Published in final edited form as:
Methods Mol Biol. 2011 ; 772: 157–178. doi:10.1007/978-1-61779-228-1_9.

SNP Discovery and Genotyping for Evolutionary Genetics Using RAD Sequencing

Paul D. Etter, **Susan Bassham**, **Paul A. Hohenlohe**, **Eric A. Johnson**, and **William A. Cresko**

Abstract

Next-generation sequencing technologies are revolutionizing the field of evolutionary biology, opening the possibility for genetic analysis at scales not previously possible. Research in population genetics, quantitative trait mapping, comparative genomics, and phylogeography that was unthinkable even a few years ago is now possible. More importantly, these next-generation sequencing studies can be performed in organisms for which few genomic resources presently exist. To speed this revolution in evolutionary genetics, we have developed *R*estriction site *A*ssociated *D*NA (RAD) genotyping, a method that uses Illumina next-generation sequencing to simultaneously discover and score tens to hundreds of thousands of single-nucleotide polymorphism (SNP) markers in hundreds of individuals for minimal investment of resources. In this chapter, we describe the core RAD-seq protocol, which can be modified to suit a diversity of evolutionary genetic questions. In addition, we discuss bioinformatic considerations that arise from unique aspects of next-generation sequencing data as compared to traditional marker-based approaches, and we outline some general analytical approaches for RAD-seq and similar data. Despite considerable progress, the development of analytical tools remains in its infancy, and further work is needed to fully quantify sampling variance and biases in these data types.

Keywords

Genetic mapping; Population genetics; Genomics; Evolution; Genotyping; Single-Nucleotide Polymorphisms; Next-generation sequencing; RAD-seq

1. Introduction

Next-generation sequencing (NGS) technologies open the possibility of gathering genomic information across multiple individuals at a genome-wide scale, both in mapping crosses and natural populations (1,2). This breakthrough technology is revolutionizing the biomedical sciences (3-5) and is becoming increasingly important for evolutionary genetics (6). The rapidly decreasing cost of NGS makes it feasible for most laboratories to address genome-wide evolutionary questions using hundreds of individuals, even in organisms for which few genomic resources presently exist (7,8). This innovation has already led to studies in QTL mapping (9), population genomics (10), and phylogeography (11, 12) that were not possible even a few years ago.

Current NGS technology theoretically allows perfect genetic information – the entire genome sequence of an individual – to be collected (2). With rapidly decreasing sequencing costs, it may soon be feasible to completely sequence genomes from the large sample of individuals necessary for many population genomic or genome-wide association studies in organisms other than humans (6). However, genome resequencing is still prohibitively

expensive for most evolutionary studies. Fortunately, for many purposes, gathering complete genomic sequence data is an unnecessary waste of resources. For example, because linkage blocks are often quite large in a quantitative trait locus (QTL) mapping cross, progeny can be adequately typed with genetic markers of sufficient density (9). Similarly, many other evolutionary genetic studies require large numbers of genetics markers, but not necessarily complete coverage of the genome (6,10, 11).

An alternative approach to whole genome resequencing is to use NGS to gather data on dense panels of genomic markers spread evenly throughout the genome. The large number of short reads, provided by platforms such as Illumina, are ideal for this applicaion (7). A methodological difficulty is focusing these large numbers of repeated reads on the same genomic regions to maximize the probability that most individuals in a study will be assayed at orthologous regions. We developed a procedure called *R*estriction site *A*ssociated *D*NA sequencing (RAL-seq) that accomplishes this goal of genome subsampling (9). By focusing the sequencing on the same subset of genomic regions across multiple individuals, RAD-seq technology allows single-nucleotide polymorphisms (SNPs) to be identified and typed for tens or hundreds of thousands of markers spread evenly throughout the genome, even in organisms for which few genomic resources presently exist. Therefore, RAL-seq provides a flexible, inexpensive platform for the simultaneous discovery of tens of thousands of genetic markers in model and nonmodel organisms alike.

In this chapter, we describe the basic protocols for generating markers specifically for sequencing using the Illumina platform. Although we provide details for Illumina sequencing, these protocols can be modified for other sequencing platforms. We also provide a general framework for the analysis of RAL-seq data and discuss several primary bioinformatic considerations that arise from unique aspects of next-generation sequencing data as compared to traditional marker-based approaches. These include inferring marker loci and genotypes *de novo* from RAD tag sequences for an organism with no sequenced genome, distinguishing SNPs from error, inferring heterozygosity in the face of sampling variance, and tracking these sources of uncertainty through further analyses. New bioinformatic advances will certainly be made in this area. We, therefore, present these analytical approaches less as a set of concrete protocols, and more as a conceptual framework with a focus on potential sources of error and bias in next-generation genotyping data produced via protocols such as RAD-seq.

2. Materials

2.1. DNA Extraction, RNase A Treatment, and Restriction Endonuclease Digestion

1. DNeasy Blood & Tissue Kit (Qiagen) (see Note [1]).
2. RNaseA (Qiagen).
3. High-quality genomic DNA from 2.1: 25 ng/μl (see Note [2]).
4. Restriction enzyme (NEB; see Note [3]).

[1]Clean, intact, high-quality DNA is required for optimal restriction endonuclease digestion and is important for the overall success of the protocol. We have found that lower quality DNA can be used, but the starting amount will likely need to be increased because a large number of DNA fragments that have a correctly ligated P1 adapter may not end up in the proper size range when the starting DNA is partially degraded. When working with heavily degraded DNA samples is the only option, we have found that parameters of the protocol can be optimized (such as using more input DNA to start with and shearing less) to create usable libraries. These libraries often do not amplify as well as ones made with intact, high molecular weight genomic DNA. The “Best Practices” sections of the most recent Illumina Sample Prep Guides are a good resource for quantification, handling, and temperature considerations.

2.2. P1 Adapter Ligation, Purification and DNA Shearing

1. NEB Buffer 2.
2. rATP (Promega): 100 mM.
3. P1 Adapter: 100 nM stocks in l× Annealing Buffer (AB). Prepare 100 μM stocks for each single-stranded oligonucleotide in 1× Elution Buffer (EB: 10 mM Tris–Cl, pH 8.5). Combine complementary adapter oligos at 10 μM each in l× AB (10× AB: 500 mM NaCI, 100 mM Tris–Cl, pH 7.5–8.0). Place in a beaker of water just off boil and cool slowly to room temperature to anneal. Alternatively, use a boil and gradual cool program in a PCR machine. Dilute to 100 nM concentration in l× AB (see Notes [4] and [5]).
4. Concentrated T4 DNA Ligase (NEB): 2,000,000 U/ml.
5. QIAquick or MinElute PCR Purification Kit (Qiagen).
6. Bioruptor, nebulizer, or Branson sonicator 450.

2.3. Size Selection/Agarose Gel Extraction

1. Agarose.
2. 5× TBE: 0.45 M Tris–Borate, 0.01 M EDTA, pH 8.3
3. 6× Orange Loading Dye Solution (Fermentas).

---

[2]We recommend using a fluorescence-based method for DNA quantification to get the most accurate concentration readings. Since they bind specifically to double-stranded DNA, the dyes used in fluorometric assays are not as affected by RNA, free nucleotides, or other contaminants commonly found in DNA preparations (which can lead to inaccurate concentration predictions when using absorbance). If using another form of DNA quantification, such as UV spectrometer 260/280 absorbance readings, be sure to confirm the concentration by comparing a known calibration sample or running the sample on an agarose gel and comparing to a known quantity of DNA or ladder. We recommend checking the integrity of samples on a gel prior to embarking on this protocol regardless of the quantification method. Genomic DNA should consist of a fairly tight high-molecular-weight band without any visible degradation products or smears.

[3]The choice of the particular restriction enzyme to use for a study is based upon several parameters such as the desired frequency of RAD sites throughout the genome, GC content, the depth of coverage necessary, and size of the genome. For example, an average restriction endonuclease with an 8-bp recognition sequence will produce 1 tag every 64 kb in an organism with equal and random frequency of cut sites. Of course, the latter part of the previous sentence is hardly ever true, so the predicted and actual number of restriction sites in a genome can be quite different. For example, the number of SbfI sites (CCTGCAGG) in the stickleback genome (predicted from genome length and assuming equal distribution of each nucleotide) is 7,069, whereas we have identified 22,829 sites found at an average distance of 20.2 kb between sites. Yet *SgrD*I (CGTCGACG), with the same nucleotides as *Sbf*I, but in a different order, has 2892 sites found at an average distance of 160.2 kb between sites, nearly an eight-fold difference. In addition, the depth of coverage for calling a genotype in an outbred sample is quite a bit higher than what is necessary for an isogenic, recombinant inbred line. In general, RAD-seq experimental design is a challenge of optimizing the number of individuals, markers, and coverage given a fixed sequencing effort due to budgetary constraints.

[4]The P1 and P2 adapters are modified Solexa© adapters (2006 Illumina, Inc., all rights reserved). The presence of some salt is necessary for the double-stranded adapters used in this protocol to hybridize and remain stable at ambient temperatures and above. At a 1 mM salt concentration, the P1 adapter, which has 64 bases of complementary double-stranded sequence (assuming a 5-bp barcode), has a T_m of approximately 41°C (depending on barcode composition). P2, which has only 24 complementary bases, has a T_m of only 27°C at the same salt concentration. At 50 mM salt the T_ms increase significantly to ~69° and 56°, respectively. Care should be taken to allow reagents to cool to ambient temperature before double-stranded adapters are ligated to digested fragments so as not to denature them. In addition, heating short AT rich RAD fragments may result in their denaturation. Since single-stranded RAD fragments will not ligate to the double-stranded adapter overhangs, these RAD tags may be underrepresented in the final library. The P2 adapter used in Hohenlohe et al. (10) is longer than the adapter we used for our first forays into RAD sequendng. The new, longer P2 adapter has a higher T_m closer to that of the P1 adapters and the added benefit of being paired-end compatible. Thus, the salt concentration in the P2 ligation becomes less important. However, the longer P2 doesn't get effectively eliminated with a column cleanup, necessitating a gel extraction after amplification no matter how well P1 adapter titration has gone (see Note 14).

[5]Below are example barcoded *EcoR*I P1 and P2 adapter sequences. "P" denotes a phosphate group, "x" refers to barcode nucleotides, and an asterisk denotes a phosphorothioate bond in the P2 adapter that is introduced to confer nuclease resistance to the double-stranded oligo (14). Phosphorothioate bonds should be added to any 3′ overhangs on P1 adapters (e.g., *Sbf*I adapters).
P1 top: 5′-AATGATACGGCGACCACCGAGATCTACACTC TTTCCCTACACGACGCTCTTCCGATCTxxxxx-3′
AATTxxxxxAGATCGGAAGAGCGTCGTGTAGGG AAAGAGTGTAGATCTCGGTGGTCGCCGT ATCATT-3′
P2 top: 5′-P-CTCAGGCATCAaCGA?TCCTCCGAGAACAA-3′
P2 bot: 5′-CAAGCAGAAGACGGCATACGACGGAGGAAT CGAGTGATGCCTGAG*T-3′

4. GeneRuler 100 bp DNA Ladder Plus (Fermentas).
5. Razor blades.
6. MinElute Gel Purification Kit (Qiagen).

2.4. End Repair and 3'-dA Overhang Addition

1. Quick Blunting Kit (NEB).
2. NEB Buffer 2.
3. dATP (Fermentas): 10 mM.
4. Klenow Fragment (3′–5′ exo$^-$, NEB): 5,000 U/ml.

2.5. P2 Adapter Ligation and RAD Tag Amplification/Enrichment

1. NEB Buffer 2.
2. rATP: 100 mM.
3. P2 Adapter: 10 μM stock in l× AB prepared as P1 adapter described above (see Notes [4] and [5]).
4. Concentrated T4 DNA Ligase.
5. Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB).
6. RAD amplification primer mix: 10 μM. Prepare 100 μM stocks for each oligonucleotide in l× EB. Mix together at 10 μM (see Note [6]).

3. Methods

The protocol described below, outlined in Fig. 1, prepares RAD tag libraries for high-throughput Illumina sequencing (see Note [7]). In short, genomic DNA is digested with a restriction enzyme and an adapter (P1) is ligated to the fragments' compatible ends (Fig. 1a). This adapter contains forward amplification and Illumina sequencing priming sites, as well as a nucleotide barcode 4 or 5 bp long for sample identification. To reduce erroneous sample assignment due to sequencing error, all barcode differ by at least three nucleotides (see Note [8]). The adapter-ligated fragments are subsequently pooled, randomly sheared, and a specific size fraction is selected following electrophoresis (Fig. 1b). DNA is then ligated to a second adapter (P2), a Y adapter (13) that has divergent ends whose two strands are complementary for only part of their length (Fig. 1c). The reverse amplification primer is unable to bind to P2 unless the complementary sequence is filled in during the first round of forward elongation originating from the P1 amplification primer. The structure of this adapter ensures that only P1 adapter-ligated RAD tags will be amplified during the final PCR

[6]RAD amplification primers are Modified Solexa Amplification primers (2006 Illumina, Inc., all rights reserved).
P1-forward primer: 5′-AATGATACGGCGACCACCG*A-3′
P2-reverse primer: 5′-CAAGCAGAAGACGGCATACG*A-3′

[7]This protocol has been modified from that used in Baird et al. (9) and now incorporates critical improvements made since publication, including ones adopted from Quail et al. (14) and Illumina library preparation protocols. Although we recommend following the protocol as described, other companies may offer superior (or cheaper) versions of reagents that come at different enzyme concentrations or activities, which should work just as well. Using them may require additional optimization, including different incubation times or reaction volumes for efficient RAD-seq library preparation. Many other brands of DNA cleanup and gel extraction columns could instead be used also, but an important consideration is the minimum required elution volume. We have successfully substituted Zymo's DNA Clean and Concentrator for Qiagen's MinElute kit and 10 Weiss units/μl Epicentre T4 ligase instead of NEB'S 2000 cohesive end units/μl ligase.

[8]Three-mismatch barcodes are optimal because although a significant number of reads will have a sequencing error in the barcode, it is very unlikely that a single read will have two errors in the same 5-bp sequence. Therefore, most of the reads that have an error in the barcode can still be assigned to the correct sample when the adapters are designed to have three mismatches, whereas with only two mismatches a read with a sequencing error in the barcode may have come from one of two samples.

amplification step (Fig. 1d). The protocol for mapping of the lateral plate locus in threespine stickleback using *EcoR*I RAD markers reported in Baird et al. (9) is described here in detail as an example of the multiplexing approach. For bulk-segregant analysis, on the contrary, samples of like phenotypes can be pooled prior to digestion and treated as a single sample labeled with a single barcode.

3.1. DNA Extraction, RNase A Treatment, and Restriction Endonuclease Digestion

1. We recommend extracting genomic DNA samples using the DNeasy Blood & Tissue Kit (Qiagen) or a similar product that produces very pure, high molecular weight, RNA-free DNA. Follow the manufacturer's instructions for extraction from your tissue type. Be sure to treat samples with RNase A following the manufacturer's instructions to remove residual RNA. The optimal concentration after elution into buffer EB is 25 ng/μl or greater (see Notes [1] and [2]).
2. Digest 1 μg of genomic DNA for each sample with the appropriate restriction enzyme in a 50 μl reaction volume, following the manufacturer's instructions. For example, for *EcoRI* digestion, combine in a microcentrifuge tube the following: 5.0 μl 10× NEB Buffer 2, 0.5 μl *EcoR*I, DNA, and H_2O to 50.0 μl (see Notes [9] and [10]).
3. Heat-inactivate the restriction enzyme following manufacturer's instructions. If the enzyme cannot be heat-inactivated, purify with a QIAquick column following the manufacturer's instructions prior to ligation. QIAquick purification should work equally well, although conceivably the representation of fragments from distantly separated sites could be reduced when using an infrequent cutter. Allow reaction to cool to ambient temperature before proceeding to ligation reaction or cleanup (see Note [4]).

3.2. P1 Adapter Ligation

1. Ligate barcoded, restriction site overhang-specific P1 adapters onto complementary compatible ends on the genomic DNA created in the previous step (see Note [11]). If you are barcoding only a few individuals or samples, choose barcodes whose sequences differ as much as possible from one another to avoid causing the Genome Analyzer software to lose cluster registry, as it is prone to do when it encounters a nonrandom assortment of nucleotides (see Note [12]).
2. To each inactivated digest, add the following: 1.0 μl 10× NEB Buffer 2, 5.0 μl Barcoded P1 Adapter (100 nM), 0.6 μl rATP (100 mM), 0.5 μl concentrated T4 DNA Ligase (2,000,000 U/ml), 2.9 μl H_2O; 60.0 μl total volume (see Note [13]). Be sure to add P1 adapters to the reaction before the ligase to avoid religation of the genomic DNA. Incubate the reaction at room temperature for 30 min to overnight. Reduce the amount of P1 used in the ligation reaction if starting with less than 1 μg

[9]"H_2O" in this text refers to water that has a resistivity of 18.2 MΩ-cm and total organic content of less than five parts per billion.
[10]Set up larger reactions if necessitated by dilute DNA and then concentrate the samples with a column before proceeding to ligation. This will cut down on throughput of the protocol, of course.
[11]In general, when making master mixes, using multichannel pipettes and working with samples in 96- or 384-well plates will speed up the restriction digest and P1 ligation steps when multiplexing multiple barcoded individuals.
[12]In Baird et al. (9) DNA samples from 96 recombinant F_2 individuals were uniquely barcoded, which allowed us to track RAD markers and associate them with differing phenotypes. For example, F_2 individuals used in the mapping analysis included 60 possessing a complete compliment of lateral plate armor and 31 with a reduced number of plates. Up to 60 uniquely barcoded samples were pooled in a single library, and all samples were sequenced in two sequencing lanes. Sequences from each individual fish were sorted out in silico by barcode, allowing the genetic mapping of variation in the lateral plates as well as variation in another skeletal trait, the pelvic structure. For the bulk-segregant analysis, DNA from F_2 individuals was pooled by phenotype prior to digestion with *SbfI*. Each digested pool was labeled with a unique barcode and then combined into a single library. In both cases, the parental samples were uniquely barcoded and combined into single libraries that were sequenced along with the F_2 libraries.

genomic DNA or if cutting with an enzyme that cuts less frequently than *EcoR*I (e.g., we used 2.5 μl of adapter when using *Sbf*I in stickleback, which has close to 23,000 *Sbf*I restriction sites in its 460 Mb reference genome). It is critical to optimize the amount of P1 adapter added when a given restriction enzyme is used for the first time in an organism (see Note [14]).

3. Heat-inactivate T4 DNA Ligase for 10 min at 65°C. Allow reaction to cool slowly to ambient temperature before shearing.

3.3. Sample Multiplexing (see Note [15]) and DNA Shearing

1. Combine barcoded samples at an equal or otherwise desired ratio. Use a 100–300 μl aliquot containing 1–2 μg DNA total to complete the protocol and freeze the rest at −20°C in case you need to optimize shearing in the next step. In Baird et al. (9), DNA from F_2 stickleback progeny that shared a phenotype was pooled and each pool was uniquely barcoded for *Sbf*I bulk-segregant analysis. These samples were then combined at equal volumes with barcoded samples from each parent to create one library. In a second experiment, *EcoR*I-based libraries were made by pooling DNA samples from F_2 fish after they were barcoded individually by P1 ligation (see Note [12]).

2. Shear DNA samples to an average size of 500 bp to create a pool of P1-ligated molecules with random, variable ends. This step requires some optimization for different DNA concentrations and for each type of restriction endonuclease. The following protocol has been optimized to shear stickleback DNA digested with either *EcoR*I or *Sbf*I using the Bioruptor and is a good starting point for any study

---

[13] NEB Buffer 2 is used in the ligation reactions instead of ligase buffer because the salt it contains (50 mM NaCl) ensures the double-stranded adapters remain annealed during the reactions (see Note 4). T4 DNA Ligase is active in all 4 NEB Buffers if supplemented with 1 mM rATP, but doesn't work at maximum efficiency in NEB 3 because of the high levels of salt in that buffer. The presence of some salt is necessary for the double-stranded adapters to remain stable at ambient temperature during the ligation, but too much may cause a problem. Be aware of the amount of salt put into the ligation reactions and adjust the concentration of the adapters accordingly (for instance, if working with a frequent cutter, use lower volumes of P1 at 1 μM instead of higher volumes at 100 nM to cut down on salt added to the reaction). If the restriction buffer used for digestion does not contain 50 mM potassium or sodium ions, or if the restriction endonuclease cannot be heat-inactivated, purify the reaction in a column prior to P1 ligation and add 6.0 μ1 NEB Buffer 2. This will negate the benefits of multiplexing somewhat but may be useful in certain RAD library applications involving one or only a few individuals.

[14] *EcoR*I has been shown to work robustly in multiple organisms in our labs. Restriction enzymes that cut less frequently create fewer RAD tags and, thus, require more input DNA and less P1 adapter to keep the molar ratio approximately equal. Libraries produced with less frequent cutters are more difficult to amplify in general and protocol parameters may take some optimization for favorable results. It is critical to perform preliminary studies to optimize the appropriate amount of P1 adapter for a given restriction enzyme that is used for the first time in an organism, unless the actual number of sites is known (i.e., if a genome sequence is available). A range of P1 adapter to DNA ratios can be used in a preliminary study, and the efficiency of ligation can then be assayed via gel visualization. Alternatively, a more precise estimate of the quantity of correctly adapted fragments in each P1-DNA ratio can be determined after ligation of both adaptors via a qPCR reaction using primers designed to the P1 and P2 adapters. Over the correct range of adapter-DNA ratios, the amount of ligated DNA should increase and then asymptote. The inflexion point of this relationship is the ideal ratio of P1 adapter to DNA. If the ratio of P1 adapter overhangs to available genomic compatible ends is too low, you can get insufficient amplification and/or biased representation of some RAD tags. However, if the ratio of P1 to genomic overhangs is too high, a contaminant band that runs around 130 bp will appear after the final PCR reaction. If this contaminant overwhelms the amplification reaction it can lead to significant adapter sequence reads in the final sequencing output (even after gel extraction following the final PCR). This phenomenon is completely dependent upon the number of actual cut sites present in that genome and the corresponding amount of P1 adapter used. Our *Sbf*I study in stickleback used 2.5 μl P1 per microgram starting material and performed very well for library construction (see Figs. 3 and 4; lanes 2 and 4); however, this is likely due to the fact that there are actually more *Sbf*I sites than expected by chance (see Note 3). Therefore, it may be preferable to start with less P1 when working on genomes with closer to the expected number of sites. In addition, the prescribed amounts of P1 adapters used in this protocol were optimized for the studies published in Baird et al. (9), using adapters lacking phosphorothioate bonds. In our hands, RAD-seq libraries created using adapters that have the phosphorothioate modifications amplify better and appear much less prone to adapter contamination. Though we have not optimized the maximum amount to use for these new adapters, P1 concentrations in the ligation reactions can be increased.

[15] This step allows multiple individually barcoded samples to be combined and processed as well as to cut down on cost, work time, and differences in amplification efficiency that may arise between different library preparations when processing many samples at once.

(see Note [16]). The goal is to create sheared product that is predominantly smaller than 1 kb in size (see Fig. 1).

3. Dilute ligation reaction to 100 μl in water (or take 100–300 μl aliquot from multiplexed samples) and shear in the Bioruptor 10 times for 30 s on high following manufacturer's instructions. Make sure that the tank water in the Bioruptor is cold (4°C) before starting. All other positions in the Bioruptor holder not filled by your sample/s should be filled with balance tubes containing an equal volume of water.

4. Clean up the sheared DNA using a MinElute column following manufacturer's instructions. This purification is performed to remove the ligase and restriction enzyme, and to concentrate the DNA so that the entire sample can be loaded in a single lane on an agarose gel. Elute in 20 μl EB.

3.4. Size Selection/Agarose Gel Extraction and End Repair

1. This step in the protocol removes free unligated or concatemerized P1 adapters and restricts the size range of tags to those that can be sequenced efficiently on an Illumina Genome Analyzer flow cell. Run the entire sheared sample in 1× Orange Loading Dye on a 1.25% agarose, 0.5× TBE gel for 45 min at 100 V, next to 2.0 μl GeneRuler 100 bp DNA Ladder Plus for size reference; run the ladder in lanes flanking the samples until the 300 and 500 bp ladder bands are sufficiently resolved from 200 to 600 bp bands (see Fig. 2; Note [17]).

2. Being careful to exclude any free P1 adapters and P1 dimers running at ~130 bp and below, use a fresh razor blade to cut a slice of the gel spanning 300–500 bp (see Note [18]). Extract DNA using MinElute Gel Purification Kit following manufacturer's instructions with the following modification: to improve representation of AT-rich sequences, melt agarose gel slices in the supplied buffer at room temperature (18–22°C) with agitation until dissolved (usually less than 30 min) (14). Elute in 20 μl EB into a microcentrifuge tube containing 2.5 μl 10× Blunting Buffer from the Quick Blunting Kit used in the following step (see Note [19]).

3. The Quick Blunting Kit protocol converts 5′ or 3′ overhangs, created by shearing, into phosphorylated blunt ends using T4 DNA Polymerase and T4 Polynucleotide Kinase.

4. To the eluate from the previous step, add 2.5 μl dNTP mix (1 mM) and 1.0 μl Blunt Enzyme Mix. Incubate at RT for 30 min.

5. Purify with a QIAquick column. Elute in 43 μl EB into a microcentrifuge tube containing 5.0 μl 10× NEB Buffer 2.

---

[16] Although we have optimized our protocol for shearing via sonication, other forms of shearing should work (the "Alternate Fragmentation Methods" sections of the Illumina Sample Prep Guides are a good resource for important considerations when choosing a shearing method).

[17] We have found that it is unwise to run more than one library sample on the same agarose gel, as is shown in the figures (unless they will be combined and sequenced in the same lane on the flow cell) since it can lead to contamination between samples. This is especially important when dealing with samples following PCR amplification. We also recommend using aerosol-resistant filter tips for all amplification and downstream steps in the protocol to avoid library contamination.

[18] A wider size range can be isolated (Subheading 3.4, step 2) to have more RAD fragments to carry through the protocol if low template amounts are evident after the P2 ligation (see Note 2).

[19] Use MinElute columns and not QIAquick columns, which require a larger elution volume.

3.5. 3'-dA Overhang Addition and P2 Adapter Ligation

1. This step in the protocol adds an "A" base to the 3′ ends of the blunt phosphorylated DNA fragments, using the polymerase activity of Klenow Fragment (3′–5′ exo$^-$). This prepares the DNA fragments for ligation to the P2 adapter, which possesses a single "T" base overhang at the 3′ end of its bottom strand.
2. To the eluate from the previous step, add: 1.0 μl dATP (10 mM), 3.0 μl Klenow (exo$^-$). Incubate at 37°C for 30 min. Allow the reaction to cool to ambient temperature.
3. Purify with a QIAquick column. Elute in 45 μl EB into a microcentrifuge tube containing 5.0 μl 10× NEB Buffer 2.
4. This step in the protocol ligates the P2 adapter, a "Y" adapter with divergent ends that contains a 3′ dT overhang, onto the ends of DNA fragments with 3′ dA overhangs to create RAD-seq library template ready for amplification.
5. To the eluate from previous step, add: 1.0 μl P2 Adapter (10 μM), 0.5 μl rATP (100 mM), 0.5 μl concentrated T4 DNA Ligase. Incubate the reaction at room temperature for 30 min to overnight.
6. Purify with a QIAquick column. Elute in 50 μl EB and quantify (see Note [18]).

3.6. RAD Tag Amplification/Enrichment

1. In this step high-fidelity PCR amplification is performed on P1 and P2 adapter-ligated DNA fragments, enriching for RAD tags that contain both adaptors, and preparing them to be hybridized to an Illumina Genome Analyzer flow cell (see Fig. 1).
2. Perform a test amplification to determine library quality. In a thin-walled PCR tube, combine: 10.5 μl H_2O, 12.5 μl Phusion High-Fidelity Master Mix, 1.0 μl RAD amplification primer mix (10 μM), 1.0 μl RAD library template (or a quantified amount; see Note [20]). Perform 18 cycles of amplification in a thermal cycler: 30 s 98°C, 18× (10 s 98°C, 30 s 65°C, 30 s 72°C), 5 min 72°C, hold 4°C. Run 5.0 μl PCR product in 1× Orange Loading Dye out on 1.0% agarose gel next to 1.0 μl RAD library template and 2.0 μl GeneRuler 100 bp DNA Ladder Plus (Fig. 3).
3. If the amplified product is at least twice as bright as the template, perform a larger volume amplification (typically 50–100 μl) but with fewer cycles (12–14, to minimize bias), to create enough to retrieve a large amount of the RAD tag library from a final gel extraction (see Note [21]). If amplification looks poor, use more library template in a second test PCR reaction. Figure 3 shows three libraries used in Baird et al. (9) that amplified well, which is apparent when comparing the amplified product to the amount of template loaded in the lane to the right of each

[20]The optimal amount of template is dependent on the restriction enzyme used and its occurrence throughout the particular genome. It is difficult to be confident of the true concentration of amplified RAD tag molecules in your final sample, which have both P1 and P2 sequences, and are, therefore, able to bind the adapter oligonucleotides present on the Illumina flow cell. Poorly amplified libraries will contain a greater number of background sheared genomic DNA fragments with only P2 adapters attached, which cannot bind to the flow cell. A more precise estimate of fragments that have correctly ligated both P1 and P2 adaptors, and can therefore form clusters on the flow cell, can be ascertained by using qPCR with primers that are specific to the amplification priming sites on each adaptor.

[21]Libraries that amplify robustly, such as those shown in Fig. 3, can be amplified with only 14 or fewer cycles of amplification to avoid skewing the representation of the library (14). The goal is to use as few PCR cycles as possible to obtain robustly amplified libraries without amplification bias.

sample. Template should appear dim, yet visible on the gel. Purify the large volume reaction with a MinElute column. Elute in 20 μl EB.

4. The following purification step is performed to eliminate any artifactual bands that may appear due to an improper ratio of P1 adapter to restriction-site compatible ends (see Note [14]). Load the entire sample in 1× Orange Loading Dye on a 1.25% agarose, 0.5× TBE gel and run for 45 min at 100 V, next to 2.0 μl GeneRuler 100 bp DNA Ladder Plus for size reference (Fig. 4). Being careful to exclude any free adapters or P1 dimers running at ~130 bp and below, use a fresh razor blade to cut a slice of the gel spanning ~350–550 bp. Extract DNA using MinElute Gel Purification Kit following manufacturer's instructions, but melt agarose gel slices in the supplied buffer at room temperature. Elute in 20 μl EB (see Note [22]).

5. Quantify the DNA using a fluorometer to accurately measure the concentration. Concentrations will range from 1 to 20 ng/μl. Determine the molar concentration of the library by examining the gel image and estimating the median size of the library smear, which should be around 450 bp. Multiply this size by 650 (the average molecular mass of a base-pair) to get the molecular weight of the library. Use this number to calculate the molar concentration of the library (see Note [23]).

6. Sequence libraries on Illumina Genome Analyzer following manufacturer's instructions (see Note [24]).

3.7. Alignment Against a Reference Genome and *De Novo* Assembly

A significant consideration for the analysis of RAD sequencing data is whether the organism of interest has a reference genome. If it does, then RAD sequences can be aligned against the genome directly, and SNPs can be called as depicted in Fig. 5 and described below. However, because the length of reads is still relatively short on most NGS platforms, at least some of the reads will fall in repetitive regions that are similar across several parts of the genome, which will be evidenced by potential assignment with equal probability to these locations. These reads can be removed from the analysis. Alternatively, paired-end sequencing can be performed to help infer the correct location of each read in the genome.

When a reference genome does not exist RAD genotyping can still be performed by assembling reads with respect to one another. For genomic regions that are unique, this assembly works as well as aligning against the genome. The aligned stack of reads can be analyzed to determine SNPs and genotypes that can be used in a genetic map or population genomic analysis. Repetitive regions are problematic as above, but the additional information provided by aligning to multiple genomic regions in the reference genome is absent, making the identification of these repetitive regions even more difficult. Stacks of reads may be abnormally deep because the focal RAD site was by chance sampled significantly more than average, or because paralogous regions with similar sequences were erroneously assembled together. For stacks that fall outside the range of the expected number of reads, the data can simply be expunged from the analysis. More problematic is a situation where two or a small number of paralogous regions are assembled, and the total

[22]For long-term storage of DNA samples, Illumina recommends a concentration of 10 nM and adding Tween-20 to the sample to a final concentration of 0.1%. This helps to prevent adsorption of the template to plastic tubes upon repeated freeze–thaw cycles, which would decrease the effective DNA concentration and, therefore, the number of sequencing clusters the library will produce.

[23]For example, a measured DNA concentration of 10 ng/μl expressed in grams per liter is 0.01 g/l. 650 g/mol/bp × 400 bp = 292,500 g/mol = 0.0002925 g/nmol. To calculate the nanomolarity, divide 0.01 g/l by 0.0002925 g/nmol to get 34.2 nmol/l or 34.2 nM.

[24]We recommend that you validate your first one or two RAD-seq libraries by cloning 1.0 μl of the gel-purified library into a blunt-end compatible sequencing vector. Sequence individual clones by conventional Sanger sequencing. Confirm that insert sequences contain the correct barcodes and restriction cut site overhang and are from the genomic source DNA.

number of reads is large, but not significantly so, because of sampling variation across all RAD loci.

In situations where paralogous regions are mistakenly assembled, one of several things can be done. First, the length of the reads can be increased, or fragments can be paired-end sequenced, with the hope of obtaining unique information that can be used to tell paralagous regions apart. These solutions increase the cost of the sequencing to an extent that may not be justified by the increase in information. If data are being collected from individuals in a population or mapping cross, tests of Hardy Weinberg Equilibrium (HWE) may allow identification of problematic, incorrectly identified "genotypes" that are really paralogous regions. For example when two monomorphic, paralogous regions are fixed for alternative nucleotides and then mistakenly assembled as a single locus, a SNP will be inferred but no homozygotes will be identified. Lastly, a network-based approach can be employed with the expectation that true SNPs should be at significant frequency and surrounded in sequence space by a constellation of low frequency sequencing errors. In most populations SNPs will be binary, and therefore, the network will consist of two high frequency alleles at the center of the constellation. In situations where paralogous regions are incorrectly assembled, the network will have the easily identifiable topology of three or more high frequency alleles each surrounding by low frequency errors. All of these solutions are imperfect, and extracting information from *de novo* RAD tags and other NGS data will be a significant area of research in the near future.

3.8. Inferring Genotypes in the Face of Sequencing Error and Sampling Variance

Another challenge of next-generation sequencing data for bioinformatic analyses is the introduction of sequencing error into many of the reads. Although the sequencing error rate is quite low, in the order of 0.1–1.0% per nucleotide, it still becomes a significant source of inferential confusion when millions of reads are considered simultaneously. Unfortunately, sequencing error compounds the problems of assembling similar paralogous regions, outlined in the previous section, by increasing the probability of misassembly. Error rates can vary across samples, RAD sites, and positions in the reads for each site. In addition, the sampling process of a heterogeneous library inherent in NGS introduces sampling variation in the number of reads observed across RAD sites as well as between alleles at a single site. These issues could be overcome by greatly increasing total sequencing depth, but of course this approach will also increase the cost of a study. A better approach to differentiating true SNPs from sequencing error is a statistical framework that accounts for the uncertainty in genotyping. Undoubtedly significant progress will be made in this area in the near future; here we present one approach as an example of a straightforward, flexible statistical model.

The following maximum-likelihood framework is based upon Hohenlohe et al. (10), designed for genotyping diploid individuals sampled from a population. The expectation is that errors should be differentiated from heterozygous SNPs by the frequency of nucleotides, with errors being represented at low frequency while alleles at heterozygous sites in an individual will be present in near equal frequencies in the total number of reads. Modifications to this approach would be required in other cases, such as haploid organisms, recombinant inbred lines, backcross mapping crosses, or single barcodes representing pools of individuals.

For a given site in an individual, let n be the total number of reads at that site, where $n = n_1 + n_2 + n_3 + n_4$, and n_i is the read count for each possible nucleotide at the site (disregarding ambiguous reads). For a diploid individual, there are ten possible genotypes (four homozygous and six (unordered) heterozygous genotypes). A multinomial sampling distribution gives the probability of observing a set of read counts (n_1, n_2, n_3, n_4) given a

particular genotype, which translates into the likelihood for that genotype. For example, the likelihoods of a homozygote (genotype 1,1) or a heterozygote (1,2) are, respectively:

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

and

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

where ε is the sequencing error rate. If we let n_1 be the count of the most observed nucleotide, and n_2 be the count of the second-most observed nucleotide, then the two equations in (1) give the likelihood of the two most likely hypotheses out of the ten possible genotypes. From here, one approach is to assign a diploid genotype to each site based on a likelihood ratio test between these two most likely hypotheses with one degree of freedom. For example, if this test is significant at the $\alpha = 0.05$ level, the most likely genotype at the site is assigned; otherwise the genotype is left unassigned for that individual. In effect this criterion removes data for which there are too few sequence reads to determine a genotype, instead of establishing a constant threshold for sequencing coverage (10). An alternative approach is to carry the uncertainty in genotyping through all subsequent analyses. This can be done by incorporating the likelihoods of each genotype in a Bayesian framework in subsequent calculation of population genetic measures, such as allele frequency, or by using genotype likelihoods in systematic resampling of the data. In either case, information on linkage disequilibrium and genotypes at neighboring loci could also be used to update the posterior probabilities of genotypes at each site.

A central parameter in the model above is the sequencing error rate ε. One option for this parameter is to assume that it is constant across all sites (15), and either estimate it from the data by maximum likelihood or calculate it from sequencing of a control sample in the sequencing run. However, there is empirical evidence that sequencing error varies among sites, and alternatively ε can be estimated independently from the data at each site. Maximum likelihood estimates of ε are calculated directly at each site by differentiation of equations (9.1). This technique has been applied successfully to RAD-seq data (10). More sophisticated models could be applied here as well, for instance assuming a probability distribution from which ε is drawn independently for each site. This probability distribution could be iteratively updated by the data, and it could also be allowed to vary by nucleotide position along each sequence read or even by cluster position on the Illumina flow cell. Further empirical work is needed to assess alternative models of sequencing error.

3.9. Future Directions and Alternative Strategies

The analytical method described in the previous section accounts for sequencing error and the random sampling variance inherent in NGS data. However, it does not account for any systematic biases in, for instance, the frequency of sequence reads for alternative alleles at a heterozygous site. For example, biased representation could occur because PCR amplification occurs more readily on one allele or barcode. Barcoding and calling genotypes separately in each individual alleviates some of this bias. In addition, sampling variation among sites and alleles that occurs early in the process can be propagated and amplified in the RAD-seq protocol. Optimizing the protocol to minimize the number of PCR cycles required is an important component of dealing with this issue. However, to date no analytical theory or tools have been developed to handle these sources of variation, and numerical simulations and empirical studies are needed to quantify them. Most simply, individuals of known sequence could be repeatedly genotyped by RAD-seq, using replicate

libraries and barcodes, to estimate and partition the resulting variance in observed read frequencies.

Some evolutionary genetic applications will dictate alternative experimental designs, including sequencing of pools of individuals to produce point estimates (with error) of allele frequencies (15). Two examples are pools of individuals from natural populations for phylogeography (11), or groups of individuals of different phenotypic classes in a quantitative trait locus (QTL) mapping cross (9). These approaches lose information about individual genotypes, but because of the volume of data produced by NGS, techniques or analyses that lose some information remain highly effective. For instance, Emerson et al. (11) estimated the fine-scale phylogeographic relationships among populations of the pitcher plant mosquito *Wyeomia smithii* that originated postglacially along the eastern seaboard of USA. Because of the small amount of DNA in each mosquito, six individuals from each population were pooled and genotyped with barcoded adaptors. Rather than directly estimating allele frequencies, which could not be done with high confidence, the analysis focused only on those SNPs for which a statistical model indicated that allele frequency differed significantly between populations. This produced a set of nucleotides that were variable among, and fixed (or nearly so) within, populations. These data were used in subsequent standard phylogenetic analyses. The majority of the potentially informative SNPs were removed from the study, but because such a large number of RAD sites were identified and typed, the remaining 3,741 sites resulted in a beautifully resolved phylogeny, with high branch support and congruity to previous biogeographic hypotheses for this species.

3.10. Summary

Ever since the integration of Mendel's laws with evolutionary theory during the Modern Synthesis of the 1930s, biologists have dreamed of a day when perfect genetic knowledge would be available for almost any organism. Nearly a century later, NGS technologies are fulfilling that promise and opening the possibility for genetic analyses that have heretofore been impossible. Perhaps the most critical aspect of these breakthroughs is the unshackling of genetic analyses from traditional model organisms, allowing genomic studies to be performed in organisms for which few genomic resources presently exist. We presented one application of NGS, RAD genotyping, a focused reduced-representation methodology that uses Illumina next-generation sequencing to simultaneously discover and score tens to hundreds of thousands markers in a very cost-efficient manner. The core RAD-seq protocol can be performed in nearly any evolutionary genetic laboratory. Undoubtedly numerous modifications of the core RAD molecular protocols can be made to suit a variety of additional research problems. Despite the ease of use of RAD and other NGS protocols, a significant challenge facing biologists is developing the appropriate analytical and bioinformatic tools for these data. Although we outline some general analytical approaches for RAD-seq, we fully anticipate that the development of bioinformatic tools for RAD and similar data will be a rich area of research for many years.

Acknowledgments

The authors thank the University of Oregon researchers who, over the past 3 years, have helped troubleshoot many preliminary versions of this protocol. This work was funded by grants from the National Institutes of Health (1R24GM079486-01A1 and Ruth L. Kirschstein National Research Service Award F32 GM078949) and the National Science Foundation (IOS-0642264 and DEB-0919090).

References

1. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008; 24:133–141. [PubMed: 18262675]

2. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotech. 2008; 26:1135–1145.
3. Asmann YW, Wallace MB, Thompson EA. Transcriptome profiling using next-generation sequencing. Gastroenterology. 2008; 135:1466–1468. [PubMed: 18848555]
4. Marguerat S, Wilhelm BT, Bahler J. Next-generation sequencing: applications beyond genomes. Biochem Soc Trans. 2008; 36:1091–1096. [PubMed: 18793195]
5. Mortazavi A, Williams BA, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5:621–628. [PubMed: 18516045]
6. Rokas A, Abbot P. Harnessing genomics for evolutionary insights. Trends Ecol Evol. 2009; 24:192–200. [PubMed: 19201503]
7. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008; 9:387–402. [PubMed: 18576944]
8. Van Tassell CP, Smith TP, Matukumalli LK, et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods. 2008; 5:247–252. [PubMed: 18297082]
9. Baird NA, Etter PD, Atwood TS, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE. 2008; 3:e3376. [PubMed: 18852878]
10. Hohenlohe P, Bassham S, Stiffler N, et al. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 2010; 6:e1000862. [PubMed: 20195501]
11. Emerson KJ, Merz CR, Catchen JM, et al. Resolving post-glacial phylogeography using high throughput sequencing. Proc Natl Acad Sci USA. 2010; 107:16196–200. [PubMed: 20798348]
12. Gompert Z, Lucas LK, Fordyce JA, et al. Secondary contact between *Lycaeides idas* and *L. melissa* in the Rocky Mountains: extensive admixture and a patchy hybrid zone. Mol Ecol. 2010; 19:3171–3192. [PubMed: 20618903]
13. Coyne KJ, Burkholder JM, Feldman RA, et al. Modified serial analysis of gene expression method for construction of gene expression profiles of microbial eukaryotic species. Appl Environ Microbiol. 2004; 70:5298–5304. [PubMed: 15345413]
14. Quail MA, Kozarewa I, Smith F, et al. A large genome center's improvements to the Illumina sequencing system. Nat Methods. 2008; 5:1005–1010. [PubMed: 19034268]
15. Lynch M. Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics. 2009; 18:295–301. [PubMed: 19293142]

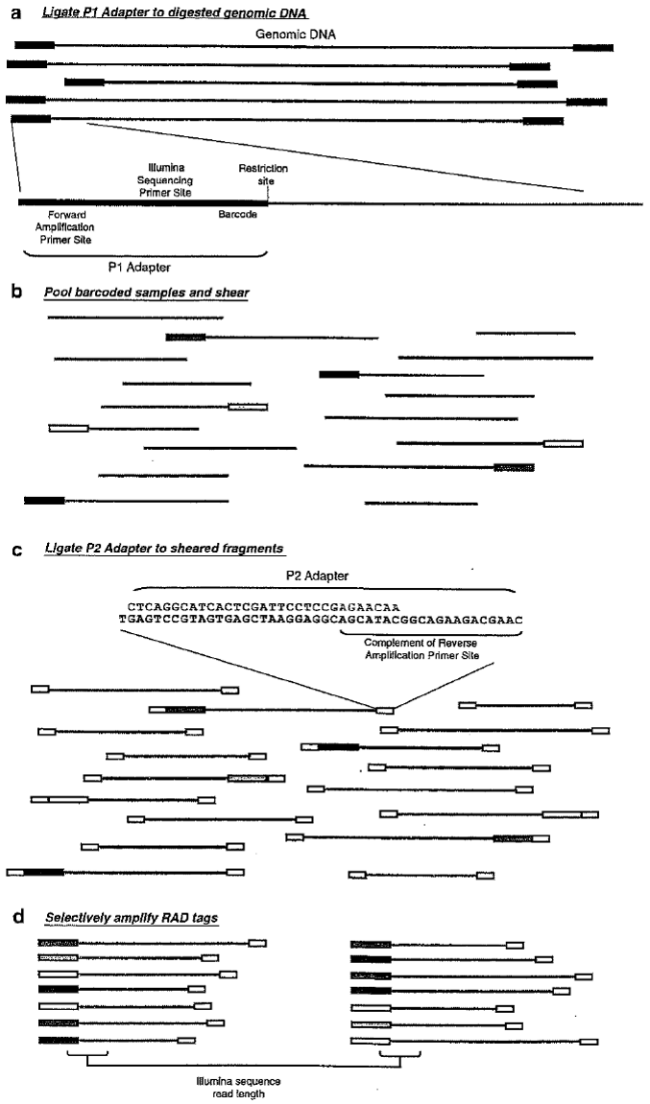

Fig. 1.

RAD tag library generation. (**a**) Genomic DNA is digested with a restriction enzyme and a barcoded P1 adapter is ligated to the fragments. The P1 adapter contains a forward amplification primer site, an Illumina sequencing primer site, and a barcode (*shaded boxes* represent P1 adapters with different barcodes). (**b**) Adapter-ligated fragments are combined (if multiplexing), sheared, and (**c**) ligated to a second adapter (P2, *white boxes*). The P2 adapter is a divergent "Y" adapter, containing the reverse complement of the P2 reverse amplification primer site, preventing amplification of genomic fragments lacking a P1 adapter. (**d**) RAD tags, which have a P1 adapter, are selectively and robustly enriched by PCR amplification (reproduced from ref. 9).

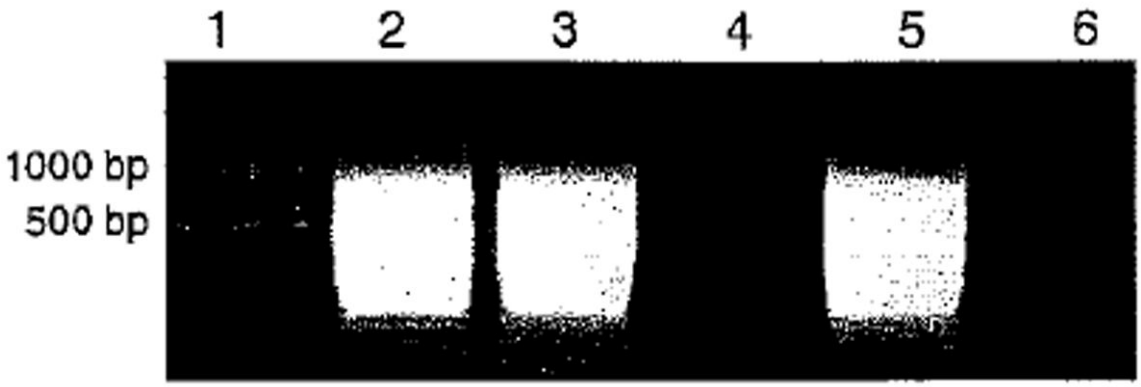

Fig. 2.

Three barcoded and multiplexed RAD tag libraries. Lanes 2, 3, and 5 each contain two DNA samples that were restriction digested, ligated to barcoded P1 adapters, combined, sheared, purified, and then loaded on an agarose gel. 2 – parental DNA samples cut with *Sbf*I. 3 – F_2 pools cut with *Sbf*I. 4 – blank. 5 – parental DNA samples cut with *EcoR*I. Libraries contain 2 μg total combined genomic DNA each. 1 and 6 – 2.0 μ1 GeneRuler 100-bp DNA Ladder Plus.

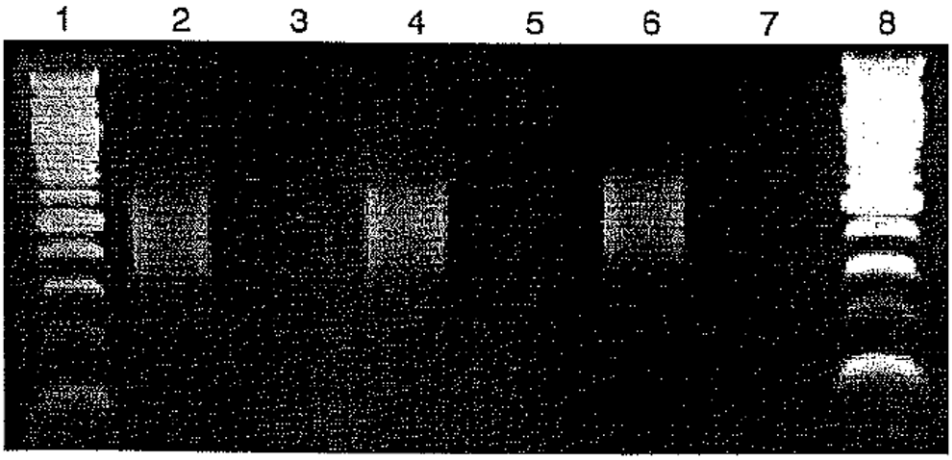

Fig. 3.

Test amplification PCR product from the three libraries shown in Fig. 2. Lanes 2, 4, and 6 contain 5.0 μl amplified PCR product. 2 – parental *Sbf*I library. 4 – F_2 *Sbf*I library. 6 – parental *EcoR*I library. Lanes 3, 5, and 7 contain 1.0 μl template used for amplification in the lane to the left. Template was loaded at 5× the amount used in the equivalent volume loaded for amplified reactions. 1 and 8 – 2.0 μl GeneRuler 100-bp DNA Ladder Plus. Libraries are 300–600 bp in size.

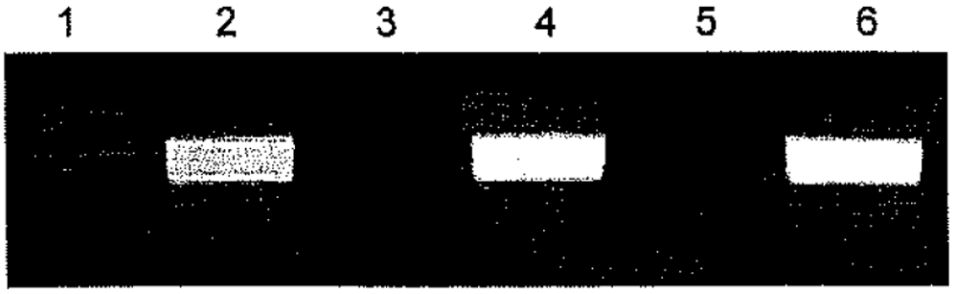

Fig. 4.

PCR product from the three libraries shown in Figs. 2 and 3 after the final large volume amplification and purification. Lanes 2, 4, and 6 each contain 20 μl purified PCR product from 100 μl amplifications. 2 – parental *Sbf*I library. 4 – F_2 *Sbf*I library. 6 – parental *EcoR*I library. 1 – 2.0 μl GeneRuler 100-bp DNA Ladder Plus. Lanes 3,5, and 7 are blank. Libraries are 300–600 bp in size.

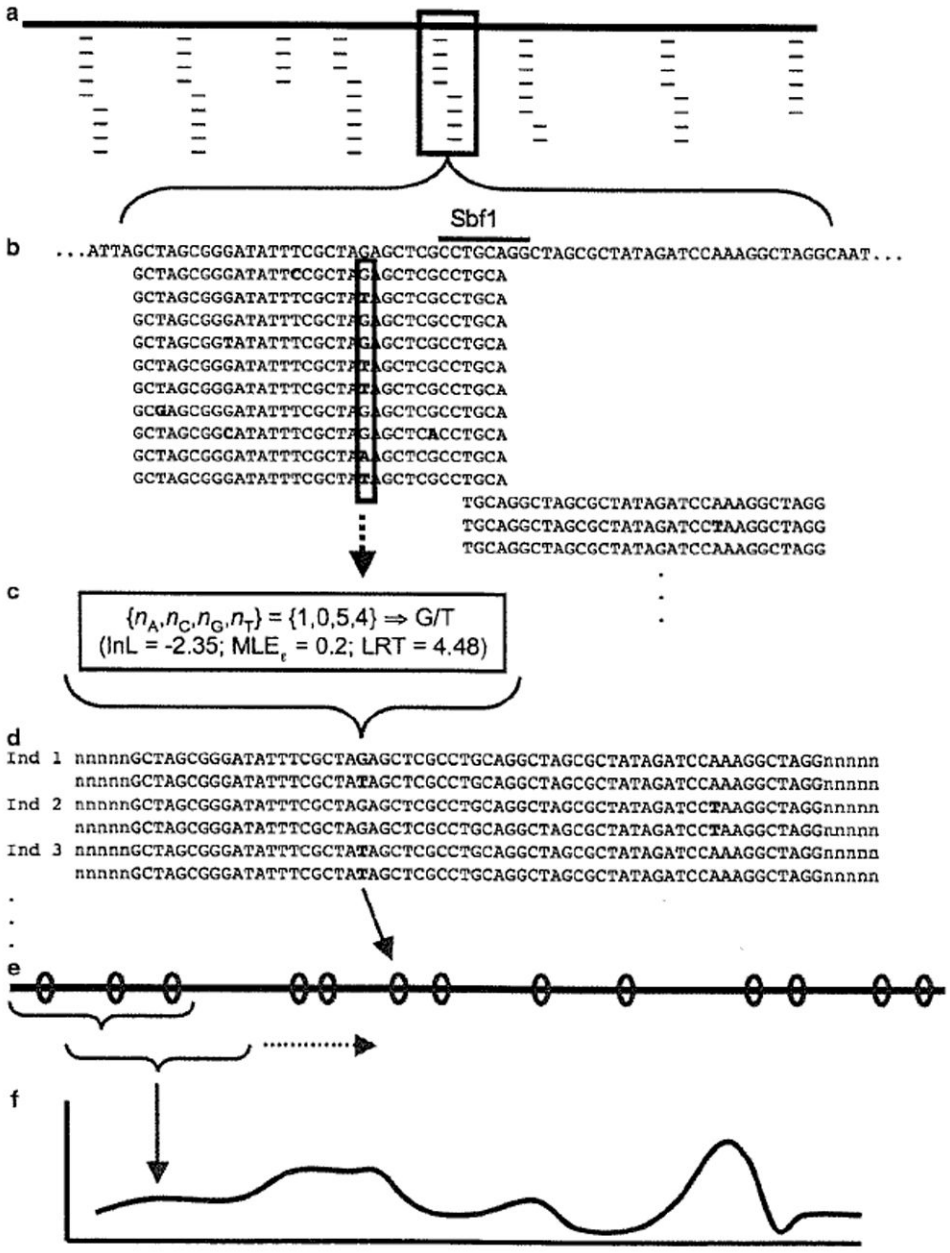

Fig. 5.

Schematic diagram of population genomic data analysis using RAD sequencing. **(a)** Following Illumina sequencing of barcoded fragments, sequence reads (*thin lines*) are aligned to a reference genome sequence (*thick line*). Depth of coverage varies across tags. Reads that do not align to the genome, or align in multiple locations, are discarded. **(b)** Sample of reads at a single RAD site. The recognition site for the enzyme *Sbf*I is indicated along the reference genome sequence (*top*), and sequence reads typically proceed in both directions from this point, at which they overlap. At each nucleotide site, reads showing each of the four possible nucleotides can be tallied (*solid box*). **(c)** Nucleotide counts at each site for each individual are used in a maximum likelihood framework to assign the diploid genotype at the site. In this example, G/T heterozygote is the most likely genotype; the method provides the log-likelihood for this genotype, a maximum-likelihood estimate for

the sequencing error rate ε, and a likelihood ratio test statistic comparing G/T to the second-most-likely genotype, G/G homozygote. **(d)** Each individual now has a diploid genotype at each nucleotide site sequenced, and single-nucleotide polymorphisms (SNPs) can be identified across populations. Note, however, that haplotype phase is still unknown across RAD tags. **(e)** SNPs (*ovals*) are distributed across the genome (*thick line*), and population genetic measures (e.g., F_{ST}) are calculated for each SNP. **(f)** A kernel smoothing average across multiple nucleotide positions is used to produce genome-wide distributions of population genetic measures.