Check for updates

Published September 26, 2014

NOTE

# The Effects of Restriction-Enzyme Choice on Properties of Genotyping-by-Sequencing Libraries: A Study in Cassava (*Manihot esculenta*)

Martha T. Hamblin* and Ismail Y. Rabbi


## ABSTRACT

Compared with other reduced-representation sequencing methods, library construction for genotyping-by-sequencing (GBS) is simpler and less expensive. However, elimination of size-selection steps results in libraries of more variable fragment size than with other reduced-representation methods, affecting several aspects of the data. To test the effect of restriction enzyme choice on library quality, we made GBS libraries with *Pst*I (6-cutter), *Pst*I/*Taq*I (4-cutter), or *Ape*KI (4.5-cutter) from the same set of DNAs from a cassava (*Manihot esculenta* Crantz) biparental population. Tag and single nucleotide polymorphism (SNP) counts were limited by the number of cut sites rather than by read number. Depth per locus was very skewed for the *Pst*I library, such that most SNPs had low read depth but a subset had very high read depth. In contrast, the *Ape*KI and *Pst*I/*Taq*I libraries had less variable distributions of read depth and yielded far more scorable SNPs. Our results suggest that 6-cutter enzymes may be most appropriate for genotyping a modest number of markers at a high multiplexing level, or for very large genomes, and perform better when used in a double digest with a 4-cutter enzyme.


M.T. Hamblin, Dep. of Plant Breeding and Genetics, Cornell Univ., Ithaca, NY; and I.Y. Rabbi, International Institute for Tropical Agriculture (IITA), Ibadan, Nigeria. This work was supported by the projects "Genomic Selection, the Next Frontier for Rapid Gains in Maize and Wheat Improvement" through funds from The Bill and Melinda Gates Foundation, and "Next Generation Cassava Breeding" through funds from The Bill and Melinda Gates Foundation and the Department for International Development of the United Kingdom. Received 27 Feb. 2014. *Corresponding author (mth3@cornell.edu).

**Abbreviations:** GBS, genotyping-by-sequencing; PCR, polymerase chain reaction; RRL, reduced representation library, SNP, single nucleotide polymorphism.

In the past, simple sequence repeats (reviewed in Varshney et al., 2005), amplified fragment length polymorphisms (Vos et al., 1995), and various kinds of SNP assays (reviewed in Gupta et al., 2008) have been used as markers for genetic mapping or association studies. Increasingly, next-generation sequencing is replacing these technologies because it allows for simultaneous discovery and scoring of large numbers of markers. However, the cost of complete genome sequencing is still high, and most applications do not require full sequence information. Next-generation sequencing of reduced-representation libraries (RRLs), which capture a specific and reproducible subset of the genome, provides the advantages of sequence data at a lower cost and without the computational burden that comes with enormous data sets.

There are numerous methods for making RRLs (reviewed in Davey et al. (2011) and Hirsch et al., 2014); many of these

Published in Crop Sci. 54:2603–2608 (2014).
doi: 10.2135/cropsci2014.02.0160
Freely available online through the author-supported open-access option.
© Crop Science Society of America | 5585 Guilford Rd., Madison, WI 53711 USA

All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.

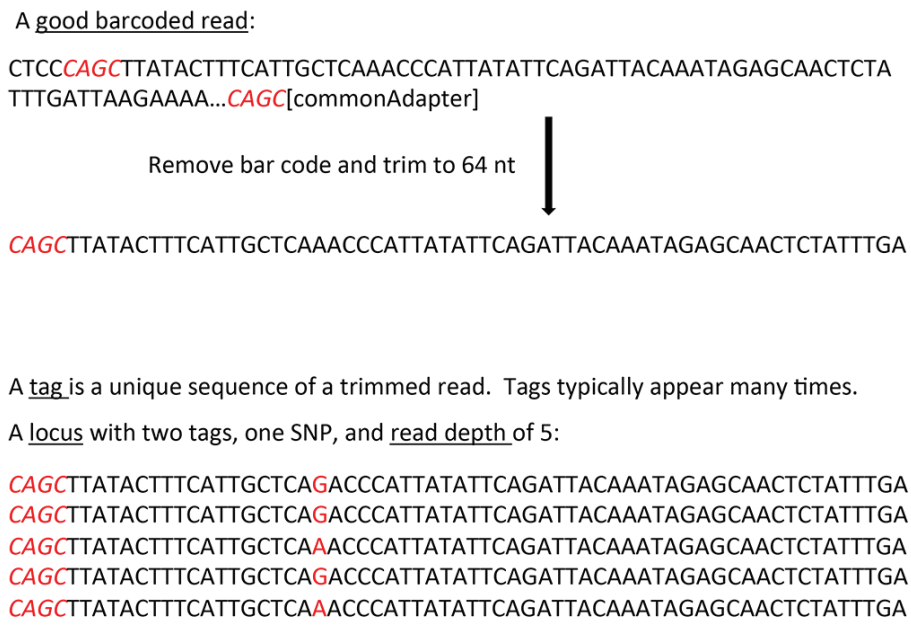


Figure 1. Illustration of terms used in bioinformatic processing of genotyping-by-sequencing reads. SNP, single nucleotide polymorphism.

rely on restriction digestion. The complexity of the methods varies, as does the cost. Genotyping-by-sequencing (Elshire et al., 2011) is an RRL method that is relatively simple and inexpensive, making it feasible to genotype large numbers of individuals. GBS has therefore become very popular, particularly for researchers working on non-model species with few genomic resources (http://www.biotech.cornell.edu/node/886; Glaubitz et al., 2014).

One of the bases of GBS's simplicity is the elimination of steps designed to generate fragments of a narrow size range; such steps are a feature of most other RRL methods (e.g., RAD sequencing; Baird et al., 2008). In GBS, the pooled, digested DNA is polymerase chain reaction (PCR)-amplified under conditions that favor amplification of smaller fragments, resulting in a sequencing library whose size distribution is less well defined than with other methods. Choice of enzyme is a critical step in GBS library development; considerations include genome size, level of inbreeding (i.e., heterozygosity), anticipated level of multiplexing, and the number of markers required for the project. The object is to find the right balance in the trade-off between genome coverage and read depth (e.g., Beissinger et al., 2013).

In making GBS libraries for an outcrossing species such as cassava, sufficient read depth is needed to accurately call heterozygous genotypes. For this reason, for our first cassava GBS libraries we chose an enzyme with a 6-base recognition sequence ("6-cutter"), *Pst*I, even though cassava has a modest-sized genome (760 Mb). Because we could score only a modest number of SNPs from those libraries (Ly et al., 2013; Rabbi et al., 2014), we subsequently tried a number of other enzymes and enzyme combinations. In the course of these experiments, we sequenced GBS libraries made with *Pst*I, *Pst*I/*Taq*I (4-cutter), or *Ape*KI (4.5-cutter) from the same set of DNAs, extracted from progeny of an outbred cassava biparental population. Important properties of the libraries varied with the frequency of cutting of the restriction enzyme(s). These results suggest that, depending on the application and genome size, predictable properties of restriction enzymes can guide the choice of enzyme in making GBS libraries.

## METHODS

Cassava DNA was prepared as described in Rabbi et al. (2014); we used two sets of 95 DNAs from a biparental mapping population with outcrossed parents. GBS libraries were constructed at the Institute for Genomic Diversity at Cornell University, according to the method of Elshire et al. (2011) for *Pst*I and *Ape*KI or Poland et al. (2012) for *Pst*I/*Taq*I. The *Pst*I/*Taq*I libraries were constructed such that only fragments with a *Pst*I site on one end and a *Taq*I site on the other were amplified and sequenced. Each library was loaded onto one lane of a flowcell of an Illumina HiSeq at the Biotechnology Resource Center at Cornell University. We processed the FASTQ files (output of the Illumina pipeline) using the Tassel pipeline 3.0 (http://tassel.bitbucket.org/TasselArchived.html) as described in Rabbi et al. (2014). Tags with a count less than 50 were eliminated because they were likely to be sequencing errors. When running the TagsToSNPByAlignmentPlugin, minimum locus coverage was set to 0.4 and minimum minor allele frequency was set to 0.1. These setting are appropriate for a biparental population in which alleles should segregate at 1:1 (parents are both heterozygous) or 3:1 (one parent is homozygous, one is heterozygous) ratios.

## RESULTS

Important properties of GBS libraries are the total number of raw reads, the number of good barcoded reads, the number of unique tags, and the read depth distribution. These terms are defined below and illustrated in Fig. 1:



14350653, 2014, 6, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.2135/cropsci2014.02.0160 by CAPES, Wiley Online Library on [17/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

**Table 1. Raw and good barcoded read counts in genotyping-by-sequencing libraries.**

| Enzyme | Plate | Flowcell | Raw reads | Good barcoded | Percent good |
|---|---|---|---|---|---|
| *Ape*KI | 1 | 1 | 150,344,791 | 130,463,262 | 87 |
| *Ape*KI | 2 | 3 | 141,989,084 | 135,125,857 | 95 |
| *Pst*I | 1 | 1 | 119,077,915 | 90,209,357 | 76 |
| *Pst*I | 2 | 2 | 112,397,029 | 90,245,256 | 80 |
| *Pst*I/*Taq*I | 1 | 3 | 184,075,809 | 167,339,637 | 91 |

**Table 2. Read and tag counts in genotyping-by-sequencing libraries made with *Pst*I or *Ape*KI.**

| | Good barcoded reads | | Tags | | Plate-specific tags | |
|---|---|---|---|---|---|---|
| Enzyme | *Ape*KI | *Pst*I | *Ape*KI | *Pst*I | *Ape*KI | *Pst*I |
| Plate 1 | 130,463,262 | 90,209,357 | 390,338 | 70,224 | 626 | 8582 |
| Plate 2 | 135,125,857 | 90,245,256 | 390,671 | 78,120 | 959 | 16,478 |
| Combined | 265,589,119 | 180,454,613 | 391,297 | 86,702 | 1585 | 25,060 |

**Table 3. Properties of genotyping-by-sequencing libraries made with three different enzymes.**

| Enzyme | Good Barcoded Reads | Tags | Median read depth per locus | Scorable SNPs† | Comments |
|---|---|---|---|---|---|
| *Pst*I‡ | 90,227,306 | 74,172 | 296 | 8584 | Fewer cut sites; larger size range |
| *Pst*I/*Taq*I | 167,339,637 | 95,289 | 1761 | 14,527 | Fewer cut sites; smaller size range |
| *Ape*KI‡ | 132,794,560 | 390,504 | 383 | 50,347 | More cut sites; smaller size range |

† SNP, single nucleotide polymorphism.

‡ Entries are the mean of two plates.

A ***read*** is a single sequence in the FASTQ output file generated from the GBS library.

A ***good, barcoded read*** is a sequence read with a perfect match to one of the barcodes and with no N's in the sequence, up to a specified length.

A ***tag*** is a unique sequence (excluding the barcode) up to a specified length from one or more "good, barcoded reads." A given tag is typically observed multiple times, unless it has been generated by a sequencing error.

A ***locus*** is a single position on the reference genome where a tag aligns. More than one tag will align at a locus if the locus is polymorphic.

***Read depth*** is the total number of reads that map to a locus, either within or across barcoded individuals.

***Read depth distribution*** is the profile of read depths across all loci.

For the five GBS libraries in our study, the number of raw reads and number of good barcoded reads varied consistently between restriction enzymes (Table 1) and was independent of the flowcell (i.e., the sequencing run). The number of tags also varied consistently between enzymes, and was not primarily limited by read number (Table 2): the two *Pst*I libraries, pooled together, had >180,000 raw reads, one third more than a single *Ape*KI library, but had only 22% of the number of tags in a single *Ape*KI library. The *Pst*I/*Taq*I library had the most reads of all, yet had far fewer unique tags than the *Ape*KI libraries. While each *Ape*KI library identified nearly all the tags, a much larger fraction of the *Pst*I tags was unique to each library.

A library made with a double digest of *Pst*I and *Taq*I had more reads than either the *Pst*I or *Ape*KI library, but only 28% more tags than the *Pst*I library (Table 3). However, the median read depth per tag for *Pst*I/*Taq*I was almost sixfold larger than for the *Pst*I library, so that 68% more SNPs could be scored. The *Ape*KI libraries had almost six times as many scorable SNPs as the *Pst*I libraries, and over three times as many as the *Pst*I/*Taq*I library (Table 3); these ratios are similar to the ratios in tag numbers, but are also affected by read depth.

The read depth distributions for these libraries are revealing: most *Pst*I tags had a low read depth, but there was a long and substantial tail of tags that had more than 5000 reads per locus (Fig. 2). In contrast, the read depth distribution for the *Ape*KI library was much less skewed (Fig. 3A; note the difference in scale of the *y* axes). If only tags with read depth over 1000 are plotted for the *Ape*KI library (Fig. 3B), the distribution looks similar to the *Pst*I library, but this plot represents a small fraction of the total reads. The read distribution for the *Pst*I/*Taq*I library (Fig. 4) was intermediate to the other two libraries.

## DISCUSSION

We sequenced GBS libraries made from the same DNAs, but prepared with different restriction enzymes, and



14350653, 2014, 6, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.2135/cropsci2014.02.0160 by CAPES, Wiley Online Library on [17/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

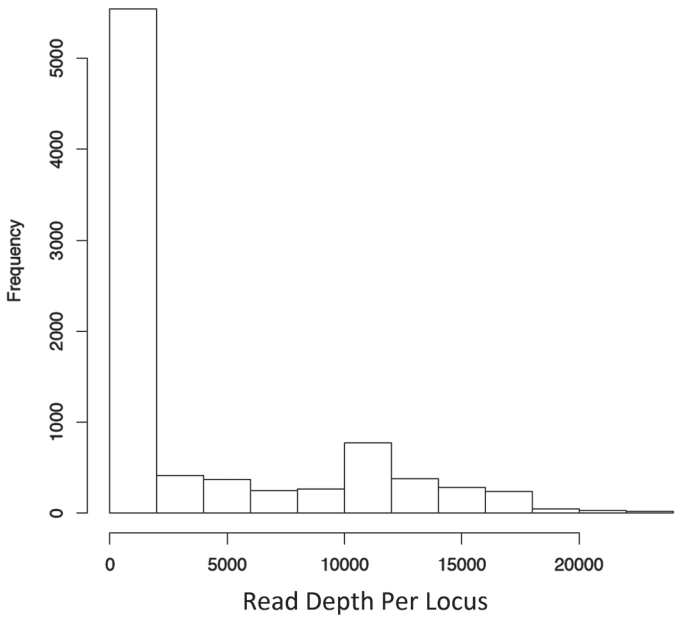


Figure 2. Read depth distribution for the genotyping-by-sequencing library made with *Pst*I. Histogram of the total number of reads for all tags at a single nucleotide polymorphism (SNP), for all scored SNPs.

analyzed the differences in library attributes. The differences that we observed can be explained by differences in the number of cut sites and the range of fragment sizes that are included in the library after PCR amplification (see "Comments" in Table 3):

1. **Number of good barcoded reads.** It is not immediately obvious why the number of reads should be affected by enzyme; read number is a function of cluster number (i.e., the number of DNA templates immobilized on the surface of the Illumina flowcell). Two factors are likely to explain this phenomenon. First, the Illumina support website (http://support.illumina.com/sequencing/sequencing_instruments/cluster_station/questions.ilmn) advises: "Short fragments tend to create smaller clusters allowing greater data density. The optimal fragment size for single-read run is 150–300 bp." The *Pst*I/*Taq*I library had fragment sizes that were closer to the optimum for increasing cluster density and thus read counts, and the *Pst*I libraries had fragment sizes farthest from the optimum. Second, calculation of the optimal dilution of the GBS library is difficult when the fragment size distribution has not been well characterized. A formula that performs well for *Ape*KI libraries will not perform well for *Pst*I libraries because the relationship between picograms of DNA and picomoles of fragments is not the same. Empirical determination of the correct dilution would lead to higher read counts for *Pst*I libraries, although the greater distribution of fragment sizes makes this challenging.

2. **Number of tags.** The higher number of tags in the *Ape*KI libraries was clearly due to the larger number of *Ape*KI cut sites in the genome for an enzyme with a 4.5-base recognition site as compared to one with a 6-base recognition site. In the *Pst*I libraries, the number of tags was limited both by the number of cut sites and by the number of fragments that could be amplified efficiently in the library. Additional digestion with *Taq*I increased the tag number, but not by a lot, because the number of *Pst*I cut sites was still limiting. However, the read depth per locus increased dramatically, because the larger *Pst*I fragments were digested (by *Taq*I) to a size that was efficiently incorporated into the GBS library.

3. **Plate-specific tags (Table 2).** The *Ape*KI libraries had a large number of tags, of which 99.6% were observed in both libraries. This suggests that the vast majority of *Ape*KI fragments were captured and sequenced in each library. In contrast, 29% of tags in the *Pst*I libraries were plate-specific. This difference likely arose from the long tail of larger fragments that were poorly represented in the *Pst*I library; by chance, a different subset of those fragments was amplified and sequenced in each library. SNPs occurring on these plate-specific tags, however, had poor read depth and high levels of missing data.

4. **Read depth per locus.** The median read depths per locus reported in Table 3 correspond to 3.1, 18.5, and 4 reads per genotype, on average, for the *Pst*I, *Pst*I/*Taq*I, and *Ape*KI libraries, respectively. The low read depth for the *Pst*I libraries arose from the phenomenon described in item 3: many fragments were too large to be efficiently amplified and thus were poorly represented in the library. The full picture is shown in the histogram (Fig. 2), with the long tail of overrepresented, shorter fragments in the *Pst*I library. While the *Ape*KI libraries had only 29% greater median read depth than the *Pst*I libraries, the greater uniformity of read depth across loci provided much higher quality genotype information. The *Pst*I/*Taq*I library had very good read depth because the fragment sizes were in the optimal range, but the number of loci, and thus SNPs, was limited by the number of *Pst*I cut sites.

## CONCLUSIONS

DNA digestion with only a 6-cutter enzyme such as *Pst*I primarily generates fragments that are larger than the desired size for a GBS library (the expected size is 4096 bp). Smaller fragments are preferentially amplified and overrepresented in the library, while the larger fragments are poorly amplified. Furthermore, the unknown and



14350653, 2014, 6, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.2135/cropsci2014.02.0160 by CAPES, Wiley Online Library on [17/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

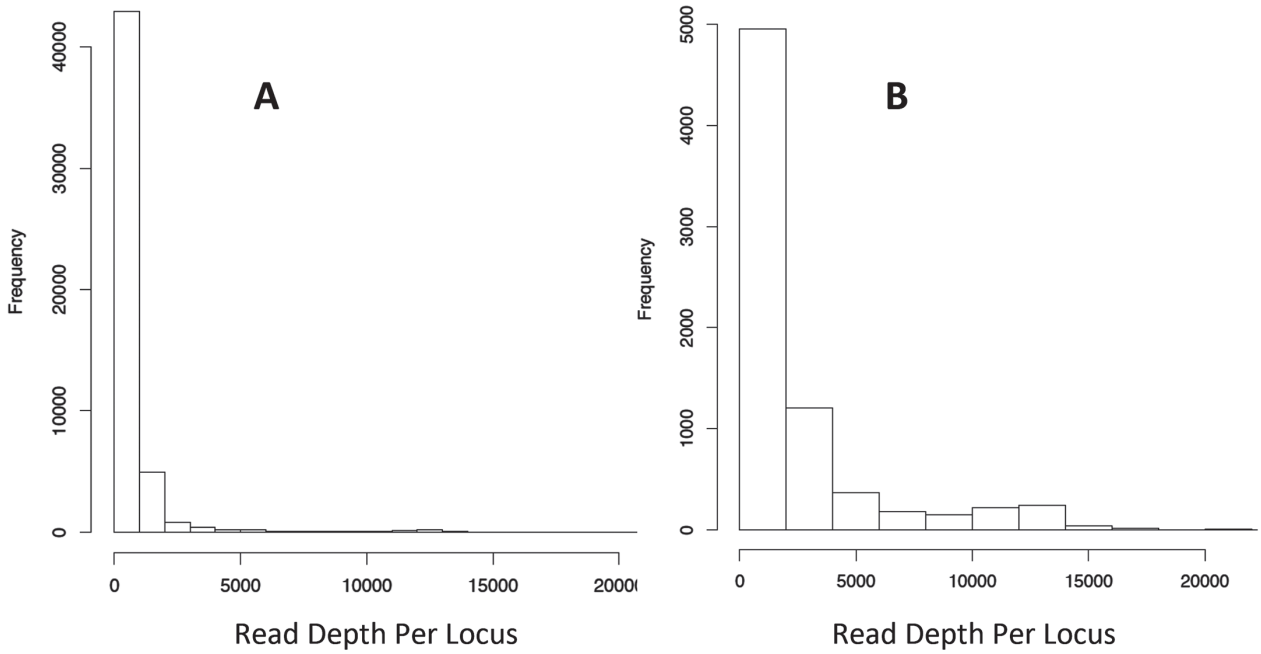


Figure 3. Read depth distribution for the genotyping-by-sequencing library made with *Ape*KI. Histogram of the total number of reads for all tags at scored single nucleotide polymorphisms (SNPs). (A) All SNPs. (B) Only SNPs with >1000 reads are shown

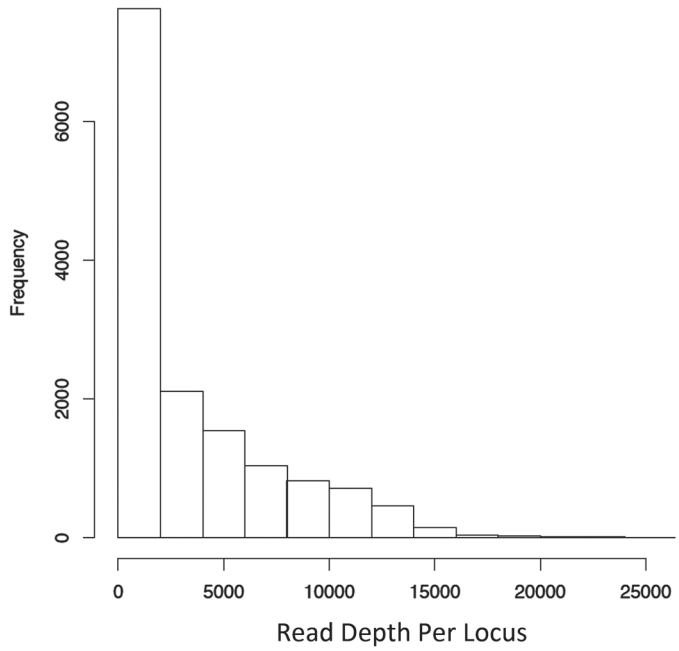


Figure 4. Read depth distribution for the genotyping-by-sequencing library made with *Pst*I/*Taq*I. Histogram of the total number of reads for all tags at a single nucleotide polymorphism (SNP), for all scored SNPs.

skewed distribution of fragment sizes causes challenges to accurate library dilution, which can result in suboptimal cluster densities and fewer reads.

At 96-plex, SNPs on the larger fragments have poor read depth and large amounts of missing data, while a small number of SNPs have far more reads than are needed to score genotypes. Thus, many reads are wasted. Furthermore, a fragment that is sequenced many times generates a large number of unique tags that arise from sequencing errors. When read depth is very high, the standard cut-off is not effective in filtering out sequencing errors. For all of these reasons, a single digestion with a 6-cutter does not generally seem to be a good choice for many applications.

In spite of the limitations of 6-cutter enzymes, there are applications for which they may be a good choice. Such applications, such as mapping in biparental populations and diversity or population structure analyses, call for a relatively small number of markers scored on a large number of individuals. The SNPs with very high read depth (on which reads were "wasted" in the 96-plex *Pst*I library) could still be scored at a high level of multiplexing.

Very large genomes, such as wheat (*Triticum aestivum* L.) and barley (*Hordeum vulgare* L.) (Poland et al., 2012) or mammals (De Donato et al., 2013), require use of an enzyme that cuts infrequently; otherwise, read depth will be unacceptable. For high levels of multiplexing, the best results are obtained when the 6-cutter is combined with a 4-cutter (e.g., Poland et al., 2012) to limit the fragment number and also reduce fragment size. De Donato et al. (2013) successfully used a *Pst*I single digest in a 48-plex library but would likely have had better results had they used a double digest.

For cassava, we have chosen to use *Ape*KI. It may seem surprising that one enzyme performs well for both maize inbred lines and outbred cassava; this is possible because the cassava genome is much smaller (760 Mb) than the maize genome (2.4 Gb). The complexity of a cassava *Ape*KI library is lower, so read depth is higher, allowing us to score heterozygous genotypes with sufficient confidence.



14350653, 2014, 6, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.2135/cropsci2014.02.0160 by CAPES, Wiley Online Library on [17/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

## Acknowledgments

We thank Lisa Blanchard for making the GBS libraries, and Qi Sun for helpful discussions early in this project.

## References

Baird, N.A., P.D. Etter, T.S. Atwood, M.C. Currey, A.L. Shiver, Z.A. Lewis, E.U. Selker, W.A. Cresko, and E.A. Johnson. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:E3376. 10.1371/journal.pone.0003376. doi:10.1371/journal.pone.0003376

Beissinger, T.M., C.N. Hirsch, R.S. Sekhon, J.M. Foerster, J.M. Johnson, G. Muttoni, B. Vaillancourt, C.R. Buell, S.M. Kaeppler, and N. de Leon. 2013. Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics 193:1073–1081. doi:10.1534/genetics.112.147710

Davey, J.W., P.A. Hohenlohe, P.D. Etter, J.Q. Boone, J.M. Catchen, and M.L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12:499–510. doi:10.1038/nrg3012

De Donato, M., S.O. Peters, S.E. Mitchell, T. Hussain, and I.G. Imumorin. 2013. Genotyping-by-sequencing (GBS): A novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. PLoS One 8:E62137. 10.1371/journal.pone.0062137. doi:10.1371/journal.pone.0062137

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:E19379. 10.1371/journal.pone.0019379. doi:10.1371/journal.pone.0019379

Glaubitz, J.C., T.M. Casstevens, F. Liu, J. Harriman, R.J. Elshire, Q. Sun, and E.S. Buckler. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS One 9:E90346. 10.1371/journal.pone.0090346. doi:10.1371/journal.pone.0090346

Gupta, P.K., S. Rustgi, and R.R. Mir. 2008. Array-based high-throughput DNA markers for crop improvement. Heredity 101:5–18. 10.1038/hdy.2008.35. doi:10.1038/hdy.2008.35

Hirsch, C.D., J. Evans, C.R. Buell, and C.N. Hirsch. 2014. Reduced representation approaches to interrogate genome diversity in large repetitive plant genomes. Brief. Funct. Genomics 13:257–267. doi: 10.1093/bfgp/elt051.

Ly, D., M.T. Hamblin, I.Y. Rabbi, M. Gedil, M. Bakare, H.G. Gauch, R. Okechukwu, A.G.O. Dixon, P. Kulakow, and J.-L. Jannink. 2013. Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: A study in cassava. Crop Sci. 53:1312–1325. doi:10.2135/cropsci2012.11.0653

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. 10.1371/journal.pone.0032253. doi:10.1371/journal.pone.0032253

Rabbi, I.Y., M.T. Hamblin, M. Gedil, P. Kulakow, M. Ferguson, A.S. Ikpan, D. Ly, and J.-L. Jannink. 2014. Genetic mapping using genotyping-by-sequencing in the clonally-propagated cassava. Crop Sci. 54:1384–1396. 10.2135/cropsci2013.07.0482.

Varshney, R.K., A. Graner, and M.E. Sorrells. 2005. Genic microsatellite markers in plants: Features and applications. Trends Biotechnol. 23:48–55. doi:10.1016/j.tibtech.2004.11.005

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. Van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and M. Zabeau. 1995. AFLP: A new technique for DNA fingerprinting. Nucleic Acids Res. 23:4407–4414. doi:10.1093/nar/23.21.4407



14350653, 2014, 6, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.2135/cropsci2014.02.0160 by CAPES, Wiley Online Library on [17/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License