STUDY DESIGNS

Genome-wide genetic marker discovery and genotyping using next-generation sequencing

John W. Davey, Paul A. Hohenlohe‡, Paul D. Etter§, Jason Q. Boone||, Julian M. Catchen‡ and Mark L. Blaxter*¶*

Abstract | The advent of next-generation sequencing (NGS) has revolutionized genomic and transcriptomic approaches to biology. These new sequencing tools are also valuable for the discovery, validation and assessment of genetic markers in populations. Here we review and discuss best practices for several NGS methods for genome-wide genetic marker development and genotyping that use restriction enzyme digestion of target genomes to reduce the complexity of the target. These new methods — which include reduced-representation sequencing using reduced-representation libraries (RRLs) or complexity reduction of polymorphic sequences (CRoPS), restriction-site-associated DNA sequencing (RAD-seq) and low coverage genotyping — are applicable to both model organisms with high-quality reference genome sequences and, excitingly, to non-model species with no existing genomic data.

**Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, King's Buildings, West Mains Road, Edinburgh, EH9 3JT, UK.*
‡Institute of Ecology and Evolution, University of Oregon, Pacific Hall, Eugene, Oregon 97403-5289, USA.
§Institute of Molecular Biology, University of Oregon, Klamath Hall, Eugene, Oregon 97403-1229, USA.
||Floragenex, Inc., 1,900 Millrace Drive, Eugene, Oregon 97403, USA.
¶The GenePool Genomics Facility, University of Edinburgh, Ashworth Laboratories, King's Buildings, West Mains Road, Edinburgh, EH9 3JT, UK.
Correspondence to J.W.D.
e-mail: john.davey@ed.ac.uk
doi:10.1038/nrg3012

Genetic markers — heritable polymorphisms that can be measured in one or more populations of individuals — lie at the heart of modern genetics and enable the study of important questions in population genetics, ecological genetics and evolution. In 2003, Luikart *et al.*[1] wrote: "The ideal molecular approach for population genomics should uncover hundreds of polymorphic markers that cover the entire genome in a single, simple and reliable experiment. Unfortunately, at present there is no such approach." Now, with the advent of next-generation sequencing (NGS), there are several such approaches, which are capable of discovering, sequencing and genotyping not hundreds but thousands of markers across almost any genome of interest in a single step[2], even in populations in which little or no genetic information is available.

This step change in marker density enables not only comprehensive genome-wide association studies for any organism, but also genome-wide studies on wild populations, with substantial benefits for conservation genetics and ecology[3]. Many biological questions can now be answered with high accuracy, for example, identifying recombination breakpoints for linkage mapping or quantitative trait locus (QTL) mapping, locating differentiated genomic regions between populations for quantitative genetics studies, genotyping large broods for marker-assisted selection or resolving the phylogeography of tens of wild populations. However, experimental design and data analysis for these new approaches can be complex[4]. Here, we aim to establish best practices for some of these approaches to simplify their application.

As with previous marker types, such as restriction fragment length polymorphisms (RFLPs)[5] and amplified fragment length polymorphisms (AFLPs)[6], many of these NGS methods depend on restriction enzymes to produce a reduced representation of a genome. We focus on the use of restriction enzymes combined with NGS for genome-wide marker discovery in new technologies such as reduced-representation sequencing, restriction-site-associated DNA sequencing (RAD-seq) and multiplexed shotgun genotyping (MSG), and we make recommendations for the use of these technologies in future studies. In order to provide detailed advice on best practice in the space available, we limit the discussion to restriction-enzyme-based methods, which define an unbiased, genome-wide set of markers and share a number of common design concerns. However, there are several methods for targeted marker development that may be of equal or greater utility for some studies. We briefly review these methods in BOX 1 and direct readers to the references there for more information.

Box 1 | **Targeted marker discovery methods**

The restriction-enzyme-based next-generation sequencing (NGS) methods described here produce a genome-wide, unbiased set of markers. There are several other NGS marker discovery methods that target particular regions of the genome, and these may provide an equally suitable or superior set of markers, depending on the application. Samples are typically pooled before sequencing, but each method can be adapted for barcoding of samples for individual genotyping[61].

RNA-seq

Next-generation cDNA sequencing (RNA-seq) makes it possible to sequence complete transcriptomes in almost any population or tissue[62,63]. Although often used to measure gene expression, RNA-seq has also been used to discover tens to hundreds of thousands of SNPs in human cell lines[64], bovine milk[65] and black cottonwood[66]. This can be done at similar costs to the restriction-enzyme-based methods discussed here, and is more likely to detect functional (for example, disease-related) SNPs[64]. However, inferring genotypes from expression data can be challenging[67], and normalizing the range of dynamic expression may be desirable[68]. Alternative transcripts can also make it difficult to infer genotypes, particularly when no reference genome is available, as *de novo* transcriptome assembly remains daunting[69,70].

Sequence capture

If the sequences of regions of interest are known, they can be targeted directly with sequence capture methods such as SureSelect, Nimblegen and Raindance[71–73]. Oligonucleotide baits are designed to bind to regions of interest and these regions are selected or enriched before sequencing. For example, whole exomes[74] or regions associated with particular diseases and traits[75] can be targeted. These methods are highly accurate when a high-quality, closely related reference sequence is available[75], but may be less suitable when the population is considerably diverged from the reference; this is because the reference-designed baits are less likely to bind strongly to the regions of interest and may result in biases against highly diverged regions.

Quantitative trait locus
(QTL). A locus that controls a quantitative phenotypic trait, identified by showing a statistical association between genetic markers surrounding the locus and phenotypic measurements.

Marker-assisted selection
The use of genetic markers to predict the inheritance of alleles at a closely linked trait locus.

Restriction fragment length polymorphism
(RFLP). A fragment-length variant that is generated through the presence or absence of a restriction enzyme recognition site. Restriction sites can be gained or lost by base substitutions, insertions or deletions.

Amplified fragment length polymorphism
(AFLP). A mapping method in which genomic DNA from different strains is PCR amplified using arbitrary primers. DNA fragments that are amplified in one strain, but not the other, are cloned, sequenced and used as polymorphic markers.

We begin by reviewing the promise of NGS and the continuing role of restriction enzymes in genetic marker development and genotyping. We then describe and compare five methods for genetic marker development using restriction enzymes and NGS, and cover best practices for the design and execution of experiments that make use of these methods. As with other applications of NGS, the quantity of data produced by these methods poses new analytical challenges. We consider these issues and the likely development of these methods in the near future.

We expect these methods to be used for several years to come. However, given the plummeting cost of NGS, in BOX 2 we consider when these methods, and perhaps genetic markers in general, might be superseded by whole-genome sequencing, and discuss the trade-offs between the two approaches for immediate projects.

Innovations in marker discovery

The impact of NGS on genetic marker technology. Traditionally, the development of markers such as microsatellites[7], RFLPs[5] and AFLPs[6] was a costly, iterative process that involved time-consuming cloning and primer design steps that could not easily be parallelized. Scoring of marker panels across target populations was also expensive and laborious. The advent of high-throughput SNP arrays removed this bottleneck from the genotyping process, but not from the discovery process: the production of a high-quality array requires a substantial investment of resources. Also, these markers are specific to the population in which they were developed, hence genotyping of new populations will be biased towards alleles present in the original survey, which is a serious problem for studies of wild or highly divergent populations.

By contrast, the NGS-based techniques described here enable the discovery, sequencing and genotyping of thousands to hundreds of thousands of markers in tens to hundreds of individuals. Such techniques can be performed directly on genomic DNA in a single sequencing step and with mostly parallelized library preparation. Genotyping of the same markers in other populations can be achieved with further sequencing runs, accurately representing the new populations and avoiding bias towards the sequence of the markers in the originally surveyed population. As the cost of SNP genotyping using sequencing-based approaches remains higher than when using existing SNP arrays, at present it may still be more economical for large consortia to develop SNP arrays that will be used in many different populations. However, for small communities the cost of sequencing is likely to be far lower than the cost of array development. Small panels of SNP or microsatellite markers can be derived using NGS, enabling traditional genotyping to be carried out on a small scale if resequencing of thousands of markers is not required by the application.

Restriction enzymes and NGS methods. Restriction enzymes have been a core tool for marker discovery and genotyping for decades, ever since the development and use of RFLPs to link many genes to human diseases (such as Huntington's disease[8] and cystic fibrosis[9]) and to construct the first complete linkage map of the human genome[10]. Restriction enzymes remain central to the genome-wide NGS methods discussed here, but rather than length polymorphisms, the developed markers are sequenced SNPs or structural variants. The diversity of restriction enzymes available (which vary in the length, symmetry or GC versus AT bias of their recognition sites, and also in their methylation-sensitivity) makes them an extremely versatile assay tool. Their flexibilities allow researchers to customize marker discovery approaches to individual projects; for example, by tailoring the approach to the genome of interest, the project goals and the budget. One can target different specific subsets of the genome by choosing different restriction enzymes. In plants one can often exclude repetitive regions by choosing a methylation-sensitive enzyme that will avoid cutting most methylated repeat elements.

NGS marker discovery and genotyping methods

Several methods have been developed for high-throughput genetic marker discovery and genotyping using restriction enzymes (FIG. 1). The published accounts of these methods differ at various steps, but many of the differences (such as the choice of sequencing platform) are not central to the methods, hence most innovations of particular methods can be broadly applied. In this section, we focus on the most substantial differences, reserving discussion of many common technical issues for the following sections.

Box 2 | **Reduced-representation or whole-genome sequencing?**

Sequencing throughput is doubling every 5 months[76], and new genomes are published regularly. So should researchers about to embark on a population study use reduced-representation methods, or sequence whole genomes? When will the sequencing of hundreds of whole genomes become commonplace?

Currently, sequencing a diploid human genome (consisting of two 3-gigabase sequences) at 30× coverage costs approximately UK£5,000 (a conservative price that includes library preparation and the costs of reagents and labour). Sequencing a population of 100 humans at 30× coverage would therefore cost £500,000 today. This coverage would be high for a human population, but may be low for a non-model species with a similar sized genome of unknown sequence. In addition, despite the promise of single-molecule, multi-kilobase reads[77] and genome scaffolding using optical mapping[78], the production of high-quality reference genomes is still a substantial project. By contrast, restriction-site-associated DNA sequencing (RAD-seq) can produce markers directly, without a costly assembly process, and could sample 200,000 markers in 100 humans at 30× coverage for around £14,000 (again, a conservative price, including library preparation and the costs of reagents and labour), a 35-fold reduction compared to whole-genome sequencing. Low coverage for multiplexed shotgun genotyping (MSG) or genotyping by sequencing (GBS) would therefore cost on the order of £1,000.

However, for species with smaller genomes, whole-genome sequencing is becoming feasible; a population of 100 animals or plants with 300-megabase genomes will cost £50,000 to sequence at depth, a figure within the reach of many laboratories. Also, draft assembly of whole-genome data may be sufficient to call a small set of markers on small-to-medium-sized contigs. Finally, if sequencing throughput continues to double every 5 months, whole-genome sequencing of the human study outlined above will cost the same as RAD-seq today after five doublings of capacity (which should occur by the end of 2013). Clearly, therefore, whole-genome sequencing of populations will soon be affordable, but we believe that reduced-representation methods will be preferable in the short-term, particularly because many research questions can be answered with a small set of markers and thus do not require every base of the genome to be sequenced.

All of the methods involve the following key steps: the digestion of multiple samples of genomic DNA (from individuals or populations) with one or more restriction enzymes; a selection or reduction of the resulting restriction fragments; and NGS of the final set of fragments, which should be less than 1 kb in size (owing to the read-length limits of current NGS platforms). Polymorphisms in the resulting sequenced fragments can be used as genetic markers. We have grouped the methods into three classes: reduced-representation sequencing, including reduced-representation libraries (RRLs) and complexity reduction of polymorphic sequences (CRoPS); RAD-seq; and low coverage genotyping, including MSG and genotyping by sequencing (GBS).

Different applications of these methods require different standards of marker data. For example, for a study of wild populations in which no reference genome is available, a large number of markers scored accurately in most individuals is desirable to ensure population parameters are estimated precisely. In these cases RAD-seq or reduced-representation methods are most appropriate. For genotyping applications such as marker-assisted selection and QTL mapping — in which broods with limited polymorphism are to be sequenced and parental genotypes are well known — low-coverage genotyping is sufficient for linkage to be inferred, provided that a reference genome is available.

Reduced-representation sequencing. Sequencing the whole genome of every individual in a population is costly and often unnecessary, as many biological questions can be answered using polymorphisms that are measured in a subset of genomic regions. RRLs and CRoPS are two methods for sampling and sequencing a small set of genome-wide regions without sequencing the entire genome.

RRLs were first used to generate a SNP map of the human genome using capillary sequencing[11]. The RRL approach has been adapted for NGS and has been used to generate tens of thousands to millions of candidate SNPs in cattle[12], swine[13–16], turkey[17], maize[18], rainbow trout[19], great tit[20], soybean[21,22], Iberian sow[23], jointed goatgrass[24] and mallard[25] using the Roche Genome Sequencer[26], the Illumina Genome Analyzer[27] and the Applied Biosystems Support Oligonucleotide Ligation Detection (SOLiD)[28] sequencing technologies individually or in combination[14,24]. Genomic DNA from multiple individuals is digested with a frequently cutting restriction enzyme of choice and pooled. The resulting restriction fragments are selected by size and then sequenced, producing partial but genome-wide coverage at a fraction of the cost of whole-genome sequencing. In its simplest form only the ends of the fragments are sequenced, but the protocol can be modified to sequence entire fragments[17].

When a high-quality reference genome is available, the reads from reduced-representation sequencing can be mapped to the reference genome and SNPs can be called as for whole-genome resequencing projects[29,30]. Without a reference genome, long reads from the Roche Genome Sequencer platform or reads from both ends of the library fragments from any NGS platform (paired-end reads) can be used to assemble the fragments *de novo* before calling SNPs. Paired-end reads also facilitate the calling of structural variations in RRLs[31].

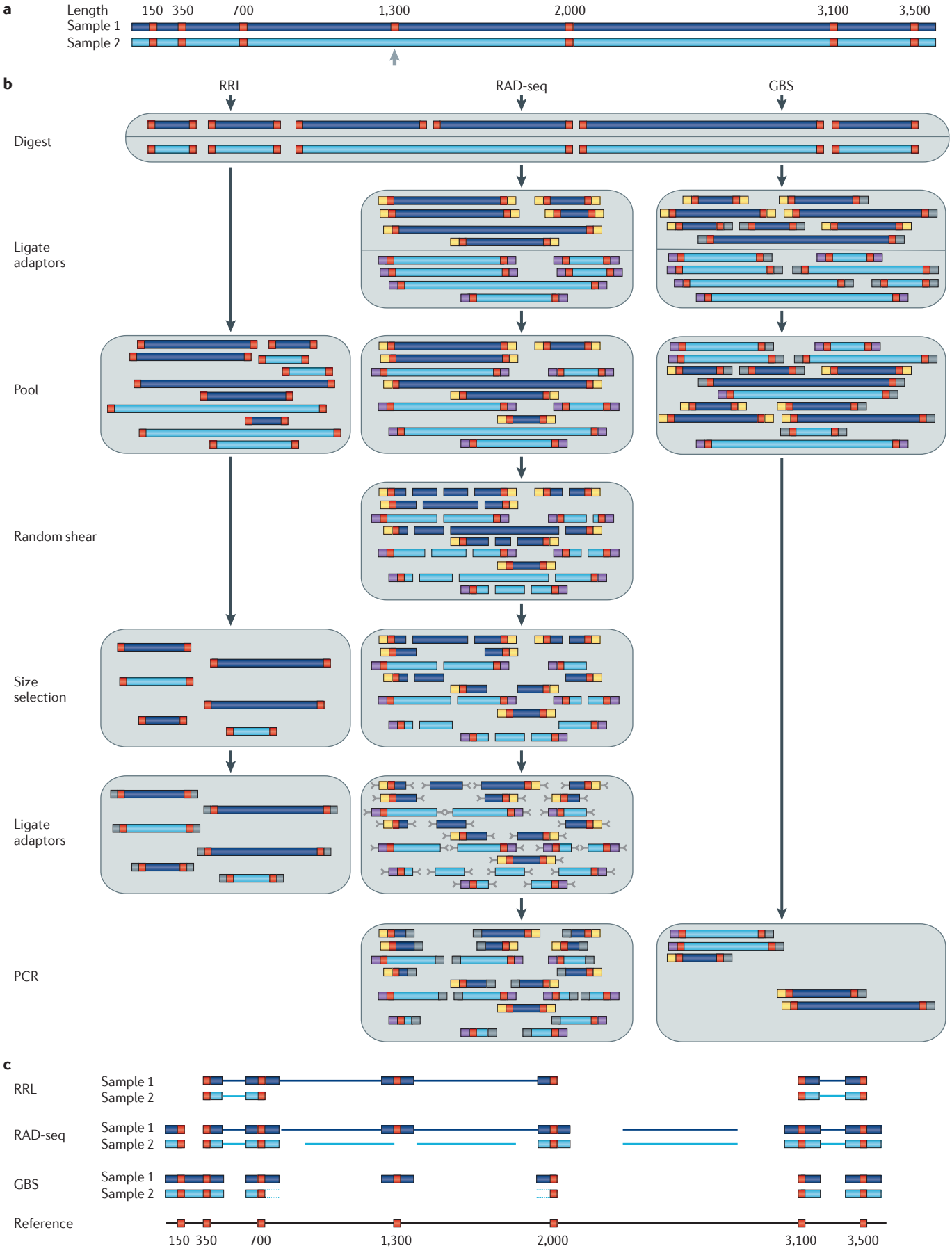
Because AFLP markers are also fragments digested using one or more restriction enzymes, they can be sequenced in a similar way by CRoPS. By adapting the amplification of AFLP fragments to enable sequencing on the Roche Genome Sequencer platform, CRoPS has been used to discover over 1,000 SNPs in maize[32,33] and to generate genome-wide F_{ST} statistics in 12 populations of *Lycaeides* butterflies[34,35].

RRLs have usually been used to sequence pools of DNA samples from multiple individuals, thus allowing the detection of polymorphisms within a population but not for each individual (see the 'Design of marker discovery experiments' section below). CRoPS was the first method to identify polymorphisms in each individual sample by incorporating short barcode identifier sequences — also known as multiplex identifier sequences (MIDs) on the Roche Genome Sequencer platform[34] — into the ligated adaptors and using an adaptor containing a different barcode for each DNA sample. The barcodes can be used to separate sequence reads for different samples bioinformatically, and they enable true population studies to be carried out in one lane of an NGS sequencing platform. RRLs can also be adapted to use barcodes. However, to make

Microsatellite
A class of repetitive DNA that is made up of repeats that are 2–8 nucleotides in length. They can be highly polymorphic and are frequently used as molecular markers in population genetics studies.

Optical mapping
A method for creating a map of a genome by stretching DNA in microfluidic channels on a slide for visualization on a fluorescent microscope. The DNA is then digested by restriction enzymes and the sizes of these fragments are inferred by the integrated intensity of the fluorescent intercalator dye.

F_{ST}
(Wright's fixation index). The fraction of the total genetic variation that is distributed among subpopulations in a subdivided population.



◀ Figure 1 | **Methods for high-throughput marker discovery. a** | An example genomic region containing restriction sites (red). A sample of DNA from each of two individuals (sample 1 is dark blue and sample 2 is light blue) is to be sequenced. Sample 2 has a variation in the cut site at 1,300 bases (grey arrow) and so this site will not be cut. **b** | Protocols for developing sequenced markers. All methods begin with a restriction enzyme digestion. Reduced-representation library (RRL; left panels): fragments from all samples are pooled and size selected to 300–700 bp. Fragments are ligated to standard sequencing adaptors (grey squares) and sequenced. In this simple case, only the ends of fragments will be sequenced, but the protocol can be modified to sequence entire fragments. Restriction-site-associated DNA sequencing (RAD-seq; middle panels): fragments are ligated to P1 adaptors (yellow for sample 1, purple for sample 2), pooled, randomly sheared and size selected to 300–700 bp. P2 adaptors with divergent ends (grey, Y-shaped) are ligated to the fragments with and without P1 adaptors. The fragments are PCR amplified with P1- and P2-specific primers. The P2 adaptor is completed when fragments containing P1 adaptors are bound by a P1 primer and copied, and the P2 primer only binds to completed P2 adaptors (grey squares). This means that only fragments with P1 and P2 adaptors (the fragments containing restriction sites) are amplified. Genotyping by sequencing (GBS; right panels): barcoded adaptors (yellow) and common adaptors (grey) are ligated to digested fragments, producing fragments with barcode+common, barcode+barcode and common+common adaptor combinations. Samples are pooled and amplified on the Illumina Genome Analyzer flowcell. Only short samples featuring a barcode+common adaptor combination are amplified for sequencing. **c** | Sequenced markers are aligned to the original reference genome. RRL: either fragment ends (thick lines) or entire fragments (thin lines) between 300 and 700 bp are sequenced. Because the site at 1,300 bases is not present in sample 2, the long fragment between 700 and 2,000 bases is filtered by size selection. RAD-seq: downstream regions of all fragments above 300 bases are sequenced, but not the fragment between 150 and 350 bases. Thin lines indicate the sequence that would be covered using paired-end sequencing. GBS: dashed lines represent regions that would be filtered during amplification, but could be imputed using (for example) the multiplexed shotgun genotyping hidden Markov model. The short fragment between 150 and 350 bases will be sequenced.

use of barcodes, the fragments from each sample must be size selected individually, before adaptor ligation and pooling.

RAD-seq. RAD markers[36] were first implemented using microarrays[37] and later adapted for NGS[38]. RAD-seq sequences short regions surrounding essentially all restriction sites for a given restriction endonuclease (assuming a sufficient sequencing depth), regardless of the length of the restriction fragments. To achieve this, the restriction fragments are randomly sheared to a length suitable for the sequencing platform of choice, and selective PCR is used to amplify for sequencing only those fragments containing a restriction site (FIG. 1). RAD-seq has been used to study population differentiation and selection in the stickleback[39], to investigate the phylogeography of pitcher plant mosquitos[40], to generate SNPs in rainbow trout[41] and to construct linkage maps in barley[42], ryegrass[43] and the diamondback moth[44].

As with RRLs and CRoPS, RAD-seq begins with digestion of a DNA sample using a chosen restriction enzyme, provided that the enzyme produces fragments with sticky end overhangs. Barcoded adaptors are then ligated onto these fragments to identify each individual in a population. However, unlike CRoPS, samples are pooled following adaptor ligation, and the remaining steps are carried out on the pooled library, reducing the labour and cost. In publications to date, RAD-seq libraries have been sequenced on the Illumina Genome Analyzer platform, generating a large depth coverage of the sequence (50–150 bases) flanking each restriction site.

For many studies, the high genome-wide marker density possible through RAD-seq will be more useful than the sequencing of large regions flanking each polymorphism by RRLs and CRoPS approaches. However, with paired-end sequencing it is possible to assemble the paired-ends for each locus into a long contig with an average length of ~500 bases[45]. This contig can be used either to anchor the markers to existing genomic resources[44] or, with sufficient coverage, to call SNPs across the whole fragment. These paired-end fragments can be used to design primers for high-throughput genotyping assays, which typically require at least 60 nucleotides of flanking DNA on either side of a SNP[45]. This can be very useful in organisms lacking a well-assembled reference genome.

Low coverage sequencing for genotyping. The above methods reduce the proportion of the genome targeted for sequencing so that each marker can be sequenced at high coverage with limited resources, thus enabling markers to be genotyped accurately across many individuals. An alternative to this approach is to sequence many target markers at low coverage per individual, accepting that a different subset of markers will be genotyped in each individual. This strategy uses markers that are sequenced at sufficient coverage, and that have known marker positions on a physical map, to impute missing genotypes and to infer recombination breakpoints. This is suitable for genotyping recombinant populations in which the parental genotypes are either known or can be assigned probabilities. This approach has been used to construct genetic maps for rice based on low-coverage whole-genome resequencing of hundreds of recombinant inbred lines (RILs)[46,47]. Another application was to generate a haplotype map of maize based on 3.3 million SNPs, using low-coverage sequencing of three RRLs that were cut with a range of different restriction enzymes[18].

However, for many studies, coverage of the whole genome is unnecessary to infer recombination breakpoints for mapping, and sequencing of sparse, short markers will be sufficient. Although low-coverage RAD-seq can be used for this, a number of methods with simpler library preparation protocols have recently been published to achieve the same aim at very low cost per sample, albeit with more complex analyses.

GBS involves the digestion of genomic DNA with a frequent cutter and the sequencing of the ends of all resulting restriction fragments[48] (FIG. 1). (In the published work[48] on maize and barley, the methylation-sensitive enzyme *Ape*KI was used.) Adaptors containing barcodes and common adaptors without barcodes are mixed and used in the ligation reaction. Not all adaptor-ligated fragments will be sequenced, because many fragments will not be efficiently bridge-amplified on an Illumina Genome Analyzer flowcell[27], either because they do not feature both a barcoded adaptor and a common adaptor, or because they are too long (>1 kb). This still

Imputation
A statistical method for handling missing data in which the missing values are replaced by estimated values.

Recombinant inbred lines
(RILs). A population of fully homozygous individuals that is obtained through the repeated selfing of F_1 hybrids, and that is comprised of 50% of each original parental genome in different combinations.

leaves a large number of shorter fragments with the correct adaptors that will be sequenced accurately, thus enabling the discovery of 25,000 SNP markers in one experiment[48].

MSG[49] follows a similar approach, except that only a barcoded adaptor is used that is ligated to both ends of each fragment, and fragments are size-selected before sequencing. MSG, in common with the approach of Xie *et al.*[47], makes use of a hidden Markov model to impute haplotypes. Soft ancestry calls are made for genomic segments, based on probabilistic calls of parental genotypes and offspring genotypes when available (BOX 3). This does not require genotyping of every marker for every individual, but it does require that markers are mapped to a relatively well-assembled reference genome (with a median scaffold size of >100 kb). This analytical framework can be applied to data from any of the methods described here.

Design of marker discovery experiments

Several interacting factors affect the choice of an optimal NGS marker discovery method. Here we discuss these factors and several technical variations and analytical challenges that apply to all methods.

Study goals. How many markers are required to achieve the goals of a study, and how complete must the genotyping be? Although NGS methods can produce tens of thousands of markers with high genotyping accuracy, this resolution and power may not be necessary to answer many biological questions. For example, for crossing studies there is little value in producing markers at substantially higher resolution than the expected spacing of crossovers (based on the recombination rate in the population), and low-coverage sequencing may produce sufficient data for calling breakpoints (as for MSG). However, it may be useful to generate more markers than required and then to select high-quality markers from these based on the needs of the study (BOX 4).

Availability of a reference genome. When a reference genome sequence is available, sequence reads produced by any of the technologies can be aligned and positioned on a physical map. The higher the quality of the reference genome assembly, the easier it is to impute missing genotypes, thus reducing the coverage that is required to genotype each individual. Reference genomes can also be used to design marker discovery experiments by simulating *in silico* the number of markers produced by different enzymes (FIG. 2). Challenges arise when a reference genome sequence is not available, or even when a reference sequence is available but is poorly assembled, comes from a distantly related taxon or is large and highly repetitive. Expected numbers of restriction sites can be crudely estimated if the genome size and GC content are known. The calculation is particularly susceptible to GC content, so it is worth making a high-quality estimate of this parameter. The task of identifying unique loci and assigning sequence reads unambiguously can be very difficult and, in practice, much data will be discarded. Criteria for filtering out spurious putative loci include excessively high read counts[40], the presence of repetitive sequence[13] or observed heterozygosity[41].

Expected degree of polymorphism. Populations with low levels of polymorphism will require the assay of more markers, and therefore require an enzyme with a higher cutting frequency to produce sufficient polymorphic markers (FIG. 2). RRLs and CRoPS, which sequence entire restriction fragments, may be more suitable for these populations. By contrast, populations with high levels of polymorphism are susceptible to problems caused by variation at restriction sites. If many sites are polymorphic, the fragment distribution will change considerably. For methods in which fragments are size-selected, a polymorphic restriction site may result in a long allele that is not included in the size selection, causing several adjacent markers to drop out (FIG. 1). However, although more markers will drop out in highly polymorphic populations, the remaining markers are more likely to be informative than in less polymorphic populations because these markers are more likely to contain a polymorphism.

Unfortunately, these methods are not a panacea for genomes with a large repetitive fraction or high ploidy. For example, using an RRL it was possible to validate 94% of a sample of putative swine SNPs[14] but only 48% of a sample of SNPs in a rainbow trout[19], largely owing to the ancestral whole-genome duplication in salmonid species. However, repetitive sequences can be partially avoided by the careful choice of restriction enzymes or by removal *in silico* by filtering putative loci, as described above.

Choice of restriction enzyme. The choice of enzyme is determined by the marker density required (FIG. 2) and should be appropriate for the species under study; for example, Van Tassell *et al.*[12] put considerable effort into choosing an enzyme that avoids common repeats in cattle, and the methylation-sensitive *Ape*KI used by Elshire *et al.*[48] may not be appropriate for other methylated genomes. All methods except RRLs use endonucleases that produce overhangs. New adaptors are required for different overhangs, but the same adaptor set can be suitable for multiple enzymes producing the same overhang, such as the 8-base cutter *Sbf*I (cut site CCTGCA^GG) and the 6-base cutter *Pst*I (cut site CTGCA^G).

DNA sample preparation. High-quality genomic DNA (free of contamination either with RNA or with DNA from other species) is crucial to the success of these protocols, given that varying efficiency of digestion, ligation and amplification can have significant effects on the final marker set. Most importantly, the quantity of DNA from different samples should be evenly balanced before pooling to avoid losing markers from some individuals owing to lack of coverage. The choice of method may also be influenced by the amount of genomic DNA starting material required (RRL, 25 µg pooled[14,29]; CRoPS, 300 ng per sample[34]; RAD-seq, 300 ng per sample[42,43]; MSG, 10 ng per sample[49]; and GBS, 100 ng per sample[48]).

Hidden Markov model
A statistical approach that is used to estimate a series of hidden states (for example, ancestry at loci along a chromosome). The method is based on observations of the states that have uncertainty (for example, the ancestral assignment of sequence reads) and the expected probability of transitions between states (for example, recombination breakpoints).

Soft ancestry calls
Assigning probabilities to ancestral (for example, parental or grandparental) genotypes, rather than making explicit, 'hard' calls. This approach appropriately propagates uncertainty (which often arises around recombination breakpoints) in individual ancestry assignments, thus enabling a more accurate inference of breakpoint location.

Scaffold
A genomic unit composed of one or more contigs that have been ordered and orientated using end-read information.

Adaptor design. Adaptors should be designed so that any barcode is at least three base pairs distinct from all others, so that reads containing an error in the barcode sequence can be uniquely assigned to a sample. When using the Illumina Genome Analyzer platform, more than two barcodes with a diversity of nucleotides represented should be combined in one lane to avoid a common issue with Illumina sequencing: low diversity at particular positions of the read can prevent the Illumina software from identifying clusters accurately[50].

Box 3 | **Statistical models for marker discovery and genotyping using next-generation sequencing**

Alleles at each locus

At each locus, the observed counts of alternative reads can be treated as a set of independent samples from a small set of possibilities, suggesting a multinomial distribution[35,39,49,79]. In the absence of error, if each sample or barcode represents a pool of individuals, the probability of observing each of the four nucleotides at a specific site is the allele frequency in the pool. If each sample is from a diploid individual, the expected probabilities are one or zero for homozygous sites and 0.5 or zero at heterozygous sites. These probabilities should be adjusted to include an error parameter to account for single-nucleotide sequencing and PCR error. Thus for a diploid individual, the probabilities of observing read counts for the four nucleotides in a homozygote (genotype 1/1) or heterozygote (genotype 1/2), and thus the likelihoods (L) of each genotype, can be modelled as:

$$L(1/1) = P(n_1,n_2,n_3,n_4|1/1) = \frac{n!}{n_1!n_2!n_3!n_4!}\left(1-\frac{3\varepsilon}{4}\right)^{n_1}\left(\frac{\varepsilon}{4}\right)^{n_2+n_3+n_4}$$

and

$$L(1/2) = P(n_1,n_2,n_3,n_4|1/2) = \frac{n!}{n_1!n_2!n_3!n_4!}\left(0.5-\frac{\varepsilon}{4}\right)^{n_1+n_2}\left(\frac{\varepsilon}{4}\right)^{n_3+n_4}$$

where n_1, n_2, n_3 and n_4 are the read counts for each nucleotide, n is the total number of reads and ε is the sequencing error rate. This error parameter can be estimated directly by maximum likelihood[35,39], drawn from the empirical output from a control sequencing lane[49] or modelled as a prior distribution.

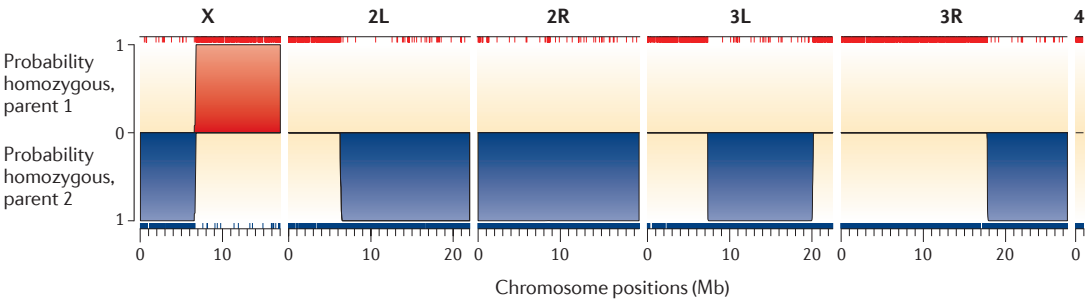
Assigning likelihoods to alternative genotypes at each locus allows uncertainty to be carried through the analysis in a Bayesian fashion; for instance, in posterior distributions around population-level statistics[32] or in 'soft ancestry' assignments to each locus in a laboratory cross[49].

Statistics along the genome

The promise of genomic approaches is to view genetic statistics — whether chromosomal ancestry in a laboratory cross or genetic differentiation in population samples — as continuous distributions along the genome. The density of markers leads to significant correlations (linkage disequilibrium) among neighbouring loci, which means that large numbers of missing genotypes can be tolerated without losing power of inference about relatively narrow genomic regions. One simple approach to take advantage of this property is sliding window averaging, so that genomic regions, rather than individual loci, can be identified as outliers in a genome scan, thus increasing the power to detect selection in natural populations[39,46,80].

Another alternative is to model the distribution of an underlying statistic, such as ancestry in a laboratory cross, along the genome using a hidden Markov model. For example, the multiplexed shotgun genotyping (MSG) technique follows the strategy of maximizing the density of markers and the number of individuals while reducing sequencing depth, combined with a high-quality reference genome sequence and a hidden Markov model[49] (see the figure).

The figure illustrates 'soft ancestry' assignment along the genome in a single backcrossed *Drosophila melanogaster* male using MSG. The posterior probability that a region is homozygous (or hemizygous on the X chromosome) for one or the other parental genotype is shown in blue or red; a high posterior probability of heterozygous genotype (on each of the four chromosomes) appears as a black line at the centre. Observed parental alleles are shown as red or blue hashmarks at the top and bottom of the plot. Note that several apparent genotyping errors are overwhelmed by marker density in the context of the hidden Markov model; that is, several loci appear to have parental genotype 1 in regions where the posterior probability for homozygous parent 2 approaches 1.0. Relatively rapid switches of posterior probability from homozygous to heterozygous regions allow narrow mapping of recombination breakpoints in this analysis. R, right chromosome arm; L, left chromosome arm. Figure is modified, with permission, from REF. 49 © (2011) Cold Spring Harbor Laboratory Press.



Sliding window averaging
The averaging of statistics, such as nucleotide diversity or F_{ST}, for all markers in a chosen size of overlapping genomic region (window). When applied across the genome, this method smoothes out variation within regions so that genome-wide patterns can be observed.

Box 4 | **Experimental design case studies**

Three case studies (discussed here and summarized in the accompanying table) highlight the factors influencing the design of high-throughput marker discovery experiments.

Case study 1: phylogeography

Emerson *et al.*[40] resolved the phylogeography of 21 populations of the pitcher plant mosquito, *Wyeomyia smithii*, using restriction-site-associated DNA sequencing (RAD-seq). No reference genome is available for this species, and the populations are wild, so ancestral genotypes are unknown. These factors make imputation of missing markers difficult, so RAD-seq is preferable to multiplexed shotgun genotyping (MSG) or genotyping by sequencing (GBS). A phylogenetic tree based on cytochrome oxidase 1 sequences, including 167 variable sites, separated the populations into northern and southern groups but could not resolve relationships between populations within each group. Using the 8-base cutter *Sbf*I (FIG. 2) to perform RAD-seq on 21 DNA pools — one pool per population and each representing six individuals — 3,741 SNPs were identified among 13,627 loci within two lanes of sequencing. Using these SNPs, four clades were identified and the relationships between the 21 populations were completely resolved. Only markers differing by one nucleotide were included, although many more markers varying by more than one nucleotide or containing insertion or deletion polymorphisms could have been developed. Restricting to markers containing 1 SNP simplified analysis and was sufficient to resolve the phylogeographic tree.

Case study 2: association mapping

Andolfatto *et al.*[49] resolved a causal locus to within 8.5 kb of its true location using MSG. A strain of *Drosophila simulans* — featuring a transgene encoding enhanced yellow fluorescent protein (EYFP) driven from a *Pax6* promoter, which causes a dominant fluorescent eye phenotype — was crossed with *Drosophila sechellia*. Ninety-six backcrossed progeny were available for genotyping. The study population was a mapping cross, a reference genome was available and ancestral genotypes could be discovered in full. Owing to these advantages, low-coverage genotyping could be used, followed by imputation of missing markers. High marker density was required to resolve the locus as accurately as possible, so the frequent 4-base cutter *Mse*I was used. Each of the 96 progeny was assigned a different barcode and individually genotyped. Reads from one lane of sequencing were mapped to the *D. simulans* reference genome and a median number of 15,070 informative markers were scored for each individual. Using the MSG hidden Markov model to infer missing genotypes, ancestry was assigned to 125,214 genomic locations for all individuals. Quantitative trait locus (QTL) analysis revealed a single significant QTL peak on the X chromosome with a maximum lod score at position 7,131,433 bp, just 8,498 bp away from the true location of the causative locus at position 7,139,931 bp.

Case study 3: measuring artificial selection

Ramos *et al.*[14] sequenced reduced-representation libraries (RRLs) from four domestic pig breeds and the wild boar. For each breed, DNA from between 23 and 36 animals was pooled, and four libraries were produced, with DNA cut with *Alu*I (and size selected to 160–200 bp or 200–240 bp), with *Hae*III or with *Msp*I. The *Alu*I libraries were sequenced with a Roche Genome Sequencer; these long reads were mapped to the pig reference genome and unique unaligned reads were assembled into contigs to represent regions missing from the reference genome. All libraries were sequenced with an Illumina Genome Analyzer to produce 36 bp reads that could be aligned to both the reference and the new contigs. In total, 372,886 SNPs were identified, 64,232 of which were chosen for inclusion on a Beadchip, of which 58,994 were polymorphic (a conversion rate of 94%). Using these SNPs, Amaral *et al.*[16] were able to measure nucleotide diversity, F_{ST} and other population genetics statistics across the whole genome, thus discovering footprints of artificial selection for coat colour, growth and indicators that the major histocompatibility complex is under balancing selection.

Factor	Case study 1	Case study 2	Case study 3
Study type	Phylogeography	QTL mapping	Artificial selection
Species	*Wyeomyia smithii*	*Drosophila simulans*	*Sus scrofus*
Genome size	850 Mb	120 Mb	2.7 Gb
Reference genome	Not available	Available	Available
Populations	Wild	Mapping cross	Domesticated
Marker density required	Low	High	Medium
Populations	21	1	5
Individuals per population	6	96	23–36
Method	RAD-seq	MSG	RRL
Restriction enzyme (cut site)	*Sbf*I (CCTGCA^GG)	*Mse*I (T^TAA)	*Alu*I (AG^CT), *Hae*III (GG^CC), *Msp*I (C^CGG)
Reads sequenced	27 million 49 bp (Illumina Genome Analyzer)	22 million 101 bp (Illumina Genome Analyzer)	380 million 36 bp (Illumina Genome Analyzer) + 4 million ≤250 bp (Roche Genome Sequencer)
Coverage	~30×	~1×	7.5×–10×
Markers identified	3,741 SNPs in 13,627 loci	15,070 scored, 125,214 imputed	~372,000 SNPs
Refs	40	49	14,16

lod score
(Base 10 'logarithm of the odds' or 'log-odds'). A statistical estimate of whether two loci are likely to lie near each other on a chromosome and are therefore likely to be inherited together. A lod score of three or more is generally considered to indicate that the two loci are close.

Major histocompatibility complex
(MHC). A complex locus on human chromosome 6p, which comprises numerous genes, including the human leukocyte antigen genes, which are involved in the immune response. MHC molecules bind peptide fragments that are derived from pathogens and display them on the cell surface for recognition by the appropriate T cells. The organizations of the MHC gene clusters are similar in many species.

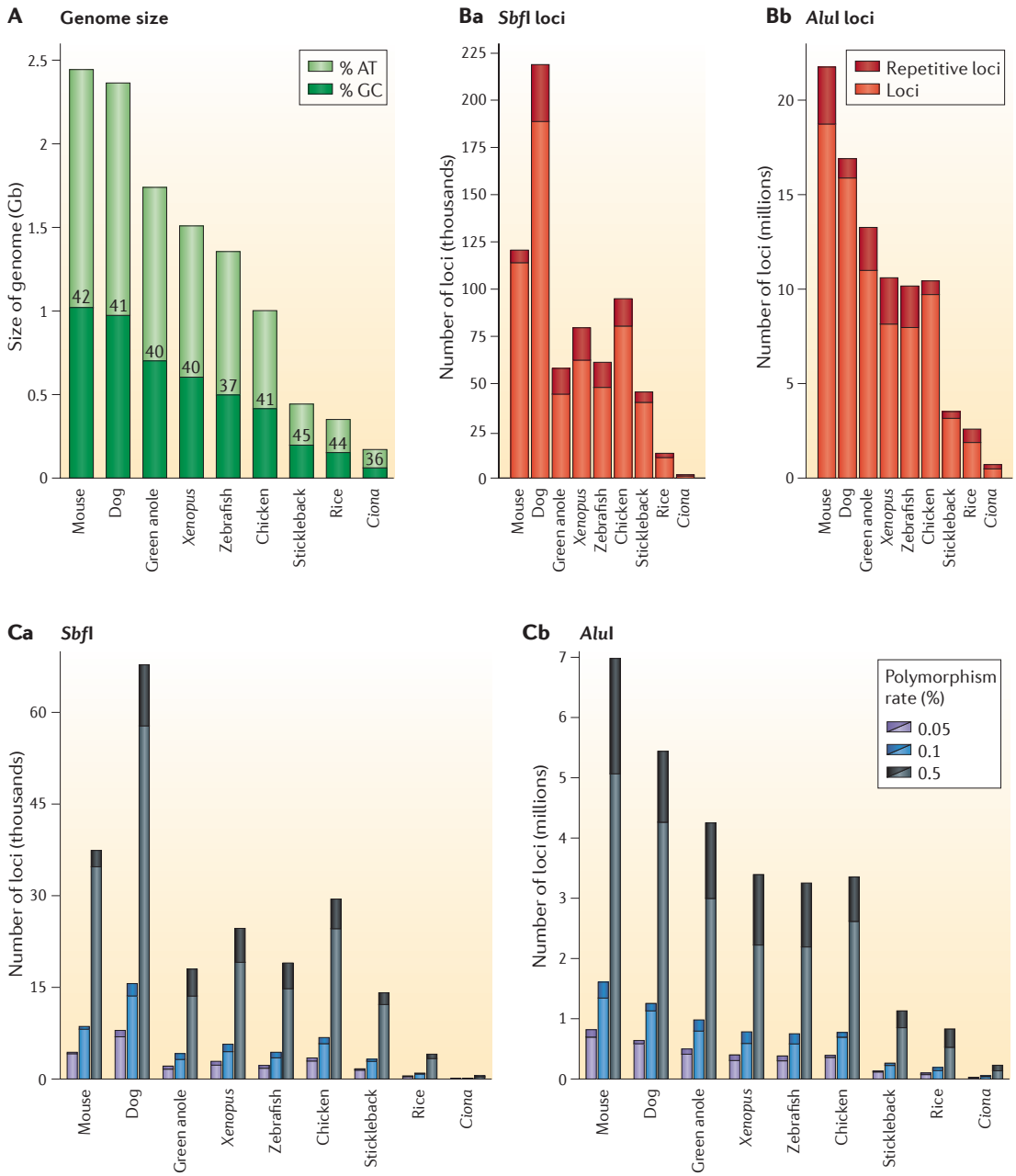


Figure 2 | **Effects of restriction enzyme selection on reference genomes of different size and with different levels of polymorphism. A** | The size and GC content of nine reference genomes, as obtained from Ensembl version 61, calculated directly from the published reference sequence. **B** | The number of loci produced from nine reference genomes, as calculated *in silico*, when digested by one of two restriction enzymes: *Sbf*I (CCTGCA^GG) (**Ba**) and *Alu*I (AG^CT) (**Bb**). Loci that are composed of repeats longer than an 80 bp read are shown in a darker shade. Note that *Alu*I targets orders of magnitude more loci than *Sbf*I. **C** | The expected number of loci containing polymorphisms under the same two restriction enzymes, *Sbf*I (**Ca**) and *Alu*I (**Cb**). We randomly applied SNPs at three different rates (0.05%, 0.1% and 0.5%) to the loci extracted *in silico* and tabulated how many would be recoverable using the software package Stacks if sequenced to sufficient depth. Averages of three replicates are shown. The darkened tops of each bar represent polymorphic loci that are confounded by repetitive sequence. Comparing the dog and mouse genomes illustrates the issues involved with selecting the most appropriate restriction enzyme. Although the two genomes are of similar size and GC composition, the dog genome produces almost twice as many *Sbf*I loci as the mouse genome. At the same time, the green anole possesses a genome larger than *Xenopus laevis*, the zebrafish and the chicken, yet contains fewer *Sbf*I loci than all three genomes. Furthermore, despite the characteristics of the genome under examination, the particular polymorphism rate of the population being studied is also important in determining how many potential genetic markers will be produced. In an organism with an unexamined genome, a pilot experiment is vital for determining the performance of a particular restriction enzyme and the allocation of sequencing resources that will be required.

Low sequence diversity is a problem with methods in which the restriction enzyme overhang appears at the same position in every read. Although using many barcodes usually avoids this problem, an innovation of GBS that can be applied to any method is the use of variable length barcodes (between 4 and 8 nucleotides long). It is important to design variable length barcodes carefully so that if a base is missed during sequencing, each barcode cannot be confused with another, shorter one. Different barcode sequences can also have a mild impact on sampling depth, but this factor has less influence than balancing the DNA quantities from different samples[49].

PCR amplification. The PCR steps in sample preparation can bias sequencing towards GC-rich, short fragments[51] and may also bias the fragment pool in other ways. Some protocols for genome and transcriptome sequencing avoid PCR altogether[52]. Unfortunately, it is difficult to avoid PCR in NGS marker methods, despite its demonstrable effect on this kind of data[48], because PCR amplification is required to ensure that adaptor-ligated fragments outcompete other fragments. If possible, the amount of input DNA should be increased, so that number of PCR cycles can be reduced.

RAD-seq has two advantages in this respect. First, because fragments are randomly sheared, rather than size-selected after digestion, fragment length is not correlated with particular loci as in the other techniques, so any amplification bias by fragment size will not affect representation across loci. Second, when paired-end sequencing is used, clonal duplicates arising from PCR can be removed by identifying fragments that have identical sequence at both ends, something which is unlikely to occur with random shearing.

The PCR step also produces adaptor dimers. The dimers should be shorter than any fragment of interest and so can be removed by gel extraction[53] or solid-phase reversible immobilization (SPRI) technology[49,54]. A titration should be performed once for each new species and enzyme combination to determine an appropriate concentration of adaptors that minimizes the formation of adaptor dimers[48,54]. The dimers can also be identified and removed bioinformatically.

Sequencing. How much sequencing is required for each of these methods? Because the methods require the identification of polymorphisms, platforms such as the Illumina Genome Analyzer and SOLiD are preferable to the Roche Genome Sequencer, although long reads from the Roche Genome Sequencer can be useful for assembling draft genomic sequence for unsequenced species[14]. The optimal mean coverage per locus varies widely across experimental goals and strategies, as it reflects trade-offs with factors such as the number of individuals, the number of groups or populations, the genome size, the density of markers across the genome and the total sequencing effort (BOX 4).

Lower coverage per locus results in a lower confidence in each genotype call or each newly discovered marker, as well as fewer and more variable numbers of individuals genotyped at each locus. These issues can be accounted for with analyses that impute across loci, taking advantage of a high-quality reference genome sequence, but this strategy requires more effort in analysis than those with higher mean coverage (BOX 3). A reduction in total sequencing depth can therefore be a false economy, because it will almost certainly increase the difficulty (and therefore cost) of downstream analysis.

At an extreme, if complete coverage of all restriction sites in a genome with full genotyping in all individuals is desired, then at least 30× coverage per locus per individual is recommended when a good reference genome is available, increasing to 60× coverage for *de novo* studies. However, this complete coverage is overkill for many studies in which sufficient markers can be developed from a subset of all restriction sites in a genome and hence incomplete genotyping is acceptable. In recombinant populations, a small number of markers with incomplete genotypes will be sufficient, provided that the parental genotypes are well understood. In mapping studies, coverage in offspring or RILs below 1×, combined with statistical imputation of haplotypes, may be the most efficient strategy. Many population genomics applications may improve statistical power by maximizing the number of individuals sampled per population[55]. Therefore, increasing the number of individuals, while accepting that a subset of markers will not be genotyped in all individuals, may provide the optimal trade-off.

Pooling individuals. Many studies use one barcode for a pool of several individuals. This can be useful to avoid a whole-genome amplification step if the amount of DNA per individual is small[40]. There is also analytical theory to suggest that such pooling improves SNP discovery and leads to better estimates of population allele frequencies[56]. However, pooling prevents genotyping of individuals after SNP discovery, rather than allowing simultaneous marker discovery and genotyping, which is a key advantage of the techniques described here. Pooling has the disadvantage of potentially missing rare variants and is highly sensitive to variation in the DNA concentration among individuals in a pool[57] (although individually barcoded sequencing also suffers from this problem). In the absence of a high-quality reference genome sequence, pooling also precludes filtering on the basis of observed heterozygosity[41]. If the sequencing resource permits, barcodes for individual samples provide greater flexibility for downstream analysis and this approach does not preclude ignoring the barcodes and pooling the samples bioinformatically.

Analytical challenges. A crucial feature of marker discovery using the techniques above is that they incorporate a multi-level sampling process. NGS sequence reads are a sample from a large, heterogeneous pool of DNA fragments. There is sampling variance in the number of reads across individuals or barcodes within the pool, across loci within each individual, and across alternative alleles at polymorphic loci. Several steps in the protocols may also exacerbate variance at all of these levels, particularly pooling of DNA, PCR amplification and

Solid-phase reversible immobilization (SPRI). The purification of nucleic acids using magnetic beads, thus avoiding gel extraction, filtration and centrifugation.

size selection. Thus NGS approaches, although producing orders of magnitude more markers than previously possible, differ from traditional marker genotyping in important ways: there is unavoidable variance in the sample sizes of individuals across loci and loci across individuals and uncertainty in genotype assignments across loci and individuals. These factors are central to the trade-offs in experimental design, such as the balance among marker density (that is, choice of restriction enzyme and fragment size range), the number of individuals and/or populations sampled and the depth of sequencing. Some progress is being made in statistically accounting for these issues (BOX 3) and several packages are now available for handling data from these methods[35,49,58–60] (see Further information). Although each of these packages was designed with a particular method in mind, there is no difficulty in processing data from one method with tools designed for another.

Implementation. The library preparation protocols for the methods discussed involve only standard molecular biology techniques (for example, restriction enzyme digestion, size selection by agarose gel extraction, shearing, ligation and PCR) and so should be accessible to any competent molecular biologist. However, in practice, the methods can produce varying results for different species, and the length of the protocol for some methods can be challenging. Therefore, it may take several attempts to produce a successful library. Library preparation services are available from several companies and sequencing facilities. A completed library can be sequenced on any NGS machine, either locally or at a sequencing facility; these libraries will be highly suitable for sequencing on bench-top sequencers. The only substantial initial cost may be the purchase of barcoded adaptors and (for some methods) primers, but the sequences for these components are all publicly available[32,38,48,49].

Future directions

We anticipate that many small improvements will be made to the protocols described here, increasing the quality and accuracy of the sequenced marker sets. New variant applications are possible; for example, RNA can be reverse transcribed into cDNA and cut with restriction enzymes, producing a small set of markers from the transcriptome that can be used to assay gene expression without the burden of transcriptome assembly. However, we expect the largest gains to come from improved analysis of the data produced by these methods. A better understanding of the variation in the data will enable more robust inference of marker identity and genotypes. We anticipate this work being of lasting value because any analytical frameworks developed will also be usable when complete genomes are available. (For example, the MSG hidden Markov model and the likelihood model outlined in BOX 3 can be used with low-coverage whole-genome shotgun sequencing.)

Rapidly increasing throughput will allow more individuals to be sequenced in a population, more markers to be sequenced per individual and each marker to be genotyped at greater depth and so with greater accuracy. We expect that it will be possible to sequence tens of thousands of markers in thousands of individuals in the near future. This will be far in excess of what is required for many studies in which a small number of markers are quite sufficient, and will be accessible using the methods that we have discussed and the recently emerging bench-top sequencing machines. Although whole-genome sequencing of populations is rapidly approaching (BOX 2), we believe that the methods described here are likely to remain invaluable for years to come in population genomics, mapping studies and reference genome sequence assembly, particularly for non-model organisms.

1. Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. The power and promise of population genomics: from genotyping to genome typing. *Nature Rev. Genet.* **4**, 981–994 (2003).
2. Stapley, J. *et al.* Adaptation genomics: the next generation. *Trends Ecol. Evol.* **25**, 705–712 (2010).
3. Allendorf, F. W., Hohenlohe, P. A. & Luikart, G. Genomics and the future of conservation genetics. *Nature Rev. Genet.* **11**, 697–709 (2010).
4. Helyar, S. J. *et al.* Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Mol. Ecol. Resour.* **11**, 123–136 (2011).
5. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
6. Vos, P. *et al.* AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* **23**, 4407–4414 (1995).
7. Jarne, P. & Lagoda, P. J. Microsatellites, from molecules to populations and back. *Trends Ecol. Evol.* **11**, 424–429 (1996).
8. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).
9. Riordan, J. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066–1073 (1989).
10. Donis-Keller, H. *et al.* A genetic linkage map of the human genome. *Cell* **51**, 319–337 (1987).
11. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
12. van Tassell, C. P. *et al.* SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods* **5**, 247–252 (2008).
The first description of the RRL approach using NGS.
13. Wiedmann, R. T., Smith, T. P. & Nonneman, D. J. SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genet.* **9**, 81 (2008).
14. Ramos, A. M. *et al.* Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS ONE* **4**, e6524 (2009).
15. Amaral, A. J. *et al.* Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. *BMC Genomics* **10**, 374 (2009).
16. Amaral, A. J. *et al.* Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS ONE* **6**, e14782 (2011).
17. Kerstens, H. H. *et al.* Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics* **10**, 479 (2009).
18. Gore, M. A. *et al.* A first-generation haplotype map of maize. *Science* **326**, 1115–1117 (2009).
An example of the simplicity and power of reduced-representation sequencing for the development of whole-genome resources.
19. Sánchez, C. *et al.* Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics* **10**, 559 (2009).
20. van Bers, N. E. M. *et al.* Genome-wide SNP detection in the great tit *Parus major* using high throughput sequencing. *Mol. Ecol.* **19** (Suppl. 1), 89–99 (2010).
21. Hyten, D. L. *et al.* High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* **11**, 38 (2010).
22. Hyten, D. L. *et al.* High-throughput SNP discovery and assay development in common bean. *BMC Genomics* **11**, 475 (2010).
23. Esteve-Codina, A. *et al.* Partial short-read sequencing of a highly inbred Iberian pig and genomics inference thereof. *Heredity* 16 Mar 2011 (doi:10.1038/hdy.2011.13).
24. You, F. M. *et al.* Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genomics* **12**, 59 (2011).
25. Kraus, R. H. S. *et al.* Genome wide SNP discovery, analysis and evaluation in mallard (*Anas platyrhynchos*). *BMC Genomics* **12**, 150 (2011).
26. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
27. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
28. Pandey, V., Nutter, R. C. & Prediger, E. in *Next Generation Genome Sequencing: Towards Personalized Medicine* (ed. Janitz, M.) 29–42 (Wiley-VCH Weinheim, 2008).

29. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
30. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Rev. Genet.* **12**, 443–451 (2011).
31. Kerstens, H. H. *et al.* Structural variation in the chicken genome identified by paired-end next-generation DNA sequencing of reduced representation libraries. *BMC Genomics* **12**, 94 (2011).
32. van Orsouw, N. J. *et al.* Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE* **2**, e1172 (2007). **The original description of the CRoPS method.**
33. Mammadov, J. A. *et al.* Development of highly polymorphic SNP markers from the complexity reduced portion of maize [*Zea mays*, L.] genome for use in marker-assisted breeding. *Theor. Appl. Genet.* **121**, 577–588 (2010).
34. Gompert, Z. *et al.* Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Mol. Ecol.* **19**, 2455–2473 (2010). **An excellent demonstration of CRoPS, with a useful analysis technique for handling large population genomics data sets.**
35. Gompert, Z. & Buerkle, C. A. A hierarchical Bayesian model for next-generation population genomics. *Genetics* **187**, 903–917 (2011).
36. Davey, J. W. & Blaxter, M. L. RADSeq: next-generation population genetics. *Brief. Funct. Genomics* **9**, 416–423 (2010).
37. 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).
38. Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376 (2008). **The original description of high-throughput RAD-seq.**
39. Hohenlohe, P. A. *et al.* Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* **6**, e1000862 (2010).
40. Emerson, K. J. *et al.* Resolving postglacial phylogeography using high-throughput sequencing. *Proc. Natl Acad. Sci. USA* **107**, 16196–16200 (2010). **A demonstration of the power of RAD-seq for the study of non-model wild populations.**
41. Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W. & Luikart, G. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Mol. Ecol. Resour.* **11**, 117–122 (2011).
42. 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).
43. Pfender, W. F., Saha, M. C., Johnson, E. A. & Slabaugh, M. B. Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *Theor. Appl. Genet.* **122**, 1467–1480 (2011).
44. Baxter, S. W. *et al.* Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE* **6**, e19315 (2011).
45. Etter, P. D., Preston, J., 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).
46. Huang, X. *et al.* High-throughput genotyping by whole-genome resequencing. *Genome Res.* **19**, 1068–1076 (2009).
47. Xie, W. *et al.* Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc. Natl Acad. Sci. USA* **107**, 10578–10583 (2010).
48. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379 (2011). **The original description of the GBS method.**
49. Andolfatto, P. *et al.* Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* **21**, 610–617 (2011). **The original description of MSG, describing the hidden Markov model approach to imputation of genotypes.**
50. Krueger, F., Andrews, S. R. & Osborne, C. S. Large scale loss of data in low-diversity Illumina sequencing libraries can be recovered by deferred cluster calling. *PLoS ONE* **6**, e16607 (2011).
51. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10**, R32 (2009). **A useful study of the accuracy of variant detection in populations on the Roche Genome Sequencer, Illumina Genome Analyzer and Applied Biosystems SOLiD platforms.**
52. Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nature Methods* **5**, 1005–1010 (2008).
53. DeAngelis, M. M., Wang, D. G. & Hawkins, T. L. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* **23**, 4742–4743 (1995).
54. Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. & Cresko, W. A. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. in *Molecular Methods for Evolutionary Genetics* (eds Orgogozo, V. & Rockman, M. V.), Humana Press, New York (in the press).
55. Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. Low coverage sequencing: implications for the design of complex trait association studies. *Genome Res.* 1 Apr 2011 (doi:10.1101/gr.117259.110).
56. Futschik, A. & Schlotterer, C. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* **186**, 207–218 (2010).
57. Cutler, D. J. & Jensen, J. D. To pool, or not to pool? *Genetics* **186**, 41–43 (2010). **A useful discussion of the advantages and disadvantages of pooling samples for SNP calling.**
58. Bradbury, P. J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
59. Kofler, R. *et al.* PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE* **6**, e15925 (2011).
60. Pandey, R. V., Kofler, R., Orozco-terWengel, P., Nolte, V. & Schlötterer, C. PoPoolation DB: a user-friendly web-based database for the retrieval of natural polymorphisms in *Drosophila*. *BMC Genet.* **12**, 27 (2011).
61. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 4 May 2011 (doi:10.1101/gr.110882.110).
62. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nature Rev. Genet.* **12**, 87–98 (2011).
63. Barbazuk, W. B. & Schnable, P. S. SNP discovery by transcriptome pyrosequencing. *Methods Mol. Biol.* **729**, 225–246 (2011).
64. Chepelev, I., Wei, G., Tang, Q. & Zhao, K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res.* **37**, e106 (2009).
65. Cánovas, A., Rincon, G., Islas-Trejo, A., Wickramasinghe, S. & Medrano, J. F. SNP discovery in the bovine milk transcriptome using RNA-Seq technology. *Mamm. Genome* **21**, 592–598 (2010).
66. Geraldes, A. *et al.* SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Mol. Ecol. Resour.* **11** (Suppl. 1), 81–92 (2011).
67. Nothnagel, M. *et al.* Statistical inference of allelic imbalance from transcriptome data. *Hum. Mutat.* **32**, 98–106 (2011).
68. Christodoulou, D. C., Gorham, J. M., Herman, D. S. & Seidman, J. G. Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. *Curr. Protoc. Mol. Biol.* **94**, 4.12.1–4.12.11 (2011).
69. Kumar, S. & Blaxter, M. L. Comparing *de novo* assemblers for 454 transcriptome data. *BMC Genomics* **11**, 571 (2010).
70. Bräutigam, A., Mullick, T., Schliesky, S. & Weber, A. P. M. Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C3 and C4 species. *J. Exp. Bot.* 11 Mar 2011 (doi:10.1093/jxb/err029).
71. Hedges, D., Guettouche, T., Yang, S. & Bademci, G. Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. *PLoS ONE* **6**, e18595 (2011).
72. Kiialainen, A. *et al.* Performance of microarray and liquid based capture methods for target enrichment for massively parallel sequencing and SNP discovery. *PLoS ONE* **6**, e16486 (2011).
73. Cheng, Y. *et al.* Identification of novel SNPs by next-generation sequencing of the genomic region containing the APC gene in colorectal cancer patients in China. *OMICS* **14**, 315–325 (2010).
74. Teer, J. K. & Mullikin, J. C. Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.* **19**, R145–R151 (2010).
75. Teer, J. K. *et al.* Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res.* **20**, 1420–1431 (2010).
76. Stein, L. D. The case for cloud computing in genome informatics. *Genome Biol.* **11**, 207 (2010).
77. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Genet.* **19**, R227–R240 (2010).
78. Neely, R. K., Deen, J. & Hofkens, J. Optical mapping of DNA: single-molecule-based methods for mapping genomes. *Biopolymers* **95**, 298–311 (2011).
79. Lynch, M. Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* **182**, 295–301 (2009).
80. Rubin, C. *et al.* Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587–591 (2010).

Acknowledgements

We are grateful to P. Andolfatto, E. Buckler, W. Cresko, R. Elshire, E. Johnson, S. Mitchell, D. Stern and four anonymous referees for reviewing and discussing drafts of this manuscript. We thank S. Bassham, S. Baxter, C. Eland, K. Gharbi, M. Liu, J. Taggart, and P. Fuentes Utrilla for discussions that have improved our understanding of these methods. J.W.D. and M.L.B. are funded by the UK Natural Environment Research Council, grant NE/H019804/1. P.A.H. and J.M.C. received funding support from the US National Institutes of Health (NIH) grant 1R24GM079486-01A1, the US National Science Foundation grant IOS-0843392 and a Keck Foundation grant to W. Cresko. J.M.C. was also funded by the NIH National Research Service Award Ruth L. Kirschstein postdoctoral fellowship 1F32GM095213-01. P.D.E. was supported by grants R21HG003834 and R21HG006036 from the US National Human Genome Research Institute awarded to E. Johnson.

Competing interests statement

The authors declare competing financial interests: see Web version for details.

FURTHER INFORMATION

Paul A. Hohenlohe's homepage: http://people.oregonstate.edu/~hohenlop
Julian M. Catchen's homepage: http://pages.uoregon.edu/jcatchen
Mark L. Blaxter's homepage: http://www.nematodes.org
Ensembl: http://www.ensembl.org/index.html
Floragenex, Inc. (commercial RAD-seq provider): http://www.floragenex.com
Genotyping by sequencing in the Buckler laboratory: http://www.maizegenetics.net/Table/Genotyping-By-Sequencing
***Nature Reviews Genetics* series on Study designs:** http://www.nature.com/nrg/series/studydesigns/index.html
***Nature Reviews Genetics* series on Applications of next-generation sequencing:** http://www.nature.com/nrg/series/nextgeneration/index.html

Software packages for analysing NGS marker data
BAMOVA: http://www.uwyo.edu/buerkle/software/bamova
MSG: http://genomics.princeton.edu/AndolfattoLab/MSG.html
PoPoolation: http://code.google.com/p/popoolation
RADtools (within the United Kingdom RAD-seq Wiki page): http://radseq.info
Stacks: http://creskolab.uoregon.edu/stacks
TASSEL: http://www.maizegenetics.net/bioinformatics

ALL LINKS ARE ACTIVE IN THE ONLINE PDF