2b-RAD: a simple and flexible method for genome-wide genotyping

Shi Wang[1,2,5], Eli Meyer[1,3,5], John K McKay[4] & Mikhail V Matz[1]

We describe 2b-RAD, a streamlined restriction site–associated DNA (RAD) genotyping method based on sequencing the uniform fragments produced by type IIB restriction endonucleases. Well-studied accessions of *Arabidopsis thaliana* were genotyped to validate the method's accuracy and to demonstrate fine-tuning of marker density as needed. The simplicity of the 2b-RAD protocol makes it particularly suitable for high-throughput genotyping as required for linkage mapping and profiling genetic variation in natural populations.

Single-nucleotide polymorphisms (SNPs) have become the genetic markers of choice in many biomedical, ecological and evolutionary studies[1,2]. Until recently, profiling a large number of SNPs was feasible only for model organisms with well-developed genomic resources. Even in such ideal circumstances, however, SNPs identified by surveying part of a population may not be informative for the rest of the population because of ascertainment bias[3]. The SNPs identified by such surveys typically represent common polymorphisms (minor allele frequency >0.05) and are therefore unlikely to be strongly associated with deleterious conditions such as disease. These two issues may explain the relative lack of success in genome-wide association studies of human disease to date[4,5]. To circumvent these problems, it would be ideal to simultaneously discover and genotype SNPs across the whole genome, especially for non-model species without well-developed genotyping platforms.

Despite dramatic reductions in sequencing cost driven by next-generation sequencing technologies, whole-genome sequencing remains costly for species with large genomes. Moreover, many studies do not require the high marker density produced by whole-genome sequencing. Restriction site–associated DNA (RAD) tag sequencing reduces genome complexity by resequencing only the stretches of DNA adjacent to recognition sites of a chosen restriction endonuclease (for a review, see ref. 6) and has proven to be a powerful tool for genetic mapping and analysis of quantitative trait loci[7,8], adaptation[9] and phylogeography[10].

Here we describe a streamlined and flexible approach for RAD genotyping, called 2b-RAD for its use of type IIB restriction enzymes. These enzymes (for example, BsaXI and AlfI) cleave genomic DNA upstream and downstream of the target site, producing tags of uniform length that are ideally suited for sequencing on existing next-generation platforms (**Fig. 1a**). To validate the new method, we prepared BsaXI libraries from a single *Arabidopsis thaliana* F_1 plant produced by crossing two resequenced accessions, Tsu-1 and Kas-1. Libraries were sequenced on both SOLiD and Illumina platforms; each platform generated more than 11 million high-quality reads (**Supplementary Table 1**) and produced even coverage across the nuclear genome except for extended gaps in centromere regions (**Fig. 1b**). To demonstrate the suitability of 2b-RAD for humans and other species with large genomes, we also sequenced and analyzed commercially obtained human DNA (**Supplementary Fig. 1**).

In *A. thaliana*, 2b-RAD reproducibly detected and genotyped the majority of unique BsaXI sites across library preparations and sequencing platforms (**Supplementary Tables 2** and **3**). Genotypes were assigned to sites with ≥20× coverage using either a simple threshold approach adjusted to maximize the agreement between SOLiD and Illumina data sets (see Online Methods) or a maximum-likelihood (ML) approach[9]. Both outputs agreed with bioinformatic predictions for the F_1 hybrid (~99.5%, **Supplementary Tables 4** and **5**). Amplicon (Sanger) sequencing of 50 polymorphic sites closely matched genotypes assigned by either method (**Fig. 2a**), confirming the accuracy of 2b-RAD genotyping (94% for threshold, 82–91% for ML). ML analysis assigned genotypes for a larger number of loci than threshold analysis overall (**Supplementary Table 5**); accordingly, the loci chosen for Sanger sequencing included 50 genotyped by ML and 34 by threshold. The additional homozygous SNPs genotyped by ML showed a higher error rate (4 errors out of 9 additional loci) than the loci genotyped by both methods (1 out of 17; Fisher's exact test; $P = 0.034$). Notably, for the loci genotyped by both methods (97% of ML calls, all threshold based), 100% identical genotypes were obtained.

Because most species studied in ecology and evolution lack an assembled genome sequence, we tested the feasibility and accuracy of *de novo* 2b-RAD analysis (**Supplementary Fig. 2**). Our procedure for deriving a reference sequence from 2b-RAD reads sequenced on the SOLiD platform is similar in principle to the Stacks software recently developed for RAD data[11], but it is more straightforward because 2b-RAD reads cannot have partial overlaps. The reference data set for mapping was created by clustering only the best-quality reads at high stringency (see Online Methods), resulting in 29,823

[1]School of Biological Sciences, University of Texas at Austin, Austin, Texas, USA. [2]College of Marine Life Sciences, Ocean University of China, Qingdao, Shandong, China. [3]Department of Zoology, Oregon State University, Corvallis, Oregon, USA. [4]Department of Bioagricultural Sciences, Colorado State University, Fort Collins, Colorado, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to S.W. (swang@ouc.edu.cn) or E.M. (eli.meyer@science.oregonstate.edu).

RECEIVED 12 SEPTEMBER 2011; ACCEPTED 20 APRIL 2012; PUBLISHED ONLINE 20 MAY 2012; DOI:10.1038/NMETH.2023

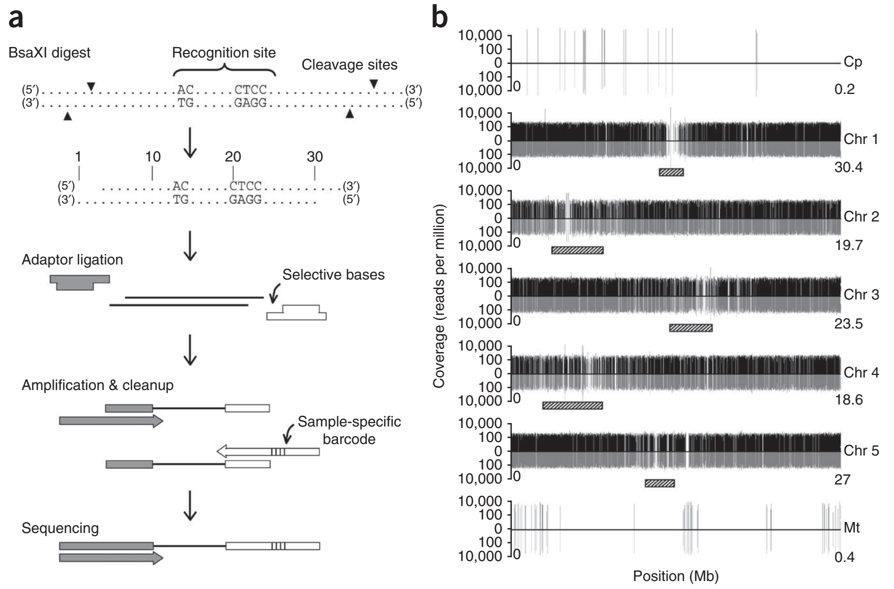

Figure 1 | Preparation and sequencing of 2b-RAD tags. (**a**) Sample preparation for 2b-RAD genotyping is accomplished by restriction digestion (BsaXI) of genomic DNA, cohesive-end ligation of partially double-stranded adaptors with compatible (NNN) overhangs, and incorporation of barcodes for multiplex sequencing by PCR. (**b**) Distribution of sequencing coverage for all genotyped BsaXI sites across the *A. thaliana* genome (Tsu-1 × Kas-1 F_1). Sequencing coverage at all genotyped sites is shown; bars extending upward represent SOLiD coverage, and bars extending downward indicate Illumina coverage. Cp, chloroplast genome; Mt, mitochondrial genome. Cross-hatched bars below each chromosome indicate repetitive centromere regions[15].

unique sequences (the cluster-derived reference, CDR) that represent individual BsaXI sites. Of the CDR sequences, 24,492 (82%) had an unambiguous match to a BsaXI site in the genome. The complete set of high-quality reads was then mapped against the CDR. Comparison of *de novo* and reference-based genotypes (**Supplementary Table 5**) revealed close agreement for homozygous calls (99.9%) and slightly weaker agreement for heterozygous calls (84.7%). This comparison provides an estimate of the small but non-negligible amount of additional genotyping errors expected as 2b-RAD is extended to outbred populations lacking a reference genome. 2b-RAD and other RAD methods would benefit from further efforts to develop new algorithms for improving genotyping accuracy in *de novo* analysis.

Reduction of marker densities allows genotyping of additional samples within the same sequencing capacity, which can be an important benefit in certain situations (for example, with large genome sizes or limited recombination events). The number of AlfI sites in the *Arabidopsis* genome is fewer than one-third that of BsaXI sites, making AlfI an attractive alternative for 2b-RAD genotyping at reduced marker density. We sequenced libraries prepared from *A. thaliana* accession Ler-1 using BsaXI and AlfI to a shallower depth, roughly proportional to the number of sites (**Fig. 2b**). AlfI produced higher coverage per sequencing effort, as expected (**Supplementary Table 6**). Between 95% and 98% of 2b-RAD genotypes agreed with the expected homozygous differences between Ler-1 and the reference Col-0 genome (**Supplementary Table 7**).

Further reductions in marker density can be achieved in 2b-RAD genotyping using modified adaptors. While the standard protocol features adaptors with fully degenerate cohesive ends (5′-NNN-3′), a subset of restriction sites can be targeted using adaptors with less degenerate 3′ overhangs (**Fig. 1a**). We sequenced such a reduced tag representation (RTR)-BsaXI library from the Ler-1 sample based on adaptors with 5′-NNG-3′ overhangs that targeted 1/16th of all BsaXI sites (**Supplementary Table 8**). We were able to genotype 1,222 (84.4%) of the targeted sites (**Fig. 2b**), for which the coverage was enriched 19-fold relative to the standard BsaXI library (**Supplementary Fig. 3**). Comparison with known homozygous SNP data for Ler-1 revealed that high genotyping accuracy (98%) was maintained in the RTR library (**Supplementary Table 7**).

Rarefaction analysis based on deep sequencing of initial BsaXI libraries revealed that genotype detection (at coverage >20×) was saturated at a sequencing depth of 3 million mapped reads, with 84.5% of unique sites genotyped at that depth and only a 5.3% gain obtained by doubling the sequencing effort (**Fig. 2c**).

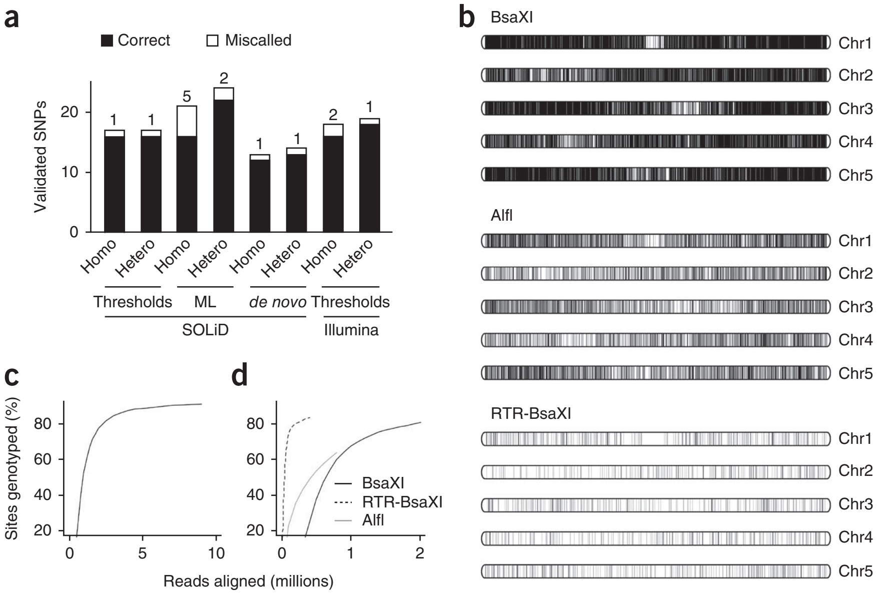

Figure 2 | Accuracy, sequencing requirements and adjustment of marker densities in 2b-RAD genotyping. (**a**) Experimental validation of 2b-RAD genotypes by Sanger sequencing. Number of disagreements is shown above each bar. (**b**) Distributions of genotyped sites (vertical lines) in the *Arabidopsis* genome (Ler-1) for BsaXI (29,493 sites), AlfI (7,726) and RTR-BsaXI libraries (1,222). (**c**) Rarefaction analysis of initial BsaXI libraries. (**d**) Rarefaction analysis of additional low-coverage libraries.

Notably, the RTR-BsaXI library required only one-tenth of the sequencing effort (80.5% of target sites genotyped at 200,000 mapped reads) to reach the genotyping efficiency achieved in the standard BsaXI library (80.4% of sites genotyped at 2.1 million mapped reads, **Fig. 2d**).

2b-RAD provides a streamlined and cost-effective alternative to existing reduced-representation genotyping methods. A detailed comparison of 2b-RAD and other such methods is shown in **Supplementary Table 9**. Like RAD, 2b-RAD allows for nearly every restriction site in the genome to be screened in parallel whereas such methods as reduced representation libraries (RRLs)[12], complexity reduction of polymorphic sequences (CRoPS)[13] and genotyping by sequencing (GBS)[14] can only target a subset of total restriction sites due to the size limit of restriction fragments (usually <500 bp) for efficient PCR amplification and sequencing. The 2b-RAD protocol is simpler than existing RAD protocols[7–10] and therefore suitable for parallel genotyping of large numbers of samples. The current version of the protocol (see Online Methods) does not include any interim purification steps and can be completed in as little as 4 h with all reactions taking place consecutively in a single well of a multi-well plate. The other advantage of 2b-RAD is the option to adjust marker density using selective adaptors; users can balance the level of genotyping detail against throughput depending on the type of study (for example, higher density for genome-wide association, lower density for linkage and quantitative trait loci mapping). This cannot be easily achieved in the existing RAD protocols[7–10]. The highly reduced 2b-RAD libraries require much less sequencing to achieve accurate genotyping (**Fig. 2d**), further increasing the suitability of 2b-RAD for high-throughput genotyping tasks with a fixed sequencing capacity.

Other RAD methods can take advantage of increasing read lengths on sequencing platforms, whereas tag lengths are constrained by the activity of type IIB restriction endonucleases in 2b-RAD, raising the concern that they may not be long enough for efficient locus discrimination in highly complex or heterogeneous genomes. In previous RAD studies, however, read lengths comparable to those of 2b-RAD tags (33–36 bp) proved effective (for example, 26 bp in ref. 7; 54 bp in ref. 9). In the *A. thaliana* genome, the great majority of 2b-RAD tags are unique (91% of AlfI and 92% of BsaXI tags), sites differing by only a single nucleotide are rare (5%) and the unambiguous best-match criterion used in the mapping excludes nearly all errors that could result from such similarities without any appreciable loss of high-quality reads. Even in the human genome, which is 26 times larger than that of *A. thaliana* and contains 1.3 million BsaXI sites, 97% of the sites are unique. *K*-mer analysis suggests that increasing read lengths beyond that of IIB restriction fragments would offer little advantage: 97.4% of 30-mers—and 98.8% of 100-mers—are unique in *A. thaliana*. Consistent with these observations, our experimental and bioinformatic validation of 2b-RAD genotypes confirms the accuracy of genotyping based on type IIB restriction fragments.

Although reference-based analysis is clearly preferable in model systems, our *de novo* analysis demonstrated that 2b-RAD genotyping achieves sufficient power and accuracy to make it applicable to organisms lacking a complete genome sequence. This highlights one of the most promising applications for RAD technology in comparative genomics: creation of dense genetic linkage maps and assembly of chromosome-wide genomic scaffolds. The 2b-RAD approach is part of a growing suite of complementary methods for genotyping based on sequencing reduced representation libraries, all of which hold great promise for studies of ecology and evolution in diverse species.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence Read Archive (SRA) database: SRP008452 and SRP009306.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grants DEB-1054766 to M.V.M. and DEB-1022196 to J.K.M., and by grants from the National Natural Science Foundation of China (31130054) and National High Technology Research and Development Program of China (2012AA10A405) to S.W. We are grateful to T.E. Juenger and D.L. Des Marais (University of Texas at Austin) for growing *Arabidopsis* samples.

AUTHOR CONTRIBUTIONS

S.W. and M.V.M. conceived and designed the study. S.W. developed the original protocol for 2b-RAD library preparation, and E.M. developed modifications for RTR. S.W. and E.M. prepared 2b-RAD libraries for SOLiD and Illumina sequencing. E.M. created bioinformatics scripts and conducted sequence analysis and genotype validation. J.K.M. provided *Arabidopsis* samples and contributed resequencing data for bioinformatic validation. S.W., E.M., J.K.M. and M.V.M. wrote the paper. M.V.M. supervised the whole study.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

**Published online at http://www.nature.com/doifinder/10.1038/nmeth.2023.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.**

1. Gray, I.C., Campbell, D.A. & Spurr, N.K. *Hum. Mol. Genet.* **9**, 2403–2408 (2000).
2. Morin, P.A., Luikart, G., Wayne, R.K. & the SNP Workshop Group. *Trends Ecol. Evol.* **19**, 208–216 (2004).
3. Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. & Nielsen, R. *Genome Res.* **15**, 1496–1502 (2005).
4. Maher, B. *Nature* **456**, 18–21 (2008).
5. Manolio, T.A. *et al. Nature* **461**, 747–753 (2009).
6. Davey, J.W. *et al. Nat. Rev. Genet.* **12**, 499–510 (2011).
7. Baird, N.A. *et al. PLoS ONE* **3**, e3376 (2008).
8. Chutimanitsakun, Y. *et al. BMC Genomics* **12**, 4 (2011).
9. Hohenlohe, P.A. *et al. PLoS Genet.* **6**, e1000862 (2010).
10. Emerson, K.J. *et al. Proc. Natl. Acad. Sci. USA* **107**, 16196–16200 (2010).
11. Catchen, J., Amores, A., Hohenlohe, P., Cresko, W. & Postlethwait, J. *G3* (*Bethesda*) **1**, 171–182 (2011).
12. Van Tassell, C.P. *et al. Nat. Methods* **5**, 247–252 (2008).
13. van Orsouw, N.J. *et al. PLoS ONE* **2**, e1172 (2007).
14. Elshire, R.J. *et al. PLoS ONE* **6**, e19379 (2011).
15. Clark, R.M. *et al. Science* **317**, 338–342 (2007).

ONLINE METHODS

DNA sources. As a main validation experiment, we profiled a single F_1 individual produced by crossing divergent accessions Tsu-1 and Kas-1 of the model plant *A. thaliana*[16]. Because *A. thaliana* accessions are typically highly inbred[17], an F_1 cross was used to create higher levels of heterozygosity than in the parental lines, allowing for validation of heterozygous genotypes. To evaluate modified protocols that reduce marker density and sequencing cost, we sampled an individual from *A. thaliana* accession Ler-1. Genomic DNA was extracted from leaf tissue using the DNeasy Plant Mini Kit (Qiagen) according to the manufacturer's instructions. Human DNA was purchased from Qiagen (catalog no. 59568).

2b-RAD sample preparation and sequencing. An overview of sample preparation for 2b-RAD genotyping is shown in **Figure 1**. A further optimized step-by-step protocol is available in the **Supplementary Protocol**. The most current version of the protocol can be downloaded from the Matz laboratory website (http://www.bio.utexas.edu/research/matz_lab/).

Library preparation began with the digestion of 100–200 ng genomic DNA in a 15-µl reaction using 4 U BsaXI (NEB) or 2 U AlfI (Fermentas) at 37 °C for 3 h. A small aliquot (~30 ng) was separated on a 1% agarose gel to verify digestion. Next, 12 µl of a ligation master mix containing 0.2 µM each of two library-specific adaptors, 1 mM ATP (NEB), and 800 U T4 DNA ligase (NEB) was added to the digestion product and incubated for 16 h (4 °C for BsaXI digests, 16 °C for AlfI digests). Then heat inactivation was performed for AlfI at 65 °C for 20 min. All adaptor sequences are provided in **Supplementary Table 10**. Standard BsaXI libraries for SOLiD used sld-ada1 and sld-ada2; BsaXI libraries for Illumina used slx-ada1 and slx-ada2. Adaptors used for AlfI libraries (sld-ada1-AlfI and sld-ada2-AlfI) featured the two-nucleotide overhang appropriate for AlfI. Reduced tag representation (RTR) BsaXI libraries used sld-rtr-ada1-BsaXI and sld-rtr-ada2-BsaXI, which feature 'NNG' overhangs in place of the usual 'NNN' to target the subset of BsaXI fragments with ends complementary to these overhangs.

Ligation products were amplified in three 20-µl reactions per sample, each composed of 7 µl ligated DNA, 0.1 µM each primer (sld-p1 and sld-p2 for SOLiD; slx-p1 and slx-p2 for Illumina), 0.3 mM dNTP, 1× Phusion HF buffer and 0.4 U Phusion high-fidelity DNA polymerase (NEB). PCR was conducted in a DNA Engine Tetrad 2 thermal cycler (Bio-Rad) with 20–22 cycles of 98 °C for 5 s, 60 °C for 20 s and 72 °C for 10 s and then a final extension of 10 min at 72 °C. The target band (SOLiD: 76 bp; Illumina: 96 bp) was excised from a 2% agarose gel, and the DNA was allowed to diffuse from the agarose into nuclease-free water for 12 h at 4 °C. Finally, barcodes were introduced by means of PCR with platform-specific barcode-bearing primers. Each 20-µl PCR reaction contained 25 ng of gel-extracted PCR product, 0.1 µM of each primer (sld-p3 and sld-p4 for SOLiD; slx-p1 and slx-p3 for Illumina), 0.3 mM dNTP, 1× Phusion HF buffer and 0.4 U Phusion high-fidelity DNA polymerase; four or five cycles of the PCR profile listed above were performed. PCR products were purified using QIAquick PCR purification kit (Qiagen) before sequencing. SOLiD sequencing was performed at the Genome Sequencing and Analysis Facility at the University of Texas at Austin (SOLiD System 3.0), and Illumina sequencing (GA-II) was performed at the Genomics Core Facility at Tufts University. All 2b-RAD sequences (SOLiD and Illumina) were archived in the Sequence Read Archive (SRA) under accession number SRP008452.

Quality filtering. Reads 35-bp long were obtained from SOLiD and Illumina platforms. Terminal tag positions were excluded from each read (positions 1 and 33, **Fig. 1**) to eliminate artifacts that might arise at ligation sites. Next, reads with ambiguous base calls (N), long homopolymer regions (>10 bp) or excessive low-quality positions (>5 positions with quality <10) that might interfere with accurate mapping were removed. The remaining trimmed, high-quality reads formed the basis for all subsequent mapping and genotyping.

Reference-based analysis. High-quality reads were aligned against the known BsaXI or AlfI sites in the *Arabidopsis* genome (TAIR9) using SHRiMP software package, version 1.2.0 (ref. 18). SOLiD reads were mapped in color space (using rmapper-cs), and Illumina reads were mapped in nucleotide space (using rmapper). Mapping parameters were similar to those described by the software's authors[18], with a spaced seed of 111100111, relaxed penalties for mismatches and gap opening (-i -90 -g -250 -q -250), increased penalties for gap extension (-e -100 -f -100), and relaxed Smith-Waterman thresholds for the full and vector searches (-h 2000 -v 1000). The resulting matches were filtered to eliminate short alignments (<28 bp) using custom Perl scripts, and to eliminate statistically weak ($P > 0.001$) and ambiguous matches (reads matching more than one site equally well) using SHRiMP's probcalc program (-n 0.8 -p 0.001). Finally, alignments were produced for each read and its matching reference site using SHRiMP's prettyprint and prettyprint-cs programs.

***De novo* analysis.** To enable 2b-RAD in organisms that lack a completed genome sequence, we developed a simple *de novo* procedure that generates a reference database from restriction sites identified in 2b-RAD reads (**Supplementary Fig. 1**). The highest-quality color-space reads (all quality scores ≥20) were translated into nucleotide sequences, and any reads lacking a perfect match to the recognition site (for BsaXI, $N_{12}ACN_5CTCCN_{10}$ or its reverse complement) were excluded. The non-redundant set of matching sequences observed more than once (analogous to the collection of alleles across all BsaXI sites) was obtained by clustering the remaining reads at 100% sequence identity using the CD-HIT software package[19]. Finally, this non-redundant collection of reproducibly observed sequences was clustered at 90% sequence identity (that is, allowing for up to three mismatches) in order to group tags derived from different alleles at the same site. The representative sequences from these clusters comprised the cluster-derived reference (CDR) and were used for mapping the reads as described in the previous paragraph. To compare the genotype calls produced by this method to the results of reference-based analysis, we matched the CDR sequences to known BsaXI sites using SHRiMP, requiring long (≥28 bp) and unambiguous matches.

Genotyping. Genotypes at the sites with ≥20× coverage were determined from sequence alignments using custom Perl scripts (available in **Supplementary Software**). The distribution of base calls was compiled for each genotyped position in every site (excluding terminal bases and enzyme recognition sites, **Fig. 1**).

Any terminal alignment regions containing ambiguous bases were excluded from genotyping, as alternate alleles in such positions cannot be detected in the alignments output from SHRiMP. Crossovers (sequencing errors deduced from di-base disagreement) and gaps were scored as N. Each position was scored as homozygous if only a single allele was present, or if a second allele was found at low coverage (≤0.5% of the total); as heterozygous if the second allele comprised ≥35% of coverage; or as undetermined otherwise. These thresholds were selected to maximize the genotypes' agreement between SOLiD and Illumina data sets while minimizing the number of undetermined positions. The performance of this simple threshold-based approach was compared to a previously described maximum likelihood method of genotype calling[9].

Validation by Sanger sequencing. Candidate sites were randomly selected for validation, including approximately equal numbers of heterozygous and homozygous SNP calls. Primers were designed using Primer 3 (primer3.sourceforge.net) based on the TAIR9 *A. thaliana* assembly to amplify an ~400-bp fragment flanking each target site (primer sequences are provided in **Supplementary Table 11**). PCR products amplified from the sample originally used for 2b-RAD were treated with ExoSAP-IT (Affymetrix) to remove unincorporated primers, and sequenced using the Sanger method (Beckman Coulter Genomics) to verify 2b-RAD genotypes.

Bioinformatic validation. Expected F_1 genotypes for the Tsu-1 × Kas-1 F_1 sample were predicted based on resequencing of the parental accessions and compared to 2b-RAD genotype calls. For Kas-1, 125 Mb of 454 genomic DNA sequences (SRA accession number SRP009306) were aligned against the *A. thaliana* genome (TAIR9) using the Roche GS Reference Mapper (version 2.51p). For Tsu-1, we aligned 4.7 Gb of Illumina sequences (SRA accession number SRX000704) against the *A. thaliana* genome (TAIR9) using the SOAP aligner version 2.20 (ref. 20). Genotypes were determined for each position in this alignment using SOAPsnp[21] with default settings and a coverage threshold of ≥3×. Ambiguous base calls (representing putatively heterozygous loci in the parental lines) were excluded from both data sets because those could not be used to unambiguously predict F_1 genotypes. Bioinformatic validation of genotypes called for *A. thaliana* accession Ler-1 involved direct comparison with genotype data for Ler-1 obtained from the 1001 genomes project (http://www.1001genomes.org/).

16. Mckay, J.K., Richards, J.H. & Mitchell-Olds, T. *Mol. Ecol.* **12**, 1137–1151 (2003).
17. Clauss, M.J., Cobban, H. & Mitchell-Olds, T. *Mol. Ecol.* **11**, 591–601 (2002).
18. Rumble, S.M. *et al. PLoS Comput. Biol.* **5**, e1000386 (2009).
19. Li, W. & Godzik, A. *Bioinformatics* **22**, 1658–1659 (2006).
20. Li, R. *et al. Bioinformatics* **25**, 1966–1967 (2009).
21. Li, R. *et al. Genome Res.* **19**, 1124–1132 (2009).

doi:10.1038/nmeth.2023