level allele frequency per SNP (based on read counts per variant site)	Population-level genotype likelihood (based on reads of multiple individuals in a population)

(Continues)

TABLE 1 (Continued)

Aspect	RAD-seq (original protocol)[a]	Pool-seq	lcWGR
Possibility to do individual-based analyses	Yes	No, individual information is missed during library preparation	No, reliable individual SNP calls cannot be obtained from low-coverage data
Scalability (+): most positive feature (−): most negative feature	High (+) Cheaper method than Pool-seq and lcWGR enabling the analysis of numerous individuals and populations (−) Low marker density limits the capacity to detect adaptive variation	High (+) Many individuals per population can be mixed in one pool for the same library preparation cost (−) Sequencing depth should be increased accordingly. More expensive than RAD-seq but cheaper than lcWGR	High (+) Low depth per individual enables the analysis of numerous individuals per population (−) More expensive than Pool-seq, especially for >50 individuals analysed per population

GB, gigabytes; MB, megabytes; CNVs, copy number variations; SVs, structural variations; INDELs, insertions and deletions; SNPs, single nucleotide polymorphisms; QC, quality control.

[a] Baird et al. (2008), as cited in (Andrews et al., 2016).

[b] When following the protocol by Therkildsen and Palumbi (2017).

[c] Price for March 2017.

[d] Hohenlohe et al. (2010).

[e] Schlötterer et al. (2014).

[f] Andrews et al. (2016).

[g] Fumagalli (2013).

[h] Additional data storage space is required for temporary files generated during data analysis, which can exceed (2–3×) the size of the final data file (e.g., for Atlantic herring, SAM file ~150 GB).

(>50×) of pooled DNA in equimolar concentration of unlabelled individuals from a population (Futschik & Schlötterer, 2010; Schlötterer et al., 2014) and (ii) lcWGR, or low-coverage individual whole-genome resequencing of multiple barcoded individuals from a population (~2–4× per individual) (Durbin et al., 2010; Nielsen et al., 2011). The general workflow for data acquisition with these approaches is presented in Box 2, and a comparison of requirements, technical aspects and expected outcomes is shown in Table 1.

Pool-seq and lcWGR have several pros and cons. The main advantage of Pool-seq is the cost reduction achieved from the preparation of a single sequencing library per pooled DNA instead of one library per individual. This allows using large sample sizes per population (Figure 2). Also, pooling equal amounts of DNA of multiple individuals facilitates the sequencing of a few chromosomes several times, leading to an improvement in SNP allele frequency estimates (Gautier et al., 2013; Ferretti, Ramos-Onsins, & Pérez-Enciso, 2013; Schlötterer et al., 2014). In Pool-seq, the detection of variant sites and the estimation of population-level allele frequencies per SNP are derived from the relative proportion of read counts of each allele within a pool (Figure 2). Pool-seq has three main limitations: First, individual genotypes are missed after mixing DNA samples in a pool. This makes it impossible to track technical errors during library preparation. Second, allele frequency estimation is susceptible to multiple factors including uneven representation of individual DNA in a pool, and sequencing and mapping errors. Finally, rare alleles are likely to be under-represented in this kind of data sets, which can lead to a truncated distribution of allele frequencies or site frequency spectrum (SFS). Pool-seq data are thus mostly biased towards the detection of frequent and large-effect alleles (Cutler & Jensen, 2010; Raineri et al., 2012). Potential solutions to these limitations are discussed in Box 3. Obtaining a complete SFS is important in population genetics, as this metric synthesizes all the sequence variation at unlinked sites in a sample. Its shape varies with different evolutionary processes including bottlenecks or range expansions (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009), and natural selection (Bustamante, Wakeley, Sawyer, & Hartl, 2001; Ronen, Udpa, Halperin, & Bafna, 2013), and is used to infer several metrics including Tajima's D and F_{ST} (Durrett, 2008; Han, Sinsheimer, & Novembre, 2015).

The main advantage of lcWGR is the low sequencing depth targeted per individual (~1–4×) that facilitates the analysis of a large number of samples per population; however, one library needs to be prepared for each individual (Figure 2). Individual library preparation is still cost restrictive for large sample sizes given the current high cost of commercial library preparation kits (Table 1). To overcome this limitation, the use of smaller reaction volumes and cheaper reagents has been advocated achieving a 6–10 times cost reduction per sample as shown for microbial genomes (<15 Mb) (Baym et al., 2015) and a teleost fish genome (~730 Mb) (Therkildsen & Palumbi, 2017). Despite the significant cost savings with this procedure, lcWGR is still slightly more expensive than Pool-seq for an equivalent sample size and sequencing depth (assuming 1× coverage per individual). The overall cost of Pool-seq and lcWGR is fairly equivalent, though, when ~50 individuals are included (~$USD 280 more in lcWGR as June 2017). Just like in Pool-seq, some of the disadvantages of lcWGR are that individual genotypes cannot be called: the low depth per individual impedes a reliable variant and genotype calling. Instead of read counts, this method detects variant sites and calculates genotype likelihoods (GLs) per site based on the accumulated sequence data of multiple individuals in a sample using a probabilistic framework that incorporates the uncertainty of the data due to sequencing, alignment and SNP calling errors. Based on GLs obtained across sites, a sample allele frequency per site is calculated from which other statistics are inferred, including the SFS (Nielsen et al., 2011, 2012; Korneliussen et al., 2014). From the sample and population-allele frequency likelihood, SNP and genotype calls can be obtained using a likelihood ratio test (Kim et al., 2011; Korneliussen et al., 2014). In theory, because SNP calling is avoided, all alleles present in a sample are considered in the GL calculation, resulting in the SFS potentially being less biased than in Pool-seq data. Another disadvantage of this method is that GLs values can vary depending on several factors, and currently, a few programs accept GLs. Potential solutions to these limitations are discussed in Box 4.

2.3 | Comparison of WGR with reduced-representation sequencing (RRS)

Reduced-representation sequencing is a general category of techniques that sequence a subset of the genome following different strategies. These techniques can be classified in three major groups: (i) Restriction site Associate DNA sequencing (RAD-seq; Andrews et al., 2016), (ii) Sequencing of cDNA obtained from mRNA (RNA-seq; Ozsolak & Milos, 2011) and (iii) Whole-exome sequencing (WES; Warr et al., 2015).

(i) RAD-seq refers to a group of methods (e.g., traditional RAD, ddRAD, ezRAD, RAD-cap, among others) that evaluate the genetic variation present around restriction cut sites. The selection of restriction enzyme (frequent or rare cutter) determines marker density (i.e., number of loci sampled per physical genomic distance unit), making these methods flexible and customizable. RAD-seq typically examine thousands of low-density genomewide SNPs located in neutral and putatively functional loci that can be genotyped by sequencing in multiple individuals and populations for a relatively low cost (reviewed by Andrews et al. (2016)). (ii) RNA-seq focuses on genetic variants in parts of the genome that are being transcribed at the time of sampling. RNA-seq is thus mostly used as a cost-effective approach for gene expression quantification but also for the comparison of variants at genes being transcribed in a particular time/tissue (reviewed by Ozsolak and Milos (2011)). (iii) WES explores genetic variants in exons of protein-coding genes using capture probes usually developed from a well-annotated reference genome (reviewed by Warr et al. (2015)). A reference genome is, however, not indispensable, as capture probes can also be designed from PCR products of targeted loci, from de novo assembly of RNA-seq transcriptomes or expressed sequence tags (ESTs), RAD-seq or WGR data and from the genome of a closely related species as functional elements are usually located in highly conserved regions (reviewed by Jones and Good (2016)). WES constitutes a cost-effective alternative to WGR for functional SNPs identification, as it screens protein-coding regions that in humans, for example, represent <2% of the total genome (a 100 times reduction of the amount of data for the same coverage) and contains 85% of mutations related to diseases in Mendelian disorders (Rabbani, Tekin, & Mahdieh, 2014).

The three RRS methods share the characteristic that they typically evaluate a small fraction (~1%–5%) of the genome, which translates into reduced sequencing cost, computing resources and storage requirements compared to WGR approaches (Ozsolak & Milos, 2011; Warr et al., 2015). RRS techniques usually provide hundreds to thousands of genomewide SNPs that are plenty for population genetics analyses. Probably the most convenient advantage of RRS approaches is that they do not rely on a reference genome for SNP calling (variant detection is based on local read alignment), which has facilitated the broad use of RRS methods in population genomics studies of nonmodel organisms (De Wit, Pespeni, Ladner, Barshis, & Palumbi, 2012; Andrews et al., 2016). The number and quality of markers can be significantly improved, though, when the reference genome is available (Ozsolak & Milos, 2011; Warr et al., 2015; Jones & Good, 2016; Andrews et al., 2016).

As experience on these techniques accumulates, so have realizations on limitations and potential sources of bias and error. For example, in RAD-seq, potential bias can be introduced during library preparation, sequencing and data analysis (reviewed in Cariou, Duret, & Charlat, 2016; Shafer et al., 2017). Also, the small fraction of the genome that is often interrogated with these methods may limit their power to detect adaptive variation (Hoban, Kelley, & Lotterhos, 2016; Lowry et al., 2017a,b). A higher marker density (and thus genome coverage) can be achieved with the use of frequent cutter enzymes (McKinney, Larson, Seeb, & Seeb, 2017; Catchen et al., 2017); however, the target marker density should be informed by the genetic architecture of phenotypic traits of interest (i.e., the genetic basis of a character that is defined by the number of genes contributing to the trait, their effect size and interactions) and by the average linkage disequilibrium block size, information that is usually unknown in nonmodel species (more on this is discussed in Box 5).

In RNA-seq studies, care needs to be taken during sampling collection, sequencing and data processing. Multiple factors can affect the gene expression profile (e.g., stress generated by organism manipulation during sampling, environmental conditions, among others), resulting in technical and biological variability that needs to be accounted for with the inclusion of replicates. Moreover, high read coverage is necessary to detect rare transcripts, and the intrinsic complexity of the transcriptome (e.g., alternative splicing) could make read alignment challenging for SNP detection (Ozsolak & Milos, 2011; Conesa et al., 2016).

A potential problem with WES is that the exon capture/PCR amplification steps can produce low coverage (limiting variant detection) when probes are poorly designed (span exon boundaries) and fail to bind to the target region (Bi et al., 2012; Jones & Good, 2016).

Unlike RRS, WGR approaches provide the highest marker density of the current genomic methods, facilitating the characterization of neutral and functional genetic variation as well as the discovery of the genetic basis of phenotypic traits (Ellegren, 2014). The proportion of the genome screened with WGR depends on read depth and length, and reference genome completeness. A comparison of the expected relative proportion of the genome covered by WGS, WES and RAD-seq is shown in Figure 3, and a comparison of requirements and outcomes of RAD-seq, Pool-seq and lcWGR is presented in Table 1. WGR is more robust than WES for the detection of exome variants as it provides a more homogeneous sequence read coverage and a better sequencing quality overall (Belkadi et al.,

BOX 5 Considerations and limitations of genome scans for detecting selection and inferring local adaptation

Genome scans are currently one of the most popular methods for the detection of selection from genomic data (Barrett & Hoekstra, 2011; Martin & Jiggins, 2013; Vitti et al., 2013; Pardo-Diaz et al., 2015; Hoban et al., 2016), in particular for the detection of directional selection.

Genome scans encompass a comprehensive survey of the genome of individuals and populations to identify molecular patterns that presumably result from selection, for example, increased linkage disequilibrium (LD; i.e., long haplotype blocks) and reduced variation around beneficial mutations, abundance of rare alleles in the population site frequency spectrum, significant allele frequency differences between populations under contrasting selection regimes, among others (Vitti et al., 2013; Ellegren, 2014; Jensen, Foll, & Bernatchez, 2016). Currently, genome scans are based on: (i) information on the physical distance (LD blocks) between loci for the detection of *selective sweeps* (Messer & Petrov, 2013; Vatsiou, Bazin, & Gaggiotti, 2016), or on (ii) knowledge of allele frequency differences between unlinked loci using (a) outlier loci tests, or (b) genome-environment association (GEA) analysis (Bernatchez, 2016; Gagnaire & Gaggiotti, 2016; Hoban et al., 2016). (i) *Selective sweeps* can be "hard," when the beneficial mutations of large effect are new and increase in frequency in the population in a short period of time, or they can be "soft," when they comprise numerous alleles of small effect that were already present in the population or resulted from recurrent independent mutational events (Messer & Petrov, 2013). The genetic signal for hard sweeps is generally easier to detect in genomic data as it includes elevated differentiation at particular loci, whereas the soft sweeps signal can be confounded with the genomic background because the genetic changes involved are more subtle. The detection of either type of selective sweeps and their distinction requires the use of specific statistical tools (reviewed by Messer and Petrov (2013) and Vatsiou et al. (2016)). (ii)(a) Outlier loci tests rely on the detection of putatively selected loci showing elevated levels of differentiation with respect to expectations under a neutral model and usually involve a window-based approach where summary statistics (e.g., F_{ST}) are estimated and averaged for all the variants present in the window (Barrett & Hoekstra, 2011). Window size choice is thus important as, by modifying the number and the range of physical separation between variants, the outcome could change. A very large window can lead to an overestimation of outlier loci due to false positives, whereas an excessively narrow window can lead to an underestimation of outlier loci by excluding from the window sections of low differentiated genomic background useful for outlier detection. Window size selection should thus account for average genomewide LD or, in its absence, for the relative genomic position and separation of variants, population polymorphism level (Hoban et al., 2016), or could also be statistically inferred as breakpoints in the genomic data (Beissinger, Rosa, Kaeppler, Gianola, & de Leon, 2015). (ii) GEA analysis uncovers putatively adaptive loci through the comparison of the genetic variation between populations adapted to contrasting environments (Barrett & Hoekstra, 2011; Martin & Jiggins, 2013; Pardo-Diaz et al., 2015). The choices of populations and environmental variables to compare are relevant as they define the power of such comparison. The population spatial resolution and the temporal resolution of environmental variables need to be considered as they will directly affect the correlations between outlier loci and environment (Hoban et al., 2016).

Other limitations of genome scans include the fact that: (i) some outlier loci may not themselves be under selection but may instead be located in the proximity of a causal mutation, implying that follow-up functional molecular studies testing the phenotypic effect of an outlier locus are needed to consider it adaptive (Barrett & Hoekstra, 2011); (ii) signals of selection can be confounded with footprints of demographic history (e.g., populations structure) (Tiffin & Ross-Ibarra, 2014); (iii) mutation and recombination rates, type and strength of selection, and the genetic architecture of adaptive traits all modulate the genomic heterogeneity of a species, restraining the capacity to detect the genetic basis of adaptation (Haasl & Payseur, 2016). Several solutions have been proposed to overcome these limitations (and others not commented here) and are described in more detail in the referenced papers.

The genetic architecture of an adaptive trait refers to the total number of genes contributing to a given character, their location, effect size and heritability, and the interactions among them (i.e., additivity, epistasis, dominance, pleiotropy) and with the environment (genome-environment interactions) (Hansen, 2006; Gagnaire & Gaggiotti, 2016). This architecture determines the range of allele frequency changes in a population responding to selection (Gagnaire & Gaggiotti, 2016). For instance, in oligogenic traits (i.e., characters underlined by a few large-effect genes) a large shift in allele frequencies is expected, whereas a small change is assumed in polygenic traits where characters result from the interaction of multiple small effect genes. The power of genome scans to identify the genetic basis of quantitative traits based on allele frequency methods therefore depends on how much of the total adaptive genetic variation of a trait is explained by the summed effect of the outlier loci (Berg & Coop, 2014; Gagnaire & Gaggiotti, 2016). Sampling design in a complex environmental landscape (Lotterhos & Whitlock, 2015), as well as sample size and number and density of markers thus play an important role in our capacity to reveal the genetic basis of adaptive traits, especially for traits of polygenic architecture (Gagnaire & Gaggiotti, 2016) which are common in nature (Rockman, 2012; Bernatchez, 2016). The choice of sequencing technique for the collection of genomic data is thus not trivial as it will define the proportion and type of genetic variation in a genome that has been sampled and thus the potential inclusion or exclusion of relevant loci (Hoban et al., 2016). The marker density required for

BOX 5 Continued

a genome scan should thus ideally account for the average LD decay to ensure that most variants contributing to a trait are surveyed (Gagnaire & Gaggiotti, 2016).

The importance of a proper planning of sequencing approaches for the study of local adaptation has been brought into sharp focus by a recent debate on the power of RAD-seq for the detection of adaptive genetic variation in natural populations. Lowry et al. (2017) used computer simulations to argue that a large proportion of putatively adaptive loci are missed by RAD-seq studies because typically only a small fraction of the genome is surveyed with loci being too widely spaced. The problem is expected to be more important for species with large census and effective population sizes which tend to exhibit short LD blocks (high LD decay and high recombination rate). On the other hand, McKinney et al. (2017) claim that a properly designed RAD-seq study that takes advantage of the flexibility of restriction enzymes, can provide enough markers to achieve high genome coverage. Authors also argue that LD and recombination are not homogeneous across the genome and adaptation signatures can frequently result in extended LD blocks (or genomic islands of divergence, Nosil, Funk, and Ortiz-Barrientos (2009)) spanning several Kb, that can be easily screened with the low marker density provided by RAD-seq, regardless of effective population size (McKinney et al., 2017). In addition, Catchen et al. (2017) argue that even with short LD blocks, RAD-seq has been successful at detecting adaptive loci (e.g., the *Eda* locus in three-spine sticklebacks), and that endangered species usually exhibit small effective population sizes for which large LD blocks should be expected. They also pointed out that some studies may be focused on detecting adaptive differentiation only at the fraction of the genome sampled and not at all adaptive loci present. Finally, Lowry et al., (2017b) emphasize that the average LD block size and variation of recombination rate along the genome is usually unknown for nonmodel species, thus, it is difficult to estimate *a priori* the minimum marker density required for a RRS approach. They propose RAD-seq studies aiming to detect adaptive loci should follow these basic principles: (i) report the limitations of a given study, (ii) in the absence of a reference genome or linkage map, efforts should be centred first on obtaining this information, (iii) complement genome scans with alternative sources of evidence (i.e., field experiments or functional molecular tests) that demonstrate the phenotypic effect of outlier loci and (iv) conduct pilot tests to assess the viability of a sequencing experiment plan.

Although previous knowledge on the extend of LD decay or recombination rate is generally lacking for nonmodel species, LD block size can be estimated from a dense genetic map or from RAD-seq data with a reference genome (Catchen et al., 2017). Figure 4 synthesizes the decision-making process for sequencing approach as a function of the existence of a reference genome; expected LD block size; whether the interest is on neutral, adaptive or both types of genetic variation; relative cost; and type of genetic variation assessed.

Sample size is another practical consideration that has not received much attention in population genomics, in part because of the perception that large sample sizes are not required because of the very large number of markers genotyped per individual. This may be fine in some cases (studies in Table 2 with small sample sizes), but in general, the establishment of a minimum sample size depends on the research question and the genetic architecture of the focal species. For example, in human genetics studies for the detection of small effect variants associated with rare diseases, even the screening of thousands of individuals has not provided enough power to detect and track such variation (Agarwala, Flannick, Sunyaev, & Altshuler, 2013; Lee, Abecasis, Boehnke, & Lin, 2014; Moutsianas et al., 2015). Therefore, a presumably large sample size may be required in studies of nonmodel species aiming to identify adaptive variation associated with polygenic traits.

In conclusion, multiple considerations need to be taken into account when planning genome scans for the detection of signatures of selection and local adaptation from genomic data. The choice of minimum sample size and sequencing technique for the collection of genomic data should respond to the research question and should be informed by the expected (or ideally verified) LD block size (or physical LD), and the genetic architecture of a given phenotypic trait. These factors will determine the marker density needed for the successful detection of putatively adaptive variation in the genome.

2015). Another advantage of WGR approaches is that they examine multiple types of genetic variation including structural variations (SVs; i.e., deletions, insertions, chromosomal rearrangements, copy number variation, Alkan, Coe, & Eichler, 2011) and mutations in regulatory elements (REs; i.e., noncoding regions that regulate gene expression and function, Wray, 2007). In contrast, RRS techniques are mostly restricted to one base changes (i.e., SNPs), and RNA-seq and WES are to variation within coding sequences.

3 | APPLICATIONS OF WGR IN CONSERVATION AND MANAGEMENT

Below, we describe contributions that WGR analysis can make to some of the main areas of interest in conservation. For each area we provide study cases that illustrate the type of questions that can be addressed with WGR. Table 2 lists key aspects of the experimental design of these studies. Within parentheses, we denote in the

TABLE 2 Key methodological aspects of case studies using WGR in different conservation biology topics

Organism	Main findings	WGR method	Sample size	Sequencing design and output	Number of markers	Software for mapping (parameters)	IR/ BQSR	SNP calling software	Ref.
Topic: Phylogenomics, hybridization and taxonomical species resolution									
Modern birds (Neoaves)	Genome-scale phylogenetic tree of 48 species representing all orders of Neoaves	huWGR	48 species, 1 ind. per species	Illumina HiSeq 2000, Roche 454, pair-end libraries of different insert sizes, 24×–160×	Whole-genome DNA sequences	SOAPdenovo[a]	–	–	Jarvis et al. (2014)
Snub-nosed monkey (*Rhinopithecus roxellana*, *R. bieti*, *R. brelichi*, *R. strykeri*)	Genome-scale phylogenetic relationships that indicate functional evolution and leaf-eating dietary adaptations	huWGR	4 species, 1 ind. per species	Illumina HiSeq, 30× huWGR, 146× de novo assembly, multiple paired-end and mate-pair libraries spanning size range of 180 bp to 20 kb	Whole-genome DNA sequences	SOAPdenovo[a]	–	–	Zhou et al. (2014)
North American wolves	Lack of unique ancestry in eastern and red wolves	huWGR	28 ind.	Illumina HiSeq, paired- and single-end sequencing libraries, average insert size of 300 to 500 bp, 4–29×	5,424,934 SNPs	Stampy[b]	NI	ANGSD[f]	vonHoldt et al. (2016)
Common milkweed (*Asclepias syriaca*)	Marker development including complete chloroplast genome, a nearly complete rDNA cistron and 5S rDNA sequence, a partial mitochondrial genome sequence and some single copy ortholog genes	lcWGR	1 ind.	Illumina GAII, 1 lane, 40 cycles, 0.5×	–	–	–	–	Straub et al. (2011)
Beardtongue plant (*Penstemon. centranthifolius*, *P. grinnellii*)	Primer design of phylogenetic markers in the plant genus based on annotation and gene prediction of more than 10,000 contigs	lcWGR	1 ind. of each species	Roche 454 platform, ~0.005×–0.007×	–	–	–	–	Blischak et al. (2014)
Yellow baboons (*Papio cynocephalus*), Anubis baboons (*P. anubis*) and hybrids	Genetic differentiation between parent taxa, complex admixture history involving intermittent but multiple hybridization events that did not indicate fitness reduction in hybrids	lcWGR	46 ind. in total	Illumina HiSeq, for huWGR, 2.09×–19.6×, paired-end 100-bp reads	~2.1 million	BWA-MEM[d] (minimum seed length of 20)	Yes/ Yes	GATK UNIFIEDGENOTYPER[g], (discarding indels)	Wall et al. (2016)
Topic: Population structure and admixture									

(Continues)

TABLE 2 (Continued)

Organism	Main findings	WGR method	Sample size	Sequencing design and output	Number of markers	Software for mapping (parameters)	IR/BQSR	SNP calling software	Ref.
Giant pandas (*Ailuropoda melanoleuca*)	Multiple demographic events including population expansion, bottlenecks and divergence, human activities most likely contributed to decline in the last 3000 years	lcWGR	34 ind.	llumina HiSeq2000 platform and 100-bp paired-end reads, 4.7×	13,020,055 SNPs	BWA-ALN[c] (NI)	NI	SOAPsnp[m]	Zhao et al. (2012)
Killer whale (*Orcinus orca*)	Differentiations between pairs of contemporary allopatric and sympatric ecotypes most likely are the consequence of ecological divergence and genetic drift resulting from bottlenecks experienced during past founder events	lcWGR	(lcWGR): 48 ind., (huWGR): 2 ind.	Illumina HiSeq 2000 platform using single-read 100-bp chemistry, 2×	603,519 variant sites	BWA[c] (NI)	NI	ANGSD[f]	Foote et al. (2016)
Western palearctic black-and-white flycatchers of the genus *Ficedula* (*F. speculigera*, *F. albicollis*, *F. hypoleuca*, *F. semitorquata*)	Most recent common ancestor of the four species dates back to 1–2 million years (Mya) ago and each species followed separated evolutionary paths involving population growth, decline (~100–200 thousand years ago) and expansion	huWGR	200 ind. in total	HiSeq 2000, paired-end 100 bp, insert size ~450 bp, ≥20×	NI	BWA[c] (NI)	Yes/Yes	SAMTOOLS[h]	Nadachowska-Brzyska et al. (2016)
Cattle (*Bos taurus*)	Historical events such as domestication or modern breeding are related with population decline	huWGR	15 to 25 ind. from each of 4 breeds	Illumina	NI	BWA[c] (NI)	NI	SAMTOOLS[h]	Boitard et al. (2016)
European dark honey bee (*Apis mellifera mellifera*) and two introduced honey bee subspecies (*A. m. carnica* and buckfast)	Genetic differentiation between subspecies that coincides with geography. Observed presence of admixed individuals in protected areas	huWGR	151 drones	Illumina HiSeq2500, pair-end 2 × 125 bp reads, 10×	3.375 million SNPs	BWA-MEM[d] (NI)	Yes/Yes	GATK-UNIFIEDGENOTYPER[g]	Parejo et al. (2016)

(Continues)

TABLE 2 (Continued)

Organism	Main findings	WGR method	Sample size	Sequencing design and output	Number of markers	Software for mapping (parameters)	IR/ BQSR	SNP calling software	Ref.
Arabidopsis halleri	Weak genetic differentiation among populations. SNPs more informative than microsatellites about genomewide genetic diversity	Pool-seq	20 ind. in 9 populations	(Pool-seq): Illumina HiSeq2000, paired-end 2× 100 bp reads, 250–300 bp insert size, 60.7× (range 52.7 to 69.3× per pool)	2,178,204 SNPs	BWA-ALN[c] and sample (allowing 10% mismatch)	NI	SAMTOOLS[h]	Fischer et al. (2017)
Almond (*Prunus dulcis*) and peach (*P. persica*) plants	Almond genomewide nucleotide diversity was ~7-fold higher than in peach, excess of rare alleles likely consistent with a recent population expansion event, no evidence of population bottleneck related with domestication and a strong genetic differentiation between almond and peach	lcWGR	13 ind. per species	Illumina HiSeq2000, 100-bp paired-end reads, depth averaged 15.8× (4.7×–34.6×) in almond and 19.7× (11.2× to 35.4× in peach	NI	BWA-MEM[d] (minimum seed length of 10 and internal seed length of 2.85)	NI	ANGSD[f]	Velasco et al. (2016)
Topic: Signatures of selection, genetic basis of phenotypic traits and local adaptation									
Darwin finches (*Geospiza* spp.)	A 240 kilobase haplotype encompassing the ALX1 gene that encodes a transcription factor affecting craniofacial development is strongly associated with beak shape	huWGR	200 ind. distributed in 15 species	Illumina Hiseq2000, 2× 100 bp paired-end reads, insert size ~400 bp, 10× coverage	44,753,624 SNPs	BWA[c] (default)	Yes/Yes	GATK-UNIFIEDGENOTYPER[g]	Lamichhaney et al. (2015)
Atlantic salmon, (*Salmon salar*)	Locus that maintains variation in age at maturity	huWGR	54 wild populations, ~500 ind., (SNP array) and 32 ind. from 7 populations (huWGR)	(huWGR): HiSeq2500, paired-end reads 2× 125 bp, average coverage 18× (8×–32×)	208,704 SNPs	BWA-MEM[d] (default)	No/No	FreeBayes[i]	Barson et al. (2015)
Atlantic salmon, (*Salmon salar*)	Locus vgll3 controls age at maturity in wild and domesticated Atlantic salmon males	Pool-seq	20 ind. per river and phenotype	Illumina HiSeq2000 platform, 12.3× per pool	4,326,591 SNPs	Bowtie2[e] (no soft clipping, end-to-end mode, seed length 18, only 1 mismatch)	NI	SAMTOOLS[h]	Ayllon et al. (2015)

(Continues)

TABLE 2 (Continued)

Organism	Main findings	WGR method	Sample size	Sequencing design and output	Number of markers	Software for mapping (parameters)	IR/ BQSR	SNP calling software	Ref.
Red siskins (*Spinus cucullata*), common canaries (*Serinus canaria*) and “red factor” canaries	Gene encoding a cytochrome P450 enzyme, CYP2J19, is the ketolase that mediates red coloration in birds	Pool-seq	12 to 39 ind. per pool	Illumina Hiseq2500, 2× 100-bp reads, 19.3× per pool	9,414,439 SNPs	BWA-MEM[d] (default)	Yes/NI	SAMTOOLS[h] and VarScan2[j]	Lopes et al. (2016)
Marine midge (*Clunio marinus*)	Locus calcium/calmodulin-dependent kinase II.1 (CaMKII.1) splice variants strongly associated with circadian timing	Pool-seq	5 populations, 100–300 ind. each	Illumina HiSeq2000, paired-end 2× 100 bp reads, 0.2–0.4Kbp insert size, 68× to 251× per pool	1,010,052 SNPs	BWA-ALN[c] and sample (maximal insert size 1500 bp)	Yes/No	Popoolation2[k]	Kaiser et al. (2016)
Domestic and wild rabbits (*Oryctolagus cuniculus*)	Mapped genes affecting brain and neuronal development likely associated with domestication	Pool-seq	10–20 ind. in 7 domestic populations and 14 wild populations	Illumina Genome Analyzer II, paired-end 2× 76 bp reads, average coverage ~10× per pool	50,165,386 SNPs	BWA[c] (default except –q 5, base quality cut-off for soft-clipping reads)	NI	SAMTOOLS[h]	Carneiro et al. (2014)
Chicken (*Gallus gallus domesticus*)	Genes associated with breed-related traits of pathogen resistance and reproductive ability	Pool-seq	16 ind. for each of 2 inbred lines	Illumina Hiseq 2000, 22–24× per pool	~4 million SNPs	BWA[c] (default)	NI	GATK-UNIFIEDGENOTYPER[g], down sampling was turned off, ploidy option was used	Fleming et al. (2016)
Yellow monkey-flower plant (*Mimulus guttatus*)	Candidate genes potentially driving the morphological, life history and salt tolerance differences between the two ecotypes	Pool-seq	“Coastal perennial” pool with 101 ind., “inland annual” pool with 92 individuals	Illumina HiSeq2500, 2× 250 bp paired-end reads,	29,693,578 SNPs	BWA[c] (NI)	NI	SNAPE[l]	Gould et al. (2017)
Atlantic salmon (*Salmon salar*)	Mapped immune-related genes	Pool-seq	30 ind. in each of 19 rivers	Illumina HiSeq2000, paired-end reads, average 26.7× per pool	~4.5 million SNPs	Bowtie2[e] (without soft clipping, end-to-end mode, seed length 18 and the interval between extracted seeds set to S,1,1.5, maximum number of mismatches per seed set to L,0,0.1)	NI	SAMTOOLS[h]	Kjærner-Semb et al. (2016)
Mosquito *Anopheles gambiae* s.l. (GOUNDRY) and *A. coluzzii*	A genomic barrier (large inversion) to gene flow between a *A. gambiae* s.l. (GOUNDRY) and *A. coluzzii*	lcWGR	11–12 ind. each	Illumina HiSeq2000, paired-end 100-bp, insert size of 500 base pairs, 9.79× to 16.44×	162–180 million SNPs per subgroup	BWA-MEM[d] (NI)	Yes/No	ANGSD[f]	Crawford et al. (2016)

(Continues)

TABLE 2 (Continued)

Organism	Main findings	WGR method	Sample size	Sequencing design and output	Number of markers	Software for mapping (parameters)	IR/ BQSR	SNP calling software	Ref.
Native sheep (*Ovis aries*)	Loci presumably involved in adaptation to high altitude and arid environments in native sheep	lcWGR	77 ind.	Illumina HiSeq 2000, 75 ind. with average depth of ~5× and ~42× for 2 samples	~21.26 million SNPs	BWA[c] (NI)	Yes/No	SAMTOOLS[h]and ANGSD[f]	Yang et al. (2016)
Shetland ponies (*Equus caballus*)	Deletions at the *SHOX* locus associated with skeletal atavism	Pool-seq	6 affected and 21 control ind.	Illumina, Hiseq2000, paired-end reads 2× 100 bp, ~56×	9,844,628 SNPs and 1,111,009 INDELs	BWA[c] (NI)	Yes/NI	GATK UNIFIEDGENOTYPER[g]	Rafati et al. (2016)
Bird ruff (*Philomachus pugnax*)	Large chromosomal inversion underlines the variety of male mating morphs	huWGR	15 independent and 9 satellite males from a single location	Illumina HiSeq 2000, 2× 125 bp paired-end reads, average fragment size ~500 bp, average ~8×	NI	BWA[c] (default)	Yes/Yes	GATK[g]	Lamichhaney et al. (2015)
Bird ruff (*Philomachus pugnax*)	Large chromosomal inversion underlines the variety of male mating morphs	huWGR	300 ind. in total	(huWGR): 80× for 5 ind. and low- (SbfI) and high- (PstI) density RAD-seq data from a pedigree population	1,068,556 SNPs	BWA-MEM[d] (default)	Yes/NI	GATK-HAPLOTYPECALLER[g] (huWGR and lo-density RAD-seq data), BCFtools (for high-density RAD-sed data)	Küpper et al. (2015)
Plant *Arabidopsis halleri*	175 genes highly associated with some of the five environmental factors tested (precipitation, slope, solar radiation, site water balance and temperature)	Pool-seq	5 populations, 20 ind. per pool	Illumina HiSeq2000, 250–300 bp insert size, 100-bp paired-end reads, average 99× per pool	~2 million SNPs	BWA-ALN[c] and sample (allowing 10% mismatch with the *A. thaliana* reference genome)	NI	SAMTOOLS[h]	Fischer et al. (2013)
Atlantic/Baltic herring (*Clupea harengus*) and Pacific herring (*Clupea pallasii*)	Genetic differences between populations spawning in different seasons and oceanic and brackish water in Europe	Pool-seq	20 populations of *C. harengus*, 1 of *C. pallasii*, 47–100 ind. per pool, 16 ind. for WGR	Illumina Hiseq2000, paired-end 2× 100 bp reads, insert size ~350 bp, ~30× per pool, ~10× per individual	8.83 million SNPs (with Pacific herring), 6.04 million among Atlantic and Baltic herring	BWA-MEM[d] (NI)	No/No	GATK-HAPLOTYPECALLER[g]	Martinez Barrio et al. (2016)
Atlantic/Baltic herring (*Clupea harengus*) and Pacific herring (*Clupea pallasii*)	6,333 SNPs showed significant allele frequency differences between spring and fall spawning populations in Canada. About 25% of these SNPs were previously observed in Baltic Sea/NE Atlantic populations	Pool-seq	(NE): Martinez Barrio et al. (2016) (NW): 6 populations of *C. harengus*, 41–50 ind. per pool, 12 ind. for WGR	(NE): Martinez Barrio et al. (2016) (NW): Hiseq2500, paired-end 2× 125 bp reads, insert size ~450–550 bp, ~40× per pool, ~10× per individual	~8,9 million SNPs	BWA-MEM[d] (default)	No/No	GATK-HAPLOTYPECALLER[g]	Lamichhaney et al. (2017)

(Continues)

TABLE 2 (Continued)

Organism	Main findings	WGR method	Sample size	Sequencing design and output	Number of markers	Software for mapping (parameters)	IR/ BQSR	SNP calling software	Ref.
Topic: Inbreeding, conservation breeding and restoration									
Mountain gorilla (eastern species: *Gorilla beringei beringei*, *G. beringei graueri*, western species: *G. gorilla gorilla*, *G. gorilla diehli*)	Extensive inbreeding (34% homozygosis) observed, indicating a steady population decline over the past 100,000 years	huWGR	44 ind. in total	Illumina HiSeq 2000, average 26×	1,649,453,084 SNPs	BWA-MEM[d] (default)	NI	FreeBayes[i]	Xue et al. (2015)
Tree *Eucalyptus grandis*	Progeny retained high and different heterozygosity percentage (52%–79%, average 66%), in disagreement with an expectation of 50% homozygosis (IBD) produced by selfing without selection	huWGR	1 outbred parent, 28 selfed offspring	Illumina HiSeq, paired-end 2× 100 bp reads, average 6.733×	308,784 heterozygous SNPs	BWA[c] (default, -q 15)	NI	SAMTOOLS[h], BAQ scores disabled.	Myburg et al. (2014)
Tree *Eucalyptus grandis*	Pseudo-overdominance most likely explains observed inbreeding depression, and it could be underlined by 100 or more genes of large effect associated with viability	huWGR	1 outbred parent, 28 selfed offspring	Illumina HiSeq, paired-end 2× 100 bp reads, average 6.733×	308,784 heterozygous SNPs	BWA[c] (default, -q 15)	NI	SAMTOOLS[h], BAQ scores disabled.	Hedrick et al. (2016)
Rice (*Oryza sativa*)	Identification of causal mutations of three phenotypic traits in inbred rice varieties	lcWGR	203 varieties	Illumina Hiseq2000, paired-end 90-bp reads, insert size 400–500 bp, average coverage 1.53×	2,288,867 SNPs	BWA[c] (default, remapping using Stampy)	Yes/NI	ANGSD[f]	Wang, Xu, et al. (2016)

IR, INDEL recalibration; BQSR, base quality score recalibration; NI, no information; SNPs, single nucleotide polymorphisms; ind., individuals; *x*, depth of coverage; huWGR, high-coverage individual whole-genome resequencing; Pool-seq, whole-genome resequencing of pooled DNA; lcWGR, low-coverage individual whole-genome resequencing.

[a] Luo et al. (2012).

[b] Lunter and Goodson (2011).

[c] Li and Durbin (2009, 2010).

[d] Li (2013).

[e] Langmead and Salzberg (2012).

[f] Korneliussen et al. (2014).

[g] McKenna et al. (2010).

[h] Li, Handsaker, et al. (2009).

[i] Garrison and Marth (2012).

[j] Koboldt et al. (2012).

[k] Kofler, Pandey, & Schlötterer (2011).

[l] Raineri et al. (2012).

[m] Li, Li, et al. (2009).

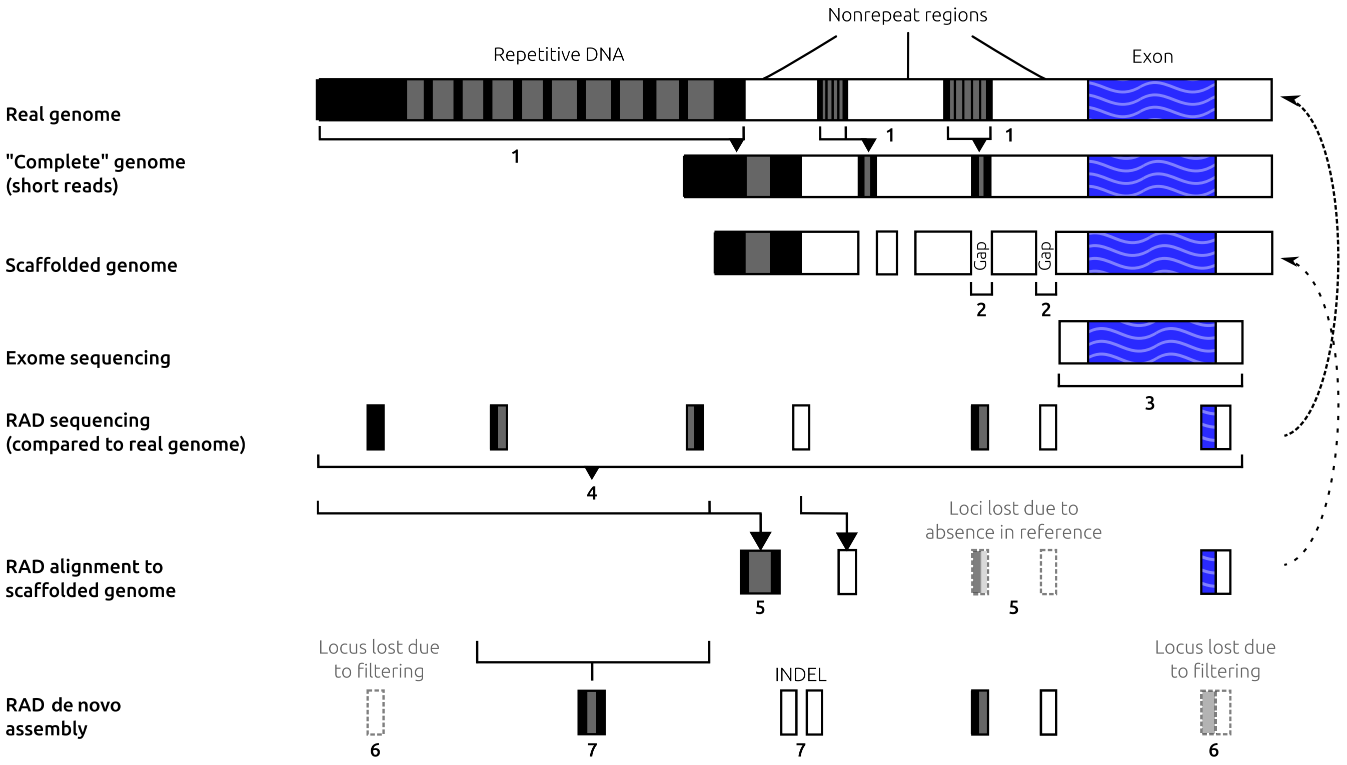

Nonrepeat regions
Repetitive DNA
Exon
Real genome
"Complete" genome (short reads)
1
1
1
Scaffolded genome
Gap
Gap
2
2
Exome sequencing
3
RAD sequencing (compared to real genome)
4
Loci lost due to absence in reference
RAD alignment to scaffolded genome
5
5
Locus lost due to filtering
INDEL
Locus lost due to filtering
RAD de novo assembly
6
7
7
6

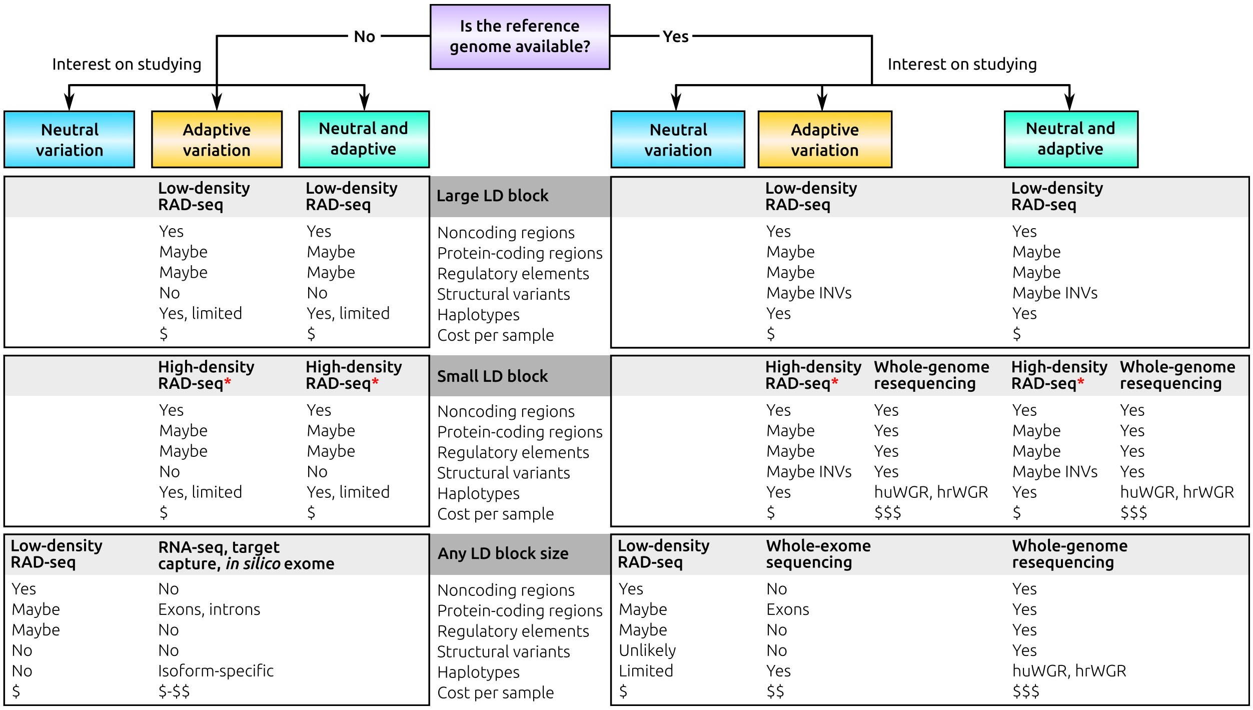
FIGURE 3 Proportion of the genome assessed by different approaches. A real genome comprises mostly repetitive and noncoding sequences with a small percentage corresponding to protein-coding regions. (1) Currently, a "complete genome" reference sequence usually misses a great proportion of repetitive regions as they are particularly challenging for base-calling and assembly algorithms based on short-read sequences; (2) A scaffold genome consists of large sequence blocks, resulting from the overlapping of short-read sequences (or contigs), that have gaps in between, corresponding to unknown parts of the sequence. A few repetitive regions are usually represented in it; (3) An exome sequence encompasses protein-coding regions (exons) and flanking sequences only, missing variation in regulatory and other noncoding regions; (4) RAD-seq randomly screens a small and dispersed amount of a real genome and includes protein-coding, noncoding and repetitive regions; (5) When RAD-seq reads are aligned onto a scaffold genome, some RAD tags are lost because they are not present in the reference sequence. Repetitive regions would be collapsed to the fraction represented in the scaffold genome; (6) When there is no reference sequence available, RAD-seq reads are assembled to each other to form contigs (known as de novo assembly). Fewer unordered loci are usually recovered with this approach than when using a reference genome, mainly during filtering or because of low read coverage. Also, (7) some contigs may be misaligned into different RAD loci when there are INDELs, and repetitive sequences may collapse into a single locus. Squares in grey and dashed lines indicate missing portions of the genome. The arrows point to the comparison being made and explained in the legend. Figure adapted from Hoban et al. (2016) with permission provided by The University of Chicago Press and Copyright Clearance Center

title of each conservation area, which WGR approach (huWGR, hrWGR, Pool-seq or lcWGR) applies.

3.1 | Phylogenomics, hybridization and taxonomical species resolution (approach: huWGR, hrWGR, lcWGR)

The successful implementation of conservation plans relies on the correct identification of the taxonomic status of organisms target of protection (Mace, 2004). Whole-genome data constitute a complete record of a species evolutionary process. By comparing large portions of the genome rather than the sequence of a few genes, as has traditionally been done, a more robust reconstruction of the evolutionary relationships among species can be achieved. This is the aim of phylogenomics (Delsuc, Brinkmann, & Philippe, 2005; Chan & Ragan, 2013). Recent studies have provided evidence of the power of whole-genome data for the reconstruction of the tree of life and for the detection of species hybridization events. More work needs to be done, however, to resolve algorithm limitations associated with the analysis of such large amount of data and to overcome intrinsic genomic challenges such as protein-coding sequence convergence, genome rearrangements, lateral gene transfer, incomplete lineage sorting, among others (Delsuc et al., 2005; Chan & Ragan, 2013).

huWGR and lcWGR approaches have been used in phylogenomics (Table 2) and in the detection of species hybridization. huWGR provides high genome coverage and taxonomic resolution whereas lcWGR offers a fragmented coverage that still can be useful for the development of phylogenetic markers (e.g., organellar genomes, ortholog genes, repetitive elements). For example, using huWGR, Jarvis, Zhang, and Li (2014) compared 48 modern bird species for the reconstruction of their phylogeny. They obtained a highly resolved tree that discriminates close relationships, identified the first divergence of Neoaves, but found difficulties attempting to resolve deep branches. Zhou et al., (2014) sequenced and assembled



	Low-density RAD-seq	Low-density RAD-seq	Large LD block		Low-density RAD-seq	Low-density RAD-seq
	Yes	Yes	Noncoding regions		Yes	Yes
	Maybe	Maybe	Protein-coding regions		Maybe	Maybe
	Maybe	Maybe	Regulatory elements		Maybe	Maybe
	No	No	Structural variants		Maybe INVs	Maybe INVs
	Yes, limited	Yes, limited	Haplotypes		Yes	Yes
	$	$	Cost per sample		$	$

	High-density RAD-seq*	High-density RAD-seq*	Small LD block		High-density RAD-seq*	Whole-genome resequencing	High-density RAD-seq*	Whole-genome resequencing
	Yes	Yes	Noncoding regions		Yes	Yes	Yes	Yes
	Maybe	Maybe	Protein-coding regions		Maybe	Yes	Maybe	Yes
	Maybe	Maybe	Regulatory elements		Maybe	Yes	Maybe	Yes
	No	No	Structural variants		Maybe INVs	Yes	Maybe INVs	Yes
	Yes, limited	Yes, limited	Haplotypes		Yes	huWGR, hrWGR	Yes	huWGR, hrWGR
	$	$	Cost per sample		$	$$$	$	$$$

Low-density RAD-seq	RNA-seq, target capture, *in silico* exome	Any LD block size	Low-density RAD-seq	Whole-exome sequencing	Whole-genome resequencing
Yes	No	Noncoding regions	Yes	No	Yes
Maybe	Exons, introns	Protein-coding regions	Maybe	Exons	Yes
Maybe	No	Regulatory elements	Maybe	No	Yes
No	No	Structural variants	Unlikely	No	Yes
No	Isoform-specific	Haplotypes	Limited	Yes	huWGR, hrWGR
$	$-$$	Cost per sample	$	$$	$$$

FIGURE 4 Sequencing alternatives for population genomics. RAD-seq, restriction-associated DNA sequencing; INVs, inversions; huWGR, high-coverage unresolved-haplotype whole-genome-resequencing, hrWGR, high-coverage resolved-haplotype WGR; LD block, Linkage disequilibrium block size. Red asterisk: RAD-seq methods commonly assess a reduced fraction of the genome, experimental adjustments are required to obtain higher marker density. Relative cost per sample: "$": affordable, "$$": more expensive than "$," and "$$$": most expensive. For each block size category (large, small, or any size), the type of genetic variation surveyed is listed underneath. If only neutral variation is of interest, the most cost-effective approach is low-density RAD-seq (any LD block size and with or without a reference genome). RAD-seq screens SNPs mostly in noncoding regions, although some variants may fall in protein-coding regions and regulatory elements. Short reads and low marker density limit the examination of large structural variants (SVs) (>50 bp). Haplotypes can be assigned though with limitations (Arnold, Corbett-Detig, Hartl, & Bomblies, 2013). In the absence of a reference genome and if only adaptive variation is of interest, there are two alternatives: RAD-seq and methods targeting protein-coding regions (i.e., RNA-seq, target capture and in silico exome (Lamichhaney et al., 2012)). The methods targeting protein-coding regions assess variation in exons and introns and may allow the reconstruction of haplotype-specific isoforms. If both, neutral and adaptive variation, are of interest, RAD-seq is the best approach. If a reference genome is available and the focus is the study of adaptive variation exclusively, there are three options, RAD-seq, whole-exome sequencing (WES), and whole-genome resequencing (WGR). RAD-seq surveys SNPs genomewide to a fraction of the genome determined by marker density adjusted to the expected LD block size. Inversions (Küpper et al., 2015) and haplotypes (Miller et al., 2012) can be assessed with this method, with haplotypes inferred from population data. WES screens SNPs located only in exons across the entire genome and may allow for the reconstruction of some haplotypes. WGR (i.e., Pool-seq, lcWGR, hrWGR, huWGR) offers the greatest marker density of all current approaches, assessing SNPs across the genome including noncoding and protein-coding regions, regulatory elements, as well as structural variants. Haplotype assignment can be achieved indirectly with huWGR and directly with hrWGR methods. WGR is the most expensive approach of all. When the interest is to evaluate both, neutral and adaptive variation, the same applies as in "adaptive variation only" and WGR is the best option for the multiple benefits it offers. Notice that the detection of adaptive variation depends not only on marker density but also on sampling size and the effect size of a locus (see Box 5)

the genome of snub-nosed monkey and compared it with those of three related monkey species. They found evidence of functional evolution and leaf-eating dietary adaptations in this group and were able to reconstruct species-specific demographic histories more precisely than previous attempts. vonHoldt et al., (2016) sequenced 28 wolf genomes to demonstrate that two endemic species of the North American wolf (the red wolf and the eastern wolf) are actually hybrids of coyote and grey wolf. With lcWGR data, Straub et al., (2011) characterized for the first time phylogenetic markers for the common milkweed (*Asclepias syriaca* L.), including the complete chloroplast genome, a partial mitochondrial genome sequence, and some single copy ortholog genes. Blischak, Wenzel, and Wolfe (2014) used ultra-low-coverage genome data (~0.005×–0.007×) for annotation and gene prediction of more than 10,000 contigs for primer design of phylogenetic markers in the plant genus *Penstemon*. Wall et al., (2016) analysed lcWGR data of yellow baboons (*Papio cynocephalus*), Anubis baboons (*P. anubis*) and their hybrids in the Amboseli ecosystem of Kenya. They found genetic differentiation between parent taxa and enough evidence to infer a complex admixture history involving intermittent but multiple hybridization events that did not indicate fitness reduction in hybrids. All these studies indicate the comparison of individual genomes can make significant contributions to conservation biology by helping resolve the phylogenetic status of species of concern and by identifying genomic regions that can be used for the development of cost-effective tools for species and hybrid identification. As phylogenomic inference relies on haplotypes, this kind of analysis cannot be performed using Pool-seq data but will be benefited by advances in hrWGR methods with long-read data.

3.2 | Demographic history and historical effective population size (approach: huWGR, hrWGR, lcWGR, Pool-seq)

The study of a species demographic history, including bottlenecks, migration patterns, range expansion and changes in historical effective population size (hN_e), is of great interest in conservation as it helps understand past historical events and their influence on the genetic makeup of contemporary populations. Such studies also allow for the testing of hypotheses regarding the effect of dated environmental events (e.g., appearance of barriers to gene flow, anthropogenic disturbance, climate change) on historical demographic processes that may have left a genetic imprint.

huWGR and lcWGR approaches have been used for the reconstruction of the demographic history of different species (Table 2). For example, Zhao et al., (2012) analysed huWGR data of 34 pandas finding genetic evidence of multiple demographic events such as population expansion, bottlenecks and divergence and inferred that human activities most likely contributed to their decline in the last 3,000 years. Foote et al. (2016) reconstructed the ancestral demography of five distinct ecotypes of the killer whale (*Orcinus orca*) based on lcWGR data of 48 individuals and huWGR data of 2 individuals. They discovered that patterns of differentiation between pairs of contemporary allopatric and sympatric ecotypes are most likely the consequence of ecological divergence and genetic drift resulting from bottlenecks experienced during past founder events.

hN_e has been estimated from huWGR data. For example, using pairwise sequentially Markovian coalescent (PSMC) analyses, Nadachowska-Brzyska, Burri, Smeds, and Ellegren (2016) obtained hN_e estimates of four species of Western Palearctic black-and-white flycatchers of the genus *Ficedula* based on whole-genome data of 200 individuals from 10 European populations. The hN_e curves indicated the most recent common ancestor of the four species dates back to 1–2 million years (Mya) and each species followed separate evolutionary paths involving population growth, decline (~100–200 thousand years ago [Kya]) and expansion. Authors suggest a mean genome coverage of ≥18× per individual, a per-site filter of ≥10 reads and no more than 25% of missing data are required for a proper inference of demographic history using PSMC and huWGR data (Nadachowska-Brzyska et al., 2016). Boitard, Rodríguez, Jay, Mona, and Austerlitz (2016) estimated N_e using an Approximate Bayesian Computation Approach and whole genomes of 15–25 individuals from each of four cattle breeds (Angus, Fleckvieh, Holstein, Jersey). They found evidence that historical domestication and modern breeding events were related to population decline. More recent statistical models promise the possibility to estimate hN_e from lcWGR and Pool-seq data, the first based on inbreeding identical-by-descent (IBD) tracts (Vieira, Albrechtsen, & Nielsen, 2016), and the latter on allele frequency changes between two temporal samples while correcting for the potential inflation of variance in allele frequency due to the two sampling steps involved in Pool-Seq experiments (i.e., during the sampling of individuals and the sequencing of pools) (Jonas, Taus, Kosiol, Schlotterer, & Futschik, 2016). In summary, the studies above illustrate how huWGR and lcWGR can facilitate inference on hN_e fluctuations and the tracing of historical demographic events that can help understand patterns of genetic diversity and structure in contemporary populations. These kinds of analyses can be extended to species of conservation interest. Haplotype-resolved genomes obtained with hrWGR approaches promise enhanced accuracy in demographic history and effective population size estimates as older long identical-by-descent portions of the genome can be assessed (Schiffels & Durbin, 2014; Snyder et al., 2015).

3.3 | Population structure and admixture (approach: huWGR, Pool-Seq, lcWGR)

One of the main goals in conservation biology is to maintain high genetic diversity in vulnerable species. Natural populations are commonly structured in local subpopulations. Genetic differences can arise among subpopulations over time as a result of the interplay between gene flow (e.g., reduced due to geographical distance or presence of barriers to dispersal), genetic drift and local adaptation (Allendorf, Luikart, & Aitken, 2013). Traditionally, the partitioning of genetic diversity within and among populations has been inferred using *F*-statistics, with F_{ST} being an estimate of the genetic differentiation among subpopulations (Holsinger & Weir, 2009). The assumption of this approach is that the average effect of neutral processes (e.g., gene flow and genetic drift) acting equally throughout the whole genome, can be estimated based on the average allele frequency at several neutral loci within subpopulations and over all subpopulations. A genomics approach, where multiple high-density loci are examined, allows instead the detection of the effect of different evolutionary forces (e.g., drift, selection) along the genome through the estimation of genetic diversity using a sliding window procedure (Allendorf, 2016). As the analysis of whole genomes provides the highest marker density, it allows the simultaneous evaluation of genomewide patterns in neutral loci that act as a record of demographic and historical events, as well as locus-specific effects that can be associated with natural selection, fitness and adaptation (Allendorf et al., 2013; Allendorf, 2016). This new perspective for the comparison of genetic diversity among populations is providing novel insights on how different evolutionary forces have affected particular loci and how differentiation could arise among natural populations despite gene flow. Below we describe some studies that estimated population structure and admixture based on neutral loci surveyed with WGR (Table 2).

Parejo et al., (2016) used huWGR data of 151 haploid drones to assess the degree of admixture between native European dark honey bee (*Apis mellifera mellifera*) and two introduced honey bee subspecies (*A. m. carnica* and buckfast) in four conservation areas of *A.*

millifera millifera in Switzerland and one in France. They found genetic differentiation between subspecies that coincided with geography and admixed individuals in protected areas. With the 50 most informative loci, they created a SNP panel for the genetic identification and monitoring of native and introduced bees. Fischer et al., (2017) compared genetic diversity and population differentiation estimates from Pool-seq and microsatellite data of 9 wild populations of the plant *A. halleri* in south-eastern Switzerland and northern Italy. They found no concordance of expected heterozygosity (H_e) estimates between marker types, and microsatellite allelic richness was a better descriptor of genomewide diversity than H_e. They found that a few thousand SNPs can provide a better estimate of genetic diversity and genetic differentiation among their populations than the 19 microsatellite loci tested. Velasco, Hough, Aradhya, and Ross-Ibarra (2016) conducted a genomewide analysis of the effects of domestication and mating system on genetic diversity of almond (*Prunus dulcis*) and peach (*P. persica*). With lcWGR data of 13 individuals from each species, they found that the genomewide nucleotide diversity was ~7-fold higher in almond than in peach, an excess of rare alleles in both species likely consistent with a recent population expansion event, no evidence of population bottleneck related to domestication and a strong genetic differentiation between species. Overall, these examples demonstrate WGR data is useful for the estimation of population structure and admixture in a variety of species.

3.4 | Signatures of selection, genetic basis of phenotypic traits and local adaptation (approach: hrWGR, huWGR, Pool-seq, lcWGR)

The identification of genomic regions involved in adaptation to local environmental conditions (local adaptation) is one of the main goals in evolutionary biology. This knowledge is crucial for conservation biology because of the importance of functional genetic diversity potentially linked with persistence in novel environments (Allendorf et al., 2010). Establishing the connection between genotype, phenotype and fitness is usually difficult though, and requires additional testing to verify the effect of presumably adaptive loci on fitness (Nielsen, 2009; Barrett & Hoekstra, 2011). We thus refer as "putatively" adaptive variants the parts of the genome that exhibit genetic signatures of selection for which the effect on fitness has not yet been tested.

There are three general strategies for the identification of loci under selection: (i) forward genetics (includes quantitative trait locus mapping and GWAS), when the phenotypic traits that underpin adaptation are known; (ii) reverse genetics (includes genome scans via genome-environment association [GEA] analyses and outlier loci tests), when the adaptive phenotype is unknown; and (iii) candidate genes examination. A complete explanation of these methods is reviewed elsewhere (Barrett & Hoekstra, 2011; Vitti, Grossman, & Sabeti, 2013; Pardo-Diaz, Salazar, & Jiggins, 2015). WGR is usually classified as reverse genetics as the traits under selection are generally unknown for nonmodel species. However, when there is particular interest in comparing contrasting phenotypes (e.g., ecotypes), a forward genetics approach following a GWAS-type comparison is possible, as it involves the directed screening of candidate genes discovered via genome scans. Genome scans are probably the most common method to detect signatures of selection in genomic data. Despite their proven power for this purpose, numerous considerations need to be accounted for when designing a genome scan experiment (see Box 5).

The advantage of WGR over other genomic approaches for the detection of loci under selection relies on the possibility to screen neutral and functional polymorphisms in high genomic resolution. Such high marker density is crucial for the identification of genetic signatures of selection such as reduced nucleotide diversity, extended linkage disequilibrium and high homozygosis (Ellegren, 2014). WGR has been used for the discovery and mapping of the genetic basis of phenotypic traits with adaptive importance (Table 2), for instance, the beak shape of Darwin finches (huWGR; Lamichhaney et al., 2015) or the age at maturity in Atlantic salmon (huWGR: Barson et al., 2015; Pool-seq: Ayllon et al., 2015). Other examples, all based on Pool-seq data, include: the red beak colour in canaries (Lopes et al., 2016), circadian timing in midges (Kaiser et al., 2016), genes affecting brain and neuronal development associated with domestication of rabbits (Carneiro et al., 2014), genes associated with breeding related traits of pathogen resistance and reproductive ability in two highly inbred chicken lines (Fleming et al., 2016), candidate genes potentially driving the morphological, life history and salt tolerance differences between ecotypes of the yellow monkey-flower plant (Gould, Chen, & Lowry, 2017), and immune-related genes in Atlantic salmon (Kjærner-Semb et al., 2016). Similarly, studies based on lcWGR data include the detection of a recent partial barrier (large inversion) to gene flow between subgroups of the mosquito *Anopheles gambiae* s.l. (Crawford, et al., 2016), and the identification of loci presumably involved in adaptation to high altitude and arid environments in native sheep (Yang et al., 2016).

Additionally, WGR allows the examination of SVs and mutations in regulatory elements (REs) (i.e., noncoding DNA sequences that control expression of neighbouring genes), providing a more complete genomewide spectrum of the amount and distribution of genetic variation. Both, SVs and mutations in REs have been shown to play an important role in the determination of phenotypic diversity, some of which could affect fitness (Wittkopp & Kalay, 2012), and the development of diseases (Melton, Reuter, Spacek, & Snyder, 2015). For example, a large deletion identified with Pool-seq data causes skeletal atavism in Shetland ponies (Rafati et al., 2016). A large chromosomal inversion discovered from huWGR data underlies the complex male mating morph diversity exhibited by the bird ruff (Küpper et al., 2015; Lamichhaney et al., 2015). And a chromosomal inversion is also responsible for the individual wing-pattern of diverse mimetic morphs in butterflies (Joron et al., 2011). Inversions play an important role because they reduce recombination, thus preventing the disruption of co-adapted gene complexes (Hoffmann & Rieseberg, 2008). Mutations in REs can affect gene expression having functional consequences in phenotypic traits (Wray, 2007; Wittkopp & Kalay, 2012). For example, mutations in the regulatory sequence of genes related to pigmentation produce different coloration patterns in cuticle, wings and abdomen of fruit flies

(*Drosophila melanogaster*; Wittkopp & Kalay, 2012). Changes in regulatory elements most likely are responsible for pelvic structure reduction in three-spine sticklebacks (*Gasterosteus aculeatus*; Shapiro et al., 2006) and for some human limb malformations (VanderMeer & Ahituv, 2011). Haplotype information obtained with hrWGR can help in the detection of haplotype-specific mutations and in the association of epigenetic factors and gene expression in tumours (Adey et al., 2013).

With WGR it is also possible to perform GEA analyses for the identification of genetic variation associated with adaptation to local conditions. For example, using Pool-seq data, Fischer et al., (2013) and Rellstab et al., (2016) evaluated the association of natural populations of *Arabidopsis halleri* with environmental factors in two time periods. For the first period, they analysed 5 populations and detected two million SNPs. They found 175 genes to be highly associated with some of the five environmental factors tested. For the second period, they extended the study to 18 populations covering a larger geographical area. Only 11 genes were found with the same association in both time periods, which could be a result of the alpine environment heterogeneity for which selection may be acting at the population level. Martinez Barrio et al., (2016) compared Pool-seq data of 20 Atlantic herring populations across the brackish Baltic Sea and the northeast Atlantic Ocean finding significant allele frequency differences at multiple loci between brackish water and oceanic populations and between spring and fall spawning populations, contrary to the common expectation of overall small genetic divergence within European populations previously observed in studies that only considered low-density loci (Larsson, Laikre, André, Dahlgren, & Ryman, 2010; Limborg et al., 2012; Teacher, André, Merilä, &Wheat, 2012; Teacher, André, Jonsson, & Merilä, 2013). Similarly, Lamichhaney et al. (2017) observed a pattern of differentiation between herring populations on both sides of the North Atlantic that comprised minute genetic differences at neutral loci but significant allele frequency differences between spring and autumn spawning populations at 6,333 SNPs, some of which are most likely associated with spawning time regulation. A total of 25% of such loci were shared between the American and European populations, and the unique loci found on each side of the ocean presumably result from local adaptation. In summary, the WGR approach is a powerful tool for the detection of signatures of selection, for uncovering the genetic basis of phenotypic traits and diseases and for the identification of signatures of local adaptation. huWGR, Pool-seq, lcWGR approaches can contribute to the understanding of the genetic variants and mechanisms underlying adaptive traits.

3.5 | Inbreeding depression, conservation breeding and restoration (approach: huWGR, potential contributions of Pool-seq and lcWGR)

Understanding the genetic basis and effects of inbreeding depression (defined as fitness reduction of the offspring resulting from the mating between closely related individuals) is a major goal in conservation biology as it affects the long-term viability of small isolated populations, whose persistence depends on targeted breeding, purging and restoration programmes (Allendorf et al., 2013). Numerous studies have tried to reveal the genetic basis underlying inbreeding depression in wild populations, however, the major obstacle for this has been the limitation to estimate the degree of individual inbreeding following traditional methods, as they require parental analysis over several generations. WGR analysis can solve these limitations by providing a large amount of genomic data per individual, which relaxes the need for parental analysis (reviewed by Kardos, Taylor, Ellegren, Luikart, and Allendorf (2016) and Hedrick, Hellsten, and Grattapaglia (2016)). For instance, Xue et al., (2015) obtained huWGR data of 44 wild individuals representing four subspecies of gorilla in Africa. They observed on average 34% homozygosis in individual genomes, which indicates extensive inbreeding most likely as a result of severe recent population decline. They also found very low genetic diversity in two of the four subspecies likely resulting from steady population declines over the past 100,000 years. Myburg et al., (2014) compared huWGR data on one outbred *Eucalyptus grandis* parent tree and 28 offspring obtained through self-

fertilization. The progeny retained high and different heterozygosity percentage (52%–79%, average 66%), in disagreement with an expectation of 50% homozygosis identical by descent (IBD) produced by selfing without selection. Hedrick et al. (2016) analysed the same data set finding that pseudo-overdominance most likely explained the observed inbreeding depression, which could be underlined by 100 or more genes of large effect associated with viability.

To our knowledge no study has thus far (June 2017) used the lcWGR approach for the study of inbreeding depression in wild populations, however, this method has been successful in the identification of causal mutations of three phenotypic traits in inbred rice varieties (Wang, Xu, et al., 2016) and the estimation of individual inbreeding coefficients (Vieira, Fumagalli, Albrechtsen, & Nielsen, 2013; Vieira et al. 2016). Thus, we envision lcWGR data is likely to be useful for this purpose in the near future. Inbreeding depression cannot be estimated from Pool-seq data as individual information is lost. However, this type of data can help in the identification of the genetic basis of phenotypic traits associated with inbreeding depression, as it has been used for the characterization of genetic variation underlying diseases (Rafati et al., 2016) and phenotypic traits in inbred organisms (Fleming et al., 2016). In conclusion, the examples presented here demonstrate WGR can help in the characterization of the genetic basis of inbreeding depression. This genetic information can be used for early diagnosis of inbreeding depression, assist in the planning of breeding programmes so as to avoid the inclusion of individuals carrying deleterious mutations that can affect recovery of captive or wild populations, and for the prediction of purging efficacy (Allendorf et al., 2010).

3.6 | Units of conservation, mixed stock analysis and genetic monitoring

Genetic patterns obtained from the analysis of neutral and adaptive genetic variation are useful for the delineation of conservation and management units (Funk et al., 2012). As previously shown, WGR approaches generate a large amount of neutral and putatively adaptive genetic markers. A subset of these loci can be used for the development of cost-effective genotyping tools suitable for the assessment of diverse aspects of interest in conservation and management (e.g., taxonomic status, hybrids, sex, carriers of genetic diseases, population structure, individual assignment and population of origin, among others). These tools can be incorporated in conservation plans of threatened species (Norman, Street, & Spong, 2013; Muñoz et al., 2015; Ivy, Putnam, Navarro, Gurr, & Ryder, 2016; Stetz et al., 2016; Fussi et al., 2016; Vandergast, 2017; Grossen, Biebach, Angelone-Alasaad, Keller, & Croll, 2017) and in management plans of commercially valuable species (Martinsohn & Ogden, 2009; Habicht et al., 2012; Bekkevold et al., 2015; Bradbury et al., 2015; Aykanat, Lindqvist, Pritchard, & Primmer, 2016; Sinclair-Waters, 2017).

4 | CONCLUDING REMARKS AND FUTURE DIRECTIONS

Our review synthesizes the advantages of WGR for addressing central questions in evolutionary biology that have not been fully answered using traditional techniques. The power of WGR resides in two main features: (i) it assesses neutral and functional genetic variation (including regulatory regions) at the highest genomic resolution among of current methods; and (ii) examines a wide variety of genetic variation, from one base changes to structural variants. The method thus facilitates the examination of the genetic basis of phenotypic traits and diseases as well as the detection of genetic signatures of natural selection, some of which can be related to local adaptation.

The scarcity of genomic resources for most species under conservation concern coupled with the still high cost of high-throughput sequencing and the elevated demand for computing resources, however, have limited the implementation of WGR in conservation biology. Some questions in conservation biology can be reasonably addressed using traditional or RRS approaches. The question then arises of when is a WGR approach justifiable? The answer depends on the research question, knowledge on the biological system, genomic resources available, the genetic architecture of phenotypic traits and ultimately on funding. If the research focus is the analysis of neutral processes, then WGR would not be necessary as RAD-seq methods would excel for an affordable price. If a highly accurate reconstruction of the species historical demography is sought, WGR would be justifiable as the estimation of coalescent events is benefited from the information provided by haplotype-resolved genomes. A good approximation to the species historical demography can also be achieved with the data generated by RRS, although a larger sample size would be necessary (Manthey, Campillo, Burns, & Moyle, 2016). The major motivation for using WGR is thus the detection of signatures of selection and the characterization of the genetic basis of phenotypic traits and diseases. RAD-seq can also be used for this purpose at the fraction of the genome screened, although its success may depend on the proportion of the genome covered. Ideally, this proportion should match the extension of linkage disequilibrium blocks, but this is usually unknown. Previous knowledge of the system and genomic resources can assist in the choice between RNA-seq, WES and WGR. For example, when there is a presumption that selection is operating on a specific tissue/life stage/time, then RNA-seq would be appropriate for assessing genetic variation in the genomic regions expressed at time of sampling. If the genes of interest are already described, then target capture and sequencing is the best strategy. When no candidate genes are known, a higher density screening such as WES or WGR would be preferable. When there is high confidence that selection is acting mostly on protein-coding parts of the genome, WES would be a cost-effective approach compared with WGR. When there is a notion that selection could be acting in regulatory elements or could be mediated by large structural variations, then WGR is likely the best choice as it provides the highest marker density and diversity in genetic variants assessed. Given that in general we do not know how selection is acting on a particular species, life stage, tissue or part of the genome, WGR should be considered as a starting point for the exploration of genomic diversity assuming sufficient funding and a reference genome are available. In the absence of reference genome, RAD-seq is an affordable alternative for the screening of neutral and putatively adaptive variation in a fraction of the genome (McKinney et al., 2017; Catchen et al., 2017) with some limitations (Hoban et al., 2016; Lowry et al., 2017a,b). Figure 4 summarizes the rationale for the selection of genomic approach as a function of availability of a reference genome, expected linkage block size and type of genetic variation of interest. Once regions of interest are detected and mapped, then a more affordable and scalable genotyping approach can be developed for massive individual genotyping on such loci.

The success of genome scans to detect adaptive variation depends on multiple factors including genetic architecture, effect size, sample size, percentage of the genome covered, effective population size and genetic drift, etc. (Box 5). Such information is usually unknown, making it hard to predict the success of RRS to reveal loci underlying a specific trait. Genome completeness, effect size and genetic architecture of a trait, and sampling design determine whether the entire genome is assessed, whether the genetic basis of a trait is traceable using genomics tools, and whether relevant individuals are included in the analysis, respectively. Also, genetic patterns resulting from demographic processes and drift may resemble those of local adaptation (Hoban et al., 2016). Therefore, outlier loci detected with genome scans should be treated as working hypothesis for further functional testing. Experimental evaluation of the effect of mutations on phenotypic traits and fitness is required to confirm and understand their adaptive nature. Such experimentation is unlikely to be performed on threatened species, although model species can be used instead, as many biochemical pathways are conserved across taxa (Andersson et al., 2012), and advances in genome-editing tools (i.e., CRISPR/Cas9) may facilitate functional testing (Varshney et al., 2015). Despite its limitations, WGR has proven to be an alternative for revealing adaptive loci and the genetic architecture of traits in a variety of organisms (Table 2), including humans (Durbin et al., 2010; Auer & Lettre, 2015; Field et al., 2016; Nielsen et al., 2017). Additionally, genomic data alone may not fully explain phenotypic variation. Epigenetic mechanisms (i.e., modifications of gene expression not due to changes in DNA sequences) play an important role in phenotype determination (Bossdorf, Richards, & Pigliucci, 2008; Richards, Bossdorf, & Pigliucci, 2010) suggesting a holistic approach would be ideal for better understanding phenotypic diversity and evolution.

Currently, there is no single WGR method fulfilling all requirements in conservation geneticists, and each method has its own limitations including sources of potential error and bias (Box 3 and 4). However, the implementation of good practices can control and minimize such biases, resulting in informative and reliable data sets that can be used for population genomics inference (details in Box 2; Fracassetti, Griffin, & Willi, 2015; Wang, Skoog, et al., 2016; Martinez Barrio et al., 2016). The field of genomics is rapidly changing, bringing new technologies and computing algorithms that promise solutions to present restrictions. For example, short-read sequences are of limited assistance for genome assembly, haplotype phasing and detection of large SVs. Long reads from third-generation sequencing can help overcome these limitations by resolving difficult parts of the genome (i.e., repetitive sequences and SVs) and by allowing the direct phasing of haplotypes. The relative low throughput, high error rate and cost have, however, restricted the use of third-generation sequencing platforms (Bleidorn, 2016), although improvements promising even higher throughput and lower cost and error rate are underway. This implies that in the future lower coverage per individual will likely be needed, high-throughput sequencing will likely be cheaper making it more accessible than otherwise, and larger sample sizes could be screened. Similarly, new computer tools and paradigms are being created, for example, graph-based genomes (The Computational Pan-genomics Consortium 2016; Paten, Novak, Eizenga, & Garrison, 2017) aim to overcome the current limitations of genome assembly which produce haplotype genome sequences, excluding a great amount of genomic variation and limiting variant detection.

For some conservation areas described above, the benefit of a dense and large number of markers is not clear. Effective population size estimation is a case in point (Waples, Larson, & Waples, 2016). Genomic information gathered via WGR can make important contributions to conservation planning and management of commercially exploited species, for instance, by helping in the delimitation and monitoring of evolutionary and/or management units and in the prioritization of imperilled populations. At the fast pace of computational and sequencing development, we can envision a not very distant future where simplified procedures, analysis and interpretation will make genomic tools accessible to managers (Garner et al., 2016; Shafer et al., 2015), the analysis of genomes will be performed in the field (Quick et al., 2016), and genomic analysis will become a routine task in many nonmodel species and fields.

DATA ACCESSIBILITY

Data retrieved in June 2017 from GenBank (https://www.ncbi.nlm.nih.gov/genome/browse/#) and shown in the plots in Figure 1b are stored in Dryad (https://doi.org/10.5061/dryad.3k8v9).

AUTHOR CONTRIBUTIONS

A.P.F.P. and D.E.R. conceived the study. A.P.F.P. compiled the papers, analysed the data and wrote the first draft of the manuscript with input from D.E.R.

ACKNOWLEDGEMENTS

We thank Leif Andersson, Sangeet Lamichhaney, Nima Rafati and Alvaro Martínez-Barrio (Uppsala University) for training on Pool-seq. Thanks to Martin Fischer, Simone Fior, Erik Garrison and Anders Albrechtsen for helpful discussions on Pool-seq and lcWGR during the 2017 Workshop on Bioinformatics for Adaptation Genomics (ETH-Zurich). Thanks to three anonymous reviewers that helped improve this manuscript. APFP thanks to the Vanier Canada Graduate Scholarship, the President's Award of Dalhousie University and the Nova Scotia Graduate Scholarship for funding. APFP and DER thank the Killam Trust. The study was funded by NSERC Discovery and Strategic grants to DER.

REFERENCES

Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., … Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, *500*, 207–211.

Agarwala, V., Flannick, J., Sunyaev, S., & Altshuler, D. (2013). Evaluating empirical bounds on complex disease genetic architecture. *Nature Genetics*, *45*, 1418–1427.

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., … Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*, R18.

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews. Genetics*, *12*, 363–376.

Allendorf, F. W. (2016). Genetics and the conservation of natural populations: Allozymes to genomes. *Molecular Ecology*, *38*, 42–49.

Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, *11*, 697–709.

Allendorf, F. W., Luikart, G., & Aitken, S. N. (2013). *Conservation and the genetics of populations*. Chichester, West Sussex: Wiley-Blackwell.

Anderson, E. C., Skaug, H. J., & Barshis, D. J. (2014). Next-generation sequencing for molecular ecology: A caveat regarding pooled samples. *Molecular Ecology*, *23*, 502–512.

Andersson, L. S., Larhammar, M., Memic, F., Wootz, H., Schwochow, D., Rubin, C.-J., … Kullander, K. (2012). Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. *Nature*, *488*, 642–646.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*, 81–92.

Angeloni, F., Wagemaker, N., Vergeer, P., & Ouborg, J. (2012). Genomic toolboxes for conservation biologists. *Evolutionary Applications*, *5*, 130–143.

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

Auer, P. L., & Lettre, G. (2015). Rare variant association studies: Considerations, challenges and opportunities. *Genome Medicine*, *7*, 16.

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., ... Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74.

Avise, J. C. (2010). Perspective: Conservation genetics enters the genomics era. *Conservation Genetics*, *11*, 665–669.

Aykanat, T., Lindqvist, M., Pritchard, V. L., & Primmer, C. R. (2016). From population genomics to conservation and management: A workflow for targeted analysis of markers identified using genome-wide approaches in Atlantic salmon *Salmo salar*. *Journal of Fish Biology*, *89* (6), 2658–2679.

Ayllon, F., Kjærner-Semb, E., Furmanek, T., Wennevik, V., Solberg, M. F., Dahle, G., ... Wargelius, A. (2015). The vgll3 locus controls age at maturity in wild and domesticated Atlantic salmon (*Salmo salar* L.) males. *PLoS Genetics*, *11*, 1–15.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, *3*(10), e3376.

Baker, M. (2012). De novo genome assembly: What every biologist should know. *Nature Methods*, *9*, 333–337.

Barrett, R. D. H., & Hoekstra, H. E. (2011). Molecular spandrels: Tests of adaptation at the genetic level. *Nature Review Genetics*, *12*, 767–780.

Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., ... Primmer, C. R. (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*, *528*, 405–408.

Baym, M., Kryazhimskiy, S., Lieberman, T. D., Chung, H., Desai, M. M., & Kishony, R. K. (2015). Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS ONE*, *10*, 1–15.

Beissinger, T. M., Rosa, G. J., Kaeppler, S. M., Gianola, D., & de Leon, N. (2015). Defining window-boundaries for genomic analyses using smoothing spline techniques. *Genetics Selection Evolution*, *47*, 30.

Bekkevold, D., Helyar, S. J., Limborg, M. T., Nielsen, E. E., Hemmer-Hansen, J., Clausen, L. A. W., & Carvalho, G. R. (2015). Gene-associated markers can assign origin in a weakly structured fish, Atlantic herring. *ICES Journal of Marine Science: Journal du Conseil*, *72*, 1790–1801.

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., ... Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*, *112*, 5473–5478.

Benestan, L. M., Ferchaud, A.-L., Hohenlohe, P. A., Garner, B. A., Naylor, G. J. P., Baums, I. B., ... Luikart, G. (2016). Conservation genomics of natural and managed populations: Building a conceptual and practical framework. *Molecular Ecology*, *25*(13), 2967–2977.

Berg, J. J., & Coop, G. (2014). A population genetic signal of polygenic adaptation (M W. Feldman, Ed.). *PLoS Genetics*, *10*, e1004412.

Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, *33*, 623–630.

Bernatchez, L. (2016). On the maintenance of genetic variation and adaptation to environmental change: Considerations from population genomics in fishes. *Journal of Fish Biology*, *89*(6), 1–38.

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, *13*, 403.

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., ... Smith, T. P. L. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, *53*, 1689–1699.

Bleidorn, C. (2016). Third generation sequencing: Technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*, *14*, 1–8.

Blischak, P. D., Wenzel, A. J., & Wolfe, A. D. (2014). Gene prediction and annotation in penstemon (Plantaginaceae): A workflow for marker development from extremely low-coverage genome sequencing. *Applications in Plant Sciences*, *2*, 1400044.

Boitard, S., Rodríguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data—An approximate Bayesian computation approach (MA Beaumont, Ed.). *PLOS Genetics*, *12*, e1005877.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). TRIMMOMATIC: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120.

Bossdorf, O., Richards, C. L., & Pigliucci, M. (2008). Epigenetics for ecologists. *Ecology Letters*, *11*, 106–115.

Bradbury, I. R., Hamilton, L. C., Dempson, B., Robertson, M. J., Bourret, V., Bernatchez, L., & Verspoor, E. (2015). Transatlantic secondary contact in Atlantic Salmon, comparing microsatellites, a single nucleotide polymorphism array and restriction-site associated DNA sequencing for the resolution of complex spatial structure. *Molecular Ecology*, *24*, 5130–5144.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., ... Korf, I. F. (2013). ASSEMBLATHON 2: Evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*, 10.

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Reviews Genetics*, *12*, 703–714.

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

Bustamante, C. D., Wakeley, J., Sawyer, S., & Hartl, D. L. (2001). Directional selection and the site-frequency spectrum. *Genetics*, *159*, 1779–1788.

Cariou, M., Duret, L., & Charlat, S. (2016). How and how much does RAD-seq bias genetic diversity estimates? *BMC Evolutionary Biology*, *16*, 240.

Carneiro, M., Rubin, C.-J., Di Palma, F., Albert, F. W., Alfoldi, J., Barrio, A. M., ... Andersson, L. (2014). Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science*, *345*, 1074–1079.

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources*, *38*, 42–49.

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, *44*, 1–12.

Chan, C. X., & Ragan, M. A. (2013). Next-generation phylogenomics. *Biology Direct*, *8*, 3.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*, 80–92.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*, 1767–1771.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13.

Crawford, J. E., Riehle, M. M., Markianos, K., Bischoff, E., Guelbeogo, W. M., Gneme, A., ... Lazzaro, B. P. (2016). Evolution of GOUNDRY, a cryptic subgroup of *Anopheles gambiae* s.l., and its impact on susceptibility to Plasmodium infection. *Molecular Ecology*, *25*, 1494–1510.

Cutler, D. J., & Jensen, J. D. (2010). To pool, or not to pool? *Genetics*, *186*, 41–43.

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., … Pereira, R. J. (2016). Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, *30*, 1–11.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158.

De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., & Palumbi, S. R. (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.

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews. Genetics*, *6*, 361–375.

Dennenmoser, S., Vamosi, S. M., Nolte, A. W., & Rogers, S. M. (2017). Adaptive genomic divergence under high gene flow between freshwater and brackish-water ecotypes of prickly sculpin (*Cottus asper*) revealed by Pool-Seq. *Molecular Ecology*, *26*, 25–42.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*, 491–498.

Durbin, R. M., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., … Africa, W. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*, 1061–1073.

Durrett, R. (2008). *Probability models for DNA sequence evolution*. New York, New York, NY: Springer.

Earl, D., Bradnam, K., St. John, J., Darling, A., Lin, D., Fass, J., … Paten, B. (2011). ASSEMBLATHON 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, *21*, 2224–2241.

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

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

Evans, J. D., Brown, S. J., Hackett, K. J. J., Robinson, G., Richards, S., Lawson, D., … Zhou, X. (2013). The i5K initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *Journal of Heredity*, *104*, 595–600.

Ferretti, L., Ramos-Onsins, S. E., & Pérez-Enciso, M. (2013). Population genomics from pool sequencing. *Molecular Ecology*, *22*, 5561–5576.

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., … Pritchard, J. K. (2016). Detection of human adaptation during the past 2000 years. *Science*, *354*, 760–764.

Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K. K., … Widmer, A. (2017). Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics*, *18*, 69.

Fischer, M. C., Rellstab, C., Tedder, A., Zoller, S., Gugerli, F., Shimizu, K. K., … Widmer, A. (2013). Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. *Molecular Ecology*, *22*, 5594–5607.

Fleming, D. S., Koltes, J. E., Fritz-Waters, E. R., Rothschild, M. F., Schmidt, C. J., Ashwell, C. M., … Lamont, S. J. (2016). Single nucleotide variant discovery of highly inbred Leghorn and Fayoumi chicken breeds using pooled whole genome resequencing data reveals insights into phenotype differences. *BMC Genomics*, *17*, 812.

Fonseca, N. A., Rung, J., Brazma, A., & Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics*, *28*, 3169–3177.

Fontanesi, L., Di Palma, F., Flicek, P., Smith, A. T., Thulin, C.-G., & Alves, P. C. (2016). LaGomiCs—Lagomorph Genomics Consortium: An International collaborative effort for sequencing the genomes of an entire mammalian order. *Journal of Heredity*, *107*, 295–308.

Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., … Wolf, J. B. W. (2016). Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nature Communications*, *7*, 11693.

Fracassetti, M., Griffin, P. C., & Willi, Y. (2015). Validation of pooled whole-genome re-sequencing in *Arabidopsis lyrata*. *PLoS ONE*, *10*, 1–15.

Frankham, R. (2010). Challenges and opportunities of genetic approaches to biological conservation. *Biological Conservation*, *143*, 1919–1927.

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

Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., & Nielsen, R. (2013). Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, *195*, 979–992.

Fumagalli, M., Vieira, F. G., Linderoth, T., & Nielsen, R. (2014). NGSTOOLS: Methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, *30*, 1486–1487.

Funk, W. C., McKay, J. K., Hohenlohe, P. A., & Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, *27*(9), 1–8.

Fussi, B., Westergren, M., Aravanopoulos, F., Baier, R., Kavaliauskas, D., Finzgar, D., … Kraigher, H. (2016). Forest genetic monitoring: An overview of concepts and definitions. *Environmental Monitoring and Assessment*, *188*, 493.

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

Gagnaire, P.-A., & Gaggiotti, O. E. (2016). Detecting polygenic selection in marine populations by combining population genomics and quantitative genetics approaches. *Current Zoology*, *62*, 1–14.

Garner, B. A., Hand, B. K., Amish, S. J., Bernatchez, L., Foster, J. T., Miller, K. M., … Luikart, G. (2015). Genomics in conservation: Case studies and bridging the gap between data and application. *Trends in Ecology & Evolution*, *31*(2), 1–3.

Garrison, E., Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907, 9.

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., … Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. *Molecular Ecology*, *22*, 3766–3779.

GIGA (2014). The Global Invertebrate Genomics Alliance (GIGA): Developing community resources to study diverse invertebrate genomes. *Journal of Heredity*, *105*, 1–18.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*, 333–351.

Gould, B. A., Chen, Y., & Lowry, D. B. (2017). Pooled ecotype sequencing reveals candidate genetic mechanisms for adaptive differentiation and reproductive isolation. *Molecular Ecology*, *26*, 163–177.

Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., … Shabalov, I. (2014). MycoCosm portal: Gearing up for 1,000 fungal genomes. *Nucleic Acids Research*, *42*, D699–D704.

Grossen, C., Biebach, I., Angelone-Alasaad, S., Keller, L. F., & Croll, D. (2017). Population genomics analyses of European ibex species show lower diversity and higher inbreeding in reintroduced populations. *Evolutionary Applications*, Accepted.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*, 1072–1075.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5* (10), e1000695.

Haasl, R. J., & Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, *25*, 5–23.

Habicht, C., Munro, A., Dann, T., Eggers, D., Templin, W., Witteveen, M., ... Volk, E. (2012). *Harvest and harvest rates of sockeye salmon stocks in fisheries of the Western Alaska Salmon Stock Identification Program (WASSIP), 2006–2008*. Anchorage, AK: Alaska Department of Fish and Game, Special Publication No. 12–24.

Haig, S. M., Miller, M. P., Bellinger, R., Draheim, H. M., Mercer, D. M., & Mullins, T. D. (2016). The conservation genetics juggling act: Integrating genetics and ecology, science and policy. *Evolutionary Applications*, *9*, 181–195.

Han, E., Sinsheimer, J. S., & Novembre, J. (2015). Fast and accurate site frequency spectrum estimation from low coverage sequence data. *Bioinformatics*, *31*, 720–727.

Hansen, T. F. (2006). The evolution of genetic architecture. *Annual Review of Ecology, Evolution, and Systematics*, *37*, 123–157.

Hatem, A., Bozdağ, D., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, *14*, 1–25.

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, *56*(2), 167–203.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*, 1–8.

Hedrick, P. W., Hellsten, U., & Grattapaglia, D. (2016). Examining the cause of high inbreeding depression: Analysis of whole-genome sequence data in 28 selfed progeny of *Eucalyptus grandis*. *New Phytologist*, *209*, 600–611.

Hedrick, P. W., & Miller, P. S. (1992). Conservation genetics: Techniques and fundamentals. *Ecological Applications*, *2*, 30–46.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., ... Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *The American Naturalist*, *188*(4), 379–397.

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.

Hohenlohe, P., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E., & Cresko, W. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, *6*, e1000862.

vonHoldt, B. M., Cahill, J. A., Fan, Z., Gronau, I., Robinson, J., Pollinger, J. P., ... Wayne, R. K. (2016). Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. *Science Advances*, *2*, e1501714.

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: Defining, estimating and interpreting F(ST). *Nature Reviews. Genetics*, *10*, 639–650.

Human Genome Sequencing Consortium I (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*, 931–945.

Ivy, J. A., Putnam, A. S., Navarro, A. Y., Gurr, J., & Ryder, O. A. (2016). Applying SNP-derived molecular coancestry estimates to captive breeding programs. *Journal of Heredity*, *107*, 403–412.

Jarvis, E., Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., ... Froman, D. P. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, *346*, 1311–1320.

Jensen, J. D., Foll, M., & Bernatchez, L. (2016). The past, present and future of genomic scans for selection. *Molecular Ecology*, *25*, 1–4.

Jonas, A., Taus, T., Kosiol, C., Schlotterer, C., & Futschik, A. (2016). Estimating the effective population size from temporal allele frequency changes in experimental evolution. *Genetics*, *204*, 723–735.

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*, 185–202.

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*, 203–206.

Kaiser, T. S., Poehn, B., Szkiba, D., Preussner, M., Sedlazeck, F. J., Zrim, A., ... Tessmar-Raible, K. (2016). The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature*, *540*, 69–73.

Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., & Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, *9*, 1205–1218.

Kim, S., Lohmueller, K., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., ... Nielsen, R. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, *12*, 231.

Kjærner-Semb, E., Ayllon, F., Furmanek, T., Wennevik, V., Dahle, G., Niemelä, E., ... Edvardsen, R. B. (2016). Atlantic salmon populations reveal adaptive divergence of immune related genes – A duplicated genome under selection. *BMC Genomics*, *17*, 610.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., ... Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*, 568–576.

Koepfli, K., Paten, B., Genome 10K Community of Scientists,, & O'Brien, S. J. (2015). The Genome 10K Project: A way forward. *Annual Review of Animal Biosciences*, *3*, 57–111.

Kofler, R., Langmuller, A. M., Nouhaud, P., Otte, K. A., & Schlotterer, C. (2016). Suitability of different mapping algorithms for genome-wide polymorphism scans with Pool-seq data. *Genes, Genomes, Genetics*, *6*, 1–20.

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, *27*(24), 3435–3436. https://doi.org/10.1093/bioinformatics/btr589.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., ... Schlötterer, C. (2011). Popoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE*, *6*(1), e15925.

Korneliussen, T. S. T., Albrechtsen, A., Nielsen, R., Nielsen, R., Paul, J., Albrechtsen, A., ... Ballinger, Dge. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*, 1–13.

Küpper, C., Stocks, M., Risse, J. E., Remedios, N., Farrell, L. L., Mcrae, B., ... Burke, T. (2015). A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Publishing Group*, *48*, 79–83.

Laehnemann, D., Borkhardt, A., & McHardy, A. C. (2016). Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Briefings in Bioinformatics*, *17*, 154–179.

Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., ... Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, *518*, 371–375.

Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoeppner, M. P., ... Andersson, L. (2015). Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nature Genetics*, *48*, 84–88.

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*, *114*(17), E3452–E3461.

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*, *109*, 19345–19350.

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

Larsson, L. C., Laikre, L., André, C., Dahlgren, T. G., & Ryman, N. (2010). Temporally stable genetic structure of heavily exploited Atlantic herring (*Clupea harengus*) in Swedish waters. *Heredity*, *104*, 40–51.

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*, *95*, 5–23.

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., . . . Schatz, M. (2016). Third-generation sequencing and the future of genomics. *bioRxiv*, doi: http://dx.doi.org/10.1101/048603

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*, **0**, 3.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The sequence alignment/map format and SAMTOOLS. *Bioinformatics*, *25*, 2078–2079.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., & Wang, J. (2009). SNP detection for massively parallel whole-genome resequencing. *Genome Research*, *19*(6), 1124–1132. https://doi.org/10.1101/gr.088013.108

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*, 1851–1858.

Li, H., & Wren, J. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, *30*, 2843–2851.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Matthew, P., Leong, J. S., . . . Vik, J. O. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, *533*, 200–205.

Limborg, M. T., Helyar, S. J., De Bruyn, M., Taylor, M. I., Nielsen, E. E., Ogden, R., . . . Bekkevold, D. (2012). Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). *Molecular Ecology*, *21*(5), 3686–3703.

Lopes, R. J., Johnson, J. D., Toomey, M. B., Ferreira, M. S., Araujo, P. M., Melo-Ferreira, J., . . . Carneiro, M. (2016). Genetic basis for red coloration in birds. *Current Biology*, *26*, 1427–1434.

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, *24*, 1031–1046.

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017a). Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, *17*, 142–152.

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017b). Responsible RAD: Striving for best practices in population genomic studies of adaptation. *Molecular Ecology Resources*, *38*, 42–49.

Lunter, G., & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*, 936–939.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., . . . Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18. https://doi.org/10.1186/2047-217x-1-18

Mace, G. M. (2004). The role of taxonomy in species conservation. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *359*, 711–719.

Macqueen, D. J., Primmer, C. R., Houston, R. D., Nowak, B. F., Bernatchez, L., Bergseth, S., . . . Yáñez, J. M. (2017). Functional annotation of all salmonid genomes (FAASG): an international initiative supporting future salmonid research, conservation and aquaculture. *BMC Genomics*, *18*(1), 484.

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Jakobsen, K. S., & Jentoft, S. (2017). Whole genome sequencing data and de novo draft assemblies for 66 teleost species. *Scientific Data*, *4*, 160132.

Manthey, J. D., Campillo, L. C., Burns, K. J., & Moyle, R. G. (2016). Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: A test in cardinalid tanagers (Aves, Genus: Piranga). *Systematic Biology*, *65*, 640–650.

Martin, M. (2011). CUTADAPT removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, *17*, 10–12.

Martin, S. H., & Jiggins, C. D. (2013). *Genomic studies of adaptation in natural populations*. In *eLS*, p. Chichester, UK: John Wiley & Sons Ltd.

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*, 1–32.

Martinsohn, J. T., & Ogden, R. (2009). FishPopTrace—Developing SNP-based population genetic assignment methods to investigate illegal fishing. *Forensic Science International: Genetics Supplement Series*, *2*, 294–296.

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*, 1297–1303.

McKinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: Comment on Breaking RAD by Lowry et al. (2016). *Molecular Ecology Resources*, *17*(3), 356–361.

McMahon, B. J., Teeling, E. C., & Höglund, J. (2014). How and why should we implement genomics into conservation? *Evolutionary Applications*, *7*, 999–1007.

Melton, C., Reuter, J. A., Spacek, D. V., & Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics*, *47*, 710–716.

Messer, P. W., & Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution*, *28*, 659–669.

Miller, M. R., Brunelli, J. P., Wheeler, P. A., Liu, S., Rexroad, C. E., Palti, Y., . . . Thorgaard, G. H. (2012). A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology*, *21*, 237–249.

Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M. A., Gaulton, K. J., . . . McCarthy, M. I. (2015). The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genetics*, *11*, 1–24.

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Verezemska, O., Isbandi, M., . . . Reddy, T. B. K. (2017). Genomes OnLine Database (GOLD) v. 6: Data updates and feature enhancements. *Nucleic Acids Research*, *45*, D446–D456.

Muñoz, I., Henriques, D., Johnston, J. S., Chavez-Galarza, J., Kryger, P., & Pinto, M. A. (2015). Reduced SNP panels for genetic identification and introgression analysis in the dark honey bee (*Apis mellifera mellifera*) (W Blenau, Ed,). *PLoS ONE*, *10*, e0124365.

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., . . . Schmutz, J. (2014). The genome of *Eucalyptus grandis*. *Nature*, *510*, 356–362.

Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Molecular Ecology*, *25*, 1058–1072.

Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., . . . Yamamoto, M. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature Communications*, *6*, 8018.

Neale, D. B., Wegrzyn, J. L., Stevens, K. A., Zimin, A. V., Puiu, D., Crepeau, M. W., . . . Langley, C. H. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, *15*, R59.

Nevado, B., Ramos-Onsins, S. E., & 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. (2009). Adaptionism - 30 years after gould and lewontin. *Evolution*, *63*, 2487–2490.

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, *541*, 302–310.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, *7*(7), e37558.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443–451.

Norman, A. J., Street, N. R., & Spong, G. (2013). De novo SNP discovery in the Scandinavian Brown Bear (*Ursus arctos*) (D Caramelli, Ed.). *PLoS ONE*, *8*, e81012.

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, *18*, 375–402.

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). QUALIMAP 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*, 292–294.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., . . . Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. *Genome Medicine*, *5*, 28.

Ouborg, N. J., Pertoldi, C., Loeschcke, V., Bijlsma, R. K., & Hedrick, P. W. (2010). Conservation genetics in transition to conservation genomics. *Trends in Genetics*, *26*, 177–187.

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. *Nature Reviews Genetics*, *12*, 87–98.

Pardo-Diaz, C., Salazar, C., & Jiggins, C. D. (2015). Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution*, *6*, 445–464.

Parejo, M., Wragg, D., Gauthier, L., Vignal, A., Neumann, P., & Neuditschko, M. (2016). Using whole-genome sequence information to foster conservation efforts for the european dark honey bee, *Apis mellifera mellifera*. *Frontiers in Ecology and Evolution*, *4*, 1–15.

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., . . . Price, A. L. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics*, *44*, 631–635.

Paten, B., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Research*, *27*, 665–676.

Pedersen, B. S., Layer, R. M., Quinlan, A. R., Li, H., Wang, K., Li, M., . . . Kang, H. (2016). VCFANNO: Fast, flexible annotation of genetic variants. *Genome Biology*, *17*, 118.

Pettersson, E., Lundeberg, J., & Ahmadian, A. (2009). Generations of sequencing technologies. *Genomics*, *93*, 105–111.

Pfeifer, S. P. (2017). From next-generation resequencing reads to a high-quality variant data set. *Heredity*, *118*, 111–124.

Phan, V., Gao, S., Tran, Q., & Vo, N. S. (2014) How genome complexity can explain the hardness of aligning reads to genomes. 2014 IEEE 4th International Conference on Computational Advances in Bio and Medical Sciences, ICCABS 2014, 16, 1–15.

Phillippy, A. M. (2017). New advances in sequence assembly. *Genome Research*, *27*, xi–xiii.

Primmer, C. R. (2009). From conservation genetics to conservation genomics. *Annals of the New York Academy of Sciences*, *1162*, 357–368.

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., . . . Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*, 228–232.

Rabbani, B., Tekin, M., & Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, *59*, 5–15.

Rafati, N., Andersson, L. S., Mikko, S., Feng, C., Pettersson, J., Janecka, J., . . . Evan, E. (2016). Large deletions at the SHOX locus in the pseudoautosomal region are associated with skeletal atavism in Shetland Ponies. *Genes, Genomes, Genetics*, *6*, 2213–2223.

Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S., & Pérez-Enciso, M. (2012). SNP calling by sequencing pooled samples. *BMC Bioinformatics*, *13*, 239.

Reinert, K., Langmead, B., Weese, D., & Evers, D. J. (2015). Alignment of next-generation sequencing reads. *Annual Review of Genomics and Human Genetics*, *16*, 133–151.

Rellstab, C., Fischer, M. C., Zoller, S., Graf, R., Tedder, A., Shimizu, K. K., . . . Gugerli, F. (2016). Local adaptation (mostly) remains local: Reassessing environmental associations of climate-related candidate SNPs in *Arabidopsis halleri*. *Heredity*, *118*, 1–9.

Richards, C. L., Bossdorf, O., & Pigliucci, M. (2010). What role does heritable epigenetic variation play in phenotypic evolution? *BioScience*, *60*, 232–237.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*, 24–26.

Rockman, M. V. (2012). The QTN program and the alleles that matter for evolution: All that's gold does not glitter. *Evolution*, *66*, 1–17.

Ronen, R., Udpa, N., Halperin, E., & Bafna, V. (2013). Learning natural selection from the site frequency spectrum. *Genetics*, *195*, 181–193.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., . . . Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*, R51.

Ruffalo, M., Koyutürk, M., Ray, S., & LaFramboise, T. (2012). Accurate estimation of short read mapping quality for next-generation genome sequencing. *Bioinformatics*, *28*, 349–355.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., . . . Roberts, M., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, *22*, 557–567.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*, 5463–5467.

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*, 919–925.

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.

Sedlackova, T., Repiska, G., Celec, P., Szemes, T., & Minarik, G. (2013). Fragmentation of DNA affects the accuracy of the DNA quantitation by the commonly used methods. *Biological Procedures Online*, *15*, 5.

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference (M Gilbert, Ed.). *Methods in Ecology and Evolution*, *8*(8), 907–917.

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., . . . Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, *30*, 78–87.

Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jónsson, B., . . . Kingsley, D. M. (2006). Corrigendum: Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, *439*, 1014.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*, 1135–1145.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews. Genetics*, *15*, 121–132.

Sinclair-Waters, M. (2017). *Genomic perspectives for conservation and management of Atlantic cod in costal Labrador*. (Unpublished Master's Thesis). Dalhousie University, Halifax, Canada.

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

Snyder, M. W., Adey, A., Kitzman, J. O., & Shendure, J. (2015). Haplotype-resolved genome sequencing: Experimental methods and applications. *Nature Reviews Genetics*, *16*, 344–358.

Snyder-Mackler, N., Majoros, W. H., Yuan, M. L., Shaver, A. O., Gordon, J. B., Kopp, G. H., … Tung, J. (2016). Efficient genome-wide sequencing and low-coverage pedigree analysis from noninvasively collected samples. *Genetics*, *203*, 699–714.

Steiner, C. C., Putnam, A. S., Hoeck, P. E. A., & Ryder, O. A. (2013). Conservation genomics of threatened animal species. *Annual Review of Animal Biosciences*, *1*, 261–281.

Stetz, J. B., Smith, S., Sawaya, M. A., Ramsey, A. B., Amish, S. J., Schwartz, M. K., & Luikart, G. (2016). Discovery of 20,000 RAD–SNPs and development of a 52-SNP array for monitoring river otters. *Conservation Genetics Resources*, *8*, 299–302.

Straub, S. C. K., Fishbein, M., Livshultz, T., Foster, Z., Parks, M., Weitemier, K., … Liston, A. (2011). Building a model: Developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics*, *12*, 211.

Teacher, A. G., André, C., Jonsson, P. R., & Merilä, J. (2013). Oceanographic connectivity and environmental correlates of genetic structuring in Atlantic herring in the Baltic Sea. *Evolutionary Applications*, *6*, 549–567.

Teacher, A. G., André, C., Merilä, J., & Wheat, C. W. (2012). Whole mitochondrial genome scan for population structure and selection in the Atlantic herring. *BMC Evolutionary Biology*, *12*, 248.

The Computational Pan-genomics Consortium (2016). Computational pan-genomics: Status, promises and challenges. *Briefings in Bioinformatics*, 1–18.

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, *17*, 194–208.

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*, 178–192.

Tiffin, P., & Ross-Ibarra, J. (2014). Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology and Evolution*, *29*, 673–680.

Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics*, *13*, 36–46.

Trowsdale, J., & Knight, J. C. (2013). Major histocompatibility complex genomics and human disease. *Annual Review of Genomics and Human Genetics*, *14*, 301–323.

Tung, J., Zhou, X., Alberts, S. C., Stephens, M., & Gilad, Y. (2015). The genetic architecture of gene expression levels in wild baboons. *eLife*, *4*, 1–22.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics* (pp. 11.10.1–11.10.33). Hoboken, NJ, USA: John Wiley & Sons Inc.

Vandergast, A. (2017). *Incorporating genetic sampling in long-term monitoring and adaptive management in the San Diego County Management Strategic Plan Area, southern California*. U.S. Geological Survey Open-File Report 2017–1061: US.

VanderMeer, J. E., & Ahituv, N. (2011). cis-regulatory mutations are a genetic cause of human limb malformations. *Developmental Dynamics*, *240*, 920–930.

Varshney, G. K., Pei, W., LaFave, M. C., Idol, J., Xu, L., Gallardo, V., … Burgess, S. M. (2015). High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Research*, *25*, 1030–1042.

Vatsiou, A. I., Bazin, E., & Gaggiotti, O. E. (2016). Detection of selective sweeps in structured populations: A comparison of recent methods. *Molecular Ecology*, *25*, 89–103.

Veeckman, E., Ruttink, T., & Vandepoele, K. (2016). Are we there yet? Reliably estimating the completeness of plant genome sequences. *The Plant Cell*, *28*, 1759–1768.

Velasco, D., Hough, J., Aradhya, M., & Ross-Ibarra, J. (2016). Evolutionary genomics of peach and almond domestication. *Genes, Genomes, Genetics*, *6*, 3985–3993.

Vieira, F. G., Albrechtsen, A., & Nielsen, R. (2016). Estimating IBD tracts from low coverage NGS data. *Bioinformatics*, *32*, 2096–2102.

Vieira, F. G., Fumagalli, M., Albrechtsen, A., & Nielsen, R. (2013). Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. *Genome Research*, *23*, 1852–1861.

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, *47*, 97–120.

von der Heyden, S., Beger, M., Toonen, R. J., van Herwerden, L., Juinio-Meñez, M. A., Ravago-Gotanco, R., … Bernardi, G. (2014). The application of genetics to marine management and conservation: Examples from the Indo-Pacific. *Bulletin of Marine Science*, *90*, 123–158.

Wall, J. D., Schlebusch, S. A., Alberts, S. C., Cox, L. A., Snyder-Mackler, N., Nevonen, K. A., … Tung, J. (2016). Genomewide ancestry and divergence patterns from low-coverage sequencing data reveal a complex history of admixture in wild baboons. *Molecular Ecology*, *25*, 3469–3483.

Wang, J., Skoog, T., Einarsdottir, E., Kaartokallio, T., Laivuori, H., Grauers, A., … Jiao, H. (2016). Investigation of rare and low-frequency variants using high-throughput sequencing with pooled DNA samples. *Scientific Reports*, *6*, 33256. https://doi.org/10.1038/srep33256.

Wang, H., Xu, X., Vieira, F. G., Xiao, Y., Li, Z., Wang, J., … Chu, C. (2016). The power of inbreeding: NGS-Based GWAS of rice reveals convergent evolution during rice domestication. *Molecular Plant*, *9*, 975–985.

Waples, R. K., Larson, W. A., & Waples, R. S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity*, *117*, 233–240.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015). Exome sequencing: Current and future perspectives. *Genes, Genomes, Genetics*, *5*, 1543–1550.

Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, *13*, 59–69.

Wong, P. B., Wiley, E. O., Johnson, W. E., Ryder, O. A., O'Brien, S. J., Haussler, D., … Murphy, R. W. (2012). Tissue sampling methods and standards for vertebrate genomics. *GigaScience*, *1*, 8.

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, *8*, 206–216.

Xue, Y., Prado-Martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., Szpak, M., … Scally, A. (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, *348*, 242–245.

Yang, J., Li, W. R., Lv, F. H., He, S. G., Tian, S. L., Peng, W. F., … Liu, M. J. (2016). Whole-genome sequencing of native sheep provides insights into rapid adaptations to extreme environments. *Molecular Biology and Evolution*, *33*, 2576–2592.

Yang, H., & Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and WANNOVAR. *Nature Protocols*, *10*, 1556–1566.

Ye, H., Meehan, J., Tong, W., & Hong, H. (2015). Alignment of short reads: A crucial step for application of next-generation sequencing data in precision medicine. *Pharmaceutics*, *7*, 523–541.

Zhang, G. (2015). Genomics: Bird sequencing project takes off. *Nature*, *522*, 34.

Zhao, S., Zheng, P., Dong, S., Zhan, X., Wu, Q., Guo, X., . . . Wei, F. (2012). Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nature Genetics*, *45*, 67–71.

Zhou, X., Wang, B., Pan, Q., Zhang, J., Kumar, S., Sun, X., . . . Li, M. (2014). Whole-genome sequencing of the snub-nosed monkey provides insights into folivory and evolutionary history. *Nature Genetics*, *46*, 1303–1310.

How to cite this article: Fuentes-Pardo AP, Ruzzante DE. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Mol Ecol*. 2017;26:5369–5406. https://doi.org/10.1111/mec.14264