Current Protocols in Plant Biology: Genotyping-by-sequencing

Jason G. Wallace[1*] and Sharon E. Mitchell[2]

[1] The University of Georgia, Department of Crop & Soil Sciences, Athens, Georgia, USA. 706-542-9696 (phone). 706-583-8120 (fax). jason.wallace@uga.edu
[2] Cornell University, Genomic Diversity Facility, Ithaca, New York, USA. 607-254-4849 (phone). sem30@cornell.edu

SIGNIFICANCE

Genotyping-by-sequencing (GBS) is a protocol to identify large numbers of genetic markers in individuals at a low cost per data point. It is designed to be easily scalable to high-throughput platforms, and many of the steps can be performed either by hand or by a liquid handling robot. The broad applicability and low per-sample cost of GBS data has made it very popular for generating marker data, especially for organisms with few existing genomic resources.

ABSTRACT

Genotyping-by-sequencing (GBS) refers to a suite of related methods that obtain genotype data from samples by using restriction enzyme digestion followed by high-throughput sequencing. GBS is a refinement of restriction site-associated DNA sequencing (RADseq) methods, with the specific aim of being able to perform library preparation quickly, cheaply, and in high-throughput. This protocol contains the steps necessary to go from purified DNA to Illumina-ready libraries. It also covers the considerations that go into planning a GBS experiment.

Keywords: Genotyping, high-throughput sequencing, GBS, RADseq, restriction enzyme

INTRODUCTION

Many questions in biology revolve around the genetic differences among individuals. Modern sequencing makes it possible to sequence the entire genome of almost any organism, but in many cases it is just as effective (and much cheaper) to get information on a subset of locations. Several “reduced representation” technologies have been developed to acquire in-depth data on parts of the genome while ignoring the rest.

One of the most popular reduced representation methods is genotyping-by-sequencing (GBS; Elshire et al., 2011). GBS uses one or more restriction enzymes to target specific locations throughout the genome for sequencing. This protocol follows the original one-enzyme GBS procedure of Elshire et al. (2011), and the general workflow is diagrammed in Figure 1. The protocol starts with purified sample DNA and finishes with a GBS library ready to be sequenced on an Illumina sequencer. Alternate Protocol 1 outlines how to adapt these methods for two-enzyme GBS.

Strategic planning:

Sample collection. The exact method for collecting DNA samples varies by organism, but should be performed in a way that yields enough DNA for analysis while minimizing contamination from other samples. Since the default GBS protocol involves relatively low coverage per individual, we do not recommend bulking heterogeneous individuals (that is, pooling DNA from multiple, genetically distinct individuals) since there will probably not be enough read depth at any given locus to accurately call allele frequencies. (See the Commentary, however, for ways this can be overcome.) Even when working with inbred organisms, we recommend sampling only a single individual, if possible, to minimize the chance of contamination.

DNA extraction. There are too many different DNA extraction protocols to consider them all here, so this unit assumes that the researcher has already isolated high-quality DNA from their samples. In our experience, column-based purification kits (such as offered by QIAGEN, Omega Bio-tek, Zymogen, or others) are the most reliable because they do the best job of removing polysaccharides and other metabolites that can inhibit downstream steps. These kits can be expensive, though, so it can be worth your time to test a cheaper method first. (Preferably on samples that are not critical to the experiment.) For example, the CTAB method (Rogers and Bendick, 1994) is a popular protocol to isolate DNA from plants using phase separation. For some plants the subsequent GBS libraries work fine, but for others they fail entirely. There may be little pattern to sample failure, and it is essentially impossible to predict in advance whether a new plant species will work with CTAB extraction or not. Also, DNA extracted from medicinal plants are often recalcitrant to RE digestion even with column-based purification. In these cases, it may be necessary to add 2.8% polyvinylpyrrolidone-40 (PVP-40) to the extraction buffer prior to column purification. (See the “Troubleshooting” section for more details.)

Restriction enzyme digestion. GBS involves digesting the sample DNA with one or more methylation sensitive restriction enzymes (REs) to limit how much of the genome is sequenced. Because these enzymes do not cut highly methylated, repetitive DNA, they enrich for the low-copy genomic regions that most researchers are interested in.

The choice of restriction enzyme is critical for a successful GBS run, since it has one of the largest impacts on the resulting data (Figure 2). Frequent-cutting REs cut at many places in the genome, generating many more sites with genotype information but with much lower sequence coverage per site (and thus more missing data). Rarer-cutting enzymes target fewer sites but generate better quality data at each of them (i.e., higher coverage per site, fewer missing SNP calls and higher confidence for calling heterozygous genotypes). Some researchers combine two enzymes with different specificities to obtain an intermediate number of sites with moderate coverage.

Genome size also plays an important role in the final dataset: using a certain restriction enzyme on a 22 Gbp loblolly pine sample will generate very different quality data than using the same enzyme on *Arabidopsis*, with a genome more than 100 times smaller. These differences are modulated by the amount of methylation in the genome, which determines how many sites are actually accessible to methylation-sensitive REs. Small, unmethylated genomes often produce more restriction fragments than expected based on genome size alone, while large, highly methylated genomes may produce fewer than expected.

GBS adapters. Each adapter consists of two oligonucleotides—a top strand and a bottom strand—that when combined contain sticky ends for ligation and Illumina sequences for PCR amplification (Figure 1). One adapter is a common adapter used in all reactions. The other contains a short “barcode” sequence unique to each adapter that is used to identify which sample each fragment came from. We strongly recommend using an existing set of adapters rather than trying to create your own, since there are many considerations that go into making good adapter sequences, especially with regard to the barcodes. (See Elshire et al., 2011.) A list of proven adapters compatible with ApeKI and PstI restriction enzymes are given in Supplemental Tables 1 and 2, respectively; these can also be used with other restriction enzymes that leave the same overhang (e.g., EcoT22I or SbfI for the PstI adapters). If you

need to use a different enzyme or combination of enzymes, we recommend doing a literature search first to see if another group has already developed appropriate adapters. If so, the sequences are often included as supplemental material, such as for the PstI-MspI double digestion of Poland et al. (2012).

The oligonucleotides for the barcoded and common adapters can be ordered in various quantities, depending on the number of reactions to be performed. Be aware that the smallest synthesis scale (usually 10 nmole) will provide enough adapter for tens of thousands of reactions. Although it is not necessary to have adapter oligos purified by HPLC, some form of additional purification (beyond standard desalting) is recommended. For example, some companies offer proprietary, column-based purification that removes truncated oligonucleotides that result from failed synthesis chemistry. Adapter oligos should not be 5′ phosphorylated; this stabilizes adapter dimers, which then become the major product in the sequencing library.

Although the ligation reaction can result in products that have the same adapter on both ends, these products do not amplify well on flowcells and so are generally not a problem (Elshire et al., 2011).

PCR amplification. One PCR reaction is used to both amplify the GBS library and add the Illumina sequencing regions to the fragments. The primers for this step are the standard paired-end Illumina primers, with sequences 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′ (Primer 1) and 5′-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT-3′ (Primer 2).

Basic protocol 1: One-enzyme genotyping-by-sequencing

This protocol involves three stages: Enzymatic digestion of DNA, ligation of barcoded sequencing adapters, and PCR amplification of pooled libraries. The reagent amounts provided are for a single 96-well plate. Preparing these libraries by hand usually results in large sample-to-sample variability in the amount of sequence data recovered (Elshire *et al.* 2011). Minimizing this variability goes a long way toward improving GBS data quality, so we strongly recommend using a liquid handling robot if one is available.

Prior to starting the protocol, the user will need to have acquired purified DNA and adjusted the samples to 30 ng/μL. Many different techniques can be used to measure DNA concentration for this adjustment, although we recommend the use of intercalating dyes (e.g., QuantiFluor or PicoGreen) over absorption-based methods (e.g., Nanodrop) because of their higher accuracy. We also recommend that a negative control (water only) is included on each 96-well plate. If multiple plates are sequenced simultaneously, place the negative control in different positions on each plate. This will help identify plates (and potential plate-swaps) after the samples are sequenced.

Reagents:

- Concentrated adapter stock plate at 3 ng/μL (see Support Protocol 1)
- 96-well plate of purified sample DNAs, adjusted to 30 ng/μL

- Primer stocks of PCR Primer 1 and Primer 2 at 50 μM in 1X TE or water
- Restriction enzyme of choice
- 10X restriction enzyme buffer
- T4 DNA ligase
- 10X T4 DNA ligase buffer with ATP
- 2X Taq polymerase master mix
- HPLC-grade, sterile water
- 1X Tris-EDTA (TE) buffer (10 mM Tris, pH 8.0 + 1 mM EDTA)
- Commercial column-based PCR cleanup kit (e.g., Qiagen QIAquick PCR Purification Kit or equivalent) with elution buffer, enough for 2 cleanups.
- Dilution buffer (10 mM Tris pH 8.0 + 0.1% Tween20)

Materials

- 1.5 ml centrifuge tube x 1
- 15 ml conical tube x 2 (or another container capable of holding 2-4 ml of master mix)
- 96-well PCR plates x 2 (1 to make a working adapter stock, 1 to make the GBS library)
- Air-permeable plate-sealing tape x 2
- Plate-sealing foil (at least 2 sheets)

Equipment:

- Thermocycler
- Vortex mixer
- Centrifuge(s) for spinning plate and single tubes
- Laminar-flow hood or SpeedVac for drying samples (Do not use a lyophilizer as sample droplets will move and possibly contaminate other wells if a strong vacuum is applied without centrifugation).

Restriction digestion of purified DNA

1. In a fresh 96-well plate, add concentrated adapter stock (Support Protocol 1) and water to create a working adapter stock. The exact dilution depends on species, though for most the amounts shown in Table 1 should work. (A complete list of what concentrations work for >300 species we have tested is in Supplemental Table 3; see also the section on “Adapter Concentration” in the Commentary.) For example, most species require 3.6 ng of adapter pairs when making libraries from ApeKI digests. Since the concentrated stock (Support Protocol 1) is at 3 ng/μL, one would add 10 μL of concentrated stock to 40 μL of water to create 50 μL of working stock at 0.6 ng/μL. This is enough to prepare 8 plates of samples.
2. In a separate 96-well plate, add 6 μL of working adapter stock to each well. Cover the plate with air-permeable tape and spin down briefly to collect the liquid. Place the plate in a laminar-flow hood or SpeedVac to dry out the liquid.
3. Remove the tape and add 100 ng of sample DNA (3.3 μL from the 30ng/uL sample plate) to each well. Cover again with air-permeable tape, spin down, and evaporate as in Step 2.
4. Create a restriction digest master mix in a 15 ml conical tube (or other container), consisting of:
    - 220 μL 10x restriction enzyme buffer

    - 100 µL (100U) restriction enzyme (Adjust RE and water volumes if the enzyme concentration is > 1U/uL. A 10-fold excess of RE should be provided, i.e., 1U enzyme/100ng DNA).
    - 1880 µL water
5. Add 20 µL of master mix to each well in the plate. Seal the plate and incubate for 2 hours at the recommended temperature for the enzyme. Transfer to 4° C. (Note: because GBS adapters are designed to not regenerate the restriction site after ligation, it is not necessary to deactivate the RE.)

<u>Ligation of samples and adapters</u>

6. Make a master mix in a 15 ml conical tube (or other container), consisting of:
    - 550 µL 10X T4 DNA ligase buffer with ATP
    - 176 µL T4 DNA ligase (= 70,400 cohesive end units)
    - 2,574 µL water
7. Add 30 µL of master mix to each well of the digestion reaction. Seal the plate and spin down to collect the liquid.
8. Incubate the reaction at 22° C for 1 hour, then shift to 65° C for 30 minutes to deactivate the ligase. Hold at 4° C.

<u>Pool and clean up samples</u>

9. Combine 5 µL of each sample into a single 1.5 mL sample tube, using a fresh tip for each well.
10. Purify the combined samples with a commercial column-based PCR cleanup kit and elute with 50 µL elution buffer.
11. Prepare a PCR reaction mix consisting of:
    - 2 µL pooled library DNA (Step 10)
    - 25 µL 2x Taq polymerase master mix
    - 1 µL of 50 µM PCR Primer 1
    - 1 µL of 50 µM PCR Primer 2
    - 21 µL water
12. Amplify by PCR using the following program:
    - 5 minutes at 72° C
    - 30 seconds at 98° C
    - 18 cycles of:
        - 10 seconds at 98° C
        - 30 seconds at 65° C
        - 30 seconds at 72° C
    - 5 minutes at 72° C
    - Hold at 4° C
13. Purify the PCR products using a column-based kit, eluting into 30 µL of elution buffer, water, or 0.1X TE buffer.
14. Run library on a capillary instrument (Experion or BioAnalyzer) to evaluate the fragment size distribution and whether repetitive DNA is amplified. Repetitive fragments will appear as discrete bands on the electropherogram.

15. The library will need to be diluted prior to sequencing. (We recommend using 10 mM Tris, pH 8.0 + 0.1% Tween20.) An explanation of the dilution calculation, along with a calculator function in Excel, are included in Supplemental Table 4. Alternatively, many sequencing facilities now perform qPCR on all samples, and this can also be used to determine the proper dilution level. Regardless of the method, after dilution the library is ready for sequencing on any Illumina sequencer.

Alternate protocol 1: Two-enzyme restriction digestion

Some researchers find a benefit to using a two-enzyme GBS protocol on their materials, so that only regions with both cut sites in proximity are selected. The usual protocol pairs a rare-cutting enzyme with a more common-cutting enzyme, which results in an intermediate number of sequenced sites relative to using either enzyme alone (Figure 2C).

Two-enzyme GBS preparation is similar to one-enzyme GBS, with two key differences:

- The barcoded adapters need to contain a different restriction site overhang than the common adapter.
- In steps 4 & 5, the master mix buffer and incubation temperature need to support digestion from both restriction enzymes.

Poland et al. (2012) recommend a Y-adapter (where part of the top and bottom strands do not match) instead of the common adapter to allow for PCR selection of only products that include one of each restriction site. Otherwise, the library would be dominated by the many small fragments produced by the frequent-cutting enzyme. They also recommend that this adapter be added at 50-100x the concentration of the barcoded adapters. (Their full protocol is available online as supplemental data for Poland et al. (2012).)

Support Protocol 1: Preparation of adapters and a concentrated stock plate

This protocol details how to make a stock plate of concentrated GBS adapters. This concentrated stock is then diluted to make a working stock for GBS library preparation (Basic Protocol 1, Step 1).

Reagents

- 1X Tris-EDTA (TE) buffer (10 mM Tris, pH 8.0 + 1 mM EDTA)
- Separate oligonucleotide stocks of all adapter components (top and bottom strand oligonucleotides for the common adapter and all barcoded adapters) at 200 μM concentration in TE buffer

Materials

- 96-well PCR plates x 2
- Deep-well (at least 1.1 mL) 96-well plate x 1
- Plate-sealing foil (2-3 sheets)
- 0.2 mL PCR tube x1
- 1.5 ml centrifuge tube x 1

Equipment

- Thermocycler
- Vortex mixer
- Centrifuge(s) for spinning plate and single tubes
- Equipment to quantify DNA (via absorption, intercalating dye, etc.)

1. Combine 25 μL of the common adapter top strand oligo, 25 μL of the common adapter bottom strand oligo, and 50 μL of TE buffer in a 0.2 mL PCR tube, for a total volume of 100 μL. In a 96-well PCR plate, repeat this for each of the barcoded adapters, making note of which adapter is in which well. Seal the plate.
2. Place the combined oligos into a thermocycler with the following program:
   - 95° C for 2 minutes
   - Ramp down to 25° C at 0.1° C per second
   - Hold at 25° C for 30 minutes
   - Hold at 4° C until transferred to refrigerator or freezer
3. In a deep-well 96-well plate, combine 5.6 μL of each annealed barcoded adapter with 995 μL of TE buffer. Seal the plate, vortex, and spin briefly in a centrifuge to collect the solution. Meanwhile, combine the 100 μL of common adapter with 900 μL of TE buffer in a 1.5 mL centrifuge tube.
4. Quantify each adapter solution according to your method of choice. We recommend using some type of intercalating dye instead of just absorbance because of the higher accuracy.
5. Transfer 300 ng of each barcoded adapter to the corresponding well of a fresh PCR plate. Add 300 ng of the common adapter to each well, plus enough TE buffer to make 200 μL. Seal the plate, vortex, and spin down to collect the liquid. This plate now contains concentrated adapter pairs at 3 ng/μL and serves as the source from which working stocks are made for each library preparation.

Support protocol 2: Optimization of adapter concentration

Optimal GBS preparations rely on an appropriate ratio of adapter to genomic DNA, which depends on the frequency of enzyme cut sites in the DNA. Too much adapter and you get lots of adapter dimers, too little and the genomic DNA ligates primarily to other genomic DNA. These unwanted processes always happen to some degree, but the goal is to minimize them.

Initially, we performed titration experiments to determine optimal amounts of adapter each time a methylation-sensitive RE was used in a different species. As more species were analyzed, it became evident that, at least for most diploids, the size of the unmethylated portion of the genome did not vary greatly, even among species with very different genome sizes. This observation is consistent with the fact that genome expansion is caused by the accumulation of repetitive DNAs (mainly transposons) and these repetitive regions are normally hypermethylated (i.e., invulnerable to digestion with methylation-sensitive REs). A list of adapter concentrations that are known to work in different species or groups is provided in Supplemental Table 3. If your species does not fall into any of these groups, or if you want to optimize adapter amounts for your species independently, use the following protocol. This is set up for testing adapters at eight different concentrations, though the researcher is free to add more if desired.

Reagents:

- High-quality genomic DNA of your species of interest, adjusted to 100 ng/μL (DNA sample can be from single or pooled individuals)

- One set of adapters (both common and barcoded) at working concentration (see Basic Protocol 1, step 1. A default working concentration is 0.3 ng/µL per adapter, or 0.6 ng/µL for the pair)
- Restriction enzyme(s) of choice
- Enzyme-appropriate restriction buffer
- 10X T4 DNA ligase buffer with ATP
- T4 DNA ligase
- Molecular biology-grade water
- 2x Taq DNA master mix
- PCR Primers 1 and 2 at 50 µM concentration.

Materials

- 0.2 ml PCR tubes x 24
- Column-based PCR purification kit (e.g.,Qiagen QIAquick PCR Purification Kit or equivalent)

Equipment

- Centrifuge
- Thermocycler
- High-sensitivity electrophoresis platform (e.g., BioRad Experion or Sage Science BluePippin)

Protocol

1. Create a master mix for digestion of genomic DNA, consisting of:
    - 18 µL genomic DNA (= 200 ng per reaction for frequent cutters. Increase to 500 ng for less frequent cutters or double digests)
    - 18 µL 10X RE buffer (adjust if your buffer stock is at a different concentration)
    - 9 µL (18U) restriction enzyme(s) (Adjust amount of RE and water if RE concentration > 2U/ µL)
    - 135 µL water
2. Aliquot the master mix into 8 PCR tubes, spin down to collect samples, and incubate for 2 hours at the temperature appropriate for the chosen enzyme(s).
3. For each of the eight tubes, mix the following reactions (all values are in microliters):

Digested DNA	20	20	20	20	20	20	20	20
10X T4 ligase buffer with ATP	5	5	5	5	5	5	5	5
Pre-mixed adapters	6	8	12	14	16	18	20	24
Water	18	16	12	10	8	6	4	0
T4 DNA ligase	1	1	1	1	1	1	1	1
Total:	*50*	*50*	*50*	*50*	*50*	*50*	*50*	*50*

Note: Add the ligase to all reactions at the same time, such as by using a multichannel pipette or putting the enzyme on the side of the tube and then spinning them all down together.

4. Spin down to collect the liquid and incubate at 22° C for 1 hour.
5. Inactivate the ligase by incubating at 65° C for 30 minutes (or according to the manufacturer’s instructions)
6. Clean up reactions with a column-based purification kit according to manufacturer’s instructions.

7. For each of the above samples, set up a PCR reaction:
    - 10 µL eluted DNA from step 6
    - 1 µL PCR primer 1 (50 µM stock)
    - 1 µL PCR primer 2 (50 µM stock)
    - 25 µL 2x Taq polymerase master mix
    - 13 µL water
8. Amplify with the following protocol:
    - 72° C for 5 minutes
    - 98° C for 30 seconds
    - 18 cycles of:
        - 98° C for 10 seconds
        - 65° C for 30 seconds
        - 72° C for 30 seconds
    - 72° C for 5 minutes
    - Hold at 4° C
9. Clean up reactions with a column-based purification kit according to manufacturer's instructions.
10. Run out each sample on a high-sensitivity size identification and quantification platform, such as a BioRad Experion, BioRad Bioanalyzer, or Sage Science BluePippin. Adapter dimers show a peak around 128 bp (Figure 3A). Identify the adapter titration that produces a good library peak but shows <u>no</u> adapter dimers (Figure 3B). Use one-half of this amount when aliquotting adapters (Basic Protocol 1, step 1), since the GBS protocol uses only 100 ng of genomic DNA per reaction (Basic Protocol 1, step 3) instead of the 200 per reaction used here.

COMMENTARY

Background information

GBS is a variation of restriction site-associated DNA sequencing (RAD-seq; Miller et al., 2007), where DNA samples are digested with restriction enzyme(s) and locations adjacent to the restriction cut sites are sequenced. The original GBS protocol of Elshire et al. (2011) simplified existing methods by removing several steps (random shearing of restriction fragments, physical size selection, and ligation of Y-adapters), thus providing a fast, inexpensive, high-throughput method for library construction. Since then, many different modifications to this protocol have been developed (reviewed in Andrews et al., 2016). Even the name "GBS" has been co-opted to mean almost any protocol that uses high-throughput sequencing to determine genotypes.

Because of the popularity of GBS, some researchers are tempted to use it indiscriminately. One should always remember that like all protocols, GBS is a tool, and it can be used well or poorly depending on how it is suited for a particular job. Before performing a GBS experiment, one should carefully consider each step to ensure that the chosen methodology will meet the needs of the final experiment.

Critical parameters:

<u>DNA quality</u>

The restriction enzyme digestion and ligation steps require higher quality DNA than for some other procedures like PCR. Because of this, DNA purity is usually the most important factor for determining whether a GBS preparation will succeed. Since GBS enriches for small restriction fragments, some DNA degradation is tolerated as long as the majority of sample fragments are larger than a few Kbp.

It is almost always worth your time (and money) to confirm that your DNA preparation step will result in DNA that can be restriction digested well, and to change your protocol if it does not. Be sure to assess digestibility using a RE that is not methylation sensitive (i.e., *Hin*dIII or *Eco*RI). Some species contain metabolites that inhibit digestion and ligation; this can usually be solved by switching from a phase-separation protocol to a column-based kit to isolate sample DNA. (See "Troubleshooting.")

Adapter concentration

Using the right amount of adapter ensures the optimal creation of library fragments instead of adapter dimers. Supplemental Table 3 contains a list of >300 species we have created GBS libraries for, along with notes on which enzyme combinations we have empirically tested and if the adapter amounts (Basic protocol 1, step 1) need to be altered. If your species does not appear on this list or if you are using an untested enzyme, we recommend using Support Protocol 2 to empirically determine how much adapter should be used in the library prep.

Size Fractionation

Because fragments larger than ~ 500 bp do not sequence efficiently on Illumina platforms, it is very important that sequencing libraries are comprised of small DNAs. Unlike RADseq, the GBS protocol does not contain a step where DNAs are physically size-separated. Instead, fragment sizes are limited by using PCR conditions that encourage the amplification of small fragments. So far, we have not encountered a species that has failed to produce a GBS library preparation with a majority of fragments in the appropriate size range (i.e., 160-500 bp).

Experion traces of some libraries may show a normal size distribution followed by a "hump" or "tail" of oversized fragments (Supplemental Figure 1). In these cases, we suggest repeating the PCR using a smaller number of PCR cycles (i.e., use 15-17 cycles instead of 18 cycles). The oversized "tail" should disappear once a better cycle number is determined. Supplemental Table 3 indicates the species that have required reduced cycle numbers.

Sequencing depth

The ultimate aim of GBS is to get genotype data. The quality of individual genotypes is determined primarily by the sequencing depth at each site. Because the molecules that are captured on the surface of each flowcell are a random selection from all possible sequences in the library, there is variation from site to site and individual to individual even with the best library preparation. A GBS experiment thus always involves tradeoffs between the number of sites investigated and the depth at each site. There are two major ways to adjust these parameters: choice of restriction enzyme(s) and choice of multiplexing level.

The restriction enzyme(s) chosen determine the number of potential sites available for GBS. Given the same amount of sequencing, a rare-cutting enzyme will have much greater depth at any given site than a common-cutting one, resulting in fewer, higher-quality genotypes (Figure 2).

Multiplexing level determines how the available sequencing is split among the samples. Putting 96 samples into a single flowcell lane will result in four times as many reads per site (on average) than if the same samples were run in 384-plex. Resequencing libraries is a variant on this, where sequencing a 384-plex library on four lanes is equivalent to having prepared four 96-plex libraries. Because of this, many researchers choose to sequence at a high multiplexing level—192- or 384-plex—and resequence the library later if they find they need greater depth.

The appropriate level of sequencing depth depends on the experiment, though there are some general guidelines. Highly inbred individuals require less sequencing depth than outcrossed ones because there is no need to accurately call heterozygous sites. Polyploid individuals usually benefit from greater depth than diploid ones to separate true SNPs from ones due to alignment of homeologous sequences. Some researchers have used GBS to look at allele frequencies in populations by bulking many individuals; in this case one should aim for very high sequencing depth to accurately call the frequencies, since at low depth the noise from stochastic sampling can overwhelm the true signal.

Troubleshooting

Both the digestion and ligation steps can be affected by contaminants in the DNA. If you are using a phase-separation method to purify DNA (e.g., CTAB or TRIzol), try using a column-based purification kit instead. If you are extracting DNA from a medicinal plant, you may need to add 2.8% polyvinylpyrrolidone-40 (PVP-40) to the extraction buffer before column purification. If you still cannot get a usable GBS library, check the expiration date of each enzyme. If the enzyme is not expired, test activity by performing test digestions or ligations of known, purified DNA standards (plasmids, lambda phage DNA, or lambda *HindIII* fragments) that can be easily obtained. In the lab, always keep restriction enzymes on ice and do not hold the bottom of the stock tube between your thumb and forefinger. Because ATP is easily degraded by repeated freeze-thaw cycles, the ligase buffer should always be tested.

If enzyme activity is not an issue, it is possible that adapters may not be well annealed. Check to make sure appropriate forward and reverse oligos were mixed (i.e., forward sequences should be combined with reverse sequences only). Oligonucleotides do not degrade as long they are frozen at high concentrations (200 uM). Although this should not be necessary, adapter stocks can be denatured in TE buffer supplemented with 50mM NaCl and reannealed prior to ligation.

Anticipated results

The result of this protocol should be a library of DNA fragments ready to be sequenced. Processing raw sequencing reads down to genotypes is beyond the scope of this protocol, especially since it involves many considerations that are experiment-specific. For general recommendations, we refer readers to a recent review by Torkamaneh et al. (2016) that compares several GBS pipelines both with and without a reference genome.

Time considerations

The entire GBS protocol can be carried out in a single day (8 hours or less). The exact time depends on how many plates one is prepping at once, and whether they are prepared by hand or using a liquid handling robot. If need be, the samples can be frozen after the ligation step (Step 8), or ligations can be done overnight at room temperature, then the protocol completed the following day. The protocol should *not* be stopped after restriction digestion because the sticky ends are vulnerable to degradation.

Bioinformatic considerations

Although a full treatment of the bioinformatics of GBS is outside the scope of this protocol, some common characteristics of GBS-derived data bear mentioning. Readers interested in a comparison of pipelines for processing GBS data are referred to Torkamaneh et al. (2016).

Missing data: The fragments output by the DNA sequencer are a random sampling of all the fragments available in the pool. This means that different sites are sequenced to different depths, and due to random sampling some sites in some individuals are never sequenced at all. GBS thus tends to have a high rate of missing data, especially in comparison to microarrays and single-sequence repeats (SSRs). The two main ways of dealing with this are to (1) increase the sequencing depth per site (and thus decrease the probability of getting zero reads from an individual at that site) and/or (2) filter out sites that have low coverage across samples. Much work has also been done on imputation, where known genotypes are used to determine the most likely value for an unknown genotype. However, imputation methods often rely on various assumptions that should be checked to ensure they will work for your material.

Sequencing errors: DNA sequencing errors look like SNPs, and this is an issue that all sequencing-based genotyping platforms face. Greater sequencing depth at each locus can increase the confidence of any individual call. Sequencing errors also tend to be rare, so simply filtering out rare SNPs (those with allele frequencies of 1-2% or less) will tend to filter out the majority of sequencing errors as well. This is most useful for materials that have known allele frequencies (such as biparental populations) but tends to be less useful for diverse materials that have many genuine SNPs in this same frequency range. Since the number of individuals with these true SNPs is by definition small, many researchers find it acceptable to remove them rather than let the data be massively contaminated with sequencing errors.

Paralogous sequences: "Paralogs" is a catch-all term for any duplicated gene sequence. They can be due to ancient or modern gene duplication, polyploidization, transposon activity, genome assembly errors, or other sources. Because these sequences tend to be highly similar, DNA sequence aligners often collapse them on top of each other, creating SNP calls that are due to differences between paralogs instead of true SNPs. The easiest way to identify these sites is to look at the sequencing depth and heterozygosity rate across sites. Because paralogous SNPs result from reads being combined across multiple loci, they tend to have higher sequencing depth than true SNPs. Similarly, they tend to have much higher rates of heterozygosity since the different genetic loci are usually fixed for small sequence differences.

ACKNOWLEDGEMENTS

This work was supported by the University of Georgia and Cornell University. We would like to thank Jeff Glaubitz, Asha M. Jain, and Mohsen Moghadam for helpful commentary and editing of the manuscript.

LITERATURE CITED

References

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., and Hohenlohe, P.A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet* 17:81-92.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379.

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., and Johnson, E.A. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res* 17:240-248.

Poland, J.A., Brown, P.J., Sorrells, M.E., and Jannink, J.L. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7:e32253.

Rogers, S.O. and Bendick, A.J. 1994. Extraction of total cellular DNA from plants, algae and fungi. *In* Plant Molecular Biology Manual (S.B. Gelvin and R.A. Schilperoort, eds.) pp. 183-190. Springer Netherlands.

Torkamaneh, D., Laroche, J., and Belzile, F. 2016. Genome-Wide SNP Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies. *PLoS One* 11:e0161333.

KEY REFERENCE(S)

- *Original GBS protocol:* Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6**:** e19379.
- *Two-enzyme GBS:* Poland, J. A., P. J. Brown, M. E. Sorrells and J. L. Jannink, 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7**:** e32253.
- *Review of various restriction site-associated DNA sequencing methods (including GBS), with pros and cons of each:* Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart and P. A. Hohenlohe, 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17**:** 81-92.
- *Comparison of GBS bioinformatics pipelines with and without a reference genome:* Torkamaneh, D., J. Laroche, and F. Belzile, 2016. Genome-wide SNP Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies.

INTERNET RESOURCES

- Software for GBS analysis
  - Fast-GBS: https://bitbucket.org/jerlar73/fastgbs
  - STACKS: http://catchenlab.life.illinois.edu/stacks/
  - TASSEL: http://www.maizegenetics.net/tassel

FIGURE LEGENDS

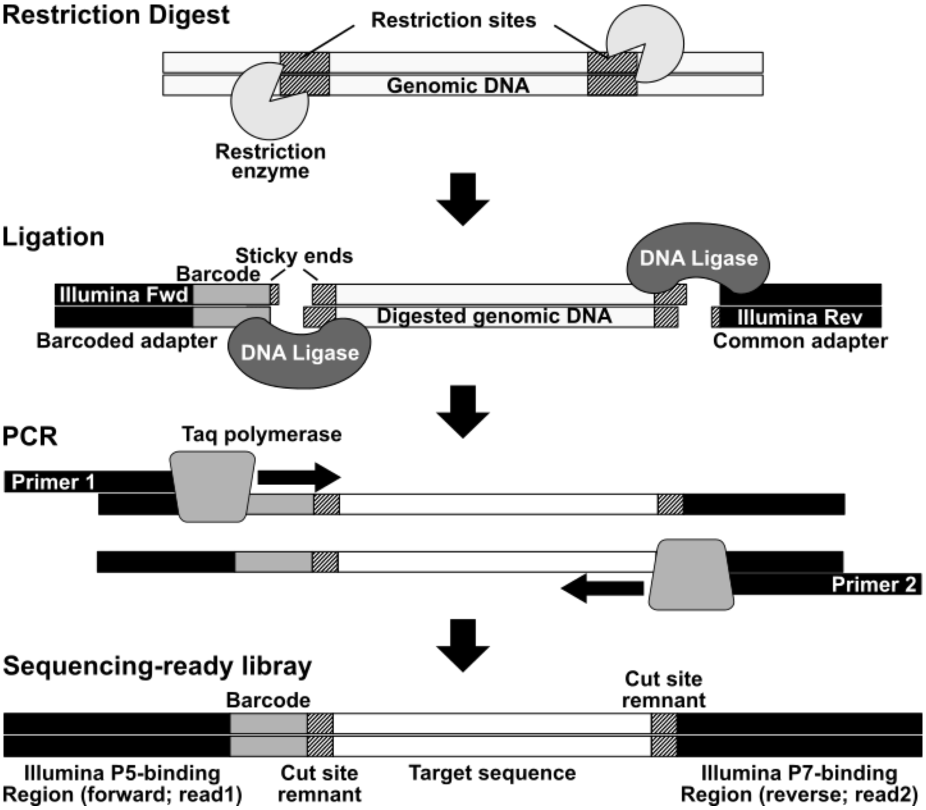

Figure 1 – GBS workflow. The basic workflow of GBS library preparation is diagrammed. Genomic DNA (top) is first digested with restriction enzymes. The digested product is then ligated to a barcoded adapter (left) and a common adapter (right). The ligated product is then amplified by PCR, which also adds the final Illumina sequencing regions.

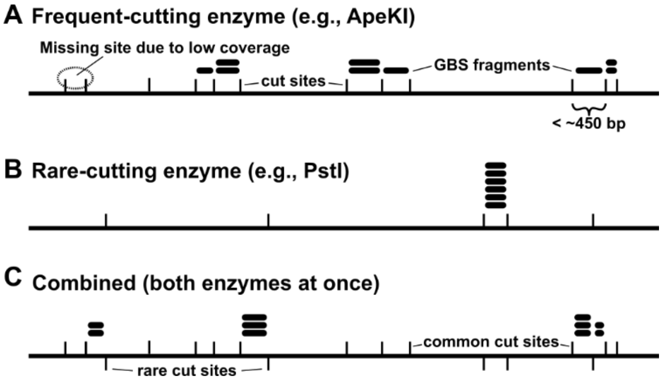
A Frequent-cutting enzyme (e.g., ApeKI)
Missing site due to low coverage
cut sites
GBS fragments
< ~450 bp
B Rare-cutting enzyme (e.g., PstI)
C Combined (both enzymes at once)
common cut sites
rare cut sites

Figure 2 – Effect of enzyme choice on the GBS library. The schematic represents a segment of genomic DNA and the GBS library that results from using different restriction enzymes. Small vertical lines represent enzyme cut sites, and the thick horizontal segments represent the resulting GBS fragments that are sequenced. Although there is no explicit size-fractionation step in the GBS protocol, the PCR amplification step (Basic Protocol 1, step 12) implicitly selects for fragments shorter than ~450 bp. (A) Digestion with a frequent-cutting enzyme results in many sites being included in the GBS library, but at a low coverage per site. Some sites may be missing due to undersampling. (B) Digestion with a rare-cutting enzyme results in far fewer sites but greater sequencing depth at each one. (C) Two-enzyme GBS (Alternative Protocol 1) targets locations where the two different enzymes cut in close proximity. Most researchers pair a common-cutting enzyme (A) with a rare-cutting one (B), resulting in an intermediate number of sites with intermediate coverage levels (C).

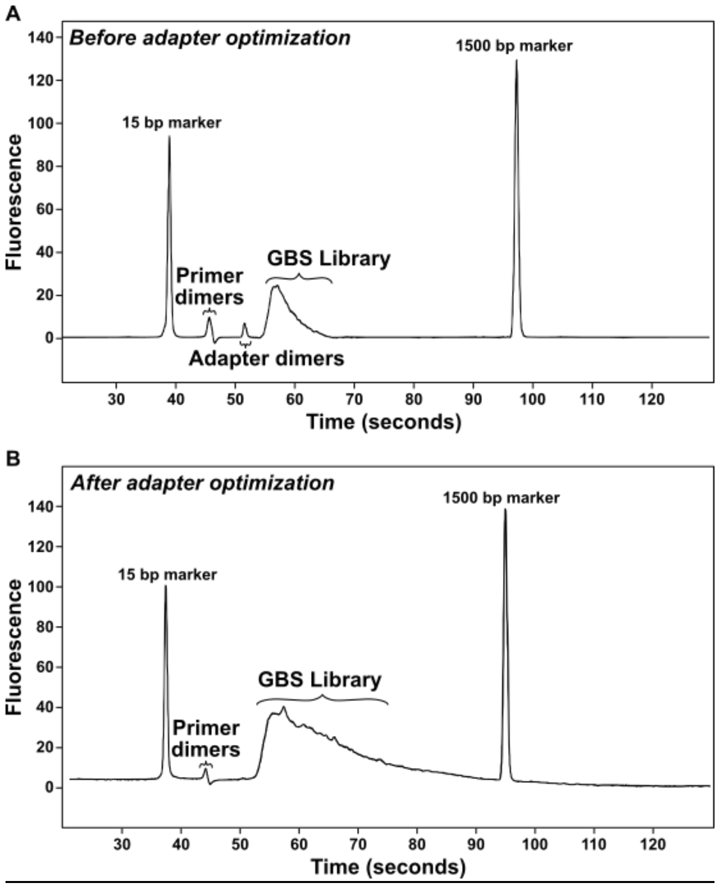

Figure 3 – Adapter optimization. (A) Without optimizing adapter concentrations for a species, adapter dimers will contaminate the GBS library prep (visible at ~128 bp). (B) Users should aim to use a concentration that produces a good library peak but contains no adapter dimers. Images based on BioRad Experion output included in Elshire et al. (2011).

TABLES

Table 1 – GBS Adapter dilutions compatible with most species we have tested

Enzyme	ApeKI	PstI or EcoT22I
Amount of adapter pairs for 100 ng of sample*	3.6 ng	1.44 ng
Concentrated stock concentration (Support Protocol 1, step 5)	3 ng/µL	3 ng/µL
Working stock concentration (Basic Protocol 1, step 1)	0.6 ng/µL	0.24 ng/µL
Dilution to create working stock from concentrated stock		
Concentrated stock	10 µL	4 µL
Water	40 µL	46 µL
Final volume of working stock (Basic Protocol 1, step 1)	50 µL	50 µL

* *See Supplemental Table 3 for a list of all species/enzyme combinations we have tested and their corresponding adapter concentrations*