STUDY DESIGNS

Harnessing the power of RADseq for ecological and evolutionary genomics

Kimberly R. Andrews[1], Jeffrey M. Good[2], Michael R. Miller[3], Gordon Luikart[4] and Paul A. Hohenlohe[5]

Abstract | High-throughput techniques based on restriction site-associated DNA sequencing (RADseq) are enabling the low-cost discovery and genotyping of thousands of genetic markers for any species, including non-model organisms, which is revolutionizing ecological, evolutionary and conservation genetics. Technical differences among these methods lead to important considerations for all steps of genomics studies, from the specific scientific questions that can be addressed, and the costs of library preparation and sequencing, to the types of bias and error inherent in the resulting data. In this Review, we provide a comprehensive discussion of RADseq methods to aid researchers in choosing among the many different approaches and avoiding erroneous scientific conclusions from RADseq data, a problem that has plagued other genetic marker types in the past.

[1]*Department of Fish and Wildlife Sciences, University of Idaho, 875 Perimeter Drive MS 1136, Moscow, Idaho 83844–1136, USA.*
[2]*University of Montana, Division of Biological Sciences, 32 Campus Dr. HS104, Missoula, Montana 59812, USA.*
[3]*Department of Animal Science, University of California, One Shields Avenue, Davis, California 95616, USA.*
[4]*Flathead Lake Biological Station, Fish and Wildlife Genomics Group, Division of Biological Sciences, University of Montana, Polson, Montana 59860, USA.*
[5]*Institute for Bioinformatics and Evolutionary Studies, Department of Biological Sciences, University of Idaho, Moscow, Idaho 83843, USA.*

Correspondence to K.R.A.
kimandrews@gmail.com

doi:10.1038/nrg.2015.28
Published online 5 Jan 2016

The development of restriction site-associated DNA sequencing (RADseq) has been deemed among the most important scientific breakthroughs in the past decade[1]. RADseq has fuelled studies in ecological, evolutionary and conservation genomics by harnessing the massive throughput of next-generation sequencing to uncover hundreds or thousands of polymorphic genetic markers across the genome in a single, simple and cost-effective experiment[2,3]. Similar to other reduced-representation sequencing approaches, RADseq targets a subset of the genome, therefore providing advantages over whole-genome sequencing, such as a greater depth of coverage per locus (and thus improved confidence in genotype calls) and the sequencing of higher numbers of samples for a given budget. Moreover, unlike many other methods for generating genome-wide data, RADseq does not require any prior genomic information for the taxa being studied. Consequently, RADseq has become the most widely used genomic approach for high-throughput single nucleotide polymorphism (SNP) discovery and genotyping in ecological and evolutionary studies of non-model organisms.

The term RADseq was originally used to describe one particular method[4] but has subsequently been adopted to refer to a range of related techniques that rely on restriction enzymes to determine the set of loci to be sequenced (BOX 1; Supplementary information S1 (figure)). These methods are also sometimes grouped under the term 'genotyping by sequencing' (GBS) techniques[5]. As with RADseq, the term GBS was originally used to describe one specific method[6]; however, this term is less descriptive than RADseq, which captures the defining feature of these methods, that is, the use of restriction enzymes to obtain DNA sequence at a genome-wide set of loci. Restriction enzymes have long been used to sample loci across the genome and to generate information on population-level variation[7,8], including genome-wide surveys for genetic variation in humans[9]. Whereas these previous techniques focused on polymorphisms within restriction cut sites or used Sanger sequencing, RADseq uses next-generation sequencing to generate sequence data adjacent to a large number of restriction cut sites[4,10,11]. RADseq loci can occur in all areas of the genome (that is, both coding and non-coding regions[2]), and individuals within or between closely related species generally share most loci due to the conservation of cut sites[12].

The many RADseq variations developed over the past several years promise to increase flexibility (for example, in the number of loci assayed) and decrease the cost and effort in ecological and evolutionary genomics studies. However, methodological differences can profoundly affect all steps of a genomic study, from study design and execution, to the resulting data output. All RADseq methods are broadly applicable across a wide range of taxa and scientific questions (BOX 2). Nonetheless, some techniques have been used more widely in certain systems, generally due to historical contingencies rather than to the relative suitability of the various approaches to different species (for example, complexity reduction of polymorphic sequences (CRoPS), GBS and reduced representation libraries (RRLs) have been primarily used in agricultural species[13]).

Box 1 | **Common RADseq methods**

Methods that sequence fragments adjacent to single restriction enzyme cut sites

Original restriction site-associated DNA sequencing (RADseq)[4,66] digests genomic DNA with one restriction enzyme, followed by mechanical shearing to reduce fragments to the appropriate length for sequencing, which (unlike other methods) creates variance in the fragment sizes at each locus. The 2bRAD[67,68] method uses type IIB restriction enzymes, which cleave DNA upstream and downstream of the recognition site, resulting in short fragments of uniform length (33–36 bp).

Methods that sequence fragments flanked by two restriction enzyme cut sites

- ***Single enzyme, indirect size selection.*** Genotyping by sequencing (GBS)[6] uses a common-cutter enzyme, and PCR preferentially amplifies short fragments. Sequence-based genotyping (SBG)[69] uses a rare cutter and one or two common cutters, and PCR preferentially amplifies short fragments.
- ***Double enzyme, indirect size selection.*** Complexity reduction of polymorphic sequences (CRoPS)[70] uses two enzymes and a proprietary library preparation kit (originally developed for 454 pyrosequencing).
- ***Single enzyme, direct size selection.*** Reduced representation libraries (RRLs)[10,71] are unique in using a blunt-end common-cutter enzyme, followed by a size selection step and a proprietary Illumina library preparation kit. Multiplexed shotgun genotyping (MSG)[56] uses one common-cutter enzyme and a size selection step. ezRAD[16] uses one or more common-cutter enzymes, and a proprietary kit for Illumina library preparation.
- ***Double enzyme, direct size selection.*** Double-digest RAD (ddRAD)[17] uses two restriction enzymes, with adaptors specific to each enzyme, and size selection by automated gel cut.

Variations on the above techniques include using methylation-sensitive enzymes[72]; adding more restriction enzymes to existing protocols to further reduce the set of loci[69,73]; adding a second digestion to eliminate adaptor dimers[14]; adapting RADseq techniques to other sequencing platforms such as Ion Torrent[73–75]; and other minor technical modifications[58,76].

In this Review, we primarily focus on the application of RADseq to ecological and evolutionary genetics in natural populations (BOX 2); however, much of our discussion is also relevant to other RADseq applications, such as trait-mapping in agricultural species[13]. We provide an overview of the diverse RADseq techniques that have been developed and highlight some of the research questions that these powerful methods can help to answer. We also discuss how technical differences among the many variant methods lead to trade-offs in experimental design and analysis, and describe general considerations for designing a RADseq study.

The RADseq family of methods

RADseq techniques share several basic steps (FIG. 1). All methods start with relatively high-molecular-weight genomic DNA[14] and begin by digesting it with one or more restriction enzymes. All methods add specific sequencing adaptors, or double-stranded oligonucleotides, that are required by all next-generation sequencing platforms. Adaptors added during RADseq protocols may contain barcodes, which are used to identify individual samples that are sequenced together (multiplexed) in a single library. Depending on the enzyme or enzymes used, RADseq protocols also reduce and/or select the sizes of DNA fragments that are optimal for next-generation sequencing.

RADseq methods differ in the order and details of enzyme digestion, adaptor ligation, barcoding and size selection, as well as the type of sequence data that can be produced at each locus. These differences can be used to categorize techniques into major groups (BOX 1). Below, we discuss important variations among methods at each step and some of the consequences for library preparation, the resulting data and subsequent bioinformatic analyses.

Starting genomic DNA. RADseq techniques have been optimized based on starting material comprised of high-molecular-weight genomic DNA, and thus these techniques can perform poorly with highly degraded genomic DNA[14]. For example, in methods without enzyme-specific adaptors (for example, ezRAD and CRoPS), smaller fragments of starting genomic DNA not adjacent to cut sites may end up in the sequencing library, thus wasting sequencing effort on non-RAD loci. The original RADseq technique[4] also requires higher molecular weight DNA than other methods, because the mechanical shearing step is most consistent and efficient with the relatively large fragments that remain after enzyme digestion (discussed below).

In general, a greater amount of starting DNA is often beneficial, as it can reduce the number of PCR cycles required and thus minimize the problem of PCR duplicates (discussed below). Some of the initial papers describing protocols recommended fairly large amounts of DNA (up to 1 µg per sample for original RADseq[15] or 5.5 µg for RRLs[10]); however, most RADseq methods are somewhat flexible in the total amount of DNA required per sample, and can often be implemented with only 50–100 ng of DNA per sample. One exception is the use of a PCR-free library preparation method, which requires large amounts of starting DNA (for example, 1–2 µg of DNA), as in one implementation of ezRAD[16].

Restriction enzyme digestion. RADseq protocols differ in the number of restriction enzymes used and the frequency with which these enzymes cut the genome, with 'common cutters' defined as restriction enzymes that cut more frequently than 'rare cutters,' generally as a result of the length of the enzyme recognition sequence (cut site). Techniques also fall into two major groups depending on how the set of sequenced loci relates to the distribution of enzyme cut sites across the genome. The original RADseq protocol and 2bRAD aim to produce sequence data at all cut sites for the restriction enzyme. By contrast, all other techniques depend on sequencing of the genomic fragments that are produced by two enzyme cut sites separated by a specified genomic distance (typically 300–600 bp apart, with the distance determined by direct or indirect size selection; see below). These cut sites may be from the same or different enzymes, depending on whether the method uses one or two enzymes (BOX 1). For each method, common-cutter or rare-cutter enzymes can be used to tailor the number of loci produced. For example, for the original RADseq protocol, a very rough estimate is that an 8-cutter will cut every $4^8 = 65{,}536$ bp, and a 6-cutter will cut every $4^6 = 4{,}096$ bp; this calculation can be adjusted to account for the GC content of the recognition sequence and the genome under study[2].

Restriction site-associated DNA sequencing
(RADseq). A method for sequencing thousands of genetic loci adjacent to restriction cut sites across the genome using massively parallel (next-generation) sequencing. Also known as genotyping by sequencing.

Next-generation sequencing
(Also known as massively parallel sequencing). Technology that first emerged around 2005 that sequences millions of DNA molecules simultaneously.

Depth of coverage
The number of sequence reads for a given locus or nucleotide site.

Adaptor ligation. RADseq techniques differ in how adaptors are constructed and ligated to DNA fragments, and also in how they are designed to ensure that only the target genomic DNA fragments (that is, those adjacent to restriction cut sites) are sequenced. In some cases, adaptors are designed to ligate only at the characteristic single-stranded sticky end that remains at restriction cut sites after digestion. Many Illumina sequencing-based RADseq protocols also use Y-adaptors that are structured to ensure that only fragments with the adaptor combinations that are required for sequencing are PCR amplified (FIG. 1). Some techniques adopt proprietary library preparation kits for adaptor ligation (for example, ezRAD, CRoPS and RRLs), which may increase the reliability as well as the cost of reagents for library construction. Using adaptors from proprietary kits can also lead to lower specificity in ligation because these adaptors do not ligate to the sticky ends, and therefore sequence

Box 2 | Ecological and evolutionary insights from RADseq data

Restriction site-associated DNA sequencing (RADseq) can be used to answer a wide variety of ecological, evolutionary and conservation-related questions.

Genomics of adaptation

Selection on colour pattern was found to be the most important factor maintaining butterfly hybrid zones by association-mapping analyses (see the figure, part **a**) and F_{ST} outlier tests (part **b**) conducted using RADseq data for two butterfly species (*Heliconius melpomene aglaope* and *H. m. amaryllis* (part **c**)); these analyses revealed that F_{ST} outliers primarily occurred in genomic regions associated with colour pattern variation. In part **a**, association scores are coloured according to the phenotypic characteristics illustrated in part **c**, and only the top 20 associated SNPs for each phenotype are shown. In part **b**, F_{ST} values are shown for all SNPs, with significant outliers in red or orange[47]. 'Unmapped' represents scaffolds that are not assigned to chromosomes in a *H. melpomene* genome assembly. Many other studies have also used RADseq to identify the genomic architecture of adaptation in other study systems (for example, REFS 24,48,77).

Inbreeding and genomic diversity

A study investigating heterozygosity–fitness correlations in seals found that genome-wide heterozygosity estimated using 14,585 RADseq SNPs had a nearly fivefold higher correlation with a fitness-associated trait than did 27 microsatellite loci[50]. RADseq genomic diversity estimates were also used to characterize the influence of social structure on autosome versus sex chromosome diversity in Tonkean macaque monkeys[78].

Effective population size (N_e)

Thousands of SNPs generated using RADseq were used to estimate N_e in salmon and smelt from western North America[79,80].

Population structure, phylogeography and conservation units

RADseq was used to develop a population-informative SNP panel to monitor stock composition in salmon and to delineate population units to harvest as discrete rather than mixed stocks[79,81]; see also REFS 82–84.

Introgression

Hohenlohe *et al.*[18] used RADseq to identify 3,180 species-diagnostic SNPs and to calculate admixture between a native and an invasive trout species; see also REFS 85,86.

Phylogenomics

RADseq data generated a highly resolved tree for 16 species of Lake Victoria cichlid fish, whereas previous analyses using amplified fragment length polymorphism (AFLP), microsatellites or a handful of sequence-based markers failed to resolve species-level relationships for these species[87].

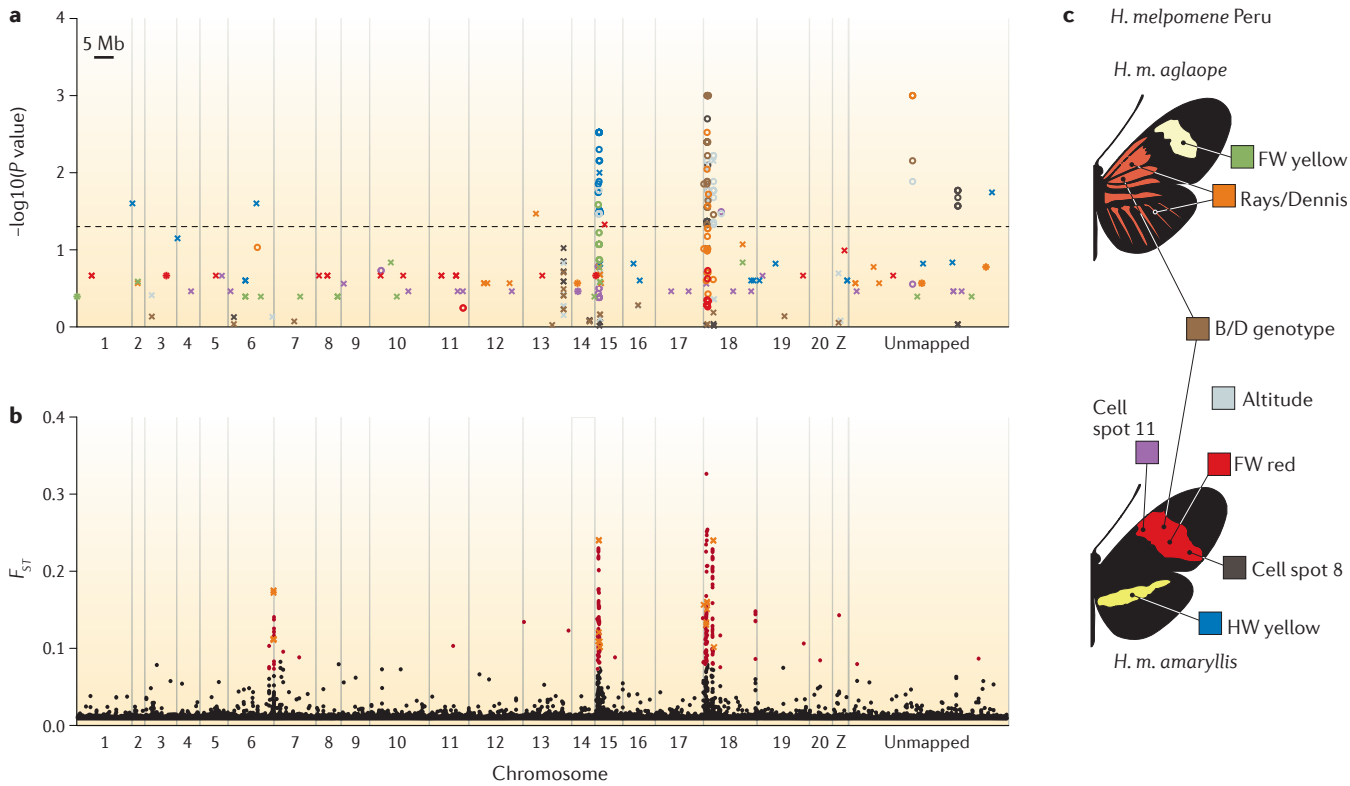

Figure reprinted from REF. 47, Cold Spring Harbor Laboratory Press.

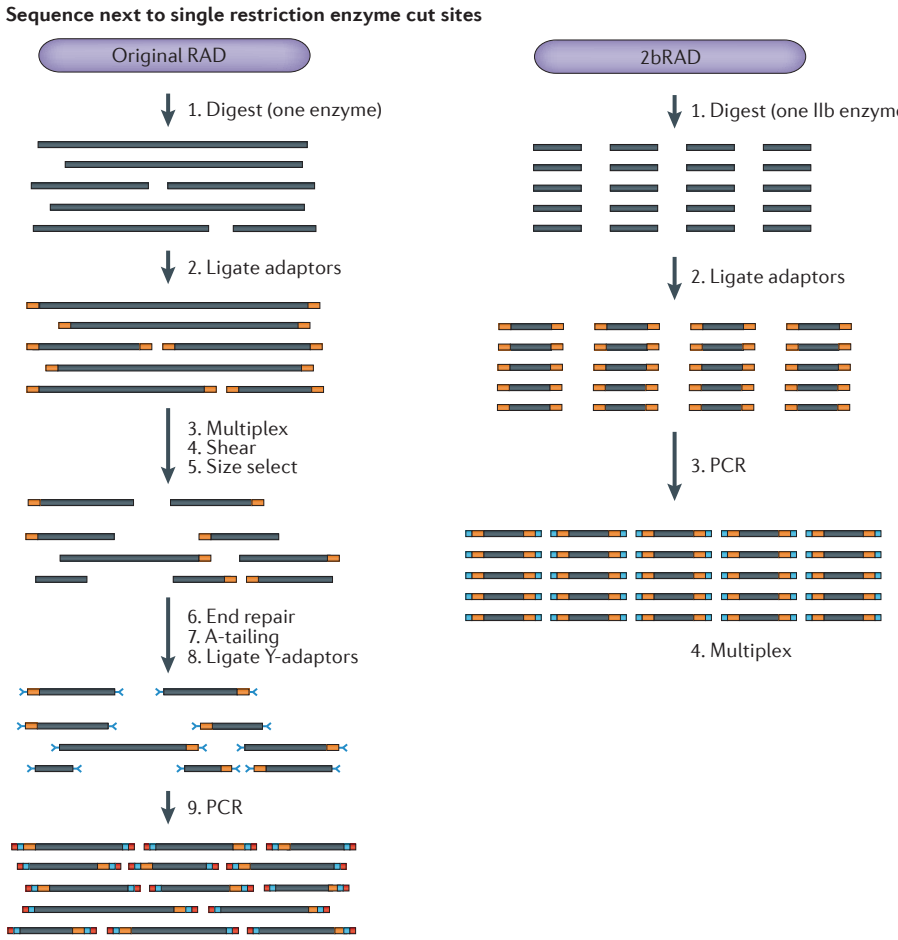

Figure 1 | **Step-by-step illustration of five RADseq library preparation protocols.** All protocols begin by digesting high-molecular-weight genomic DNA with one or more restriction enzymes. For most protocols, the sequencing adaptors (oligonucleotides) are added in two stages, with one set of oligonucleotides added during a ligation step early in the protocol, and a second set of oligonucleotides incorporated during a final PCR step. The second set of oligonucleotides extends the length of the total fragment to produce the entire Illumina adaptor sequences. By contrast, the original RADseq adds adaptors in three stages. For Illumina sequencing, the adaptors on either end of each DNA fragment must differ, and therefore some protocols (for example, original RADseq, double digest RAD (ddRAD) and ezRAD) use Y-adaptors that are structured to ensure that only fragments with different adaptors at either end are PCR-amplified (illustrated here as Y-shaped adaptors). Other protocols (for example, genotyping by sequencing (GBS)) simply rely on the fact that fragments without the correct adaptors will not be sequenced. To generate fragments of an ideal length for sequencing, most methods use common-cutter enzymes (for example, 4–6 bp cutters) to generate a wide range of fragment sizes, followed by a direct size selection (gel-cutting or magnetic beads, for example, ezRAD and ddRAD) or an indirect size selection (as a consequence of PCR amplification or sequencing efficiency, for example, GBS).

Box 3 | **Pooling**

The pooling of samples without individual barcoding during restriction site-associated DNA sequencing (RADseq) library preparation can enable the estimation of population allele frequencies at a reduced cost[63,88,89]. However, several sources of error are unique or magnified for pooled sequencing. Unequal representation of DNA from individual samples could lead to inaccurate allele frequency estimates[90,91], a problem that is exacerbated by PCR duplicates[89]. In addition, the identification of allele dropout, paralogues, mapping errors and hidden population structure is more difficult or even impossible for pooled data[63,89]. Similarly, distinguishing sequencing error from low-frequency alleles is more difficult for pooled data.

Errors that are caused by the unequal representation of individual samples in pooled sequencing libraries can be substantially reduced by using large per-pool sample sizes and depth of coverage, and by the removal of PCR duplicates[89,92,93]. The prevalence of PCR duplicates can be reduced by using a small number of PCR cycles, which should be feasible for pooled sequencing with a large starting amount of genomic DNA. Generating and comparing sequence data for replicate pools for each population can also help to identify and correct for the unequal representation of individual samples[89]. Nonetheless, this does not mitigate problems with identifying paralogues or allele drop-out.

Researchers should also be aware of restrictions to the analyses that can be conducted with pooled sequence data. Analyses requiring individual genotypes — such as assignment tests (for example, Bayesian clustering analyses with STRUCTURE[94]), relatedness tests or estimates of inbreeding coefficients — are not possible with this type of data. Several approaches for inferring population history or detecting selection depend on accurate estimates of linkage disequilibrium (LD)[95,96], and although there is limited power to estimate LD with the typically unphased data that results from individually barcoded RADseq data, it is not at all possible with pooled data. More fundamentally, pooling assumes that all samples in a pool are from a single well-mixed population, and cryptic population structure will be obscured if multiple groups are unknowingly combined within a pool.

data could be generated from fragments of degraded DNA that are not adjacent to restriction cut sites[16].

Size selection. For most protocols, the restriction digest reduces genomic DNA to a wide range of fragment lengths, and a size selection step is then used to isolate the fragments of ideal lengths for sequencing. This approach leads to key distinctions among RADseq protocols (BOX 1): for all the methods that sequence DNA fragments that are flanked by two cut sites the set of loci to be genotyped is further reduced by this size selection, because each potential locus has a characteristic fragment size that is determined by the distance between cut sites. In these techniques, size selection is done either indirectly, as a consequence of PCR amplification or sequencing efficiency (for example, GBS and CRoPS), or directly, using manual or automated gel cutting techniques or magnetic beads (for example, RRLs, multiplexed shotgun genotyping (MSG), ezRAD and double digest RAD (ddRAD)). For these methods, the consistency of size selection across libraries is crucial for producing data on a comparable set of loci across samples; inconsistency can lead to different sets of loci appearing in different libraries, resulting in wasted sequencing effort and high levels of missing genotypes.

By contrast, the original RADseq protocol and 2bRAD do not use size selection to reduce the set of loci to be sequenced; instead, all loci adjacent to restriction cut sites are targeted by these two methods. The original RADseq method follows digestion by a single enzyme with a mechanical shearing step to produce fragments that are appropriate for Illumina sequencing. This approach means that each sequenced fragment has a cut site at one end and a randomly sheared end at the other, and a range of fragment sizes is produced at each locus. As a result, the size selection step does not further reduce the set of loci but is used only to optimize Illumina sequencing efficiency and remove adaptor dimers. The 2bRAD method is unique among the RADseq protocols in that it uses IIB restriction enzymes to produce short fragments that are of equal size across all loci (33–36 bp).

Barcoding. The use of barcodes built into the adaptors allows the multiplexing of individual samples early in library preparation for some of the protocols; this multiplexing is sometimes called pooling, but should not be confused with the pooling of individuals into one barcode (BOX 3). During library preparation, as soon as barcoded adaptors are ligated to each sample, the samples can be multiplexed, which can greatly reduce the time and expense of subsequent steps in studies with large numbers of samples. The multiplexing of samples early in the library preparation requires the use of in-line barcodes. Adaptors from proprietary kits do not have in-line barcodes, and therefore custom-made adaptors are required for in-line barcoding. Many techniques can also be used with combinatorial barcoding, in which DNA fragments from each sample are identified by a unique combination of two different identifiers, typically one in-line barcode and one Illumina index (6–8 bp located near the middle of the adaptor) added at the PCR stage to the opposite end of the DNA fragment (for example, see the method used in Peterson *et al.*[17]). An alternative combinatorial barcoding strategy would be to use two Illumina indexes, one on each side of the DNA fragment. However, this strategy would not allow the multiplexing of samples early in the library preparation. Another alternative would be to use in-line barcodes on both sides of the DNA fragment; however, all Illumina libraries have at least one index, meaning that this approach would waste sequencing effort on a redundant in-line barcode. Combinatorial barcoding decreases the total number of adaptors required to distinguish individual samples, thus, for example, a set of 24 barcoded adaptors and 16 indexes can uniquely identify 384 samples in a sequencing lane.

Type of sequence data. Most RADseq techniques currently use Illumina sequencing. Illumina machines offer a range of sequence read lengths (currently 50–300 bp, and likely to increase further in the future) and also the option of either single-end sequencing, which produces one 'forward' read per DNA fragment, or paired-end sequencing, which produces one forward read and one 'reverse' read per fragment. These options can be applied to all RADseq libraries, although paired-end sequencing would not be beneficial for 2bRAD, which produces very short fragments (33–36 bp). For all other methods, forward reads begin from the restriction enzyme cut site, and longer reads typically capture more genomic sequence. For all the methods that target loci flanked by two cut sites, reverse reads begin at the second cut site and therefore these reads will align at identical locations in the genome for each locus.

Adaptors
Double-stranded oligonucleotides that must be ligated to DNA fragments before next-generation sequencing. Illumina adaptors contain regions that anneal to the flow cell, an 'index' sequence that act as a barcode to identify individual samples, and primer binding sites for bridge amplification and sequencing of the DNA fragment and indexes.

Barcodes
(Also known as in-line barcodes). Short unique sequences (typically 6–12 bp) used to identify individual samples. Occur at the end of the adaptor that is immediately adjacent to the genomic DNA fragment after adaptor ligation. The barcode is sequenced immediately before sequencing of the DNA fragment, and thus the barcode sequence will appear at the beginning of the sequence reads.

Sequencing library
DNA prepared for next-generation sequencing. The DNA must be an appropriate length for sequencing and must have sequencing adaptors ligated.

Sticky end
(Also known as DNA overhang). The string of single-stranded DNA that remains on the end of a DNA fragment that has been digested with a restriction enzyme. Some restriction enzymes produce blunt ends (double-stranded ends) rather than sticky ends.

IIB restriction enzymes
Restriction enzymes that cut DNA on both sides of the recognition site.

Pooling
Combining multiple individual samples into a DNA library with only one unique identifier (for example, one barcode or one index).

Combinatorial barcoding
Using two different barcoding methods, usually a standard Illumina index and an inline barcode. This method can reduce the number of adaptors that must be purchased, thus reducing library preparation cost.

Illumina index
A unique 6 bp or 8 bp sequence incorporated into Illumina adaptors that functions as a barcode to identify individual samples.

Single-end sequencing
Illumina sequencing of only one end of each DNA fragment.

Paired-end sequencing
Illumina sequencing of both ends of each DNA fragment.

Contigs
A group of overlapping sequence reads assembled to form a longer sequence.

Paralogues
Sequences originating through duplication within the genome.

Filtering
Removing unwanted sequence reads from a data set owing to low sequence quality, low depth of coverage, evidence for paralogy and other reasons.

By contrast, paired-end sequencing using the original RADseq protocol produces a very different type of data. The forward reads begin at the cut site but the reverse reads start from the randomly sheared end, typically 400–700 bp away. Therefore, the reverse reads at any given locus are staggered in length[18], and these data can be used to assemble long contigs. For example, these contigs can be as long as 1 kb if library fragments are tailored to be this length[15,19]. These RAD contigs improve the identification of paralogues[20], provide more sequence for BLAST searching of functionally important loci[18] and could provide haplotype data for genealogical or phylogenetic analysis. Longer contig sequences also enable the design of PCR primers or sequence capture probes to target loci of interest for further study[21,22].

For all methods, the read pairs produced by paired-end sequencing can overlap depending on read length and fragment size range, so that if fragments are less than 200–300 bp long (for example, some fragments produced using GBS with a common-cutter enzyme), increasing read lengths or using paired-end sequencing may not gain any genomic sequence information. However, overlapping read pairs can be used to improve genotyping accuracy towards the ends of the reads, which tend to have higher rates of sequencing error[23].

Bioinformatic analyses. Post-sequencing analyses will generally share several basic steps for data generated using all RADseq methods. Initial analyses include de-multiplexing and trimming of barcodes (if present), filtering reads based on the presence of the expected restriction enzyme cut site and sequence quality, and possibly trimming if quality declines towards the end of reads. For some RADseq methods, PCR duplicates can be removed during the initial analyses to improve the downstream genotyping accuracy (see below). If a reference genome is available, loci can then be identified by the alignment of sequence reads to this reference genome. Alternatively, loci can be assembled *de novo* by clustering similar sequence reads together and assuming that variation among reads at a locus represents either sequencing error or allelic variation. After locus discovery, long contigs can be generated for paired-end data obtained using the original RADseq (see above). Genotyping can be conducted using maximum likelihood[24] or Bayesian approaches[25,26]; maximum likelihood methods can require higher depth of coverage than Bayesian methods, particularly if Bayesian approaches make use of population-level allele frequencies to set prior probabilities on genotypes.

Several programs specifically designed for analysing RADseq data are available (for example, Stacks[27], pyRAD[28] and UNEAK[29], in addition to other publicly available scripts and pipelines). Stacks contains a number of flexible modules to conduct all parts of the analysis, from quality filtering and locus identification (either reference-aligned or *de novo*) to genotyping and calculating population genetic statistics. Specifically designed for phylogenetic applications, pyRAD conducts quality filtering and *de novo* locus identification and genotyping, with the advantage that it can handle insertion–deletion variation among alleles and may thus be better suited to studies with a broader taxonomic scale. UNEAK is part of the TASSEL pipeline for association mapping with GBS data[30] and uses a network-based SNP detection algorithm but is somewhat less flexible than other software in certain aspects such as read trimming and parameters for *de novo* locus identification. RADseq data can also be analysed using more generic software tools for quality filtering, alignment to a reference genome and genotyping.

Following genotyping, further filtering is typically recommended to remove loci and/or individual samples with large proportions of missing data. The appropriate level of filtering at this stage depends on the study goals and the subsequent analyses to be conducted, as these vary in their sensitivity to missing data and the sample size of individuals and loci. Several recent publications have highlighted how the details of RADseq data analysis, particularly the parameters used in *de novo* locus identification, can considerably affect analytical results[31–33]. Some of this work provides explicit recommendations for how to apply bioinformatic tools to RADseq data. Overall, it is crucial for researchers to vary the parameters used in all steps of the analysis, from quality filtering to locus identification and genotyping, to critically evaluate the sensitivity of the results and to optimize the analysis depending on the study goals.

Sources of error and bias

RADseq methods share some sources of sequencing and genotyping errors with all next-generation sequencing methods[34]. In addition, there are several unique potential sources of error and bias in RADseq methods, the impact of which can vary across library preparation protocols and statistical analyses.

Allele dropout and null alleles. Allele dropout manifests in RADseq when a polymorphism occurs at a restriction enzyme recognition site, resulting in a failure to cut the genomic DNA at that location. Alleles that lack the complete recognition site will not be sequenced and are therefore null alleles. If a SNP occurs within a null allele, the failure to sequence the allele could cause genotyping errors, with individuals heterozygous for the null allele appearing as homozygotes. The absence of a restriction cut site could also drive allele dropout for loci at neighbouring cut sites, because the post-digestion fragment lengths may fall outside the selected size range for methods that use size selection to reduce the set of loci (FIG. 2a).

The frequency of allele dropout increases with the cumulative length of the restriction enzyme recognition sites, owing to an increase in the probability of mutations in longer sequences[35]. Simulation studies also indicate that allele dropout increases with overall levels of polymorphism in the study system, and has a greater effect on data generated by ddRAD than on data generated by original-RADseq because the loci depend on the presence of two cut sites rather than one[35,36].

Genotyping errors that are caused by allele dropout can bias population genetic statistics through the underestimation of genomic diversity, the overestimation of F_{ST},

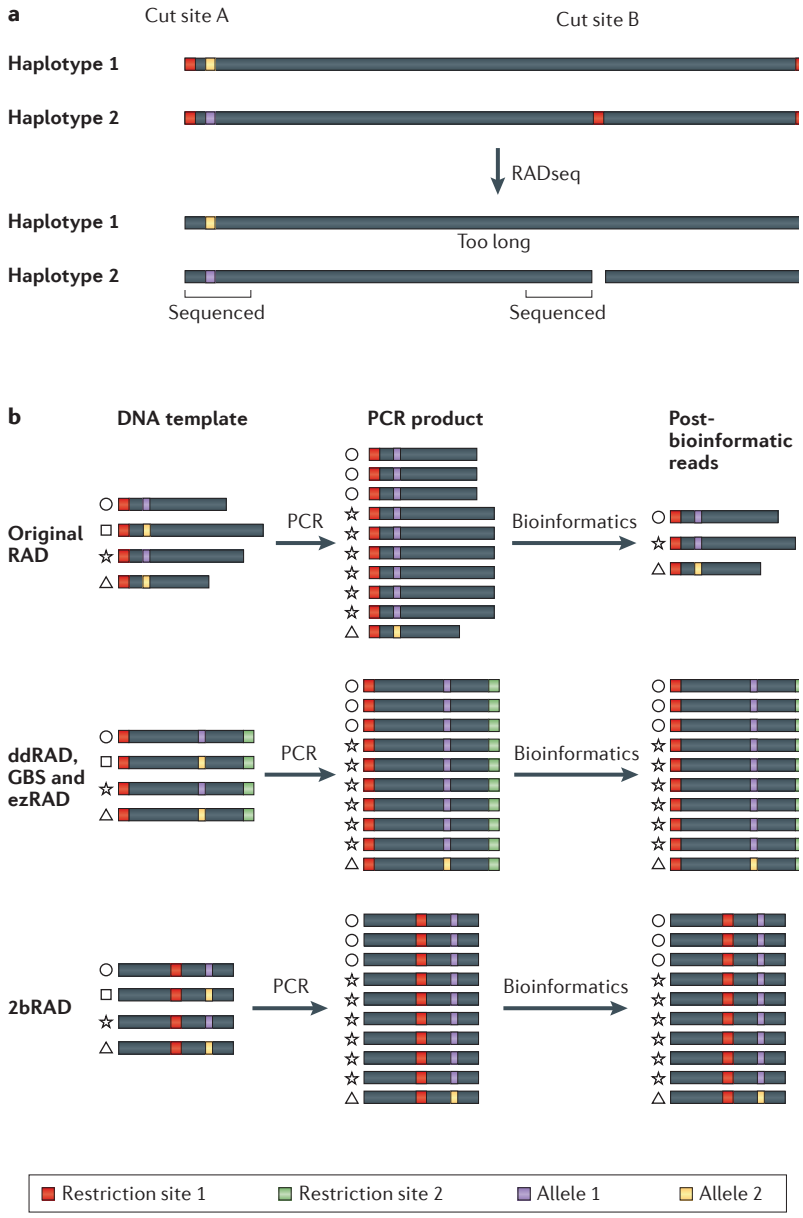

Figure 2 | **Sources of error and bias in RADseq data. a** | An example of allele dropout for a restriction site-associated DNA sequencing (RADseq) protocol that uses size selection to reduce the number of loci to be sequenced. Grey lines represent chromosomes within one individual, red squares represent restriction cut sites, coloured squares represent heterozygous SNPs, and square brackets represent genomic regions that are sequenced. Mutation in cut site B for haplotype 1 makes the post-digestion fragment containing the SNP too long to be retained during size selection for haplotype 1, eliminating the possibility of sequencing of any loci on that fragment, and causing the individual to appear homozygous at the heterozygous SNP. **b** | An example of fragments produced after PCR for one heterozygous locus for different RADseq protocols, and the reads retained after bioinformatic analyses. PCR duplicates are shown with the same symbol (circle, square, asterisk or triangle) as the parent fragment from the original template DNA. By chance, some alleles will amplify more than others during PCR. For all protocols, PCR duplicates will be identical in sequence composition and length to the original template molecule. For the original RADseq, this feature (that is, identical length) can be used to identify and remove PCR duplicates bioinformatically, because original template molecules for a given locus will not be identical in length. For alternative RADseq methods, this feature cannot be used to identify PCR duplicates, because all original template molecules for a given locus are identical in length. High frequencies of PCR duplicates can cause heterozygotes to appear as homozygotes or can cause PCR errors to appear as true diversity. Part **b** is adapted with permission from REF. 37, Wiley.

and an increase in false positives and false negatives in F_{ST} outlier tests[35,36]. However, there is evidence that the impact of these biases may be limited unless effective population sizes are large (N_e>10^5)[35]. F_{ST} biases can be largely compensated by removing the loci with null alleles from the data set. In theory, loci with null alleles should be identifiable by the high variance in depth of coverage across individual samples, as some individuals will lack one or both copies at the locus. However, many other factors also cause variance in depth of coverage (see below), so this is not always a reliable indicator of null alleles. Nevertheless, loci with a high prevalence of null alleles will be removed by many standard filtering practices that retain only loci that are successfully genotyped across a minimum percentage of individual samples. Although the removal of loci with null alleles should mostly compensate for biased F_{ST} estimates, it may do little to compensate for biased diversity estimates. Loci with null alleles are expected to occur more frequently in genomic regions with higher mutation rates and/or levels of standing genetic diversity, and thus the absence of these loci from the data set will tend to lead to the systematic underestimation of overall genomic diversity[36].

PCR duplicates and genotyping errors. Most next-generation sequencing library preparation protocols have a PCR step during which clonal DNA fragments (known as PCR duplicates) are generated from the original genomic DNA fragments (known as the parent fragments)[37,38]. During PCR, stochastic processes can cause one allele to amplify more than the other allele at a given locus in an individual sample. This potential skew can lead to downstream genotyping errors because heterozygotes can appear as homozygotes (FIG. 2a), or alleles that contain PCR errors can appear as true alleles (FIG. 2b). Studies report that PCR duplicates can occur at high frequencies in RADseq data (for example, in 20–60% of reads[18,37,38]). In theory, PCR should not systematically favour one allele over another at a given locus, and therefore parameters estimated from a large number of loci are unlikely to be substantially biased. However, analyses that require high genotyping accuracy at individual loci, such as outlier tests or parentage assignments, could produce erroneous results if PCR duplicates are present.

For sequence data generated using most next-generation sequencing protocols, PCR duplicates can be identified and removed bioinformatically to improve genotyping accuracy. This is possible in protocols with a mechanical or random enzymatic fragmentation step, as PCR duplicates can be identified as fragments that start and end at identical positions in the genome. Because of the mechanical shearing step, this method can also be used to identify PCR duplicates in sequence data generated using original RADseq with paired-end sequencing (FIG. 2b). In some circumstances (when the distance between forward and reverse reads is very short or local coverage is very high), this filter will remove fragments that are not duplicates but that, by chance, have the same start and end points. However, this should occur only rarely and should be conservative with respect to

genotyping accuracy. This method cannot be used to identify PCR duplicates in any RADseq protocols other than the original RADseq, because all fragments for a given locus will have identical start and stop positions[2].

Another recently developed method has shown promise for identifying PCR duplicates through the use of degenerate base regions within the sequencing adaptors to tag parent fragments before PCR[38–40]. This method could be incorporated into any protocol that uses custom-designed adaptors. An alternative method for dealing with PCR duplicates is to eliminate the PCR step of library prep altogether, as in ezRAD with Illumina PCR-free kits[16]. However, PCR-free kits are currently much more expensive and require a greater quantity of genomic DNA (1 μg) than other RADseq protocols.

Variance in depth of coverage among loci. Whereas PCR duplicates and allele dropout can cause genotyping errors as a result of the preferential sequencing of certain alleles within RADseq loci, several other phenomena can cause the preferential sequencing of certain loci over other loci. These phenomena should not cause genotyping errors, but will require greater overall sequencing effort to obtain sufficient depth for the loci that sequence less frequently. One well-known phenomenon is the preferential amplification of fragments based on GC content during PCR[2,41–43], and this bias should affect all RADseq methods that include a PCR step equally. Another phenomenon is the preferential amplification of shorter fragments over longer fragments. This issue will affect all RADseq methods that sequence fragments flanked by two cut sites (BOX 1), because each locus has a characteristic fragment length. This issue will not affect either 2bRAD because all loci are uniform in length or the original RADseq because each locus is represented by a variety of fragment lengths.

Another phenomenon that influences variance in depth of coverage among loci is driven by the mechanical shearing step in the original RAD. Fragments of <10 kb shear with lower efficiency, and therefore loci that originate from shorter restriction fragments will yield fewer reads than loci that originate from longer fragments[41]. However, this phenomenon should have less influence on the majority of original RADseq studies, which typically use rare cutters that digest genomic DNA to fragments >10 kb.

When coverage varies widely among loci, obtaining sufficient numbers of reads to accurately genotype the low-coverage loci will require an increase in the average depth of coverage across all loci. To accomplish this, the number of individuals multiplexed per sequence lane must be decreased, and this will increase the cost of the research project or decrease the number of individual samples that can be analysed. Alternatively, low-coverage loci could simply be removed from the data set if sufficient data can be obtained from high-coverage markers, and in practice this is common.

How to design a RADseq study

Designing a RADseq study for a particular application requires several major considerations to be taken into account regarding the most appropriate RADseq method, sampling and sequencing strategies, budget and other methodological details. The trade-offs among selected methods are summarized in TABLE 1.

Number of loci. The number of loci identified and genotyped by RADseq methods depends on the genome size, the frequency of the restriction cut sites in the genome and the number of cut sites that are targeted for sequencing. Computational tools are available to estimate the number of loci expected for each protocol[42,44]. RADseq methods that target all cut sites (original RAD and 2bRAD) or that use common-cutter enzymes without a direct size selection step (GBS) generally provide more loci, but the number can be adjusted by the choice of enzyme. By contrast, protocols that involve an explicit size-selection step (for example, ddRAD and ezRAD) can adjust the number of loci not only by choice of enzyme or enzymes, but also by changing the size range selected, and thus they typically have more flexibility to provide a smaller number of loci. Alternatively, another method of reducing the number of loci in any RADseq protocol is to design probes for a subset of informative RADseq loci and use these to capture and sequence selected loci (that is, RAD capture or Rapture[22]).

The optimal number of loci depends on the goals of the study. Studies that are focused on estimating neutral or genome-wide processes, such as phylogenetic relationships, geographic population structure, gene flow, introgression and individual inbreeding (identity by descent) often require only several hundred to a few thousand SNP-containing RADseq loci to adequately sample the genome[12,18,45,46]. By contrast, studies that seek to characterize functionally important regions across the entire genome, such as those exhibiting signatures of selection, require a larger set of markers (for example, up to tens or even hundreds of thousands of RADseq loci)[24,47,48]. In mapping studies, the optimal number of RADseq loci depends on the expected extent of linkage disequilibrium along the chromosomes and recombination patterns. For example, a laboratory F_2 cross or a very recently admixed population would require fewer loci than an outbred population, although statistical power may be increased with large numbers of progeny and more markers. For association mapping in an outbred population, many more markers would be required. Quantifying diversity patterns along chromosomal stretches (for example, runs of homozygosity) to estimate recent and historical effective population size and inbreeding also requires tens of thousands of loci[46,49,50].

Some biological factors can also increase the number of loci that should be targeted. Bottlenecked or small populations with low genomic variation may require the sequencing of more loci to accurately quantify the levels of variation. Genomes with a history of whole-genome or gene duplication, such as salmonid fish[51] or many plants[52], or genomes with high levels of transposable elements or other repeat sequences, such as some plant species[53], may also require large numbers of loci to compensate for the stringent filtering (removal) of problematic loci.

Allele dropout
Failure of an allele present in a sample to be detected by sequencing.

Null alleles
Alleles present in a sample that fail to be identified by genotyping. The presence of a null allele leads to allele dropout.

Linkage disequilibrium
Nonrandom association of alleles at different loci.

Table 1 | **Summary of trade-offs among five RADseq methods**

	Original RAD	2bRAD	GBS	ddRAD	ezRAD
Options for tailoring number of loci	Change restriction enzyme	Change restriction enzyme	Change restriction enzyme	Change restriction enzyme or size selection window	Change restriction enzyme or size selection window
Number of loci per 1 Mb of genome size*	30–500	50–1,000	5–40	0.3–200	10–800
Length of loci	≤1kb if building contigs; otherwise ≤300 bp‡	33–36 bp	<300 bp‡	≤300 bp‡	≤300 bp‡
Cost per barcoded or indexed sample	Low	Low	Low	Low	High
Effort per barcoded or indexed sample§	Medium	Low	Low	Low	High
Use of proprietary kit	No	No	No	No	Yes
Identification of PCR duplicates	With paired-end sequencing	No	With degenerate barcodes	With degenerate barcodes	No
Specialized equipment needed	Sonicator	None	None	Pippin Prep‖	Pippin Prep‖
Suitability for large or complex genomes¶	Good	Poor	Moderate	Good	Good
Suitability for *de novo* locus identification (no reference genome)#	Good	Poor	Moderate	Moderate	Moderate
Available from commercial companies	Yes	No	Yes	Yes	No

*Estimated as follows: original restriction site-associated DNA sequencing (RADseq), assuming either a 6-cutter or an 8-cutter; 2bRAD, assuming type IIB enzymes with recognition sites containing 5–7 specific nucleotides; genotyping by sequencing (GBS), data from Elshire *et al.*[6]; double digest RAD (ddRAD), data from Table 1 in Peterson *et al.*[17] and allowing for up to double the size range; ezRAD, data from Toonen *et al.*[16] for species with reference genomes. ‡Based on current single-end read-length limits in sequencing technology. §Assumes individual barcoding of many samples. ‖Can alternatively be used with standard gel equipment. ¶Based on the ability to reduce the total number of loci and lengths of loci. #Based on the lengths of loci to distinguish paralogues and duplicate sequence.

Type of sequence reads. Longer sequence reads and/or paired-end sequencing reads provide many advantages, including improved locus identification, discrimination of paralogous or repetitive sequence and BLAST searching for functionally important loci. For most RADseq protocols, sequence length is primarily limited by sequencing technology (for example, typically up to 150 bp reads with Illumina, but up to 300 bp in some cases). Many research questions can be sufficiently addressed with relatively short reads (for example, 100 bp) and single-end sequencing. However, as described above, longer RADseq loci can be obtained by assembling contigs from paired-end sequence reads with the original RAD (up to 1 kb[18]), and this method can be particularly advantageous for complex genomes in the absence of a reference genome. Of all the methods, 2bRAD produces the shortest reads (33–36 bp), so this technique is not recommended for *de novo* locus identification or in the case of large and complex genomes (for example, the human genome[54]), as the read length is essentially too short to enable reliable mapping.

Prior genomic resources. Prior reference sequence can provide numerous advantages for RADseq studies. A reference genome sequence, a poorly assembled set of genomic scaffolds or even a set of previously identified RAD loci can greatly improve the ability to filter paralogous or repetitive sequences, identify insertion–deletion variation and remove non-target DNA sequence (for example, bacterial contamination)[55]. A well-assembled reference genome provides further advantages. For example, mapping studies can use information on the physical positions of loci to infer haplotypes across larger chromosomal regions that cover multiple loci[56]. The GBS and MSG methods have been used in this way for trait mapping in model species, in which chromosomal blocks of parental ancestry are fairly large. Population genomic studies can use a reference genome assembly to conduct sliding window analyses and increase the statistical power to detect genomic regions of interest, such as regions under divergent selection between populations[24,48]. In the absence of a reference genome, long contigs generated with the original RADseq protocol should provide the greatest ability to distinguish paralogous or repetitive sequences[15,18,19].

Depth of sequencing coverage. Libraries from all RADseq methods can be sequenced to produce different depths of coverage, and the ideal depth for individually barcoded samples varies widely across studies. At one extreme, laboratory mapping studies with a well-assembled reference genome can be most efficient with very low coverage (<1×)[57]. Much higher coverage is required (for example, 10–20×) for confident *de novo* locus discovery and genotyping in diploids, although lower depth (for example, 5×) can be used if *de novo* assembly is conducted by combining reads from multiple

Sliding window analyses
Analyses in which summary statistics are calculated within a chromosomal segment (window), as the window is incrementally advanced along each chromosome.

Box 4 | **Alternatives to RADseq**

Two major alternative reduced representation next-generation sequencing methods to restriction site-associated DNA sequencing (RADseq) are transcriptome sequencing (RNA-seq) and targeted (probe-based) capture.

Transcriptome sequencing (RNA-seq)

RNA-seq sequences transcribed regions of the genome using RNA as a starting point for library preparation.

Advantages. RNA-seq can be used to quickly sequence thousands of functional genomic regions in almost any species with limited or no genomic resources[97]. Most transcripts can be annotated against existing genome databases[98], providing a much stronger functional context when compared with anonymous RADseq loci.

Disadvantages. RNA-seq provides limited opportunities to dynamically scale sequencing effort based on question or experimental design. Individual transcripts may differ by several orders of magnitude in relative abundance[99], complicating genotyping[100] and increasing sequencing costs. Functional annotation may be limited in taxonomic groups with poor database representation. RNA-seq requires high-quality samples, which can limit its feasibility for many studies.

Targeted (probe-based) capture

Targeted (probe-based) capture sequences pre-selected genomic regions using a DNA probe to isolate regions of interest.

Advantages. Targeted capture is highly scalable and able to sequence a single locus[101] or hundreds of thousands of loci[102,103]. Technical performance is typically very high[104], with low variance in sequencing coverage across regions and individuals[35,41,105]. Capture can be applied across moderate-to-deep evolutionary timescales[106–108] and on degraded DNA samples, making it popular for phylogenetic[33,109,110] and ancient DNA studies[111–116].

Disadvantages. Primary limitations for capture are the availability of genomic resources for designing probes, and the generally higher cost compared with RADseq or RNA-seq[60].

samples (although reads must then be separated by individual before genotyping). Higher coverage would be required in polyploid taxa because the coverage per haploid genome is reduced for an equal number of reads. Alternatively, in some cases, individuals may be pooled into single barcodes (BOX 3), with much lower coverage per individual because individual genotypes are not assigned.

Budget. The major expense in producing RADseq data is often the sequencing itself. The total sequencing effort is divided among the number of loci, the number of samples and populations, and the desired coverage per locus per individual. However, the different protocols can also considerably differ in the expense of library preparation, and in the way in which library preparation costs scale with the number of samples. For example, although the original RADseq protocol has a relatively large number of steps, samples are multiplexed early in the protocol and the subsequent steps are conducted on mixtures of up to 96 or more barcoded samples, so the marginal cost of increasing samples is minimized in terms of both time and money. By contrast, the cost of ezRAD scales roughly linearly with samples because multiplexing does not occur until the end, so this method may be most appropriate for small numbers of samples or pools of samples[16]. Some RADseq protocols also require an initial financial investment in specialized barcoded adaptors, although a single set of such oligonucleotides is often sufficient for a large number of libraries. In addition, some RAD protocols can require the purchase of specialized laboratory equipment. Original RADseq requires the use of a DNA sonicator, and RADseq protocols that use a direct size selection step (for example, ddRAD and ezRAD) can increase the precision and consistency of size selection, and can decrease the possibility of cross-contamination, by using a Pippin Prep[17] (Sage Science, Beverly, USA).

Comparability of data. A final consideration when designing a RADseq study is the consistency of the data across sequencing runs and across laboratories. Inconsistency in size selection could produce variation among libraries for methods that use size selection to reduce the set of loci. The consistency of different size selection techniques (automated or manual gel extraction versus bead-based selection) has not been rigorously quantified, but magnetic beads are probably much less consistent[58]. Methods that target every cut site (original RAD and 2bRAD) are generally expected to be more consistent across libraries; however, these methods are prone to other sources of error (discussed above). There can be some consistency in the loci genotyped even across methods, depending on the choice of restriction enzymes. For example, the loci sequenced using SbfI and EcoRI in a ddRAD protocol should be a subset of those sequenced using SbfI with original RAD.

Alternative or complementary approaches. Although RADseq has many benefits as a tool for SNP genotyping and discovery, it is not the best method of choice for every ecological and evolutionary study. Transcriptome sequencing (RNA-seq)[59] and targeted (probe-based) capture[60] are two major alternative reduced representation approaches that take advantage of next-generation sequencing (BOX 4). Whole-genome re-sequencing and whole-genome pooled sequencing are other alternatives that provide much more genomic information than reduced representation techniques[61–63]. However, despite the increasing feasibility of whole-genome re-sequencing for population studies, many ecological and evolutionary questions stand to gain little from such an increase in genome-wide data. For example, a RADseq study using several to tens of thousands of markers to detect selection based on allele frequency or linkage disequilibrium is more likely to be limited by the number of individuals sampled than by the density of markers.

Alternative genomic approaches can also be used to complement RADseq for more comprehensive or flexible investigation in a particular system. For example, the development of *de novo* reference genomes for non-model species is becoming increasingly feasible as sequencing and assembly technologies continue to improve[64,65], and such a reference provides numerous advantages for the analysis of RADseq data from population-level sampling[24,47,48,55]. Transcriptome sequencing can also complement RADseq data by targeting coding (and presumably functional) sequence, whereas RADseq interrogates both coding and non-coding loci. RADseq can also be used as the first step in a larger study to focus on significant loci. For example, RADseq can provide a genome-wide scan to identify candidate loci of interest, and sequence data at these loci can then be used to

design probes for sequence capture. Subsequent targeted sequencing could then be conducted on a large number of samples at greatly reduced cost per sample, and with poorer quality DNA.

Conclusions

RADseq techniques have enormous power and versatility for SNP discovery and genotyping in ecological and evolutionary genomics, but researchers should use careful consideration when choosing and applying these methods. Numerous RADseq protocols have been developed that differ not only in the technical details and cost of the library prep, but also in the types of data produced and the sources of genotyping error and bias. Therefore, protocols will differ in their suitability depending on the research questions, study systems and budget. Despite rapid changes in sequencing technology and costs, we anticipate that reduced-representation sequencing approaches such as RADseq will continue to be a crucial tool for genomics studies of natural populations for the foreseeable future. When implemented appropriately, RADseq approaches provide efficient, flexible and cost-effective avenues to unleash the power of next-generation sequencing technologies for gaining new insights into ecological, evolutionary and conservation-related questions.

1. [No authors listed]. Breakthrough of the year. Scorecard. *Science* **330**, 1608–1609 (2010).
2. Davey, J. W. *et al.* Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510 (2011).
Reviews methods for genomic marker discovery and genotyping using next-generation sequencing methods.
3. Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* **4**, 981–994 (2003).
4. Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3,** e3376 (2008).
Introduces one of the most widely used RADseq methods, which we describe as original RAD throughout.
5. Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R. & Hohenlohe, P. A. Genotyping-by-sequencing in ecological and conservation genomics. *Mol. Ecol.* **22**, 2841–2847 (2013).
6. Elshire, R. J. *et al.* A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. *PLoS ONE* **6,** e19379 (2011).
Introduces GBS, one of the most widely used RADseq methods.
7. Avise, J. C., Lansman, R. A. & Shade, R. O. Use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus *Peromyscus*. *Genetics* **92**, 279–295 (1979).
8. Brown, W. M. Polymorphism in mitochondrial DNA of humans as revealed by restricion endonuclease analysis. *Proc. Natl Acad. Sci. USA* **77**, 3605–3609 (1980).
9. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
10. Van Tassell, C. P. *et al.* SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* **5**, 247–252 (2008).
11. Wiedmann, R. T., Smith, T. P. L. & Nonneman, D. J. SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genet.* **9**, 81 (2008).
12. Cariou, M., Duret, L. & Charlat, S. Is RAD-seq suitable for phylogenetic inference? An *in silico* assessment and optimization. *Ecol. Evol.* **3**, 846–852 (2013).
13. Poland, J. A. & Rife, T. W. Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* **5**, 92–102 (2012).
14. Graham, C. *et al.* Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Mol. Ecol. Resour.* **15**, 1304–1315 (2015).
15. Etter, P. D., Preston, J. L., Bassham, S., Cresko, W. A. & Johnson, E. A. Local *de novo* assembly of RAD paired-end contigs using short sequencing reads. *PLoS ONE* **6,** e18561 (2011).
Introduces a method for generating long contigs from paired-end RADseq data.
16. Toonen, R. J. *et al.* ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ* **1**, e203 (2013).
17. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. & Hoekstra, H. E. Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7,** e37135 (2012).
Introduces ddRAD, one of the most widely used RADseq methods.
18. Hohenlohe, P. A. *et al.* Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Mol. Ecol.* **22**, 3002–3013 (2013).
19. Willing, E.-M., Hoffmann, M., Klein, J. D., Weigel, D. & Dreyer, C. Paired-end RAD-seq for *de novo* assembly and marker design without available reference. *Bioinformatics* **27**, 2187–2193 (2011).
20. Waples, R. K., Seeb, L. W. & Seeb, J. E. Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). *Mol. Ecol. Resour.* http://dx.doi.org/10.1111/1755-0998.12394 (2015).
21. Amish, S. J. *et al.* RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays. *Mol. Ecol. Resources* **12**, 653–660 (2012).
22. Ali, O. A. *et al.* RAD capture (Rapture): flexible and efficient sequence-based genotyping. *BioRxiv* http://dx.doi.org/10.1101/028837 (2015).
Extends RADseq with the addition of a sequence-capture step to target a subset of RAD loci, and also presents a substantially revised new version of the original RADseq protocol.
23. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
24. Hohenlohe, P. A. *et al.* Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* **6,** e1000862 (2010).
An early application of RADseq for population genomics, identifies loci under selection in multiple, independently derived freshwater stickleback populations.
25. Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y. & Wang, J. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE* **7**, e37558 (2012).
Introduces Bayesian methods for SNP-calling using the sample allele frequency spectra estimated from next-generation sequencing data.
26. Fumagalli, M. *et al.* Quantifying population genetic differentiation from next-generation sequencing data. *Genetics* **195**, 979–992 (2013).
27. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. & Cresko, W. A. Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–3140 (2013).
Introduces Stacks, a widely used software package for locus discovery, genotyping and population genomic analysis using RADseq data.
28. Eaton, D. A. R. PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. *Bioinformatics* **30**, 1844–1849 (2014).
29. Lu, F. *et al.* Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.* **9,** e1003215 (2013).
30. Bradbury, P. J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
31. Ilut, D. C., Nydam, M. L. & Hare, M. P. Defining loci in restriction-based reduced representation genomic data from nonmodel species: sources of bias and diagnostics for optimal clustering. *Biomed. Res. Int.* **2014**, 675158 (2014).
32. Mastretta-Yanes, A. *et al.* Gene duplication, population genomics, and species-level differentiation within a tropical mountain shrub. *Genome Biol. Evol.* **6**, 2611–2624 (2014).
33. Leaché, A. D. *et al.* Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated dna sequencing. *Genome Biol. Evol.* **7**, 706–719 (2015).
34. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
35. Gautier, M. *et al.* The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* **22**, 3165–3178 (2013).
Uses computer simulations to investigate the influence of allele dropout on population genomic statistics for RADseq data.
36. Arnold, B., Corbett-Detig, R. B., Hartl, D. & Bomblies, K. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* **22**, 3179–3190 (2013).
37. Andrews, K. R. *et al.* Trade-offs and utility of alternative RADseq methods: reply to Puritz *et al.* 2014. *Mol. Ecol.* **23**, 5943–5946 (2014).
38. Schweyen, H., Rozenberg, A. & Leese, F. Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (dbr) in sequencing adapters. *Biol. Bull.* **227**, 146–160 (2014).
39. Casbon, J. A., Osborne, R. J., Brenner, S. & Lichtenstein, C. P. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res.* **39,** e81 (2011).
40. Tin, M. M. Y., Rheindt, F. E., Cros, E. & Mikheyev, A. S. Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Mol. Ecol. Resour.* **15**, 329–336 (2015).
41. Davey, J. W. *et al.* Special features of RAD Sequencing data: implications for genotyping. *Mol. Ecol.* **22**, 3151–3164 (2013).
42. DaCosta, J. M. & Sorenson, M. D. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS ONE* **9,** e106713 (2014).
43. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40,** e72 (2012).
44. Lepais, O. & Weir, J. T. SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Mol. Ecol. Resour.* **14**, 1314–1321 (2014).
45. Cruaud, A. *et al.* Empirical assessment of RAD sequencing for interspecific phylogeny. *Mol. Biol. Evol.* **31**, 1272–1274 (2014).
46. Kardos, M., Luikart, G. & Allendorf, F. W. Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity* **115**, 63–72 (2015).
47. Nadeau, N. J. *et al.* Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. *Genome Res.* **24**, 1316–1333 (2014).

48. Ruegg, K., Anderson, E. C., Boone, J., Pouls, J. & Smith, T. B. A role for migration-linked genes and genomic islands in divergence of a songbird. *Mol. Ecol.* **23**, 4757–4769 (2014).
49. Kirin, M. *et al.* Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE* **5,** e13996 (2010).
50. Hoffman, J. I. *et al.* High-throughput sequencing reveals inbreeding depression in a natural population. *Proc. Natl Acad. Sci. USA* **111**, 3775–3780 (2014).
51. Allendorf, F. & Thorgaard, G. in *Evolutionary Genetics of Fishes Monographs in Evolutionary Biology* Ch. 1 (ed. Turner, B. J.) 1–53 (Springer, 1984).
52. Adams, K. L. & Wendel, J. F. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**, 135–141 (2005).
53. Charles, M. *et al.* Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. *Genetics* **180**, 1071–1086 (2008).
54. Palmieri, N. & Schloetterer, C. Mapping accuracy of short reads from massively parallel sequencing and the implications for quantitative expression profiling. *PLoS ONE* **4,** e6323 (2009).
55. Hand, B. K. *et al.* Genomics and introgression: discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Curr. Zool.* **61**, 146–154 (2015).
56. Andolfatto, P. *et al.* Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* **21**, 610–617 (2011).
57. Swarts, K. *et al.* Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. *Plant Genome* http://dx.doi.org/10.3835/plantgenome2014.05.0023 (2014).
58. Heffelfinger, C. *et al.* Flexible and scalable genotyping-by-sequencing strategies for population studies. *BMC Genomics* **15,** 979 (2014).
59. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
60. Jones, M. & Good, J. Targeted capture in evolutionary and ecological genomics. *Mol. Ecol.* http://dx.doi.org/10.1111/mec.13304 (2015).
61. Ellegren, H. *et al.* The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**, 756–760 (2012).
62. Kardos, M. *et al.* Whole genome resequencing uncovers molecular signatures of natural and sexual selection in wild bighorn sheep. *Mol. Ecol.* **24**, 5616–5632 (2015).
63. Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**, 749–763 (2014).
64. Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* **24**, 688–696 (2014).
65. Putnam, N. *et al.* Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *arXiv* http://arxiv.org/abs/1502.05331 (2015).
66. Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A. & Johnson, E. A. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* **17**, 240–248 (2007).
67. Wang, S., Meyer, E., McKay, J. K. & Matz, M. V. 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat. Methods* **9**, 808–812 (2012).
68. Guo, Y. *et al.* An improved 2b-RAD approach (I2b-RAD) offering genotyping tested by a rice (*Oryza sativa* L.) F2 population. *BMC Genomics* **15,** 956 (2014).
69. Truong, H. T. *et al.* Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS ONE* **7,** e37565 (2012).
70. van Orsouw, N. J. *et al.* Complexity reduction of polymorphic sequences (CRoPS (TM)): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE* **2**, e1172 (2007).
71. Greminger, M. P. *et al.* Generation of SNP datasets for orangutan population genomics using improved reduced-representation sequencing and direct comparisons of SNP calling algorithms. *BMC Genomics* **15,** 16 (2014).
72. Schield, D. R. *et al.* EpiRADseq: scalable analysis of genomewide patterns of methylation using next-generation sequencing. *Methods Ecol. Evol.* http://dx.doi.org/10.1111/2041-210X.12435 (2015).
73. Stolle, E. & Moritz, R. F. A. RESTseq — efficient benchtop population genomics with RESTriction fragment SEQuencing. *PLoS ONE* **8,** e63960 (2013).
74. Pukk, L. *et al.* Less is more: extreme genome complexity reduction with ddRAD using ion torrent semiconductor technology. *Mol. Ecol. Resour.* **15**, 1145–1152 (2015).
75. Recknagel, H., Jacobs, A., Herzyk, P. & Elmer, K. R. Double-digest RAD sequencing using Ion Proton semiconductor platform (ddRADseq-ion) with nonmodel organisms. *Mol. Ecol. Resour.* **15**, 1316–1329 (2015).
76. Chen, Q. *et al.* Genotyping by genome reducing and sequencing for outbred animals. *PLoS ONE* **8,** e67500 (2013).
77. Chutimanitsakun, Y. *et al.* Construction and application for QTL analysis of a restriction site associated DNA (RAD) linkage map in barley. *BMC Genomics* **12,** 4 (2011).
78. Evans, B. J., Zeng, K., Esselstyn, J. A., Charlesworth, B. & Melnick, D. J. Reduced representation genome sequencing suggests low diversity on the sex chromosomes of Tonkean macaque monkeys. *Mol. Biol. Evol.* **31**, 2425–2440 (2014).
79. Larson, W. A., Seeb, J. E., Pascal, C. E., Templin, W. D. & Seeb, L. W. Single-nucleotide polymorphisms (SNPs) identified through genotyping-by-sequencing improve genetic stock identification of Chinook salmon (*Oncorhynchus tshawytscha*) from western Alaska. *Can. J. Fisheries Aquat. Sci.* **71**, 698–708 (2014).
80. Candy, J. R. *et al.* Population differentiation determined from putative neutral and divergent adaptive genetic markers in Eulachon (*Thaleichthys pacificus*, Osmeridae), an anadromous Pacific smelt. *Mol. Ecol. Resourc.* **15**, 1421–1434 (2015).
81. Dann, T. H., Habicht, C., Baker, T. T. & Seeb, J. E. Exploiting genetic diversity to balance conservation and harvest of migratory salmon. *Can. J. Fisheries Aquat. Sci.* **70**, 785–793 (2013).
82. Emerson, K. J. *et al.* Resolving postglacial phylogeography using high-throughput sequencing. *Proc. Natl Acad. Sci. USA* **107**, 16196–16200 (2010).
83. Combosch, D. J. & Vollmer, S. V. Trans-Pacific RAD-Seq population genomics confirms introgressive hybridization in Eastern Pacific *Pocillopora* corals. *Mol. Phylogenet. Evol.* **88**, 154–162 (2015).
84. Gaither, M. R. *et al.* Genomic signatures of geographic isolation and natural selection in coral reef fishes. *Mol. Ecol.* **24**, 1543–1557 (2015).
85. Eaton, D. A. R. & Ree, R. H. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Syst. Biol.* **62**, 689–706 (2013).
86. Ford, A. G. P. *et al.* High levels of interspecific gene flow in an endemic cichlid fish adaptive radiation from an extreme lake environment. *Mol. Ecol.* **24**, 3421–3440 (2015).
87. Wagner, C. E. *et al.* Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol. Ecol.* **22**, 787–798 (2013).
88. Futschik, A. & Schlöetterer, C. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* **186**, 207–218 (2010).
89. Gautier, M. *et al.* Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Mol. Ecol.* **22**, 3766–3779 (2013).
90. Anderson, E. C., Skaug, H. J. & Barshis, D. J. Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Mol. Ecol.* **23**, 502–512 (2014).
91. Zhu, Y., Bergland, A. O., Gonzalez, J. & Petrov, D. A. Empirical validation of pooled whole genome population re-sequencing in *Drosophila melanogaster*. *PLoS ONE* **7,** e41901 (2012).
92. Lynch, M., Bost, D., Wilson, S., Maruki, T. & Harrison, S. Population-genetic inference from pooled-sequencing data. *Genome Biol. Evol.* **6**, 1210–1218 (2014).
93. Ferretti, L., Ramos-Onsins, S. E. & Perez-Enciso, M. Population genomics from pool sequencing. *Mol. Ecol.* **22**, 5561–5576 (2013).
94. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
95. Kayser, M., Brauer, S. & Stoneking, M. A genome scan to detect candidate regions influenced by local natural selection in human populations. *Mol. Biol. Evol.* **20**, 893–900 (2003).
96. Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**, 1566–1575 (2005).
97. Ekblom, R. & Galindo, J. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**, 1–15 (2011).
98. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
99. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
100. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93**, 641–651 (2013).
101. Briggs, A. W. *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**, 318–321 (2009).
102. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
103. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
104. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111–118 (2010).
105. Henning, F., Lee, H. J., Franchini, P. & Meyer, A. Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: benefits and pitfalls of using RAD markers for dense linkage mapping. *Mol. Ecol.* **23**, 5224–5240 (2014).
106. Good, J. M. *et al.* Comparative population genomics of the ejaculate in humans and the Great Apes. *Mol. Biol. Evol.* **30**, 964–976 (2013).
107. Hedtke, S. M., Morgan, M. J., Cannatella, D. C. & Hillis, D. M. Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. *PLoS ONE* **8,** e67908 (2013).
108. Bi, K. *et al.* Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* **13**, 403 (2012).
109. Faircloth, B. C. *et al.* Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* **61**, 717–726 (2012).
110. McCormack, J. E. *et al.* Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res.* **22**, 746–754 (2012).
111. Burbano, H. A. *et al.* Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* **328**, 723–725 (2010).
112. Bos, K. I. *et al.* A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* **478**, 506–510 (2011).
113. Avila-Arcos, M. C. *et al.* Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Sci. Rep.* **1,** 74 (2011).
114. Bos, K. I. *et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514**, 494–497 (2014).
115. Carpenter, M. L. *et al.* Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.* **93**, 852–864 (2013).
116. Castellano, S. *et al.* Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl Acad. Sci. USA* **111**, 6666–6671 (2014).

Acknowledgements
The authors thank M. Gaither, E. Carroll, A. Moura, R. Bracewell and M. Jones for helpful discussions. K.R.A. was supported by the University of Idaho College of Natural Resources, USA. P.A.H. received support from US National Institutes of Health (NIH) grant P30 GM103324 and NSF grant 1316549. J.M.G. is supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R01HD73439) and the National Institute of General Medical Sciences (R01GM098536) of the National Institutes of Health. G.L. was supported by grants from US National Science Foundation (DEB-0742181 and DEB-1067613) and NASA-(NNX14AB84G).

Competing interests statement
The authors declare no competing interests.

SUPPLEMENTARY INFORMATION
See online article: S1 (figure)
ALL LINKS ARE ACTIVE IN THE ONLINE PDF