MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 1264–1278 doi: 10.1111/1755-0998.12566

SPECIAL ISSUE: SEQUENCE CAPTURE

# RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data

SANDRA L. HOFFBERG,* TROY J. KIERAN,† JULIAN M. CATCHEN,‡ ALISON DEVAULT,§ BRANT C. FAIRCLOTH,¶ RODNEY MAURICIO* and TRAVIS C. GLENN*,†

*Department of Genetics, University of Georgia, Athens, GA 30602, USA, †Department of Environmental Health Science, University of Georgia, Athens, GA 30602, USA, ‡Department of Animal Biology, University of Illinois, Urbana, IL 61801, USA, §MycroArray, 5692 Plymouth Rd., Ann Arbor, MI 48105, USA, ¶Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA

## Abstract

**Molecular ecologists seek to genotype hundreds to thousands of loci from hundreds to thousands of individuals at minimal cost per sample. Current methods, such as restriction-site-associated DNA sequencing (RADseq) and sequence capture, are constrained by costs associated with inefficient use of sequencing data and sample preparation. Here, we introduce RADcap, an approach that combines the major benefits of RADseq (low cost with specific start positions) with those of sequence capture (repeatable sequencing of specific loci) to significantly increase efficiency and reduce costs relative to current approaches. RADcap uses a new version of dual-digest RADseq (3RAD) to identify candidate SNP loci for capture bait design and subsequently uses custom sequence capture baits to consistently enrich candidate SNP loci across many individuals. We combined this approach with a new library preparation method for identifying and removing PCR duplicates from 3RAD libraries, which allows researchers to process RADseq data using traditional pipelines, and we tested the RADcap method by genotyping sets of 96–384 *Wisteria* plants. Our results demonstrate that our RADcap method: (i) methodologically reduces (to <5%) and allows computational removal of PCR duplicate reads from data, (ii) achieves 80–90% reads on target in 11 of 12 enrichments, (iii) returns consistent coverage (≥4×) across >90% of individuals at up to 99.8% of the targeted loci, (iv) produces consistently high occupancy matrices of genotypes across hundreds of individuals and (v) costs significantly less than current approaches.**

*Keywords*: 3RAD, ddRAD, Illumina, reduced representation libraries, target enrichment, *Wisteria*

*Received 19 April 2016; revision received 6 July 2016; accepted 11 July 2016*

## Introduction

Massively parallel sequencing is changing molecular ecology and other life science disciplines (Rogers & Venter 2005; Tautz *et al.* 2010). While the costs of whole-genome sequencing and genome resequencing have declined, the time investment, cost and computational complexity of genome assembly and genome resequencing remain significant drawbacks. Fortunately, many biological hypotheses can be tested with a fraction of the genome, from several hundred to several thousand variable loci (Cariou *et al.* 2013; Pante *et al.* 2015). Although genome reduction techniques that collect data from hundreds or thousands of loci are an appealing and inexpensive proxy for whole-genome resequencing, the matter of how best to collect genotypes from many loci across hundreds or thousands of individuals remains (Harvey *et al.* 2016).

Genome reduction techniques fall into a broad class of so-called 'reduced representation' approaches, which collect data from a small and repeatable fraction of the genome across a population of individuals, enabling the population under study to be compared at homologous loci (Altshuler *et al.* 2000; Novaes *et al.* 2008; Wiedmann *et al.* 2008). Sequence capture (Okou *et al.* 2007; Gnirke *et al.* 2009) and restriction-site-associated DNA sequencing (RADseq; Miller *et al.* 2007; Baird *et al.* 2008; Davey & Blaxter 2010; Davey *et al.* 2011; Peterson *et al.* 2012) are two widely used types of reduced representation approaches for massively parallel sequencing. Although both methods have advantages and disadvantages (Harvey *et al.* 2016), neither is entirely capable of

Correspondence: Sandra Hoffberg, Fax: 706-542-3910; E-mail: sandra@hoffberg.org and Travis C. Glenn, Fax: 706-542-7472; E-mail: travisg@uga.edu

achieving a primary goal of many population genetic studies: consistently obtaining a set of hundreds or thousands of putatively unlinked single-nucleotide polymorphisms (SNPs) from hundreds to thousands of individuals at low cost (e.g. <$10/sample).

Sequence capture combines a custom set of long, biotinylated, oligonucleotide baits with in-solution hybridization to enrich any number of genomic regions of nearly any size (Gnirke *et al.* 2009; Saintenac *et al.* 2011; Cao *et al.* 2013). Sequence capture requires prior sequence information to design capture baits (Gnirke *et al.* 2009). Several groups have designed bait sets that target conserved sequences (Bi *et al.* 2012; Faircloth *et al.* 2012; Lemmon *et al.* 2012), which allow sets of baits to be used across many species. Sequence capture is constrained by high library preparation costs, expensive baits and randomness of where the collected sequences start and stop, and off-target sequence reads (Harvey *et al.* 2016).

RADseq methods reduce the genome by sequencing many thousands of DNA fragments that are located near restriction enzyme cut sites (Miller *et al.* 2007; Baird *et al.* 2008; Davey *et al.* 2011). Various RADseq derivatives (Andrews *et al.* 2016) have been developed based on the original RADseq method (Miller *et al.* 2007; Baird *et al.* 2008; Davey *et al.* 2011), including our 3RAD variant (Graham *et al.* 2015), and we use the term 'RADseq' to generically refer to any of the derivative forms of RADseq. Compared to sequence capture, RADseq methods generally have lower library preparation costs and do not explicitly require genomic information from the taxa of interest (Harvey *et al.* 2016; Heyduk *et al.* 2016).

The quality of RADseq data sets is often diminished due to missing data from stochastic variation (mutation and methylation) and molecular and bioinformatic protocols (see Mastretta-Yanes *et al.* (2015) for a review of RADseq limitations). Errors introduced to RADseq libraries during PCR are particularly problematic. Incorporation errors that occur early during the PCR reaction can be amplified to high coverage as PCR proceeds (Tin *et al.* 2015), and PCR duplication of loci can give falsely high confidence in the accuracy of downstream variant calls (Casbon *et al.* 2011; Schweyen *et al.* 2014; Tin *et al.* 2015). For example, many RADseq processing pipelines use coverage to validate the accuracy of SNP calls even though PCR duplicates can comprise 20–90% of reads in RADseq libraries (Andrews *et al.* 2014; Schweyen *et al.* 2014; Tin *et al.* 2015; Ali *et al.* 2016).

The traditional approach for distinguishing duplicates in standard genomic libraries, which are randomly sheared on both ends, and RADseq libraries that are randomly sheared on one end is to identify duplicate reads as those having identical start and stop positions when aligned to a reference sequence. However, this technique cannot be applied to ddRAD-type approaches, where all sequence reads from a RAD locus are identical (Andrews *et al.* 2014). Single-molecule tagging has been employed to identify and remove PCR duplicates in a variety of approaches (Miner *et al.* 2004; Kivioja *et al.* 2012; Smith *et al.* 2014), including RADseq and ddRAD, by incorporating degenerate bases in adapters (Casbon *et al.* 2011; Schweyen *et al.* 2014; Tin *et al.* 2015), but all of the methods have limitations in their general implementation.

Here, we introduce RADcap, a novel method that combines the benefits of single-molecule tagging with 3RAD and sequence capture to collect a consistent and repeatable sample of hundreds of loci across hundreds of individuals, remove PCR duplicates from the resulting data and call SNPs using a probabilistic base-calling pipeline (GATK; DePristo *et al.* 2011; McKenna *et al.* 2010). The RADcap workflow begins with a pilot experiment using 3RAD to collect genetic information from a small sample of individuals. After processing the resulting sequence reads using STACKS (Catchen *et al.* 2011, 2013) to identify variable RAD loci, the workflow proceeds by designing a set of biotinylated ssRNA baits targeting a subset of the variable RAD loci and enriching the targeted loci from a pool of DNA libraries prepared using our inexpensive 3RAD library preparation process. To ameliorate the problem of false confidence in genotype calls bolstered by PCR duplicates, the RADcap approach incorporates a random 8-nucleotide (nt) sequence tag in place of the iTru5 primer index (Fig. 1 and Fig. S1, Supporting information) into each library molecule, which allows researchers to distinguish PCR duplicates from unique template molecules during postprocessing of the sequence data. Finally, following a GATK workflow, we created a RADcap data processing package, which calls SNPs in the duplicate-free reads using a 'radnome' (those RAD loci we targeted with capture baits) as a reference sequence. We empirically tested the RADcap method by measuring genetic diversity of 96 samples of *Wisteria* collected across an urban centre as well as 203 greenhouse-grown seedlings.

## Methods

### *Study system and experimental design*

*Wisteria* is a genus of flowering plants in the family Fabaceae that includes a number of woody perennial climbing vines that reproduce sexually and vegetatively (Valder 1995). In the southeastern United States, two species of *Wisteria*, *W. floribunda* and *W. sinensis*, were introduced from East Asia in the early 19th century (Wilson 1916; Wyman 1949) as ornamentals. Most individuals of introduced *Wisteria* growing in the

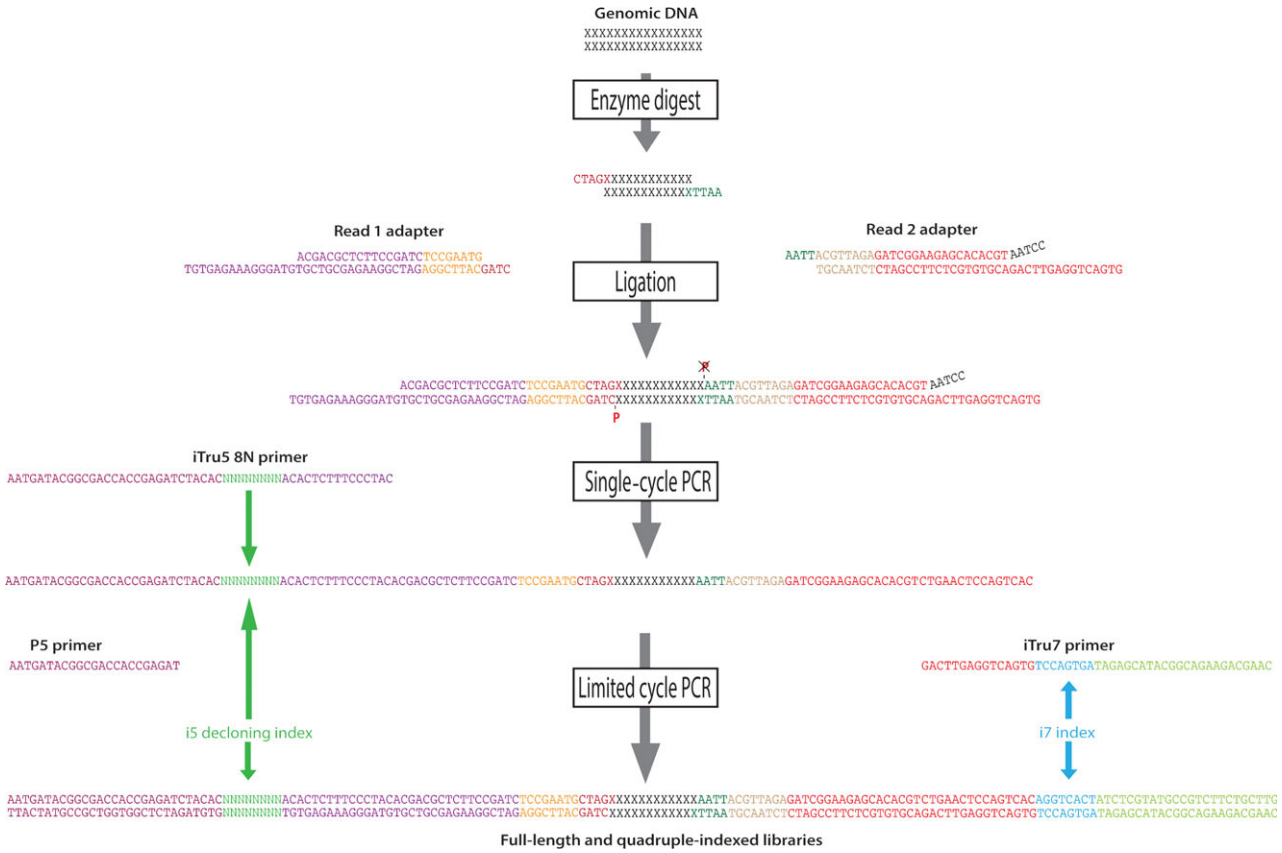


**Fig. 1** The RADcap workflow illustrating components of the library molecule and the sequence of the ends of library molecules. Genomic DNA is digested with enzymes that leave enzyme-specific sticky ends, to which we ligate adapters. The Read 1 adapter is comprised of four bases that bind to the XbaI restriction-site overhang (dark red), a sample-specific internal sequence tag, used to identify the sample (orange), and a Read 1 sequencing primer that is partially single stranded to facilitate annealing of the iTru5 primer (purple). The Read 2 adapter is a y-yoke adapter composed of the four bases that bind to the EcoRI restriction-site overhang (dark green), a sample-specific internal sequence tag (tan) and the Read 2 sequencing primer (red). During the one-cycle PCR, the iTru5 primer is added to the library: the partial library is denatured, the primer anneals to the Read 1 sequencing primer overhang (purple), and extends, thereby adding the degenerate barcode with 8 N bases (green) and the P5 primer (maroon) which anneals to the Illumina flow cell. After cleaning up the reaction, a limited cycle PCR is performed to add the iTru7 primer, comprised of the Read 2 sequencing primer (red) which anneals to the single-stranded adapter added earlier, a sample-specific barcode (blue) and P7 primer (light green) which anneals to the Illumina flow cell.

southeast today are hybrids of *W. sinensis* and *W. floribunda* (Trusty *et al.* 2007, 2008). While currently available genetic markers can distinguish species, there are no markers available with enough resolution to distinguish among individuals from the same population. Understanding the population genetics of this introduced species requires many more markers and is crucial to understanding how *Wisteria* is spreading. We compare estimates of genetic diversity of 96 wild-collected samples from Athens, GA, obtained from RADcap to genetic diversity of the same samples prepared via 3RAD. We calculate the efficiency with which we sequence these loci by also genotyping 203 greenhouse-grown *Wisteria* seedlings (see Appendix S1, Supporting information).

### *3RAD SNP discovery and bait design*

We collected sequence information for sequence capture baits from a pilot 3RAD study of four individual *Wisteria* plants: three samples collected around Athens, GA (wist69-3, wist124-1 and wist276-4), and one sample collected from a greenhouse-grown seedling (Wmat9-7-P5-S1). We prepared samples using 3RAD (Graham *et al.* 2015), which we summarize below and explain, in detail, in Appendix S1 (Supporting information). We added short forward and reverse adapters with inline barcodes to extracted DNA from each of the four samples, and we performed a restriction digest of this solution using XbaI, EcoRI-HF and NheI-HF (Fig. 1; see Appendix S1, Supporting information, for

enzyme selection). Following initial digestion, we added T4 DNA ligase to the digested DNA without disabling the restriction enzymes, and we cycled temperatures to sequentially promote ligation of adapters followed by digestion of chimeras and dimers. Chimeras and dimers are preferentially digested because they re-create the enzyme cut site, whereas adapters ligated to target DNA do not. We cleaned the resulting reactions with NaCl-PEG diluted SpeedBeads (Rohland & Reich 2012), and we completed the adapter sequences using PCR with iTru5 and iTru7 primers (see Glenn *et al.* (2016) for details about primers). We pooled the resulting libraries, size-selected fragments of 550 bp (±10%) with a Pippin Prep (Sage Science, Inc.), and performed a final round of low-cycle PCR using the P5 and P7 primers to increase the concentration of fragments in the desired size range. We sequenced samples on an ILLUMINA NEXTSEQ v2 300 cycle kit to obtain paired-end 150-nt (PE150) reads (Fig. S1, Supporting information).

In 3RAD (and RADcap) libraries, the forward and reverse adapters with internal indices are used to distinguish samples, and the iTru7 primer is used to distinguish plates. We used the *process_radtags* program in STACKS v1.29 (Catchen *et al.* 2011, 2013) to clean and demultiplex the resulting sequence data. We 'rescued' sequence tags and RAD tags within 2 bp of their expected sequence; otherwise, we removed reads with an uncalled base or containing the wrong adapter or wrong cut site. Because our 3RAD adapter sequences vary in length, and because STACKS requires all reads to be the same length, we used *process_radtags* to truncate reads to 140 bp, removing 0–3 bases of sequence per read. We parallel-merged the mates of paired-end reads (paste command in Unix). We ran the STACKS pipeline with the following modifications: in the ustacks program, we removed highly repetitive stacks, we used the deleveraging algorithm, and we set the maximum distance between stacks (M) to 3; in the cstacks program, we set the number of mismatches allowed between sample tags when generating the catalog (n) to 4; in the program *populations*, we required at least three individuals to have reads to retain a given locus (r), and we set the minimum stack depth required for individuals at a locus (m) to 3. We output the full sequence from each allele identified across our pilot samples in FASTA format. We selected loci that were polymorphic, but had less than five SNPs across both paired-end reads and that were present in three or four of the samples, which resulted in 1740 paired reads (candidate loci) for bait design.

We selected bait sequences to minimize target redundancy and bait-to-bait hybridization, which can compromise the synthesis of ssRNA baits as well as the target capture hybridization reaction. To perform these steps, we subjected sequences to self-analysis using BLAST 2.2.19 (*filter query sequence = false, word size = 11, e-value = 1e-13, number of sequences to show alignments for = 2000*; Boratyn *et al.* 2013). We discarded any locus with one or both sequences having a BLAST hit of at least 140 bp to another sequence (682 loci). Next, we subjected sequences to a same-strand self-analysis in BLAST (as above, *query strand = bottom*). We discarded 94 additional loci in which one or both paired sequences had a BLAST hit to another sequence, leaving 964 loci. Then, we designed two sets of 90 mer baits targeting these 964 loci. In the first set, we chose a single bait from both mates of paired-end sequences for every locus, and we positioned baits to start at the 20th base of their parent sequences (creating 1928 *Wisteria* baits, 2 per locus; Appendix S2: Wist-Probes-Set1.fasta, Supporting information). In the second set, we added additional baits from both sequences corresponding to a random subset of 200 loci (creating 400 additional baits; Appendix S3: Wist-Probes-Set2-SUBSET-400.fasta, Supporting information), and we positioned these baits to start at the 40th base of their parent sequences. The two sets produced a total of 2328 baits targeting 964 *Wisteria* library molecules. To reduce synthesis costs, we combined this bait set design with a similar number of baits designed in the same way for another species (*Pueraria montana* var. *lobata*, kudzu; see Discussion). We subjected the bait sequences for both species to a final same-strand self-analysis using BLAST (same process as above), and we did not find evidence of additional bait-to-bait hybridization. Before bait synthesis and because MYbaits cannot be synthesized with a mixture of bases, we replaced any variable positions in any bait sequence with a random candidate base, and we replaced all unknown ('N') positions with a thymine. We created a custom set of biotinylated RNA baits by having them synthesized as a MYbaits-1 kit (MYcroarray, Ann Arbor, MI, USA).

### *Library preparation, experimental treatments and sequencing*

We provide detailed sample collection and sample preparation methods in Appendix S1 (Supporting information). Briefly, we randomly arranged 191 of the 202 greenhouse-grown *Wisteria* samples in two plates (RADcap_Plate1 and RADcap_Plate2; Table S1, Supporting information) with one sample included on both plates. We placed the remaining 11 greenhouse-grown samples into a third plate, along with four samples that are duplicated and 41 samples that are triplicated from plates 1 and 2 (RADcap_Plate3; Table S1, Supporting information). This arrangement allowed us to reprocess 133 libraries independently prepared from the same extracted DNA, and we used these replicates to compute

the amount of missing data between replicate samples that was not caused by genetic variation. We arranged the 192 samples from wild-collected individuals by DNA concentration in a fourth and fifth plate (RADcap_Plate4 and RADcap_Plate5; Table S1, Supporting information). We normalized DNA concentration across plates, then we digested the plated DNA with XbaI, NheI-HF and EcoRI-HF in a reaction that included forward and reverse adapters. As before, we added T4 DNA ligase to the digested DNA without disabling the restriction enzymes, and we cycled temperatures to sequentially promote ligation of adapters followed by digestion of chimeras and dimers (Fig. 1; see Appendix S1, Supporting information, for enzyme selection). Following adapter ligation, we combined approximately 33% of the ligation volume from each sample in each plate into plate-specific pools, we cleaned each pool with SpeedBeads, and we resuspended cleaned pools in 33 μL of TLE. For plates 1–4, we split each pool into three aliquots of 20 μL, 10 μL and 3 μL, and we used these aliquots to test the effect of different PCR conditions on the efficiency of RADcap (described below; Table 1). To tag and track duplicate reads that resulted from the PCR amplification process, we designed a new iTru5 (Glenn *et al.* 2016) primer that incorporated a random 8 nt sequence tag (i.e., the i5 index sequence was specified as NNNNNNNN when ordering the iTru5-8N primer). This resulted in the synthesis of a mixture of 65 536 iTru5 primers with unique 8 nt index sequences. In the experimental treatments, below, we incorporated these uniquely tagged iTru5-8N primers in to our DNA library constructs using different PCR conditions to determine what methods produce the fewest PCR duplicates.

*Treatment 1: One-primer, one-cycle amplification.* Following adapter ligation and cleaning, we split each 20 μL aliquot into two tubes to increase the total PCR volume possible, and we performed a single-cycle, one-primer PCR (Fig. 1). Each reaction contained 10 μL template DNA and the iTru5-8N primer. Because we amplified each reaction using only one cycle, the primers did not denature from the library molecules and re-anneal to different library molecules. We pooled the two resulting reactions and cleaned them with SpeedBeads, and we split them into two tubes for a 6-cycle PCR where we included the P5 primer and the plate-specific iTru7 primer (Table S2, Supporting information). This second reaction completed the library construct, added the plate-specific i7 index sequence to each library construct and increased the total amount of library available for capture. We called the plates in this treatment RADcap_1cycle_Plate1–4.

*Treatment 2: Two-primer, five-cycle amplification.* For the aliquots of 10 μL, we performed four PCRs for each pooled plate with 2 μL template DNA in each. We included both the iTru5-8N primer and the iTru7 primer (Table S2, Supporting information) in each PCR, and we ran PCR for five cycles. Because we included the iTru5-8N primer in the PCR reaction for multiple cycles, newly synthesized molecules could receive new iTru5 tags (Casbon *et al.* 2011), and thus, a single template DNA molecule could generate multiple library constructs with unique i5 sequence tags (i.e. this method produced ≤10 undetectable PCR duplicates per template molecule). Because we used libraries in these treatments that were identical to those used above (i.e. one-primer amplification with a single-cycle PCR), this experiment allowed us to determine the effect of low-efficiency first-strand replication and test how additional PCR cycles affect the identification of PCR duplicates and subsequent variant calling. We called the plates in this treatment RADcap_5-cycle_Plate1–4.

*Treatment 3: Low-template, one-primer amplification.* It is thought that using less template DNA can exacerbate the problem of PCR read duplication (Casbon *et al.* 2011). We used the 3 μL aliquot from plate 1 to determine the effect of low DNA concentrations on PCR duplication and subsequent variant calling. We added the iTru5-8N primer to 3 μL of template from plate 1,

**Table 1** Overview of PCR conditions for each treatment, DNA plates in each treatment and how plates were grouped for analyses

| Treatment | Name | Plate IDs | Cycles with iTru5-8N | iTru5-8N reaction volume (μL) | All PCR duplicates tagged? | Captured? | Analysis groups (plate ID's) |
|---|---|---|---|---|---|---|---|
| 1 | RADcap_1cycle | 1–4 | 1 | 100 | Yes | Yes | 1; 1–4 |
| 2 | RADcap_5cycle | 1–4 | 5 | 100 | No | Yes | 1; 1–4 |
| 3 | RADcap_Low_Template | 1 | 1 | 25 | Yes | Yes | 1 |
| 4 | RADcap_optimized | 5 | 1 | 300* | Yes | Yes | 5 |
| 5 | 3RAD_SizeSelect | 5 | 1 | 300* | Yes | No | 5 |

*The iTru5-8N reaction volume for the size-selected and optimized treatment represents the same reactions, as these treatments were split after the single-primer PCR and clean-up.

and we a performed a single-cycle PCR. As we described for the one-primer amplifications above, we cleaned the resulting PCR product and performed another six-cycle PCR to add the iTru7 primer (Table S2, Supporting information). We called this treatment RADcap_Low_Template_Plate1.

*RAD locus capture.* Following all final PCRs described above, we pooled the replicate PCRs, cleaned each plate pool with SpeedBeads and performed a separate capture hybridization reaction on each pool from each treatment (three plates with two treatments and one plate with three treatments, for nine total captures) according to the MYcroarray MYbaits v3.0 protocol, with a hybridization temperature of 65 °C for 21 h. Following capture, we split each capture into three tubes and amplified loci with the P5 and P7 primers in an 18 cycle PCR recovery. PCRs from each capture were pooled and cleaned with SpeedBeads. Following PCR and clean-up, we quantified the nine experimental treatments, and we pooled these libraries with unrelated libraries from other experiments at a ratio that would return 20% of the reads from an Illumina sequencing run (Fig. S3, Supporting information). We sequenced the pooled libraries using an ILLUMINA NEXTSEQ HIGH OUTPUT v2 150 cycle kit to achieve PE75 reads (Fig. S1, Supporting information).

*Treatments 4 and 5: Optimized RADcap vs. size selection.* After sequencing, the data from plates 1–4 included eight loci with an average coverage 20× higher than other loci in the one-primer treatment, 10× higher than other loci in the two-primer treatment and 28× higher than other loci in the low-template treatment. To block the overenrichment of these loci, we designed and ordered 29 custom oligonucleotides (Table S3, Supporting information) between 26 and 60 bp long that were complementary to the baits targeting these eight loci and which had a DNA to RNA $T_m$ >70 °C. We also optimized the PCR for plate 5, based on the one-primer treatment above, by increasing reaction volumes threefold for the PCR to add the iTru5-8N primer and 1.3-fold for the PCR to add the iTru7 primer, and we included the locus-specific bait blockers during the hybridization reaction. To compare the results of capturing RAD loci to those of size selection normally performed in 3RAD (and other RADseq protocols), we split the plate 5 pool in half following the one-primer PCR and SpeedBead clean-up, and we captured loci from one-half as described above. We called this treatment RADcap_optimized_Plate5. We size-selected the remaining half of the plate 5 pool as described above. We called this treatment 3RAD_SizeSelect_Plate5. We pooled these two libraries with unrelated libraries to obtain 7% of the reads on a second ILLUMINA NEXTSEQ run (Fig. S3, Supporting information) using conditions described above for the other RADcap libraries.

## Data analysis

*Modification of STACKS software.* STACKS (Catchen *et al.* 2011, 2013) had previously been modified to identify the variable-length internal tags that distinguish individual samples in 3RAD data. However, no software program existed to properly identify and remove the PCR duplicates from RADseq data. We developed new code as part of the *clone_filter* module within STACKS v1.35 to remove PCR duplicates. *Clone_filter* can be used before or after *process_radtags* and can use any combination of inline or index sequence tags, in addition to using read sequences, to reduce duplicated reads to a single representative in the output. Importantly, *clone_filter* does not modify FASTQ headers, allowing repeated use of *process_radtags* and *clone_filter* for read demultiplexing and duplicate removal.

*Data processing.* After sequencing, we converted BCL files to FASTQ format using BCL2FASTQ2 v2.16.0.10 (Illumina, Inc.), and we modified the default parameters to create a separate FASTQ file for index reads (Fig. S1 and Appendix S4: Example_scripts.md, Supporting information). We demultiplexed and removed PCR duplicates from the FASTQ data using STACKS v 1.35. First, we demultiplexed reads originating from different plates by iTru7 tag (Table S2, Supporting information) using *process_radtags*. We discarded reads with an uncalled base, reads having low quality (using default settings) or reads having a sequence tag or RAD tag more than two bases distant from the expected sequence. We rescued reads having sequence tags or RAD tags within two bases of the expected sequence. This initial demultiplexing produced paired-end files corresponding to each plate in each treatment. We ran *process_radtags* again on each plate of samples, with the same parameters, to truncate reads to 64 bp and demultiplex reads by inner adapter, which produced paired-end files for each individual in each plate. Finally, we used the *clone_filter* program to remove any read having the same combination of random iTru5 tag and RAD sequence, which probably represent duplicates created during PCR amplification.

We created a FASTA-formatted 'radnome' file that contained the 964 paired sequences from which we designed baits, and we used this file as a reference sequence for read alignment and SNP calling (Appendix S5: wisteria_reference.fasta, Supporting information). Within this FASTA file, paired reads were separate entries given arbitrary locus names, and we inserted 20 Ns between the sequences for Read 1 and

Read 2. We aligned RADcap reads to the reference using BWA v 0.7.7 (Li & Durbin 2009) with the mem algorithm and shorter split hits marked as secondary (M), and we called SNPs using an automated pipeline (https://github.com/faircloth-lab/radcap) that incorporates BWA, PICARD and the open-source GATK-LITE package (McKenna *et al.* 2010; DePristo *et al.* 2011). Following automated BWA alignment, the pipeline merged individual alignments, re-aligned BAM files around indels, called SNPs and indels, and filtered problematic or low-quality SNP calls from the total set of raw SNP calls to create a passing file of SNPs.

Variant calling is an inherently population-based process in that errors can be distinguished from variants at a specific position by considering that position in all individuals in the population (Craig *et al.* 2008; Bansal *et al.* 2010; Catchen *et al.* 2011). Therefore, the detection and statistical properties of variant genotypes are dependent on how the population under study is sampled, with fewer variant sites recovered with lower statistical support from smaller populations. To mimic this effect of population sampling and to facilitate comparisons among our experimental treatments, we called SNPs in two ways. First, we treated all 384 individuals from plates 1–4 as a single population, and we called SNPs separately in each of the one-primer ($n$ = 384 individuals) and two-primer experimental treatments ($n$ = 384 individuals). Second, we treated the 96 samples in plate 1 as a single population, and we called SNPs for the plate 1 population in the one-primer, two-primer and low-template treatments, as well as the plate 5 optimized and size-selected treatments (Table 1). After SNP calling, we filtered the resulting VCF files using VCFTOOLS v0.1.12b (Danecek *et al.* 2011) to exclude sites with more than 50%, 20% or 10% missing data (i.e. 50%, 80% or 90% complete data), and we computed summary statistics across captured loci and variant sites using a program from the RADCAP software package (Appendix S4: Example_scripts.md, Supporting information).

*RADcap assessment.* Because PCR can be biased by the composition of certain primers, we wanted to estimate how well our iTru5-8N primers were incorporated into our library constructs. Using the FASTQ file of index reads as input, we determined the count of each iTru5-8N sequence tag using FASTX v0.0.14 (Gordon & Hannon 2010; Appendix S4: Example_scripts.md, Supporting information). We plotted the cumulative count of iTru5-8N sequence tags incorporated to DNA libraries for all possible sequence tag combinations, except for iTru5-8N tags that do not return a signal on the NEXTSEQ (GGGGGGGG), those DNA inserts that have no apparent i5 sequence tag (AGATCTCG) and those iTru5 sequence tags of other libraries on the sequencing run.

We expected sequence capture to be more efficient than size selection and that the resulting data from captured RAD loci would include fewer off-target reads, have higher coverage at target loci, and consistently recover a larger number of target loci from reads. To investigate these parameters, we computed the coverage of each position in each sample from BAM files using SAMTOOLS v1.2 (Li *et al.* 2009; Appendix S4: Example_scripts.md, Supporting information). For this analysis, we used the BAM files produced directly from BWA to avoid effects of the BAM re-alignment on our coverage computations and because we wanted to assess which loci were present in the data set (where coverage of loci in the radnome reference was >0), despite being monomorphic or having errors. We report the average coverage for bases 32 and 190 of each reference locus, representing the middle base of Read 1 and Read 2, in all samples within each plate, normalized by million reads per sample. To determine whether the variation in coverage between loci in a treatment decreased in the optimized treatment, we plotted the log-transformed coverage of each locus and tested whether the optimized treatment had less variation in log-transformed coverage using a one-sided Siegel-Tukey test for equality in variability with adjusted medians in DESCTOOLS (Signorell 2015) in R. We then calculated the average coverage per locus per million reads per sample for loci with at least 4× coverage in plates RADcap_1cycle_Plate1, RADcap_5cycle_Plate1, RADcap_Low_Template_Plate1, RADcap_optimized_Plate5, and 3RAD_SizeSelect_Plate5. As a measure of consistency and to see whether the same loci were recovered in each treatment, we identified the loci with at least 4× coverage in 90% of samples from each treatment and determined the loci in common between treatments using VENNDIAGRAM (Chen & Boutros 2011) in R (Appendix S6: Venn_diagram_code.R, Supporting information). In addition, we plotted the density kernel of the coverage for Read 1 and Read 2 for each of the five treatments and compared the distributions of coverage between treatments in a one-sided two-sample Kolmogorov–Smirnov test in R. We compared coverage of loci targeted by 4 baits to coverage of loci targeted by 2 baits at base 32 and 190 by performing a $t$-test.

To determine how many reads were necessary to recover all of the targeted loci at reasonable coverage, we plotted the number of loci at or above 4× coverage and the number of reads for each sample. To estimate coverage at lower read numbers, we took the median coverage for all samples at each locus and divided that to get corresponding coverage between 1000 and 250 000 reads per sample. We plotted the number of loci at or above 4×, 10× and 20× coverage as a function of the reads per sample. We calculated the frequency of missing data between replicate samples within the one-primer and

two-primer treatments by converting VCF output files to GENEPOP format in PGDSPIDER v. 2.0.9.1 (Lischer & Excoffier 2012) and counting the number of SNPs at which one sample had a base called while another did not. Because there were no replicate samples within plate 5, we could not assess the amount of missing data. Instead, we compared estimates of genetic diversity of *Wisteria* in plates RADcap_optimized_Plate5 and 3RAD_SizeSelect_Plate5 from 80% filled matrices in GENALEX v6.502 (Peakall & Smouse 2006, 2012). For each plate, we report the average number of samples genotyped (of 96) across all loci, the number of alleles identified, the effective number of alleles, Shannon's Information Index, the observed and expected heterozygosity and the fixation index, along with standard error estimates for each parameter.

## Results

### *Initial 3RAD SNP discovery for bait design*

Following SNP discovery using four *Wisteria* samples, we obtained 1.4–2.5 million PE150 reads per sample, and we retained an average of 83.7% of reads after quality filtering. We identified 31 686 STACKS catalog loci, 1350 of which were sequenced in all four samples, and 3483 of which were sequenced in three samples. Of the loci recovered in at least three samples, 2573 loci were polymorphic and contained a total of 6531 variant sites. After filtering these loci, there were 1428 putative variants in the 964 loci we used to design our capture baits.

### *Random tagging at the i5 index position allows removal of PCR duplicates*

For the RADcap samples in plates 1–5, we obtained 3–40 million reads per plate (average 17 million), and we retained >94% of reads after quality filtering (Table 2). This left an average of 85 000 reads per sample for plates other than 3RAD_SizeSelect_Plate5, RADcap_5cycle_Plate3 and RADcap_Low_Template_Plate1 (which had 200 000, 11 000 and 16 000 reads per sample, respectively; Fig. S3 and Table S4, Supporting information). We incorporated and sequenced all 65 536 of the expected iTru5 random sequence tags in both of the ILLUMINA NEXTSEQ runs performed to generate our data (Fig. S2, Supporting information). All plates from which we collected data using RADcap had a similar per cent of reads retained after quality filtering by *process_radtags* in STACKS (Table 2). We retained an average of 68.9% of reads after decloning (range 20.4–95.7%; Tables 2 and S4, Supporting information), with the most reads retained from the optimized PCR protocol (which we performed on RADcap_optimized_Plate5 and 3RAD_SizeSelect_Plate5) and RADcap_5cycle_Plate3.

### *Optimizing RADcap efficiency and coverage*

All but one of the capture treatments yielded ≥80% of reads on target (Table 2), while the optimized treatment (RADcap_optimized_Plate5) yielded the highest proportion of reads on target (90%). More traditional 3RAD with size selection (3RAD_SizeSelect_Plate5) yielded 15% of reads on target. Similarly, the optimized and two-primer treatments had the highest average coverage, at 928 and 942 reads per locus per million reads per sample, respectively (Table 2), but we note that the two-primer coverage is inflated with undetected duplicate sequences from multiple rounds of PCR. The one-primer and low-template treatments had slightly lower average coverages, at 783 and 629 reads per locus per million reads per sample, respectively. The size-selected treatment had the lowest average coverage at 155 reads per locus per million reads per sample.

The coverage per locus per million reads was much higher among the RADcap samples than traditional 3RAD size-selected samples (Fig. S4, Supporting information). The variation in coverage per locus per million reads sequenced per sample was lower for RADcap_optimized_Plate5 than 3RAD_SizeSelect_Plate5 and RADcap_1cycle_Plate1 (Siegel–Tukey test, d.f. = 963, $P < 0.0083$ in both cases; Fig. S4, Supporting information), despite loci targeted by four baits having a significantly higher coverage than loci targeted by 2 baits in RADcap_optimized_Plate5 (1086 and 886, respectively; *t*-test with equal variances, $t = 4.61$, d.f. = 1926, $P = 2.1 \times 10^{-6}$). The variation in coverage for the RADcap_optimized_Plate5 did not differ from the RADcap_5cycle_Plate1 or the RADcap_Low_Template_Plate1 (Siegel–Tukey test, d.f. = 963, $P > 0.37$ in both cases) even though loci targeted by four baits had higher coverage than loci targeted by two baits in RADcap_5cycle_Plate1 (1105 and 900, *t*-test with unequal variances, $t = 3.50$, d.f. = 913, $P = 2.4 \times 10^{-4}$). Average coverages between loci with two and four baits did not differ in RADcap_Low_Template_Plate1, RADcap_1cycle_Plate1, and 3RAD_SizeSelect_Plate5. The increased performance of RADcap is also apparent when the coverage per locus is plotted as a density distribution (Fig. S5, Supporting information).

### *RADcap effectively and consistently enriched target loci and produces dense SNP matrices*

We consistently recovered more targeted loci within RADcap treatments than traditional 3RAD with size selection after analysis with GATK (Fig. 2a; Table S5, Supporting information). Specifically, the optimized treatment performed the best, with 912 loci recovered at 50% matrix occupancy, 880 recovered at 80% occupancy

**Table 2** The total reads per plate, per cent retained after quality filtering, per cent retained in paired and 'remainder' files after decloning, per cent mapped and the average coverage per million reads sequenced per sample of base 32 and 190 of each locus

| Plate | Number of reads | % Retained after quality filtering | % Retained after decloning | % Reads that map to reference | Average coverage |
|---|---|---|---|---|---|
| RADcap_1cycle_Plate1 | 19 397 440 | 94.9 | 25.1 | 79.6 | 783 |
| RADcap_1cycle_Plate2 | 14 703 752 | 95.0 | 20.4 | 85.7 | – |
| RADcap_1cycle_Plate3 | 15 865 294 | 94.8 | 67.2 | 81.5 | – |
| RADcap_1cycle_Plate4 | 17 907 048 | 95.0 | 63.3 | 83.8 | – |
| RADcap_5cycle_Plate1 | 18 045 032 | 95.0 | 83.3 | 84.8 | 942 |
| RADcap_5cycle_Plate2 | 17 968 264 | 95.0 | 86.9 | 85.5 | – |
| RADcap_5cycle_Plate3 | 2 332 154 | 94.1 | 95.7 | 84.3 | – |
| RADcap_5cycle_Plate4 | 18 455 900 | 95.1 | 86.1 | 84.1 | – |
| RADcap_Low_Template_Plate1 | 3 285 096 | 95.3 | 41.0 | 65.8 | 629 |
| RADcap_optimized_Plate5 | 17 929 096 | 95.5 | 94.2 | 90.1 | 928 |
| 3RAD_SizeSelect_Plate5 | 39 543 602 | 95.9 | 94.8 | 15.1 | 155 |

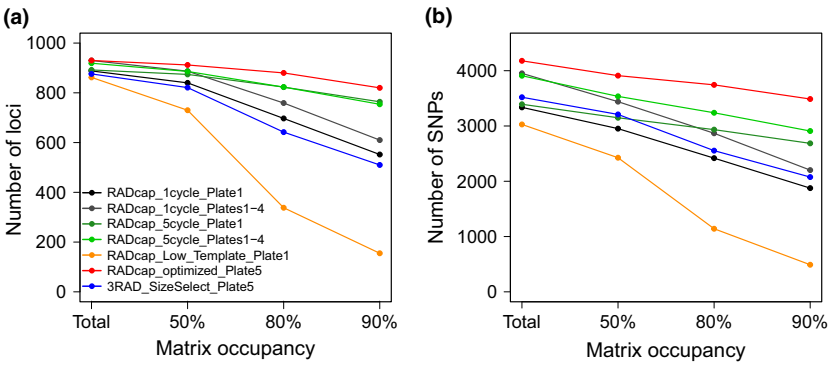


**Fig. 2** The number of loci (a) and SNPs (b) retained at various levels of matrix occupancy for different treatments, analysed with GATK.

and 820 recovered at 90% occupancy. The 96 samples analysed from the one-primer and two-primer treatments performed slightly poorer, with 840 and 874 loci recovered at 50% matrix occupancy, 697 and 823 loci recovered at 80% matrix occupancy and 552 and 764 loci recovered at 90% matrix occupancy, respectively. Traditional 3RAD with size selection returned 821, 642 and 510 loci at the same levels of matrix occupancy. As expected, RADcap_Low_Template_Plate1 showed the poorest performance, returning 730, 338 and 155 loci at the same levels of occupancy. The number of SNPs called within loci showed similar patterns (Fig. 2b), with RADcap_optimized_Plate5 performing better than all other treatments. The effects of population size on the number of SNPs called are apparent in the differences we observed between RADcap_5cycle_Plate1 and RADcap_5cycle_Plate1–4 and between RADcap_1cycle_Plate1 and RADcap_1cycle_Plate1–4.

Another important metric for most researchers is the consistency with which reduced representation approaches collect data across plates or from all individuals in a population. In a 90% filled matrix of loci with 4× or higher coverage, more than half of the loci (516 of 964; 54%) were shared between all treatments except low template, an additional 286 loci (30%) were shared among all three RADcap treatments, and an additional 125 loci (13%) were shared among the RADcap_5cycle_Plate1, RADcap_optimized_Plate5 and 3RAD_SizeSelect_Plate5 (Fig. S6, Supporting information). Impressively, RADcap_optimized_Plate5 contained data at 4× coverage for 962 of the 964 loci (99.8%; Fig. S6, Supporting information). Solely 36 loci (3.7%) were present in only one or two treatments (Fig. S6, Supporting information). Thus, most loci were present in most samples regardless of which treatment they originated.

In both RADcap_optimized_Plate5 and RADcap_5cycle_Plate1, we recovered most of the 964 loci in most samples regardless of the sequencing depth (Fig. 3). By comparison, in the size-selected treatment, even samples with the largest number of reads did not include as many loci as these RADcap treatments. When we modelled a reduced number of reads over all samples for each locus in the optimized treatment, we found that 20 000–30 000 reads were sufficient to capture all loci with at least 4× coverage, and 60 000 reads per sample were sufficient to achieve 10× coverage at all loci (Fig. 3). To achieve 20× or higher coverage at all loci, we estimated that ≥200 000 reads per sample were required.

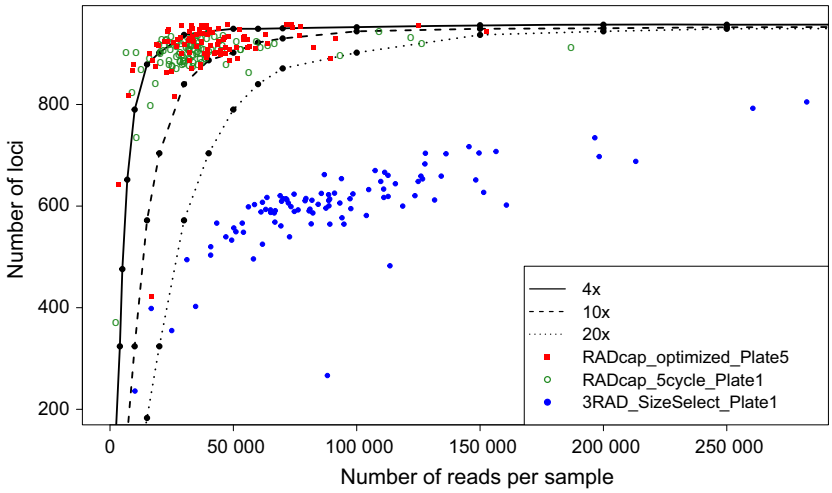


**Fig. 3** The number of loci recovered at different depths of sequencing coverage for RADcap and 3RAD library preparations. The points represent the number of loci sequenced to ≥4× coverage in each sample relative to the number of reads sequenced for that sample. The lines represent the number of loci that should be recovered with the optimized treatment at various read depths for a minimum coverage of 4×, 10× and 20×.

### *Error rate and genetic diversity of Wisteria*

The amount of missing data in samples replicated within a treatment was effectively equal between one-primer and two-primer treatments (7.20% and 7.61%, respectively). Because the optimized and size-selected treatments had the same samples and were filtered to have the same occupancy, we show the effect of estimating diversity with a smaller data set. In the 80% occupancy matrices, we recovered 3744 SNPs in the optimized treatment and 2554 SNPs in the size-selected treatment. On average, two more samples were genotyped in the optimized treatment than size-selected treatment for each SNP (Table S6, Supporting information). The number of alleles, number of effective alleles per SNP and Shannon's Information Index were higher for size-selected samples than optimized samples (2.018, 1.159 and 0.184, respectively, for size-selected and 2.014, 1.143 and 0.169, respectively, for optimized samples). $F_{IS}$ was higher for size-selected samples than optimized samples (0.227 and 0.180, respectively). Although observed heterozygosity was the same, expected heterozygosity was higher for size-selected samples (0.104 vs. 0.095 for optimized samples; Table S6, Supporting information).

## Discussion

RADcap represents a significant improvement to current reduced representation library approaches by efficiently sampling a consistent portion of the genome from large numbers of individuals at low cost (Table 3). Our optimized protocol achieves ≥4× coverage for ≥90% of samples at 99.8% of targeted loci with <187 000 reads per sample. However, only 20 000 reads per sample would be necessary for this matrix occupancy in the optimized treatment. Although a cut-off of 4× coverage is commonly used (Catchen *et al.* 2013; Pegadaraju *et al.* 2013; Graham *et al.* 2015; Mastretta-Yanes *et al.* 2015; Ali *et al.* 2016), it is well known that 4× coverage will often lead to inaccurate genotypes and that deeper sequencing is needed for consistent and accurate genotyping (DePristo *et al.* 2011; Sims *et al.* 2014). Fortunately, RADcap is sufficiently efficient that 10–20× coverage can be obtained for 90% complete matrices with affordable amounts of sequencing. This lower sequencing depth provides similar measures of genetic diversity as RADseq, but with higher confidence, because we sample loci more consistently and deeply.

### *Initial 3RAD SNP discovery and bait design*

STACKS performed well for the task of identifying SNPs and polymorphic loci from pilot 3RAD samples—using four individuals for initial SNP discovery yielded 2573 polymorphic loci, higher than our goal of 2000 loci from which to design baits. However, using a small sample size limited our ability to identify polymorphic loci that result from biological variation relative to those that arose from sequencing errors, and we probably failed to detect rare alleles due to ascertainment bias (Nielsen 2000; Clark *et al.* 2005). For future RADcap projects, we recommend using 16–96 individuals for SNP discovery. Another constraint of our current approach is that the pilot-scale 3RAD experiment requires a significant amount of time to complete, including several weeks to synthesize baits. If a genome sequence is available for the focal organism, the genome could be digested *in silico* and loci with mapped SNPs could be used for bait design.

### *Random tagging at the i5 index position allows removal of PCR duplicates*

We used the iTru5-8N tag to successfully remove PCR duplicates from our data using new additions to the

**Table 3** Major processes and reagents of RADcap, costs for this study and potential improvements to reduce costs and/or increase throughput

| | Reagent cost ($US) | Alternatives and potential improvements |
|---|---|---|
| Process* | | |
| Isolating DNA | 2.00 | Speed-bead based DNA isolation |
| Normalizing DNA | 0.20 | Robots, Sequal-Prep or similar |
| Digesting DNA & Ligating Adapters | 1.00 | Reduce amount of enzymes used |
| Single cycle degenerate iTru5 | 0.20 | |
| Size selection (SNP discovery) | 1.00† | SpeedBeads or gel-cut |
| NEXTSEQ PE150 (SNP discovery) | 21.20† | HiSeq PE100–PE150 |
| Sequence capture | 1.20 | – |
| NEXTSEQ PE75 | 1.20 | |
| **RADcap genotyping total per sample** | **5.80** | |
| Major reagents‡ | | |
| MYcroarray MYbaits | 2400.00§ | Single bait per locus, increase number of projects pooled for baits |
| 3RAD adapters | 370.00† | Purchase aliquots or share |
| iTru primers | 345.00† | Purchase aliquots or share |

*Calculated on a per-sample basis and assumes full 96-well plates; ‡Batch cost per project; §Included or †excluded in the total genotyping cost per sample.

STACKS codebase. Although there are $4^8$ = 65 536 possible iTru5-8N sequence tags for each locus, false duplicates (i.e. independent DNA molecules with the same iTru5-8N sequence tag) will be encountered at much lower coverage (McKinney 1966; Schweyen *et al.* 2014). The number of iTru5-8N sequence tags that we used to identify duplicates is much larger than tag pools used in the past (e.g. Schweyen *et al.* 2014; Tin *et al.* 2015), allowing more than sufficient depth of coverage after duplicate removal. The approach we created also does not require researchers to anneal complementary oligos within a large pool of oligos containing degenerate tags (cf. Schweyen *et al.* 2014), which will produce a preponderance of double-stranded adapters with mismatches.

While our approach is simple and powerful, it is constrained by an upper limit of 65 536 possible iTru5-8N sequence tags. Thus, the method we implemented in STACKS uses the iTru5-8N sequence plus the corresponding read sequence to define duplicates—otherwise ≤65 536 reads would be retained from any library. In addition, it is critical to use conditions that promote high efficiency of first-strand synthesis (i.e. our optimized treatment, and not the one-primer, two-primer or low-template treatments) to avoid high levels of PCR duplication. Finally, STACKS is the only software we have extensively tested to remove duplicates from these types of data, although IPYRAD recently added similar support functions (ipyrad.readthedocs.io).

Casbon *et al.* (2011) found that reduced numbers of template molecules going into PCR increased the rate of PCR duplication. In contrast, our treatment with the least amount of input DNA, RADcap_Low_Template_Plate1, had fewer duplicates than RADcap_1cycle_Plate1. Thus, the specific conditions used for first-strand extension are critical to producing a diverse RAD library and can be even more important than the amount of template used. The low-level of duplicates in the size-selected and optimized protocols (which have only one cycle of first-strand extension) demonstrates that high levels of duplicates are not inevitable and that careful optimization of reaction conditions can keep duplicates to quite low proportions. However, the only way to know what percentage of the reads are duplicates is to implement a strategy to detect duplicates, which also facilitates their removal. Thus, tagging and removing duplicates is prudent for all RADcap and RADseq experiments.

### *Optimizing RADcap efficiency*

Our modifications of the one-primer treatment to the optimized (also single-primer) treatment included increasing the PCR volume and adding locus-specific bait blockers for the eight loci that were overabundant in the first set of RADcap reads. These modifications decreased the variation in coverage among loci and increased the number of loci we recovered. The bait blockers had a modest effect (the reads attributed to the blocked loci decreased from 14.7% in the one-primer treatment and 8.1% in the two-primer treatment to 3.6% in the optimized treatment). Thus, we surmise that increasing the PCR volume used for first-strand synthesis was far more important. Because we tested the one-primer and two-primer treatments on plates 1–4 and the increased PCR volume using plate 5, it is unclear whether the increased volume of the PCR decreased the rate of PCR duplication or whether the initial quality of

the DNA or ligations for plate 5 was higher than the quality of plates 1–4.

### *RADcap captures nearly all loci targeted and produces dense SNP matrices*

We sequenced most of the targeted loci in RADcap_optimized_Plate5 and only slightly fewer loci in RADcap_5-cycle_Plate1. The 99.8% overlap between RADcap _optimized_Plate5 and RADcap_5cycle_Plate1 illustrates the strength of using sequence capture to collect RAD loci: we were able to recover most of the same loci across at least 90% of 192 samples that we prepared several weeks apart.

GATK worked well on these data, recovering and retaining large numbers of loci and SNPs at 50%, 80% and 90% matrix occupancy. The number of loci and SNPs called followed predictable patterns within and among data sets from all treatments. GATK is in common use across a variety of genotyping studies and performs well for moderate- to large-scale data sets (Cornish & Guda 2015); however, GATK is unsuited to most ddRAD data sets because of read duplication; thus, our system for removing duplicates was critical for meeting the assumptions of GATK. We could have used STACKS or other SNP-calling software packages (e.g. FREEBAYES, IPYRAD, SAMTOOLS), and subsequent work will provide a detailed comparison among SNP-calling software packages on RADcap data.

### *RADcap adds relatively few errors to Illumina sequences*

Errors in RADseq data may derive from library preparation because high-fidelity DNA polymerases introduce $2.8 \times 10^{-7}$ errors per nucleotide incorporated (KAPA, Boston, MA, USA). If these errors occurred in a single cycle of PCR, it would result in 4683 errors in the 223 million PE64 reads in the present data set, which could be amplified to high coverage. A much larger problem is the 0.1% substitution error rate made by Illumina machines (Glenn 2011), which results in an additional 28 544 000 incorrect bases. Decloning facilitated by the random iTru5-8N primer tags does not prevent PCR or sequencing errors, but the use of probabilistic base-calling algorithms can help to reduce the likelihood of a base introduced by these errors from being called as a true variant.

### *RADcap works with mixtures of baits from different organisms diluted to 1× concentration*

The MYbaits-1 synthesis allows ~20 000 baits and the smallest synthesis scale is sufficient for 12 captures. If fewer than 20 000 baits are designed, the concentration of baits is increased proportionately. We synthesized 2328 baits from *Wisteria* and 2624 baits from an other organism (kudzu); therefore, we were able to reduce bait costs by 50% for each project. Because we synthesized approximately one-quarter of the maximum allowed number of baits, we could do 12 × 4 captures with the smallest MYbaits-1 synthesis. This demonstrates the flexibility of RADcap; researchers can order baits for a variety of taxa and capture different numbers of samples from each taxa (see Heyduk *et al.* (2016) for examples). Baits for both species were present in all captures, despite DNA from only one species being present in any given capture. The large proportion of loci captured suggests that there was no meaningful interference from the additional baits during sequence capture and that the concentration of baits we applied to each sample pool was sufficient.

### *Comparison of RADcap to Rapture*

Other groups with different goals have also combined RADseq with sequence capture (Jones & Good 2015), such as Suchan *et al.*'s (2016) use of RADseq fragments as baits. While completing this manuscript, a separate group published a method similar to RADcap (Rapture; Ali *et al.* 2016). Rapture is an enrichment-based, RAD sequencing approach that uses a two-step protocol to capture RADseq loci. RADcap and Rapture both require DNA isolation, restriction enzyme digests, ligation of adapters, pooling, clean-up, capture and sequencing. Both methods are significant advances that increase the density and consistency of genotype matrices while simultaneously reducing costs for large-scale projects.

There are significant differences in cost, flexibility, duplicate detection and sequence coverage between RADcap and Rapture as a result of RADcap's integration with 3RAD and the Adapterama system (Glenn *et al.* 2016). The 3RAD adapters require 8 phosphorylated oligos and 32 plain oligos to achieve 96 combinations (Graham *et al.* 2015), whereas Rapture requires 96 biotinylated oligos plus 96 phosphorylated oligos, making Rapture adapters significantly more expensive ($370 for RADcap vs. $3750 for Rapture). Adding or switching to different enzymes in Rapture requires additional sets of adapters at additional cost, whereas 3RAD facilitates the use of 72 different possible enzymes and combinations of enzymes (Graham *et al.* 2015) with fewer sets of interchangeable adapters. Rapture detects duplicates based on the starting position of Read 2, which may be anywhere along ~500 bases (following shearing and size selection), whereas RADcap uses an 8-bp tag. Because RADcap has 65 536 tags, whereas Rapture has ~500, RADcap data will have fewer falsely identified

duplicates than Rapture. Finally, dual-digest RADcap increases coverage at both ends of the library molecule, and we show that fewer reads per sample are required to achieve the same coverage with RADcap than with Rapture (20 000 vs. 50 000 for at least 4× coverage, respectively). On the other hand, Rapture's use of random shearing increases the length of the genomic region that is sequenced, which may be an advantage worth the trade-off in decreased coverage, depending upon the goals of the project.

### *Future improvements and extensions*

RADcap opens the door to a variety of additional research opportunities. One of the most important is the option of using the capture baits from RAD loci on randomly sheared genomic libraries (i.e. standard genomic libraries). Such work will facilitate direct comparisons between RAD loci and other loci commonly used for sequence capture (exons, UCEs, anchored loci, etc.). Although preparing randomly sheared genomic libraries for RADcap increases the cost per sample, it will allow the following: (i) assembling contigs at captured loci so that more sequence is available to facilitate a better understanding of the sequence context for the RADcap loci; (ii) investigating rates of divergence at restriction sites; (iii) collecting RAD loci from samples with deeper divergences than is feasible with restriction sites (i.e. for phylogenetics) and (iv) using PHYLUCE (Faircloth 2016) and other analytical tools that have been developed for sequence capture systems. Capture baits also facilitate using RADseq for degraded and contaminated samples (cf. Graham *et al.* 2015) and focusing on microsatellite loci present in RADseq libraries, either through the use of locus-specific baits that target the flanking regions or via generic baits to the repeats (cf. Glenn & Schable 2005). Additionally, given the high efficiency we observe with two baits per locus, future work should investigate whether a single bait per locus is sufficient.

### *Summary*

We present a novel protocol to cheaply sequence a specific set of hundreds to thousands of loci in hundreds to thousands of samples deeply enough to obtain high-quality genotypes. We demonstrate a generalizable method for identifying PCR duplicates in Illumina libraries. We show that it is possible to reduce PCR duplicates to 5% of the total library, to routinely achieve >80% on-target reads and to achieve dense matrices of genotypes from hundreds of individuals. We strongly recommend that researchers adopt methods that yield high coverage and dense matrices of high-confidence genotypes, and we hope that RADcap allows other scientists to obtain high-quality data and make more robust conclusions about their study systems.

## Acknowledgements

We thank Todd Pierson, Kerin Bentley and Natalia Bayona-Vásquez. This work was supported by grants DEB-1242260 and DEB-1146440 from the U.S. National Science Foundation and the U.S. National Science Foundation Partnership for International Research and Education (PIRE) program (OISE 0730218). This study was also supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology.

## Conflict of interest

The authors declared conflict of interest. AD is employed by MYcroarray, a for-profit business that sells customized MYbaits kits. TJK and TCG are partially supported through cost-recovery projects in the EHS DNA laboratory, including projects that use 3RAD, and both are likely to use RADcap in the future.

## References

Ali OA, O'Rourke SM, Amish SJ *et al.* (2016) RAD Capture (Rapture): flexible and efficient sequence-based genotyping. *Genetics*, **202**, 389–400.

Altshuler D, Pollara VJ, Cowles CR *et al.* (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Andrews KR, Hohenlohe PA, Miller MR *et al.* (2014) Trade-offs and utility of alternative RADseq methods: reply to Puritz et al. 2014. *Molecular Ecology*, **23**, 5943–5946.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**, 81–92.

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Bansal V, Harismendy O, Tewhey R *et al.* (2010) Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Research*, **20**, 537–545.

Bi K, Vanderpool D, Singhal S *et al.* (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, **13**, 1–14.

Boratyn GM, Camacho C, Cooper PS *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Research*, **41**, W29–W33.

Cao H, Wu J, Wang Y *et al.* (2013) An integrated tool to study MHC region: accurate SNV detection and HLA genes typing in human MHC region using targeted high-throughput sequencing. *PLoS One*, **8**, e69388.

Cariou M, Duret L, Charlat S (2013) Is RAD-seq suitable for phylogenetic inference? An *in silico* assessment and optimization. *Ecology and Evolution*, **3**, 846–852.

Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research*, **39**, 1–8.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci *de novo* from short-read sequences. *G3-Genes Genomes Genetics*, **1**, 171–182.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Chen H, Boutros PC (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*, **12**, 35.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, **15**, 1496–1502.

Cornish A, Guda C (2015) A comparison of variant calling pipelines using genome in a bottle as a reference. *BioMed Research International*, **2015**, 456479.

Craig DW, Pearson JV, Szelinger S *et al.* (2008) Identification of genetic variants using barcoded multiplexed sequencing. *Nature Methods*, **5**, 887–893.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Davey JL, Blaxter MW (2010) RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, **9**, 416–423.

Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Faircloth BC (2016) PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, **32**, 786–788.

Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Glenn TC, Schable NA (2005) Isolating microsatellite DNA loci. *Methods in Enzymology*, **395**, 202–222.

Glenn TC, Nilsen R, Kieran TJ *et al.* (2016) Adapterama I: universal stubs and primers for thousands of dual-indexed Illumina libraries (iTru & iNext). *bioRxiv*, 049114. doi: 10.1101/049114.

Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.

Gordon A, Hannon G (2010) Fastx-toolkit. In: *FASTQ/A Short-Reads Preprocessing Tools*. Available from: http://hannonlab.cshl.edu/fastx_toolkit/index.html.

Graham CF, Glenn TC, McArthur AG *et al.* (2015) Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Molecular Ecology Resources*, **15**, 1304–1315.

Harvey MG, Smith BT, Glenn TC, Faircloth BC, Brumfield RT (2016) Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology*, syw036. doi: 10.1093/sysbio/syw036.

Heyduk K, Stephens JD, Faircloth BC, Glenn TC (2016) Targeted DNA region re-sequencing. In: *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing* (eds Aransay AM, Trueba JLL), pp. 43–68. Springer International Publishing, Switzerland.

Jones MR, Good JM (2015) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185–202.

Kivioja T, Vaharautio A, Karlsson K *et al.* (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, **9**, 72–74.

Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, **61**, 727–744.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2015) Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology Resources*, **15**, 28–41.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

McKinney EH (1966) Generalized birthday problem. *The American Mathematical Monthly*, **73**, 385–387.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Miner BE, Stöger RJ, Burden AF, Laird CD, Hansen RS (2004) Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Research*, **32**, e135.

Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**, 931–942.

Novaes E, Drost DR, Farmerie WG *et al.* (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics*, **9**, 1–14.

Okou DT, Steinberg KM, Middle C *et al.* (2007) Microarray-based genomic selection for high-throughput resequencing. *Nature Methods*, **4**, 907–909.

Pante E, Abdelkrim J, Viricel A *et al.* (2015) Use of RAD sequencing for delimiting species. *Heredity*, **114**, 450–459.

Peakall R, Smouse PE (2006) GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*, **28**, 2537–2539.

Pegadaraju V, Nipper R, Hulke B, Qi L, Schultz Q (2013) *De novo* sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. *BMC Genomics*, **14**, 556.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One*, **7**, e37135.

Rogers Y-H, Venter JC (2005) Genomics: massively parallel sequencing. *Nature*, **437**, 326–327.

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, **22**, 939–946.

Saintenac C, Jiang D, Akhunov ED (2011) Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biology*, **12**, R88.

Schweyen H, Rozenberg A, Leese F (2014) Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *Biological Bulletin*, **227**, 146–160.

Signorell A (2015) DescTools: Tools for descriptive statistics. R package version 0.99.15.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, **15**, 121–132.

Smith E, Jepsen K, Khosroheidari M *et al.* (2014) Biased estimates of clonal evolution and subclonal heterogeneity can arise from PCR duplicates in deep sequencing experiments. *Genome Biology*, **15**, 420.

Suchan T, Pitteloud C, Gerasimova N *et al.* (2016) Hybridization capture using RAD probes (hyRAD) a new tool for performing genomic analyses on collection specimens. *PLoS ONE*, **11**, e0151651.

Tautz D, Ellegren H, Weigel D (2010) Next generation molecular ecology. *Molecular Ecology*, **19**, 1–3.

Tin MMY, Rheindt FE, Cros E, Mikheyev AS (2015) Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Molecular Ecology Resources*, **15**, 329–336.

Trusty JL, Johnson KJ, Lockaby BG, Goertzen LR (2007) Bi-parental cytoplasmic DNA inheritance in *Wisteria* (Fabaceae): evidence from a natural experiment. *Plant and Cell Physiology*, **48**, 662–665.

Trusty JL, Lockaby BG, Zipperer WC, Goertzen LR (2008) Horticulture, hybrid cultivars and exotic plant invasion: a case study of *Wisteria* (Fabaceae). *Botanical Journal of the Linnean Society*, **158**, 593–601.

Valder P (1995) *Wisterias: A Comprehensive Guide*. Timber Press, Inc., Portland, Oregon.

Wiedmann RT, Smith TPL, Nonneman DJ (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genetics*, **9**, 1–7.

Wilson EH (1916) The *Wisterias* of China and Japan. *The Gardeners' Chronicle*, **1545**, 61–62.

Wyman D (1949) The Wisterias. *Arnoldia*, **9**, 17–28.

T.C.G. and B.C.F. conceived RADcap; S.L.H. and R.M. conceived *Wisteria* project; J.M.C. conceived and implemented decloning software; S.L.H. and T.J.K. prepared samples; A.D. designed baits; S.L.H. and B.C.F. analysed the data; S.L.H., B.C.F., T.C.G. and R.M. wrote the manuscript; T.C.G. and R.M. contributed funding and other resources; and all authors edited and approved of the final version.

## Data accessibility

Raw sequencing reads and final sets of SNPs: http://dx.doi.org/10.5061/dryad.ss6c9. SNP-calling pipeline: https://github.com/faircloth-lab/radcap.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Table S1** The layout of samples in each plate and inline barcodes used for each sample

**Table S2** The iTru7 primer outer tag sequence for each plate

**Table S3** Sequences of locus-specific bait blockers added to the HYB mix during sequence capture for the optimized treatment

**Table S4** The number of quality-filtered reads, per cent retained after removing duplicates, and per cent mapped to reference for each sample in each treatment

**Table S5** The total number of loci and SNPs obtained by GATK and the number of loci and SNPs in at least 50%, 80% and 90% of samples within each treatment

**Table S6** A comparison of population genetic summary statistics for *Wisteria* samples in plate 5 prepared via RADcap and 3RAD

**Fig. S1** Full-length declonable 3RAD library molecule and sequencing reads.

**Fig. S2** Cumulative frequency of 65 536 degenerate i5 tags in two different ILLUMINA NEXTSEQ HIGH OUTPUT v2 150 cycle runs.

**Fig. S3** The number of quality-filtered reads per sample for each plate before removing PCR duplicates.

**Fig. S4** The average log-transformed coverage per million reads per sample for each locus for 96 samples plotted on a log scale for the *y*-axis.

**Fig. S5** The density of the log coverage of loci in each treatment.

**Fig. S6** Venn diagram of the number of loci with at least 4× coverage shared across at least 90% (86) of samples in a single plate of the one-primer, two-primer, optimized and size-selected treatments.

**Appendix S1.** Methods detailing sample collection and library preparation.

**Appendix S2.** Bait sequences targeting Read 1 and Read 2 of 964 *Wisteria* loci.

**Appendix S3.** Bait sequences targeting Read 1 and Read 2 of a subset of 200 *Wisteria* loci, resulting in 4 baits per locus for these loci.

**Appendix S4.** Step by step analysis and example scripts to demultiplex and declone sequencing reads, call variants, and analyze data to produce Fig. S2, cumulative frequency of i5 tags, and Figs 3, S4, S5, S6, which include the coverage of loci.

**Appendix S5.** Radnome reference produced from pilot 3RAD loci, used to align reads in BWA.

**Appendix S6.** Code run to produce Fig. S6, the Venn diagram of loci shared between plates.